Bayesian Analysis (2022) **17**, Number 4, pp. 1193–1217

Bayesian Sparse Spiked Covariance Model with a Continuous Matrix Shrinkage Prior*

Fangzheng Xie[†,¶], Joshua Cape[‡], Carey E. Priebe[§], and Yanxun Xu[¶]

Abstract. We propose a Bayesian methodology for estimating spiked covariance matrices with a jointly sparse structure in high dimensions. The spiked covariance matrix is reparameterized in terms of the latent factor model, where the loading matrix is equipped with a novel matrix spike-and-slab LASSO prior, which is a continuous shrinkage prior for modeling jointly sparse matrices. We establish the rate-optimal posterior contraction for the covariance matrix with respect to the spectral norm as well as that for the principal subspace with respect to the projection spectral norm loss. We also study the posterior contraction rate of the principal subspace with respect to the two-to-infinity norm loss, a novel loss function measuring the distance between subspaces that is able to capture entry-wise eigenvector perturbations. We show that the posterior contraction rate with respect to the two-to-infinity norm loss is tighter than that with respect to the routinely used projection spectral norm loss under certain low-rank and bounded coherence conditions. In addition, a point estimator for the principal subspace is proposed with the rate-optimal risk bound with respect to the projection spectral norm loss. The numerical performance of the proposed methodology is assessed through synthetic examples and the analysis of a real-world face data example.

MSC2020 subject classifications: Primary 62H25, 62C10; secondary 62H12.

Keywords: joint sparsity, latent factor model, matrix spike-and-slab LASSO, rate-optimal posterior contraction, two-to-infinity norm loss.

1 Introduction

In contemporary statistics, datasets are typically collected with high-dimensionality, where the dimension p can be significantly larger than the sample size n. For example, in genomics studies, the number of genes is typically much larger than the number of subjects (The Cancer Genome Atlas Network et al., 2012). In computer vision, the number of pixels in each image can be comparable to or exceed the number of images when the resolution of these images is relatively high (Georghiades et al., 2001; Lee et al., 2005). When dealing with such high-dimensional datasets, covariance matrix estimation plays a central role in understanding the complex structure of the data and has received significant attention in various contexts, including latent factor models (Geweke and

arXiv: 1808.07433

*The work of Xu was supported by NSF 1918854 and NSF 1940107.

[†]Department of Statistics, Indiana University. Corresponding author, fxie@iu.edu

[‡]Department of Statistics, University of Pittsburgh, joshua.cape@pitt.edu

[§]Department of Applied Mathematics and Statistics, Johns Hopkins University, cep@jhu.edu

[¶]Department of Applied Mathematics and Statistics, Johns Hopkins University, yanxun.xu@jhu.edu

 https://doi.org/10.1214/21-BA1292

Zhou, 1996; Bernardo et al., 2003), Gaussian graphical models (Wainwright and Jordan, 2008; Liu et al., 2012), etc. However, in the high-dimensional setting, additional structural assumptions are often necessary in order to address challenges associated with statistical inference (Johnstone and Lu, 2009). For example, sparsity is introduced for sparse covariance/precision matrix estimation (Friedman et al., 2008; Cai and Zhou, 2012; Cai et al., 2016), and low-rank structures are enforced in spiked covariance models (Johnstone, 2001; Cai et al., 2015). The readers can refer to Cai et al. (2016) for a recent literature review.

In this paper, we focus on the sparse spiked covariance models under the Gaussian sampling distribution assumption. The spiked covariance models, originally named in Johnstone (2001), are a class of models that can be described as follows: The observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independently collected from a p-dimensional mean-zero normal distribution with covariance matrix $\mathbf{\Sigma}$ of the form

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} + \sigma^2\mathbf{I}_p, \tag{1.1}$$

where $\mathbf{U}$ is a $p \times r$ matrix with orthonormal columns, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_r)$ is an $r \times r$ diagonal matrix with $\lambda_1 \geq \ldots \geq \lambda_r > 0$, and $r < p$. Since the eigenvalues of the covariance matrix are $\lambda_1 + \sigma^2 \geq \ldots \geq \lambda_r + \sigma^2 > \sigma^2 = \cdots = \sigma^2$ and $\lambda_r > 0$, there exists an eigen-gap $\lambda_r(\mathbf{\Sigma}) - \lambda_{r+1}(\mathbf{\Sigma}) = (\lambda_r + \sigma^2) - \sigma^2 = \lambda_r > 0$, where $\lambda_k(\mathbf{\Sigma})$ denotes the k-th largest eigenvalue of $\mathbf{\Sigma}$ for $k = 1, \ldots, p$. Therefore, the first r leading eigenvalues of $\mathbf{\Sigma}$ can be regarded as "spike" or signal eigenvalues, and the remaining $p - r$ eigenvalues may be treated as "bulk" or noise eigenvalues. We further assume that the eigenvector matrix $\mathbf{U}$ is jointly sparse, the formal definition of which is deferred to Section 2.1. Roughly speaking, the joint sparsity refers to a significant amount of rows in $\mathbf{U}$ being zero, which allows for feature selections and brings straightforward interpretation in many applications. For example, in the analysis of face images, a classical method to extract common features among different facial characteristics, expressions, illumination conditions, etc., is to obtain the eigenvectors of these face data, referred to as eigenfaces. Each coordinate of these eigenvectors corresponds to a specific pixel in the image. Nonetheless, the number of pixels (features) is typically much larger than the number of images (samples). It is often desirable to gain insights into the face information via a relatively small number of pixels, referred to as key pixels.

The literature on sparse spiked covariance matrix estimation in high-dimensions from a frequentist perspective is quite rich. In Johnstone and Lu (2009), it is shown that the classical principal component analysis can fail when $p \gg n$. In Cai et al. (2013) and Vu and Lei (2013), the minimax estimation of the principal subspace (*i.e.*, the linear subspace spanned by the eigenvector matrix $\mathbf{U}$) with respect to the projection Frobenius norm loss under various sparse structures on $\mathbf{U}$ is considered. Under the joint sparsity assumption, Cai et al. (2015) further provide minimax estimation procedures of the principal subspace with respect to the projection spectral norm loss.

In contrast, there is comparatively limited literature on Bayesian estimation of sparse spiked covariance matrices that provides theoretical guarantees. Recently, Pati et al. (2014), Gao and Zhou (2015), and Ning (2021) explore the posterior contraction rates for Bayesian estimation of sparse spiked covariance models. In particular, in Pati et al.

(2014), the authors discuss the posterior contraction behavior of the covariance matrix $\mathbf{\Sigma}$ with respect to the spectral norm loss under the Dirichlet-Laplace shrinkage prior (Bhattacharya et al., 2015), but the contraction rates are sub-optimal when the number of spikes r grows with the sample size; In Gao and Zhou (2015), the authors propose a carefully designed prior on $\mathbf{U}$ that yields the rate-optimal posterior contraction of the principal subspace with respect to the projection Frobenius norm loss, but their approach is computationally intractable except for the posterior mean as a point estimator. Some efficient computation algorithms are developed in Ning (2021) with a spike-and-slab prior, including the expectation-maximization algorithm and the variational inference algorithm. Ning (2021) also discuss the contraction rate of the exact posterior and the variational posterior.

We propose a continuous matrix shrinkage prior, referred to as the matrix spike-and-slab LASSO prior, to model the joint sparsity of the eigenvector matrix $\mathbf{U}$ of the spiked covariance matrix. The matrix spike-and-slab LASSO prior is a novel continuous shrinkage prior that generalizes the classical spike-and-slab LASSO prior for vectors (Rockova, 2018; Rockova and George, 2018) to jointly sparse rectangular matrices. The major contribution of this work is two-fold: Firstly, we establish the rate-optimal posterior contraction for the entire covariance matrix $\mathbf{\Sigma}$ with respect to the spectral norm loss as well as that for the principal subspace with respect to the projection spectral norm loss; Secondly, we also focus on the two-to-infinity norm loss, a novel loss function measuring the entrywise behavior between linear subspaces. This loss function can detect entrywise perturbations of the eigenvector matrix $\mathbf{U}$ spanning the principal subspace. Under certain low-rank and bounded coherence conditions on $\mathbf{U}$, we obtain a tighter posterior contraction rate for the principal subspace with respect to the two-to-infinity norm loss than that with respect to the projection spectral norm loss. Besides the contraction of the full posterior distribution, the Bayes procedure also leads to a point estimator for the principal subspace with a rate-optimal risk bound.

The rest of the paper is organized as follows. In Section 2, we briefly review the sparse spiked covariance models, introduce the relevant loss functions, and propose the matrix spike-and-slab LASSO prior. Section 3 elaborates on our theoretical contributions of the matrix spike-and-slab LASSO prior and the posterior contraction rates under various loss functions. The numerical performance of the proposed methodology is presented in Section 4 through synthetic examples and the analysis of a real-world computer vision dataset. Further discussion is included in Section 5.

Notations Let p and r be positive integers. We adopt the shorthand notation $[p] = \{1, \ldots, p\}$. For any finite set S, we use $|S|$ to denote the cardinality of S. The symbols $\lesssim$ and $\gtrsim$ mean the inequality up to a universal constant, *i.e.*, $a \lesssim b$ ($a \gtrsim b$, resp.) if $a \leq Cb$ ($a \geq Cb$) for some absolute constant $C > 0$. We write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. The $p \times r$ zero matrix is denoted by $\mathbf{0}_{p\times r}$ and the p-dimensional zero column vector is denoted by $\mathbf{0}_p$. When the dimension is clear, the zero matrix is simply denoted by $\mathbf{0}$. The $p \times p$ identity matrix is denoted by $\mathbf{I}_p$ and the subscript p is sometimes omitted when the dimension is clear. An orthonormal r-frame in $\mathbb{R}^p$ is a $p \times r$ matrix $\mathbf{U}$ with orthonormal columns, *i.e.*, $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{r\times r}$. The set of all orthonormal r-frames in $\mathbb{R}^p$ is

denoted by $\mathbb{O}(p,r)$. When $p=r$, we write $\mathbb{O}(r)=\mathbb{O}(r,r)$. For a p-dimensional vector $\mathbf{x}\in\mathbb{R}^p$, we use x_j to denote its jth component, $\|\mathbf{x}\|_1=\sum_{j=1}^p|x_j|$ to denote its ℓ_1-norm, $\|\mathbf{x}\|_2$ to denote its ℓ_2-norm, and $\|\mathbf{x}\|_\infty=\max_{j\in[p]}|x_j|$ to denote its ℓ_∞-norm. For a symmetric square matrix $\mathbf{\Sigma}\in\mathbb{R}^{p\times p}$, we use $\lambda_k(\mathbf{\Sigma})$ to denote the kth-largest eigenvalue of $\mathbf{\Sigma}$. For a matrix $\mathbf{A}\in\mathbb{R}^{p\times r}$, we use $\mathbf{A}_{j*}$ to denote the row vector formed by the jth row of $\mathbf{A}$, $\mathbf{A}_{*k}$ to denote the column vector formed by the kth column of $\mathbf{A}$, the lower case letter a_{ij} to denote the (i,j)-th element of $\mathbf{A}$, $\|\mathbf{A}\|_{\mathrm{F}}=\sqrt{\sum_{j=1}^p\sum_{k=1}^r a_{jk}^2}$ to denote the Frobenius norm of $\mathbf{A}$, $\|\mathbf{A}\|_2=\sqrt{\lambda_1(\mathbf{A}^{\mathrm{T}}\mathbf{A})}$ to denote the spectral norm of $\mathbf{A}$, $\|\mathbf{A}\|_{2\to\infty}=\max_{\|\mathbf{x}\|_2=1}\|\mathbf{A}\mathbf{x}\|_\infty$ to denote the two-to-infinity norm of $\mathbf{A}$, and $\|\mathbf{A}\|_\infty=\max_{\|\mathbf{x}\|_\infty=1}\|\mathbf{A}\mathbf{x}\|_\infty$ to denote the (matrix) infinity norm of $\mathbf{A}$. We remark that the matrix infinity norm can be equivalently written as $\|\mathbf{A}\|_\infty=\max_{j\in[p]}\sum_{k=1}^r|a_{jk}|$, which differs from the maximum absolute value of the entries of $\mathbf{A}$. The prior and posterior distributions appearing in this paper are denoted by Π. The density of Π with respect to the underlying sigma-finite measure is denoted by π.

