A Bayesian Sparse Factor Model with Adaptive Posterior Concentration[*]

Ilsang Ohn[†], Lizhen Lin[‡] and Yongdai Kim[§]

Abstract. In this paper, we propose a new Bayesian inference method for a high-dimensional sparse factor model that allows both the factor dimensionality and the sparse structure of the loading matrix to be inferred. The novelty is to introduce a certain dependence between the sparsity level and the factor dimensionality, which leads to adaptive posterior concentration while keeping computational tractability. We show that the posterior distribution asymptotically concentrates on the true factor dimensionality, and more importantly, this posterior consistency is adaptive to the sparsity level of the true loading matrix and the noise variance. We also prove that the proposed Bayesian model attains the optimal detection rate of the factor dimensionality in a more general situation than those found in the literature. Moreover, we obtain a near-optimal posterior concentration rate of the covariance matrix. Numerical studies are conducted and show the superiority of the proposed method compared with other competitors.

MSC2020 subject classifications: Primary 62F15; secondary 62G20.

Keywords: adaptive inference, sparse factor models, factor dimensionality, covariance matrix estimation, optimal posterior concentration rate.

1 Introduction

In this paper, we propose a novel Bayesian method for learning a high-dimensional sparse linear factor model and study asymptotic concentration properties of the posterior distribution. We consider the following linear factor model where p-dimensional random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are distributed as

$$\mathbf{Y}_i | (\mathbf{Z}_i = \mathbf{z}_i) \stackrel{\text{ind}}{\sim} \mathrm{N}(\mathbf{B}\mathbf{z}_i, \psi\mathbf{I}), \quad \mathbf{Z}_i \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}) \tag{1.1}$$

for $i \in [n]$ with $\mathbf{B}$ representing a $p \times \xi$ factor loading matrix, $\psi > 0$ a noise variance and $\mathbf{Z}_i$ a ξ-dimensional (latent) factor related to datum $\mathbf{Y}_i$, where $\xi \in \{1, \ldots, p-1\}$. Under this model, the marginal distribution of the data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is given by

$$\mathbf{Y}_i \stackrel{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} := \mathbf{B}\mathbf{B}^\top + \psi\mathbf{I}.$$

[†]Department of Statistics, Inha University

[‡]Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame

[§]Department of Statistics, Seoul National University. Corresponding author, ydkim0903@gmail.com

Therefore, correlations among the observed variables in each $\mathbf{Y}_i$ in the above factor model are explained by a low rank matrix $\mathbf{BB}^\top$ which leads to a substantial but efficient reduction of the model complexity. The factor model has been used in a broad range of high-dimensional inference tasks including covariance estimation (Fan et al., 2008, 2011, 2018b), linear regression (Bai and Ng, 2006; Kneip and Sarda, 2011; Stock and Watson, 2002), multiple testing under arbitrary dependence (Fan et al., 2012, 2018a; Leek and Storey, 2008) and other supervised learning tasks (Fan et al., 2017; Silva, 2011).

For high-dimensional data where the dimension p is much large than the sample size n, we need a low dimensional structure on $\mathbf{B}$ to have a consistent estimator and *sparsity* is a popularly used condition, which assumes that true loading matrix $\mathbf{B}^\star$ is sparse in the sense that only few entries of $\mathbf{B}^\star$ is nonzero and the other entries are zero. There are various Bayesian models for sparse linear factor models including Bhattacharya and Dunson (2011); Srivastava et al. (2017); Xie et al. (2022); Ning (2021).

Along with considering the sparsity, determining the *factor dimensionality* ξ is an important and practical topic in factor modeling. From a theoretical point of view, an appropriate estimation of the factor dimensionality is required to optimize the bias-variance trade-off in a factor model. The factor dimensionality is also of practical interest, especially when it has a physical interpretation e.g., the number of interacting pathways in genomics (Carvalho et al., 2008) and the number of personality traits in psychology (Caprara et al., 1993).

Frequentist approaches typically adopt a two-step procedure where the factor dimensionality is chosen or estimated before estimating the parameters in the model. Many consistent model selection methods have been proposed, which fit the factor models for different values of ξ and select the best ξ based on their choice of model selection criteria (Bai and Ng, 2002, 2007). Alternatively, the eigenvalues of the empirical covariance or correlation matrix can be used to estimate the factor dimensionality. Several procedures related to this approach have been proposed and yielded consistency (Ahn and Horenstein, 2013; Fan et al., 2020; Lam and Yao, 2012; Onatski, 2010). The estimation of the factor dimensionality for high-dimensional sparse factor models has also been considered by Cai et al. (2013, 2015).

On the Bayesian side, prior distributions which put prior mass on ξ directly have been popularly used to infer the factor dimensionality. Examples are the spike and slab prior with the Poisson prior on ξ (Pati et al., 2014) and spike and slap priors with the Indian buffet process (IBP) (Chen et al., 2010; Knowles and Ghahramani, 2011; Rockova and George, 2016; Ohn and Kim, 2022). Recently, a few works have provided theoretical results for the posterior distribution of the factor dimensionality under nonparametric priors. Rockova and George (2016) considered a spike and slab prior with the one-parameter IBP and proved that the posterior probability of the factor dimensionality being upper bounded by a certain quantity converges to 1. But the upper bound is much larger than the true factor dimensionality. Ohn and Kim (2022) derived posterior consistency of the factor dimensionality under a spike and slab prior with the two-parameter IBP. However, their result is nonadaptive in the sense that the choice of hyperparameters of the prior distribution relies on the information or knowledge of the true sparsity of the loading matrix, which is unknown in practice.

Another promising theoretical result was provided by Gao and Zhou (2015). The authors studied the Bayesian factor model in the context of sparse principle component analysis (PCA) in which the prior distribution concentrates its mass on the orthogonal loading matrix, i.e., $\mathbf{B}^\top \mathbf{B}$ is diagonal, and established the adaptive posterior consistency of the factor dimensionality. However, the orthogonal constraint makes it hard to compute posterior distribution and Gao and Zhou (2015) only succeeded in implementing a posterior sampler for the factor model with a one-dimensional factor i.e., $\xi = 1$.

We propose a novel Bayesian model that overcomes the theoretical and practical limitations of the existing Bayesian approaches. A key feature of the proposed Bayesian model is that the sparsity and the factor dimensionality are negatively correlated under the prior distribution, and this a priori negative correlation between them helps to prevent overestimating the true factor dimensionality. This is a critical difference between the proposed prior and the widely used IBP-type priors. Yet, posterior computation can be carried out through a simple and efficient Monte Carlo Markov chain (MCMC) algorithm. Our numerical studies show that the developed MCMC algorithm can apply to high-dimensional data without many hampers.

We thoroughly investigate the theoretical properties of the posterior distribution of the proposed Bayesian model. We prove that the posterior distribution of the factor dimensionality converges to the true one. In particular, we prove that the proposed Bayesian model attains the optimal detection rate for the eigengap (i.e., the size of the smallest eigenvalue of the low rank part $\mathbf{B}^\star (\mathbf{B}^\star)^\top$ of the true covariance matrix $\boldsymbol{\Sigma}^\star = \mathbf{B}^\star (\mathbf{B}^\star)^\top + \psi^\star \mathbf{I}$) for the consistency of the factor dimensionality. We also show that the posterior distribution of the covariance matrix concentrates around the truth at a near optimal rate. The novelty of our results lies in that it does not require any prior knowledge of the true sparsity and noise variance and hence all the optimal theoretical properties of the posterior are *adaptive* to the sparsity and noise variance.

It should be noted that the proposed Bayesian model has theoretical advantages over not only other Bayesian factor models but also existing frequentist's estimators of the factor dimensionality. The estimator of Cai et al. (2013) is adaptive to the true sparsity but requires a larger detection rate for the eigengap than the optimal one. On the other hand, the estimator of Cai et al. (2015) achieves the optimal detection rate but is not adaptive to the true sparsity. Moreover, both estimators assume the known noise variance, which limits their applicability.

The rest of the paper is organized as follows. In Section 2, we introduce the proposed prior distribution and develop an efficient MCMC algorithm for sampling from the posterior distribution. In Section 3, asymptotic properties of the posterior distribution are derived. In Section 4, we conduct simulation studies and real data analysis. Section 5 concludes the paper.

1.1 Notation

Let $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$ be the sets of real numbers, positive numbers and natural numbers, respectively. Let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of 0's and of 1's, respectively, where the dimensions of such vectors can differ according to the context. For a positive integer p, we let

$[p] := \{1, 2, \ldots, p\}$. For a real number x, $\lfloor x \rfloor$ 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 $a_n \lesssim b_n$ or equivalently $b_n \gtrsim a_n$ 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 $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We denote by $\mathbb{1}(\cdot)$ the indicator function.

