The Annals of Statistics
2019, Vol. 47, No. 1, 319–348
https://doi.org/10.1214/18-AOS1689

POSTERIOR GRAPH SELECTION AND ESTIMATION CONSISTENCY FOR HIGH-DIMENSIONAL BAYESIAN DAG MODELS

BY XUAN CAO, KSHITIJ KHARE AND MALAY GHOSH

University of Florida

Covariance estimation and selection for high-dimensional multivariate datasets is a fundamental problem in modern statistics. Gaussian directed acyclic graph (DAG) models are a popular class of models used for this purpose. 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 variables. A variety of priors have been developed in recent years for Bayesian inference in DAG models, yet crucial convergence and sparsity selection properties for these models have not been thoroughly investigated. Most of these priors are adaptations/generalizations of the Wishart distribution in the DAG context. In this paper, we consider a flexible and general class of these "DAG-Wishart" priors with multiple shape parameters. Under mild regularity assumptions, we establish strong graph selection consistency and establish posterior convergence rates for estimation when the number of variables p is allowed to grow at an appropriate subexponential rate with the sample size n.

1. Introduction. One of the major challenges in modern day statistics is to formulate models and develop inferential procedures to understand the complex multivariate relationships present in high-dimensional datasets, where the number of variables is much larger than the number of samples. The covariance matrix, denoted by Σ, is one of the most fundamental objects that quantifies relationships between variables in multivariate datasets. 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 parameter 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 Σ), concentration graph models (sparsity in $\Omega = \Sigma^{-1}$) and directed acyclic graph (DAG) models (sparsity in the Cholesky parameter of Ω).

In this paper, we focus on Gaussian DAG models. In particular, suppose we have i.i.d. observations $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ from a p-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. Let $\Omega = LD^{-1}L^T$ be the modified Cholesky

Received May 2017; revised January 2018.

MSC2010 subject classifications. Primary 62F15; secondary 62G20.

Key words and phrases. Posterior consistency, high-dimensional data, Bayesian DAG models, covariance estimation, graph selection.

decomposition of the inverse covariance matrix $\Omega = \Sigma^{-1}$, that is, L is a lower triangular matrix with unit diagonal entries and D is a diagonal matrix with positive diagonal entries. For a DAG model, this normal distribution is assumed to be Markov with respect to a given directed acyclic graph $\mathscr{D}$ with vertices $\{1, 2, \dots, p\}$ (edges directed from larger to smaller vertices). This is equivalent to saying that $L_{ij} = 0$ whenever $\mathscr{D}$ does not have a directed edge from i to j (these concepts are discussed in detail in Section 2). Hence, a Gaussian DAG model restricts Σ (and Ω) to a lower- dimensional space by imposing sparsity constraints encoded in $\mathscr{D}$ on L.

On the frequentist side, a variety of penalized likelihood methods for sparse estimation of L exist in the literature; see [2, 11, 14, 20, 22, 23, 27]. Some of these methods, such as those in [20, 27], constrain the sparsity pattern in L to be banded, whereas others, such as those in [11, 14, 23], put no constraints on the sparsity pattern. Most of the above methods derive asymptotic estimation and model selection consistency properties for the resulting estimator in a an appropriate high-dimensional regime; see Section 7 for more details. On the Bayesian side, the first class of priors on the restricted space of covariance matrices corresponding to a Gaussian DAG model was initially developed in [10, 24]. As pointed out in [5], the priors in [10] can be considered as analogs of the $\mathcal{G}$-Wishart distribution for concentration graph models (inducing sparsity in Ω). In fact, for the special case of perfect DAGs, the priors in [10] are same as the $\mathcal{G}$-Wishart priors. As with the $\mathcal{G}$-Wishart priors, the priors in [10] have a single shape parameter. Letac and Massam [16] introduced a flexible class of priors with multiple shape parameters which facilitate differential shrinkage in high-dimensional settings. However, these priors are defined only for perfect DAG models. Recently, Ben-David et al. [5] introduce a class of DAG-Wishart distributions with multiple shape parameters. This class of distributions is defined for arbitrary DAG models, and is identical to the Letac–Massam IW_{P_G} priors for the special case of perfect DAG models. Thus, this class of DAG-Wishart distributions offers a flexible framework for Bayesian inference in Gaussian DAG models, and generalizes previous Wishart-based priors for DAG models.

The priors above are specified for a known DAG $\mathscr{D}$, and provide a Bayesian approach for estimating the covariance matrix. However, if the underlying DAG is not known and needs to be selected, one can easily extend this framework by specifying a prior on the space of DAGs, and looking at the posterior probabilities of the DAGs given the data. Such an approach was used in the context of concentration graph models in [4]. The utility of this Bayesian approach in substantially improving finite sample graph selection performance (as compared to existing penalized likelihood methods) has been demonstrated in [5]. We discuss and demonstrate this further in Sections 7 and 8.

Despite the developments in Bayesian methods for analyzing Gaussian DAG models, a comprehensive evaluation of the high-dimensional consistency properties of these methods has not been undertaken to the best of our knowledge. Assuming the data comes from a “true” DAG model, two aspects of the asymptotic behavior of the posterior are of interest: (a) assigning high posterior probability to the "true" underlying graph (graph selection consistency), and (b) estimating the "true" covariance matrix accurately (estimation consistency).

Gaussian concentration graph models, which induce sparsity in the inverse covariance matrix Ω, are a related but markedly different class of models as compared to Gaussian DAG models. The two classes of models intersect only at perfect DAGs, which are equivalent to decomposable concentration graph models. In the context of concentration graph models, high-dimensional posterior estimation consistency has been explored in recent work [3, 4, 26]. In [3, 26], estimation consistency is established for the decomposable concentration graph models when the underlying concentration graph is known, and the number of variables p is allowed to increase at an appropriate subexponential rate relative to the sample size n. Banerjee and Ghosal [4] get rid of the assumption of decomposability and do not assume the true concentration graph is known. They use independent Laplace priors for the off-diagonal entries of the inverse covariance matrix, and use independent Bernoulli priors for the edges of the concentration graph. In this framework, estimation consistency is established in [4] under suitable regularity assumptions when $\sqrt{(p+s)\log p/n} \to 0$ (s denotes the total number of nonzero off-diagonal entries in the "true" inverse covariance matrix). The authors do not address model selection consistency, but provide high-dimensional Laplace approximations for the marginal posterior probabilities for the graphs, along with a proof of the validity of these approximations.

In this paper, our goal is to explore both model selection and estimation consistency in a high-dimensional setting for Gaussian DAG models. In particular, we consider a hierarchical Gaussian DAG model with DAG-Wishart priors on the covariance matrix and independent Bernoulli priors for each edge in the DAG. Under standard regularity assumptions, which include letting p increase at an appropriate subexponential rate with n, we establish *posterior ratio consistency* (Theorem 4.1), that is, the ratio of the maximum marginal posterior probability assigned to a "nontrue" DAG to the posterior probability assigned to the "true" DAG converges to zero in probability under the true model. In particular, this implies that the true DAG will be the mode of the posterior DAG distribution with probability tending to 1 as $n \to \infty$. An almost sure version of posterior ratio consistency is established in Theorem 4.2. Next, under the additional assumption that the prior over DAGs is restricted to graphs with edge size less than an appropriate function of the sample size n, we show *strong graph selection consistency* (Theorem 4.3) and establish a posterior convergence rate for estimation of the inverse covariance matrix (Theorem E.1 in the Supplementary Material [7]). Strong graph selection consistency implies that under the true model, the posterior probability of the true graph converges in probability to 1 as $n \to \infty$. As pointed out in Remark 2, the assumption of restricting the prior over models with appropriately bounded parameter size has been used in [17] for regression models, and in [4] for concentration graph models.

Narisetty and He [17] establish strong model selection consistency of high-dimensional regression models with spike and slab priors. While there are some connections between our model and the one in [17] since the entries of L can be interpreted as appropriate regression coefficients, there are fundamental differences between the two models and the corresponding analyses. A detailed explanation of this is provided in Remark 1.

In recent work, Altamore et al. [1] develop a class of objective nonlocal priors for Gaussian DAG models. This class of priors is structurally different from the DAG Wishart priors of [5], and we also investigate posterior model selection consistency under these nonlocal priors. In fact, we show under almost identical assumptions to the DAG Wishart setting that under the true model, the posterior probability of the true graph converges in probability to 1 as $n \to \infty$ (Theorem 6.1). Another recent paper [8] tackles the problem of covariate-adjusted DAG selection, that is, estimating a sparse DAG based covariance matrix in the presence of covariates. Establishing consistency in this more complex setup is beyond the scope of our paper, and will be an excellent topic for future research.

The rest of the paper is structured as follows. Section 2 provides background material from graph theory and Gaussian DAG models. In Section 3, we provide the hierarchical Bayesian DAG model. Graph selection consistency results are stated in Section 4, and the proofs are provided in Section 5. In Section 6, we establish graph selection consistency for nonlocal priors. A detailed discussion and comparison of the Bayesian approach of [5] and existing penalized likelihood approaches is undertaken in Section 7. In Section 8, we use simulation experiments to illustrate the posterior ratio consistency result, and demonstrate the benefits of the Bayesian approach for graph selection vis-a-vis existing penalized likelihood approaches.

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) $\mathscr{D} = (V, E)$ consists of the vertex set $V = \{1, \dots, p\}$ and an edge set E such that there is no directed path starting and ending at the same vertex. As in [5], we will without loss of generality assume a parent ordering, where that all the edges are directed from larger vertices to smaller vertices. The set of parents of i, denoted by $\mathrm{pa}_i(\mathscr{D})$, is the collection of all vertices which are larger than i and share an edge with i. Similarly, the set of children of i, denoted by $\mathrm{chi}_i(\mathscr{D})$, is the collection of all vertices which are smaller than i and 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 which obey the directed Markov property with respect to a DAG $\mathscr{D}$. In particular, if $\boldsymbol{y} = (y_1, \dots, y_p)^T \sim N_p(0, \Sigma)$ and $N_p(0, \Sigma) \in \mathscr{N}_{\mathscr{D}}$, then $y_i \perp \boldsymbol{y}_{\{i+1,\dots,p\} \setminus \mathrm{pa}_i(\mathscr{D})} | \boldsymbol{y}_{\mathrm{pa}_i(\mathscr{D})}$ 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 (see, e.g., [19]). 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 $i \notin \mathrm{pa}_j(\mathscr{D})$. In other words, the structure of the DAG $\mathscr{D}$ is reflected in the Cholesky factor of the inverse covariance matrix. In light of this, it is often more convenient to reparametrize in terms of the Cholesky parameter of the inverse covariance matrix as follows.

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 $i \notin \mathrm{pa}_j(\mathscr{D})$, and 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 $(D, L) \in \Theta_{\mathscr{D}}$ 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 the directed edges of $\mathscr{D}$ by undirected ones. A DAG $\mathscr{D}$ is said to be perfect if the parents of all vertices are adjacent. An undirected graph is called decomposable if it has no induced cycle of length $n \geq 4$, excluding the loops. It is known that if $\mathscr{D}$ is a perfect directed acyclic graph (DAG), then $\mathscr{D}^u$ is a decomposable graph. Conversely, given an undirected decomposable graph, one can always direct the edges so that the resulting graph is a perfect DAG. This fact can be used to show that the class of normal distributions satisfying the directed Markov property with respect to $\mathscr{D}$ (DAG models, sparsity in L) is identical to the class of normal distributions satisfying the undirected Markov property with respect to $\mathscr{D}^u$ (concentration graph models, sparsity in Ω) *if and only if* $\mathscr{D}$ is a perfect DAG (see [18]).

