Bayesian Analysis (2022) **17**, Number 3, pp. 901–929

On Posterior Consistency of Bayesian Factor Models in High Dimensions*

Yucong Ma† and Jun S. Liu‡

Abstract. As a principled dimension reduction technique, factor models have been widely adopted in applications. However, conducting a proper Bayesian factor analysis can be subtle in high-dimensional settings since it requires both a careful prescription of the prior distribution and a suitable computational strategy. We analyze issues of posterior inconsistency and sensitivity under different priors for high-dimensional sparse normal factor models, and show why adopting the $\sqrt{n}$-orthonormal factor assumption can resolve these issues and lead to a more robust and efficient Bayesian analysis. We also provide an efficient Gibbs sampler to conduct the required computation, and show that it can be orders of magnitude more efficient than compared existing algorithms.

Keywords: factor analysis, high dimensional data, posterior consistency, orthogonality, Gibbs sampling.

1 Introduction

Factor models, which assume that the information in high-dimensional observations can be captured by a few latent factors, have been widely adopted in social science, economics, bioinformatics, and many other fields that need interpretable dimension reduction for their data. In this article, we consider the following *normal factor* formulation: each G-dimensional vector observation $\mathbf{y}_i$ (e.g., daily returns of ∼3000 U.S. stocks) is related to a K-dimensional vector of latent factors $\boldsymbol{\omega}_i$ (e.g., 20 market factors) through a skinny tall factor loading matrix $\mathbf{B}$, plus idiosyncratic errors:

$$\mathbf{y}_i \mid \boldsymbol{\omega}_i, \mathbf{B}, \boldsymbol{\Sigma} \overset{i.i.d.}{\sim} \mathcal{N}_G(\mathbf{B}\boldsymbol{\omega}_i, \boldsymbol{\Sigma}),\ i = 1, \ldots, n, \tag{1.1}$$

and the idiosyncratic variance matrix $\boldsymbol{\Sigma}$ is assumed to be diagonal as in the literature. In matrix form, we denote the observations as $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$, which is a $G \times n$ matrix, and the factors as a $K \times n$ matrix $\boldsymbol{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$. The factors are usually assumed to follow the standard Gaussian independently: $\boldsymbol{\omega}_i \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$.

People are often interested in estimating the $G \times K$ loading matrix $\mathbf{B}$ in order to gain insight in the correlation structure of the observations. Marginalizing out $\boldsymbol{\omega}_i$, we have $[\mathbf{y}_i \mid \mathbf{B}, \boldsymbol{\Sigma}] \sim \mathcal{N}_G(\mathbf{0}, \mathbf{B}\mathbf{B}^T + \boldsymbol{\Sigma})$, implying that the loading matrix $\mathbf{B}$ is only identifiable up to a right orthogonal transformation (rotationally invariant). It is thus rather difficult

*This research is supported in part by the National Science Foundation of USA Grants DMS-1613035, DMS-1712714 and DMS-1903139.

†Department of Statistics, Harvard University. Science Center 715, One Oxford Street, Cambridge, MA 02138, yucongma@g.harvard.edu

‡Department of Statistics, Harvard University. Science Center 715, One Oxford Street, Cambridge, MA 02138, jliu@stat.harvard.edu

 https://doi.org/10.1214/21-BA1281

to pinpoint the factor loading matrix consistently, to determine the dimensionality of the latent factors, or to design efficient algorithms to conduct a proper full Bayesian analysis of the model.

In recent years, considerable progresses have been made in the realm of sparse Bayesian factor analysis under different prior settings (Bhattacharya and Dunson, 2011; Ročková and George, 2016; Fruehwirth-Schnatter and Lopes, 2018). Both Ročková and George (2016) and Fruehwirth-Schnatter and Lopes (2018) employed spike-and-slab (SpSL) priors, i.e., mixtures of either a concentrated distribution with a small variance (continuous) or a point mass (discrete), and a diffuse distribution, for elements in $\mathbf{B}$. While Fruehwirth-Schnatter and Lopes (2018) imposed discrete SpSL priors conditional on the feature allocation, Ročková and George (2016) employed independent continuous SpSL priors, under which a fast posterior mode-detecting strategy was proposed. The identifiability of sparse factor models was discussed in Fruehwirth-Schnatter and Lopes (2018), who also designed an efficient Markov chain Monte Carlo (MCMC) procedure to simulate from the posterior distribution of an over-parameterized sparse factor model under the discrete SpSL prior. Bhattacharya and Dunson (2011) introduced the multiplicative gamma process (MGP) shrinkage prior, which allows for infinitely many factors and proposed an adaptive Gibbs sampler for automatic factor number selection.

Our work focuses on how to make 'correct' Bayesian inference for a sparse Bayesian factor model in high dimensions. By 'correct' we mean that the posterior distribution of the covariance matrix ($\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}$) should concentrate at the underlying true matrix and should contract towards the truth asymptotically as both sample size and dimensionality increases under certain conditions. This is a subtle problem since a prior that induces posterior consistency in low dimensions ($n \gg G$) may fail to do so in a high dimensional ($G \gg n$) regime. The "incorrectness" is most prominently shown when the seemingly innocent SpSL prior is used as in Ročková and George (2016). When the average number s of nonzero elements in each column of the loading matrix $\mathbf{B}$ is no less than the sample size n, we observe from simulations a 'magnitude inflation' phenomenon. That is, posterior samples of the loading matrix (also the covariance matrix) are inflated in the matrix norm compared to the true data-generating loading matrix. The extent of inflation is affected by the variance of the slab part of the SpSL prior — the more diffuse the slab prior the more inflation we observe. Interestingly, however, this inflation is not reflected in the posterior mode as Ročková and George (2016) showed that the *maximum a posteriori* (MAP) estimate of the loading matrix is consistent (up to trivial rotations) under the same simulation setup. This $s \geq n$ setting is not unusual in practice, as shown by the real example in Section 7.

The inflation phenomenon is closely related to the nearly non-identifiable nature of model (1.1): $\mathbf{B} \times \boldsymbol{\omega}_i = (\mathbf{B}\mathbf{C}) \times (\mathbf{C}^{-1}\boldsymbol{\omega}_i)$ for any non-singular $\mathbf{C} \in \mathcal{R}^{K \times K}$. Requiring that the $\boldsymbol{\omega}_i$ are *i.i.d.* normal alleviates the identifiability issue, but is not enough to "tie down" $\mathbf{B}$ in the posterior distribution if too many independent diffuse priors are imposed on its elements. Problems with the use of diffuse priors in general Bayesian inference when observation sample sizes are small relative to the number of parameters being estimated have been studied in the literature (Efron, 1973; Kass and Wasserman, 1996; Natarajan and McCulloch, 1998). Such issues for Bayesian factor analysis were

also noted in Ghosh and Dunson (2009) and a practical solution was proposed without further theoretical investigations.

The Ghosh-Dunson model allows each factor to have an unknown variance that follows an inverse Gamma prior and imposes the standard normal distribution on the loading matrix's elements. If one reallocates the variances of the factors to the loading matrix side, this model is equivalent to reformatting the elements of the loading matrix $\mathbf{B}$ as $\beta_{jk} = q_{jk} \cdot r_k$ with r_k being a column-wise parameter following an inverse-Gamma distribution and q_{jk} being *i.i.d.* standard normal. When a diffuse prior is imposed on r_k, the marginal prior for β_{jk} is also diffuse. Consequently, this hierarchical prior construction resolves the magnitude inflation problem by imposing a prior dependency among the elements in each column of $\mathbf{B}$ and reducing the number of diffuse parameters. Bhattacharya and Dunson (2011) follows this idea and thus is free of the inflation phenomenon as well. But this prior is observed to have a mild "twisting" effect on the posterior of the loading matrix.

The magnitude inflation phenomenon is closely related to the weak identifiability nature of the normal factor model in high dimensions. As a consequence, informative priors can easily become too influential and thus dominate the posterior distribution. Though it is not obvious, the independent SpSL prior is a very informative prior, as the dimension s grows faster than n. Ghosh and Dunson (2009)'s solution for controlling the prior effect is effective empirically. But it is still an open problem to show theoretically that the Ghosh-Dunson model is free from the prior dominance.

To search for a more principled strategy for resolving the prior dominance issue, we here study asymptotic behaviors of the posterior distributions under an independent SpSL prior for elements of the loading matrix and a right-rotational invariant distribution for the factor matrix $\mathbf{\Omega}$ (i.e., $\mathbf{\Omega}$ and $\mathbf{\Omega R}$ follows the same distribution for all $n \times n$ orthogonal matrix $\mathbf{R}$). In doing so, we are able to connect the observed inflation phenomena of the posterior distribution with the factor assumption and show that employing a stronger control over $\mathbf{\Omega\Omega}^T/n$ through the factor assumption can result in a consistent posterior distribution for the loading matrix under high dimensions ($s \gg n \to \infty$) provided that the sparsity pattern is known.

Insights revealed by our theoretical analyses are: (i) the normal factor model assumption controls $\mathbf{\Omega\Omega}^\top/n$ too weakly, thus inducing a weakly identifiable high dimensional model; (ii) the posterior distribution of $\mathbf{B}$ becomes too sensitive to the prior because of the weak identifiability and high dimensionality. Consequently, we propose the following *$\sqrt{n}$-orthonormal factor model* as a remedy: under the same relationship as in (1.1), we assume that $\mathbf{\Omega}/\sqrt{n}$ is uniform on the Stiefel manifold $St(K, n)$, which is the set of all orthonormal K-frames in $\mathbb{R}^n$, or, equivalently, the first K rows of a $n \times n$ Haar-distributed random orthogonal matrix (Meckes, 2014).

The proposed model should be viewed as an inferential model instead of a generative model, which is analogous to the "standardization" idea often employed in data analysis. Technically, whenever the data are generated from the normal factor model, i.e., $\mathbf{Y} = \mathbf{B\Omega} + \mathbf{\Delta}$ as in (1.1) with $\mathbf{\Omega}$ being a generated standard normal matrix, it can also be viewed as being generated by a $\sqrt{n}$-orthonormal factor model $\mathbf{Y} = (\mathbf{BK}(\mathbf{\Omega})/\sqrt{n}) \times (\sqrt{n} \cdot \mathbf{V}(\mathbf{\Omega})) + \mathbf{\Delta}$ with loading matrix being $(\mathbf{BK}(\mathbf{\Omega})/\sqrt{n})$. Here

$\mathbf{K}(\boldsymbol{\Omega})$ and $\mathbf{V}(\boldsymbol{\Omega})$ are from the LQ decomposition $\boldsymbol{\Omega} = \mathbf{K}(\boldsymbol{\Omega})\mathbf{V}(\boldsymbol{\Omega})$. The LQ decomposition can be done by the Gram–Schmidt orthogonalization starting from the first row of $\boldsymbol{\Omega}$, resulting in a $K \times K$ lower triangular matrix $\mathbf{K}(\boldsymbol{\Omega})$ and a $K \times n$ orthonormal matrix $\mathbf{V}(\boldsymbol{\Omega})$. By the Bartlett decomposition theorem (Muirhead, 2009), $\mathbf{V}(\boldsymbol{\Omega})$ is uniform on the Stiefel manifold $St(K, n)$ if $\boldsymbol{\Omega}$ is a matrix with i.i.d. standard Gaussian elements. The new loading matrix $(\mathbf{BK}(\boldsymbol{\Omega})/\sqrt{n})$ inherits the same generalized lower triangular structure (Fruehwirth-Schnatter and Lopes, 2018) from $\mathbf{B}$ (if it possesses any) and they are identical in the asymptotic sense as $n \to \infty$. Thus, the $\sqrt{n}$-orthonormal factor model reallocates the magnitude variability of $\boldsymbol{\Omega}$ to the loading matrix so that the posterior inference of the new loading matrix, and thus the whole model, becomes less sensitive to the prior. Besides having the same model interpretability, the $\sqrt{n}$-orthonormal factor model is shown by simulations (under various prior setups) to have two major advantages:

(a) **Robustness.** The posterior distribution is robust against the choice of the prior distribution for elements of the loading matrix in the "Large s, Small n" scenario. The posterior consistency can hold for a broader set of prior choices including the one from Ročková and George (2016).

(b) **Efficiency.** Gibbs samplers for the normal factor model can be easily adapted to handle $\sqrt{n}$-orthonormal factors by only modifying the conditional sampling step for $\boldsymbol{\Omega}$. This modification requires negligible computational cost, but leads to a significant efficiency gain in MCMC sampling. We demonstrate this improvement with both simulations and a real data example.

Compared to other methods for boosting MCMC (e.g., Fruehwirth-Schnatter and Lopes (2018)), our approach is also more straightforward and easier to implement. For these reasons, we suggest to use the $\sqrt{n}$-orthonormal factor model in place of the normal factor model before specifying priors and conducting the downstream Bayesian analysis in high-dimensional settings.

All consistency and convergence concepts in our work are in the frequentist (repeated-sampling) sense. Take the loading matrix for example. If for any open neighborhood $\mathcal{N}$ of an entry of the true loading matrix (the magnitude of entries is at the constant order), the probability for a random draw from the posterior distribution of that entry to fall in $\mathcal{N}$, as a function of the data in the repeated sampling sense, converges to 1 almost surely as n and G go to infinity, we say that the posterior inference of the loading matrix is consistent, or simply that *"the posterior sample of the loading matrix converges to the truth."*

The article is structured as follows. Section 2 introduces Bayesian sparse factor models of Ročková and George (2016), Ghosh and Dunson (2009) and Bhattacharya and Dunson (2011). Under these frameworks, Section 3 illustrates by a synthetic example the posterior inconsistency problem in high dimensions, especially the 'magnitude inflation' phenomenon under the SpSL prior from Ročková and George (2016). Section 4 provides theoretical explanations for the phenomenon. Section 5 reveals the connection between the posterior inconsistency and the factor modeling assumption, and proposes the $\sqrt{n}$-orthonormal factor model whose posterior consistency can be guaranteed. Section 6

numerically verifies the robustness and efficiency gain of using the $\sqrt{n}$-orthonormal factor model for Bayesian inference. Section 7 presents a real-data application, and Section 8 concludes with a short discussion.