For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For a p-dimensional vector $\boldsymbol{\beta} := (\beta_j)_{j \in [p]}$, let $\|\boldsymbol{\beta}\|_r := \left(\sum_{j=1}^{p} |\beta_j|^r \right)^{1/r}$ for $r \geq 1$ and $\|\boldsymbol{\beta}\|_0 := \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0)$. For a set $\mathcal{S} \subset \{1, \ldots, p\}$, define $\boldsymbol{\beta}_{[\mathcal{S}]} := (\beta_j)_{j \in \mathcal{S}}$. For a $p \times q$-dimensional matrix $\mathbf{A} := (a_{jk})_{j \in [p], k \in [q]}$, we denote the spectral norm of the matrix $\mathbf{A}$ by $\|\mathbf{A}\|$ and the Frobenius norm by $\|\mathbf{A}\|_{\mathrm{F}}$, that is, $\|\mathbf{A}\| := \sup_{\mathbf{x} \in \mathbb{R}^q : \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2$ and $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\mathrm{Tr}(\mathbf{A}^\top \mathbf{A})}$. Let $\|\mathbf{A}\|_1$ be the vector ℓ_1 norm of $\mathbf{A}$, i.e., $\|\mathbf{A}\|_1 := \sum_{j=1}^{p} \sum_{k=1}^{q} |a_{jk}|$. For sets $\mathcal{S} \subset [p]$ and $\mathcal{K} \subset [q]$, we let $\mathbf{A}_{[\mathcal{S}, \mathcal{K}]} := (a_{jk})_{j \in \mathcal{S}, k \in \mathcal{K}}$ which is the submatrix of $\mathbf{A}$ taking the rows in $\mathcal{S}$ and columns in $\mathcal{K}$. For notational simplicity, we write $\mathbf{A}_{[:, \mathcal{K}]} := \mathbf{A}_{[[p], \mathcal{K}]}$ and $\mathbf{A}_{[\mathcal{S}, :]} := \mathbf{A}_{[\mathcal{S}, [q]]}$. Furthermore, let $\mathbf{A}_{[j, :]} := \mathbf{A}_{[\{j\}, :]}$ and $\mathbf{A}_{[:, k]} := \mathbf{A}_{[:, \{k\}]}$, which denote the j-th row and k-th column of $\mathbf{A}$, respectively. We let $\lambda_1(\boldsymbol{\Sigma}) \geq \lambda_2(\boldsymbol{\Sigma}) \cdots \geq \lambda_p(\boldsymbol{\Sigma})$ be the ordered eigenvalues and $|\boldsymbol{\Sigma}|$ be the determinant of a $p \times p$-dimensional matrix $\boldsymbol{\Sigma}$. Let $\mathbb{S}_{++}^p$ be a set of $p \times p$ symmetric positive definite matrices.

For a given probability measure G, let P_G denote the probability or the expectation operator under the probability measure G. We denote by p_G the probability density function of G with respect to the Lebesgue measure if exists. For convenience, we write $\mathsf{P}_{\boldsymbol{\Sigma}} := \mathsf{P}_{\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})}$ and $\mathsf{p}_{\boldsymbol{\Sigma}} := \mathsf{p}_{\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})}$ for a normal distribution $\mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$. For $n \in \mathbb{N}$, let $\mathsf{P}_G^{(n)}$ be the probability or the expectation under the product measure and, if exists, $\mathsf{p}_G^{(n)}$ its density function.

2   Proposed prior and MCMC algorithm

In this section, we design a novel prior tailored for the loading matrix in a factor model and develop a computationally efficient MCMC algorithm for sampling the posterior distribution.

2.1   Prior distribution

The proposed prior on the loading matrix first samples the "sparse structure" of the loading matrix and then samples nonzero elements. Let $\mathbf{u} := (u_1, \ldots, u_p)^\top \in \Delta_p := \{0, 1\}^p \setminus \{\mathbf{0}\}$ and, for a positive integer $q \in \mathbb{N}$ we have chosen, $\mathbf{v} := (v_1, \ldots, v_q)^\top \in \Delta_q := \{0, 1\}^q \setminus \{\mathbf{0}\}$. They are latent indicator variables that determine nonzero rows and columns of the loading matrix $\mathbf{B}$, respectively. Note that the sparsity of columns determines the factor dimensionality ξ, i.e., $\xi = \|\mathbf{v}\|_0$ and q is a pre-specified upper bound of the factor dimensionality. We impose the prior distribution on $\mathbf{u}$ and $\mathbf{v}$ such

that

$$\Pi(\mathbf{u}, \mathbf{v}) = \frac{Q_A(\|\mathbf{u}\|_0, \|\mathbf{v}\|_0)}{\sum_{\mathbf{u}' \in \Delta_p} \sum_{\mathbf{v}' \in \Delta_q} Q_A(\|\mathbf{u}'\|_0, \|\mathbf{v}'\|_0)} \mathbb{1}(\mathbf{u} \in \Delta_p, \mathbf{v} \in \Delta_q)$$

$$\text{with } Q_A(\omega, \xi) := Q_{A,p,q,n}(\omega, \xi) := \frac{1}{\binom{p}{\omega}\binom{q}{\xi}} \exp\big(-A\omega\xi \log(p \vee n)\big) \tag{2.1}$$

for some $A > 0$. Under this prior distribution, the non-sparsity $\omega = \|\mathbf{u}\|_0$ and the factor dimensionality $\xi = \|\mathbf{v}\|_0$ are negatively correlated in the sense that ω becomes smaller when ξ is large and vice versa. Note that $\|\mathbf{u}\|_0$ and $\|\mathbf{v}\|_0$ are negatively correlated in the proposed prior (2.1), which sharply contrasts with existing IBP-type priors (Rockova and George, 2016; Ohn and Kim, 2022) that assume the independence of $\|\mathbf{u}\|_0$ and $\|\mathbf{v}\|_0$. The IBP-type priors, however, are known to have the posterior consistency only when the true sparsity is known. We devise the prior (2.1) to achieve an optimal posterior concentration rate even when the true sparsity is unknown.

Conditional on $\mathbf{u}$ and $\mathbf{v}$, we then impose the prior distribution of the loading matrix $\mathbf{B}$ such that

$$\Pi(\mathbf{B} \in \mathcal{B}|\mathbf{u}, \mathbf{v}) = \int_{\mathcal{B}} \prod_{j=1}^{p} \prod_{k=1}^{q} \big[\delta_0(\beta_{jk})\big]^{1-u_j v_k} \Big[\mathsf{p}_{\mathtt{Lap}(1)}(\beta_{jk})\Big]^{u_j v_k} \mathrm{d}\beta_{jk} \tag{2.2}$$

for any measurable set $\mathcal{B} \subset \mathbb{R}^{p \times q}$, where δ_0 denotes the Dirac-delta function at 0 and $\mathtt{Lap}(1)$ does the Laplace distribution with scale 1. That is, independently for each loading β_{jk}, we consider a spike and slab type prior distribution with the Dirac spike and Laplace slab. The use of the Laplace slab, which is more diffused than the normal distribution, is commonly used in Bayesian sparse factor models (Rockova and George, 2016; Xie et al., 2022; Ning, 2021; Ohn and Kim, 2022) in order to reduce bias in the estimation of large loadings.

Lastly, we consider the inverse Gamma prior distribution for ψ such that

$$\Pi(\psi \in \mathcal{B}') = \mathsf{P}_{\mathtt{IG}(\mathbf{a})}(\psi \in \mathcal{B}'). \tag{2.3}$$

for any measurable set $\mathcal{B}' \subset \mathbb{R}$, where $\mathtt{IG}(\mathbf{a})$ with $\mathbf{a} := (a_1, a_2) \in \mathbb{R}_+^2$ denotes the inverse gamma distribution with shape a_1 and rate a_2.

The proposed prior is carefully designed to attain the posterior consistency of the factor dimensionality adaptively to the sparsity, therefore we call the prior defined through (2.1), (2.2) and (2.3) the *adaptive spike and slab (AdaSS)* prior.

2.2 Posterior computation

In this section, we develop an MCMC algorithm to compute the posterior distribution of the parameters $\mathbf{u} := (u_j)_{j \in [p]}$, $\mathbf{v} := (v_k)_{k \in [q]}$, $\mathbf{B} := (\beta_{jk})_{j \in [p], k \in [q]}$, ψ and the latent factors $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. We first introduce additional notations. Let Y_{ij} be the j-th element of $\mathbf{Y}_i$ and Z_{ik} be the k-th element of $\mathbf{Z}_i$. Let $\mathcal{S} := \{j \in [p] : u_j = 1\}$ and $\mathcal{K} := \{k \in [q] :$

$v_k = 1\}$. We use the notation $f(Y|X = x)$ to denote the conditional density of Y given $X = x$.

To make posterior sampling of the factor loading β_{jk} easy, we employ the scale mixture representation of the Laplace distribution. Note that if $\beta_{jk}|\tau_{jk} \sim \mathtt{N}(0, \tau_{jk})$ and $\tau_{jk} \sim \mathtt{Exp}(1/2)$, then marginally we have $\beta_{jk} \sim \mathtt{Lap}(1)$, where $\mathtt{Exp}(1/2)$ stands for the exponential distribution with mean 2. In the MCMC algorithm, we introduce auxiliary scale parameters τ_{jk} for $j \in [p]$ and $k \in [q]$.

Then a single iteration of our proposed MCMC sampler goes as follows.

Sample β_{jk} for $j \in [p]$ and $k \in [q]$: We sample β_{jk} from

$$
\beta_{jk}|- \sim \begin{cases} \mathtt{N}(\hat{\beta}_{jk}, \hat{\tau}_{jk}) & \text{if } u_j = 1 \text{ and } v_k = 1 \\ \delta_0 & \text{otherwise} , \end{cases}
$$

where

$$
\widehat{\beta}_{jk} := \widehat{\tau}_{jk} \left\{ \psi^{-1} \sum_{i=1}^n Z_{ik} \left(Y_{ij} - \sum_{h \in [q]: h \neq k} Z_{ih}\beta_{jh} \right) \right\} \tag{2.4}
$$

$$
\widehat{\tau}_{jk} := \left(\psi^{-1} \sum_{i=1}^n Z_{ik}^2 + \tau_{jk}^{-1} \right)^{-1} . \tag{2.5}
$$

Sample τ_{jk} for $j \in [p]$ and $k \in [q]$: We sample τ_{jk} from

$$
\tau_{jk}|- \sim \begin{cases} \mathtt{GIG}(1, \beta_{jk}^2, \frac{1}{2}) & \text{if } u_j = 1, \\ \mathtt{Exp}(\frac{1}{2}) & \text{otherwise,} \end{cases} \tag{2.6}
$$

where $\mathtt{GIG}(a, b, c)$ denotes the generalized inverse Gaussian distribution with density $\mathtt{p}_{\mathtt{GIG}(a,b,c)}(z) \propto z^{c-1}\mathrm{e}^{-(az+b/z)/2}\mathbb{1}(z > 0)$.

