Bayesian Factor-adjusted Sparse Regression

Jianqing Fan*, Bai Jiang*, and Qiang Sun†

May 28, 2020

Abstract

Many sparse regression methods are based on the assumption that covariates are weakly correlated, which unfortunately do not hold in many economic and financial datasets. To address this challenge, we model the strongly-correlated covariates by a factor structure: strong correlations among covariates are explained by common factors and the remaining variations are interpreted as idiosyncratic components. We then propose a factor-adjusted sparse regression model with both common factors and idiosyncratic components as decorrelated covariates and develop a semi-Bayesian method. Parameter estimation rate-optimality and model selection consistency are established by non-asymptotic analyses. We show on simulated data that the semi-Bayesian method outperforms its Lasso analogue, manifests insensitivity to the overestimates of the number of common factors, pays a negligible price when covariates are not correlated, scales up well with increasing sample size, dimensionality and sparsity, and converges fast to the equilibrium of the posterior distribution. Numerical results on a real dataset of U.S. bond risk premia and macroeconomic indicators also lend strong supports to the proposed method.

keywords: factor model, Bayesian sparse regression, posterior contraction rate, model selection.

1 Introduction

High-dimensional linear regression models are useful for a wide array of economic problems (Fan et al., 2011b; Belloni et al., 2012). A typical form of these models is given by

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \sigma\boldsymbol{\varepsilon}_{n\times 1}, \tag{1}$$

where $\mathbf{Y}$ is an n-dimensional response vector, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is a design matrix of n observations and p covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of regression coefficients, σ is (unknown) standard deviation, and $\boldsymbol{\varepsilon}$ is an n-dimensional vector of standard Gaussian noises, independent with $\mathbf{X}$. Both the response vector $\mathbf{Y}$ and covariates $\mathbf{X}_j$ are assumed to be centered without loss of generality, and thus no intercept term is included in the model. Of interest is the high-dimensional regime in which the dimensionality p is much larger than the sample size n. A crucial prerequisite to estimate this model in the high-dimensional regime is the sparseness of $\boldsymbol{\beta}$. That is, the number of non-zero regression coefficients $s = \|\boldsymbol{\beta}\|_0$, called sparsity, is much smaller than the dimensionality p. Model (1) is thereafter referred to as the *sparse regression model* in the rest of this paper.

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544; E-mail: jqfan@princeton.edu, baij@princeton.edu. Fan and Jiang's research was supported by NSF grant DMS-1662139 and NIH grant 2R01-GM072611-14

†Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3; E-mail: qsun@utstat.toronto.edu.

Popular procedures to identify and estimate the non-zero regression coefficients are regularized M-estimation methods (Tibshirani, 1996; Fan and Li, 2001; Candes and Tao, 2007; Fan and Lv, 2008; Zhang and Huang, 2008; Fan and Lv, 2010; Su and Candes, 2016, among others). Meanwhile, Bayesian methods, including those exploiting shrinkage priors (e.g., Park and Casella, 2008; Polson and Scott, 2012; Armagan et al., 2013; Bhattacharya et al., 2015; Ročková and George, 2018) and those exploiting spike-and-slab priors (e.g., Narisetty and He, 2014; Castillo et al., 2015), has been developed.

Much work in this branch of statistical literature is based on the condition that covarites are weakly correlated (Fan and Lv, 2010). Specific types of the weak correlation condition include the mutual coherence condition (Donoho and Huo, 2001; Donoho and Elad, 2003; Donoho et al., 2006; Bunea et al., 2007), the irrepresentable condition (Zhao and Yu, 2006), the restricted eigenvalue condition (Bickel et al., 2009; Fan et al., 2018), the uniform compatibility condition (Bühlmann and van de Geer, 2011, page 157), and the sparse eigenvalue condition (Castillo et al., 2015; Song and Liang, 2017; Fan et al., 2018).

However, many real datasets, especially those in economic and financial studies, are featured by strongly correlated covariates. In an economic or financial dataset, covariates are usually stock returns or macroeconomic indicators over a period of time, which are often influenced by similar economic fundamentals and are thus heavily correlated due to the existence of co-movement patterns (Forbes and Rigobon, 2002; Stock and Watson, 2002a,b; Ludvigson and Ng, 2009).

The above argument shows the necessity to take the underlying correlation structure of covariates into account of the sparse regression analysis, and adjust the weak correlation condition accordingly. For this purpose, we consider factor models (Stock and Watson, 2002a,b; Bai and Ng, 2002; Bai, 2003; Fan et al., 2008, 2011a), in which each observation (row) $\boldsymbol{x}_i \in \mathbb{R}^p$ in the design matrix $\mathbf{X}_{n\times p}$ is decomposable as

$$\boldsymbol{x}_i = \mathbf{B}_{p\times k}\boldsymbol{f}_i + \boldsymbol{u}_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{f}_i$ is a vector of k common factors, $\mathbf{B}$ is a $p \times k$ matrix of factor loading coefficients, and $\boldsymbol{u}_i$ is a vector of p idiosyncratic components, uncorrelated with $\boldsymbol{f}_i$. Let $\mathbf{F} = [\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n]^{\mathrm{T}}$ be the $n \times k$ matrix formed by piling up $\boldsymbol{f}_i$'s, and $\mathbf{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]^{\mathrm{T}}$ be the $n \times p$ matrix formed by piling up $\boldsymbol{u}_i$'s, then the matrix form of the factor model is written as

$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times k}\mathbf{B}^{\mathrm{T}}_{p\times k} + \mathbf{U}_{n\times p}. \tag{2}$$

Each covariate $\mathbf{X}_j$ is now decomposable as the strong correlation part $\mathbf{F}\boldsymbol{b}_j$ and the idiosyncratic component $\mathbf{U}_j$, where $\boldsymbol{b}_j$ is the vector of factor loading coefficients of covariate $\mathbf{X}_j$, i.e., the j-th row of $\mathbf{B}$. Both common factors $\mathbf{F}$ and idiosyncratic components $\mathbf{U}$ are assumed latent, but they are estimatable by using Principal Component Analysis (PCA) (Bai and Ng, 2002; Bai, 2003; Fan et al., 2013; Wang and Fan, 2017). Model (2) embraces the well-known CAPM model (Sharpe, 1964; Lintner, 1975) and Fama-French model (Fama and French, 1993), in which common factors are observable.

If variables $[\mathbf{F}, \mathbf{U}]$ in the factor model (2) are observable or estimable at a high accuracy given $\mathbf{X}$, cross-fertilizing the sparse regression model (1) and the factor model (2) leads to a substantial improvement

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}. \tag{3}$$

where strong correlation parts $\mathbf{F}\boldsymbol{b}_j$'s of covariates $\mathbf{X}_j$'s contribute to the response aggregately. We further propose to drop the constraint $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ and use the following *factor-adjusted sparse regression model* instead.

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \tag{4}$$

We favor this model over model (3) for three reasons. First, model (3) is nested in model (4), thus any method that consistently estimates model (4) would consistently estimate model (3). Second, with no constraint, model (4) is more flexible to explain the variation of the response vector in the regression analysis than model (3). Third, if $\mathbf{F}$, $\mathbf{U}$ are observed or estimated at a high accuracy given $\mathbf{X}$, it is possible to extend the current framework of sparse regression methods towards a systematic methodology for model (4). In contrast, it is inconvenient to enforce the constraint $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ on the sparse high-dimensional vector $\boldsymbol{\beta}$ in iterative optimization algorithms for regularized M-estimation methods or posterior computation algorithms for Bayesian methods.

In the factor-adjust model (4), the weak correlation condition shall be imposed on idiosyncratic components $\mathbf{U}$ rather than original covariates $\mathbf{X}$. Each idiosyncratic component $\mathbf{U}_j$ is the "decorrelated" version of its original covariate $\mathbf{X}_j$ excluding the strong correlation part $\mathbf{F}\boldsymbol{b}_j$. Consequently, idiosyncratic components $\mathbf{U}$ comply with the weak correlation condition more likely than original covariates $\mathbf{X}$ do. Take the sparse eigenvalue (SE) condition, a specific type of the weak correlation condition, as example. A non-vanishing sparse eigenvalue of covariates $\mathrm{SE}(\mathbf{X})$ is required by Bayesian methods to ensure the statistical consistency (Castillo et al., 2015; Song and Liang, 2017) and the computational efficiency (Yang et al., 2016). As we will prove in Section 2,

$$\mathrm{SE}(\mathbf{X}) \leq \mathrm{SE}(\mathbf{U}) \times \max_{j=1}^{p} \frac{\|\mathbf{U}_j\|^2}{\|\mathbf{X}_j\|^2} \times \mathrm{R}(\mathbf{U}),$$

where $\mathrm{R}(\mathbf{U}) \asymp 1$ is a quantity related to the restricted isometry property of $\mathbf{U}$ (Candes and Tao, 2007). Random matrix theories can verify the constant order of this quantity for a broad range of random matrices arising from the field of sparse regression. Clearly, if the strong correlation part $\mathbf{F}\boldsymbol{b}_j$ of each covariate $\mathbf{X}_j$ dominates the individual component $\mathbf{U}_j$ and explains a large portion of the total variation $\|\mathbf{X}_j\|^2 = \|\mathbf{F}\boldsymbol{b}_j + \mathbf{U}_j\|^2 \approx \|\mathbf{F}\boldsymbol{b}_j\|^2 + \|\mathbf{U}_j\|^2$, then $\mathrm{SE}(\mathbf{X})$ would be much smaller than $\mathrm{SE}(\mathbf{U})$.

We also remark that the sparseness assumption on $\boldsymbol{\beta}$ has been implicitly adjusted by model (4). A non-zero β_j in model (4) means that covariate $\mathbf{X}_j$, excluding strong correlations with other covariates, has a specific effect on the response. This is conceptually more reasonable than the original sparseness assumption when covariates are factor-structured. In model (1), if covariates are strong correlated, it does not make sense to assume that any particular covariate $\mathbf{X}_j$ for some $1 \leq j \leq p$ processes a specific influence on the response variable, meanwhile many other covariates that are strongly correlated with $\mathbf{X}_j$ do not.

The factor-adjusted model (4) considered in this paper differs from the factor-augmented models of Stock and Watson (2002b,a); Bai and Ng (2006) in the form

$$\mathbf{Y} = \mathbf{F}_{n\times k}\boldsymbol{\alpha} + \mathbf{W}_{n\times q}\boldsymbol{\gamma} + \sigma\boldsymbol{\varepsilon}. \tag{5}$$

In model (5), common factors $\mathbf{F}$ are extracted from a large panel of data $\mathbf{X}_{n\times p}$ via PCA, yet the q other covariates $\mathbf{W}$ are introduced from outside of the panel. These models are typically low-dimensional with small q. In model (4), covariates $\mathbf{U}$ other than common factors $\mathbf{F}$ are created internally from the panel of data $\mathbf{X}$, allowing to explore an additional explanatory power of the panel. Moreover, the analysis of high-dimensional model (4) in this paper is applicable to the low-dimensional model (5), as model (4) can easily incorporate external variables $\mathbf{W}$ as part of $\mathbf{F}$ and/or $\mathbf{U}$. For simplicity of presentation, we omit the details.

Kneip and Sarda (2011) gave an insightful discussion on the limitation of traditional sparse regression methods on model (1) with factor-structured covariates, and proposed another factor-augmented

model in the form

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha}' + \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}. \tag{6}$$

This model can be transformed as model (4) by the reparameterization $\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$. However, model (6) still requires the weak correlation condition on original covariates $\mathbf{X}$, which, as we have discussed before, is more restrictive than that on decorrelated covariates $\mathbf{U}$.

Fan et al. (2020a) pointed out the failure of regularized M-estimation methods on model (1) with factor-structured covariates and proposed to use the factor-adjusted model (4). They estimated latent variables $[\mathbf{F}, \mathbf{U}]$ in the factor model by PCA, and then performed Lasso to with estimates $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ in place of true variables $[\mathbf{F}, \mathbf{U}]$. Similarly to this paper, they imposed the weak correlation condition on idiosyncratic components $\mathbf{U}$ rather than original covariates $\mathbf{X}$.

We are curious if any Bayesian method consistently identify and estimate the nonzero regression coefficients in the factor-adjusted model (4), and to what extent the factor adjustment and the latent variable estimation decline the performance of the Bayesian method. Given theoretical results on the factor-adjusted Lasso method (Fan et al., 2020a), both questions are still challenging, because the definition of the parameter estimation error rate, the definition of model selection consistency and technical conditions of Bayesian methods are significantly different from those of frequentist methods (Castillo et al., 2015). Even if a Bayesian method is theoretically sound in the asymptotic regime, it is unclear whether it performs better or worse than frequentist methods on finite data.

This paper proposes a semi-Bayesian approach for model (4). As detailed in Section 3, a full-Bayesian approach cannot work easily due to the involvement of latent variables $[\mathbf{F}, \mathbf{U}]$ in the posterior computation. Inspired by Fan et al. (2020a), we consider estimating latent variables by PCA and performing a Bayesian spike-and-slab method with these estimates as covariates. This semi-Bayesian approach results in a pseudo posterior distribution. Theoretical analyses reveal that the pseudo posterior distribution achieves the rate-optimality of parameter estimation and adapts to the unknown sparsity s and unknown standard deviation σ. For these results, we only need the sparse eigenvalue condition on idiosyncratic components $\mathbf{U}$ and the estimation error rate $\sqrt{\log p/n}$ of latent variables $[\mathbf{F}, \mathbf{U}]$. The first condition is easy to hold since $\mathbf{U}$ have been decorrelated, and the second condition is examined under generic conditions of the factor model. Moreover, under a commonly-seen beta-min condition in the literature, the pseudo-posterior distribution correctly identifies the non-zero regression coefficients. Interestingly, although the factor adjustment does not change the estimation error rate of the Bayesian method, it does result in larger constant factors of estimation errors and require stronger sparse eigenvalue and beta-min conditions.

The rest of this paper proceeds as follows. Section 2 compares the sparse eigenvalues of original covariates $\mathbf{X}$ and decorrelated covarites $\mathbf{U}$. Section 3 presents the semi-Bayesian approach for the factor-adjusted model (4). Section 4 establishes the estimation error rate $\sqrt{\log p/n}$ of latent variables in the factor model. Section 5 follows to investigate the parameter estimation error rate and model selection consistency of the pseudo-posterior distribution. Section 6 collects experimental results on simulated datasets. Section 7 evaluates the proposed method on a real dataset of U.S. bond risk premia. Section 8 concludes the paper with a brief discussion. Technical proofs and algorithmic implementation details are detailed in the appendices.

Notation. For an index set ξ, write $|\xi|$ as its cardinality and ξ^c as its complement. For two index sets ξ, ξ', write $\xi \setminus \xi'$ as the set difference. For a vector $\boldsymbol{v}$, let $\boldsymbol{v}_\xi$ denote the sub-vector assembling components indexed by ξ, let $\|\boldsymbol{v}\|$ denote the ℓ_2 norm, and let $\|\boldsymbol{v}\|_0$ denote the number of non-zero

entries. For a matrix $\mathbf{A}_{m_1\times m_2} = [a_{ij}]_{1\le i\le m_1, 1\le j\le m_2}$, write uppercase $\mathbf{A}_j$ for its j-th column, and lowercase $\boldsymbol{a}_i$ for its i-th row. Let $\mathbf{A}_\xi = [\mathbf{A}_j : j \in \xi]$ be the sub-matrix of $\mathbf{A}$ assembling the columns indexed by $\xi \subseteq \{1, \ldots, m\}$. Let $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$ be its element-wise maximum norm, $\|\mathbf{A}\|$ be its operator norm (induced by the ℓ_2 norm of vectors), and $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i,j} a_{ij}^2}$ be its Frobenius norm. Let $\mathrm{vec}(\mathbf{A})$ be the vectorization of $\mathbf{A}$ formed by concatenating column vectors of $\mathbf{A}$. For a symmetric matrix $\mathbf{A}$, write its largest eigenvalue as $\lambda_{\max}(\mathbf{A})$, its smallest eigenvalue as $\lambda_{\min}(\mathbf{A})$, and its trace as $\mathrm{trace}(\mathbf{A})$. Write $\mathrm{diag}(a_1, \ldots, a_m)$ for a diagonal matrix of elements $a_1, \ldots, a_m$. For two positive sequences a_n, b_n, $a_n \succcurlyeq b_n$ (or $b_n \preccurlyeq a_n$) means $b_n = O(a_n)$; $a_n \succ b_n$ (or $b_n \prec a_n$) means $b_n = o(a_n)$; and $a_n \asymp b_n$ means both $a_n \succcurlyeq b_n$ and $a_n \preccurlyeq b_n$.

2 Sparse Eigenvalue of Covariates

This section compares the sparse eigenvalues of original covariates $\mathbf{X}$ and decorrelated covariates $\mathbf{U}$ and evidences that the sparse eigenvalue condition on $\mathbf{U}$ in model (4) holds more likely than that on $\mathbf{X}$ in models (1) and (6) does.

Definition 1 (Sparse Eigenvalue, Definition 2.3 of Castillo et al. (2015))**.** *The $\bar{s}$-order sparse eigenvalue of (the scaled Gram matrix) of the design matrix $\mathbf{X}_{n\times p}$ is defined as*

$$\mathrm{SE}(\mathbf{X}; \bar{s}) = \frac{\min_{\xi:\ |\xi|\le \bar{s}}\ \lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi)}{\max_{j=1}^{p} \|\mathbf{X}_j\|^2}.$$

Definition 2 (Restriced Isometry Property, Definition 10.5.8 of Vershynin (2018))**.** *An $n \times p$ matrix $\mathbf{U}$ satisfies the $\bar{s}$-order restricted isometry property (RIP) with parameters κ_0, κ_1 if*

$$\kappa_0\|\boldsymbol{\beta}\| \le \|\mathbf{U}\boldsymbol{\beta}\| \le \kappa_1\|\boldsymbol{\beta}\|$$

for all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\|\boldsymbol{\beta}\|_0 \le \bar{s}$.

Lemma 1. *In the factor model (2), for any integer $\bar{s} > k$,*

$$\mathrm{SE}(\mathbf{X}; \bar{s}) \le \mathrm{SE}(\mathbf{U}; \bar{s}) \times \max_{j=1}^{p} \frac{\|\mathbf{U}_j\|^2}{\|\mathbf{X}_j\|^2} \times \mathrm{R}(\mathbf{U}; \bar{s}), \quad \textit{with } \mathrm{R}(\mathbf{U}; \bar{s}) = \max_{\xi:\ |\xi|\le \bar{s}} \frac{\lambda_{\max}(\mathbf{U}_\xi^{\mathrm{T}}\mathbf{U}_\xi)}{\lambda_{\min}(\mathbf{U}_\xi^{\mathrm{T}}\mathbf{U}_\xi)}.$$

If $\mathbf{U}$ satisfies $\bar{s}$-order RIP (Definition 2) with parameters κ_0, κ_1 then $\mathrm{R}(\mathbf{U}; \bar{s}) \le \kappa_1^2/\kappa_0^2$.

Proof. From Cauchy Interlacing Theorem it follows that $\lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi) \le \lambda_{\min}(\mathbf{X}_{\xi'}^{\mathrm{T}}\mathbf{X}_{\xi'})$ for two nested models $\xi \supseteq \xi'$, implying that $\lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi)$ of model ξ with size $\xi \le \bar{s}$ achieves the minimum value at some model of size $\bar{s}$. That is,

$$\min_{\xi:\ |\xi|\le \bar{s}}\ \lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi) = \min_{\xi:\ |\xi|= \bar{s}}\ \lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi).$$

Let $s_{\min}(\mathbf{A})$ denote the smallest singular values of matrix $\mathbf{A}$ and let $\boldsymbol{b}_\xi = [\boldsymbol{b}_j :\ j \in \xi]$. From Weyl's theorem on perturbed singular values and the fact that $\mathbf{F}\boldsymbol{b}_\xi$ is of rank at most k, it follows that

$$\lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi) = s_{\min}^2(\mathbf{X}_\xi) \le (s_{\min}(\mathbf{F}\boldsymbol{b}_\xi) + \|\mathbf{U}_\xi\|)^2 = \|\mathbf{U}_\xi\|^2 = \lambda_{\max}(\mathbf{U}_\xi^{\mathrm{T}}\mathbf{U}_\xi) \le \lambda_{\min}(\mathbf{U}_\xi^{\mathrm{T}}\mathbf{U}_\xi) \times \mathrm{R}(\mathbf{U})$$

for each model ξ of size $\bar{s} > k$. On the other hand,

$$\max_{j=1}^{p} \|\mathbf{U}_j\| = \max_{j=1}^{p} \|\mathbf{X}_j\| \times \frac{\|\mathbf{U}_j\|}{\|\mathbf{X}_j\|} \le \max_{j=1}^{p} \|\mathbf{X}_j\| \times \max_{j=1}^{p} \frac{\|\mathbf{U}_j\|}{\|\mathbf{X}_j\|}.$$

Therefore,

$$\mathrm{SE}(\mathbf{X}) = \frac{\min_{\xi:\ |\xi|=\bar{s}}\ \lambda_{\min}(\mathbf{X}_\xi^{\mathrm{T}}\mathbf{X}_\xi)}{\max_{j=1}^p \|\mathbf{X}_j\|^2} \le \frac{\min_{\xi:\ |\xi|=\bar{s}}\ \lambda_{\min}(\mathbf{U}_\xi^{\mathrm{T}}\mathbf{U}_\xi)}{\max_{j=1}^p \|\mathbf{U}_j\|^2} \times \max_{j=1}^p \frac{\|\mathbf{U}_j\|^2}{\|\mathbf{X}_j\|^2} \times \mathrm{R}(\mathbf{U}),$$

proving the first claim. The second claim is trivial. □

The concept of RIP, first introduced by a seminar work of Candes and Tao (2007) in compressed sensing, plays an important role in the recovery of the nonzero regression coefficients by ℓ_1 minimization in place of ℓ_0 minimization, and guarantees the estimation consistency of the Lasso method (Vershynin, 2018, Sections 10.5 and 10.6). In general, a matrix with the concentration of measure property is a good restricted isometry with κ_0/κ_1 being of constant order (Baraniuk et al., 2008). Concrete examples include subgaussian random matrices with independent rows (Vershynin, 2012, Theorem 5.65). To ensure $\bar{s}$-order RIP, these examples require $n \succcurlyeq \bar{s}\log(p/\bar{s})$, which is usually satisfied in the sparse regression setup.