2 Bayesian sparse factor models and inference

2.1 Prior settings for loading coefficient selection

In order to enhance model identifiability and interpretability, one often imposes a sparsity assumption for the loading matrix. Two typical types of priors for encoding sparsity are SpSL priors and continuous shrinkage priors (see more reviews and discussions in Shin and Liu (2021)). In this article, we consider the prior setups for the loading matrix discussed in Ročková and George (2016), Ghosh and Dunson (2009), and Bhattacharya and Dunson (2011), with a primary focus on the first one due to both theoretical convenience and its prominent and typical posterior behavior under high dimensions. The idiosyncratic variance matrix $\mathbf{\Sigma}$ is assumed to be diagonal with elements σ_j^2 following a conjugate prior: $\sigma_1^2, \cdots, \sigma_G^2 \overset{i.i.d.}{\sim}$ Inverse-Gamma$(\eta/2, \eta\varepsilon/2)$ in all considered prior setups.

The SpSL-IBP prior (Ročková and George, 2016) Let β_{jk} denote the (j,k)-th element of the loading matrix $\mathbf{B}$. Then, *a priori*, the β_{jk}'s follow a SpSL prior and are mutually independent given the hyper-parameters, i.e.,

$$[\beta_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1] = (1-\gamma_{jk})\psi(\beta_{jk}|\lambda_0) + \gamma_{jk}\psi(\beta_{jk}|\lambda_1), \quad \lambda_0 \gg \lambda_1, \tag{2.1}$$

where $[\cdot \mid \cdot]$ is a generic notation for conditional distributions, $\psi(\beta \mid \lambda) = \frac{\lambda}{2}\exp(-\lambda|\beta|)$ denotes the Laplace (λ) distribution, and the binary indicator γ_{jk} follows

$$\gamma_{jk} \mid \theta_k \overset{ind}{\sim} \text{Bernoulli}(\theta_k) \quad \text{and} \quad \theta_k = \prod_{l=1}^{k} \nu_l, \quad \nu_l \overset{i.i.d.}{\sim} \text{Beta}(\alpha, 1). \tag{2.2}$$

Note that θ_k decreases with respect to k. The prior on the θ_k's is known as the Indian Buffet process (IBP) prior, which is employed to select the true factor dimensionality adaptively. Ročková and George (2016) proposed a PXL-EM (parameter expanded likelihood EM) algorithm, a Bayesian variant of parameter-expanded EM (Liu et al., 1998) for posterior mode detection, which converges dramatically faster than the EM (expectation-maximization) algorithm (Dempster et al., 1977) in finding the *maximum a posteriori* (MAP) estimator (i.e., $\hat{\mathbf{B}}, \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Theta}}$ that maximizes $\pi(\mathbf{B}, \mathbf{\Sigma}, \mathbf{\Theta} \mid \mathbf{Y})$ where $\mathbf{\Theta}$ is the vector formed by θ_k's) and also demonstrated the consistency of the MAP estimator in estimating the loading matrix under the "Large s, Small n" setting.

The modified Ghosh-Dunson prior We modify the prior setting introduced by Ghosh and Dunson (2009) to a SpSL form. More precisely, each element β_{jk} is expressed as the product of a column-wise magnitude parameter r_k and the 'normalized' loading element q_{jk}, i.e., $\beta_{jk} = r_k q_{jk}$ with

$$[r_k \mid \lambda] = \varphi(r_k|\lambda) \quad \text{and} \quad [q_{jk} \mid \gamma_{jk}, \lambda_0, \lambda_1] = (1-\gamma_{jk})\varphi(q_{jk}|\lambda_0) + \gamma_{jk}\varphi(q_{jk}|\lambda_1), \tag{2.3}$$

where $\varphi(\cdot \mid \lambda)$ is the normal density with mean 0 and precision λ, and $\lambda_0 \gg \lambda_1 \gg \lambda$. We assign γ_{jk} the same prior as in (2.2). Ghosh and Dunson (2009)'s original model corresponds to assuming $\lambda_1 = 1$, $\gamma_{jk} \equiv \theta_k \equiv 1$, i.e., a normal instead of normal mixture prior for the q_{jk}. With this dependent prior specification, the number of the "slab parameters" is greatly reduced since all elements in each column of $\mathbf{B}$ share a common slab parameter r_k. This idea of reducing the number of slab parameter is useful for curbing the influence of the priors in a high-dimensional setting.

The multiplicative gamma process (MGP) shrinkage prior Bhattacharya and Dunson (2011) consider a shrinkage-type prior with the degree of shrinkage increasing across the column index as follows,

$$\beta_{jk} \mid \phi_{jk}, \tau_k \sim \mathcal{N}\left(0, \phi_{jk}^{-1}\tau_k^{-1}\right), \quad \phi_{jk} \sim \text{Gamma}(v_1/2, v_2/2), \tag{2.4}$$

$$\tau_k = \prod_{l=1}^{k} \delta_l, \quad \delta_1 \sim \text{Gamma}\,(a_1, 1)\,, \quad \delta_l \sim \text{Gamma}\,(a_2, 1)\,, \quad l \geqslant 2, \quad a_2 > 1. \tag{2.5}$$

Here τ_k is a global shrinkage parameter for the k-th column and the ϕ_{jk}'s are local shrinkage parameters for elements in the k-th column. Note that $\beta_{jk}\tau_k^{1/2}$ plays the same role as q_{jk} in the modified Ghosh-Dunson prior and $\tau_k^{-1/2}$ corresponds to the magnitude parameter r_k. The two priors differ in the representation of sparsity (the SpSL vs. the continuous shrinkage form) and how factor dimensionality is selected. Bhattacharya and Dunson (2011) decide the true factor dimensionality using the MGP prior (2.5) that induces an increasing shrinkage effect on β_{jk} as k grows.

The aforementioned priors all have posterior consistency (for $\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}$) guaranteed under a fixed G and $n \to \infty$ setting. However, they induce distinctive posterior behaviors in the "Large s, Small n" setting when the factors are assumed to be standard normal in the inferential model.

2.2 Standard Gibbs sampling procedures

We explore the posterior distributions under different prior settings via Gibbs sampling (Gelfand and Smith, 1990; Liu, 2008; Tanner and Wong, 1987). Take the SpSL-IBP prior as an example, the full posterior distribution of the parameters, $(\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta})$, can be written generically as

$$\pi(\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta} \mid \mathbf{Y}) \propto f(\mathbf{Y}|\mathbf{B}, \mathbf{\Omega}, \mathbf{\Sigma}) f(\mathbf{\Omega}) p(\mathbf{B}|\mathbf{\Gamma}) p(\mathbf{\Gamma}|\mathbf{\Theta}) p(\mathbf{\Theta}) p(\mathbf{\Sigma}), \tag{2.6}$$

where $\pi(\cdot)$ is a generic notation representing the posterior density, $f(\cdot)$ stands for the model likelihood, and $p(\cdot)$ denotes the imposed prior density. As explained earlier, $\mathbf{\Omega}$ is the $K \times n$ factor matrix with column vectors $\boldsymbol{\omega}_i$, $\mathbf{\Gamma}$ is the $G \times K$ *feature allocation matrix* with entries given by γ_{jk}, $\mathbf{\Sigma}$ is the diagonal matrix of idiosyncratic variances, and $\mathbf{\Theta}$ is the K-dimensional feature *sparsity probability vector* formed by the θ_k's. Observation $\mathbf{Y}$ is formatted as a $G \times n$ matrix with columns $\mathbf{y}_i$. A standard Gibbs sampler for sampling from the full posterior distribution (2.6) iteratively updates each component according to the corresponding conditional distributions (see Appendix B (Ma and Liu, 2021) for a detailed prescription).

Due to multimodality of the posterior distribution caused by the invariance of the likelihood function under matrix rotations (therefore only the sparsity prior can provide information to differentiate different modes) and the strong ties between the factor loading and common factors, this basic Gibbs sampler is very "sticky" and can only explore a small neighborhood of the initial values. By initializing the sampler from a "good" value, such as the MAP estimate found by the PXL-EM algorithm, however, this sampler appears to be a reasonable tool for revealing the local posterior behavior around the initial value. More dramatic global MCMC transition moves are required in order to have a fully functional MCMC sampler (see Appendix C (Ma and Liu, 2021)).

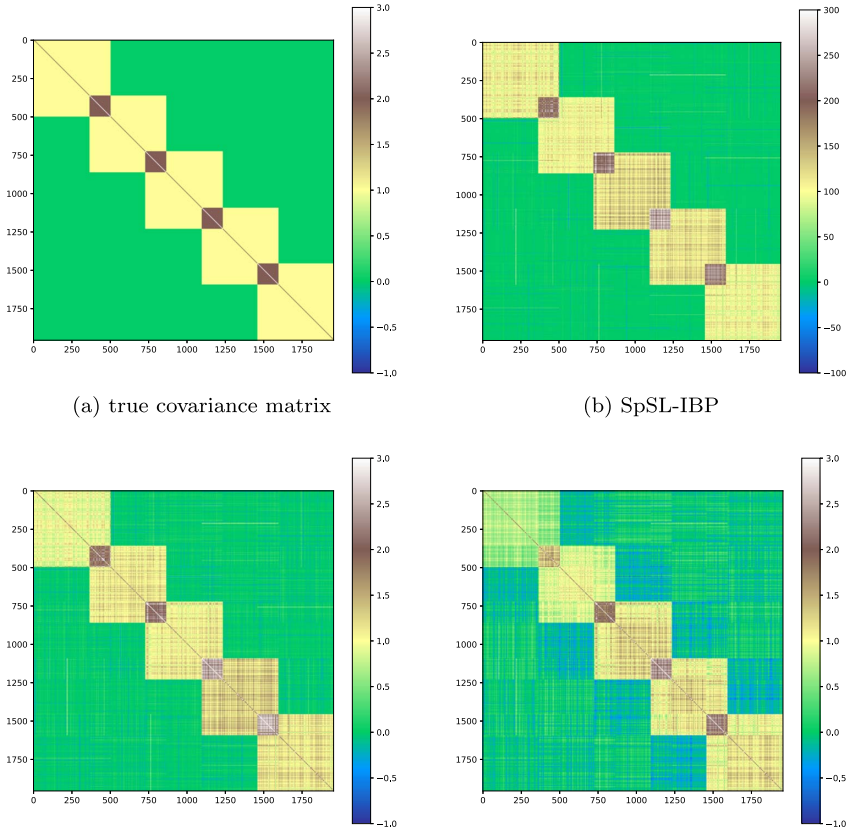
3 The posterior inconsistency phenomenon

3.1 A synthetic example

We generate a dataset from model (1.1) similar to that of Ročková and George (2016), which consists of $n = 100$ observations, $G = 1956$ responses, $K_{true} = 5$ factors drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_5)$, and $\mathbf{\Sigma}_{true} = \mathbf{I}_G$. The true loading matrix is a block diagonal matrix as shown in the leftmost sub-figure of Figure 2, where a black entry stands for 1 and a blank one for 0 (thus $s = 500 > n$). With the synthetic dataset, we run the basic Gibbs sampler for each model (using different priors with the normal factor assumption) under the following tuning setups: (i) $\alpha = 1/G, \lambda_0 = 20, \lambda_1 = 0.1$ for the SpSL-IBP prior; (ii) $\alpha = 1/G, \lambda = 0.001, \lambda_0 = 200, \lambda_1 = 1$ for the modified Ghosh-Dunson prior; (iii) $v_1 = v_2 = 3, a_1, a_2 \sim \text{Gamma}(2, 1)$ for the MGP prior. We set $K = 8$ (run the samplers with K factors) and $\eta = \epsilon = 1$ by default for all the three priors. The loading matrix $\mathbf{B}$ and the idiosyncratic variance matrix $\mathbf{\Sigma}$ are initialized at their MAP estimates from the PXL-EM algorithm.

Heat-maps of the estimated covariance matrix $\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}$ from three different approaches are presented in Figure 1. The first panel plots the data generating covariance matrix as a reference. The other three panels show the posterior mean estimates of the covariance matrix, $\bar{\mathbf{B}}\bar{\mathbf{B}}^T + \bar{\mathbf{\Sigma}}$, under the three priors discussed in Section 2.1. Here, $\bar{\mathbf{B}}$ and $\bar{\mathbf{\Sigma}}$ are the posterior means based on 1500 posterior samples obtained from the Gibbs sampler. We use this form of the estimates in order to maintain the factor model structure. The three priors induce quite different posterior behaviors, among which the posterior estimate under the modified Ghosh-Dunson prior is the closest to the data generating true value. For the MGP prior, posterior estimate of some zeros elements (blue area in panel (d) of Figure 1) are "twisted"—not penalized to values close to the truth 0. The posterior estimate under the SpSL-IBP prior exhibits an interesting "magnitude inflation" phenomenon (the range of the color-bar in panel (b) is 100 times larger than the others), although the relative magnitudes after rescaling look most similar to the true ones.

Under the SpSL-IBP prior with $\lambda_0 = 20$ and $\lambda_1 = 0.1$, we show in Figure 2 ten snapshots of the heat-map of $|\mathbf{B}|$ in a Gibbs sampling trajectory with all model parameters initialized at their true values. We can observe that the direction of each column vector in the loading matrix is well preserved during Gibbs iterations, whereas the absolute value of every non-zero element increases over the iteration and eventually stabilizes around



(a) true covariance matrix

(b) SpSL-IBP

(c) modified Ghosh-Dunson

(d) MGP

Figure 1: Heat-maps of the estimated covariance matrix $\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}$. Panel (a) plots the data generating covariance matrix. Panels (b), (c), and (d) show the covariance matrix estimated using 1500 posterior samples under each prior setup. Note that the colorbar in panel (b) is 100 times larger than others, indicating the much inflated covariance matrix estimate under the SpSL-IBP prior.

a value much larger than the true one. As a demonstration of the inflation, Figure 3(a) further displays the trace plot of $\log(|\beta_{1,1}|)$ with $\lambda_1 = 0.001$ and 0.1, respectively, which also indicates the slow convergence of the basic Gibbs sampler under a small λ_1. As detailed in Appendix C (Ma and Liu, 2021), the convergence can be dramatically improved by adding a few scaling group moves (Liu and Wu, 1999; Liu and Sabatti, 2000) to the Gibbs sampler. The degree of inflation is influenced by the ratio of the number of observations n over the average number of nonzero elements of each column in the true factor loading matrix, s, as well as the choice of independent slab priors. For example, as in Figure 3(b), when n is increased from 100 to 1000, the posterior samples of the loading matrix stabilize around somewhere closer to the true loading matrix.