2 Sparse Bayesian spiked covariance models

2.1 Background and loss functions

In the spiked covariance model (1.1), the matrix $\mathbf{\Sigma}$ has the form $\mathbf{\Sigma}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}+\sigma^2\mathbf{I}_p$. We focus on the case where the leading r eigenvectors of $\mathbf{\Sigma}$ (the columns of $\mathbf{U}$) are jointly sparse (Cai et al., 2015; Vu and Lei, 2013). Formally, the row support of $\mathbf{U}$ is defined as $\mathrm{supp}(\mathbf{U})=\left\{j\in[p]:(\mathbf{U}_{j*})^{\mathrm{T}}\neq\mathbf{0}_r\right\}$. We say $\mathbf{U}$ is jointly s-sparse if $|\mathrm{supp}(\mathbf{U})|\leq s$. Heuristically, this assumption asserts that the signal comes from at most s features among all p features. Geometrically, the joint sparsity has the interpretation that at most s coordinates of $\mathbf{y}_i$ generate the subspace $\mathrm{Span}\{\mathbf{U}_{*1},\ldots,\mathbf{U}_{*r}\}$ (Vu and Lei, 2013). We also remark that $s\geq r$ due to the orthonormal constraint on the columns of $\mathbf{U}$.

This paper studies a Bayesian approach for estimating the covariance matrix $\mathbf{\Sigma}$. We quantify how well the proposed methodology can estimate the entire covariance matrix $\mathbf{\Sigma}$ and the principal subspace $\mathrm{Span}\{\mathbf{U}_{*1},\cdots,\mathbf{U}_{*r}\}$ in a high-dimensional and jointly sparse setup. Leaving the Bayesian methodology for a moment, we first introduce some necessary background and the related loss functions for the sparse spiked covariance models in general. Throughout the paper, we write $\mathbf{\Sigma}_0=\mathbf{U}_0\mathbf{\Lambda}_0\mathbf{U}_0^{\mathrm{T}}+\sigma_0\mathbf{I}_p$ to be the true covariance matrix that generates the data $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]^{\mathrm{T}}$ from the p-dimensional multivariate Gaussian distribution $\mathrm{N}_p(\mathbf{0}_p,\mathbf{\Sigma}_0)$, where $\mathbf{\Lambda}_0=\mathrm{diag}(\lambda_{01},\cdots,\lambda_{0r})$, $\lambda_{01}\geq\ldots\geq\lambda_{0r}>0$, and $\sigma_0^2>0$. The parameter space for $\mathbf{\Sigma}$ is

$$\Theta(p,r,s)=\left\{\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}+\sigma^2\mathbf{I}_p:\mathbf{U}\in\mathbb{O}(p,r),|\mathrm{supp}(\mathbf{U})|\leq s,\lambda_1\geq\ldots\geq\lambda_r>0,\sigma^2>0\right\}.$$

When $(s\log p)/n\to 0$ and $\lambda_{01},\lambda_{0r}$ are bounded away from 0 and $+\infty$, Cai et al. (2015) establish the minimax rate of convergence for estimating $\mathbf{\Sigma}\in\Theta(p,r,s)$:

$$\inf_{\widehat{\mathbf{\Sigma}}}\sup_{\mathbf{\Sigma}_0\in\Theta(p,r,s)}\mathbb{E}_{\mathbf{\Sigma}_0}\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}_0\|_2^2\asymp\frac{s\log p}{n}. \tag{2.1}$$

For the estimation of the principal subspace $\mathrm{Span}\{\mathbf{U}_{*1}, \ldots, \mathbf{U}_{*r}\}$, a standard loss function is the projection spectral norm loss $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$, which is equivalent to the spectral sine-theta distance between subspaces (Stewart and Sun, 1990). The corresponding minimax rate of convergence for $\mathbf{U}\mathbf{U}^{\mathrm{T}}$ is given by Cai et al. (2015):

$$\inf_{\widehat{\mathbf{U}}} \sup_{\boldsymbol{\Sigma}_0 \in \Theta(p,r,s)} \mathbb{E}_{\boldsymbol{\Sigma}_0} \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2^2 \asymp \frac{s \log p}{n}. \tag{2.2}$$

Motivated by the recent papers Cape et al. (2019a) and Cape et al. (2019b), which presents a collection of technical tools for the analysis of entrywise eigenvector perturbation bounds, we also focus on the following two-to-infinity norm loss

$$\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty} \tag{2.3}$$

for estimating $\mathrm{Span}\{\mathbf{U}_{*1}, \ldots, \mathbf{U}_{*r}\}$ in addition to the projection spectral norm loss, where $\mathbf{W}_{\mathbf{U}} = \arg\inf_{\mathbf{W}\in\mathbb{O}(r)} \|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}\|_{\mathrm{F}}$. Here, $\mathbf{W}_{\mathbf{U}}$ corresponds to the orthogonal alignment of $\mathbf{U}_0$ so that $\widehat{\mathbf{U}}$ and $\mathbf{U}_0\mathbf{W}_{\mathbf{U}}$ are close in the Frobenius norm sense. As pointed out in Cape et al. (2019b), the use of $\mathbf{W}_{\mathbf{U}}$ as the orthogonal alignment matrix is preferred over the alignment matrix $\mathbf{W}_{2\to\infty} = \arg\inf_{\mathbf{W}\in\mathbb{O}(r)} \|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}\|_{2\to\infty}$ because $\mathbf{W}_{2\to\infty}$ is not analytically computable in general, whereas $\mathbf{W}_{\mathbf{U}}$ can be explicitly computed (Stewart and Sun, 1990). Specifically, if $\mathbf{U}_0^{\mathrm{T}}\widehat{\mathbf{U}}$ admits the singular value decomposition $\mathbf{U}_0^{\mathrm{T}}\widehat{\mathbf{U}} = \widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{V}}^{\mathrm{T}}$, then $\mathbf{W}_{\mathbf{U}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^{\mathrm{T}}$.

The following lemma formalizes the connection between the projection spectral norm loss and the two-to-infinity norm loss.

Lemma 2.1. *Let* $\mathbf{U}$ *and* $\mathbf{U}_0$ *be two orthonormal* r*-frames in* $\mathbb{R}^p$*, where* $2r < p$*. Then there exists an orthonormal* $2r$*-frame* $\mathbf{V}_{\mathbf{U}}$ *in* $\mathbb{R}^p$ *depending on* $\mathbf{U}$ *and* $\mathbf{U}_0$*, such that*

$$\|\mathbf{U} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty} \le \|\mathbf{V}_{\mathbf{U}}\|_{2\to\infty} \left(\|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2 + \|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2^2\right),$$

where $\mathbf{W}_{\mathbf{U}} = \arg\inf_{\mathbf{W}\in\mathbb{O}(r)} \|\mathbf{U} - \mathbf{U}_0\mathbf{W}\|_{\mathrm{F}}$ *is the Frobenius orthogonal alignment matrix.*

When the projection spectral norm loss $\|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$ is much smaller than one, Lemma 2.1 states that the two-to-infinity norm loss $\|\mathbf{U} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty}$ can be upper bounded by the product of the projection spectral norm loss and $\|\mathbf{V}_{\mathbf{U}}\|_{2\to\infty}$, where $\mathbf{V}_{\mathbf{U}} \in \mathbb{O}(p, 2r)$ is an orthonormal $2r$-frame in $\mathbb{R}^p$. In particular, under the sparse spiked covariance models in high dimensions, the number of spikes r can be much smaller than the dimension p (*i.e.*, $\mathbf{V}_{\mathbf{U}}$ is a "tall and thin" rectangular matrix), and hence the factor $\|\mathbf{V}_{\mathbf{U}}\|_{2\to\infty}$ can be much smaller than $\max_{\mathbf{V}\in\mathbb{O}(p,2r)} \|\mathbf{V}\|_2 = 1$.

2.2 A continuous matrix shrinkage prior for joint sparsity

The recent decade has witnessed the development of a collection of continuous shrinkage priors that introduce sparse structures in various statistical contexts. For an incomplete list of references, see Bhattacharya et al. (2015), Pati et al. (2014), Carvalho et al. (2010), Rockova and George (2018), Rockova and George (2016), Rockova (2018), Shin

et al. (2018), and Shin et al. (2020). In this section, we first illustrate the general Bayesian strategies in modeling the sparsity occurring in high-dimensional models and then elaborate on the proposed prior model. Consider a simple yet canonical sparse normal mean problem. Suppose we observe independent normal data $y_i \sim \mathrm{N}(\beta_i, 1)$, $i = 1, \ldots, n$ and want to estimate the mean vector $\boldsymbol{\beta}_n = (\beta_i)_{i=1}^n$, which is assumed to be sparse in the sense that $\sum_{i=1}^n \mathbb{1}(|\beta_i| \neq 0) \leq s_n$ with the sparsity level $s_n = o(n)$ as $n \to \infty$. To model the sparsity of $\boldsymbol{\beta}$, classical Bayesian methods impose the following spike-and-slab prior on $\boldsymbol{\beta}$: for any measurable set $A \subset \mathbb{R}$,

$$\begin{aligned} \Pi(\beta_i \in A \mid \lambda, \xi_i) &= (1 - \xi_i)\delta_0(A) + \xi_i \int_A \psi(\beta \mid \lambda)\mathrm{d}\beta, \\ (\xi_i \mid \theta) &\sim \text{Bernoulli}(\theta), \end{aligned} \tag{2.4}$$

where ξ_i is the indicator that $\beta_i \neq 0$, $\theta \in (0, 1)$ represents the prior probability of β_i being non-zero, δ_0 is the point-mass at 0 (called the "spike" distribution), and $\psi(\cdot \mid \lambda)$ is the density of an absolutely continuous distribution (called the "slab" distribution) with respect to the Lebesgue measure on $\mathbb{R}$ governed by some hyperparameter λ. Theoretical justifications for the use of the spike-and-slab prior (2.4) for the sparse normal means and the sparse Bayesian factor models have been established in Castillo and van der Vaart (2012) and Pati et al. (2014), respectively. Therein, the spike-and-slab prior (2.4) involves point-mass mixtures, which can be daunting in terms of the posterior simulations (Pati et al., 2014). To address this issue, the spike-and-slab LASSO prior (Rockova, 2018) is designed as a continuous relaxation of (2.4):

$$\begin{aligned} \pi(\beta_i \mid \lambda_0, \lambda, \xi_i) &= (1 - \xi_i)\psi(\beta_i \mid \lambda_0) + \xi_i \psi(\beta_i \mid \lambda), \\ (\xi_i \mid \theta) &\sim \text{Bernoulli}(\theta), \end{aligned} \tag{2.5}$$

where $\psi(\beta \mid \lambda) = (\lambda/2)\exp(-\lambda|\beta|)$ is the Laplace distribution with mean 0 and variance $2/\lambda^2$. When $\lambda_0 \gg \lambda$, the spike-and-slab LASSO prior (2.5) closely resembles the spike-and-slab prior (2.4). The continuity feature of the spike-and-slab LASSO prior (2.5), in contrast to the classical spike-and-slab prior (2.4), is highly desired in high-dimensional settings in terms of the computational efficiency.