2.2. *DAG-Wishart distribution.* In this section, we specify the multiple shape parameter DAG-Wishart distributions introduced in [5]. First, we provide required notation. Given a directed graph $\mathscr{D} = (V, E)$, with $V = \{1, \dots, p\}$, and a $p \times p$ matrix A, denote the column vectors $A^{>}_{\mathscr{D}.i} = (A_{ij})_{j\in\mathrm{pa}_i(\mathscr{D})}$ and $A^{\geq}_{\mathscr{D}.i} = (A_{ii}, (A^{>}_{\mathscr{D}.i})^T)^T$. Also, let $A^{>i}_{\mathscr{D}} = (A_{kj})_{k,j\in\mathrm{pa}_i(\mathscr{D})}$,

$$A^{\geq i}_{\mathscr{D}} = \begin{bmatrix} A_{ii} & (A^{>}_{\mathscr{D}.i})^T \\ A^{>}_{\mathscr{D}.i} & A^{>i}_{\mathscr{D}} \end{bmatrix}.$$

In particular, $A^{\geq}_{\mathscr{D}.p} = A^{\geq p}_{\mathscr{D}} = A_{pp}$.

The DAG-Wishart distributions in [5] corresponding to a DAG $\mathscr{D}$ are defined on the Cholesky space $\Theta_{\mathscr{D}}$. Given a positive definite matrix U and a p-dimensional vector $\boldsymbol{\alpha}(\mathscr{D})$, the (unnormalized) density of the DAG-Wishart distribution on $\Theta_{\mathscr{D}}$ is given by

$$\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(LD^{-1}L^T\right)U\right)\right\}\prod_{i=1}^{p} D_{ii}^{-\frac{\alpha_i(\mathscr{D})}{2}}, \tag{2.1}$$

for every $(D, L) \in \Theta_{\mathscr{D}}$. Let $\nu_i(\mathscr{D}) = |\mathrm{pa}_i(\mathscr{D})| = |\{j : j > i, (j, i) \in E(\mathscr{D})\}|$. If $\alpha_i(\mathscr{D}) - \nu_i(\mathscr{D}) > 2$, for all $1 \leq i \leq p$, the density in (2.1) can be normalized to a probability density, and the normalizing constant is given by

$$\begin{aligned} &z_{\mathscr{D}}\left(U, \boldsymbol{\alpha}(\mathscr{D})\right) \\ &\quad = \prod_{i=1}^{p} \frac{\Gamma(\frac{\alpha_i(\mathscr{D})}{2} - \frac{\nu_i(\mathscr{D})}{2} - 1)2^{\frac{\alpha_i(\mathscr{D})}{2}-1}(\sqrt{\pi})^{\nu_i(\mathscr{D})}\det(U_{\mathscr{D}}^{>i})^{\frac{\alpha_i(\mathscr{D})}{2}-\frac{\nu_i(\mathscr{D})}{2}-\frac{3}{2}}}{\det(U_{\mathscr{D}}^{\geq i})^{\frac{\alpha_i(\mathscr{D})}{2}-\frac{\nu_i(\mathscr{D})}{2}-1}}. \end{aligned} \tag{2.2}$$

In this case, we define the DAG-Wishart density $\pi_{U,\boldsymbol{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}$ on the Cholesky space $\Theta_{\mathscr{D}}$ by

$$\pi_{U,\boldsymbol{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}(D, L) = \frac{1}{z_{\mathscr{D}}(U, \boldsymbol{\alpha}(\mathscr{D}))}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left(LD^{-1}L^T\right)U\right)\right\}\prod_{i=1}^{p} D_{ii}^{-\frac{\alpha_i(\mathscr{D})}{2}}$$

for every $(D, L) \in \Theta_{\mathscr{D}}$. 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 differential shrinkage of the variables in high-dimensional settings.

The class of densities $\pi_{U,\boldsymbol{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}$ form 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 $\pi_{U,\boldsymbol{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ are independent, identically distributed $N_p(\mathbf{0}, (L^T)^{-1}DL^{-1})$ random vectors, then the posterior distribution of (D, L) is $\pi_{\tilde{U},\tilde{\boldsymbol{\alpha}}(\mathscr{D})}^{\Theta_{\mathscr{D}}}$, where $S = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Y}_i\boldsymbol{Y}_i^T$ denotes the sample covariance matrix, $\tilde{U} = U + nS$, and $\tilde{\boldsymbol{\alpha}}(\mathscr{D}) = (n + \alpha_1(\mathscr{D}), \ldots, n + \alpha_p(\mathscr{D}))$.

3. Model specification. Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n \in \mathbb{R}^p$ be the observed data. The class of DAG-Wishart distributions in Section 2 can be used for Bayesian covariance estimation and DAG selection through the following hierarchical model:

$$\begin{aligned} \boldsymbol{Y}|\left((D, L), \mathscr{D}\right) &\sim N_p\left(\mathbf{0}, \left(LD^{-1}L^T\right)^{-1}\right), \\ (D, L)|\mathscr{D} &\sim \pi_{U,\boldsymbol{\alpha}(\mathscr{D})}^{\Theta_{\mathscr{D}}}, \\ \pi(\mathscr{D}) &= \prod_{i=1}^{p-1} q^{\nu_i(\mathscr{D})}(1-q)^{p-i-\nu_i(\mathscr{D})}. \end{aligned} \tag{3.1}$$

The prior density for DAGs above corresponds to an Erdős–Renyi type of distribution on the space of DAGs, where each directed edge is present with probability q independently of the other edges. In particular, similar to [4], define $\gamma_{ij} = \mathbb{I}\{(i,j) \in E(\mathscr{D})\}$, $1 \le i < j \le p$ to be the edge indicator. Let γ_{ij}, $1 \le i < j < p$ be independent identically distributed Bernoulli(q) random variables. It follows that

$$\pi(\mathscr{D}) = \prod_{(i,j):1\le i<j\le p} q^{\gamma_{ij}}(1-q)^{1-\gamma_{ij}} = \prod_{i=1}^{p-1} q^{\nu_i(\mathscr{D})}(1-q)^{p-i-\nu_i(\mathscr{D})}.$$

The model in (3.1) has three hyperparameters: the scale matrix U (positive definite), the shape parameter vector $\boldsymbol{\alpha}(\mathscr{D})$ and the edge probability q.

The hierarchical model in (3.1) can be used to estimate a DAG as follows. By (3.1) and Bayes' rule, the (marginal) posterior DAG probabilities are given by

$$\begin{aligned}
&\pi(\mathscr{D}|\boldsymbol{Y}) \\
&\quad = \int_{\Theta_{\mathscr{D}}} \frac{\pi(\boldsymbol{Y}|\mathscr{D},(L,D))\pi^{\Theta_{\mathscr{D}}}_{U,\boldsymbol{\alpha}(\mathscr{D})}((L,D))\pi(\mathscr{D})}{\pi(\boldsymbol{Y})}\,dL\,dD \\
&\quad = \frac{\pi(\mathscr{D})}{\pi(\boldsymbol{Y})}\int_{\Theta_{\mathscr{D}}} \pi\bigl(\boldsymbol{Y}|\mathscr{D},(L,D)\bigr)\pi\bigl((L,D)|\mathscr{D}\bigr)\,dL\,dD \qquad (3.2)\\
&\quad = \frac{\pi(\mathscr{D})}{\pi(\boldsymbol{Y})}\int_{\Theta_{\mathscr{D}}} \frac{\exp(-\frac{1}{2}\operatorname{tr}(LD^{-1}L^T(U+nS)))\prod_{i=1}^{p} D_{ii}^{-(n+\frac{\alpha_i(\mathscr{D})}{2})}}{z_{\mathscr{D}}(U,\boldsymbol{\alpha}(\mathscr{D}))}\,dL\,dD \\
&\quad = \frac{\pi(\mathscr{D})}{\pi(\boldsymbol{Y})(\sqrt{2\pi})^n}\frac{z_{\mathscr{D}}(U+nS,n+\boldsymbol{\alpha}(\mathscr{D}))}{z_{\mathscr{D}}(U,\boldsymbol{\alpha}(\mathscr{D}))}.
\end{aligned}$$

Hence, the marginal posterior density $\pi(\mathscr{D}|\boldsymbol{Y})$ is available in closed form [up to the multiplicative constant $\pi(\boldsymbol{Y})$]. In particular, these posterior probabilities can be used to select a DAG by computing the posterior mode defined by

$$\hat{\mathscr{D}} = \arg\max_{\mathscr{D}} \pi(\mathscr{D}|\boldsymbol{Y}). \qquad (3.3)$$

4. DAG selection consistency: Main results. In this section, we will explore the high-dimensional asymptotic properties of the Bayesian DAG selection approach specified in Section 3. For this purpose, we will work in a setting where the dimension $p = p_n$ of the data vectors, and the edge probabilities $q = q_n$ vary with the sample size n. We assume that the data is actually being generated from a true model which can be specified as follows. Let $\boldsymbol{Y}_1^n, \boldsymbol{Y}_2^n, \ldots, \boldsymbol{Y}_n^n$ be independent and identically distributed p_n-variate Gaussian vectors with mean 0 and covariance matrix $\Sigma_0^n = (\Omega_0^n)^{-1}$. Let $\Omega_0^n = L_0^n(D_0^n)^{-1}(L_0^n)^T$ be the modified Cholesky decomposition of Ω_0^n. Let $\mathscr{D}_0^n$ be the true underlying DAG, that is, $L_0^n \in \mathscr{L}_{\mathscr{D}_0^n}$. Denote d_n as the maximum number of nonzero entries in any column of L_0^n,

$s_n = \min_{1\le j\le p, i\in \mathrm{pa}_j(\mathscr{D}_0^n)} |(L_0^n)_{ji}|$. Let $\bar{P}$ and $\bar{E}$, respectively, denote the probability measure and expected value corresponding to the "true" Gaussian DAG model presented above.

In order to establish our asymptotic results, we need the following mild regularity assumptions. Each assumption below is followed by an interpretation/discussion. Note that for a symmetric $p \times p$ matrix $A = (A_{ij})_{1\le i,j\le p}$, let $\mathrm{eig}_1(A) \le \mathrm{eig}_2(A) \le \ldots \le \mathrm{eig}_p(A)$ denote the ordered eigenvalues of A.

ASSUMPTION 1. There exists $\varepsilon_{0,n} \le 1$, such that for every $n \ge 1$, $0 < \varepsilon_{0,n} \le \mathrm{eig}_1(\Omega_0^n) \le \mathrm{eig}_{p_n}(\Omega_0^n) \le \varepsilon_{0,n}^{-1}$, where $\frac{(\frac{\log p}{n})^{\frac{1}{2}-\frac{1}{2+k}}}{\varepsilon_{0,n}^4} \to 0$, as $n \to \infty$, for some $k > 0$.

This is a much weaker assumption for high-dimensional covariance asymptotics than, for example, [3, 4, 6, 9, 26]. Here, we allow the lower and upper bounds on the eigenvalues to depend on p and n.

ASSUMPTION 2. $d_n^{2+k}\sqrt{\frac{\log p_n}{n}} \to 0$, and $(\sqrt{\frac{\log p_n}{n}})^{\frac{k}{2(k+2)}} \log n \to 0$, as $n \to \infty$.

This assumption essentially states that the number of variables p_n has to grow slower than $e^{n/d_n^{4+2k}}$ (and also $e^{n/(\log n)^{2+k}}$). Again, similar assumptions are common in high-dimensional covariance asymptotics; see, for example, [3, 4, 6, 26].

ASSUMPTION 3. Let $q_n = e^{-\eta_n n}$ in (3.1), where $\eta_n = d_n(\frac{\log p_n}{n})^{\frac{1/2}{1+k/2}}$. Hence, $q_n \to 0$, as $n \to \infty$.

This assumption provides the rate at which the edge probability q_n needs to approach zero. A similar assumption can be found in [17] in the context of linear regression. This can be interpreted as a priori penalizing graphs with a large number of edges.

