The Annals of Statistics
2019, Vol. 47, No. 6, 3413–3437
https://doi.org/10.1214/18-AOS1783

MINIMAX POSTERIOR CONVERGENCE RATES AND MODEL SELECTION CONSISTENCY IN HIGH-DIMENSIONAL DAG MODELS BASED ON SPARSE CHOLESKY FACTORS

BY KYOUNGJAE LEE[*,1], JAEYONG LEE[†,2] AND LIZHEN LIN[‡,1]

Inha University,[*] Seoul National University[†] and University of Notre Dame[‡]

In this paper we study the high-dimensional sparse directed acyclic graph (DAG) models under the empirical sparse Cholesky prior. Among our results, strong model selection consistency or graph selection consistency is obtained under more general conditions than those in the existing literature. Compared to Cao, Khare and Ghosh [*Ann. Statist.* (2019) **47** 319–348], the required conditions are weakened in terms of the dimensionality, sparsity and lower bound of the nonzero elements in the Cholesky factor. Furthermore, our result does not require the irrepresentable condition, which is necessary for Lasso-type methods. We also derive the posterior convergence rates for precision matrices and Cholesky factors with respect to various matrix norms. The obtained posterior convergence rates are the fastest among those of the existing Bayesian approaches. In particular, we prove that our posterior convergence rates for Cholesky factors are the minimax or at least nearly minimax depending on the relative size of true sparseness for the entire dimension. The simulation study confirms that the proposed method outperforms the competing methods.

1. Introduction. Detecting the dependence structure of multivariate data is one of important and challenging tasks, especially when the number of variables is much larger than the sample size. Due to advancements in technology, such data are routinely collected in a wide range of areas including genomics, climatology, proteomics and neuroimaging. The estimation of the covariance (or precision) matrix is crucial to reveal the dependence structure. Under the high-dimensional setting, however, the traditional sample covariance matrix is no longer a consistent estimator of the true covariance matrix [Johnstone and Lu (2009)]. For the consistent estimation of the high-dimensional covariance or precision matrices, various restrictive matrix classes have been proposed to reduce the number of effective parameters. They include the bandable matrices [Banerjee and Ghosal (2014), Bickel

Received February 2018; revised October 2018.

[1] Supported by NSF Grants IIS 1663870 and DMS Career 1654579, and DARPA Grant N66001-17-1-4041.

[2] Supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2018R1A2A3074973).

MSC2010 subject classifications. Primary 62C20; secondary 62F15, 62C12.

Key words and phrases. DAG model, precision matrix, Cholesky factor, posterior convergence rate, strong model selection consistency.

and Levina (2008), Cai and Yuan (2012), Cai, Zhang and Zhou (2010)], sparse matrices [Banerjee and Ghosal (2015), Cai and Zhou (2012a, 2012b)] and low-dimensional structural matrices such as the sparse spiked covariance [Cai, Ma and Wu (2015), Gao and Zhou (2015)] and sparse factor models [Fan, Fan and Lv (2008), Pati et al. (2014)]. When the class of sparse matrices is of interest, the sparsity pattern can be encoded in many different ways. Sparsity can be imposed on the covariance matrix, precision matrix or Cholesky factor, which lead to different graph models. In this paper we focus on imposing sparsity on the Cholesky factor of the precision matrix.

Consider a sample of data $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma_n)$, where $N_p(\mu, \Sigma)$ is the p-dimensional normal distribution with the mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. For every positive definite matrix $\Omega_n = \Sigma_n^{-1}$, the modified Cholesky decomposition (MCD) guarantees the existence of unique Cholesky factor A_n and diagonal matrix D_n such that $\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n)$. The sparsity of a Gaussian directed acyclic graph (DAG) can be uniquely encoded by the Cholesky factor A_n through the structure of the graph. In this paper we assume that the parent ordering of the variables is known, which is a common assumption used in the literature such as in Ben-David et al. (2015), Khare et al. (2016), Yu and Bien (2017) and Cao, Khare and Ghosh (2019). The details on this concept will be provided in Section 2.2. For the estimation of Cholesky factor A_n, the banded assumption and the sparsity assumption are two popular assumptions. Under the banded assumption, the elements of the matrix far from the diagonal are assumed to be all zero, while under the sparsity assumption, there is no constraint on the zero-pattern other than assuming most of the entries are zero. In recent years, various penalized likelihood estimators have been proposed with the sparsity assumption on A_n [Huang et al. (2006), Khare et al. (2016), Rothman, Levina and Zhu (2010), Shojaie and Michailidis (2010), van de Geer and Bühlmann (2013)] and banded assumption on A_n [Yu and Bien (2017)].

On the Bayesian side, relatively few works have dealt with asymptotic properties of the posteriors of high-dimensional Gaussian DAG models. Posterior convergence rates for the precision matrices with G-Wishart priors [Roverato (2000)] were derived by Banerjee and Ghosal (2014) and Xiang, Khare and Ghosh (2015), where G is a decomposable graph. Note that a decomposable graph can be converted to a perfect DAG, a special case of the DAGs, by ignoring directions. Lee and Lee (2017) obtained the posterior convergence rates and minimax lower bounds for the precision matrices, but only bandable Cholesky factors were considered. Posterior convergence rates for the precision matrices as well as strong model selection consistency were recently derived by Cao, Khare and Ghosh (2019) for sparse DAG models. However, their results are not adaptive to the unknown sparsity s_0, and the conditions required for obtaining such results are somewhat restrictive.

In this paper we consider high-dimensional sparse Gaussian DAG models where sparsity is imposed via the sparse Cholesky factor. We adopt an empirical Bayes approach with a fractional likelihood. The empirical Bayes approach is justified by showing desirable asymptotic properties of the induced posterior such as strong model selection consistency and optimal posterior convergence rates. In addition, our theoretical results are adaptive to the unknown sparsity s_0.

There are four main contributions of this work. First we show strong model selection consistency under much more general conditions than those in the literature. Specifically, the required conditions on the dimensionality, sparsity, structure of the Cholesky factor A_n and the lower bound of the nonzero elements in A_n (the *beta-min* condition, which will be described later) are significantly weakened. Second, we derive the minimax or nearly minimax posterior convergence rates for the Cholesky factors under two scenarios: with or without the beta-min condition for the true Cholesky factor. We show that at least one of the posterior convergence rates is minimax depending on the relative size of true sparseness for the entire dimension. To the best of our knowledge, this is the first result on minimax posterior convergence rates in high-dimensional DAG models. Third, we obtain the posterior convergence rates for precision matrices with respect to the spectral norm and matrix ℓ_∞ norm, which is the fastest among those of existing Bayesian approaches. Compared to Cao, Khare and Ghosh (2019), we achieve faster posterior convergence rate under more general conditions, except the bounded eigenvalue condition. Furthermore, their results depend on the unknown sparsity s_0, whereas ours do not. Fourth, our method significantly improves the model selection performance in practice. In particular, our method outperforms the other state-of-the-art methods in a simulation study. The theoretical choice of hyperparameters provided good guidelines for practical performance. Note that the choice of the hyperparameter, the individual edge probability q_n, in the hierarchical DAG-Wishart prior [Cao, Khare and Ghosh (2019)] can be problematic in practice, as the posterior with the theoretical choice of q_n tends to be stuck at very small size models.

The rest of paper is organized as follows. In Section 2, we introduce notation, Gaussian DAG models, the empirical sparse Cholesky prior, the fractional posterior and the parameter class for the precision matrices. In Section 3, strong model selection consistency, posterior convergence rates and minimax lower bounds for the Cholesky factor and posterior convergence rates for the precision matrices are established. A simulation study focusing on the model selection property are represented in Section 4. The proofs of the main results are provided in the supplemental article [Lee, Lee and Lin (2018)].

2. Preliminaries.

2.1. *Norms and notation.* For any $a, b \in \mathbb{R}$, we denote $a \vee b$ and $a \wedge b$ as the maximum and minimum of a and b, respectively. For any $a \in \mathbb{R}$, we denote $\lfloor a \rfloor$ as the largest integer equal to or smaller than a. For any sequences a_n and b_n, $a_n = o(b_n)$ denotes $a_n / b_n \to 0$ as $n \to \infty$. We denote $a_n = O(b_n)$, or equivalently $a_n \lesssim b_n$, if $a_n \le C b_n$ for some constant $C > 0$, where C is an universal constant.

We denote the indicator function for a set A as $I(\cdot \in A)$ or $I_A(\cdot)$. For a given p-dimensional vector $u = (u_1, \dots, u_p)^T$ and set $S \subseteq \{1, \dots, p\}$, we define $u_S = (u_j)^T_{j\in S} \in \mathbb{R}^{|S|}$, where $|S|$ is the cardinality of S. For given index sets S, $S' \subseteq \{1, \dots, p\}$ and real matrix $B \in \mathbb{R}^{p\times p}$, we denote $B_{(S,S')}$ as the $|S| \times |S'|$ submatrix consisting only of Sth columns and S'th rows of B, and let $B_S = B_{(S,S)}$. For a real matrix B, we denote S_B as the index set for nonzero elements of B and call S_B the *support* of B. We define $\mathcal{C}_p$ as the class of all $p \times p$ dimensional positive definite matrices. For any $p \times p$ symmetric matrix B, $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ are the minimum and maximum eigenvalues of B, respectively.