Motivated by the spike-and-slab LASSO prior, we propose a continuous matrix shrinkage prior to model the joint sparsity in the sparse spiked covariance models (1.1). The orthonormal constraint on the columns of $\mathbf{U}$ makes it challenging to incorporate prior distributions. Instead, we consider the following reparameterization of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \left(\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}}\right)\left(\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}}\right)^{\mathrm{T}} + \sigma^2\mathbf{I}_p = \mathbf{B}\mathbf{B}^{\mathrm{T}} + \sigma^2\mathbf{I}_p, \tag{2.6}$$

where $\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{p\times r}$, and $\mathbf{V} \in \mathbb{O}(r)$ is an arbitrary orthogonal matrix. Clearly, in contrast to the orthonormal constraint on $\mathbf{U}$, there is no constraint on $\mathbf{B}$ except that $\mathrm{rank}(\mathbf{B}) = r$. Furthermore, $\mathbf{B}$ inherits the joint sparsity from $\mathbf{U}$: For $|\mathrm{supp}(\mathbf{U})| = s \geq r$, there exists some permutation matrix $\mathbf{P} \in \mathbb{R}^{p\times p}$ and $\mathbf{U}^\star \in \mathbb{O}(s, r)$, such that

$$\mathbf{U} = \mathbf{P}\begin{bmatrix} \mathbf{U}^\star \\ \mathbf{0}_{(p-s)\times r} \end{bmatrix}.$$

It follows directly that

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}} = \mathbf{P}\begin{bmatrix}\mathbf{U}^{\star}\\ \mathbf{0}_{(p-s)\times r}\end{bmatrix}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}} = \mathbf{P}\begin{bmatrix}\mathbf{U}^{\star}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}}\\ \mathbf{0}_{(p-s)\times r}\end{bmatrix},$$

implying that $|\mathrm{supp}(\mathbf{B})| \leq s$. Therefore, working with $\mathbf{B}$ allows us to circumvent the orthonormal constraint while maintaining the jointly sparse structure of $\mathbf{U}$.

We are now in a position to formalize the proposed continuous matrix shrinkage prior on $\mathbf{B} = [b_{jk}]_{p\times r}$. Here we assume that the rank r of $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$ is known. When r is unknown, one can apply the diagonal thresholding method in Cai et al. (2013) to estimate r consistently. For any $\alpha, \lambda > 0$, let $\psi_\alpha(x \mid \lambda)$ be the density function of the following double Gamma distribution:

$$\psi_\alpha(x \mid \lambda) = \frac{\lambda^{1/\alpha}}{2\Gamma(1/\alpha)}|x|^{1/\alpha-1}\exp(-\lambda|x|), \quad -\infty < x < \infty.$$

For each $j \in [p]$, we assign the following hierarchical prior distribution on $\mathbf{B}_{1*}, \ldots, \mathbf{B}_{p*}$:

$$\begin{aligned}(\mathbf{B}_{1*}, \ldots, \mathbf{B}_{p*} \mid \lambda_0, \theta) &\overset{\text{i.i.d.}}{\sim} (1-\theta)\prod_{k=1}^{r}\psi_r(b_{jk} \mid \lambda+\lambda_0) + \theta\prod_{k=1}^{r}\psi_1(b_{jk} \mid \lambda),\\ \lambda_0 &\sim \mathrm{IG}(1/p^2, 1),\\ \theta &\sim \mathrm{Beta}\left(1, p^{1+\kappa}\right),\end{aligned} \tag{2.7}$$

where $\mathrm{IG}(a, b)$ is the inverse Gamma distribution with density $\pi(\lambda_0) \propto \lambda_0^{-a-1}\exp(-b/\lambda_0)$, and $\kappa, \lambda > 0$ are tuning parameters. Clearly, with probability $1-\theta$, each row $\mathbf{B}_{j*}$ follows a multivariate distribution with density $\prod_{k=1}^{r}\psi_r(b_{jk} \mid \lambda+\lambda_0)$, and hence, $\|\mathbf{B}_{j*}\|_1 \sim \mathrm{Exp}(\lambda+\lambda_0)$ for each j with probability $1-\theta$. The exponential distribution $\mathrm{Exp}(\lambda+\lambda_0)$ with a large λ_0 shrinks the entire row $\mathbf{B}_{j*}$ towards zero and therefore promotes the joint sparsity on $\mathbf{B}$.

We refer to the hierarchical prior (2.7) on $\mathbf{B}$ as the matrix spike-and-slab LASSO prior and denote $\mathbf{B} \sim \mathrm{MSSL}_{p\times r}(\lambda, 1/p^2, p^{1+\kappa})$. The hyperparameter λ is fixed throughout. In the single-spike case ($r = 1$), we observe that $\psi_1(b_{jk} \mid \lambda) = (\lambda/2)\exp(-\lambda|b_{jk}|)$ reduces to the density function of the Laplace distribution, and hence the matrix spike-and-slab LASSO prior coincides with the spike-and-slab LASSO prior (Rockova, 2018). Here, $1-\theta$ represents the proportion of the rows B_{j*}'s that are close to zero. Therefore, $\theta \sim \mathrm{Beta}(1, p^{1+\kappa})$ indicates that the matrix spike-and-slab LASSO prior favors a large proportion of rows of $\mathbf{B}$ being close to $\mathbf{0}$. We also remark that the first term $\prod_{k=1}^{r}\psi_r(b_{jk} \mid \lambda+\lambda_0)$ closely resembles the "spike" component in the spike-and-slab distribution (2.4), whereas the second term $\prod_{k=1}^{r}\psi_1(b_{jk} \mid \lambda)$ is a multivariate generalization of the "slab" component in (2.4). We complete the prior specification by imposing $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$ for some $a_\sigma, b_\sigma > 0$ for the sake of conjugacy.

Lastly, we remark that the parameterization (2.6) of the spiked covariance models (1.1) has the following interpretation. The sampling model $\mathbf{y}_i \sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$ can be equivalently represented in terms of the latent factor model

$$\mathbf{y}_i = \mathbf{B}\mathbf{z}_i + \boldsymbol{\varepsilon}_i, \quad \mathbf{z}_i \sim \mathrm{N}_r(\mathbf{0}_r, \mathbf{I}_r), \quad \boldsymbol{\varepsilon}_i \sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2\mathbf{I}_p), \quad i = 1, \ldots, n, \tag{2.8}$$

where $\mathbf{z}_i$, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) r-dimensional latent factors, $\mathbf{B}$ is a $p \times r$ factor loading matrix, and $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$ are i.i.d. noisy vectors. Since $\mathbf{B}$ is also sparse by our earlier discussion, this formulation is related to the sparse Bayesian factor models presented in Bhattacharya and Dunson (2011) and Pati et al. (2014), the differences being the sparsity pattern of $\mathbf{B}$ and the prior specification. In addition, the latent factor formulation (2.8) is convenient for the posterior simulation through a Markov chain Monte Carlo (MCMC) sampler, as discussed in Section 3.1 of Bhattacharya and Dunson (2011).

3 Theoretical properties

3.1 Properties of the matrix spike-and-slab LASSO prior

The theoretical properties of the classical spike-and-slab LASSO prior have been partially explored by Rockova (2018) and Rockova and George (2018) in the context of the sparse linear models and the sparse normal means problems, respectively. It is not clear whether the properties of the spike-and-slab LASSO priors adapt to other statistical contexts, including the sparse spiked covariance matrix models, the high-dimensional multivariate regression (Bai and Ghosh, 2018), etc. In this subsection, we present a collection of theoretical properties of the matrix spike-and-slab LASSO prior that not only are useful for deriving posterior contraction under the spiked covariance matrix models but also may be of independent interest for other statistical tasks, *e.g.*, sparse Bayesian linear regression with multivariate response (Ning and Ghosal, 2018).

Let $\mathbf{B} \in \mathbb{R}^{p\times r}$ be a $p \times r$ matrix and let $\mathbf{B}_0 \in \mathbb{R}^{p\times r}$ be a jointly s-sparse $p \times r$ matrix with $r \leq s \leq p$. Here $\mathbf{B}_0$ corresponds to the underlying truth. In the sparse spiked covariance matrix models, $\mathbf{B}$ represents the scaled eigenvector matrix $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$ up to an orthonormal matrix in $\mathbb{O}(r)$, but for the sake of generality, we do not impose the statistical context in this subsection. A fundamental measure of goodness for various prior models with high dimensionality is the prior mass assignment on a small neighborhood around the true but unknown value of the parameter. This is referred to as the *prior concentration* in the literature of Bayes theory. Formally, we consider the prior probability of the non-centered ball $\{\|\mathbf{B} - \mathbf{B}_0\|_{\mathrm{F}} < \eta\}$ under the prior distribution for small values of η.

Lemma 3.1. *Suppose* $\mathbf{B} \sim \mathrm{MSSL}_{p\times r}(\lambda, 1/p^2, p^{1+\kappa})$ *for some fixed positive constants* λ *and* κ*, and* $\mathbf{B}_0 \in \mathbb{R}^{p\times r}$ *is jointly* s*-sparse, where* $1 \leq r \leq s \leq p/2$*. Then for small values of* $\eta \in (0, 1)$ *with* $\eta \geq 1/p^{\gamma}$ *for some* $\gamma > 0$*, it holds that*

$$\Pi\left(\|\mathbf{B} - \mathbf{B}_0\|_{\mathrm{F}} < \eta\right) \geq \exp\left[-C_1 \max\left\{\lambda^2 s \|\mathbf{B}_0\|^2_{2\to\infty}, sr\left|\log\frac{\lambda\eta}{\sqrt{sr}}\right|, s\log p\right\}\right]$$

for some absolute constant $C_1 > 0$.

