Joint Bayesian Variable and DAG Selection Consistency for High-dimensional Regression Models with Network-structured Covariates

Xuan Cao
Department of Mathematical Sciences, University of Cincinnati

Kyoungjae Lee
Department of Statistics, Inha university

May 29, 2019

Abstract

We consider the joint sparse estimation of regression coefficients and the covariance matrix for covariates in a high-dimensional regression model, where 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 inference in identifying such network-structured predictors in regression setting, yet crucial sparsity selection properties for these models have not been thoroughly investigated. In this paper, 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 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.

*Corresponding author.
1 Introduction

In modern day statistics, datasets where the number of variables is much larger than the number of samples are more pervasive than they have ever been. One of the major problems is high-dimensional variable selection, where the challenge is to select a subset of predictor variables which significantly affect a given response. The literature on Bayesian variable selection in linear regression is vast and rich. George and McCulloch [1993] propose the stochastic search variable selection 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 with 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 model selection consistency of the approach in a high-dimensional 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 datasets. 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 either in the covariance matrix, its inverse, or the Cholesky factor of the inverse. The sparsity patterns in these matrices can be uniquely encoded in terms of appropriate graphs. Hence the corresponding models are often referred to as covariance graph models (sparsity in $\Sigma$), concentration graph models (sparsity in $\Omega = \Sigma^{-1}$), and directed acyclic graph (DAG) models (sparsity in the Cholesky factor of $\Omega$).

In this work, we will work in a high-dimensional regression setting, where 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 variable and DAG selection, and to establish the selection consistency in a high-dimensional regime. One popular motivation for this type of problem comes from genomic studies: the mechanism for an effect on an outcome such as a quantitative molecular phenotypes including gene expression, proteomics, or metabolomics data often displays a coordinated change along a pathway, and the impact of one single genotype may not be apparent. In this setting, our proposed method can incorporate and highlight unknown pathways or regulatory networks that impact the response. 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 also linked through a DAG.

When the underlying graph structure is known, several approaches including both frequentist and Bayesian methods have been proposed and studied in the literature to solve the variable selection problem. Li and Li [2008a, 2010] study a graph-constrained 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 $L_\gamma$-norm that smooths the regression coefficients of the predictors over the available network. On the Bayesian side, 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 the selection of both pathways and genes within them based on 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] estimate a network among relevant predictors by first performing a stochastic search in the regression setting to identify possible subsets of predictors, then applying a Bayesian model averaging method to estimate a dependency network. Liu et al. [2014] develop a Bayesian method for regularized regression, which provides inference on the inter-relationship between variables by explicitly modeling through a graph Laplacian matrix. Peterson et al. [2016] simultaneously infer a sparse network among the predictors and perform variable selection using this network as guidance by incorporating it into a prior favoring selection of connected variables based on a Gaussian graphical model among the predictors, which provides a sparse and interpretable representation of the conditional dependencies found in the data.

Despite the developments in Bayesian methods for jointly performing variable and graph selection, a rigorous investigation of the high-dimensional consistency properties of these methods has not been undertaken to the best of our knowledge. Hence, our goal was to investigate if joint selection consistency results could be established in the 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, and the massive parameter space to be explored for both the coefficients and the graph.

In this paper, 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 a MRF prior linking the variable indicators to the graph structure. Under high-dimensional settings, we establish posterior ratio consistency, following the nomenclature in Cao et al. [2019a] and Narisetty and He [2014], for both the variable and the DAG with given DAG and variable, respectively (Theorem 4.1 and Theorem 4.2). In Theorem 4.3 and Theorem 4.4, we also establish the posterior ratio consistency and the strong selection consistency for any pair of the DAG and variable. In particular, the strong selection consistency implies that under the true model, the posterior probability of the true variable indicator and the true graph converges in probability to 1 as $n \to \infty$. Finally, through simulation studies, we demonstrate that the models studied in this paper can outperform existing state-of-
the-art methods including both penalized likelihood and Bayesian approaches in several settings.

The rest of paper is organized as follows. Section 2 provides background material regarding Gaussian DAG model and the DAG-Wishart distribution. In Section 3, we introduce our hierarchical Bayesian model. Model selection consistency results are stated in Section 4 with proofs provided in Section 6. In Section 5 we use simulation experiments to illustrate the proposed hierarchical approach, and demonstrate the benefits of our Bayesian method and computation procedures for identifying network-structured predictors vis-a-vis existing Bayesian and penalized likelihood approaches. We end our paper with a discussion session in Section 7.

2 Preliminaries

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

2.1 Gaussian DAG models

Throughout this paper, a directed acyclic graph (DAG) $\mathcal{D} = (V, E)$ consists of the 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. The set of parents of $i$, denoted by $\text{pa}_i(\mathcal{D})$, is the collection of all vertices which are larger than $i$ and share an edge with $i$. A Gaussian DAG model over a given DAG $\mathcal{D}$, denoted by $N_\mathcal{D}$, consists of all multivariate Gaussian distributions which obey the directed Markov property with respect to a DAG $\mathcal{D}$. In particular, if $x = (x_1, \ldots, x_p)^T \sim N_p(0, \Sigma)$ and $N_p(0, \Sigma) \in N_\mathcal{D}$, then $x_i \perp x_{i+1, \ldots, p} \mid \text{pa}_i(\mathcal{D})$ for each $i$.

Any positive definite matrix $\Omega$ 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 $\Omega$ (see for example Pourahmadi [2007]). It is well-known that if $\Omega = LD^{-1}L^T$ is the modified Cholesky decomposition of $\Omega$, then $N_p(0, \Omega^{-1}) \in N_\mathcal{D}$ if and only if $L_{ij} = 0$ whenever $i \notin \text{pa}_j(\mathcal{D})$. In other words, the structure of the DAG $\mathcal{D}$ is reflected in the Cholesky factor of the inverse covariance matrix.

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

The skeleton of $\mathcal{D}$, denoted by $\mathcal{D}^n = (V, E^*)$, can be obtained by replacing all the directed edges of
$\mathcal{D}$ by undirected ones. We define the adjacency matrix of $\mathcal{D}$ to be a $(0,1)$-matrix such that the elements of the matrix indicate whether pairs of vertices are adjacent or not in $\mathcal{D}$, i.e., 1 representing adjacent, 0 representing not adjacent.

2.2 DAG-Wishart Distribution

In this section, we revisit the multiple shape parameter DAG-Wishart distributions introduced in Ben-David et al. [2016]. First, we provide required notation. Given a directed graph $\mathcal{D} = (V,E)$, with $V = \{1,\ldots,p\}$, and a $p \times p$ matrix $A$, denote the column vectors $A^\triangleright_{i,\mathcal{D}} = (A_{ij})_{j \in \text{pa}_i(\mathcal{D})}$ and $A^\preceq_{i,\mathcal{D}} = (A_{ii}, (A^\triangleright_{i,\mathcal{D}})^T)$. Also, let $A^\triangleright_{i,\mathcal{D}} = (A_{kj})_{k,j \in \text{pa}_i(\mathcal{D})}$.

$$A^\triangleright_{i,\mathcal{D}} = \begin{bmatrix} A_{ii} & (A^\triangleright_{i,\mathcal{D}})^T \\ A^\triangleright_{i,\mathcal{D}} & A^\preceq_{i,\mathcal{D}} \end{bmatrix}.$$ 

In particular, $A^\triangleright_{\mathcal{D},p} = A^\triangleright_p = A_{pp}$.

The DAG-Wishart distributions in Ben-David et al. [2016] corresponding to a DAG $\mathcal{D}$ are defined on the Cholesky space $\Theta_{\mathcal{D}}$. Given a positive definite matrix $U$ and a $p$-dimensional vector $\alpha(\mathcal{D})$, the (unnormalized) density of the DAG-Wishart distribution on $\Theta_{\mathcal{D}}$ is given by $\exp\left\{-\frac{1}{2} \text{tr}\left((LD^{-1}LT)U\right)\right\} \prod_{i=1}^p D^2_{ii} \frac{\alpha_i(\mathcal{D})}{2}$, for every $(D,L) \in \Theta_{\mathcal{D}}$. Let $\nu_i(\mathcal{D}) = |\text{pa}_i(\mathcal{D})| = |\{j : j > i, (j,i) \in E(\mathcal{D})\}|$. If $\alpha_i(\mathcal{D}) - \nu_i(\mathcal{D}) > 2$, for all $1 \leq i \leq p$, the density in (1) can be normalized to a probability density, and the normalizing constant is given by $z_{\mathcal{D}}(U,\alpha(\mathcal{D})) = \prod_{i=1}^p \frac{\Gamma(\frac{\nu_i(\mathcal{D})}{2} + \frac{\alpha_i(\mathcal{D})}{2} - 1)}{\sqrt{(\nu_i(\mathcal{D}))!} \sqrt{\prod_{j=1}^p \frac{\alpha_j(\mathcal{D})}{2}}}$.

In this case, we define the DAG-Wishart density $\pi_{U,\alpha_i(\mathcal{D})}^\Theta_{\mathcal{D}}(D,L)$ by

$$\pi_{U,\alpha_i(\mathcal{D})}^\Theta_{\mathcal{D}}(D,L) = \frac{1}{z_{\mathcal{D}}(U,\alpha(\mathcal{D}))} \exp\left\{-\frac{1}{2} \text{tr}\left((LD^{-1}LT)U\right)\right\} \prod_{i=1}^p D^2_{ii} \frac{\alpha_i(\mathcal{D})}{2}$$

for every $(D,L) \in \Theta_{\mathcal{D}}$. The above density has the same form as the classical Wishart density, but is defined on the lower dimensional space $\Theta_{\mathcal{D}}$ and has $p$ shape parameters $\{\alpha_i(\mathcal{D})\}_{i=1}^p$ which can be used for differential shrinkage of the variables in high-dimensional settings.

The class of densities $\pi_{U,\alpha_i(\mathcal{D})}^\Theta_{\mathcal{D}}$ form a conjugate family of priors for the Gaussian DAG model $\mathcal{N}(\mathcal{D})$. In particular, if the prior on $(D,L) \in \Theta_{\mathcal{D}}$ is $\pi_{U,\alpha(\mathcal{D})}^\Theta_{\mathcal{D}}$ and $X_1,\ldots,X_n$ are independent, identically distributed $N_p(0,(LT)^{-1}DL^{-1})$ random vectors, then the posterior distribution of $(D,L)$ is $\pi_{U,\tilde{\alpha}(\mathcal{D})}^\Theta_{\mathcal{D}}$, where $S = \frac{1}{n} \sum_{i=1}^n X_iX_i^T$ denotes the sample covariance matrix, $U = \tilde{U} + nS$, and $\tilde{\alpha}(\mathcal{D}) = (n + \alpha_1(\mathcal{D}),\ldots, n + \alpha_p(\mathcal{D}))$. 


