Statistica Sinica Preprint No: SS-2019-0202								
TitleJoint Bayesian Variable and DAG Selection Consistenc								
	for High-dimensional Regression Models with							
	Network-structured Covariates							
Manuscript ID	ID SS-2019-0202							
URL	http://www.stat.sinica.edu.tw/statistica/							
DOI	10.5705/ss.202019.0202							
Complete List of Authors	Xuan Cao and							
	Kyoungjae Lee							
Corresponding Author	Kyoungjae Lee							
E-mail	leekjstat@gmail.com							

Statistica Sinica

JOINT BAYESIAN VARIABLE AND DAG SELECTION CONSISTENCY FOR HIGH-DIMENSIONAL REGRESSION MODELS WITH NETWORK-STRUCTURED COVARIATES

Xuan Cao and Kyoungjae Lee

University of Cincinnati and Inha University

Abstract: We consider the joint sparse estimation of the regression coefficients and the covariance matrix for covariates in a high-dimensional regression model. Here, the predictors are both relevant to a response variable of interest and functionally related to one another via a Gaussian directed acyclic graph (DAG) model. Gaussian DAG models introduce sparsity in the Cholesky factor of the inverse covariance matrix, and the sparsity pattern in turn corresponds to specific conditional independence assumptions on the underlying predictors. A variety of methods have been developed in recent years for Bayesian inferences that identify such network-structured predictors in a regression setting. However, crucial sparsity selection properties for these models have not been thoroughly investigated. Therefore, we consider a hierarchical model with spike and slab priors on the regression coefficients, and a flexible and general class of DAG– Wishart distributions with multiple shape parameters on the Cholesky factors of the inverse covariance matrix. Under mild regularity assumptions, we establish the joint selection consistency for both the variable and the underlying DAG of the covariates when the dimension of the predictors is allowed to grow much larger than the sample size. We demonstrate that our method outperforms existing methods in selecting network-structured predictors in several simulation settings.

Key words and phrases: DAG-Wishart prior, Posterior ratio consistency, Strong selection consistency.

1. Introduction

In practice, we often encounter data sets in which the number of variables is much larger than the number of samples. Here, a major problem is that of high-dimensional variable selection, where the challenge is to select a subset of predictor variables that significantly affect a given response. The literature on Bayesian variable selection in linear regression is vast and rich. For example, George and McCulloch (1993) propose the stochastic search variable selection method, which uses the Gaussian distribution with a zero mean and a small, but fixed variance as the spike prior, and another Gaussian distribution with a large variance as the slab prior. Ishwaran, Kogalur, and Rao (2005) also use Gaussian spike and slab priors, but use continuous bimodal priors for the variance of the regression coefficient to alleviate the difficulty of choosing specific prior parameters. Narisetty and He (2014) introduce shrinking and diffusing priors as spike and slab priors, and establish the model selection consistency of the approach in a highdimensional setting.

Another important problem is how to formulate models and develop inferential procedures to understand the complex relationships and multivariate dependencies in these high-dimensional data sets. A covariance matrix is one of the most fundamental objects that quantifies these relationships. A common and effective approach for covariance estimation in sample-starved settings is to induce sparsity in the covariance matrix, its inverse, or the Cholesky factor of the inverse. The sparsity patterns in these matrices can be encoded uniquely using appropriate graphs. Hence, the corresponding models are often referred to as covariance graph models (sparsity in Σ), concentration graph models (sparsity in $\Omega = \Sigma^{-1}$), or directed acyclic graph (DAG) models (sparsity in the Cholesky factor of Ω).

In this work, we focus on a high-dimensional regression setting in which the predictors are both relevant to a response variable of interest and functionally related to one another via a Gaussian DAG model. Our goal is to jointly perform the variable and DAG selection, and then to establish the selection consistency in a high-dimensional regime. The advantage of joint modeling is that we can borrow information from the DAG structure to improve the performance of the variable selection. A popular motivation for this type of problem comes from genomic studies: the mechanism for an effect on an outcome such as quantitative molecular phenotypes, including gene expression, proteomics, or metabolomics data, often displays a coordinated change along a pathway, and the impact of a single genotype may not be apparent. In this setting, our proposed method can incorporate and highlight unknown pathways or regulatory networks that affect the response, which may improve the performance of the variable selection by borrowing information from the network structure. To uncover these relationships, we develop a Bayesian hierarchical model that favors the inclusion of variables that are not only relevant to the outcome of interest, but are also linked through a DAG.

Several approaches have been proposed for a known underlying graph structure, including both frequentist and Bayesian methods, when solving the variable selection problem. Li and Li (2008, 2010) study a graphconstrained regularization procedure and its theoretical properties to take into account the neighborhood information of the variables measured on a known graph. Pan et al. (2010) propose a grouped penalty based on the L_{γ} -norm that smooths the regression coefficients of the predictors over the available network. From a Bayesian perspective, Li and Zhang (2010) and Stingo and Vannucci (2010) incorporate a graph structure in the Markov random field (MRF) prior on indicators of variable selection, encouraging the joint selection of predictors with known relationships. Stingo et al. (2011) and Peng et al. (2013) propose selecting both the pathways and the genes within them using prior knowledge on gene–gene interactions or functional relationships.

However, when the underlying graph is unknown and needs to be selected, comparatively fewer methods have been proposed. Dobra (2009) estimates a network among relevant predictors by first performing a stochastic search in the regression setting to identify possible subsets of predictors. Then, a Bayesian model averaging method is applied to estimate a dependency network. Liu et al. (2014) develop a Bayesian method for a regularized regression that provides an inference on the inter-relationship between the variables by explicitly modeling a graph Laplacian matrix. Peterson et al. (2016) simultaneously infer a sparse network among the predictors and perform variable selection. They use this network as guidance by incorporating it into a prior that favors the selection of connected variables based on a Gaussian graphical model among the predictors. This, in turn, provides a sparse and interpretable representation of the conditional dependencies found in the data. In a slightly different context, Chekouo et al. (2015) and Chekouo et al. (2016) relate two sets of covariates via a DAG to integrate multiple genomic platforms and select the most relevant features. Given the ordering of the variables, they use a mixture of a non-local prior (Johnson and Rossell, 2012) and a point mass at zero to infer the DAG structure.

To the best of our knowledge, despite the developments in Bayesian methods for joint variable and graph selection, no rigorous investigations of the high-dimensional consistency properties of these methods have been undertaken. Hence, our goal is to investigate whether joint selection consistency results can be established in a high-dimensional regression setting with network-structured predictors. This is a challenging goal, particularly because of the interaction between the regression coefficients and the graph in the posterior analysis, as well as the massive parameter space to be explored for both the coefficients and the graph.

We consider a hierarchical multivariate regression model with DAG– Wishart priors on the covariance matrix for the predictors, spike and slab priors on regression coefficients, independent Bernoulli priors for each edge in the DAG, and an MRF prior linking the variable indicators to the graph structure. Under high-dimensional settings, we establish the posterior ratio consistency, following Cao et al. (2019c) and Narisetty and He (2014), for both the variable and the DAG, with a given DAG and variable, respectively (Theorems 1 and 2). In Theorems 3 and 4, we establish the posterior ratio consistency and the strong selection consistency for any DAG and variable pair. In particular, the strong selection consistency implies that under the true model, the posterior probability of the true variable indicator and the true graph converge in probability to one as $n \to \infty$. Finally, using simulation studies, we demonstrate that the proposed models outperform existing state-of-the-art methods, including both the penalized likelihood and the Bayesian approaches, in several settings.

The rest of paper is organized as follows. Section 2 provides background material on the Gaussian DAG model and the DAG–Wishart distribution. In Section 3, we introduce our hierarchical Bayesian model. The model selection consistency results are presented in Section 4, with proofs provided in the Supplementary Material. In Section 5, we conduct simulation experiments to illustrate the performance of the proposed method. The benefits of our Bayesian method for identifying network-structured predictors are demonstrated vis-a-vis existing Bayesian and penalized likelihood approaches. Section 6 concludes the paper.

2. Preliminaries

In this section, we provide the necessary background material on graph theory, Gaussian DAG models, and DAG–Wishart distributions.

