Bayesian Analysis (2023) **18**, Number 3, pp. 1017–1040

Post-Processed Posteriors for Banded Covariances[*] [†]

Kwangmin Lee[‡,§], Kyoungjae Lee[¶] and Jaeyong Lee[‖]

Abstract. We consider Bayesian inference of banded covariance matrices and propose a post-processed posterior. The post-processing of the posterior consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior, which does not satisfy any structural restrictions. In the second step, the posterior samples are transformed to satisfy the structural restriction through a post-processing function. The conceptually straightforward procedure of the post-processed posterior makes its computation efficient and can render interval estimators of functionals of covariance matrices. We show that it has nearly optimal minimax rates for banded covariances among all possible pairs of priors and post-processing functions. Additionally, we provide a theorem on the credible set of the post-processed posterior under the finite dimension assumption. We prove that the expected coverage probability of the $100(1-\alpha)\%$ highest posterior density region of the post-processed posterior is asymptotically $1-\alpha$ with respect to any conventional posterior distribution. It implies that the highest posterior density region of the post-processed posterior is, on average, a credible set of conventional posterior. The advantages of the post-processed posterior are demonstrated by a simulation study and a real data analysis.

Keywords: minimax rate, inverse-Wishart, Bayesian, high-dimensional analysis.

MSC2020 subject classifications: Primary 62C20, 62F15; secondary 62H10.

1 Introduction

In this paper, we propose a new Bayesian procedure for banded covariance matrices. The banded matrices are the matrices whose entries farther than a certain distance from the diagonal are all zeros. Banded covariance matrices arise in modelling marginal dependence structures of variables with natural ordering such as time series data. The banded sample covariance has been applied to the autoregressive and moving average models (Wu and Pourahmadi, 2009) and the time-varying autoregressive-moving-average models (Wiesel and Globerson, 2012).

arXiv: 2011.12627

[*] Kwangmin Lee was supported by Chonnam National University (Grant number: 2023-0482) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00211979). Kyoungjae Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1F1A1063905). Jaeyong Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1A2A3074973 and 2020R1A4A1018207).

[†] Funding information has been updated on 31 May 2023.

[‡] Department of Big Data Convergence, Chonnam National University, klee564@jnu.ac.kr

[§] Department of Mathematics and Statistics, Chonnam National University

[¶] Department of Statistics, Sungkyunkwan University, leekjstat@gmail.com

[‖] Department of Statistics, Seoul National University, leejyc@gmail.com

https://doi.org/10.1214/22-BA1333

When p is small relative to n, the inverse-Wishart prior is the most commonly used conjugate prior for the covariance of the multivariate normal model. We denote $\Sigma \sim IW_p(\Lambda, \nu)$, if it has density $\pi(\Sigma) \propto |\Sigma|^{-\nu/2} \exp\{-\mathrm{tr}(\Sigma^{-1}\Lambda)/2\}$, for any $p \times p$ positive definite matrix Σ, where $\nu > 2p$ is the degree of freedom, and $|\Sigma|$ is the determinant of Σ. The inverse-Wishart prior has many nice properties under the traditional setting of a small p. The posterior induced by the inverse-Wishart prior attains the optimal minimax rate when $p \leq cn,\ 0 \leq c < 1$, under the spectral norm (Lee and Lee, 2018). The Jeffreys prior for covariance matrices (Yang and Berger, 1996) can be expressed as the limit of the inverse-Wishart prior as the degree of freedom and the scale matrix converge to $p+1$ and the $p \times p$ zero matrix, respectively. When the degree of freedom is $2p+1$ and the scale matrix is a diagonal matrix, the marginal distribution of each correlation induced by the inverse-Wishart prior follows a uniform distribution over the interval $[-1, 1]$ (Huang and Wand, 2013); thus it can be viewed as a non-informative prior for correlations.

When $p \geq n$, however, Lee and Lee (2018) showed that the degenerate prior δ_{I_p}, an obviously inadequate prior, attains the optimal minimax rate, implying that without further assumptions, the inference of the covariance matrix is hopeless. This is expected because, without any constraint, the number of parameters, $p(p+1)/2$, in the covariance matrix is much larger than the sample size n. To reduce the number of effective parameters, several matrix classes have been proposed including bandable matrices (Cai and Zhou, 2010; Banerjee and Ghosal, 2014), sparse matrices (Cai and Liu, 2011; Cai et al., 2016; Lee et al., 2019) and low-dimensional structural matrices (Cai et al., 2013; Pati et al., 2014; Gao and Zhou, 2015).

In this paper, we focus on the banded covariance assumption. The banded covariance assumption is a popular structural assumption to reduce the number of effective parameters, especially when there is a natural ordering between variables. From the frequentist side, the banded covariance estimator has been studied extensively. Since the banded covariance structure is an example of the Gaussian covariance graph model, the methods of Kauermann (1996) and Chaudhuri et al. (2007) can be used for the estimation of the banded covariance. However, the two methods are originally designed for the case of $p < n$, and need a modification of the sample covariance matrix for the case of $p \geq n$. Bickel and Levina (2008) focused on the bandable covariance structure and obtained the convergence rate of the banded sample covariance. Despite of a few point estimation methods, there is no frequentist interval estimation method in high-dimensional settings. Chaudhuri et al. (2007) suggested an interval estimation of banded covariances under the asymptotic normality of the maximum likelihood estimator, which is valid only for fixed p. Also, there is no minimax rate result in the literature for banded covariance matrices, although Cai and Zhou (2010, 2012a) showed the tapering estimator satisfies the optimal minimax rate for bandable covariances, which are the matrices whose entries are getting smaller as they are more distant from the diagonal.

Compared to frequentist methods, Bayesian methods have a natural advantage of producing interval estimators automatically. However, Bayesian methods for banded covariance matrices that are scalable and supported by theoretical properties in high-dimensional settings are scarce. This is due to the difficulty of inventing a tractable prior distribution on the space of banded covariances. Khare and Rajaratnam (2011) and Silva and Ghahramani (2009) proposed prior distributions for covariance graphical model, which can be used for banded covariance matrices, but there are no minimax optimality results for these methods. This is partly because there are no closed forms of normalizing constants for these priors, which prevents direct investigation of posterior asymptotics. It is also mathematically challenging to apply traditional posterior consistency and contraction theorems (Ghosal and Van der Vaart, 2017), which are applicable when it is hard to tract posterior directly.

In summary, there are no Bayesian or frequentist methods for banded covariance matrices, which (1) are computationally efficient, (2) produce interval estimators for functionals of covariance matrices, and (3) have optimal or nearly optimal minimax rate. In this paper, we propose a new Bayesian method that has the above three properties. In particular, we propose *post-processed posteriors* for banded covariance matrices.

The construction of the post-processed posterior consists of two steps, *the initial posterior computing step* and *the post-processing step*. In the initial posterior computing step, posterior samples are generated from the initial posterior, the conjugate inverse-Wishart posterior for covariance, without any structural restrictions. In the post-processing step, the initial posterior samples are transformed through a function $f(\Sigma)$ whose range belongs to a space of banded covariances. We call the distribution of the transformed posterior samples the post-processed posterior, which will be rigorously defined in Section 2.

The idea of transforming the posterior samples has been suggested in various settings. Posterior projection methods (Dunson and Neelon, 2003; Gunn and Dunson, 2005; Lin and Dunson, 2014; Patra et al., 2018) are proposed for various problems, which project the posterior samples onto the constrained parameter space to obtain the projected posterior. Our proposal is the same as the posterior projection method in spirit, but the choice of posterior transformation is determined through asymptotic consideration, while the posterior projection method uses the projection on the constrained space. In fact, our proposal is the posterior projection method on the space of banded covariances with the Frobenius norm. Recently, Bashir et al. (2018) proposed a support recovery method for sparse precision matrices based on post-processing of the posterior samples.

The post-processed posterior is conceptually straightforward and computationally fast. This is advantageous when the data set is huge and the dimension of the observations is high. The existing Bayesian method can be slow at times especially in high-dimensional settings. Through the simulation study, we will show that the post-processed posterior significantly reduces the computation time compared to the covariance graphical models proposed by Silva and Ghahramani (2009) and Khare and Rajaratnam (2011). Furthermore, the post-processed posterior attains the nearly optimal minimax rate for the class of banded covariance matrices. This is the first minimax result for banded covariance matrices in both Bayesian and frequentist sides.

The banded covariance matrices have been investigated as a case of covariance graphical model, but the minimax lower bound for covariance graphical model is absent in the literature. Methods for obtaining minimax lower bound, e.g., Le Cam's method and Assouad's lemma, are based on the testing problem of δ-separated sets as described in Proposition 15.1 of Wainwright (2019). Since patterns of parameter spaces of covariance graphical model differ by the graphs, it is not easy to choose representative separated sets for arbitrary graphical structures. Instead, we focus on the banded covariance structure and could choose appropriate separated sets. We also show that the post-processed posterior has the nearly optimal minimax rate for the class of bandable covariances, which is given in the supplementary material (Lee et al., 2022).

It is worth mentioning that there are substantial differences between banded covariance and precision matrices. For banded precision matrices, G-Wishart priors (Banerjee and Ghosal, 2014) or banded Cholesky priors (Lee and Lee, 2021) can be used, and the normalizing constants are available in a closed form. However, by contrast, imposing a prior on the Cholesky factor of a banded covariance matrix will result in a non-conjugate posterior whose normalizing constant is intractable. Intuitively, in the Bayesian framework, constraints on precision matrices are more manageable than those on covariance matrices because the precision matrix is a natural parameter of multivariate normal distributions as an exponential family. In other words, the likelihood function of the covariance is expressed through the precision matrix. Thus, Bayesian banded covariance matrix estimation is more challenging than banded precision matrix estimation.