3 Model Specification

In this section, we specify our hierarchical model to facilitate the purpose of joint variable and DAG selection for regression models with network-structured predictors. We start by considering the standard Gaussian linear regression model with \( p \) coefficients and by introducing some required notation. Similar to Peterson et al. [2016] and Li and Li [2008a], 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, i.e., \( X_i \sim N_p(0, (LD^{-1}L^T)^{-1}) \), for \( i = 1, 2, \ldots, n \), where \( \beta \) is a \( p \times 1 \) vector of regression coefficients, and \((L, D)\) represents the Cholesky parameter corresponding to a DAG \( \mathcal{D} \). Let matrix \( G \) represent the adjacency matrix corresponding to DAG \( \mathcal{D} \). Our goal is both (i) the variable selection, i.e., to correctly identify all the non-zero regression coefficients, and (ii) network estimation, i.e., to precisely recover the sparsity pattern in \( \mathcal{D} \).

For variable selection, we denote a variable indicator \( \gamma = \{\gamma_1, \ldots, \gamma_p\} \), such that \( \gamma_j = 1 \) if and only if \( \beta_j \neq 0 \), for \( 1 \leq j \leq p \), and let \( \beta_\gamma \) be the vector formed by the active components in \( \beta \) corresponding to a model \( \gamma \). 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 \( X_\gamma \) denote the design matrix formed from the columns of \( X \) corresponding to model \( \gamma \). For the rest of the paper, simply let \( \gamma = \sum_{j=1}^{p} \gamma_j \) represent the size of model \( \gamma \) for notational convenience. For the network estimation, the class of DAG-Wishart distribution in Section 2.2 can be used for joint variable and DAG selection through the following hierarchical model.

\[
Y | X_\gamma, \beta_\gamma \sim N_n(X_\gamma \beta_\gamma, \sigma^2 I_n), \tag{2}
\]

\[
X_i | (L, D), \mathcal{D} \sim N_p(0, (LD^{-1}L^T)^{-1}), \quad \text{for } i = 1, 2, \ldots, n, \tag{3}
\]

\[
(L, D) | \mathcal{D} \sim \pi_{L_0, \alpha(\mathcal{D})}(D, L), \tag{4}
\]

\[
\beta_\gamma | \gamma \sim N_p(0, \tau^2 \sigma^2 I_p), \tag{5}
\]

\[
\pi(\mathcal{D}) \propto \prod_{j=1}^{p-1} q^{\nu_j(\mathcal{D})}(1 - q)^{p-j-\nu_j(\mathcal{D})}, \tag{6}
\]

\[
\pi(\gamma | \mathcal{D}) \propto \exp \left(-a^T \gamma + b^T G_\gamma \right). \tag{7}
\]

for some constants \( \sigma, \tau, a, \) and \( b \geq 0 \). Thus, we assume that \( \sigma \) in (2) is a known constant for simplicity. However, it can be extended to unknown \( \sigma \) case by imposing an inverse-gamma prior, which will be shown in Corollary 4.1. Note that in (5), we are essentially imposing a spike and slab prior on the regression coefficients, where \( \tau^2 \) indicates the variance of the slab part. See Narisetty and He [2014] and Yang et al. [2016] and the references therein. Prior (6) corresponds to an Erdos-Renyi type of prior over the space of DAGs. In particular, similar to Cao et al. [2019a], define \( \gamma_{ji} = \mathbb{I}\{(j, i) \in E(\mathcal{D})\} \), \( 1 \leq j < i \leq p \)
to be the edge indicator. Let \( \gamma_{ji}, 1 \leq i < j < p \) be independent identically distributed Bernoulli\((q)\) random variables. Recall \( \nu_j(\mathcal{D}) = |pa_j(\mathcal{D})| \) is the cardinality of the parent set of vertex \( j \). It follows that

\[
\pi(\mathcal{D}) = \prod_{(j,i):1 \leq j < i \leq p} q^{\gamma_{ji}} (1 - q)^{1 - \gamma_{ji}} = \prod_{j=1}^{p-1} q^{\nu_j(\mathcal{D})}(1 - q)^{p - j - \nu_j(\mathcal{D})}.
\]

**Remark 1.** In (7), we are imposing a Markov Random Field (MRF) prior on the variable indicator \( \gamma \) that favors the inclusion of variables inked to other variables in the associated DAG. This type of MRF priors have also been used in the 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 (7) controls the variable inclusion probability, with smaller 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 (2) to (7) can be used to estimate a DAG as follows. By (1) and Bayes’ rule, the following lemma gives the (marginal) join posterior probabilities with proof provided in Section 6.

**Lemma 3.1.** Under the hierarchical model in (2) to (7), the (marginal) join variable and DAG posterior is given by,

\[
\pi(\gamma, \mathcal{D} | Y, X) \propto \pi(\gamma | \mathcal{D}) \pi(\mathcal{D}) z_{\mathcal{D}}(U + XTX, n + \alpha(\mathcal{D})) \frac{z_{\mathcal{D}}(U, \alpha(\mathcal{D}))}{z_{\mathcal{D}}(U + XTX, n + \alpha(\mathcal{D}))} \times \det \left( \tau^2 X^T X + I_n \right)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T \left( I_n + \tau^2 X^T X \right)^{-1} Y \right) \right\},
\]

where \( z_{\mathcal{D}}(\cdot, \cdot) \) is the normalized constant in the DAG-Wishart distribution.

Hence, after integrating out \( \beta_\gamma \), 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 variable and DAG by computing the posterior mode defined by

\[
(\hat{\gamma}, \hat{\mathcal{D}}) = \arg \max_{(\gamma, \mathcal{D})} \pi(\gamma, \mathcal{D} | Y, X).
\]

**4 Joint Selection Consistency**

In this section we will explore the high-dimensional asymptotic properties of the Bayesian joint variable and DAG selection approach specified in Section 3. For this purpose, we will 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

\[ Y = X\beta_0 + \epsilon_n, \]

where \( Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n \), \( X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p} \), \( X_i \overset{i.i.d.}{\sim} N_p(0, \Sigma_0^n) \) and \( \epsilon_n \sim N_n(0, \sigma_0^2 I_n) \). Here \( \beta_0 \) is the true \( p \)-dimensional vector of regression coefficients, and the entries of \( \epsilon_n \) are i.i.d Gaussian with mean zero and variance \( \sigma_0^2 \). As in the usual context of variable selection, we assume that the true vector of regression coefficients is sparse, i.e., all the entries of \( \beta_0 \) are zero except those corresponding to the active entries in the true variable indicator \( \gamma_0 \) [Castillo et al., 2015, Yang et al., 2016, Narisetty and He, 2014]. For ease of exposition, we also assume that the true quantities \( \gamma_0, \beta_0, \sigma_0^2 \) do not vary with \( n \). However, the results in this paper can be easily extended to the case where \( \gamma_0, \beta_0 \), and entries of \( \gamma_0, \beta_0 \), and \( \sigma_0^2 \) vary with \( n \) but stay bounded. 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 \( \Omega_0^n \). Let \( G_0 \) be the true underlying DAG with structure corresponding to the sparsity pattern in \( L_0^n \), i.e., \( L_0^n \in \mathcal{L}_{\mathcal{D}_0^n} \), and let \( G_0 \) be the adjacency matrix for \( G_0 \). Denote \( d_n \) as the maximum number of non-zero entries in any column of \( L_0^n \) and \( s_n = \min_{1 \leq j \leq p, \ i \in \text{pa}_j(G_0)} |(L_0^n)_{ij}| \). Let \( \bar{p} \) 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 for every \( n \geq 1 \), \( \epsilon_0 \leq \text{eig}_1(\Omega_0^n) \leq \text{eig}_p(\Omega_0^n) \leq \epsilon_0^{-1} \).

This is a standard assumption for high dimensional covariance asymptotic consistency, both in the frequentist and 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. [2019a] relax this assumption by allowing the lower and upper bounds on the eigenvalues to depend on \( p \) and \( n \).

**Assumption 2.** \( d_n \sqrt{\log p/n} \to 0 \) and \( d_n \log p/(s_n^2 n) \to 0 \) as \( n \to \infty \).

This assumption resembles the dimension assumption in Cao et al. [2019b], and is a much weaker assumption for high dimensional covariance asymptotic than for example, Xiang et al. [2015], Banerjee and Ghosal [2014, 2015] and Cao et al. [2019a]. Here we essentially allow the dimension of our covariance matrix to grow slower than \( \exp(n/d_n^2) \). Recall that \( s_n \) is the smallest (in absolute value) non-zero 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 that is needed for establishing consistency. This type of condition has been used for the exact support recovery of the high-dimensional linear regression models as well as Gaussian DAG models. See for example Yang et al. [2016], Khare et al. [2017], Lee et al. [2018] and Cao et al. [2019a].
Assumption 3. Let $q_n = O(p^{-\alpha_1})$ and $R_n \sim n/\log n$ for some constant $\alpha_1 > 0$. We assume that $\pi(\mathcal{D}) = 0$ and $\pi(\gamma|\mathcal{D}) = 0$ for all $(\mathcal{D}, \gamma)$ satisfying $\max_{1 \leq j \leq p-1} \nu_i(\mathcal{D}) \geq R_n$ and $|\gamma| > R_n$.

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 $2^p$ possible models, places zero mass on unrealistically large models.

Note that our $q_n$ is of slower rate approaching zero compared to the one in Cao et al. [2019a], which helps avoiding the potential computation limitation such as simulation results always favor the most sparse model.

This assumption also states that the MRF prior on the space of the $2^n$ possible models, places zero mass on unrealistically large models (see similar assumptions in Shin et al. [2018], Narisetty and He [2014] in the context of regression).

Assumption 4. For every $n \geq 1$, the hyperparameters for the DAG-Wishart prior $\pi^{\Theta_{n, \mathcal{D}_n}}_{U_n, \alpha(\mathcal{D}_n)}$ satisfy (i) $2 < \alpha_i(\mathcal{D}_n) - \nu_i(\mathcal{D}_n) < c$ for every $\mathcal{D}_n$ and $1 \leq i \leq q_n$, and (ii) $0 < \delta_1 \leq \text{eig}_1(U_n) \leq \text{eig}_p(U_n) \leq \delta_2 < \infty$. Here $c, \delta_1, \delta_2$ are constants not depending on $n$.

This assumption provides mild restrictions on the hyperparameters for the DAG-Wishart distribution. The assumption $2 < \alpha_i(\mathcal{D}) - \nu_i(\mathcal{D})$ establishes prior propriety. The assumption $\alpha_i(\mathcal{D}) - \nu_i(\mathcal{D}) < c$ implies that the shape parameter $\alpha_i(\mathcal{D})$ can only differ from $\nu_i(\mathcal{D})$ (number of parents of $i$ in $\mathcal{D}$) 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 $n$.

Assumption 5. The hyperparameters in model (5) and the MRF prior (7) satisfy $\tau^2 \sim \sqrt{\log p}$, $a \sim \alpha_1 \log p$, and $bR_n^2 = o(\log p)$.

Recall that the parameter $a$ in (7) controls the variable inclusion probability, and $b$ reflects that how strongly the inclusion probability of a variable is affected by the inclusion of its neighbors in the DAG. In the Bayesian variable selection literature, similar priors corresponding to $a = C \log p$ 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 on the the variance of the slab prior, $\tau^2$, is required to approach infinity is also stated here to ensure desired model selection consistency.

