Posterior Consistency of Factor Dimensionality in High-Dimensional Sparse Factor Models[*]

Ilsang Ohn[†] and Yongdai Kim[‡]

Abstract. Factor models aim to describe a dependence structure among high-dimensional random variables in terms of a low-dimensional unobserved random vector called a factor. One of the major practical issues of applying the factor model is to determine the factor dimensionality. In this paper, we propose a computationally feasible nonparametric prior distribution which achieves the posterior consistency of the factor dimensionality. We also derive the posterior contraction rate of the covariance matrix which is optimal when the factor dimensionality of the true covariance matrix is bounded. We conduct numerical studies that illustrate our theoretical results.

Keywords: covariance matrix, factor dimensionality, factor model, Indian buffet process, posterior consistency, posterior contraction rate.

1 Introduction

Factor models describe a dependence structure among a high-dimensional correlated random vector in terms of a low-dimensional unobserved random vector called a latent factor or just factor. To be specific, the (linear) factor model considered in this paper assumes that a p-dimensional random vector $\mathbf{Y}$ is distributed as

$$\mathbf{Y}|\mathbf{Z} = \mathbf{z} \sim \mathsf{N}(\mathbf{Bz}, \sigma^2\mathbf{I}), \quad \mathbf{Z} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}), \qquad (1.1)$$

where $\mathbf{B}$ is a $p \times K$ factor loading matrix, $\mathbf{Z}$ is a K-dimensional factor with $K < p$, and $\sigma^2 > 0$ is a noise variance. Under this model, the marginal distribution of $\mathbf{Y}$ is given by

$$\mathbf{Y} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} := \mathbf{BB}^\top + \sigma^2\mathbf{I}.$$

That is, the distribution of $\mathbf{Y}$ is determined by the structured covariance matrix $\mathbf{BB}^\top + \sigma^2\mathbf{I}$. This decomposition of the covariance matrix leads to the substantial reduction of the model complexity, and thus the factor model has been applied to a broad range of areas including high-dimensional covariance estimation (Fan et al., 2008, 2011, 2018), high-dimensional supervised learning (Fan et al., 2017; Kneip and Sarda, 2011; Silva, 2011; Stock and Watson, 2002) and multiple testing under arbitrary dependence (Fan et al., 2012, 2019; Leek and Storey, 2008), and popularly used in various application fields such as economy, psychology and gene expression studies (e.g., Bernanke et al., 2005; Carvalho et al., 2008; Forni et al., 2003; Hochreiter et al., 2006; McCrae and John, 1992).

[†]Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, USA, iohn@nd.edu
[‡]Department of Statistics, Seoul National University, South Korea, ydkim0903@gmail.com

A major practical issue of using the factor model is to determine the *factor dimensionality* K. We need to select an appropriate factor dimensionality to optimize the bias-variance trade-off. In addition, the factor dimensionality is of interest itself in practice since each factor could have a physical interpretation (e.g., the number of interacting pathways in genomics and the number of personality traits in psychology).

Frequentist approaches typically choose the factor dimensionality before estimating the loading matrix. One of the widely used methods is to fit the factor models for different values of K and to select the best K based on a model selection criterion (Bai and Ng, 2002, 2007). Alternatively, the factor dimensionality can be chosen based on the eigenvalues of the empirical covariance matrix (Ahn and Horenstein, 2013; Lam and Yao, 2012; Onatski, 2010).

Various priors have been developed for Bayesian analysis of the factor model with unknown factor dimensionality. Examples are spike and slab priors with the Indian buffet process (IBP) (Chen et al., 2010; Knowles et al., 2011; Rockova and George, 2016) and shrinkage type priors with the degree of shrinkage increasing across the column index (Bhattacharya and Dunson, 2011; Srivastava et al., 2017).

Large sample properties of the posterior distribution of the factor model also have received much attention. Pati et al. (2014) investigated the posterior contraction rate of the covariance matrix with respect to the spectral norm for a sparse factor model where most of the entries in the factor loading matrix are zero. They showed that the derived posterior contraction rate is near-optimal in the minimax sense, up to a logarithm factor even when p is much larger than n. An improved contraction rate was obtained by Xie et al. (2018) and Ning (2021) proved that the variational posterior can also achieve this improved rate.

However, these studies focused on the covariance matrix but not on the factor dimensionality. Rockova and George (2016) considered the Bayesian factor model with a spike and slab prior with the IBP proposed by Ghahramani and Griffiths (2005), and proved that the posterior probability that the factor dimensionality is bounded by a certain quantity converges to one. But, the posterior consistency of the factor dimensionality is still unsolved.

Gao and Zhou (2015) studied a Bayesian sparse principal component analysis (PCA) model, which is equivalent to the factor model with the constraint that the columns of the loading matrix are orthogonal to each other. They derived posterior contraction rates of the covariance matrix and the principal subspace estimation with respect to the spectral norm and proved the posterior consistency of the rank of the covariance matrix. Due to the orthogonality of the loading matrix, the rank of the covariance matrix is equal to the factor dimensionality. But there is no easy computational method to approximate the posterior distribution mainly because of the orthogonality constraint on the loading matrix.

The main contribution of this paper is to propose a Bayesian factor model which is computationally tractable and at the same time achieves the posterior consistency of the factor dimensionality. For this purpose, we consider a spike and slab prior with

the two-parameter IBP which is an extension of the one-parameter IBP in Ghahramani and Griffiths (2005). Bayesian factor models with two-parameter IBP have been studied by Chen et al. (2010) and Paisley and Carin (2009) but no theoretical study has been done. In this paper, we develop an easily implementable Markov chain Monte Carlo (MCMC) algorithm for our proposed Bayesian model, which is similar to that for the model of Ghahramani and Griffiths (2005), and prove the posterior consistency of the factor dimensionality. In addition, as a by-product, we derive the posterior contraction rate of the covariance matrix which is similar to or sometimes better than other Bayesian factor models.

This paper is organized as follows. In Section 2, we introduce the assumptions on the true model. Then we explain the proposed prior and its properties. In Section 3, we provide posterior asymptotic results. In Section 4, we provide an MCMC algorithm for our proposed Bayesian model. In Section 5, we conduct simulation studies to supplement our asymptotic results and show superiority of the proposed method. Real data analysis is conducted in Section 6. Concluding remarks follow in Section 7.

1.1 Notation

Let $\mathbb{R}$ be the set of real numbers and $\mathbb{N}$ be the set of natural numbers. For the positive integer p, we let $[p] := \{1, 2, \ldots, p\}$. For a real number x, let $\lfloor x \rfloor$ denote the largest integer less than or equal to x and $\lceil x \rceil$ denote the smallest integer larger than or equal to x. For two real numbers a and b, we write $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For two positive sequences $\{a_n\}_{n \in \mathbb{N}}$ and $\{b_n\}_{n \in \mathbb{N}}$, we write $a_n \lesssim b_n$ if there exists a positive constant $C > 0$ such that $a_n \leq C b_n$ for any $n \in \mathbb{N}$. Moreover, we write $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We denote by $\mathbb{1}(\cdot)$ the indicator function. Let $\Gamma(a)$ denote the gamma function and $\mathsf{B}(b, c)$ denote the beta function, where a, b and c are positive constants. Let $\mathbf{0}$ and $\mathbf{1}$ denote the vector of 0's and the one of 1's, respectively.

For a set S, $|S|$ denotes its cardinality. For a p-dimensional vector $\boldsymbol{\beta} \equiv (\beta_j)_{j \in [p]}$, let $\|\boldsymbol{\beta}\|_0 := \sum_{j=0}^{p} \mathbb{1}(\beta_j \neq 0)$. We denote $\boldsymbol{\beta}^S := (\beta_j : j \in S)$ for a subset $S \subset \{1, \ldots, p\}$. We let $\lambda_1(\boldsymbol{\Sigma}) \geq \lambda_2(\boldsymbol{\Sigma}) \geq \cdots \geq \lambda_p(\boldsymbol{\Sigma})$ be the ordered eigenvalues of a $p \times p$ matrix $\boldsymbol{\Sigma}$. For a $p \times k$ matrix $\mathbf{A} \equiv (a_{jh})_{j \in [p], h \in [k]}$, we denote the spectral norm of the matrix $\mathbf{A}$ by $\|\mathbf{A}\|$ and the Frobenius norm by $\|\mathbf{A}\|_{\mathrm{F}}$. That is, $\|\mathbf{A}\| := \sup_{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \lambda_1^{1/2}(\mathbf{A}^\top \mathbf{A})$ and $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\mathsf{tr}(\mathbf{A}^\top \mathbf{A})}$. Also we use $\|\mathbf{A}\|_1$ to denote ℓ_1 norm of $\mathrm{vec}(\mathbf{A})$, i.e., $\|\mathbf{A}\|_1 := \sum_{j=1}^{p} \sum_{h=1}^{k} |a_{jh}|$.

2 Assumptions and prior distribution

2.1 Assumptions

Throughout this paper, we assume that for each $n \in \mathbb{N}$, we observe n independent and identically distributed p_n dimensional observations $\mathbf{Y}_{1:n} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ from the normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{0n}$ and we model the data using the model (1.1).