2.1 Gaussian DAG models

Throughout this paper, a DAG $\mathscr{D} = (V, E)$ consists of a vertex set $V = \{1, \ldots, p\}$ and an edge set E, such that there is no directed path starting and ending at the same vertex. As in Ben-David et al. (2016) and Cao et al. (2019c), we assume a parent ordering in which all edges are directed from larger vertices to smaller vertices. Thus, the ordering of variables is assumed to be known. The set of parents of i, denoted by $pa_i(\mathscr{D})$, is the collection of all vertices larger than i that share an edge with i. A Gaussian DAG model over a given DAG \mathscr{D} , denoted by $\mathscr{N}_{\mathscr{D}}$, consists of all multivariate Gaussian distributions that obey the directed Markov property with respect to the DAG \mathscr{D} . In particular, if $x = (x_1, \ldots, x_p)^T \sim N_p(0, \Sigma)$ and $N_p(0, \Sigma) \in \mathscr{N}_{\mathscr{D}}$, then $x_i \perp x_{\{i+1,\ldots,p\}\setminus pa_i(\mathscr{D})} | x_{pa_i}(\mathscr{D})$ for each i.

Any positive-definite matrix Ω can be uniquely decomposed as $\Omega = LD^{-1}L^T$, where L is a lower triangular matrix with unit diagonal entries, and D is a diagonal matrix with positive diagonal entries. This decomposition is known as the modified Cholesky decomposition of Ω (e.g., see Pourahmadi (2007)). It is well known that if $\Omega = LD^{-1}L^T$ is the modified Cholesky decomposition of Ω , then $N_p(0, \Omega^{-1}) \in \mathscr{N}_{\mathscr{D}}$ if and only if $L_{ij} = 0$ whenever $i \notin pa_j(\mathscr{D})$. In other words, the structure of the DAG \mathscr{D} is reflected in the Cholesky factor L of the inverse covariance matrix.

Given a DAG \mathscr{D} on p vertices, denote $\mathscr{L}_{\mathscr{D}}$ as the set of lower triangular matrices with unit diagonals, and $L_{ij} = 0$ if $i \notin pa_j(\mathscr{D})$. Furthermore, let \mathscr{D}^p_+ be the set of strictly positive diagonal matrices in $\mathbb{R}^{p \times p}$. We refer to $\Theta_{\mathscr{D}} = \mathscr{D}^p_+ \times \mathscr{L}_{\mathscr{D}}$ as the Cholesky space corresponding to \mathscr{D} , and to $(D, L) \in \Theta_{\mathscr{D}}$ as the Cholesky parameter corresponding to \mathscr{D} . In fact, the relationship between the DAG and the Cholesky parameter implies that $\mathscr{N}_{\mathscr{D}} = \{N_p(0, (L^T)^{-1}DL^{-1}) : (D, L) \in \Theta_{\mathscr{D}}\}.$

The skeleton of \mathscr{D} , denoted by $\mathscr{D}^u = (V, E^u)$, can be obtained by replacing all directed edges of \mathscr{D} with undirected edges. We define the adjacency matrix of \mathscr{D} as a (0,1)-matrix, such that the elements of the matrix indicate whether or not pairs of vertices are adjacent in \mathscr{D} ; adjacent vertices are denoted by one, and all others by zero.

2.2 DAG–Wishart distribution

In this section, we revisit the multiple shape parameter DAG–Wishart distributions introduced in Ben-David et al. (2016). Given a directed graph $\mathscr{D} = (V, E)$, with $V = \{1, \ldots, p\}$, and a $p \times p$ matrix A, denote the column vectors $A_{\mathscr{D}.i}^{\geq} = (A_{ij})_{j \in pa_i(\mathscr{D})}^T$ and $A_{\mathscr{D}.i}^{\geq} = (A_{ii}, (A_{\mathscr{D}.i}^{\geq})^T)^T$. In addition,

$$A_{\mathscr{D}}^{\geq i} = \begin{bmatrix} A_{ii} & (A_{\mathscr{D},i}^{\geq})^T \\ A_{\mathscr{D},i}^{\geq} & A_{\mathscr{D}}^{\geq i} \end{bmatrix},$$

where $A_{\mathscr{D}}^{>i} = (A_{kj})_{k,j \in pa_i(\mathscr{D})}$. In particular, we have $A_{\mathscr{D},p}^{\geq} = A_{\mathscr{D}}^{\geq p} = A_{pp}$. Let $\nu_i(\mathscr{D}) = |pa_i(\mathscr{D})| = |\{j : j > i, (j,i) \in E(\mathscr{D})\}|.$

The DAG-Wishart distributions in Ben-David et al. (2016) corresponding to a DAG \mathscr{D} are defined on the Cholesky space $\Theta_{\mathscr{D}}$. Given a $p \times p$ positive-definite matrix U and a p-dimensional vector $\boldsymbol{\alpha}(\mathscr{D}) = (\alpha_1(\mathscr{D}), \ldots, \alpha_p(\mathscr{D}))$, with $\min_{1 \leq i \leq p} \{\alpha_i(\mathscr{D}) - \nu_i(\mathscr{D})\} > 2$, the probability density of the DAG– Wishart distribution is given by

$$\pi_{U,\alpha(\mathscr{D})}^{\Theta_{\mathscr{D}}}(D,L) = \frac{1}{z_{\mathscr{D}}(U,\boldsymbol{\alpha}(\mathscr{D}))} \exp\{-\frac{1}{2} \operatorname{tr}((LD^{-1}L^{T})U)\} \prod_{i=1}^{p} D_{ii}^{-\frac{\alpha_{i}(\mathscr{D})}{2}} \mathbb{I}\Big((D,L) \in \Theta_{\mathscr{D}}\Big),$$
(2.1)

where

$$z_{\mathscr{D}}(U, \boldsymbol{\alpha}(\mathscr{D})) = \prod_{i=1}^{p} \frac{\Gamma(\frac{\alpha_{i}(\mathscr{D})}{2} - \frac{\nu_{i}(\mathscr{D})}{2} - 1)2^{\frac{\alpha_{i}(\mathscr{D})}{2} - 1}(\sqrt{\pi})^{\nu_{i}(\mathscr{D})}det(U_{\mathscr{D}}^{>i})^{\frac{\alpha_{i}(\mathscr{D})}{2} - \frac{\nu_{i}(\mathscr{D})}{2} - \frac{3}{2}}}{det(U_{\mathscr{D}}^{\geq i})^{\frac{\alpha_{i}(\mathscr{D})}{2} - \frac{\nu_{i}(\mathscr{D})}{2} - 1}},$$

and $\mathbb{I}(\cdot)$ denotes the indicator function. The above density has the same form as the classical Wishart density, but is defined on the lower dimensional space $\Theta_{\mathscr{D}}$ and has p shape parameters $\{\alpha_i(\mathscr{D})\}_{i=1}^p$, which can be used for the differential shrinkage of variables in high-dimensional settings. The class of densities $\pi_{U,\alpha(\mathscr{D})}^{\Theta_{\mathscr{D}}}$ forms a conjugate family of priors for the Gaussian DAG model $\mathscr{N}(\mathscr{D})$. In particular, if the prior on $(D, L) \in \Theta_{\mathscr{D}}$ is $\pi_{U,\alpha(\mathscr{D})}^{\Theta_{\mathscr{D}}}$ and $X_1, \ldots, X_n \mid D, L, \mathscr{D} \stackrel{i.i.d.}{\sim} N_p(0, (L^T)^{-1}DL^{-1})$, then the resulting posterior distribution of (D, L) is $\pi_{\tilde{U},\tilde{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}$, where $S = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$, $\tilde{U} =$ U + nS, and $\tilde{\alpha}(\mathscr{D}) = (n + \alpha_1(\mathscr{D}), \ldots, n + \alpha_p(\mathscr{D})).$

3. Model Specification

In this section, we specify our hierarchical model for joint variable and DAG selection in regression models with network-structured predictors. We start by considering the standard Gaussian linear regression model with p coefficients and introducing some required notation. Similarly to Peterson et al. (2016) and Li and Li (2008), we consider both the response $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{n \times 1}$ and the predictors $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$ to be random variables. In particular, $Y \sim N_n (X\beta, \sigma^2 I_n)$, and the predictors are assumed to obey a multivariate Gaussian distribution; that is, $X_i \stackrel{i.i.d.}{\sim} N_p \left(0, (LD^{-1}L^T)^{-1}\right)$, for $i = 1, 2, \ldots, n$, where $\beta \in \mathbb{R}^{p \times 1}$ is a vector of regression coefficients, and (L, D) represents the Cholesky parameter corresponding to a DAG \mathcal{D} . Let the symmetric matrix $G = (G_{ij})_{1 \leq i,j \leq p}$ represent the adjacency matrix corresponding to DAG \mathcal{D} , where $G_{ij} = G_{ji} = 1$ if and only if there is an edge between vertex i and vertex j, and $G_{ij} = G_{ji} = 0$

otherwise. Our goal is both (i) variable selection (i.e., correctly identifying all nonzero regression coefficients) and (ii) network estimation (i.e., precisely recovering the sparsity pattern in \mathscr{D}).