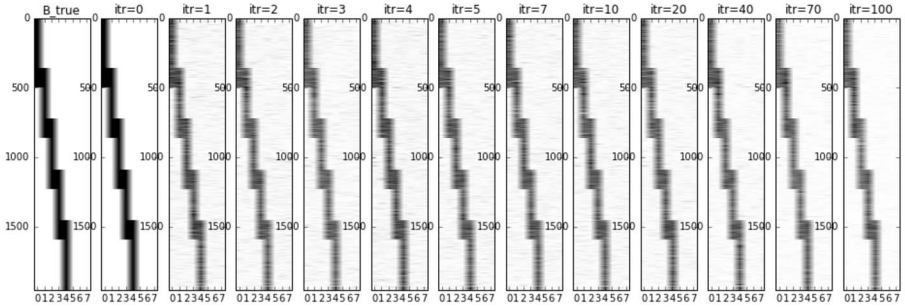
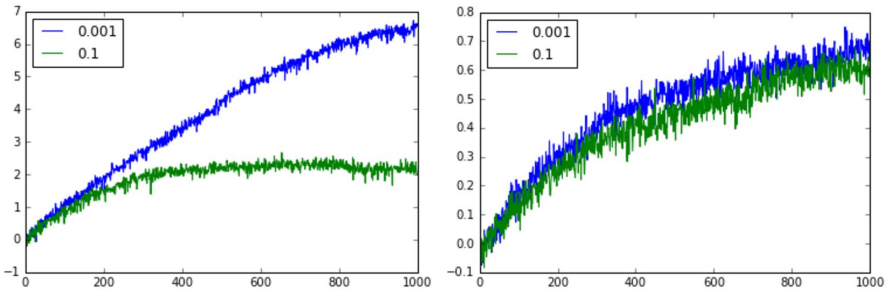


Figure 2: Heat-maps of $|\mathbf{B}|$ in 100 iterations from the Gibbs sampler under the SpSL-IBP prior setup. The gray scale of an entry β_{jk} in one subplot is decided by $|\beta_{jk}|/\max_{j,k}\{|\beta_{jk}|\}$. The darker entries imply a larger ratio. The directions of the columns of the loading matrix are well preserved throughout the Gibbs iterations.



(a) $n = 100$ (b) $n = 1000$

Figure 3: Trace plots of $\log(|\beta_{1,1}|)$ from a Gibbs sampler under the SpSL-IBP prior with $\lambda_0 = 20$, and $\lambda_1 = 0.001$ and 0.1, respectively, for data sets of size (a) $n = 100$, and (b) $n = 1000$. The samples of $\beta_{1,1}$ stabilize around a much larger value than its true value 1. The inflation of the samples is more severe when n is smaller or the variance of slab priors is larger.

3.2 Magnitude inflation and direction consistency

Our numerical results revealed some perplexing consequences of using independent (conditional on the feature allocation $\mathbf{\Gamma}$) SpSL priors for a Bayesian factor model when $s \geq n$, which can be summarized as "magnitude inflation" and "direction consistency". While the former means that the posterior draws of the loading matrix are inflated entry-wise compared with the true loading matrix with the inflation magnitude dependent on how diffuse the slab prior is, the latter says that the direction of columns of posterior samples of the loading matrix somehow still converges to the true direction as $n, s \to \infty$. Intuitively, when the number of independent slab priors employed grows at a faster rate than the number of observations, these priors will overwhelm the signal from data. The

interesting observation is that the overdose of independent slab priors only dilutes the signal for the magnitude part in the loading matrix but has little impact on the identification of the column space. It is also worth mentioning, regardless of the occurrence of "magnitude inflation", the posterior distribution of the idiosyncratic variance matrix $\mathbf{\Sigma}$ still has a nice concentration around the truth.

Traditional literature tends to ignore the inflation problem by treating it as a consequence of the lack of enough observations (i.e., n is too small compared to s) to guarantee posterior sample consistency. However, we notice that, with the same amount of observations, the inflation problem does not occur when using the priors prescribed in Ghosh and Dunson (2009) and Bhattacharya and Dunson (2011) with hyper-parameters within a reasonable range. Their priors impose an additional hierarchical structure on elements in the loading matrix. Moreover, the MAP estimator is rather precise in estimating the true loading matrix and directions of columns of the loading matrix are well captured by the posterior samples (under the SpSL-IBP prior), as in the synthetic example. This suggests that the data provide sufficient information for recovering the true loading up to trivial rotations. Thus, the magnitude inflation phenomena or other posterior inconsistency problems may be caused by some modeling issues.

One likely explanation is that the normal factor model is only weakly identifiable, and thus the posterior distribution is sensitive to the prior on the high-dimensional loading matrix (Figure 1), if not well controlled. The magnitude inflation phenomenon is an expression of the dominating influence of the independent SpSL prior. Although a diffuse prior (such as normal with a large variance) is uninformative in low dimensions, it becomes highly informative in magnitude when many parameters follow this prior independently (i.e., as s grows faster than n in our setting). Ghosh and Dunson (2009) provides a strategy to control the impact of the prior by imposing a hierarchical dependency structure for elements in each column of the loading matrix. In this article, we give another solution though enhancing the model identifiability by changing the factor assumption. This idea is motivated by the theoretical study of the posterior behavior under the independent SpSL prior, which we will introduce in the next two sections. In Section 4 and 5, we focus on the SpSL-IBP prior and follow the notations from Section 2.

4 Characterization of the magnitude inflation

It is generally recognized that in a Bayesian factor model using an improper flat prior on elements of the loading matrix can be dangerous, and will lead to an improper posterior distribution when $G \geq n$. This is in fact not very intuitive, so we illustrate this point with a very simple example with $K = 1$ factor, $n = 2$ observations, and independent noises. Let the two vector observations be $\mathbf{y}_1$ and $\mathbf{y}_2$, each of G-dimensional. We can therefore write $\mathbf{y}_1 = \mathbf{v}_1 + \boldsymbol{\epsilon}_1$, and $\mathbf{y}_2 = \mathbf{v}_2 + \boldsymbol{\epsilon}_2$, with $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_G)$, which is very much like the canonical Normal means problem, with only one additional requirement: $\mathbf{v}_1 = \omega_1 \mathbf{b}$ and $\mathbf{v}_2 = \omega_2 \mathbf{b}$. Here, the model assumes that the factor $\omega_j \sim \mathcal{N}(0, 1)$, and $\mathbf{b}$ is a G-dimensional loading matrix (vector). Thus, marginally we have $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{I}_G + \mathbf{b}\mathbf{b}^T), i = 1, 2$.

A peculiar thing is that in the canonical Normal means problem, if we assign flat priors to $\mathbf{v}_1$ and $\mathbf{v}_2$, their posterior distributions are simply $\mathcal{N}(\mathbf{y}_1, \mathbf{I}_G)$ and $\mathcal{N}(\mathbf{y}_2, \mathbf{I}_G)$, respectively, which are still proper although they yield inadmissible estimators for $\mathbf{v}_1$ and $\mathbf{v}_2$ when $G \geq 3$. However, with the factor model assumptions, which effectively reduces the number of parameters from $2G$ to G, the posterior distribution for $\mathbf{b}$ becomes improper if we assign $\mathbf{b}$ a flat prior and $G \geq 2$.

Mathematically equivalent phenomena occur even in the simple univariate Gaussian mean estimation: let $y \sim \mathcal{N}(\alpha\beta, 1)$. If we assume that $\alpha \sim \mathcal{N}(0, 1)$, then, when assuming a flat prior, the posterior distribution of β is proportional to

$$(\beta^2 + 1)^{-1/2} \exp\left\{-(2(\beta^2 + 1))^{-1} y^2\right\},$$

which is a non-integrable function, thus improper. But if we assume a proper prior on β, its posterior distribution becomes proper but its posterior variance relies heavily on its prior variance. A simple fix of the problem is to realize that we cannot identify both parameters simultaneously and have to let α take a fixed value. These phenomena also happen for the general factor models in certain settings, and our goal is to understand how these issues play out in high dimensional factor models and whether certain intuitive remedies work both theoretically and computationally for these more complex cases.

For the general factor model, we can similarly marginalize out the factor variables and derive the posterior distribution of the loading matrix under the flat prior:

$$\pi(\mathbf{B} \mid \mathbf{Y}, \mathbf{\Sigma}) \propto |\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma})^{-1}(\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i^T)\right]\right\},$$

where the exponential term is both upper and lower bounded by some functions of $\mathbf{Y}$ and $\mathbf{\Sigma}$. Term $|\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}|^{-n/2}$ is lower bounded by $(||\mathbf{B}||_F^2 + \lambda_{max}(\mathbf{\Sigma}))^{-\frac{n \times K}{2}}$, where $||\mathbf{B}||_F$ represents the Frobenius norm of $\mathbf{B}$, and $\lambda_{max}(\mathbf{\Sigma})$ denotes the largest eigenvalue of $\mathbf{\Sigma}$. When the dimension of $\mathbf{B}$, which is $G \times K$, is no smaller than $n \times K$, $\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma})$ will integrate to infinity in the complement region of any bounded set in $\mathcal{R}^{G \times K}$, leading to an improper posterior distribution. If we impose a proper but diffuse slab prior instead of the improper flat prior on elements of $\mathbf{B}$, the posterior distribution can still be very sensitive to the variance of slab prior, as seen in Figure 3.

To formalize this intuition for general Bayesian factor models, we provide the following theorem on the divergence of the posterior distribution of the loading matrix if we use a sequence of increasingly diffuse "slab" priors. Note that for theorems in Section 4, we do not require $\mathbf{\Sigma}$ to be diagonal. To cover generic prior choices, we replace (2.1) with

$$[\beta_{jk} \mid \gamma_{jk}] = (1 - \gamma_{jk})\psi(\beta_{jk}) + \gamma_{jk}\phi(\beta_{jk}), \tag{4.1}$$

where ψ denotes the spike prior density and ϕ denotes the slab prior density.

Theorem 4.1. *Let $\{\phi_m\}_{m=1,\cdots}$ be a sequence of densities such that $lim_{m\to\infty}\phi_m(\beta) = 0$ for every $\beta \in \mathcal{R}$ and there exists a constant $C \in (0, 1)$ such that $\phi_m(\beta) > C\ max_\beta(\phi_m(\beta))$ holds for every β in some non-decreasing Borel sets S_m that converges to $\mathcal{R}$ as $m \to \infty$. If $s = ||\mathbf{\Gamma}||_F^2/K \geq n$, then for any fixed finite-measure Borel set S, $\lim_{m\to\infty} P(\mathbf{B} \in$*

$S|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m) = 0$, *where* $\mathbf{B} \mid \mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$ *is based on the posterior distribution from model (1.1) with normally distributed factors and* ϕ_m *as the slab part in the SpSL prior on loading matrix elements.*

Theorem 4.1 partially explains the magnitude inflation and the dependence of the inflation rate on the choice of the slab prior. Let S be any fixed $G \times K$ dimensional ball. The theorem implies that the probability of a posterior sample $\mathbf{B}$, conditional on $\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$, having a matrix norm smaller than any constant goes to zero as we use a series of slab priors $\{\phi_m\}_{m=1,2,\cdots}$ that is increasingly diffuse. In a general sense, it can also be understood as the convergence in distribution of $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$ towards $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty$ (conditional posterior of B with flat slab prior), which is a point mass at infinity when $s \geq n$. For cases such that $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty$ is indeed proper, e.g., when $s \ll n$ or the assumed distribution on the factors is changed, we strictly have the convergence of $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$ towards $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty$ in distribution as stated in the next theorem. Therefore, if the posterior distribution of the loading matrix is proper under a flat slab prior and the Bayesian consistency is justified in this situation, we have approximately the same consistency when employing a reasonably diffuse slab prior.

Theorem 4.2. *Consider model (1.1) without the normality assumption for the factors. Let* $\{\phi_m\}_{m=1,\cdots}$ *be a sequence of prior densities maximized at 0 such that,* $\forall \beta \in \mathbb{R}$, $\lim_{m\to\infty} \phi_m(\beta)\phi_m^{-1}(0) = 1$. *Let* $\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m)$ *denote the conditional posterior density of* $\mathbf{B}$ *under a SpSL prior for its elements, with the spike density* ψ *and the slab density* ϕ_m, *and let* $\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty)$ *be the one corresponding to the flat slab prior (this is appropriate since the indicator matrix* $\mathbf{\Gamma}$ *is conditioned on). If* $\pi(\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty)$ *is integrable, then* $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, m$ *converges to* $\mathbf{B}|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}, \infty$ *in distribution as* $m \to \infty$.

5 Model modifications and posterior consistency

To concentrate on the magnitude inflation and direction consistency problems, we study behaviors of the posterior distribution of the Bayesian factor model assuming that the diagonal idiosyncratic covariance matrix $\mathbf{\Sigma}$ and the true number of factors (for the basic factor model) or the true feature allocation matrix $\mathbf{\Gamma}$ (for the sparse factor model) are known. In contrast to the solution provided by Ghosh and Dunson (2009), which imposes dependency among the magnitudes of loading matrix elements within the same column through prior setup, we restrict ourselves to a special class of SpSL priors for loading matrix elements, which have a point mass at zero as the spike and a flat (limit of a sequence of increasingly diffuse distributions) slab part. This is a natural choice for being non-informative and is always appropriate when considering the conditional posterior distributions given $\mathbf{\Gamma}$. We focus on studying the connection between posterior consistency and the factor assumption, and demonstrate why the $\sqrt{n}$-orthonormal factor model is a natural choice under high dimensions.