We next formally characterize how the matrix spike-and-slab LASSO prior promotes the joint sparsity of $\mathbf{B}$ using a probabilistic argument. Unlike the classical spike-and-slab prior, which enables exact zeros in the mean vector with a positive probability,

the matrix spike-and-slab LASSO prior is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{p\times r}$, implying that $|\mathrm{supp}(\mathbf{B})| = p$ with prior probability one. Rather than forcing the elements of $\mathbf{B}$ to be exactly zero, the matrix spike-and-slab LASSO prior shrinks the elements of $\mathbf{B}$ toward zero. This behavior suggests the following generalization of the row support of a matrix $\mathbf{B}$: For $\delta > 0$ taken to be small, we define $\mathrm{supp}_\delta(\mathbf{B}) = \{j \in [p] : \|\mathbf{B}_{j*}\|_2 > \delta\}$. Namely, $\mathrm{supp}_\delta(\mathbf{B})$ consists of the row indices of $\mathbf{B}$ whose Euclidean norms are greater than δ. Intuitively, one should expect that under the matrix spike-and-slab LASSO prior, $|\mathrm{supp}_\delta(\mathbf{B})|$ should be small with large probability. The following lemma formally confirms this intuition.

Lemma 3.2. *Suppose* $\mathbf{B} \sim \mathrm{MSSL}_{p\times r}(\lambda, 1/p^2, p^{1+\kappa})$ *for some fixed positive constants* λ *and* $\kappa \le 1$, $1 \le r \le p$. *Let* $\delta \in (0,1)$ *be a small number with* $\delta > 1/p^\gamma$ *for some* $\gamma > 0$, *and let* s *be an integer such that* $(s\log p)/p$ *is sufficiently small. Then for any* $\beta > 4\gamma\exp(1)$, *it holds that*

$$\Pi\left(|\mathrm{supp}_\delta(\mathbf{B})| > \beta s\right) \le 2\exp\left\{-\min\left(\frac{\beta\kappa}{2}, \frac{\beta}{2\mathrm{e}} - 2\gamma\right) s\log p\right\}.$$

Lastly, we provide the following tail probability inequality for the matrix spike-and-slab LASSO prior.

Lemma 3.3. *Suppose* $\mathbf{B} \sim \mathrm{MSSL}_{p\times r}(\lambda, 1/p^2, p^{1+\kappa})$ *for some fixed positive* λ *and* $\kappa < 1$, *and* $\mathbf{B}_0 \in \mathbb{R}^{p\times r}$ *is jointly* s*-sparse, where* $r\log n \lesssim \log p$, *and* $(s\log p)/p$ *is sufficiently small. Let* $(\delta_n)_{n=1}^\infty$ *and* $(t_n)_{n=1}^\infty$ *be positive sequences such that* $1/p^\gamma \le \delta_n \le 1$ *and* $t_n/(sr) \to \infty$. *Then for sufficiently large* n *and for all* $\beta > 4\gamma\exp(1)$, *it holds that*

$$\begin{aligned}&\Pi\left[\sum_{j=1}^p \|\mathbf{B}_{j*}\|_1 \mathbb{1}\{j \in \mathrm{supp}_{\delta_n}(\mathbf{B}) \cup \mathrm{supp}(\mathbf{B}_0)\} \ge t_n\right] \\ &\quad\le 2\exp\left[-C_2\min\left\{\left(\frac{t_n}{\beta sr}\right)^2, \left(\frac{t_n}{r}\right)^2, \frac{t_n}{r}\right\}\right] + 3\exp\left\{-\min\left(\frac{\beta\kappa}{2}, \frac{\beta}{2\mathrm{e}} - 2\gamma\right) s\log p\right\}\end{aligned}$$

for some absolute constant $C_2 > 0$.

3.2 Posterior contraction for the sparse Bayesian spiked covariance model

We now elaborate on the posterior contraction rates for the proposed Bayesian sparse spiked covariance models with respect to various loss functions, which are the main results of this paper. We first present a collection of necessary assumptions.

Assumption 3.1 (Joint sparsity)**.** $|\mathrm{supp}(\mathbf{U}_0)| \le s$ and $1 \le r \le s \le p$.

Assumption 3.2 (Bounded spectra)**.** λ_{01} and λ_{0r} are bounded away from 0 and ∞.

Assumption 3.3 (High-dimensionality and consistency)**.** $p/n \to \infty$, $(s\log p)/n \to 0$.

Assumption 3.4 (Low-rank assumption)**.** $r\log n \lesssim \log p$.

Assumption 3.5 (Prior specification)**.** $\mathbf{B} \sim \mathrm{MSSL}_{p\times r}(\lambda, 1/p^2, p^{1+\kappa})$ with some $\lambda > 0$ and $\kappa \leq 1$, and $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$ for some $a_\sigma, b_\sigma \geq 1$.

Remark 3.1. Several remarks concerning the assumptions above are in order. Assumptions 3.1 and 3.2 are standard for the sparse spiked covariance models. Assumption 3.3 states the high-dimensional nature of the problem (namely, a large p small n scenario) and ensures the consistency with regard to the spectral norm $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_2$. Assumption 3.4 is a mild low-rank condition that also appeared in Gao and Zhou (2015) and Ning (2021). Loosely speaking, it guarantees that the posterior contraction rate under the intrinsic metric of the normal covariance model is equivalent to the spectral norm. Assumption 3.5 specifies the prior distribution of $\boldsymbol{\Sigma}$ through the priors of $\mathbf{B}$ and σ^2.

Below, Theorem 3.1 and Theorem 3.2 state that the posterior contraction rates with respect to $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_2$ and $\|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$ are minimax-optimal, respectively.

Theorem 3.1. *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n \sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_0)$ *independently with* $\boldsymbol{\Sigma}_0 = \mathbf{U}_0\boldsymbol{\Lambda}_0\mathbf{U}_0^{\mathrm{T}} + \sigma_0^2\mathbf{I}_p$ *and assume Assumptions 3.1–3.5 hold. Then there exists some constants* $M_0, R_0, C_0 > 0$ *depending on* $\sigma_0, \lambda_{01}, \lambda_{0r}$*, and the hyperparameters, such that*

$$\mathbb{E}_0\left\{\Pi\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_2 > M\sqrt{\frac{s\log p}{n}} \;\middle|\; \mathbf{Y}_n\right)\right\} \leq R_0\exp(-C_0 s\log p) \tag{3.1}$$

for all $M \geq M_0$ *when* n *is sufficiently large.*

Theorem 3.2. *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n \sim \mathrm{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_0)$ *independently with* $\boldsymbol{\Sigma}_0 = \mathbf{U}_0\boldsymbol{\Lambda}_0\mathbf{U}_0^{\mathrm{T}} + \sigma_0^2\mathbf{I}_p$ *and assume Assumptions 3.1–3.5 hold. Let* $M_0, R_0, C_0 > 0$ *be the constants given by Theorem 3.1. For each* $\mathbf{B}$*, let* $\mathbf{U}_{\mathbf{B}} \in \mathbb{O}(p, r)$ *be the left-singular vector matrix of* $\mathbf{B}$*. Then the following holds for all* $M \geq M_0$ *and sufficiently large* n*:*

$$\mathbb{E}_0\left\{\Pi\left(\|\mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2 > \frac{2M}{\lambda_{0r}}\sqrt{\frac{s\log p}{n}} \;\middle|\; \mathbf{Y}_n\right)\right\} \leq R_0\exp(-C_0 s\log p). \tag{3.2}$$

Remark 3.2. We briefly compare the posterior contraction rates obtained in Theorem 3.1 and Theorem 3.2 with some related results in the literature. In Pati et al. (2014), the authors consider the posterior contraction with respect to the spectral norm loss $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0\|_2$ of the entire covariance matrix, while in Gao and Zhou (2015), the authors consider the posterior contraction with respect to the projection Frobenius norm loss $\|\mathbf{U}\mathbf{U}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_{\mathrm{F}}$ for estimating $\mathrm{Span}\{\mathbf{U}_{*1}, \ldots, \mathbf{U}_{*r}\}$. Our result is similar to Ning (2021), who derive the posterior contraction rates with a spike-and-slab prior on the entries of the loading matrix $\mathbf{B}$. This differs from the current work as we adopt a continuous matrix shrinkage prior. In Pati et al. (2014), the notion of sparsity is slightly different than the joint sparsity notion presented here. They assume that under the latent factor model representation (2.8), the individual supports of columns of $\mathbf{B}$ are not necessarily the same. When $r = O(1)$, the assumption in Pati et al. (2014) coincides the joint sparsity and our rate $\epsilon_n = \sqrt{(s\log p)/n}$ improves the rate $\sqrt{(s\log p\log n)/n}$ obtained in Pati et al. (2014) by a logarithmic factor. The assumptions in Gao and Zhou (2015) are the same as those in Pati et al. (2014), and Gao and Zhou (2015) focus on

designing a prior that yields the rate-optimal posterior contraction with respect to the Frobenius norm loss of the projection matrices as well as the adaptation to the sparsity level s and the rank r. Our result in equation (3.2), which focuses on the projection spectral norm loss, serves as a complement to the rate-optimal posterior contraction for the principal subspace under the joint sparsity assumption in contrast to Gao and Zhou (2015).

The posterior contraction rate (3.2) also leads to the following risk bound for a point estimator of the principal subspace $\mathrm{Span}\{\mathbf{U}_{*1}, \ldots, \mathbf{U}_{*r}\}$:

Theorem 3.3. *Assume the conditions in Theorem 3.1 hold. Let*

$$\widehat{\mathbf{\Omega}} = \int \mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}}\Pi(\mathrm{d}\mathbf{B} \mid \mathbf{Y}_n)$$

be the posterior mean of the projection matrix $\mathbf{U}_{\mathbf{B}}\mathbf{U}_{\mathbf{B}}^{\mathrm{T}}$ and $\widehat{\mathbf{U}} \in \mathbb{O}(p, r)$ be the orthonormal r-frame in $\mathbb{R}^p$ with columns being the eigenvectors corresponding to the first r-largest eigenvalues of $\widehat{\mathbf{\Omega}}$. Let M_0, R_0 be the constants in Theorem 3.1. Then the following risk bound holds for $\widehat{\mathbf{U}}$ for sufficiently large n:

$$\mathbb{E}_0\left(\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2\right) \leq \left(\frac{4M_0}{\lambda_{0r}} + 4\sqrt{R_0}\right)\sqrt{\frac{s\log p}{n}}.$$

To derive the posterior contraction rate for the principal subspace with respect to the two-to-infinity norm loss, we need the posterior contraction result for $\mathbf{\Sigma}$ with respect to the stronger matrix infinity norm. This contraction rate is obtained in the following lemma.

Lemma 3.4. *Assume the conditions in Theorem 3.1 hold. Further assume that the eigenvector matrix $\mathbf{U}_0$ exhibits the bounded coherence: $\|\mathbf{U}_0\|_{2\to\infty} \leq C_\mu\sqrt{r/s}$ for some constant $C_\mu \geq 1$, and the number of spikes r is sufficiently small in the sense that $r^3/s = O(1)$. Let R_0, C_0 be the constants in Theorem 3.1. Then there exists some constant $M_\infty > 0$ depending on σ_0, $\lambda_{01}, \lambda_{0r}$, and the hyperparameters, such that the following posterior contraction for $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}} + \sigma^2\mathbf{I}_p$ holds for all $M \geq M_\infty$ when n is sufficiently large:*

$$\mathbb{E}_0\left\{\Pi\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_\infty > Mr\sqrt{\frac{s\log p}{n}} \;\middle|\; \mathbf{Y}_n\right)\right\} \leq R_0\exp(-C_0 s\log p). \tag{3.3}$$

We are now in a position to present the posterior contraction of the principal subspace with regard to the two-to-infinity norm loss in Theorem 3.4 below.