We introduce some regularity conditions on the sequence of the true covariance matrices $\{\boldsymbol{\Sigma}_{0n}\}_{n\in\mathbb{N}}$. First, we consider sparse loading matrices. Since we want to deal with a high dimensional case where p_n is much larger than n, we assume that the true loading matrix is sufficiently sparse to make the factor loading estimable. Define the class of $p \times k_0$ matrices with the sparsity condition that the number of nonzero elements of each column is at most s,

$$\mathcal{B}(p, k_0, s) := \left\{ \mathbf{B} \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k_0}) \in \mathbb{R}^{p \times k_0} : \max_{1 \leq h \leq k_0} \|\boldsymbol{\beta}_h\|_0 \leq s \right\}. \tag{2.1}$$

The above class of sparse loading matrices is also considered by Pati et al. (2014) and Gao and Zhou (2015). The parameter space for the true covariance matrix we consider is given by

$$\mathcal{C}(p, k_0, s, c) := \left\{ \boldsymbol{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} : \mathbf{B} \in \mathcal{B}(p, k_0, s), \sigma^2 > c_0, \|\boldsymbol{\Sigma}\| \leq c \right\} \tag{2.2}$$

for some universal constant $c_0 > 0$. Here the lower bound $\sigma^2 > c_0$ is introduced to prevent that $\boldsymbol{\Sigma}_{0n}$ becomes ill-conditioned.

We assume that for each sample size $n \in \mathbb{N}$, the true covariance matrix $\boldsymbol{\Sigma}_{0n}$ belongs to the class $\mathcal{C}_{0n} := \mathcal{C}(p_n, s_n, k_{0n}, c_n)$, where the sequence of the classes $\{\mathcal{C}_{0n}\}_{n\in\mathbb{N}}$ satisfies the following assumptions:

(A1) $p_n > n$.

(A2) $c_n^2 s_n^2 k_{0n} \log p_n / n = o(1)$.

(A3) $1 \leq k_{0n} < p_n/2$.

(A4) $c_n \lesssim s_n$.

The assumption (A1) means that we consider high dimensional models. However, our theoretical analysis can be applied to low dimensional cases where $n > p_n$. In the low dimensional cases, the $\log p_n$ factors in the results are replaced with $\log n$. The quantity $c_n^2 s_n^2 k_{0n} \log p_n / n$ in (A2) is equal to the square of the posterior contraction rate of the covariance matrix under our proposed prior with respect to the spectral norm. In (A3), we assume that the true model has at least one factor and the factor dimensionality does not exceed half of the number of observed variables for technical reasons.

The assumption (A4) implies that we allow the largest eigenvalue of $\boldsymbol{\Sigma}_{0n}$ to grow with sample size. The bound $c_n \lesssim s_n$ is mild in view of random matrix theory. Suppose that $\tilde{B}$ is a $s_n \times k_{0n}$ random matrix whose entries are independent centered random variables with finite fourth moments. Then by Theorem 2 of Latala (2005), $\mathbb{E}\|\tilde{B}\| \lesssim \sqrt{s_n} + \sqrt{k_{0n}}$. If $k_{0n} \lesssim s_n$, we have that $\mathbb{E}\|\tilde{B}\tilde{B}^\top\| \lesssim s_n$. Pati et al. (2014) and Rockova and George (2016) assumed the same condition, while the other works on the Bayesian covariance estimation assumed the stronger condition that the largest eigenvalue of $\boldsymbol{\Sigma}_{0n}$ is bounded (Gao and Zhou, 2015; Xie et al., 2018).

2.2 Prior and its properties

Let β_{jk} be the (j, k) entry of the $p \times \infty$-dimensional loading matrix $\mathbf{B}$. We consider the following prior

$$\beta_{jk}|\xi_{jk} \overset{\text{ind}}{\sim} (1 - \xi_{jk})\delta_0 + \xi_{jk}\mathsf{Laplace}(1), \ j \in [p], \ k \in \mathbb{N}, \tag{2.3}$$

$$\xi_{jk}|\theta_k \overset{\text{ind}}{\sim} \mathsf{Bernoulli}(\theta_k), \ j \in [p], \ k \in \mathbb{N}, \tag{2.4}$$

$$\theta_k := \prod_{h=1}^{k} \nu_h, \ k \in \mathbb{N} \text{ where } \nu_h \overset{\text{iid}}{\sim} \mathsf{Beta}(\alpha, \kappa + 1), \ h \in \mathbb{N}, \tag{2.5}$$

where $\alpha > 0$ and $\kappa \geq 0$ are hyperparameters. Note that $\{\xi_{jk}\}$ is the stick-breaking representation of the two-parameter IBP (Teh et al., 2007). Thus, we refer to the above distribution on $\mathbf{B}$ as $\mathsf{SSIBP}_p(\alpha, \kappa)$, which is an abbreviation of *spike and slab Indian buffet process*.

The (one-parameter) IBP introduced by Ghahramani and Griffiths (2005), which is the prior on $\{\xi_{jk}\}$ with $\kappa = 0$, has been popularly used as a prior on the nonzero entries of the loading matrix. The prior distribution of Knowles et al. (2011) for $\mathbf{B}$ is almost the same as the SSIBP prior except that $\kappa = 0$ and the Laplace distribution in (2.3) is replaced by the normal distribution. Rockova and George (2016) used the IBP for the prior of ξ_{jk}, but they replace δ_0 in the first line of (2.3) with the Laplace distribution with a very small dispersion. This replacement enables us to use the fast and scalable expectation-maximization algorithm that estimates a posterior mode. We consider the two parameter IBP in order to get the posterior consistency of the factor dimensionality by choosing κ appropriately. Although Chen et al. (2010) and Paisley and Carin (2009) used the two-parameter IBP, no theoretical study has been done.

In the following three subsections, we provide some theoretical properties of the SSIBP prior. All the proofs of the results in this section are deferred to Appendix A.1 in the supplementary material (Ohn and Kim, 2021).

Prior distribution of the factor dimensionality

We derive an upper bound of the tail probability of the factor dimensionality induced by the SSIBP prior. We first define the factor dimensionality of the $p \times \infty$ dimensional loading matrix $\mathbf{B}$.

Definition 2.1. For a given $p \times \infty$ loading matrix $\mathbf{B} \equiv (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots)$, we define the *factor dimensionality* $\mathsf{K}^+(\mathbf{B})$ as the number of nonzero columns of $\mathbf{B}$, i.e.,

$$\mathsf{K}^+(\mathbf{B}) := \sum_{k=1}^{\infty} \mathbb{1}\left(\|\boldsymbol{\beta}_k\|_0 \geq 1\right),$$

where $\boldsymbol{\beta}_k$ denotes the k-th column of $\mathbf{B}$.

The following lemma shows that the tail probability of the factor dimensionality is exponentially decaying as the factor dimensionality increases.

Lemma 2.1. *If* $\mathbf{B} \sim \mathsf{SSIBP}_p(\alpha, \kappa)$ *for* $\alpha > 0$ *and* $\kappa \geq 0$, *then for any* $k \in \mathbb{N}$

$$\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq C_{\alpha,\kappa} p \left(\frac{\alpha}{\alpha + \kappa + 1} \right)^{k+1}, \tag{2.6}$$

where $C_{\alpha,\kappa} := 2((\alpha + \kappa + 1)/(\kappa + 1) + 4/3)$. *In particular, if* $\alpha \leq \kappa + 1$, *then*

$$\Pi(\mathsf{K}^+(\mathbf{B}) > k) \leq 6p \left(\frac{\alpha}{\alpha + \kappa + 1} \right)^{k+1}. \tag{2.7}$$

Prior distribution of row-wise sparsity

Recall that we assume column-wise maximum sparsity of the true loading matrix. In this section, for the ease of mathematical expositions, we introduce a notion of stronger sparsity called row-wise sparsity.

Definition 2.2. For a $p \times \infty$ loading matrix $\mathbf{B} \equiv (\beta_{jh})_{j \in [p], h \in \mathbb{N}}$ and a positive integer $k \in \mathbb{N}$, we define the *row-support up to k-th column* of $\mathbf{B}$ as

$$\mathsf{supp}_k(\mathbf{B}) := \left\{ j \in [p] : \sum_{h=1}^{k} |\beta_{jh}| > 0 \right\}.$$

$\mathbf{B}$ is said to be *row-wise s-sparse up to k* if $|\mathsf{supp}_k(\mathbf{B})| \leq s$.

Note that for any $\mathbf{B} \in \mathbb{R}^{p \times \infty}$ and any $k \in \mathbb{N}$,

$$\max_{1 \leq h \leq k} \|\boldsymbol{\beta}_h\|_0 \leq |\mathsf{supp}_k(\mathbf{B})|,$$

where $\boldsymbol{\beta}_h$ denotes the h-th column of $\mathbf{B}$.

Throughout this paper, we set $\kappa = p^{1+\delta}$ for a fixed constant $\delta > 0$. This choice of the hyperparameter is intended to put most of the prior mass concentrating on sufficiently sparse loading matrices, which is shown in the following lemma.

