Bayesian Analysis (2018) **13**, Number 4, pp. 1215–1233

Optimal Bayesian Minimax Rates for Unconstrained Large Covariance Matrices

Kyoungjae Lee* and Jaeyong Lee†‡

Abstract. We obtain the optimal Bayesian minimax rate for the unconstrained large covariance matrix of multivariate normal sample with mean zero, when both the sample size, n, and the dimension, p, of the covariance matrix tend to infinity. Traditionally the posterior convergence rate is used to compare the frequentist asymptotic performance of priors, but defining the optimality with it is elusive. We propose a new decision theoretic framework for prior selection and define *Bayesian minimax rate*. Under the proposed framework, we obtain the optimal Bayesian minimax rate for the spectral norm for all rates of p. We also considered Frobenius norm, Bregman divergence and squared log-determinant loss and obtain the optimal Bayesian minimax rate under certain rate conditions on p. A simulation study is conducted to support the theoretical results.

MSC 2010 subject classifications: Primary 62C10, 62C20; secondary 62F15.

Keywords: Bayesian minimax rate, convergence rate, decision theoretic prior selection, unconstrained covariance.

1 Introduction

Estimating covariance matrix plays a fundamental role in multivariate data analysis. Many statistical methods in multivariate data analysis such as the principle component analysis, canonical correlation analysis, linear and quadratic discriminant analysis require the estimated covariance matrix as the starting point of the analysis. In the risk management and the longitudinal data analysis, the covariance matrix estimation is a crucial part of the analysis. The log-determinant of covariance matrix is used for constructing hypothesis test or quadratic discriminant analysis Anderson (2003).

Suppose we observe a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, $X_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, from the p-dimensional normal distribution with mean zero and covariance matrix Σ, i.e.

$$X_1, \ldots, X_n \mid \Sigma \overset{iid}{\sim} N_p(0, \Sigma).$$

We assume the zero mean and focus on the covariance matrix.

With advance of technology, data arising from various areas such as climate prediction, image processing, gene association study, and proteomics, are often high dimensional. In such high dimensional settings, it is often natural to assume that the dimension

*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, USA, klee25@nd.edu

†Department of Statistics, Seoul National University, South Korea, leejyc@gmail.com

‡Supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2011-0030811).

 https://doi.org/10.1214/18-BA1094

of the variable p tends to infinity as the sample size n gets larger, i.e. $p = p_n \longrightarrow \infty$ as $n \longrightarrow \infty$. This assumption can be justified as follows. First, when p is large in comparison with n, often the limiting scenario with p tending to infinity approximates closer to the reality than that with p fixed. Second, in many cases we can postulate the reality is infinitely complex and involves infinitely many variables, and with limited resources and time, we can collect only a portion of variables and observations. If we have more resources to collect more data, it is natural to collect more observations as well as more variables, i.e. to increase both n and p.

When p tends to infinity as $n \longrightarrow \infty$, the traditional covariance estimator is not optimal Johnstone and Lu (2009). The sparsity or bandable assumptions on large matrices have been used frequently in the literature. Many researchers have studied the large sample properties under the restrictive matrix classes. Bickel and Levina (2008b) considered the bandable covariance/precision classes and studied the convergence rate of banding estimator on those classes. Verzelen (2010) derived the convergence rate for precision matrices via sparse Cholesky factors and showed that it is the minimax rate under the Frobenius norm. In addition, the minimax convergence rates for the sparse or bandable covariance matrices were established by Cai et al. (2010), Cai and Zhou (2012a,b) and Xue and Zou (2013). For a comprehensive review on the convergence rate for the covariance and precision matrices, see Cai et al. (2016).

The posterior convergence rate has been investigated by Pati et al. (2014), Banerjee and Ghosal (2014), and Gao and Zhou (2015). Pati et al. (2014) showed that their continuous shrinkage priors are optimal for the sparse covariance estimation under the spectral norm in the sense that the posterior convergence rate is quite close to the frequentist minimax rate. They achieved a nearly minimax rate up to a $\sqrt{\log n}$ term under the spectral norm and sparse assumption even when $n = o(p)$. Banerjee and Ghosal (2014) considered Bayesian banded precision matrix estimation using graphical models. They obtained the posterior convergence rate of the precision matrix under matrix ℓ_∞ norm when $\log p = o(n)$. Gao and Zhou (2015) developed a prior distribution for the sparse principal component analysis (PCA) and showed that it achieves the minimax rate under the Frobenius norm. They also derived the posterior convergence rate under the spectral norm.

Most of the previous works on the Bayesian estimation of large covariance matrix concentrate on the constrained covariance or precision matrix. To the best of our knowledge, only Gao and Zhou (2016) considered asymptotic results for large unconstrained covariance matrix under the "large p and large n" setting. However, they attained the Bernstein-von Mises theorems under somewhat restrictive assumptions on the dimension p.

In this paper, we fill the gap in the literature. At first, we propose a new decision theoretic framework to define Bayesian minimax rate. The posterior convergence rate is the primary concept when the asymptotic optimality is studied in the Bayesian sense. But it is not completely satisfactory. The following is a quote from Ghosal and van der Vaart (2017) which they write just after defining the posterior convergence rate.

> 'We defined "a" rather than *the* rate of contraction, and hence logically any rate slower than a contraction rate is also a contraction rate. Naturally we

> are interested in a fastest decreasing sequence ϵ_n, but in general this may not exist or may be hard to establish. Thus our rate is an upper bound for a targeted rate, and generally we are happy if our rate is equal to or close to an "optimal" rate. With an abuse of terminology we often make statements like "ϵ_n is *the* rate of contraction." '

In the proposed new decision theoretic framework, a probability measure on the parameter space is an action and a prior is a decision rule for it gives a probability measure (the posterior) for a given data set. In this setup, we define the convergence rate and the Bayesian minimax rate.

We investigate the Bayesian minimax rates for unconstrained large covariance matrix. We consider four losses for the covariance inference: spectral norm, Frobenius norm, Bregman divergence and squared log-determinant loss. For the spectral norm, we have the complete result of the Bayesian minimax rate. We show that the Bayesian minimax rate is $\min(p/n, 1)$ for all rates of p. For the Frobenius norm and Bregman divergence, we show the Bayesian minimax lower bound is $p \cdot \min(p, \sqrt{n})/n$ for all rates of p, but obtained the upper bound under the constraint $p \leq \sqrt{n}$. Thus, under the condition $p \leq \sqrt{n}$, the Bayesian minimax rate is p^2/n. We also show that the Bayesian minimax rate under the squared log-determinant loss is p/n when $p = o(n)$.

The rest of the paper is organized as follows. In Section 2, we define the model, the covariance classes we consider, and introduce some notations. We propose the new decision theoretical framework and define the Bayesian minimax rate. The Bayesian minimax rates under the spectral norm, the Frobenious norm, the Bregman matrix divergence, and the squared log-determinant loss are presented in Section 3. A simulation study is given in section 4. The discussion is given in Section 5, and the proofs are given in Supplementary Material (Lee and Lee (2017)).

2 Preliminaries

2.1 The Model and the Inverse-Wishart Prior

Suppose we observe a random sample from the p-dimensional normal distribution

$$X_1, \cdots, X_n \mid \Sigma_n \overset{iid}{\sim} N_p(0, \Sigma_n), \tag{1}$$

where Σ_n is a $p \times p$ positive definite matrix, and p is a function of n such that $p = p_n \longrightarrow \infty$ as $n \longrightarrow \infty$. The true value of the covariance matrix is denoted by Σ_0 or Σ_{0n}, which is dependent on n.