For the rest of this paper, $p_n, \Omega_0^n, \Sigma_0^n, L_0^n, D_0^n, \mathcal{D}_0^n, d_n, q_n, \tau_n, A_n$ will be denoted as $p, \Omega_0, \Sigma_0, L_0, D_0, \mathcal{D}_0, d, q, \tau, A$ as needed 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 $\gamma$ and $D$

In this section, we show that our method guarantees the posterior ratio consistency of $\gamma$ and $D$. Although Peterson et al. [2016] consider a similar network-structured regression model, theoretical properties of Bayesian models such as posterior ratio consistency and joint selection consistency have not been established yet up to our knowledge. We first establish the posterior ratio consistency with respect to $D$ under the true variable indicator $\gamma_0$. Theorem 4.1 says that the true DAG will be the mode of the posterior distribution with probability tending to 1 as $n \to \infty$ under fixed $\gamma_0$.

**Theorem 4.1.** Under Assumptions 1-4,

$$\max_{D \neq \hat{D}_0} \frac{\pi(\gamma_0, \hat{D}|Y, X)}{\pi(\gamma, \hat{D}_0|Y, X)} \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{4.1}$$

The next theorem establishes the posterior ratio consistency with respect to $\gamma$ under DAG $D$. This notion of consistency implies that the true variable indicator $\gamma_0$ will be the mode of the posterior distribution with probability tending to 1 as $n \to \infty$ under fixed $\hat{D}$.

**Theorem 4.2.** Under Assumptions 1-5, the following holds:

$$\max_{(\gamma, D) \neq (\gamma_0, \hat{D}_0)} \frac{\pi(\gamma, \hat{D}|Y, X)}{\pi(\gamma_0, \hat{D}_0|Y, X)} \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{4.2}$$

From Theorem 4.1, Theorem 4.2 and the fact that

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

we can obtain the joint posterior ratio consistency with respect to both $\gamma$ and $D$. It implies that the true variable indicator and DAG, $(\gamma_0, \hat{D}_0)$, will be the mode of the posterior distribution with probability tending to 1.

**Theorem 4.3.** Under Assumptions 1-5, the following holds:

$$\max_{(\gamma, D) \neq (\gamma_0, \hat{D}_0)} \frac{\pi(\gamma, \hat{D}|Y, X)}{\pi(\gamma_0, \hat{D}_0|Y, X)} \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{4.3}$$

which implies that

$$\hat{P}((\hat{\gamma}, \hat{D}) = (\gamma_0, \hat{D}_0)) \to 1, \quad \text{as } n \to \infty.$$
4.2 Strong selection consistency of $\gamma$ and $\mathcal{D}$

In this section, we establish the joint strong selection consistency with respect to both $\gamma$ and $\mathcal{D}$. Theorem 4.4 shows that the posterior probability assigned to the true variable indicator $\gamma_0$ and the true underlying graph $\mathcal{D}_0$ grows to 1 as $n \to \infty$. We call this property the joint strong selection consistency. Note that the result given in Theorem 4.3 does not guarantee the joint strong selection consistency.

**Theorem 4.4.** Under Assumptions 1-5, if we further assume $\alpha_1 > 2\kappa$ for some constant $\kappa > 1$, the following holds:

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

We would like to point out that the condition on $\alpha_1$, which controls the rate of independent Bernoulli probability specified in Assumption 3, is only needed for strong selection consistency (Theorem 4.4). Similar restrictions on the hyperparameters have been considered for establishing consistency properties in the regression setup [Yang et al., 2016, Lee et al., 2018, Cao et al., 2019c]. The model selection consistency for the posterior mode in Theorem 4.3 does not require any restriction on $\alpha_1$.

All the aforementioned theorems are based on known $\sigma^2$, which tends to be not flexible enough, as in real applications, the underlying true variance often remains unavailable. Therefore, we introduce the following corollary for a fully Bayesian hierarchical approach, where an appropriate inverse-gamma prior is imposed on $\sigma^2$. It turns out that even with the unknown $\sigma^2$, strong model selection consistency still holds under the same conditions given in Theorem 4.4.

**Corollary 4.1.** Suppose $\sigma^2$ is unknown and a proper inverse gamma density with some positive constant parameters $(a_0, b_0)$ is placed on $\sigma^2$. Under Assumptions 1-5, and $\alpha_1 > 2\kappa$ for some constant $\kappa > 1$, the following holds:

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

5 Numerical Studies

5.1 Posterior inference

For positive real values $a_0$ and $b_0 > 0$, let $IG(a_0, b_0)$ be the inverse-gamma distribution with the shape parameter $a_0$ and scale parameter $b_0$. Then, similar to (8), the joint posterior distribution of $\gamma$ and $\mathcal{D}$ based on (2)–(7) and $\sigma^2 \sim IG(a_0, b_0)$ is

$$
\pi(\gamma, \mathcal{D} | Y, X)
$$
\begin{align*}
\alpha = \pi(\gamma | \mathcal{D}) \pi(\mathcal{D}) \frac{z_{\mathcal{D}}(U + X^T X, n + \alpha(\mathcal{D}))}{z_{\mathcal{D}}(U, \alpha(\mathcal{D}))} \det \left( I_n + \tau^2 X_\gamma X_\gamma^T \right)^{-\frac{1}{2}} \left\{ b_0 + \frac{1}{2} Y^T (I_n + \tau^2 X_\gamma X_\gamma^T) Y \right\}^{-\frac{n+2n_0}{2}}.
\end{align*}

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

1. Set the initial values \( \gamma^{(1)} \) and \( \mathcal{D}^{(1)} \).

2. For each \( s = 2, \ldots, S \),
   (a) sample \( \gamma^{\text{new}} \sim q_{\gamma}(\cdot | \gamma^{(s-1)}) \);
   (b) set \( \gamma^{(s)} = \gamma^{\text{new}} \) with the probability
   \[
   p_{\text{acc}, \gamma} = \min \left\{ 1, \frac{\pi(\gamma^{\text{new}} | \mathcal{D}^{(s-1)}, Y, X) q_{\gamma}^{(s-1)} | \mathcal{D}^{(s-1)}, Y, X) q_{\gamma}^{(\gamma^{new}) | \gamma^{(s-1)}}}{\pi(\gamma^{(s-1)} | \mathcal{D}^{(s-1)}, Y, X) q_{\gamma}^{(\gamma) | \gamma^{(s-1)}}} \right\},
   \]
   otherwise set \( \gamma^{(s)} = \gamma^{(s-1)} \);
   (c) sample \( \mathcal{D}^{\text{new}} \sim q_{\mathcal{D}}(\cdot | \mathcal{D}^{(s-1)}) \);
   (d) set \( \mathcal{D}^{(s)} = \mathcal{D}^{\text{new}} \) with the probability
   \[
   p_{\text{acc}, \mathcal{D}} = \min \left\{ 1, \frac{\pi(\mathcal{D}^{\text{new}} | \gamma^{(s)}, Y, X) q_{\mathcal{D}}^{(\mathcal{D}^{(s-1)}) | \mathcal{D}^{(s-1)}, \gamma^{(s)}, Y, X) q_{\mathcal{D}}^{(\mathcal{D}^{\text{new}}) | \mathcal{D}^{(s-1)}}}{\pi(\mathcal{D}^{(s-1)} | \gamma^{(s)}, Y, X) q_{\mathcal{D}}^{(\mathcal{D}^{(s-1)}) | \mathcal{D}^{(s-1)}, \gamma^{(s)}, Y, X) q_{\mathcal{D}}^{(\mathcal{D}^{\text{new}}) | \mathcal{D}^{(s-1)}}} \right\},
   \]
   otherwise set \( \mathcal{D}^{(s)} = \mathcal{D}^{(s-1)} \).

The inference for the DAG \( \mathcal{D} \), the steps 2-(c) and 2-(d) in the above algorithm, can be parallelized for each column. For more details, we refer to Cao et al. [2019a] and Lee et al. [2018]. We used the proposal kernel \( q_{\gamma}(\cdot | \gamma') \) which gives a new set \( \gamma^{\text{new}} \) by changing a randomly chosen nonzero component in \( \gamma' \) to 0 with probability 0.5 or by changing a randomly chosen zero component to 1 randomly with probability 0.5. The same proposal kernels were used for each column of \( \mathcal{D} \).

### 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 that we have \( X_i = (X_{i1}, \ldots, X_{ip}) \sim_{i.i.d.} N_p(0, \Sigma_0), i = 1, \ldots, n \), 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 in this case. Among 240 genes, we assume that there are 40 transcription factors (TFs) and each TF regulates 5 genes. Let \( TF_j \) be the index for the \( j \)th TF and \( (TF_1, TF_2, \ldots, TF_{40}) = (6, 12, \ldots, 240) \). It corresponds to the DAG \( \mathcal{D}_0 \), the support of \( L_0 \), such that \( pa_{TF_j-k}(\mathcal{D}_0) = \{ TF_j \} \) for \( j = 1, \ldots, 40 \) and \( k = 1, \ldots, 5 \). Suppose that the TFs independently follow the normal distribution, that is, \( X_{TF_j} \sim_{i.i.d.} N(0, d_{TF_j}), \) where \( d_{TF_j} \sim_{i.i.d.} \text{Unif}(3, 5), \) for \( j = 1, \ldots, 40 \). We further
assume that, given \( X_{TF_j} \), the conditional distribution of the gene \( X_j \) that \( TF_j \) regulates is \( N(X_{TF_j}, d_j) \), where \( d_j \sim Unif(3, 5) \) for \( j = 1, \ldots, 240 \). It corresponds to the true modified Cholesky parameter \((L_0, D_0)\) such that \((L_0)_{TF_j,TF_i-k} = 1\) and \(D_0 = \text{diag}(d_j)\), where \( d_j \sim Unif(3, 5) \) for \( j = 1, \ldots, 40 \) and \( k = 1, \ldots, 5 \). We simulate the data from

\[
Y = X\beta_0 + \epsilon,
\]

where \( X = (X_1, \ldots, X_n)^T \) and \( \epsilon \sim N_n(0, \sigma^2 I_n) \) and \( \sigma^2 = \|\beta_0\|^2/4 \). We investigate four settings for the true coefficient vector \( \beta_0 \) as described in Li and Li [2008b] and Peterson et al. [2016]. In the first setting, it is assumed that \( \beta_{0,TF_{1:4}} = (5, -5, 3, -3)^T \), \( \beta_{0,TF_{1:5}} = \beta_{0,TF_{j}}/\sqrt{10} \) for \( j = 1, 2, 3, 4 \) and \( k = 1, \ldots, 5 \), and \( \beta_{0,j} = 0 \) for \( j = 25, \ldots, 240 \). This setting implies the genes in the same cluster have the same signs for the coefficients. In the second setting, the true coefficient \( \beta_0 \) is the same as the first setting except that the signs are reversed for the two genes that \( TF_j \) regulates, i.e., \( \beta_{0,TF_{1:4}} = -\beta_{0,TF_{1:4}}/\sqrt{10} \) for \( j = 1, 2, 3, 4 \) and \( k = 1, 2 \). This setting implies that the genes in the same cluster might have different signs for the coefficients.

The third and fourth settings the same as the first and second settings except for considering \( 10 \) instead of \( \sqrt{10} \). Thus, they consider smaller signals. We call this simulation setting the Scenario 1.