There is difference in the estimation methods of covariance and precision matrices in the frequentist literature as well. The sparse covariance estimation is typically based on banding or thresholding the sample covariance (Cai and Zhou, 2012b) while the sparse precision matrix estimation is often based on the penalized likelihood approach (Cai et al., 2011; Zhang and Zou, 2014). The difference is due to the form of the likelihood function as well as the singularity of the sample covariance matrix. Contrary to the sample covariance matrix, the sample precision matrix, the inverse of the sample covariance matrix, does not exist when $p > n$, which prevents thresholding the sample precision matrix. One could choose a small constant $\epsilon > 0$ to make $S_n + \epsilon I_p$ invertible and use $(S_n + \epsilon I_p)^{-1}$ instead of S_n^{-1} as the sample precision matrix, but it can be computationally unstable especially when ϵ is small.

The rest of the paper is organized as follows. In Section 2, the post-processed posterior is introduced for the banded covariances. In Section 3, it is shown that the banding post-processed posterior attains the nearly optimal minimax rate for banded covariance matrices, and the expected coverage probability of the $(1-\alpha)100\%$ highest posterior density region of the post-processed posterior is asymptotically $1-\alpha$ with respect to a conventional posterior distribution. In Section 4, the post-processed posterior is demonstrated via simulation studies and a real data analysis. The supplementary material contains the proofs of the theorems in the paper, a minimax result for bandable covariance matrices and more numerical studies.

2 Post-Processed Posterior

Suppose $X_1, \ldots, X_n$ are independent and identically distributed samples from $N_p(0, \Sigma)$, the p-dimensional normal distribution with zero mean vector and covariance matrix

$\Sigma = (\sigma_{ij}) > 0$. We write $B > 0$ ($B \geq 0$) if B is a positive (nonnegative) definite matrix. When the variables have a natural ordering such as time or causal relationship, it is commonly assumed that the covariance satisfies a band structure. In this paper, we assume that Σ is banded:

$$\Sigma \in \tilde{\mathcal{B}}_{p,k} := \Big\{ \Sigma \in \mathcal{C}_p : \sigma_{ij} = 0 \text{ if } |i-j| > k, \forall i,j \in [p] \Big\},$$

where k is a natural number, $[p] = \{1, 2, \dots, p\}$ and $\mathcal{C}_p$ is the set of all $p \times p$ positive definite matrices.

We propose a computationally efficient and theoretically supported Bayesian method for banded covariance matrices. The proposed method consists of two steps: the initial posterior computing step and the post-processing step. We describe these two steps in detail below.

Step 1. (Initial posterior computing step)

In the initial posterior step, a conjugate posterior for the parameter space without any structural restriction is obtained. We take the inverse-Wishart prior $IW_p(B_0, \nu_0)$. We say this is the *initial prior* π^i for Σ. By conjugacy, the *initial posterior* is then

$$\Sigma \mid \mathbb{X}_n \sim IW_p(B_0 + nS_n, \nu_0 + n),$$

where $\mathbb{X}_n = (X_1, \dots, X_n)^T$ and $S_n = n^{-1} \sum_{i=1}^n X_i X_i^T$ is the sample covariance matrix. We sample $\Sigma^{(1)}, \Sigma^{(2)}, \dots, \Sigma^{(N)}$ from the initial posterior, $\pi^i(\Sigma | \mathbb{X}_n)$.

Step 2. (Post-processing step)

Let the function $B_k(B)$ denote the k-band operation,

$$B_k(B) = \{b_{ij} I(|i-j| \leq k)\}$$

for any $B = (b_{ij}) \in \mathbb{R}^{p \times p}$. In the second step, we post-process the samples from the initial posterior to obtain those from the post-processed posterior. The samples from the post-processed posterior, $\Sigma_{(i)}$'s, are defined by

$$\begin{aligned} \Sigma_{(i)} &= f(\Sigma^{(i)}) = B_k^{(\epsilon_n)}(\Sigma^{(i)}) \qquad (1) \\ &:= \begin{cases} B_k(\Sigma^{(i)}) + \Big[\epsilon_n - \lambda_{\min}\{B_k(\Sigma^{(i)})\}\Big] I_p, & \text{if } \lambda_{\min}\{B_k(\Sigma^{(i)})\} < \epsilon_n, \\ B_k(\Sigma^{(i)}), & \text{otherwise,} \end{cases} \end{aligned}$$

where ϵ_n is a small positive number decreasing to 0 as $n \to \infty$, for $i = 1, \dots, N$. There is no guarantee that $B_k(\Sigma^{(i)})$ is positive definite, so the second term of (1) is added to make $\Sigma_{(i)}$ positive definite. The resulting post-processed samples, $(\Sigma_{(1)}, \dots, \Sigma_{(N)})$, are banded positive definite matrices. We suggest using the samples from the post-processed posterior for Bayesian inference of banded covariance matrices.

We call the posterior distribution of (1) the k-banding post-processed posterior to emphasize that the k-band operation B_k is used; however, other operations can be used to obtain the desired structure. We call the function f represented by (1) the *post-processing function*, and the post-processed posterior with the post-processing function f is denoted by $\pi^{pp}(\cdot|\mathbb{X}_n; f)$ or simply $\pi^{pp}(\cdot|\mathbb{X}_n)$ if f is understood in the context.

3 Properties of Post-Processed Posterior

3.1 Minimax Convergence Rates

In this section, we show that the proposed post-processed posterior procedure is nearly optimal in the minimax sense among all possible post-processed posterior procedures, the pairs of initial priors and post-processing functions. A conventional Bayesian procedure can be considered as a post-processed posterior procedure, which has a prior supported on $\tilde{\mathcal{B}}_{p,k}$ and an identity post-processing function. Thus, the proposed post-processed posterior is nearly optimal even compared with conventional Bayesian procedures.

Lee and Lee (2018) proposed a decision-theoretic framework for comparison of priors. In this framework, a posterior and the space of all probability measures on the parameter space are considered as an action and the action space, respectively. A prior is a decision rule in this setting because a prior combined with data generates a posterior. The posterior-loss (P-loss) and posterior-risk (P-risk) (Lee and Lee, 2018) are the loss and risk functions.

The decision-theoretic framework of Lee and Lee (2018) can be modified for the study of the minimax properties of post-processed posterior. In this setting, a post-processed posterior is an action and a post-processed posterior procedure, a pair of an initial prior and a post-processing function, is a decision rule. We define the P-loss and P-risk of the post-processed posterior as follows:

$$\begin{aligned}\mathcal{L}\{\Sigma_0, \pi^{pp}(\cdot \mid \mathbb{X}_n; f)\} &:= \mathrm{E}^{\pi^{pp}}(||\Sigma_0 - \Sigma|| \mid \mathbb{X}_n) \\ &= \mathrm{E}^{\pi^i}\{||\Sigma_0 - f(\Sigma)|| \mid \mathbb{X}_n\}, \\ \mathcal{R}(\Sigma_0, \pi^{pp}) &:= \mathrm{E}_{\Sigma_0}[\mathcal{L}\{\Sigma_0, \pi^{pp}(\cdot \mid \mathbb{X}_n; f)\}] \\ &= \mathrm{E}_{\Sigma_0}[\mathrm{E}^{\pi^i}\{||\Sigma_0 - f(\Sigma)|| \mid \mathbb{X}_n\}],\end{aligned}$$

where E^{π^i} and E_{Σ_0} denote expectations with respect to $\Sigma \sim \pi^i$ and random samples $X_1, \ldots, X_n$ from $N_p(0, \Sigma_0)$, respectively, Σ_0 is the true parameter of the Σ, and $||A|| := \{\lambda_{\max}(AA^T)\}^{1/2}$ is the spectral norm of a symmetric matrix A. We define the maximum P-risk given a parameter space Θ_0, in which the true parameter is believed to reside, as

$$\sup_{\Sigma_0 \in \Theta_0} \mathrm{E}_{\Sigma_0}[\mathcal{L}\{\Sigma_0, \pi^{pp}(\cdot \mid \mathbb{X}_n; f)\}].$$

We now define the minimax rate and convergence rate for post-processed posteriors. Let

$$\Pi^* = \{\pi^{pp}(\cdot; f) = (\pi, f) : \pi \in \Pi, f \in \mathcal{F}\}$$

be the space of all possible post-processing procedures, where Π is the space of all priors on $\mathcal{C}_p$, and $\mathcal{F}$ is the space of all possible post-processing functions, for example, $\mathcal{F}_k^B = \{f : \mathcal{C}_p \to \tilde{\mathcal{B}}_{p,k}\}$.

Before we give some definitions of minimax rates, we introduce some notation. For any positive sequences a_n and b_n, we denote $a_n = o(b_n)$ if $a_n/b_n \longrightarrow 0$ as $n \to \infty$, and $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all sufficiently large n. We denote $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

A sequence r_n is said to be the minimax rate for Π^* over Θ_0 if

$$\inf_{(\pi,f)\in\Pi^*} \sup_{\Sigma_0\in\Theta_0} \mathrm{E}_{\Sigma_0}[\mathcal{L}\{\Sigma_0, \pi^{pp}(\cdot|\mathbb{X}_n; f)\}] \asymp r_n,$$

and a post-processing procedure $(\pi, f) \in \Pi^*$ is said to have P-risk convergence rate a_n if

$$\sup_{\Sigma_0\in\Theta_0} \mathrm{E}_{\Sigma_0}[\mathcal{L}\{\Sigma_0, \pi^{pp}(\cdot|\mathbb{X}_n; f)\}] \lesssim a_n.$$

If $a_n \asymp r_n$ and r_n is the P-risk minimax rate, $(\pi, f) \in \Pi^*$ is said to attain the P-risk minimax rate.

We are now ready to state that the banding post-processed posterior attains nearly minimax rate in terms of the P-risk over banded covariance matrices. Suppose that we observe the data $X_1, \ldots, X_n$ from p-dimensional normal distribution, $N_p(0, \Sigma_0)$, with $\Sigma_0 \in \mathcal{B}_{p,k_0}$, where

$$\mathcal{B}_{p,k_0} = \big\{\Sigma \in \tilde{\mathcal{B}}_{p,k_0} : \lambda_{\max}(\Sigma) < M_0, \lambda_{\min}(\Sigma) > M_1\big\},$$

where $0 < M_1 \leq M_0 < \infty$, $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ are the minimum and maximum eigenvalues of Σ, respectively, and k_0 represents the bandwidth of the true covariance.