Sample u_j for $j \in [p]$: If $\mathcal{S} \setminus \{j\} = \emptyset$, we set $u_j = 1$. Otherwise, we sample u_j from $\mathtt{Bernoulli}(\rho_j^{\mathrm{row}}/(1 + \rho_j^{\mathrm{row}}))$ with

$$
\rho_j^{\mathrm{row}} := \frac{\Pi(u_j = 1|-)}{\Pi(u_j = 0|-)}
$$

$$
= \frac{Q_A(|\mathcal{S} \setminus \{j\}| + 1, |\mathcal{K}|)}{Q_A(|\mathcal{S} \setminus \{j\}|, |\mathcal{K}|)} \frac{\prod_{i=1}^n f(\mathbf{Y}_i|u_j = 1, \mathbf{v}, \mathbf{T}_{j,\mathcal{K}}, \mathbf{B}_{[-j,:]}, \psi, \mathbf{Z}_i)}{\prod_{i=1}^n f(\mathbf{Y}_i|u_j = 0, \mathbf{B}_{[-j,:]}, \psi, \mathbf{Z}_i)}
$$

$$
= (p \vee n)^{-A|\mathcal{K}|} \frac{\binom{p}{|\mathcal{S}\setminus\{j\}|}}{\binom{p}{|\mathcal{S}\setminus\{j\}|+1}}
$$

$$
\times \frac{\int \prod_{i=1}^n f(\mathbf{Y}_i|\mathbf{B}, \psi, \mathbf{Z}_i) \prod_{k=1}^q \left\{ \mathtt{p}_{\mathtt{N}(0,\tau_{jk})}(\beta_{jk}) \right\}^{v_k} \left\{ \delta_0(\beta_{jk}) \right\}^{1-v_k} \mathrm{d}\beta_{jk}}{\prod_{i=1}^n f(\mathbf{Y}_i|\mathbf{B}_{[j,:]} = \mathbf{0}, \mathbf{B}_{[-j,:]}, \psi, \mathbf{Z}_i)}
$$

$$= (p \vee n)^{-A|\mathcal{K}|} \frac{|\mathcal{S} \setminus \{j\}| + 1}{p - |\mathcal{S} \setminus \{j\}|} \sqrt{\frac{|\hat{\mathbf{T}}_{j,\mathcal{K}}|}{|\mathbf{T}_{j,\mathcal{K}}|}} \exp\left(\frac{1}{2} \hat{\boldsymbol{\beta}}_{j,\mathcal{K}}^\top \hat{\mathbf{T}}_{j,\mathcal{K}}^{-1} \hat{\boldsymbol{\beta}}_{j,\mathcal{K}}\right), \tag{2.7}$$

where $\mathbf{B}_{[-j,:]} := \mathbf{B}_{[[p]\setminus\{j\},:]}$, $\mathbf{T}_{j,\mathcal{K}} := \text{DIAG}((\tau_{jk})_{k\in\mathcal{K}})$, $\hat{\mathbf{T}}_{j,\mathcal{K}} := \text{DIAG}((\hat{\tau}_{jk})_{k\in\mathcal{K}})$ and $\hat{\boldsymbol{\beta}}_{j,\mathcal{K}} := (\hat{\beta}_{jk})_{k\in\mathcal{K}}$.

Sample v_k for $k \in [q]$. If $\mathcal{K} \setminus \{k\} = \emptyset$, we set $v_k = 1$. Otherwise, we sample v_k from $\texttt{Bernoulli}(\rho_k^{\text{col}}/(1 + \rho_k^{\text{col}}))$ with

$$
\begin{aligned}
\rho_k^{\text{col}} &:= \frac{\Pi(v_k = 1|-)}{\Pi(v_k = 0|-)} \\
&= \frac{Q_A(|\mathcal{S}|, |\mathcal{K} \setminus \{k\}| + 1)}{Q_A(|\mathcal{S}|, |\mathcal{K} \setminus \{k\}|)} \frac{\prod_{i=1}^n f(\mathbf{Y}_i | v_k = 1, \mathbf{u}, \mathbf{T}_{\mathcal{S},k}, \mathbf{B}_{[:,-k]}, \psi, \mathbf{Z}_i)}{\prod_{i=1}^n f(\mathbf{Y}_i | v_k = 0, \mathbf{B}_{[:,-k]}, \psi, \mathbf{Z}_i)} \\
&= (p \vee n) p^{-A|\mathcal{S}|} \frac{\binom{q}{|\mathcal{K}\setminus\{k\}|}}{\binom{q}{|\mathcal{K}\setminus\{k\}|+1}} \\
&\quad \times \frac{\int \prod_{i=1}^n f(\mathbf{Y}_i | \mathbf{B}, \psi, \mathbf{Z}_i) \prod_{j=1}^p \left\{ \mathtt{p}_{\mathtt{N}(0,\tau_{jk})}(\beta_{jk}) \right\}^{u_j} \left\{ \delta_0(\beta_{jk}) \right\}^{1-u_j} \, \mathrm{d}\beta_{jk}}{\prod_{i=1}^n f(\mathbf{Y}_i | \mathbf{B}_{[:,k]} = \mathbf{0}, \mathbf{B}_{[:,-k]}, \psi, \mathbf{Z}_i)} \\
&= (p \vee n)^{-A|\mathcal{S}|} \frac{|\mathcal{K} \setminus \{k\}| + 1}{q - |\mathcal{K} \setminus \{k\}|} \sqrt{\frac{|\hat{\mathbf{T}}_{\mathcal{S},k}|}{|\mathbf{T}_{\mathcal{S},k}|}} \exp\left(\frac{1}{2} \hat{\boldsymbol{\beta}}_{\mathcal{S},k}^\top \hat{\mathbf{T}}_{\mathcal{S},k}^{-1} \hat{\boldsymbol{\beta}}_{\mathcal{S},k}\right), \tag{2.8}
\end{aligned}
$$

where $\mathbf{B}_{[:,-k]} := \mathbf{B}_{[:,[q]\setminus\{k\}]}$, $\mathbf{T}_{\mathcal{S},k} := \text{DIAG}((\tau_{jk})_{j\in\mathcal{S}})$, $\hat{\mathbf{T}}_{\mathcal{S},k} := \text{DIAG}((\hat{\tau}_{jk})_{j\in\mathcal{S}})$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S},k} := (\hat{\beta}_{jk})_{j\in\mathcal{S}}$.

Sample $\mathbf{Z}_i$ for $i \in [n]$: We sample $\mathbf{Z}_i := (Z_{ik})_{k\in[q]}$ from

$$(Z_{ik})_{k\in\mathcal{K}}|- \sim \mathtt{N}\left(\hat{\boldsymbol{\alpha}}_{\mathcal{K}}^\top \mathbf{Y}_i, \hat{\boldsymbol{\Xi}}_{\mathcal{K}}\right)$$
$$(Z_{ik})_{k\in[q]\setminus\mathcal{K}}|- \sim \mathtt{N}\left(\mathbf{0}, \mathbf{I}\right),$$

where

$$\hat{\boldsymbol{\alpha}}_{\mathcal{K}} := \psi^{-1} \hat{\boldsymbol{\Xi}}_{\mathcal{K}} \mathbf{B}_{[:,\mathcal{K}]}^\top \tag{2.9}$$

$$\hat{\boldsymbol{\Xi}}_{\mathcal{K}} := \left(\psi^{-1} \mathbf{B}_{[:,\mathcal{K}]}^\top \mathbf{B}_{[:,\mathcal{K}]} + \mathbf{I}\right)^{-1}. \tag{2.10}$$

Sample ψ: We sample ψ from

$$\psi|- \sim \mathtt{IG}\left(a_1 + \frac{np}{2}, a_2 + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p \left(Y_{ij} - \sum_{k\in\mathcal{K}} Z_{ik}\beta_{jk}\right)^2\right).$$

We provide several remarks on possible extensions of our Bayesian factor models.

Remark 1 (Extension to heterogeneous noise variances)**.** The AdaSS prior can be easily modified for a factor model with heterogeneous noise variances, under which the covariance matrix of the observed variable $\mathbf{Y}_i$ is decomposed as $\text{VAR}(\mathbf{Y}_i) = \mathbf{BB}^\top + \mathbf{\Psi}$ with $\mathbf{\Psi} := \text{DIAG}(\psi_1, \ldots, \psi_p)$. In this situation, a standard choice of the prior distribution on $\mathbf{\Psi}$ is a product of the inverse gamma distributions, that is,

$$\psi_j \overset{\text{ind}}{\sim} \texttt{IG}(\mathbf{a}_j)$$

for some $\mathbf{a}_j := (a_{j1}, a_{j2}) \in \mathbb{R}_+^2$ for $j \in [p]$. Then the conditional posterior of ψ_j is given by

$$\psi_j|- \sim \texttt{IG}\left(a_{j1} + \frac{n}{2}, a_{j2} + \frac{1}{2}\sum_{i=1}^n \left(Y_{ij} - \sum_{k \in \mathcal{K}} Z_{ik}\beta_{jk}\right)^2\right).$$

For posterior sampling of other parameters, calculations given in (2.4)-(2.10) are modified as $\widehat{\beta}_{jk} := \widehat{\tau}_{jk}\left\{\psi_j^{-1}\sum_{i=1}^n Z_{ik}\left(Y_{ij} - \sum_{h \in []:h \neq k} Z_{ih}\beta_{jh}\right)\right\}$, $\widehat{\tau}_{jk} := \left(\psi_j^{-1}\sum_{i=1}^n Z_{ik}^2 + \tau_{jk}^{-1}\right)^{-1}$, $\widehat{\boldsymbol{\alpha}}_{\mathcal{K}} := \widehat{\boldsymbol{\Xi}}_{\mathcal{K}}\mathbf{B}_{[:,\mathcal{K}]}^\top\mathbf{\Psi}^{-1}$ and $\widehat{\boldsymbol{\Xi}}_{\mathcal{K}} := \left(\mathbf{B}_{[:,\mathcal{K}]}^\top\mathbf{\Psi}^{-1}\mathbf{B}_{[:,\mathcal{K}]} + \mathbf{I}\right)^{-1}$, respectively. □