For any p-dimensional vector $u = (u_1, \dots, u_p)^T$, we define vector norms $\|u\|_1 = \sum_{j=1}^p |u_j|$, $\|u\|_2 = (\sum_{j=1}^p u_j^2)^{1/2}$ and $\|u\|_{\max} = \max_{1\le j\le p} |u_j|$. For any $p \times p$ matrix $B = (b_{ij})$, we define the spectral norm, matrix ℓ_1 norm, matrix ℓ_∞ norm and Frobenius norm by

$$\|B\| = \sup_{\substack{x\in\mathbb{R}^p \\ \|x\|_2=1}} \|Bx\|_2 = \left(\lambda_{\max}(B^T B)\right)^{1/2},$$

$$\|B\|_1 = \sup_{\substack{x\in\mathbb{R}^p \\ \|x\|_1=1}} \|Bx\|_1 = \max_{1\le j\le p} \sum_{i=1}^p |b_{ij}|,$$

$$\|B\|_\infty = \sup_{\substack{x\in\mathbb{R}^p \\ \|x\|_{\max}=1}} \|Bx\|_{\max} = \max_{1\le i\le p} \sum_{j=1}^p |b_{ij}|, \quad \text{and}$$

$$\|B\|_F = \left(\sum_{i=1}^p \sum_{j=1}^p b_{ij}^2\right)^{1/2},$$

respectively.

For a given positive integer m, we denote χ^2_m as the chi-square distribution with degrees of freedom m. For any random variables Y_1, Y_2 and Y_3, we denote $Y_1 \overset{d}{\equiv} Y_2 \oplus Y_3$ if the distribution of Y_1 is equal to that of $Y_2 + Y_3$, and Y_2 and Y_3 are independent. For given positive numbers a and b, $\mathrm{Gamma}(a, b)$ and $\mathrm{IG}(a, b)$ are the gamma distribution and inverse-gamma distribution with shape parameter a and rate parameter b, respectively. $\mathrm{Beta}(a, b)$ is the beta distribution whose density function at $x \in (0, 1)$ is proportional to $x^{a-1}(1 - x)^{b-1}$. We denote $N_p(X \mid \mu, \Sigma)$ as the density function of $N_p(\mu, \Sigma)$ at $X \in \mathbb{R}^p$. We denote the inverse-Wishart distribution by $\mathrm{IW}_p(\nu, \Phi)$, where the degree of freedom and scale matrix are $\nu > p - 1$ and $\Phi \in \mathcal{C}_p$, respectively.

2.2. *Gaussian DAG models.* We consider the model

$$X_1, \dots, X_n \mid \Omega_n \overset{\text{i.i.d.}}{\sim} N_p(0, \Omega_n^{-1}), \tag{1}$$

where $\Omega_n = \Sigma_n^{-1}$ is a $p \times p$ precision matrix and $X_i = (X_{i1}, \dots, X_{ip})^T \in \mathbb{R}^p$ for all $i = 1, \dots, n$. The MCD guarantees that there exists unique lower triangular matrix $A_n = (a_{jl})$ and diagonal matrix $D_n = \mathrm{diag}(d_j)$ such that

$$\Omega_n = (I_p - A_n)^T D_n^{-1} (I_p - A_n), \tag{2}$$

where $a_{jj} = 0$ and $d_j > 0$ for all $j = 1, \dots, p$. Let S_{A_n} be the support of the Cholesky factor A_n, and S_j be the support of the jth row of A_n. Let $\mathbb{P}_{\Omega_n}$ and $\mathbb{E}_{\Omega_n}$ be the probability measure and expectation corresponding to the model (1), respectively.

The model (1) can be interpreted as a Gaussian DAG model depending on the sparsity pattern of A_n. For a set of vertices $V = \{1, \dots, p\}$ and a set of directed edges E, a graph $\mathcal{D} = (V, E)$ is said to be a DAG if there is no directed cycles. It is also called the Bayesian network or belief network. In this paper we assume that the variables have a known natural ordering in which no edges exist from larger vertices to smaller vertices. It has been commonly assumed in literature including Shojaie and Michailidis (2010), Ben-David et al. (2015), Khare et al. (2016), Yu and Bien (2017) and Cao, Khare and Ghosh (2019). There are relatively few works on DAG models when the ordering of variables is unknown [Kalisch and Bühlmann (2007), Rütimann and Bühlmann (2009), van de Geer and Bühlmann (2013)]. As discussed in van de Geer and Bühlmann (2013), when the ordering is unknown, a very different technique is needed relative to the known ordering case.

For $i \in V$, define the set of all i's parents as the subset of V smaller than i and sharing an edge with i and denote it as $\mathrm{pa}_i(\mathcal{D})$. Any multivariate Gaussian distribution that obeys the directed Markov property with respect to a DAG $\mathcal{D}$ is said to be a *Gaussian DAG model over* $\mathcal{D}$. To be specific, if $X = (X_1, \dots, X_p)^T \sim N_p(0, \Omega^{-1})$ and $N_p(0, \Omega^{-1})$ belongs to a Gaussian DAG model over $\mathcal{D}$, then

$$X_j \perp \{X_{j'}\}_{j' < j, j' \notin \mathrm{pa}_j(\mathcal{D})} | (X)_{\mathrm{pa}_j(\mathcal{D})},$$

for each $j = 1, \dots, p$. Furthermore, if we adopt the MCD as in (2), with the known ordering of variables, $N_p(0, \Omega^{-1})$ belongs to a Gaussian DAG model over $\mathcal{D}$ if and only if $a_{jl} = 0$ whenever $l \notin \mathrm{pa}_j(\mathcal{D})$ [Cao, Khare and Ghosh (2019)]. In other words, *the support of A uniquely determines a DAG $\mathcal{D}$* under the known ordering assumption. The model $X = (X_1, \dots, X_p)^T \sim N_p(0, \Omega^{-1})$ given S_A is equivalent to a Gaussian DAG model, and it can be represented as a linear autoregressive model,

$$\begin{aligned} &X_1 \mid d_1 \sim N(0, d_1), \\ &X_j \mid a_{S_j}, d_j, S_j \overset{\text{ind}}{\sim} N\Bigl(\sum_{l \in S_j} X_l a_{jl}, d_j\Bigr), \qquad j = 2, \dots, p, \end{aligned} \tag{3}$$

where $a_{S_j} = a_{j,S_j} = (a_{jj'})^T_{j' \in S_j}$. For more details on the expression (3), refer to Bickel and Levina (2008) and Ben-David et al. (2015). The autoregressive model interpretation enables us to adopt the priors introduced in the linear regression literature. Since a_{S_j} corresponds to nonzero elements among $a_j = (a_{j1}, \dots, a_{j,j-1})^T$, one can use a prior designed for sparse regression coefficient vectors for a_j, which is our strategy introduced in Section 2.3.

In this paper we consider the high-dimensional setting where $p = p_n$ is a function of n increasing to infinity as $n \to \infty$ and $p \geq n$. We assume that the data were generated from a true precision matrix Ω_{0n}, where $\Sigma_{0n} = \Omega_{0n}^{-1}$ is the true covariance matrix. Denote the MCD (2) of the true precision matrix by $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1}(I_p - A_{0n})$, where $A_{0n} = (a_{0,jl})$, $a_{0j} = (a_{0,j1}, \dots, a_{0,j,j-1})^T$ and $D_{0n} = \operatorname{diag}(d_{0j})$. For notational convenience, let $\mathbb{P}_0 = \mathbb{P}_{\Omega_{0n}}$ and $\mathbb{E}_0 = \mathbb{E}_{\Omega_{0n}}$.

We now define some notation related to the data set. Let $\mathbf{X}_n = (X_1, \dots, X_n)^T \in \mathbb{R}^{n \times p}$ be the data of size n, and $\tilde{X}_j = (X_{1j}, \dots, X_{nj})^T \in \mathbb{R}^n$ be the jth column of $\mathbf{X}_n$. For a given index set $S \subseteq \{1, \dots, p\}$, let $\mathbf{X}_S = (\tilde{X}_j)_{j \in S} \in \mathbb{R}^{n \times |S|}$ be the data matrix consisting only of Sth columns of $\mathbf{X}_n$. Let $Z_{ij} = (X_{i1}, \dots, X_{i,j-1})^T \in \mathbb{R}^{j-1}$ and $\tilde{Z}_j = (Z_{1j}, \dots, Z_{nj})^T \in \mathbb{R}^{n \times (j-1)}$ for all $j = 2, \dots, p$.

For a given positive integer $1 \leq s \leq p$, we define $\Psi_{\max}(s)^2 = \sup_{S: 0 < |S| \leq s} \lambda_{\max}(\mathbf{X}_S^T \mathbf{X}_S)$ and $\Psi_{\min}(s)^2 = \inf_{S: 0 < |S| \leq s} \lambda_{\min}(\mathbf{X}_S^T \mathbf{X}_S)$, where the supremum and infimum are taken over all index sets $S \subseteq \{1, \dots, p\}$. *We say that the restricted eigenvalue condition is met for some integer s if $n^{-1}\Psi_{\min}(s)^2$ is bounded away from zero uniformly for all large n.* This condition has been used in the high-dimensional regression literature to control the behavior of the design matrix. The autoregressive model representation (3) connects the eigenvalues of the precision matrix Ω_{0n} with those of the design matrix in (3) because the quantity $\mathbf{X}_{S_j}$ corresponds to the design matrix based on the representation. Thus, the bounded eigenvalue assumption (A1) in Section 2.5 essentially corresponds to the restricted eigenvalue condition.