For the variable selection, we denote a variable indicator $\gamma = \{\gamma_1, \ldots, \gamma_p\}$, where $\gamma_j \in \{0, 1\}$ for $1 \leq j \leq p$, and $|\gamma| = \sum_{j=1}^p \gamma_j$. Let $\beta_\gamma = (\beta_j)_{\{j:\gamma_j=1\}}^T \in \mathbb{R}^{|\gamma| \times 1}$ be the vector formed by the active components in β corresponding to a model γ . For any $n \times p$ matrix A, let A_k represent the submatrix formed from the columns of A corresponding to model k. In particular, Let X_γ denote the design matrix formed from the columns of X corresponding to model γ . For the network estimation, the class of DAG–Wishart distributions in Section 2.2 can be used for joint variable and DAG selection with the following hierarchical model:

$$Y|X_{\gamma},\beta_{\gamma} \sim N_n \left(X_{\gamma}\beta_{\gamma},\sigma^2 I_n\right), \qquad (3.1)$$

$$X_i|(L,D), \mathscr{D} \stackrel{i.i.d.}{\sim} N_p\left(0, (LD^{-1}L^T)^{-1}\right), \text{ for } i = 1, 2, \dots, n,$$
 (3.2)

$$(L,D)|\mathscr{D} \sim \pi_{U,\alpha(\mathscr{D})}^{\Theta_{\mathscr{D}}}(D,L), \qquad (3.3)$$

$$\beta_{\gamma}|\gamma \sim N_{|\gamma|} \left(0, \tau^2 \sigma^2 I_{\gamma}\right), \qquad (3.4)$$

$$\pi(\mathscr{D}) \propto \prod_{j=1}^{p-1} q^{\nu_j(\mathscr{D})} (1-q)^{p-j-\nu_j(\mathscr{D})} \mathbb{I}\left\{ \max_{1 \le j \le p-1} \nu_i(\mathscr{D}) < R \right\}, \quad (3.5)$$

$$\pi(\gamma|\mathscr{D}) \propto \exp\left(-a1^T\gamma + b\gamma^T G\gamma\right) \mathbb{I}\left\{|\gamma| < R\right\},\tag{3.6}$$

for some constants $\sigma, \tau, a > 0, b \ge 0, 0 < q < 1$, and a positive integer

 $0 \leq R \leq p$. Here, we assume that σ in (3.1) is a known constant, for simplicity. However, it can be extended to the unknown case by imposing an inverse-gamma prior; see Corollary 1. Note that in (3.4), we are essentially imposing a spike and slab prior on the regression coefficients, where τ^2 indicates the variance of the slab part; see Narisetty and He (2014), Yang et al. (2016), and the references therein. Prior (3.5) corresponds to an Erdos–Renyi type of prior over the space of DAGs. In particular, similarly to Cao et al. (2019c), we define $e_{ji} = \mathbb{I}\{(j,i) \in E(\mathscr{D})\}$, for $1 \leq j < i \leq p$, to be the edge indicator. Let e_{ji} , where $1 \leq i < j < p$, be independent and identically distributed (i.i.d.) Bernoulli(q) random variables. Recall $\nu_j(\mathscr{D}) = |pa_j(\mathscr{D})|$ is the cardinality of the parent set of vertex j. It follows that $\pi(\mathscr{D}) = \prod_{(j,i):1 \le j < i \le p} q^{e_{ji}} (1-q)^{1-e_{ji}} = \prod_{j=1}^{p-1} q^{\nu_j(\mathscr{D})} (1-q)^{p-j-\nu_j(\mathscr{D})}$. In (3.5) and (3.6), the positive integer R is an upper bound on the DAG and regression complexity. Note that to obtain our desired asymptotic consistency results, we introduce appropriate conditions for the hyperparameters τ, R, a, b and the edge probability q in Section 4.

Remark 1. In (3.6), given a DAG \mathscr{D} , we are imposing an MRF prior on the variable indicator γ that favors the inclusion of variables linked to other variables in the associated DAG. MRF priors have also been used in a variable selection setting in Peterson et al. (2016), Li and Zhang (2010), and Stingo and Vannucci (2010). In particular, as indicated in Peterson et al. (2016), the parameter a in (3.6) controls the variable inclusion probability, with larger values of a corresponding to sparser models, while b essentially determines how strongly the inclusion probability of a variable is affected by the inclusion of its neighbors in the DAG.

The hierarchical model in (3.1)–(3.6) can be used to estimate a pair of a variable and a DAG, as follows. By (2.1) and Bayes' rule, the following lemma gives the (marginal) joint posterior probabilities; the proof is provided in the Supplementary Material.

Lemma 1. Under the hierarchical model in (3.1)-(3.6), the (marginal) joint variable and DAG posterior is given by

$$\pi \left(\gamma, \mathscr{D} | Y, X\right)$$

$$\propto \pi(\gamma | \mathscr{D}) \pi(\mathscr{D}) \frac{z_{\mathscr{D}}(U + X^T X, n + \alpha(\mathscr{D}))}{z_{\mathscr{D}}(U, \alpha(\mathscr{D}))}$$

$$\times \det \left(\tau^2 X_{\gamma}^T X_{\gamma} + I_{|\gamma|}\right)^{-\frac{1}{2}} \exp \left\{-\frac{1}{2\sigma^2} \left(Y^T \left(I_n + \tau^2 X_{\gamma} X_{\gamma}^T\right)^{-1} Y\right)\right\},$$
(3.7)

where $z_{\mathscr{D}}(\cdot, \cdot)$ is the normalized constant in the DAG–Wishart distribution.

Hence, after integrating out β_{γ} , we have the joint posterior available in closed form (up to the multiplicative constant $\pi(X, Y)$). In particular,

these posterior probabilities can be used to select a pair of a variable and a DAG by computing the posterior mode defined by

$$(\hat{\gamma}, \hat{\mathscr{D}}) = \operatorname*{argmax}_{(\gamma, \mathscr{D})} \pi \left(\gamma, \mathscr{D} | Y, X\right).$$
(3.8)

4. Joint Selection Consistency

In this section, we explore the high-dimensional asymptotic properties of the Bayesian joint variable and DAG selection approach specified in Section 3. For this purpose, we work in a setting where the number of regression coefficients $p = p_n$ increases with the sample size n. The true data-generating mechanism is given by

 $Y = X\beta_0^n + \epsilon_n,$

where $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p_n}$, $X_i \stackrel{i.i.d.}{\sim} N_{p_n}(0, \Sigma_0^n)$, and $\epsilon_n \sim N_n(0, \sigma_0^2 I_n)$. Here, β_0^n is the true p_n -dimensional vector of regression coefficients, and Σ_0^n is the true covariance matrix. As in the usual context of variable selection, we assume that the true vector of regression coefficients is sparse; that is, all entries of β_0^n are zero, except those corresponding to the active entries in the true variable indicator γ_0^n (Castillo et al., 2015; Yang et al., 2016; Narisetty and He, 2014). Denote $\rho_{1n} = \min_{j \in \gamma_0^n} |\beta_{0j}^n|$ and $\rho_{2n} = \max_{j \in \gamma_0^n} |\beta_{0j}^n|$ as the min-

imum and maximum magnitudes, respectively, of the nonzero entries in β_0^n . We assume that the true quantities $|\gamma_0^n|$, ρ_{1n} , and ρ_{2n} vary with n. Let $\Omega_0^n = (\Sigma_0^n)^{-1} = L_0^n (D_0^n)^{-1} (L_0^n)^T$, where (D_0^n, L_0^n) denotes the modified Cholesky parameter of Ω_0^n . Let \mathscr{D}_0^n be the true underlying DAG with structure corresponding to the sparsity pattern in L_0^n (i.e, $L_0^n \in \mathscr{L}_{\mathscr{D}_0^n}$), and let G_0^n be the adjacency matrix for \mathscr{D}_0^n . Denote d_n as the maximum number of nonzero entries in any column of L_0^n , and $s_n = \min_{1 \le j \le p_n, i \in pa_j(\mathscr{D}_0^n)} |(L_0^n)_{ij}|$ as the minimum magnitude of the nonzero off-diagonal entries in L_0^n . Let \overline{P} denote the probability measure corresponding to the true model presented above. In order to establish the desirable consistency results, we need the following mild assumptions. Each assumption is followed by an interpretation/discussion.