Lemma 2.2. *If* $\mathbf{B} \sim \mathsf{SSIBP}_p(\alpha, p^{1+\delta})$ *with* $\alpha \in (0, 1)$ *and* $\delta > 0$, *then for any* $k \in \mathbb{N}$, *and* $t \geq 1$,

$$\Pi(|\mathsf{supp}_k(\mathbf{B})| > t) \leq (k+1)\mathrm{e}^{-(\delta/12)t \log p}. \tag{2.8}$$

The above lemma trivially implies that

$$\Pi\left(\left\{ \mathbf{B} \equiv (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots) \in \mathbb{R}^{p \times \infty} : \max_{1 \leq h \leq k} \|\boldsymbol{\beta}_h\|_0 > t \right\} \right) \leq (k+1)\mathrm{e}^{-(\delta/12)t \log p}.$$

Prior concentration near the true loading matrix

In this subsection, we show that the SSIBP prior puts sufficiently large mass near the truth. Since the true loading matrix $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ depends on n, we let also the

hyperparameters $\alpha = \alpha_n$ and $\kappa = \kappa_n$ also depend on n. We let Π_n denote the prior distribution $\mathsf{SSIBP}_{p_n}(\alpha_n, \kappa_n) \times \mathsf{IG}(a, b)$ on $(\mathbf{B}, \sigma^2)$, where $\mathsf{IG}(a, b)$ denotes the inverse gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Unless there is a confusion, we understand the loading matrix $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ as the $p_n \times \infty$ dimensional matrix $(\mathbf{B}_{0n}, \mathbf{0}_{p_n \times \infty})$, if necessary, where $\mathbf{0}_{p_n \times \infty}$ denotes $p_n \times \infty$ dimensional matrix of 0's.

Lemma 2.3. *Suppose that $\mathbf{B}_{0n} \in \mathcal{B}(p_n, k_{0n}, s_n)$. If $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ for $\alpha_n > 0$ and $\delta > 0$, then for any $n \in \mathbb{N}$ and $\eta > 0$,*

$$\Pi_n\big(\|\mathbf{B} - \mathbf{B}_{0n}\|_\mathrm{F} \leq \eta\big) \geq \alpha_n^{k_{0n}} \mathrm{e}^{-\|\mathbf{B}_{0n}\|_1 - \eta k_{0n} - C_1 s_n k_{0n} \log(p_n \vee \eta^{-1})} \tag{2.9}$$

for some universal constant $C_1 > 0$ depending only on δ. Moreover, if η_n is such that $p_n^{-m} \lesssim \eta_n \lesssim 1$ for some $m > 0$ and the assumption (A4) holds, we have

$$\Pi_n\big(\|\mathbf{B} - \mathbf{B}_{0n}\|_\mathrm{F} \leq \eta_n\big) \geq \alpha_n^{k_{0n}} \mathrm{e}^{-C_2 s_n k_{0n} \log p_n} \tag{2.10}$$

for some universal constant $C_2 > 0$ depending only on δ.

Using Lemma 2.3, we can obtain the following prior concentration result for the covariance matrix with respect to the Frobenius norm.

Corollary 2.4. *Suppose that $\boldsymbol{\Sigma}_{0n} \in \mathcal{C}_{0n} := \mathcal{C}(p_n, k_{0n}, s_n, c_n)$, where $\{\mathcal{C}_{0n}\}_{n \in \mathbb{N}}$ satisfies the assumptions (A1)–(A4). If $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ and $\sigma^2 \sim \mathsf{IG}(a, b)$ for $\alpha_n > 0$, $\delta > 0$, $a > 0$ and $b > 0$, then*

$$\Pi_n\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\|_\mathrm{F} \leq \sqrt{\frac{s_n k_{0n}}{n}}\right) \geq \alpha_n^{k_{0n}} \mathrm{e}^{-C_1 s_n k_{0n} \log p_n} \tag{2.11}$$

for some universal constant $C_1 > 0$ depending only on δ, a and b.

Most priors for high dimensional sparse factor models have the lower bound $\exp(-C s_n k_{0n} \log p_n)$ for the prior concentration (Gao and Zhou, 2015; Pati et al., 2014; Rockova and George, 2016). The lower bound in (2.11) is similar to them, but the key difference of (2.11) is that the lower bound depends explicitly on the hyperparameter α_n. For the posterior consistency of the factor dimensionality, controlling α_n is indispensable.

3 Asymptotic properties of the posterior distribution

In this section we study asymptotic properties of the posterior distributions of the covariance matrix and the factor dimensionality in the sparse factor model, respectively. We denote by $\Pi_n(\cdot | \mathbf{Y}_{1:n})$ the posterior distribution induced by the prior Π_n and the data $\mathbf{Y}_{1:n} \equiv (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. For a given sample size n and covariance matrix $\boldsymbol{\Sigma}$, we let $\mathbb{P}_{\boldsymbol{\Sigma}}$ and $\mathbb{E}_{\boldsymbol{\Sigma}}$ denote the probability measure and the expectation operator under the law $(\mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}))^n$, where we suppress the dependence on n for simplicity.

The proofs of the results in this section are deferred to Appendix A.2 in the supplementary material (Ohn and Kim, 2021).

3.1 Posterior contraction rate of covariance matrix

We let $\{\mathcal{C}_{0n}\}_{n\in\mathbb{N}}$ with $\mathcal{C}_{0n} := \mathcal{C}(p_n, k_{0n}, s_n, c_n)$ be the sequence of the classes of covariance matrices satisfying the assumptions (A1)–(A4). The next theorem derives the posterior contraction rate of the covariance matrix.

Theorem 3.1. *A priori, let* $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(\alpha_n, p_n^{1+\delta})$ *with* $0 < \alpha_n < 1$ *and* $\sigma^2 \sim \mathsf{IG}(a, b)$ *for* $\delta > 0$, $a > 0$ *and* $b > 0$. *Then*

$$\sup_{\boldsymbol{\Sigma}_{0n}\in\mathcal{C}_{0n}} \mathbb{E}_{\boldsymbol{\Sigma}_{0n}}\left[\Pi_n\left(\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0n}\| > M\epsilon_n \Big| \mathbf{Y}_{1:n}\right)\right] = o(1) \tag{3.1}$$

for any sufficiently large $M > 0$, *where*

$$\epsilon_n := c_n\sqrt{\frac{k_{0n}}{n}\max\left\{s_n\log p_n, \log\left(\frac{1}{\alpha_n}\right)\right\}}. \tag{3.2}$$

If we set the hyperparameter α_n to satisfy $\log(\frac{1}{\alpha_n}) \lesssim s_n\log p_n$, for instance, $\alpha_n = \alpha_0$ for some $\alpha_0 \in (0,1)$ or $\alpha_n = p_n^{-1}$ as in Rockova and George (2016), the posterior contraction rate becomes

$$\epsilon_n = c_n\sqrt{\frac{s_n k_{0n}\log p_n}{n}}. \tag{3.3}$$

We compare this rate with those in other related Bayesian studies. Pati et al. (2014) derived a similar posterior contraction rate with their own prior and assumptions. The posterior contraction rate of Pati et al. (2014) is $c_n\sqrt{s_n k_{0n}^3\log p_n/n}\sqrt{\log n}$. We remove $k_{0n}\sqrt{\log n}$ factor by using the improved test construction developed by Gao and Zhou (2015). Gao and Zhou (2015) obtained the posterior contraction rate $\sqrt{s_n k_{0n}\log p_n/n}$ under the assumption that the largest eigenvalue of the true covariance matrix is bounded. The posterior distribution of Xie et al. (2018) and the variational posterior distribution of Ning (2021) enjoy the contraction rate $\sqrt{s_n\log p_n/n}$ under the assumptions that both the true largest eigenvalue and the true factor dimensionality are bounded. Our posterior contraction rate in (3.3) recovers those rates.

Whether the convergence rate ϵ_n in (3.3) is minimax optimal is still open problem. The closely related minimax result is studied by Cai et al. (2015). They derived the minimax convergence rate of a covariance matrix with a row-wise sparse loading matrix, i.e., the covariance matrix in the class given by

$$\bar{\mathcal{C}}(p, k, s, c) := \left\{\boldsymbol{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2\mathbf{I} : \mathbf{B} \in \mathbb{R}^{p\times k}, |\mathsf{supp}_k(\mathbf{B})| \le s, \sigma^2 \ge c_0, \|\boldsymbol{\Sigma}\| \le c\right\}. \tag{3.4}$$

Under the additional assumption that $\sigma^2 = 1$, Cai et al. (2015) proved that

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\boldsymbol{\Sigma}_{0n}\in\bar{\mathcal{C}}(p_n,k_{0n},s_n,c_n):\sigma_{0n}^2=1} \mathbb{E}_{\boldsymbol{\Sigma}_{0n}}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{0n}\| \gtrsim \sqrt{c_n}\sqrt{\frac{s_n\log(p_n/s_n)}{n}}, \tag{3.5}$$

which is faster than the posterior convergence rate in Theorem 3.1 by the factor $\sqrt{c_n}$. However, if we drop the assumption of $\sigma^2 = 1$, we have the minimax lower bound

$$\inf_{\widehat{\boldsymbol{\Sigma}}} \sup_{\boldsymbol{\Sigma}_{0n}\in\bar{\mathcal{C}}(p_n,k_{0n},s_n,c_n)} \mathbb{E}_{\boldsymbol{\Sigma}_{0n}}\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{0n}\| \gtrsim c_n\sqrt{\frac{s_n\log(p_n/s_n)}{n}}. \tag{3.6}$$

See Appendix A.3 in the supplementary material (Ohn and Kim, 2021) for derivation. Note that $\mathcal{C}(p_n, k_{0n}, s_n, c_n) \subset \bar{\mathcal{C}}(p_n, k_{0n}, s_n k_{0n}, c_n)$ but $\mathcal{C}(p_n, k_{0n}, s_n, c_n) \not\subset \bar{\mathcal{C}}(p_n, k_{0n}, q, c_n)$ for any $q < s_n k_{0n}$. Thus, our posterior contraction rate ϵ_n in (3.3) would be expected to be (nearly) minimax optimal, which we leave as a future work.