ASSUMPTION 4. $\frac{\eta_n d_n}{\varepsilon_{0,n} s_n^2} \to 0$ as $n \to \infty$.

Recall that s_n is the smallest (in absolute value) nonzero off-diagonal entry in L_0^n, and can be interpreted as the "signal size." Hence, this assumption provides a lower bound for the signal size that is needed for establishing consistency.

ASSUMPTION 5. For every $n \ge 1$, the hyperparameters for the DAG-Wishart prior $\pi^{\Theta_{\mathscr{D}_n}}_{U_n,\boldsymbol{\alpha}(\mathscr{D}_n)}$ in (3.1) satisfy (i) $2 < \alpha_i(\mathscr{D}_n) - \nu_i(\mathscr{D}_n) < c$ for every $\mathscr{D}_n$ and $1 \le i \le p_n$, and (ii) $0 < \delta_1 \le \mathrm{eig}_1(U_n) \le \mathrm{eig}_{p_n}(U_n) \le \delta_2 < \infty$. Here, c, δ_1, δ_2 are constants not depending on n.

This assumption provides mild restrictions on the hyperparameters for the DAG-Wishart distribution. The assumption $2 < \alpha_i(\mathscr{D}) - \nu_i(\mathscr{D})$ 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. While the authors in [5] do not specifically discuss hyperparameter choice, they do provide some recommendations in their experiments section. For the shape parameters, they recommend $\alpha_i(\mathscr{D}_n) = c\nu_i(\mathscr{D}_n) + b$. They mostly use $c = 1$ in which case Assumption 5 is satisfied. The also use $c \in (2.5, 3.5)$ in some examples, in which case Assumption 5 is not satisfied. Also, they choose the scale matrix to be a constant multiple of the identity matrix, which clearly satisfies Assumption 5.

For the rest of this paper, p_n, Ω_0^n, Σ_0^n, L_0^n, D_0^n, $\mathscr{D}_0^n$, $\hat{\mathscr{D}}^n$, $\mathscr{D}^n$, d_n, q_n, A_n will be denoted as p, Ω_0, Σ_0, L_0, D_0, $\mathscr{D}_0$, $\hat{\mathscr{D}}$, $\mathscr{D}$, d, q, A as needed for notational convenience and ease of exposition.

We now state and prove the main DAG selection consistency results. Our first result establishes what we call as posterior ratio consistency. This notion of consistency implies that the true DAG will be the mode of the posterior DAG distribution with probability tending to 1 as $n \to \infty$.

THEOREM 4.1 (Posterior ratio consistency). *Under Assumptions 1–5, the following holds*:

$$\max_{\mathscr{D} \neq \mathscr{D}_0} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} \xrightarrow{\bar{P}} 0 \qquad \text{as } n \to \infty.$$

A proof of this result is provided in Section 5. If one was interested in a point estimate of the underlying DAG using the Bayesian approach considered here, the most obvious choice would be the posterior mode $\hat{\mathscr{D}}$ defined in (3.3). From a frequentist point of view, it would be natural to inquire if we have model selection consistency, that is, if $\hat{\mathscr{D}}$ is a consistent estimate of $\mathscr{D}_0$. In fact, the model selection consistency of the posterior mode follows immediately from posterior ratio consistency established in Theorem 4.1, by noting that

$$\max_{\mathscr{D} \neq \mathscr{D}_0} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} < 1 \quad \Rightarrow \quad \hat{\mathscr{D}} = \mathscr{D}_0.$$

We state this result formally in the corollary below.

COROLLARY 4.1 (Model selection consistency for posterior mode). *Under Assumptions 1–5, the posterior mode $\hat{\mathscr{D}}$ is equal to the true DAG $\mathscr{D}_0$ with probability tending to* 1, *that is*,

$$\bar{P}(\hat{\mathscr{D}} = \mathscr{D}_0) \to 1 \qquad \text{as } n \to \infty.$$

If p is of a larger order than a positive power of n, then a stronger version of the posterior ratio consistency in Theorem 4.1 can be established.

THEOREM 4.2 (Almost sure posterior ratio consistency). *If $p/n^{\widetilde{k}} \to \infty$ for some $\widetilde{k} > 0$, then under Assumptions 1–5 the following holds*:

$$\max_{\mathscr{D} \neq \mathscr{D}_0} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} \to 0 \qquad \textit{almost surely } \bar{P},$$

as $n \to \infty$.

Next, we establish another stronger result (compared to Theorem 4.1) which implies that the posterior mass assigned to the true DAG $\mathscr{D}_0$ converges to 1 in probability (under the true model). Following [17], we refer to this notion of consistency as strong selection consistency. To establish this stronger notion of consistency, we restrict our analysis to DAGs with total number of edges bounded by an appropriate function of n (see also Remark 2).

THEOREM 4.3 (Strong selection consistency). *Under Assumptions 1–5, if we restrict only to DAGs with number of edges at most $\frac{1}{8}d(\frac{n}{\log p})^{\frac{1+k}{2+k}}$, the following holds*:

$$\pi(\mathscr{D}_0|\boldsymbol{Y}) \xrightarrow{\bar{P}} 1 \qquad \textit{as } n \to \infty.$$

REMARK 1. In the context of linear regression, Narisetty and He [17] consider the following hierarchical Bayesian model:

$$\begin{aligned}
&\boldsymbol{Y}|X, \boldsymbol{\beta}, \qquad \sigma^2 \sim N(X\boldsymbol{\beta}, \sigma^2 I),\\
&\beta_i|\sigma^2, \qquad Z_i = 0 \sim N(0, \sigma^2\tau_{0,n}^2),\\
&\beta_i|\sigma^2, \qquad Z_i = 1 \sim N(0, \sigma^2\tau_{1,n}^2),\\
&P(Z_i = 1) = 1 - P(Z_i = 0) = q_n,\\
&\sigma^2 \sim \text{Inverse-Gamma}(\alpha_1, \alpha_2).
\end{aligned}$$

In particular, they put an independent spike and slab prior on each linear regression coefficient (conditional on the variance parameter σ^2), and an inverse Gamma prior on the variance. Also, each regression coefficient is present in the model with a probability q_n. In this setting, the authors in [17] establish strong selection consistency for the regression coefficients (assuming the prior is constrained to leave out unrealistically large models). There are similarities between the models and the consistency analysis in [17] and this paper. Note that the off-diagonal entries in the ith column of L are the linear regression coefficients corresponding to fitting the ith variable against all variables with label greater than i, and in our model (3.1) each coefficient is present independently with a given probability q_n. Also, similar to [17], in terms of proving posterior consistency, we bound the ratio of posterior probabilities for a nontrue model and the true model by a "prior term" which is a power of $q_n/(1-q_n)$, and a "data term." The consistency proof is then a careful exercise in balancing these two terms against each other on a case-by-case basis. However, despite these similarities, there are some fundamental differences in the two models and the corresponding analysis. First, the DAG-Wishart prior does not in general correspond to assigning an independent spike and slab prior to each entry of L. The columns of L are independent of each other under this prior, but it introduces correlations among the entries in each column of L. Also, the DAG-Wishart prior introduces exact sparsity in L, which is not the case in [17] as $\tau_{0,n}^2$ is assumed to be strictly positive. Hence, it is structurally different than the prior in [17]. Second, the "design" matrices corresponding to the regression coefficients in each column of L are random (they are functions of the sample covariance matrix S) and are correlated with each other. In particular, this leads to major differences and further challenges in analyzing the ratio of posterior graph probabilities (a crucial step in establishing consistency).

REMARK 2. We would like to point out that posterior ratio consistency (Theorems 4.1 and 4.2) does not require any restriction on the maximum number of edges; this requirement is only needed for strong selection consistency (Theorem 4.3). Similar restrictions on the prior model size have been considered for establishing consistency properties in other contexts. For concentration graph models, Banerjee and Ghosal [4] use a hierarchical prior where each edge of the concentration graph is independently present with a given probability q. For establishing high-dimensional posterior convergence rates, they restrict the prior to graphs with total number of edges bounded by an appropriate fixed constant. A variation where the upper bound on the number of edges is a random variable with subexponential tails is also considered. For linear regression, Narisetty and He [17] too restrict the prior model size to an appropriate function of n (number of nonzero regression coefficients) for establishing strong selection consistency (when the variance parameter is random).

5. Proofs of Theorems 4.1, 4.2 and 4.3.

The proof of Theorems 4.1, 4.2 and 4.3 will be broken up into various steps. We begin by presenting a useful lemma that provides an upper bound for the ratio of posterior DAG probabilities.

LEMMA 5.1. *Under Assumption 5, for a large enough constant M and large enough n, the ratio of posterior probabilities of any DAG $\mathscr{D}$ and the true DAG $\mathscr{D}_0$*

satisfies

$$\frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} \le \prod_{i=1}^{p} M\left(\frac{\delta_2}{\delta_1}\right)^{\frac{d}{2}} n^{2c}\left(\sqrt{\frac{\delta_2}{n}}\frac{q}{1-q}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)} \frac{|\tilde{S}^{\ge i}_{\mathscr{D}_0}|^{\frac{1}{2}}}{|\tilde{S}^{\ge i}_{\mathscr{D}}|^{\frac{1}{2}}} \frac{(\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)})^{\frac{n+c_i(\mathscr{D}_0)-3}{2}}}{(\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})})^{\frac{n+c_i(\mathscr{D})-3}{2}}}$$

$$\triangleq \prod_{i=1}^{p} B_i(\mathscr{D}, \mathscr{D}_0),$$

where $c_i(\mathscr{D}) = \alpha_i(\mathscr{D}) - \nu_i(\mathscr{D}), c_i(\mathscr{D}_0) = \alpha_i(\mathscr{D}_0) - \nu_i(\mathscr{D}_0)$, $\tilde{S} = S + \frac{U}{n}$ *and* $\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})} = \tilde{S}_{ii} - (\tilde{S}^{>}_{\mathscr{D}\cdot i})^T(\tilde{S}^{>i}_{\mathscr{D}})^{-1}\tilde{S}^{>}_{\mathscr{D}\cdot i}$.

The proof of this lemma is provided in the Supplementary Material [7]. Our goal is to find an upper bound (independent of $\mathscr{D}$ and i) for $B_i(\mathscr{D}, \mathscr{D}_0)$, such that the upper bound converges to 0 as $n \to \infty$. By Lemma 5.1, this will be enough to establish Theorem 4.1. Before we undertake this goal, we present a proposition that will be useful in further analysis. Note that for any positive definite matrix A, and $M \subseteq \{1, 2\dots, p\} \setminus \{i\}$, we denote $A_{i|M} = A_{ii} - A_{iM}A_{MM}^{-1}A_{Mi}$.

PROPOSITION 5.2. *Given a DAG $\mathscr{D}$ with p vertices*:

(a) *If* $\mathrm{pa}_i(\mathscr{D}) \supseteq \mathrm{pa}_i(\mathscr{D}_0)$, *then* $(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D})} = (D_0)_{ii} = (\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)}$ *does not depend on $\mathscr{D}$.*

(b) *If* $\mathrm{pa}_i(\mathscr{D}) \subseteq \mathrm{pa}_i(\mathscr{D}_0)$, *then* $(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D})} - (\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)} \ge \varepsilon_{0,n}(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D}))s^2$, *where* $\varepsilon_{0,n} > 0$ *and* $s = \min_{j\in\mathrm{pa}_i(\mathscr{D}_0)} |(L_0)_{ji}|$.