The following theorems say that the P-risk of the banding post-processed posterior is nearly minimax optimal.

Theorem 3.1. *Let the initial prior π^i of Σ be $IW_p(A_n, \nu_n)$. If $A_n \in \mathcal{B}_{p,k}$, $n/4 \geq (M_0^{1/2} M_1^{-1} \log p) \vee k \vee ||A_n|| \vee (\nu_n - 2p)$, $\nu_n > 2p$ and $||A_n|| \vee (\nu_n - 2p) = o(n)$, then*

$$\sup_{\Sigma_0\in\mathcal{B}_{p,k_0}} \mathrm{E}_{\Sigma_0}\{\mathrm{E}^{\pi^i}(||B_k^{(\epsilon_n)}(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \leq C\{(\log k)^2 \frac{k + \log p}{n} + (k_0 - k)^2 I(k_0 > k)\},$$

where the post-processing function $B_k^{(\epsilon_n)}$ is defined in (1), $\epsilon_n^2 = O\{(\log k)^2(k + \log p)/n\}$, and C is a constant that depends on M_0 and M_1.

Theorem 3.2. *If $n/2 \geq [\min\{(M_0 - M_1)^2, 1\} \log p] \vee k_0$ and $\nu_n > 2p$, then*

$$\inf_{(\pi,f)\in\Pi^*} \sup_{\Sigma_0\in\mathcal{B}_{p,k_0}} \mathrm{E}_{\Sigma_0}\{\mathrm{E}^{\pi^i}(||f(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \geq C \frac{k_0 + \log p}{n},$$

where C is a constant that depends on M_0 and M_1.

Theorem 3.1 gives the convergence rate of the P-risk of the banding post-processed posterior for a class of banded covariance matrices $\mathcal{B}_{p,k_0}$, while a minimax lower bound is given in Theorem 3.2. Note that we distinguish the true bandwidth k_0 from bandwidth k in the post-processing function. If $k = k_0$, the obtained convergence rate in Theorem 3.1 matches the minimax lower bound up to a $(\log k_0)^2$ factor. When $k_0 \le k \le C_1 k_0$ for some positive constant C_1, the obtained convergence rates are asymptotically equivalent to the nearly minimax rate, $(\log k_0)^2(k_0 + \log p)/n$. While Theorem 3.1 gives the convergence rate of the maximum P-risk over the parameter space $\mathcal{B}_{p,k_0}$, we also consider the convergence rate given a fixed $\Sigma_0 \in \mathcal{B}_{p,k_0}$ in the following remark.

Remark. *Given a fixed $\Sigma_0 \in \mathcal{B}_{p,k_0}$, we have the convergence rate as below:*

$$\mathrm{E}_{\Sigma_0}\{\mathrm{E}^{\pi^i}(||B_k^{(\epsilon_n)}(\Sigma) - \Sigma_0||^2 \mid \mathbb{X}_n)\} \le C\{(\log k)^2 \frac{k + \log p}{n} + ||B_k(\Sigma_0) - \Sigma_0||^2\},$$

where C is some positive constant depending on M_0 and M_1. Even when $k < k_0$, if $||B_k(\Sigma_0) - \Sigma_0||^2$ is small enough so that $||B_k(\Sigma_0) - \Sigma_0||^2 = O((\log k_0)^2(k_0 + \log p)/n)$, the convergence rate is still nearly minimax.

In practice, an outcome of the post-processed posterior may not be an element of $\mathcal{B}_{p,k_0}$. In the following Theorem 3.3, we show that the probability that the post-processed posterior sample belongs to $\mathcal{B}_{p,k_0}$ converges to 1 as $n \longrightarrow \infty$. Thus, when n is large enough, a post-processed posterior sample resides in the parameter space $\mathcal{B}_{p,k_0}$ with a large probability tending to 1.

Theorem 3.3. *Suppose $X_1, \ldots, X_n$ are generated from $N_p(0, \Sigma_0)$ with $\Sigma_0 \in \mathcal{B}_{p,k_0}$. Let the prior π^i of Σ be $IW_p(A_n, \nu_n)$. If $A_n \in \mathcal{B}_{p,k_0}$, $n/4 \ge (M_0^{1/2} M_1^{-1} \log p) \vee k_0 \vee ||A_n|| \vee (\nu_n - 2p)$, $\nu_n > 2p$, $||A_n|| \vee (\nu_n - 2p) \vee k_0 \vee \log p = o(n)$ and $\epsilon_n^2 = O\{(\log k_0)^2(k_0 + \log p)/n\}$, then*

$$\mathrm{Pr}^{\pi^i}(B_{k_0}^{(\epsilon_n)}(\Sigma) \in \mathcal{B}_{p,k_0} \mid \mathbb{X}_n) \xrightarrow{p} 1,$$

as $n \longrightarrow \infty$, where $\xrightarrow{p}$ means convergence in probability.

Proof. We show $\mathrm{Pr}_{\Sigma_0}[\mathrm{Pr}^{\pi^i}(B_{k_0}^{(\epsilon_n)}(\Sigma) \notin \mathcal{B}_{p,k_0} \mid \mathbb{X}_n) > \delta] \longrightarrow 0$, as $n \longrightarrow \infty$ for all $\delta > 0$. Let $\delta_2 = (M_0 - \lambda_{\max}(\Sigma_0)) \wedge (\lambda_{\min}(\Sigma_0) - M_1)$. We have

$$\begin{aligned}\mathrm{Pr}^{\pi^i}(B_{k_0}^{(\epsilon_n)}(\Sigma) \notin \mathcal{B}_{p,k_0} \mid \mathbb{X}_n) &\le \mathrm{Pr}^{\pi^i}(||B_{k_0}^{(\epsilon_n)}(\Sigma) - \Sigma_0||_2 > \delta_2/2 \mid \mathbb{X}_n) \\ &\le 4\delta_2^{-2}\mathrm{E}^{\pi^i}(||B_{k_0}^{(\epsilon_n)}(\Sigma) - \Sigma_0||_2^2 \mid \mathbb{X}_n),\end{aligned}$$

for all sufficiently small $\epsilon_n > 0$. By Theorem 3.1, we get

$$\begin{aligned}&\mathrm{Pr}_{\Sigma_0}[\mathrm{Pr}^{\pi^i}(B_{k_0}^{(\epsilon_n)}(\Sigma) \notin \mathcal{B}_{p,k_0} \mid \mathbb{X}_n) > \delta] \\ &\le 4\delta^{-1}\delta_2^{-2}\mathrm{E}_{\Sigma_0}\mathrm{E}^{\pi^i}(||B_{k_0}^{(\epsilon_n)}(\Sigma) - \Sigma_0||_2^2 \mid \mathbb{X}_n) \\ &\le 4\delta^{-1}\delta_2^{-2}(\log k_0)^2\frac{k_0 + \log p}{n},\end{aligned}$$

for a positive constant C. Thus, if $k_0 \vee \log p = o(n)$, then the upper bound goes to zero as $n \longrightarrow \infty$. □

3.2 Interval Estimation

In this subsection, we show that the $(1-\alpha)100\%$ highest posterior density region of the post-processed posterior is asymptotically on the average an $(1-\alpha)100\%$ credible set of the conventional posterior. By the conventional Bayesian method, we mean the Bayesian method imposing a prior distribution on banded covariance matrices directly. Thus, the post-processed posterior provides approximations to the credible regions of the conventional posterior.

For a given integer $0 < k \le p$ and $\Sigma \in \mathcal{C}_p$, let $\theta_1 = \theta_1(\Sigma) = (\sigma_{ij}, |i-j| \le k)$ and $\theta_2 = \theta_2(\Sigma) = (\sigma_{ij}, |i-j| > k)$. Let $\pi^c(\theta_1)$ be a prior for k-banded covariance matrices. We use the bracket notation for the distribution or density of random variables. For examples, the joint distribution of $h(X)$ and $g(Y)$ and conditional distribution of $h(X)$ given $g(Y)$ are denoted by $[h(X), g(Y)]$ and $[h(X)|g(Y)]$, respectively. Probability that $h(X) \in A$ will be denoted by $[h(X) \in A|g(Y)]$ where A is a set. Subscripts to the brackets are used to distinguish different joint distributions of (X, Y).

Define

$$\begin{aligned}
[\theta_1 \mid \mathbb{X}_n]_{PPP,0} &= \int \pi^i(\theta_1, \theta_2 \mid \mathbb{X}_n) d\theta_2, \\
&\propto \int \pi^i(\theta_1, \theta_2) p\{\mathbb{X}_n \mid \Sigma(\theta_1, \theta_2)\} d\theta_2 \\
[\theta_1 \mid \mathbb{X}_n]_C &= \pi^c(\theta_1 \mid \mathbb{X}_n) \\
&\propto \pi^c(\theta_1) p\{\mathbb{X}_n \mid \Sigma(\theta_1, 0)\},
\end{aligned}$$

where $p(\mathbb{X}_n \mid \Sigma)$ is the probability density function of $\mathbb{X}_n$ when X_i's follow $N_p(0, \Sigma)$. In the above, $[\theta_1 \mid \mathbb{X}_n]_{PPP,0}$ and $[\theta_1 \mid \mathbb{X}_n]_C$ denote the post-processed posterior with only the k-band operation B_k and the posterior of the conventional Bayesian method, respectively. Note that, in $[\theta_1 \mid \mathbb{X}_n]_{PPP,0}$, we use subscript 0 to distinguish it from the post-processed posterior defined in (1), which we will denote as $[\theta_1 \mid \mathbb{X}_n]_{PPP}$.