Remark 2 (Extension to correlated factors)**.** The factor model we investigated assumes that all components of the latent factor are independent. A more general model would be the correlated factor model such that $\mathbf{Y}_i|(\mathbf{Z}_i = \mathbf{z}_i) \overset{\text{ind}}{\sim} \texttt{N}(\mathbf{Bz}_i, \psi\mathbf{I})$ and $\mathbf{Z}_i \overset{\text{iid}}{\sim} \texttt{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{z})$ for some $\mathbf{\Sigma}_\mathbf{z} \in \mathbb{S}_{++}^q$, and our AdaSS prior can be easily modified for this model. If we impose the inverse Wishart prior $\texttt{IW}(\mathbf{A}, \nu)$ with scale matrix $\mathbf{A} \in \mathbb{S}_{++}^q$ and the degrees of freedom $\nu > 0$ on $\mathbf{\Sigma}_\mathbf{z}$, then we can sample $\mathbf{\Sigma}_\mathbf{z}$ from the conditional posterior

$$\mathbf{\Sigma}_\mathbf{z}|- \sim \texttt{IW}\left(\mathbf{A} + \sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top, \nu + n\right),$$

while the other posterior sampling schemes remain the same. Unfortunately, the theoretical results in this paper could not be applied directly to the correlated factor model since the correlated factor does not guarantee the required sparsity pattern of the loading matrix to achieve optimal posterior concentration rates. □

Remark 3 (Post-processing for estimating loading matrices)**.** The loading matrix $\mathbf{B}$ is not identifiable since for any $q \times q$ orthogonal matrix $\mathbf{Q}$ with $\mathbf{QQ}^\top = \mathbf{I}$ the transformed loading matrix $\mathbf{BQ}$ yields the exact same likelihood as that of $\mathbf{B}$. Consequently, we need additional effort if we are interested in estimating the loading matrix. A number of methods have been proposed to resolve this identifiability issue for Bayesian factor analysis. One approach is to impose a prior distribution on the loading matrix satisfying certain identifiability constraints such as the positive diagonal, lower triangular (PLT) constraint (Lopes and West, 2004; Ghosh and Dunson, 2009; Leung and Drton, 2016; Man and Culpepper, 2022). But as pointed out by Carvalho et al. (2008); Aßmann et al. (2016), the posterior distribution obtained under the PLT constraint may not be invariant to the ordering of the observed variables. An alternative approach is to post-process MCMC samples to make $\mathbf{B}$ identifiable (Aßmann et al., 2016; Papastamoulis and

Ntzoufras, 2022), which can be directly applicable to our Bayesian model. We illustrate the effectiveness of this approach by analyzing a toy example in Appendix C.2 in the supplementary material (Ohn et al., 2023)

$\square$

3 Asymptotic properties of the posterior distribution

We study frequentist properties of the posterior distribution induced by the proposed AdaSS prior distribution. Throughout this section, we assume that the number of columns q of the loading matrix $\mathbf{B}$ of our Bayesian model is taken to be sufficiently large so that it is at least the true factor dimensionality. Given data $\mathbf{Y}_{1:n} := (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, we denote by $\Pi(\cdot|\mathbf{Y}_{1:n})$ the posterior distribution under the AdaSS prior Π. Proofs of all results in this section are deferred to Appendix A in the supplementary material (Ohn et al., 2023).

3.1 Class of covariance matrices

We first define a class of matrices to which the true covariance matrix belongs. We denote by r the *true factor dimensionality*. To deal with very high dimensional cases where the dimension p is much larger than the sample size n, we impose sparsity on the loading matrix. Specifically, for a loading matrix $\mathbf{B}^\star \in \mathbb{R}^{p \times r}$, we define its (row) support by

$$\mathrm{SUPP}(\mathbf{B}^\star) := \left\{ j \in [p] : \mathbf{B}^\star_{[j,:]} \neq \mathbf{0} \right\}.$$

We say that the loading matrix $\mathbf{B}^\star$ is *s-sparse* if $|\mathrm{SUPP}(\mathbf{B}^\star)| \leq s$ and let

$$\mathcal{B}(p,r,s) := \left\{ \mathbf{B}^\star \in \mathbb{R}^{p \times r} : |\mathrm{SUPP}(\mathbf{B}^\star)| \leq s \right\}$$

be a set of $p \times r$-dimensional *s*-sparse (loading) matrices. The parameter space for the covariance matrix we consider throughout the paper is given by

$$\mathcal{C}(p,r,s,\lambda,\zeta) := \left\{ \mathbf{\Sigma}^\star = \mathbf{B}^\star(\mathbf{B}^\star)^\top + \psi^\star \mathbf{I} : \mathbf{B}^\star \in \mathcal{B}(p,r,s), \lambda_1(\mathbf{\Sigma}^\star) \leq \lambda, \psi \geq \psi_0, \right.$$
$$\left. \min \left\{ \lambda_r(\mathbf{B}^\star(\mathbf{B}^\star)^\top), \min_{j \in \mathrm{SUPP}(\mathbf{B}^\star)} \left\| \mathbf{B}^\star \left(\mathbf{B}^\star_{[j,:]} \right)^\top \right\|_2 \right\} \geq \zeta \right\}$$
$$(3.1)$$

for some arbitrarily small constant $\psi_0 > 0$. We discuss the implications of the conditions determining the class $\mathcal{C}(p,r,s,\lambda,\zeta)$.

- As we will show in Section 3.2 (Theorem 3), the posterior concentration rate of the covariance matrix depends on the dimension p, sparsity s, factor dimensionality r and the upper bound of the largest eigenvalue λ, but not on ζ, which is needed for the consistency of the factor dimensionality, and ψ_0.

- Our parameter space (3.1) includes loading matrices whose row support is sparse, which is also considered in Cai et al. (2013, 2015); Xie et al. (2022); Ning (2021).

On the other hand, Pati et al. (2014); Gao and Zhou (2015); Rockova and George (2016); Ohn and Kim (2022) consider the sparsity of the column support, which means that the nonzero entries of each column vector are less than or equal to s. Note that the s column support sparsity implies sr row support sparsity and hence the column and row support sparsities have the same order of s as $s \to \infty$ when r is bounded.

- The lower bound $\psi \geq \psi_0$, which is assumed in Pati et al. (2014); Ohn and Kim (2022), is introduced to avoid ill-conditioned covariance matrices. In contrast, Cai et al. (2013, 2015); Gao and Zhou (2015); Rockova and George (2016); Xie et al. (2022); Ning (2021) assume that ψ is fixed in their theories.

- The condition $\lambda_r(\mathbf{B}^\star(\mathbf{B}^\star)^\top) \geq \zeta$ yields the eigengap between spikes and noises, which prevents underestimation of the factor dimensionality. The same condition is assumed by Cai et al. (2013, 2015); Gao and Zhou (2015); Ohn and Kim (2022).

- Note that $\left\|\mathbf{B}^\star(\mathbf{B}^\star_{[j,:]})^\top\right\|_2 \geq \left\|\mathbf{B}^\star_{[j,:]}\right\|_2^2$, thus the condition $\min_{j \in \mathrm{SUPP}(\mathbf{B})}\left\|\mathbf{B}^\star(\mathbf{B}^\star_{[j,:]})^\top\right\|_2 \geq \zeta$ is met when the magnitudes of the nonzero rows do not vanish too quickly. This condition enables accurate estimation of the sparsity level of the true loading matrix. This condition is similar to the beta-min condition (see Section 7.4 of Bühlmann and Van De Geer (2011)) in high-dimensional sparse regression models. We believe that this condition is indispensable unless the column vectors are assumed to be orthogonal (Cai et al., 2013, 2015; Gao and Zhou, 2015) or the true sparsity level s is known (Ohn and Kim, 2022).

Note that we allow the model architecture parameters p, s, r λ and ζ to depend on n, but we do not specify the subscript n to those quantities, e.g., keep using p instead of p_n, for notational simplicity.

3.2 Posterior consistency of the factor dimensionality

In this section, we explore asymptotic properties of the posterior distribution of the factor dimensionality in the sparse factor model. For a loading matrix $\mathbf{B} \in \mathbb{R}^{p \times q}$, we define the factor dimensionality corresponding to $\mathbf{B}$ as

$$\xi := \xi(\mathbf{B}) := \left|\left\{k \in [q] : \mathbf{B}_{[:,k]} \neq \mathbf{0}\right\}\right|, \tag{3.2}$$

that is, the factor dimensionality $\xi(\mathbf{B})$ is equal to the number of nonzero columns of $\mathbf{B}$. The following theorem shows that the posterior distribution of the factor dimensionality behaves nicely.

Theorem 1. *Assume that $r \leq p/2$, $\lambda \lesssim s$ and $\epsilon := \lambda\sqrt{sr\log(p \vee n)/n} = o(1)$. Moreover, assume that $\zeta \geq c_0\epsilon$ for sufficiently large $c_0 > 0$. Then for any $\delta > 0$, there exists a constant $A_\delta > 0$ depending only on δ such that*

$$\inf_{\boldsymbol{\Sigma}^\star \in \mathcal{C}(p,r,s,\lambda,\zeta)} \mathsf{P}_{\boldsymbol{\Sigma}^\star}^{(n)}\left[\Pi\left(r \leq \xi(\mathbf{B}) \leq (1+\delta)r|\mathbf{Y}_{1:n}\right)\right] \to 1 \tag{3.3}$$

for the prior distribution defined through (2.1), (2.2) *and* (2.3) *with* $A > A_\delta$, *and* $\mathbf{a} \in \mathbb{R}_+^2$. *Furthermore, if* $r \lesssim \log(p \vee n)/\log n$, *then there exists an universal constant* $A_0 > 0$ *such that*

$$\inf_{\mathbf{\Sigma}^\star \in \mathcal{C}(p,r,s,\lambda,\zeta)} \mathsf{P}_{\mathbf{\Sigma}^\star}^{(n)} \left[\Pi \left(\xi(\mathbf{B}) = r | \mathbf{Y}_{1:n} \right) \right] \to 1, \tag{3.4}$$

whenever $A > A_0$.

Theorem 1 implies that a posteriori the factor dimensionality ξ is not much larger than the true one r, and ξ concentrates on r asymptotically if r is not too large. .0 Gao and Zhou (2015) attains posterior consistency of the factor dimensionality of the orthogonal loading matrix under a mildly growing regime $r \lesssim \log p$. Our condition $r \lesssim \log(p \vee n)/\log n$ for the posterior consistency is slightly stronger than Gao and Zhou (2015). However, note that we do not impose the orthogonality constraint on the loading matrix which makes posterior computation difficult.