3 Model and Methodology

The goal of this paper is to study the factor-adjusted sparse regression model (4), in which common factors and idiosyncratic components $[\mathbf{F}, \mathbf{U}]$ are latent, but $\mathbf{X}$ are observed through the factor model (2). Each datum (row) $\boldsymbol{x}_i$ in $\mathbf{X}$ is assumed decomposable as $\boldsymbol{x}_i = \mathbf{B}\boldsymbol{f}_i + \boldsymbol{u}_i$ with $\mathbb{E}\boldsymbol{f}_i = \mathbf{0}$, $\mathbb{E}\boldsymbol{u}_i = \mathbf{0}$, and $\mathbb{E}[\boldsymbol{f}_i\boldsymbol{u}_i^{\mathrm{T}}] = \mathbf{0}$. Note that $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{1\le i\le n}$ are not necessarily identically or independently distributed. $\mathbb{E}[\mathbf{F}^{\mathrm{T}}\mathbf{F}/n]$ is normalized as $\mathbf{I}$ without loss of generality, and $\mathbb{E}[\mathbf{U}^{\mathrm{T}}\mathbf{U}/n]$ is denoted by $\boldsymbol{\Sigma}$. The Gaussian noises $\boldsymbol{\varepsilon}$ are independent of $\mathbf{F}$ and $\mathbf{U}$. The number of common factors k is fixed, but the dimensionality p of $\mathbf{U}$ and the sparsity s of its regression coefficients $\boldsymbol{\beta}$ may grow as n increases. Assume $p \succ n$ but $s\log p \prec n$ so that the desired estimation error rate $\epsilon_n = \sqrt{s\log p/n} \to 0$ as $n \to \infty$.

The first step is to estimate latent variables $[\mathbf{F}, \mathbf{U}]$ given $\mathbf{X}$. We follow Bai and Ng (2002); Bai (2003); Fan et al. (2013); Wang and Fan (2017) to use a PCA-based method for this task. Let $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_n$ be the eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$ in the descending order. It is natural to estimate the column space of $\mathbf{F}$ by the eigenspace corresponding to the k largest eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$. Write the eigenequation as

$$\frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{np} \times \frac{\widehat{\mathbf{F}}}{\sqrt{n}} = \frac{\widehat{\mathbf{F}}}{\sqrt{n}} \times \widehat{\boldsymbol{\Lambda}}, \quad \frac{\widehat{\mathbf{F}}^{\mathrm{T}}\widehat{\mathbf{F}}}{n} = \mathbf{I},$$

where $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k)$ is the diagonal matrix of the k largest eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$, and $\widehat{\mathbf{F}}$ is $\sqrt{n}$ times their corresponding eigenvectors. Further, $\mathbf{B}$ and $\mathbf{U}$ are estimated as

$$\widehat{\mathbf{B}} = \frac{\mathbf{X}^{\mathrm{T}}\widehat{\mathbf{F}}}{n}, \quad \widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{F}}\widehat{\mathbf{B}}^{\mathrm{T}} = \left(\mathbf{I} - \frac{\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}}{n}\right)\mathbf{X}.$$

If the number of common factors k is unknown, one may estimate it by

$$\widehat{k} = \operatorname*{argmax}_{1\le j\le \bar{k}} \frac{\widehat{\lambda}_j}{\widehat{\lambda}_{j+1}}, \tag{7}$$

where $\bar{k}$ is a prescribed upper bound for k (Luo et al., 2009; Lam and Yao, 2012; Ahn and Horenstein, 2013). Another viable method for estimating unknown k is by Bai and Ng (2002).

Given estimates $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ for latent variables $[\mathbf{F}, \mathbf{U}]$, we propose a Bayesian spike-and-slab method for parameter estimation and model selection tasks. Let $\xi = \{j : \beta_j \neq 0\}$ be the support of $\boldsymbol{\beta}$. A hierarchical prior $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta})$ with a slab prior on $\boldsymbol{\alpha}$ and a spike-and-slab prior on $\boldsymbol{\beta}$ is assigned.

$$\begin{aligned} \sigma^2 &\sim g(\sigma^2), \\ \boldsymbol{\alpha} &\sim \prod_{j=1}^{k} h_1(\alpha_j), \\ 1\{j \in \xi\} &\sim \texttt{Bernoulli}(s_0/p), \\ \boldsymbol{\beta}_\xi &\sim \prod_{j \in \xi} h_2(\beta_j/\tau_j)/\tau_j, \quad \boldsymbol{\beta}_{\xi^c} = 0, \end{aligned} \tag{8}$$

where g is a positive continuous density function on $(0, \infty)$, e.g., the inverse-gamma density; h_1 and h_2 are "slab" positive density functions on $(-\infty, +\infty)$, e.g., the Gaussian density $e^{-z^2/2}/\sqrt{2\pi}$ or the Laplace density $e^{-|z|/2}/2$; hyperparameters $\tau_1, \ldots, \tau_p$ control the scales of regression coefficients $\beta_1, \ldots, \beta_p$; and hyperparameter s_0 controls the sparsity of model ξ. For the scaling hyperparameters, we set $\tau_j^{-1} = \|\widehat{\mathbf{U}}_j/\sqrt{n}\|$ so that the effects of possibly heterogeneous scales of $\widehat{\mathbf{U}}_j$'s are appropriately adjusted. For the sparsity hyperparameter, we simply set $s_0 = 1$ in the simulation experiments. On a real dataset, one could make an informative choice of s_0 according to expertise knowledge in the specific area, or tune s_0 by sophisticated cross-validation or empirical Bayes procedures.

Combining the prior (8) with the pseudo data generating process $\mathbf{Y} = \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$ results in a pseudo posterior distribution

$$\widehat{\pi}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta} | \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) \propto \pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) \mathcal{N}(\mathbf{Y} | \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}, \sigma^2 \mathbf{I}), \tag{9}$$

where $\mathcal{N}(\mathbf{y} | \boldsymbol{\mu}, \sigma^2 \mathbf{I})$ is the n-dimensional Gaussian density function with mean $\boldsymbol{\mu}$ and covariance $\sigma^2 \mathbf{I}$. This pseudo posterior distribution (9) differs from the exact posterior distributions $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{F}, \mathbf{U}, \mathbf{Y})$, obtained by a Bayesian procedure with true variables $[\mathbf{F}, \mathbf{U}]$ as covariates, and $\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Y})$, obtained by a full-Bayesian procedure.

It is worth noting that, even in the simple setup where $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{1 \leq i \leq n}$ are identically and independently distributed (i.i.d.) and $\boldsymbol{f}_i \sim \mathcal{P}_f, \boldsymbol{u}_i \sim \mathcal{P}_u$ are jointly independent, the exact posterior distribution

$$\pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta} | \mathbf{X}, \mathbf{Y}) \propto \pi(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) \int \mathcal{N}(\mathbf{Y} | \mathbf{F}\boldsymbol{\alpha} + (\mathbf{X} - \mathbf{F}\mathbf{B}^{\mathrm{T}})\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \prod_{i=1}^{n} \mathcal{P}_f(\boldsymbol{f}_i) \mathcal{P}_u(\boldsymbol{x}_i - \mathbf{B}\boldsymbol{f}_i) d\boldsymbol{f}_i,$$

is computationally intractable due to the involvement of latent variables in the integral. Thus a full-Bayesian procedure does not estimate model (4) easily.

4 Theoretical Results on Factor Model

Section 4.1 establishes the estimation error rate $\sqrt{\log p/n}$ of the PCA-based method for latent common factors $\mathbf{F}$. Two conditions are needed. The first (Assumption 1) concerns convergence rates of the sample covariance matrices $\mathbf{F}^{\mathrm{T}}\mathbf{F}/n$, $\mathbf{F}^{\mathrm{T}}\mathbf{U}/n$, $\mathbf{U}^{\mathrm{T}}\mathbf{U}/n$ towards their ideal counterparts $\mathbf{I}$, $\mathbf{0}$ and $\boldsymbol{\Sigma}$. The second (Assumption 2) concerns the eigen (or singular) structures of factor loading coefficients $\mathbf{B}$ and the covariance matrix $\boldsymbol{\Sigma}$. Section 4.2 proceeds to estimate each idiosyncratic component $\mathbf{U}_j$ under an additional condition (Assumption 3) on the magnitudes of entries in $\mathbf{B}$ and $\boldsymbol{\Sigma}$. Section 4.3 highlights the technical novelty of these results.

4.1 Estimation of Common Factors

We summarize assumptions and results for the estimation error rate of $\mathbf{F}$ first, and commend on them later.

Assumption 1 (On Sample Covariance Matrices)**.** *There exists constant L_0 such that*

$$\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\|_{\max} \le L_0\sqrt{\log p/n},$$
$$\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\|_{\max} \le L_0\sqrt{\log p/n},$$
$$\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \mathbf{\Sigma}\|_{\max} \le L_0\sqrt{\log p/n}.$$

with high probability at least $1 - o_n$.

Assumption 2 (On Eigen Structures of $\mathbf{B}$ and $\mathbf{\Sigma}$)**.**

(i) Let $\lambda_1 \ge \lambda_2 \cdots \ge \lambda_k$ be the eigenvalues of $\mathbf{B}^{\mathrm{T}}\mathbf{B}/p$. For each $1 \le j \le k$, $\lambda_j \asymp 1$.

(ii) $\|\mathbf{\Sigma}\| \preccurlyeq p\sqrt{\log p/n}$.

(iii) $\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}) \prec p^2 \log p/n$.

Theorem 1. *Under Assumptions 1-2, the following statements hold.*

(a) Recall that $\widehat{\lambda}_j$, $j = 1, \ldots, n$ are eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$. There exists constant L_1 such that

$$\max_{1\le j\le k} |\widehat{\lambda}_j - \lambda_j| \le L_1\sqrt{\log p/n}, \qquad \max_{k+1\le j\le n} |\widehat{\lambda}_j - 0| \le L_1\sqrt{\log p/n}$$

with high probability at least $1 - o_n$.

(b) Let $\mathbf{\Pi}$ and $\widehat{\mathbf{\Pi}}$ be the projection matrices onto the column spaces of $\mathbf{F}$ and $\widehat{\mathbf{F}}$, respectively. There exists constant L_2 such that the sin-theta distance between two column spaces is bounded as

$$\|(\mathbf{I} - \mathbf{\Pi})\widehat{\mathbf{\Pi}}\|_{\mathrm{F}} = \|(\mathbf{I} - \widehat{\mathbf{\Pi}})\mathbf{\Pi}\|_{\mathrm{F}} = \|\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\|_{\mathrm{F}}/\sqrt{2} \le L_2\sqrt{\log p/n}$$

with high probability at least $1 - o_n - \frac{\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B})}{p^2 \log p/n}$.

(c) There exist constant L_3 and some orthogonal matrix $\mathbf{H}_{k\times k}$ such that

$$\|\widehat{\mathbf{F}}\mathbf{H}/\sqrt{n} - \mathbf{F}/\sqrt{n}\|_{\mathrm{F}} \le L_3\sqrt{\log p/n}$$

with high probability at least $1 - o_n - \frac{\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B})}{p^2 \log p/n}$.

Assumption 1 is a high-level condition not requiring samples $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i=1}^n$ to be identically or independently distributed. Kneip and Sarda (2011, Assumption A2 and Proposition 1) assumed the same error rate $\sqrt{\log p/n}$ for the factor-augmented sparse regression model (6) and provided a sufficient condition for i.i.d. samples $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i=1}^n$. Fan et al. (2013) derived this error rate for stationary and weakly-correlated time-series $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i\ge1}$. Our recent works on concentration inequalities for Markov chains (Jiang et al., 2018; Fan et al., 2019) can verify this error rate in case that $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i\ge1}$ are functions of ergodic Markov chains. Below are two concrete examples in which Assumption 1 holds. Their proofs are provided in the appendix.

Example 1. *Let* $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i=1}^n$ *be independent (not necessarily identically distributed) samples with* $\mathbb{E}\boldsymbol{f}_i = \mathbf{0}$, $\mathbb{E}\boldsymbol{u}_i = \mathbf{0}$, *and* $\mathbb{E}[\boldsymbol{f}_i\boldsymbol{u}_i^{\mathrm{T}}] = \mathbf{0}$. *If* f_{ij}*'s and* u_{ij}*'s have subgaussian norms at most* c, *then Assumption 1 holds. Note that a mean-zero variable bounded by* $c\log(2)$ *or a Gaussian variable with mean zero and variance less than* $c^2/2$ *has a subgaussian norm at most* c.

Example 2. *Let* $\{(\boldsymbol{f}_i, \boldsymbol{u}_i)\}_{i\geq 1}$ *be functions of a stationary, general-state-space Markov chain* $\{\boldsymbol{Z}_i\}_{i\geq 1}$, *i.e.,* $f_{ij} = f_{ij}(\boldsymbol{Z}_i)$ *and* $u_{ij} = u_{ij}(\boldsymbol{Z}_i)$, *with* $\mathbb{E}\boldsymbol{f}_i(\boldsymbol{Z}_i) = \mathbf{0}$, $\mathbb{E}\boldsymbol{u}_i(\boldsymbol{Z}_i) = \mathbf{0}$, *and* $\mathbb{E}[\boldsymbol{f}_i(\boldsymbol{Z}_i)\boldsymbol{u}_i(\boldsymbol{Z}_i)^{\mathrm{T}}] = \mathbf{0}$. *If the Markov chain admits a non-zero* $\mathcal{L}_2$*-spectral gap, and there exist envelop functions* $\bar{f}(\boldsymbol{z}), \bar{u}(\boldsymbol{z})$ *such that* $\max_{i,j}|f_{ij}(\boldsymbol{z})| \leq \bar{f}(\boldsymbol{z})$, $\max_{i,j}|u_{ij}(\boldsymbol{z})| \leq \bar{u}(\boldsymbol{z})$ *for any* $\boldsymbol{z}$ *in the state space of the Markov chain and* $\mathbb{E}[\bar{f}^4(\boldsymbol{Z}_1)] \leq c^4, \mathbb{E}[\bar{u}^4(\boldsymbol{Z}_1)] \leq c^4$, *then Assumption 1 holds.*

Assumption 2(i) concerns the eigen spectrum of $\mathbf{B}\mathbf{B}^{\mathrm{T}}/p$. The positive definiteness of $\mathbf{B}^{\mathrm{T}}\mathbf{B}/p$ indicates that each factor $\mathbf{F}_j$ makes a non-trivial contribution to the variations of covariates $\mathbf{X}$. This condition is commonly seen in the literature of factor models. Assumption 2(ii) allows the operator norm (largest eigenvalue) of $\boldsymbol{\Sigma}$ to grow with increasing n, p. Assumption 2(iii) amounts to

$$\mathrm{vec}(\boldsymbol{\Sigma})^{\mathrm{T}}\mathrm{vec}(\mathbf{B}\mathbf{B}^{\mathrm{T}}) \prec p^2\log p/n,$$

where $\mathrm{vec}(\mathbf{A})$ denotes the vector formed by concatenating column vectors of matrix $\mathbf{A}$. As $\boldsymbol{\Sigma}$ contain p^2 entries, this condition actually encourages the sparseness of $\boldsymbol{\Sigma}$ and weak correlations among idiosyncratic components $\mathbf{U}_j$'s by anchoring most entries of $\boldsymbol{\Sigma}$ around zero.

Assumption 2 ensures the pervasiveness of latent factors by characterizing a "low-rank spike plus sparse" eigen structure of the covariance matrix of covariates

$$\mathbb{E}[\mathbf{X}^{\mathrm{T}}\mathbf{X}/n] = \mathbf{B}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Sigma}.$$

All non-zero eigenvalue of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$ are of order $\Omega(p)$ due to Assumption 2(i), while all eigenvalues of $\boldsymbol{\Sigma}$ is of order $o(p)$ due to Assumption 2(ii). This large gap between eigenvalues is crucial for estimating the column space of $\mathbf{F}$ through PCA (Wang and Fan, 2017; Fan et al., 2020b). In contrast, if this gap is relatively small compared to the eigenvalues of $\boldsymbol{\Sigma}$, PCA may result in inconsistent estimation (Johnstone and Lu, 2009). Conditions on $\mathbf{B}, \boldsymbol{\Sigma}$ used in previous works (Bai and Ng, 2002; Fan et al., 2020a) are special cases of Assumption 2.

Example 3. *In addition to Assumption 2(i), Bai and Ng (2002) assumed that* $\max_j\|\boldsymbol{b}_j\| \leq c_1$, $\max_j \boldsymbol{\Sigma}_{jj} \leq c_2$, *and* $\sum_{i,j}|\boldsymbol{\Sigma}_{ij}| \leq c_3 p$. *Their conditions imply that* $\|\boldsymbol{\Sigma}\| \leq \sqrt{c_2 c_3 p}$ *and* $\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}) \leq c_1^2 c_3 k p$.

Example 4. *In addition to Assumption 2(i), Fan et al. (2020b) assumed that* $\|\boldsymbol{\Sigma}\| \leq c_4$. *Their conditions imply* $\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}) \leq \mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\|\boldsymbol{\Sigma}\| \leq c_4(\lambda_1 + \cdots + \lambda_k)p$.

4.2 Estimation of Idiosyncratic Components

Estimating each idiosyncratic component $\mathbf{U}_j$ is more challenging than estimating common factors. We need an additional assumption to control the magnitudes of entries of $\mathbf{B}$ and $\boldsymbol{\Sigma}$.

Assumption 3 (On Magnitudes of Entries of $\mathbf{B}$ and $\boldsymbol{\Sigma}$)**.** $\max_{j=1}^p\|\boldsymbol{b}_j\| \preccurlyeq 1$, *where* $\boldsymbol{b}_j$ *is the* j*-th row of* $\mathbf{B}$, *and* $\max_{j=1}^p \boldsymbol{\Sigma}_{jj} \preccurlyeq 1$, *where* $\boldsymbol{\Sigma}_{jj}$ *is the* j*-th diagonal entry of* $\boldsymbol{\Sigma}$.

Corollary 1. *Suppose Assumptions 1 to 3 hold. There exists constant L_4 such that*

$$\max_{j=1}^{p} \|\widehat{\mathbf{U}}_j/\sqrt{n} - \mathbf{U}_j/\sqrt{n}\| \le L_4\sqrt{\log p/n}$$

with high probability at least $1 - o_n - \frac{\operatorname{trace}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B})}{p^2 \log p/n}$.

To motivate Assumption 3, let us have a close look at the estimation error

$$\widehat{\mathbf{U}}_j - \mathbf{U}_j = (\boldsymbol{\Pi} - \widehat{\boldsymbol{\Pi}})\mathbf{X}_j - \boldsymbol{\Pi}\mathbf{U}_j,$$

where $\boldsymbol{\Pi}$ and $\widehat{\boldsymbol{\Pi}}$ are projection matrices onto the column spaces of $\mathbf{F}$ and $\widehat{\mathbf{F}}$ introduced by Theorem 1(b). It follows that

$$\|\widehat{\mathbf{U}}_j/\sqrt{n} - \mathbf{U}_j/\sqrt{n}\| \le L_2\sqrt{2\log p/n}\|\mathbf{X}_j/\sqrt{n}\| + \|\boldsymbol{\Pi}\mathbf{U}_j/\sqrt{n}\|.$$

Assumption 3 is intended to bound the term $\|\mathbf{X}_j/\sqrt{n}\|^2 \approx \|\boldsymbol{b}_j\|^2 + \boldsymbol{\Sigma}_{jj}$. The term $\|\boldsymbol{\Pi}\mathbf{U}_j/\sqrt{n}\|$, the projection of $\mathbf{U}_j/\sqrt{n}$ onto the column space of $\mathbf{F}$, is small due to the weak correlation between $\mathbf{F}$ and $\mathbf{U}$. Bai and Ng (2002) used Assumption 3 to estimate common factors $\mathbf{F}$ (see Example 3). Here we only need it for estimating idiosyncratic components $\mathbf{U}$. Without this assumption, Theorem 1 for estimating common factors $\mathbf{F}$ still stands.

4.3 Technical Novelty

Theorem 1(b) measures the estimation error of the column space of $\mathbf{F}$ by the sin-theta distance, a metric in the matrix perturbation theory (Stewart, 1990) to quantify the difference between two linear spaces.