3.2 Posterior consistency of the factor dimensionality

For the posterior consistency of the factor dimensionality, we need the additional assumption on the smallest eigenvalue of the low rank matrix $\mathbf{B}_{0n}\mathbf{B}_{0n}^\top$. Define

$$\mathcal{C}^\star(p, k_0, s, c, \zeta) := \left\{ \mathbf{\Sigma} \equiv \mathbf{B}\mathbf{B}^\top + \sigma^2 \mathbf{I} \in \mathcal{C}(p, k_0, s, c) : \lambda_{k_0}(\mathbf{B}\mathbf{B}^\top) > \zeta \right\} \qquad (3.7)$$

for $\zeta > 0$. We assume that the true covariance matrix $\mathbf{\Sigma}_{0n}$ belongs to the class $\mathcal{C}_{0n}^\star := \mathcal{C}^\star(p_n, k_{0n}, s_n, c_n, \zeta_n)$, where the sequence of the classes $\{\mathcal{C}_{0n}^\star\}_{n \in \mathbb{N}}$ satisfies the following assumption:

(A5) $\zeta_n > C_0 c_n \sqrt{s_n^2 k_{0n} \log p_n / n}$ for a sufficiently large constant $C_0 > 0$

in addition to the assumptions (A1)–(A4).

The assumption (A5) assumes that ζ_n is larger than the posterior contraction rate of the covariance matrix, which is necessary for not underestimating the factor dimensionality k_{0n}. The assumption (A5) plays a similar role as the beta-min condition which has been popularly used to prove variable section consistency in high-dimensional regression (Castillo and van der Vaart, 2012; Martin et al., 2017).

We let $\{\mathcal{C}_{0n}^\star\}_{n \in \mathbb{N}}$ with $\mathcal{C}_{0n}^\star := \mathcal{C}^\star(p_n, k_{0n}, s_n, c_n, \zeta_n)$ be the sequence of covariance matrices satisfying the assumptions (A1)–(A5). The following theorem proves that the posterior consistency of the factor dimensionality.

Theorem 3.2. *A priori, let* $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(p_n^{-As_n^2}, p_n^{1+\delta})$ *for sufficiently large* $A > 0$ *and* $\sigma^2 \sim \mathsf{IG}(a, b)$ *for* $\delta > 0$, $a > 0$ *and* $b > 0$. *Then*

$$\sup_{\mathbf{\Sigma}_{0n} \in \mathcal{C}_{0n}^\star} \mathbb{E}_{\mathbf{\Sigma}_{0n}} \left[\Pi_n \left(\mathsf{K}^+(\mathbf{B}) \neq k_{0n} \Big| \mathbf{Y}_{1:n} \right) \right] = o(1). \qquad (3.8)$$

Rockova and George (2016) proved that $\mathbb{E}_{\mathbf{\Sigma}_{0n}}[\Pi_n(\mathsf{K}^+(\mathbf{B}) > M s_n k_{0n} | \mathbf{Y})] = o(1)$ for sufficiently large $M > 0$ for our prior with $\kappa_n = 0$ (i.e., the one-parameter IBP). This result is much weaker than ours when s_n diverges. Also Rockova and George (2016) did not consider the posterior probability of the underestimation of the factor dimensionality. We introduce a diverging value of κ_n and the assumption (A5) to prevent the underestimation.

For posterior consistency of the factor dimensionality, the condition (A5) plays a crucial role. When the assumption (A5) does not hold, that is, the true covariance matrix $\mathbf{\Sigma}_{0n}$ belongs to the larger class $\mathcal{C}(p_n, k_{0n}, s_n, c_n)$ instead of $\mathcal{C}_{0n}^\star$, we only ensure that

$$\mathbb{E}_{\mathbf{\Sigma}_{0n}} \left[\Pi_n \left(\tilde{k}_n \leq \mathsf{K}^+(\mathbf{B}) \leq k_{0n} \Big| \mathbf{Y}_{1:n} \right) \right] \to 1, \qquad (3.9)$$

where $\tilde{k}_n := \sup\{\tilde{k} \in [k_{0n}] : \lambda_{\tilde{k}}(\mathbf{B}_{0n}\mathbf{B}_{0n}^\top) > \zeta_n\} \vee 0$ with ζ_n satisfying (A5). The proof of (3.9) is given at the end of Appendix A.3 in the supplementary material (Ohn and Kim, 2021).

Cai et al. (2015) proved that if the true covariance matrix is of the form $\mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \mathbf{I}$ with $\mathbf{B}_{0n} \in \mathbb{R}^{p_n \times k_{0n}}$ and $|\mathsf{supp}_{k_{0n}}(\mathbf{B}_{0n})| \leq q_n$, the necessary and sufficient condition for consistent estimation of the factor dimensionality is given by

$$\zeta_n > C\sqrt{\frac{q_n \log p_n}{n}}$$

for a sufficiently large constant $C > 0$ under regularity conditions. Compared to the optimal lower bound of Cai et al. (2015), our lower bound of the eigengap in (A5) is larger by $\sqrt{s_n}$ factor. Moreover, for α_n in Theorem 3.2, the posterior contraction rate of the covariance matrix becomes

$$\epsilon_n = c_n\sqrt{\frac{s_n^2 k_{0n} \log p_n}{n}}, \tag{3.10}$$

which is also $\sqrt{s_n}$ times slower than the posterior contraction rate in (3.3). This suboptimality would be mainly due to the $\mathsf{SSIBP}_{p_n}(p_n^{-As_n^2}, p_n^{1+\delta})$ prior but not an inherent problem of Bayesian analysis. There is a prior which achieves the posterior consistency of the factor dimensionality and the fast posterior contraction rate of the covariance matrix as (3.3) simultaneously. An example is the prior of Gao and Zhou (2015). However, note that the SSIBP prior is computationally tractable.

Remark 1. When the true noise variance σ_{0n}^2 is known, we can derive a posterior contraction rate of the covariance matrix with respect to the Frobenius norm from Theorems 3.1 and 3.2. This is because if $\mathsf{K}^+(\mathbf{B}) = k_{0n}$ holds, we have $\mathsf{rank}(\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top) \leq 2k_{0n}$ and so $\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\|_\mathrm{F} = \|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\|_\mathrm{F} \leq \sqrt{2k_{0n}}\|\mathbf{B}\mathbf{B}^\top - \mathbf{B}_{0n}\mathbf{B}_{0n}^\top\|$. Hence the posterior contraction rate of $\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\|_\mathrm{F}$ is at most $\sqrt{k_{0n}}\epsilon_n = c_n s_n k_{0n}\sqrt{\log p_n/n}$.

3.3   Bounded factor dimensionality

The suboptimality of the posterior contraction rate of the covariance matrix with the SSIBP prior can be improved when the true factor dimensionality k_{0n} is bounded. The bounded factor dimensionality is not too restrictive in some situations such as macroeconomic applications (Onatski, 2010; Li et al., 2017). We introduce the following two assumptions instead of (A3) and (A5):

(A3′) $1 \leq k_{0n} \leq \bar{K}$ for some absolute constant $\bar{K} > 0$.

(A5′) $\zeta_n > C_0 c_n\sqrt{s_n \log p_n/n}$ for a sufficiently large constant $C_0 > 0$.

Let $\{\tilde{\mathcal{C}}_{0n}\}_{n \in \mathbb{N}}$ with $\tilde{\mathcal{C}}_{0n} := \mathcal{C}(p_n, k_{0n}, s_n, c_n)$ be the sequence of the classes of covariance matrices satisfying the assumptions (A1), (A2), (A3′), (A4). Furthermore, let $\tilde{\mathcal{C}}_{0n}^\star := \mathcal{C}^\star(p_n, k_{0n}, s_n, c_n, \zeta_n) \subset \tilde{\mathcal{C}}_{0n}$ with ζ_n satisfying (A5′).

Lemma 3.3. *A priori, let $\mathbf{B} \sim \mathsf{SSIBP}_{p_n}(p_n^{-As_n}, p_n^{1+\delta})$ for sufficiently large $A > 0$ and $\sigma^2 \sim \mathsf{IG}(a, b)$ for $\delta > 0$, $a > 0$ and $b > 0$. Then*

$$\sup_{\mathbf{\Sigma}_{0n} \in \tilde{\mathcal{C}}_{0n}} \mathbb{E}_{\mathbf{\Sigma}_{0n}}\left[\Pi_n\left(\|\mathbf{\Sigma} - \mathbf{\Sigma}_{0n}\| > Mc_n\sqrt{\frac{s_n \log p_n}{n}}\Big|\mathbf{Y}_{1:n}\right)\right] = o(1), \qquad (3.11)$$

for any sufficiently large $M > 0$ and

$$\sup_{\mathbf{\Sigma}_{0n} \in \tilde{\mathcal{C}}_{0n}^{\star}} \mathbb{E}_{\mathbf{\Sigma}_{0n}}\left[\Pi_n\left(\mathsf{K}^+(\mathbf{B}) \neq k_{0n}\Big|\mathbf{Y}_{1:n}\right)\right] = o(1). \qquad (3.12)$$

The detailed proof of Lemma 3.3 is almost similar to Theorems 3.1 and 3.2 and thus omitted.