Assumption 1. There exists $0 < \epsilon_0 \leq 1$, such that $\epsilon_0 \leq eig_1(\Omega_0^n) \leq eig_{p_n}(\Omega_0^n) \leq \epsilon_0^{-1}$, for every $n \geq 1$, where $eig_1(\Omega_0^n)$ and $eig_{p_n}(\Omega_0^n)$ are the minimum and maximum eigenvalues, respectively, of Ω_0^n .

This is a standard assumption for high-dimensional covariance asymptotic consistency, in both the frequentist and the Bayesian paradigms; see, for example, Bickel and Levina (2008), El Karoui (2008), Banerjee and Ghosal (2014), Xiang et al. (2015), and Banerjee and Ghosal (2015). Cao et al. (2019c) relax this assumption by allowing the lower and upper bounds on the eigenvalues to depend on p_n and n.

Assumption 2. For the true DAG, $d_n \sqrt{\log p_n/n} \to 0$ and $d_n \log p_n/(s_n^2 n) \to 0$. 0. For the true regression coefficient, $|\gamma_0^n| \sqrt{\log p_n/n} \to 0$, $\log n \log p_n/(n\rho_{1n}^2) \to 0$, and $\rho_{2n}/\sqrt{\log p_n} \to 0$ as $n \to \infty$.

This assumption resembles the dimension assumption in Cao et al. (2019a), and is a much weaker assumption for a high-dimensional covariance asymptotic than those of, for example, Xiang et al. (2015), Banerjee and Ghosal (2014), Banerjee and Ghosal (2015), and Cao et al. (2019c). Here, we essentially allow the dimension of our covariance matrix to grow more slowly than $\exp(n/d_n^2)$. Recall that s_n is the smallest (in absolute value) nonzero off-diagonal entry in L_0^n , so the second condition in Assumption 2 can also be interpreted as the lower bound for the signal size. This assumption, also known as the "beta-min" condition, provides a lower bound for the signal size needed to establish consistency. This type of condition has been used for the exact support recovery of high-dimensional linear regression and Gaussian DAG models; see, for example, Yang et al. (2016), Khare et al. (2017), Lee et al. (2018), and Cao et al. (2019c). Assumption 2 also allows the complexity of γ_0^n and the nonzero entries of β_0^n to grow with n, while staying uniformly bounded by a function of n and p_n . In addition, the assumption on ρ_{1_n} can be viewed as the beta-min condition in the regression context.

Assumption 3. The hyperparameters in model (3.4) and the MRF prior (3.6) satisfy $\tau^2 \sim \sqrt{\log p_n}$, $a \sim \alpha_1 \log p_n$, and $bn^2/\{(\log n)^2 \log p_n\} \to 0$ as $n \to \infty$, where for any positive sequences a_n and b_n , $a_n \sim b_n$ implies that there exist positive constants c and C, such that $c \leq \min(a_n/b_n, b_n/a_n) \leq \max(a_n/b_n, b_n/a_n) \leq C$.

Recall that the parameter a in (3.6) controls the variable inclusion probability, and b reflects how strongly this probability is affected by the inclusion of its neighbors in the DAG. In Section 4.3, we investigate the behavior of the posterior probability evaluated at the true model under b > 0 and b = 0. In the Bayesian variable selection literature, similar priors corresponding to $a = C \log p_n$, for some constant C > 0, and b = 0 have been commonly used to obtain selection consistency (Narisetty and He, 2014; Castillo et al., 2015; Yang et al., 2016). The assumption that the variance of the slab prior, τ^2 , is required to approach infinity is also stated here to ensure the desired model selection consistency.

Assumption 4. Let $q_n = O(p_n^{-\alpha_1})$, for some constant $\alpha_1 > 0$, and R_n in model (3.5) and (3.6) satisfy $R_n \sim n/\log n$ and $bR_n^2/\log p_n \to 0$ as $n \to \infty$.

This assumption provides the rate at which the edge probability q_n needs

to approach zero. It also states that the prior on the space of the $2^{\binom{p_n}{2}}$ possible models places zero mass on unrealistically large models. Note that q_n approaches zero more slowly than in Cao et al. (2019c), which helps avoid potential computation limitations, such as the simulation results always favoring the most sparse model. This assumption also states that the MRF prior on the space of the 2^{p_n} possible models places zero mass on unrealistically large models (see similar assumptions in Shin et al. (2018) and Narisetty and He (2014) in the context of regression).

Assumption 5. For every $n \geq 1$, the hyperparameters for the DAG-Wishart prior $\pi_{U_n,\alpha(\mathscr{D}_n)}^{\Theta_{\mathscr{D}_n}}$ satisfy (i) $2 < \alpha_i(\mathscr{D}_n) - \nu_i(\mathscr{D}_n) < c$, for every \mathscr{D}_n and $1 \leq i \leq q_n$, and (ii) $0 < \delta_1 \leq eig_1(U_n) \leq eig_p(U_n) \leq \delta_2 < \infty$. Here, c, δ_1 , and δ_2 are constants that do not depend on n.

This assumption provides mild restrictions on the hyperparameters for the DAG–Wishart distribution. The assumption $2 < \alpha_i(\mathscr{D}) - \nu_i(\mathscr{D})$ establishes prior propriety. The assumption $\alpha_i(\mathscr{D}) - \nu_i(\mathscr{D}) < c$ implies that the shape parameter $\alpha_i(\mathscr{D})$ can only differ from $\nu_i(\mathscr{D})$ (number of parents of i in \mathscr{D}) by a constant, which does not vary with n. Additionally, the eigenvalues of the scale matrix U_n are assumed to be uniformly bounded in n.

For the rest of this paper, $p_n, \Omega_0^n, \Sigma_0^n, L_0^n, D_0^n, \mathscr{D}_0^n, \mathscr{D}_n^n, d_n, q_n, \beta_n, \gamma_n, \tau_n$, and A_n will be denoted as $p, \Omega_0, \Sigma_0, L_0, D_0, \mathscr{D}_0, \mathscr{D}, d, q, \beta, \gamma, \tau$, and A, respectively, for notational convenience and ease of exposition. We now state and prove the main joint variable and DAG selection consistency results.

4.1 Posterior ratio consistency of γ and \mathscr{D}

In this section, we show that our method guarantees the posterior ratio consistency of γ and \mathscr{D} . Although Peterson et al. (2016) consider a similar network-structured regression model, to the best of our knowledge, theoretical properties of Bayesian models such as posterior ratio consistency and joint selection consistency have not yet been established. We first establish the posterior ratio consistency with respect to \mathscr{D} under the true variable indicator γ_0 . Theorem 1 states that the true DAG is the mode of the posterior distribution with probability tending to one as $n \to \infty$ under fixed γ_0 .

Theorem 1. Under Assumptions 1 2, 4, and 5,

$$\max_{\mathscr{D}\neq\mathscr{D}_0}\frac{\pi(\gamma_0,\mathscr{D}|Y,X)}{\pi(\gamma_0,\mathscr{D}_0|Y,X)} \xrightarrow{\bar{P}} 0, \quad as \ n \to \infty.$$

Remark 2. Note that the posterior ratio consistency of a DAG is achieved under a given parent ordering in which all edges are directed from larger vertices to smaller vertices. For several applications in genetics and environmental sciences, a location- or time-based ordering of variables is naturally

4.1 Posterior ratio consistency of γ and \mathscr{D}

available. For temporal data, a natural ordering of variables is provided by the time at which they are observed. In quantitative molecular applications, the variables can be genes or SNPs located on a chromosome, and their spatial location provides a natural ordering; see Huang et al. (2006), Shojaie and Michailidis (2010), Yu and Bien (2016), Khare et al. (2017), and the references therein.