Definition 3 (Principal Angles and Sin-Theta Distance)**.** *Let $\widehat{\boldsymbol{\Psi}}_{n\times k}$ and $\boldsymbol{\Psi}_{n\times k}$ be orthonormal bases of two linear subspaces $\widehat{\mathcal{L}}$ and $\mathcal{L}$ of rank k in $\mathbb{R}^n$. The principal or canonical angle between two linear subspaces $\widehat{\mathcal{L}}$ and $\mathcal{L}$ is defined as*

$$\angle(\widehat{\mathcal{L}}, \mathcal{L}) = (\cos^{-1} s_1, \ldots, \cos^{-1} s_k)^{\mathrm{T}},$$

where $s_1, \ldots, s_k \in [0, 1]$ are the singular values of $\widehat{\boldsymbol{\Psi}}^{\mathrm{T}}\boldsymbol{\Psi}$ or $\boldsymbol{\Psi}^{\mathrm{T}}\widehat{\boldsymbol{\Psi}}$. The sin-theta distance between two linear subspaces $\widehat{\mathcal{L}}$ and $\mathcal{L}$ is defined as

$$\|\sin\angle(\widehat{\mathcal{L}}, \mathcal{L})\| = \sqrt{\sum_{j=1}^{k} \sin^2(\cos^{-1} s_j)}.$$

Equivalently, with $\widehat{\boldsymbol{\Pi}} = \widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Psi}}^{\mathrm{T}}$ and $\boldsymbol{\Pi} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}$ being the projection matrices onto the two linear subspaces,

$$\|\sin\angle(\widehat{\mathcal{L}}, \mathcal{L})\|^2 = \|(\mathbf{I} - \boldsymbol{\Pi})\widehat{\boldsymbol{\Pi}}\|_{\mathrm{F}}^2 = \|(\mathbf{I} - \widehat{\boldsymbol{\Pi}})\boldsymbol{\Pi}\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{\Pi}} - \boldsymbol{\Pi}\|_{\mathrm{F}}^2/2.$$

To devise the proof of Theorem 1(b), we develop a novel extension of the Davis-Kahan theorem (Davis and Kahan, 1970; Yu et al., 2014). The eigendecomposition of $\mathbf{B}^{\mathrm{T}}\mathbf{B}/p = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{\mathrm{T}}$ deduces that

$$\frac{\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}}}{np} \times \frac{\mathbf{F}\mathbf{R}}{\sqrt{n}} = \frac{\mathbf{F}\mathbf{R}}{\sqrt{n}} \times \boldsymbol{\Lambda} + \boldsymbol{\Delta}, \quad \text{where } \boldsymbol{\Delta} = \frac{\mathbf{F}}{\sqrt{n}} \times \frac{\mathbf{B}^{\mathrm{T}}\mathbf{B}}{p} \times \left(\frac{\mathbf{F}^{\mathrm{T}}\mathbf{F}}{n} - \mathbf{I}\right)\mathbf{R}.$$

We call this equation a "$\mathbf{\Delta}$-approximate" eigenequation of $\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}}/np$ and compare it with the exact eigenequation

$$\frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{np} \times \frac{\widehat{\mathbf{F}}}{\sqrt{n}} = \frac{\widehat{\mathbf{F}}}{\sqrt{n}} \times \widehat{\mathbf{\Lambda}}.$$

A novel variant of Davis-Kahan theorem (Lemma A3) relates the difference between $\widehat{\mathbf{F}} \approx \mathbf{F}\mathbf{R}$ to differences between $\mathbf{X} \approx \mathbf{F}\mathbf{B}^{\mathrm{T}}$, $\widehat{\mathbf{\Lambda}} \approx \mathbf{\Lambda}$, $\mathbf{0} \approx \mathbf{\Delta}$. This variant of Davis-Kahan theorem gives a clear insight on roles of eigen structures of $\mathbf{B}, \mathbf{\Sigma}$ and concentration properties of sample covariance matrices in the estimation of factor models. It may be potential applicable to the analyses of PCA-based methods for other problems.

Theorem 1(c) follows from Theorem 1(b). Both sides of the equation are divided by a factor $\sqrt{n}$ such that $\widehat{\mathbf{F}}/\sqrt{n}$ and $\mathbf{F}/\sqrt{n}$ are (nearly) orthogonal. This result can be viewed as the non-asymptotic version of Bai and Ng (2002, Theorem 1). The former gives a non-asymptotic error bound $\sqrt{\log p/n}$ with a precise characterization of the tail probability, while the latter gives an asymptotic error bound $O_p(\sqrt{1/n})$. The additional factor $\log p$ arises from the essential difference between non-asymptotic analyses and asymptotic analyses. The starting point of the proof of Theorem 1 is to deduce from Assumption 1 non-asymptotic error bounds

$$\begin{aligned}
\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\| &\le k\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\|_{\max} \le L_0 k\sqrt{\log p/n}, \\
\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\| &\le \sqrt{kp}\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\|_{\max} \le L_0\sqrt{kp\log p/n}, \\
\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \mathbf{\Sigma}\| &\le p\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \mathbf{\Sigma}\|_{\max} \le L_0 p\sqrt{\log p/n},
\end{aligned}$$

The assumptions of Bai and Ng (2002) can deduce asymptotic error bounds

$$\begin{aligned}
\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\| &= O_p(\sqrt{1/n}), \\
\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\| &= O_p(\sqrt{p/n}), \\
\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \mathbf{\Sigma}\| &= O_p(p\sqrt{1/n}).
\end{aligned}$$

Using the asymptotic error bounds instead of the non-asymptotic error bounds in the technical proof of Theorem 1 would reproduce the asymptotic result of Bai and Ng (2002, Theorem 1).

5 Theoretical Results on Bayesian Sparse Regression

This section summarizes the theoretical properties of the pseudo-posterior distribution given by (9). The ℓ_2-error rate $\epsilon_n = \sqrt{s\log p/n}$ is achieved for regression coefficients $\boldsymbol{\alpha}^\star$, $\boldsymbol{\beta}^\star$ under commonly-seen assumptions for Bayesian sparse regression. This rate is so far the best achievable rate by Bayesian methods even with true variables $[\mathbf{F}, \mathbf{U}]$ (Castillo et al., 2015; Song and Liang, 2017). Byproducts of the analysis include the adaptivity to the unknown sparsity s and unknown standard deviation $\sigma^\star$. When the beta-min condition holds, the pseudo-posterior distribution consistently selects the true sparse model $\xi^\star = \{j : \beta_j^\star \neq 0\}$. Interestingly, although the factor adjustment does not change the order $\epsilon_n = \sqrt{s\log p/n}$ of the estimation error, it does require a stronger sparse eigenvalue condition and result in larger constant factors of the estimation error.

Section 5.1 makes three assumption. The first is the sparse eigenvalue condition on $\mathbf{U}$, the second is a high-level condition controlling the estimation error of $\widehat{\mathbf{F}} \approx \mathbf{F}$, $\widehat{\mathbf{U}} \approx \mathbf{U}$ in the factor model, and the last is on the magnitude of the true regression coefficients. Sections 5.2 and 5.3 present the main theorem and its sketch of proof.

5.1 Assumptions

Following the literature of sparse regression, we assume $p \succ n$ but $s \log p \prec n$ such that the desired estimation error rate $\epsilon_n = \sqrt{s \log p / n} \to 0$ as $n \to \infty$. Other assumptions are stated as follows.

Assumption 4 (On Sparse Eigenvalue)**.** *There exist constants $M_0 > 0$, κ_0, κ_1 such that*

$$\min_{\xi:\ |\xi| \le (1+M_0)s} \lambda_{\min}(\mathbf{U}_\xi^{\mathrm{T}} \mathbf{U}_\xi / n) \ge \kappa_0^2,$$

and that $\max_{j=1}^p \|\mathbf{U}_j / \sqrt{n}\| \le \kappa_1$ with high probability. Therefore, the $(1+M_0)s$-order sparse eigenvalue (Definition 1) of $\mathbf{U}$ is at least κ_0^2 / κ_1^2.

Variants of this sparse eigenvalue condition have been imposed on original covariates $\mathbf{X}$ by Bayesian sparse regression methods to ensure both estimation consistency (Castillo et al., 2015; Song and Liang, 2017) and computational efficiency (Yang et al., 2016). Here this condition is imposed on decorrelated covariates $\mathbf{U}$. As discussed in Section 2, on decorrelated covariates $\mathbf{U}$ rather than original covariates $\mathbf{X}$ this condition holds more likely. When $\mathbf{U}$ consists of independent subgaussian rows, random matrix theories (Vershynin, 2012, Theorem 5.39) can verify Assumption 4.

Assumption 5 (On Estimation of Factor Model)**.** *There exist constants L_3, L_4 such that*

$$\left\| \frac{\widehat{\mathbf{F}}\mathbf{H}}{\sqrt{n}} - \frac{\mathbf{F}}{\sqrt{n}} \right\|_{\mathrm{F}} \le L_3 \sqrt{\frac{\log p}{n}}, \qquad \max_{1 \le j \le p} \left\| \frac{\widehat{\mathbf{U}}_j}{\sqrt{n}} - \frac{\mathbf{U}_j}{\sqrt{n}} \right\| \le L_4 \sqrt{\frac{\log p}{n}},$$

with high probability, where $\mathbf{H}$ is some $k \times k$ rotation (orthogonal) matrix.

The estimation error rate $\sqrt{\log p / n}$ of latent variables has been verified by Theorem 1 and Corollary 1. Note that $\widehat{\mathbf{F}} / \sqrt{n}$ is an orthonormal basis whose span approximates the column space of $\mathbf{F}$, and $\widehat{\mathbf{F}}\mathbf{H} / \sqrt{n}$ spans the same linear space.

Assumption 6 (On True Parameters)**.** *$\sigma^\star > 0$ is fixed, $\|\boldsymbol{\alpha}^\star\| \preccurlyeq 1$, $\|\boldsymbol{\beta}^\star\| \preccurlyeq 1$.*

The assumed constant orders of $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta}^\star$ are not restrictive. When Assumption 5 is in place, both vectors of regression coefficients have bounded ℓ_2-norms if the response variable has a bounded variance. To see this point, write

$$\mathbb{E}[\|\mathbf{Y}\|^2 / n] = \|\boldsymbol{\alpha}^\star\|^2 + (\boldsymbol{\beta}^\star_{\xi^\star})^{\mathrm{T}} \mathbb{E}[\mathbf{U}_{\xi^\star}^{\mathrm{T}} \mathbf{U}_{\xi^\star} / n] \boldsymbol{\beta}^\star_{\xi^\star} + \sigma^{\star 2} \ge \|\boldsymbol{\alpha}^\star\|^2 + (1 - o(1)) \kappa_0^2 \|\boldsymbol{\beta}^\star\|^2 + \sigma^{\star 2}.$$

Assumptions 5 and 6 together control the difference between the true data generating process $\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star + \sigma^\star \boldsymbol{\varepsilon}$ and the pseudo data generating process $\mathbf{Y} = \widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star + \sigma^\star \boldsymbol{\varepsilon}$ in terms of the deviation between their conditional means

$$\|(\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| \le L_5 \sigma^\star \sqrt{n} \epsilon_n, \quad \text{with } L_5 = L_3 \|\boldsymbol{\alpha}^\star / \sigma^\star\| + L_4 \|\boldsymbol{\beta}^\star / \sigma^\star\|.$$

We remark that, when more accurate estimation methods of latent variables than the PCA-based method are available, larger magnitudes of regression coefficients are allowed.

5.2 Main Results

Before presenting main results, we formally define the estimation error rate in the Bayesian setup, which is different to that in the frequentist setup. The following definition of the posterior contraction rate is adopted from the Bayesian literature (Ghosal et al., 2000; Shen and Wasserman, 2001; Castillo et al., 2015; Song and Liang, 2017).

Definition 4 (Posterior Contraction Rate)**.** *Consider a parametric model* $\{\mathcal{P}_{\boldsymbol{\theta}} : \theta \in \Theta\}$*. Let* $\mathcal{D}_n$ *for* $n \geq 1$ *be a sequence of data generations from* $\mathcal{P}_{\boldsymbol{\theta}^\star}$*. Let* $\boldsymbol{\gamma}(\boldsymbol{\theta})$ *be a function of* $\boldsymbol{\theta}$*, and* $\ell(\boldsymbol{\gamma}(\boldsymbol{\theta}), \boldsymbol{\gamma}^\star)$ *be a loss function between the estimate* $\boldsymbol{\gamma}(\boldsymbol{\theta})$ *and the estimand* $\boldsymbol{\gamma}^\star$*. A sequence of posterior distributions (random measures)* $\pi(\boldsymbol{\theta}|\mathcal{D}_n)$ *for* $n \geq 1$ *is said to achieve the contraction rate* ϵ_n *of estimation error* $\ell(\boldsymbol{\gamma}(\boldsymbol{\theta}), \boldsymbol{\gamma}^\star)$ *if*

$$\pi(\ell(\boldsymbol{\gamma}(\boldsymbol{\theta}), \boldsymbol{\gamma}^\star) \leq M\epsilon_n | \mathcal{D}_n) \to 1$$

in $\mathbb{P}_{\boldsymbol{\theta}^\star}$*-probability as* $n \to \infty$ *for some constant* $M > 0$*.*

In this paper, we consider

$$\mathcal{D}_n = (\mathbf{X}, \mathbf{Y}), \quad \boldsymbol{\theta} = (\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}), \quad \boldsymbol{\gamma}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \quad \boldsymbol{\gamma}^\star = \begin{pmatrix} \mathbf{H}\boldsymbol{\alpha}^\star \\ \boldsymbol{\beta}^\star \end{pmatrix},$$

where $\mathbf{H}$ is the rotation matrix introduced in the estimation of the factor model. The objective is to show that the pseudo-posterior distribution $\widehat{\pi}(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})$ given by (9) achieves the contraction rate $\epsilon_n = \sqrt{s \log p / n}$ in terms of ℓ_2 error

$$\ell(\boldsymbol{\gamma}(\boldsymbol{\theta}), \boldsymbol{\gamma}^\star) = \|\boldsymbol{\gamma}(\boldsymbol{\theta}) - \boldsymbol{\gamma}^\star\| = \left\| \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} - \begin{pmatrix} \mathbf{H}\boldsymbol{\alpha}^\star \\ \boldsymbol{\beta}^\star \end{pmatrix} \right\|.$$

Note that $\widehat{\mathbf{F}}$ and $\mathbf{F}$ span almost the same linear space, and $\widehat{\mathbf{F}}\mathbf{H} \approx \mathbf{F}$ element-wisely. Thus the pseudo-posterior distribution is expected to concentrate around $\boldsymbol{\alpha} \approx \mathbf{H}\boldsymbol{\alpha}^\star$ such that $\widehat{\mathbf{F}}\boldsymbol{\alpha} \approx \widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star \approx \mathbf{F}\boldsymbol{\alpha}^\star$. Now we are ready to present the main results of the paper.

Theorem 2. *Define an "*ϵ_n*-neighborhood" of parameter* $(\sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}')$ *as*

$$A(\sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}', M_0, M_1, M_2, \epsilon_n) = \{(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \text{ Eq. (10)}\}$$

$$\begin{cases} |\xi \setminus \xi'| \leq M_0 s, \\ \dfrac{\sigma^2}{\sigma'^2} \in \left[\dfrac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n}, \dfrac{1 + M_1\epsilon_n}{1 - M_1\epsilon_n} \right], \\ \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\| \leq M_2\sigma'\epsilon_n, \\ \|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \leq M_3\sigma'\epsilon_n/\kappa_0, \end{cases} \tag{10}$$

where M_0*,* M_1*,* M_2*,* M_3 *are absolute constants,* ξ *and* ξ' *are supports of* $\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}'$*, respectively. Suppose Assumptions 4 to 6 holds with* $M_0 - 2 > L_5^2$*, where* $L_5 = L_3\|\boldsymbol{\alpha}^\star/\sigma^\star\| + L_4\|\boldsymbol{\beta}^\star/\sigma^\star\|$*. The following statements hold with some constants* M_1*,* M_2*,* M_3*,* M_4*.*

(a) (Estimation Error) For any constant $C_1 < M_0 - 2 - L_5^2$*,*

$$\widehat{\pi}\left(A(\sigma^\star, \mathbf{H}\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n)|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \geq 1 - e^{-C_1 s \log p}$$

with high probability.

(b) *(Prediction Error) For any* $C_2 < M_0 - 2 - L_5^2$,

$$\widehat{\pi}\left(\|(\widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| \le M_4\sigma^\star\sqrt{n}\epsilon_n | \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \ge 1 - e^{-C_2 s\log p}$$

with high probability.

(c) *(Model Selection) Suppose the beta-min condition* $\min_{j\in\xi^\star}|\beta_j^\star| \succ \epsilon_n$ *holds in addition. For any* $C_3 < M_0 - 2 - L_5^2$,

$$\widehat{\pi}\left(A(\sigma^\star, \mathbf{H}\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n) \cap \{\xi \supseteq \xi^\star\} | \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \ge 1 - e^{-C_3 s\log p}$$

with high probability, implying that

$$\widehat{\pi}\left(|\xi \setminus \xi^\star| \le M_0 s,\ \xi \supseteq \xi^\star | \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \ge 1 - e^{-C_3 s\log p}$$
$$\widehat{\pi}\left(\left\{j : |\beta_j| \ge 2M_3\sigma\sqrt{|\xi|\log p/n}\right\} = \xi^\star \middle| \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \ge 1 - e^{-C_3 s\log p}.$$

Part (a) establishes the posterior contraction rate ϵ_n in terms of ℓ_2 error for regression coefficients $\boldsymbol{\alpha}^\star$ (up to some rotation matrix $\mathbf{H}$) and $\boldsymbol{\beta}^\star$. It also asserts that the posterior model ξ overshoots the true sparse model $\xi^\star$ by no more than a constant factor M_0, and that the relative estimation error of the standard deviation $\sigma_\star$ is $M_1\epsilon_n$. Part (b) shows that $\widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}$ predicts the true conditional mean $\mathbb{E}[\mathbf{Y}|\mathbf{F}, \mathbf{U}] = \mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star$ with mean squared error $M_4\epsilon_n$ for each single datum instance on average.

The beta-min condition in part (c) has been used by Bayesian sparse regression methods to achieve the model selection consistency (Castillo et al., 2015; Song and Liang, 2017). Without this condition, the Bayesian methods cannot tell whether a nearly-zero regression coefficient $\beta_j \preccurlyeq \epsilon_n$ is truly non-zero or faked by the randomness of data generations. The first implication of part (c) asserts that the pseudo-posterior distribution selects all variables in $\xi^\star$ and at most $M_0 s$ false positives. In simulation experiments, the pseudo posterior distribution overestimates the true model size $s = |\xi^\star|$ by less than 5%. The second implication of part (c) enables a posterior model selection rule. Simply speaking, one can consistently select the true model $\xi^\star$ by filtering out coefficients β_j's larger than threshold $2M_3\sigma\sqrt{|\xi|\log p/n}$. In simulation experiments, the majority of pseudo-posterior samples of ξ are exactly the true model $\xi^\star$ even without the posterior model selection rule.

Recall that the constant L_5 relates to estimation errors of latent variables in the factor model. The constant M_0 indicates the strength of the sparse eigenvalue condition (Assumption 4), and determines the quality of the posterior distribution (if true variables $[\mathbf{F}, \mathbf{U}]$ are used). The constraint $M_0 - 2 > L_5^2$ in Theorem 2 arises when the Bayesian sparse regression method copes with the estimated latent variables. A less accurate estimation of latent variables in the factor model would result in large L_5. Consequently, the factor-adjusted Bayesian method would need a stronger sparse eigenvalue condition with larger M_0.

5.3 Sketch of Proof

Let $\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$ and $\widehat{\mathbb{P}}_{(\sigma^\star, \mathbf{H}\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$ be the probability measures associated with the true data generating process $\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star + \sigma^\star\boldsymbol{\varepsilon}$ and the pseudo data generating process $\mathbf{Y} = \widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star + \sigma^\star\boldsymbol{\varepsilon}$, respectively. Fig. 1 illustrates the paradigm of proof of Theorem 2. The analysis is first moved from the probability space of the true data generating process into that of the pseudo data generating process by conditioning on a realization of $\mathbf{F}$, $\mathbf{U}$, $\widehat{\mathbf{F}}$, $\widehat{\mathbf{U}}$, $\mathbf{H}$. In the space of the pseudo data generating

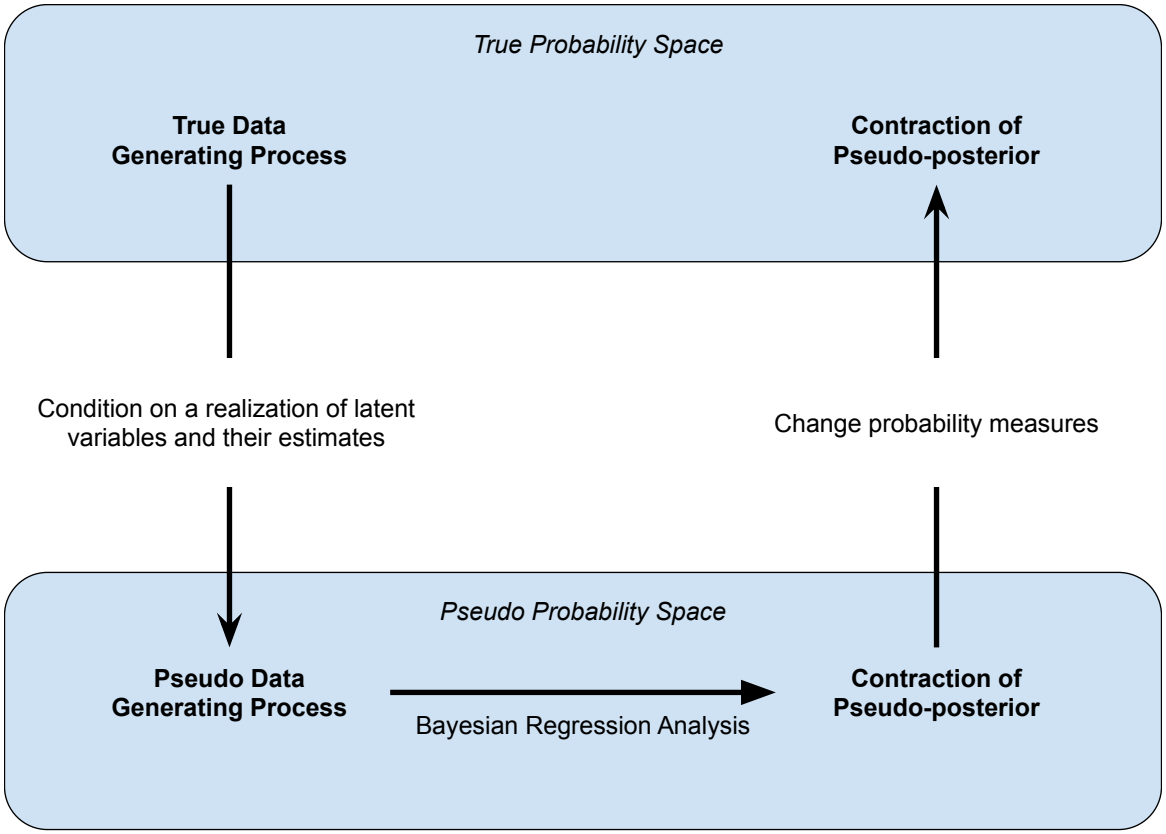


Figure 1: paradigm of Proof of Theorem 2

process, theoretical properties of the pseudo-posterior distribution $\widehat{\pi}(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})$ are established. These theoretical properties are then translated back to the probability space of $\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$.