Theorem 3.4. *Assume the conditions in Lemma 3.4 hold. Let R_0, C_0 be the constants in Theorem 3.1. For each $\mathbf{B}$, let $\mathbf{U}_{\mathbf{B}} \in \mathbb{O}(p, r)$ be the left-singular vector matrix of $\mathbf{B}$. Then there exists some large constant $M_{2\to\infty} > 0$ depending on σ_0, $\lambda_{01}, \lambda_{0r}$, and the hyperparameters, such that the following posterior contraction for $\mathbf{U}_{\mathbf{B}}$ holds for all $M \geq$*

$M_{2\to\infty}$:

$$\mathbb{E}_0\left[\Pi\left\{\|\mathbf{U_B}-\mathbf{U}_0\mathbf{W_U}\|_{2\to\infty} > M\left(\sqrt{\frac{r^3\log p}{n}}\vee\frac{s\log p}{n}\right)\right\}\right] \le 2R_0\exp(-C_0 s\log p), \tag{3.4}$$

where $\mathbf{W_U} = \arg\inf_{\mathbf{W}\in\mathbb{O}(r)}\|\mathbf{U_B}-\mathbf{U}_0\mathbf{W}\|_{\mathrm{F}}$ *is the Frobenius orthogonal alignment matrix.*

Remark 3.3. The bounded coherence condition that $\|\mathbf{U}_0\|_{2\to\infty}\le C_\mu\sqrt{r/s}$ originates from the delocalization of eigenvectors in low-rank matrix recovery (Candès and Recht, 2009). Loosely speaking, when $\mathbf{U}_0$ is a tall-and-thin rectangular matrix, the bounded coherence of $\mathbf{U}_0$ states that the orthonormal columns of $\mathbf{U}_0$ are different from the highly localized standard basis vectors. The bounded coherence condition is also related to the pervasive assumption appearing in the econometrics and financial applications (Fan et al., 2008; Pati et al., 2014). Specifically, $\mathbf{U}_0$ represents the left singular vector matrix of the factor loading matrix $\mathbf{B}_0 := \mathbf{U}_0\mathbf{\Lambda}_0^{1/2}\mathbf{V}^{\mathrm{T}}$ under the latent factor model representation (2.8) for some $\mathbf{V}\in\mathbb{O}(r)$. By the random matrix theory, $\mathbf{B}_0$ is pervasive when the non-zero rows of $\mathbf{B}_0$ are random realizations from a bounded random vector (Fan et al., 2013). By Proposition 3 in Fan et al. (2016), $\mathbf{U}_0$ satisfies the bounded coherence condition when $\mathbf{B}_0$ is pervasive. In addition, the assumption that $r^3/s = O(1)$ also appeared in the (dense) covariance estimation problem in Cape et al. (2019b).

Remark 3.4. We also present some remarks concerning the posterior contraction with respect to the two-to-infinity norm loss $\|\mathbf{U}-\mathbf{U}_0\mathbf{W_U}\|_{2\to\infty}$. Cape et al. (2019b) show that

$$\|\mathbf{U}-\mathbf{U}_0\mathbf{W_U}\|_{2\to\infty}\le\|\mathbf{U}-\mathbf{U}_0\mathbf{W_U}\|_2\le\sqrt{2}\|\mathbf{U}\mathbf{U}^{\mathrm{T}}-\mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2,$$

meaning that $\|\mathbf{U}-\mathbf{U}_0\mathbf{W_U}\|_{2\to\infty}$ can be coarsely upper bounded by the projection spectral norm loss $\|\mathbf{U}\mathbf{U}^{\mathrm{T}}-\mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$. This naive bound immediately yields

$$\mathbb{E}_0\left\{\Pi\left(\|\mathbf{U_B}-\mathbf{U}_0\mathbf{W_U}\|_{2\to\infty} > M\sqrt{\frac{s\log p}{n}}\;\middle|\;\mathbf{Y}_n\right)\right\}\le R_0\exp(-C_0 s\log p)$$

for some appropriately selected large constant M, which is the same as (3.2). Our result (3.4) improves this rate by a factor of $\{\sqrt{r^3/s}\vee\sqrt{(s\log p)/n}\}$, resulting in a tighter posterior contraction rate with respect to the two-to-infinity norm loss. In particular, when $r\ll s$ (*i.e.*, $\mathbf{U}_0$ is a "tall and thin" rectangular matrix), the factor $\sqrt{r^3/s}$ can be much smaller than 1.

4 Numerical examples

4.1 Synthetic examples

We evaluate the numerical performance of the proposed Bayesian method for estimating the sparse spiked covariance models via simulation studies. We set the sample size

$n = 100$ and the number of features $p = 200$. The support size s of the eigenvector matrix $\mathbf{U}_0$ ranges over $\{8, 12, 20, 40\}$ and the number of spikes r takes values in $\{1, 4\}$. The indices of the non-zero rows of $\mathbf{U}_0$ are uniformly sampled from $\{1, \ldots, p\}$ and we set the diagonal elements of $\mathbf{\Lambda}_0$ to be equally spaced over the interval $[10, 20]$, with $\lambda_{01} = 20$ and $\lambda_{0r} = 10$ if $r = 4$. The non-zero rows of $\mathbf{U}_0$, themselves forming an orthonormal r-frame in $\mathbb{R}^s$, denoted by $\mathbf{U}_0^\star$, are taken as the left singular vector matrix of a $s \times r$ matrix $\mathbf{L}$ whose entries are generated from $\mathrm{Unif}(1, 2)$ independently.

The posterior inference is carried out using a standard Metropolis-within-Gibbs sampler. We take the first 1000 iterations of the MCMC sampler as the burn-in phase and collect the subsequent 4000 iterations as the post-burn-in samples. We set $\lambda = 1$, $a_\sigma = b_\sigma = 1$, and $\kappa = 1$ in all numerical examples. The convergence diagnostics of the MCMC chains are provided in the Supplementary Material (Xie et al., 2022). We then take the posterior mean $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ as the point estimator for $\mathbf{\Sigma}$, and $\mathrm{Span}(\widehat{\mathbf{U}})$ given by Theorem 3.3 as the point estimator for the principal subspace $\mathrm{Span}\{\mathbf{U}_{*1}, \ldots, \mathbf{U}_{*r}\}$.

For comparison, several competitors are considered, including the sparse Bayesian factor model with the multiplicative Gamma process shrinkage prior (MGPS, Bhattacharya and Dunson, 2011), the principal orthogonal complement thresholding method (POET, Fan et al., 2013), the sparse principal component analysis (SPCA, Zou et al., 2006), and the adaptive sparse principal component analysis (ASPCA, Cai et al., 2013). In each simulation setup (*i.e.*, each (r, s) pair), 50 independent replicates of the synthetic datasets are generated. For each synthetic dataset, we compute the point estimators $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{U}}$ as well as those offered by the three competing approaches, the spectral norm loss ($\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_2$), the two-to-infinity norm loss ($\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty}$), and the projection spectral norm loss ($\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$). We then compute the medians of these losses across the 50 replicates. The exception here is ASPCA, which only provides a point estimator for the principal subspace rather than the whole covariance matrix. We only obtain the projection spectral norm loss and the two-to-infinity norm loss for the principal subspace for ASPCA. The results are tabulated in Table 1.

The numerical results in Tables 1(a) and 1(b) indicate that the proposed Bayesian approach yields the smallest spectral norm losses for $\mathbf{\Sigma}$ and the smallest projection spectral norm losses for the subspace estimation, respectively, among others, except for ASPCA. While ASPCA outperforms the proposed Bayesian method in terms of the projection spectral norm loss for small values of s ($s \in \{8, 12\}$) when $r = 4$, its performance deteriorates rapidly as soon as the number of non-zero rows of $\mathbf{U}_0$ increases ($s \in \{20, 40\}$) when $r = 1$. In terms of the two-to-infinity norm loss for the subspace estimation, Table 1(c) shows that the point estimates $\widehat{\mathbf{U}}$ using the proposed approach yield smaller losses compared to the competitors when $s \in \{8, 12, 20\}$ for both $r = 1$ and $r = 4$, while POET is more accurate for the single-spike cases when $s = 40$. The comparison between the two losses for the subspace estimation is also visualized in Figure 1, suggesting that the two-to-infinity norm loss is less sensitive to the row support size s than the projection spectral norm loss as s increases.

We further evaluate the estimation performance for the principal subspace when $s = 20$, $r = 1$ and $s = 40$, $r = 4$ through a single replicate in Figures 2, 3, and 4, respectively. For the visualization of recovering $\mathbf{U}_0$ across different methods, we rotate

(a) The spectral norm loss $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_2$

s	8		12		20		40	
r	1	4	1	4	1	4	1	4
MSSL	**1.81**	**6.48**	**2.06**	**6.37**	**2.80**	**7.55**	**4.63**	**9.62**
MGPS	9.86	16.63	9.88	17.63	9.88	18.56	9.89	19.04
POET	7.54	11.17	7.47	11.10	7.61	11.60	7.60	10.97
SPCA	8.08	18.03	8.09	18.04	8.11	18.07	8.17	18.10

(b) The squared projection spectral norm loss $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2^2$

s	8		12		20		40	
r	1	4	1	4	1	4	1	4
MSSL	**0.010**	0.028	**0.014**	0.032	**0.029**	**0.038**	**0.10**	**0.060**
MGPS	0.18	0.29	0.19	0.33	0.18	0.29	0.19	0.22
POET	0.18	0.21	0.18	0.20	0.19	0.20	0.18	0.20
SPCA	0.05	0.092	0.068	0.11	0.10	0.15	0.18	0.22
ASPCA	0.022	**0.015**	0.083	**0.027**	0.24	0.057	0.82	0.11

(c) The squared two-to-infinity norm loss $\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty}^2$

s	8		12		20		40	
r	1	4	1	4	1	4	1	4
MSSL	**0.0037**	**0.012**	**0.0044**	**0.011**	**0.0061**	**0.010**	0.017	**0.011**
MGPS	0.0088	0.12	0.0088	0.085	0.0082	0.080	0.0088	0.055
POET	0.0082	0.013	0.0082	0.013	0.0086	0.012	**0.0088**	0.013
SPCA	0.024	0.027	0.022	0.040	0.022	0.039	0.025	0.038
ASPCA	0.0079	0.0067	0.048	0.012	0.046	0.018	0.090	0.029

Table 1: Loss functions for the simulation study: The spectral norm loss $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}_0\|_2$, the squared projection spectral norm loss $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2^2$, and the squared two-to-infinity norm loss $\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty}^2$. The medians of the loss function values across 50 replicates of synthetic datasets are tabulated. MSSL stands for the sparse Bayesian spiked covariance matrix model with the matrix spike-and-slab LASSO prior; MGPS refers the sparse Bayesian factor model with the multiplicative Gamma process shrinkage prior; POET refers to the principal orthogonal complement thresholding method; SPCA refers to the sparse principal component analysis; ASPCA refers to the adaptive sparse principal component analysis.