Notations: Let H_n denote the Haar measure (i.e., uniform distribution) on the space of $n \times n$ orthogonal matrices and let m_n be the uniform measure on the Stiefel manifold $St(K, n)$. Let $\mathbf{M}_{i\cdot}$ and $\mathbf{M}_{\cdot j}$ denote the i-th row and the j-th column of matrix $\mathbf{M}$, respectively, as column vectors, and let $\mathbf{M}_{i,j}$ denote the element at i-th row and j-th column of $\mathbf{M}$. $\mathbf{M}_{i_1:i_2}$ denotes the sub-matrix formed by row i_1-th to i_2 and $\mathbf{M}_{i_1:i_2,j_1:j_2}$

denote the sub-matrix formed by rows i_1-th to i_2 and columns j_1 to j_2. Notation $\mathbf{M}^{\perp}$ represents an orthogonal complement (not unique) of $\mathbf{M}$ when $\mathbf{M}$ is not a square matrix, $\mathcal{P}_{(\cdot)}$ represents the projection mapping towards the row vector space of a matrix and $\mathbf{P}_{(\cdot)}$ is the projection matrix of the mapping. Let $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ denote the largest and smallest singular values of a matrix, and let $\lambda_k(\cdot)$ denote the k-th largest singular values. The L_2 norm is denoted by $\|\cdot\|$, the Frobenius norm is denoted by $\|\cdot\|_F$, and the outer product is "$\otimes$".

5.1 The basic Bayesian factor model

We show the posterior consistency of the loading matrix by first studying the posterior consistency of the factor matrix $\mathbf{\Omega}$ (defined in Section 2.2). It is easy to see that, with a flat prior on every element of $\mathbf{B}$, the posterior distribution of $\mathbf{B}$ and $\mathbf{\Omega}$ can be written as:

$$\mathbf{B}_{j\cdot}|\mathbf{Y},\mathbf{\Omega},\mathbf{\Sigma} \overset{ind}{\sim} \mathcal{N}((\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_{j\cdot}, \sigma_j^2(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}), \tag{5.1}$$

$$\pi(d\mathbf{\Omega}|\mathbf{Y},\mathbf{\Sigma}) \propto |\mathbf{\Omega}\mathbf{\Omega}^T|^{-G/2} \exp\left(\sum_{j=1}^{G} \frac{1}{2\sigma_j^2}\mathbf{Y}_{j\cdot}^T\mathbf{\Omega}^T(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_{j\cdot}\right) p_{\mathbf{\Omega}}(d\mathbf{\Omega}), \tag{5.2}$$

where $p_{\mathbf{\Omega}}$ denotes the prior distribution of $\mathbf{\Omega}$ and "$\overset{ind}{\sim}$" means that the $\mathbf{B}_{j\cdot}$'s are mutually independent.

In Section 5, we no longer restrict the factors in $\mathbf{\Omega}$ to follow the standard Normal distribution (i.e., the normal factor assumption), only requiring its distribution $p_{\mathbf{\Omega}}$ to satisfy the following two conditions: (a) $cov(\boldsymbol{\omega}_i) = \mathbf{I}_K$, so as to keep the marginal covariance structure of $\mathbf{Y}$ unchanged; (b) right rotational-invariant (i.e., $\mathbf{\Omega}$ and $\mathbf{\Omega}\mathbf{R}$ follow the same distribution $\forall$ $n \times n$ orthogonal matrix $\mathbf{R}$). Two non-Gaussian examples are: (i) each row of $\mathbf{\Omega}$ follows independently a uniform distribution on the $\sqrt{n}$-radius sphere; (ii) $\mathbf{\Omega}/\sqrt{n}$ is uniform on the Stiefel manifold $St(K,n)$, i.e., $\mathbf{\Omega}/\sqrt{n}$ is the first K rows of a Haar-distributed $n \times n$ orthogonal random matrix. We emphasize here again that this generalization should be viewed as an inference model, instead of a generative model. It is analogous to analyses of contingency tables conditional on the marginal sums.

A straightforward characterization of condition (b) can be made through the LQ decomposition (the transpose of the QR decomposition). Suppose the LQ decomposition of $\mathbf{\Omega} = \mathbf{K}(\mathbf{\Omega})\mathbf{V}(\mathbf{\Omega})$ is done by the Gram–Schmidt orthogonalization starting from the first row of $\mathbf{\Omega}$, resulting in a $K \times K$ lower triangular matrix $\mathbf{K}(\mathbf{\Omega})$ and a $K \times n$ orthonormal matrix $\mathbf{V}(\mathbf{\Omega})$. Then, requirement (b) enables us to generate $\mathbf{\Omega}$ from $p_{\mathbf{\Omega}}$ by generating a pair of $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ from two independent distributions—a marginal distribution on $\mathbf{K}(\mathbf{\Omega})$ (denoted as $p_{\mathbf{K}}$) and a uniform distribution on the Stiefel manifold $St(K,n)$ for $\mathbf{V}(\mathbf{\Omega})$.

Using the LQ decomposition, we can rewrite expression (5.2) as

$$\begin{aligned}\pi(d\mathbf{\Omega}|\mathbf{Y},\mathbf{\Sigma}) \propto &\left(|\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2} p_{\mathbf{K}}(d\mathbf{K}(\mathbf{\Omega}))\right) \\ &\times \left(\exp\Big(\sum_{j=1}^{G} \frac{1}{2\sigma_j^2}\|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_{j\cdot})\|^2\Big) m(d\mathbf{V}(\mathbf{\Omega}))\right)\end{aligned} \tag{5.3}$$

since $|\mathbf{\Omega}\mathbf{\Omega}^T| = |\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|$, and $\mathbf{Y}_{j\cdot}^T\mathbf{\Omega}^T(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}\mathbf{\Omega}\mathbf{Y}_{j\cdot}$ is the square of the length of $\mathbf{Y}_{j\cdot}$'s projection on the row space of $\mathbf{\Omega}$. Therefore, $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ are independent *a posteriori*, and

$$\pi(d\mathbf{K}(\mathbf{\Omega})|\mathbf{Y},\mathbf{\Sigma}) \propto |\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2} p_{\mathbf{K}}(d\mathbf{K}(\mathbf{\Omega})), \tag{5.4}$$

$$\pi(d\mathbf{V}(\mathbf{\Omega})|\mathbf{Y},\mathbf{\Sigma}) \propto \exp\Big(\sum_{j=1}^{G} \frac{1}{2\sigma_j^2}\|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_{j\cdot})\|^2\Big) m(d\mathbf{V}(\mathbf{\Omega})). \tag{5.5}$$

Equation (5.4) implies that $\mathbf{K}(\mathbf{\Omega})$ may have an improper posterior distribution because the likelihood term $|\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T|^{-G/2}$ creates "attractors" when the determinant of $\mathbf{K}(\mathbf{\Omega})\mathbf{K}(\mathbf{\Omega})^T$ is close to 0. Therefore, with large enough G, the right-hand side of (5.4) explodes to infinity fast enough around the attractors and becomes non-integrable, thus leading to an improper posterior distribution for $\mathbf{K}(\mathbf{\Omega})$. In contrast, since $\exp\left(\sum_{j=1}^{G} \frac{1}{2\sigma_j^2}\|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})}(\mathbf{Y}_{j\cdot})\|^2\right)$ is upper bounded by $\exp\left(\sum_{j=1}^{G} \frac{1}{2\sigma_j^2}\|\mathbf{Y}_{j\cdot}\|^2\right)$, the posterior distribution (5.5) for $\mathbf{V}(\mathbf{\Omega})$ is always proper, based on which we can further derive posterior consistency of the row vector space of $\mathbf{\Omega}$.

Consistency of the row vector space of the factor matrix

The consistency of row vector space of $\mathbf{\Omega}$ is intuitive from (5.5) for the noiseless case (i.e., $\mathbf{Y} = \mathbf{B}_0\mathbf{\Omega}_0$), since the exponential term in (5.5) is uniquely maximized when the row vector spaces of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ coincide. As in an annealing algorithm, the exponential term enforces the growing contraction towards the maximum point (where row spaces of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ coincide) as G increases. On the other hand, the prior measure in a neighborhood of the row vector space of $\mathbf{\Omega}_0$ (defined as $p_{\mathbf{\Omega}}(\{\mathbf{\Omega} : ||\mathbf{V}(\mathbf{\Omega}_0)^{\perp}\mathbf{V}(\mathbf{\Omega})^T||_F < \epsilon\})$) gets more diffuse as n grows. Therefore, in an asymptotic regime with $G, n \to \infty$, and under some mild conditions on the growing rate of G and n to ensure that the diffusion is slower than the contraction, the consistency of the row vector space of $\mathbf{\Omega}$ follows immediately as summarized below. Detailed proofs of the lemma and theorem can be found in Appendix D.3 and D.4 (Ma and Liu, 2021).

Lemma 5.1. *Let $\mathbf{B}_{0,G}$ be a $G \times K$ matrix, $\mathbf{\Omega}_{0,n}$ be a $K \times n$ matrix, and $\mathbf{\Sigma}_G$ be a known $G \times G$ diagonal matrix. Suppose noiseless data generated as $\mathbf{Y} = \mathbf{B}_{0,G}\mathbf{\Omega}_{0,n}$ are given. We, however, model each column of $\mathbf{Y}$ as mutually independent and $\mathbf{Y}_{\cdot i} \sim \mathcal{N}_G(\mathbf{B}\mathbf{\Omega}_{\cdot i}, \mathbf{\Sigma}_G)$, $i = 1, \cdots, n$. With a flat prior on each of $\mathbf{B}$'s elements and a right-rotational invariant prior on $\mathbf{\Omega}$, we have the following inequality for the posterior distribution of $\mathbf{\Omega}$:*

$$P(||\mathbf{V}(\mathbf{\Omega}_{0,n})^{\perp}\mathbf{V}(\mathbf{\Omega})^T||_F > \epsilon|\mathbf{Y},\mathbf{\Sigma}_G)$$
$$\leq \left(1 + m_n(\{\mathbf{V} : ||\mathbf{V}_0\mathbf{V}^T||_F < \frac{\epsilon}{L}\}) \times \exp(\frac{3}{8}\epsilon^2\lambda_{min}(\mathbf{\Sigma}_G^{-1/2}\mathbf{B}_{0,G}\mathbf{K}(\mathbf{\Omega}_{0,n}))^2)\right)^{-1},$$

where $L = 2\lambda_{max}(\mathbf{\Sigma}_G^{-1/2}\mathbf{B}_{0,G}\mathbf{K}(\mathbf{\Omega}_{0,n}))/\lambda_{min}(\mathbf{\Sigma}_G^{-1/2}\mathbf{B}_{0,G}\mathbf{K}(\mathbf{\Omega}_{0,n}))$ and $\mathbf{V}_0$ is any fixed $K \times n$ orthonormal matrix.

Lemma 5.1 provides a probability bound between $\mathbf{V}(\mathbf{\Omega})$ sampled from the posterior distribution and $\mathbf{V}(\mathbf{\Omega}_{0,n})$ when there is no noise in the observation $\mathbf{Y}$. Since $||\mathbf{V}(\mathbf{\Omega}_{0,n})^{\perp}\mathbf{V}(\mathbf{\Omega})^T||_F^2$ equals to the sum of squared sine canonical angles between the row space of $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$, lemma 5.1 implies the convergence of these canonical angles towards 0 as $n, G = s \to \infty$ (i.e. the Bayesian consistency of row vector space of $\mathbf{\Omega}$) when $-\log(m_n(\{\mathbf{V} : ||\mathbf{V}_0^{\perp}\mathbf{V}^T||_F < \frac{\epsilon}{L}\})) = o(\epsilon^2\lambda_{min}(\mathbf{\Sigma}_G^{-1/2}\mathbf{B}_{0,G}\mathbf{K}(\mathbf{\Omega}_{0,n}))^2)$, which is the technical requirement that ensures the dilution is "covered up" by the contraction. Base on this lemma, we generalize the consistency of row vector space of $\mathbf{\Omega}$ to the noisy observation case under the "Large G(s), Small n" paradigm.

Definition 5.1. *Let $\mathbf{B}_0$ be a countable array, or a bivariate function of the form $\mathbf{B}_0(j,k)$, with $j = 1, \cdots, \infty$ and $k = 1, \cdots, K$. Intuitively, this is an $\infty \times K$ matrix. We say that $\mathbf{B}_0$ is a regular infinite loading matrix if there are two universal constants $C_1, C_2 > 0$ such that, $\|(\mathbf{B}_0)_{j\cdot}\| \leq C_1$ and $\lambda_{min}((\mathbf{B}_0)_{1:j})/\sqrt{j} \geq C_2$ for $j = 1, \cdots, \infty$.*

Theorem 5.2. *Suppose $\mathbf{B}_0$ is a regular infinite loading matrix. Let $\mathbf{\Omega}_{0,n}$ be a $K \times n$ matrix with linear independent rows and let $\mathbf{\Sigma} = diag(\sigma_1^2, \cdots)$ be a known infinite diagonal matrix in which σ_j, $\forall j$, is bounded below and above by constants $c_3 > 0$ and $c_4 < \infty$, respectively. Let $\mathbf{Y}$ be an $\infty \times n$ matrix, whose j-th row is generated from $\mathcal{N}_n((\mathbf{B}_0)_{j\cdot}\mathbf{\Omega}_{0,n}, \sigma_j^2\mathbf{I}_n)$, independently. For every fixed G, consider modeling the i-th column of $\mathbf{Y}_{1:G}$ by $\mathcal{N}_G(\mathbf{B}\mathbf{\Omega}_{\cdot i}, \mathbf{\Sigma}_G)$ for $i = 1, \ldots, n$ with $\mathbf{\Sigma}_G = diag(\sigma_1^2, \cdots, \sigma_G^2)$. With a flat prior on each of $\mathbf{B}$'s elements and a proper right-rotational invariant prior on $\mathbf{\Omega}$, we have, for a random draw $\mathbf{\Omega}$ from its posterior distribution, almost surely (with respect to the randomness in $\mathbf{Y}$) that*

$$||\mathbf{V}(\mathbf{\Omega}_{0,n})^{\perp}\mathbf{V}(\mathbf{\Omega})^T||_F \mid \mathbf{Y}_{1:G}, \mathbf{\Sigma}_G \to 0 \text{ in probability as } G \to \infty.$$

Posterior distribution of the loading matrix

From (5.5), it is clear that data only provide information on the row vector space of $\mathbf{V}(\mathbf{\Omega})$, the posterior distribution of $\mathbf{V}(\mathbf{\Omega})$ conditioned on its row vector space is uniform among all the $K \times n$ orthonormal matrices within the row space. Utilizing the posterior consistency of the row space provided by Theorem 5.2, we can approximate an $\mathbf{V}(\mathbf{\Omega})$ drawn from its posterior by another random variable of the form $\mathbf{O}\mathbf{V}(\mathbf{\Omega}_{0,n})$, where $\mathbf{O}$ is a $K \times K$ uniform (Haar distributed) random orthogonal matrix (see Appendix D.5 (Ma and Liu, 2021) for details).