We also investigate a different simulation scenario, say Scenario 2, where the signals in \( \beta_0 \) are small. In this case, there are \( p = 150 \) genes, \( 30 \) TFs and \( 4 \) regularized genes for each TF. The precision matrix \( \Sigma_0^{-1} = L_0(D_0)^{-1}L_0^T \) is generated by \( d_j \sim Unif(2, 5) \) and \( (L_0)_{TF_j,TF_j-k} \sim Unif(0.3, 0.7) \). The variance of \( \epsilon \) is chosen as \( \sigma^2 = \|\beta_0\|^2/2 \). We consider four settings for the true coefficient vector \( \beta_0 \). In the first and third settings, \( \beta_0 \) is generated by \( \beta_{0,j} \sim Unif(0.5, 1) \) and \( \beta_{0,j} \sim Unif(0.2, 1) \) for \( j = 1, \ldots, 20 \), respectively, and \( \beta_{0,j} = 0 \) for \( j = 21, \ldots, 150 \). In the second and fourth settings, we only change the signs of nonzero entries of \( \beta_0 \) randomly.

We compare the performance of our joint selection method with other 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 in Lasso and elastic net were chosen by 10-fold cross-validation. For 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 default. The other hyperparameters were set at \( a_0 = 0.1, b_0 = 0.01 \) and \( \tau^2 = 1 \). The initial state for \( \gamma \) was set at \( p \)-dimensional zero vector, i.e., the empty model, while the initial state for \( \mathcal{D} \) was chosen by the CSCS method [Khare et al., 2017]. For posterior inference, 5,000 posterior samples were drawn with a burn-in period of 5,000. The indices having posterior inclusion probability larger than 0.5 were included in the final model.

To evaluate the performance of variable selection, the sensitivity, specificity, Matthews correlation coeffi-
Table 1: The summary statistics for Scenario 1 are represented for each setting. Different setting means different choice of the true coefficient $\beta_0$. MCC, #Error and PMSE are Matthews correlation coefficient, the number of errors and mean-squared prediction error, 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.

<table>
<thead>
<tr>
<th>Setting 1</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.8750</td>
<td>0.9861</td>
<td>0.8611</td>
<td>6</td>
<td>69.1445</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.8750</td>
<td>0.9861</td>
<td>0.8611</td>
<td>6</td>
<td>71.0443</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.0000</td>
<td>0.8056</td>
<td>0.5412</td>
<td>42</td>
<td>45.5522</td>
</tr>
<tr>
<td>Elastic</td>
<td>1.0000</td>
<td>0.9352</td>
<td>0.7085</td>
<td>14</td>
<td>41.8631</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 2</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.8750</td>
<td>0.9954</td>
<td>0.9049</td>
<td>4</td>
<td>56.4885</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.7500</td>
<td>0.9954</td>
<td>0.8282</td>
<td>7</td>
<td>73.7870</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.7083</td>
<td>0.8519</td>
<td>0.4170</td>
<td>39</td>
<td>106.0526</td>
</tr>
<tr>
<td>Elastic</td>
<td>0.8750</td>
<td>0.8426</td>
<td>0.5122</td>
<td>37</td>
<td>92.6665</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 3</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.2083</td>
<td>0.9907</td>
<td>0.3549</td>
<td>21</td>
<td>42.5213</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.2500</td>
<td>0.9907</td>
<td>0.4023</td>
<td>20</td>
<td>40.3569</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.8000</td>
<td>0.8241</td>
<td>0.5648</td>
<td>38</td>
<td>40.3191</td>
</tr>
<tr>
<td>Elastic</td>
<td>1.0000</td>
<td>0.9444</td>
<td>0.7085</td>
<td>12</td>
<td>29.3908</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 4</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.7500</td>
<td>1.0000</td>
<td>0.8174</td>
<td>6</td>
<td>20.9925</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.6500</td>
<td>1.0000</td>
<td>0.7854</td>
<td>7</td>
<td>15.4705</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.9500</td>
<td>0.8154</td>
<td>0.5754</td>
<td>25</td>
<td>11.6528</td>
</tr>
<tr>
<td>Elastic</td>
<td>0.9500</td>
<td>0.8923</td>
<td>0.6912</td>
<td>15</td>
<td>10.7742</td>
</tr>
</tbody>
</table>

Table 2: The summary statistics for Scenario 2 are represented for each setting. Different setting means different choice of the true coefficient $\beta_0$.

<table>
<thead>
<tr>
<th>Setting 1</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.7500</td>
<td>0.9907</td>
<td>0.8174</td>
<td>6</td>
<td>20.9925</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.6500</td>
<td>1.0000</td>
<td>0.7854</td>
<td>7</td>
<td>15.4705</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.0000</td>
<td>0.8308</td>
<td>0.6290</td>
<td>22</td>
<td>14.8092</td>
</tr>
<tr>
<td>Elastic</td>
<td>0.9500</td>
<td>0.9077</td>
<td>0.7201</td>
<td>13</td>
<td>18.9942</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Setting 2</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>#Error</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint.CL</td>
<td>0.6000</td>
<td>1.0000</td>
<td>0.7518</td>
<td>8</td>
<td>15.8789</td>
</tr>
<tr>
<td>Joint.P</td>
<td>0.5000</td>
<td>1.0000</td>
<td>0.6814</td>
<td>10</td>
<td>19.2450</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.9000</td>
<td>0.7692</td>
<td>0.4877</td>
<td>32</td>
<td>13.4260</td>
</tr>
<tr>
<td>Elastic</td>
<td>0.8000</td>
<td>0.8615</td>
<td>0.5371</td>
<td>32</td>
<td>13.4260</td>
</tr>
</tbody>
</table>

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 = \( \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\hat{Y}_i - Y_{test,i})^2 \),

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively. Here we denote \( \hat{Y}_i = X_i^T \hat{\beta} \), where \( \hat{\beta} \) is the estimated coefficient based on each method. For Bayesian methods, the usual least square estimates based on the selected support were used as \( \hat{\beta} \). We generated test samples \( Y_{test,1}, \ldots, Y_{test,n_{test}} \) with \( n_{test} = 100 \) to calculate the MSPE.
We notice that 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. It leads to relatively larger number of errors for the regularization methods compared with those for the Bayesian joint selection methods. We also found that the Bayesian joint selection method proposed in this paper (Joint.CL) works better than that proposed by Peterson et al. [2016] (Joint.P). In fact, the two Bayesian joint selection methods are quite similar to each other except 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 would be preferable in this case. Furthermore, based on asymptotic results, one can expect that our method will give accurate inference results as we have more observations, while asymptotic properties of the Bayesian method proposed by Peterson et al. [2016] are still in question.

6 Proofs

In this section, we provide proofs for Lemma 3.1, Theorems 4.1 to 4.4, and Corollary 4.1.

Proof of Lemma 3.1. It follows from the hierarchical models in (2) to (7), we have

\[
\pi (\gamma, \mathcal{D}|Y, X) \\
= \int \pi (Y|\gamma, \beta) \prod_{i=1}^{n} \pi (X_i|(L, D)) \pi^{\Theta, \alpha(\mathcal{D})}((L, D)) \\
\times \pi (\beta|\gamma) \pi (\mathcal{D}) \, d\beta, d(L, D) \\
= \pi(\gamma) \pi(\mathcal{D}) \int \pi (Y|\gamma, \beta) \pi (\beta|\gamma) \, d\beta \\
\times \int \prod_{i=1}^{n} \pi (X_i|(L, D)) \pi^{\Theta, \alpha(\mathcal{D})}((L, D)) \, d(L, D).
\]

(10)

First, note that by the conjugacy of the DAG-Wishart distribution, we have

\[
\int \prod_{i=1}^{n} \pi (X_i|(L, D)) \pi^{\Theta, \alpha(\mathcal{D})}((L, D)) \, d(L, D) \\
= \frac{z_{\mathcal{D}}(U + X^T X, n + \alpha(\mathcal{D}))}{z_{\mathcal{D}}(U, \alpha(\mathcal{D}))},
\]
where $z_{\phi}(\cdot, \cdot)$ is the normalized constant for the DAG-Wishart distribution. Next, note that

$$
\int \pi(Y|\gamma, \beta_\gamma) \pi(\beta_\gamma|\gamma) \, d\beta_\gamma \\
\propto \int (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} (Y - X_\gamma \beta_\gamma)^T (Y - X_\gamma \beta_\gamma) \right\} \\
\times (\tau^2 \sigma^2)^{-\frac{1}{2} |\gamma|} \exp\left\{-\frac{1}{2\tau^2 \sigma^2} \beta_\gamma^T \beta_\gamma \right\} \, d\beta_\gamma \\
\propto (\tau^2)^{-\frac{1}{2} |\gamma|} (\sigma^2)^{-\frac{n+\frac{1}{2} |\gamma|}{2}} \exp\left\{-\frac{1}{2\sigma^2} \left( \beta_\gamma^T \left( X_\gamma^T X_\gamma + \frac{1}{\tau^2} I \right) \beta_\gamma - 2 \beta_\gamma^T X_\gamma Y \right) \right\} \, d\beta_\gamma \\
\propto (\tau^2)^{-\frac{1}{2} |\gamma|} (\sigma^2)^{-\frac{n+\frac{1}{2} |\gamma|}{2}} \exp\left\{-\frac{1}{2\tau^2 \sigma^2} \beta_\gamma^T \left( X_\gamma^T X_\gamma + \frac{1}{\tau^2} I \right) \beta_\gamma \right\} \\
\propto \left( \frac{1}{\sigma^2} \right)^{-\frac{n}{2}} \det \left( \tau^2 X_\gamma^T X_\gamma + I \right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} \left( Y^T (I_n + \tau^2 X_\gamma^T X_\gamma)^{-1} Y \right) \right\},
$$

where the last term follows from the Woodbury matrix identity. Therefore, by (10), under the proposed hierarchical model and known $\sigma^2$, we have

$$
\pi(\gamma, \mathcal{D}|Y, X) \\
\propto \pi(\gamma|\mathcal{D}) \pi(\mathcal{D})z_{\phi}(U + X^T X, \alpha(\mathcal{D})) \\
\times \det \left( \tau^2 X_\gamma^T X_\gamma + I \right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} \left( Y^T (I_n + \tau^2 X_\gamma^T X_\gamma)^{-1} Y \right) \right\},
$$

where $z_{\phi}(\cdot, \cdot)$ is the normalized constant in the DAG-Wishart distribution. \hfill \Box

**Proof of Theorem 4.1.** It follows from Assumption 5 and model (7) that, for large enough $n > N$,

$$
\frac{\pi(\gamma_0|\mathcal{D})}{\pi(\gamma_0|\mathcal{D}_0)} = \exp \left( b_0^T (G - G_0) \gamma_0 \right) \\
\leq \exp \left( b |\gamma_0|^2 \right) \leq \exp(|\gamma_0|^2).
$$

(13)

Let $S = \frac{1}{n} X^T X$ denote the sample covariance matrix of $X$. It follows from (8), (13), and Lemma 5.1 in
Define the event $E$ and $M$ (Lemma 6.1-6.3). Recall that our goal is to find an upper bound (independent of $\nu_4(\mathcal{D}) - \nu_1(\mathcal{D})$)

For all the following analyses, we will restrict ourselves to the event $E_n$ (Lemma A.3 of [Bickel and Levina, 2008], Hanson-Wright inequality from [Rudelson and Vershynin, 2013] and the union-sum inequality, there exists constants $c$, $m_1$, $m_2$, such that

$$P \left( \| \tilde{S} - \Sigma_0 \|_{\max} \geq c' \sqrt{\frac{\log p}{n}} \right) \leq m_1 p^{2-m_2(c')^2/4} \rightarrow 0.$$  

For all the following analyses, we will restrict ourselves to the event $E_n$.