Suppose that the true covariance matrix Σ_0 has the k-banded structure. Let $(\hat{\theta}_1^*, \hat{\theta}_2^*) = \operatorname{argmax}_{\theta_1,\theta_2} \log p\{\mathbb{X}_n \mid \Sigma(\theta_1, \theta_2)\}$ and $\hat{\theta}_1 = \operatorname{argmax}_{\theta_1} \log p\{\mathbb{X}_n \mid \Sigma(\theta_1, 0)\}$ be the maximum likelihood estimators. Furthermore, we denote the Fisher-information matrix by

$$\begin{aligned}
\mathcal{I}(\theta_1, \theta_2) &= -\mathrm{E}_{\Sigma(\theta_1,\theta_2)}\Big\{\Big[\frac{\partial}{\partial\theta}\log p\{\mathbb{X}_n|\Sigma(\theta_1, \theta_2)\}\Big]^T\Big[\frac{\partial}{\partial\theta}\log p\{\mathbb{X}_n|\Sigma(\theta_1, \theta_2)\}\Big]\Big\}, \\
&= \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}
\end{aligned}$$

and $\mathcal{I}_{11\cdot 2}(\theta_1, \theta_2) = \mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21}$.

For the theorem on the credible interval of post-processed posterior, we make assumptions based on the Bernstein-von-Mises theorem and the regularity conditions on the maximum likelihood estimator. We assume that the total variation distance version of Bernstein von-Mises theorem holds for $[\theta_1 \mid \mathbb{X}_n]_{PPP,0}$ and $[\theta_1 \mid \mathbb{X}_n]_C$, i.e.,

A1. (Bernstein-von Mises condition)

$$\lim_{n\longrightarrow\infty} \mathrm{E}_{\Sigma_0}||[n^{1/2}(\theta_1(\Sigma) - \hat{\theta}_1^*) \mid \mathbb{X}_n]_{PPP,0} - N_p(0, \mathcal{I}_{11\cdot 2}^{-1}\{\theta_1(\Sigma_0), 0\})||_{TV} = 0,$$
$$\lim_{n\longrightarrow\infty} \mathrm{E}_{\Sigma_0}||[n^{1/2}(\theta_1(\Sigma) - \hat{\theta}_1) \mid \mathbb{X}_n]_{C} - N_p(0, \mathcal{I}_{11}^{-1}\{\theta_1(\Sigma_0), 0\})||_{TV} = 0.$$

Using a slight abuse of notation, we let $N_p(0, \mathcal{I}^{-1})$ denote the probability measure of the multivariate normal distribution with zero mean vector and covariance matrix $\mathcal{I}^{-1}$. For any probability measures P and Q on a σ-field $\mathcal{M}$, $||P - Q||_{TV}$ is defined by $\sup_{A\in\mathcal{M}} |P(A) - Q(A)|$. The total variation distance version of the Bernstein von-Mises theorem is given in Van der Vaart (2000) and Ghosal and Van der Vaart (2017).

Furthermore, we assume that the following regularity conditions hold. Let $\xrightarrow{d}$ and $\xrightarrow{P}$ denote the convergence in distribution and in probability, respectively.

A2. As $n \longrightarrow \infty$,

$$\begin{aligned} n^{-1/2}L_n'\{\theta_1(\Sigma_0), 0\} &\xrightarrow{d} N_p[0, \mathcal{I}\{\theta_1(\Sigma_0), 0\}], \\ (\hat{\theta}_1^*, \hat{\theta}_2^*) &\xrightarrow{P} (\theta_1(\Sigma_0), 0), \\ \hat{\theta}_1 &\xrightarrow{P} \theta_1(\Sigma_0), \\ \sup_{t:||t-\theta_1(\Sigma_0)||\le\epsilon_n} \frac{1}{n}||L_n''(t, 0) - L_n''\{\theta_1(\Sigma_0), 0\}|| &\xrightarrow{P} 0 \quad \text{as } \epsilon_n \to 0, \\ -n^{-1}L_n''\{\theta_1(\Sigma_0), 0\} &\xrightarrow{P} \mathcal{I}\{\theta_1(\Sigma_0), 0\}, \end{aligned} \tag{2}$$

$\mathcal{I}\{\theta_1(\Sigma_0), 0\}$ is positive-definite, and $L_n''\{\theta_1(\Sigma_0), 0\}$ is continuous, where

$$\begin{aligned} L_n(\theta_1, \theta_2) &= \log p\{\mathbb{X}_n \mid \Sigma(\theta_1, \theta_2)\}, \\ L_n'\{\theta_1(\Sigma_0), 0\} &= \partial L_n\{\theta_1(\Sigma_0), 0\}/\partial(\theta_1, \theta_2), \\ L_n''\{\theta_1(\Sigma_0), 0\} &= \partial^2 L_n\{\theta_1(\Sigma_0), 0\}/\partial(\theta_1, \theta_2)^2. \end{aligned}$$

Theorem 3.4 shows that, under the regularity conditions, the highest posterior density region based on the post-processed posterior is, on average, a credible region of the conventional Bayesian method for banded covariance matrices. The regularity conditions, A1 and A2, are not generally satisfied in the high-dimensional settings, while they hold for a fixed p. However, we would like to emphasize that it is the first result on credible sets of post-processed posteriors (or projected posteriors) up to our knowledge. Note that the idea of transforming posterior samples has been used in the various settings, but they do not provide theoretical results on credible sets constructed from the transformed posterior sample.

Theorem 3.4. *Suppose A1 and A2 hold. If $C_{1-\alpha,n}$ is the highest posterior density regions of $[\theta_1 \mid \mathbb{X}_n]_{PPP}$ and p is fixed, then*

$$\lim_{n\longrightarrow\infty} \mathrm{E}_{\Sigma_0}\{[\theta_1(\Sigma) \in C_{1-\alpha,n} \mid \mathbb{X}_n]_C\} = 1 - \alpha.$$

4 Numerical Studies

4.1 Choice of Post-Processing Parameters

The post-processed posterior procedure requires the banding parameter k and the positive-definiteness adjustment parameter ϵ_n in the post-processing step (1). We suggest using the Bayesian leave-one-out cross-validation (LOOCV) method (Gelman et al., 2014) to choose these parameters. We define the log-predictive density given k and ϵ_n as

$$R(k, \epsilon_n) = \sum_{i=1}^{n} \log \int p\{X_i \mid B_k^{(\epsilon_n)}(\Sigma)\} \pi^i(\Sigma \mid \mathbb{X}_{n,-i}) d\Sigma,$$

where $\mathbb{X}_{n,-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $p\{\cdot \mid B_k^{(\epsilon_n)}(\Sigma)\}$ is the multivariate normal density with zero mean and the covariance matrix $B_k^{(\epsilon_n)}(\Sigma)$. Then, using Monte Carlo method, we obtain the estimated log-predictive density as

$$\hat{R}(k, \epsilon_n) = \sum_{i=1}^{n} \log \frac{1}{S} \sum_{s=1}^{S} p\{X_i \mid B_k^{(\epsilon_n)}(\Sigma_{i,s})\}, \tag{3}$$

where $\Sigma_{i,s}$ is the sth sample from $\pi^i(\cdot \mid \mathbb{X}_{n,-i})$, and S is the number of the posterior samples. We choose the parameters as $(\hat{k}, \hat{\epsilon}_n) = \text{argmin}_{k \in [p], \epsilon_n > 0} \hat{R}(k, \epsilon_n)$. We optimize k and ϵ_n using the grid search method. In this simulation study, we consider the set of ϵ_n as $\{0.5, 0.3, 0.1, 0.05, 0.01, 0.005, 0.001\}$. When calculating $\hat{R}(k, \epsilon_n)$, we set $S = 500$, i.e., 500 initial posterior samples are used for the Monte Carlo integration.

We apply the Bayesian LOOCV method to simulation data. We consider four banded covariances $\Sigma_0^{(1)}$, $\Sigma_0^{(1')}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$ as the true covariance matrices. Let $\Sigma_{(\rho,\alpha)}^{(1)*} = (\sigma_{(\rho,\alpha),ij}^{(1)})_{p \times p}$, where

$$\sigma_{(\rho,\alpha),ij}^{(1)} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho |i-j|^{-(\alpha+1)}, & 1 \le i \ne j \le p. \end{cases}$$

Then we define $\Sigma_0^{(1)} = B_{k_0}(\Sigma_{(0.6,1)}^{(1)*}) + [0.5 - \{\lambda_{\min}(B_{k_0}(\Sigma_{(0.6,1)}^{(1)*}))\}] I_p$ and $\Sigma_0^{(1')} = B_{k_0}(\Sigma_{(5,0.01)}^{(1)*}) + [0.5 - \{\lambda_{\min}(B_{k_0}(\Sigma_{(5,0.01)}^{(1)*}))\}] I_p$, where k_0 is the bandwidth. They are defined by banding $\Sigma_{(\rho,\alpha)}^{(1)*}$ and adding an identity matrix multiplied by a positive number to make the minimum eigenvalue of the resulting matrix be 0.5. Let $\Sigma_0^{(2)*} = (\sigma_{0,ij}^{(2)})_{p \times p}$, where $\sigma_{0,ij}^{(2)} = \{1 - |i-j|/(k_0+1)\} \wedge 0$ for any $1 \le i, j \le p$. Then we set $\Sigma_0^{(2)} = \Sigma_0^{(2)*} + [0.5 - \{\lambda_{\min}(\Sigma_0^{(2)*})\}] I_p$. Let $\Sigma_0^{(3)*} = L_0 D_0 L_0^T$ and $\Sigma_0^{(3)} = \Sigma_0^{(3)*} + [0.5 - \{\lambda_{\min}(\Sigma_0^{(3)*})\}] I_p$, where

$$L_{ij}^0 = \begin{cases} 1, & 1 \le i = j \le p \\ l_{ij}, & 0 < i - j \le k_0 \\ 0, & \text{otherwise}, \end{cases}$$