Let $\mathbf{B}_{0,G}$ denotes the matrix formed by the first G rows of $\mathbf{B}_0$. By plugging $\mathbf{V}(\mathbf{\Omega}) = \mathbf{O}\mathbf{V}(\mathbf{\Omega}_{0,n})$ into the matrix form of (5.1), which can be written as

$$\mathbf{B} \mid \mathbf{Y}, \mathbf{\Omega}, \mathbf{\Sigma} \sim \mathcal{N}_{K\times G}(\mathbf{Y}\mathbf{\Omega}^T(\mathbf{\Omega}\mathbf{\Omega}^T)^{-1}, (\mathbf{\Omega}\mathbf{\Omega}^T)^{-1} \otimes \mathbf{\Sigma}),$$

we obtain a decomposition for the posterior samples of $\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ as:

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbf{B}\mathbf{K}(\mathbf{\Omega}) \mid \mathbf{Y}, \mathbf{\Sigma} \quad \sim \quad & \mathbf{B}_{0,G}(\mathbf{K}(\mathbf{\Omega}_{0,n})/\sqrt{n})\mathbf{O}^T + ((\mathbf{Y} - \mathbf{B}_{0,G}\mathbf{\Omega}_{0,n})/\sqrt{n})\mathbf{V}(\mathbf{\Omega}_{0,n})^T\mathbf{O}^T \\ & + \mathcal{N}_{G\times K}(\mathbf{0}, \frac{1}{n}\mathbf{I}_K \otimes \mathbf{\Sigma}). \end{aligned} \tag{5.6}$$

For a considerable large n and normal true factor matrix $\mathbf{\Omega}_{0,n}$, $\mathbf{K}(\mathbf{\Omega}_{0,n})/\sqrt{n}$, as the Cholesky factor of $\mathbf{\Omega}_{0,n}\mathbf{\Omega}_{0,n}^T/n$, approaches the identity matrix, so the first term of the right hand side of (5.6) approaches $\mathbf{B}_{0,G}\mathbf{O}^T$. Meanwhile, the second term $((\mathbf{Y} - \mathbf{B}_{0,G}\mathbf{\Omega}_{0,n})/\sqrt{n})\mathbf{V}(\mathbf{\Omega}_{0,n})^T\mathbf{O}^T$ is the row projection of the idiosyncratic noise matrix $(\mathbf{Y}-\mathbf{B}_{0,G}\mathbf{\Omega}_{0,n})$ to a K dimensional space, divided by $\sqrt{n}$, which converges in probability to 0 entry-wise as $n \to \infty$. The third term is a centered normal (independent with $\mathbf{O}$) with variance shrinking to 0 as n increases. This implies that under $G = s \gg n \to \infty$ regime, posterior samples of $\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ can be asymptotically expressed as the true loading matrix times an uniform random orthogonal matrix.

Factor assumption and consistency Posterior distributions of $\mathbf{B}$ and $\mathbf{K}(\mathbf{\Omega})$ are coupled. A "deflation" problem of $\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ occurs when the factors in $\mathbf{\Omega}$ of the inferential model are assumed to be normal and $n = O(G)$, in which case the posterior distribution of $\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ can be derived in closed form by the Bartlett decomposition as:

$$\begin{aligned} \frac{1}{\sqrt{n}}(\mathbf{K}(\mathbf{\Omega}))_{k,k}|\mathbf{Y},\mathbf{\Sigma} \sim \frac{1}{\sqrt{n}}\chi_{n-k+1-G},\ k = 1,\cdots,K, \\ \frac{1}{\sqrt{n}}(\mathbf{K}(\mathbf{\Omega}))_{k',k}|\mathbf{Y},\mathbf{\Sigma} \sim \mathcal{N}(0,\frac{1}{n}),\ 1 \le k < k' \le K, \end{aligned} \tag{5.7}$$

where χ_ν denotes the Chi distribution with ν degrees of freedom. Posterior samples of the loading matrix, therefore, have to be inflated correspondingly. Ideally, we desire the convergence of the posterior distribution of $\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ towards a point mass at the identity matrix to guarantee the posterior consistency (up to rotations) of the loading matrix, and can indeed achieve this by imposing a stronger control over $\mathbf{\Omega}\mathbf{\Omega}^T/n$ through the assumption on $p_{\mathbf{\Omega}}$. A particular simple strategy is to require that all factors are orthogonal and have equal norm, which implies that $\mathbf{\Omega}/\sqrt{n}$ is uniform in the Stiefel manifold $St(K,n)$. More discussions are deferred to the end of Section 5.2.

5.2 Sparse Bayesian factor model

With a special feature allocation design, $\mathbf{V}(\mathbf{\Omega})$ is identifiable so that the consistency of the row space of the factor matrix can be generalized to the consistency of $\mathbf{V}(\mathbf{\Omega})$. We impose a *generalized lower triangular structure* (Fruehwirth-Schnatter and Lopes, 2018) on the feature allocation matrix $\mathbf{\Gamma}$ to cope with the rotational invariance problem of the loading matrix. We call $\mathbf{\Gamma}$ a generalized lower triangular matrix if the row index of the top nonzero entry in the k-th column l_k (define $l_0 = 1$, $l_{K+1} = G+1$) increases with k and $\gamma_{jk} = 1$ if and only if $j \ge l_k$. Under the flat SpSL prior (use a mixture of point mass at zero and flat distribution as prior) on entries of $\mathbf{B}$ in the Sparse Bayesian factor model, we can derive the conditional distributions of $\mathbf{B}$ and $\mathbf{\Omega}$: for $j = l_k,\cdots,l_{k+1}-1$,

$$\mathbf{B}_{j,1:k}|\mathbf{Y},\mathbf{\Omega},\mathbf{\Sigma},\mathbf{\Gamma} \overset{ind}{\sim} \mathcal{N}((\mathbf{\Omega}_{1:k}{\mathbf{\Omega}_{1:k}}^T)^{-1}\mathbf{\Omega}_{1:k}\mathbf{Y}_{j\cdot},\ \sigma_j^2(\mathbf{\Omega}_{1:k}{\mathbf{\Omega}_{1:k}}^T)^{-1}), \tag{5.8}$$

$$\pi(d\mathbf{\Omega}|\mathbf{Y},\mathbf{\Sigma},\mathbf{\Gamma}) \propto \prod_{k=1}^{K}|\mathbf{\Omega}_{1:k}{\mathbf{\Omega}_{1:k}}^T|^{-(l_{k+1}-l_k)/2}\exp\left(\sum_{k=1}^{K}\sum_{j=l_k}^{l_{k+1}-1}\frac{1}{2\sigma_j^2}\|\mathcal{P}_{\mathbf{\Omega}_{1:k}}(\mathbf{Y}_{j\cdot})\|^2\right)p_{\mathbf{\Omega}}(d\mathbf{\Omega}), \tag{5.9}$$

where $\mathbf{B}_{j,1:k} = (\beta_{j1}, \beta_{j2}, \cdots, \beta_{jk})^T$ and $p_{\mathbf{\Omega}}$ denotes the distribution assumed on $\mathbf{\Omega}$ such that condition (a) and (b) in Section 5.1 holds.

Given the LQ decomposition $\mathbf{\Omega} = \mathbf{K}(\mathbf{\Omega})\mathbf{V}(\mathbf{\Omega})$ and

$$\mathbf{\Omega}_{1:k} = \mathbf{K}(\mathbf{\Omega})_{1:k}\mathbf{V}(\mathbf{\Omega}) = \mathbf{K}(\mathbf{\Omega})_{1:k,1:k}\mathbf{V}(\mathbf{\Omega})_{1:k},$$

since $\mathbf{K}(\mathbf{\Omega})$ is lower triangular, $\mathbf{\Omega}_{1:k}{\mathbf{\Omega}_{1:k}}^T = \mathbf{K}(\mathbf{\Omega})_{1:k,1:k}{\mathbf{K}(\mathbf{\Omega})_{1:k,1:k}}^T$ is a function of $\mathbf{K}(\mathbf{\Omega})$. $\mathcal{P}_{\mathbf{\Omega}_{1:k}}(\mathbf{Y}_{j\cdot})$ is the projection of $\mathbf{Y}_{j\cdot}$ towards the row vector space of $\mathbf{\Omega}_{1:k}$, which is a function of $\mathbf{V}(\mathbf{\Omega})$. The adoption of the generalized lower triangular structure on feature allocation matrix ensures a separation in likelihood of (5.9) so that the determinant part is connected to $\mathbf{\Omega}$ only through $\mathbf{K}(\mathbf{\Omega})$ and the exponential part only through $\mathbf{V}(\mathbf{\Omega})$. We thus can derive that $\mathbf{K}(\mathbf{\Omega})$ and $\mathbf{V}(\mathbf{\Omega})$ are independent *a posteriori* and that:

$$\pi(d\mathbf{K}(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}) \propto \prod_{k=1}^{K} \mathbf{K}(\mathbf{\Omega})_{k,k}^{-(G-l_k+1)} p_K(d\mathbf{K}(\mathbf{\Omega})), \tag{5.10}$$

$$\pi(d\mathbf{V}(\mathbf{\Omega})|\mathbf{Y}, \mathbf{\Sigma}, \mathbf{\Gamma}) \propto \exp\left(\sum_{k=1}^{K}\sum_{j=l_k}^{l_{k+1}-1} \frac{1}{2\sigma_j^2}\|\mathcal{P}_{\mathbf{V}(\mathbf{\Omega})_{1:k}}(\mathbf{Y}_{j\cdot})\|^2\right) m(d\mathbf{V}(\mathbf{\Omega})). \tag{5.11}$$

Expression (5.11) gives a proper posterior for $\mathbf{V}(\mathbf{\Omega})$, and for the noiseless case (i.e. $\mathbf{Y} = \mathbf{B}_0\mathbf{\Omega}_0$), the density is maximized when the row vector space of $\mathbf{V}(\mathbf{\Omega})_{1:k}$ and $\mathbf{V}(\mathbf{\Omega}_0)_{1:k}$ coincide for $k = 1, \cdots, K$, based on which we can generalize theorem 5.2 to the consistency (up to sign permutations) of $\mathbf{V}(\mathbf{\Omega})$.

Consistency of $\mathbf{V}(\mathbf{\Omega})$

Definition 5.2. *Let $\mathbf{B}_0$ be an $\infty \times K$ matrix with nonzero rows and let $\mathbf{\Gamma}_0$ be a binary matrix of the same shape. We call $\mathbf{\Gamma}_0$ a generalized lower triangular feature allocation matrix of $\mathbf{B}_0$ if it satisfies*

1. *$\mathbb{I}_{(\mathbf{B}_0)_{j,k}\neq 0} \leq (\mathbf{\Gamma}_0)_{j,k}$ holds for $j = 1, \cdots, \infty$, $k = 1, \cdots, K$, where $\mathbb{I}$ is the indicator function;*
2. *$(\mathbf{\Gamma}_0)_{j,k_1} \leq (\mathbf{\Gamma}_0)_{j,k_2}$ holds for $j = 1, \cdots, \infty$, $K \geq k_1 > k_2 \geq 1$.*

Furthermore, for every fixed dimension G, let ψ_G denote the unique permutation of $(1, \cdots, G)$, so that $\psi_G(j_1) < \psi_G(j_2)$ if and only if either (i) $(\sum_k \mathbf{\Gamma}_{j_1,k}) < (\sum_k \mathbf{\Gamma}_{j_2,k})$ or (ii) $(\sum_k \mathbf{\Gamma}_{j_1,k}) = (\sum_k \mathbf{\Gamma}_{j_2,k})$ but $j_1 < j_2$.

Definition 5.3. *Let $\mathbf{B}_0$ be a $\infty \times K$ matrix with nonzero rows and let $\mathbf{\Gamma}_0$ be a generalized lower triangular feature allocation matrix of $\mathbf{B}_0$. The two $G \times K$ matrices $\mathbf{B}_{0,G}$ and $\mathbf{\Gamma}_{0,G}$ are formed by permuting the first G rows of $\mathbf{B}_0$ and $\mathbf{\Gamma}_0$ according to ψ_G (the j-th row of $\mathbf{B}_0$ is the $\psi_G(j)$-th row of $\mathbf{B}_{0,G}$, see Figure 4 for an example). Let $l_{0,k}$ be the row index of the top nonzero entry in the k-th column of the generalized lower triangular matrix $\mathbf{\Gamma}_{0,G}$ (define $l_{0,0} = 1$, $l_{0,K+1} = G+1$), and let $\mathbf{B}_{0,G}^{(k)}$ be the submatrix of $\mathbf{B}_{0,G}$ formed by*

rows indexed from $l_{0,k}$ *to* $l_{0,k+1}-1$ *and columns indexed from 1 to k. We call* $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ *a regular infinite loading pair if there are two universal constants* $C_1, C_2 > 0$ *such that,* $\|(\mathbf{B}_0)_{j\cdot}\| \leq C_1$ *and* $min_k \lambda_{min}(\mathbf{B}_{0,j}^{(k)})/\sqrt{j} \geq C_2$ *for* $j = 1, \cdots, \infty$.

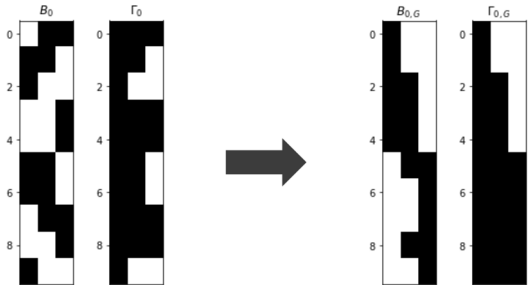


Figure 4: An example of $\mathbf{B}_0$, $\mathbf{\Gamma}_0$, and $\mathbf{B}_{0,G}$, $\mathbf{\Gamma}_{0,G}$ after ψ_G permutation.