We now analyze the behavior of $PR_i(\mathcal{D}, \mathcal{D}_0)$ under different scenarios in a sequence of three lemmas (Lemma 6.1-6.3). Recall that our goal is to find an upper bound (independent of $\mathcal{D}$ and $i$) for $PR_i(\mathcal{D}, \mathcal{D}_0)$, such that the upper bound converges to 0 as $n \rightarrow \infty$.

**Lemma 6.1.** If $pa_i(\mathcal{D}) \supset pa_i(\mathcal{D}_0)$, then there exists $N_1$ (not depending on $i$ or $\mathcal{D}$) such that for $n \geq N_1$ we have $PR_i(\mathcal{D}, \mathcal{D}_0) \leq 2p^{-\Delta \nu_4(\mathcal{D}) - \nu_1(\mathcal{D})}$, for any constant $\kappa > 1$.

**Proof of Lemma 6.1.** Since $pa_i(\mathcal{D}) \supset pa_i(\mathcal{D}_0)$, we can write $|\tilde{S}_{\mathcal{D}_0}^\perp| = |\tilde{S}_{\mathcal{D}_0}^\perp| R_{\mathcal{D}_0}^{-1}$. Here $R_{\mathcal{D}_0}^{-1}$ is the Schur complement of $\tilde{S}_{\mathcal{D}_0}^\perp$, defined by $R_{\mathcal{D}_0}^{-1} = D - B^T (\tilde{S}_{\mathcal{D}_0}^\perp)^{-1} B$, for appropriate sub matrices $B$ and $D$ of $\tilde{S}_{\mathcal{D}_0}^\perp$.

Since $\tilde{S}_{\mathcal{D}_0}^\perp \geq (\frac{\sqrt{n}}{\sigma_2})^2$, and $R_{\mathcal{D}_0}^{-1}$ is a principal submatrix of $(\tilde{S}_{\mathcal{D}_0}^\perp)^{-1}$, the largest eigenvalue of $R_{\mathcal{D}_0}^{-1}$ is bounded above by $\frac{n}{\sigma_2}$. Therefore,

$$\left( \frac{|\tilde{S}_{\mathcal{D}_0}^\perp|}{|\tilde{S}_{\mathcal{D}_0}^\perp|} \right)^{1/2} \leq \left( \frac{n}{\sigma_2} \right)^{\nu_4(\mathcal{D}) - \nu_1(\mathcal{D})}.$$  

(17)

\footnote{For matrices $A$ and $B$, we say $A \geq B$ if $A - B$ is positive semi-definite}
Denote $S_{j|\mathcal{G}_i} = S_{jj} - (S_{\mathcal{G}_i \mathcal{G}_j})^T (S_{\mathcal{G}_j \mathcal{G}_j})^{-1} S_{\mathcal{G}_i \mathcal{G}_j}$. It immediately follows that

$$\hat{S}_{i|\mathcal{P}_a, (\mathcal{G})} \geq S_{i|\mathcal{P}_a, (\mathcal{G})}. \quad (18)$$

Since we are restricting ourselves to the event $E_n^\infty$, it follows from (15) that

$$||S_{\mathcal{G}_0}^{\geq i} - (\Sigma_0)^{\geq i}_{\mathcal{G}_0}||_{(2,2)} \leq (\nu_t(\mathcal{G}_0) + 1)c' \sqrt{\frac{\log p}{n}}. \quad (19)$$

Therefore,

$$\begin{align*}
||((S_{\mathcal{G}_0}^{\geq i})^{-1} - ((\Sigma_0)^{\geq i}_{\mathcal{G}_0})^{-1}||_{(2,2)}
&=||((S_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)}||S_{\mathcal{G}_0}^{\geq i} - (\Sigma_0)^{\geq i}_{\mathcal{G}_0}||_{(2,2)}||((\Sigma_0)^{\geq i}_{\mathcal{G}_0})^{-1}||_{(2,2)}
&\leq(||((S_{\mathcal{G}_0}^{\geq i})^{-1} - ((\Sigma_0)^{\geq i}_{\mathcal{G}_0})^{-1}||_{(2,2)} + \frac{1}{\epsilon_0})(\nu_t(\mathcal{G}_0) + 1)c' \sqrt{\frac{\log p}{n}}. \quad (19)
\end{align*}$$

Recall $d = \max_{1 \leq i \leq p-1} \nu_t(\mathcal{G}_0)$. By the assumption that $d \sqrt{\frac{\log p}{n}} \to 0$ and (19), for large enough $n$, we have

$$||((S_{\mathcal{G}_0}^{\geq i})^{-1} - ((\Sigma_0)^{\geq i}_{\mathcal{G}_0})^{-1}||_{(2,2)} \leq \frac{4c'}{\epsilon_0} \sqrt{\frac{\log p}{n}} = o(1) \quad (20)$$

and

$$\frac{1}{S_{i|\mathcal{P}_a, (\mathcal{G}_0)}} = \left[ (S_{\mathcal{G}_0}^{\geq i})^{-1} \right]_{ii} \geq \frac{\epsilon_0}{2}. \quad (21)$$

Note that for any $\mathcal{G}$, $||S_{\mathcal{G}_0}^{\geq i} - S_{\mathcal{G}}^{\geq i}||_{\max} \leq \frac{\delta_2}{n}$ gives us $||\hat{S}_{\mathcal{G}_0}^{\geq i} - S_{\mathcal{G}_0}^{\geq i}||_{(2,2)} \leq (\nu_t(\mathcal{G}_0) + 1)\frac{\delta_2}{n}$. Therefore,

$$\begin{align*}
||((\hat{S}_{\mathcal{G}_0}^{\geq i})^{-1} - (S_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)}
&=||((\hat{S}_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)}||\hat{S}_{\mathcal{G}_0}^{\geq i} - S_{\mathcal{G}_0}^{\geq i}||_{(2,2)}||S_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)}
&\leq(||((\hat{S}_{\mathcal{G}_0}^{\geq i})^{-1} - (S_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)} + ||S_{\mathcal{G}_0}^{\geq i})^{-1} - ((\Sigma_0)^{\geq i}_{\mathcal{G}_0})^{-1}||_{(2,2)} + \frac{1}{\epsilon_0}) \times \left( \frac{1}{\epsilon_0} + o(1) \right)
\times (\nu_t(\mathcal{G}_0) + 1)\frac{\delta_2}{n}. \quad (22)
\end{align*}$$

Following from (20), (21), and $\frac{d}{n} \to 0$, for large enough $n$, (22) yields

$$||((\hat{S}_{\mathcal{G}_0}^{\geq i})^{-1} - (S_{\mathcal{G}_0}^{\geq i})^{-1}||_{(2,2)} \leq \frac{8\delta_2}{\epsilon_0} \frac{d}{n} \text{ and } \frac{1}{S_{i|\mathcal{P}_a, (\mathcal{G})}} = \left[ (\hat{S}_{\mathcal{G}}^{\geq i})^{-1} \right]_{ii} \geq \frac{\epsilon_0}{4}. \quad (23)$$
Hence, it follow from (23) and (21) that,

\[
\left| \frac{1}{S_{i[p_a_i(D_0)]}} - \frac{1}{\tilde{S}_{i[p_a_i(D_0)]}} \right| \leq \frac{8\delta_2}{\epsilon_0^2} \frac{n}{M}
\]  \tag{24}

and

\[
|S_{i[p_a_i(D_0)]} - \tilde{S}_{i[p_a_i(D_0)]}| \leq c_1 \frac{d}{n}
\]  \tag{25}

where \(c_1 = 64\delta_2/\epsilon_0^4\) is a constant.

Further note that when \(pa_i(D_0) \subset pa_i(D)\), \(n(D_0)_{ii}^{-1}S_{i[p_a_i(D_0)]} \sim \chi^2_{n-\nu_i(D)}\) and \(n(D_0)_{ii}^{-1}S_{i[p_a_i(D_0)]} \overset{d}{=} n(D_0)_{ii}^{-1}S_{i[p_a_i(D)]} \oplus \chi^2_{\nu_i(D)} - \nu_i(D_0)\) under the true model. By Lemma 4.1 in [Cao et al., 2019c], we get

\[
P\left[ |n(D_0)_{ii}^{-1}S_{i[p_a_i(D_0)]} - (n - \nu_i(D))| > \sqrt{(n - \nu_i(D)) \log p} \right] \leq 2p^{-\frac{1}{8}} \to 0,
\]  \tag{26}

and

\[
P\left[ |n(D_0)_{ii}^{-1}S_{i[p_a_i(D_0)]} - n(D_0)_{ii}^{-1}S_{i[p_a_i(D)]} - (\nu_i(D) - \nu_i(D_0))| \right.
\]
\[
> \sqrt{(\nu_i(D) - \nu_i(D_0)) \log p}
\]
\[
\leq 2p^{-\frac{1}{8}} \to 0,
\]  \tag{27}

Following from (17), (18), (23), (25), (26), (27), and Assumption 3, for larger enough \(n > N_1\) and some constant \(M'\), we have

\[
PR_i(D, D_0)
\]
\[
\leq M' \left( \frac{\delta_2}{\delta_1} \right)^{\frac{1}{2}} n^{2c} (2p^{-\alpha_1})^{\nu_i(D) - \nu_i(D_0)}
\times \left( \frac{1 + \frac{n(D_0)_{ii}^{-1}S_{i[p_a_i(D_0)]} - n(D_0)_{ii}^{-1}S_{i[p_a_i(D)]} + c_1 \frac{d}{(D_0)_{ii}}}{n(D_0)_{ii}^{-1}S_{i[p_a_i(D)]}} \right) ^{2n + c - 3}
\leq M' \left( \frac{\delta_2}{\delta_1} \right)^{\frac{1}{2}} n^{2c} (2p^{-\alpha_1})^{\nu_i(D) - \nu_i(D_0)}
\times \exp \left\{ \frac{\nu_i(D) - \nu_i(D) + \sqrt{\nu_i(D) - \nu_i(D_0) \log p + c_1 \frac{d}{(D_0)_{ii}} n + c - 3}}{n - \nu_i(D) - \sqrt{(n - \nu_i(D)) \log p}} \right\}
\leq (2p)^{-\frac{1}{2} \kappa (\nu_i(D) - \nu_i(D_0))}, \text{ for any constant } \kappa > 1.
\]  \tag{28}
The second inequality follows from $\frac{d}{\log p} \to 0$ and $\frac{\nu(\varrho)}{n} \to 0$, as $n \to \infty$ and $\frac{c}{2} \leq (D_0)_{ii} \leq \frac{2}{c_0}$. 