The convergence rate in (3.11) is minimax optimal (Pati et al., 2014). That is, under the bounded factor dimensionality setup, we can attain both posterior consistency of the factor dimensionality and the optimal convergence of the covariance matrix. As we have shown in Corollary 2.4, the prior concentration is lower bounded by $\alpha^{k_{0n}} \exp(-C_1 s_n \log p_n)$ for some $C_1 > 0$ in this case, thus the upper bound of the posterior probability $\mathbb{E}_{\mathbf{\Sigma}_{0n}}\Pi_n(\mathsf{K}^+(\mathbf{B}) > k_{0n}|\mathbf{Y}_{1:n})$ is given by $\alpha_n \exp(C_2 s_n \log p_n)$ for some $C_2 > 0$. Therefore the choice of the hyperparameter $\alpha_n = p_n^{-As_n}$ for large $A > C_2$, instead of the choice $\alpha_n = p_n^{-As_n^2}$ in Theorem 3.2, is sufficient to make the upper bound go to zero. Moreover, the choice $\alpha_n = p_n^{-As_n}$ does not hurt the prior concentration, i.e., $\alpha^{k_{0n}} \exp(-C_1 s_n \log p_n) \asymp \exp(-C_3 s_n \log p_n)$ for some $C_3 > 0$, thus the results of Theorem 3.3 can be obtained.

Remark 2. A referee raised a question whether the same results of Lemma 3.3 hold when the upper bound $\bar{K}$ of the factor dimensionality grows as n goes to infinity. Unfortunately, the answer is negative. If $k_{0n} \leq \bar{K}_n$ for some positive sequence $\{\bar{K}_n\}_{n \in \mathbb{N}}$, it can be shown from our proof that the posterior consistency of the factor dimensionality is obtained with the hyperparameter $\alpha_n \asymp p_n^{-As_n\bar{K}_n}$ for sufficiently large $A > 0$, but this yields the posterior contraction rate $c_n\sqrt{s_n k_{0n}\bar{K}_n \log p_n/n}$ of the covariance matrix, which is larger than the optimal one $c_n\sqrt{s_n k_{0n} \log p_n/n}$ when $\bar{K}_n \to \infty$. That is, with the current proof technique, we cannot obtain the optimal convergence of the covariance matrix and the consistency of the factor dimensionality simultaneously if the factor dimensionality is not bounded.

4   Posterior computation

In this section, we provide an MCMC algorithm for posterior computation under the prior distribution $\mathsf{SSIBP}_p(\alpha, \kappa) \times \mathsf{IG}(a, b)$ on $(\mathbf{B}, \sigma^2)$. We report computation times of the MCMC sampler for some large-scale synthetic data sets in Appendix C.1 in the supplementary material (Ohn and Kim, 2021) to show its tractability. In addition, we provide trace, autocorrelation and partial autocorrelation plots of the posterior samples of some randomly selected parameters in Appendix C.2 in the supplementary material (Ohn and Kim, 2021) to check the convergence of generated MCMC samples.

For notational simplicity, let $K^* := \mathsf{K}^+(\mathbf{B})$ be the number of nonzero columns of the loading matrix. The detailed algorithm is as follows:

- For posterior sampling of the factor loading β_{jk}, we use the scale mixture representation of the Laplace distribution. If $\beta_{jk}|\tau_{jk} \sim \mathsf{N}(0, \tau_{jk})$ and $\tau_{jk} \sim \mathsf{Exp}(1/2)$, then marginally we have $\beta_{jk} \sim \mathsf{Laplace}(1)$, where $\mathsf{Exp}(1/2)$ denotes the exponential distribution with mean 2. In the MCMC algorithm, we introduce auxiliary scale parameters τ_{jk} for $j \in [p]$ and $k \in [K^*]$ to generate β_{jk} from its conditional posterior distribution easily. For $j \in [p]$ and $k \in [K^*]$, the factor loading β_{jk} is sampled from the conditional posterior

$$
\beta_{jk}|- \sim \begin{cases} \mathsf{N}(\widehat{\beta}_{jk}, \widehat{\tau}_{jk}) & \text{if } \xi_{jk} = 1, \\ \delta_0 & \text{if } \xi_{jk} = 0, \end{cases}
$$

where

$$
\widehat{\tau}_{jk} := \left(\sigma^{-2} \sum_{i=1}^{n} Z_{ik}^2 + \tau_{jk}^{-1} \right)^{-1},
$$

$$
\widehat{\beta}_{jk} := \widehat{\tau}_{jk} \left\{ \sigma^{-2} \sum_{i=1}^{n} Z_{ik} \left(Y_{ij} - \sum_{h \in [K^*]: h \neq k} Z_{ih} \beta_{jh} \right) \right\}.
$$

- For $j \in [p]$ and $k \in [K^*]$, the auxiliary scale parameter τ_{jk} is sampled from

$$
\tau_{jk}|- \sim \begin{cases} \mathsf{GIG}(1, \beta_{jk}^2, \frac{1}{2}) & \text{if } \xi_{jk} = 1, \\ \mathsf{Exp}(\frac{1}{2}) & \text{if } \xi_{jk} = 0, \end{cases}
$$

where $\mathsf{GIG}(a, b, p)$ denotes the generalized inverse Gaussian (GIG) distribution with the density $f(z)$ proportional to $f(z) \propto z^{p-1} \mathrm{e}^{-(az+b/z)/2}$.

- For $j \in [p]$, the indicator parameters $(\xi_{jk} : k \in \mathbb{N})$ are sampled as follows. For $k \in [K^*]$, ξ_{jk} is sampled with probability

$$
\frac{\Pi(\xi_{jk} = 1|-)}{\Pi(\xi_{jk} = 0|-)} = \frac{p_{jk}}{\kappa + p - p_{jk}} \sqrt{\frac{\widehat{\tau}_{jk}}{\tau_{jk}}} \exp\left(\frac{1}{2\widehat{\tau}_{jk}} \widehat{\beta}_{jk}^2 \right),
$$

where $p_{jk} := \sum_{l \in [p]: l \neq j} \xi_{lk}$. For $k > K^*$, we first propose $K_j^* \in \mathbb{N} \cup \{0\}$ and $\boldsymbol{\beta}_j^* := (\beta_{j,1}^*, \ldots, \beta_{j,K_j^*}^*) \in \mathbb{R}^{K_j^*}$ from the proposal distribution

$$
J(K_j^*) J(\beta_j^* | K_j^*) = \mathsf{Poisson}(1) \mathsf{Laplace}(1)^{K_j^*}.
$$

Then we accept the proposal with probability

$$
\max\left\{ 1, |\mathbf{M}_j|^{-n/2} \mathrm{e}^{u_j^2 (\boldsymbol{\beta}_j^*)^\top \mathbf{M}_j^{-1} \boldsymbol{\beta}_j^* / 2} \left(\frac{\alpha}{p} \right)^{K_j^*} \right\},
$$

where we denote $\mathbf{M}_j := \sigma^{-2}\boldsymbol{\beta}_j^*(\boldsymbol{\beta}_j^*)^\top + \mathbf{I}$ and $u_j := \sigma^{-2}\sum_{i=1}^n(Y_{ij} - \sum_{k=1}^{K^*} Z_{ik}\beta_{jk})$. If the proposal is accepted, we update

$$\mathbf{B} \leftarrow (\mathbf{B}, (\beta_{j,k}^* \mathbb{1}(\ell = j))_{\ell \in [p], k \in [K_j^*]}),$$
$$K^* \leftarrow K^* + K_j^*.$$

- For $i \in [n]$, the latent variable $\mathbf{Z}_i$ is sampled from

$$\mathbf{Z}_i|- \sim \mathsf{N}\big(\sigma^{-2}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}}\mathbf{B}^\top\mathbf{Y}_i, \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}}\big),$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z}} := (\sigma^{-2}\mathbf{B}^\top\mathbf{B} + \mathbf{I})^{-1}$.

- The noise variance parameter σ^2 is sampled from

$$\sigma^2|- \sim \mathsf{IG}\left(a + \frac{np}{2}, b + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^p\left(Y_{ij} - \sum_{k=1}^{K^*} Z_{ik}\beta_{jk}\right)^2\right).$$

5 Simulation study

We conduct numerical experiments that illustrate our theoretical findings. The posterior distribution is estimated by the MCMC algorithm presented in Section 4. For each posterior computation, we run the MCMC sampler for 3,000 draws, discard the first 500 samples as burn-in and save every 10th sample.

5.1 Choice of hyperparmeters

We compare the concentrations of the posterior distributions of the factor dimensionality and the covariance matrix with various choices of the hyperparameters α_n and κ_n by simulation.

We let the sample size n vary among $\{100, 200, 300\}$. For each sample size n, we let the dimension of the data be equal to $p_n = 10n$ (i.e., $p_n = 1000, 2000, 3000$) and the column-wise maximum sparsity be equal to $s_n = 3\lfloor\log p_n\rfloor$ (i.e., $s_n = 18, 21, 24$). We consider the bounded factor dimensionality cases where the true factor dimensionality k_{0n} is either 1 or 5. For given s_n and k_{0n}, the $p_n \times k_{0n}$ true loading matrix $\mathbf{B}_{0n}$ is generated by selecting s_n nonzero elements randomly of each column and then sampling their values from $\{-2, 2\}$ randomly. A noise variance σ_{0n}^2 is set to be 1. Simulated data are generated from the Gaussian distribution with the mean 0 and covariance matrix $\boldsymbol{\Sigma}_{0n} := \mathbf{B}_{0n}\mathbf{B}_{0n}^\top + \mathbf{I}$.