The proof of this proposition is provided in the Supplementary Material [7]. Next, we show that in our setting, the sample and population covariance matrices are sufficiently close with high probability. It follows by Assumptions 1, 2, 5, Lemma A.3 of [6] and the Hanson–Wright inequality from [21] that there exists constants m_1, m_2 and δ depending on $\varepsilon_{0,n}$ only such that for $1 \le i, j \le p$, we have

$$\bar{P}(|S_{ij} - (\Sigma_0)_{ij}| \ge t) \le m_1 \exp\{-m_2 n(t\varepsilon_{0,n})^2\}, \qquad |t| \le \delta.$$

By the union-sum inequality, for a large enough c' such that $2 - m_2(c')^2/4 < 0$, we get that

$$\bar{P}\left(\|\tilde{S} - \Sigma_0\|_{\max} \ge c'\sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}}\right) \le m_1 p^{2-m_2(c')^2/4} \to 0. \tag{5.1}$$

Define the event C_n as

$$C_n = \left\{\|\tilde{S} - \Sigma_0\|_{\max} \ge c'\sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}}\right\}. \tag{5.2}$$

It follows from (5.1) and (5.2) that $\bar{P}(C_n) \to 0$ as $n \to \infty$.

We now analyze the behavior of $B_i(\mathscr{D}, \mathscr{D}_0)$ under different scenarios in a sequence of five lemmas (Lemmas 5.3–5.7). Recall that our goal is to find an upper bound (independent of $\mathscr{D}$ and i) for $B_i(\mathscr{D}, \mathscr{D}_0)$, such that the upper bound converges to 0 as $n \to \infty$. *For all these lemmas, we will restrict ourselves to the event C_n^c.*

LEMMA 5.3. *If $\mathrm{pa}_i(\mathscr{D}) \supset \mathrm{pa}_i(\mathscr{D}_0)$ and $\nu_i(\mathscr{D}) \le 3\nu_i(\mathscr{D}_0) + 2$, then there exists N_1 (not depending on i or $\mathscr{D}$) such that for $n \ge N_1$ we have $B_i(\mathscr{D}, \mathscr{D}_0) \le \varepsilon_{1,n}$, where $\varepsilon_{1,n} = 2e^{-\frac{\eta_n}{2}n}$.*

PROOF. Since $\mathrm{pa}_i(\mathscr{D}_0) \subset \mathrm{pa}_i(\mathscr{D})$, we can write $|\tilde{S}_{\mathscr{D}}^{\ge i}| = |\tilde{S}_{\mathscr{D}_0}^{\ge i}||R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}}|$. Here, $R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}}$ is the Schur complement of $\tilde{S}_{\mathscr{D}_0}^{\ge i}$, defined by

$$R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}} = D - B^T (\tilde{S}_{\mathscr{D}_0}^{\ge i})^{-1} B$$

for appropriate submatrices B and D of $\tilde{S}_{\mathscr{D}}^{\ge i}$. Since $\tilde{S}_{\mathscr{D}}^{\ge i} \ge (\frac{U}{n})_{\mathscr{D}}^{\ge i}$,[1] and $R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}}^{-1}$ is a principal submatrix of $(\tilde{S}_{\mathscr{D}}^{\ge i})^{-1}$, it follows from Assumption 5 that the largest eigenvalue of $R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}}^{-1}$ is bounded above by $\frac{n}{\delta_1}$. Therefore,

$$\left(\frac{|\tilde{S}_{\mathscr{D}_0}^{\ge i}|}{|\tilde{S}_{\mathscr{D}}^{\ge i}|}\right)^{\frac{1}{2}} = |R_{\tilde{S}_{\mathscr{D}_0}^{\ge i}}^{-1}|^{1/2} \le \left(\sqrt{\frac{n}{\delta_1}}\right)^{\nu_i(\mathscr{D}) - \nu_i(\mathscr{D}_0)}. \tag{5.3}$$

Since we are restricting ourselves to the event C_n^c, it follows by (5.1) that

$$\|\tilde{S}_{\mathscr{D}_0}^{\ge i} - (\Sigma_0)_{\mathscr{D}_0}^{\ge i}\|_{(2,2)} \le (\nu_i(\mathscr{D}_0) + 1)c' \sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}}.$$

Therefore,

$$\begin{aligned}
&\|(\tilde{S}_{\mathscr{D}_0}^{\ge i})^{-1} - ((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} \\
&\quad = \|(\tilde{S}_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} \|(\tilde{S}_{\mathscr{D}_0}^{\ge i})^{-1} - ((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} \|((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} \qquad (5.4) \\
&\quad \le \left(\|(\tilde{S}_{\mathscr{D}_0}^{\ge i})^{-1} - ((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} + \frac{1}{\varepsilon_{0,n}}\right)(\nu_i(\mathscr{D}_0) + 1)c' \sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}} \frac{1}{\varepsilon_{0,n}}.
\end{aligned}$$

By Assumptions 1, 2 and $d > 0$, we have

$$\frac{d\sqrt{\frac{\log p_n}{n}}}{\varepsilon_{0,n}^4} \to 0 \qquad \text{as } n \to \infty. \tag{5.5}$$

[1] For matrices A and B, we say $A \ge B$ if $A - B$ is positive semidefinite.

Hence, there exists N_1' such that for $n \geq N_1'$,

$$\frac{c'}{\varepsilon_{0,n}^2}(d+1)\sqrt{\frac{\log p}{n}} < \frac{1}{2} \quad \text{and} \quad 2\frac{c'}{\varepsilon_{0,n}^3}(d+1)\sqrt{\frac{\log p}{n}} < \frac{1}{\varepsilon_{0,n}}.$$

Since $\nu_i(\mathscr{D}_0) \leq d$, it follows by (5.4) and Assumption 1 that

$$\|(\tilde{S}_{\mathscr{D}_0}^{\geq i})^{-1}\|_{(2,2)} \leq \frac{2}{\varepsilon_{0,n}}$$

and

$$\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)}} = [(\tilde{S}_{\mathscr{D}_0}^{\geq i})^{-1}]_{ii} \leq \frac{2}{\varepsilon_{0,n}} \tag{5.6}$$

for $n \geq N_1'$. Since, $\mathrm{pa}_i(\mathscr{D}_0) \subset \mathrm{pa}_i(\mathscr{D})$, we get

$$\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)} \geq \tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}.$$

Let N_1'' be such that for $n \geq N_1''$, $q \leq \frac{\sqrt{\delta_1}}{2\sqrt{\delta_2}} \leq \frac{1}{2}$. Using $2 < c_i(\mathscr{D}), c_i(\mathscr{D}_0) < c$, (5.3), (5.6) and Lemma 5.1, we get

$$\begin{aligned} B_i(\mathscr{D},\mathscr{D}_0) &\leq M\left(\frac{\delta_2}{\delta_1}\right)^{\frac{d}{2}} n^{2c}\left(\sqrt{\frac{\delta_2}{\delta_1}}\frac{q}{1-q}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\left(\frac{2}{\varepsilon_{0,n}}\right)^{c} \\ &\quad \times \left(\frac{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)}}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}}\right)^{\frac{n+c-3}{2}} \qquad (5.7) \\ &\leq M\left(\frac{2}{\varepsilon_{0,n}}\right)^{c}\left(\frac{\delta_2}{\delta_1}\right)^{\frac{d}{2}} n^{2c}\left(\sqrt{\frac{\delta_2}{\delta_1}}2q\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\left(\frac{\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}}}{\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)}}}\right)^{\frac{n+c-3}{2}}, \end{aligned}$$

for $n \geq \max(N_1', N_1'')$. We would like to note that the arguments leading up to (5.7) only require the assumption $\mathrm{pa}_i(\mathscr{D}_0) \subset \mathrm{pa}_i(\mathscr{D})$. This observation enables us to use (5.7) in the proof of Lemmas 5.4 and 5.5.

By following exactly the same sequence of arguments leading up to (5.4), and replacing $\mathscr{D}$ by $\mathscr{D}_0$, we get

$$\begin{aligned} &\|(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\geq i})^{-1}\|_{(2,2)} \qquad (5.8) \\ &\quad \leq \left(\|(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\geq i})^{-1}\|_{(2,2)} + \frac{1}{\varepsilon_{0,n}}\right)(\nu_i(\mathscr{D})+1)c'\sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}}\frac{1}{\varepsilon_{0,n}}. \qquad (5.9) \end{aligned}$$

By (5.5), there exists N_1''' such that for $n \geq N_1'''$,

$$\frac{c'}{\varepsilon_{0,n}^2}(3d+3)\sqrt{\frac{\log p}{n}} < \frac{1}{2} \quad \text{and} \quad 2\frac{c'}{\varepsilon_{0,n}^3}(3d+3)\sqrt{\frac{\log p}{n}} < \frac{\varepsilon_{0,n}}{2}. \tag{5.10}$$

Note that by hypothesis $\nu_i(\mathscr{D}) + 1 \le 3\nu_i(\mathscr{D}_0) + 3 \le 3d + 3$. It follows from (5.8) that

$$\|(\tilde{S}_{\mathscr{D}}^{\ge i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\ge i})^{-1}\|_{(2,2)} \le 2\frac{c'}{\varepsilon_{0,n}^3}(3d+3)\sqrt{\frac{\log p}{n}} \tag{5.11}$$

for $n \ge N_1'''$. Using $\nu_i(\mathscr{D}) - \nu_i(\mathscr{D}_0) \ge 1$, (5.7), (5.11), Proposition 5.2(a) and the definition of q_n, it follows that for $n \ge \max(N_1', N_1'', N_1''')$,

$$\begin{aligned}
&B_i(\mathscr{D}, \mathscr{D}_0)\\
&\quad \le 2\tilde{M}\frac{1}{\varepsilon_{0,n}^c}\left(\frac{\delta_2}{\delta_1}\right)^{\frac{d}{2}} n^{2c} q \left(\frac{\|((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} + 2\frac{c'}{\varepsilon_{0,n}^3}(3d+3)\sqrt{\frac{\log p}{n}}}{\|((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} - 2\frac{c'}{\varepsilon_{0,n}^3}(3d+3)\sqrt{\frac{\log p}{n}}}\right)^{\frac{n-c+3}{2}}\\
&\quad \le 2\exp\left\{-d\left(\frac{\log p}{n}\right)^{\frac{1/2}{1+k/2}} n + d\log\left(\frac{\delta_2}{\delta_1}\right)\right.\\
&\qquad \left. + 2c\log n + \frac{c}{4}\log\left(\frac{1}{\varepsilon_{0,n}^4}\right) + \log\tilde{M}\right\}\\
&\qquad \times \left(1 + \frac{2\frac{c'}{\varepsilon_{0,n}^3}(3d+3)\sqrt{\frac{\log p}{n}}}{\|((\Sigma_0)_{\mathscr{D}_0}^{\ge i})^{-1}\|_{(2,2)} - \frac{\varepsilon_{0,n}}{2}}\right)^n\\
&\quad \le 2\exp\left\{-d\left(\frac{\log p}{n}\right)^{\frac{1/2}{1+k/2}} n + d\log\left(\frac{\delta_2}{\delta_1}\right)\right.\\
&\qquad \left. + 2c\log n + \frac{c}{4}\log\left(\frac{1}{\varepsilon_{0,n}^4}\right) + \log\tilde{M}\right\}\\
&\qquad \times \exp\left\{\frac{12c'}{\varepsilon_{0,n}^4}(d+1)\sqrt{n\log p}\right\},
\end{aligned}$$

where $\tilde{M} = M2^c\sqrt{\frac{\delta_2}{\delta_1}}$. Since $\eta_n = d(\frac{\log p}{n})^{\frac{1/2}{1+k/2}}$ has a strictly larger order than $\frac{d}{n}$, $\frac{\log n}{n}$, $\frac{\log(\frac{1}{\varepsilon_{0,n}^4})}{n}$ and $d\frac{\sqrt{\frac{\log p}{n}}}{\varepsilon_{0,n}^4}$[2] by Assumptions 1 and 2, it follows that there exists N_1'''' such that for $n \ge N_1''''$, the expression in the exponent is dominated by $-\frac{\eta_n}{2}$. It follows that

$$B_i(\mathscr{D}, \mathscr{D}_0) \le 2e^{-\frac{\eta_n}{2}n}$$

for $n \ge N_1 \overset{\Delta}{=} \max(N_1', N_1'', N_1''', N_1'''')$. □

[2] We say a_n is of a larger order than b_n if $\frac{b_n}{a_n} \to 0$ as $n \to \infty$.