More precisely, the objective is to show events

$$\begin{aligned}
\mathcal{E}_1: &\ \widehat{\pi}\left(A^c(\sigma^\star, \mathbf{H}\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n)|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \geq e^{-C_1 s \log p}, \\
\mathcal{E}_2: &\ \widehat{\pi}\left(\|(\widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| > M_4\sigma^\star\sqrt{n}\epsilon_n|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \geq e^{-C_2 s \log p}, \\
\mathcal{E}_3: &\ \widehat{\pi}\left(A^c(\sigma^\star, \mathbf{H}\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n) \cup \{\xi \not\supseteq \xi^\star\}|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \geq e^{-C_3 s \log p}
\end{aligned}$$

happen with vanishing probability under $\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$. The first step (Appendix B.1) is to show that

$$\mathcal{F}: \begin{cases} \min_{\xi:\ |\xi| \leq (M_0+1)s} \lambda_{\min}(\widehat{\mathbf{U}}_\xi^{\mathrm{T}}\widehat{\mathbf{U}}_\xi/n) \geq \kappa_0^2/4 \\ \max_{j=1}^p \|\widehat{\mathbf{U}}_j/\sqrt{n}\| \leq 2\kappa_1 \\ \|(\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| \leq L_5\sigma^\star\sqrt{n}\epsilon_n \end{cases} \tag{11}$$

happens with high $\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$-probability. Set $M_4 \geq 2L_5$ and define another event

$$\mathcal{E}_2': \ \widehat{\pi}\left(\|(\widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}) - (\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star)\| > M_4\sigma^\star\sqrt{n}\epsilon_n/2|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}\right) \geq e^{-C_2 s \log p}.$$

Evidently, $\mathcal{E}_2 \cap \mathcal{F} \subseteq \mathcal{E}_2' \cap \mathcal{F}$, Write

$$\begin{aligned}
&\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_1) \leq \mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_1|\mathcal{F})\mathbb{P}(\mathcal{F}) + \mathbb{P}(\mathcal{F}^c), \\
&\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_2) \leq \mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_2|\mathcal{F})\mathbb{P}(\mathcal{F}) + \mathbb{P}(\mathcal{F}^c) \leq \mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_2'|\mathcal{F})\mathbb{P}(\mathcal{F}) + \mathbb{P}(\mathcal{F}^c), \\
&\mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_3) \leq \mathbb{P}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}(\mathcal{E}_3|\mathcal{F})\mathbb{P}(\mathcal{F}) + \mathbb{P}(\mathcal{F}^c).
\end{aligned}$$

Since the event $\mathcal{F}$ happens with high $\mathbb{P}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}$-probability, it suffices to bound conditional probabilities of events $\mathcal{E}_1$, $\mathcal{E}_2'$ and $\mathcal{E}_3$ given event $\mathcal{F}$. The second step (Appendix B.2) is to show that

$$\begin{aligned}
\mathbb{P}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_1|\mathbf{F},\mathbf{U},\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le \left[\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_1|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H})\right]^{1/2}\times e^{L_5^2 s\log p/2},\\
\mathbb{P}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_2'|\mathbf{F},\mathbf{U},\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le \left[\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_2'|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H})\right]^{1/2}\times e^{L_5^2 s\log p/2},\\
\mathbb{P}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_3|\mathbf{F},\mathbf{U},\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le \left[\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_3|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H})\right]^{1/2}\times e^{L_5^2 s\log p/2},
\end{aligned} \tag{12}$$

for any realization of $(\mathbf{F},\mathbf{U},\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H})$ belonging to event $\mathcal{F}$. In (12), the term $e^{L_5^2 s\log p/2}$ relates to the estimation of latent variables in the factor model, and the terms $\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_i|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H})$ relates to the quality of Bayesian sparse regression. Given $\min_{j\in\xi^\star}|\beta_j| \ge \sqrt{32M_0}\sigma_\star\epsilon_n/\kappa_0$, the third step (Appendix B.3) is to show that

$$\begin{aligned}
\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_1|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le e^{-C_1' s\log p}\\
\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_2'|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le e^{-C_2' s\log p}\\
\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_3|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) &\le e^{-C_3' s\log p}
\end{aligned} \tag{13}$$

for any constants $C_1' < M_0 - 2 - C_1$, $C_2' < M_0 - 2 - C_2$, $C_3' < M_0 - 2 - C_3$ if M_1, M_2, M_3, M_4 are sufficiently large. The theorem is concluded by choosing suitable $C_i' \in (L_5^2, M_0 - 2 - C_i)$ for $i = 1, 2, 3$.

As mentioned above, the proof critically depends on the non-asymptotic error bounds characterizing the contraction rate of the pseudo-posterior distribution. Classical works in Bayesian sparse regression (Narisetty and He, 2014; Castillo et al., 2015) are inadequately quantitative for the analysis in this paper. Our technique is inspired by a recent non-asymptotic analysis of Bayesian shrinkage methods (Song and Liang, 2017). However, given their results on Bayesian shrinkage methods, the analysis of Bayesian spike-and-slab methods in this paper is still challenging.

6 Simulation Experiments

This section harvests experimental results on simulated data. The default setting of experiments is as follows. For the data generating process, $(n,p,s,k) = (200,500,5,3)$, $\boldsymbol{f}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0},\mathbf{I})$, $\boldsymbol{u}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0},\mathbf{I})$, $\mathbf{B} \sim \text{Uniform}[-3.0,+3.0]^{p\times k}$. For the true parameters, $\sigma^{\star 2} = 0.5$, $\xi^\star = \{1,2,3,4,5\}$, $\boldsymbol{\beta}^\star_{\xi^\star} = (3.0,3.0,3.0,3.0,3.0)^{\mathrm{T}}$, and $\boldsymbol{\alpha}^\star = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}^\star$.

For prior (8), we choose the inverse-gamma density g with shape 1 and scale 1, the Gaussian densities $h_1(z) = \mathcal{N}(z|0,10^2)$, $h_2(z) = \mathcal{N}(z|0,1)$ and hyperparameters $s_0 = 1$ and $\tau_j^{-1} = \|\widehat{\mathbf{U}}_j\|/\sqrt{n}$. Starting from $(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta}) = (1.0,\mathbf{0},\mathbf{0})$, we iterate a Gibbs sampler $T = 20$ times and drop the first $T/2 = 10$ iterations as the burn-in period. The implementation details of the Gibbs sampler are put in the appendix.

The pseudo-posterior distribution are evaluated in terms of five metrics. The posterior mean of $\boldsymbol{\beta}$ is compared to $\boldsymbol{\beta}^\star$ in terms of ℓ_2 error. The model selection rate, the portion of the posterior samples that select the true model (i.e., $\xi = \xi^\star$) and the sure screening rate, the portion of the posterior samples that select all sparse coefficients (i.e., $\xi \supseteq \xi^\star$) are computed. To evaluate the adaptivity to unknown sparsity s, the average model size $|\xi|$ is computed. To evaluate the adaptivity to unknown standard deviation $\sigma^\star$, the posterior mean of σ^2 is compared to $\sigma^{\star 2}$ in terms of relative error. These metrics are evaluated and averaged over 100 replicates of the datasets.

The factor-adjusted Bayesian method is compared to the routine Bayesian method, the routine Lasso method (Friedman et al., 2010, R package `glmnet`) and the factor-adjusted Lasso method (Fan et al., 2020a, R package `FarmSelect`). The ℓ_1-penalty hyperparameters of the Lasso methods are optimized by ten-fold cross-validation. Note that the Bayesian/Lasso methods with covariates $\mathbf{X}$ can be seen as the factor-adjusted Bayesian/Lasso methods with the underestimated number of common factors $\widehat{k} = 0$.

6.1 Insensitivity to Overestimates of k

Table 1 summarizes the performances of four methods in the default setting. The factor-adjusted Bayesian method outperforms other three methods on both parameter estimation and model selection tasks. Its performance is insensitive to overestimated numbers of common factors $\widehat{k} = 6, 9, 12$. The factor-adjusted Lasso method tends to select two or three more covariates other than covariates of the true model. This issue is alleviated when larger nonzero coefficients are set.

Method	$\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|$	$\xi = \xi^\star$	$\xi \supseteq \xi^\star$	$\|\xi\|$	$\|\sigma^2/\sigma^{\star 2} - 1\|$
Lasso, $\widehat{k} = 0$	0.914	0%	100%	18.37	1.697
Factor-adjusted Lasso, $\widehat{k} = 3$	0.409	31%	100%	6.90	0.311
Factor-adjusted Lasso, $\widehat{k} = 6$	0.409	23%	100%	7.14	0.304
Factor-adjusted Lasso, $\widehat{k} = 9$	0.410	24%	100%	7.39	0.292
Factor-adjusted Lasso, $\widehat{k} = 12$	0.411	23%	100%	7.62	0.285
Bayes, $\widehat{k} = 0$	0.189	42.2%	100.0%	5.80	0.110
Factor-adjusted Bayes, $\widehat{k} = 3$	0.125	84.5%	100.0%	5.16	0.080
Factor-adjusted Bayes, $\widehat{k} = 6$	0.128	83.9%	100.0%	5.18	0.084
Factor-adjusted Bayes, $\widehat{k} = 9$	0.135	83.7%	100.0%	5.18	0.082
Factor-adjusted Bayes, $\widehat{k} = 12$	0.133	85.6%	100.0%	5.16	0.086

Table 1: Experimental results in the default setting.

In case that covariates $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are not correlated, the factor-adjusted Bayesian method performs slightly worse than the Bayesian method (Table 2).

Method	$\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|$	$\xi = \xi^\star$ (%)	$\xi \supseteq \xi^\star$ (%)	$\|\xi\|$	$\|\sigma^2/\sigma^{\star 2} - 1\|$
Lasso, $\widehat{k} = 0$	0.311	0%	100%	31.27	0.134
Factor-adjusted Lasso, $\widehat{k} = 3$	0.414	32%	100%	6.53	0.317
Factor-adjusted Lasso, $\widehat{k} = 6$	0.415	27%	100%	6.67	0.310
Factor-adjusted Lasso, $\widehat{k} = 9$	0.417	24%	100%	7.06	0.305
Factor-adjusted Lasso, $\widehat{k} = 12$	0.419	20%	100%	7.16	0.297
Bayes, $\widehat{k} = 0$	0.119	85.7%	100.0%	5.16	0.091
Factor-adjusted Bayes, $\widehat{k} = 3$	0.123	85.7%	100.0%	5.16	0.091
Factor-adjusted Bayes, $\widehat{k} = 6$	0.124	84.4%	100.0%	5.17	0.090
Factor-adjusted Bayes, $\widehat{k} = 9$	0.127	84.6%	100.0%	5.17	0.091
Factor-adjusted Bayes, $\widehat{k} = 12$	0.129	85.2%	100.0%	5.16	0.091

Table 2: Experimental results in the setting with no common factor.

The model selection rate for the Bayesian methods has a different meaning to that for the Lasso methods. For example, 50% model selection rate given by a Lasso method means that the true sparse model is selected in 50 out of 100 replicates of the dataset. 90% model selection rate given by a Bayesian method means that every 9 of 10 posterior samples select the true sparse model in each replicate of the dataset on average. In the experiments summarized by Tables 1 and 2, at least every 7 of 10 pseudo-posterior samples obtained by the factor-adjusted Bayesian method select the true sparse model in each of 100 replicates of the dataset. A majority voting rule would definitely enhance the model selection rate of the factor-adjusted Bayesian methods.

6.2 Impacts of Correlations among Covariates

As discussed in the introduction, the sparse regression methods on model (1) fail to work when the covariates are strongly correlated, and the factor adjustment are intended to address the issue. To showcase this issue, we vary the magnitude of factor loading coefficients $\mathbf{B}$ in the default setting and draw $\mathbf{B} \sim \text{Uniform}[-B_{\max}, +B_{\max}]^{p\times k}$ with $B_{\max} = 2.0, 2.5, 3.0, 3.5, 4.0, 4.5$. A larger $B_{\max}$ indicates a smaller sparse eigenvalue of $\mathbf{X}$. Neither the routine Lasso method nor the routine Bayesian method works when $B_{\max} \geq 4.0$ (Fig. 2).

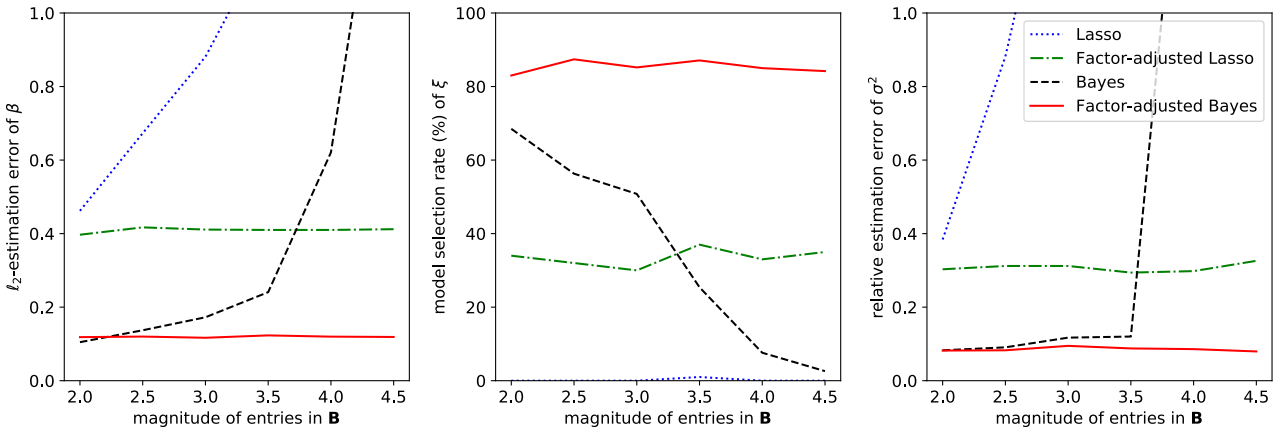


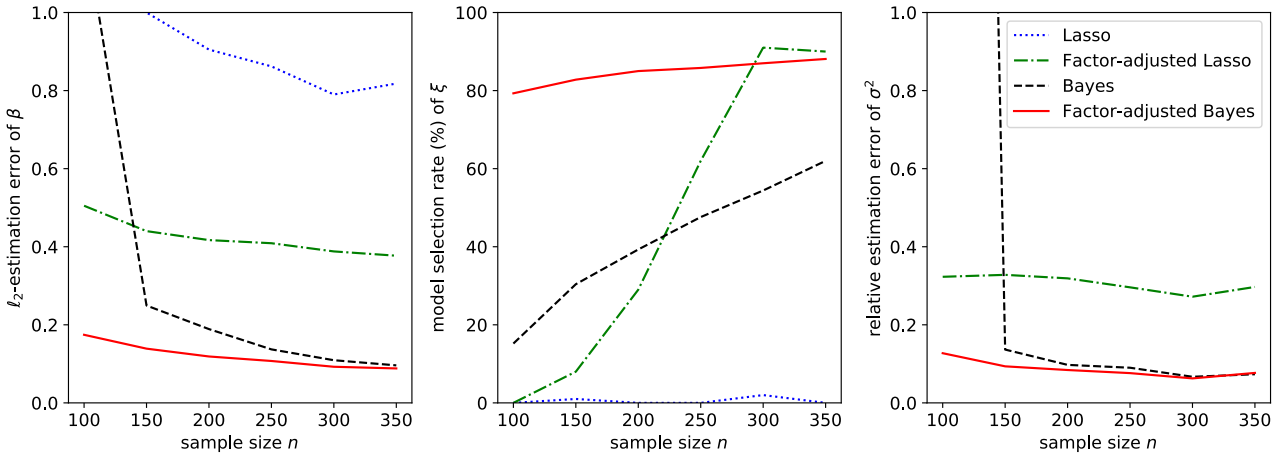
Figure 2: ℓ_2-estimation error of $\boldsymbol{\beta}$ (left), model selection rate of ξ (middle) and relative estimation error of σ^2 (right) influenced by the magnitude of entries in $\mathbf{B}$. Factor-adjusted methods use $\widehat{k} = k = 3$.

6.3 Scalability as n, p, s Increase

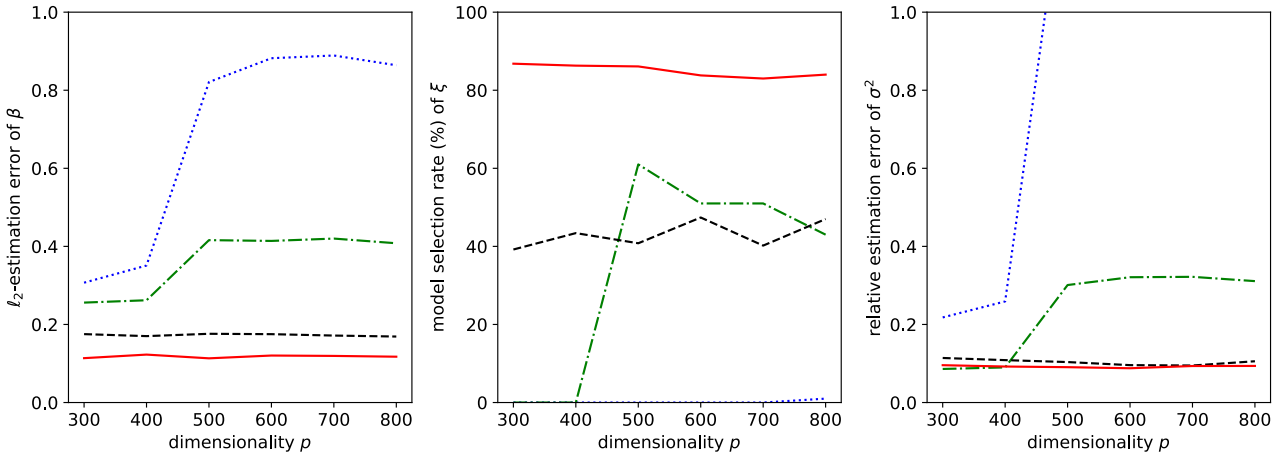
The proposed method is tested with various setups of the sample size n, the dimensionality p and the sparsity s. In Fig. 3(a), $p = 500$ and $s = 5$ are fixed, and n is varied. In Fig. 3(b), $n = 200$ and $s = 5$ are fixed, and p is varied. In Fig. 3(c), $n = 200$ and $p = 500$ are fixed, and s is varied. For factor-adjusted methods, $\widehat{k} = k = 3$ are used. Overall, the factor-adjusted Bayesian method outperforms the other three methods on both parameter estimation and model selection tasks under most combinations of (n, p, s).

6.4 Convergence Diagnostics for Gibbs Sampler

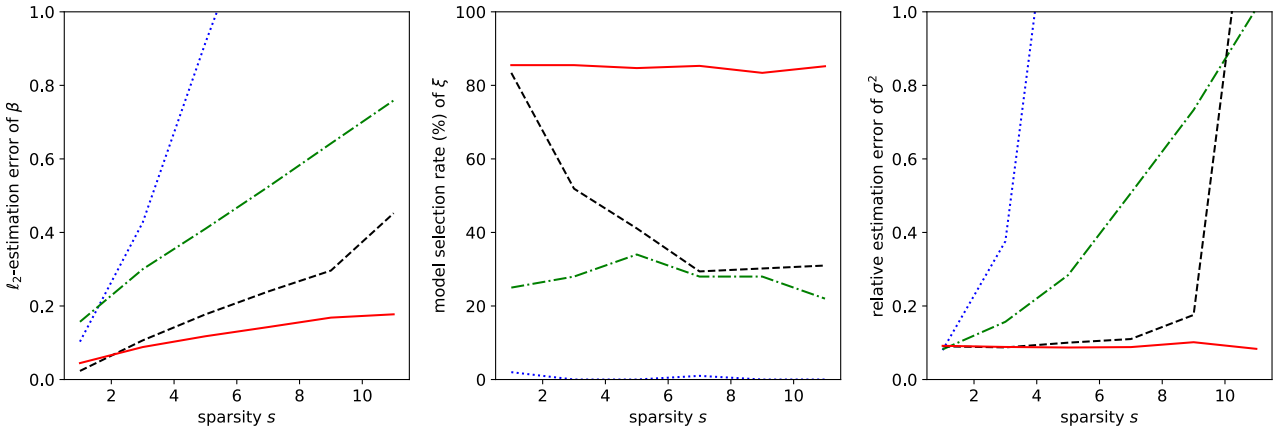
A Gibbs sampler is designed for the posterior computation of the factor-adjusted Bayesian method. We provide a graphics tool to diagnose the convergence of this Gibbs sampler towards the target



(a)



(b)



(c)

Figure 3: ℓ_2-estimation error of $\boldsymbol{\beta}$ (left), model selection rate of ξ (middle) and relative estimation error of σ^2 (right) influenced by (a) sample size n, (b) dimensionality p and (c) sparsity s. Factor-adjusted methods use $\widehat{k} = k = 3$.

distributions (Fig. 4). At each iteration t, the current regression coefficients $\boldsymbol{\beta}^{(t)}$ is compared to the previous regression coefficients $\boldsymbol{\beta}^{(t-1)}$ in terms of the Euclidean distance, and the current model $\xi^{(t)}$ is compared to the previous model $\xi^{(t-1)}$ in terms of Jaccard distance $1 - \frac{|\xi^{(t)} \cap \xi^{(t-1)}|}{|\xi^{(t)} \cup \xi^{(t-1)}|}$. In Fig. 4, the Gibbs

sampler converges to the target distribution for the factor-adjust Bayesian method after 6 iterations. However, it does converge for the routine Bayesian method after 20 iterations, because the routine Bayesian method is performed on strongly correlated covariates with a small sparse eigenvalue. A small sparse eigenvalue often leads to slow convergence speeds of Bayesian sparse regression methods (Yang et al., 2016).