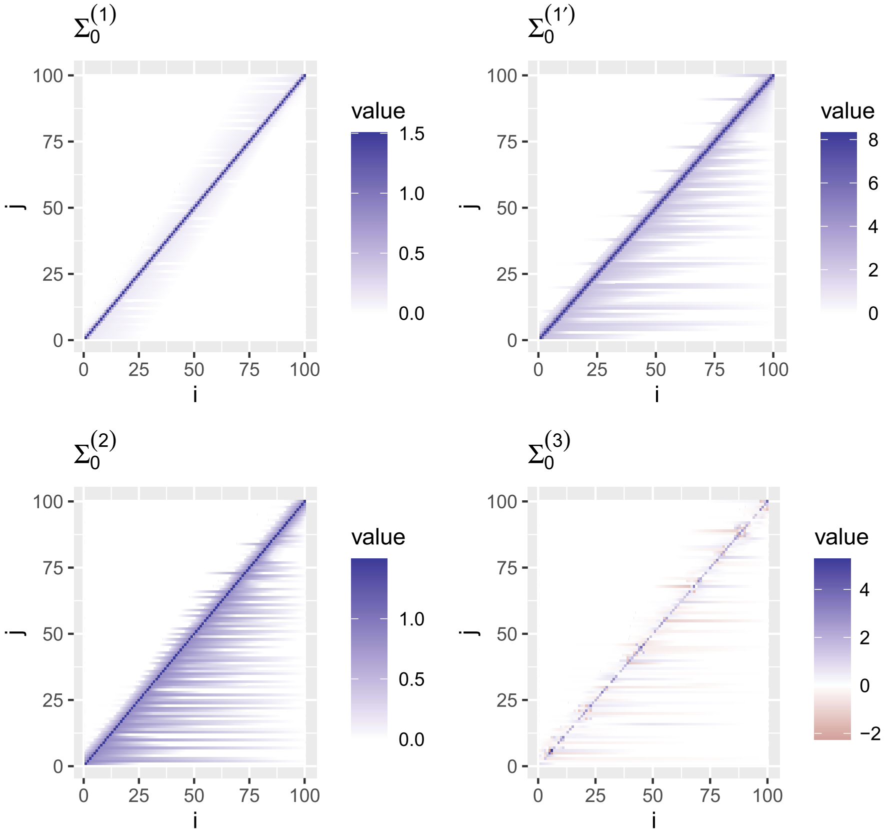

Figure 1: Visualization of true banded covariances.

l_{ij} are independent sample from $N(0, 1)$, and $D_0 = diag(d_{ii})$ is a diagonal matrix where d_{ii} is independent sample from $IG(5, 1)$, the inverse-gamma distribution with the shape parameter 5 and the scale parameter 1. The true covariance matrices with $p = 100$ and $k_0 = 5$ are plotted in Figure 1.

For each banded covariance with $k_0 = 5$, we generate the data $X_1, \ldots, X_n$ from $N_p(0, \Sigma_0^{(t)})$ independently, where $t = 1, 2, 3, 4$, $n = 25, 50, 100$ and $p = 100$. For the initial prior of the post-processed posterior, we choose $IW_p(A_0, \nu_0)$ with $\nu_0 = 2p+k+1$ and $A_0 = I_p$.

We examine estimated $\hat{k}$ for 100 repetitions of the simulated data and present the results in Table 1. We also investigate the estimation error of the post-processed posterior with $\hat{k}$. Let $\hat{\Sigma}_{k_0}$ and $\hat{\Sigma}_{\hat{k}}$ be the posterior means of the post-processed posterior with bandwidth k_0 and $\hat{k}$, respectively. The average of the error ratio, $||\hat{\Sigma}_{\hat{k}} - \Sigma_0|| / ||\hat{\Sigma}_{k_0} - \Sigma_0||$, are reported in Table 2. For $\Sigma_0^{(1')}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$, the error ratios are close to 1, while the estimated bandwidths $\hat{k}$ are concentrated around k_0. On the other hand, when the

		$\hat{k}$							
		1	2	3	4	5 ($= k_0$)	6	7	8
$\Sigma_0^{(1)}$	$n = 25$	99	1	0	0	0	0	0	0
	$n = 50$	100	0	0	0	0	0	0	0
	$n = 100$	0	100	0	0	0	0	0	0
$\Sigma_0^{(1')}$	$n = 25$	3	23	31	3	39	1	0	0
	$n = 50$	0	0	6	0	88	6	0	0
	$n = 100$	0	0	0	0	95	5	0	0
$\Sigma_0^{(2)}$	$n = 25$	0	0	0	48	38	14	0	0
	$n = 50$	0	0	0	14	75	11	0	0
	$n = 100$	0	0	0	11	80	9	0	0
$\Sigma_0^{(3)}$	$n = 25$	0	0	5	27	61	7	0	0
	$n = 50$	0	0	0	14	84	2	0	0
	$n = 100$	0	0	0	1	89	10	0	0

Table 1: Number of times bandwidth $\hat{k}$ are chosen out of 100 repetitions. The first and second columns represent the true covariance and the sample sizes, respectively. For example, when the true covariance is $\Sigma_0^{(1)}$ and $n = 25$, $\hat{k} = 1$ is chosen for 99 data sets and $\hat{k} = 2$ for one data set.

	$n = 25$	$n = 50$	$n = 100$
$\Sigma_0^{(1)}$	0.821	0.797	0.825
$\Sigma_0^{(1')}$	1.016	1.003	1.003
$\Sigma_0^{(2)}$	1.009	1.009	1.005
$\Sigma_0^{(3)}$	1.009	1.003	1.005

Table 2: The average of error ratios $||\hat{\Sigma}_{\hat{k}} - \Sigma_0||/||\hat{\Sigma}_{k_0} - \Sigma_0||$ for 100 repetitions of the simulation data. The first column and first row represent the true covariances and the sample sizes, respectively.

true covariance matrix is $\Sigma_0^{(1)}$, $\hat{k}$ tends to underestimate k_0 and $\hat{\Sigma}_{\hat{k}}$ gives similar or smaller estimation errors than $\hat{\Sigma}_{k_0}$. See Table 2. A referee pointed out that this may be because $\Sigma_0^{(1)}$ is too diagonally dominant. We also agree with the referee. We have examined values σ_{ij}/σ_{ii}, where σ_{ij} is the (i,j) element of $\Sigma_0^{(1)}$, as

$$(\frac{\sigma_{i,i+1}}{\sigma_{ii}}, \frac{\sigma_{i,i+2}}{\sigma_{ii}}, \frac{\sigma_{i,i+3}}{\sigma_{ii}}, \frac{\sigma_{i,i+4}}{\sigma_{ii}}, \frac{\sigma_{i,i+5}}{\sigma_{ii}}) = (0.396, 0.100, 0.044, 0.025, 0.016).$$

The off-diagonal elements are small compared to the diagonal element. Especially, when the index difference $|i - j|$ is larger than 2, σ_{ij}/σ_{ii} is smaller than 0.05. Thus, using smaller bandwidth can increase the accuracy.

For the choice of the bandwidth, we also regard the adaption of the post-processed posterior given a prior on k. The remark below presents the posterior distribution $\pi(k \mid \mathbb{X}_n)$ given a prior on k and the simulation data analysis.

Remark. *The proposed post-processed posterior framework can be adapted to the case where a prior on k is specified. We split n observations of $\mathbb{X}_n$ into $\mathbb{X}^{(1)}$ and $\mathbb{X}^{(2)}$, which consist of n_1 and n_2 observations, respectively, where $n_1 + n_2 = n$. Then, the posterior distribution $\pi(k \mid \mathbb{X}_n)$ is derived as*

$$\pi(k \mid \mathbb{X}_n) \propto \pi(k)\Big\{\frac{c_2}{c}\int \frac{p(\mathbb{X}^{(1)} \mid \Sigma)}{p(\mathbb{X}_n \mid B_k^{(\epsilon_n)}(\Sigma))}\pi^i(\Sigma \mid \mathbb{X}^{(2)})d\Sigma\Big\}^{-1}, \tag{4}$$

where $c = \int p(\mathbb{X}_n \mid \Sigma)\pi^i(\Sigma)d\Sigma$, $c_2 = \int p(\mathbb{X}^{(2)} \mid \Sigma)\pi^i(\Sigma)d\Sigma$, and $\pi^i(\Sigma \mid \mathbb{X}^{(2)})$ is the density function of the initial posterior distribution given data $\mathbb{X}^{(2)}$. Note that c and c_2 are obtained analytically. We can estimate $\pi(k \mid \mathbb{X}_n)$ by applying the Monte Carlo integration, which is denoted by $\hat{\pi}(k \mid \mathbb{X}_n)$. For the derivation of (4) and estimation of $\hat{\pi}(k \mid \mathbb{X}_n)$, see the supplementary material.

We examine $\hat{\pi}(k \mid \mathbb{X}_n)$ using the simulation data. For the simulation data analysis, we set $n_1 = n/2$, $\epsilon_n = (\log k)(k + \log p)^{1/2}/n^{1/2}$, a prespecified value of ϵ_n based on Theorem 3.1, and use the uniform prior $\pi(k) \propto 1$, for $k = 1, \ldots, 10$. We have tested values of n_1 including 0, n and $n/2$ and empirically found that $n_1 = n/2$ works fine for most cases. Figure 2 shows the average of the estimated marginal posterior $\hat{\pi}(k \mid \mathbb{X}_n)$ based on 100 repetitions. For the true covariances $\Sigma_0^{(1')}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$, when $n = 100$, the estimated posteriors are concentrated on the true bandwidth $k_0 = 5$. However, when the true covariance is $\Sigma_0^{(1)}$, the estimated posteriors tend to spread over values smaller than the true bandwidth $k_0 = 5$. We believe that this is because $\Sigma_0^{(1)}$ is too diagonally dominant as discussed after Tables 1 and 2.