Theorem 5.3. *Let* $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ *be a regular infinite loading pair with* $\mathbf{\Gamma}_0$ *known, let* $\mathbf{\Omega}_{0,n}$ *be a* $K \times n$ *matrix with linearly independent rows, and let* $\mathbf{\Sigma} = diag(\sigma_1^2, \cdots)$ *be a known infinite diagonal matrix such that* $C_3 \leq \sigma_j^2 \leq C_4$ *holds for* $j = 1, \cdots$, *with constants* $C_3, C_4 > 0$. *The* j*-th row of* $\infty \times n$ *matrix* $\mathbf{Y}$ *is generated by* $\mathcal{N}_n((\mathbf{B}_0)_{j\cdot}\mathbf{\Omega}_{0,n}, \sigma_j^2 \mathbf{I}_n)$. *For every fixed* G, *let* $\mathbf{Y}_{1:G}$ *denote the matrix formed by permuting the first* G *rows of* $\mathbf{Y}$ *according to* ψ_G *and consider modeling the* i*-th column of* $\mathbf{Y}_{1:G}$ *by* $\mathcal{N}_G(\mathbf{B}\mathbf{\Omega}_{\cdot i}, \mathbf{\Sigma}_G)$ *for* $i = 1, \ldots, n$ *with* $\mathbf{\Sigma}_G = diag(\sigma^2_{\psi_G^{-1}(1)}, \cdots, \sigma^2_{\psi_G^{-1}(G)})$. *With a flat prior on each of* $\mathbf{B}$*'s non-zero element according to the feature allocation matrix* $\mathbf{\Gamma}_{0,G}$ *and a prior on* $\mathbf{\Omega}$ *that is invariant under right orthogonal transformations, for a random draw* $\mathbf{\Omega}$ *from its posterior distribution, we have almost surely (with respect to the randomness in* $\mathbf{Y}$*) that*

$$||\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^{\perp} \mathbf{V}(\mathbf{\Omega})_{1:k}^{T}||_F | \mathbf{Y}_{1:G}, \mathbf{\Sigma}_G, \mathbf{\Gamma}_{0,G} \to 0,$$

for $k = 1, \cdots, K$ *as* $G \to \infty$.

Theorem 5.3 is understood as the consistency (up to sign permutations) of $\mathbf{V}(\mathbf{\Omega})$ for fixed n and $G \asymp s \to \infty$, in the sense that $||\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}^{\perp} \mathbf{V}(\mathbf{\Omega})_{1:k}^{T}||_F$ converges to 0 for all k, which implies that the canonical angles between the row space of $\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}$ and that of $\mathbf{V}(\mathbf{\Omega})_{1:k}$ converge to 0 as $G \to \infty$. When these angles are all equal to 0, $\mathbf{V}(\mathbf{\Omega})$ differs from $\mathbf{V}(\mathbf{\Omega}_{0,n})$ only by a sign for each row. Since the data provides no information on the signs, in the asymptotic regime with $G \asymp s \gg n \to \infty$, we can approximate $\mathbf{V}(\mathbf{\Omega})$ drawn from its posterior distribution by a random sign diagonal matrix $\mathbf{S}$, i.e., a diagonal matrix with *i.i.d.* random signs on the diagonal, times $\mathbf{V}(\mathbf{\Omega}_{0,n})$.

Posterior sample consistency

Recall that from Section 5.1, for the basic Bayesian factor model with $G = s \gg n \to \infty$, $\mathbf{BK}(\mathbf{\Omega})/\sqrt{n}$ drawn from the posterior distribution can be asymptotically represented as

the true loading matrix times a uniform random orthogonal matrix. If the true feature allocation matrix is lower triangular, we have

$$\begin{aligned}\mathbf{B}^{(k)}\mathbf{K}(\mathbf{\Omega})_{1:k}/\sqrt{n}|\mathbf{Y},\mathbf{\Omega},\mathbf{\Sigma}_G,\mathbf{\Gamma}_{0,G} \sim\ & \mathbf{B}^{(k)}_{0,G}(\mathbf{K}(\mathbf{\Omega}_{0,n})_{1:k,1:k}/\sqrt{n})\mathbf{V}(\mathbf{\Omega}_{0,n})_{1:k}\mathbf{V}(\mathbf{\Omega})^T_{1:k} \\ & + ((\mathbf{Y}_{l_k:l_{k+1}-1} - \mathbf{B}^{(k)}_{0,G}(\mathbf{\Omega}_{0,n})_{1:k})/\sqrt{n})\mathbf{V}(\mathbf{\Omega})^T_{1:k} \\ & + \mathcal{N}_{(l_{k+1}-l_k)\times k}(\mathbf{0}, \frac{1}{n}\mathbf{I}_k \otimes \mathbf{\Sigma}^{(k)}_G), \end{aligned} \tag{5.12}$$

whose right hand side converges entry-wise in probability to $\mathbf{B}^{(k)}_{0,G}\mathbf{S}^T_{1:k,1:k}$ under the $G \asymp s \gg n \to \infty$ setting (by similar argument as in Section 5.1). Note that $\mathbf{B}^{(k)}\mathbf{K}(\mathbf{\Omega})_{1:k} = \mathbf{B}_{l_k:l_{k+1}-1}\mathbf{K}(\mathbf{\Omega})$, we can therefore summarize the convergence of $\mathbf{B}^{(k)}\mathbf{K}(\mathbf{\Omega})_{1:k}/\sqrt{n}$ to derive the convergence of posterior samples of $\mathbf{B}\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$ towards $\mathbf{B}_{0,G}\mathbf{S}^T$.

The posterior sample consistency (up to sign permutations) of the loading matrix is immediate once we have $\mathbf{K}(\mathbf{\Omega})/\sqrt{n}$, or equivalently $\mathbf{\Omega}\mathbf{\Omega}^T/n$, from its posterior distribution converging in probability to the identity matrix. The density in (5.10) indicates that the posterior distribution of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ is contributed by two terms: the determinant $\prod_{k=1}^K \mathbf{K}(\mathbf{\Omega})_{k,k}^{-(G-l_k+1)}$ and the model assumption represented by $p_{\mathbf{\Omega}}$. The determinant term creates singularities when $\mathbf{K}(\mathbf{\Omega})_{k,k} = 0$ and the order of these "poles" $\sim s$. When this term dominates, we observe the inflation phenomenon of posterior samples of the loading matrix. Meanwhile, the model assumption term can bound $\mathbf{K}(\mathbf{\Omega})$ away from these singularities by assigning little probability measure in their neighborhoods and also induces the convergence of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ towards the identity matrix (through requirement (a) introduced in Section 5.1). Consequently, the posterior behavior of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ is influenced by both the increasing rate of n, s and the choice of distribution $p_{\mathbf{\Omega}}$. Those $p_{\mathbf{\Omega}}$ that bounds away singularities with high probability and forces a fast convergence of $\mathbf{\Omega}\mathbf{\Omega}^T/n$ towards the identity matrix can permit a faster rate of s going to infinity comparing to n, to guarantee the posterior consistency of the loading matrix. Although our analysis regarding the relation between the factor assumption and the posterior consistency is specific to the SpSL-IBP prior, it reveals the connection between the factor assumption and the strength of posterior contraction towards the truth.

A simple and effective strategy to cope with the posterior inconsistency is to adopt the $\sqrt{n}$*-orthonormal factor model* for Bayesian inference. That is, we assume *a priori* that $\mathbf{\Omega}/\sqrt{n}$ is uniform in the Stiefel manifold $St(K, n)$. With this choice, we have $\mathbf{\Omega}\mathbf{\Omega}^T/n = \mathbf{I}_K$ and that the posterior sample consistency of $\mathbf{B}$ naturally holds even when n has a rather slow growing rate compared with s. Under the standard normal factor generative model, the effect of this modified model is to condition the Bayesian inference on $\mathbf{\Omega}\mathbf{\Omega}^T = \mathbb{E}(\mathbf{\Omega}\mathbf{\Omega}^T)$. While employing the prior structures proposed by Ghosh and Dunson (2009) and Bhattacharya and Dunson (2011) work reasonably well empirically, we have not been able to find rigorous theoretical supports for them. In contrast, using $\sqrt{n}$-orthonormal factors with independent SpSL prior is guaranteed to be consistent in high dimensions given the true feature allocation. In the next section, we also show that numerically the resulting posterior estimates have the smallest MAEs compared with those based on other priors discussed in Section 2.1 (with a normal factor inferential model) and the induced Gibbs sampling algorithm is the most efficient.

6 Numerical results

6.1 Sample $\sqrt{n}$-orthonormal factors

In Section 5, we justify the adoption of the $\sqrt{n}$-orthonormal factor model in the "Large s, Small n" paradigm (i.e., the factor matrix $\mathbf{\Omega}$ scaled by $1/\sqrt{n}$ is uniform in the Stiefel manifold $St(K, n)$). To construct a Gibbs sampler under this new factor model, we only need to revise the conditional sampling of $\mathbf{\Omega} \mid \mathbf{Y}, \mathbf{B}, \mathbf{\Sigma}$ in the basic Gibbs samplers.

Let $\mathbf{\Omega}_{k\cdot}$ denote the k-th row of the factor matrix and $\mathbf{\Omega}_{-k}$ denote the remaining rows, all as column vectors. The conditional distribution $\mathbf{\Omega}_{k\cdot} \mid \mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}$ is altered from a multivariate normal distribution to:

$$\pi(d\mathbf{\Omega}_{k\cdot}|\mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}) \propto f(\mathbf{\Omega}_{k\cdot};\ \bar{\mathbf{\Omega}}_{k\cdot}, \bar{\sigma}_k^2\mathbf{I}_n) \times p_{\mathbf{\Omega}_{-k}}(d\mathbf{\Omega}_{k\cdot}), \tag{6.1}$$

where $p_{\mathbf{\Omega}_{-k}}$ is the uniform measure on the centred $\sqrt{n}$-radius sphere in the orthogonal space of $\mathbf{\Omega}_{-k}$, and $f(\mathbf{\Omega}_{k\cdot};\ \bar{\mathbf{\Omega}}_{k\cdot}, \bar{\sigma}_k^2\mathbf{I}_n)$ is the multivariate normal density function with mean $\bar{\mathbf{\Omega}}_{k\cdot}$ and covariance matrix $\bar{\sigma}_k^2\mathbf{I}_n$, with

$$\bar{\mathbf{\Omega}}_{k\cdot} = (\mathbf{B}_{\cdot k}^T\mathbf{\Sigma}^{-1}\mathbf{B}_{\cdot k})^{-1}(\mathbf{Y} - \sum_{t\neq k}\mathbf{B}_{\cdot t}\mathbf{\Omega}_{t\cdot}^T)^T\mathbf{\Sigma}^{-1}\mathbf{B}_{\cdot k}, \quad \bar{\sigma}_k^2 = (\mathbf{B}_{\cdot k}^T\mathbf{\Sigma}^{-1}\mathbf{B}_{\cdot k})^{-1}.$$

To sample from (6.1), we cut this $\sqrt{n}$-radius sphere by hyperplanes that are orthogonal to vector $\bar{\mathbf{\Omega}}_{k\cdot}$ and denote this collection of intersections of the sphere and hyperplanes as $\{S_d \mid d \in (-\sqrt{n}, \sqrt{n})\}$, where d is the Euclidean distance between the origin and the hyperplane. Essentially, $\{S_d\}$ are $(n\text{-}k)$-dimensional spheres and every point in the same S_d has the same multivariate normal density $f(\cdot; \bar{\mathbf{\Omega}}_{k\cdot}, \bar{\sigma}_k^2\mathbf{I}_n)$, so we can sample $\mathbf{\Omega}_{k\cdot}$ from (6.1) by first sampling d from its marginal distribution and then uniformly sample from sphere S_d given the sampled d. Using the area formula of sphere, we can deduce the marginal distribution for d as

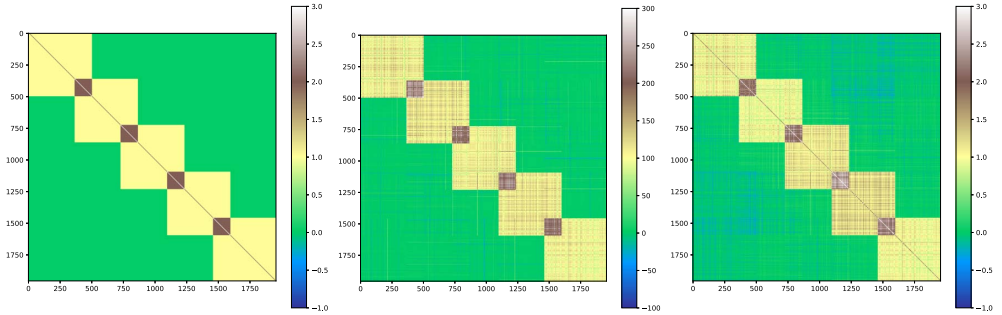
$$\pi(d \mid \mathbf{Y}, \mathbf{\Omega}_{-k}, \mathbf{B}, \mathbf{\Sigma}) \propto (n - d^2)^{(n-K-2)/2} \exp(\|\mathcal{P}_{\mathbf{\Omega}_{-k}^{\perp}}(\bar{\mathbf{\Omega}}_{k\cdot})\| d/\bar{\sigma}_k^2) \tag{6.2}$$

and sample from this unimodal distribution using the Metropolis algorithm. The additional computational cost incurred by the model revision only comes from the Metropolis algorithm and is almost negligible.

6.2 SpSL-orthonormal factor model

We denote the $\sqrt{n}$-orthonormal factor model with the SpSL-IBP prior as the SpSL-orthonormal factor model. We revisit the synthetic example in Section 3.1 to check if the magnitude inflation problem is resolved by employing the $\sqrt{n}$-orthonormal factor inferential model. The tuning parameters are the same as in Section 3.1 and parameters $\mathbf{B}$ and $\mathbf{\Sigma}$ are initialized at the MAP estimate.