the estimates according to the Frobenius orthogonal alignment. To be more specific, for a point estimator $\widehat{\mathbf{U}}$ obtained using a frequentist method, we first compute the orthogonal alignment matrix $\mathbf{W}_{\widehat{\mathbf{U}}} = \arg\inf \mathbf{W} \in \mathbb{O}(r)\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}\|_{\mathrm{F}}$ and then use $\widehat{\mathbf{U}}\mathbf{W}_{\widehat{\mathbf{U}}}^{\mathrm{T}}$ as the estimator for $\mathbf{U}_0$. For the Bayesian method, we compute the orthogonal alignment $\mathbf{W}_{\mathbf{U_B}} = \arg\inf_{\mathbf{W}\in\mathbb{O}(r)}\|\mathbf{U_B} - \mathbf{U}_0\mathbf{W}\|_{\mathrm{F}}$ for each posterior sample $\mathbf{U_B}$ and then take $\mathbf{U_B}\mathbf{W}_{\mathbf{U_B}}^{\mathrm{T}}$ as the posterior estimate for $\mathbf{U}_0$. It can clearly be seen that POET is able to capture the signal but fails to recover the row support of the principal subspace, whereas SPCA is able to recover the subspace support but is not accurate in estimating

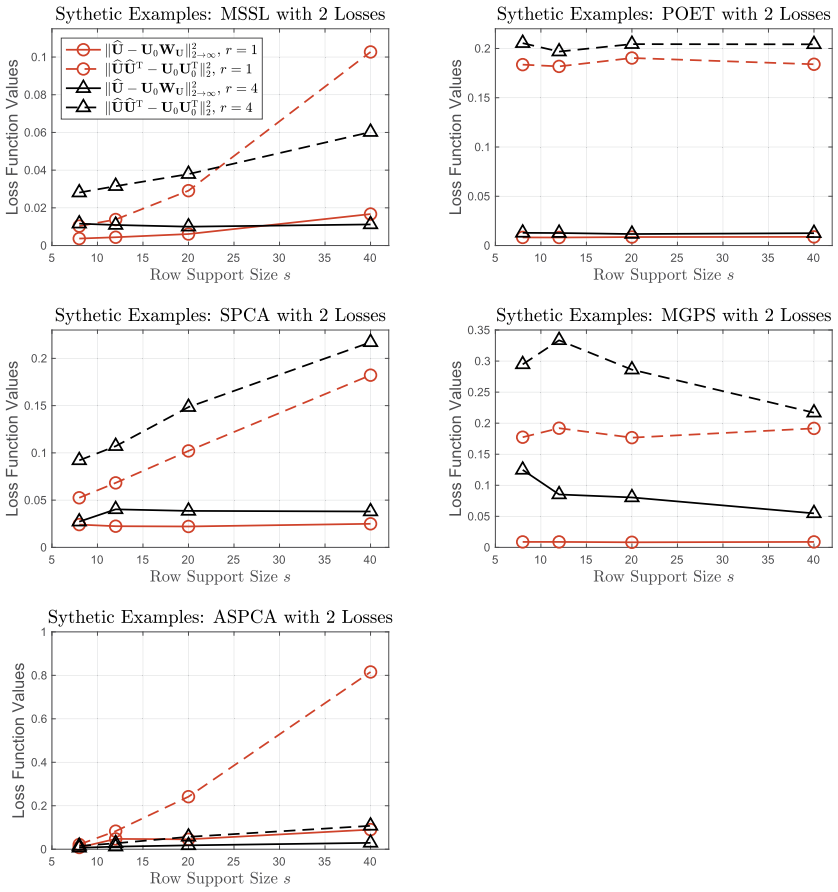

Figure 1: Comparison of the two-to-infinity norm loss ($\|\widehat{\mathbf{U}} - \mathbf{U}_0\mathbf{W}_{\mathbf{U}}\|_{2\to\infty}$) and the projection spectral norm loss ($\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}} - \mathbf{U}_0\mathbf{U}_0^{\mathrm{T}}\|_2$) for the synthetic examples. MSSL refers to the posterior means of $\mathbf{\Sigma}$ under the matrix spike-and-slab LASSO prior; MGPS refers the sparse Bayesian factor model with the multiplicative Gamma process shrinkage prior; POET refers to the principal orthogonal complement thresholding method; SPCA refers to the sparse principal component analysis; ASPCA refers to the adaptive sparse principal component analysis.

the signal. MGPS performs similarly to POET but results in wider credible intervals than those using the proposed approach. The performance of ASPCA is satisfactory when $r = 4$ but it severely under-estimates the number of non-zero rows of $\mathbf{U}_0$ when $r = 1$ and results in unsatisfactory estimates.

In addition, we report the running times of the five methods using one synthetic dataset for each setup in Table 2. The proposed MSSL is faster than the other Bayesian method, MGPS. Although POET is significantly faster than the proposed approach and

s	8		12		20		40	
r	1	4	1	4	1	4	1	4
MSSL	6.79 s	12.35 s	6.55 s	12.41 s	6.93 s	13.82 s	7.81 s	13.07 s
MGPS	8.82 s	26.26 s	8.83 s	25.59 s	8.27 s	25.60 s	7.99 s	25.34 s
POET	2.14 s	1.83 s	1.75 s	1.77 s	1.78 s	1.82 s	1.82 s	1.83 s
SPCA	0.18 s	3.76 s	0.08 s	5.56 s	0.11 s	8.60 s	0.14 s	17.21 s
ASPCA	0.01 s	0.008 s	0.005 s	0.004 s	0.006 s	0.004 s	0.006 s	0.005 s

Table 2: Runtime comparison for the simulation study. MSSL refers to the posterior means of $\mathbf{\Sigma}$ under the matrix spike-and-slab LASSO prior; MGPS refers the sparse Bayesian factor model with the multiplicative Gamma process shrinkage prior; POET refers to the principal orthogonal complement thresholding method; SPCA refers to the sparse principal component analysis; ASPCA refers to the adaptive sparse principal component analysis.

MGPS when $r = 4$, and sparse PCA outperforms all the other methods when $r = 1$, we will see in Section 4.2 that when the dimension p becomes large, POET and sparse PCA fail to produce results within 20 hours. ASPCA is the fastest approach among all methods considered here. However, it is not stable when $r = 1$, as presented in Table 1 and Figure 2. Overall, the proposed sparse Bayesian spiked covariance model is able to estimate the signals accurately and efficiently, recover the row support of $\mathbf{U}_0$, and provide better uncertainty quantification with narrower credible intervals for the synthetic datasets.

4.2 A face data example

The joint sparsity of the eigenvector matrix $\mathbf{U}$ is often desired in the feature extraction for some high-dimensional data. In this subsection, we illustrate how the proposed Bayesian approach is able to extract the key features through a real data example in computer vision.

We consider a subset of the Extended Yale Face Database B (Georghiades et al., 2001; Lee et al., 2005). It consists of the face images for 38 subjects, and for each subject, 64 aligned images of size 192×168 are taken under different illumination conditions. Here we focus on the 22nd subject and reduce the size of each image to 96×84 (8064 pixels in total), following the preprocessing step in She (2017). In doing so, we obtain a data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^{\mathrm{T}}$ of size 64×8064.

In computer vision, the principal component analysis has been broadly applied to obtain the low-dimensional features, known as the eigenfaces, from high-dimensional face image data. Under the proposed Bayesian framework, we perform the posterior inference by implementing a Metropolis-within-Gibbs sampler with 1000 burn-in iterations and 4000 post-burn-in samples. The number of spikes r is estimated using the diagonal thresholding method proposed in Cai et al. (2013). For comparison, we also implement MGPS (Bhattacharya and Dunson, 2011). Instead of obtaining the eigenfaces, we focus on the extraction of the key pixels via thresholding the obtained estimated eigenvector

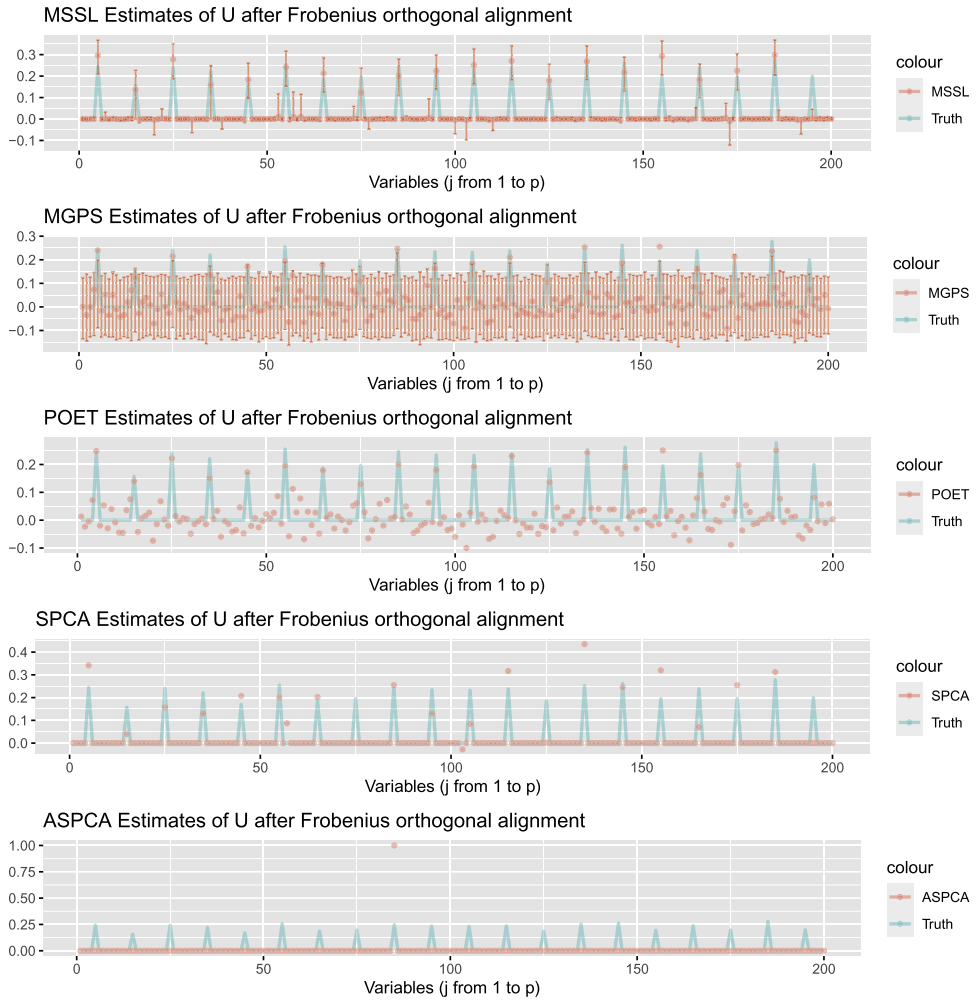

Figure 2: Simulation performance for a single replicate with $s = 20$ and $r = 1$. The estimates are rotated to the simulation truth $\mathbf{U}_0$ according to the Frobenius orthogonal alignment. The red bars in the top two panels are estimated 95% credible intervals using the proposed approach and MGPS, respectively.

matrix $\widehat{\mathbf{U}}$. Specifically, for the proposed approach, the estimate $\widehat{\mathbf{U}}$ can be computed according to Theorem 3.3, and for MGPS, $\widehat{\mathbf{U}}$ can be obtained by computing the left singular vectors of the loading matrix. The key pixels are then obtained by finding $\{j \in [8064] : \|\widehat{\mathbf{U}}_{j*}\|_1/r > \tau\}$ for some small tolerance $\tau > 0$. The other three competitors are unable to produce valid results for the following reasons. Neither POET nor sparse PCA was able to produce results within 20 hours. ASPCA encountered numerical instability when we attempted to implement it and reported errors.