4.2 Comparative Study

We compare the post-processed posterior with other methods: the banded sample covariance (Bickel and Levina, 2008) and covariance graphical model methods. Since the k-banded covariance structure corresponds to a graph model, covariance graphical model methods can be used for the banded covariance estimation. As frequentist methods of the covariance graphical model, we consider dual maximum likelihood estimator (Kauermann, 1996), and the maximum likelihood estimator by iterative conditional fitting (Chaudhuri et al., 2007). As Bayesian methods, we consider G-inverse Wishart distribution (Silva and Ghahramani, 2009) and Wishart distributions for covariance graph (Khare and Rajaratnam, 2011). Additionally, we conduct the *dual post-processed posterior*, which is a post-processing posterior based on the dual algorithm (Kauermann, 1996) instead of the banding post-processing function $B_k^{(\epsilon_n)}$. We obtain a posterior sample of the dual post-processed posterior as follows:

Step 1. (Initial posterior computing step) For $l = 1, 2, \ldots$, sample $\Sigma^{(l)}$ from the initial posterior,

$$\Sigma^{(l)} \mid \mathbb{X}_n \sim IW_p(B_0 + nS_n, \nu_0 + n).$$

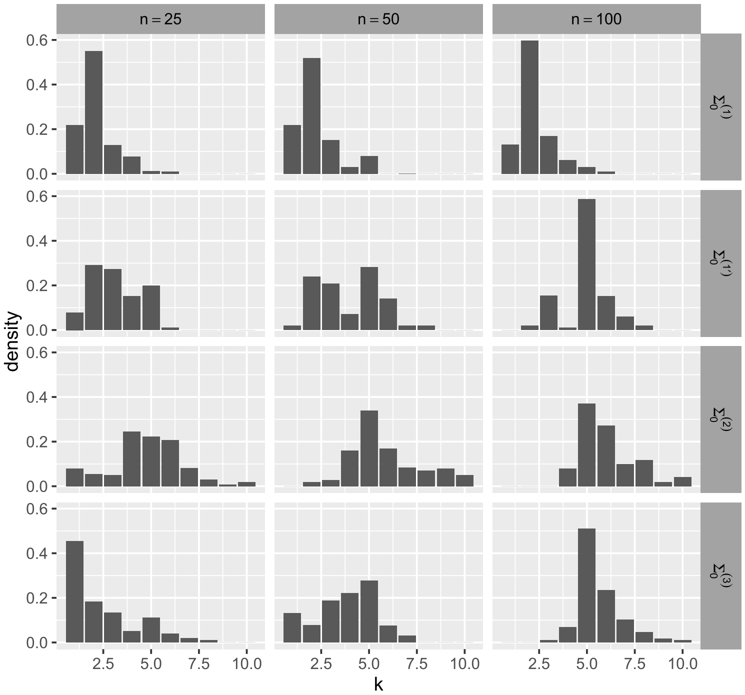

Figure 2: The average of the estimated marginal posterior probabilities of the bandwidth, $\pi(k \mid \mathbb{X}_n)$, for various simulation settings.

Step 2. (Post-processing step) Obtain $\Sigma_{(l)}^D$ as the solution of the simultaneous equations:

$$\{(\Sigma_{(l)}^D)^{-1}\}_{ij} = \{(\Sigma^{(l)})^{-1}\}_{ij},$$

for $|i-j| \le k$ and $(\Sigma_{(l)}^D)_{ij} = 0$ for $|i-j| > k$.

In this comparative study, we fix the bandwidth k as the true bandwidth k_0 since the covariance graphical model methods do not consider the choice of the graph structure. For the Wishart distribution for covariance graph (Khare and Rajaratnam, 2011), we used $\alpha_i = 2k_0 + 5$ and $U = I_p$ as they suggested. Similarly, we set $\delta = 5$ and $U = I_p$ for the G-inverse Wishart distribution (Silva and Ghahramani, 2009) as they suggested. For both methods, the initial values of the Σ for the Markov chain Monte Carlo algorithms were set at the identity matrix and 500 posterior samples were drawn with 500 burn-in sample.

For the dual maximum likelihood estimator and the maximum likelihood estimator, $S_n + \epsilon_n I_p$ is used in place of the sample covariance matrix because these algorithms need a positive definite sample covariance matrix. The adjustment parameter ϵ_n is chosen as the minimizer of $\hat{R}_f(\epsilon_n)$, which is defined as

$$\hat{R}_f(\epsilon_n) = \sum_{i=1}^{n} \log p\{X_i \mid h(\mathbb{X}_{n,-i};\epsilon_n)\}, \tag{5}$$

where $h(\mathbb{X}_n;\epsilon_n)$ is a frequentist estimator of Σ based on $\mathbb{X}_n$ and an adjustment parameter ϵ_n.

We compare the methods using the simulation data in Section 4.1. For the comparison, we consider two respects: the spectral norm error and the coverage probability for a functional of covariance when interval estimation is available.

Comparison of Spectral Norm Error

For each simulation setting, 100 sets of samples were generated. The performance of each estimator is measured by the mean spectral norm error

$$\frac{1}{100}\sum_{s=1}^{100} ||\Sigma_0 - \hat{\Sigma}^{(s)}||,$$

where $\hat{\Sigma}^{(s)}$ is a point estimate based on the sth simulated data set. For Bayesian methods, we use the posterior mean as the point estimator. Table 3 shows the mean spectral norm error of each method.

The post-processed posterior performs reasonably well in all cases. While the performance of the maximum likelihood estimator and the banded sample covariance is similar to that of the post-processed posterior, when $n = 100$, the maximum likelihood estimator and the banded sample covariance have larger mean spectral norm errors when $n = 25$.

	$n = 25$			$n = 50$			$n = 100$		
	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$
Post-processed posterior	2.96	3.63	4.32	1.89	2.64	3.13	1.35	1.89	2.13
G-inverse Wishart	2.98	5.79	6.80	2.72	5.23	6.16	2.33	4.39	5.10
Wishart for covariance graph	3.90	6.86	5.88	1.73	4.36	4.78	1.47	2.94	5.06
Dual post-processed posterior	3.24	6.46	7.71	3.23	6.42	7.68	3.06	6.02	7.19
Banded sample covariance	3.42	4.44	5.17	2.06	2.89	3.40	1.44	1.97	2.23
Dual MLE	3.17	6.33	7.59	3.08	6.09	7.38	2.64	4.96	6.04
MLE	4.59	4.09	5.70	2.24	2.51	3.27	1.48	1.67	2.12

Table 3: Spectral norm-based errors for $\Sigma_0^{(1)}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$. MLE means the maximum likelihood estimator.

Comparison of Coverage Probability

We investigate the performance of interval estimation for a functional of covariances in this section. There is no valid frequentist interval estimator for functionals of banded or bandable covariances in the high-dimensional covariance. However, if one assumes p is fixed, the interval estimator for functionals of banded covariances can be derived from the Fisher information matrix given in Chaudhuri et al. (2007). Define $\mathrm{vecb}(\Sigma) := \mathrm{vecb}(\Sigma; k) = \mathrm{vec}(\{\sigma_{ij} : i \le j, |i-j| \le k\})$ and $Q \in \mathbb{R}^{p^2 \times p^*}$ such that $\mathrm{vec}(\Sigma) = Q \times \mathrm{vecb}(\Sigma; k)$, where vec is the column-wise vectorization operation, and p^* is the dimension of $\mathrm{vecb}(\Sigma; k)$. By asymptotic normality of maximum likelihood estimators and the Fisher information matrix in Chaudhuri et al. (2007), we obtain

$$n^{1/2}\{\mathrm{vecb}(\Sigma^{MLE}) - \mathrm{vecb}(\Sigma_0)\} \xrightarrow{d} N_{p^*}[0, 2\{Q^T(\Sigma_0^{-1} \otimes \Sigma_0^{-1})Q\}^{-1}],$$

as $n \longrightarrow \infty$, where Σ^{MLE} is obtained by the iterative conditional fitting (Chaudhuri et al., 2007). Let $\phi\{\mathrm{vecb}(\Sigma)\}$ and $\nabla\phi\{\mathrm{vecb}(\Sigma)\}$ be a functional and its derivative, respectively. By the delta method, we obtain

$$n^{1/2}[\phi\{\mathrm{vecb}(\Sigma^{MLE})\} - \phi\{\mathrm{vecb}(\Sigma_0)\}] \xrightarrow{d} N(0, \sigma^2_{0,\phi}),$$

as $n \longrightarrow \infty$, where $\sigma^2_{0,\phi} = 2\nabla\phi\{\mathrm{vecb}(\Sigma_0)\}\{Q^T(\Sigma_0^{-1} \otimes \Sigma_0^{-1})Q\}^{-1}\nabla^T\phi\{\mathrm{vecb}(\Sigma_0)\}$. Then, we induce an $(1-\alpha)100\%$ confidence interval of the functional as

$$\phi(\mathrm{vecb}(\Sigma^{MLE})) \pm z_{\alpha/2}\frac{\sigma_{0,\phi}}{n^{1/2}}.$$

Since $\sigma_{0,\phi}$ depends on the true covariance matrix, we use an estimated value as

$$\hat{\sigma}^2_\phi = 2\nabla\phi\{\mathrm{vecb}(\Sigma^{MLE})\}\{Q^T((\Sigma^{MLE})^{-1} \otimes (\Sigma^{MLE})^{-1})Q\}^{-1}\nabla^T\phi\{\mathrm{vecb}(\Sigma^{MLE})\}.$$

For Bayesian methods, we obtain credible intervals using the posterior samples. For posterior sample $\Sigma_1, \ldots, \Sigma_S$, the $(1-\alpha)100\%$ credible interval for a functional $\phi(\Sigma)$ can be obtained based on $\phi(\Sigma_1), \ldots, \phi(\Sigma_S)$. We set $S = 500$ in the simulation.

In the numerical experiment, we focus on the conditional mean for the prediction problem as a functional of covariances. When $X_i = (X_{i,1}, \ldots, X_{i,p})^T \sim N_p(0, \Sigma)$, the conditional mean given $X_{-p} = (X_1, \ldots, X_{p-1})^T$ is

$$\mathrm{cm}(\Sigma; X_{-p}) := \mathrm{E}(X_p \mid X_{-p}) = \Sigma_{p,-p}\Sigma^{-1}_{-p,-p}X_{-p}.$$

We compare the coverage probabilities and the lengths of intervals for 95% credible intervals of $\mathrm{cm}(\Sigma; X_{-p})$ in Table 4.