2.3. *Empirical sparse Cholesky prior.* In this paper we suggest the following prior distribution for our model:

$$
\begin{aligned}
a_{S_j} \mid d_j, S_j &\overset{\text{ind}}{\sim} N_{|S_j|}\left(\widehat{a}_{S_j}, \frac{d_j}{\gamma}(\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1}\right), \qquad j = 2, \dots, p, \\
\pi(d_j) &\overset{\text{i.i.d.}}{\propto} d_j^{-\nu_0/2-1}, \qquad j = 1, \dots, p, \\
\pi_j(S_j = S_j') &\propto \binom{j-1}{|S_j'|}^{-1} f_{nj}(|S_j'|), \qquad j = 2, \dots, p, S_j' \subseteq \{1, \dots, j-1\}, \\
f_{nj}(|S_j'|) &\propto c_1^{-|S_j'|} p^{-c_2|S_j'|} I(0 \leq |S_j'| \leq R_j \wedge (j-1)), \qquad j = 2, \dots, p,
\end{aligned}
\tag{4}
$$

for some positive constants $\nu_0, c_1, c_2, R_2, \dots, R_p$ and γ, where f_{nj} is a probability mass function on $\{0, 1, \dots, R_j \wedge (j-1)\}$ and $\widehat{a}_{S_j} = (\mathbf{X}_{S_j}^T \mathbf{X}_{S_j})^{-1} \mathbf{X}_{S_j}^T \tilde{X}_j$. The proposed prior is empirical in the sense that it depends on the data, so we call the prior (4) the empirical sparse Cholesky (ESC) prior. To obtain the desired asymptotic properties, appropriate conditions for hyperparameters in (4) will be introduced in Section 3. Note that the prior for d_j can be generalized to the proper prior $\mathrm{IG}(\nu_0/2, \nu_0')$ for some constant $\nu_0' > 0$ and the results in Section 3 also hold for this prior choice. However, for computational convenience, we describe and prove the main results with the improper prior $\pi(d_j) \propto d_j^{-\nu_0/2-1}$.

For the conditional prior of a_j given d_j, we first introduce zero components through the prior π_j and impose the Zellner's g-prior [Zellner (1986)] on the nonzero components, a_{S_j}. The use of Zellner's g-prior simplifies the calculation of the marginal posterior for S_j. Martin, Mess and Walker (2017) suggested a similar prior in the high-dimensional linear regression model. Also note that the ESC prior has a connection to the DAG-Wishart prior [Ben-David et al. (2015), Cao, Khare and Ghosh (2019)]. Theorem 7.3 in Ben-David et al. (2015) shows that the DAG-Wishart prior on (A_n, D_n) given a DAG implies the inverse-gamma distribution on d_j and multivariate normal distribution on the nonzero elements of a_j given d_j, where (a_j, d_j) are mutually independent for all $j = 1, \dots, p$. Thus, the ESC prior (4) is quite close to the DAG-Wishart prior when the support of A_n is given.

Cao, Khare and Ghosh (2019) used the DAG-Wishart prior to recover the sparse DAG and estimate the precision matrix in high-dimensional settings. Thus, their prior on (A_n, D_n) is quite close to ours, and can be viewed as a set of priors for autoregressive model (3) as discussed in the previous paragraph. For the support of DAGs, they imposed the elementwise sparsity using independent Bernoulli distributions with the hyperparameter q_n, which has a nice interpretation as the individual edge probability. Based on the hierarchical DAG-Wishart prior, they obtained the strong model selection consistency for the DAG and the posterior convergence rate for the precision matrix with respect to the spectral norm. However, they did not directly adopt the autoregressive model interpretation as in (3), which is different from our approach. By using the ESC prior, we can adopt state-of-the-art techniques on selection consistency for the regression coefficient [Martin, Mess and Walker (2017)] and achieve the strong model selection consistency under much weaker conditions than those in Cao, Khare and Ghosh (2019). Furthermore, compared to the existing literature, we obtain faster posterior convergence rates for precision matrices and Cholesky factors under weaker conditions using the techniques introduced by Lee and Lee (2017), Lee and Lee (2018) and Martin, Mess and Walker (2017). Indeed, the posterior convergence rates for Cholesky factors are nearly or exactly optimal depending on the relative size of true sparseness for the entire dimension.

2.4. *α-posterior distribution.* We suggest adopting the fractional likelihood with power $\alpha \in (0, 1)$,

$$L_n(A_n, D_n)^\alpha = \prod_{i=1}^{n} \{N_p(X_i \mid 0, (I_p - A_n)^{-1} D_n((I_p - A_n)^T)^{-1})\}^\alpha. \tag{5}$$

The use of fractional likelihood has received increased attention in recent years [Martin and Walker (2014), Syring and Martin (2016), Miller and Dunson (2018)]. In this paper we use the fractional likelihood mainly because of its appealing theoretical properties under relatively weaker conditions compared to the actual posterior [Bhattacharya, Pati and Yang (2019)]. In the proof of the main results of this paper, the use of the fractional likelihood enables us to effectively deal with the ratio of estimated residual variances $\widehat{d}_{S_j}$ (the proof of Theorem 3.1) and the ratio of likelihood $L_{nj}(a_j, d_j)$ (the proof of Lemma 7.2), where $\widehat{d}_{S_j}$ and $L_{nj}(a_j, d_j)$ will be defined later.

Here we give a more detailed justification of using the fractional likelihood. The proposed conditional prior for a_{S_j} in (4) tracks the data closely because it is centered at the least square estimate. It may cause the unexpected inconsistency [Walker and Hjort (2001)]. The fractional likelihood approach can prohibit it by preventing the posterior from following the data too closely. To be more specific, the use of fractional likelihood can be interpreted as an empirical Bayes procedure by considering

$$L_n(A_n, D_n)^\alpha \pi(A_n, D_n) = L_n(A_n, D_n) \frac{\pi(A_n, D_n)}{L_n(A_n, D_n)^{1-\alpha}}.$$

Hence, the resulting posterior consists of an ordinary likelihood function and a data-dependent prior $\pi(A_n, D_n)/L_n(A_n, D_n)^{1-\alpha}$. Note that the power α *only appears in the prior*. From this point of view, the prior is rescaled by a fractional likelihood which has an effect of penalizing parameter values that track the data too closely, while the penalty effect is controlled by the *hyperparameter α*.

The choice of α can be important from a practitioner's point of view even though *theoretical results in this paper hold for any choice of* $0 < \alpha < 1$. In practice, we suggest using α close to 1 to mimic the usual likelihood in finite sample scenario if there is no suspect of model failure, that is, misspecification. As long as one chooses α sufficiently close to 1, for example, $\alpha = 0.999$ or $\alpha = 0.9999$, our experience confirms that the α-posterior can be hardly distinguishable from the "usual" posterior even in a finite sample scenario.

REMARK 2.1. Grünwald and van Ommen (2017) suggested using I-log-SafeBayes (or R-log-SafeBayes) to determine α, which gives the minimizer $\hat{\alpha}$ of the posterior-expected posterior-randomized loss of prediction (or its variant). The induced posterior is robust to model misspecification in some cases [Grünwald and van Ommen (2017)].

The prior (4) and fractional likelihood (5) lead to the following joint posterior distribution:

$$
\begin{aligned}
&a_{S_j} \mid d_j, S_j, \mathbf{X}_n \overset{\text{ind}}{\sim} N_{|S_j|}\Big(\widehat{a}_{S_j}, \frac{d_j}{(\alpha+\gamma)}(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}\Big), \qquad j = 2, \dots, p, \\
&d_j \mid S_j, \mathbf{X}_n \overset{\text{ind}}{\sim} \mathrm{IG}\Big(\frac{\alpha n + \nu_0}{2}, \frac{\alpha n}{2}\widehat{d}_{S_j}\Big), \qquad j = 1, \dots, p, \qquad (6)\\
&\pi_\alpha(S_j \mid \mathbf{X}_n) \propto \pi_j(S_j)\Big(1 + \frac{\alpha}{\gamma}\Big)^{-\frac{|S_j|}{2}} (\widehat{d}_{S_j})^{-\frac{\alpha n + \nu_0}{2}}, \qquad j = 2, \dots, p,
\end{aligned}
$$

where $\widehat{d}_{S_j} = n^{-1}\tilde{X}_j^T(I_n - \tilde{P}_{S_j})\tilde{X}_j$ and $\tilde{P}_{S_j} = \mathbf{X}_{S_j}(\mathbf{X}_{S_j}^T\mathbf{X}_{S_j})^{-1}\mathbf{X}_{S_j}^T$. We refer to the posterior (6) as the α-posterior and denote it by $\pi_\alpha(\cdot \mid \mathbf{X}_n)$ to clarify that we consider the α-fractional likelihood. Throughout the paper, $\alpha \in (0, 1)$ is a fixed constant.

2.5. *Parameter class.* For given positive constants $0 < \alpha < 1$, $0 < \epsilon_0 < 1/2$, C_{bm} and a sequence of positive integers s_0, we introduce conditions (A1)–(A4) for the true precision matrix:

(A1) $\epsilon_0 \le \lambda_{\min}(\Omega_{0n}) \le \lambda_{\max}(\Omega_{0n}) \le \epsilon_0^{-1}$.
(A2) $\max_{1\le j\le p}\sum_{l=1}^p I(a_{0,jl} \ne 0) \le s_0$.
(A3)

$$
\min_{(j,l):a_{0,jl}\ne 0} |a_{0,jl}|^2 \ge \frac{16}{\alpha(1-\alpha)\epsilon_0^2(1-2\epsilon_0)^2} C_{\mathrm{bm}} \frac{\log p}{n}.
$$

(A4) $\max_{1\le l\le p}\sum_{j=2}^p I(a_{0,jl} \ne 0) \le s_0$.

Condition (A1) ensures that the eigenvalues of Ω_{0n} are bounded by fixed constants, which has been commonly used for the estimation of the high-dimensional precision matrices [Banerjee and Ghosal (2015), Cai, Liu and Zhou (2016), Ren et al. (2015)] as well as the high-dimensional DAGs [Khare et al. (2016), Yu and Bien (2017)]. In this paper condition (A1) is mainly used to get upper bounds of d_{0j}, d_{0j}^{-1} and $\|A_{0n}\|$.

Condition (A2) restricts the number of nonzero elements in each row of A_{0n} to be smaller than s_0. Note that s_0 may increase to infinity as n gets larger. In our setting, it is equivalent to say that the cardinality of $\mathrm{pa}_j(\mathcal{D}_0)$ is less than s_0 for any $j = 2, \dots, p$, where $\mathcal{D}_0$ is the DAG corresponding to A_{0n}.