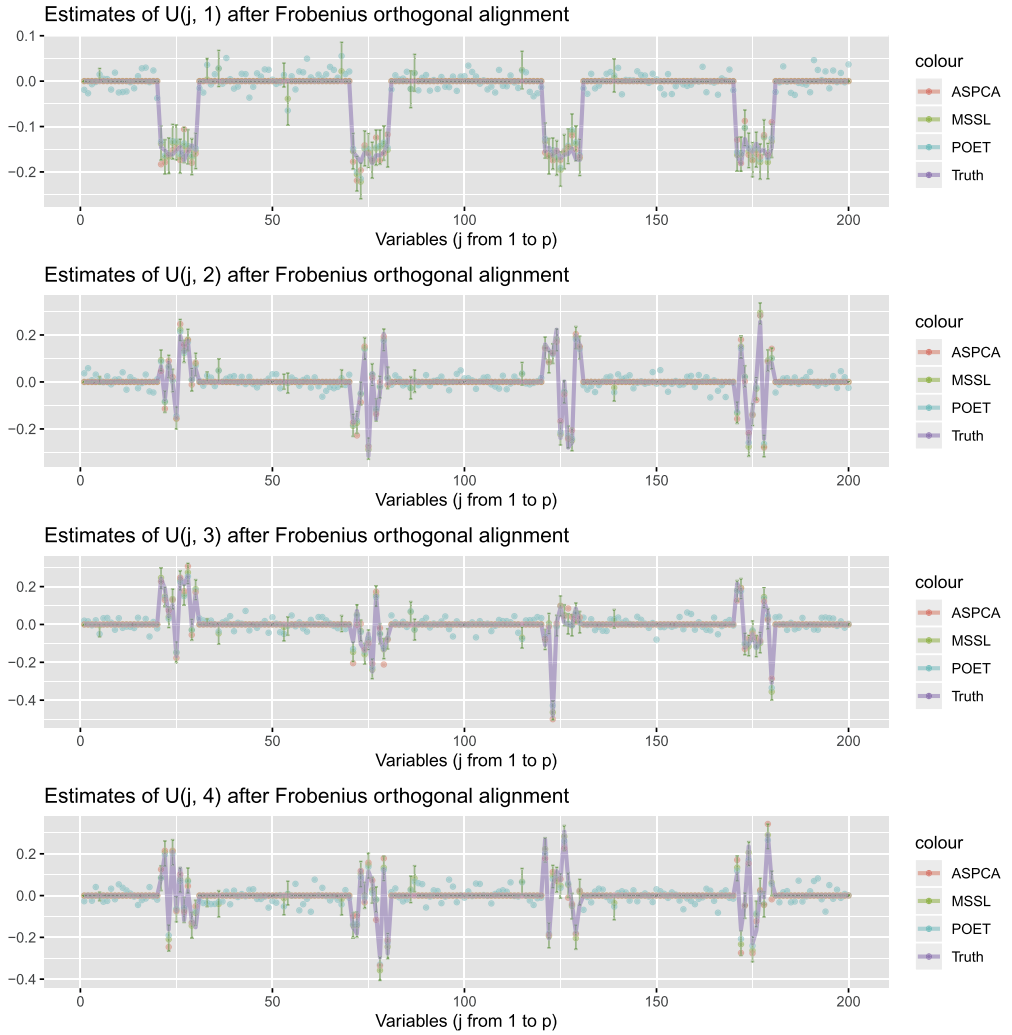

Figure 3: Simulation performance for a single replicate with $s = 40$ and $r = 4$. The estimates are rotated to the simulation truth $\mathbf{U}_0$ according to the Frobenius orthogonal alignment. The green bars in the four panels are estimated 95% credible intervals using the proposed approach.

We present the sample images of the 22nd subject in the first row of Figure 5. The key pixels of sample image #1 extracted using the two models with different threshold values of τ are provided in the second and the third rows of Figure 5. We recover the pixels with higher values (corresponding to eyes, lips, and nose tips of the subject) using the proposed model and MGPS. This observation is also in accordance with the conclusion from She (2017). Nevertheless, as the threshold value τ increases, the

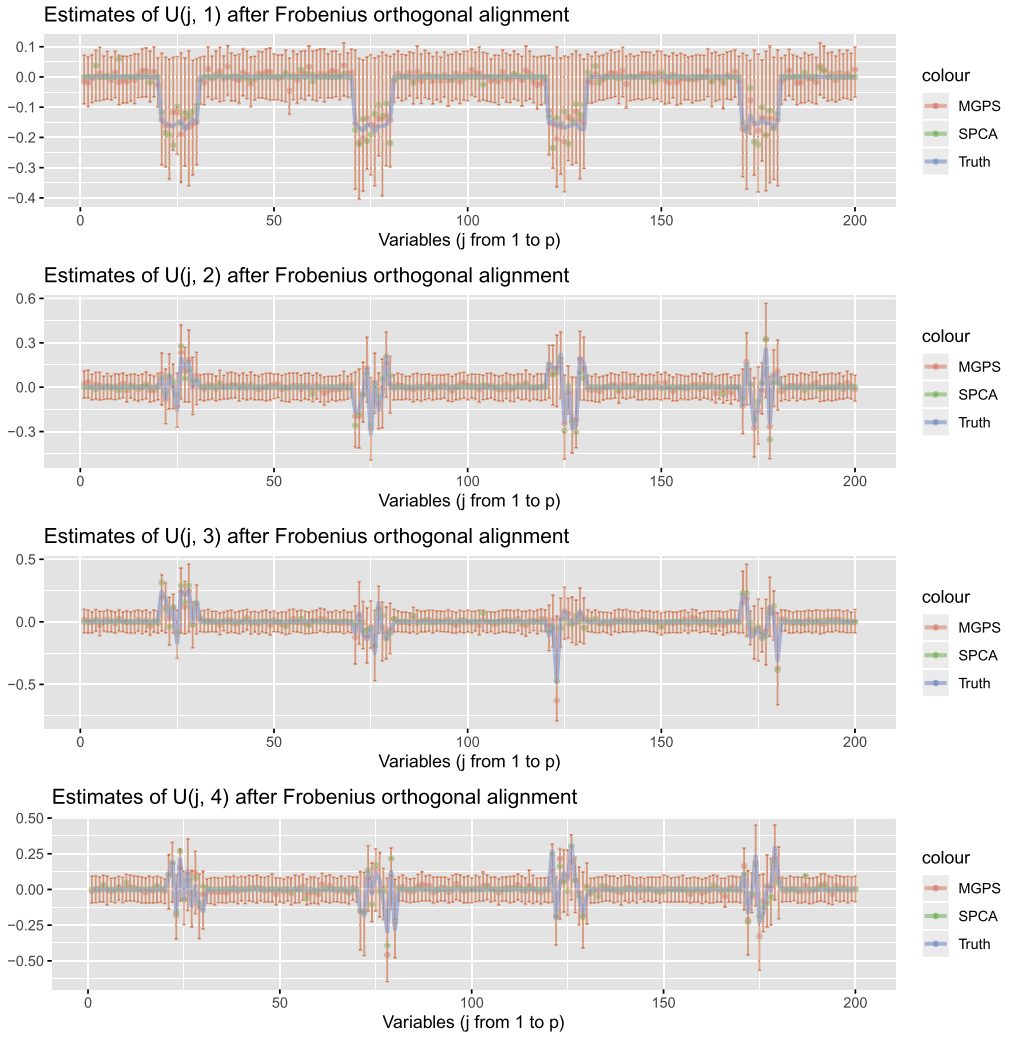

Figure 4: Simulation performance for a single replicate with $s = 40$ and $r = 4$. The estimates are rotated to the simulation truth $\mathbf{U}_0$ according to the Frobenius orthogonal alignment. The red bars in the four panels are estimated 95% credible intervals for MGPS.

number of key pixels captured using MGPS decreases significantly, whereas the proposed approach is more robust to the threshold value τ and maintains the key pixels that are sensitive to illumination. This phenomenon is expected since, unlike the matrix spike-and-slab LASSO prior, MGPS is not designed to model the joint sparsity and the feature extraction but rather column-specific sparsity for each individual factor loading.

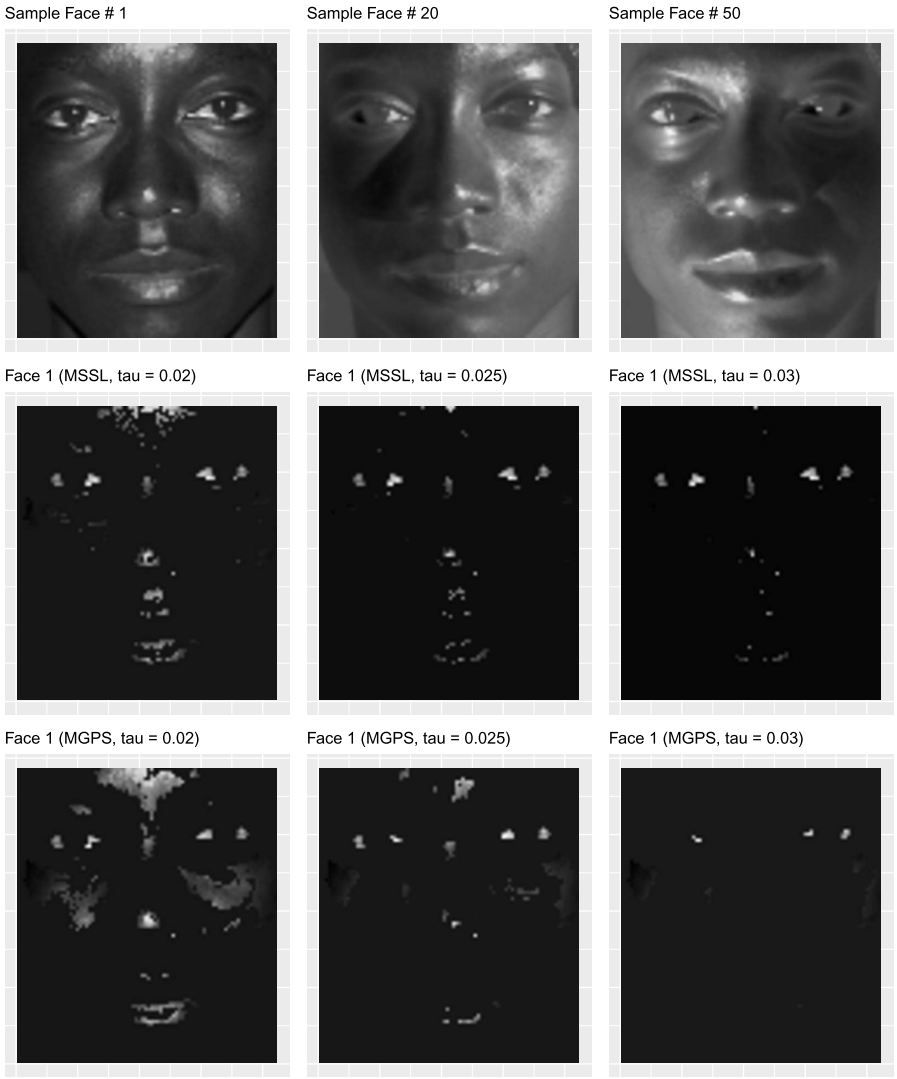

Figure 5: The face data example: The first row corresponds to sample images of the 22nd subject (image number 1, 20, and 50, respectively). The second and the third rows are the key pixels of the #1 image using the proposed Bayesian approach with the matrix spike-and-slab LASSO prior (MSSL) and MGPS with different threshold values τ.