The post-processed posterior performs well overall. It appears that the post-processed posterior and the Wishart for covariance graph produce practically reasonable interval estimates, but the coverage probabilities of Wishart for covariance graph tend to be smaller than the nominal coverage. The G-inverse Wishart and the dual post-processed posterior have much smaller coverage probabilities than the nominal probability. The maximum likelihood estimator tends to produce wide (thus conservative) confidence intervals when $n = 25$, which makes it less meaningful in practice.

Additionally, we compare computation times of the Bayesian methods in Table 5. The post-processed posterior is faster than G-inverse Wishart distribution and Wishart distribution for covariance graph methods. The dual post-processed posterior method is the fastest because it does not have the cross-validation step for the adjustment parameter ϵ_n, but its mean spectral norm errors in Table 3 sometimes shows poor performance.

		$n = 25$	
	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$
Post-processed posterior	94.7% (2.55)	92.9% (2.36)	93.9% (3.90)
G-inverse Wishart	50.4% (1.09)	48.0% (0.93)	45.3% (1.63)
Wishart for covariance graph	99.3% (2.86)	99.9% (3.24)	98.2% (3.78)
Dual post-processed posterior	70.9% (0.89)	61.5% (0.82)	43.3% (1.14)
Maximum likelihood estimator	99.5% (3.77)	99.8% (5.71)	99.9% (15.83)
		$n = 50$	
	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$
Post-processed posterior	97.0% (2.15)	97.0% (1.84)	96.4% (3.27)
G-inverse Wishart	62.3% (0.72)	63.4% (0.66)	60.1% (1.03)
Wishart for covariance graph	96.0% (1.51)	98.9% (1.72)	91.2% (1.89)
Dual post-processed posterior	73.5% (0.81)	73.3% (0.74)	54.9% (1.09)
Maximum likelihood estimator	97.9% (1.69)	99.2% (2.13)	99.6% (4.28)
		$n = 100$	
	$\Sigma_0^{(1)}$	$\Sigma_0^{(2)}$	$\Sigma_0^{(3)}$
Post-processed posterior	94.9% (1.26)	96.5% (1.52)	97.1% (2.77)
G-inverse Wishart	74.6% (0.59)	71.3% (0.54)	70.7% (0.84)
Wishart for covariance graph	94.2% (0.94)	97.0% (1.06)	85.3% (1.16)
Dual post-processed posterior	48.8% (0.78)	51.7% (0.70)	48.5% (1.09)
Maximum likelihood estimator	97.7% (1.08)	98.4% (1.23)	99.6% (2.65)

Table 4: Coverage probabilities and lengths of interval estimates of the conditional mean for banded covariances $\Sigma_0^{(1)}$, $\Sigma_0^{(2)}$ and $\Sigma_0^{(3)}$. The average lengths of intervals are represented in parentheses.

	1-quantile	mean	median	3-quantile
Post-processed posterior	40.45	40.63	40.63	40.78
G-inverse Wishart	205.47	206.67	207.32	208.23
Wishart for covariance graph	353.91	355.14	356.31	357.08
Dual post-processed posterior	10.60	10.73	10.67	10.78

Table 5: The summary statistics of computing times (unit: sec) for Bayesian methods, when $p = 100$ and $n = 50$. In the computing times of the post-processed posterior method, the step of Bayesian leave-one-out cross-validation for ϵ_n is involved.

4.3 Application to Call Center Data

We apply the post-processed posterior to analyze the call center data set, which is used in Huang et al. (2006) and Bickel and Levina (2008). The data set consists of the number of phone calls for 239 days, and the numbers of calls are recorded for 17 hours from 7:00 and divided into 10-minute intervals. We denote the number of calls in the jth time index of the ith day as N_{ij} $(i = 1, \ldots, 239; j = 1, \ldots, 102)$, and define $x_{i,j} = (N_{ij} + 1/4)^{1/2}$ so that its distribution is similar to the normal distribution. Furthermore, to focus on covariance estimation, we center the data.

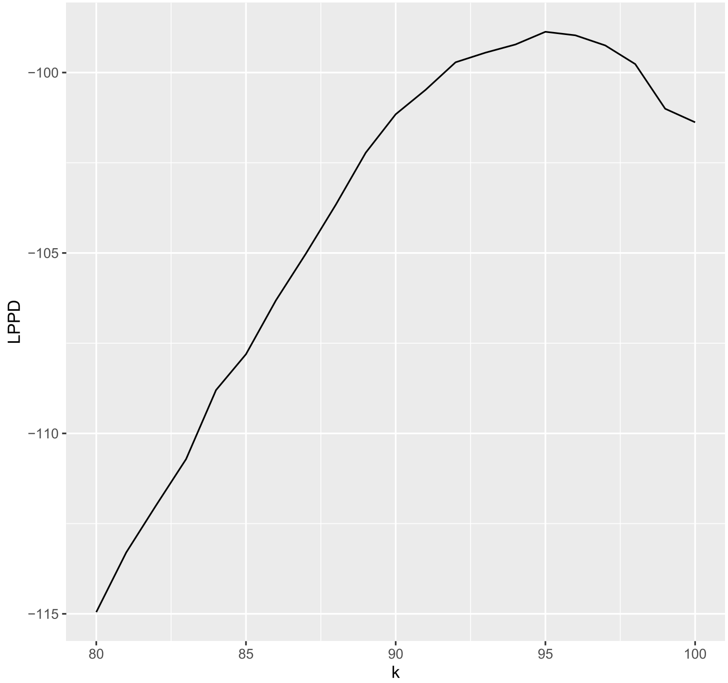

Figure 3: The estimated log posterior predictive densities via the Bayesian leave-one-out cross-validation. The estimated log posterior predictive densities are calculated for the bandwidth k from 80 to 100.

Using the covariance estimators by the centered data, we predict the numbers of calls at $j = 71, \dots, 102$ time points given those at the other time points. Let $x_i^{(1)} = (x_{i,1}, \dots, x_{i,70})^T$, $x_i^{(2)} = (x_{i,71}, \dots, x_{i,102})^T$, then we obtain estimated conditional mean of $x_i^{(2)}$ given $x_i^{(1)}$ as

$$x_i^{(2)}(\Sigma, x_i^{(1)}) = \Sigma_{21}\Sigma_{11}^{-1}x_i^{(1)},$$

where $\Sigma_{ab} = E\{x_i^{(a)}(x_i^{(b)})^T\}$ for any $a, b \in \{1, 2\}$. The first 205 days ($i = 1, \dots, 205$) were used as a training data to estimate Σ, and the last 34 days ($i = 206, \dots, 239$) were used as a test data. We measure the accuracy of the methods based on the mean square error,

$$(34)^{-1} \sum_{i=206}^{239} ||x_i^{(2)} - \hat{x}_i^{(2)}||^2, \tag{6}$$

where $\hat{x}_i^{(2)} \equiv x_i^{(2)}(\hat{\Sigma}, x_i^{(1)})$ is an estimator for $x_i^{(2)}$. Here, $\hat{\Sigma}$ is an estimator of Σ, where posterior means are used based on 500 posterior samples for Bayesian methods.

We choose bandwidth k and the positive-definiteness adjustment parameter ϵ_n using the Bayesian leave-one-out cross validation method given in Section 4.1. Let $\hat{\epsilon}(k) = \operatorname{argmin}_{\epsilon_n > 0} \hat{R}(k, \epsilon_n)$. We examine $\hat{R}(k, \hat{\epsilon}(k))$ in Figure 3. Based on Figure 3, we choose

Method	error	LPPD
Post-processed posterior	0.88	−84.60
Inverse-Wishart posterior	1.19	−88.28
Dual post-processed posterior	1.18	−91.65
Banded sample covariance	0.95	–
Dual maximum likelihood estimator	0.96	–
Sample covariance	0.99	–

Table 6: Mean square error between observations and estimated conditional mean. The error column represents the mean square error of prediction values, and the LPPD column represents the log posterior predictive density.

$\hat{k} = 95$. For the frequentist methods, we select the bandwidth based on the leave-one-out cross-validation similar to (5).

We give the prediction error (6) in Table 6. Additionally, we compare the Bayesian methods using the log posterior predictive density (LPPD), which is defined as

$$\log \sum_{i=1}^{S} p(\mathbb{X}^{test} \mid \Sigma_i),$$

where Σ_i is the ith posterior sample, S is the number of posterior sample and $\mathbb{X}^{test}$ is the test data.

The post-processed posterior outperforms the other methods in prediction error and the log posterior predictive density. By the definition of $x_i^{(2)}(\Sigma, x_i^{(1)})$, Bayesian methods naturally induce interval estimators based on posterior samples of Σ. We visualize the estimators as well as 95% credible intervals from the post-processed posterior for the 1st subject in the test data in Figure 4.

5 Discussion

In this paper, we have proposed a non-traditional Bayesian procedure called the post-processed posterior. It is conceptually straightforward and computationally fast. It attains a nearly minimax convergence rate over all possible pairs of post-processing functions and initial priors, including conventional Bayesian posteriors. Also, its highest density credible sets are asymptotically credible sets of the conventional posteriors on average, and thus its credible sets can be viewed as approximations to the credible sets of the conventional posteriors.