Condition (A3) is the well-known *beta-min* condition, which determines the lower bound for the nonzero signals. The beta-min condition has been used for the exact support recovery of the high-dimensional linear regression coefficients [Bühlmann and van de Geer (2011), Castillo, Schmidt-Hieber and van der Vaart (2015), Martin, Mess and Walker (2017), Wainwright (2009a), Yang, Wainwright and Jordan (2016)] as well as the high-dimensional DAGs [Cao, Khare and Ghosh (2019), Khare et al. (2016), Yu and Bien (2017)].

Condition (A4) restricts the number of nonzero elements in each column of A_{0n} to be smaller than s_0. In other words, the number of edges directed from any vertex is less than s_0. This assumption is required to deal with the posterior probability of $\|A_n - A_{0n}\|_1$. Note that if we consider only the banded structure for the Cholesky factor as in Yu and Bien (2017), conditions (A2) and (A4) automatically hold for some s_0.

Now we define a class of precision matrices

$$\mathcal{U}_p = \mathcal{U}_p(\epsilon_0, s_0, \alpha, C_{\mathrm{bm}}) = \{\Omega \in \mathcal{C}_p : \Omega \text{ satisfies (A1)–(A3)}\}.$$

In Section 3, we show that one can achieve the strong model selection consistency for any $\Omega_{0n} \in \mathcal{U}_p$. Furthermore, we derive the posterior convergence rates for A_{0n} and show that these are optimal or nearly optimal for the class $\mathcal{U}_p$ (or $\mathcal{U}_p$ without condition (A3)).

REMARK 2.2. Cao, Khare and Ghosh (2019) weakened the bounded eigenvalue condition (A1) by replacing a constant ϵ_0 with a sequence $\epsilon_{0,n}$, which can go to zero at certain rate. Our results also still hold under the similar weakened bounded eigenvalue condition with $\epsilon_{0,n}$, but it will sacrifice the other conditions. For example, by using a sequence $\epsilon_{0,n}$ in place of a fixed ϵ_0 in the proof of Theorem 3.1, one can see that $s_0 \log p \le Cn\epsilon_{0,n}^2$ for some $C > 0$ and the beta-min condition (A3) with $\epsilon_{0,n}$ in place of ϵ_0 are required.

3. Main results.

We introduce Condition (P) on the hyperparameters in the ESC prior (4), which is necessary for the results in this section. Note that this condition is for the hyperparameters of the prior distribution, which does not affect the true parameter space.

CONDITION (P). Assume that $\nu_0 = o(n)$, $c_1 = O(1)$, $c_2 \ge 2$ and $\gamma = O(1)$. For given positive constants $0 < \alpha < 1$ and $0 < \epsilon_0 < 1/2$ used in conditions (A1) and (A3), assume that $R_j = \lfloor n(\log p)^{-1}\{(\log n)^{-1} \vee c_3\}\rfloor$ for any $j = 2, \dots, p$ and some small constant $0 < c_3 < (\epsilon')^2\epsilon_0^2/\{128(1+2\epsilon_0)^2\}$, where $\epsilon' = \{(1-\alpha)/10\}^2$.

The condition $c_2 \ge 2$ is similar to the condition $\kappa \ge 2$ in Yang, Wainwright and Jordan (2016). Note that the constants c_1 and c_2 in the ESC prior control the row-wise sparsity of the Cholesky factor A_n: large values of them make the posterior prefer small values for $|S_j|$. Thus, the above condition means that we need certain amount of penalty on $|S_j|$ to achieve desirable asymptotic properties. The condition on R_j means that R_j is of order $n(\log p)^{-1}$ and smaller than $n(\log p)^{-1}(\epsilon')^2\epsilon_0^2/\{128(1+2\epsilon_0)^2\}$, so it can be replaced by the condition $R_j = \lfloor n(\log p)^{-1}(\epsilon')^2\epsilon_0^2/\{128(1+2\epsilon_0)^2\}\rfloor$. To assure $s_0 \le R_n$, we will assume that $s_0 \le n(\log p)^{-1}c_3/2$ later. In general, assuming $s_0 = O(n(\log p)^{-1})$ or even $s_0 = o(n(\log p)^{-1})$ is essential to prove theoretical properties such as selection consistency and convergence rates. However, it can be unrealistically small for some finite sample size n. More importantly, the quantity ϵ_0 is unknown in typical applications, so it is desirable to make the prior work for any choice of ϵ_0. Condition (P) argues that there is such a prior. We suggest choosing a small enough c_3 so that R_j can be regarded as $R_j = \lfloor n(\log p \cdot \log n)^{-1} \rfloor$ for finite samples.

REMARK 3.1. Yang, Wainwright and Jordan (2016) suggested a prior for the linear model similar to the ESC prior but for the mean vector of the prior $\pi(a_{S_j} \mid d_j, S_j)$, they used zero mean vector while we used $\widehat{a}_{S_j}$. There are two consequences from the use of the data-dependent mean $\widehat{a}_{S_j}$. First, we do not need an upper bound condition for $\|\mathbf{X}_{S_{0j}} a_{0,S_{0j}}\|_2$ or $\|a_{0,S_{0j}}\|_2$, while Yang, Wainwright and Jordan (2016) assumed $\|\mathbf{X}_{S_{0j}} a_{0,S_{0j}}\|_2 \le g d_{0j} \log p$, where $g = \gamma^{-1}$ in this paper. It is known that this type of condition is required if we use the Zellner's g-prior with zero mean [Shang and Clayton (2011)]. Second, to prove model selection consistency, Yang, Wainwright and Jordan (2016) assumed $g = p^{2c}$ for some $c \ge 1/2$ corresponding to $\gamma = p^{-2c}$ in our notation. This is the so-called information paradox of Zellner's g-priors [Liang et al. (2008)]. We do not require this condition and just assume $\gamma = O(1)$.

3.1. *Strong model selection consistency.* When the recovery of the DAG is of interest, it is desirable to use a Bayesian procedure that guarantees the strong model selection consistency. We show that the α-posterior warrants this property under mild conditions. As mentioned earlier, the Gaussian DAG model has an interpretation as a sequence of autoregressive model (3), which enables us to adopt the state-of-the-art techniques for the selection consistency of the regression coefficient in Martin, Mess and Walker (2017).

To use the results in Martin, Mess and Walker (2017), there are two main issues that need to be addressed. The first is the *restricted eigenvalue condition* for the design matrix. In our setting, the design matrices consist of columns of data matrix $\mathbf{X}_n$, thus each row follows a multivariate normal distribution. We show that under the bounded eigenvalue condition (A1), the restricted eigenvalue condition for any integer $R = o(n)$ automatically holds on some *large* set N^c having $\mathbb{P}_0$-probability tending to 1 (Lemma 6.1 in the Supplementary Material). A similar result appears in Narisetty and He (2014). The second issue is more challenging than the first. Martin, Mess and Walker (2017) considered only the known (fixed) residual variance case, which corresponds to the known d_{0j} case in our setting. The assumption on the known residual variance results in a relatively straightforward proof for selection consistency. We extended their techniques to the unknown residual variance case by applying (noncentral) chi-square concentration inequalities for the estimated residual variances $\widehat{d}_{S_j}$ for some index set S_j, which is motivated by Shin, Bhattacharya and Johnson (2018). It reveals that the ratio of the marginal posteriors $\pi_\alpha(S_j \mid \mathbf{X}_n)/\pi_\alpha(S_{0j} \mid \mathbf{X}_n)$ actually behaves like the ratio of the conditional posteriors given d_{0j}, $\pi_\alpha(S_j \mid d_{0j}, \mathbf{X}_n)/\pi_\alpha(S_{0j} \mid d_{0j}, \mathbf{X}_n)$, with $\mathbb{P}_0$-probability tending to 1, where S_{0j} is the index set for the nonzero elements in the jth row of A_{0n}.

We also note here that unlike the Lasso type (or its variants) of results with the random design matrix [Wainwright (2009b)], our theory does not require the *irrepresentable condition* on the true covariance matrix. For example, Yu and Bien (2017) and Khare et al. (2016) require the irrepresentable condition for the asymptotic properties of estimators in DAG models. See Section IV of Wainwright (2009b) for more details on the irrepresentable condition.

THEOREM 3.1 (Strong model selection consistency). *For given positive constants $0 < \alpha < 1$, $0 < \epsilon_0 < 1/2$, $C_{\rm bm} > c_2 + 2$ and an integer s_0, assume that Ω_{0n} satisfies conditions (A1), (A2) and (A3), that is, $\Omega_{0n} \in \mathcal{U}_p$. Consider model (1) and the ESC prior (4) with Condition (P). If $s_0 \log p \le nc_3/2$,*

$$\sup_{\Omega_{0n} \in \mathcal{U}_p} \mathbb{E}_0\big[\pi_\alpha(S_{A_n} \ne S_{A_{0n}} \mid \mathbf{X}_n)\big] = o(1).$$

The assumption $s_0 \log p = o(n)$ or $s_0 \log p \le cn$ for some constant $c > 0$ is widely used in the high-dimensional sparse covariance or precision matrix estimation literature. In Theorem 3.1, we assume less restrictive condition $s_0 \log p \le nc_3/2$, which automatically guarantees $s_0 \le R_j$ for all $j = 2, \dots, p$. Note that the constant c_3 is defined in Condition (P).