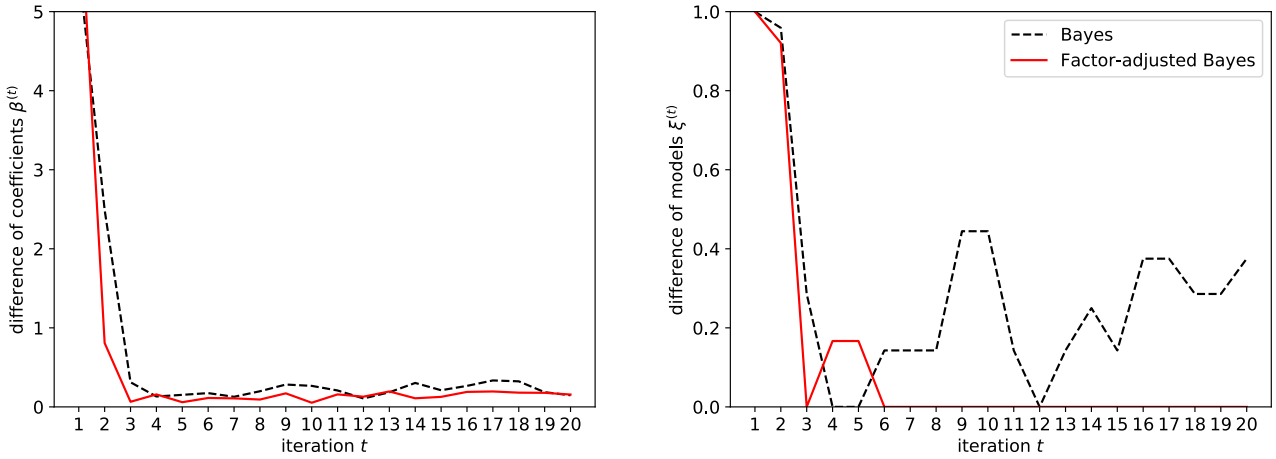


Figure 4: Convergence diagnostics for the Gibbs sampler.

7 Predicting U.S. Bond Risk Premia

This section applies our method to predict U.S. bond risk premia with a large panel of macroeconomic variables. The response variables are monthly U.S. bond risk premia with maturity of $m = 2, 3, 4, 5$ years spanning the period from January, 1964 to December, 2003 (Ludvigson and Ng, 2009). The m-year bond risk premium at period $i+1$ is defined as the (log) holding return from buying an m-year bond at period i and selling it as an $(m-1)$-year bond at period $i+1$, exceeding the (log) return on one-year bond bought at period i. The covariates are $p = 131$ macroeconomic variables collected in the FRED-MD database (McCracken and Ng, 2016) during the same period. The scree plot of PCA of these covariates (Fig. 5) shows the strong correlations among $p = 131$ covariates. The first principal component accounts for 55.9% of the total variation of the covariates, and that the first 5 principal components account for 89.7% of the total variation of the covariates.

The rolling window regression and next value prediction are considered. Specifically, each of two-year, three-year, four-year and five-year U.S. bond risk premia is regressed on the macroeconomic variables in the previous month. For each time window of size $n = 120$ ahead of month $t = n + 2, \ldots, 480$, fit the sparse regression model

$$y_i = f(\boldsymbol{x}_{i-1}) + \sigma\varepsilon_i, \quad i = t-n, \ldots, t-1,$$

and give an out-of-sample prediction $\widehat{y}_t = \widehat{f}(\boldsymbol{x}_{t-1})$. The standard sparse regression model (1) and the factor-adjusted model (4) are considered, and corresponding Bayesian and Lasso methods are performed. For the factor-adjusted methods, the number of common factors k is estimated by the maximum eigenvalue ratio method as (7). For the Bayesian methods, we set $s_0 = 20$ in the prior distribution (8). The principal component regression method (Wehrens and Mevik, 2007) is also

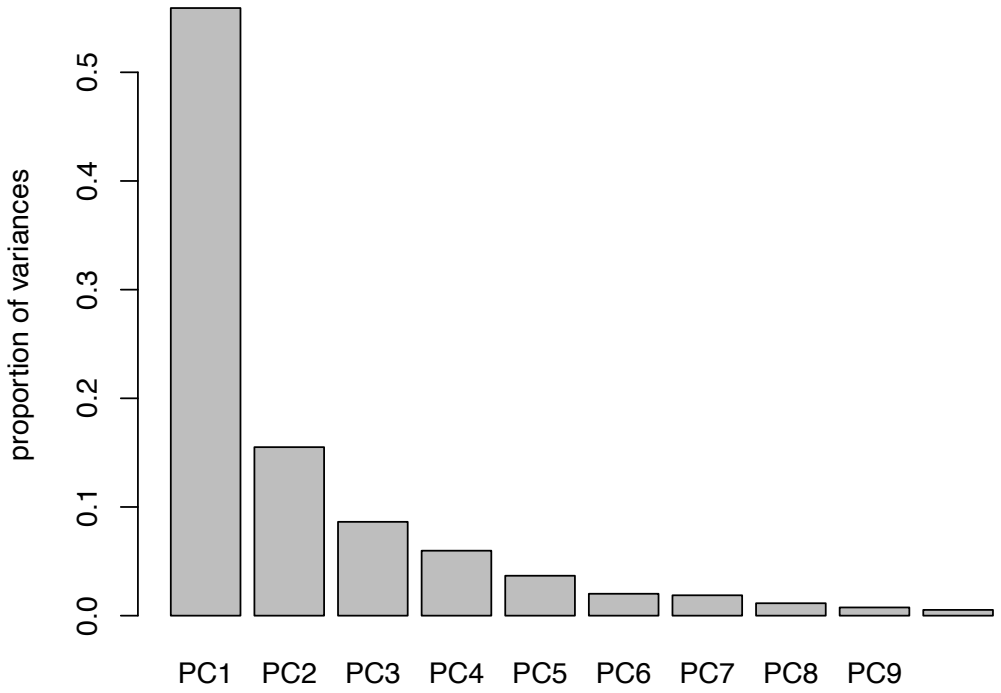


Figure 5: Proportion of variances explained by the first 10 principal components.

included for comparison. In a similar vein to (Ludvigson and Ng, 2009), the top eight principal components are taken from PCA to be covariates in the regression analysis.

The performances of these regression methods are evaluated in terms of the out-of-sample R^2-value.

$$R^2 = 1 - \frac{\sum_{t=n+2}^{480}(\widehat{y}_t - y_t)^2}{\sum_{t=n+2}^{480}(\bar{y}_t - y_t)^2},$$

where y_t is one of two-year, three-year, four-year and five-year U.S. bond risk premia, $\widehat{y}_t$ is the prediction of y_t given by the fitted regression model, and $\bar{y}_t$ is the average of $\{y_{t-n}, \ldots, y_{t-1}\}$. Tables 3 and 4 collect the out-of-sample R^2 values and the average model sizes the five methods achieve on the dataset of U.S. bound risk premia. The factor-adjusted Bayesian method achieves the highest out-of-sample R^2 value and select the sparsest models among all methods in comparison.

Method	2-yr bond	3-yr bond	4-yr bond	5-yr bond
Principal Component Regression	0.646	0.603	0.568	0.540
Lasso	0.728	0.721	0.703	0.685
Factor-adjusted Lasso	0.761	0.751	0.736	0.719
Bayes	0.737	0.715	0.698	0.674
Factor-adjusted Bayes	0.765	0.763	0.752	0.728

Table 3: Out-of-sample R^2 values achieved on the dataset of U.S. bond risk premia.

8 Discussion

We propose a factor-adjusted sparse regression model (4) to handle highly correlated covariates. We decompose the covariates into strong correlation parts driven by common factors and idiosyncratic

Method	2-yr bond	3-yr bond	4-yr bond	5-yr bond
Lasso	24.32	24.77	25.99	26.27
Factor-adjusted Lasso	26.47	26.94	27.29	26.44
Bayes	19.74	22.55	24.43	25.04
Factor-adjusted Bayes	14.92	19.36	21.48	22.38

Table 4: The average sizes of sparse models selected for the dataset of U.S. bond risk premia.

components, where the common factors explain most of the variations. All common factors but a small number of idiosyncratic components are assumed to contribute to the response. The corresponding Bayesian methodology is then developed for estimating such a model. Theoretical results suggest that the proposed methodology can consistently identify and estimate nonzero regression coefficients.

In the factor-adjusted model, sparse regression methods require the weak correlation condition on idiosyncratic components $\mathbf{U}$, which is easier to hold than that on original covariates $\mathbf{X}$ in the sparse regression model (1) and the factor-augmented regression model (6). Section 2 makes this intuition precise by quantitatively characterizing the ratio between sparse eigenvalues of $\mathbf{U}$ and $\mathbf{X}$. When covariates are strongly correlated, the factor-adjusted Bayesian method outperforms the routine Bayesian method (Table 1). When covariates are not correlated (although it is unlike the case in practice), the factor-adjusted Bayesian method pays a negligible price for model misspecification (Table 2). In case of extremely strong correlation among covariates, both routine Bayesian and Lasso methods fail to work, but the factor-adjust Bayesian and Lasso methods perform robustly (Fig. 2). The factor adjustment also enhances the computational efficiency of the Bayesian method (Fig. 4).

The factor-adjusted model covers the standard sparse regression model as a sub-model. Thus it provides more flexibility in the regression analysis and potentially explores more explanatory power from the data. On the dataset of U.S. bond risk premia, the factor-adjusted Bayesian method achieves 2.8%-5.4% more out-of-sample R^2 values with 3-5 less variables (Tables 3 and 4). We hereby recommend the factor-adjusted model over the standard model for regression analyses on real datasets with highly correlated covariates.

Acknowledgement

We would like to thank Drs Qifan Song, Faming Liang, Yun Yang for helpful discussions on the properties of Bayesian sparse regression methods.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

Allen-Zhu, Z. and Li, Y. (2016). Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*.

Armagan, A., Dunson, D. B. and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica* **23** 119.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150.

Baraniuk, R., Davenport, M., DeVore, R. and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* **28** 253–263.

Barron, A. R. (1998). Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. *Bayesian Statistics* **6** 27–52.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429.

Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics* **36** 2577–2604.

Bickel, P. J., Ritov, Y., Tsybakov, A. B. et al. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics* **37** 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer, New York.

Bunea, F., Tsybakov, A., Wegkamp, M. et al. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Annals of Statistics* **43** 1986–2018.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis* **7** 1–46.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences* **100** 2197–2202.

Donoho, D. L., Elad, M. and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory* **52** 6–18.

Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory* **47** 2845–2862.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33** 3–56.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147** 186–197.

Fan, J., Guo, Y. and Jiang, B. (2019). Adaptive huber regression on markov-dependent data. *Stochastic Processes and their Applications* to appear.

Fan, J., Ke, Y. and Wang, K. (2020a). Decorrelation of covariates for high dimensional sparse regression. *Journal of Econometrics* **216** 71–85.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2011a). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* **39** 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 603–680.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics* **46** 814–841.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20** 101–148.

Fan, J., Lv, J. and Qi, L. (2011b). Sparse high-dimensional models in economics. *Annual Review of Economics* **3** 291–317.

Fan, J., Wang, K., Zhong, Y. and Zhu, Z. (2020b). Robust high dimensional factor models with applications to statistical machine learning. *Statistical Science* to appear.

Forbes, K. J. and Rigobon, R. (2002). No contagion, only interdependence: measuring stock market comovements. *Journal of Finance* **57** 2223–2261.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531.

Jiang, B., Sun, Q. and Fan, J. (2018). Bernstein's inequality for general markov chains. *arXiv preprint arXiv:1805.10721* .

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

Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research* **13** 1037–1057.

Kneip, A. and Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *Annals of Statistics* **39** 2410–2447.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics* **40** 694–726.

Lancaster, P. and Tismenetsky, M. (1985). *The theory of matrices: with applications*. Elsevier.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

Lintner, J. (1975). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. In *Stochastic Optimization Models in Finance*. Elsevier, 131–155.

Ludvigson, S. C. and Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies* **22** 5027–5067.

Luo, R., Wang, H. and Tsai, C.-L. (2009). Contour projected dimension reduction. *Annals of Statistics* **37** 3743–3778.

McCracken, M. W. and Ng, S. (2016). FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics* **34** 574–589.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics* **42** 789–817.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103** 681–686.

Pelekis, C. (2016). Lower bounds on binomial and poisson tails: an approach via tail conditional expectations. *arXiv preprint arXiv:1609.06651* .

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 287–311.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association* **113** 431–444.

Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk. *Journal of Finance* **19** 425–442.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics* **29** 687–714.

Song, Q. and Liang, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964* .

Stewart, G. W. (1990). *Matrix perturbation theory*. Academic Press.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97** 1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* **20** 147–162.

Su, W. and Candes, E. (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics* **44** 1038–1068.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58** 267–288.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *Annals of Statistics* **45** 1342–1374.

Wehrens, R. and Mevik, B.-H. (2007). The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software* **18** 1–24.

Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *Annals of Statistics* **44** 2497–2532.

Yu, Y., Wang, T. and Samworth, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. *Biometrika* **102** 315–323.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics* **36** 1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7** 2541–2563.

Appendices

All notation used in the appendices are listed as follows. Some of them may have been defined in the main body of the paper.

For an index set ξ, write $|\xi|$ as its cardinality and ξ^c as its complement (with respect to the whole index set $\{1,\dots,p\}$). For two index sets ξ, ξ', write $\xi \setminus \xi'$ as the set difference. For a vector $\boldsymbol{v}$, $\boldsymbol{v}_\xi$ denotes the sub-vector assembling components indexed by ξ, $\|\boldsymbol{v}\|$ denotes the ℓ_2 norm, and $\|\boldsymbol{v}\|_0$ denotes the number of non-zero entries.

For a matrix $\mathbf{A}_{m_1\times m_2} = [a_{ij}]_{1\le i\le m_1, 1\le j\le m_2}$, write uppercase $\mathbf{A}_j$ for its j-th column, and lowercase $\boldsymbol{a}_i$ for its i-th row. Let $\mathbf{A}_\xi = [\mathbf{A}_j : j \in \xi]$ be the sub-matrix of $\mathbf{A}$ assembling the columns indexed by $\xi \subseteq \{1,\dots,m\}$. Let $\|\mathbf{A}\|_{\max} = \max_{i,j}|a_{ij}|$ be its element-wise maximum norm, $\|\mathbf{A}\|$ be its operator norm induced by the ℓ_2 norm of vectors, and $\|\mathbf{A}\|_{\mathrm{F}}$ be its Frobenius norm. Let $\mathrm{vec}(\mathbf{A})$ be the vectorization of $\mathbf{A}$ formed by concatenating column vectors of $\mathbf{A}$. For a matrix $\mathbf{A}$ of full column rank, write $\mathbf{A}^\dagger = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ as its left pseudo-inverse, then $\mathbf{A}\mathbf{A}^\dagger$ is the projection matrix on the column space of $\mathbf{A}$.

For a symmetric matrix $\mathbf{A}$, write its largest eigenvalue as $\lambda_{\max}(\mathbf{A})$, its smallest eigenvalue as $\lambda_{\min}(\mathbf{A})$, and its trace as $\mathrm{trace}(\mathbf{A})$. Write $\mathrm{diag}(a_1,\dots,a_m)$ for a diagonal matrix of elements $a_1,\dots,a_m$. For two squared matrices $\mathbf{A},\mathbf{B}$ of the same dimension, we write $\mathbf{A}\ge\mathbf{B}$ (or $\mathbf{B}\le\mathbf{A}$) if $\mathbf{A}-\mathbf{B}$ is positive semidefinite.

For two positive sequences a_n, b_n, $a_n \succcurlyeq b_n$ (or $b_n \preccurlyeq a_n$) means $b_n = O(a_n)$; $a_n \succ b_n$ (or $b_n \prec a_n$) means $b_n = o(a_n)$; and $a_n \asymp b_n$ means both $a_n \succcurlyeq b_n$ and $a_n \preccurlyeq b_n$. $a_n \gtrsim b_n$ (or $b_n \lesssim a_n$) means that $a_n > b_n$ for sufficiently large n.

A Technical Proofs for Factor Model Estimation

This appendix collects technical proofs for Theorem 1, Corollary 1, and Examples 1 and 2 concerning the estimation of factor models.

A.1 Proof of Theorem 1

We first prepare four preliminary results as Lemmas A1 to A4 and then prove Theorem 1 and Corollary 1.

Lemma A1. *Suppose Assumption 1 holds. With high probability at least* $1-o_n$,

$$\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\| \le \|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n - \mathbf{I}\|_{\mathrm{F}} \le L_0 k\sqrt{\log p/n},$$
$$\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\| \le \|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n - \mathbf{0}\|_{\mathrm{F}} \le L_0\sqrt{kp\log p/n},$$
$$\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \boldsymbol{\Sigma}\| \le \|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \boldsymbol{\Sigma}\|_{\mathrm{F}} \le L_0 p\sqrt{\log p/n}.$$

Proof. The proof uses merely the relations between the operator norm, the Frobenius norm, and the element-wise maximum norm of matrices. □

Lemma A2. *With probability at least* $1-\frac{\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B})}{p^2\log p/n}$,

$$\|\mathbf{U}\mathbf{B}\|_{\mathrm{F}} \le p\sqrt{\log p}.$$

Proof. Applying Markov's inequality yields

$$\mathbb{P}\left(\|\mathbf{UB}\|_{\mathrm{F}} \geq p\sqrt{\log p}\right) = \mathbb{P}\left(\|\mathbf{UB}\|_{\mathrm{F}}^2 \geq p^2 \log p\right) \leq \frac{\mathbb{E}\left[\|\mathbf{UB}\|_{\mathrm{F}}^2\right]}{p^2 \log p}.$$

And,

$$\mathbb{E}\left[\|\mathbf{UB}\|_{\mathrm{F}}^2\right] = \mathbb{E}\left[\mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UB})\right] = \mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbb{E}\left[\mathbf{U}^{\mathrm{T}}\mathbf{U}\right]\mathbf{B}) = n \times \mathrm{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}).$$

□

Lemma A3 (Variant of Davis-Kahan Theorem)**.** *Let $\widehat{\mathbf{A}}$ be an $n \times n$ symmetric matrix with eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n$ and corresponding eigenvectors $\widehat{\psi}_1, \ldots, \widehat{\psi}_n$. Fix $1 \leq l \leq r \leq n$ and assume that $\min\{\widehat{\lambda}_{l-1} - \widehat{\lambda}_l, \widehat{\lambda}_r - \widehat{\lambda}_{r+1}\} > 0$, where $\widehat{\lambda}_0 := +\infty$ and $\widehat{\lambda}_{n+1} := -\infty$. Let $\widehat{\mathbf{\Lambda}}$ be the diagonal matrix of eigenvalues $\widehat{\lambda}_l, \ldots, \widehat{\lambda}_r$, and $\widehat{\mathbf{\Lambda}}_c$ be the diagonal matrix of other eigenvalues. Let $\widehat{\mathbf{\Psi}}$ and $\widehat{\mathbf{\Psi}}_c$ be their corresponding eigenvectors of $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_c$, respectively. Let $\mathbf{A}$ be an $n \times n$ matrix with "$\mathbf{\Delta}$-approximate" eigenequation*

$$\mathbf{A}\mathbf{\Psi} = \mathbf{\Psi}\mathbf{\Lambda} + \mathbf{\Delta},$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_l, \ldots, \lambda_r)$ and $\mathbf{\Psi} = (\psi_l, \ldots, \psi_r)$ consists of $k = l - r + 1$ (not necessarily orthonormal) vectors. Then

$$\|\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\|_{\mathrm{F}} \leq \frac{\|\mathbf{\Delta}\|_{\mathrm{F}} + \|(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{\Psi}\|_{\mathrm{F}} + \|\mathbf{\Psi}\|\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\mathrm{F}}}{\min\{\widehat{\lambda}_{l-1} - \widehat{\lambda}_l, \widehat{\lambda}_r - \widehat{\lambda}_{r+1}\}}.$$

Proof. The $\mathbf{\Delta}$-approximate eigenequation derives that

$$\mathbf{\Delta} = \mathbf{A}\mathbf{\Psi} - \mathbf{\Psi}\mathbf{\Lambda} = \widehat{\mathbf{A}}\mathbf{\Psi} - \mathbf{\Psi}\widehat{\mathbf{\Lambda}} + (\mathbf{A} - \widehat{\mathbf{A}})\mathbf{\Psi} - \mathbf{\Psi}(\mathbf{\Lambda} - \widehat{\mathbf{\Lambda}}),$$

implying

$$\|\widehat{\mathbf{A}}\mathbf{\Psi} - \mathbf{\Psi}\widehat{\mathbf{\Lambda}}\|_{\mathrm{F}} \leq \|\mathbf{\Delta}\|_{\mathrm{F}} + \|(\widehat{\mathbf{A}} - \mathbf{A})\mathbf{\Psi}\|_{\mathrm{F}} + \|\mathbf{\Psi}\|\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_{\mathrm{F}}.$$