**Lemma 6.2.** If $pa_i(\mathcal{D}) \subset pa_i(\mathcal{D}_0)$, then there exists $N_2$ (not depending on $i$ or $\mathcal{D}$) such that for $n \geq N_1$ we have $PR_i(\mathcal{D}, \mathcal{D}_0) \leq p^{-\frac{2\Delta_i}{d}}$.

**Proof of Lemma 6.2.** Now we move to discuss the scenario when $pa_i(\mathcal{D})$ is a subset of $pa_i(\mathcal{D}_0)$, i.e., $pa_i(\mathcal{D}) \subset pa_i(\mathcal{D}_0)$. Since $pa_i(\mathcal{D}_0) \supset pa_i(\mathcal{D})$, we can write $|\tilde{S}_{\mathcal{D}}^{\geq i}| = |\tilde{S}_{\mathcal{D}_0}^{\geq i}| |R_{\tilde{S}_{\mathcal{D}}^{\geq i}}|$, where $R_{\tilde{S}_{\mathcal{D}}^{\geq i}}$ denotes the Schur complement of $\tilde{S}_{\mathcal{D}}^{\geq i}$, defined by $R_{\tilde{S}_{\mathcal{D}}^{\geq i}} = \tilde{D} - \tilde{B}^T(\tilde{S}_{\mathcal{D}}^{\geq i})^{-1}\tilde{B}$ for appropriate sub matrices $\tilde{B}$ and $\tilde{D}$ of $\tilde{S}_{\mathcal{D}_0}^{\geq i}$.

It follows by (19) that if restrict to $E_c$, we have $||(\tilde{S}_{\mathcal{D}_0}^{\geq i})^{-1} - ((\Sigma_0)^{\geq i})_{\mathcal{D}_0, \mathcal{D}}^{-1}||_{(2,2)} \leq 4c' d \sqrt{\log p \over n}$ and $||R_{\tilde{S}_{\mathcal{D}}^{\geq i}}^{-1} = \tilde{D} - B^T((\Sigma_0)^{\geq i})_{\mathcal{D}_0, \mathcal{D}}^{-1}B$ for appropriate sub matrices $\tilde{B}$ and $\tilde{D}$ of $(\Sigma_0)^{\geq i}_{\mathcal{D}_0, \mathcal{D}}$. Hence, there exists $N_2'$ such that

$$
\left(\frac{|\tilde{S}_{\mathcal{D}}^{\geq i}|}{|\tilde{S}_{\mathcal{D}_0}^{\geq i}|}\right)^{\frac{1}{2}} = \frac{1}{|R_{\tilde{S}_{\mathcal{D}}^{\geq i}}^{-1}|^{1/2}} \leq \frac{1}{\lambda \min \left(\frac{R_{\tilde{S}_{\mathcal{D}}^{\geq i}}^{-1}}{(\Sigma_0)^{\geq i}_{\mathcal{D}_0, \mathcal{D}}^{-1}} - K \frac{4\sqrt{\log p}}{n^2} \right)} \leq \left(\frac{\nu(\varrho)}{\epsilon_0 / 2}\right)^{1/2} \quad \text{for } n > N_2'.
$$

Since $pa_i(\mathcal{D}) \subset pa_i(\mathcal{D}_0)$, we get $\tilde{S}_{i|pa_i(\mathcal{D}_0)} \leq \tilde{S}_{i|pa_i(\mathcal{D})}$.

Let $K_1 = 4c'/\epsilon_0^2$. By (14) and $2 < c_i(\mathcal{D}), c_i(\mathcal{D}_0) < c$, it follows that there exists $N_2'''$ such that for $n \geq N_2'''$, we get

$$
PR_i(\mathcal{D}, \mathcal{D}_0)
\leq \left(M' \sqrt{\frac{2n}{\delta_2(\varrho)}} \right)^{1/2} 2^{c(n - 1)} \cdot \left(\frac{1}{(\Sigma_0)^{\geq i}_i|pa_i(\mathcal{D})|} + K_1 \sqrt{\frac{\log p}{n}}\right)^{n - \frac{2 - 3}{2}}
\leq \left\{ \exp \left\{ \frac{2\log M' + d \log \left(\frac{2d}{\epsilon_0^2}\right) + 4c \log n}{n - 1} + \frac{2\log \nu_i(\mathcal{D}_0) - \nu_i(\mathcal{D})}{n - 1} \right\} \right\}^{n - 1}
\times \left(1 + \frac{1}{(\Sigma_0)^{\geq i}_{i|pa_i(\mathcal{D})}} - 2K_1 \sqrt{\frac{\log p}{n}}\right)^{-n - 1}
$$

It then follows from Proposition 5.2 in [Cao et al., 2019a] that,

$$
PR_i(\mathcal{D}, \mathcal{D}_0)
$$

20
It follows by (29) and the above arguments that

\[
PR_i(\mathcal{D}, \mathcal{R}_0) \leq \left( \frac{1 + \frac{c^2}{\pi^2} s_n^2}{1 + \frac{c^2}{4} s_n^2} \right)^\frac{n-1}{2} \leq \exp \left\{ -\left( \frac{\frac{c^2}{\pi^2} s_n^2}{1 + \frac{c^2}{4} s_n^2} \right) \frac{n-1}{2} \right\}
\]

for \( n \geq \max(N'_2, N''_2, N'''_2) \). Since there exist a \((L_0)_{ji}\) such that \( s_n^2 \leq (L_0)_{ji}^2 \leq \frac{1}{\tilde{c}_0} \left( \frac{(L_0)_{ji}^2}{(D_0)_{ij}} \right) \leq \frac{(\Omega)_{ji}^2}{\tilde{c}_0} \leq \frac{1}{\tilde{c}_0} \) and \( c^2 s_n^2 \leq 1 \), it follows that there exists \( N_2 = \max(N'_2, N''_2, N'''_2) \) such that for \( n \geq N_2 \), such that

\[
PR_i(\mathcal{D}, \mathcal{R}_0) \leq \left( \frac{1 + \frac{c^2}{\pi^2} s_n^2}{1 + \frac{c^2}{4} s_n^2} \right)^\frac{n-1}{2} \leq \exp \left\{ -\left( \frac{\frac{c^2}{\pi^2} s_n^2}{1 + \frac{c^2}{4} s_n^2} \right) \frac{n-1}{2} \right\}
\]

\[
\leq e^{-\frac{1}{\pi^2} \frac{c^2}{\pi^2} s_n^2 (n-1)/2} \leq p^{-\frac{2n+1}{n+1}}. \tag{31}
\]

The last inequality follows from \( \frac{\log p}{n^2} \to 0 \), as \( n \to \infty \).

\textbf{Lemma 6.3.} If \( pa_i(\mathcal{D}) \) is not necessarily a superset or a subset of \( pa_i(\mathcal{R}_0) \), i.e. \( pa_i(\mathcal{R}_0) \neq pa_i(\mathcal{D}) \), \( pa_i(\mathcal{R}_0) \nsubseteq pa_i(\mathcal{D}) \), and \( pa_i(\mathcal{R}_0) \nsubseteq pa_i(\mathcal{D}) \), then there exists \( N_3 \) (not depending on \( i \) or \( \mathcal{D} \)) such that for \( n \geq N_3 \) we have \( PR_i(\mathcal{D}, \mathcal{R}_0) \leq (2p)^{-\frac{n+1}{n+1}} \nu_i(\mathcal{D}) \).

\textbf{Proof of Lemma 6.3.} Next consider the scenario when \( pa_i(\mathcal{D}) \) is not necessarily a superset or a subset of \( pa_i(\mathcal{R}_0) \), i.e. \( pa_i(\mathcal{R}_0) \neq pa_i(\mathcal{D}) \), \( pa_i(\mathcal{R}_0) \nsubseteq pa_i(\mathcal{D}) \), and \( pa_i(\mathcal{R}_0) \nsubseteq pa_i(\mathcal{D}) \). Let \( \mathcal{D}^* \) be an arbitrary DAG...
with \( pa_i(\mathcal{D}^*) = pa_i(\mathcal{D}) \cap pa_i(\mathcal{D}_0) \). Immediately we get \( pa_i(\mathcal{D}^*) \subset pa_i(\mathcal{D}_0) \) and \( pa_i(\mathcal{D}^*) \subset pa_i(\mathcal{D}) \). It follows from (14) that

\[
PR_i(\mathcal{D}, \mathcal{D}_0) \leq M \left( \frac{\delta_2}{\delta_1} \right)^{\frac{d}{2}} n^{2c} \left( \sqrt{\frac{\delta_2}{n}} \frac{\nu_i(\mathcal{D}) - \nu_i(\mathcal{D}^*)}{\nu_i(\mathcal{D}_0)} \right) \left( \frac{\hat{S}_{|pa_i(\mathcal{D})}}{\hat{S}_{|pa_i(\mathcal{D}^*)}} \right)^{n + 4(\mathcal{D}^*) - 2} \times \left( \frac{\delta_2}{\delta_1} \right)^{\frac{d}{2}} n^{2c} \left( \sqrt{\frac{\delta_2}{n}} \frac{\nu_i(\mathcal{D}) - \nu_i(\mathcal{D}_0)}{\nu_i(\mathcal{D})} \right) \left( \frac{\hat{S}_{|pa_i(\mathcal{D})}}{\hat{S}_{|pa_i(\mathcal{D})}} \right)^{n + 4(\mathcal{D}_0) - 2}
\]

\[
\leq PR_i(\mathcal{D}, \mathcal{D}^*) \times PR_i(\mathcal{D}^*, \mathcal{D}_0).
\]  

(32)

Note that \( pa_i(\mathcal{D}^*) \subset pa_i(\mathcal{D}) \). It follows from (28) that

\[
PR_i(\mathcal{D}, \mathcal{D}^*) \leq (2p)^{-\frac{2p}{\pi} (\nu_i(\mathcal{D}) - \nu_i(\mathcal{D}^*))}, \text{ for any constant } \kappa > 1 \text{ and } n \geq N_1.
\]  

(33)

Following from (31) and the fact that \( pa_i(\mathcal{D}^*) \subset pa_i(\mathcal{D}_0) \), we have

\[
PR_i(\mathcal{D}, \mathcal{D}^*) \leq p^{-\frac{2n_1 d}{n}}, \text{ for } n \geq N_2.
\]  

(34)

By (32) and \( \nu_i(\mathcal{D}^*) < d \), we get

\[
PR_i(\mathcal{D}, \mathcal{D}^*) \leq (2p)^{-\frac{2n_1}{n} (\nu_i(\mathcal{D}) - \nu_i(\mathcal{D}^*))} p^{-\frac{2n_1 d}{n}} < (2p)^{-\frac{2n_1}{n} \nu_i(\mathcal{D})}, \text{ for } n > N_3 = \max\{N_1, N_2\}.
\]  

(35)

The proof of Theorem (4.1) immediately follows after these three lemmas. For any \( \mathcal{D} \neq \mathcal{D}_0 \), there exists at least one \( 1 \leq i \leq p - 1 \), such that \( pa_i(\mathcal{D}) \neq \mathcal{D}_0 \). It follows from Lemmas 6.1, 6.2 and 6.3 that, for large enough \( n \geq N_3 \), under the true variable indicator \( \gamma_0 \),

\[
\max_{\mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma_0, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} \to 0, \text{ as } n \to \infty.
\]  