The next theorem establishes the posterior ratio consistency with respect to γ under DAG \mathscr{D} . This notion of consistency implies that the true variable indicator γ_0 is the mode of the posterior distribution with probability tending to one as $n \to \infty$ under fixed \mathscr{D} .

Theorem 2. Under Assumptions 1–5, the following holds:

$$\max_{\substack{(\gamma,\mathscr{D})\neq(\gamma_0,\mathscr{D}_0)}}\frac{\pi(\gamma,\mathscr{D}|Y,X)}{\pi(\gamma_0,\mathscr{D}|Y,X)} \xrightarrow{\bar{P}} 0, \quad as \ n \to \infty.$$

Remark 3. By carefully examining the proof of Theorem 2, we find that even under a DAG with a mis-specified ordering, the consistency result for γ under fixed \mathscr{D} still holds. We also investigate the performance of the proposed method under a mis-specified ordering in Section 5. The results suggest that our method recovers the true variable indicator γ_0 well, even in the mis-specified case.

4.2 Strong selection consistency of γ and \mathscr{D}

From Theorem 1, Theorem 2, and the fact that

$$\frac{\pi(\gamma, \mathscr{D}|Y, X)}{\pi(\gamma_0, \mathscr{D}_0|Y, X)} = \frac{\pi(\gamma_0, \mathscr{D}|Y, X)}{\pi(\gamma_0, \mathscr{D}_0|Y, X)} \times \frac{\pi(\gamma, \mathscr{D}|Y, X)}{\pi(\gamma_0, \mathscr{D}|Y, X)},$$

we can obtain the joint posterior ratio consistency with respect to both γ and \mathscr{D} . It implies that the true variable indicator and DAG, $(\gamma_0, \mathscr{D}_0)$, will be the mode of the posterior distribution with probability tending to one.

Theorem 3. Under Assumptions 1–5, the following holds:

$$\max_{(\gamma,\mathscr{D})\neq(\gamma_0,\mathscr{D}_0)}\frac{\pi(\gamma,\mathscr{D}|Y,X)}{\pi(\gamma_0,\mathscr{D}_0|Y,X)} \xrightarrow{\bar{P}} 0 \quad as \ n \to \infty,$$

which implies that

$$\bar{P}((\hat{\gamma}, \hat{\mathscr{D}}) = (\gamma_0, \mathscr{D}_0)) \to 1, \ as \ n \to \infty.$$

4.2 Strong selection consistency of γ and \mathscr{D}

In this section, we establish the joint strong selection consistency with respect to both γ and \mathscr{D} . Theorem 4 shows that the posterior probability assigned to the true variable indicator γ_0 and the true underlying graph \mathscr{D}_0 grows to one as $n \to \infty$. We call this property the joint strong selection consistency. Note that the result given in Theorem 3 does not guarantee this property. **Theorem 4.** Under Assumptions 1–5, if we further assume $\alpha_1 > 2$, then the following holds:

$$\pi(\gamma_0, \mathscr{D}_0 | Y, X) \xrightarrow{\bar{P}} 1 \quad as \ n \to \infty.$$

Note that the condition on α_1 , which controls the rate of the independent Bernoulli probability specified in Assumption 4, is only needed for strong selection consistency (Theorem 4). Similar restrictions on the hyperparameters have been considered in order to establish the consistency properties in the regression setup (Yang et al., 2016; Lee et al., 2018; Cao et al., 2019b). The model selection consistency for the posterior mode in Theorem 3 does not require a restriction on α_1 .

The aforementioned theorems are based on known σ^2 . However, in real applications, the underlying true variance is often unavailable. Therefore, we introduce the following corollary for a fully Bayesian hierarchical approach, where an appropriate inverse-gamma prior is imposed on σ^2 . It turns out that even with the unknown σ^2 , strong model selection consistency still holds under the same conditions given in Theorem 4.

Corollary 1. Suppose σ^2 is unknown and a proper inverse-gamma density with some positive constant parameters (a_0, b_0) is placed on σ^2 . Under Assumptions 1–5, and $\alpha_1 > 2$, the following holds:

$$\pi(\gamma_0, \mathscr{D}_0 | Y, X) \xrightarrow{P} 1 \quad as \ n \to \infty.$$

4.3 Behavior of the posterior probability when b = 0

In this section, we examine the behavior of the posterior probability for $(\gamma_0, \mathscr{D}_0)$ corresponding to two scenarios, that is, when the MRF prior parameter b > 0 and b = 0, respectively. The goal is to show that under a certain assumption on the connection between the sparsity patterns in γ_0 and \mathscr{D}_0 , by borrowing the graph information from the MRF prior, the posterior probability assigned to $(\gamma_0, \mathscr{D}_0)$ will increase. In particular, we introduce the following condition with respect to the true sparsity patterns encoded in both the variable indicator and the graph.

Condition 1. The true adjacency matrix G_0 and the true variable indicator γ_0 satisfy $\gamma_{0i} = \gamma_{0j} = 1$ whenever $(G_0)_{ij} = 1$, for $1 \le i, j \le p$.

Condition 1 essentially assumes that the variables connected through the underlying true DAG are active. Under this condition, compared with modeling the variable and the DAG separately (i.e. b = 0), incorporating network information into the variable selection through the MRF prior with b > 0 increases the posterior probability assigned to $(\gamma_0, \mathcal{D}_0)$, as illustrated in the following theorem. The proof for Theorem 5 is provided in the Supplementary Material.

Theorem 5. Let $\pi_1(\gamma_0, \mathscr{D}_0 | Y, X)$ be the posterior probability evaluated at $(\gamma_0, \mathscr{D}_0)$ under b > 0, and let $\pi_2(\gamma_0, \mathscr{D}_0 | Y, X)$ be the posterior probability evaluated at $(\gamma_0, \mathscr{D}_0)$ under b = 0. Then, the following holds:

$$\pi_1(\gamma_0, \mathscr{D}_0 \mid Y, X) > \pi_2(\gamma_0, \mathscr{D}_0 \mid Y, X).$$

Theorem 5 implies that, under Condition 1, our method achieves joint strong selection consistency without the condition on b stated in Assumption 3, which means the hyperparameter b in the MRF prior does not need to go to zero.

5. Numerical Studies

5.1 Posterior inference

For given positive real values a_0 and $b_0 > 0$, let $IG(a_0, b_0)$ be the inversegamma distribution with the shape parameter a_0 and scale parameter b_0 . Then, similarly to (3.7), the joint posterior distribution of γ and \mathscr{D} based

5.1 Posterior inference

on (3.1)–(3.6) and $\sigma^2 \sim IG(a_0, b_0)$ is

$$\pi(\gamma, \mathscr{D} \mid Y, X)$$

$$\propto \pi(\gamma \mid \mathscr{D}) \pi(\mathscr{D}) \frac{z_{\mathscr{D}}(U + X^T X, n + \alpha(\mathscr{D}))}{z_{\mathscr{D}}(U, \alpha(\mathscr{D}))}$$

$$\times \det \left(I_{\mid \gamma \mid} + \tau^2 X_{\gamma}^T X_{\gamma} \right)^{-\frac{1}{2}} \left\{ b_0 + \frac{1}{2} Y^T \left(I_n + \tau^2 X_{\gamma} X_{\gamma}^T \right) Y \right\}^{-\frac{n+2a_0}{2}}$$

We suggest using a Metropolis–Hastings within Gibbs sampling for the posterior inference:

- 1. Set the initial values $\gamma^{(1)}$ and $\mathscr{D}^{(1)}$.
- 2. For each s = 2, ..., S,
 - (a) sample $\gamma^{new} \sim q_{\gamma}(\cdot \mid \gamma^{(s-1)});$
 - (b) set $\gamma^{(s)} = \gamma^{new}$ with the probability

$$p_{acc,\gamma} = \min\left\{1, \frac{\pi(\gamma^{new} \mid \mathscr{D}^{(s-1)}, Y, X)}{\pi(\gamma^{(s-1)} \mid \mathscr{D}^{(s-1)}, Y, X)} \frac{q_{\gamma}(\gamma^{(s-1)} \mid \gamma^{new})}{q_{\gamma}(\gamma^{new} \mid \gamma^{(s-1)})}\right\},$$

otherwise set $\gamma^{(s)} = \gamma^{(s-1)};$

(c) sample $\mathscr{D}^{new} \sim q_{\mathscr{D}}(\cdot \mid \mathscr{D}^{(s-1)});$

(d) set $\mathscr{D}^{(s)} = \mathscr{D}^{new}$ with the probability

$$p_{acc,\mathscr{D}} = \min\left\{1, \frac{\pi(\mathscr{D}^{new} \mid \gamma^{(s)}, Y, X)}{\pi(\mathscr{D}^{(s-1)} \mid \gamma^{(s)}, Y, X)} \frac{q_{\mathscr{D}}(\mathscr{D}^{(s-1)} \mid \mathscr{D}^{new})}{q_{\mathscr{D}}(\mathscr{D}^{new} \mid \mathscr{D}^{(s-1)})}\right\},$$

otherwise set $\mathscr{D}^{(s)} = \mathscr{D}^{(s-1)}$.