It is left to show that

$$\|\widehat{\mathbf{A}}\mathbf{\Psi} - \mathbf{\Psi}\widehat{\mathbf{\Lambda}}\|_{\mathrm{F}} \geq \min\{\widehat{\lambda}_{l-1} - \widehat{\lambda}_l, \widehat{\lambda}_r - \widehat{\lambda}_{r+1}\}\|\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\|_{\mathrm{F}}.$$

To this end, from the facts that $\widehat{\mathbf{A}} = \widehat{\mathbf{\Psi}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Psi}}^{\mathrm{T}} + \widehat{\mathbf{\Psi}}_c\widehat{\mathbf{\Lambda}}_c\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}$ and that $\mathbf{I} = \widehat{\mathbf{\Psi}}\widehat{\mathbf{\Psi}}^{\mathrm{T}} + \widehat{\mathbf{\Psi}}_c\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}$ it follows that

$$\widehat{\mathbf{A}}\mathbf{\Psi} - \mathbf{\Psi}\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Psi}}\underbrace{(\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Psi}}^{\mathrm{T}}\mathbf{\Psi} - \widehat{\mathbf{\Psi}}^{\mathrm{T}}\mathbf{\Psi}\widehat{\mathbf{\Lambda}})}_{\mathbf{S}_1} + \widehat{\mathbf{\Psi}}_c\underbrace{(\widehat{\mathbf{\Lambda}}_c\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi} - \widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\widehat{\mathbf{\Lambda}})}_{\mathbf{S}_2}.$$

Further,

$$\begin{aligned}\|\widehat{\mathbf{A}}\mathbf{\Psi} - \mathbf{\Psi}\widehat{\mathbf{\Lambda}}\|_{\mathrm{F}}^2 = \|\widehat{\mathbf{\Psi}}\mathbf{S}_1 + \widehat{\mathbf{\Psi}}_c\mathbf{S}_2\|_{\mathrm{F}}^2 &= \mathrm{trace}\left[(\widehat{\mathbf{\Psi}}\mathbf{S}_1 + \widehat{\mathbf{\Psi}}_c\mathbf{S}_2)^{\mathrm{T}}(\widehat{\mathbf{\Psi}}\mathbf{S}_1 + \widehat{\mathbf{\Psi}}_c\mathbf{S}_2)\right] \\ &= \mathrm{trace}\left[\mathbf{S}_1^{\mathrm{T}}\mathbf{S}_1 + \mathbf{S}_2^{\mathrm{T}}\mathbf{S}_2\right] \geq \mathrm{trace}\left[\mathbf{S}_2^{\mathrm{T}}\mathbf{S}_2\right] = \|\mathbf{S}_2\|_{\mathrm{F}}^2.\end{aligned}$$

Proceed to lower bound the term $\|\mathbf{S}_2\|_{\mathrm{F}}$. For real matrices $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3$, we write $\mathrm{vec}(\mathbf{T}_1)$ as the vectorization of $\mathbf{T}_1$ formed by concatenating column vectors of $\mathbf{T}_1$, and denote by $\mathbf{T}_1 \otimes \mathbf{T}_2$ the Kronecker product of matrices $\mathbf{T}_1$ and $\mathbf{T}_2$. Using the identity $\mathrm{vec}(\mathbf{T}_1\mathbf{T}_2\mathbf{T}_3) = \mathbf{T}_3^{\mathrm{T}} \otimes \mathbf{T}_1\mathrm{vec}(\mathbf{T}_2)$ for any matrices $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3$ with appropriate dimensions, we have

$$\begin{aligned}\|\mathbf{S}_2\|_{\mathrm{F}} &= \|\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\widehat{\mathbf{\Lambda}} - \widehat{\mathbf{\Lambda}}_c\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\|_{\mathrm{F}} = \|\mathrm{vec}(\mathbf{I}_{n-k}\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\widehat{\mathbf{\Lambda}}) - \mathrm{vec}(\widehat{\mathbf{\Lambda}}_c\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\mathbf{I}_k)\| \\ &= \|\widehat{\mathbf{\Lambda}} \otimes \mathbf{I}_{n-k}\mathrm{vec}(\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}) - \mathbf{I}_k \otimes \widehat{\mathbf{\Lambda}}_c\mathrm{vec}(\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi})\| \\ &\geq \min\{\widehat{\lambda}_{l-1} - \widehat{\lambda}_l, \widehat{\lambda}_r - \widehat{\lambda}_{r+1}\}\|\mathrm{vec}(\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi})\| = \min\{\widehat{\lambda}_{l-1} - \widehat{\lambda}_l, \widehat{\lambda}_r - \widehat{\lambda}_{r+1}\}\|\widehat{\mathbf{\Psi}}_c^{\mathrm{T}}\mathbf{\Psi}\|_{\mathrm{F}}.\end{aligned}$$

This concludes the proof. □

Lemma A4. *Suppose Assumption 1 holds. Let $\mathbf{D}_{k\times k}$ be the diagonal matrix of k singular values of $\mathbf{F}/\sqrt{n}$. With high probability at least $1-o_n$,*

$$\|\mathbf{D}^2-\mathbf{I}\|_{\mathrm{F}} \le L_0 k\sqrt{\log p/n}.$$

Proof. This claim immediately follows from Lemma A1. Indeed, let $\mathbf{V}$ be the right singular vectors of $\mathbf{F}$ then

$$\|\mathbf{D}^2-\mathbf{I}\|_{\mathrm{F}} = \|\mathbf{V}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{F}\mathbf{V}/n-\mathbf{I}\|_{\mathrm{F}} = \|\mathbf{V}^{\mathrm{T}}(\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I})\mathbf{V}\|_{\mathrm{F}} = \|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I}\|_{\mathrm{F}}.$$

□

Proof of Theorem 1(a). It suffices to bound

$$\|\mathbf{X}^{\mathrm{T}}\mathbf{X}/n-\mathbf{B}\mathbf{B}^{\mathrm{T}}\| \le L_1 p\sqrt{\log p/n}$$

for some constant L_1, then Weyl's theorem on perturbed eigenvalues can apply. Indeed,

$$\begin{aligned}\mathbf{X}^{\mathrm{T}}\mathbf{X}/n-\mathbf{B}\mathbf{B}^{\mathrm{T}} &= (\mathbf{F}\mathbf{B}^{\mathrm{T}}+\mathbf{U})^{\mathrm{T}}(\mathbf{F}\mathbf{B}^{\mathrm{T}}+\mathbf{U})/n-\mathbf{B}\mathbf{B}^{\mathrm{T}}\\ &= \mathbf{B}(\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I})\mathbf{B}^{\mathrm{T}}+\mathbf{U}^{\mathrm{T}}\mathbf{F}\mathbf{B}^{\mathrm{T}}/n+\mathbf{B}\mathbf{F}^{\mathrm{T}}\mathbf{U}/n+\mathbf{U}^{\mathrm{T}}\mathbf{U}/n.\end{aligned}$$

Each of four terms can be bounded by Lemma A1 and Assumption 2(i)(ii). Precisely,

$$\begin{aligned}\|\mathbf{B}(\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I})\mathbf{B}^{\mathrm{T}}\| &\le \|\mathbf{B}\|^2\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I}\| \le \lambda_1 p\times L_0\sqrt{k^2\log p/n},\\ \|\mathbf{U}^{\mathrm{T}}\mathbf{F}\mathbf{B}^{\mathrm{T}}/n\| = \|\mathbf{B}\mathbf{F}^{\mathrm{T}}\mathbf{U}/n\| &\le \|\mathbf{B}\|\|\mathbf{F}^{\mathrm{T}}\mathbf{U}/n\| \le \sqrt{\lambda_1 p}\times L_0\sqrt{kp\log p/n},\\ \|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n\| &\le \|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n-\mathbf{\Sigma}\|+\|\mathbf{\Sigma}\| = L_0\sqrt{p^2\log p/n}+O(p\sqrt{\log p/n}).\end{aligned}$$

□

Proof of Theorem 1(b). Write the eigendecomposition of $\mathbf{B}^{\mathrm{T}}\mathbf{B}/p$ as $\mathbf{R}\mathbf{\Lambda}\mathbf{R}^{\mathrm{T}}$ with $\mathbf{\Lambda}=\mathrm{diag}(\lambda_1,\dots,\lambda_k)$ being the diagonal matrix of eigenvalues of $\mathbf{B}^{\mathrm{T}}\mathbf{B}/p$, then

$$\frac{\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}}}{np}\times\frac{\mathbf{F}\mathbf{R}}{\sqrt{n}} = \frac{\mathbf{F}\mathbf{R}}{\sqrt{n}}\times\mathbf{\Lambda}+\mathbf{\Delta}, \quad \text{with } \mathbf{\Delta}=(\mathbf{F}/\sqrt{n})(\mathbf{B}^{\mathrm{T}}\mathbf{B}/p)\mathbf{\Lambda}(\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I})\mathbf{R}.$$

Recall that $\widehat{\mathbf{\Lambda}}=\mathrm{diag}(\widehat{\lambda}_1,\dots,\widehat{\lambda}_k)$ is the diagonal matrix of k largest eigenvalues of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$, and write the eigenequation of $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$ as

$$\frac{\mathbf{X}\mathbf{X}^{\mathrm{T}}}{np}\times\frac{\widehat{\mathbf{F}}}{\sqrt{n}} = \frac{\widehat{\mathbf{F}}}{\sqrt{n}}\times\widehat{\mathbf{\Lambda}}.$$

Let $\widehat{\mathbf{F}}_c$ be $\sqrt{n}$ times other $n-k$ eigenvectors $\mathbf{X}\mathbf{X}^{\mathrm{T}}/np$ that are orthogonal to those in $\widehat{\mathbf{F}}/\sqrt{n}$. By the variant of Davis-Kahan theorem (Lemma A3) and the orthogonality of $\mathbf{R}$,

$$\|\widehat{\mathbf{F}}_c^{\mathrm{T}}\mathbf{F}/n\|_{\mathrm{F}} \le \frac{\|\mathbf{\Delta}\|_{\mathrm{F}}+\|(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}})\mathbf{F}\|_{\mathrm{F}}/(n^{3/2}p)+\|\mathbf{F}/\sqrt{n}\|\|\mathbf{\Lambda}-\widehat{\mathbf{\Lambda}}\|_{\mathrm{F}}}{\widehat{\lambda}_k-\widehat{\lambda}_{k+1}}.$$

Proceed to bound each term in the quotient.

(a) For the term $\|\mathbf{F}/\sqrt{n}\|$, from Lemma A1 it follows that

$$\left|\|\mathbf{F}/\sqrt{n}\|^2-1\right| = |\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n\|-\|\mathbf{I}\|| \le \|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I}\| \le L_0 k\sqrt{\log p/n}.$$

(b) For the term $\|\mathbf{\Delta}\|_{\mathrm{F}}$, from Lemma A1 and Assumption 2(i) it follows that

$$\begin{aligned}\|\mathbf{\Delta}\|_{\mathrm{F}} &\le \|\mathbf{F}/\sqrt{n}\|_{\mathrm{F}}\|\mathbf{B}^{\mathrm{T}}\mathbf{B}/p\|\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I}\| \le \sqrt{k}\|\mathbf{F}/\sqrt{n}\|\|\mathbf{\Lambda}\|\|\mathbf{F}^{\mathrm{T}}\mathbf{F}/n-\mathbf{I}\| \\ &\le \sqrt{k}\left(1+L_0k\sqrt{\log p/n}\right)\times\lambda_1\times L_0k\sqrt{\log p/n} \preccurlyeq \sqrt{\log p/n}.\end{aligned}$$

(c) For the term $\|(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}})\mathbf{F}\|_{\mathrm{F}}$, write

$$(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathbf{F}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{F}^{\mathrm{T}})\mathbf{F} = \mathbf{X}\mathbf{U}^{\mathrm{T}}\mathbf{F}+\mathbf{U}\mathbf{B}\mathbf{F}^{\mathrm{T}}\mathbf{F}.$$

By part (a) and Lemma A1 ,

$$\begin{aligned}\|\mathbf{X}\mathbf{U}^{\mathrm{T}}\mathbf{F}\|_{\mathrm{F}} &\le \|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{F}^{\mathrm{T}}\mathbf{U}\| \le \sqrt{n}\|\mathbf{X}\|\|\mathbf{F}^{\mathrm{T}}\mathbf{U}\| = \sqrt{np\widehat{\lambda}_1}\times\|\mathbf{F}^{\mathrm{T}}\mathbf{U}\| \\ &\le \sqrt{np(\lambda_1+L_1\sqrt{\log p/n})}\times\sqrt{knp\log p} \preccurlyeq np\sqrt{\log p}.\end{aligned}$$

By part (a), Lemmas A1 and A2,

$$\|\mathbf{U}\mathbf{B}\mathbf{F}^{\mathrm{T}}\mathbf{F}\|_{\mathrm{F}} \le \|\mathbf{U}\mathbf{B}\|_{\mathrm{F}}\|\mathbf{F}\|^2 = p\sqrt{\log p}\times n\left(1+L_0k\sqrt{\log p/n}\right) \preccurlyeq np\sqrt{\log p}.$$

(d) For the term $\widehat{\lambda}_{k+1}-\widehat{\lambda}_k$, from part (a) it follows that

$$|\widehat{\lambda}_{k+1}-0| \le L_1\sqrt{\log p/n}, \quad |\widehat{\lambda}_k-\lambda_k| \le L_1\sqrt{\log p/n}.$$

Collecting these four pieces together yields that, for some constant L_2',

$$\|\widehat{\mathbf{F}}_c^{\mathrm{T}}\mathbf{F}/n\|_{\mathrm{F}} \le L_2'\sqrt{\log p/n}.$$

Next, recall the singular value decomposition $\mathbf{F}/\sqrt{n} = (\widetilde{\mathbf{F}}/\sqrt{n})\mathbf{D}\mathbf{V}^{\mathrm{T}}$ in Lemma A4, and write

$$\|\widehat{\mathbf{F}}_c^{\mathrm{T}}\widetilde{\mathbf{F}}/n\|_{\mathrm{F}} = \|\widehat{\mathbf{F}}_c^{\mathrm{T}}\mathbf{F}\mathbf{V}\mathbf{D}^{-1}/n\|_{\mathrm{F}} \le \|\widehat{\mathbf{F}}_c^{\mathrm{T}}\mathbf{F}/n\|_{\mathrm{F}}\|\mathbf{D}^{-1}\| \le L_2'\sqrt{\frac{\log p/n}{1-L_0k\sqrt{\log p/n}}} \le L_2\sqrt{\log p/n}.$$

for some constant L_2. This derives the desired result, as $\mathbf{\Pi} = \widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^{\mathrm{T}}/n$, $\widehat{\mathbf{\Pi}} = \widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n$ and

$$\|\widehat{\mathbf{F}}_c^{\mathrm{T}}\widetilde{\mathbf{F}}/n\|_{\mathrm{F}} = \|(\mathbf{I}-\mathbf{\Pi})\widehat{\mathbf{\Pi}}\|_{\mathrm{F}} = \|(\mathbf{I}-\widehat{\mathbf{\Pi}})\mathbf{\Pi}\|_{\mathrm{F}} = \|\widehat{\mathbf{\Pi}}-\mathbf{\Pi}\|_{\mathrm{F}}/\sqrt{2}.$$

□

Proof of Theorem 1(c). Recall the singular value decomposition $\mathbf{F}/\sqrt{n} = (\widetilde{\mathbf{F}}/\sqrt{n})\mathbf{D}\mathbf{V}^{\mathrm{T}}$ in Lemma A4. Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be left and right singular vectors of $\widehat{\mathbf{F}}^{\mathrm{T}}\widetilde{\mathbf{F}}/n$, respectively. Set $\mathbf{H} = \mathbf{V}_1\mathbf{V}_2^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}$ then

$$\begin{aligned}\|\widehat{\mathbf{F}}\mathbf{H}-\mathbf{F}\|_{\mathrm{F}} &\le \|\widehat{\mathbf{F}}\mathbf{V}_1\mathbf{V}_2^{\mathrm{T}}(\mathbf{I}-\mathbf{D})\mathbf{V}^{\mathrm{T}}\|_{\mathrm{F}} + \|(\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2)\mathbf{V}_2^{\mathrm{T}}\mathbf{D}\mathbf{V}^{\mathrm{T}}\|_{\mathrm{F}} \\ &\le \|\widehat{\mathbf{F}}\|\|\mathbf{I}-\mathbf{D}\|_{\mathrm{F}} + \|\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2\|_{\mathrm{F}}\|\mathbf{D}\|.\end{aligned}$$

Since $\|\widehat{\mathbf{F}}\| = \sqrt{n}$ and all entries in the diagonal matrix $\mathbf{D}$ are $O(\sqrt{\log p/n})$-close to 1, it is left to bound the term $\|\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2\|_{\mathrm{F}}$. Let $s_1,\ldots,s_k$ be singular values of $\widehat{\mathbf{F}}^{\mathrm{T}}\widetilde{\mathbf{F}}/n$. Clearly, all of them are bounded by $\|\widehat{\mathbf{F}}^{\mathrm{T}}\widetilde{\mathbf{F}}/n\| \le \|\widehat{\mathbf{F}}/\sqrt{n}\|\|\widetilde{\mathbf{F}}/\sqrt{n}\| = 1$. Write

$$\|\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2\|_{\mathrm{F}}^2 = \mathrm{trace}\left[(\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2)^{\mathrm{T}}(\widehat{\mathbf{F}}\mathbf{V}_1-\widetilde{\mathbf{F}}\mathbf{V}_2)\right] = 2n\left(k-\sum_{j=1}^k s_j\right) \le 2n\left(k-\sum_{j=1}^k s_j^2\right),$$

where $k-\sum_{j=1}^k s_j^2$ is the sin-theta distance (Definition 3) between column spaces of $\widehat{\mathbf{F}}$ and $\mathbf{F}$, which has been bounded by part (b). □

Proof of Corollary 1. Recall that $\mathbf{\Pi}$ and $\widehat{\mathbf{\Pi}}$ are projection matrices onto the column spaces of $\mathbf{F}$ and $\widehat{\mathbf{F}}$, respectively, in Theorem 1(b). By the construction of $\widehat{\mathbf{U}}$, the estimation error of $\widehat{\mathbf{U}}_j$ for $\mathbf{U}_j$ is written as

$$\widehat{\mathbf{U}}_j - \mathbf{U}_j = (\mathbf{\Pi} - \widehat{\mathbf{\Pi}})\mathbf{X}_j - \mathbf{\Pi}\mathbf{U}_j.$$

Putting it together with Theorem 1(b) yields

$$\|\widehat{\mathbf{U}}_j/\sqrt{n} - \mathbf{U}_j/\sqrt{n}\| \le L_2\sqrt{2\log p/n}\|\mathbf{X}_j/\sqrt{n}\| + \|\mathbf{\Pi}\mathbf{U}_j/\sqrt{n}\|.$$

For the first term,

$$\begin{aligned}\|\mathbf{X}_j/\sqrt{n}\|^2 &= \|\mathbf{F}\boldsymbol{b}_j/\sqrt{n}\|^2 + \|\mathbf{U}_j/\sqrt{n}\|^2 + \boldsymbol{b}_j^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{U}_j/n \\ &\le (1 + L_0 k\sqrt{\log p/n})\|\boldsymbol{b}_j\|^2 + (\mathbf{\Sigma}_{jj} + L_0\sqrt{\log p/n}) + \|\boldsymbol{b}_j\| \times \sqrt{k}L_0\sqrt{\log p/n},\end{aligned}$$

where $\boldsymbol{b}_j$ is the j-th row of $\mathbf{B}$. For the second term, recall that the singular value decomposition of $\mathbf{F}/\sqrt{n}$ is given by $\widetilde{\mathbf{F}}/\sqrt{n}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ in Lemma A4. Write

$$\|\mathbf{\Pi}\mathbf{U}_j/\sqrt{n}\| = \|(\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^{\mathrm{T}}/n)\mathbf{U}_j/\sqrt{n}\| = \|(\widetilde{\mathbf{F}}/\sqrt{n})\mathbf{V}\mathbf{D}^{-1}(\mathbf{F}^{\mathrm{T}}\mathbf{U}_j/n)\| \le \|\mathbf{D}^{-1}\|\|\mathbf{F}^{\mathrm{T}}\mathbf{U}_j/n\|,$$

where eigenvalues (diagomal entries) of $\mathbf{D}$ are $O(\sqrt{\log p/n})$-close to 1, and $\|\mathbf{F}^{\mathrm{T}}\mathbf{U}_j/n\| \le L_0\sqrt{k\log p/n}$ due to Assumption 1. □

A.2 Proof of Example 1

This example is a consequence of the properties of subexponential and subgaussian random variables, which are commonly seen in the literature of high-dimensional statistics.

Definition 5 (Subexponential Random Variable, also Definition 2.7.5 of Vershynin (2018))**.** *The subexponential norm of a random variable Z is defined as*

$$\|Z\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}e^{-|Z|/t} \le 2\}.$$

A random variable is said subexponential if its subexponential norm is finite.

Definition 6 (Subgaussian Random Variable, also Definition 2.5.6 of Vershynin (2018))**.** . *The subgaussian norm of a random variable Z is defined as*

$$\|Z\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}e^{Z^2/t^2} \le 2\}.$$

A random variable is said subgaussian if its subgaussian norm is finite.