Ohn and Kim (2022) obtains posterior consistency without any condition on the growth rate of the true factor dimensionality r by using a prior that strongly regularizes ξ. However, this strong regularization sacrifices the convergence rate of the covariance matrix by a factor of $\sqrt{s}$ compared to the optimal rate. Another critical drawback of their prior is that the knowledge of the true sparsity level is required to select the hyper-parameters in the prior. In contrast, the AdaSS prior attains the posterior consistency of the factor dimensionality *without knowing the true sparsity level*.

On the frequentist side, Cai et al. (2013, 2015) proposed consistent estimators of the factor dimensionality for sparse factor models. However, Cai et al. (2013) requires a $\sqrt{s}$ times larger detection rate for the eigengap (i.e., the lower bound of ζ) than ours and Cai et al. (2015) is nonadaptive to the true sparsity. Moreover, a known and fixed noise variance ψ is required for the consistency of both estimators, while our consistency result is *adaptive to the unknown noise level*.

Remark 4. One should set $q > r$ to correctly estimate r, but r is unknown. A naive strategy would be to set q very large, e.g., $q = p-1$, so that $q \geq r$. However, unnecessarily large q requires huge computation. A better strategy for choosing q is to set $q = \sqrt{n}$. This choice is based on our posterior contraction rate $\lambda\sqrt{sr\log(p \vee n)/n} = o(1)$ of the covariance matrix given in Theorem 3 in the next section. Since $r = \text{RANK}(\mathbf{B}^\star(\mathbf{B}^\star)^\top) \leq s$ for the true loading matrix $\mathbf{B}^\star$, if $\lambda \log(p \vee n) \gtrsim 1$ we have $r = o(\sqrt{n})$. Therefore, asymptotically, the upper bound $q = \sqrt{n}$ does not underdetermine the true factor dimensionality. A similar problem occurs in Gao and Zhou (2015), where the authors assume $r \lesssim \log p$ and set $q \asymp p^b$ for some $b > 0$ so that $q \gg r$. When p is large, our choice $q = \sqrt{n}$ is much smaller than that of Gao and Zhou (2015), which leads to more efficient computation □

Remark 5. In Theorem 1, we assume an upper bound for the largest eigenvalue such that $\lambda \lesssim s$. This bound is mild in view of the random matrix theory. Suppose that $\tilde{\mathbf{B}} \in \mathbb{R}^{s \times r}$ is a random matrix whose entries are independent centered random variables with finite fourth moments. Then by Theorem 2 of Latala (2005), since $r \leq s$, we have $\mathsf{E}\|\tilde{\mathbf{B}}\| \lesssim \sqrt{s} + \sqrt{r} \lesssim \sqrt{s}$. Therefore, $\mathsf{E}[\lambda_1(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top)] = \mathsf{E}\|\tilde{\mathbf{B}}\tilde{\mathbf{B}}^\top\| \lesssim s$. Pati et al.

(2014) and Rockova and George (2016) assumed the same condition as ours, while other studies on the Bayesian covariance estimation (Gao and Zhou, 2015; Xie et al., 2022) used a stronger condition that the largest eigenvalue of the true covariance matrix is bounded. □

In Theorem 1, we show that the true factor dimensionality is almost consistently recovered whenever the eigengap ζ is larger than the *detection rate* $\lambda\sqrt{sr\log p/n}$ by a sufficiently large constant $c_0 > 0$. As shown in the next proposition, this detection rate is optimal when $r \lesssim 1$ in the sense that any method cannot consistently estimate the factor dimensionality when the eigengap ζ is less than $a_0\lambda\sqrt{s\log(p/s)/n}$ for some constant $a_0 > 0$. This result is an extension of Theorem 5 of Cai et al. (2015) for unknown ψ and diverging λ.

Proposition 2. *Assume that $s\log p/n = o(1)$. Then there exists a constant $a_0 > 0$ such that if $0 \leq \zeta \leq a_0\lambda\sqrt{s\log(p/s)/n}$,*

$$\inf_{\hat{r}:\mathbb{R}^{p\times n}\mapsto\mathbb{N}} \sup_{\boldsymbol{\Sigma}\in\mathcal{C}(p,r,s,\lambda,\zeta)} \mathsf{P}_{\boldsymbol{\Sigma}}^{(n)}\left(\hat{r}(\mathbf{Y}_{1:n}) \neq r\right) \geq \frac{1}{4} \tag{3.5}$$

for all but finitely many n, where the infimum runs over all possible estimator $\hat{r}$ of r.

3.3 Posterior concentration rate of the covariance matrix

In the linear factor model, the covariance matrix $\mathbf{B}\mathbf{B}^\top + \psi\mathbf{I}$ determines its distribution. In this section, we prove that the posterior distribution of the covariance matrix in our Bayesian model concentrates around the true covariance matrix at a near-optimal rate, which is summarized in the next theorem.

Theorem 3. *Assume that $\lambda \lesssim s$. Then there exists a constant $M > 0$ such that*

$$\sup_{\boldsymbol{\Sigma}^\star\in\mathcal{C}(p,r,s,\lambda,0)} \mathsf{P}_{\boldsymbol{\Sigma}^\star}^{(n)}\left[\Pi\left(\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\| > M\lambda\sqrt{\frac{sr\log(p\vee n)}{n}}\Big|\mathbf{Y}_{1:n}\right)\right] \to 0 \tag{3.6}$$

for the prior distribution defined through (2.1)*,* (2.2) *and* (2.3) *with $A > 0$ and $\mathbf{a} \in \mathbb{R}_+^2$.*

Note that the lower bound of the eigengap ζ is set to 0 in Theorem 3 while it should be larger than a certain rate in Theorem 1, that is, the eigengap condition is required only for consistent estimation of the factor dimensionality but not for the contraction of the covariance matrix. This difference implies that the optimal estimation of the covariance matrix does not require the consistent estimation of the factor dimensionality.

Our posterior concentration rate in (3.6) is near optimal when $r \lesssim 1$ as shown in the following proposition, which is a direct consequence of Theorem 5.4 of Pati et al. (2014).

Proposition 4. *Assume that $s \log p/n = o(1)$ and $r \lesssim 1$. Then*

$$\inf_{\hat{\mathbf{\Sigma}}:\mathbb{R}^{p \times n} \mapsto \mathbb{S}_{++}^p} \sup_{\mathbf{\Sigma}^\star \in \mathcal{C}(p,r,s,\lambda,0)} \mathsf{P}_{\mathbf{\Sigma}^\star}^{(n)} \left[\left\| \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star \right\| \right] \gtrsim \lambda \sqrt{\frac{s \log(p/s)}{n}}, \qquad (3.7)$$

for all but finitely many n, where the infimum runs over all possible estimator $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}^\star$.

4 Numerical examples

We evaluate the empirical performance of the proposed Bayesian model with the AdaSS prior through simulation studies and real data analysis. For each posterior computation, we run the MCMC sampler described in Section 2.2 for 3,000 iterations discarding the first 500 as burn-in, and by thinning every 5, we obtain the final 500 MCMC samples from the posterior. We give the convergence diagnostics via trace, autocorrelation and partial autocorrelation plots of some randomly selected parameters in Appendix C.1 in the supplementary material (Ohn et al., 2023), which confirm that the MCMC sampler converges well.

4.1 Simulation study

In this section, we conduct an extensive numerical study to compare the performance of the AdaSS prior for estimating the factor dimensionality and the covariance matrix with various competitors. Throughout the simulation study, we set the number of columns of the loading matrix $q = \lceil \sqrt{n} \rceil$ for a sample size n and the hyperparameters $A = 0.1$ and $\mathbf{a} = (0.01, 0.01)$.

Posterior distribution of the factor dimensionality

We first compare the AdaSS prior and the spike and slab with the two-parameter IBP prior of Ohn and Kim (2022) for evaluating the concentration behaviors of their posterior distributions of the factor dimensionality. We only consider the prior of Ohn and Kim (2022) since other Bayesian models either do not infer the factor dimensionality (Ning, 2021; Xie et al., 2022) or do not achieve posterior consistency of the factor dimensionality (Rockova and George, 2016; Bhattacharya and Dunson, 2011; Srivastava et al., 2017) or are purely theoretical (i.e., do not have a posterior computation algorithm) (Gao and Zhou, 2015).

There are two hyperparameters of the spike and slab with the two-parameter IBP prior of Ohn and Kim (2022). The first hyperparameter denoted by α controls the factor dimensionality and the second hyperparameter denoted by κ does the sparsity of the loading matrix. For κ, we choose $p^{1.1}$ as recommended by Ohn and Kim (2022). For α, we consider three values: p^{-30}, p^{-25} and p^{-20}. Ohn and Kim (2022) proved that using $\alpha = p^{-As}$ for a constant $A > 0$ and the true sparsity s can lead to the posterior consistency. Assuming that $s = 30$, these choices of α correspond to the choices of $A \in \{4/6, 5/6, 1\}$. We use the MCMC sampler used in Knowles and Ghahramani (2011); Ohn and Kim (2022) for approximating the posterior.

We generate a data set consisting of n synthetic random vectors from the multivariate normal distribution with mean $\mathbf{0}$ and variance $\mathbf{B}^{\star}(\mathbf{B}^{\star})^{\top} + 2\mathbf{I}$ independently, where $\mathbf{B}^{\star}$ is a s-sparse $p \times r$ loading matrix. For the true loading matrix, we randomly select the location of s nonzero rows and sample the elements in the nonzero rows from $\{-2, 2\}$ randomly. We set $(n, p) = (100, 1000)$ and let the sparsity s and factor dimensionality r vary among $s \in \{10, 30, 50\}$ and $r \in \{1, 3, 5\}$, respectively.

