Bayesian Analysis (2023) **18**, Number 1, pp. 25–47

Scalable Bayesian High-dimensional Local Dependence Learning[*]

Kyoungjae Lee[†] and Lizhen Lin[‡]

Abstract. In this work, we propose a scalable Bayesian procedure for learning the local dependence structure in a high-dimensional model where the variables possess a natural ordering. The ordering of variables can be indexed by time, the vicinities of spatial locations, and so on, with the natural assumption that variables far apart tend to have weak correlations. Applications of such models abound in a variety of fields such as finance, genome associations analysis and spatial modeling. We adopt a flexible framework under which each variable is dependent on its neighbors or predecessors, and the neighborhood size can vary for each variable. It is of great interest to reveal this local dependence structure by estimating the covariance or precision matrix while yielding a consistent estimate of the varying neighborhood size for each variable. The existing literature on banded covariance matrix estimation, which assumes a fixed bandwidth cannot be adapted for this general setup. We employ the modified Cholesky decomposition for the precision matrix and design a flexible prior for this model through appropriate priors on the neighborhood sizes and Cholesky factors. The posterior contraction rates of the Cholesky factor are derived which are nearly or exactly minimax optimal, and our procedure leads to consistent estimates of the neighborhood size for all the variables. Another appealing feature of our procedure is its scalability to models with large numbers of variables due to efficient posterior inference without resorting to MCMC algorithms. Numerical comparisons are carried out with competitive methods, and applications are considered for some real datasets.

Keywords: selection consistency, optimal posterior convergence rate, varying bandwidth.

1 Introduction

The problem of covariance matrix or precision matrix estimation has been extensively studied over the last few decades. A typical model setup is to assume $X = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ follows a p-dimensional normal distribution $N_p(0, \Sigma)$, with mean zero and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. The dependence structure among the variables is encoded by the covariance matrix Σ or its inverse $\Omega = \Sigma^{-1}$. In high-dimensional settings where p can be much larger than the sample size, the traditional sample covariance matrix or inverse-Wishart prior leads to inconsistent estimates of Σ or Ω, see Johnstone and Lu

[*] This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018207). We also acknowledge the generous support of NSF grants DMS CAREER 1654579 and DMS 2113642.

[†] Department of Statistics, Sungkyunkwan University, South Korea, leekjstat@gmail.com

[‡] Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, USA, lizhen.lin@nd.edu

https://doi.org/10.1214/21-BA1299

(2009) and Lee and Lee (2018). Restricted matrix classes with a banded or sparse structure are often imposed on the covariance or precision matrices as a common practice for consistent estimation (see, e.g., Bickel and Levina, 2008, Cai et al., 2016 and Lee et al., 2019).

In this paper, we focus on investigating the local dependence structure in a high-dimension model where the variables possess a natural ordering. An ordering on variables is often encountered for example in time series or genome data, where variables close to each other in time or location are more likely to be correlated than variables located far apart. More specifically, we assume that each variable depends on its neighboring variables or predecessors and more importantly, the *size of the neighborhood or bandwidth can vary with each variable.* Therefore, the existing literature on banded covariance matrix estimation which deals with a fixed bandwidth cannot be adapted for this set up.

Our work employs the modified Cholesky decomposition (MCD) (Pourahmadi, 1999) of the precision matrix, which provides an efficient way to learn the local dependence structure of the data. The MCD has been widely used for covariance or precision matrix estimation including in Rütimann and Bühlmann (2009), Shojaie and Michailidis (2010) and van de Geer and Bühlmann (2013), just to name a few. In more details, assume the variables $X_1, \ldots, X_p$ are arranged according to a known ordering. For any $p \times p$ positive definite matrix Ω, there uniquely exists a lower triangular matrix $A = (a_{jl}) \in \mathbb{R}^{p \times p}$ and a diagonal matrix $D = diag(d_j) \in \mathbb{R}^{p \times p}$such that $\Omega = (I_p - A)^T D^{-1} (I_p - A)$, where $a_{jj} = 0$ and $d_j > 0$ for all $j = 1, \ldots, p$. It is called the MCD of Ω, and we call A the Cholesky factor. Based on the MCD, $X = (X_1, \ldots, X_p)^T \sim N_p(0, \Omega^{-1})$ is equivalent to a set of linear regression models, $X_1 \sim N(0, d_1)$ and

$$X_j \mid X_1, \ldots, X_{j-1} \sim N\Big(\sum_{l=1}^{j-1} a_{jl} X_l,\ d_j \Big), \quad j = 2, \ldots, p. \tag{1}$$

In this paper, we assume *local dependence structure* by considering

$$X_j \mid X_{j-k_j}, \ldots, X_{j-1} \sim N\Big(\sum_{l=j-k_j}^{j-1} a_{jl} X_l,\ d_j \Big), \quad j = 2, \ldots, p \tag{2}$$

for some k_j instead of (1), which implies $a_{jl} = 0$ for any $l = 1, \ldots, j - k_j - 1$ and $j = 2, \ldots, p$. It leads to a varying bandwidth structure for the Cholesky factor which means that X_j is dependent only on k_j closest variables, $X_{j-k_j}, \ldots, X_{j-1}$, and conditionally independent of others. Our goal is to learn the local dependence structure of data based on model (2) by learning a_{jl}, d_j as well as the varying bandwidth or neighborhood size k_j. Allowing a varying bandwidth in the above model is a more realistic assumption as the range and pattern of dependence can vary over time or spatial locations. See Figure 1 for an illustration of the Cholesky structure with varying bandwidth.

A number of work have been proposed for estimating sparse Cholesky factors based on penalized likelihood approaches, including Rothman et al. (2010), van de Geer and

Bühlmann (2013) and Khare et al. (2019). However, these methods are not suitable for modeling local dependence because they allow arbitrary sparsity patterns for Cholesky factors. For example, X_j can depend on X_1 but not on $X_2, \ldots, X_{j-1}$ in their framework, which is not desirable in the context of local dependence. An et al. (2014) considered a banded Cholesky factor and proposed a consistent test for bandwidth selection, but they assumed the same local dependence structure for all variables, that is, assumed a common k instead of k_j in (2). The most relevant work is Yu and Bien (2017) which developed a penalized likelihood approach for estimating the banded Cholesky factor in (2) using a hierarchical group lasso penalty. Their approach is formulated as a convex optimization problem, and theoretical properties such as selection consistency and convergence rates were established. However, they required relatively strong conditions to obtain theoretical results, which will be discussed in detail in Section 3.

From a Bayesian perspective, Banerjee and Ghosal (2014) and Lee and Lee (2021) suggested G-Wishart priors and banded Cholesky priors for learning local dependence structure with a common bandwidth k, and posterior convergence rates for precision matrices are derived assuming k is known. Lee et al. (2019) proposed the empirical sparse Cholesky prior for sparse Cholesky factors. Under their model each variable can be dependent on distant variables, so it is not suitable for local dependence structure. Lee and Lin (2020) suggested a prior distribution for banded Cholesky factors. Although bandwidth selection consistency as well as consistency of Bayes factors were established in high-dimensional settings, again a common bandwidth k is assumed in their model. Achieving selection consistency simultaneously for bandwidths of all the variables is a much more challenging problem as the number of bandwidths as well as the size of the bandwidth can diverge as the number of variables p goes to infinity.

In this paper, we propose a prior tailored to Cholesky factors with local dependence structure. The proposed prior allows a flexible learning of local dependence structure based on varying bandwidth k_j. This paper contributes in both theoretical and practical developments. For theoretical advancement, we prove selection consistency for varying bandwidth and nearly optimal posterior convergence rates for Cholesky factors. This is the first Bayesian method for local dependence learning with varying bandwidth in high-dimensional settings with theoretical guarantees. Furthermore, we significantly weaken required conditions to obtain theoretical properties compared with those in Yu and Bien (2017). On the other hand, from a practical point of view, the induced posterior allows fast computations, enabling scalable inference for large data sets. The posterior inference does not require Markov chain Monte Carlo (MCMC) algorithms and is easily parallelizable. Furthermore, our simulation studies show that the proposed method outperforms other competitors in various settings. We find that the proposed cross-validation for the proposed method is much faster than the contenders and selects nearly optimal hyperparameter in terms of specificity and sensitivity. Finally, it is worth mentioning that posterior inference for sparse Cholesky factors with an arbitrary sparsity pattern is computationally much more expensive than that for banded Cholesky factors, developing a statistical method for banded Cholesky factors is therefore of great importance independent of existing methods for sparse Cholesky factors.

The rest of paper is organized as follows. In Section 2, model assumptions, the proposed local dependence Cholesky prior and the induced fractional posterior are introduced. The main results including bandwidth selection consistency and optimal posterior convergence rates are established in Section 3. In Section 4, the performance of the proposed method is illustrated based on simulated data and real data analysis. Concluding remarks and discussions are given in Section 5, while the proofs of main results and additional simulation results are provided in the supplementary material (Lee and Lin, 2022). R codes for implementation of our empirical results are available at https://github.com/leekjstat/LANCE.