Proof of Example 1. We present the proof of the third inequality in Assumption 1 here. The proofs of the other two inequalities are similar. For each $1 \le j \le p$ and each $1 \le l \le p$, write

$$[\mathbf{U}^{\mathrm{T}}\mathbf{U}/n - \mathbf{\Sigma}]_{jl} = \frac{1}{n}\sum_{i=1}^{n}\left(u_{ij}u_{il} - \mathbb{E}[u_{ij}u_{il}]\right).$$

By Vershynin (2018, Lemma 2.7.7), the product of two subgaussian random variables is subexponential. Formally, $\|u_{ij}u_{il}\|_{\psi_1} \le \|u_{ij}\|_{\psi_2}\|u_{il}\|_{\psi_2} = c_1$. By Vershynin (2018, Exercise 2.7.10), the centered version of $u_{ij}u_{il}$ is still subexponential. Formally, $\|u_{ij}u_{il} - \mathbb{E}[u_{ij}u_{il}]\|_{\psi_1} \le c_2$ for some constant c_2. By

Bernstein's inequality for independent sub-exponential random variables (Vershynin, 2018, Theorem 2.8.2),

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(u_{ij}u_{il}-\mathbb{E}[u_{ij}u_{il}])>\epsilon\right)\leq 2\exp\left(-c_3\min\left\{\frac{n\epsilon^2}{c_2^2},\frac{n\epsilon}{c_2}\right\}\right)$$

for some constant c_3. The union bound for all pairs of $1\leq j,l\leq p$ gives that

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n-\mathbf{\Sigma}\|_{\max}>\epsilon\right)&=\mathbb{P}\left(\cup_{j=1}^{p}\cup_{l=1}^{p}\left\{\frac{1}{n}\sum_{i=1}^{n}(u_{ij}u_{il}-\mathbb{E}[u_{ij}u_{il}])>\epsilon\right\}\right)\\&\leq 2p^2\exp\left(-c_3\min\left\{\frac{n\epsilon^2}{c_2^2},\frac{n\epsilon}{c_2}\right\}\right)\end{aligned}$$

Setting $\epsilon=\sqrt{3c_2^2\log p/c_3 n}$ yields that

$$\|\mathbf{U}^{\mathrm{T}}\mathbf{U}/n-\mathbf{\Sigma}\|_{\max}\leq\sqrt{3c_2^2\log p/c_3 n}$$

with probability at least $1-2/p$. □

A.3 Proof of Example 2

This example is proven by the truncation technique in the literature of high-dimensional matrix estimation (Bickel and Levina, 2008; Fan et al., 2011a, 2013) and a generalized version of Bernstein's inequality for general-state-space Markov chains (Jiang et al., 2018). A proof is provided here for convenience of readers, although it is almost the same with that of Fan et al. (2019, Lemma 1),

Proof of Example 2. We present the proof of the third inequality in Assumption 1 here. The proofs of the other two inequalities are similar. Let the $\mathcal{L}_2$-spectral gap of the Markov chain be $1-\gamma$. Define a truncation operator

$$\mathcal{T}_t(w)=\begin{cases}-t & \text{if } w<-t\\ w & \text{if } |w|\leq t\\ +t & \text{if } w>+t.\end{cases}\tag{14}$$

For each $1\leq j\leq p$ and each $1\leq l\leq p$, write

$$[\mathbf{U}^{\mathrm{T}}\mathbf{U}/n-\mathbf{\Sigma}]_{jl}=\frac{1}{n}\sum_{i=1}^{n}(u_{ij}u_{il}-\mathbb{E}[u_{ij}u_{il}])\leq D_{1jl}+D_{2jl}+D_{3jl},$$

where

$$\begin{aligned}D_{1jl}&=\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathcal{T}_t(u_{ij}u_{il})-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[u_{ij}u_{il}]\right|,\\D_{2jl}&=\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{T}_t(u_{ij}u_{il})-\frac{1}{n}\sum_{i=1}^{n}u_{ij}u_{il}\right|,\\D_{3jl}&=\left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{T}_t(u_{ij}u_{il})-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathcal{T}_t(u_{ij}u_{il})]\right|.\end{aligned}$$

Using the fact that $|\mathcal{T}_t(w)-w|\leq|w|1\{|w|>t\}\leq|w|^2/t$, Cauchy-Schwarz inequality and the assumption that $|u_{ij}(z)|\leq\bar{u}(z)$,

$$\max_{j,l}D_{1jl}\leq\max_{j,l}\frac{1}{tn}\sum_{i=1}^{n}\mathbb{E}[|u_{ij}u_{il}|^2]\leq\max_{j,l}\frac{1}{tn}\sum_{i=1}^{n}\mathbb{E}[\bar{u}_i^4]\leq\frac{c^4}{t}.$$

Similarly,

$$\max_{j,l} D_{2jl} \le \max_{j,l} \frac{1}{tn} \sum_{i=1}^{n} \bar{u}_i^4 \le \frac{c^4 + o_p(1)}{t}.$$

where $\frac{1}{n}\sum_{i=1}^{n} \bar{u}_i^4 \to \mathbb{E}[\bar{u}_1^4] \le c^4$ almost surely by the Strong Law of Large Number for Markov Chains. Note that $|\mathcal{T}_t[W] - \mathbb{E}\mathcal{T}_t[W]| \le 2t$. Applying the Bernstein's inequality for Markov chains (Jiang et al., 2018, Theorem 1.1) yields

$$\mathbb{P}\left(D_{3jl} > \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2 \cdot \frac{1+\gamma}{1-\gamma} \cdot V_{n,t} + 10t\epsilon}\right),$$

with

$$V_{n,t} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\{\mathcal{T}_t[u_{ij}u_{ik}]\} \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[u_{ij}^2 u_{ik}^2] \le c^4.$$

The union bound of all pairs of $1 \le j, l \le p$ gives that

$$\mathbb{P}\left(\max_{j,k} D_{3jk} > \epsilon\right) \le 2p^2 \exp\left(-\frac{n\epsilon^2}{2 \cdot \frac{1+\gamma}{1-\gamma} \cdot c^4 + 10t\epsilon}\right).$$

Let $t = \frac{c^2}{30}\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{n}{\log p}}$, and $\epsilon = 3c^2\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log p}{n}}$. Then

$$\max_{j,k} D_{3jk} \le 3c^2\sqrt{\frac{1+\gamma}{1-\gamma} \cdot \frac{\log p}{n}}$$

with probability at least $1 - 2/p$. Putting upper bounds for $\max_{j,l} D_{1jl}$, $\max_{j,l} D_{2jl}$ and $\max_{j,l} D_{3jl}$ together completes the proof. □

B Technical Proofs for Bayesian Sparse Regression

This appendix details the proof of Theorem 2. Throughout the proof, let $\mathbb{P}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}$ and $\widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}$ denote the probability measures associated with the data generating processes $\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$ and $\mathbf{Y} = \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$, respectively.

B.1 Proof of (11)

Suppose Assumption 5 holds. For any model ξ of size at most $(1 + M_0)s$,

$$\|\widehat{\mathbf{U}}_\xi - \mathbf{U}_\xi\| \le \|\widehat{\mathbf{U}}_\xi - \mathbf{U}_\xi\|_{\mathrm{F}} \le \sqrt{(1+M_0)s}\max_{j=1}^{p}\|\widehat{\mathbf{U}}_j - \mathbf{U}_j\| \le L_4\sqrt{(1+M_0)s\log p}.$$

implying

$$\min_{\xi:\ |\xi|\le(1+M_0)s} \lambda_{\min}(\widehat{\mathbf{U}}_\xi^{\mathrm{T}}\widehat{\mathbf{U}}_\xi/n) \ge \left(\kappa_0 - L_4\sqrt{(1+M_0)s\log p/n}\right)^2 \gtrsim \kappa_0^2/4,$$
$$\max_{j=1}^{p}\|\widehat{\mathbf{U}}_j/\sqrt{n}\| \le \max_{j=1}^{p}\|\mathbf{U}_j/\sqrt{n}\| + \max_{j=1}^{p}\|\widehat{\mathbf{U}}_j/\sqrt{n} - \mathbf{U}_j/\sqrt{n}\| \le \kappa_1 + L_4\sqrt{\log p/n} \lesssim 2\kappa_1.$$

The last bound is derived as follows.

$$\begin{aligned}\|(\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| &\le \|\widehat{\mathbf{F}}\mathbf{H} - \mathbf{F}\|_{\mathrm{F}}\|\boldsymbol{\alpha}^\star\| + \max_{j=1}^{p}\|\widehat{\mathbf{U}}_j - \mathbf{U}_j\|\|\boldsymbol{\beta}^\star\|_1 \\ &\le L_3\|\boldsymbol{\alpha}^\star\|\sqrt{\log p} + L_4\|\boldsymbol{\beta}^\star\|\sqrt{s\log p} \le L_5\sigma^\star\sqrt{n}\epsilon_n.\end{aligned}$$

B.2 Proof of (12)

Let $\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the density function of the multivariate normal distribution. Using a change-of-measure trick and Cauchy-Schwarz inequality, write

$$
\begin{aligned}
&\mathbb{P}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_1|\mathbf{F},\mathbf{U},\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{H}) \\
&= \int 1\left\{\widehat{\pi}(A^c|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{y}) \geq e^{-C_1 s\log p}\right\}\mathcal{N}(\mathbf{y}|\mathbf{F}\boldsymbol{\alpha}^\star+\mathbf{U}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})d\mathbf{y} \\
&= \int 1\left\{\widehat{\pi}(A^c|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{y}) \geq e^{-C_1 s\log p}\right\}\frac{\mathcal{N}(\mathbf{y}|\mathbf{F}\boldsymbol{\alpha}^\star+\mathbf{U}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}{\mathcal{N}(\mathbf{y}|\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}\times\mathcal{N}(\mathbf{y}|\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})d\mathbf{y} \\
&\leq \left[\int 1^2\left\{\widehat{\pi}(A^c|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{y}) \geq e^{-C_1 s\log p}\right\}\mathcal{N}(\mathbf{y}|\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})d\mathbf{y}\right]^{1/2} \\
&\quad\times\left[\int\left(\frac{\mathcal{N}(\mathbf{y}|\mathbf{F}\boldsymbol{\alpha}^\star+\mathbf{U}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}{\mathcal{N}(\mathbf{y}|\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}\right)^2\mathcal{N}(\mathbf{y}|\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})d\mathbf{y}\right]^{1/2}, \\
&= \left[\widehat{\mathbb{P}}_{(\sigma^\star,\mathbf{H}\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}(\mathcal{E}_1|\widehat{\mathbf{F}},\widehat{\mathbf{U}})\right]^{1/2}\times\exp\left(\|(\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star)-(\mathbf{F}\boldsymbol{\alpha}^\star+\mathbf{U}\boldsymbol{\beta}^\star)\|^2/2\sigma^{\star 2}\right)
\end{aligned}
$$

The second term is bounded by $e^{L_5^2 s\log p/2}$, due to the last bound of (11). This concludes the first bound of (12). Other bounds of (12) are proven similarly.

B.3 Proof of (13)

The below theorem concern the estimation error rate, the prediction error rate and the model selection consistency of Bayesian sparse regression for the data generating process $\mathbf{Y} = \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$ with fixed design $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ and true parameters $(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$. Substituting $\boldsymbol{\alpha}^\star$, κ_0, κ_1, M_2, M_3, M_4 in this theorem with $\mathbf{H}\boldsymbol{\alpha}^\star$, $\kappa_0/2$, $2\kappa_1$, $M_2/2$, $M_3/2$, $M_4/2$, respectively, proves (13).

Theorem 3 (Bayesian Factor-adjusted Sparse Regression with Fixed Design)**.** *Consider data generating process* $\mathbf{Y} = \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$ *with fixed design* $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ *and true parameters* $(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$*. Let* $\widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}$ *denote the probability measure associated with the data generating process. Suppose*

$$
\begin{aligned}
&\widehat{\mathbf{F}}^{\mathrm{T}}\widehat{\mathbf{F}}/n = \mathbf{I}, \quad \widehat{\mathbf{F}}^{\mathrm{T}}\widehat{\mathbf{U}}/n = \mathbf{0} \\
&\min_{\xi:\ |\xi|\leq(M_0+1)s}\lambda_{\min}(\widehat{\mathbf{U}}_\xi^{\mathrm{T}}\widehat{\mathbf{U}}_\xi/n) \geq \kappa_0^2 \\
&\max_{j=1}^{p}\|\widehat{\mathbf{U}}_j\| \leq \kappa_1 \\
&\|\boldsymbol{\alpha}^\star\| \preccurlyeq 1, \quad \|\boldsymbol{\beta}^\star\| \preccurlyeq 1.
\end{aligned}
\tag{15}
$$

The following statements hold with some constants M_1*,* M_2*,* M_3*,* M_4*.*

(a) (Estimation Error) For any constants C_1, C_1' *such that* $C_1 + C_1' < M_0 - 2$*,*

$$
\widehat{\pi}(A^c(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star,M_0,M_1,M_2,M_3,\epsilon_n)|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{Y}) \geq e^{-C_1 s\log p}
$$

with $\widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}$*-probability at most* $e^{-C_1' s\log p}$*.*

(b) (prediction error rate) For any constants C_2, C_2' *such that* $C_2 + C_2' < M_0 - 2$*,*

$$
\widehat{\pi}(\|(\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta})-(\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star)\| > M_4\sigma^\star\sqrt{n}\epsilon_n|\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{Y}) \geq e^{-C_2 s\log p}.
$$

with $\widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}$*-probability at most* $e^{-C_2' s\log p}$*.*

(c) *(model selection consistency) Suppose* $\min_{j\in\xi^\star}|\beta_j^\star| \geq \sqrt{8M_0}\sigma_\star\epsilon_n/\kappa_0$ *in addition. For any constants* C_3, C_3' *such that* $C_3 + C_3' < M_0 - 2$,

$$\widehat{\pi}(A^c(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n) \cup \{\xi \not\supseteq \xi^\star\} | \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) \geq e^{-C_3 s \log p}$$

with $\widehat{\mathbb{P}}_{(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)}$*-probability at most* $e^{-C_3' s \log p}$.

Next lemma, borrowed from Barron (1998, Lemma 6) and Song and Liang (2017, Lemma A4), is the central technique to prove Theorem 3.

Lemma B1. *Consider a parametric model* $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. *Let* Θ_{0n} *and* Θ_n *be two subsets of the parameter space. Let* $\{\mathcal{D}_n\}_{n\geq 1}$ *be a sequence of data generations according to true parameter* $\boldsymbol{\theta}^\star$. *Let* $\pi(\boldsymbol{\theta})$ *be a prior distribution over the parameter space. If*

(1) $\pi(\Theta_{0n}) \leq \delta_{0n}$,

(2) *there exists a test function* $\phi_n(\mathcal{D}_n)$ *such that*

$$\sup_{\boldsymbol{\theta}\in\Theta_n} \mathbb{E}_{\boldsymbol{\theta}}(1 - \phi_n) \leq \delta_{1n}, \quad \mathbb{E}_{\boldsymbol{\theta}^\star}\phi_n \leq \delta_{1n}',$$

(3) *and*

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(\frac{\int \pi(\boldsymbol{\theta}) P_{\boldsymbol{\theta}}(\mathcal{D}_n) d\boldsymbol{\theta}}{P_{\boldsymbol{\theta}^\star}(\mathcal{D}_n)} \leq \delta_{2n}\right) \leq \delta_{2n}',$$

then for any δ_{3n},

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(\pi(\Theta_{0n} \cup \Theta_n | \mathcal{D}_n) \geq \frac{\delta_{0n} + \delta_{1n}}{\delta_{2n}\delta_{3n}}\right) \leq \delta_{1n}' + \delta_{2n}' + \delta_{3n}.$$

The intuition of this lemma is that any less preferred parameter guess $\boldsymbol{\theta} \in \Theta_{0n} \cup \Theta_n$ should either excluded by the prior (for $\boldsymbol{\theta} \in \Theta_{0n}$) or distinguished from the true parameter $\boldsymbol{\theta}^\star$ by a uniformly powerful test ϕ_n (for $\boldsymbol{\theta} \in \Theta_n$). We are going to set up suitable Θ_n and ϕ_n for each part of Theorem 3 and apply Lemma B1.

Lemmas B2 to B5 are useful to verify three conditions of Lemma B1 in the setup of Theorem 3. Lemma B2 is a novel tail probability bound for the Binomial distribution taken from Pelekis (2016, Theorem 1). Proving Lemmas B3 and B4 takes a substantial amount of work. We postpone their proofs to the next subsection.

Lemma B2 (Theorem 1.1 of Pelekis (2016))**.** *For random variable* $Z \sim \text{Binomial}(p, q)$, *if* $pq < t \leq p-1$ *then*

$$\mathbb{P}(Z \geq t) \leq \frac{\mu^{2(\widetilde{t}+1)}}{2} \binom{p}{\widetilde{t}+1} \Big/ \binom{t}{\widetilde{t}+1},$$

where $\widetilde{t} = \lfloor (t - pq)/(1 - q) \rfloor < m$.

Lemma B3. *Let* o *stand for any small constant. In the setup of Theorem 3, the following statements hold.*

(a) *Let*

$$\Theta_{1n} = \left\{ (\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \begin{array}{c} |\xi \setminus \xi^\star| \leq M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \notin \left[\dfrac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n}, \dfrac{1 + M_1\epsilon_n}{1 - M_1\epsilon_n}\right] \end{array} \right\},$$

$$\phi_{1n} = 1\left\{ \max_{\xi:\ |\xi\setminus\xi^\star| \leq M_0 s} \left| \mathbf{Y}^{\mathrm{T}} \left[\mathbf{I} - \widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n - \widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\right] \mathbf{Y}/n\sigma^{\star 2} - 1 \right| \geq M_1\epsilon_n \right\},$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{1n} \lesssim \exp(-(M_1^2/8 - M_0 - o)s\log p),$$
$$\sup_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{1n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{1n}) \lesssim \exp(-(M_1^2/8 - o)s\log p).$$

(b) Let

$$\Theta_{2n} = \left\{ (\sigma^2,\boldsymbol{\alpha},\boldsymbol{\beta}) : \begin{array}{c} |\xi\setminus\xi^\star| \le M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \in \left[\dfrac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \dfrac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right], \\ \|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star\| > M_2\sigma^\star\epsilon_n, \end{array} \right\},$$
$$\phi_{2n} = 1\left\{ \|\widehat{\mathbf{F}}^{\mathrm{T}}\mathbf{Y}/n - \boldsymbol{\alpha}^\star\| \ge M_2\sigma^\star\epsilon_n/2 \right\},$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{2n} \lesssim \exp(-(M_2^2/8 - o)s\log p),$$
$$\sup_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{2n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{2n}) \lesssim \exp(-(M_2^2/8 - o)s\log p).$$

(c) Let

$$\Theta_{3n} = \left\{ (\sigma^2,\boldsymbol{\alpha},\boldsymbol{\beta}) : \begin{array}{c} |\xi\setminus\xi^\star| \le M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \in \left[\dfrac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \dfrac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right], \\ \|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star\| \le M_2\sigma^\star\epsilon_n, \\ \|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\| > M_3\sigma^\star\epsilon_n/\kappa_0 \end{array} \right\},$$
$$\phi_{3n} = 1\left\{ \max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s} \|\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\mathbf{Y} - \boldsymbol{\beta}^\star_{\xi\cup\xi^\star}\| \ge M_3\sigma^\star\epsilon_n/2\kappa_0 \right\}.$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} \lesssim \exp(-(M_3^2/8 - M_0 - o)s\log p),$$
$$\sup_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{3n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}) \lesssim \exp(-(M_3^2/8 - o)s\log p).$$

(d) Let

$$\Theta_{4n} = \left\{ (\sigma^2,\boldsymbol{\alpha},\boldsymbol{\beta}) : \begin{array}{c} |\xi\setminus\xi^\star| \le M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \in \left[\dfrac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \dfrac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right] \\ \|(\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta}) - (\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star)\| > M_4\sigma^\star\sqrt{n}\epsilon_n \end{array} \right\},$$
$$\phi_{4n} = 1\left\{ \max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s} \left\| \left[\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\dagger + \widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\right]\mathbf{Y} - \left(\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star\right) \right\| \ge M_4\sigma^\star\sqrt{n}\epsilon_n/2 \right\},$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{4n} \lesssim \exp(-(M_4^2/8 - M_0 - o)s\log p),$$
$$\sup_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{4n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{4n}) \lesssim \exp(-(M_4^2/8 - o)s\log p).$$

(e) Suppose $\min_{j\in\xi^\star}|\beta_j^\star| \geq M_5\sigma^\star\epsilon_n/\kappa_0$ *in addition. Let*

$$\Theta_{5n} = \left\{ (\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \begin{array}{c} |\xi \setminus \xi^\star| \leq M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \in \left[\dfrac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \dfrac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right] \\ \|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star\| \leq M_2\sigma^\star\epsilon_n, \\ \xi \not\supseteq \xi^\star \end{array} \right\},$$

$$\phi_{5n} = 1\left\{ \min_{\xi\not\supseteq\xi^\star:\ |\xi\setminus\xi^\star|\leq M_0 s} \left\| \left(\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi\widehat{\mathbf{U}}_\xi^\dagger\right)\mathbf{Y}\right\| \leq M_5\sigma^\star\sqrt{n}\epsilon_n/2 \right\},$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{5n} \lesssim \exp(-(M_5^2/8-o)s\log p),$$
$$\sup_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{5n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{5n}) \lesssim \exp(-(M_5^2/8-o)s\log p).$$

(f) Let

$$\Theta_{6n} = \left\{ (\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \begin{array}{c} |\xi \setminus \xi^\star| \leq M_0 s, \\ \dfrac{\sigma^2}{\sigma^{\star 2}} \in \left[\dfrac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \dfrac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right] \\ \xi \supseteq \xi^\star, \\ \|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star\| \leq M_2\sigma^\star\epsilon_n, \\ \|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\| > M_3\sigma^\star\epsilon_n/\kappa_0 \end{array} \right\},$$

$$\phi_{6n} = 1\left\{ \max_{\xi\supseteq\xi^\star:\ |\xi\setminus\xi^\star|\leq M_0 s} \|\widehat{\mathbf{U}}_\xi^\dagger\mathbf{Y} - \boldsymbol{\beta}_\xi^\star\| \geq M_3\sigma^\star\epsilon_n/2\kappa_0 \right\},$$

then

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{6n} \lesssim \exp(-(M_3^2/8-M_0-o)s\log p),$$
$$\sup_{(\sigma,\alpha,\beta)\in\Theta_{6n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{6n}) \lesssim \exp(-(M_3^2/8-o)s\log p).$$