Figure 1 presents the posterior distribution of the factor dimensionality for the AdaSS prior and spike and slab with the two-parameter IBP prior with different α. The posterior distribution under the AdaSS prior concentrates at the true factor dimensionality quite well for all nine cases, while the performance of the two-parameter IBP prior depends heavily on the choice of the hyperparameter α. If α is not sufficiently small, the resulting posterior distribution apparently overestimates the true factor dimensionality. The smallest choice of $\alpha = p^{-30}$ estimates the factor dimensionality consistently for some cases, but also severe underestimation occurs in other cases. The results of this simulation show that there is no universally good choice of the hyperparameter α
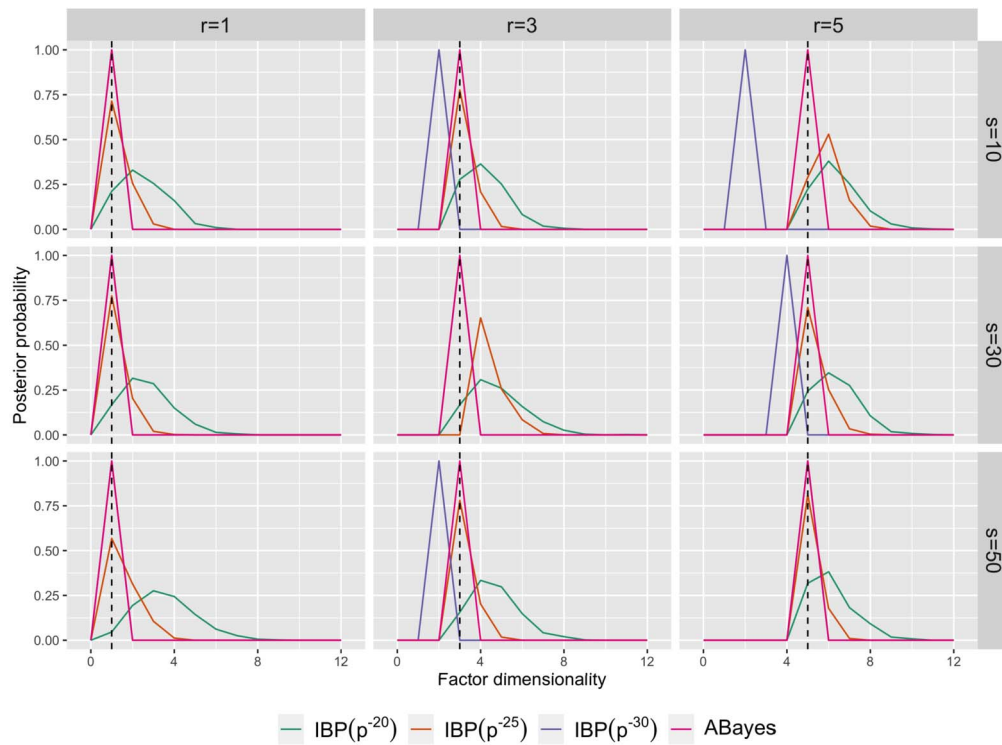

Figure 1: Posterior distributions of the factor dimensionality for the proposed adaptive Bayesian (AdaSS) prior and spike and slab with the two-parameter IBP prior with various α (IBP(α)). The black dashed vertical lines indicate the true factor dimensionality r.

in the two-parameter IBP across different levels of the sparsity, while the AdaSS prior performs consistently well with a single choice of the hyperparameter.

Comparison with frequentist estimators

In this simulation, we compare the performance of the AdaSS prior with some frequentist estimators for point estimation of the factor dimensionality. For our Bayesian model, we use the mode of the posterior distribution of the factor dimensionality as a point estimator. We consider the following five frequentist estimators as competitors: with $\mathbf{S}$ and $\mathbf{R}$ being the sample covariance and correlation matrices, respectively, and $r_{\max} \in [p]$ pre-specified,

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

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

 where $w := 2^{2/3}(2^{2/3} - 1)$.

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

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

- Growth ratio estimator (GR, Ahn and Horenstein (2013)):

$$\hat{r}_{\mathrm{GR}} := \operatorname*{argmax}_{j \in [r_{\max}]} \frac{\log\left(L_{j-1}/L_j\right)}{\log\left(L_j/L_{j+1}\right)},$$

 where $L_j := \sum_{t=j+1}^{p} \lambda_t(\mathbf{S})$.

- Adjusted correlation threshold estimator (ACT, Fan et al. (2020)):

$$\hat{r}_{\mathrm{ACT}} := \max\left\{j \in [r_{\max}] : \lambda_j^{\dagger} > 1 + \sqrt{p/(n-1)}\right\},$$

 where

$$\lambda_j^{\dagger} := \left[\frac{1-w_j}{\hat{\lambda}_j} - \frac{w_j}{p-j}\left\{\sum_{t=j+1}^{p} (\hat{\lambda}_t - \hat{\lambda}_j)^{-1} + 4(\hat{\lambda}_{j+1} - \hat{\lambda}_j)^{-1}\right\}\right]^{-1}$$

 with $\hat{\lambda}_j := \lambda_j(\mathbf{R})$ and $w_j := (p-j)/(n-1)$ for $j \in [p]$.

- Diagonal thresholding estimator (DT, Cai et al. (2013)):

$$\hat{r}_{\mathrm{DT}} := \max\left\{j \in [r_{\max}] : \lambda_j(\check{\mathbf{S}}_{[J,J]}) > 2\left(1 + \sqrt{|J|/n} + \sqrt{(2(1+|J|)\log(\mathrm{e}p) + 6\log n)/n}\right)^2\right\},$$

where $\check{\mathbf{S}} := n^{-1} \sum_{i=1}^{n} (\mathbf{Y}_i + \mathbf{Z}_i^0)(\mathbf{Y}_i + \mathbf{Z}_i^0)^\top$ with $\mathbf{Z}_1^0, \ldots, \mathbf{Z}_n^0 \overset{\text{iid}}{\sim} \mathtt{N}(\mathbf{0}, \mathbf{I})$ and

$$J := \left\{ j \in [p] : \check{\mathbf{S}}_{[j,j]} \geq 2 + 6\sqrt{\log p / n} \right\}.$$

We fix $r_{\max} = 10$ throughout the simulation study.

A synthetic data set of size n is generated from the p-dimensional normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{B}^\star (\mathbf{B}^\star)^\top + \mathbf{I}$, where the true loading matrix $\mathbf{B}^\star \in \mathbb{R}^{p \times r}$ is s-sparse. The true loading matrix $\mathbf{B}^\star$ is generated as follows: we first select s nonzero rows and sample the elements in the nonzero rows from the uniform distribution on $[-4/\sqrt{s}, -3/\sqrt{s}] \cup [3/\sqrt{s}, 4/\sqrt{s}]$. We take two sample sizes $n = 50$ and $n = 100$. We fix the dimension $p = 1,000$ and let the sparsity s and factor dimensionality r vary among $s \in \{10, 30, 50\}$ and $r \in \{1, 3, 5\}$, respectively.

The simulation results based on 100 synthetic data sets of size $n = 50$ and $n = 100$ are summarized in Tables 1 and 2, respectively. We see that the proposed Bayesian model with the AdaSS prior outperforms the other competitors in the estimation of the factor dimensionality. The AdaSS prior has the highest proportion of the correct estimation for 17 out of the total 18 simulation setups. In particular, there are considerable performance gaps between the AdaSS prior and the other competitors when sparsity is small ($s = 10$) or the factor dimensionality is large ($r = 5$).

Covariance matrix estimation

In this simulation study, we compare the AdaSS prior with other competitors for covariance matrix estimation. For competitors, we consider the principal orthogonal complement thresholding method (POET, Fan et al. (2013)), the variational inference method for Bayesian sparse PCA (SPCA-VI, Ning (2021)), the Bayesian sparse factor models with multiplicative gamma process shrinkage prior (MGPS, Bhattacharya and Dunson (2011)) and two maximum a posteriori estimators that employ the multi-scale generalized double Pareto prior (MDP, (Srivastava et al., 2017)) and the spike-and-slab lasso with Indian buffet process prior (SSL-IBP, (Rockova and George, 2016)), respectively. For the POET and SPCA-VI, the factor dimensionality must be selected in advance and we use the true factor dimensionality for this. We use the posterior mean of the covariance matrix as the point estimator for the MGPS and AdaSS priors.

We generate 100 synthetic data sets with sample size $n = 50$ and $n = 100$, respectively, and we report the averages of the scaled spectral norm losses $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star\| / \|\mathbf{\Sigma}^\star\|$ between the point estimate $\hat{\mathbf{\Sigma}}$ of each estimator and the true covariance matrix $\mathbf{\Sigma}^\star$ obtained over 100 synthetic data sets in Tables 3 and 4. The AdaSS prior performs generally well, while the POET, MGPS and MDP are significantly inferior. SSL-IBP is not much worse and performs best for the setups with $s = 50$.