2 Preliminaries

2.1 Notation

For any a and $b \in \mathbb{R}$, we denote $a \wedge b$ and $a \vee b$ as the minimum and maximum of a and b, respectively. For any positive sequences a_n and b_n, $a_n = o(b_n)$ denotes $a_n/b_n \longrightarrow 0$ as $n \to \infty$. We denote $a_n = O(b_n)$, or equivalently $a_n \lesssim b_n$, if there exists a constant $C > 0$ such that $a_n < Cb_n$ for all large n. For any matrix $A = (a_{ij}) \in \mathbb{R}^{p\times p}$, we denote $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum eigenvalues, respectively. Furthermore, we define the matrix ℓ_∞-norm, Frobenius norm and element-wise maximum norm as follows: $\|A\|_\infty = \max_{1\le i\le p}\sum_{j=1}^p |a_{ij}|$, $\|A\|_F = (\sum_{i=1}^p\sum_{j=1}^p a_{ij}^2)^{1/2}$ and $\|A\|_{\max} = \max_{1\le i,j\le p}|a_{ij}|$. We denote $IG(a,b)$ as the inverse-Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$.

2.2 High-dimensional Gaussian Models

Throughout the paper, we assume a high-dimensional setting with $p = p_n \ge n$, and that the variables have a known natural ordering. Specifically, we assume we observe a sample of data with sample size n from a p-dimensional Gaussian model,

$$X_1,\ldots,X_n \mid \Omega_n \overset{i.i.d.}{\sim} N_p(0, \Omega_n^{-1}), \tag{3}$$

where $\Omega_n = \Sigma_n^{-1} \in \mathbb{R}^{p\times p}$ is a precision matrix. For the rest of the paper, we use subscript n for any $p \times p$ matrices to indicate that the dimension $p = p_n$ grows as $n \to \infty$. Let $\mathbf{X}_n = (X_1,\ldots,X_n)^T \in \mathbb{R}^{n\times p}$ be a data matrix, and $X_i = (X_{i1},\ldots,X_{ip})^T \in \mathbb{R}^p$ for all $i = 1,\ldots,n$. We denote the Cholesky factor and the diagonal matrix from the MCD as A_n and D_n, respectively, i.e., $\Omega_n = (I_p - A_n)^T D_n^{-1}(I_p - A_n)$. Model (3) is related to a directed acyclic graph (DAG) model depending on the sparsity pattern of A_n (van de Geer and Bühlmann, 2013; Lee et al., 2019), but we will not go into detail on DAG models.

In this paper, we assume $A_n = (a_{jl})$ has a banded structure with *varying bandwidths*, $\{k_2,\ldots,k_p\}$, which satisfies $0 \le k_j \le j-1$ and $\sum_{l:|j-l|>k_j}|a_{jl}| = 0$ for each $j = 2,\ldots,p$. Let $\tilde{X}_j \in \mathbb{R}^n$ and $\mathbf{X}_{j(k_j)} \in \mathbb{R}^{n\times k_j}$ be sub-matrices of $\mathbf{X}_n$ consisting of the jth and $(j-k_j),\ldots,(j-1)$th columns, respectively. Under the varying bandwidths assumption, model (3) can be represented as

$$\begin{aligned}\tilde{X}_1 \mid d_1 &\sim N_n(0, d_1 I_n),\\ \tilde{X}_j \mid \mathbf{X}_{j(k_j)}, a_j^{(k_j)}, d_j, k_j &\sim N_n\big(\mathbf{X}_{j(k_j)} a_j^{(k_j)}, d_j I_n\big), \quad j = 2, \ldots, p,\end{aligned}$$

where $a_j^{(k_j)} = (a_{jl})_{(j-k_j)\le l \le (j-1)} \in \mathbb{R}^{k_j}$. The above representation implies that to predict the jth variable, $\tilde{X}_j$, it suffices to know its k_j-nearest predecessors, $\mathbf{X}_{j(k_j)}$. Thus, the varying bandwidths assumption of A_n directly induces the *local dependence* structure between variables. As mentioned before, this is often natural given commonly encountered ordered variables in time series or genome data sets, and a more realistic and flexible assumption than the common bandwidth assumption.

2.3 Local Dependence Cholesky Prior

To conduct Bayesian inference, we need to impose a prior on A_n and D_n as well as k_j, $j = 2, \ldots, p$, which restricts A_n to have a local dependence structure. We propose the following prior

$$\begin{aligned} a_j^{(k_j)} \mid d_j, k_j &\overset{ind.}{\sim} N_{k_j}\Big(\widehat{a}_j^{(k_j)}, \frac{d_j}{\gamma}\big(\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)}\big)^{-1}\Big), \; j = 2, \ldots, p,\\ \pi(d_j) &\propto d_j^{-\nu_0/2-1}, \; j = 1, \ldots, p, \\ \pi(k_j) &\propto c_1^{-k_j} p^{-c_2 k_j} I(0 \le k_j \le \{R_j \wedge (j-1)\}), \; j = 2, \ldots, p, \end{aligned} \tag{4}$$

for some positive constants $\gamma, c_1, c_2, R_2, \ldots, R_p$ and ν_0, where $\widehat{a}_j^{(k_j)} = (\mathbf{X}_{j(k_j)}^T \mathbf{X}_{j(k_j)})^{-1} \times \mathbf{X}_{j(k_j)}^T \tilde{X}_j$. We call the above prior LANCE (LocAl depeNdence CholEsky) prior. The conditional prior for $a_j^{(k_j)}$ is a version of the Zellner's g-prior (Zellner, 1986) and depends on the data. By using the prior $\pi(a_j^{(k_j)} \mid d_j, k_j)$ with sub-Gaussian tails and data-dependent center, we can obtain theoretical properties of posteriors without having to use priors with heavier tails or introducing redundant upper bound conditions on $\|a_j^{(k_j)}\|_2$ as discussed by Lee et al. (2019). The prior for d_j is improper and includes the Jeffreys' prior (Jeffreys, 1946), $\pi(d_j) \propto d_j^{-1}$, as a special case. In fact, the proper prior $d_j \sim IG(\nu_0/2, \nu_0')$, for some constant $\nu_0' > 0$, can be used. However, we proceed with the above improper prior to reduce the number of hyperparameters.

For posterior inference, we suggest using the fractional posterior, which has received increased attention recently (Martin and Walker, 2014; Martin et al., 2017; Lee et al., 2019). Let $L(A_n, D_n)$ be the likelihood function of model (3). For a given constant $0 < \alpha < 1$, the α-fractional posterior is defined by $\pi_\alpha(A_n, D_n \mid \mathbf{X}_n) \propto L(A_n, D_n)^\alpha \pi(A_n, D_n)$. Thus, α-fractional posterior is a posterior distribution updated with the likelihood function raised to a power of α instead of the usual likelihood. We denote the α-fractional posterior by $\pi_\alpha(\cdot \mid \mathbf{X}_n)$ to indicate the use of α-fractional likelihood. Theoretical properties of fractional posterior can often be established under weaker conditions compared with the usual posterior (Bhattacharya et al., 2019). Under our model, the α-fractional posterior has the following closed form:

$$\begin{aligned} a_j^{(k_j)} \mid d_j, k_j, \mathbf{X}_n &\overset{ind.}{\sim} N_{k_j}\Big(\widehat{a}_j^{(k_j)}, \frac{d_j}{\alpha+\gamma}\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)}\big)^{-1}\Big), \ j=2,\ldots,p, \\ d_j \mid k_j, \mathbf{X}_n &\overset{ind.}{\sim} IG\Big(\frac{\alpha n+\nu_0}{2}, \frac{\alpha n}{2}\widehat{d}_j^{(k_j)}\Big), \ j=1,\ldots,p, \\ \pi_\alpha(k_j \mid \mathbf{X}_n) &\propto \pi(k_j)\Big(1+\frac{\alpha}{\gamma}\Big)^{-\frac{k_j}{2}}\big(\widehat{d}_j^{(k_j)}\big)^{-\frac{\alpha n+\nu_0}{2}}, \ j=2,\ldots,p, \end{aligned} \tag{5}$$

where $\widehat{d}_j^{(k_j)} = n^{-1}\tilde{X}_j^T(I_n - \tilde{P}_{jk_j})\tilde{X}_j$ and $\tilde{P}_{jk_j} = \mathbf{X}_{j(k_j)}(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)})^{-1}\mathbf{X}_{j(k_j)}^T$. Based on (5), one can notice that posterior inference for each $j = 2,\ldots,p$ is parallelizable. Furthermore, a MCMC algorithm is not needed because direct posterior sampling from (5) is possible. Note that the three posterior distributions in (5) form the joint posterior distributions of $(a_j^{(k_j)}, d_j, k_j)$, which are independent for each $j = 1,\ldots,p$. Thus, LANCE prior leads to a fast and scalable posterior inference even in high-dimensions.

3 Main Results

In this section, we show that LANCE prior accurately unravels the local dependence structure. More specifically, it is proved that LANCE prior attains bandwidth selection consistency for all bandwidths and nearly minimax posterior convergence rates for Cholesky factors. To obtain desired asymptotic properties of posteriors, we assume the following conditions:

(A1) $\lambda_{\max}(\Omega_{0n})\log p/\lambda_{\min}(\Omega_{0n}) = o(n)$.