It is worthwhile to compare our result to those of Cao, Khare and Ghosh (2019), Yu and Bien (2017) and Khare et al. (2016). Note that in these works it is also assumed that the ordering of variables is known. Cao, Khare and Ghosh (2019) showed the strong model selection consistency using the hierarchical DAG-Wishart prior. They assumed variants of conditions (A1), (A2) and (A3). First, they relaxed condition (A1) by letting $\epsilon_{0,n} \to 0$ such that $(\log p/n)^{1/2 - 1/(2+k)} = o(\epsilon_{0,n}^4)$ for some $k > 0$, instead of a fixed $\epsilon_0 > 0$. Second, they assumed the same condition (A2) but further assumed $s_0^{2+k}\sqrt{\log p/n} = o(1)$ and $(\log p/n)^{k/(4k+8)} \log n = o(1)$ and considered only the DAGs with the total number of edges at most $8^{-1} s_0 (n/\log p)^{(1+k)/(2+k)}$, which can be restrictive. Note that, when $p \ge n$, it does not include the banded Cholesky factor having s_0 nonzero elements for each row. Third, they assumed somewhat strong beta-min condition compared with (A3), which requires $\min_{j,l: a_{0,jl} \ne 0} |a_{0,jl}|^2 \ge M_n^2 s_0^2 \epsilon_{0,n}^{-1} (\log p/n)^{1/(2+k)}$ for some $k > 0$ and some sequence $M_n \to \infty$. Thus, their assumptions on the tuple (n, p, s_0) as well as the parameter class are much more restrictive than ours, except for the bounded eigenvalue condition. Furthermore, the choice of hyperparameter in the hierarchical DAG-Wishart prior depends on the unknown sparsity parameter s_0, thus it is not adaptive to the unknown parameter. More specifically, the hyperparameter q_n in the hierarchical DAG-Wishart prior should be set at $q_n = s_0(\log p/n)^{1/(2+k)}$ for some $k > 0$ to achieve the strong model selection consistency.

Yu and Bien (2017) suggested a penalized maximum likelihood estimation for the Cholesky factor of the precision matrix and proved the exact signed support recovery under the condition $\rho^{-2}\|D_{0n}\|\epsilon_0^{-1}(12\pi^2 s_0 + 32)\log p < n$. They considered the class of precision matrices satisfying condition (A1) and having a banded structure with the row-specific bandwidths $s_{0j} = |S_{0j}|$ such that $a_{0,jl} = 0$ for all $1 \le l < j - s_{0j}$ and $2 \le j \le p$. Thus, by taking $s_0 = \max_j s_{0j}$, their class satisfies conditions (A2) and (A4). They also assumed the beta-min condition, $\min_{j,l:a_{0,jl}\neq 0} |d_{0j}^{-1/2} a_{0,jl}| \ge 8\rho^{-1}\sqrt{2\|D_{0n}\|\log p/n}(4\max_j \|\Sigma_{0n,S_{0j}}^{-1}\|_\infty + 5\epsilon_0^{-1})$. In general, it holds that $\|\Sigma_{0n,S_{0j}}^{-1}\|_\infty = O(s_{0j}^{1/2})$ without further assumption, thus the above condition implies that the minimum nonzero $|d_{0j}^{-1/2} a_{0,jl}|$ is bounded below by $\sqrt{s_0 \log p/n}$ with respect to a constant multiple, thus stronger than condition (A3). Furthermore, they assumed the irrepresentable condition

$$\max_{2\le j\le p} \max_{\substack{1\le l\le j\\ l\in S_{0j}^c}} \|(\Sigma_{0n})_{(l,S_{0j})}(\Sigma_{0n,S_{0j}})^{-1}\|_1 \le \frac{6(1-\rho)}{\pi^2}$$

for some constant $\rho \in (0, 1]$. Therefore, they only considered the banded Cholesky factor and used somewhat strong beta-min condition as well as the irrepresentable condition. However, the comparison with our result (Theorem 3.1) is not straightforward because their exact signed support recovery property is stronger than the selection consistency proved in Theorem 3.1.

Khare et al. (2016) proved the signed support recovery property of the convex sparse Cholesky selection (CSCS) method when the data vectors $X_1, \dots, X_n$ are random sample from a sub-Gaussian distribution. They assumed condition (A1) as well as the (stronger) variants of conditions (A2) and (A3): they assumed $\sum_{j=2}^p s_{0j} = o(n/\log n)$ (which is stronger than $s_0 \log p \le nc_3/2$) and $\min_{j,l:a_{0,jl}\neq 0} |a_{0,jl}|^2 \ge M_n s_0^2 \log n/n$ for some $M_n \to \infty$. Furthermore, they considered only the moderate high-dimensional setting, that is, $p = O(n^c)$ for some constant $c > 0$. They also required the irrepresentable condition similar to those in Yu and Bien (2017).

3.2. *Posterior convergence rates for Cholesky factors.* In this subsection, we derive the posterior convergence rates for the Cholesky factors in two different scenarios depending on the existence of the beta-min condition (A3). At first, under the beta-min condition, we show the posterior convergence rates and minimax lower bounds with respect to the matrix ℓ_∞ norm and Frobenius norm. The obtained posterior convergence rates are *nearly* minimax, and become exactly minimax if $\log p = O(s_0)$ and $\log j = O(s_{0j})$ for all $j = 2, \dots, p$. We also derive the posterior convergence rate and minimax lower bound with respect to the matrix

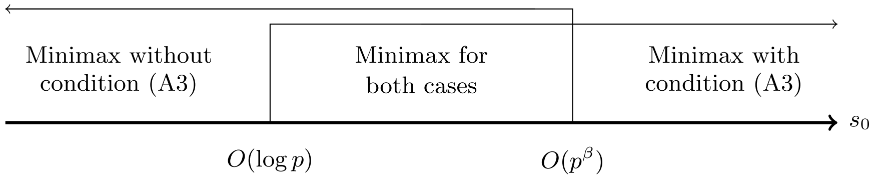

FIG. 1. *For a given* $0 < \beta < 1$, *it describes the range for* s_0 *in which the minimax rate for the Cholesky factor can be obtained.* (A3) *means the beta-min condition.*

ℓ_∞ norm without the beta-min condition. The obtained posterior convergence rate turns out to be nearly minimax, and it will be exactly minimax if $s_0 \le p^\beta$ for some $0 < \beta < 1$. Note that regardless of the relation between s_0 and p, at least one of the scenarios achieves the minimax rate. Especially, we attain the minimax rate for both scenarios if $C \log p \le s_0 \le p^\beta$ for some constant $C > 0$. Figure 1 describes the range for s_0 in which the minimax rate can be obtained.

3.2.1. *Posterior convergence rates for Cholesky factors under Beta-min condition.* Define $\widehat{A}_n = (\widehat{a}_{jl})$, where $(\widehat{a}_{jl})_{l \in S_{0j}} = \widehat{a}_{S_{0j}}$ and $(\widehat{a}_{jl})_{l \in S_{0j}^c} = 0$. Thus, $\widehat{A}_n$ is the empirical estimates of A_{0n} with true support $S_{A_{0n}}$. To obtain the posterior convergence rate for the Cholesky factor, we use a divide and conquer strategy that is similar to Lee and Lee (2017), Lee and Lee (2018). Specifically, we decompose the posterior contraction probability into two parts as follows:

$$\begin{aligned} &\pi_\alpha\big(\|A_n - A_{0n}\| \ge 2\epsilon_n' \mid \mathbf{X}_n\big) \\ &\quad \le \pi_\alpha\big(\|A_n - \widehat{A}_n\| \ge \epsilon_n' \mid \mathbf{X}_n\big) + \pi_\alpha\big(\|\widehat{A}_n - A_{0n}\| \ge \epsilon_n' \mid \mathbf{X}_n\big) \end{aligned} \tag{7}$$

for some positive sequence ϵ_n'. As in Section 3.1, we concentrate on a *large* set N^c allowing us to handle the posterior contraction probability easily. The first part of the right-hand side of (7) describes how the posterior distribution concentrates around the empirical estimate $\widehat{A}_n$. We use the selection consistency result in Theorem 3.1, and we focus only on the set $S_{A_n} = S_{A_{0n}}$. It enables us to deal with the posterior distribution for A_n easily, but with a cost of the beta-min condition (A3) which is usually not essential for the convergence rate results. Through the posterior distribution (6) given $S_{A_n} = S_{A_{0n}}$, we can obtain the contraction probability for $\|A_n - \widehat{A}_n\|$ using the concentration inequality for the chi-square random variables. By taking expectation to the second part of the right-hand side of (7), it gives the contraction probability of $\widehat{A}_n$, $\mathbb{P}_0[\|\widehat{A}_n - A_{0n}\| \ge \epsilon_n']$.

THEOREM 3.2 (Posterior convergence rates for A_{0n} with beta-min condition). *For given positive constants* $0 < \alpha < 1$, $0 < \epsilon_0 < 1/2$, $C_{\text{bm}} > c_2 + 2$ *and an integer* s_0, *assume that* Ω_{0n} *satisfies conditions* (A1), (A2) *and* (A3), *that is*, $\Omega_{0n} \in \mathcal{U}_p$.

Consider model (1) *and the ESC prior* (4) *with Condition* (P). *If* $s_0 \log p = o(n)$,

$$\sup_{\Omega_{0n}\in\mathcal{U}_p} \mathbb{E}_0\left[\pi_\alpha\left(\|A_n - A_{0n}\|_\infty \ge K_{\text{chol}}\sqrt{s_0}\left(\frac{s_0+\log p}{n}\right)^{1/2} \middle| \mathbf{X}_n\right)\right] = o(1),$$

$$\sup_{\Omega_{0n}\in\mathcal{U}_p} \mathbb{E}_0\left[\pi_\alpha\left(\|A_n - A_{0n}\|_F^2 \ge K_{\text{chol}}\frac{\sum_{j=2}^p (s_{0j}+\log j)}{n} \middle| \mathbf{X}_n\right)\right] = o(1)$$

for some constant $K_{\text{chol}} > 0$.