We consider two different choices of α_n. The first one is $\alpha_n = p_n^{-1}$, which is used by Rockova and George (2016), and the other one is $\alpha_n = p_n^{-s_n}$ motivated by our theoretical result in Lemma 3.3. For κ_n, we consider two different choices: $\kappa_n = 0$, which corresponds to the one-parameter IBP, and $\kappa_n = p_n^{1+\delta}$ with $\delta = 0.1$, which is the prior considered in this paper.

We investigate the two quantities: the posterior probability of the correct selection of the factor dimensionality (i.e., $\Pi_n(\mathsf{K}^+(\mathbf{B}) = k_{0n}|\mathbf{Y}_{1:n})$) and the scaled spectral norm

k_{0n}	(n, p_n)	$\alpha_n = p_n^{-1}$		$\alpha_n = p_n^{-s_n}$	
		$\kappa_n = 0$	$\kappa_n = p_n^{1.1}$	$\kappa_n = 0$	$\kappa_n = p_n^{1.1}$
	$(100, 1000)$	0 (0.001)	0.005 (0.007)	0.134 (0.105)	0.256 (0.128)
1	$(200, 2000)$	0 (0)	0.001 (0.002)	0.461 (0.106)	0.54 (0.137)
	$(300, 3000)$	0 (0)	0 (0.001)	0.961 (0.142)	0.967 (0.093)
	$(100, 1000)$	0.008 (0.007)	0.043 (0.016)	0.12 (0.047)	0.228 (0.082)
5	$(200, 2000)$	0.004 (0.004)	0.015 (0.008)	0.504 (0.077)	0.558 (0.109)
	$(300, 3000)$	0.002 (0.003)	0.004 (0.004)	0.927 (0.113)	0.98 (0.067)

Table 1: The averages and standard errors of the posterior probabilities of the correct selection of the factor dimensionality over 100 simulations.

k_{0n}	(n, p_n)	$\alpha_n = p_n^{-1}$		$\alpha_n = p_n^{-s_n}$	
		$\kappa_n = 0$	$\kappa_n = p_n^{1.1}$	$\kappa_n = 0$	$\kappa_n = p_n^{1.1}$
	$(100, 1000)$	0.158 (0.071)	0.17 (0.069)	0.157 (0.069)	0.167 (0.075)
1	$(200, 2000)$	0.099 (0.051)	0.103 (0.058)	0.093 (0.049)	0.098 (0.057)
	$(300, 3000)$	0.098 (0.06)	0.099 (0.065)	0.094 (0.067)	0.097 (0.063)
	$(100, 1000)$	0.216 (0.064)	0.238 (0.058)	0.222 (0.053)	0.238 (0.056)
5	$(200, 2000)$	0.165 (0.047)	0.175 (0.053)	0.159 (0.049)	0.177 (0.05)
	$(300, 3000)$	0.152 (0.043)	0.153 (0.046)	0.151 (0.043)	0.158 (0.044)

Table 2: The averages and standard errors of the scaled spectral norm losses of the covariance matrix over 100 simulations.

loss of the covariance matrix (i.e., $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}_{0n}\|/\|\mathbf{\Sigma}_{0n}\|$, where $\hat{\mathbf{\Sigma}}$ denotes the posterior mean of the covariance matrix). We generate 100 independent data sets to have 100 measures of these two quantities and report the averages of those quantities in Tables 1 and 2, respectively.

Table 1 presents the averaged posterior probability of the factor dimensionality being equal to the true one by the four choices of the hyperparameters α_n and κ_n. The results indicate that the choice of both α_n and κ_n is important for the posterior consistency of the factor dimensionality, but the choice of α_n is more important. For the choice of $\alpha_n = p_n^{-1}$, we found that the posterior distribution puts most of its mass to larger values of the factor dimensionality than the true one. Table 2 presents the averaged scaled spectral norm losses for covariance matrix estimation. The results are similar for all the choices of the hyperparameters. It supports our theoretical results that the posterior contraction rates are the same for the four choices.

5.2 Comparison with other methods

In this simulation, we compare finite sample performance of the SSIBP prior with other methods. We use the hyperparameters $\alpha_n = p_n^{-25}$ and $\kappa_n = p_n^{1.1}$ for the SSIBP prior. For factor dimensionality estimation, we use the posterior mode as the point estimator and for covariance matrix estimation, we take the posterior mean as the point estimator.

For synthetic data generation, we consider the following two challenging scenarios.

Case 1 (Low sample size) Generate data with sample size $n = 50$ and dimension $p_n = 1000$. For a true covariance matrix, we consider two values of the column-wise maximum sparsity $s_n \in \{5, 20\}$ and three values of the factor dimensionality $k_{0n} \in \{1, 3, 5\}$. As in Section 5.1, nonzero loading values are sampled from $\{-2, 2\}$ and the noise variance is set to be 1.

Case 2 (Large noise variance) Generate data with larger sample size $n = 150$ and dimension $p_n = 1000$. But in this case the noise variance is set to 3, three times larger than that is Case 1. The sparsity and true factor dimensionality values are the same as the Case 1.

For each setup, we repeat the data generation 100 times and compute the point estimators of the factor dimensionality and the covariance matrix by the SSIBP prior and other competing methods for each synthetic data set.

For factor dimensionality estimation, we consider the following three estimators as competitors:

- Eigenvalue threshold estimator (ET, Onatski (2010))

$$\hat{k}_{\mathrm{ET}} := \sum_{j=1}^{p} \mathbb{1}\big(\lambda_j(\mathbf{S}) > w\lambda_{k_{\max}+1}(\mathbf{S}) + (1-w)\lambda_{2k_{\max}+1}(\mathbf{S})\big),$$

  with $w := 2^{2/3}(2^{2/3} - 1)$, where $\mathbf{S}$ denotes the sample covariance matrix and $k_{\max} \in [p]$ does the pre-specified upper bound of the factor dimensionality.

- Eigenvalue ratio estimator (ER, Ahn and Horenstein (2013); Lam and Yao (2012)):

$$\hat{k}_{\mathrm{ER}} := \operatorname*{argmax}_{j \in [k_{\max}]} \lambda_j(\mathbf{S})/\lambda_{j+1}(\mathbf{S}).$$

- Multiplicative gamma process shrinkage prior (MGPS, Bhattacharya and Dunson (2011)): The posterior mode of the sparse Bayesian factor model with the multiplicative gamma process shrinkage prior.

For covariance estimation, we consider the following three competitors: the principal orthogonal complement thresholding method (POET, Fan et al. (2013)), the variational inference method for Bayesian sparse PCA (VI, Ning (2021)) as well as the MGPS (Bhattacharya and Dunson, 2011). For the POET and VI, an user must select the factor dimensionality a priori. We set the true value of the factor dimensionality for this. We use the posterior mean of the covariance matrix as the point estimator for the MGPS.

The simulation results of factor dimensionality estimation for Cases 1 and 2 are summarized in Table 3 and Table 4, respectively. The SSIBP prior performs very well for all the setups. The frequentist method ET also performs well in general but does not for the setup with $s_n = 5$, $k_{0n} = 5$ and $\sigma_{0n}^2 = 3$. The frequentist method ER tends to underestimate the factor dimensionality for the extremely sparse setups (i.e., $s_n = 5$) in both Cases 1 and 2. The Bayesian method MGPS is suboptimal in all the setups.

k_{0n}	s_n		ET	ER	MGPS	SSIBP
1	5	True	89	100	0	100
		Over	11	0	100	0
		Under	0	0	0	0
		Ave	1.12	1	15.34	1
	20	True	83	100	0	97
		Over	17	0	100	3
		Under	0	0	0	0
		Ave	1.17	1	15.25	1.03
3	5	True	97	59	1	100
		Over	1	0	99	0
		Under	2	41	0	0
		Ave	2.99	2.38	15.32	3
	20	True	99	98	0	79
		Over	1	0	100	21
		Under	0	2	0	0
		Ave	3.01	2.98	14.75	3.21
5	5	True	80	37	3	99
		Over	0	0	92	0
		Under	20	63	5	1
		Ave	4.79	3.13	16.05	4.99
	20	True	100	100	5	86
		Over	0	0	87	14
		Under	0	0	8	0
		Ave	5	5	14.18	5.14

Table 3: Proportions of the estimated factor dimensionality over 100 simulations for Case 1. "True", "Over" and "Under" denote proportions of correct estimation, overestimation and underestimation, respectively. "Ave" is the average of the estimated factor dimensionality.

Table 5 and Table 6 present the simulation results of covariance matrix estimation for Cases 1 and 2, respectively. For each method, we report the average of the scaled spectral norm losses $\|\hat{\Sigma} - \Sigma_0\|/\|\Sigma_0\|$ for the true covariance matrix Σ_0 and the estimate $\hat{\Sigma}$ over 100 simulations. The results show that the SSIBP prior outperforms the other competing methods. The VI method behaves reasonably well, however the overall performance is inferior to the SSIBP prior.

6   Real data analysis

In this section, we analyze the big five personality traits (Goldberg, 1990) data set which we obtained from https://www.kaggle.com/tunguz/big-five-personality-test. This data set contains answers of 50 questions about personality with the five-level Likert scale. The 50 questions can be grouped into five groups which represent the five personality traits of extraversion, agreeableness, openness, conscientiousness, and neuroticism.

k_{0n}	s_n		ET	ER	MGPS	SSIBP
1	5	True	84	100	3	100
		Over	16	0	97	0
		Under	0	0	0	0
		Ave	1.16	1	2.82	1
	20	True	89	100	3	100
		Over	11	0	97	0
		Under	0	0	0	0
		Ave	1.11	1	3.66	1
3	5	True	97	60	17	98
		Over	3	0	27	0
		Under	0	40	56	2
		Ave	3.03	2.44	3.5	2.98
	20	True	98	100	15	100
		Over	2	0	11	0
		Under	0	0	74	0
		Ave	3.02	3	2.44	3
5	5	True	59	42	5	98
		Over	0	0	13	0
		Under	41	58	82	2
		Ave	4.57	3.36	3.5	4.98
	20	True	100	100	0	100
		Over	0	0	4	0
		Under	0	0	96	0
		Ave	5	5	2.6	5

Table 4: Same as Table 3 but for Case 2.