(A2) For some constant $M_{\rm bm} > c_2 + 1$,

$$\min_{(j,l):a_{0,jl}\neq 0}\frac{a_{0,jl}^2}{d_{0j}} \geq \frac{10M_{\rm bm}\ \lambda_{\max}(\Omega_{0n})}{(\alpha+\nu_0/n)(1-\alpha-\nu_0/n)}\frac{\log p}{n}.$$

(A3) $k_{0j} \leq R_j$ for any $j = 2,\ldots,p$, $\max_j R_j \log p \leq n(1+5\sqrt{\epsilon})/\{C_{\rm bm}(1-2\epsilon)^2\}$, where $\epsilon = \{(1-\alpha)/10\}^2$ and $C_{\rm bm} = 10M_{\rm bm}/\{(\alpha+\nu_0/n)(1-\alpha-\nu_0/n)\}$

(A4) The hyperparameters satisfy $\nu_0 = O(1)$, $\gamma = O(1)$, $c_1 = O(1)$, $c_2 > 1$ and $0.6 \leq \alpha < 1$.

Condition (A1) allows the condition number $\lambda_{\max}(\Omega_{0n})/\lambda_{\min}(\Omega_{0n})$ to grow to infinity at a rate slower than $n/\log p$. This condition is weaker than the bounded eigenvalue conditions used in Banerjee and Ghosal (2015), Khare et al. (2019) and Lee et al. (2019), which assume $c < \lambda_{\min}(\Omega_{0n}) \leq \lambda_{\max}(\Omega_{0n}) < C$ for some constants c and $C > 0$. We note here that we require the bounded eigenvalue conditions for the (nearly) minimaxity of the posterior convergence rates in Theorems 3.2 and 3.3.

Condition (A2) determines the lower bound for the nonzero signals, $a_{0,jl}^2/d_{0j}$. This is called the *beta-min* condition. It has been well known that the beta-min condition is essential for variable selection (Bühlmann and van de Geer, 2011; Martin et al., 2017) and support recovery for sparse matrices (Yu and Bien, 2017; Cao et al., 2019). Condition (A2) implies that the rate of nonzero $a_{0,jl}^2/d_{0j}$ should, at least, be $\lambda_{\max}(\Omega_{0n})\log p/n$, which has the same rate with $\log p/n$ under the bounded eigenvalue condition.

Condition (A3), together with condition (A4), provides an upper bound for the true bandwidth k_{0j}: they allow the maximum bandwidth, $\max_j k_{0j}$, to grow to infinity at a rate not faster than $n/\log p$. The rest of condition (A4) presents sufficient conditions for hyperparameters to obtain theoretical properties. The conditions on c_1 and c_2 control the strength of penalty for large models, i.e., large bandwidths. Note that condition $c_2 > 1$ is weaker than the condition $c_2 \geq 2$ used in Lee et al. (2019), which implies that the local dependence assumption requires a weaker penalty compared with arbitrary sparsity patterns. In Section 4, we will give a practical guidance for the choice of hyperparameters.

Theorem 3.1 (Bandwidth selection consistency)**.** *Consider model* (3) *and LANCE prior* (4)*. Under conditions (A1)–(A4), we have*

$$\mathbb{E}_0\Big\{\pi_\alpha\big(k_j \neq k_{0j} \text{ for at least one } 2 \leq j \leq p \mid \mathbf{X}_n\big)\Big\} \longrightarrow 0 \qquad \text{as } n \to \infty.$$

Theorem 3.1 says that the posterior probability of incorrectly estimating local dependence structure, i.e., bandwidths, converges to zero in probability as $n \to \infty$. Thus, the proposed method can consistently recover local dependence structure for each variable.

We compare the above result with existing theoretical results in other work. First of all, we note here that Khare et al. (2019), Cao et al. (2019) and Lee et al. (2019) assumed arbitrary dependence structures for Cholesky factors, thus their methods are not tailored to local dependence structure considered in this paper. Although Bickel and Levina (2008), Banerjee and Ghosal (2014) and Lee and Lee (2021) focused on banded Cholesky factors, which result in banded precision matrices, they assumed a common dependence structure, i.e., a common bandwidth, for each variable, and did not provide bandwidth selection consistency.

To the best of our knowledge, the state-of-the-art theoretical result for estimating local dependence structure is obtained by Yu and Bien (2017). They proposed a penalized likelihood approach and obtained bandwidth selection consistency in a high-dimensional setting. They assumed that the eigenvalues of Ω_{0n} lie in $[\kappa^{-2}, \kappa^2]$ and the beta-min condition,

$$\min_{(j,l):a_{0,jl}\neq 0} \frac{|a_{0,jl}|}{\sqrt{d_{0j}}} \geq 8\rho^{-1}(4\max_j \|\Sigma_{0n,k_{0j}}^{-1}\|_\infty + 5\kappa^2)\sqrt{2\|D_{0n}\|_\infty \frac{\log p}{n}},$$

for some constants $\kappa > 1$ and $\rho \in (0,1]$, where $\Sigma_{0n,k_{0j}} = (\sigma_{0,il})_{j-k_{0j}\leq i,l\leq j}$ denotes the sub-matrix of the true covariance matrix. Note that these conditions are more restrictive than our conditions (A1) and (A2). For example, $\|\Sigma_{0n,k_{0j}}^{-1}\|_\infty = O(k_{0j}^{1/2})$ holds under the bounded eigenvalue condition, so the beta-min condition in Yu and Bien (2017) implies that the minimum nonzero $a_{0,jl}^2/d_{0j}$ is bounded below by $\max_j k_{0j}\log p/n$ with respect to a constant multiple; in contrast, the rate of the lower bound in condition (A2) is $\log p/n$ under the bounded eigenvalue condition. Furthermore, they assumed the so-called *irrepresentable* condition,

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

which is typically required for the lasso type methods with a random design matrix (e.g., see Wainwright (2009) and Khare et al. (2019)). Yu and Bien (2017) proved the exact signed support recovery property under the above conditions and $n > \rho^{-2}\|D_{0n}\|_\infty\kappa^2$ $(12\pi^2\max_j k_{0j} + 32)\log p$. Note that condition (A3) together with (A4) implies $n > C\max_j k_{0j}\log p$ for some constant $C > 0$. Hence, stronger conditions are also required by Yu and Bien (2017) to establish bandwidth selection consistency compared with Theorem 3.1.

Next, we show that LANCE prior achieves nearly minimax posterior convergence rates for Cholesky factors. The posterior convergence rates are obtained with or without beta-min condition (A2). Under the beta-min condition, Theorem 3.2 presents posterior convergence rates based on various matrix norms.

Theorem 3.2 (Posterior convergence rates with beta-min condition)**.** *Suppose that the conditions in Theorem 3.1 hold. If $k_0\log p = o(n)$, where $k_0 = \max_j k_{0j}$, we have*

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_{\max} \ge K_{\mathrm{chol}}\frac{\lambda_{\max}(\Omega_{0n})^2}{\lambda_{\min}(\Omega_{0n})^2}\Big(\frac{k_0+\log p}{n}\Big)^{1/2} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_\infty \ge K_{\mathrm{chol}}\frac{\lambda_{\max}(\Omega_{0n})^2}{\lambda_{\min}(\Omega_{0n})^2}\sqrt{k_0}\Big(\frac{k_0+\log p}{n}\Big)^{1/2} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_F^2 \ge K_{\mathrm{chol}}\frac{\lambda_{\max}(\Omega_{0n})^2}{\lambda_{\min}(\Omega_{0n})^2}\frac{\sum_{j=2}^p(k_{0j}+\log j)}{n} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

for some constant $K_{\mathrm{chol}} > 0$ not depending on unknown parameters.

Estimating each row of Cholesky factor A_n can be considered as a linear regression problem with a random design matrix. By assuming beta-min condition (A2), we can use the selection consistency result in Theorem 3.1, although a beta-min condition is usually not essential for obtaining convergence rates. With a cost of the beta-min condition, we can focus only on the set $a_j^{(k_j)} = a_j^{(k_{0j})}$ for all j in the posterior with high probability tending to 1. Then the above posterior convergence rates for various matrix norms boil down to those related to $\|a_j^{(k_{0j})} - a_{0j}^{(k_{0j})}\|$ for various vector norms $\|\cdot\|$, which makes the problem simpler.

Assume that $\epsilon_0 \le \lambda_{\min}(\Omega_{0n}) \le \lambda_{\max}(\Omega_{0n}) \le \epsilon_0^{-1}$ for some small constant $0 < \epsilon_0 < 1/2$. Then, the obtained posterior convergence rates in Theorem 3.2 under the matrix ℓ_∞-norm and Frobenius norm are minimax if $\log p = O(k_0)$ and $\log j = O(k_{0j})$ for all $j = 2,\dots,p$, respectively, by Theorem 3.3 in Lee et al. (2019). Therefore, the above posterior convergence rates are nearly or exactly minimax depending on the dimensionality p and bandwidths k_{0j} under bounded eigenvalue conditions on Ω_{0n}.

Now, we establish posterior convergence rate of Cholesky factors without beta-min condition (A2). To obtain the desired result, we consider a *modified* LANCE prior using $d_j \sim IG(\nu_0/2, \nu_0')$ for some constant $\nu_0' > 0$ instead of $\pi(d_j) \propto d_j^{-\nu_0/2-1}$ in (4). This modification is mainly used to derive a lower bound of the likelihood ratio appearing in the denominator of posteriors (see Lemma 7.1 in Lee et al., 2019). Theorem 3.3 shows the posterior convergence rate under various matrix norms.