4.2   Real data analysis

In this section, we analyze gene expression data on aging in mice from the AGEMAP (Atlas of Gene Expression in Mouse Aging Project) database Zahn et al. (2007). We

s	r		ET	ER	GR	ACT	DT	AdaSS
10	1	True	85	97	97	96	100	100
		Over	14	3	3	4	0	0
		Under	1	0	0	0	0	0
		Ave	1.14	1.03	1.03	1.04	1	1
	3	True	24	3	3	0	0	97
		Over	1	0	0	0	0	0
		Under	75	97	97	100	100	3
		Ave	2.16	1.35	1.37	1.01	1	2.97
	5	True	0	0	0	0	0	77
		Over	0	0	0	0	0	15
		Under	100	100	100	100	100	8
		Ave	2.97	1.54	1.55	1.01	1.1	5.06
30	1	True	86	96	97	97	100	100
		Over	12	4	3	3	0	0
		Under	2	0	0	0	0	0
		Ave	1.1	1.1	1.07	1.03	1	1
	3	True	46	15	15	0	0	94
		Over	0	0	0	0	0	0
		Under	54	85	85	100	100	6
		Ave	2.43	1.57	1.57	1.09	1	2.93
	5	True	0	1	1	0	0	87
		Over	0	0	0	0	0	0
		Under	100	99	99	100	100	13
		Ave	3.2	1.72	1.77	1.06	1	4.86
50	1	True	80	98	97	95	100	100
		Over	17	2	3	5	0	0
		Under	3	0	0	0	0	0
		Ave	1.17	1.04	1.05	1.05	1	1
	3	True	58	9	10	6	0	73
		Over	2	0	0	0	0	0
		Under	40	91	90	94	100	27
		Ave	2.62	1.52	1.56	1.4	1	2.67
	5	True	2	2	2	0	0	61
		Over	0	0	0	0	0	0
		Under	98	98	98	100	100	39
		Ave	3.39	1.94	1.99	1.25	1	4.54

Table 1: Proportions of correct estimation ("True"), overestimation ("Over") and underestimation ("Under") of the estimated factor dimensionalities for various sparsity s and true factor dimensionality r obtained on 100 synthetic data sets of size $n = 50$. "Ave" is the average of the estimated factor dimensionalities.

obtained this data from the online website http://statweb.stanford.edu/~owen/data/ AGEMAP. There are 5 female and 5 male mice in each age group, where there are 4

s	r		ET	ER	GR	ACT	DT	AdaSS
10	1	True	78	100	100	93	100	100
		Over	22	0	0	7	0	0
		Under	0	0	0	0	0	0
		Ave	1.24	1	1	1.07	1	1
	3	True	66	9	10	0	0	86
		Over	2	0	0	0	0	13
		Under	32	91	90	100	100	1
		Ave	2.7	1.55	1.57	1.02	1.25	3.12
	5	True	6	0	0	0	0	20
		Over	0	0	0	0	0	77
		Under	94	100	100	100	100	3
		Ave	3.63	1.83	1.86	1	2.13	5.93
30	1	True	83	100	100	93	100	100
		Over	17	0	0	7	0	0
		Under	0	0	0	0	0	0
		Ave	1.18	1	1	1.09	1	1
	3	True	92	33	34	5	0	99
		Over	4	0	0	2	0	1
		Under	4	67	66	93	100	0
		Ave	3	1.96	1.98	1.56	1	3.01
	5	True	46	3	4	0	0	68
		Over	1	0	0	0	0	32
		Under	53	97	96	100	100	0
		Ave	4.45	2	2.1	1.43	1	5.32
50	1	True	85	100	100	94	100	100
		Over	15	0	0	6	0	0
		Under	0	0	0	0	0	0
		Ave	1.16	1	1	1.06	1	1
	3	True	99	59	61	64	0	98
		Over	1	0	0	2	0	0
		Under	0	41	39	34	100	2
		Ave	3.01	2.36	2.38	2.67	1	2.98
	5	True	68	15	20	2	0	91
		Over	0	0	0	0	0	6
		Under	32	85	80	98	100	3
		Ave	4.68	2.36	2.65	2.61	1	5.03

Table 2: Proportions of correct estimation ("True"), overestimation ("Over") and underestimation ("Under") of the estimated factor dimensionalities for various sparsity s and true factor dimensionality r obtained on 100 synthetic data sets of size $n = 100$. "Ave" is the average of the estimated factor dimensionalities.

age groups of 1, 6, 16 and 24 months. Thus there are 40 mice in total. From each of 40 mice, 16 microarrays obtained from 16 different tissues were prepared, and from each

s	r	POET	SPCA-VI	MGPS	MDP	SSL-IBP	ABayes
	1	2.366 (0.151)	0.245 (0.108)	2.591 (2.04)	2.103 (0.135)	0.689 (0.062)	**0.233 (0.103)**
10	3	1.83 (0.254)	0.398 (0.12)	1.45 (0.863)	1.583 (0.212)	0.646 (0.094)	**0.301 (0.113)**
	5	1.59 (0.232)	0.422 (0.111)	1.195 (0.575)	1.271 (0.191)	0.696 (0.101)	**0.335 (0.107)**
	1	2.375 (0.155)	0.772 (0.102)	1.945 (1.292)	2.104 (0.14)	0.699 (0.067)	**0.624 (0.152)**
30	3	2.073 (0.202)	0.674 (0.117)	2.078 (1.309)	1.839 (0.184)	0.696 (0.063)	**0.609 (0.172)**
	5	1.868 (0.192)	0.644 (0.086)	1.551 (0.87)	1.649 (0.175)	0.684 (0.052)	**0.631 (0.138)**
	1	2.345 (0.146)	0.901 (0.039)	2.018 (1.52)	2.072 (0.13)	**0.759 (0.065)**	0.847 (0.134)
50	3	2.145 (0.194)	0.762 (0.102)	1.996 (1.168)	1.901 (0.175)	**0.695 (0.069)**	0.966 (0.208)
	5	2.013 (0.2)	0.744 (0.098)	1.519 (0.752)	1.786 (0.182)	**0.709 (0.067)**	1.049 (0.228)

Table 3: The averages and standard errors of the scaled spectral norm losses of the estimators of the covariance matrix obtained on 100 synthetic data sets with $n = 50$.

s	r	POET	SPCA-VI	MGPS	MDP	SSL-IBP	ABayes
	1	1.437 (0.108)	**0.164 (0.064)**	2.374 (1.123)	1.322 (0.101)	0.523 (0.048)	0.17 (0.066)
10	3	1.136 (0.146)	0.274 (0.074)	1.565 (1.129)	1.044 (0.141)	0.483 (0.066)	**0.218 (0.074)**
	5	0.999 (0.151)	0.281 (0.069)	1.051 (0.407)	0.881 (0.135)	0.464 (0.057)	**0.244 (0.075)**
	1	1.437 (0.09)	0.366 (0.105)	2.489 (1.124)	1.317 (0.085)	0.534 (0.053)	**0.317 (0.088)**
30	3	1.28 (0.138)	0.449 (0.083)	2.187 (0.946)	1.175 (0.133)	0.533 (0.055)	**0.353 (0.105)**
	5	1.178 (0.116)	0.457 (0.079)	1.723 (1.618)	1.084 (0.109)	0.538 (0.059)	**0.4 (0.087)**
	1	1.428 (0.095)	0.707 (0.112)	2.357 (1.226)	1.306 (0.09)	0.57 (0.065)	**0.536 (0.133)**
50	3	1.328 (0.102)	0.628 (0.102)	2.236 (1.072)	1.219 (0.1)	**0.547 (0.058)**	0.587 (0.133)
	5	1.261 (0.119)	0.57 (0.072)	1.729 (0.97)	1.161 (0.112)	**0.564 (0.059)**	0.652 (0.123)

Table 4: The averages and standard errors of the scaled spectral norm losses of the estimators of the covariance matrix obtained on 100 synthetic data sets with $n = 100$.

microarray, gene expression levels of $8,932$ probes were measured. In this paper, we focus only on the microarray data from the cerebrum tissue, for which the rotation test of Perry and Owen (2010) provided strong evidence for the presence of the latent factor. We will call this one tissue data set with sample size $n = 40$ and dimension $p = 8,932$ the AGEMAP data for simplicity.

We preprocessed the AGEMAP data following the regression model of Perry and Owen (2010). We obtained the mean-centered data by regressing out an intercept, sex, and age effects on each of the 8,932 outcomes. Then the factor dimensionality is estimated based on the mean-centered data set. We consider the factor model with heterogeneous noise variances and impose the AdaSS prior presented in Remark 1. We set $q = 10 \geq \sqrt{n} = \sqrt{40}$, $A = 0.1$ and $\mathbf{a}_j = (0.01, 0.01)$ for every $j \in [p]$ in the prior. Then we take the posterior mode of the factor dimensionality as the point estimate. For comparison, we also considered the five frequentist methods described in Section 4.1, i.e., ET, ER, GR, ACT and DT.

Table 5 provides the factor dimensionality estimates by the proposed Bayesian model and the five competing frequentist methods. The four methods including the AdaSS prior estimate the factor dimensionality by 1. The presence of the one-dimensional latent factor was advocated by the rotation test of Perry and Owen (2010).

Figure 2a shows the histogram of the posterior means of the latent factors

ET	ER	GR	ACT	DT	AdaSS
8	1	1	10	1	1

Table 5: Estimated factor dimensionality for the AGEMAP data.

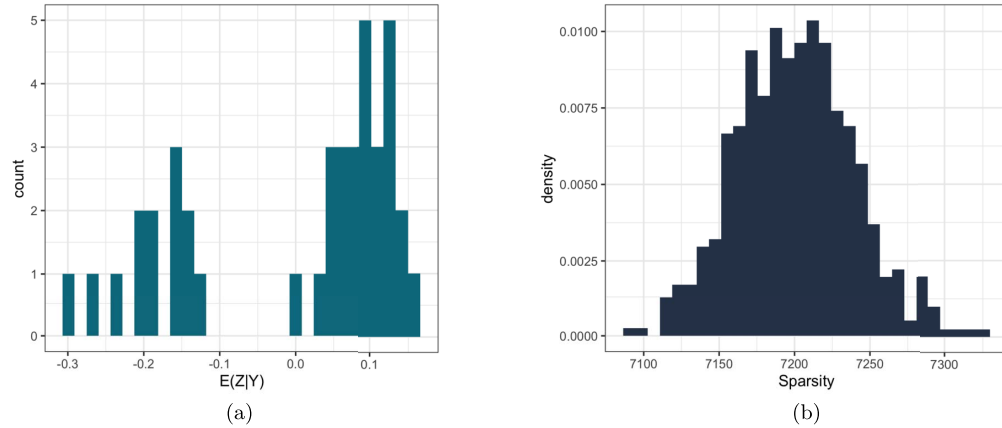(a)   (b)

Figure 2: (a) Histogram of the posterior means of the latent factors for each observation; (b) Posterior distribution of the sparsity of the loading matrix for the AGEMAP data.