Therefore we expect that the true factor dimensionality is 5 and each element of the factor represents the one of these five personality traits. The total number of responses are 1,015,342 but the respondents who completely answered are 874,434.

We repeat the following procedure 1,000 times. We first randomly select 200 responses and normalize them to have the mean of each question be 0. We then estimate the factor dimensionality by the four methods we considered in Section 5.2, i.e., eigenvalue threshold estimator (ET, Onatski (2010)), eigenvalue ratio estimator (ER, Ahn and Horenstein (2013); Lam and Yao (2012)), multiplicative gamma process shrinkage prior (MGPS, Bhattacharya and Dunson (2011)) and the SSIBP prior. The hyperparameters of the SSIBP prior are set to be $\alpha_n = p_n^{-20}$ and $\kappa_n = p_n^{1.1}$. As in Section 5.2, posterior modes are used as the point estimator of the factor dimensionality for the Bayesian methods.

The estimation results of the factor dimensionality by the four methods are summarized in Figure 1. In view of the prior knowledge that the true factor dimensionality is 5, we can see that the ET and SSIBP prior perform better than the other two methods. The ER method has the large proportion of underestimation and the MGPS always yields factor dimensionality estimates larger than 5.

k_{0n}	s_n	POET	VI	MGPS	SSIBP
1	5	1.642 (0.14)	0.183 (0.105)	1.906 (1.005)	0.182 (0.091)
	20	0.662 (0.124)	0.196 (0.111)	0.612 (0.258)	0.23 (0.122)
3	5	1.794 (0.146)	0.435 (0.119)	2.039 (1.191)	0.239 (0.078)
	20	0.787 (0.137)	0.451 (0.125)	0.691 (0.394)	0.364 (0.126)
5	5	1.96 (0.198)	0.722 (0.16)	2.092 (0.979)	0.297 (0.108)
	20	0.917 (0.151)	0.69 (0.125)	0.791 (0.468)	0.356 (0.092)

Table 5: The averages and standard errors of the scaled spectral norm losses of the covariance matrix over 100 simulations for Case 1.

k_{0n}	s_n	POET	VI	MGPS	SSIBP
1	5	1.708 (0.071)	0.115 (0.056)	0.975 (0.622)	0.12 (0.053)
	20	0.667 (0.09)	0.142 (0.073)	1.026 (0.706)	0.154 (0.06)
3	5	1.795 (0.104)	0.318 (0.1)	1.108 (0.668)	0.189 (0.114)
	20	0.751 (0.079)	0.341 (0.088)	0.923 (0.56)	0.198 (0.06)
5	5	1.83 (0.145)	0.535 (0.11)	1.142 (0.595)	0.21 (0.106)
	20	0.775 (0.072)	0.545 (0.087)	0.847 (0.343)	0.209 (0.05)

Table 6: Same as Table 5 but for Case 2.

In Figure 2, we compare the estimated loading matrices of the two Bayesian methods. In the figure, we plot the posterior means of the loading matrix given that the factor dimensionality is equal to the posterior mode, that is, $\mathbb{E}(\mathbf{B}|\mathsf{K}^+(\mathbf{B}) = \hat{k}, \mathbf{Y}_{1:n})$ where $\hat{k} = \operatorname{argmax}_{k \in \mathbb{N}} \Pi_n(\mathsf{K}^+(\mathbf{B}) = k|\mathbf{Y}_{1:n})$. Recall that the total 50 questions in the data set are grouped so that 10 questions in each group correspond to each personality trait. The SSIBP prior captures this structure quite well, while the MGPS prior does not recover it.

7 Concluding remarks

In this paper, we proposed a computationally tractable prior which has desirable large sample properties. The proposed prior achieves the posterior consistency of both the factor dimensionality and induced covariance matrix. We also derived the posterior contraction rate of the covariance matrix, which is a near-optimal rate. Moreover, if the true factor dimensionality is bounded, we obtained the posterior consistency of the factor dimensionality and the optimal convergence rate of the covariance matrix simultaneously. In addition, we provided the MCMC algorithm and showed that it works well with (moderately) large data sets.

Our MCMC algorithm is applicable without much hamper to moderately large data but would be difficult to use for very large data. Computationally more efficient inferential methods such as variational inference for big data are worth pursuing. Ning (2021) proposed the variational inference algorithm for the sparse Bayesian factor model. However, the algorithm of Ning (2021) does not infer the factor dimensionality, but it fixes the factor dimensionality given by an user. Designing a variational inference algorithm which can infer the factor dimensionality would be an interesting future work.
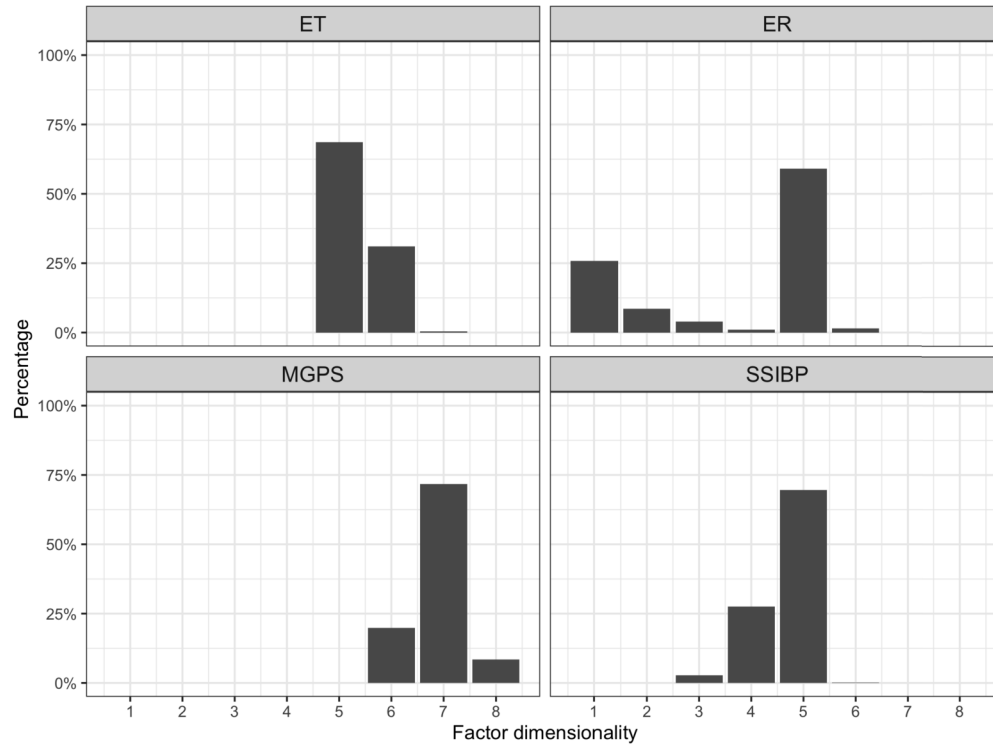
Figure 1: Distributions of the factor dimensionality estimates over 1,000 data sets of size 200 randomly selected from the big five personality traits data.

The selection consistency of entries in the loading matrix is also an important problem. Since it is devised for selecting the factor dimensionality, our prior may not achieve the entry-level selection consistency. Once the true factor dimensionality is known, the problem is similar to the high-dimensional regression problem, but the latent factor is random and not observable. Thus techniques to prove the selection consistency of high-dimensional regression models such as those of Castillo et al. (2015) may not be applicable directly. To check the entry-level selection consistency of our prior or to develop a prior which achieves the entry-level selection consistency would require completely new researches.

The proposed prior distribution is nonadaptive in the sense that the true sparsity level s_n should be known in advance to choose the hyperparameters optimally. In practice, the hyperparameters could be selected by the maximum marginal likelihood (Rousseau and Szabo, 2017), which would be computationally demanding and not theoretically guaranteed. It would be an interesting future work to develop a computationally tractable prior which is adaptive to the true sparsity level s_n.

In an ultra high dimensional setup in which the dimension p_n grows exponentially in n, the number of nonzero rows s_n of the loading matrix should be of order $\log p_n$.