Theorem 3.3 (Posterior convergence rates without beta-min condition)**.** *Consider model* (3) *and the modified LANCE prior described above. Suppose that the conditions (A1), (A3) and (A4) hold. If* $k_0 \log p = o(n)$*,* $\lambda_{\max}(\Omega_{0n})/\lambda_{\min}(\Omega_{0n}) = O(p)$ *and* $\lambda_{\max}(\Omega_{0n}) = o(n)$*, we have*

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_{\max} \ge K_{\text{chol}}\Big(\frac{\lambda_{\max}(\Omega_{0n})}{\lambda_{\min}(\Omega_{0n})}\frac{k_0\log p}{n}\Big)^{1/2} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_{\infty} \ge K_{\text{chol}}k_0\Big(\frac{\lambda_{\max}(\Omega_{0n})}{\lambda_{\min}(\Omega_{0n})}\frac{\log p}{n}\Big)^{1/2} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

$$\mathbb{E}_0\Big[\pi_\alpha\Big\{\|A_n - A_{0n}\|_F^2 \ge K_{\text{chol}}\frac{\lambda_{\max}(\Omega_{0n})}{\lambda_{\min}(\Omega_{0n})}\frac{\sum_{j=2}^p k_0\log j}{n} \mid \mathbf{X}_n\Big\}\Big] = o(1),$$

for some constant $K_{\text{chol}} > 0$ *not depending on unknown parameters.*

If we assume bounded eigenvalue conditions on Ω_{0n}, the above posterior convergence rates are slightly slower than those in Theorem 3.2 due to the absence of the beta-min condition. Suppose $\epsilon_0 \le \lambda_{\min}(\Omega_{0n}) \le \lambda_{\max}(\Omega_{0n}) \le \epsilon_0^{-1}$ for some small constant $0 < \epsilon_0 < 1/2$. Under these bounded eigenvalue conditions, Yu and Bien (2017) (Lemma 17) obtained the convergence rates $\zeta_\Gamma\sqrt{\log p/n}$, $\zeta_\Gamma(k_0+1)(\log p/n)^{1/2}$ and $\zeta_\Gamma\sqrt{(\sum_{j=2}^p k_{0j}+p)\log p/n}$ under the element-wise maximum norm, matrix ℓ_∞-norm and Frobenius norm, respectively, where $\zeta_\Gamma = 8(2\|D_{0n}\|_\infty)^{1/2}\rho^{-1}(4\max_j\|\Sigma_{0n,k_{0j}}^{-1}\|_\infty + 5\epsilon_0^{-1})$ for some constant $\rho \in (0,1]$. Note that, as mentioned before, it holds that $\|\Sigma_{0n,k_{0j}}^{-1}\|_\infty = O(k_{0j}^{1/2})$ without further assumption. Thus, their convergence rates are slower than or comparable to ours. We would also like to mention that, under bounded eigenvalue conditions on Ω_{0n}, the posterior convergence rate under the matrix ℓ_∞-norm in Theorem 3.3 is minimax if $\log p \asymp \log(p/k_0)$ by Theorem 3.5 in Lee et al. (2019). For example, it is the case if $k_0 = O(p^\beta)$ for some $0 < \beta < 1$.

Remark 1. *By carefully modifying the proof of Theorem 3.6 in Lee et al. (2019), we can probably obtain posterior convergence rates for precision matrices. However, it is not clear whether the obtained posterior convergence rates are minimax optimal. Although Liu and Ren (2020) showed minimax rates for precision matrices with bandable Cholesky factors, they considered slightly different parameter spaces for Cholesky factors from what we consider. To the best of our knowledge, minimax rates for precision matrices based on Cholesky factors with varying bandwidths have not been established.*

4 Numerical Studies

4.1 Choice of Hyperparameters

The proposed Bayesian method has hyperparameters that need to be determined, and we provide the following guidelines. It is reasonable to use the hyperparameter α close to 1 unless there is a strong evidence of model misspecification. Thus, in our numerical studies, $\alpha = 0.99$ is used. We use the hyperparameter $\nu_0 = 0$, which leads to the Jeffreys' prior (Jeffreys, 1946) for d_j. The upper bound for model sizes, R_j, is set at $R_j = \lfloor n/2 \rfloor - 2$. The rest of hyperparameters γ, c_1 and c_2 control the penalty for large models through $\pi_\alpha(k_j \mid \mathbf{X}_n)$: as the values of γ, c_1 and c_2 increase, $\pi_\alpha(k_j \mid \mathbf{X}_n)$ prefers smaller values of k_j. We suggest to determine c_2 using the Bayesian cross-validation method (Gelman et al., 2014), with the other hyperparameters $\gamma = 0.1$ and $c_1 = 1$ fixed. Compared with the Bayesian cross-validation method for choosing (γ, c_1, c_2), this approach significantly reduces computational time, while achieving nice performance in our simulation study.

To conduct Bayesian cross-validation, we repeatedly split the data $n_{\rm cv}$ times into a training set and a test set with size $n_1 = \lceil n/2 \rceil$ and $n_2 = \lfloor n/2 \rfloor$, respectively. Let $I_1(\nu)$ and $I_2(\nu)$ be indices for the νth training set and test set, respectively, i.e., $|I_1(\nu)| = n_1$, $|I_2(\nu)| = n_2$ and $I_1(\nu) \cup I_2(\nu) = \{1, \ldots, n\}$, for any $\nu = 1, \ldots, n_{\rm cv}$. Denote $\mathbf{X}_{I_1(\nu)}$ and $\mathbf{X}_{I_2(\nu)}$ as $\{X_i\}_{i \in I_1(\nu)}$ and $\{X_i\}_{i \in I_2(\nu)}$, respectively. Then for a given hyperparameter c_2, the estimated out-of-sample log predictive density, ${\rm lpd}_{\rm cv, c_2}$, is

$$\begin{aligned} {\rm lpd}_{\rm cv,c_2} &= \sum_{\nu=1}^{n_{\rm cv}} \log f_{c_2}(\mathbf{X}_{I_2(\nu)} \mid \mathbf{X}_{I_1(\nu)}) \\ &= \sum_{\nu=1}^{n_{\rm cv}} \log \Big\{ \sum_k f(\mathbf{X}_{I_2(\nu)} \mid k) \pi_{\alpha, c_2}(k \mid \mathbf{X}_{I_1(\nu)}) \Big\}, \end{aligned}$$

where $k = (k_2, \ldots, k_p)$, $f(\mathbf{X}_{I_2(\nu)} \mid k)$ is the marginal likelihood for k given $\mathbf{X}_{I_2(\nu)}$, and $\pi_{\alpha,c_2}(k \mid \mathbf{X}_{I_1(\nu)})$ is the fractional posterior based on $\mathbf{X}_{I_1(\nu)}$ and the hyperparameter c_2. The aim of the Bayesian cross-validation is to find the optimal c_2 maximizing ${\rm lpd}_{\rm cv,c_2}$.

The marginal likelihood $f(\mathbf{X}_{I_2(\nu)} \mid k)$ is available in a closed form:

$$f(\mathbf{X}_{I_2(\nu)} \mid k) = \prod_{j=2}^{p} \left[(2\pi)^{-n_2/2} \Gamma\Big(\frac{n_2 + \nu_0}{2}\Big) \Big(1 + \frac{1}{\gamma}\Big)^{-k_j/2} \Big\{ \widehat{d}_j^{(k_j)}(I_2(\nu))/2 \Big\}^{-(n_2+\nu_0)/2} \right],$$

where $\widehat{d}_j^{(k_j)}(I_2(\nu))$ is the estimated variance $\widehat{d}_j^{(k_j)}$ using $\mathbf{X}_{I_2(\nu)}$. The closed form of the marginal posterior $\pi_{\alpha,c_2}(k_j \mid \mathbf{X}_{I_1(\nu)})$ is also available in (5), which requires the calculation of $\widehat{d}_j^{(k_j)}(I_1(\nu))$. When calculating ${\rm lpd}_{\rm cv,c_2}$, the main computational burden comes from calculating $\widehat{d}_j^{(k_j)}(I_1(\nu))$ and $\widehat{d}_j^{(k_j)}(I_2(\nu))$ for each $k_j = 0, 1, \ldots, R_j \wedge (j-1)$ and $j = 2, \ldots, p$. Note that these quantities do not vary from different choices of c_2. For a given randomly split data, these quantities only need to be calculated once regardless of the value of c_2. Therefore, LANCE prior enables scalable cross-validation-based inference even in high-dimensions. Throughout the numerical study, we split the data $n_{\rm cv} = 5$ times.