LEMMA 5.4. *Assume* $\mathrm{pa}_i(\mathscr{D}) \supset \mathrm{pa}_i(\mathscr{D}_0)$, $\nu_i(\mathscr{D}) > 3\nu_i(\mathscr{D}_0) + 2$ *and* $\frac{1}{\varepsilon_{0,n}^2}(\nu_i(\mathscr{D})+1)\sqrt{\frac{\log p}{n}} \leq \frac{1}{2c'}$, *then there exists* N_2 (*not depending on* i *or* $\mathscr{D}$), *such that for* $n \geq N_2$, $B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_{2,n}$, *where* $\varepsilon_{2,n} = e^{-\eta_n n}$.

PROOF. By following exactly the same sequence of arguments leading up to (5.4), and replacing $\mathscr{D}$ by $\mathscr{D}_0$, we get

$$\begin{aligned}&\|(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\geq i})^{-1}\|_{(2,2)}\\ &\quad\leq \biggl(\|(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\geq i})^{-1}\|_{(2,2)} + \frac{1}{\varepsilon_{0,n}}\biggr)(\nu_i(\mathscr{D})+1)c'\sqrt{\frac{\log p}{n\varepsilon_{0,n}^2}}\frac{1}{\varepsilon_{0,n}}.\end{aligned}$$

Using $\frac{1}{\varepsilon_{0,n}^2}(\nu_i(\mathscr{D})+1)\sqrt{\frac{\log p}{n}} \leq \frac{1}{2c'}$, $\nu_i(\mathscr{D}_0) < \nu_i(\mathscr{D})$, (5.4) and (5.5), for large enough $n \geq N_2'$, we get

$$\|(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}}^{\geq i})^{-1}\|_{(2,2)} \leq \frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D})+1)\sqrt{\frac{\log p}{n}}, \tag{5.12}$$

$$\|(\tilde{S}_{\mathscr{D}_0}^{\geq i})^{-1} - ((\Sigma_0)_{\mathscr{D}_0}^{\geq i})^{-1}\|_{(2,2)} \leq \frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D}_0)+1)\sqrt{\frac{\log p}{n}} \tag{5.13}$$

and

$$\frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D}_0)+1)\sqrt{\frac{\log p}{n}} \leq \frac{\varepsilon_{0,n}}{2}. \tag{5.14}$$

Note that the arguments leading up to (5.7) only use $\mathrm{pa}_i(\mathscr{D}_0) \subset \mathrm{pa}_i(\mathscr{D})$. It follows from (5.7), Proposition 5.2, (5.12) and (5.13) that these exists N_2'' such that

$$\begin{aligned}&B_i(\mathscr{D}, \mathscr{D}_0)\\ &\quad\leq \exp\biggl\{d\log\biggl(\frac{\delta_2}{\delta_1}\biggr) + 2c\log n + \frac{c}{4}\log\biggl(\frac{1}{\varepsilon_{0,n}^4}\biggr) + \log\tilde{M}\biggr\}\biggl(2q\sqrt{\frac{\delta_1}{\delta_2}}\biggr)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\\ &\qquad\times\left(\frac{\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}}}{\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)}}}\right)^{\frac{n+c-3}{2}}\\ &\quad\leq \exp\biggl\{d\log\biggl(\frac{\delta_2}{\delta_1}\biggr) + 3c\log n\biggr\}\biggl(2q\sqrt{\frac{\delta_1}{\delta_2}}\biggr)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\\ &\qquad\times\left(\frac{\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)}} + \frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D})+1)\sqrt{\frac{\log p}{n}}}{\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)}} - \frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D}_0)+1)\sqrt{\frac{\log p}{n}}}\right)^{\frac{n+c-3}{2}}\end{aligned}$$

for $n \geq N_2''$. Note that $\nu_i(\mathscr{D}) > 3\nu_i(\mathscr{D}_0) + 2$ implies $\nu_i(\mathscr{D}) + \nu_i(\mathscr{D}_0) + 2 \leq 2(\nu_i(\mathscr{D}) - \nu_i(\mathscr{D}_0))$. It follows by Assumption 1, (5.14) and $q = q_n = e^{-\eta_n n}$ that

$$\begin{aligned}
&B_i(\mathscr{D}, \mathscr{D}_0)\\
&\quad\leq \exp\left\{d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\left(2q\sqrt{\frac{\delta_1}{\delta_2}}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\\
&\qquad\times\left(1 + \frac{\frac{2c'}{\varepsilon_{0,n}^3}(\nu_i(\mathscr{D}) + \nu_i(\mathscr{D}_0) + 2)\sqrt{\frac{\log p}{n}}}{\varepsilon_{0,n}/2}\right)^{\frac{n+c-3}{2}}\\
&\quad\leq \exp\left\{d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\left(2q\sqrt{\frac{\delta_1}{\delta_2}}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}\\
&\qquad\times \exp\left\{\frac{8c'}{\varepsilon_{0,n}^4}(\nu_i(\mathscr{D}) - \nu_i(\mathscr{D}_0))\sqrt{n \log p}\right\}\\
&\quad\leq \left(2\sqrt{\frac{\delta_1}{\delta_2}} \exp\left\{-\eta_n n + \frac{8c'}{\varepsilon_{0,n}^4}\sqrt{n \log p} + d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)}.
\end{aligned}$$

Since η_n has a strictly larger order than $\frac{\sqrt{\frac{\log p}{n}}}{\varepsilon_{0,n}^4}$, $\frac{d}{n}$ and $\frac{\log n}{n}$, there exists N_2 such that

$$B_i(\mathscr{D}, \mathscr{D}_0) \leq \left(e^{-\frac{\eta_n}{2} n}\right)^{\nu_i(\mathscr{D})-\nu_i(\mathscr{D}_0)} \leq e^{-\eta_n n}$$

for $n \geq N_2$. □

LEMMA 5.5. *If* $\mathrm{pa}_i(\mathscr{D}) \supset \mathrm{pa}_i(\mathscr{D}_0)$, $\nu_i(\mathscr{D}) > 3\nu_i(\mathscr{D}_0) + 2$ *and* $\frac{1}{\varepsilon_{0,n}^2}(\nu_i(\mathscr{D}) + 1)\sqrt{\frac{\log p}{n}} > \frac{1}{2c'}$, *then there exists* N_3 (*not depending on* i *or* $\mathscr{D}$), *such that for* $n \geq N_3$, $B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_{3,n}$, *where* $\varepsilon_{3,n} = (\frac{1}{\delta_1 n})^n$.

PROOF. Since $\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}} = [(\tilde{S}_{\mathscr{D}}^{\geq i})^{-1}]_{ii}$ and $\tilde{S}_{\mathscr{D}}^{\geq i} \geq (\frac{U}{n})_{\mathscr{D}}^{\geq i}$, we get

$$\frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}} \leq \frac{n}{\delta_1}.$$

By (5.4) and (5.5), there exists N_3' such that for $n \geq N_3'$,

$$\frac{c'}{\varepsilon_{0,n}^2}(d+1)\sqrt{\frac{\log p}{n}} < \frac{1}{2} \quad \text{and} \quad 2\frac{c'}{\varepsilon_{0,n}^3}(d+1)\sqrt{\frac{\log p}{n}} < \frac{\varepsilon_{0,n}}{2}.$$

Since $\nu_i(\mathscr{D}_0) \leq d$, it follows by (5.4) and Assumption 1 that

$$\|(\tilde{S}_{\mathscr{D}_0}^{\geq i})^{-1}\|_{(2,2)} \geq \frac{\varepsilon_{0,n}}{2} \quad \text{and} \quad \frac{1}{\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)}} = [(\tilde{S}_{\mathscr{D}_0}^{\geq i})^{-1}]_{ii} \geq \frac{\varepsilon_{0,n}}{2}$$

for $n \geq N_3'$. Note that by hypothesis, we have

$$\nu_i(\mathscr{D}) > \frac{\varepsilon_{0,n}^2}{2c'}\sqrt{\frac{n}{\log p}} - 1.$$

Since the arguments leading up to (5.7) only require $\mathrm{pa}_i(\mathscr{D}) \supset \mathrm{pa}_i(\mathscr{D}_0)$, using the above facts along with Assumption 2, there exists N_3'' such that for $n \geq N_3''$, we get

$$\begin{aligned}
&B_i(\mathscr{D}, \mathscr{D}_0)\\
&\quad \leq \exp\left\{d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\left(2q\sqrt{\frac{\delta_1}{\delta_2}}\right)^{\frac{\varepsilon_{0,n}^2}{2c'}\sqrt{\frac{n}{\log p}}-2d}\left(\frac{2n}{\delta_1 \varepsilon_{0,n}}\right)^{\frac{n+c-3}{2}}\\
&\quad \leq \exp\left\{d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\left(2q\sqrt{\frac{\delta_1}{\delta_2}}\right)^{\frac{\varepsilon_{0,n}^2}{2c'}\sqrt{\frac{n}{\log p}}-2d}\left(\frac{2n}{\delta_1 \varepsilon_{0,n}}\right)^{n}\\
&\quad \leq \left(\frac{n^2}{\varepsilon_{0,n}\delta_1}\right)^n (2q)^{\frac{\varepsilon_{0,n}^2}{2c'}\sqrt{\frac{n}{\log p}}-2d} \exp\left\{d \log\left(\frac{\delta_2}{\delta_1}\right) + 3c \log n\right\}\\
&\quad = \left(\frac{1}{\delta_1}\exp\left\{-\frac{\varepsilon_{0,n}^2}{2c'}\eta_n\sqrt{\frac{n}{\log p}} + 2\eta_n d + (2+3c)\log n + d \log\left(\frac{\delta_2}{\delta_1}\right)\right.\right.\\
&\qquad \left.\left. + \frac{1}{4}\log\left(\frac{1}{\varepsilon_{0,n}^4}\right)\right\}\right)^n.
\end{aligned}$$

By Assumption 1, we have $\frac{1}{\varepsilon_{0,n}^2} = o((\frac{\log p}{n})^{-\frac{1}{2}(\frac{1}{2}-\frac{1}{2+k})})$. Then, by (5.5), Assumptions 2 and 3, we obtain $\varepsilon_{0,n}^2 \eta_n \sqrt{\frac{n}{\log p}}$ has a larger order than $\eta_n d$, $\log n$ and $\log(\frac{1}{\varepsilon_{0,n}^4})$. It follows that there exists N_3 such that

$$\begin{aligned}
B_i(\mathscr{D}, \mathscr{D}_0) &\leq \left(\frac{1}{\delta_1}\exp\left\{-\frac{\varepsilon_{0,n}^2}{4c'}\eta_n\sqrt{\frac{n}{\log p}}\right\}\right)^n\\
&\leq \left(\frac{1}{\delta_1}\exp\{-\log n\}\right)^n\\
&= \left(\frac{1}{\delta_1 n}\right)^n
\end{aligned}$$

for $n \geq N_3$. □

LEMMA 5.6. *If* $\mathrm{pa}_i(\mathscr{D}) \subset \mathrm{pa}_i(\mathscr{D}_0)$, *then there exists* N_4 (*not depending on* i *or* $\mathscr{D}$), *such that for* $n \geq N_4$, $B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_{4,n}$, *where* $\varepsilon_{4,n} = e^{-d\eta_n n}$.