(36)

\[\square\]
Proof of Theorem 4.2. Now for the fixed \( \mathcal{D} \) case, it follows from (8) and model (7) that

\[
\frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}|Y, X)} = \frac{\pi(\gamma|\mathcal{D})}{\pi(\gamma_0|\mathcal{D})} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X_\gamma^T X_\gamma)^{-1} Y \right) \right\}
\]

\[
= \exp \left\{ \frac{-a1^T \gamma + b_\gamma^T G_\gamma}{-a1^T \gamma_0 + b_\gamma^T G_\gamma_0} \right\} \det \left( \tau^2 X_\gamma^T X_\gamma + I_\gamma \right) \frac{-1/2}{\det \left( \tau^2 X_{\gamma_0}^T X_{\gamma_0} + I_{\gamma_0} \right) \frac{-1/2}{\det \left( \tau^2 X_{\gamma_0}^T X_{\gamma_0} + I_{\gamma_0} \right) \frac{-1/2}{\det \left( \tau^2 X_{\gamma_0}^T X_{\gamma_0} + I_{\gamma_0} \right) \frac{-1/2}} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X_{\gamma_0}^T X_{\gamma_0})^{-1} Y \right) \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X_{\gamma_0}^T X_{\gamma_0})^{-1} Y \right) \right\}.
\]

For any model \( \gamma \) presenting the variable space, denote \( Q_\gamma = \det \left( \tau^2 X_\gamma^T X_\gamma + I_\gamma \right)^{-1/2}, P_\gamma = X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T \),

\[ R_\gamma^* = Y^T (I_n + \tau^2 (X_\gamma X_\gamma^T))^{-1} Y \text{ and } R_\gamma = Y^T (I_n - P_\gamma) Y. \]

Our method of proving variable selection consistency involves utilizing properties of \( R_\gamma \) and approximating \( R_\gamma^* \) and \( R_{\gamma_0}^* \) with \( R_\gamma \) and \( R_{\gamma_0} \) respectively. Using the Woodbury matrix identity, we have

\[ R_{\gamma_0}^* = Y^T (I_n + \tau^2 (X_{\gamma_0} X_{\gamma_0}^T))^{-1} Y = Y^T \left( I_n - X_{\gamma_0} (I_n/\tau^2 + X_{\gamma_0} X_{\gamma_0})^{-1} X_{\gamma_0}^T \right) Y. \]

Note that for \( 1 \leq i \leq p \),

\[ R_{\gamma_0}^* = Y^T (I_n + \tau^2 (X_{\gamma_0} X_{\gamma_0}^T))^{-1} Y = Y^T \left( I_n - X_{\gamma_0} (I_n/\tau^2 + X_{\gamma_0} X_{\gamma_0})^{-1} X_{\gamma_0}^T \right) Y \]

and \( R_{\gamma_0} = Y^T \left( I_n - X_\gamma (X_\gamma X_\gamma^T)^{-1} X_\gamma^T \right) Y. \) It follows that

\[ R_{\gamma_0}^* - R_{\gamma_0} \geq 0 \quad (38) \]

and

\[
R_{\gamma_0}^* - R_{\gamma_0} = Y^T X_{\gamma_0} (X_{\gamma_0}^T X_{\gamma_0})^{-\frac{1}{2}} \left( I_n - (I_n + (X_{\gamma_0}^T X_{\gamma_0})^{-1/\tau^2})^{-1} \right) (X_{\gamma_0}^T X_{\gamma_0})^{-\frac{1}{2}} X_{\gamma_0} Y \leq \frac{1}{1 + n\epsilon_0 \tau^2/2} Y^T P_{\gamma_0} Y. \quad (39)
\]
Note that \( \frac{R_{\gamma 0}}{\sigma^2} \sim \chi^2_{n-|\gamma_0|} \) and \( \frac{Y^T P_{\gamma 0} Y}{\sigma^2} \sim \chi^2_{|\gamma_0|} \left( \frac{\beta_0^T X_{\gamma 0}^T X_{\gamma 0} \beta_0}{\sigma^2} \right) \). Here we denote \( \chi^2_m \) as the centered chi-squared distribution with degrees of freedom \( m > 0 \) and \( \chi^2_m(\lambda) \) as the noncentral chi-squared distribution with noncentral parameter \( \lambda \). It follows from Lemma 4.1 and 4.2 in [Cao, Khare, and Ghosh, 2019c] that

\[
P \left[ \frac{R_{\gamma 0}}{\sigma^2} - (n - |\gamma_0|) \right] > \sqrt{(n - |\gamma_0|) \log p} \leq 2p^{-\frac{\epsilon_0}{2}} \to 0, \quad \text{as } n \to \infty,
\]

and

\[
P \left[ \frac{Y^T P_{\gamma 0} Y}{\sigma^2} - \left( |\gamma_0| + \frac{\beta_0^T X_{\gamma 0}^T X_{\gamma 0} \beta_0}{\sigma^2} \right) > n \log p - |\gamma_0| - \frac{\beta_0^T X_{\gamma 0}^T X_{\gamma 0} \beta_0}{\sigma^2} \right]
\leq \exp \left\{ - \frac{|\gamma_0|}{2} \left\{ \frac{np}{\sigma^2} - \log \left( 1 + \frac{n \log p}{\sigma^2} \right) \right\} \right\}
\leq \exp \left\{ - \frac{|\gamma_0|}{4} \left\{ \frac{\log p}{1 + \frac{\epsilon_0}{2 \sigma^2} \frac{\beta_0^T \beta_0}{\sigma^2}} \right\} \right\} \leq p^{-c|\gamma_0|} \to 0, \quad \text{as } n \to \infty. \tag{41}
\]

Further note that,

\[
R_{\gamma} - R_\gamma = Y^T X_{\gamma} (X_{\gamma}^T X_{\gamma})^{-1/2} \left( I_n - (I_n + (X_{\gamma}^T X_{\gamma})^{-1/2})^{-1} \right) (X_{\gamma}^T X_{\gamma})^{-1/2} X_{\gamma}^T Y
\geq \frac{\epsilon_0}{\epsilon_0 + 2n \sigma^2} Y^T P_{\gamma} Y. \tag{42}
\]

In the case when all the active elements of the true model \( \gamma_0 \) are contained in model \( \gamma \), it follows that \( \frac{R_{\gamma 0} - R_{\gamma}}{\sigma^2} \sim \chi^2_{|\gamma| - |\gamma_0|} \). Again, by Lemma 4.1 in [Cao, Khare, and Ghosh, 2019c], it follows that

\[
P \left[ \frac{Y^T (P_\gamma - P_{\gamma 0}) Y}{\sigma^2} - (|\gamma| - |\gamma_0|) \right] > \sqrt{(|\gamma| - |\gamma_0|) \log p} \leq 2p^{-\frac{\epsilon_0}{2}} \to 0, \tag{43}
\]

and

\[
P \left[ \frac{R_{\gamma 0} - R_{\gamma}}{\sigma^2} - (|\gamma| - |\gamma_0|) \right] > \sqrt{(|\gamma| - |\gamma_0|) \log p} \leq 2p^{-\frac{\epsilon_0}{2}} \to 0, \tag{44}
\]

as \( n \to \infty \). Hence, by (37), (39), (41), (44) and \( R^*_\gamma - R_\gamma \geq 0 \), we have

\[
\frac{\exp \left\{ - \frac{1}{2 \sigma^2} \left( Y^T (I + \sigma^2 X_{\gamma}^T X_{\gamma})^{-1} Y \right) \right\}}{\exp \left\{ - \frac{1}{2 \sigma^2} \left( Y^T (I + \sigma^2 X_{\gamma 0}^T X_{\gamma 0})^{-1} Y \right) \right\}}
\]
Therefore, it follows from Assumption 5, \( \gamma \geq \gamma_0 \), (37) and (45) that, for large enough \( n \geq N_4 \),

\[
\frac{\pi (\gamma, \mathcal{D}|Y, X)}{\pi (\gamma_0, \mathcal{D}|Y, X)} \leq \frac{\exp \left( -a_1 T \gamma + b T G \gamma \right)}{\exp (a_1 T \gamma_0 + b_0 T G \gamma_0)} \left( Q_{\gamma_0} \right) \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \sigma^2 T^2 X^T \gamma) \right)^{-1} Y \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \sigma^2 T^2 X^T \gamma_0) \right)^{-1} Y \right\} \frac{1}{2\sigma^2} \left( |\gamma| - |\gamma_0| + \sqrt{(|\gamma| - |\gamma_0|) \log p + \frac{1}{1 + \eta_0 \sigma^2/2} n \log p} \right) \}
\]

(46)

Next, when \( \gamma \subset \gamma_0 \), let \( Z \) be a standard normal distribution. When \( \gamma \subset \gamma_0 \), it follows from Lemma L.1 in [Cao, Khare, and Ghosh, 2019c] and the relation between noncentral chi-squared and normal distribution that,

\[
P \left( \frac{R^*_{\gamma} - R^*_{\gamma_0}}{\sigma^2} < 4|\gamma_0| \log p \right) < P \left( (Z - \sqrt{\lambda})^2 < 4|\gamma_0| \log p \right) < e^{-\frac{n \eta_0 \sigma^2}{4\pi^2}} \to 0, \text{ as } n \to \infty.
\]

(47)

where \( \delta = \frac{1}{2} \min_{j \in \gamma_0} |\beta_0 j| \) and \( \lambda = \frac{\beta^T \gamma \left( X^T_{\gamma_0} P_{X_{\gamma_0}} X_{\gamma_0} \right) \beta_0}{\sigma^2} > \frac{n \eta_0 \delta^2}{\sigma^2} \). it again follows from (37), (39), (47) and \( R^*_{\gamma} - R_{\gamma} \geq 0 \), with probability tending to 1, we have

\[
\exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \sigma^2 T^2 X^T \gamma) \right)^{-1} Y \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \sigma^2 T^2 X^T \gamma_0) \right)^{-1} Y \right\} \frac{1}{2\sigma^2} \left( |\gamma| - |\gamma_0| + \sqrt{(|\gamma| - |\gamma_0|) \log p + \frac{1}{1 + \eta_0 \sigma^2/2} n \log p} \right) = \exp \left\{ -\frac{1}{2\sigma^2} \left( R^*_{\gamma} - R^*_{\gamma_0} \right) \right\}
\]

25
Theorem (4.1) immediately follows after (46), (49) and (51). For any $\gamma$, next consider the scenario when $\gamma \subset \gamma_0$, (37) and (48) that, for large enough $n \geq N_5$, with probability tending to 1,

$$\pi (\gamma, \mathcal{D}|Y, X) \leq \exp \left\{ -\frac{1}{2\sigma^2} \left( R_\gamma - \left( R_{\gamma_0} + \frac{1}{1 + n\epsilon_0\tau^2/2} Y^T P_{\gamma_0} Y \right) \right) \right\} \leq \exp \left\{ -4|\gamma_0| \log p + \frac{1}{1 + n\epsilon_0\tau^2/2} n \log p \right\}.$$