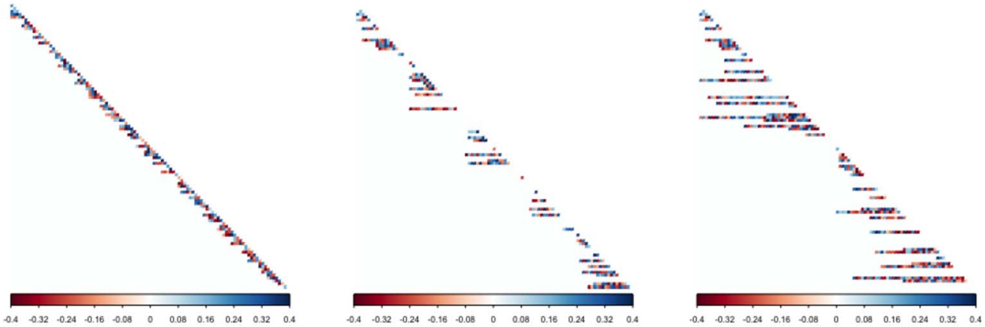

Figure 1: The true Cholesky factors for Model 1 (left), Model 2 (middle) and Model 3 (right) with $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $p = 100$.

4.2 Simulated Data

Throughout the numerical studies in this section, we focus on the bandwidth selection performance, while additional simulation studies focusing on the estimation performance are given in the supplementary material. We generated the true precision matrix $\Omega_{0n} = (I_p - A_{0n})^T D_{0n}^{-1} (I_p - A_{0n})$ for simulation studies. The diagonal entries of $D_{0n} = diag(d_{0j})$ were drawn independently from $Unif(2, 5)$. Next, the lower triangular entries of $A_{0n} = (a_{0,jl})$ were generated as follows:

- Model 1: For $2 \leq v \leq p$, the bandwidth of the vth row of A_{0n} is sampled from $Unif\{1, \ldots, \min(v-1, 5)\}$. This produces a sparse Cholesky factor with the maximum bandwidth size 5. Each nonzero element in A_{0n} is drawn independently from $Unif(A_{0,\min}, A_{0,\max})$, where the positive of negative sign is assigned with probability 0.5.

- Model 2: A_{0n} is a block diagonal matrix consisting of 5 blocks with size $p/5$, while the maximum size of bandwidths is 40. This setting produces a moderately sparse Cholesky factor. The length of the bandwidth of the vth row in each block follows a mixture distribution, $0.5 \times Unif\{1, \ldots, \min(v-1, 40)\} + 0.5 \times \delta_0$, where δ_0 is a point mass at zero. Each nonzero element in A_{0n} is drawn independently from $Unif(A_{0,\min}, A_{0,\max})$, where the positive or negative sign is assigned with probability 0.5. This setting corresponds to Model 2 in Yu and Bien (2017).

- Model 3: A_{0n} is a block diagonal matrix consisting of 2 blocks with size $p/2$, and the rest of the generation process is similar to that of Model 2. This setting produces a denser Cholesky factor compared with Model 2. This setting corresponds to Model 3 in Yu and Bien (2017).

Figure 1 shows a simulated true Cholesky factors for Settings 1 and 2 with $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $p = 100$.

We compare the performance of the proposed LANCE prior with the penalized likelihood approach in Yu and Bien (2017), which we call YB method. Since the unweighted version outperformed the weighted version in the simulation studies in Yu and Bien (2017), we used the unweighted version of the penalized likelihood approach. Furthermore, the convex sparse Cholesky selection (CSCS) (Khare et al., 2019), which is designed for sparse Cholesky factors, is also considered as a contender. Let $\widehat{A}_n^{YB}$ and $\widehat{A}_n^{CSCS}$ be the estimated Cholesky factor based on the penalized likelihood approach proposed by Yu and Bien (2017) and Khare et al. (2019), respectively. To estimate the local dependence structure, we set all of the estimated entries in $\widehat{A}_n^{YB}$ whose absolute values are below 0.1^{10} to zero, as suggested by Yu and Bien (2017). The receiver operating characteristic (ROC) curves for LANCE prior and the penalized likelihood approaches were drawn based on 100 hyperparameters c_2 selected from $[-1.5, 5]$ and 100 tuning parameters λ selected from $[0.01, 4]$, respectively. Using these hyperparameter values, we also compare the performance of cross-validation for LANCE prior and YB method. The Bayesian cross-validation described in Section 4.1 was used for LANCE prior. For YB method, we used `varband_cv` function in `R` package `varband`. The 5-fold cross-validation was used based on the unweighted version of the penalty. We did not conduct a cross-validation for CSCS method due to heavy computation. The three methods, the LANCE, YB and CSCS methods, can be run in parallel, although we did not use parallel computing in the numerical study.

Figures 2, 3 and 4 represent ROC curves based on 10 simulated data sets for Model 1, Model 2 and Model 3, respectively. For Model 3 with $p = 300$, we omit the results for CSCS method, because it did not converge for several days and caused a convergence problem. As expected, CSCS method does not work well compared with other two methods tailored to local dependence structure. The main reason for this phenomenon is that CSCS method does not guarantee local dependence structure. Based on the simulation results, the performances of LANCE prior and YB method are comparable in Model 1 (i.e., when bandwidths are smaller than 5), but LANCE prior tends to give larger area under the curves than those of YB method in Model 2 (i.e., when bandwidths are moderately large). Especially in Model 3 (i.e., when bandwidths are large), LANCE prior significantly outperforms YB method especially for large p. Thus, it seems that YB method tends to work better with smaller bandwidths, which is consistent with the observations in Yu and Bien (2017). Furthermore, we found that LANCE prior works better under large signals, $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$, compared with small signals, $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$. This makes sense because it is expected that large signals will easily satisfy the beta-min condition (A2).

The dots in Figures 2, 3 and 4 show the results based on cross-validation, where red dots and black dots represent those of LANCE prior and YB method, respectively. We found that cross-validation based on LANCE prior gives nearly optimal result in the sense that the result of the cross-validation method is located close to $(1, 1)$ on the ROC curve. On the other hand, cross-validation based on YB method tends to produce high false positive, which results in low specificity. In many cases, even when ROC curves of the two methods are similar, the performances of our cross-validation-based inference are much better than those of YB method.

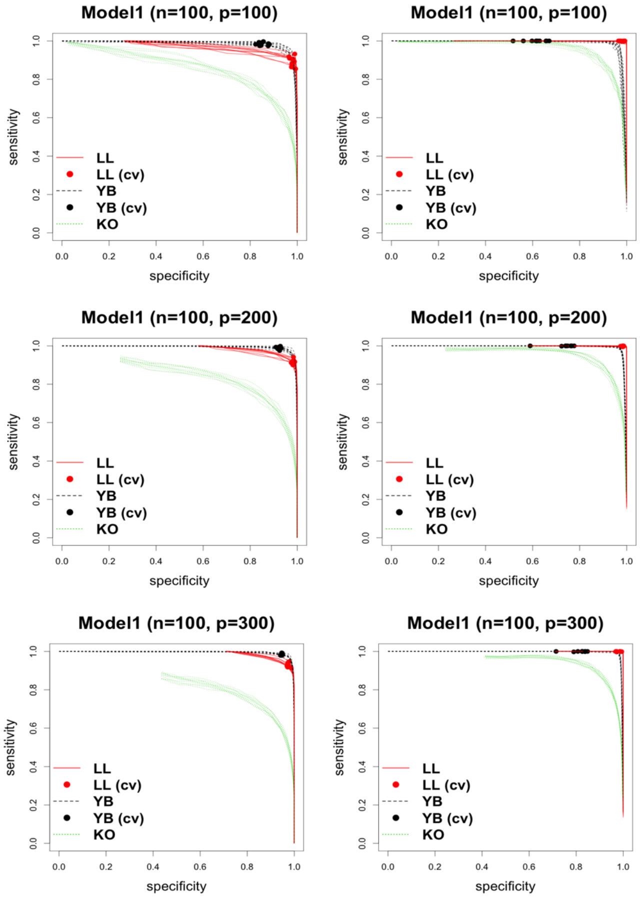

Figure 2: ROC curves are represented based on 10 simulated data sets from Model 1 with $n = 100$ and $p \in \{100, 200, 300\}$. Left column and right column show the results for $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$, respectively. LL and YB represent the methods proposed in this paper and Yu and Bien (2017), respectively.