PROOF. Since $\mathrm{pa}_i(\mathscr{D}_0) \supset \mathrm{pa}_i(\mathscr{D})$, we can write $|\tilde{S}^{\geq i}_{\mathscr{D}_0}| = |\tilde{S}^{\geq i}_{\mathscr{D}}||R_{\tilde{S}^{\geq i}_{\mathscr{D}}}|$. Here, $R_{\tilde{S}^{\geq i}_{\mathscr{D}}}$ is the Schur complement of $\tilde{S}^{\geq i}_{\mathscr{D}}$, defined by

$$R_{\tilde{S}^{\geq i}_{\mathscr{D}}} = \tilde{D} - \tilde{B}^T \big(\tilde{S}^{\geq i}_{\mathscr{D}}\big)^{-1} \tilde{B}$$

for appropriate submatrices $\tilde{B}$ and $\tilde{D}$ of $\tilde{S}^{\geq i}_{\mathscr{D}_0}$. It follows by (5.4) that if restrict to C_n^c,

$$\big\|\big(\tilde{S}^{\geq i}_{\mathscr{D}_0}\big)^{-1} - \big((\Sigma_0)^{\geq i}_{\mathscr{D}_0}\big)^{-1}\big\|_{(2,2)} \leq \frac{4c'}{\varepsilon^3_{0,n}} d \sqrt{\frac{\log p}{n}},$$

for $n > N_4'$. It follows that

$$\big\|R^{-1}_{\tilde{S}^{\geq i}_{\mathscr{D}}} - R^{-1}_{(\Sigma_0)^{\geq i}_{\mathscr{D}}}\big\|_{(2,2)} \leq \frac{4c'}{\varepsilon^3_{0,n}} d \sqrt{\frac{\log p}{n}},$$

where $R_{(\Sigma_0)^{\geq i}_{\mathscr{D}}}$ represents the Schur complement of $(\Sigma_0)^{\geq i}_{\mathscr{D}}$ defined by

$$R_{(\Sigma_0)^{\geq i}_{\mathscr{D}}} = \bar{D} - \bar{B}^T \big((\Sigma_0)^{\geq i}_{\mathscr{D}}\big)^{-1} \bar{B}$$

for appropriate submatrices $\bar{B}$ and $\bar{D}$ of $(\Sigma_0)^{\geq i}_{\mathscr{D}_0}$. Let $\lambda_{min}(A)$ denote the smallest eigenvalue of a positive definite matrix A. By Assumptions 1 and 2, it follows that there exists N_4'' such that

$$\begin{aligned}\left(\frac{|\tilde{S}^{\geq i}_{\mathscr{D}_0}|}{|\tilde{S}^{\geq i}_{\mathscr{D}}|}\right)^{\frac{1}{2}} &= \frac{1}{|R^{-1}_{\tilde{S}^{\geq i}_{\mathscr{D}}}|^{1/2}} \leq \frac{1}{(\lambda_{\min}(R^{-1}_{(\Sigma_0)^{\geq i}_{\mathscr{D}}}) - K\frac{d}{\varepsilon^3_{0,n}}\sqrt{\frac{\log p}{n}})^{\frac{\nu_i(\mathscr{D}_0)-\nu_i(\mathscr{D})}{2}}} \\ &\leq \left(\frac{1}{\varepsilon_{0,n}/2}\right)^{\frac{\nu_i(\mathscr{D}_0)-\nu_i(\mathscr{D})}{2}} \qquad \text{for large enough } n\end{aligned}$$

for $n \geq N_4''$. Since $\mathrm{pa}_i(\mathscr{D}) \subset \mathrm{pa}_i(\mathscr{D}_0)$, we get

$$\tilde{S}_{i|\mathrm{pa}_i(\mathscr{D}_0)} \leq \tilde{S}_{i|\mathrm{pa}_i(\mathscr{D})}.$$

Let $K_1 = 4c'$. By Lemma 5.1 and Proposition 5.2, and $2 < c_i(\mathscr{D}), c_i(\mathscr{D}_0) < c$, it follows that there exists N_4''' such that for $n \geq N_4'''$, we get

$$\begin{aligned}&B_i(\mathscr{D}, \mathscr{D}_0) \\ &\quad\leq M\left(\frac{2}{\varepsilon_{0,n}}\right)^c \left(\frac{\delta_2}{\delta_1}\right)^{\frac{d}{2}} n^{2c}\left(\sqrt{\frac{2n}{\delta_2\varepsilon_{0,n}}}q^{-1}\right)^{\nu_i(\mathscr{D}_0)-\nu_i(\mathscr{D})} \\ &\qquad\times\left(\frac{\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D})}} + K_1\frac{d}{\varepsilon^3_{0,n}}\sqrt{\frac{\log p}{n}}}{\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)}} - K_1\frac{d}{\varepsilon^3_{0,n}}\sqrt{\frac{\log p}{n}}}\right)^{\frac{n+2-3}{2}}\end{aligned}$$

$$\begin{aligned}
&\le \Bigg(\exp\Bigg\{\frac{2d\log(\frac{\delta_2}{\delta_1}) + 6c\log n + (c+d)\log(\frac{1}{\varepsilon_{0,n}})}{n-1} \\
&\qquad + 8\eta_n\big(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D})\big)\Bigg\}\Bigg)^{\frac{n-1}{2}} \\
&\qquad \times \left(1 + \frac{(\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D}_0)}} - \frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D})}}) - 2K_1\frac{d}{\varepsilon_{0,n}^3}\sqrt{\frac{\log p}{n}}}{\frac{1}{(\Sigma_0)_{i|\mathrm{pa}_i(\mathscr{D})}} + K_1\frac{d}{\varepsilon_{0,n}^3}\sqrt{\frac{\log p}{n}}}\right)^{-\frac{n-1}{2}} \\
&\le \Bigg(\exp\Bigg\{\frac{2d\log(\frac{\delta_2}{\delta_1}) + 6c\log n + (c+d)\log(\frac{1}{\varepsilon_{0,n}})}{n-1} \\
&\qquad + 8\eta_n\big(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D})\big)\Bigg\}\Bigg)^{\frac{n-1}{2}} \\
&\qquad \times \left(1 + \frac{\varepsilon_{0,n}s_n^2(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D})) - 2K_1\frac{d}{\varepsilon_{0,n}^3}\sqrt{\frac{\log p}{n}}}{2/\varepsilon_{0,n}}\right)^{-\frac{n-1}{2}}.
\end{aligned}\tag{5.15}$$

Note that, by Assumptions 2, 3 and 4, $\frac{d\eta_n}{\varepsilon_{0,n}^2 s_n^2} \to 0$,

$$\frac{2d\log(\frac{\delta_2}{\delta_1}) + 6c\log n + (c+d)\log(\frac{1}{\varepsilon_{0,n}})}{(n-1)\varepsilon_{0,n}^2 s_n^2} \to 0,$$

and $\frac{\eta_n}{s_n^2} \to 0$ as $n \to \infty$. Since $e^x \le 1 + 2x$ for $x < \frac{1}{2}$, by Assumptions 1 and 4, there exists N_4'''' such that for $n \ge N_4''''$,

$$2K_1\frac{d}{\varepsilon_{0,n}^3}\sqrt{\frac{\log p}{n}} \le \varepsilon_{0,n}\eta \le \frac{\varepsilon_{0,n}s_n^2}{2}$$

and

$$\begin{aligned}
&\exp\Bigg\{\frac{2d\log(\frac{\delta_2}{\delta_1}) + 6c\log n + (c+d)\log(\frac{1}{\varepsilon_{0,n}})}{n-1} + 4\eta\big(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D})\big)\Bigg\} \\
&\quad \le 1 + 8\eta\big(\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D})\big) + \frac{4d\log(\frac{\delta_2}{\delta_1}) + 12c\log n + 2(c+d)\log(\frac{1}{\varepsilon_{0,n}})}{n-1} \\
&\quad \le 1 + \frac{\varepsilon_{0,n}^2 s_n^2}{8}.
\end{aligned}$$

It follows by (5.15) and the above observations that

$$B_i(\mathscr{D}, \mathscr{D}_0) \le \left(\frac{1 + \frac{\varepsilon_{0,n}^2}{8}s_n^2}{1 + \frac{\varepsilon_{0,n}^2}{4}s_n^2}\right)^{\frac{n-1}{2}}$$

for $n \geq \max(N_4', N_4'', N_4''', N_4'''')$. The last step follows by noting that $\nu_i(\mathscr{D}_0) - \nu_i(\mathscr{D}) \geq 1$. Since $\mathrm{pa}_i(\mathscr{D}_0)$ is nonempty by hypothesis, $\exists j \in \nu_i(\mathscr{D}_0)$ such that $|(L_0)_{ji}| \geq s_n$, which implies that $s_n^2 \leq (L_0)_{ji}^2 \leq \frac{1}{\varepsilon_{0,n}}(\frac{[(L_0)_{ji}]^2}{(D_0)_{ii}}) \leq \frac{(\Omega_0)_{jj}}{\varepsilon_{0,n}} \leq \frac{1}{\varepsilon_{0,n}^2}$ and $\varepsilon_{0,n}^2 s_n^2 \leq 1$. Hence, following from Assumption 4, there exists N_4 such that for $n \geq N_4$, we get

$$B_i(\mathscr{D}, \mathscr{D}_0) \leq \left(1 - \frac{\frac{\varepsilon_{0,n}^2}{8} s_n^2}{1 + \frac{\varepsilon_{0,n}^2}{4} s_n^2}\right)^{\frac{n-1}{2}} \leq \exp\left\{-\left(\frac{\frac{\varepsilon_{0,n}^2}{8} s_n^2}{1 + \frac{\varepsilon_{0,n}^2}{4} s_n^2}\right)\left(\frac{n-1}{2}\right)\right\}$$
$$\leq e^{-\frac{1}{10}\varepsilon_{0,n}^2 s_n^2 (\frac{n-1}{2})} \leq e^{-d\eta_n n}. \qquad \square$$

LEMMA 5.7. *Suppose* $\mathscr{D}$ *is such that* $\mathrm{pa}_i(\mathscr{D}_0) \neq \mathrm{pa}_i(\mathscr{D})$, $\mathrm{pa}_i(\mathscr{D}_0) \nsubseteq \mathrm{pa}_i(\mathscr{D})$, *and* $\mathrm{pa}_i(\mathscr{D}_0) \nsupseteq \mathrm{pa}_i(\mathscr{D})$, *then, for* $n \geq N_5$ (*not depending on* i *or* $\mathscr{D}$), $B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_{5,n}$, *where* $\varepsilon_{5,n} = \max(\varepsilon_{1,n}, \varepsilon_{2,n}, \varepsilon_{3,n})\varepsilon_{4,n}$.

The proof of this lemma is provided in the Supplementary Material [7]. With these lemmas in hand, Theorem 4.1 can be proved as follows. By Lemmas 5.3–5.7, if we restrict to the event C_n^c, and $\mathrm{pa}_i(\mathscr{D}) \neq \mathrm{pa}_i(\mathscr{D}_0)$, then $B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_n^*$ for every $n \geq \max(N_1, N_2, N_3, N_4)$, where $\varepsilon_n^* \triangleq \max\{\varepsilon_{1,n}, \varepsilon_{2,n}, \varepsilon_{3,n}, \varepsilon_{4,n}, \varepsilon_{5,n}\}$ converges to 0 as $n \to \infty$ (by Assumption 3). Note that if $\mathscr{D} \neq \mathscr{D}_0$, then there exists at least one i, such that $\mathrm{pa}_i(\mathscr{D}) \neq \mathrm{pa}_i(\mathscr{D}_0)$. It follows by Lemma 5.1, that if we restrict to C_n^c, then

$$\max_{\mathscr{D} \neq \mathscr{D}_0} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} \leq \max_{\mathscr{D} \neq \mathscr{D}_0} \prod_{i=1}^{p} B_i(\mathscr{D}, \mathscr{D}_0) \leq \varepsilon_n^* \tag{5.16}$$

for every $n \geq \max(N_1, N_2, N_3, N_4)$. By (5.1), $P(C_n^c) \to 1$ as $n \to \infty$. Theorem 4.1 follows immediately.