Khare et al. (2016) obtained the convergence rate $\sum_{j=2}^p s_{0j}\lambda_n$ for estimating the Cholesky factor under the spectral norm in a moderately high-dimensional setting, that is, $p = O(n^c)$ for some constant $c > 0$, where λ_n is the tuning parameter in CSCS method. They also assumed condition (A1) as well as the (stronger) variants of conditions (A2) and (A3) as described in Section 3.1. Because they assumed $\sqrt{\sum_{j=2}^p s_{0j}\log p/n} = o(\lambda_n)$, $\sum_{j=2}^p s_{0j}\lambda_n$ is strictly slower than $(\sum_{j=2}^p s_{0j})^{3/2}\sqrt{\log p/n}$ in term of the rate, which implies that their convergence rate is slower than the posterior convergence rate obtained in this paper.

In fact, it turns out that the posterior convergence rates in Theorem 3.2 are nearly optimal. Theorem 3.3 describes that the rates of the frequentist minimax lower bounds for the class $\mathcal{U}_p$, which are of independent interests. Note that the rates of Theorem 3.2 are exactly optimal if $\log p = O(s_0)$ and $\log j = O(s_{0j})$ for all $j = 2, \ldots, p$ matching the minimax rates of Theorem 3.3. The key idea for proving the minimax lower bounds is to break down the model (1) into a set of linear regression models.

THEOREM 3.3 (Minimax lower bounds for A_{0n} with beta-min condition). *For given positive constants* $0 < \alpha < 1$, ϵ_0, C_{bm} *and an integer* s_0, *assume that* Ω_{0n} *satisfies conditions* (A1), (A2) *and* (A3), *that is*, $\Omega_{0n} \in \mathcal{U}_p$. *Consider model* (1). *Then*

$$\inf_{\widehat{A}_n}\sup_{\Omega_{0n}\in\mathcal{U}_p} \mathbb{E}_0\|\widehat{A}_n - A_{0n}\|_\infty \ge c\cdot\frac{s_0}{\sqrt{n}},$$

$$\inf_{\widehat{A}_n}\sup_{\Omega_{0n}\in\mathcal{U}_p} \mathbb{E}_0\|\widehat{A}_n - A_{0n}\|_F^2 \ge c\frac{\sum_{j=2}^p s_{0j}}{n}$$

for some constant $c > 0$, *where the infimum is taken over all possible estimators* $\widehat{A}_n$.

3.2.2. *Posterior convergence rates for Cholesky factors without Beta-min condition.* For a given positive constant ϵ_0 and a sequence of positive integers s_0, we define a class of precision matrices,

$$\mathcal{U}_p^0 = \mathcal{U}_p^0(\epsilon_0, s_0) = \{\Omega \in \mathcal{C}_p : \Omega \text{ satisfies (A1) and (A2)}\}.$$

Note that in the definition of $\mathcal{U}_p^0$, we *do not require the beta-min condition.* Theorem 3.4 gives the posterior convergence rate for the class $\mathcal{U}_p^0$. For the Theorem 3.4, we use the ESC prior (4) but let $d_j \sim \mathrm{IG}(\nu_0/2, \nu_0')$ for some constant $\nu_0' > 0$ instead of $\pi(d_j) \propto d_j^{-\nu_0/2-1}$. We call this the modified ESC (MESC) prior. As mentioned before, Theorems 3.1, 3.2 and 3.6 in Section 3 also hold for the MESC prior, but we describe Theorems 3.1, 3.2 and 3.6 with the ESC prior for the computational convenience.

We consider the denominator and numerator of the posterior probability $\pi_\alpha(\|A_n - A_{0n}\|_\infty \ge \epsilon_n')$ separately, for some positive sequence ϵ_n'. For any $j = 2, \dots, p$, let $R_{nj}(a_j, d_j) = L_{nj}(a_j, d_j)/L_{nj}(a_{0j}, d_{0j})$ be the likelihood ratio, where

$$L_{nj}(a_j, d_j) = (2\pi d_j)^{-n/2} \exp\{-\|\tilde{X}_j - \tilde{Z}_j a_j\|_2^2/(2d_j)\}.$$

Dealing with the likelihood ratio $R_{nj}(a_j, d_j)$ is one of the main tasks for proving Theorem 3.4. Lemma 7.1, Lemma 7.2 and Lemma 7.3 in the Supplementary Material describe how we can deal with the likelihood ratio $R_{nj}(a_j, d_j)$ in the denominator and numerator.

THEOREM 3.4 (Posterior convergence rate for A_{0n} without beta-min condition). *For a given positive constant* $0 < \alpha < 1$, $0 < \epsilon_0 < 1/2$ *and an integer* s_0, *assume that* Ω_{0n} *satisfies conditions* (A1) *and* (A2), *that is*, $\Omega_{0n} \in \mathcal{U}_p^0$. *Consider model* (1) *with the MESC prior with Condition* (P). *If* $s_0 \log p = o(n)$ *and* $\nu_0 = O(1)$, *then*

$$\sup_{\Omega_{0n} \in \mathcal{U}_p^0} \mathbb{E}_0\left[\pi_\alpha\left(\|A_n - A_{0n}\|_\infty \ge K_{\mathrm{chol}}' s_0 \left(\frac{\log p}{n}\right)^{1/2} \middle| \mathbf{X}_n\right)\right] = o(1)$$

for some constant $K_{\mathrm{chol}}' > 0$.

Yu and Bien (2017) obtained the convergence rate $\max_j \|\Sigma_{0n,S_{0j}}^{-1}\|_\infty \cdot \|A_{0n}\|_\infty s_0 \sqrt{\log p/n} + \max_j \|\Sigma_{0n,S_{0j}}^{-1}\|_\infty^2 s_0^2 \log p/n$ for the Cholesky factor with respect to the matrix ℓ_∞ norm. As stated before, they assumed condition (A1), the banded Cholesky factor structure (which corresponds to conditions (A2) and (A4) in this paper) and the irrepresentable condition. Note that their convergence rate coincides with ours only if $\|A_{0n}\|_\infty$ and $\max_j \|\Sigma_{0n,S_{0j}}^{-1}\|_\infty$ are bounded and $s_0^2 \log p = O(n)$.

To the best of our knowledge, it is the first result on the posterior convergence rate for the high-dimensional sparse Cholesky factor without the beta-min condition. Interestingly, the obtained posterior convergence rate is the same with the minimax convergence rate for the s_0-sparse coefficient vector in the regression models when $s_0 \le p^\beta$ for some $0 < \beta < 1$. Note that the condition $s_0 \le p^\beta$ is not restrictive in the high-dimensional setting, because this condition is met if $n \le p^\beta$. Theorem 3.5 confirms that the above posterior convergence rate is nearly minimax for any $\Omega_{0n} \in \mathcal{U}_p^0$. Similar to Theorem 3.3, the key idea for proving Theorem 3.5 is to break down the model into a set of linear regression models.

THEOREM 3.5 (Minimax lower bound for A_{0n} without beta-min condition). *For a given constant ϵ_0 and an integer s_0, assume that Ω_{0n} satisfies conditions (A1) and (A2), that is, $\Omega_{0n} \in \mathcal{U}_p^0$. Consider model (1). Then*

$$\inf_{\widehat{A}_n} \sup_{\Omega_{0n} \in \mathcal{U}_p^0} \mathbb{E}_0 \|\widehat{A}_n - A_{0n}\|_\infty \geq c \cdot s_0 \left(\frac{\log(p/s_0)}{n} \right)^{1/2}$$

for some constant $c > 0$.

REMARK 3.2. If we assume $s_0 \leq p^\beta$ for some $0 < \beta < 1$, then $\log(p/s_0)$ has the same rate with $\log p$, and the rate of the mininax lower bound in Theorem 3.5 becomes $s_0\sqrt{\log p/n}$. This assumption is reasonable in the high-dimensional setting.

3.3. *Posterior convergence rates for precision matrices.* In this subsection, we derive the posterior convergence rates for the precision matrices with respect to various matrix norms. Define $\widehat{\Omega}_n = (I_p - \widehat{A}_n)^T \widehat{D}_n^{-1}(I_p - \widehat{A}_n)$, where $\widehat{A}_n$ and $\widehat{D}_n = \operatorname{diag}(\widehat{d}_{S_{0j}})$ are the empirical estimates of A_{0n} and D_{0n} with the true support $S_{A_{0n}}$. Similar to the previous subsection, we use the divide and conquer strategy to deal with the posterior probability. For the recovery of $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1}(I_p - A_{0n})$, we further assume condition (A4). For given positive constants ϵ_0, C_{bm} and a sequence of positive integers s_0, define the parameter class as follows:

$$\mathcal{U}_p^* = \mathcal{U}_p^*(\epsilon_0, s_0, \alpha, C_{\mathrm{bm}}) = \{\Omega \in \mathcal{C}_p : \Omega \text{ satisfies (A1)–(A4)}\}.$$

Theorem 3.6 shows the posterior convergence rates for the precision matrix with respect to the spectral norm and matrix ℓ_∞ norm.