Note that the inference for the DAG \mathscr{D} , that is, steps 2-(c) and 2-(d) in the above algorithm, can be parallelized for each column. For further detail, refer to Cao et al. (2019c) and Lee et al. (2018). We used the proposal kernel $q_{\gamma}(\cdot | \gamma')$, which gives a new set γ^{new} by changing a randomly chosen nonzero component in γ' to zero with probability 0.5, or by changing a randomly chosen zero component to one randomly with probability 0.5. The same kernels were used for each column of \mathscr{D} .

5.2 Simulation studies

In this section, we demonstrate the performance of the proposed method in various settings. We closely follow, but slightly modify the simulation settings in Peterson et al. (2016).

Suppose we have $X_i = (X_{i1}, \ldots, X_{ip})^T \overset{i.i.d.}{\sim} N_p(0, \Sigma_0)$, for $i = 1, \ldots, n$, where $\Sigma_0^{-1} = L_0(D_0)^{-1}L_0^T$, n = 100, and p = 240. If we consider p as the number of genes, we have 240 genes. We assume there are 40 transcription factors (TFs), and that each TF regulates five genes. Let TF_j be the index for the *j*th TF and $(TF_1, TF_2, \ldots, TF_{40}) = (6, 12, \ldots, 240)$. This corresponds to the DAG \mathscr{D}_0 , the support of L_0 , such that $pa_{TF_j-k}(\mathscr{D}_0) =$ $\{TF_j\}$, for $j = 1, \ldots, 40$ and $k = 1, \ldots, 5$. Suppose the TFs independently follow a normal distribution; that is, $X_{TF_j} \overset{ind}{\sim} N(0, d_{TF_j})$, where $d_{TF_j} \overset{i.i.d.}{\sim}$ Unif(3,5), for j = 1, ..., 40. We further assume that, given X_{TF_j} , the conditional distribution of the gene X_j regulated by $TF_{j'}$ is $N(X_{TF_{j'}}, d_j)$, where $d_j \stackrel{i.i.d.}{\sim} Unif(3,5)$, for j = 1, ..., 240. This corresponds to the true modified Cholesky parameter (L_0, D_0) , such that $(L_0)_{TF_j, TF_j-k} = 1$ and $D_0 = diag(d_j)$, for j = 1, ..., 40 and k = 1, ..., 5. We simulate the data from

$$Y = X\beta_0 + \epsilon,$$

where $X = (X_1, \ldots, X_n)^T$, $\epsilon \sim N_n(0, \sigma_\epsilon^2 I_n)$, and $\sigma_\epsilon^2 = \|\beta_0\|_2^2/4$. We investigate four settings for the true coefficient vector β_0 , as described in Li and Li (2008) and Peterson et al. (2016). In the first setting, it is assumed that $\beta_{0,TF_{1:4}} = (5, -5, 3, -3)^T$, $\beta_{0,TF_j-k} = \beta_{0,TF_j}/\sqrt{10}$ for j = 1, 2, 3, 4 and $k = 1, \ldots, 5$, and $\beta_{0,j} = 0$ for $j = 25, \ldots, 240$. This setting implies that the coefficients of genes in the same cluster have the same signs. In the second setting, the true coefficient β_0 is the same as the first setting, except that the signs are reversed for the two genes regulated by TF_j ; that is, $\beta_{0,TF_j-k} = -\beta_{0,TF_j}/\sqrt{10}$, for j = 1, 2, 3, 4 and k = 1, 2. This setting implies that the coefficients of genes in the same cluster might have different signs. The third and fourth settings the same as the first and second settings, respectively, expect that we consider 10 instead of $\sqrt{10}$. Thus, they consider smaller signals. We call this simulation setting Scenario 1.

We also investigate a simulation scenario in which the signals in β_0 are small. In this case, there are p = 150 genes, 30 TFs, and four regularized genes for each TF. The precision matrix $\Sigma_0^{-1} = L_0(D_0)^{-1}L_0^T$ is generated by $d_j \stackrel{i.i.d.}{\sim} Unif(2,5)$ and $(L_0)_{TF_j,TF_j-k} \stackrel{i.i.d.}{\sim} Unif(0.3, 0.7)$. The variance of ϵ is chosen as $\sigma_{\epsilon}^2 = \|\beta_0\|_2^2$. We consider four settings for the true coefficient vector β_0 . In the first and third settings, β_0 is generated by $\beta_{0,j} \stackrel{i.i.d.}{\sim} Unif(0.5, 1)$ and $\beta_{0,j} \stackrel{i.i.d.}{\sim} Unif(0.2, 1)$, respectively, for $j = 1, \ldots, 20$, and $\beta_{0,j} = 0$ for $j = 21, \ldots, 150$. In the second and fourth settings, we randomly change the signs of the nonzero entries of β_0 . We call this simulation setting Scenario 2.

Lastly, we consider a setting in which the network structure of the covariate X is an undirected graph. We generate the covariates $\tilde{X}_i \stackrel{i.i.d.}{\sim} N_p(0, \Sigma_0)$, for $i = 1, \ldots, n$, where $n = 100, p = 150, \Sigma_0 = \tilde{\Sigma}_0 + \{0.01 - eig_1(\tilde{\Sigma}_0)\}I_p$, and

$$(\tilde{\Sigma}_0)_{ij} = \begin{cases} 2 \max\left(1 - \frac{|i-j|}{10}, 0\right), & \text{if } |i-j| \le 5\\ 0, & \text{otherwise.} \end{cases}$$

Note that Σ_0 is positive definite. Furthermore, to consider the mis-specified ordering case, we randomly shuffle the columns of $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_n)^T$ to construct X. We simulate the data from $Y = X\beta_0 + \epsilon$, where $\epsilon \sim$ $N_n(0, \sigma_{\epsilon}^2 I_n)$ and $\sigma_{\epsilon}^2 = \|\beta_0\|_2^2/4$. Two settings for the true coefficient vector β_0 are considered. In the first setting, β_0 is generated by $\beta_{0,j} \stackrel{i.i.d.}{\sim} Unif(0.5, 1)$ for j = 1, ..., 10 and $\beta_{0,j} = 0$ for j = 11, ..., 150. In the second setting, we randomly change the signs of the nonzero entries of β_0 . We call this simulation setting Scenario 3; the simulation results for this setting are reported in Table 3.

We compare the performance of our joint selection method with that of existing variable selection methods: Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), and the Bayesian joint selection method proposed by Peterson et al. (2016). The tuning parameters for the Lasso and elastic net were chosen by 10-fold cross-validation. For the Bayesian methods, as discussed by Peterson et al. (2016), we suggest using the hyperparameters a = 2.75 and b = 0.5 for the MRF prior as the default. Furthermore, to show the benefits of joint modeling, we tried b = 0, which corresponds to the Bayesian method that models the variable and the DAG separately. The other hyperparameters were set at $a_0 = 0.1, b_0 = 0.01, \tau^2 = 1, q =$ $0.005, U = I_p$, and $\alpha_i(\mathcal{D}) = \nu_i(\mathcal{D}) + 10$, for all $i = 1, \ldots, p$. The initial state for γ was set as a p-dimensional zero vector; that is, the empty model, while the initial state for \mathscr{D} was chosen using the convex sparse Cholesky selection method (Khare et al., 2017). For the posterior inference, 5,000 posterior samples were drawn after a burn-in period of 5,000. Indices with a posterior inclusion probability larger than 0.5 were included in the final model. The resulting model is called the median probability model. Note that when the posterior probability is larger than 1/2, the model coincides with the posterior mode Barbieri and Berger (2004). Because we have proved the joint strong selection consistency (Theorem 4), the two models are asymptotically equivalent in our setting. Thus, although other approaches (e.g., see Scott and Carvalho (2008)) can be adapted to give a reasonable estimate of the posterior mode, we use the median probability model as a convenient, but asymptotically equivalent alternative.