Figure 5 shows the comparison between the posterior estimates of the covariance matrix under the normal factor model and the $\sqrt{n}$-orthonormal factor model, using the same SpSL-IBP prior. By comparing panel (c) with panel (b), we conclude that the mag-



(a) truth (b) normal factors (c) $\sqrt{n}$-orthonormal factor

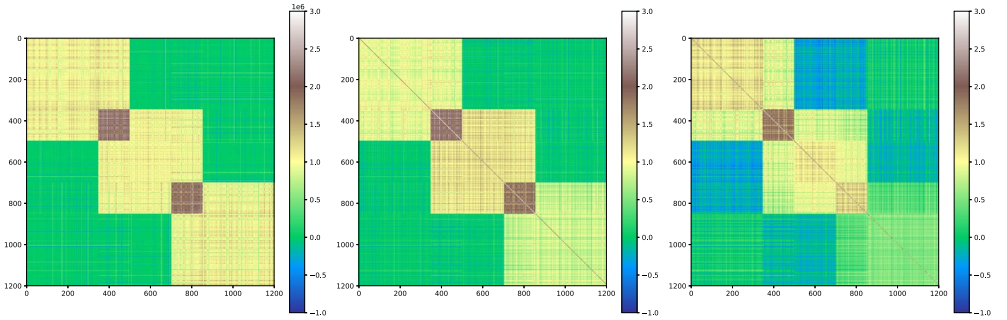
Figure 5: Heat-maps of (a) the true covariance matrix, $\mathbf{B}\mathbf{B}^T + \mathbf{\Sigma}$, of model (1.1); and posterior estimates of the covariance matrix under (b), the normal factor model, and (c), the $\sqrt{n}$-orthonormal factor model, with the SpSL-IBP prior. The estimates are based on 1500 posterior samples. Note that the range of colorbar in Panel (b) is 100 times larger than the other two.

nitude inflation problem is completely resolved through imposing the $\sqrt{n}$-orthonormal restriction on the factors, and the resulting posterior samples of the nonzero elements in the loading matrix are distributed around the truth (see the supplemental figures in Appendix E.3 (Ma and Liu, 2021) for a 90% credible interval of the loading matrix elements).

Under the SpSL-orthonormal factor model, we further show that the posterior distributions of the loading matrix elements are robust against the choice of λ_1 and K (K is the factor number when running the sampler). We conduct two simulation experiments: (i) $K = 8,\ \lambda_0 = 20$ and $\lambda_1 \in \{0.001, 0.01, 0.1, 0.5\}$; and (ii) $\lambda_0 = 20, \lambda_1 = 0.001$ and $K \in \{5, 6, 7, 8\}$. The posterior density of 15 zero elements and 15 nonzero elements (estimated by averaging over the conditional posterior densities) are demonstrated in the supplemental figures in Appendix E.3 (Ma and Liu, 2021). We also observe a similar robustness against the tuning parameter choices in priors of Ghosh and Dunson (2009) and Bhattacharya and Dunson (2011) under the $\sqrt{n}$-orthonormal factor model. Figure 10 of Appendix E.3 depicts a comparison between setting $v_1 = v_2 = 0.5$ versus $v_1 = v_2 = 3$ in the MGP prior for a simulated dataset.

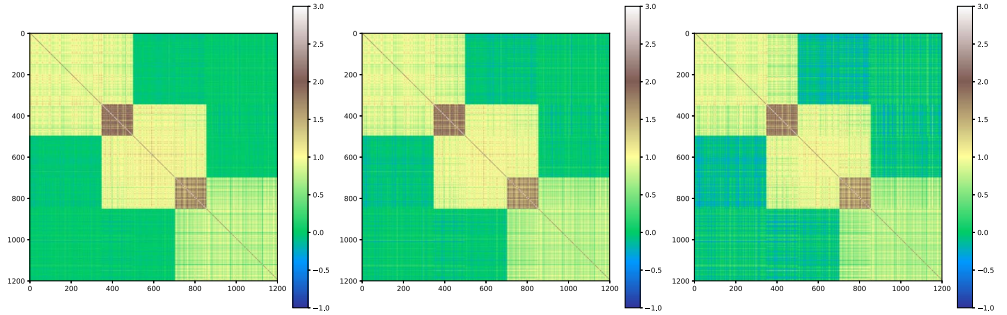
6.3 Robustness and efficiency

We here demonstrate the general robustness of the $\sqrt{n}$-orthonormal factor model for different prior choices by generating datasets from the standard normal factor model (1.1) with $(G, s, n) = (1200, 500, 100)$, corresponding to a $G > s \gg n$ case. The true loading matrix is a block-diagonal matrix with $\beta_{j,k} = 1$ for $k = 1, 2, 3$ and $j = 350 \times (k-1)+1, \cdots 350 \times (k-1)+500$, and $\beta_{j,k} = 0$ otherwise. The elements of the true factor matrix are generated from standard normal. We applied the three priors discussed in Section 2.1 with the following tuning parameters: $\alpha = 1/G, \lambda_0 = 20, \lambda_1 = 0.001$ for the



(a) SpSL-IBP (b) modified Ghosh-Dunson (c) MGP

Figure 6: Heat-maps of the posterior estimates of the covariance matrix resulting from the three priors under the normal factor model with $G > s \gg n$. Note that the range of colorbar for panel (a) is 10^6 times larger than the others.



(a) SpSL-IBP (b) modified Ghosh-Dunson (c) MGP

Figure 7: Heat-maps of the posterior estimates of the covariance matrix resulting from the three priors under the $\sqrt{n}$-orthonormal factor model with $G > s \gg n$.

SpSL-IBP prior; $\alpha = 1/G, \lambda = 0.001, \lambda_0 = 200, \lambda_1 = 1$ for the modified Ghosh-Dunson prior; $v_1 = v_2 = 3$, $a_1, a_2 \sim \text{Gamma}(2, 1)$ for the MGP prior. We set $K = 5$ and $\eta = \epsilon = 1$ as default for all three priors.

Figure 6 shows the estimated covariance matrix under the three different priors in Section 2.1 for a normal factor inferential model. Similar as in Section 3.1, these posterior distributions are quite different from each other, indicating an overly strong influence of the prior specification on the posterior inference. However, as shown in Figure 7, when the $\sqrt{n}$-orthonormal factor inferential model is adopted, the three priors result in similar posterior distributions, which are all closer to the truth.

To quantify the improved estimation efficiencies under the $\sqrt{n}$-orthonormal factor model, we further show in Table 1 the entry-wise mean absolute error (MAE) between each estimated covariance matrix and the true one under three priors and two factor

	SpSL-IBP	modified G&D	MGP
normal factors	$5.7\times10^5/5.5\times10^5/5.5\times10^5$	0.117/0.116/0.104	0.248/0.242/0.229
$\sqrt{n}$-orthonormal factors	0.106/0.099/0.093	0.123/0.111/0.102	0.162/0.146/0.134

Table 1: Mean absolute error of the estimated covariance matrix (computed from 1500 posterior samples of the Gibbs sampler) under different priors and factor assumptions in three generated datasets for the $G > s \gg n$ case.

model specifications, with three independent replications. We can see that the MAEs under the $\sqrt{n}$-orthonormal factor model are either significantly smaller than or comparable to their corresponding ones under the normal factor model. Furthermore, the estimation based on the SpSL-IBP prior under the $\sqrt{n}$-orthonormal factor gives the smallest MAE in all replications among all settings. We observe the same phenomena even in the case with $s \ll n$ as shown later.

Another advantage of adopting the $\sqrt{n}$-orthonormal factor model is to improve MCMC sampling efficiency. In Table 2, we compute the effective sample size (ESS) of the model parameters using 1500 iterations of the Gibbs sampler (after burn-in) under different schemes in the three simulated datasets. The ESS results are averaged across dimensions. For example, $\beta_{\cdot,1}$ indicates ESSs computed by using the average auto-correlations of all nonzero elements in the first column of the loading matrix $\mathbf{B}$. For the normal factor model, successive posterior samples from Gibbs sampling of each nonzero element in $\mathbf{B}$ and each element in the factor matrix $\mathbf{\Omega}$ are strongly positively correlated, which causes high auto-correlations among the MCMC samples. The strong tie between the magnitudes of $\mathbf{B}$ and $\mathbf{\Omega}$ is completely removed under the $\sqrt{n}$-orthonormal factor model, thus resulting in much lower auto-correlations among the MCMC samples and a significant boost of the ESS. Note that the ESS under the normal factor model is quite variable across the three replications, further demonstrating the poor mixing of the Gibbs sampler under this model. Interestingly, the ESS of the σ_j^2's are not impacted by the factor assumption and remains very stable across different priors and different replications.

	normal factors			$\sqrt{n}$-orthonormal factors		
	SpSL-IBP	modified G&D	MGP	SpSL-IBP	modified G&D	MGP
$\beta_{\cdot1}$	250.1/27.8/196.6	24.0/16.0/35.6	41.9/22.6/22.9	1396.6/1445.1/1478.2	1489.9/1507.0/1503.2	1416.7/1407.7/1408.7
$\beta_{\cdot2}$	69.3/73.8/177.4	18.2/29.8/16.1	140.8/19.9/38.9	1422.9/1469.3/1438.5	1470.8/1487.0/1462.5	1399.1/1414.7/1357.1
$\beta_{\cdot3}$	286.5/65.3/74.0	50.2/17.1/39.1	38.9/48.7/19.9	1134.4/1200.6/1475.6	1461.6/1476.8/1483.1	1361.3/1383.0/1403.6
$\Omega_{1\cdot}$	104.2/17.1/83.5	13.8/12.5/21.0	23.3/14.4/14.0	515.7/838.5/438.9	212.8/826.7/495.1	386.4/107.2/115.3
$\Omega_{2\cdot}$	31.6/38.0/75.0	13.2/17.9/11.1	30.2/11.7/12.8	233.5/471.5/392.1	145.4/343.4/403.6	179.4/439.4/237.3
$\Omega_{3\cdot}$	106.1/32.5/33.9	25.1/12.3/19.9	19.5/19.7/12.7	495.2/639.2/219.5	145.9/418.7/416.4	338.8/110.8/127.7
σ^2	1139.2/1242.6/1208.0	1124.8/971.0/1224.5	1072.9/1084.5/1062.9	1194.6/1217.9/1209.4	1043.5/1283.1/1176.1	866.7/1183.0/1000.6

Table 2: ESS of parameters (computed from 1500 posterior samples) using Gibbs sampling under different schemes in three generated datasets of the $G > s \gg n$ case.

We further consider an experiment with $s \ll n < G$. We generate three datasets from model (1.1) with $(G, s, n) = (1200, 10, 200)$ and normal factors. The true loading matrix is a block-diagonal matrix with $\beta_{j,k} = 1$ for $k = 1, 2, 3$ and $j = 7 \times (k - 1) + 1, \cdots 7 \times (k - 1) + 10$, and $\beta_{j,k} = 0$ otherwise. We apply the three priors with the following tuning parameters: $\alpha = 1/G, \lambda_0 = 200, \lambda_1 = 0.001$ for the SpSL-IBP prior; $\alpha = 1/G, \lambda = 0.001, \lambda_0 = 10^4, \lambda_1 = 1$ for the modified Ghosh-Dunson prior; $v_1 = 3, v_2 = 0.003$, $a_1, a_2 \sim \text{Gamma}(2, 1)$ for the MGP prior. We again set $K = 5$

and $\eta = \epsilon = 1$ as default for all three priors. The posterior estimates of $\mathbf{B}$ based on MCMC samples under the three priors are similar to each other and close to the truth under both factor modeling assumptions (see supplemental figures in Appendix E.1 (Ma and Liu, 2021)). This is consistent with the posterior consistency result for using a discrete SpSL prior under the $G > n > s$ setting (Pati et al., 2014). Nevertheless, we still observe from Table 3 an efficiency gain of ESS (for the nonzero loading matrix elements) when using the $\sqrt{n}$-orthonormal factor model. We also see from Table 4 that the SpSL-IBP prior under the $\sqrt{n}$-orthonormal factor model gives the most accurate posterior estimate (with the smallest MAE). In contrast, the MGP prior appears to provide the least accurate estimate.

	normal factors			$\sqrt{n}$-orthonormal factors		
	SpSL-IBP	modified G&D	MGP	SpSL-IBP	modified G&D	MGP
$\beta_{\cdot 1}$	255.0/144.6/268.4	324.2/228.6/250.4	259.4/165.1/303.5	948.7/912.0/857.3	769.3/1017.7/823.4	747.6/805.1/450.5
$\beta_{\cdot 2}$	146.6/257.2/265.8	296.4/242.5/168.2	267.1/237.9/292.8	635.7/691.2/568.4	476.4/724.3/464.8	492.5/575.2/337.3
$\beta_{\cdot 3}$	270.2/278.4/336.8	320.2/346.3/278.6	189.6/212.0/62.7	993.1/843.1/893.5	841.6/862.7/827.8	720.2/690.6/731.1
$\Omega_{1\cdot}$	1129.0/1016.8/1184.5	1070.5/992.2/876.4	951.5/862.6/956.8	1058.5/1113.2/1003.8	969.2/868.5/723.7	852.7/866.5/730.9
$\Omega_{2\cdot}$	1147.9/1149.7/1137.5	708.9/993.2/864.4	851.7/920.1/971.9	988.1/969.7/870.1	801.2/795.1/731.7	739.6/771.7/641.5
$\Omega_{3\cdot}$	1113.5/1189.5/1240.1	1130.3/1157.0/1070.5	880.0/979.7/511.3	1129.9/1127.0/1091.2	1014.8/1072.0/897.9	865.1/885.5/831.3
$\sigma^2_{\cdot}$	1492.8/1482.8/1486.4	1435.8/1441.6/1423.2	1200.7/1210.5/1242.2	1484.3/1472.7/1490.4	1403.3/1395.6/1423.9	1246.3/1199.0/1195.6

Table 3: ESS of parameters (computed from 1500 posterior samples) using Gibbs sampling under different schemes in three generated datasets of the $s \ll n < G$ case.