To prove Theorem 4.2, note that if $p/n^{\widetilde{k}} \to \infty$, then one can choose c' in (5.1) such that $m_2(c')^2/4 = 2 + 2/\widetilde{k}$. It follows that $P(C_n) \leq m_1/n^2$ for large enough n. The result follows by (5.16) and the Borel–Cantelli lemma.

We now move on to the proof of Theorem 4.3, and only consider DAGs with number of edges at most $h = \frac{1}{8}d(\frac{n}{\log p})^{\frac{1+k}{2+k}}$. By Lemmas 5.3–5.7, it follows that if we restrict to C_n^c, then

$$\frac{1 - \pi(\mathscr{D}_0|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} = \sum_{\mathscr{D} \neq \mathscr{D}_0, \mathscr{D} \text{ has atmost } h \text{ edges}} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})}$$
$$\leq \sum_{i=0}^{h} \binom{\binom{p}{2}}{i} \max_{\mathscr{D} \neq \mathscr{D}_0} \frac{\pi(\mathscr{D}|\boldsymbol{Y})}{\pi(\mathscr{D}_0|\boldsymbol{Y})} \tag{5.17}$$
$$\leq p^{3h} e^{-\frac{\eta_n n}{2}} = e^{3h \log p - \frac{\eta_n n}{2}} = e^{-\frac{1}{8}dn^{\frac{1+k}{2+k}} (\log p)^{\frac{1}{2+k}}}$$

for $n \geq \max(N_1, N_2, N_3, N_4)$. Theorem 4.3 follows immediately.

6. Results for nonlocal priors.

In [1], the authors present an alternative to the Wishart-based Bayesian framework for Gaussian DAG models by using nonlocal priors. Nonlocal priors were first introduced in [12] as densities that are identically zero whenever a model parameter is equal to its null value in the context of hypothesis testing (compared to local priors, which still preserve positive values at null parameter values). Nonlocal priors tend to discard spurious covariates faster as the sample size n grows, while preserving exponential learning rates to detect nonzero coefficients as indicated in [12]. These priors were further extended to Bayesian model selection problems in [13] by imposing nonlocal prior densities on a vector of regression coefficients. The nonlocal prior based approach for Gaussian DAG models proposed in [1], adapted to our notation and framework, can be described as follows:

$$\begin{aligned} \boldsymbol{Y} \mid ((D, L), \mathscr{D}) &\sim N_p(\mathbf{0}, (LD^{-1}L^T)^{-1}), \\ \pi_{NL}((D, L) \mid \mathscr{D}) &\propto \prod_{j=1}^{p} \frac{1}{D_{jj}} \Big(\prod_{i \in \mathrm{pa}_j(\mathscr{D})} L_{ij}^{2r} \Big), \\ \pi_{NL}(\mathscr{D}) &= \prod_{i=1}^{p-1} q^{\nu_i(\mathscr{D})} (1-q)^{p-i-\nu_i(\mathscr{D})}, \end{aligned} \tag{6.1}$$

where r is a fixed positive integer. Note that the prior on (D, L) is an improper objective prior. If $\max_{1 \leq j \leq p} \mathrm{pa}_j(\mathscr{D}) > n$, then this objective prior leads to an improper posterior for (D, L) as well. In such cases, the authors in [1] propose using fractional Bayes factors. However, for the purposes of proving strong selection consistency, similar to Theorem 4.3, we will restrict the prior on the space of DAGs to graphs whose total number of edges is appropriately bounded (leaving out unrealistically large models, in the terminology of [17]). This will ensure that the posterior impropriety issue never arises.

The next result establishes strong selection consistency for the objective nonlocal prior based approach of [1]. The proof is provided in the Supplementary Material [7].

THEOREM 6.1 (Strong selection consistency for nonlocal priors). *Consider the nonlocal prior based model described in* (6.1). *Under Assumptions* 1–4, *if we restrict the prior to DAGs with total number of edges at most* $d(\frac{n}{\log p})^{\frac{1}{2(2+k)}}$, *the following holds*:

$$\pi_{NL}(\mathscr{D}_0 \mid \boldsymbol{Y}) \xrightarrow{\bar{P}} 1,$$

as $n \to \infty$.

Note that the only difference between the assumptions needed for Theorem 4.3 (for DAG-Wishart priors) and Theorem 6.1 (for nonlocal priors) is that in Theorem 6.1 we restrict to DAGs with number of edges at most $d(\frac{n}{\log p})^{\frac{1}{2(2+k)}}$ [as opposed to $\frac{1}{8}d(\frac{n}{\log p})^{\frac{1+k}{2+k}}$ for Theorem 4.3]. All the remaining assumptions (Assumptions 1–4) are identical. Assumption 5 relates to the hyperparameters of the DAG-Wishart distribution, and hence is not relevant for the nonlocal prior setting.

7. Discussion: Comparison of penalized likelihood and Bayesian approaches. As mentioned in the Introduction, several penalized likelihood approaches for sparse estimation in Gaussian DAG models have been proposed in the literature. In this section, we compare and contrast these methods with the Bayesian approach of Ben-David et al. [5] considered in this paper.

For this discussion, we will focus on the approaches in [11, 14, 23], because these do not put any restrictions on the resulting sparsity pattern and focus on DAG models with ordering similar to the work of Ben-David et al. [5]. For several applications in genetics, finance and climate sciences, a location or time based ordering of variables is naturally available. For temporal data, a natural ordering of variables is provided by the time at which they are observed. In genetic datasets, the variables can be genes or SNPs located contiguously on a chromosome, and their spatial location provides a natural ordering. More examples can be found in [11, 14, 23, 27].

The more complex case of DAG models where a domain-specific ordering of the vertices is not known has also been studied in the literature; see [2, 22, 25] and the references therein. In [22], the authors first recover the underlying conditional independence relationships and find a DAG representative of these conditional independences. Then a covariance matrix obeying the conditional independence relationships in this DAG is estimated. In [2], the authors simultaneously estimate the DAG and the covariance matrix using a penalized regression approach.

7.1. *Comparison*: *Graph search complexity and accuracy.* For all the penalized likelihood methods in [11, 14, 23], a user-specified penalty parameter controls the level of sparsity of the resulting estimator. Varying values of the penalty parameter provide a range of possible DAG models to choose from. This set of graphs is referred to as the solution path. The choice of the penalty parameter is typically made by assigning a "score" to each DAG on the solution path using the Bayesian Information Criterion (BIC) or cross-validation, and choosing the DAG with the highest score. For the Bayesian approach, the posterior probabilities naturally assign a "score" for *all the* $2^{\binom{p}{2}}$ *DAGs*, not just the graphs on the solution path produced by the penalized likelihood methods. Of course, the entire space of DAGs is prohibitively large to search in high-dimensional settings. To address this, Ben-David et al. [5] develop a computationally feasible approach which searches around the graphs on the penalized likelihood solution path by adding or removing edges, and demonstrate that significant improvement in accuracy can be obtained by searching beyond the penalized likelihood solution paths using posterior probabilities. Hence, this Bayesian procedure maintains the advantage of being able to do a principled broader search (for improved accuracy) in a computationally feasible way. One can extend this procedure (see Section 8.2) by also searching on and around the solution paths of other methods, such as the CSCS method in [14], and choose the graph with the maximum posterior probability.

7.2. *Comparison*: *Uncertainty quantification and prior information*. A natural benefit of Bayesian approaches is the ability to incorporate prior knowledge and naturally provide uncertainty quantification through the posterior distribution. Prior knowledge can be incorporated in a principled way only when the hyperparameters are interpretable, and the class of priors is flexible. The distributional and moment results in [5] provide a natural interpretability for the hyperparameters U and $\boldsymbol{\alpha}$. As mentioned in [5], a separate shape parameter α_i for each variable allows for differential shrinkage. Also, the results in this paper provide justification for the asymptotic accuracy of the posterior distribution corresponding to the Bayesian approach of [5] in high-dimensional settings. Uncertainty quantification for estimates produced by penalized likelihood methods can be achieved through a CLT or through resampling methods such as bootstrap. To the best of our knowledge, a high-dimensional CLT, or results establishing high-dimensional accuracy of the bootstrap in this context are not available for the penalized likelihood based estimators in [11, 14, 23].

7.3. *Comparison*: *Convergence rates for estimation of* Ω. In this section, we will undertake a comparison of the assumptions and convergence rates between the Ω-estimate using the CSCS procedure in [14] and the posterior distribution convergence rate for Ω in Theorem E.1 in the Supplementary Material [7]. To the best of our knowledge, high-dimensional asymptotic convergence rates are not available for the estimates obtained from the procedure in [11]. We start with a point-by-point comparison of the parallel/related assumptions used for these high-dimensional asymptotic results:

1. For CSCS $p = p_n$ is assumed to be bounded above by a polynomial in n, whereas in this paper p_n can grow much faster than a polynomial in n (see Assumption 2). For CSCS, the eigenvalues of the Ω_0 are assumed to be uniformly bounded in n, whereas in this paper we allow the eigenvalues of Ω to grow with n (see Assumption 1).
2. As with any ℓ_1-penalized method, [14] use an incoherence condition for their asymptotic results. This condition is algebraically complex and hard to interpret. We do not need any such assumption for our asymptotic results.
3. For CSCS mild assumptions are specified regarding the rate at which the penalty parameter λ_n goes to zero. In this paper, we need to make analogous mild assumptions on the prior parameters q_n, U_n and $\boldsymbol{\alpha}(\mathscr{D}_n)$ (see Assumptions 3 and 5).

4. Recall that s_n is the smallest (in absolute value) nonzero off-diagonal entry of L_0. For CSCS, it is assumed that $\frac{s_n}{\sqrt{d_n \lambda_n}} \to \infty$ as $n \to \infty$, where λ_n is the penalty parameter, whereas we assume that $\frac{s_n}{\sqrt{\eta_n d_n}} \to \infty$ (see Assumption 4 with $\varepsilon_{0,n}$ as a constant for a fair comparison with CSCS). There are other assumptions in [14] regarding the rate at which λ_n goes to 0, but these do not enable a direct comparison of the two rates for s_n.

The convergence rate of the CSCS estimate of Ω is $m_n \lambda_n$ (m_n is the number of nonzeros in the Cholesky factor of the true concentration matrix) whereas the posterior convergence rate for Ω in Theorem E.1 in the Supplementary Material [7] is $d_n^2 \sqrt{\frac{\log p_n}{n}}$ (treating $\varepsilon_{0,n}$ as a constant for a fair comparison with CSCS). Using other assumptions regarding λ_n in [14], it can be shown that $m_n^{3/2} \sqrt{\frac{\log p_n}{n}} = o(m_n \lambda_n)$. Hence, if $m_n^{3/2} > d_n^2$, the Bayesian approach leads to a faster convergence rate. Of course, one can construct situations where $m_n^{3/2} < d_n^2$ and choose λ_n such that CSCS would lead to a faster convergence rate than the Bayesian approach. Since m_n is the *total* number of nonzeros in the true Cholesky factor one would expect that for a large majority of graphs, Theorem E.1 would lead to a faster convergence rate than CSCS.