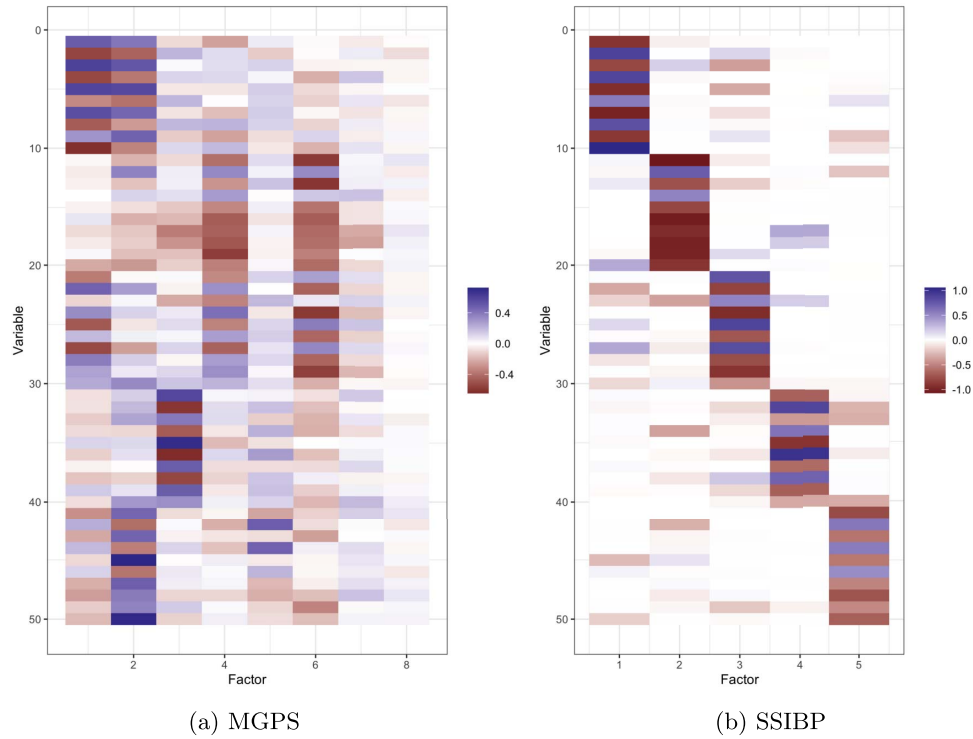(a) MGPS   (b) SSIBP

Figure 2: Posterior means of the loading matrix for a subsampled big five personality traits data set.

This order would be too small compared to the dimension p_n because $(p_n - \log p_n)$ many entries of the random vector $\mathbf{Y}$ should be pure noises. To resolve this problem, a G-block covariance model by Bunea et al. (2020) could be an alternative, but there is no Bayesian inference method with theoretical guarantees. We leave this problem as a future work.

Supplementary Material

Supplementary Material to "Posterior consistency of factor dimensionality in high-dimensional sparse factor models" (DOI: 10.1214/21-BA1261SUPP; .pdf).

References

Ahn, S. C. and Horenstein, A. R. (2013). "Eigenvalue ratio test for the number of factors." *Econometrica*, 81(3): 1203–1227. MR3064065. doi: https://doi.org/10.3982/ECTA8968.   492, 505, 507

Bai, J. and Ng, S. (2002). "Determining the number of factors in approximate factor models." *Econometrica*, 70(1): 191–221. MR1926259. doi: https://doi.org/10.1111/1468-0262.00273. 492

Bai, J. and Ng, S. (2007). "Determining the number of primitive shocks in factor models." *Journal of Business & Economic Statistics*, 25(1): 52–60. MR2338870. doi: https://doi.org/10.1198/073500106000000413. 492

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach." *The Quarterly Journal of Economics*, 120(1): 387–422. 491

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 492, 505, 507

Bunea, F., Giraud, C., Luo, X., Royer, M., and Verzelen, N. (2020). "Model-assisted variable clustering: minimax-optimal recovery and algorithms." *The Annals of Statistics*. MR4065155. doi: https://doi.org/10.1214/18-AOS1794. 510

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. 498, 500

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). "High-dimensional sparse factor modeling: applications in gene expression genomics." *Journal of the American Statistical Association*, 103(484): 1438–1456. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 491

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 509

Castillo, I. and van der Vaart, A. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 499

Chen, B., Chen, M., Paisley, J., Zaas, A., Woods, C., Ginsburg, G. S., Hero, A., Lucas, J., Dunson, D., and Carin, L. (2010). "Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies." *BMC Bioinformatics*, 11(1): 552. 492, 493, 495

Fan, J., Fan, Y., and Lv, J. (2008). "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics*, 147(1): 186–197. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 491

Fan, J., Han, X., and Gu, W. (2012). "Estimating false discovery proportion under arbitrary covariance dependence." *Journal of the American Statistical Association*, 107(499): 1019–1035. MR3010887. doi: https://doi.org/10.1080/01621459.2012.720478. 491

Fan, J., Ke, Y., Sun, Q., and Zhou, W. X. (2019). "FarmTest: Factor-Adjusted Ro-

bust Multiple Testing With Approximate False Discovery Control." *Journal of the American Statistical Association*, 114(528): 1880–1893. MR4047307. doi: https://doi.org/10.1080/01621459.2018.1527700. 491

Fan, J., Liao, Y., and Mincheva, M. (2011). "High dimensional covariance matrix estimation in approximate factor models." *The Annals of Statistics*, 39(6): 3320–3356. MR3012410. doi: https://doi.org/10.1214/11-AOS944. 491

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). 505

Fan, J., Liu, H., Wang, W., et al. (2018). "Large covariance estimation through elliptical factor models." *The Annals of Statistics*, 46(4): 1383–1414. MR3819104. doi: https://doi.org/10.1214/17-AOS1588. 491

Fan, J., Xue, L., and Yao, J. (2017). "Sufficient forecasting using factor models." *Journal of Econometrics*, 201(2): 292–306. MR3717565. doi: https://doi.org/10.1016/j.jeconom.2017.08.009. 491

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). "Do financial variables help forecasting inflation and real activity in the euro area?" *Journal of Monetary Economics*, 50(6): 1243–1255. 491

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. 492, 494, 497, 498, 500

Ghahramani, Z. and Griffiths, T. L. (2005). "Infinite latent feature models and the Indian buffet process." In *Advances in Neural Information Processing Systems*, 475–482. 492, 493, 495

Goldberg, L. R. (1990). "An alternative "description of personality": the big-five factor structure." *Journal of Personality and Social Psychology*, 59(6): 1216. MR0094249. 506

Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). "A new summarization method for Affymetrix probe level data." *Bioinformatics*, 22(8): 943–949. 491

Kneip, A. and Sarda, P. (2011). "Factor models and variable selection in high-dimensional regression analysis." *The Annals of Statistics*, 39(5): 2410–2447. MR2906873. doi: https://doi.org/10.1214/11-AOS905. 491

Knowles, D., Ghahramani, Z., et al. (2011). "Nonparametric Bayesian sparse factor models with application to gene expression modeling." *The Annals of Applied Statistics*, 5(2B): 1534–1552. MR2849785. doi: https://doi.org/10.1214/10-AOAS435. 492, 495

Lam, C. and Yao, Q. (2012). "Factor modeling for high-dimensional time series: inference for the number of factors." *The Annals of Statistics*, 40(2): 694–726. MR2933663. doi: https://doi.org/10.1214/12-AOS970. 492, 505, 507

Latala, R. (2005). "Some estimates of norms of random matrices." *Proceedings of the American Mathematical Society*, 133(5): 1273–1282. MR2111932. doi: https://doi.org/10.1090/S0002-9939-04-07800-1. 494

Leek, J. T. and Storey, J. D. (2008). "A general framework for multiple testing dependence." *Proceedings of the National Academy of Sciences*, 105(48): 18718–18723. 491

Li, Z., Wang, Q., Yao, J., et al. (2017). "Identifying the number of factors from singular values of a large sample auto-covariance matrix." *The Annals of Statistics*, 45(1): 257–288. MR3611492. doi: https://doi.org/10.1214/16-AOS1452. 500

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 499

McCrae, R. R. and John, O. P. (1992). "An introduction to the five-factor model and its applications." *Journal of Personality*, 60(2): 175–215. 491

Ning, B. (2021). "Spike and slab Bayesian sparse principal component analysis." *arXiv preprint* arXiv:2102.00305. 492, 498, 505, 508

Ohn, I. and Kim, Y. (2021). "Supplementary Material to "Posterior consistency of factor dimensionality in high-dimensional sparse factor models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1261SUPP. 495, 497, 499, 500, 501

Onatski, A. (2010). "Determining the number of factors from empirical distribution of eigenvalues." *The Review of Economics and Statistics*, 92(4): 1004–1016. 492, 500, 505, 507

Paisley, J. and Carin, L. (2009). "Nonparametric factor analysis with beta process priors." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 777–784. ACM. 493, 495

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. 492, 494, 497, 498, 501

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. 492, 494, 495, 497, 498, 499, 503

Rousseau, J. and Szabo, B. (2017). "Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator." *The Annals of Statistics*, 45(2): 833–865. MR3650402. doi: https://doi.org/10.1214/16-AOS1469. 509

Silva, A. P. D. (2011). "Two-group classification with high-dimensional correlated data: A factor model approach." *Computational Statistics & Data Analysis*, 55(11): 2975–2990. MR2813060. doi: https://doi.org/10.1016/j.csda.2011.05.002. 491

Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017). "Expandable factor analysis." *Biometrika*, 104(3): 649–663. MR3694588. doi: https://doi.org/10.1093/biomet/asx030.   492

Stock, J. H. and Watson, M. W. (2002). "Forecasting using principal components from a large number of predictors." *Journal of the American Statistical Association*, 97(460): 1167–1179. MR1951271. doi: https://doi.org/10.1198/016214502388618960.   491

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). "Stick-breaking construction for the Indian buffet process." In *Artificial Intelligence and Statistics*, 556–563.   495

Xie, F., Xu, Y., Priebe, C. E., and Cape, J. (2018). "Bayesian estimation of sparse spiked covariance matrices in high dimensions." *arXiv preprint* arXiv:1808.07433.   492, 494, 498