For the prior of the covariance matrix Σ_n in model (1), we consider the inverse-Wishart prior

$$\Sigma_n \sim IW_p(\nu_n, A_n), \tag{2}$$

where $\nu_n > p - 1$, A_n is a $p \times p$ positive definite matrix for a proper prior. The mean of Σ_n is $A_n/(\nu_n - p - 1)$. The condition $\nu_n > p - 1$ is needed for the distribution to have a density in the space of $p \times p$ positive definite matrices. If ν_n is an integer with

$\nu_n \leq p-1$, (2) defines a singular distribution on the space of $p \times p$ positive semidefinite matrices Uhlig (1994).

We also consider the truncated inverse-Wishart prior. The inverse-Wishart prior with parameter ν and A whose eigenvalues are restricted in $[K_1, K_2]$ with $0 < K_1 < K_2$ is denoted by $IW_p(\nu, A, K_1, K_2)$. The truncated inverse-Wishart prior was adopted for technical reason. By Lemma E.1, to connect the Frobenius norm with Bregman matrix divergence, the eigenvalues of argument matrices have to be bounded. The truncated inverse-Wishart prior guarantees that the posterior covariance matrix has bounded eigenvalues.

2.2 Matrix Norms and Notations

We define the spectral norm (or matrix ℓ_2 norm) for matrices by

$$\|A\| := \sup_{\|x\|_2=1} \|Ax\|_2,$$

where $\|\cdot\|_2$ denotes the vector ℓ_2 norm defined by $\|x\|_2 := (\sum_{i=1}^p x_i^2)^{1/2}$, $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ and A is $p \times p$ matrix. The spectral norm is the same as $\sqrt{\lambda_{max}(A^T A)}$ or $\lambda_{max}(A)$ if A is symmetric, where $\lambda_{max}(B)$ denotes the largest eigenvalue of B.

The Frobenius norm is defined by

$$\|A\|_F := \left(\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2 \right)^{\frac{1}{2}},$$

where $A = (a_{ij})$ is a $p \times p$ matrix. It is the same as $\sqrt{tr(A^T A)}$, where $tr(B)$ denotes the trace of B. The Frobenius norm is the vector ℓ_2 norm with $p \times p$ matrices treated as p^2-dimensional vectors.

The Bregman divergence Bregman (1967) is originally defined for vectors, but it can be extended to the real symmetric matrices. Let ϕ be a differentiable and strictly convex function that maps real symmetric $p \times p$ matrices to $\mathbb{R}$. The Bregman divergence with ϕ between two real symmetric matrices is defined as

$$D_\phi(A, B) := \phi(A) - \phi(B) - tr[(\nabla\phi(B))^T (A - B)],$$

where A and B are real symmetric matrices and $\nabla\phi$ is the gradient of ϕ, i.e., $\nabla\phi(B) = (\partial\phi(B)/\partial B_{i,j})$.

In this paper, we consider a class of ϕ such that $\phi(X) = \sum_{i=1}^p \varphi(\lambda_i)$ where φ is a differentiable and strictly convex real-valued function and λ_i's are the eigenvalues of A. Furthermore, we assume that φ satisfies the following properties for some constant $\tau_1 > 0$:

(i) φ is a twice differentiable and strictly convex function over $\lambda \in (\tau_1, \infty)$;

(ii) there exist some constants $C > 0$ and $r \in \mathbb{R}$ such that $|\varphi(\lambda)| \leq C\lambda^r$ for all $\lambda \in (\tau_1, \infty)$; and

(iii) for any positive constants $\tau > \tau_1$, there exist some positive constants M_L and M_U such that $M_L \leq \varphi''(\lambda) \leq M_U$ for all $\lambda \in [\tau_1, \tau]$.

The above class of Bregman matrix divergences includes the squared Frobenius norm, von Neumann divergence and Stein's loss. For their use in statistics and mathematics, see Cai and Zhou (2012b), Dhillon and Tropp (2007) and Kulis et al. (2009).

If $\varphi(\lambda) = \lambda^2$, the Bregman divergence is the squared Frobenius norm $D_\phi(A, B) = \|A - B\|_F^2$. If $\varphi(\lambda) = \lambda \log \lambda - \lambda$, it is the von Neumann divergence $D_\phi(A, B) = tr(A \log A - A \log B - A + B)$, where $\log A$ is the matrix logarithm, i.e., $A = VDV^T$ is mapped to $\log A = V \log DV^T$. Here, $D = diag(d_i)$ is a $p \times p$ diagonal matrix where d_i is the ith eigenvalue of A, and $V = [V_1, \cdots, V_p]$ is a $p \times p$ orthogonal matrix where V_i is an eigenvector of A corresponding to the eigenvalue d_i. If $\varphi(\lambda) = -\log \lambda$, the Bregman divergence is the Stein's loss $D_\phi(A, B) = tr(AB^{-1}) - \log \det(AB^{-1}) - p$. The Stein's loss is the Kullback–Leibler divergence between two multivariate normal distributions with means zero and covariance matrices A and B, respectively.

Finally, we introduce some notations for asymptotic analysis which will be used subsequently. For any positive sequences a_n and b_n, we say $a_n \asymp b_n$ if there exist positive constants c and C such that $c \leq a_n/b_n \leq C$ for all sufficiently large n. We define $a_n = o(b_n)$, if $a_n/b_n \to 0$ as $n \to \infty$ and $a_n = O(b_n)$, if there exist positive constants N and M such that $|a_n| \leq M|b_n|$ for all $n \geq N$. For any random variables X_n and X, $X_n \xrightarrow{d} X$ means the convergence in distribution. For any real symmetric matrix A, $A > 0$ ($A \geq 0$) means that the matrix A is positive definite (nonnegative definite). We denote δ_A as the dirac measure at A.

2.3 A Class of Covariance Matrices

Let $\mathcal{C}_p$ denote the set of all $p \times p$ covariance matrices. For any positive constants τ, τ_1 and τ_2, define the class of covariance matrix

$$\begin{aligned} \mathcal{C}(\tau) &= \mathcal{C}_p(\tau) := \{\Sigma \in \mathcal{C}_p : \|\Sigma\| \leq \tau, \Sigma \geq 0\}, \\ \mathcal{C}(\tau_1, \tau_2) &= \mathcal{C}_p(\tau_1, \tau_2) := \{\Sigma \in \mathcal{C}_p : \lambda_{min}(\Sigma) \geq \tau_1, \|\Sigma\| \leq \tau_2\}, \end{aligned}$$

where $\lambda_{min}(\Sigma)$ is the smallest eigenvalue of Σ. Throughout the paper, we consider the model (1) and assume that the true covariance matrix belongs to $\mathcal{C}(\tau)$ or $\mathcal{C}(\tau_1, \tau_2)$.

Often the subgaussian property is used to relax the Gaussian distribution assumption. The distribution of random vector X has subgaussian property with variance factor $\tau > 0$, if

$$P(|v^T(X - \mathrm{E}X)| > t) \leq e^{-t^2/(2\tau)}$$

for all $t > 0$ and $\|v\| = 1$. The subgaussian property with variance factor τ implies $\|\mathrm{Var}(X)\| \leq 2\tau$. In the literature, the subgaussian distribution is frequently used as a basic assumption, for examples, Cai et al. (2010), Cai and Zhou (2012a,b) and Xue and Zou (2013). If X follows a multivariate normal distribution, $\|\Sigma\| \leq \tau$ is a sufficient condition for X to have the subgaussian property.