	SpSL-IBP	modified G&D	MGP
normal factors	$1.24\times10^{-4}/1.24\times10^{-4}/1.14\times10^{-4}$	$1.64\times10^{-4}/1.57\times10^{-4}/3.35\times10^{-4}$	$7.07\times10^{-4}/5.84\times10^{-4}/5.83\times10^{-4}$
$\sqrt{n}$-orthonormal factors	$1.14\times10^{-4}/1.16\times10^{-4}/1.11\times10^{-4}$	$1.84\times10^{-4}/1.66\times10^{-4}/2.59\times10^{-4}$	$7.57\times10^{-4}/5.66\times10^{-4}/6.96\times10^{-4}$

Table 4: Mean absolute error of the estimated covariance matrix (computed from 1500 posterior samples of Gibbs sampler) under different priors and factor assumptions in three generated datasets of the $s \ll n < G$ case.

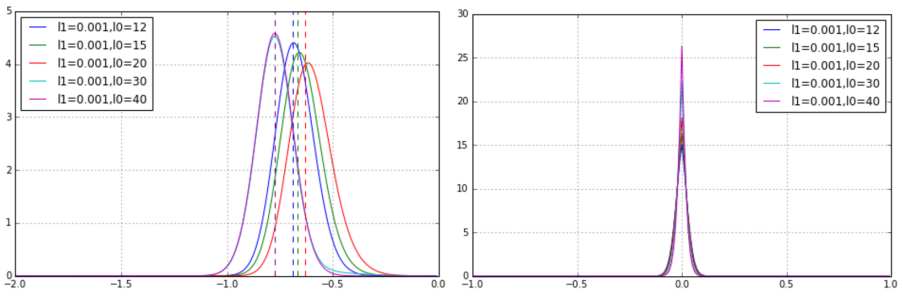
7 Dynamic exploration with application

Although the $\sqrt{n}$-orthonormal factor model can be coupled with general prior assignments on the loading matrix, in this application, we focus on the setup from Ročková and George (2016) (i.e., the SpSL-orthonormal factor model), under which posterior consistency can be theoretically guaranteed given the feature allocation. The application of our Gibbs sampler requires a successful implementation of the PXL-EM algorithm (Ročková and George, 2016) to search for a posterior mode that can serve to initialize the sampler. When applying this framework to real data, the choice of the factor dimensionality K as well as parameters λ_0 and λ_1 for the SpSL prior is crucial. For the choice of K, we make two recommendations: (i) use the estimated number of factors from PXL-EM as a plug-in estimator; (ii) choose K to be sufficiently large initially and discard the useless factors (whose corresponding $\{\gamma_{jk}\}_{j=1,\cdots,G}$ are all zero) in the sampling process, which is similar to the idea of choosing the number of factors adaptively from Bhattacharya and Dunson (2011). The computational complexity of the Gibbs sampler scales linearly with the factor dimensionality in sampler, K.

Parameters λ_0 and λ_1 determine the threshold for a loading matrix's element to follow either a spike or a slab prior. For the PXL-EM algorithm, Ročková and George (2016) proposed a dynamic posterior exploration process to help find the MAP in a se-

quence of prior settings and determine an appropriate value for these hyper-parameters. Initially, they fix λ_1 at a small value and gradually increase λ_0 until the solution path is stabilized. The solution given by the PXL-EM under the final value of λ_0 approximates the MAP estimate under a flat and point mass mixture prior on loading matrix elements and is proposed as the estimator for the parameters. The same procedure can be applied to the full posterior inference based on the SpSL-orthonormal factor model. We observed a similar stabilization of the posterior distributions of every nonzero loading element when performing dynamic exploration for the SpSL-orthonormal factor model, which is illustrated in the application of our method to the cerebrum microarray data from AGEMAP (Atlas of Gene Expression in Mouse Aging Project) database of Zahn et al. (2007). This dataset was also analyzed by Ročková and George (2016) using their PXL-EM algorithm. For every individual mouse in this dataset (5 males and 5 females, at four age periods), cerebrum microarray expression data from 8932 genes are recorded, observations $\mathbf{y}_i, i = 1, \cdots, 40$ for the factor model are taken to be the residuals of the expression values for each of the 8932 genes regressed on age and gender with an intercept.

We ran a Gibbs sampler initialized at the MAP detected by the PXL-EM algorithm with $\lambda_1 = 0.001, \alpha = 1/G$, and λ_0 gradually increasing in the sequence of 12, 15, 20, 30, and 40. As the detected factor dimensionality by the PXL-EM algorithm is 1, we specify K to be 1 in our framework. Figure 8 demonstrates the evolution of the posterior densities of $\beta_{2873,1}$ and $\beta_{1,1}$ as λ_0 changes.

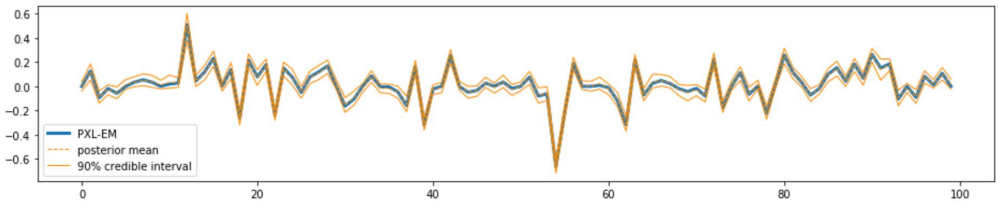


(a) $\beta_{2873,1}$ (b) $\beta_{1,1}$

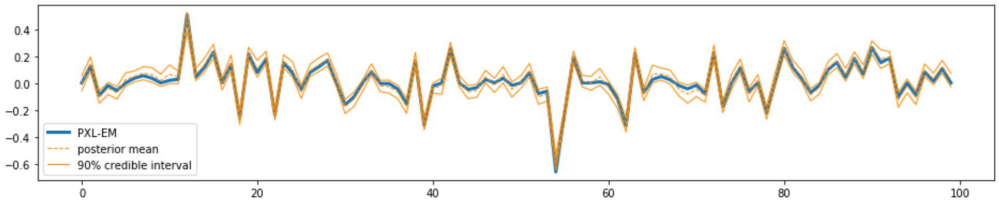
Figure 8: Posterior pdf of (a) $\beta_{2873,1}$ and (b) $\beta_{1,1}$ under SpSL-orthonormal factor model with increasing λ_0.

The posterior distribution of $\beta_{1,1}$ centers at 0 and becomes more and more spiky as λ_0 increases. For the nonzero element $\beta_{2873,1}$, its posterior distributions resemble the normal distribution with a relative stable variance. The posterior mean of $\beta_{2873,1}$ first moves towards zero and then away and stabilizes. This change of direction is caused by the change of its slab indicator $\gamma_{2873,1}$ from 0 to 1 in its posterior samples, in which case the posterior distribution of $\beta_{j,k}$ is only influenced by the slab, but not the spike prior. Vertical dotted lines are the MAP estimates, which are close to the posterior means. Having recognized that the stabilization of the MAP estimates and the posterior distributions occurs almost simultaneously as λ_0 increases, in practice we can find the ideal pair of penalty parameters such that the posterior distribution is

stabilized by looking for the stabilization of the MAP estimates instead of sampling from the posterior with λ_0 on multiple levels. More summary and comparative figures of the posterior simulation are illustrated in Appendix E.2 (Ma and Liu, 2021) with $\lambda_0 = 30$.



(a) The SpSL-orthonormal factor model



(b) The modified Ghosh-Dunson model

Figure 9: Posterior mean and credible interval of $\beta_{1,1}, \cdots, \beta_{100,1}$ estimated from samples of specified model and the MAP estimate from PXL-EM algorithm.

Figure 9 provides a comparison between the posterior inference results from the SpSL-orthonormal factor model ($\lambda_0 = 30$, $\lambda_1 = 0.001$) and the modified Ghosh-Dunson model ($\lambda = 0.001$, $\lambda_0 = 200$, $\lambda_1 = 1$), which shows that the two models give very similar posterior credible intervals (computed using 1000 posterior samples after burn-in) for the loading matrix, and both posterior means are also very close to the MAP estimate from the PXL-EM algorithm. Additionally, the Gibbs sampler for the SpSL-orthonormal factor model results in a much larger ESS compared to that for the Ghosh-Dunson model (e.g., the ESS for $\beta_{55,1}$ are 905.0 and 42.7 for the two methods, respectively). We omit scientific interpretations of the inference results since our goals are only to verify that our procedure gives similar results as those in Ročková and George (2016) based on point estimates under the normal factor model, and to show how to conduct a proper full Bayesian analysis efficiently for this dataset.

In summary, we can start our Bayesian inference for the SpSL-orthonormal factor model by first choosing a small λ_1 and a sequence of increasing λ_0, denoted as $\{\lambda_0^{(t)}\}_{t=1,\cdots}$. We then run the PXL-EM algorithm sequentially with λ_1 and $\lambda_0^{(t)}$ for $t = 1, \cdots$, with parameters initialized at the MAP estimate found in the previous round. The process is terminated when the difference between the new MAP estimate and the one from the previous round is below a chosen threshold. Afterward, we run our Gibbs sampler under the SpSL-orthonormal factor model using the final pair of penalty parameters with $\mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}$ and K initialized at the MAP estimate and $\boldsymbol{\Omega}, \boldsymbol{\Gamma}$ initialized with random draws from their domains.

8 Discussion

A primary intention of this work is to advocate the use the $\sqrt{n}$-orthonormal factor inferential model, which not only enables us to conduct a more robust full Bayesian analysis, but also results in a more efficient posterior sampling algorithm for high-dimensional factor models. We arrive at this point by first numerically revealing the posterior inconsistency of a seemingly standard Bayesian analysis of the normal factor model in high dimensions, and then demonstrating analytically that, under the normal factor model, the posterior distribution is generally too sensitive to the prior specification and such a sensitivity is tied to the weakly identifiable nature of the normal factor assumption. We propose to enforce the $\sqrt{n}$-orthonormal factor assumption as a practical remedy, which should be treated as an inferential model and is analogous to the multinomial model used in the analysis of a contingency table conditional on its marginal sums.

Besides our proposed solution, Bernardo et al. (2003) and Ghosh and Dunson (2009) provided another perspective, which is to reduce the dimensionality of the prior distribution by enforcing certain relationships among the parameters so as to ensure that the prescribed prior does not overwhelm the data. In this article, we provide a further modification of their model by imposing a SpSL prior on the normalized loading matrix's elements, which allows for a greater flexibility in handling sparsity in high dimensions.

Using the SpSL prior employed by Ročková and George (2016), we are able to show theoretically that the adoption of a strict $\sqrt{n}$-orthonormal factor assumption can ensure posterior consistency given the true feature allocation. But this type of rigorous analysis for other models, including the Ghosh-Dunson model and its modification, still evades our vigorous attempts. Interests for future exploration may be focused on efficient posterior sampling algorithms as well as theoretical guarantees of posterior consistency when using such priors. The $\sqrt{n}$-orthonormal factor model itself is also interesting, since the posterior consistency under this model is empirically more robust against prior specification of the loading matrix in the high dimensional setting. It would be interesting to see a mathematical formulation of this empirical result in future works.

Supplementary Material

Supplementary Material for "On Posterior Consistency of Bayesian Factor Models in High Dimensions" (DOI: 10.1214/21-BA1281SUPP; .pdf). The supplementary material provides additional discussions regarding the modified Ghosh-Dunson model, a Gibbs sampler for the SpSL factor model and the scaling group moves. It also includes the proofs of theorems and additional summary figures of simulations and the real application.

References

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." *Bayesian statistics*, 7: 733–742. MR2003537. 927

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 902, 903, 904, 905, 906, 910, 919, 921, 924

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. 905

Efron, B. E. (1973). "Discussion of "Marginalization Paradoxes in Bayesian and Structural Inference"." *Journal of the Royal Statistical Society*. doi: https://doi.org/10.1111/j.2517-6161.1973.tb00952.x. 902

Fruehwirth-Schnatter, S. and Lopes, H. F. (2018). "Sparse Bayesian Factor Analysis when the Number of Factors is Unknown." *arXiv preprint* arXiv:1804.04231. 902, 904, 916

Gelfand, A. E. and Smith, A. F. (1990). "Sampling-based approaches to calculating marginal densities." *Journal of the American statistical association*, 85(410): 398–409. MR1141740. 906

Ghosh, J. and Dunson, D. B. (2009). "Default prior distributions and efficient posterior computation in Bayesian factor analysis." *Journal of Computational and Graphical Statistics*, 18(2): 306–320. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 903, 904, 905, 906, 910, 912, 919, 921, 927

Kass, R. E. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91(435): 1343–1370. 902

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media. MR2401592. 906

Liu, C., Rubin, D. B. and Wu, Y. (1998). "Parameter expansion to accelerate EM: The PX-EM algorithm." *Biometrika*, 85(4): 755–770. doi: https://doi.org/10.1093/biomet/85.4.755. 905

Liu, J. S. and Sabatti, C. (2000). "Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation." *Biometrika*, 87(2): 353–369. MR1782484. doi: https://doi.org/10.1093/biomet/87.2.353. 908

Liu, J. S. and Wu, Y. N. (1999). "Parameter Expansion for Data Augmentation." *Publications of the American Statistical Association*, 94(448): 1264–1274. MR1731488. doi: https://doi.org/10.2307/2669940. 908

Ma, Y. and Liu, J. S. (2021). "Supplementary Material for "On Posterior Consistency of Bayesian Factor Models in High Dimensions"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1281SUPP. 906, 907, 908, 914, 915, 921, 924, 926

Meckes, E. (2014). "Concentration of measure and the compact classical matrix groups." 903

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons. MR0652932. 904

Natarajan, R. and McCulloch, C. E. (1998). "Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?" *Journal of Computational and Graphical Statistics*, 7(3): 267–277. 902

Pati, D., Bhattacharya, A., Pillai, N. S., Dunson, D., et al. (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. 924

Ročková, 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. 902, 904, 905, 907, 924, 925, 926, 927

Shin, M. and Liu, J. S. (2021). "Neuronized priors for Bayesian sparse linear regression." *Journal of the American Statistical Association*, 1–43. doi: https://doi.org/10.1080/01621459.2021.1876710. 905

Tanner, M. A. and Wong, W. H. (1987). "The calculation of posterior distributions by data augmentation." *Journal of the American statistical Association*, 82(398): 528–540. MR0898357. 906

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., et al. (2007). "AGEMAP: a gene expression database for aging in mice." *PLoS genetics*, 3(11): e201. 925