5 Discussion

We have shown that the two-to-infinity norm loss for the principal subspace estimation can capture the entrywise perturbations of the eigenvector matrix $\mathbf{U}$ in contrast to the routinely used projection spectral norm loss. A novel matrix shrinkage prior that

extends the continuous spike-and-slab LASSO due to Rockova and George (2018) and Rockova (2018) has been developed. We have obtained the contraction rate of the full posterior distribution for the principal subspace under the two-to-infinity norm loss, which is sharper than the rate under the usual projection spectral norm loss, provided that $\mathbf{U}$ exhibits certain low-rank and bounded coherence conditions.

In future work, we intend to study whether a point estimator can be found from the posterior distribution with a risk bound that coincides with the posterior contraction rate under the two-to-infinity norm loss. In addition, it is also worth exploring the minimax-optimal rates of convergence with respect to the two-to-infinity norm loss. Throughout the paper, we assume that the number of spikes r is known. When r is unknown, a convenient approach is to estimate r using a frequentist method (e.g., the diagonal thresholding method as in Cai et al., 2013) first and then apply our Bayesian method using the estimated r. Alternatively, it is feasible to adaptively estimate r in the literature of Bayesian latent factor models (see, for example, Bhattacharya and Dunson, 2011; Gao and Zhou, 2015; Pati et al., 2014). Hence, exploring a rank-adaptive Bayesian procedure and obtain attractive theoretical properties or computation tractability could be interesting extensions as well.

The low-rank assumption (Assumption 3.4) requires that $r \log n \lesssim \log p$ and guarantees that the minimax rate for estimating the covariance matrix $\mathbf{\Sigma}$ under the Frobenius norm coincides with that under the spectral norm. More precisely, the minimax rate with regard to $\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_{\mathrm{F}}$ is $\sqrt{(rs + s \log p)/n}$, whereas the minimax rate under $\|\mathbf{\Sigma} - \mathbf{\Sigma}_0\|_2$ is $\sqrt{(s \log p)/n}$ and does not depend on the rank r. When Assumption 3.4 is violated, the two rates differ from each other, and the proof technique adopted in this work is no longer applicable to establish the rate-optimal posterior contraction under the (non-intrinsic) spectral norm (Hoffmann et al., 2015). In the recent technical report (Xie, 2021), the author partially addressed the rate-optimal posterior contraction under the spectral norm without assuming $r \log n \lesssim \log p$, but the other assumptions there were more restrictive. The posterior contraction rates under non-intrinsic metrics in general high-dimensional models remain relatively underexplored.

Markov chain Monte Carlo can be computationally intensive for high-dimensional settings in general. In this paper, we applied a standard Metropolis-within-Gibbs sampler for Bayesian computation of the sparse spiked covariance models. Inspired by Ning (2021), it would be desirable to develop computationally efficient methods, such as an expectation-maximization algorithm for the maximum *a posteriori* estimation instead of computing the full posterior distribution (Rockova and George, 2016) or a penalized least squares method (She, 2017), and explore the underlying theoretical guarantees.

Supplementary Material

Supplementary Material for "Bayesian Sparse Spiked Covariance Model With a Continuous Matrix Shrinkage Prior" (DOI: 10.1214/21-BA1292SUPP; .pdf). The supplementary material contains the proofs of the theoretical results, additional technical results, and additional numerical results.

References

Bai, R. and Ghosh, M. (2018). "High-dimensional multivariate posterior consistency under global-local shrinkage priors." *Journal of Multivariate Analysis*. URL https://www.sciencedirect.com/science/article/pii/S0047259X17306905. MR3830639. doi: https://doi.org/10.1016/j.jmva.2018.04.010. 1200

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." *Bayesian Statistics*, 7: 733–742. MR2003537. 1193

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 1200, 1205, 1208, 1213

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet-Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. PMID: 27019543. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 1195, 1197

Cai, T., Ma, Z., and Wu, Y. (2015). "Optimal estimation and rank detection for sparse spiked covariance matrices." *Probability Theory and Related Fields*, 161(3-4): 781–815. MR3334281. doi: https://doi.org/10.1007/s00440-014-0562-z. 1194, 1196, 1197

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. 1194, 1199, 1205, 1208, 1213

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation." *Electron. J. Statist.*, 10(1): 1–59. MR3466172. doi: https://doi.org/10.1214/15-EJS1081. 1194

Cai, T. T. and Zhou, H. H. (2012). "Optimal rates of convergence for sparse covariance matrix estimation." *Ann. Statist.*, 40(5): 2389–2420. MR3097607. doi: https://doi.org/10.1214/12-AOS998. 1194

Candès, E. J. and Recht, B. (2009). "Exact matrix completion via convex optimization." *Foundations of Computational Mathematics*, 9(6): 717–772. MR2565240. doi: https://doi.org/10.1007/s10208-009-9045-5. 1204

Cape, J., Tang, M., and Priebe, C. E. (2019a). "Signal-plus-noise matrix models: Eigenvector deviations and fluctuations." *Biometrika*, 106(1): 243–250. MR3912394. doi: https://doi.org/10.1093/biomet/asy070. 1197

Cape, J., Tang, M., and Priebe, C. E. (2019b). "The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics." *The Annals of Statistics*, 47(5): 2405–2439. MR3988761. doi: https://doi.org/10.1214/18-AOS1752. 1197, 1204

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for

sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1197

Castillo, I. and van der Vaart, A. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." *Ann. Statist.*, 40(4): 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 1198

Fan, J., Fan, Y., and Lv, J. (2008). "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics*, 147(1): 186–197. Econometric modelling in finance and risk management: An overview. URL https://www.sciencedirect.com/science/article/pii/S0304407608001346. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 1204

Fan, J., Liao, Y., and Mincheva, M. (2013). "Large covariance estimation by thresholding principal orthogonal complements." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4): 603–680. MR3091653. doi: https://doi.org/10.1111/rssb.12016. 1204, 1205

Fan, J., Wang, W., and Zhong, Y. (2016). "An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation." *arXiv preprint arXiv:1603.03516*. MR3827095. 1204

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical LASSO." *Biostatistics*, 9(3): 432–441. doi: https://doi.org/10.1093/biostatistics/kxm045. 1194

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. 1194, 1195, 1202, 1203, 1213

Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. (2001). "From few to many: illumination cone models for face recognition under variable lighting and pose." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 643–660. 1193, 1208

Geweke, J. and Zhou, G. (1996). "Measuring the pricing error of the arbitrage pricing theory." *The Review of Financial Studies*, 9(2): 557–587. 1193

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). "On adaptive posterior concentration rates." *The Annals of Statistics*, 43(5): 2259–2295. MR3396985. doi: https://doi.org/10.1214/15-AOS1341. 1213

Johnstone, I. M. (2001). "On the distribution of the largest eigenvalue in principal components analysis." *Annals of Statistics*, 295–327. MR1863961. doi: https://doi.org/10.1214/aos/1009210544. 1194

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. PMID: 20617121. MR2751448. doi: https://doi.org/10.1198/jasa.2009.0121. 1194

Lee, K.-C., Ho, J., and Kriegman, D. J. (2005). "Acquiring linear subspaces for face

recognition under variable lighting." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5): 684–698. 1193, 1208

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). "High-dimensional semiparametric Gaussian copula graphical models." *Ann. Statist.*, 40(4): 2293–2326. MR3059084. doi: https://doi.org/10.1214/12-AOS1037. 1194

Ning, B. (2021). "Spike and slab Bayesian sparse principal component analysis." *arXiv preprint arXiv:2102.00305*. 1194, 1195, 1202, 1213

Ning, B. and Ghosal, S. (2018). "Bayesian linear regression for multivariate responses under group sparsity." *arXiv preprint arXiv:1807.03439*. MR4091112. doi: https://doi.org/10.3150/20-BEJ1198. 1200

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. 1194, 1197, 1198, 1200, 1202, 1204, 1213

Rockova, V. (2018). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *The Annals of Statistics*, 46(1): 401–437. MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 1195, 1197, 1198, 1199, 1200, 1213

Rockova, V. and George, E. I. (2016). "Fast Bayesian factor analysis via automatic rotations to sparsity." *Journal of the American Statistical Association*, 111(516): 1608–1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 1197, 1213

Rockova, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 1195, 1197, 1200, 1213

She, Y. (2017). "Selective factor extraction in high dimensions." *Biometrika*, 104(1): 97–110. MR3626471. doi: https://doi.org/10.1093/biomet/asw059. 1208, 1210, 1213

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). "Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings." *Statistica Sinica*, 28(2): 1053. MR3791100. 1197

Shin, M., Bhattacharya, A., and Johnson, V. E. (2020). "Functional horseshoe priors for subspace shrinkage." *Journal of the American Statistical Association*, 115(532): 1784–1797. MR4189757. doi: https://doi.org/10.1080/01621459.2019.1654875. 1198

Stewart, G. W. and Sun, J.-g. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Boston, MA: Academic Press. URL https://www.worldcat.org/title/matrix-perturbation-theory/oclc/908946968. MR1061154. 1197

The Cancer Genome Atlas Network et al. (2012). "Comprehensive genomic characterization of squamous cell lung cancers." *Nature*, 489(7417): 519–525. 1193

Vu, V. Q. and Lei, J. (2013). "Minimax sparse principal subspace estimation in high dimensions." *The Annals of Statistics*, 41(6): 2905–2947. MR3161452. doi: https://doi.org/10.1214/13-AOS1151. 1194, 1196

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning*, 1(1–2): 1–305. doi: https://doi.org/10.1561/2200000001. 1194

Xie, F. (2021). "Euclidean representation of low-rank matrices and its statistical applications." *arXiv preprint arXiv:2103.04220*. 1213

Xie, F., Cape, J., Priebe, C. E. and Xu, Y. (2022). "Supplementary material for: Bayesian Sparse Spiked Covariance Model with a Continuous Matrix Shrinkage Prior." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1292SUPP. 1205

Zou, H., Hastie, T., and Tibshirani, R. (2006). "Sparse principal component analysis." *Journal of Computational and Graphical Statistics*, 15(2): 265–286. MR2252527. doi: https://doi.org/10.1198/106186006X113430. 1205