2.4 Decision Theoretic Prior Selection

Let $d(\Sigma, \Sigma')$ be a pseudo-metric that measures the discrepancy between two covariance matrices Σ and Σ'. A sequence $\epsilon_n \longrightarrow 0$ is called a posterior convergence rate at the true parameter Σ_0 if for any $M_n \longrightarrow \infty$,

$$\pi(d(\Sigma, \Sigma_0) \geq M_n \epsilon_n \mid \mathbf{X}_n) \quad \longrightarrow \quad 0$$

in $\mathbb{P}_{\Sigma_0}$-probability as $n \longrightarrow \infty$. The convergence rate is measured by the rate of ϵ_n, which allows that the posterior contraction probability converges to zero in probability $\mathbb{P}_{\Sigma_0}$, where $\mathbb{P}_{\Sigma_0}$ is the distribution for random sample $(X_1, \ldots, X_n) \overset{iid}{\sim} N_p(0, \Sigma_0)$. In the literature, the posterior is said to achieve the minimax rate if its convergence rate is the same as the frequentist minimax rate (Pati et al. (2014); Gao and Zhou (2015); Hoffmann et al. (2015)). Since the posterior convergence rate cannot be faster than the frequentist minimax rate (Hjort et al. (2010)), it is often called the optimal rate of posterior convergence (Shen and Ghosal (2015); Rocková (2017)). However, its definition is elusive as the quote from Ghosal and van der Vaart (2017) indicates.

As an alternative framework for the evaluation of the prior and the posterior, we take a frequentist decision theoretical approach. For each n, the parameter space is $\mathcal{C}_p$ and the action space is the set of all probability measures on $\mathcal{C}_p$. After the data $\mathbf{X}_n$ is collected, the posterior $\pi(\cdot|\mathbf{X}_n)$ is computed for the given prior π and the posterior takes a value in the action space. In this setup, the prior can be considered as a decision rule, because the prior and observations together produce the posterior. A probability measure in the action space will be used as a posterior for the inference, but it does not have to be generated from a prior. We define the loss and risk function of the parameter Σ_0 and the prior π as

$$\begin{aligned} \mathcal{L}(\Sigma_0, \pi(\cdot|\mathbf{X}_n)) &:= \mathrm{E}^{\pi}\big(d(\Sigma, \Sigma_0)|\mathbf{X}_n\big), \\ \mathcal{R}(\Sigma_0, \pi) &:= \mathrm{E}_{\Sigma_0}\mathcal{L}(\Sigma_0, \pi(\cdot|\mathbf{X}_n)) = \mathrm{E}_{\Sigma_0}\mathrm{E}^{\pi}\big(d(\Sigma, \Sigma_0)|\mathbf{X}_n\big). \end{aligned}$$

Note that the risk function measures the performance of the prior π. To distinguish them from the usual loss and risk, we call the above loss and risk as *posterior loss (P-loss)* and *posterior risk (P-risk)*. The P-risk itself is not new. For example, the P-risk was also used in Castillo (2014) for density estimation on the unit interval.

There are a couple of benefits of the proposed decision theoretic prior selection. First, the decision theoretic prior selection makes the definition of the minimax rate of the posterior mathematically concrete. Although the minimax rate of the posterior is used frequently, it has been used without a rigorous definition. The frequentist minimax rate is used as a proxy of the desired concept. Second, in the study of the posterior convergence rate, the scale of the loss function needs to be carefully chosen so that the posterior consistency holds. But in the proposed decision theoretic prior selection, the inconsistent priors can be compared without any conceptual difficulty. Thus, the scale of the loss function does not need to be chosen.

We now define the minimax rate and convergence rate for P-loss. Let Π_n be the class of all priors on Σ_n. A sequence r_n is said to be the *minimax rate for P-loss (P-loss*

minimax rate) or simply *the Bayesian minimax rate* for the class $\mathcal{C}_p^* \subset \mathcal{C}_p$ and the space of the prior distributions $\Pi_n^* \subset \Pi_n$, if

$$\inf_{\pi \in \Pi_n^*} \sup_{\Sigma_0 \in \mathcal{C}_p^*} \mathrm{E}_{\Sigma_0} \mathcal{L}(\Sigma_0, \pi(\cdot|\mathbf{X}_n)) \asymp r_n.$$

A prior π^* is said to have a *convergence rate for P-loss (P-loss convergence rate)* or *convergence rate* a_n, if

$$\sup_{\Sigma_0 \in \mathcal{C}_p} \mathrm{E}_{\Sigma_0} \mathcal{L}(\Sigma_0, \pi^*(\cdot|\mathbf{X}_n)) \lesssim a_n,$$

and, if $a_n \asymp r_n$ where r_n is the minimax rate for P-loss, π^* is said to attain the minimax rate for P-loss or the Bayesian minimax rate. If it is clear from context, we will drop P-loss and refer them as the minimax rate and the convergence rate. For a given inference problem, we wish to find a prior π^* which attains the minimax rate for P-loss.

Remark. The P-loss convergence rate implies the posterior convergence rate by Proposition A.1 in Supplementary Material (Lee and Lee (2017)). By obtaining the P-loss convergence rate, we also get the traditional posterior convergence rate. The converse may not be true, because for certain loss functions, the P-loss may not even converge to 0 while the posterior convergence rate converges to 0.

Remark. The P-loss convergence rate is slower than or equal to the frequentist minimax rate by Proposition A.2 in Supplementary Material (Lee and Lee (2017)). To obtain a P-loss minimax lower bound, the mathematical tools for frequentist minimax lower bound can be used.

Remark. If we assume that the prior class Π_n includes the data dependent priors, the P-loss minimax rate is the same as the frequentist minimax rate. Take $\pi = \delta_{\hat{\Sigma}^*}$ where $\hat{\Sigma}^*$ is an estimator attaining the frequentist minimax rate. Then, π attains the frequentist minimax rate and thus attains the Bayesian minimax rate. However, the data-dependent prior is not acceptable for legitimate Bayesian analysis unless the prior is dependent on ancillary statistics. Even if Π_n does not contain data-dependent priors, in most cases the frequentist and P-loss minimax rates are the same.

However, if we consider a restricted class of priors, the P-loss minimax rate might differ from the usual frequentist minimax rate. In such cases, the frequentist minimax rate will not be a natural concept to study the asymptotic properties of the posterior. See Remark in subsection 3.2.

3 Bayesian Minimax Rates under Various Matrix Loss Functions

3.1 Bayesian Minimax Rate under Spectral Norm

In this subsection, we show that the Bayesian minimax rate for covariance matrix under the spectral norm is $\min(p/n, 1)$. We also show that the prior

$$\pi_n(\Sigma_n) \quad = \quad IW_p(\Sigma_n \mid \nu_n, A_n) I\left(p \leq \frac{n}{2}\right) + \delta_{I_p}(\Sigma_n) I\left(p > \frac{n}{2}\right) \tag{3}$$

attains the Bayesian minimax rate for the class $\mathcal{C}(\tau_1, \tau_2)$ under the spectral norm, where $IW_p(\Sigma \mid \nu_n, A_n)$ is the inverse-Wishart distribution, $\nu_n > p-1$ and A_n is a $p \times p$ positive definite matrix. We have the complete result for all values of n and p. The Bayesian minimax rate holds for any n and p, regardless of their relationship. The number $1/2$ in the prior (3) can be replaced by any number in $(0, 1)$ and the prior still renders the minimax rate.