We also found that the proposed method is much faster than YB method in most settings. Figure 5 shows box plots of the computation times for cross-validation using 100 hyperparameters. For each method, box plots were drawn based on 10 simulated data sets with $n = 100$ and $p = 300$. The relative computational gain of LANCE prior can be summarized by dividing the computation time for the LANCE prior by that for YB method. In our simulation settings, the mean and median of the relative computational gain of LANCE prior were 1539 and 208, respectively, which clearly show a computational advantage of the proposed method. Also note that YB method was conducted using the `R` package `varband` providing `C++` implementations, while LANCE prior was implemented using only `R`. The main reason for this observation is that YB method requires solving a penalized likelihood problem for each value of the tuning

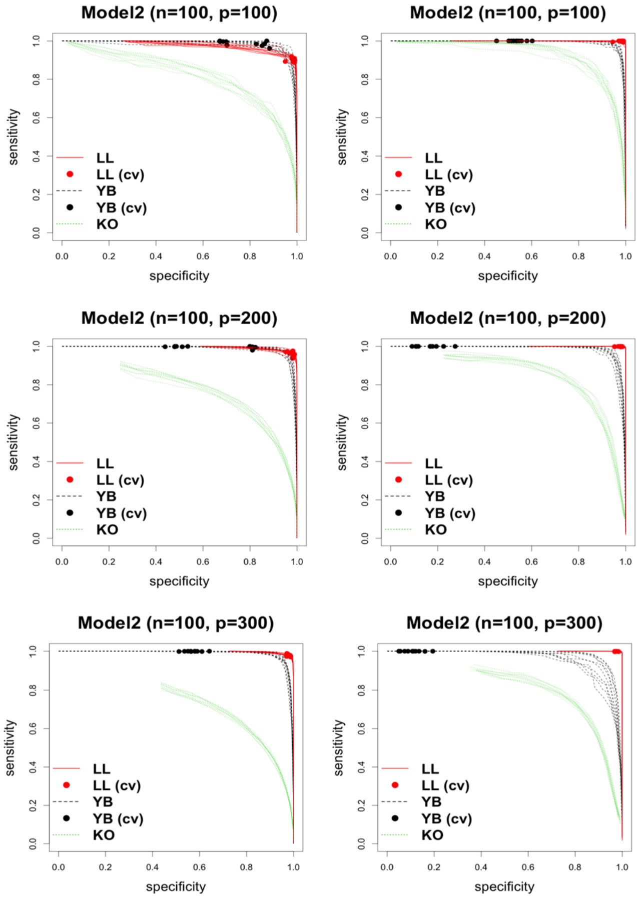

Figure 3: ROC curves are represented based on 10 simulated data sets from Model 2 with $n = 100$ and $p \in \{100, 200, 300\}$. Left column and right column show the results for $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$, respectively.

parameter λ. On the other hand, for LANCE prior, once the estimated error variance, $\widehat{d}_j^{(k_j)}$, is calculated, there is no need to recalculate it for various values of c_2. As a result, the cross-validation for LANCE prior is much faster than the state-of-the-art contender, thus enables scalable inference even in high-dimensions.

4.3 Real Data Analysis: Phone Call Center and Gun Point Data

We demonstrate the practical performance of LANCE prior by applying our model to two real data examples. We first consider the telephone call center data set, which was analyzed by Huang et al. (2006) and Bickel and Levina (2008). This data set consists of phone calls for 239 days in 2002 from a call center of a major U.S. financial

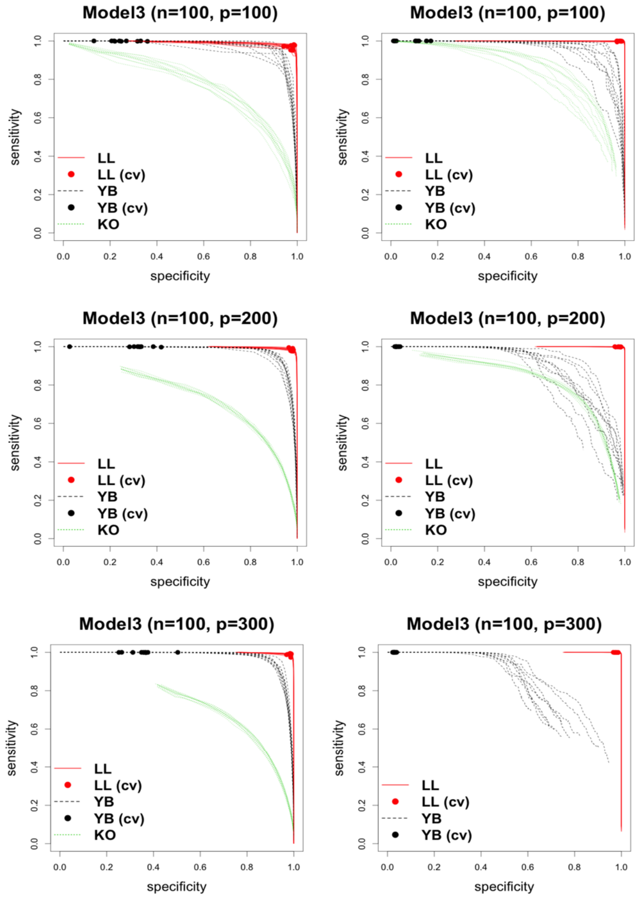

Figure 4: ROC curves are represented based on 10 simulated data sets from Model 3 with $n = 100$ and $p \in \{100, 200, 300\}$. Left column and right column show the results for $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$, respectively.

organization. The phone calls were recorded from 7:00 am until midnight for every 10 minutes, resulting in 102 intervals for each day, except holidays, weekends and days when the recording system did not work properly. The number of calls on the jth time interval of the ith day is denoted by N_{ij} for $i = 1, \ldots, 239$ and $j = 1, \ldots, 102$ ($p = 102$). The first 205 days are used as a training set ($n = 205$), and the last 34 days are used a test set. The data were transformed to $X_{ij} = \sqrt{N_{ij} + 1/4}$ as in Huang et al. (2006) and Bickel and Levina (2008). The data were centered after the transformation. Note that the data has a natural time ordering between the variables making it appropriate to apply LANCE prior.

The primary purpose is to predict the number of phone calls during a certain time period. We divide 102 time intervals for each day into two groups: those before the 51st

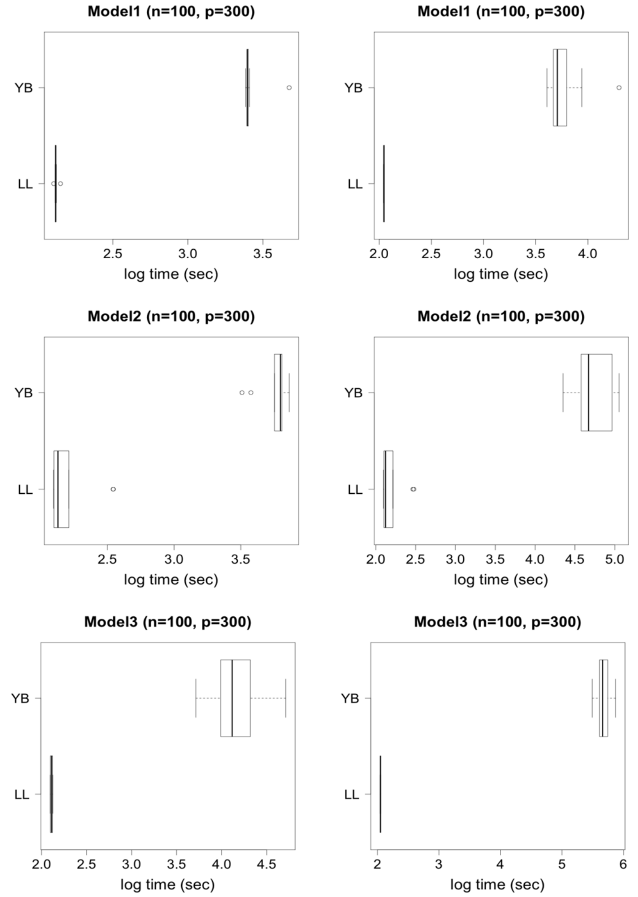

Figure 5: Logarithm of computation times for each method based on 10 simulated data sets. Left column and right column show the results for $(A_{0,\min}, A_{0,\max}) = (0.1, 0.4)$ and $(A_{0,\min}, A_{0,\max}) = (0.4, 0.6)$, respectively.

interval and those after 52nd interval. For each $j = 52, \ldots, 102$, we predict X_{ij} using the best linear predictor based on $X^j := (X_{i1}, \ldots, X_{i,j-1})^T$,

$$\hat{X}_{ij} = \mu_j + \Sigma_{(j,1:(j-1))}\{\Sigma_{(1:(j-1),1:(j-1))}\}^{-1}(X^j - \mu^j),$$

where $\mu_j = \mathbb{E}(X_{1j})$, $\mu^j = (\mu_1, \ldots, \mu_{j-1})^T$ and $\Sigma_{S_1,S_2} = (\sigma_{ij})_{i\in S_1, j\in S_2}$. The unknown parameters are estimated by $\hat{\mu}_j = \sum_{i=1}^{205} X_{ij}/205$ and $\hat{\Sigma} = \hat{\Omega}^{-1}$, where $\hat{\Omega}$ are estimated using various methods including LANCE prior: LANCE prior, YB method (Yu and Bien, 2017), ESC prior (Lee et al., 2019) and CSCS method (Khare et al., 2019). For LANCE prior, $\widehat{a}_j^{(\hat{k}_j)}$ and $\widehat{d}_j^{(\hat{k}_j)}$ are used to construct $\hat{\Omega}$, where $\hat{k}_j$ is the posterior mode. For ESC prior, instead of varying bandwidths, the posterior sample-based mode of the support of the Cholesky factor is used.