To evaluate the performance of the variable selection, the sensitivity, specificity, area under the curve (AUC), Matthews correlation coefficient (MCC), number of errors (#Error), and mean-squared prediction error (MSPE) are reported in Tables 1, 2, and 3. The criteria are defined as follows:

Sensitivity =
$$\frac{TP}{TP + FN}$$
,
Specificity = $\frac{TN}{TN + FP}$,
MCC = $\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$,
#Error = $FP + FN$,
MSPE = $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{Y}_i - Y_{\text{test},i})^2$,

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively. The AUC is calculated based on the true positive rate (Sensitivity) and the false positive rate (1–Specificity) for Bayesian methods with varying thresholds. To draw the AUC, for each threshold, indices with a posterior inclusion probability larger than a given threshold were included in the final model. The AUCs for the regularization methods are omitted. We denote $\hat{Y}_i = X_i^T \hat{\beta}$, where $\hat{\beta}$ is the estimated coefficient based on each method. For the Bayesian methods, the usual least square estimates based on the selected support were used as $\hat{\beta}$. We generated test samples $Y_{\text{test},1}, \ldots, Y_{\text{test},n_{\text{test}}}$, with $n_{\text{test}} = 100$, to calculate the MSPE.

Tables 1 and 2 show that the Bayesian joint selection methods tend to have better specificity and MCC, while the regularization methods (Lasso and elastic net) have better sensitivity. As discussed by Peterson et al. (2016), this seems natural because the regularization methods based on cross-validation tend to include many redundant variables. This leads to a relatively larger number of errors for the regularization methods compared with those for the Bayesian joint selection methods. Furthermore, the proposed joint Bayesian selection method (Joint.CL (b = 1/2)) outperforms that proposed by Peterson et al. (2016) (Joint.P) in terms of all measures

Table 1: The summary statistics for Scenario 1 are represented for each setting. Each setting denotes a different choice of the true coefficient β_0 . Sens and Spec are sensitivity and specificity, respectively. Joint.CL: the Bayesian joint selection method proposed in this paper. Joint.P: the Bayesian joint selection method suggested by Peterson et al. (2016). Elastic: elastic net.

	Setting 1							Setting 2					
	Sens	Spec	AUC	MCC	#Error	MSPE	Sens	Spec	AUC	MCC	#Error	MSPE	
Joint.CL $(b = \frac{1}{2})$	0.8750	0.9861	0.9937	0.8611	6	69.1445	0.8750	0.9954	0.9894	0.9049	4	56.4885	
Joint.CL $(b = 0)$	0.7500	0.9815	0.9601	0.7605	10	96.9889	0.3333	1.0000	0.9058	0.5571	16	142.2708	
Joint.P	0.8750	0.9861	0.9838	0.8611	6	71.0443	0.7500	0.9954	0.9958	0.8282	7	73.7870	
Lasso	1.0000	0.8056		0.5412	42	45.5522	0.7083	0.8519		0.4170	39	106.0526	
Elastic	1.0000	0.9352	•	0.7685	14	41.8631	0.8750	0.8426		0.5122	37	92.6665	
			Set	ting 3			Setting 4						
	Sens	Spec	AUC	MCC	#Error	MSPE	Sens	Spec	AUC	MCC	#Error	MSPE	
Joint.CL $(b = \frac{1}{2})$	0.2083	0.9907	0.8493	0.3549	21	42.5213	0.3750	1.0000	0.7373	0.5922	15	30.3394	
Joint.CL $(b = 0)$	0.1667	0.9907	0.7117	0.3025	22	42.7116	0.3333	0.9954	0.7619	0.5191	17	35.0479	
Joint.P	0.2500	0.9907	0.8559	0.4023	20	40.3569	0.2917	0.9954	0.8954	0.4797	18	35.7181	
Lasso	1.0000	0.8241		0.5648	38	32.1919	0.6667	0.8102		0.3362	49	40.7437	
Elastic	1.0000	0.9444		0.7935	12	29.3908	0.6250	0.8935		0.4261	32	34.9673	

in Tables 1 and 2 except the AUC. In fact, the two Bayesian joint selection methods are quite similar, except for the graph structure they consider. In these simulation scenarios, the DAG structure seems more appropriate because clearly there are parents (TFs genes) and children (regularized genes for each TF). Thus, our method is preferable in this case. Lastly, the results show that our joint modeling (Joint.CL (b = 1/2)) significantly improves the performance of the variable selection compared with modeling the vari-

Table 2: The summary statistics for Scenario 2 are represented for each

setting. I	Each	setting	denotes a	different	choice of	of the	true	coefficient	β_0 .
------------	------	---------	-----------	-----------	-----------	--------	------	-------------	-------------

			Set	ting 1			Setting 2								
	Sens	Spec	AUC	MCC	#Error	MSPE	Sens	Spec	AUC	MCC	#Error	MSPE			
Joint.CL $(b = \frac{1}{2})$	0.7500	0.9923	0.9362	0.8174	6	20.9925	0.6000	1.0000	0.8933	0.7518	8	15.8789			
Joint.CL $(b = 0)$	0.6000	1.0000	0.9200	0.7518	8	29.6691	0.6500	0.9923	0.8790	0.7506	8	23.3007			
Joint.P	0.6500	1.0000	0.9842	0.7854	7	15.4705	0.5000	1.0000	0.9081	0.6814	10	19.2450			
Lasso	1.0000	0.8308		0.6290	22	14.8092	0.9000	0.7692		0.4877	32	13.4260			
Elastic	0.9500	0.9077		0.7201	13	18.9942	0.8000	0.8615		0.5371	22	14.5779			
	Setting 3									Setting 4					
			Set	ting 3					Set	ting 4					
	Sens	Spec	Set AUC	ting 3 MCC	#Error	MSPE	Sens	Spec	Set AUC	ting 4 MCC	#Error	MSPE			
Joint.CL $(b = \frac{1}{2})$	Sens 0.7500	Spec 1.0000	Set AUC 0.9537	ting 3 MCC 0.8498	#Error	MSPE 6.6246	Sens 0.6500	Spec 1.0000	Set AUC 0.8398	ting 4 MCC 0.7854	#Error 7	MSPE 7.4111			
Joint.CL $(b = \frac{1}{2})$ Joint.CL $(b = 0)$	Sens 0.7500 0.4000	Spec 1.0000 1.0000	Set AUC 0.9537 0.9631	ting 3 MCC 0.8498 0.6051	#Error 5 12	MSPE 6.6246 20.3681	Sens 0.6500 0.3000	Spec 1.0000 1.0000	Set AUC 0.8398 0.7962	ting 4 MCC 0.7854 0.5204	#Error 7 14	MSPE 7.4111 12.7521			
Joint.CL $(b = \frac{1}{2})$ Joint.CL $(b = 0)$ Joint.P	Sens 0.7500 0.4000 0.6500	Spec 1.0000 1.0000 1.0000	Set AUC 0.9537 0.9631 0.9811	ting 3 MCC 0.8498 0.6051 0.7854	#Error 5 12 7	MSPE 6.6246 20.3681 11.6528	Sens 0.6500 0.3000 0.4500	Spec 1.0000 1.0000 1.0000	Set AUC 0.8398 0.7962 0.9057	ting 4 MCC 0.7854 0.5204 0.6441	#Error 7 14 11	MSPE 7.4111 12.7521 9.2049			
Joint.CL $(b = \frac{1}{2})$ Joint.CL $(b = 0)$ Joint.P Lasso	Sens 0.7500 0.4000 0.6500 0.9500	Spec 1.0000 1.0000 1.0000 0.8154	Set AUC 0.9537 0.9631 0.9811	ting 3 MCC 0.8498 0.6051 0.7854 0.5754	#Error 5 12 7 25	MSPE 6.6246 20.3681 11.6528 8.2451	Sens 0.6500 0.3000 0.4500 0.8500	Spec 1.0000 1.0000 1.0000 0.7462	Set AUC 0.8398 0.7962 0.9057	ting 4 MCC 0.7854 0.5204 0.6441 0.4299	#Error 7 14 11 36	MSPE 7.4111 12.7521 9.2049 7.7223			

 Table 3:
 The summary statistics for Scenario 3 are represented for each

setting. Each setting denotes a different choice of the true coefficient β_0 .