The main result of the section is given in Theorem 1 whose proof is given in Supplementary Material (Lee and Lee (2017)). We divide the proof into two parts: lower bound and upper bound parts. First, we show that the lower bound of the frequentist minimax rate is $\min(p/n, 1)$, which may be of interest in its own right, and it in turn implies that $\min(p/n, 1)$ is a Bayesian minimax lower bound. After that, the P-loss convergence rate with the prior (3) is derived, which is the same as the Bayesian minimax lower bound when $\nu_n^2 = O(np)$ and $A_n = S_n$. Consequently, we obtain the following theorem by combining these two results. Throughout the paper, Π_n is the class of all priors on $\Sigma_n \in \mathcal{C}_p$ as we have defined in subsection 2.4.

Theorem 1. *Consider the model (1). For any positive constants $\tau_1 < \tau_2$,*

$$\inf_{\pi \in \Pi_n} \sup_{\Sigma_0 \in \mathcal{C}(\tau_1, \tau_2)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n - \Sigma_0\|^2 \mid \mathbf{X}_n) \asymp \min\left(\frac{p}{n}, 1\right).$$

Furthermore, the prior (3) with $\nu_n^2 = O(np)$ and $\|A_n\|^2 = O(np)$ attains the Bayesian minimax rate.

Remark. The proof for the lower bound holds even for τ_1 and τ_2 depending on n and possibly for $\tau_1 \longrightarrow 0$ and $\tau_2 \longrightarrow \infty$ as $n \longrightarrow \infty$. In such cases, the rate of the minimax lower bound is $\tau_2^2 \cdot \min(p/n, 1)$. For details, see Theorem B.1 in the Supplementary Material (Lee and Lee (2017)). Note that τ_2 affects the minimax lower bound, while τ_1 does not. A similar phenomenon occurs for estimation of sparse spiked covariance matrices. See Theorem 4 of Cai et al. (2016).

We have complete results of the Bayesian minimax rate under the spectral norm. In words, the results above do not have any condition on the rate of p and n. For a given rate of p, we obtained the Bayesian minimax rate. When p grows the same rate as n, the above theorem shows that estimating the covariance under the spectral norm is hopeless. Indeed, this can be seen from the form of the prior (3). When $p \geq n/2$, the point mass prior δ_{I_p} gives the Bayesian minimax rate. In words, you can not do better than the useless point mass prior δ_{I_p}.

Applying techniques used in the proof of the upper bound, one can show that the prior (3) also gives the same P-loss convergence rate for precision matrix.

Corollary 1. *Consider the model (1) and prior (3) with $\nu_n^2 = O(np)$ and $\|A_n\|^2 = O(np)$. For any positive constants $\tau_1 < \tau_2$,*

$$\sup_{\Sigma_0 \in \mathcal{C}(\tau_1, \tau_2)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n^{-1} - \Sigma_0^{-1}\|^2 \mid \mathbf{X}_n) \leq c \cdot \min\left(\frac{p}{n}, 1\right)$$

for all sufficiently large n and some constant $c > 0$.

We remark here that Gao and Zhou (2016) derived a posterior convergence rate for unconstrained covariance matrix under the spectral norm when $p = o(n)$. In this paper, we obtained a P-loss convergence rate which implies the stronger convergence than a posterior convergence rate, for any n and p. Gao and Zhou (2016) also attained a posterior convergence rate for precision matrix under $p^2 = o(n)$. In this paper, Corollary 1 gives a P-loss convergence rate for any n and p.

3.2 Bayesian Minimax Rate under Frobenius Norm

Throughout this subsection, $\tau > 0$ can depend on n and possibly $\tau \longrightarrow \infty$ as $n \longrightarrow \infty$. In this subsection, we show that the rate of the Bayesian minimax lower bound for covariance matrix under Frobenius norm is $\tau^2 \cdot \min(p, \sqrt{n}) \cdot p/n$ for the class $\mathcal{C}(\tau)$, and the inverse-Wishart prior attains the Bayesian minimax lower bound when $p \leq \sqrt{n}$.

The following theorem gives the Bayesian minimax lower bound. The proof of Theorem 2 is given in Supplementary Material (Lee and Lee (2017)). In the proof of the theorem, we prove that the lower bound of the frequentist minimax rate is $\tau^2 \cdot \min(p, \sqrt{n}) \cdot p/n$ as a by-product.

Theorem 2. *Consider the model* (1)*. For any* $\tau > 0$*,*

$$\inf_{\pi \in \Pi_n} \sup_{\Sigma_0 \in \mathcal{C}(\tau)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n - \Sigma_0\|_F^2 \mid \mathbf{X}_n) \geq c \cdot \tau^2 \cdot \frac{p}{n} \cdot \min(p, \sqrt{n})$$

for all sufficiently large n *and some constant* $c > 0$*.*

Theorem 3. *Consider the model* (1) *and prior* (2) *with* $\nu_n > 0$ *and* $A_n > 0$ *for all* n*. If* $\nu_n = p$ *and* $\|A_n\|^2 = O(n)$*, for any* $\tau > 0$*,*

$$\sup_{\Sigma_0 \in \mathcal{C}(\tau)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n - \Sigma_0\|_F^2 \mid \mathbf{X}_n) \leq c \cdot \tau^2 \cdot \frac{p^2}{n}$$

for some constant $c > 0$ *and all sufficiently large* n*. Furthermore, if* $p \leq \sqrt{n}$*,* $\nu_n^2 = O(np)$ *and* $\|A_n\|^2 = O(np)$ *is the necessary and sufficient condition for achieving the rate* p^2/n*.*

Note that if $\tau > 0$ is a fixed constant, from the relationship between the spectral norm and Frobenius norm, one can obtain a P-loss convergence rate $\min(p, n) \cdot p/n$ instead of p^2/n in Theorem 3. However, in this case, one should restrict the parameter space to $\mathcal{C}(\tau_1, \tau_2)$ instead of the more general parameter space $\mathcal{C}(\tau)$.

In practice, we recommend using $\nu_n = p$ and small A_n such as $A_n = O_p$ or $A_n = I_p$, where O_p denotes a $p \times p$ zero matrix because it guarantees the rate p^2/n regardless of the relation between n and p. Note that the Jeffreys prior Jeffreys (1961)

$$\pi(\Sigma_n) \quad \propto \quad det(\Sigma_n)^{-(p+2)/2},$$

the independence-Jeffreys prior Sun and Berger (2007)

$$\pi(\Sigma_n) \quad \propto \quad det(\Sigma_n)^{-(p+1)/2}$$

and the prior proposed by Geisser and Cornfield (1963)

$$\pi(\Sigma_n) \quad \propto \quad det(\Sigma_n)^{-p}$$

satisfy the above conditions. They can be viewed as inverse-Wishart priors, $IW(\nu_n, A_n)$, with parameters $(1, O_p)$, $(0, O_p)$ and $(p-1, O_p)$, respectively. Furthermore, the $IW(p+1, S_n)$ prior, whose mean is S_n, also satisfies the conditions in Theorem 3.

By Theorem 3 and Theorem 2, we have the Bayesian minimax rate $\tau^2 \cdot p^2/n$ for covariance matrix under the Frobenius norm when $p \leq \sqrt{n}$. Thus, with the inverse-Wishart prior, we attain the Bayesian minimax rate under the Frobenius norm.