8. Experiments.

8.1. *Simulation I*: *Illustration of posterior ratio consistency.* In this section, we illustrate the DAG selection consistency result in Theorems 4.1 and 4.2 using a simulation experiment. We consider 10 different values of p ranging from 250 to 2500, and choose $n = p/5$. Then, for each fixed p, we construct a $p \times p$ lower triangular matrix with diagonal entries 1 and off-diagonal entries 0.5. Then each lower triangular entry is independently set to zero with a certain probability such that the expected value of nonzero entries for each column does not exceed 3. We refer to this matrix as L_0. The matrix L_0 also gives us the true DAG $\mathscr{D}_0$. Next, we generate n i.i.d. observations from the $N(\mathbf{0}_p, (L_0^{-1})^T L_0^{-1})$ distribution, and set the hyperparameters as $U = I_p$ and $\alpha_i(\mathscr{D}) = \nu_i(\mathscr{D}) + 10$ for $i = 1, 2, \dots, p$. The above process ensures Assumptions 1–5 are satisfied. We then examine posterior ratio consistency under four different cases by computing the log posterior ratio of a "nontrue" graph $\mathscr{D}$ and $\mathscr{D}_0$ as follows:

1. Case 1: $\mathscr{D}$ is a subgraph of $\mathscr{D}_0$ and the number of total edges of $\mathscr{D}$ is exactly half of $\mathscr{D}_0$, that is, $|E(\mathscr{D})| = \frac{1}{2}|E(\mathscr{D}_0)|$.
2. Case 2: $\mathscr{D}$ is a supergraph of $\mathscr{D}_0$ and the number of total edges of $\mathscr{D}$ is exactly twice of $\mathscr{D}_0$, that is, $|E(\mathscr{D})| = 2|E(\mathscr{D}_0)|$.
3. Case 3: $\mathscr{D}$ is not necessarily a subgraph of $\mathscr{D}_0$, but the number of total edges in $\mathscr{D}$ is half the number of total edges in $\mathscr{D}_0$.

TABLE 1
Log of posterior probability ratio for $\mathscr{D}$ and $\mathscr{D}_0$ for various choices of the "nontrue" DAG $\mathscr{D}$. Here, $\mathscr{D}_0$ denotes the true underlying DAG

		$\mathscr{D} \subset \mathscr{D}_0$	$\mathscr{D} \supset \mathscr{D}_0$		
p	n	$\lvert E(\mathscr{D})\rvert = \frac{1}{2}\lvert E(\mathscr{D}_0)\rvert$	$\lvert E(\mathscr{D})\rvert = 2\lvert E(\mathscr{D}_0)\rvert$	$\lvert E(\mathscr{D})\rvert = \frac{1}{2}\lvert E(\mathscr{D}_0)\rvert$	$\lvert E(\mathscr{D})\rvert = 2\lvert E(\mathscr{D}_0)\rvert$
250	50	38,553	−133,007	24,723	−139,677
500	100	93,634	−458,799	51,553	−438,377
750	150	41,935	−784,866	60,731	−1,042,449
1000	200	249,342	−1,118,384	−28,657	−1,791,276
1250	250	18,847	−1,787,260	−245,769	−2,633,731
1500	300	−79,566	−2,603,779	−452,125	−3,873,151
1750	350	−512,894	−2,971,286	−455,941	−5,808,992
2000	400	−443,457	−4,082,005	−1,388,037	−7,139,952
2250	450	−558,718	−4,533,967	−1,883,472	−8,744,044
2500	500	−571,653	−4,708,833	−2,644,104	−9,910,277

4. Case 4: $\mathscr{D}$ is not necessarily a supergraph of $\mathscr{D}_0$, but the number of total edges in $\mathscr{D}$ is twice the number of total edges in $\mathscr{D}_0$.

The log of the posterior probability ratio for various cases is provided in Table 1. As expected the log of the posterior probability ratio eventually decreases as n becomes large in all four cases, thereby providing a numerical illustration of Theorems 4.1 and 4.2.

8.2. *Simulation II*: *Illustration of graph selection.* In this section, we perform a simulation experiment to illustrate the potential advantages of using the hybrid Bayesian graph selection approach outlined in Section 7.1. We consider 7 values of p ranging from 2500 to 4000, with $n = p/5$. For each fixed p, the Cholesky factor L_0 of the true concentration matrix, and the subsequent dataset, is generated by the same mechanism as in Section 8.1. Then we perform graph selection using the four procedures outlined below:

1. *Lasso-DAG BIC path search*: We implement the Lasso-DAG approach in [23]. The penalty parameter is varied on a grid so that the resulting graphs range from approximately three times the edges compared to the true graph with approximately one-third edges compared to the true graph. We then select the best graph according to the "BIC"-like measure defined as

$$\mathrm{BIC}(\lambda) = n\,\mathrm{tr}(S\hat{\Omega}) - n \log|\hat{\Omega}| + \log n * E, \tag{8.1}$$

where $\widehat{L}$ is the resulting estimator from Lasso-DAG, E denotes the total numbers of nonzero entries in $\hat{L}$ and $\hat{\Omega} = \hat{L}^T\hat{L}$.
2. *Lasso-DAG with quantile based tuning*: We again implement the Lasso-DAG approach in [23], but choose penalty parameters (separate for each variable i)

given by $\lambda_i(\alpha) = 2n^{-\frac{1}{2}} Z^*_{\frac{0.1}{2p(i-1)}}$, where Z^*_q denotes the $(1-q)$th quantile of the standard normal distribution. This choice is justified in [23] based on asymptotic considerations.

3. *CSCS BIC path search*: We implement the CSCS approach in [14]. The penalty parameter is varied on a grid so that the resulting graphs range from three times the edges compared to the true graph with one-third edges compared to the true graph. The best graph us selected using the "BIC"-like measure in (8.1).
4. *Bayesian approach*: We construct two sets of candidate graphs as follows:
   (a) All the graphs on the solution paths for Lasso-DAG and CSCS are included in the candidate set. To increase the search range, we generate additional graphs by thresholding the modified Cholesky factor of $(S + 0.5I)^{-1}$ (S is the sample covariance matrix) to get a sequence of 300 additional graphs, and include them in the candidate set. We then search around all the above graphs using Shotgun Stochastic Search to generate even more candidate graphs. Then we implement Algorithm A.8 in [15], the Greedy Hill-climbing algorithm, to our candidate graphs. For each graph, this particular search procedure first generates a new DAG by adding one random edge and only chooses it if the new DAG has a higher posterior score. Then we generate another graph by deleting one random edge from the chosen DAG and select the one with higher score. We repeat the whole process 20 times for every graph in the previous candidate set and all the chosen DAGs are included in the candidate set.
   (b) We combine Algorithm 18.1 in [15] and the idea of cross-validation to form our second set of candidate graphs. The original data set of n observations is randomly partitioned into 10 equal sized subsets. Of the 10 subsets, a single subset is excluded, and the remaining 9 subsets are used as our new sample. The same thresholding procedure to generate 300 graphs is performed for the new sample covariance matrix. The process is then repeated 10 times, with each of the 10 subsamples removed exactly once. We then have a total of 3000 graphs as the second candidate set.

   The log posterior probabilities are computed for all graphs in the two candidate sets, and the graph with the highest probability is chosen.

The model selection performance of these four methods is then compared using several different measures of structure such as positive predictive value, true positive rate and false positive rate (average over 20 independent repetitions). Positive Predictive Value (PPV) represents the proportion of true edges among all the edges detected by the given procedure, True Positive Rate (TPR) measures the proportion of true edges detected by the given procedure among all the edges from the true graph and False Positive Rate (FPR) represents the proportion of false edges detected by the given procedure among all the nonedges in the true graph. One would like the PPV and TPR values to be as close to 1 as possible, and the FPR value to

TABLE 2
Model selection performance table

		Lasso-DAG			Lasso-DAG			CSCS			Bayesian		
		BIC path search			Quantile-based lambdas			BIC path search			Log-score path search		
p	n	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR	PPV	TPR	FPR
2500	500	0.082	0.201	0.003	0.103	0.182	0.002	0.086	0.205	0.003	0.993	0.396	3.76×10^{-6}
2750	550	0.053	0.173	0.003	0.065	0.163	0.003	0.082	0.185	0.002	0.988	0.458	6.75×10^{-6}
3000	600	0.060	0.187	0.003	0.071	0.171	0.002	0.078	0.183	0.002	0.993	0.453	3.26×10^{-6}
3250	650	0.062	0.171	0.002	0.071	0.156	0.002	0.080	0.184	0.002	0.969	0.502	1.73×10^{-5}
3500	700	0.067	0.177	0.002	0.078	0.165	0.002	0.081	0.188	0.002	0.972	0.525	1.31×10^{-5}
3750	750	0.068	0.174	0.002	0.079	0.164	0.002	0.076	0.175	0.002	0.978	0.524	9.47×10^{-6}
4000	800	0.068	0.187	0.002	0.077	0.176	0.002	0.073	0.165	0.002	0.957	0.565	1.91×10^{-5}

be as close to 0 as possible. The results are provided in Table 2. It is clear that the Bayesian approach outperforms the penalized likelihood approaches based on all measures. The PPV values for the Bayesian approach are all above 0.95, while the ones for the penalized likelihood approaches are around 0.1. The TPR values for the Bayesian approach are all above 0.39, while the ones for the penalized likelihood approaches are all below 0.21. The FPR values for the Bayesian approach are all significantly smaller than the penalized approaches. Overall, this experiment illustrates that the Bayesian approach can be used for a broader yet computationally feasible graph search, and can lead to a significant improvement in graph selection performance.

Acknowledgments. The authors are grateful to anonymous referees and an Associate Editor for their encouraging and helpful comments which substantially improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models" (DOI: 10.1214/18-AOS1689SUPP; .pdf). This supplemental file contains additional proofs for theorems and technical lemmas.

REFERENCES

[1] ALTOMARE, D., CONSONNI, G. and LA ROCCA, L. (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics* **69** 478–487. MR3071066

[2] ARAGAM, B., AMINI, A. and ZHOU, Q. (2015). Learning directed acyclic graphs with penalized neighbourhood regression. Available at https://arxiv.org/abs/1511.08963.

[3] BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. MR3273620

[4] BANERJEE, S. and GHOSAL, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. MR3321485

[5] BEN-DAVID, E., LI, T., MASSAM, H. and RAJARATNAM, B. (2016). High dimensional Bayesian inference for Gaussian directed acyclic graph models. Technical report. Available at http://arxiv.org/abs/1109.4371.

[6] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[7] CAO, X., KHARE, K. and GHOSH, M. (2019). Supplement to "Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models." DOI:10.1214/18-AOS1689SUPP.

[8] CONSONNI, G., LA ROCCA, L. and PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.* **44** 741–764. MR3687971

[9] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012

[10] GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324

[11] HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

[12] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. MR2830762

[13] JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

[14] KHARE, K., OH, S., RAHMAN, S. and RAJARATNAM, B. (2017). A convex framework for high-dimensional sparse cholesky based covariance estimation in gaussian dag models. Technical report. Available at https://arxiv.org/abs/1610.02436.

[15] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA. MR2778120

[16] LETAC, G. and MASSAM, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706

[17] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987

[18] PAULSEN, V. I., POWER, S. C. and SMITH, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85** 151–178. MR1005860

[19] POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94** 1006–1013. MR2376812

[20] ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. MR2672482

[21] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. MR3125258

[22] RÜTIMANN, P. and BÜHLMANN, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electron. J. Stat.* **3** 1133–1160. MR2566184

[23] SHOJAIE, A. and MICHAILIDIS, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** 519–538. MR2672481

[24] SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97** 1141–1153. MR1951266

[25] VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113

[26] XIANG, R., KHARE, K. and GHOSH, M. (2015). High dimensional posterior convergence rates for decomposable graphical models. *Electron. J. Stat.* **9** 2828–2854. MR3439186

[27] YU, G. and BIEN, J. (2017). Learning local dependence in ordered data. *J. Mach. Learn. Res.* **18** Paper No. 42, 60. MR3655307

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
102 GRIFFIN-FLOYD HALL
GAINESVILLE, FLORIDA 32611
USA
E-MAIL: caoxuan@ufl.edu
kdkhare@stat.ufl.edu
ghoshm@ufl.edu