$\mathsf{E}(Z_{ik^*}|\mathbf{Y}_{1:n})$ for $i \in [n]$ obtained under the AdaSS prior, where k^* denotes the index of the nonzero column of the loading matrix, i.e., $\mathbf{B}_{[:,k^*]} \neq \mathbf{0}$ under the posterior distribution. The bimodality of the histogram is clearly shown, which is also confirmed by Perry and Owen (2010). Figure 2b presents the posterior distribution of the sparsity $|\text{supp}(\mathbf{B})|$ of the loading matrix, which ranges from 79.4% to 82%. A similar 78% sparsity of the estimated factor model was reported by Rockova and George (2016).

5   Concluding remarks

In this paper, we proposed a novel prior distribution, called AdaSS, to infer high-dimensional sparse factor models. We proved that the resulting posterior distribution asymptotically concentrates at the true factor dimensionality without knowing the sparsity level of the true loading matrix. This adaptivity to the sparsity is a remarkable advantage of the proposed method over other theoretically consistent estimators such as the point estimator in Cai et al. (2015) and Bayesian posterior distribution in Ohn and Kim (2022). We also showed that the proposed model attained the optimal detection rate of the eigengap for consistent estimation of the factor dimensionality. Moreover, the concentration rate of the posterior distribution of the covariance matrix is optimal when the true factor dimensionality is bounded and equal or faster than those of other Bayesian models otherwise. Our numerical studies amply confirmed the theoretical results and provided strong empirical support to the proposed AdaSS prior.

With our prior, nonzero row vectors of the loading matrix $\mathbf{B}$ are not sparse. That

is, when s'-many nonzero row vectors and the factor dimensionality are given, all the entries of the corresponding $s' \times \xi$ sub-matrix are all nonzero. In practice, one may want to have sparsity in nonzero row vectors. Our prior can be modified easily to ensure such sparsity without hampering the asymptotic properties, which we will report somewhere else.

There are several promising directions for future work. In this paper, we consider the *static* factor model where the observations are assumed to be identically distributed. However, this static factor model may be inadequate to capture the dependence structure in some types of data, e.g., time series data. As an alternative, we may consider a dynamic factor model, where the covariance matrix as well as the factor dimensionality can be different at each time point. It would of interest to study the posterior consistency of the factor dimensionality which possibly varies over time. Another promising avenue of research is to develop the Bayesian factor model which deals with non-Gaussian or mixed-type observed variables. We believe that the proposed Bayesian model can be easily extended to those types of data using the Gaussian copula factor model developed by Murray et al. (2013). It would be interesting to investigate the theoretical properties of such a non-Gaussian extension of the proposed Bayesian model.

Supplementary Material

Supplement to "A Bayesian sparse factor model with adaptive posterior concentration" (DOI: 10.1214/23-BA1392SUPP; .pdf). The supplementary material contains the proofs of the main results and additional numerical results.

References

Seung C Ahn and Alex R Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013. MR3064065. doi: https://doi.org/10.3982/ECTA8968. 1278, 1291

Christian Aßmann, Jens Boysen-Hogrefe, and Markus Pape. Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. *Journal of Econometrics*, 192(1):190–206, 2016. MR3463672. doi: https://doi.org/10.1016/j.jeconom.2015.10.010. 1284

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002. MR1926259. doi: https://doi.org/10.1111/1468-0262.00273. 1278

Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006. MR2238213. doi: https://doi.org/10.1111/j.1468-0262.2006.00696.x. 1278

Jushan Bai and Serena Ng. Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, 25(1):52–60, 2007. MR2338870. doi: https://doi.org/10.1198/073500106000000413. 1278

Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 1278, 1289, 1292

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. MR2807761. doi: https://doi.org/10.1007/978-3-642-20192-9. 1286

Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013. MR3161458. doi: https://doi.org/10.1214/13-AOS1178. 1278, 1279, 1285, 1286, 1287, 1291

Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3-4):781–815, 2015. MR3334281. doi: https://doi.org/10.1007/s00440-014-0562-z. 1278, 1279, 1285, 1286, 1287, 1288, 1296

Gian Vittorio Caprara, Claudio Barbaranelli, Laura Borgogni, and Marco Perugini. The "big five questionnaire": A new questionnaire to assess the five factor model. *Personality and individual Differences*, 15(3):281–288, 1993. 1278

Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 1278, 1284

Bo Chen, Minhua Chen, John Paisley, Aimee Zaas, Christopher Woods, Geoffrey S Ginsburg, Alfred Hero, Joseph Lucas, David Dunson, and Lawrence Carin. Bayesian inference of the number of factors in gene-expression analysis: application to human virus challenge studies. *BMC bioinformatics*, 11(1):552, 2010. 1278

Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008. MR2472991. doi: https://doi.org/10.1016/j.jeconom.2008.09.017. 1278

Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320, 2011. MR3012410. doi: https://doi.org/10.1214/11-AOS944. 1278

Jianqing Fan, Xu Han, and Weijie Gu. Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035, 2012. MR3010887. doi: https://doi.org/10.1080/01621459.2012.720478. 1278

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4), 2013. MR3091653. doi: https://doi.org/10.1111/rssb.12016. 1292

Jianqing Fan, Lingzhou Xue, and Jiawei Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292–306, 2017. MR3717565. doi: https://doi.org/10.1016/j.jeconom.2017.08.009. 1278

Jianqing Fan, Yuan Ke, Qiang Sun, and Wen-Xin Zhou. Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association*, to appear, 2018a. MR4047307. doi: https://doi.org/10.1080/01621459.2018.1527700. 1278

Jianqing Fan, Han Liu, and Weichen Wang. Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383, 2018b. MR3819104. doi: https://doi.org/10.1214/17-AOS1588. 1278

Jianqing Fan, Jianhua Guo, and Shurong Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, to appear, 2020. MR4436317. doi: https://doi.org/10.1080/01621459.2020.1825448. 1278, 1291

Chao Gao and Harrison H Zhou. Rate-optimal posterior contraction for sparse PCA. *The Annals of Statistics*, 43(2):785–818, 2015. MR3325710. doi: https://doi.org/10.1214/14-AOS1268. 1279, 1286, 1287, 1288, 1289

Joyee Ghosh and David B Dunson. Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 1284

Alois Kneip and Pascal Sarda. Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics*, 39(5):2410–2447, 2011. MR2906873. doi: https://doi.org/10.1214/11-AOS905. 1278

David Knowles and Zoubin Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, pages 1534–1552, 2011. MR2849785. doi: https://doi.org/10.1214/10-AOAS435. 1278, 1289

Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012. MR2933663. doi: https://doi.org/10.1214/12-AOS970. 1278, 1291

Rafal Latala. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005. MR2111932. doi: https://doi.org/10.1090/S0002-9939-04-07800-1. 1287

Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. 1278

Dennis Leung and Mathias Drton. Order-invariant prior specification in Bayesian factor analysis. *Statistics & Probability Letters*, 111:60–66, 2016. MR3474783. doi: https://doi.org/10.1016/j.spl.2016.01.006. 1284

Hedibert Freitas Lopes and Mike West. Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67, 2004. MR2036762. 1284

Albert Xingyi Man and Steven Andrew Culpepper. A mode-jumping algorithm for Bayesian factor analysis. *Journal of the American Statistical Association*,

117(537):277–290, 2022. MR4399085. doi: https://doi.org/10.1080/01621459.2020.1773833. 1284

Jared S Murray, David B Dunson, Lawrence Carin, and Joseph E Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013. MR3174649. doi: https://doi.org/10.1080/01621459.2012.762328. 1297

Bo Ning. Spike and slab Bayesian sparse principal component analysis. *arXiv preprint arXiv:2102.00305*, 2021. 1278, 1281, 1285, 1286, 1289, 1292

Ilsang Ohn and Yongdai Kim. Posterior consistency of factor dimensionality in high-dimensional sparse factor models. *Bayesian Analysis*, 17(2):491–514, 2022. MR4483228. doi: https://doi.org/10.1214/21-ba1261. 1278, 1281, 1286, 1287, 1289, 1296

Ilsang Ohn, Lizhen Lin, and Yongdai Kim Supplementary Material for "A Bayesian Sparse Factor Model with Adaptive Posterior Concentration". *Bayesian Analysis*, 2023. doi: https://doi.org/10.1214/23-BA1392SUPP. 1285, 1289

Alexei Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010. 1278, 1291

Panagiotis Papastamoulis and Ioannis Ntzoufras. On the identifiability of Bayesian factor analytic models. *Statistics and Computing*, 32(2):23, 2022. MR4394853. doi: https://doi.org/10.1007/s11222-022-10084-4. 1284

Debdeep Pati, Anirban Bhattacharya, Natesh S Pillai, and David Dunson. Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130, 2014. MR3210997. doi: https://doi.org/10.1214/14-AOS1215. 1278, 1286, 1287, 1288

Patrick O Perry and Art B Owen. A rotation test to verify latent structure. *Journal of Machine Learning Research*, 11(2), 2010. MR2600622. 1295, 1296

Veronika Rockova and Edward I George. Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516):1608–1622, 2016. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 1278, 1281, 1286, 1288, 1289, 1292, 1296

A Pedro Duarte Silva. Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistics & Data Analysis*, 55(11):2975–2990, 2011. MR2813060. doi: https://doi.org/10.1016/j.csda.2011.05.002. 1278

Sanvesh Srivastava, Barbara E Engelhardt, and David B Dunson. Expandable factor analysis. *Biometrika*, 104(3):649–663, 2017. MR3694588. doi: https://doi.org/10.1093/biomet/asx030. 1278, 1289, 1292

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002. MR1951271. doi: https://doi.org/10.1198/016214502388618960. 1278

Fangzheng Xie, Joshua Cape, Carey E Priebe, and Yanxun Xu. Bayesian sparse spiked covariance model with a continuous matrix shrinkage prior. *Bayesian Analysis*, 17(4):1193–1217, 2022. MR4506026. doi: https://doi.org/10.1214/21-ba1292. 1278, 1281, 1285, 1286, 1288, 1289

Jacob M Zahn, Suresh Poosala, Art B Owen, Donald K Ingram, Ana Lustig, Arnell Carter, Ashani T Weeraratna, Dennis D Taub, Myriam Gorospe, Krystyna Mazan-Mamczarz, et al. Agemap: a gene expression database for aging in mice. *PLoS Genet*, 3(11):e201, 2007. 1292