Theorem 4. *Consider the model* (1)*. If* $p \leq \sqrt{n}$*, for any* $\tau > 0$*,*

$$\inf_{\pi \in \Pi_n} \sup_{\Sigma_0 \in \mathcal{C}(\tau)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n - \Sigma_0\|_F^2 \mid \mathbf{X}_n) \asymp \tau^2 \cdot \frac{p^2}{n}.$$

Furthermore, $\nu_n^2 = O(np)$ *and* $\|A_n\|^2 = O(np)$ *is the necessary and sufficient condition for the prior* (2) *to achieve the Bayesian minimax rate when* $p \leq \sqrt{n}$*.*

Remark. In Section 2.4, we have said that the Bayesian minimax rate can be different from the frequentist minimax rate when a restricted prior class is considered, and that the frequentist minimax rate will not be a natural concept to address the asymptotic properties of the posteriors from a restricted prior class. We give an example here. Consider a prior class $\Pi_n^* = \{\pi \in IW_p(\nu_n, A_n) : \nu_n \geq n,\ A_n \in \mathcal{C}_p\}$ and assume $p \leq \sqrt{n}$. It is easy to check that

$$\inf_{\pi \in \Pi_n^*} \sup_{\Sigma_0 \in \mathcal{C}(\tau)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(\|\Sigma_n - \Sigma_0\|_F^2 \mid \mathbf{X}_n) \quad \asymp \quad \tau^2 \cdot p,$$

from the proof of Theorem 3. Note that the obtained P-loss minimax rate differs from the usual frequentist minimax rate, $\tau^2 \cdot p^2/n$.

3.3 Bayesian Minimax Rate under Bregman matrix Divergence

In this section, we obtain the Bayesian minimax rate under a certain class of Bregman matrix divergences. Let Φ be the class of differentiable and strictly convex real-valued functions satisfying (i)–(iii) conditions in the subsection 2.2, and let $\mathcal{D}_\Phi$ be the class of Bregman matrix divergences D_ϕ where $\phi(X) = \sum_{i=1}^p \varphi(\lambda_i)$ for symmetric matrix X and $\varphi \in \Phi$.

To achieve the Bayesian minimax convergence rate for Bregman matrix divergences, we use the truncated inverse-Wishart distribution $IW_p(\nu_n, A_n, K_1, K_2)$ whose eigenvalues are all in $[K_1, K_2]$ for some positive constants $K_1 < K_2$. The density function of $IW_p(\nu_n, A_n, K_1, K_2)$ is given by

$$\pi^{n, K_1, K_2}(\Sigma_n) \quad = \quad \frac{\det(\Sigma_n)^{-(\nu+p+1)/2} e^{-\frac{1}{2} tr(A_n \Sigma_n^{-1})} I(\Sigma_n \in \mathcal{C}(K_1, K_2))}{\int_{\mathcal{C}(K_1, K_2)} \det(\Sigma_n')^{-(\nu+p+1)/2} e^{-\frac{1}{2} tr(A_n \Sigma_n'^{-1})} d\Sigma_n'}, \tag{4}$$

where $\nu_n > p - 1$ and A_n is a $p \times p$ positive definite matrix.

Theorem 5. *Consider the model* (1). *If* $p \leq \sqrt{n}$, *for any positive constants* $\tau_1 < \tau_2$

$$\inf_{\pi \in \Pi_n} \sup_{\Sigma_0 \in \mathcal{C}(\tau_1, \tau_2)} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}(D_{\phi}(\Sigma_n, \Sigma_0) \mid \mathbf{X}_n) \asymp \frac{p^2}{n}$$

for all $D_{\phi} \in \mathcal{D}_{\Phi}$. *Furthermore, the prior* (4) *with* $\nu_n^2 = O(np)$, $\|A_n\|^2 = O(np)$, $K_1 < \tau_1$ *and* $K_2 > \tau_2$ *achieves the Bayesian minimax rate when* $p \leq \sqrt{n}$.

To extend the minimax result for the squared Frobenius norm to the Bregman matrix divergence, the posterior distribution for Σ_n and the true covariance Σ_0 should be included in the class $\mathcal{C}(K_1, K_2)$ and $\mathcal{C}(\tau_1, \tau_2)$, respectively, for some positive constants $K_1 < \tau_1$ and $K_2 > \tau_2$. The truncated inverse-Wishart prior was needed to restrict the posterior distribution for Σ_n within the class $\mathcal{C}(K_1, K_2)$. In practice, we recommend using sufficiently small K_1 and large K_2. According to the above theorem, the minimax convergence rate for the class $\mathcal{D}_{\Phi}$ is equivalent to that for the Frobenius norm if we consider the parameter space $\mathcal{C}(\tau_1, \tau_2)$. Moreover, the truncated inverse-Wishart prior $IW_p(\nu_n, A_n, K_1, K_2)$ achieves the Bayesian minimax rate. The proof of the theorem is given in Supplementary Material (Lee and Lee (2017)).

3.4 Bayesian Minimax Rate of Log Determinant of Covariance Matrix

In this subsection, we establish the Bayesian minimax rate for the log-determinant of the covariance matrix under squared error loss. The frequentist minimax lower bound was derived by Cai et al. (2015). We prove that the inverse-Wishart prior achieves the Bayesian minimax rate when $p = o(n)$.

The estimator of the log-determinant of the covariance matrix can be used as a basic ingredient for constructing hypothesis test or the quadratic discriminant analysis Anderson (2003). The log-determinant of the covariance matrix is needed to compute the quadratic discriminant function for multivariate normal distribution

$$-\frac{1}{2} \log \det \Sigma - \frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu),$$

where x is the random sample from $N_p(\mu, \Sigma)$. Furthermore, the differential entropy of $N_p(\mu, \Sigma)$ is given by

$$\frac{p}{2} + \frac{p \log(2\pi)}{2} + \frac{\log \det \Sigma}{2},$$

so the estimation of the differential entropy is equivalent to estimation of the log-determinant of the covariance matrix, when we consider the multivariate normal distribution. The differential entropy has various applications including independent component analysis (ICA), spectroscopy, image analysis, and information theory. See Beirlant et al. (1997), Dudewicz and Mommaerts (1991), Hyvärinen (1998) and Cover and Thomas (1991).

Cai et al. (2015) showed that the minimax rate for the log-determinant of the covariance matrix under squared error loss is p/n and their estimator achieves this optimal rate when $p = o(n)$.

On the Bayesian side, Srivastava and Gupta (2008) and Gupta and Srivastava (2010) suggested a Bayes estimator for log-determinant of the covariance matrix of the multivariate normal. They proposed using the inverse-Wishart prior and showed that the posterior mean minimizes expected Bregman divergence. In this subsection, we support their argument by showing that the inverse-Wishart prior achieves the P-loss minimax rate for log-determinant of the covariance matrix under squared error loss. Thus, we show that the inverse-Wishart prior gives the optimal result in the Bayesian sense. We also show the sufficient conditions for achieving the Bayesian minimax rate. The following theorem presents the Bayesian minimax rate for the log-determinant of the covariance matrix under the squared error loss. The proof of the theorem is given in Supplementary Material (Lee and Lee (2017)).