(48)

Therefore, it follows from $\tau^2 \sim \sqrt{\log p}$, $a \sim \log p$, $\gamma \subset \gamma_0$, (37) and (48) that, for large enough $n \geq N_5$, with probability tending to 1,

$$\frac{\pi (\gamma, \mathcal{D}|Y, X)}{\pi (\gamma_0, \mathcal{D}|Y, X)} = \exp \left\{ -a1^T \gamma + b\gamma^T G\gamma \right\} Q_{\gamma_0} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X^T X) -1 \right) Y \right\} \leq \exp \left\{ \frac{1}{2} |\gamma_0| \log \left( n \sqrt{\log pe_0} \right) \right\} \times \exp \left\{ -4|\gamma_0| \log p + \frac{1}{1 + n\epsilon_0\tau^2/2} n \log p \right\} \leq \exp \left\{ -2|\gamma_0| \log p \right\}.$$

(49)

Next, consider the scenario when $\gamma \not\subset \gamma_0$ and $\gamma \not\subset \gamma_0$. Denote $\gamma' = \gamma \cap \gamma_0$. It follows from (37) that

$$\frac{\pi (\gamma, \mathcal{D}|Y, X)}{\pi (\gamma_0, \mathcal{D}|Y, X)} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X^T X) -1 \right) Y \right\} \leq \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X^T X) -1 \right) Y \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X^T X) -1 \right) Y \right\} \times \pi (\gamma' | \mathcal{D}) Q_{\gamma_0} \exp \left\{ -\frac{1}{2\sigma^2} \left( Y^T (I + \tau^2 X^T X) -1 \right) Y \right\} \pi (\gamma | \mathcal{D}) Q_{\gamma_0}.$$

(50)

Since $\gamma' \subset \gamma$ and $\gamma' \subset \gamma_0$, following the same arguments leading up to (46) and (49), we have for large enough $n > \max \{N_4, N_5\}$, with probability tending to 1,

$$\frac{\pi (\gamma, \mathcal{D}|Y, X)}{\pi (\gamma_0, \mathcal{D}|Y, X)} \leq \exp \left\{ -\frac{\alpha_1}{\kappa} (|\gamma| - |\gamma'|) \log p \right\} \exp \left\{ -2|\gamma_0| \log p \right\} \leq \exp \left\{ -\frac{\alpha_1}{\kappa} (|\gamma| - |\gamma'|) \log p - 2|\gamma_0| \log p \right\}.$$

(51)

Theorem (4.1) immediately follows after (46), (49) and (51). For any $\gamma \neq \gamma_0$, for large enough $n >
Next, it follows from Lemmas 6.1 - 6.3 that if we restrict to $E_\alpha$, then for large enough constant $N > \max\{N_4, N_5\}$, we have

$$\max_{(\gamma, \mathcal{D}) \neq (\gamma_0, \mathcal{D}_0)} \pi(\gamma, \mathcal{D}|Y, X) \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (52)$$

**Proof of Theorem 4.4.** We now move on to the proof of Theorem 4.4. We have

$$1 - \frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}|Y, X)} = \sum_{(\gamma, \mathcal{D}) \neq (\gamma_0, \mathcal{D}_0)} \frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)}$$

$$= \sum_{\gamma \neq \gamma_0} \frac{\pi(\gamma, \mathcal{D}_0|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} + \sum_{\mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma_0, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} + \sum_{\gamma \neq \gamma_0, \mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)}. \quad (53)$$

Note that it follows from the proof of Theorem 4.2 that for large enough constant $N > \max\{N_4, N_5\}$,

$$\sum_{\gamma \neq \gamma_0} \frac{\pi(\gamma, \mathcal{D}_0|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)}$$

$$\leq \sum_{\gamma \neq \gamma_0} \frac{\pi(\gamma, \mathcal{D}_0|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} + \sum_{\mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma_0, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} + \sum_{\gamma \neq \gamma_0, \mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)}$$

$$\leq \sum_{|\gamma_0| = 1} \frac{|\gamma_0|}{|\gamma|} \exp \left\{ -\frac{\alpha_1}{\kappa} |\gamma_0| \log p \right\} + \sum_{|\gamma_0| + 1} R_n \left( \frac{p - |\gamma_0|}{|\gamma| - |\gamma_0|} \right) \exp \left\{ -2(|\gamma| - |\gamma_0|) \log p \right\}$$

$$+ \sum_{|\gamma_0| = 1} R_n \left( \frac{p}{|\gamma|} \right) \exp \left\{ -\frac{\alpha_1}{\kappa} (|\gamma| - |\gamma'|) \log p - 2|\gamma_0| \log p \right\}.$$ 

Further note that the upper bound of the binomial coefficient satisfies $\binom{p}{k} \leq p^k$, for any $1 \leq k \leq p$. It follows that when $\alpha_1 > 2\kappa$ for some $\kappa > 1$,

$$\sum_{\gamma \neq \gamma_0} \frac{\pi(\gamma, \mathcal{D}_0|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (54)$$

Next, it follows from Lemmas 6.1 - 6.3 that if we restrict to $E_n$, then for large enough constant $N > N_3$, we have

$$\sum_{\mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma_0, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)} \leq \sum_{j=1}^{p-1} \sum_{\mathcal{D} \neq \mathcal{D}_0} \frac{\pi(\gamma_0, \mathcal{D}_0|Y, X)}{\pi(\gamma_0, \mathcal{D}_0|Y, X)}$$

27
Proof of Corollary 4.1. Note that with the extra layer of inverse gamma distribution on $\sigma^2$, by integrating out $\sigma^2$ in the proof of Lemma 3.1, the (marginal) joint posterior distribution is given by

\[
\pi(\gamma, \mathcal{D}|Y, X) = \int \pi(Y|\gamma, \beta) \prod_{i=1}^{n} \pi(X_i|(L, D)) \frac{\pi_{U, \alpha(\mathcal{D})}(L, D)}{\pi(\gamma, \mathcal{D}|Y, X)} \pi(\mathcal{D}|\gamma, \beta) \pi(\sigma^2)d\beta d(L, D)d(\sigma^2) \\
\times \alpha(\gamma|\mathcal{D}) \pi(\mathcal{D}) \frac{z_{\mathcal{D}}(U + X^T \gamma, n + \alpha(\mathcal{D}))}{z_{\mathcal{D}}(U, \alpha(\mathcal{D}))} Q_\gamma \\
\times \left( \frac{1}{2} \left( Y^T (I + \tau^2 X^T \gamma)^{-1} Y \right) + b_0 \right)^{-\frac{1}{2} + a_0},
\]

where $Q_\gamma = \det \left( \tau^2 X^T X + I \right)^{-\frac{1}{2}}$. The proofs for Lemma 6.1 to Lemma 6.3 will go through with the new
posterior. For the variable selection consistency, it follows from (58) that,

$$\frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}|Y, X)} = \frac{\exp\left(-a_1^T \gamma + b\gamma^T G\gamma\right)}{\exp\left(-a_1^T \gamma_0 + b\gamma_0^T G\gamma_0\right)} \frac{Q_\gamma}{Q_{\gamma_0}} \times \left(\frac{1}{2} \left(Y^T (I + \tau^2 X^T X_\gamma)^{-1} Y\right) + b_0\right)^{-\left(\frac{a}{2} + a_0\right)}$$

$$= \frac{\exp\left(-a_1^T \gamma + b\gamma^T G\gamma\right)}{\exp\left(-a_1^T \gamma_0 + b\gamma_0^T G\gamma_0\right)} \frac{Q_\gamma}{Q_{\gamma_0}} \left(\frac{R^*_\gamma + 2b_0}{R^*_\gamma_0 + 2b_0}\right)^{-\left(\frac{a}{2} + a_0\right)}. \quad (59)$$

It follows from the arguments leading up to (49) and $1 + x \leq e^x$ that when $\gamma \supset \gamma_0$, we have

$$\frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}|Y, X)} \leq \exp\left\{-a(|\gamma| - |\gamma_0|) + bR^2_\gamma\right\} \frac{1}{n\tau^2 e_0} \left(\frac{R^*_\gamma - R^*_\gamma_0}{R^*_\gamma_0 + 2a_0}\right)^{\frac{a}{2} + b_0}$$

$$\leq \exp\left\{-a(|\gamma| - |\gamma_0|) + bR^2_\gamma\right\} \exp\left\{\frac{1}{2} |\gamma_0| \log \left(n\sqrt{\log pe_0}\right)\right\}$$

$$\times \exp\left\{\left(\frac{a}{2} + a_0\right) \left(|\gamma| - |\gamma_0| + \sqrt{(|\gamma| - |\gamma_0|) \log p} + \frac{1}{n + 4e_0} \log p + 2b_0\right)\right\}$$

$$\leq \exp\left\{\frac{-a_1}{K} (|\gamma| - |\gamma_0|) \log p\right\}. \quad (60)$$

Next, when $\gamma \subset \gamma_0$, it follows by the arguments leading up to (49) and $1 - x \leq e^{-x}$ that,

$$\frac{\pi(\gamma, \mathcal{D}|Y, X)}{\pi(\gamma_0, \mathcal{D}|Y, X)} = \frac{\exp\left(-a_1^T \gamma + b\gamma^T G\gamma\right)}{\exp\left(-a_1^T \gamma_0 + b\gamma_0^T G\gamma_0\right)} \left(\frac{R^*_\gamma - R^*_\gamma_0}{R^*_\gamma_0 + 2a_0}\right)^{\frac{a}{2} + a_0}$$

$$\leq \exp \{a|\gamma_0|\} \exp\left\{\frac{1}{2} |\gamma_0| \log \left(n\sqrt{\log pe_0}\right)\right\}$$

$$\times \exp\left\{\left(\frac{a}{2} + a_0\right) \left(-4|\gamma_0| \log p + \frac{1}{n + 4e_0} \log p\right)\right\}$$

$$\leq \exp \{-2|\gamma_0| \log p\}. \quad (61)$$

When $\gamma \notin \gamma_0$ and $\gamma \notin \gamma_0$, the exact same results as the previous case without the inverse gamma prior can be obtained by following the arguments leading up to (51). Similarly, Corollary 4.1 can be acquired from the same arguments leading up to (57).
7 Discussion

In this paper, we work in a regression setting, where 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 regression coefficients, independent Bernoulli priors for each edge in the DAG, and a MRF prior linking the variable indicators to the graph structure. Under high-dimensional settings and standard regularity assumptions, when the underlying variance $\sigma^2$ is available, we establish both posterior ratio consistency and strong selection consistency for estimating the variable and the graph for the covariates jointly. When the underlying response variance is unknown and an appropriate inverse gamma prior is placed on $\sigma^2$, we also establish the joint selection consistency under the same regularity conditions. Finally, through simulation studies, we demonstrate that the model studied in this paper can outperform existing state-of-the-art methods in selecting network-structured predictors including both penalized likelihood and Bayesian approaches in several settings. For future studies, we intend to explore other types of priors over the graph space and on the regression coefficients to see if the consistency and better simulation performance can both be achieved under weakened assumptions.

References