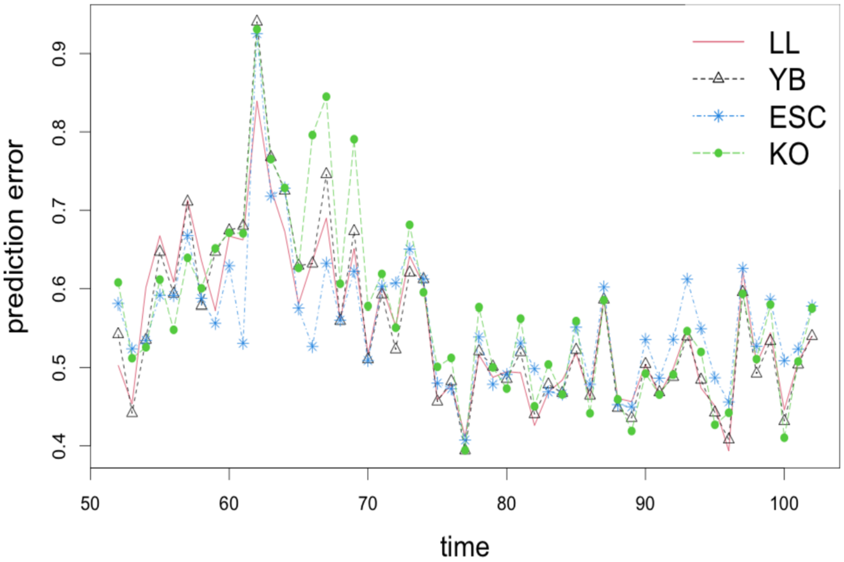

Figure 6: Prediction errors at each time point for each method. LL, YB, ESC and KO represent the methods proposed in this paper, Yu and Bien (2017), Lee et al. (2019) and Khare et al. (2019), respectively.

The absolute prediction error is calculated by $PE_j = \sum_{i=206}^{239} |X_{ij} - \hat{X}_{ij}|/34$ for each $j = 52, \ldots, 102$, and the average of prediction errors, $\sum_{j=52}^{102} PE_j$, is used to evaluate the performance of each method. The hyperparameter in each method is chosen based on cross-validation, except ESC prior. Because applying cross-validation to ESC prior is prohibitive due to heavy computation, we set the hyperparameters in ESC prior at $\gamma = 0.1$, $\nu_0 = 0$, $c_1 = 0.0005$ and $c_2 = 1$ as suggested by Lee et al. (2019). Figure 6 represents prediction errors at each time point. Averages of prediction errors for the methods proposed in this paper, Yu and Bien (2017), Lee et al. (2019) and Khare et al. (2019) are 0.5502, 0.5531, 0.5576 and 0.5708, respectively. It suggests that a local dependence structure is more suitable for the call center data than an arbitrary dependence structure, which makes sense due to the nature of the data.

We further illustrate the performance of LANCE prior in a classification problem. The GunPointAgeSpan data set, which is publicly available at http://timeseriesclassification.com, is used to conduct the quadratic discriminant analysis (QDA). This consists of the two GunPoint data sets released in 2003 and 2018, respectively, and each year, the same two actors (one male and one female) participated in the experiment. There are two classes in a data set: Gun and Point. For the Gun class, the actors hold a gun and point the gun toward a target, while they point with just their fingers (without a gun) in the Point class. The x-axis coordinates of centroid of the hand are recorded over five seconds of the movement based on 30 frames per second, which results in 150 frames ($p = 150$) per action. Thus, this data set also has a time ordering between the variables. The GunPointAgeSpan has 135 training set ($n = 135$)

	LL	YB	ESC	KO
Error	0.2310	0.3956	0.2437	0.3704

Table 1: Classification errors for the test set in the GunPointAgeSpan data.

and 316 test set, and the purpose of the analysis is to classify test observations into the two classes (Gun and Point). The numbers of observations corresponding to the Gun class are 68 and 160 in the training and test data, respectively. The training and test data sets are centered.

For each data x in the test set, the quadratic discriminant score $\delta_k(x)$ is calculated as follows:

$$\delta_k(x) = \frac{1}{2}\log\det(\Omega_k) - \frac{1}{2}(x-\mu_k)^T\Omega_k(x-\mu_k) + \log\left(\frac{n_k}{n}\right), \quad k = 1, 2,$$

where μ_k and Ω_k are the mean vector and precision matrix for the class $k = 1, 2$, respectively. Here, n_k is the number of observations for the class k, thus we have $n_1 = 68$ and $n_2 = 67$. To conduct the QDA, we estimate the unknown parameters μ_k and Ω_k using the training data set. They are plugged into $\delta_k(x)$ similarly to the phone call center data example, and x is then classified as the class $\hat{k} = \mathrm{argmax}_k\, \delta_k(x)$. The performances of LANCE prior, YB method, ESC prior and CSCS method are compared, where cross-validation is used for each method except ESC prior.

Table 1 represents classification errors for the test set based the QDA using $\hat{\Omega}_k$ estimated by each method. For this data set, Bayesian methods seem to achieve lower classification errors than the penalized likelihood approaches, while LANCE prior achieves the lowest classification error. Despite similar performance, LANCE prior has a clear advantage over ESC prior by enabling a scalable cross-validation to select the hyperparameter. In practice, there is no guideline for choosing the hyperparameters in ESC prior, which can dramatically affect the performance.

5 Discussion

In this paper, we propose a Bayesian procedure for high-dimensional local dependence learning, where variables close to each other are more likely to be correlated. The proposed prior, LANCE prior, allows an exact computation of posteriors, which enables scalable inference even in high-dimensional settings. Furthermore, it provides a scalable Bayesian cross-validation to choose the hyperparameters. We establish selection consistency for the local dependence structure and posterior convergence rates for the Cholesky factor. The required conditions for these theoretical results are significantly weakened compared with the existing literature. Simulation studies in various settings show that LANCE prior outperforms other contenders in terms of the ROC curve, cross-validation-based analysis and computation time. Two real data analyses based on the phone call center and gun point data illustrate the satisfactory performance of the proposed method in linear prediction and classification problems, respectively.

It is worth mentioning that LANCE prior is only applicable when k_j is smaller than n due to $(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)})^{-1}$ term in the conditional prior for $a_j^{(k_j)}$. Although we rule out this situation by introducing condition (A3), in practice, we can modify LANCE prior to allow $k_j > n$ when needed. Specifically, we can modify the conditional prior for $a_j^{(k_j)}$ to

$$a_j^{(k_j)} \mid d_j, k_j \overset{ind.}{\sim} N_{k_j}\Big(\tilde{a}_j^{(k_j)}, \frac{d_j}{\gamma}\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + cI_{k_j}\big)^{-1}\Big),\ j = 2, \ldots, p, \tag{6}$$

for some constant $c > 0$, where $\tilde{a}_j^{(k_j)} = \big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + cI_{k_j}\big)^{-1}\mathbf{X}_{j(k_j)}^T\tilde{X}_j$. Note that the above prior (6) has a variance stabilizing term cI_{k_j}. Then, the resulting α-fractional posterior has the following closed form:

$$\begin{aligned}
a_j^{(k_j)} \mid d_j, k_j, \mathbf{X}_n \overset{ind.}{\sim} N_{k_j}\Big(&\big\{\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + \frac{c\gamma}{\alpha+\gamma}I_{k_j}\big\}^{-1}\mathbf{X}_{j(k_j)}^T\tilde{X}_j, \\
&d_j\big\{(\alpha+\gamma)\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + c\gamma I_{k_j}\big\}^{-1}\Big),\ j = 2, \ldots, p, \\
d_j \mid k_j, \mathbf{X}_n \overset{ind.}{\sim} IG\Big(&\frac{\alpha n + \nu_0}{2}, \frac{\alpha n}{2}\tilde{d}_j^{(k_j)}\Big),\ j = 1, \ldots, p, \\
\pi_\alpha(k_j \mid \mathbf{X}_n) \propto \pi(k_j)\Big\{&\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + cI_{k_j}\big)^{-1}\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + \frac{c\gamma}{\alpha+\gamma}I_{k_j}\big)\Big\}^{-\frac{k_j}{2}} \\
&\times\Big(1 + \frac{\alpha}{\gamma}\Big)^{-\frac{k_j}{2}}\big(\tilde{d}_j^{(k_j)}\big)^{-\frac{\alpha n+\nu_0}{2}},\ j = 2, \ldots, p,
\end{aligned}$$

where

$$\begin{aligned}
\tilde{d}_j^{(k_j)} &= n^{-1}\tilde{X}_j^T\big\{I_n - \mathbf{X}_{j(k_j)}\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + \frac{c\gamma}{\alpha+\gamma}I_{k_j}\big)^{-1}\mathbf{X}_{j(k_j)}^T\big\}\tilde{X}_j \\
&+ \tilde{X}_j^T\mathbf{X}_{j(k_j)}\big\{\big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + \frac{c\gamma}{\alpha+\gamma}I_{k_j}\big)^{-1} - \big(\mathbf{X}_{j(k_j)}^T\mathbf{X}_{j(k_j)} + cI_{k_j}\big)^{-1}\big\}\mathbf{X}_{j(k_j)}^T\tilde{X}_j\,\gamma/(\alpha n).
\end{aligned}$$

Thus, the proposed LANCE prior can be modified to allow large bandwidths such that $k_j > n$. A sufficiently small $c > 0$ would give similar results with (5) in practice, although theoretical properties of the resulting α-fractional posterior should be further investigated.