Theorem 6. *Consider the model* (1)*. If* $p = o(n)$*, we have*

$$\inf_{\pi \in \Pi_n} \sup_{\Sigma_0 \in \mathcal{C}_p} \mathbb{E}_{\Sigma_0} \mathbb{E}^{\pi}((\log\det\Sigma_n - \log\det\Sigma_0)^2 \mid \mathbf{X}_n) \asymp \frac{p}{n}.$$

Furthermore, prior (2) *with* $\nu_n^2 = O(n/p)$ *and* $A_n = O_p$ *attains the Bayesian minimax rate.*

Remark. One can also show that the optimal minimax convergence rate is achieved by using the prior (2) with $\nu_n^2 = O(n/p)$, $A_n = c_n S_n$ and $c_n^2 = O(n/p)$.

Remark. Gao and Zhou (2016) showed the Bernstein-von Mises result for the log-determinant of covariance, which implies a posterior convergence rate. However, they considered a restrictive parameter space $\mathcal{C}(\tau_1, \tau_2)$ and the stronger condition $p^3 = o(n)$. In this paper, the more general parameter space $\mathcal{C}_p$ and weaker condition $p = o(n)$ are sufficient for the stronger result, a P-loss convergence rate.

4 Simulation study

In this section, we support our theoretical results by a simulation study. The simulations for three loss functions, spectral norm, square of scaled Frobenius norm and squared log-determinant loss, were conducted. We compare the performance of the minimax priors with those of some frequentist estimators.

We choose the posterior mean as a Bayesian estimator. The posterior mean obtained from the minimax prior attains the minimax rate in Theorem B.2, Theorem 3 and Theorem 6 by the Jensen's inequality.

We generated dataset $X_1, \ldots, X_n$ from $N_p(0, \Sigma_0)$ where true covariance matrix Σ_0 was either diagonal or full covariance matrix. A full covariance matrix is a covariance matrix which does not have any restriction on its elements such as sparsity or banding. In the diagonal covariance setting, the true covariance is $\Sigma_0 = diag(\sigma_{0,ii})$ where $\sigma_{0,ii} \overset{iid}{\sim}$

$Unif(0,5)$. In the full covariance setting, we made the true covariance $\Sigma_0 = V^T V$ where $V = (v_{ij})$ is a $p \times p$ matrix with $v_{ij} \overset{iid}{\sim} N(0, 5/p)$. In the simulation study, the dimensions of the true covariance matrices are $25, 50, 100$ and 200, and the numbers of data n are either $n = p^2$ or $n = \lceil p^{3/2} \rceil$. For each setting, we generated a true covariance once for which we generated 100 data sets and calculated estimators of the covariance.

For the spectral norm and square of scaled Frobenius norm loss, we computed the posterior mean of the inverse-Wishart prior, $IW(\nu_n, A_n)$, for comparison. We chose $\nu_n = 2, \sqrt{n/p}, p$ and n to see the effect of the ν_n, but fixed $A_n = O_p$ to remove the prior effect on the structure of the covariance estimate. By Theorems B.2 and 3, when $n = p^2$, the inverse-Wishart prior with $\nu_n = 2, \sqrt{n/p}$ and p are minimax priors, while that with $\nu_n = n$ is not. We also computed the sample covariance S_n and the tapering estimator $\hat{\Sigma}_k$ Cai et al. (2010) for comparison. As mentioned before, the sample covariance matrix is a Bayesian estimator using inverse-Wishart prior with $\nu_n = p + 1$ and $A_n = O_p$, which satisfies the conditions in Theorem 3. We used $k = \sqrt{n}$ as the threshold of tapering estimator. It corresponds to $\alpha = 0$ in Cai et al. (2010), which gives the minimal sparse constraint for the covariance matrix in their class.

Figure 1 summarizes the simulation results for the spectral norm. Each point of the plot was calculated by

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

where $\widehat{\Sigma}_n^{(s)}$ is the estimate of the true covariance Σ_0 in s-th simulation. The first and second rows of Figure 1 show the results when the true covariance matrix is a diagonal and full covariance, respectively; the left and right columns are the results when $n = p^2$ and $n = \lceil p^{3/2} \rceil$, respectively.

The inverse-Wishart prior with $\nu_n = p$ and the sample covariance performed well in all cases. They are either the best or comparable to the best. When $n = \lceil p^{3/2} \rceil$, the truncated inverse-Wishart prior with $\nu_n = n$ is not minimax, and the simulation results show that it performed the worst or the second to the worst. The inverse-Wishart priors with $\nu_n = 2$ and $\sqrt{n/p}$ are minimax, and thus their risks decrease as $n \longrightarrow \infty$ in all cases, but their performance are slightly worse than that with $\nu_n = p$. The tapering estimator $\hat{\Sigma}_k$ performed the best in diagonal settings because it gives zero to many of upper and lower diagonal elements or shrink them toward zero. However, in the full covariance settings, it performed the worst or close to the worst for the same reason.

Figure 2 summarizes the simulation results for Frobenius norm. Each point of the plot was calculated by

$$\frac{1}{100} \sum_{s=1}^{100} \frac{1}{p} \|\Sigma_0 - \widehat{\Sigma}_n^{(s)}\|_F^2,$$

where $\widehat{\Sigma}_n^{(s)}$ is the estimate of the true covariance Σ_0 in s-th simulation. The results are quite similar to the spectral norm case.

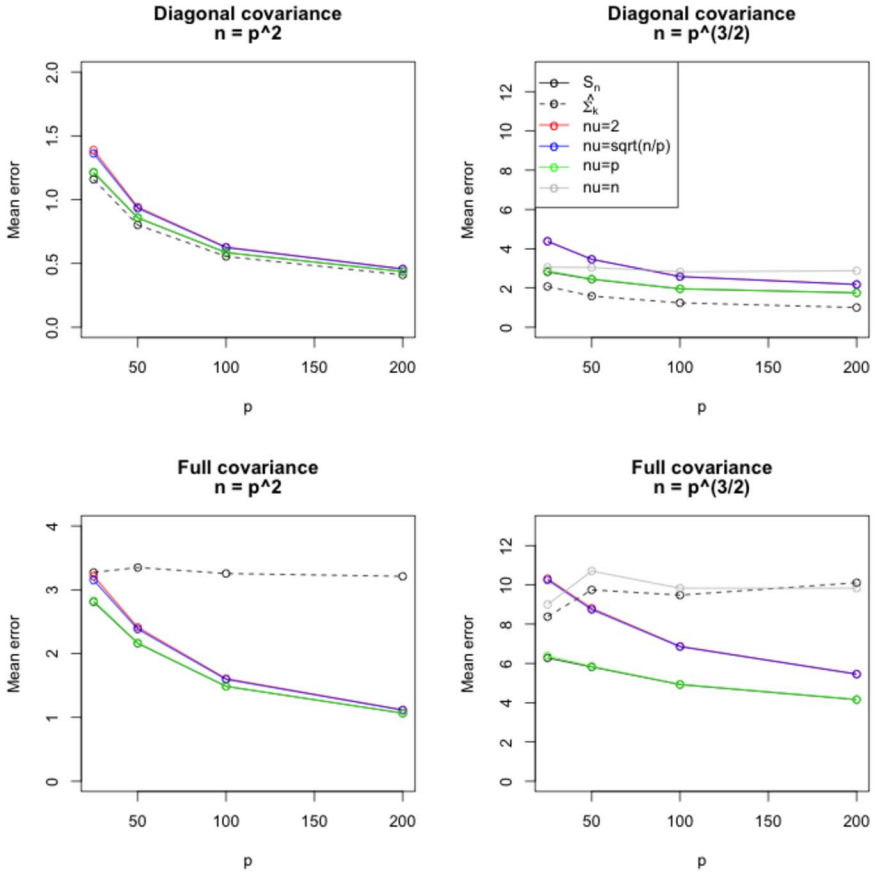

Figure 1: The risks for the Bayes estimator with $IW(\nu_n, O_p, K)$, the sample covariance S_n and tapering estimator $\hat{\Sigma}_k$ under the spectral norm loss function. The true covariances were generated in diagonal setting (top row) and full covariance setting (bottom row). The number of the observation was chosen by either $n = p^2$ (left column) or $n = \lceil p^{3/2} \rceil$ (right column).