	Setting 1					Setting 2						
	Sens	Spec	AUC	MCC	#Error	MSPE	Sens	Spec	AUC	MCC	#Error	MSPE
Joint.CL $(b = \frac{1}{2})$	1.0000	0.9357	0.9964	0.7018	9	1.6875	1.0000	0.9500	0.9821	0.7475	7	1.6469
Joint.CL $(b = 0)$	1.0000	0.9429	0.9786	0.7237	8	1.7331	1.0000	0.9429	0.9786	0.7237	8	1.7331
Joint.P	1.0000	0.9500	0.9857	0.7475	7	1.6838	1.0000	0.9500	0.9821	0.7475	7	1.6838
Lasso	1.0000	0.3571		0.1890	90	1.8121	1.0000	0.3571		0.1890	90	1.8121
Elastic	1.0000	0.6714		0.3463	46	1.6770	1.0000	0.6286		0.3184	52	1.6969

able and DAG separately (Joint.CL (b = 0)). These findings suggest that the proposed joint modeling approach actually improves variable selection performance by borrowing information from the DAG structure.

Table 3 shows the results for Scenario 3, where the true network struc-

ture for X is an undirected graph and the ordering is mis-specified. Even in this case, our joint modeling method provides comparable performance to that of Peterson et al. (2016), which is designed for undirected graphs. Similarly to Scenarios 1 and 2, the regularization methods do not work well compared with the Bayesian methods in our settings.

6. Conclusion

We examine a regression setting in which the predictors are both relevant to a response variable of interest and functionally related to one another via a Gaussian DAG model. In particular, we consider a hierarchical multivariate regression model with DAG–Wishart priors on the covariance matrix for the predictors, spike and slab priors on the regression coefficients, independent Bernoulli priors for each edge in the DAG, and an MRF prior linking the variable indicators to the graph structure. Under high-dimensional settings and standard regularity assumptions, for a known underlying variance σ^2 , we establish both the posterior ratio consistency and the strong selection consistency in order to jointly estimate the variable and the graph for the covariates. When the underlying response variance is unknown and an appropriate inverse-gamma prior is placed on σ^2 , we also establish the joint selection consistency under the same regularity conditions. Finally, we use

BAYESIAN JOINT SELECTION

simulation studies to demonstrate that the proposed model outperforms existing state-of-the-art methods in terms of selecting network-structured predictors, including both penalized likelihood and Bayesian approaches, in several settings. In future work, we intend to explore other types of priors over the graph space and on the regression coefficients to determine whether the consistency and better simulation performance can be achieved under weaker assumptions.

Supplementary Materials

Supplementary material includes the proofs for main results and other auxiliary results.

Acknowledgments

We thank Dr. Christine Peterson for sharing the code to implement the joint Bayesian variable and graph selection method in Peterson et al. (2016). We would like to thank two referees for their valuable comments. This research was supported by the Simons Foundation's collaboration grant (No.635213), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2019R1F1A1059483) and INHA UNIVER-SITY Research Grant.

References

- Banerjee, S. and S. Ghosal (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* 8, 2111–2137.
- Banerjee, S. and S. Ghosal (2015). Bayesian structure learning in graphical models. J. Multivariate Anal. 136, 147–162.
- Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. Ann. Statist. 32(3), 870–897.
- Ben-David, E., T. Li, H. Massam, and B. Rajaratnam (2016). High dimensional Bayesian inference for gaussian directed acyclic graph models. *Technical Report http://arxiv.org/abs/1109.4371*.
- Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. Ann. Statist. 36, 199–227.
- Cao, X., K. Khare, and M. Ghosh (2019a). Consistent Bayesian sparsity selection for high-dimensional Gaussian DAG models with multiplicative and beta-mixture priors. https://arxiv.org/abs/1903.03531.
- Cao, X., K. Khare, and M. Ghosh (2019b). High-dimensional posterior consistency for hierarchical non-local priors in regression. *Bayesian Anal.*, to appear.
- Cao, X., K. Khare, and M. Ghosh (2019c). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. Ann. Statist. 47(1), 319–348.

- Castillo, I., J. Schmidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. Ann. Statist. 43, 1986–2018.
- Chekouo, T., F. C. Stingo, J. D. Doecke, and K.-A. Do (2015). miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics* 71(2), 428–438.
- Chekouo, T., F. C. Stingo, M. Guindani, and K.-A. Do (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. Ann. Appl. Stat. 10(3), 1547–1571.
- Dobra, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics* 10(4), 621–639.
- El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. Ann. Statist. 36, 2757–2790.
- George, E. I. and R. E. McCulloch (1993). Variable selection via Gibbs sampling. J. Amer. Statist. Assoc 88(423), 881–889.
- Huang, J., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance selection and estimation via penalised normal likelihood. *Biometrika* 93, 85–98.
- Ishwaran, H., U. B. Kogalur, and J. S. Rao (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. Ann. Statist. 33(2), 730–773.
- Johnson, V. E. and D. Rossell (2012). Bayesian model selection in high-dimensional settings. Journal of the American Statistical Association 107(498), 649–660.

- Khare, K., S. Oh, S. Rahman, and B. Rajaratnam (2017). A convex framework for highdimensional sparse Cholesky based covariance estimation in Gaussian DAG models. *Preprint, Department of Statisics, University of Florida.*
- Lee, K., J. Lee, and L. Lin (2018). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *Ann. Statist., to appear.*
- Li, C. and H. Li (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24(9), 1175–1182.
- Li, C. and H. Li (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann. Appl. Stat. 4(3), 1498–1516.
- Li, F. and R. Zhang (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J. Amer. Statist. Assoc 105(491), 1202–1214.
- Liu, F., S. Chakraborty, F. Li, Y. Liu, and A. C. Lozano (2014). Bayesian regularization via graph Laplacian. *Bayesian Anal.* 9(2), 449–474.
- Narisetty, N. and X. He (2014). Bayesian variable selection with shrinking and diffusing priors. Ann. Statist. 42, 789–817.
- Pan, W., B. Xie, and X. Shen (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* 66(2), 474–484.
- Peng, B., D. Zhu, B. P. Ander, X. Zhang, F. Xue, F. R. Sharp, and X. Yang (2013). An

REFERENCES

integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways. *PLOS ONE* 8(7), 1–16.

- Peterson, C. B., F. C. Stingo, and M. Vannucci (2016). Joint Bayesian variable and graph selection for regression models with network-structured predictors. *Stat. Med.* 35(7), 1017– 1031.
- Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance–correlation parameters. *Biometrika* 94, 1006–1013.
- Scott, J. G. and C. M. Carvalho (2008). Feature-inclusion stochastic search for Gaussian graphical models. J. Comput. Graph. Statist. 17(4), 790–808.
- Shin, M., A. Bhattacharya, and V. Johnson (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* 28, 1053–1078.
- Shojaie, A. and G. Michailidis (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* 97, 519–538.
- Stingo, F. C., Y. A. Chen, M. G. Tadesse, and M. Vannucci (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. Ann. Appl. Stat. 5(3), 1978–2002.
- Stingo, F. C. and M. Vannucci (2010). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* 27(4), 495–501.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B

Stat. Methodol. 58(1), 267-288.

- Xiang, R., K. Khare, and M. Ghosh (2015). High dimensional posterior convergence rates for decomposable graphical models. *Electron. J. Stat. 9*, 2828–2854.
- Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high-dimensional Bayesian variable selection. Ann. Statist. 44(6), 2497–2532.

Yu, G. and J. Bien (2016). Learning local dependence in ordered data. arXiv:1604.07451.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67(2), 301–320.

University of Cincinnati

E-mail: (caox4@ucmail.uc.edu)

Inha University

E-mail: (leekjstat@gmail.com)