An important and interesting future direction is to consider the estimation of covariance or precision structures with an unknown topological ordering on the variables of a DAG. Assume one specifies an arbitrary label of the variables as $x_1, \ldots, x_p$ which does not necessarily correspond to the ordered variables. Let $\sigma \in \mathcal{S}_p$ be an element in the permutation group of p variables and G_σ is the corresponding true DAG structure. If one is interested in learning both the DAG G_σ and the resulting covariance structure, a natural attempt is to impose a prior on the DAG G along with a prior on the covariance structure given σ or the DAG, and then perform posterior inference. However, some identifiability conditions need to be imposed for the identifiability of the DAG structure. Please see Theorem 2.2 of Park and Kim (2020) for a state-of-the-art result on the identifiability of a Gaussian DAG, which essentially says that a Gaussian DAG is identifiable if the uncertainty level of a node j is smaller than that of its descendants, given the non-descendants. One can potentially design a prior for the variances (d_js in our notation) that satisfies this ordered constraint. It would be interesting to study the posterior contraction and model selection properties of this model and design efficient MCMC algorithms.

If one does not impose any identifiability conditions (e.g., through the prior), another direction to deal with the unknown ordering case is to learn the equivalent class (or a representative of the class) of a DAG for all σs, the so-called *structure learning of DAGs* which will learn the covariance structure of an equivalent class without learning the topological ordering of the variable or the underlying DAG structure. A recent work Zhou and Chang (2021) is one such example. The key ideas are the following: Let $N_p(0, \Sigma^*)$ be the distribution of some true Gaussian DAG model G^*. One can show that (see definition 7 in Zhou and Chang (2021) for example), for any $\sigma \in \mathcal{S}_p$, one can have the Cholesky factor matrix A^*_σ and the positive diagonal matrix D^*_σ such that $\Sigma^* = (I_p - A^*_\sigma)^{-1} D^*_\sigma (I_p - (A^*_\sigma)^T)^{-1}$. That is, for any σ, one can find a unique pair (A^*_σ, D^*_σ) which corresponds to some DAG G^*_σ called minimal I-map of the equivalent classes which has the same covariance structure as G^*. This minimal I-map can be viewed as a representative point of the equivalent class which can be uniquely constructed. The structure learning problems boil down to the learnings of $\{G^*_\sigma, \sigma \in \mathcal{S}_p\}$ essentially. One can design appropriate priors that can obtain posterior consistency over $\{G^*_\sigma, \sigma \in \mathcal{S}_p\}$.

Another work that deals with unknown ordering is Cao and Zhang (2020) which proposed to sample K permutation matrices or σs under the MCD model. They obtained the posterior estimates of the Cholesky factors and diagonal matrices for each permutation which are then averaged to obtain a final estimate of the precision matrix.

Supplementary Material

Supplementary to "Scalable Bayesian high-dimensional local dependence learning". (DOI: 10.1214/21-BA1299SUPP; .pdf). It contains the proofs of the main results in this paper and additional simulation results.

References

An, B., Guo, J., and Liu, Y. (2014). "Hypothesis testing for band size detection of high-dimensional banded precision matrices." *Biometrika*, 101(2): 477–483. MR3215361. doi: https://doi.org/10.1093/biomet/asu006. 27

Banerjee, S. and Ghosal, S. (2014). "Posterior convergence rates for estimating large precision matrices using graphical models." *Electronic Journal of Statistics*, 8(2): 2111–2137. MR3273620. doi: https://doi.org/10.1214/14-EJS945. 27, 31

Banerjee, S. and Ghosal, S. (2015). "Bayesian structure learning in graphical models." *Journal of Multivariate Analysis*, 136: 147–162. MR3321485. doi: https://doi.org/10.1016/j.jmva.2015.01.015. 30

Bhattacharya, A., Pati, D., Yang, Y., et al. (2019). "Bayesian fractional posteriors." *The Annals of Statistics*, 47(1): 39–66. MR3909926. doi: `https://doi.org/10.1214/18-AOS1712`. 29

Bickel, P. J. and Levina, E. (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, 36(1): 199–227. MR2387969. doi: `https://doi.org/10.1214/009053607000000758`. 26, 31, 38, 39

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg. MR2807761. doi: `https://doi.org/10.1007/978-3-642-20192-9`. 30

Cai, T. T., Liu, W., and Zhou, H. H. (2016). "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation." *The Annals of Statistics*, 44(2): 455–488. MR3476606. doi: `https://doi.org/10.1214/13-AOS1171`. 26

Cao, X., Khare, K., and Ghosh, M. (2019). "Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models." *The Annals of Statistics*, 47(1): 319–348. MR3909935. doi: `https://doi.org/10.1214/18-AOS1689`. 31

Cao, X. and Zhang, S. (2020). "A permutation-based Bayesian approach for inverse covariance estimation." *Communications in Statistics – Theory and Methods*, 49(14): 3557–3571. MR4107619. doi: `https://doi.org/10.1080/03610926.2019.1590601`. 44

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and Computing*, 24(6): 997–1016. MR3253850. doi: `https://doi.org/10.1007/s11222-013-9416-2`. 34

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika*, 93(1): 85–98. MR2277742. doi: `https://doi.org/10.1093/biomet/93.1.85`. 38, 39

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007): 453–461. MR0017504. doi: `https://doi.org/10.1098/rspa.1946.0056`. 29, 34

Johnstone, I. M. and Lu, A. Y. (2009). "On consistency and sparsity for principal components analysis in high dimensions." *J. Amer. Statist. Assoc.*, 104(486): 682–693. MR2751448. doi: `https://doi.org/10.1198/jasa.2009.0121`. 25

Khare, K., Oh, S.-Y., Rahman, S., and Rajaratnam, B. (2019). "A scalable sparse Cholesky based approach for learning high-dimensional covariance matrices in ordered data." *Machine Learning*, 108(12): 2061–2086. MR4026660. doi: `https://doi.org/10.1007/s10994-019-05810-5`. 27, 30, 31, 32, 36, 40, 41

Lee, K. and Lee, J. (2018). "Optimal Bayesian minimax rates for unconstrained large covariance matrices." *Bayesian Analysis*, 13(4): 1215–1233. MR3855369. doi: `https://doi.org/10.1214/18-BA1094`. 26

Lee, K. and Lee, J. (2021). "Estimating large precision matrices via modified Cholesky decomposition." *Statistica Sinica*, 31(1): 173–196. MR4270383. doi: https://doi.org/10.5705/ss.20. 27, 31

Lee, K., Lee, J., and Lin, L. (2019). "Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors." *The Annals of Statistics*, 47(6): 3413–3437. MR4025747. doi: https://doi.org/10.1214/18-AOS1783. 26, 27, 28, 29, 30, 31, 32, 33, 40, 41

Lee, K. and Lin, L. (2020). "Bayesian bandwidth test and selection for high-dimensional banded precision matrices." *Bayesian Anal.*, 15(3): 737–758. MR4132648. doi: https://doi.org/10.1214/19-BA1167. 27

Lee, K. and Lin, L. (2022). "Supplementary material for: Scalable Bayesian High-dimensional Local Dependence Learning." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1299SUPP. 28

Liu, Y. and Ren, Z. (2020). "Minimax estimation of large precision matrices with bandable Cholesky factor." *The Annals of Statistics*, 48(4): 2428–2454. MR4134801. doi: https://doi.org/10.1214/19-AOS1893. 33

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 29, 30

Martin, R. and Walker, S. G. (2014). "Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector." *Electronic Journal of Statistics*, 8(2): 2188–2206. MR3273623. doi: https://doi.org/10.1214/14-EJS949. 29

Park, G. and Kim, Y. (2020). "Identifiability of Gaussian structural equation models with homogeneous and heterogeneous error variances." *Journal of the Korean Statistical Society volume*, 49: 276–292. MR4122465. doi: https://doi.org/10.1007/s42952-019-00019-7. 43

Pourahmadi, M. (1999). "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation." *Biometrika*, 86(3): 677–690. MR1723786. doi: https://doi.org/10.1093/biomet/86.3.677. 26

Rothman, A. J., Levina, E., and Zhu, J. (2010). "A new approach to Cholesky-based covariance regularization in high dimensions." *Biometrika*, 97(3): 539–550. MR2672482. doi: https://doi.org/10.1093/biomet/asq022. 26

Rütimann, P. and Bühlmann, P. (2009). "High dimensional sparse covariance estimation via directed acyclic graphs." *Electronic Journal of Statistics*, 3: 1133–1160. MR2566184. doi: https://doi.org/10.1214/09-EJS534. 26

Shojaie, A. and Michailidis, G. (2010). "Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs." *Biometrika*, 97(3): 519–538. MR2672481. doi: https://doi.org/10.1093/biomet/asq038. 26

van de Geer, S. and Bühlmann, P. (2013). "ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs." *The Annals of Statistics*, 41(2): 536–567. MR3099113. doi: https://doi.org/10.1214/13-AOS1085. 26, 28

Wainwright, M. J. (2009). "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)." *IEEE Transactions on Information Theory*, 55(5): 2183–2202. MR2729873. doi: https://doi.org/10.1109/TIT.2009.2016018. 32

Yu, G. and Bien, J. (2017). "Learning local dependence in ordered data." *Journal of Machine Learning Research*, 18(42): 1–60. MR3655307. 27, 31, 32, 33, 35, 36, 37, 40, 41

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 29

Zhou, Q. and Chang, H. (2021). "Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes." 44

Acknowledgments

We are very grateful to the Associate Editor and the two reviewers for their valuable comments which have led to great improvement in our paper. We would also like to thank Guo (Hugo) Yu for very helpful discussions.