For the square of log-determinant loss, we chose the maximum likelihood estimator (MLE) $\log\det S_n$ and the uniformly minimum variance unbiased estimator (UMVUE) for comparison. The UMVUE of $\log\det\Sigma$ is given by

$$\log\det S_n + p\log\left(\frac{n}{2}\right) - \sum_{j=0}^{p-1}\psi\left(\frac{n-k}{2}\right),$$

where ψ is the digamma function which is defined by $\psi(x) = d/dz \log\Gamma(z)|_{z=x}$ where Γ is the gamma function. See Ahmed and Gokhale (1989) for more details. We tried the same settings for inverse-Wishart prior as before. Note that for $n = p^2$ and $n = \lceil p^{3/2} \rceil$,

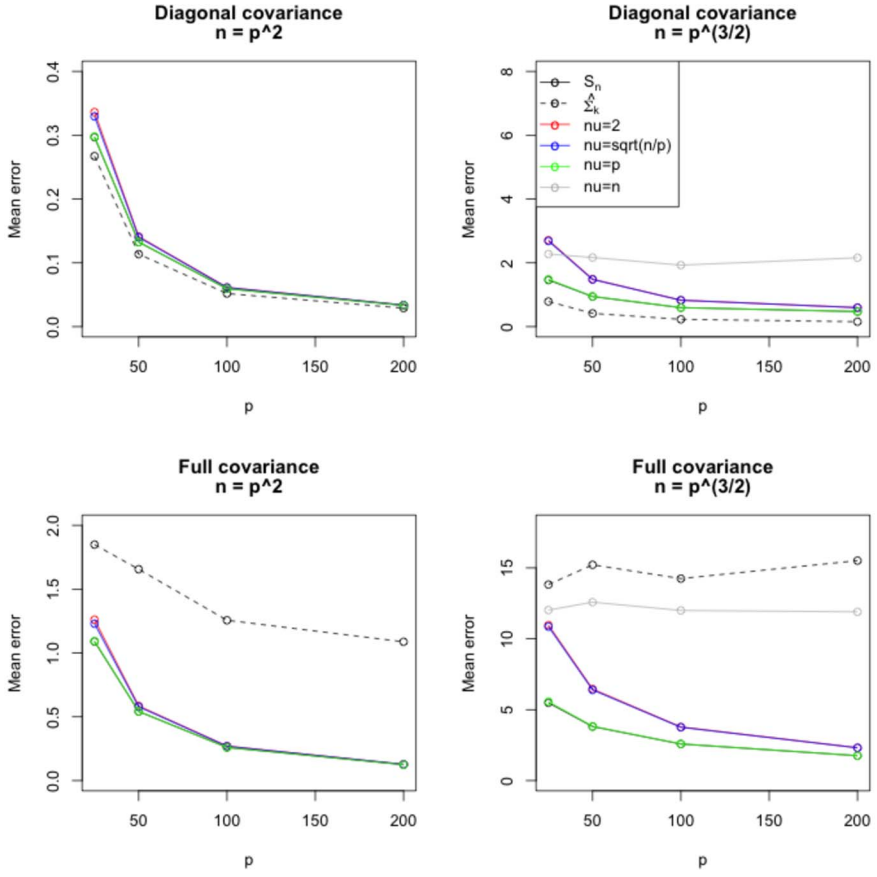

Figure 2: The risks for the Bayes estimator with $IW(\nu_n, O_p, K)$, the sample covariance S_n and tapering estimator $\hat{\Sigma}_k$ under the squared Frobenius norm loss function. The true covariances were generated in diagonal setting (top row) and full covariance setting (bottom row). The number of the observation was chosen by either $n = p^2$ (left column) or $n = \lceil p^{3/2} \rceil$ (right column).

the choices $\nu_n = 2$ and $\sqrt{n/p}$ satisfy the sufficient condition in Theorem 6 while $\nu_n = p$ and n do not. The posterior mean of the log-determinant for the inverse-Wishart prior is

$$\log\det\left(S_n + \frac{A_n}{n}\right) + p\log\left(\frac{n}{2}\right) - \sum_{j=0}^{p-1}\psi\left(\frac{n+\nu_n-k}{2}\right).$$

Thus, the UMVUE is the same as the Bayesian estimator using inverse-Wishart prior with $\nu_n = 0$ and $A_n = O_p$, which satisfies the sufficient condition in Theorem 6.

Figure 3 summarizes the simulation results for log-determinant. Each point of the plot was calculated by

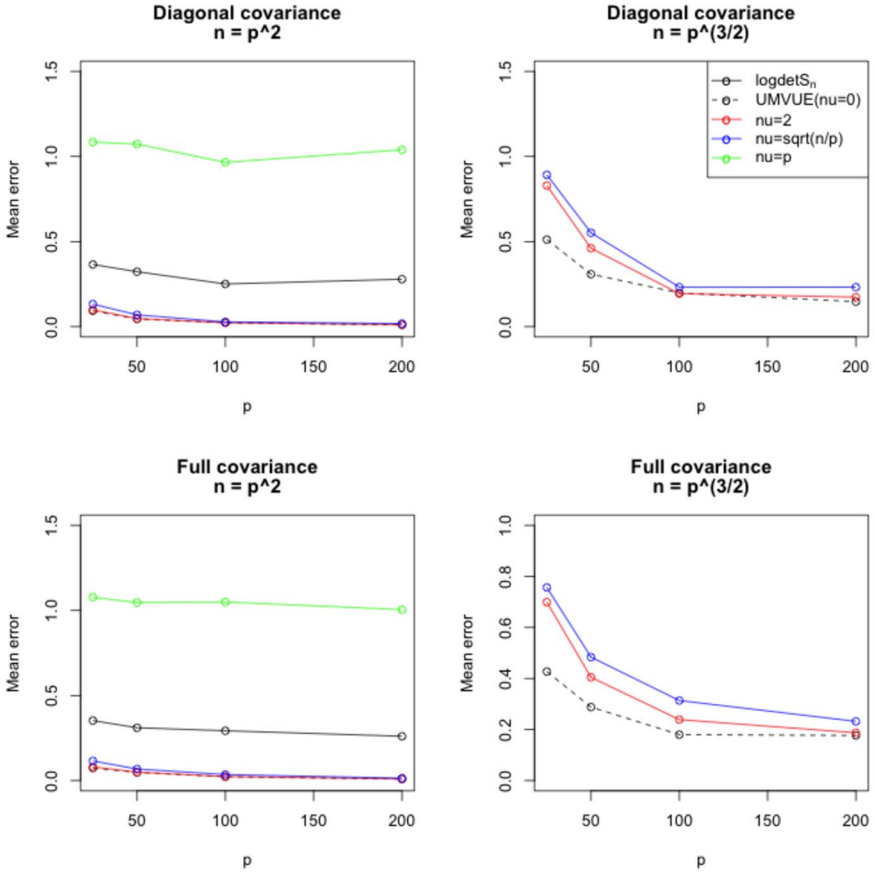

Figure 3: The squared log-determinant loss function plot. The true covariances were generated in diagonal setting (top row) and full covariance setting (bottom row). The number of the observation was chosen by either $n = p^2$ (left column) or $n = \lceil p^{3/2} \rceil$ (right column).

$$\frac{1}{100}\sum_{s=1}^{100}(\log\det\Sigma_0 - \widehat{\log\det}\,\Sigma_n^{(s)})^2,$$

where $\widehat{\log\det}\,\Sigma_n^{(s)}$ is the estimate of $\log\det\Sigma$ in s-th simulation and Σ_0 is the true covariance. The top and bottom rows are for the diagonal and full true covariance cases, respectively; the left and right columns are for $n = p^2$ and $\lceil p^{3/2} \rceil$, respectively.