The post-processed posterior can be used for the Bayesian inference on the bandable covariance. The class of bandable covariance matrices is defined as

$$\Big\{\Sigma = (\sigma_{ij}) \in \mathcal{C}_p : \sum_{(i,j):|i-j|\geq k} |\sigma_{ij}| \leq Mk^{-\alpha}, \forall k \geq 1, \lambda_{\max}(\Sigma) \leq M_0, \lambda_{\min}(\Sigma) \geq M_1\Big\},$$

where $\alpha, M > 0$ and $0 < M_1 < M_0$. We have shown that the post-processed posterior is also nearly optimal in the minimax sense for the class of bandable covariance matrices

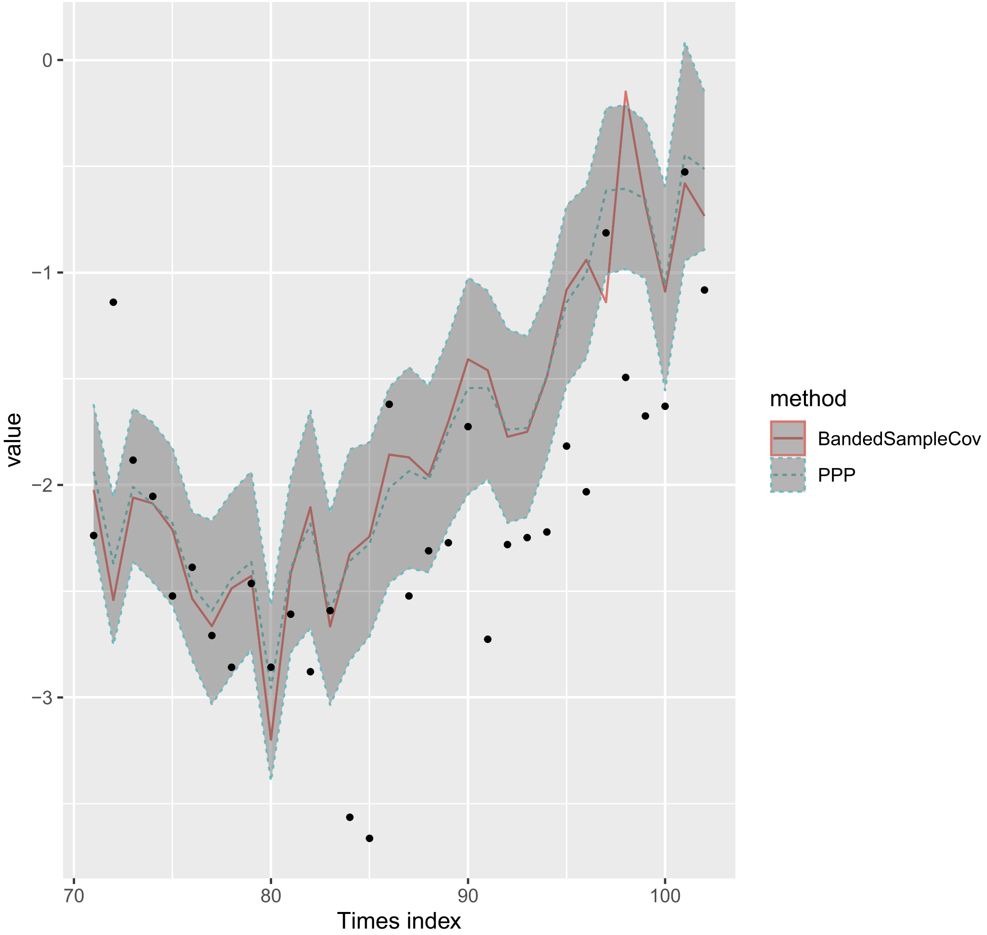

Figure 4: The estimated conditional mean from the 71st to 102nd time indexes of the 1st subject in the test data. The red line and green dashed-line represents the estimated conditional mean of the banded sample covariance and the post-processed posterior, respectively. For the post-processed posterior distribution, 95% credible intervals of the conditional mean are also represented as shade. The black dots represents the observations from the 71st to 102nd time indexes of the 1st subject in the test data.

(see Section S1 in the supplementary material). It is somewhat surprising that the banding post-processed posterior is nearly optimal minimax rate because it was believed that the banding estimator gives the sub-optimal convergence rate (Bickel and Levina, 2008) for bandable covariance matrices.

The idea of post-processing can be used in other covariance structures. For example, the method can be applied to the class of sparse covariance matrices. We are investigating the theoretical properties of the approach. We also believe the post-processing idea can be applied to other problems like sparse linear regression models and high-dimensional nonparametric regression models. The open question is to set the boundary of the post-processing posterior idea: when it has solid theoretical support.

Supplementary Material

Supplementary Material for “Post-Processed Posteriors for Banded Covariances” (DOI: 10.1214/22-BA1333SUPP; .pdf). We show the result of minimax convergence rate for bandable covariance case and more simulation results. We also give an example of application to Linear Discriminant Analysis. The proofs of lemmas and theorems in the main paper are represented. We provide the R code of the post-processed posterior procedure in the Github repository `https://github.com/KwangminLee564/bandPPP`.

References

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`. 1018, 1020

Bashir, A., Carvalho, C. M., Hahn, P. R., and Jones, M. B. (2018). “Post-processing posteriors over precision matrices to produce sparse graph estimates.” *Bayesian Analysis*. MR4044846. doi: `https://doi.org/10.1214/18-BA1139`. 1019

Bickel, P. J. and Levina, E. (2008). “Regularized estimation of large covariance matrices.” *The Annals of Statistics*, 199–227. MR2387969. doi: `https://doi.org/10.1214/009053607000000758`. 1018, 1030, 1034, 1037

Cai, T. and Liu, W. (2011). “Adaptive thresholding for sparse covariance matrix estimation.” *Journal of the American Statistical Association*, 106(494): 672–684. MR2847949. doi: `https://doi.org/10.1198/jasa.2011.tm10560`. 1018

Cai, T., Liu, W., and Luo, X. (2011). “A constrained ℓ_1 minimization approach to sparse precision matrix estimation.” *Journal of the American Statistical Association*, 106(494): 594–607. MR2847973. doi: `https://doi.org/10.1198/jasa.2011.tm10155`. 1020

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`. 1018

Cai, T. T., Ma, Z., and Wu, Y. (2013). “Sparse PCA: Optimal rates and adaptive estimation.” *The Annals of Statistics*, 41(6): 3074–3110. MR3161458. doi: `https://doi.org/10.1214/13-AOS1178`. 1018

Cai, T. T. and Zhou, H. H. (2010). “Optimal rates of convergence for covariance matrix estimation.” *The Annals of Statistics*, 38(4): 2118–2144. MR2676885. doi: `https://doi.org/10.1214/09-AOS752`. 1018

Cai, T. T. and Zhou, H. H. (2012a). "Minimax estimation of large covariance matrices under ℓ_1-norm." *Statistica Sinica*, 1319–1349. MR3027085. 1018

Cai, T. T. and Zhou, H. H. (2012b). "Optimal rates of convergence for sparse covariance matrix estimation." *The Annals of Statistics*, 40(5): 2389–2420. 1020

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). "Estimation of a covariance matrix with zeros." *Biometrika*, 94(1): 199–216. MR2307904. doi: `https://doi.org/10.1093/biomet/asm007`. 1018, 1030, 1033

Dunson, D. B. and Neelon, B. (2003). "Bayesian inference on order-constrained parameters in generalized linear models." *Biometrics*, 59(2): 286–295. MR1987395. doi: `https://doi.org/10.1111/1541-0420.00035`. 1019

Gao, C. and Zhou, H. H. (2015). "Rate-optimal posterior contraction for sparse PCA." *The Annals of Statistics*, 43(2): 785–818. MR3325710. doi: `https://doi.org/10.1214/14-AOS1268`. 1018

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`. 1027

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press. MR3587782. doi: `https://doi.org/10.1017/9781139029834`. 1019, 1026

Gunn, L. H. and Dunson, D. B. (2005). "A transformation approach for incorporating monotone or unimodal constraints." *Biostatistics*, 6(3): 434–449. 1019

Huang, A. and Wand, M. P. (2013). "Simple marginally noninformative prior distributions for covariance matrices." *Bayesian Analysis*, 8(2): 439–452. MR3066948. doi: `https://doi.org/10.1214/13-BA815`. 1018

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`. 1034

Kauermann, G. (1996). "On a dualization of graphical Gaussian models." *Scandinavian Journal of Statistics*, 105–116. MR1380485. 1018, 1030

Khare, K. and Rajaratnam, B. (2011). "Wishart distributions for decomposable covariance graph models." *The Annals of Statistics*, 39(1): 514–555. MR2797855. doi: `https://doi.org/10.1214/10-AOS841`. 1019, 1030, 1031

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`. 1018, 1022

Lee, K. and Lee, J. (2021). "Estimating large precision matrices via modified Cholesky decomposition." *Statistica Sinica*, 31(2021): 173–196. MR4270383. doi: `https://doi.org/10.5705/ss.20`. 1020

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. 1018

Lee, K., Lee, K., and Lee, J. (2022). "Supplementary Material for "Post-processed posteriors for banded covariances"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1333SUPP. 1020

Lin, L. and Dunson, D. B. (2014). "Bayesian monotone regression using Gaussian process projection." *Biometrika*, 101(2): 303–317. MR3215349. doi: https://doi.org/10.1093/biomet/ast063. 1019

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices." *The Annals of Statistics*, 42(3): 1102–1130. MR3210997. doi: https://doi.org/10.1214/14-AOS1215. 1018

Patra, S., Sen, D., and Dunson, D. (2018). "Constrained Bayesian inference through posterior projections." *arXiv e-prints*, arXiv:1812.05741. MR4035485. 1019

Silva, R. and Ghahramani, Z. (2009). "The hidden life of latent variables: Bayesian learning with mixed graph models." *Journal of Machine Learning Research*, 10(Jun): 1187–1238. MR2520804. 1019, 1030, 1031

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 1026

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press. MR3967104. doi: https://doi.org/10.1017/9781108627771. 1020

Wiesel, A. and Globerson, A. (2012). "Covariance estimation in time varying ARMA processes." In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*, 357–360. IEEE. 1017

Wu, W. B. and Pourahmadi, M. (2009). "Banding sample autocovariance matrices of stationary processes." *Statistica Sinica*, 1755–1768. MR2589209. 1017

Yang, R. and Berger, J. O. (1996). *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University. 1018

Zhang, T. and Zou, H. (2014). "Sparse precision matrix estimation via lasso penalized D-trace loss." *Biometrika*, 101(1): 103–120. MR3180660. doi: https://doi.org/10.1093/biomet/ast059. 1020