THEOREM 3.6 (Posterior convergence rates for Ω_{0n}). *For given positive constants $0 < \alpha < 1$, $0 < \epsilon_0 < 1/2$, $C_{\mathrm{bm}} > c_2 + 2$ and an integer s_0, assume that Ω_{0n} satisfies conditions (A1)–(A4), that is, $\Omega_{0n} \in \mathcal{U}_p^*$. Consider model (1) and the ESC prior (4) with Condition (P) and $\nu_0^2 = O(n \log p)$. If $s_0^{3/2}(s_0 + \log p) = o(n)$, then*

$$\sup_{\Omega_{0n} \in \mathcal{U}_p^*} \mathbb{E}_0 \left[\pi_\alpha \left(\|\Omega_n - \Omega_{0n}\| \geq K_{\mathrm{conv}} s_0^{3/4} \left(\frac{s_0 + \log p}{n} \right)^{1/2} | \mathbf{X}_n \right) \right] = o(1),$$

and, if $s_0(s_0 + \log p) = o(n)$, then

$$\sup_{\Omega_{0n} \in \mathcal{U}_p^*} \mathbb{E}_0 \left[\pi_\alpha \left(\|\Omega_n - \Omega_{0n}\|_\infty \geq K_{\mathrm{conv}} \cdot \|I_p - A_{0n}\|_\infty s_0 \left(\frac{s_0 + \log p}{n} \right)^{1/2} \middle| \mathbf{X}_n \right) \right]$$
$$= o(1)$$

for some constant $K_{\mathrm{conv}} > 0$.

It is worthwhile to compare our result to other existing results. Cao, Khare and Ghosh (2019) obtained the posterior convergence rate, $s_0^2 \epsilon_{0,n}^{-2} \sqrt{\log p/n}$, for the precision matrix with respect to the spectral norm. As discussed in Section 3.1, they assumed variants of conditions (A1), (A2) and (A3). They further assumed the condition (A4). Although they did not state clearly that condition (A4) was used, this condition is necessary to use Lemma 3.1 of Xiang, Khare and Ghosh (2015) in their proof. If we assume the bounded eigenvalue condition (A1), their convergence rate becomes $s_0^2 \sqrt{\log p/n}$, which is slower than the convergence rate in Theorem 3.6. Note that they assumed $s_0^{2+k} \sqrt{\log p/n} = o(1)$ for some constant $k > 0$, which is stronger than our assumption $s_0^{3/2}(s_0 + \log p) = o(n)$. Thus, we obtain the faster posterior convergence rates under more general condition on the tuple (n, p, s_0) and parameter class, except for the bounded eigenvalue condition.

Yu and Bien (2017) considered the parameter class they used to prove the strong model selection consistency, but dropped the beta-min condition. They derived the convergence rate

$$\max_j \|(\Sigma_{0n,S_{0j}})^{-1}\|_\infty \|D_{0n}^{-1/2}(I_p - A_{0n})\|_\infty s_0 \left(\frac{\log p}{n}\right)^{1/2}$$

for the precision matrix with respect to the matrix ℓ_∞ norm. Note that this convergence rate depends on the rate of $\max_j \|(\Sigma_{0n,S_{0j}})^{-1}\|_\infty \|D_{0n}^{-1/2}(I_p - A_{0n})\|_\infty$. In general, it holds that $\max_j \|(\Sigma_{0n,S_{0j}})^{-1}\|_\infty = O(s_{0j}^{1/2})$. Thus, their convergence rate is slower than the posterior convergence rate in Theorem 3.6, without a further assumption on Σ_{0n} guaranteeing $\max_j \|(\Sigma_{0n,S_{0j}})^{-1}\|_\infty = O(\sqrt{(s_0/\log p)} + 1)$.

4. Numerical results. The use of the ESC prior not only guarantees optimal or near optimal asymptotic properties but also allows us to conduct the posterior inference easily. In this section, we carry out simulation studies to illustrate the model selection performance of our method. For the comparison, we chose state-of-the-art methods for high-dimensional sparse DAG models and measured the performance of each method. The simulation study confirms that our ESC prior outperforms the other existing methods.

4.1. *Metropolis–Hastings algorithm.* Recall that, by (6), the marginal posterior distribution for $S_j \subseteq \{1, \dots, j-1\}$ can be derived analytically as

$$\pi_\alpha(S_j \mid \mathbf{X}_n) \propto \pi_j(S_j)\left(1 + \frac{\alpha}{\gamma}\right)^{-\frac{|S_j|}{2}} (\widehat{d}_{S_j})^{-\frac{\alpha n + \nu_0}{2}} \tag{8}$$

for all $j = 2, \dots, p$, up to some normalizing constants. Thus, we can run the Rao–Blackwellized Metropolis–Hastings algorithm for each $j = 2, \dots, p$ in parallel.

Here we briefly summarize the algorithm used for the inference, where L is the number of posterior samples:

Run the following steps for $j = 2, \dots, p$:

(a) Set the initial value $S_j^{(1)}$.

(b) For each $l = 2, \dots, L$,

i. sample $S_j^{\mathrm{new}} \sim q(\cdot \mid S_j^{(l-1)})$;

ii. compute the acceptance probability

$$p_{\mathrm{acc}} = \min\left\{1, \frac{\pi_\alpha(S_j^{new} \mid \mathbf{X}_n) q(S_j^{(l-1)} \mid S_j^{new})}{\pi_\alpha(S_j^{(l-1)} \mid \mathbf{X}_n) q(S_j^{new} \mid S_j^{(l-1)})}\right\},$$

and set $S_j^{(l)} = S_j^{\mathrm{new}}$ with probability p_{acc}; otherwise, set $S_j^{(l)} = S_j^{(l-1)}$.

We chose the kernel $q(S' \mid S)$ which forms a new set S' by changing a randomly selected nonzero component to 0 with probability 0.5 or by changing a randomly selected zero component to 1 with probability 0.5.

The marginal posterior for S_j is controlled by the prior $\pi_j(S_j)$, the penalty term $(1 + \alpha/\gamma)^{-|S_j|/2}$ and the estimated residual variance $\widehat{d}_{S_j}$. The data favor to minimize the estimated residual while the prior and penalty term give more weight to the simpler models. The marginal posterior of S_j will find the model that balances data tracking and model complexity.

To use the Metropolis–Hastings algorithm, we need to choose the tuning parameters. Apart from the impact on theoretical results, the choice of tuning parameters also influences the practical performance. As Martin, Mess and Walker (2017) suggested, we set $\alpha = 0.999$ to mimic the Bayesian model with the ordinary likelihood. In the simulation study, as long as $1 - \alpha$ is close to zero, the performance was not dependent on the choice of α. The hyperparameters were chosen as $\gamma = 0.1$, $\nu_0 = 0$, $c_1 = 0.0005$ and $c_2 = 2$ to satisfy Condition (P).

4.2. *Simulation setting.* For the simulation study, we considered the sparse Cholesky settings similar to those used in Khare et al. (2016). We randomly chose 3% or 4% of the lower triangular entries of the Cholesky factor A_{0n} and sampled their values from a uniform distribution on $[-0.7, -0.3] \cup [0.3, 0.7]$. The remaining entries were set to zero. The entries of the diagonal matrix D_{0n} were sampled from a uniform distribution on $[2, 5]$. Given the precision matrix $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1} (I_p - A_{0n})$, the data sets were generated from the multivariate normal distribution $N_p(0, \Omega_{0n}^{-1})$ with $(n = 100, p = 300)$ and $(n = 200, p = 500)$.

4.3. *Other competing methods.* We compared the model selection performance of our method with those of other existing methods: the empirical Bayes (EB) procedure in Martin, Mess and Walker (2017), which we will denote as

EB.M, hierarchical DAG-Wishart (DAG-W) prior [Cao, Khare and Ghosh (2019)] and convex sparse Cholesky selection (CSCS) [Khare et al. (2016)].

1. (EB.M): Because EB.M is originally proposed for the regression coefficient estimation, it can be applied independently to estimate each a_{0j} for $j = 2, \dots, p$. We set the hyperparameters α, γ, c_1 and c_2 to be the same as those in our setting for a fair comparison. Note that Martin, Mess and Walker (2017) used $\gamma = 0.001$, $c_1 = 1$ and $c_2 = 0.05$ in their simulation study, but in our simulations, these choices did not yield better results: they tended to make unacceptably large FDR values. The key difference between our method and EB.M is on how to infer the diagonal matrix D_n. Martin, Mess and Walker (2017) suggested plugging in the cross-validation based Lasso residual sum of squares estimate [Reid, Tibshirani and Friedman (2016)] of d_{0j}, while we impose a prior on d_j and integrate it out to obtain the marginal posterior for S_j. Thus, EB.M ignores the uncertainty of d_j and replaces it with a plug-in estimate.

2. (DAG-W): The hierarchical DAG-Wishart prior [Cao, Khare and Ghosh (2019)] enables one to calculate the marginal posterior for the DAG analytically. Note that, in Cao, Khare and Ghosh (2019), they conducted *log-posterior score search algorithm* instead of Markov chain Monte Carlo (MCMC) algorithm. Basically, they generated sets of candidate graphs by using frequentist approaches and thresholding the modified Cholesky factor of $(n^{-1}\mathbf{X}_n^T\mathbf{X}_n + 0.5I_p)^{-1}$, and the graph which maximizes the log-posterior was chosen as the final estimate. In our simulation study, we implemented the log-posterior score search algorithm as well as Metropolis–Hastings algorithm, using the marginal posterior for the DAG, for a comprehensive comparison. For the implementation, we set the shape parameters at $\alpha_j(\mathcal{D}) = S_j + 10$ and the scale matrix at $U_n = I_p$ as they suggested, where $\mathcal{D}$ is the DAG corresponding to $\{S_j\}_{j=2}^p$. The critical part is the choice of the hyperparameter q_n, which is the individual edge probability. It was shown that the choice of $q_n = e^{-\eta_n n}$ leads to strong model selection consistency, where $\eta_n = s_0(\log p/n)^{1/(2+k)}$ for some $k > 0$. Thus, the theoretical choice of q_n depends on the unknown parameter s_0 and constant $k > 0$. Furthermore, even with $s_0 = 1$ and $k = 0$, the resulting q_n is too small, which does not allow the posterior to explore the model space efficiently. We observed that the choice $q_n = e^{-\eta_n n}$ makes the posterior stuck in very small size models and not able to detect the true model. For example, for the setting ($n = 100$, $p = 300$) with the sparsity 3%, the corresponding posterior with $q_n = e^{-\eta_n n}$ concluded that the true Cholesky factor is a zero matrix, that is, it never selected any nonzero variable. Thus, in our simulation study, we conducted the simulation only for two choices, $q_n = 0.01$ and $q_n = 0.001$, although they might not guarantee the strong model selection consistency. For the log-posterior score search, we chose $q_n = e^{-\eta_n n}$ as in Cao, Khare and Ghosh (2019).