For the squared log-determinant loss, the inverse-Wishart priors with $\nu_n = 2$ and $\sqrt{n/p}$ are minimax, while those with $\nu_n = p$ and n are not. The UMVUE or the Bayes estimator of the inverse-Wishart priors with $\nu_n = 0$ performed the best in all cases. The inverse-Wishart priors with $\nu_n = 2$ and $\sqrt{n/p}$ performed comparable to the UMVUE.

Interestingly, the inverse-Wishart priors with $\nu_n = p$, which was the best under the spectral norm, performed worst in all cases. When $n = \lceil p^{3/2} \rceil$, the results for $\nu_n = p$ do not appear in the Figure 3 because of its large risk values. This signifies the fact that we need to choose different prior parameter for different loss function.

5 Discussion

In this paper, we develop a new framework for the Bayesian minimax theory, and introduce Bayesian minimax rate and P-loss convergence rate. The proposed decision theoretic framework gives an alternative way to distinguish the good priors from the inadequate ones and makes the definition of the minimax rate of the posterior clear. We obtain the Bayesian minimax rates for the normal covariance model under the various loss functions: spectral norm, the squared Frobenius norm, Bregman matrix divergence and squared log-determinant loss for large covariance estimation. We show that the inverse-Wishart prior or truncated inverse-Wishart prior attains the Bayesian minimax rate. The simulation results support the theory obtained.

Supplementary Material

Supplementary Material for "Optimal Bayesian Minimax Rates for Unconstrained Large Covariance Matrices" (DOI: 10.1214/18-BA1094SUPP; .pdf).

References

Ahmed, N. A. and Gokhale, D. (1989). "Entropy expressions and their estimators for multivariate distributions." *IEEE Transactions on Information Theory*, 35(3): 688–692. 1228

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley. URL https://books.google.com.pe/books?id=1Ts4nwEACAAJ 1215, 1225

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. 1216

Beirlant, J., Dudewicz, E. J., Györfi, L., and van der Meulen, E. C. (1997). "Nonparametric entropy estimation: An overview." *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39. 1225

Bickel, P. J. and Levina, E. (2008b). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, 36(1): 199–227. MR2387969. doi: https://doi.org/10.1214/009053607000000758. 1216

Bregman, L. M. (1967). "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming." *USSR Computational Mathematics and Mathematical Physics*, 7(3): 200–217. 1218

Cai, T. T., Liang, T., and Zhou, H. H. (2015). "Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions." *Journal of Multivariate Analysis*, 137: 161–172. 1225

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation." *Electronic Journal of Statistics*, 10(1): 1–59. 1216, 1222

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). "Optimal rates of convergence for covariance matrix estimation." *The Annals of Statistics*, 38(4): 2118–2144. 1216, 1219, 1227

Cai, T. T. and Zhou, H. H. (2012a). "Minimax estimation of large covariance matrices under l1 norm." *Statistica Sinica*, 22(4): 1319–1378. 1216, 1219

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. 1216, 1219

Castillo, I. (2014). "On Bayesian supremum norm contraction rates." *The Annals of Statistics*, 42(5): 2058–2091. 1220

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience. 1225

Dhillon, I. S. and Tropp, J. A. (2007). "Matrix nearness problems with Bregman divergences." *SIAM Journal on Matrix Analysis and Applications*, 29(4): 1120–1146. 1219

Dudewicz, E. J. and Mommaerts, W. (1991). "Maximum entropy methods in modern spectroscopy: a review and an empiric entropy approach." In *conference proceedings on The frontiers of statistical scientific theory & industrial applications (Vol. II)*, 115–160. American Sciences Press. 1225

Gao, C. and Zhou, H. H. (2015). "Rate-optimal posterior contraction for sparse PCA." *The Annals of Statistics*, 43(2): 785–818. 1216, 1220

Gao, C. and Zhou, H. H. (2016). "Bernstein-von Mises theorems for functionals of the covariance matrix." *Electronic Journal of Statistics*, 10(2): 1751–1806. 1216, 1223, 1226

Geisser, S. and Cornfield, J. (1963). "Posterior distributions for multivariate normal parameters." *Journal of the Royal Statistical Society: Series B*, 25: 368–376. 1224

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. 1216, 1220

Gupta, M. and Srivastava, S. (2010). "Parametric Bayesian estimation of differential entropy and relative entropy." *Entropy*, 12(4): 818–843. 1226

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. URL https://books.google.co.kr/books?id=0GUzMF59AsgC 1220

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). "On adaptive posterior concentration rates." *The Annals of Statistics*, 43(5): 2259–2295. 1220

Hyvärinen, A. (1998). "New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit." In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10*, NIPS '97, 273–279. Cambridge, MA, USA: MIT Press. URL http://dl.acm.org/citation.cfm?id=302528.302606 1225

Jeffreys, H. (1961). *Theory of Probability*. Oxford, England: Oxford, third edition. 1223

Johnstone, I. M. and Lu, A. Y. (2009). "On consistency and sparsity for principal components analysis in high dimensions." *Journal of the American Statistical Association*, 104(486): 682–693. 1216

Kulis, B., Sustik, M. A., and Dhillon, I. S. (2009). "Low-rank kernel learning with Bregman matrix divergences." *Journal of Machine Learning Research*, 10: 341–376. 1219

Lee, K. and Lee, J. (2017). "Supplementary material for "Optimal Bayesian minimax rates for unconstrained large covariance matrices"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1094SUPP. 1217, 1221, 1222, 1223, 1225, 1226

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. 1216, 1220

Rocková, V. (2017). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *The Annals of Statistics*, 1–34. To appear. 1220

Shen, W. and Ghosal, S. (2015). "Adaptive Bayesian procedures using random series priors." *Scandinavian Journal of Statistics*, 42(4): 1194–1213. 1220

Srivastava, S. and Gupta, M. R. (2008). "Bayesian estimation of the entropy of the multivariate Gaussian." In *2008 IEEE International Symposium on Information Theory*, 1103–1107. IEEE. 1226

Sun, D. and Berger, J. O. (2007). "Objective Bayesian analysis for the multivariate normal model." *Bayesian Statistics*, 8: 525–547. MR2433206. 1223

Uhlig, H. (1994). "On singular Wishart and singular multivariate beta distributions." *The Annals of Statistics*, 22(1): 395–405. 1218

Verzelen, N. (2010). "Adaptive estimation of covariance matrices via cholesky decomposition." *Electronic Journal of Statistics*, 4: 1113–1150. 1216

Xue, L. and Zou, H. (2013). "Minimax optimal estimation of general bandable covariance matrices." *Journal of Multivariate Analysis*, 116: 45–51. 1216, 1219