3. (CSCS): We chose the CSCS method [Khare et al. (2016)] as a state-of-the-art frequentist competitor. The tuning parameter λ_n in the CSCS method was

selected by the Bayesian Information Criterion (BIC)-like measure which is defined in Section 2.3 of Khare et al. (2016). We calculated the values of BIC-like measure for λ_n from 0.1 to 5.1 with an increment of 0.1. The value of λ_n minimizing the BIC-like measure was chosen for the estimation.

4.4. *Results.* We ran the Metropolis–Hastings algorithm for each data set to conduct posterior inferences. Every MCMC chain ran for 24,000 iterations with a burn-in period of 4000, so we obtained 20,000 posterior samples. We used the models selected by the CSCS method as the initial states for MCMC chains. We constructed the final model by collecting indices with inclusion probabilities, $\pi(a_{jl} \neq 0 \mid \mathbf{X}_n)$, exceeding 0.5.

To measure the model selection performance, the number of errors, false discovery rate (FDR), true positive rate (TPR) and inclusion probabilities were reported. We calculated the mean inclusion probability for zero entries in A_{0n} and denote it by $\bar{p}_0$. Similarly, the mean inclusion probability for nonzero entries in A_{0n} is denoted by $\bar{p}_1$. More specifically, we calculated

$$\bar{p}_0 = \frac{1}{\sum_{j=2}^p (j-1-s_{0j})} \sum_{j=2}^p \sum_{l \notin S_{0j}} \pi(a_{jl} \neq 0 \mid \mathbf{X}_n),$$

$$\bar{p}_1 = \frac{1}{\sum_{j=2}^p s_{0j}} \sum_{j=2}^p \sum_{l \in S_{0j}} \pi(a_{jl} \neq 0 \mid \mathbf{X}_n).$$

The simulation results are summarized in Table 1. The ESC prior performs generally better than the other competing methods. The EB.M works reasonably well, but the overall performance is worse than that of ESC prior. The DAG-Wishart prior tends to have low TPR and mean inclusion probability $\bar{p}_1$ when $q_n = 0.001$. Note that when $q_n = 0.01$, which is chosen to be close to the unknown true sparsity level, the DAG-Wishart prior performs reasonably well, but the ESC prior still works better. However, the true sparsity is in general unknown, so fitting q_n close to the true sparsity is a challenging task in practice. The log-posterior score search algorithm for DAG-Wishart is computationally efficient even for large p, but tends to have low FDR as well as TPR in our settings. The CSCS method has high TPR values, but at the same time, it has high FDR values. Thus, from the simulation study, we confirm that our ESC prior not only has nice theoretical properties but also practically outperforms the other existing methods.

Acknowledgment. We thank Ryan Martin for helpful discussions about the techniques for proving selection consistency. We would like to thank two referees for their valuable comments which have led to improvements of an earlier version of the paper. We would also like to thank Kshitij Khare and Syed Rahman for sharing with us their code to implement the CSCS method [Khare et al. (2016)]. Kyoungjae Lee thanks Xuan Cao for sharing her code to implement the log-posterior score search algorithm and helpful discussions about the DAG-Wishart prior.

TABLE 1

ESC, EB.M, DAG-W and CSCS denote our method (empirical sparse Cholesky prior), the empirical Bayes procedure proposed by Martin, Mess and Walker (2017), the hierarchical Bayesian model using DAG-Wishart prior (Cao, Khare and Ghosh (2019)) and Convex Sparse Cholesky Selection (Khare et al. (2016)), respectively. Sp: sparsity; FDR: false discovery rate; TPR: true positive rate; $\bar{p}_0$: the mean inclusion probability for zero entries in A_{0n}; $\bar{p}_1$: the mean inclusion probability for nonzero entries in A_{0n}

(n, p, Sp)	Method	# of errors	FDR	TPR	$\bar{p}_0$	$\bar{p}_1$
(100, 300, 3%)	ESC	264	0.0361	0.8349	0.0071	0.8321
	EB.M	419	0.1083	0.7836	0.0041	0.7828
	DAG-W($q_n = 0.01$)	285	0.0208	0.8052	0.0024	0.8036
	DAG-W($q_n = 0.001$)	462	0.0122	0.6647	0.0006	0.6688
	DAG-W(log-score)	1194	0.0065	0.1130	·	·
	CSCS	2188	0.6433	0.7799	·	·
(100, 300, 4%)	ESC	389	0.0494	0.8261	0.0084	0.8194
	EB.M	325	0.0347	0.7866	0.0020	0.7815
	DAG-W($q_n = 0.01$)	422	0.0295	0.7887	0.0032	0.7873
	DAG-W($q_n = 0.001$)	644	0.0216	0.6555	0.0011	0.6556
	DAG-W(log-score)	1619	0.0056	0.0981	·	·
	CSCS	4025	0.7766	0.8045	·	·
(200, 500, 3%)	ESC	103	0.0118	0.9842	0.0039	0.9796
	EB.M	212	0.0075	0.9506	0.0005	0.9509
	DAG-W($q_n = 0.01$)	98	0.0049	0.9786	0.0010	0.9773
	DAG-W($q_n = 0.001$)	182	0.0022	0.9535	0.0002	0.9519
	DAG-W(log-score)	4285	0.0000	0.1412	·	·
	CSCS	10,214	0.7397	0.9388	·	·
(200, 500, 4%)	ESC	153	0.0061	0.9754	0.0043	0.9650
	EB.M	281	0.0038	0.9473	0.0005	0.9457
	DAG-W($q_n = 0.01$)	163	0.0041	0.9713	0.0011	0.9684
	DAG-W($q_n = 0.001$)	295	0.0017	0.9425	0.0002	0.9416
	DAG-W(log-score)	4341	0.0000	0.1301	·	·
	CSCS	14,632	0.7550	0.9285	·	·

SUPPLEMENTARY MATERIAL

Minimax Posterior Convergence Rates and Model Selection Consistency in High-dimensional DAG Models based on Sparse Cholesky Factors (DOI: 10.1214/18-AOS1783SUPP; .pdf). We present the proofs for the main results and other auxiliary results.

REFERENCES

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

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

BEN-DAVID, E., LI, T., MASSAM, H. and RAJARATNAM, B. (2015). High dimensional Bayesian inference for Gaussian directed acyclic graph models. Available at arXiv:1109.4371v5.

BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Bayesian fractional posteriors. *Ann. Statist.*. **47** 39–66. MR3909926

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

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761

CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. MR3476606

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

CAI, T. T. and YUAN, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CAI, T. T. and ZHOU, H. H. (2012a). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22** 1319–1349. MR3027084

CAI, T. T. and ZHOU, H. H. (2012b). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

CAO, X., KHARE, K. and GHOSH, M. (2019). Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *Ann. Statist.* **47** 319–348. MR3909935

CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

GAO, C. and ZHOU, H. H. (2015). Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** 785–818. MR3325710

GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979

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

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

KHARE, K., OH, S., RAHMAN, S. and RAJARATNAM, B. (2016). A convex framework for high-dimensional sparse Cholesky based covariance estimation. Preprint. Available at arxiv:1610.02436.

LEE, K. and LEE, J. (2017). Estimating large precision matrices via modified cholesky decomposition. Available at arXiv:1707.01143.

LEE, K. and LEE, J. (2018). Optimal Bayesian minimax rates for unconstrained large covariance matrices. *Bayesian Anal.* **13** 1211–1229. MR3855369

LEE, K., LEE, J. and LIN, L. (2019). Supplement to "Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors." DOI:10.1214/18-AOS1783SUPP.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879

MARTIN, R. and WALKER, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. MR3273623

MILLER, J. W. and DUNSON, D. B. (2018). Robust Bayesian inference via coarsening. *J. Amer. Statist. Assoc.* DOI:10.1080/01621459.2018.1469995.

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

PATI, D., BHATTACHARYA, A., PILLAI, N. S. and DUNSON, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. MR3210997

REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2016). A study of error variance estimation in Lasso regression. *Statist. Sinica* **26** 35–67. MR3468344

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

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

ROVERATO, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87** 99–112. MR1766831

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

SHANG, Z. and CLAYTON, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *J. Statist. Plann. Inference* **141** 3463–3474. MR2817355

SHIN, M., BHATTACHARYA, A. and JOHNSON, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* **28** 1053–1078. MR3791100

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

SYRING, N. A. and MARTIN, R. (2016). *Scaling the Gibbs Posterior Credible Regions*. Preprint. Available at arxiv:1509.00922.

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

WAINWRIGHT, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

WAINWRIGHT, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. MR1872068

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

YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44** 2497–2532. MR3576552

YU, G. and BIEN, J. (2017). Learning local dependence in ordered data. *J. Mach. Learn. Res.* **18** 42. MR3655307

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*. *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

K. LEE
INHA UNIVERSITY
100 INHA-RO, MICHUHOL-GU
INCHEON 22212
SOUTH KOREA
E-MAIL: leekjstat@gmail.com

J. LEE
DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
1 GWANAK-RO, GWANAK-GU
SEOUL 08826
SOUTH KOREA
E-MAIL: leejyc@gmail.com

L. LIN
DEPARTMENT OF APPLIED AND COMPUTATIONAL
MATHEMATICS AND STATISTICS
UNIVERSITY OF NOTRE DAME
NOTRE DAME, INDIANA 46556
USA
E-MAIL: lizhen.lin@nd.edu