Lemma B4. *In the setup of Theorem 3, for any constants* $C_4 > 2$ *and* $C_4' > 0$,

$$\widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\int \frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})} d\pi(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta}) \leq e^{-C_4 s\log p}\right) \lesssim e^{-C_4' s\log p}.$$

Lemma B5. *For parameter subspaces* Θ_j, $j=1,\ldots,m$ *and test functions* ϕ_j, $j=1,\ldots,m$,

$$\sup_{\theta\in\cup_{j=1}^m\Theta_j} \mathbb{E}_\theta\left(1-\max_{j=1}^m\phi_j\right) \leq \max_{j=1}^m\left\{\sup_{\theta\in\Theta_j}\mathbb{E}_\theta(1-\phi_j)\right\}.$$

Proof of Lemma B5.

$$\begin{aligned}
\sup_{\theta\in\cup_{j=1}^m\Theta_j} \mathbb{E}_\theta\left(1-\max_{j=1}^m\phi_j\right) &= \max_{j=1}^m\left\{\sup_{\theta\in\Theta_j}\mathbb{E}_\theta\left(1-\max_{k=1}^m\phi_k\right)\right\} \\
&= \max_{j=1}^m\left\{\sup_{\theta\in\Theta_j}\mathbb{E}_\theta\left(\min_{k=1}^m(1-\phi_k)\right)\right\} \\
&\leq \max_{j=1}^m\left\{\sup_{\theta\in\Theta_j}\mathbb{E}_\theta\left(1-\phi_j\right)\right\}.
\end{aligned}$$

□

Proof of Theorem 3(a). Verify the three conditions of Lemma B1 with

$$\Theta_{0n} = \{(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) : |\xi \setminus \xi^\star| > M_0 s\}, \quad \Theta_n = \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{3n}, \quad \phi_n = \max\{\phi_{1n}, \phi_{2n}, \phi_{3n}\},$$

where $\Theta_{1n}, \Theta_{2n}, \Theta_{3n}, \phi_{1n}, \phi_{2n}, \phi_{3n}$ are defined in Lemma B3(a)(b)(c). Evidently,

$$\Theta_{0n} \cup \Theta_n = \Theta_{0n} \cup \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{3n} = A^c(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n).$$

Applying Lemma B2 yields that

$$\pi(\Theta_{0n}) \le \pi(|\xi| > M_0 s) \lesssim \frac{1}{2}\left(\frac{s_0}{p}\right)^{2(M_0 s - s_0 + 1)} \binom{p}{M_0 s - s_0 + 1} \le \frac{p^{-(M_0 s - s_0 + 1)}}{2} \lesssim \delta_{0n} = e^{-(M_0 - o) s \log p}.$$

From Lemma B3(a)(b)(c) and Lemma B5, it follows that

$$\sup_{\Theta_n} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_n) \le \max_{i=1,2,3} \sup_{\Theta_{in}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_{in}) \le \delta_{1n} = e^{-(\min\{M_1^2, M_2^2, M_3^2\}/8 - o) s \log p},$$
$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_n \le \sum_{i=1,2,3} \widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{in} \le \delta'_{1n} = e^{-\min\{M_1^2, M_2^2 + 8M_0, M_3^2\}/8 - M_0 - o) s \log p}.$$

By Lemma B4, the third condition in Lemma B1 hold with

$$\delta_{2n} = e^{-C_4 s \log p}, \quad \delta'_{2n} = e^{-C'_4 s \log p}$$

for any $C_4 > 2$ and $C'_4 > 0$. Setting sufficiently large M_1, M_2, M_3, C'_4 and suitable C'_1, C_4, δ_{3n} such that

$$\frac{\delta_{0n} + \delta_{1n}}{\delta_{2n}\delta_{3n}} \le e^{-(M_0 - C_4 - C'_1) s \log p}, \quad \delta'_{1n} + \delta'_{2n} + \delta_{3n} \le e^{-C'_1 s \log p}$$

completes the proof. □

Proof of Theorem 3(b). Verify the three conditions of Lemma B1 with

$$\Theta_n = \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{4n}, \quad \phi_n = \max\{\phi_{1n}, \phi_{2n}, \phi_{4n}\},$$

where $\Theta_{1n}, \Theta_{2n}, \Theta_{4n}, \phi_{1n}, \phi_{2n}, \phi_{4n}$ are defined in Lemma B3(a)(b)(d). Evidently,

$$\Theta_{0n} \cup \Theta_n = \Theta_{0n} \cup \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{4n} \supseteq \{(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \|(\widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}) - (\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star)\| > M_4 \sigma^\star \sqrt{n} \epsilon_n\}.$$

From Lemma B3(a)(b)(d) and Lemma B5, it follows that

$$\sup_{\Theta_n} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_n) \le \max_{i=1,2,4} \sup_{\Theta_{in}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_{in}) \le \delta_{1n} := e^{-(\min\{M_1^2, M_2^2, M_4^2\}/8 - o) s \log p}$$
$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_n \le \widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{1n} + \widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{4n}, \le \delta'_{1n} := e^{-(\min\{M_1^2, M_2^2 + M_0, M_4^2\}/8 - M_0 - o) s \log p}.$$

The other two conditions of Lemma B1 have been verified in the proof of part (a). □

Proof of Theorem 3(c). Verify the three conditions of Lemma B1 with

$$\Theta_n = \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{5n} \cup \Theta_{6n}, \quad \phi_n = \max\{\phi_{1n}, \phi_{2n}, \phi_{5n}, \phi_{6n}\},$$

where $\Theta_{1n}, \Theta_{2n}, \Theta_{5n}, \Theta_{6n}, \phi_{1n}, \phi_{2n}, \phi_{5n}, \phi_{6n}$ are defined in Lemma B3(a)(b)(e)(f). Evidently,

$$\Theta_{0n} \cup \Theta_n = \Theta_{0n} \cup \Theta_{1n} \cup \Theta_{2n} \cup \Theta_{5n} \cup \Theta_{6n} = A^c(\sigma^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star, M_0, M_1, M_2, M_3, \epsilon_n) \cup \{\xi \not\supseteq \xi^\star\}.$$

From Lemma B3(a)(b)(e)(f) and Lemma B5, it follows that

$$\sup_{\Theta_n} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_n) \le \max_{i=1,2,5,6} \sup_{\Theta_{in}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1 - \phi_{in}) \le \delta_{1n} := e^{-\min\{M_1^2, M_2^2, M_3^2, M_5^2\}/8 - o) s \log p}$$
$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_n \le \sum_{i=1,2,5,6} \widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{in} \le \delta'_{1n} := e^{-(\min\{M_1^2, M_2^2 + 8M_0, M_3^2, M_5^2 + 8M_0\}/8 - M_0 - o) s \log p}.$$

The other two conditions of Lemma B1 have been verified in the proof of part (a). □

B.4 Technical Proofs of Lemmas

The proofs of Lemmas B3 and B4 use two preliminary results as follows.

Lemma B6 (Probability bounds of chi-squared random variables)**.** *Let χ^2_d be a chi-squared random variable of degree d, and o stands for any small constant.*

(a) For any ϵ_n such that $n\epsilon_n > d_n$,

$$\mathbb{P}(\chi^2_{n-d_n}/n \geq 1+\epsilon_n) \leq e^{-\min\left\{\frac{(n\epsilon_n+d_n)^2}{8(n-d_n)}, \frac{n\epsilon_n+d_n}{8}\right\}},$$
$$\mathbb{P}(\chi^2_{n-d_n}/n \leq 1-\epsilon_n) \leq e^{-\min\left\{\frac{(n\epsilon_n-d_n)^2}{8(n-d_n)}, \frac{n\epsilon_n-d_n}{8}\right\}},$$

In addition, if $\epsilon_n \to 0$ but $n\epsilon_n \succ d_n$,

$$\mathbb{P}(\chi^2_{n-d_n}/n \geq 1+\epsilon_n) \lesssim e^{-(1/8-o)n\epsilon_n^2}$$
$$\mathbb{P}(\chi^2_{n-d_n}/n \geq 1+\epsilon_n) \lesssim e^{-(1/8-o)n\epsilon_n^2}.$$

(b)

$$\mathbb{P}(\chi^2_{d_n} \geq t_n) \leq e^{-\left(\sqrt{2t_n-d_n}-\sqrt{d_n}\right)^2/4}.$$

In addition, if $t_n \succ d_n$ then for any $\widetilde{t}_n$ such that $\widetilde{t}_n/t_n \to 1$

$$\mathbb{P}(\chi^2_{d_n} \geq t_n) \lesssim e^{-(1/2-o)\widetilde{t}_n}.$$

Proof. For part (a), the first assertion follows from the sub-exponential tail of chi-squared distributions, and the second assertion is due to

$$(1/8-o)n\epsilon_n^2 \lesssim \frac{(n\epsilon_n+d_n)^2}{8(n-d_n)} \lesssim \frac{n\epsilon_n+d_n}{8}$$
$$(1/8-o)n\epsilon_n^2 \lesssim \frac{(n\epsilon_n-d_n)^2}{8(n-d_n)} \lesssim \frac{n\epsilon_n-d_n}{8}$$

For part (b), the first assertion is a corollary of Laurent and Massart (2000, Lemma 1), and the second assertion follows from

$$(1/2-o)\widetilde{t}_n \lesssim \left(\sqrt{2t_n-d_n}-\sqrt{d_n}\right)^2/4.$$

□

Lemma B7 (Lancaster and Tismenetsky (1985, p. 294))**.** *Suppose a $p \times p$ symmetric matrix $\mathbf{S}$ has the partitioned form*

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix},$$

where $\mathbf{S}_{11}$ is a non-singular principal submatrix of $\mathbf{S}$. Then

$$\lambda_{\min}(\mathbf{S}_{22}-\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}) \geq \lambda_{\min}(\mathbf{S}).$$

Proof of Lemma B3(a). Under the null hypothesis,

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{1n} = \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\leq M_0 s} |\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^{\dagger}_{\xi\cup\xi^\star})\boldsymbol{\varepsilon}/n-1| \geq M_1\epsilon_n\right).$$

Projection matrices $\widehat{\mathbf{U}}_{\xi'\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi'\cup\xi^\star} \le \widehat{\mathbf{U}}_{\xi''\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi''\cup\xi^\star}$ for nested models $\xi' \subseteq \xi''$, and thus the term $\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon}$ achieves its maximum value at some ξ with $|\xi| = M_0 s$ and $\xi \setminus \xi^\star = \emptyset$ and its minimum value at $\xi = \emptyset$. Thus

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{1n} \le & \sum_{\xi:\ |\xi|=M_0 s,\ \xi\setminus\xi^\star=\emptyset} \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star})\boldsymbol{\varepsilon}/n \le 1-M_1\epsilon_n\right)\\ & + \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi^\star})\boldsymbol{\varepsilon}/n \ge 1+M_1\epsilon_n\right)\\ = & \binom{p-s}{M_0 s}\mathbb{P}\left(\chi^2_{n-k-(1+M_0)s}/n \le 1-M_1\epsilon_n\right)+\mathbb{P}\left(\chi^2_{n-k-s}/n \ge 1+M_1\epsilon_n\right).\end{aligned}$$

Applying Lemma B6(a) yields

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{1n} \lesssim (p^{M_0 s}+1)\times e^{-(M_1^2/8-o)n\epsilon_n^2} \lesssim e^{-(M_1^2/8-M_0-o)n\epsilon_n^2}.$$

Under the alternative hypothesis, write $\phi_{1n} = \max_{\xi':\ |\xi'\setminus\xi^\star|\le M_0 s}\ \phi_{1n}^{\xi'}$ with

$$\phi_{1n}^{\xi'} = 1\left\{\left|\mathbf{Y}^{\mathrm{T}}\left[\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\right]\mathbf{Y}/n\sigma^{\star 2}-1\right| \ge M_1\epsilon_n\right\},$$

then, by Lemma B5,

$$\sup_{\Theta_{1n}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{1n}) \le \max_{\xi':\ |\xi'\setminus\xi^\star|\le M_0 s}\ \sup_{\Theta_{1n}\cap\{\xi=\xi'\}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{1n}^{\xi'}).$$

On each partition $\Theta_{1n}\cap\{\xi=\xi'\}$ of Θ_{1n}, due to the restriction $\frac{\sigma^2}{\sigma^{\star 2}} \notin \left[\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}, \frac{1+M_1\epsilon_n}{1-M_1\epsilon_n}\right]$ of Θ_{1n},

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{1n}^{\xi'}) &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\left|\boldsymbol{\varepsilon}^{\mathrm{T}}\left[\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\right]\boldsymbol{\varepsilon}/n\times(\sigma^2/\sigma^{\star 2})-1\right| < M_1\epsilon_n\right)\\ &\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\left[\mathbf{I}-\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n-\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\right]\boldsymbol{\varepsilon}/n \notin [1-M_1\epsilon_n, 1+M_1\epsilon_n]\right)\\ &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\chi^2_{n-k-|\xi\cup\xi^\star|}/n \notin [1-M_1\epsilon_n, 1+M_1\epsilon_n]\right)\\ &\le \mathbb{P}\left(\chi^2_{n-k-(1+M_0)s}/n < 1-M_1\epsilon_n\right)+\mathbb{P}\left(\chi^2_{n-k-s}/n > 1+M_1\epsilon_n\right)\end{aligned}$$

This bound holds for any ξ' such that $|\xi'\setminus\xi^\star| \le M_0 s$ and any $(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Theta_{1n}\cap\{\xi=\xi'\}$. Applying Lemma B6(a) yields

$$\sup_{\Theta_{1n}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{1n}) \lesssim e^{-(M_1^2/8-o)n\epsilon_n^2}.$$

□

Proof of Lemma B3(b). Under the null hypothesis,

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{2n} = \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\|\widehat{\mathbf{F}}^{\mathrm{T}}\boldsymbol{\varepsilon}/n\| \ge M_2\epsilon_n/2\right) = \mathbb{P}\left(\chi^2_k \ge M_2^2 n\epsilon_n^2/4\right) \le e^{-(M_2^2/8-o)n\epsilon_n^2},$$

where the last step uses Lemma B6(b). Under the alternative hypothesis,

$$\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{2n}) = \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star+\sigma\widehat{\mathbf{F}}^{\mathrm{T}}\boldsymbol{\varepsilon}/n\| < M_2\sigma^\star\epsilon_n/2\right).$$

Using the restrictions $\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^\star\| > M_2\sigma^\star\epsilon_n$ and $\frac{\sigma^{\star 2}}{\sigma^2} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}$ of Θ_{2n} and Lemma B6(b),

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{2n}) &\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|\widehat{\mathbf{F}}^{\mathrm{T}}\boldsymbol{\varepsilon}/n\| > \sqrt{\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}}\times M_2\epsilon_n/2\right)\\ &= \mathbb{P}\left(\chi^2_k > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_2^2 n\epsilon_n^2/4\right) \lesssim e^{-(M_2^2/8-o)n\epsilon_n^2}.\end{aligned}$$

□

Proof of Lemma B3(c). Under the null hypothesis, write

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} &= \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s} \|\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon}\| \ge M_3\epsilon_n/2\kappa_0\right)\\ &= \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s} \boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}^{\dagger\mathrm{T}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon} \ge M_3^2\epsilon_n^2/4\kappa_0^2\right)\end{aligned}$$

Using the fact that $\widehat{\mathbf{U}}^{\dagger\mathrm{T}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star} \le \left(n\kappa_0^2\right)^{-1}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}$,

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} \le \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s} \boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon} \ge M_3^2 n\epsilon_n^2/4\right).$$

Projection matrices $\widehat{\mathbf{U}}_{\xi'\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi'\cup\xi^\star} \le \widehat{\mathbf{U}}_{\xi''\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi''\cup\xi^\star}$ for nested models $\xi'\subseteq\xi''$, and thus the term $\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon}$ achieves its maximum value at some ξ with $|\xi|=M_0 s$ and $\xi\setminus\xi^\star=\emptyset$. Thus

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} &\le \sum_{\xi:\ |\xi|=M_0 s,\ \xi\setminus\xi^\star=\emptyset} \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon} \ge M_3^2 n\epsilon_n^2/4\right)\\ &= \binom{p-s}{M_0 s}\mathbb{P}\left(\chi^2_{(1+M_0)s} \ge M_3^2 n\epsilon_n^2/4\right)\end{aligned}$$

Applying Lemma B6(b) yields

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} \lesssim p^{M_0 s}e^{-(M_3^2/8-o)n\epsilon_n^2} = e^{-(M_3^2/8-M_0-o)n\epsilon_n^2}.$$

Under the alternative hypothesis, write $\phi_{3n} = \max_{\xi':\ |\xi'\setminus\xi^\star|\le M_0 s}\phi_{3n}^{\xi'}$ with

$$\phi_{3n}^{\xi'} = 1\left\{\|\widehat{\mathbf{U}}^\dagger_{\xi'\cup\xi^\star}\mathbf{Y} - \boldsymbol{\beta}^\star_{\xi'\cup\xi^\star}\| \ge M_3\sigma^\star\epsilon_n/2\kappa_0\right\},$$

then, by Lemma B5,

$$\sup_{\Theta_{3n}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}) \le \max_{\xi':\ |\xi'\setminus\xi^\star|\le M_0 s}\ \sup_{\Theta_{3n}\cap\{\xi=\xi'\}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}^{\xi'}).$$

On each partition $\Theta_{3n}\cap\{\xi=\xi'\}$ of Θ_{3n}, due to the constraints $\|\boldsymbol{\beta}_{\xi\cup\xi^\star}-\boldsymbol{\beta}^\star_{\xi\cup\xi^\star}\| = \|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\| > M_3\sigma^\star\epsilon_n/\kappa_0$ and $\frac{\sigma^{\star 2}}{\sigma^2} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}$ of Θ_{3n},

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}^{\xi'}) &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|\boldsymbol{\beta}_{\xi\cup\xi^\star}-\boldsymbol{\beta}^\star_{\xi\cup\xi^\star}+\sigma\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon}\| < M_3\sigma^\star\epsilon_n/2\kappa_0\right)\\ &\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon}\| > \sqrt{\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}}\times M_3\epsilon_n/2\kappa_0\right)\\ &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}^{\dagger\mathrm{T}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_3^2\epsilon_n^2/4\kappa_0^2\right)\end{aligned}$$

Using the fact that $\widehat{\mathbf{U}}^{\dagger\mathrm{T}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star} \le \left(n\kappa_0^2\right)^{-1}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}$ again,

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}^{\xi'}) &\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi\cup\xi^\star}\boldsymbol{\varepsilon} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_3^2\epsilon_n^2/4\kappa_0^2\right)\\ &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\chi^2_{|\xi\cup\xi^\star|} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_3^2 n\epsilon_n^2/4\right)\\ &\le \mathbb{P}\left(\chi^2_{(1+M_0)s} > \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_3^2 n\epsilon_n^2/4\right).\end{aligned}$$

This bound holds for any ξ' such that $|\xi'\setminus\xi^\star| \le M_0 s$ and any $(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{3n}\cap\{\xi=\xi'\}$. Applying Lemma B6(b) yields

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} \lesssim e^{-(M_3^2/8-o)n\epsilon_n^2}.$$

□

Proof of Lemma B3(d). Under the null hypothesis,

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{4n} &= \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s}\left\|\left[\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\right]\boldsymbol{\varepsilon}\right\|\ge M_4\sqrt{n}\epsilon_n/2\right)\\ &= \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\max_{\xi:\ |\xi\setminus\xi^\star|\le M_0 s}\boldsymbol{\varepsilon}^{\mathrm{T}}\left[\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\right]\boldsymbol{\varepsilon}\ge M_4^2 n\epsilon_n^2/4\right)\end{aligned}$$

Projection matrices $\widehat{\mathbf{U}}_{\xi'\cup\xi^\star}\widehat{\mathbf{U}}_{\xi'\cup\xi^\star}^{\dagger}\le\widehat{\mathbf{U}}_{\xi''\cup\xi^\star}\widehat{\mathbf{U}}_{\xi''\cup\xi^\star}^{\dagger}$ for nested models $\xi'\subseteq\xi''$, and thus the term $\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\boldsymbol{\varepsilon}$ achieves its maximum value at some ξ with $|\xi|=M_0 s$ and $\xi\setminus\xi^\star=\emptyset$. Thus

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{4n} &\le \sum_{\xi:\ |\xi|=M_0 s,\ \xi\setminus\xi^\star=\emptyset}\widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\left[\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\right]\boldsymbol{\varepsilon}\ge M_4^2 n\epsilon_n^2/4\right)\\ &= \binom{p-s}{M_0 s}\mathbb{P}\left(\chi^2_{k+(1+M_0)s}\ge M_4^2 n\epsilon_n^2/4\right)\end{aligned}$$

Applying Lemma B6(b) yields

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{4n}\lesssim p^{M_0 s}e^{-(M_4^2/8-o)n\epsilon_n^2}=e^{-(M_4^2/8-M_0-o)s\log p}.$$

Under the alternative hypothesis, write $\phi_{4n}=\max_{\xi':\ |\xi'\setminus\xi^\star|\le M_0 s}\phi_{4n}^{\xi'}$ with

$$\phi_{4n}^{\xi'}=1\left\{\left\|\left[\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}\right]\mathbf{Y}-\left(\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star\right)\right\|\ge M_4\sigma^\star\sqrt{n}\epsilon_n/2\right\},$$

then, by Lemma B5,

$$\sup_{\Theta_{4n}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{4n})\le\max_{\xi'\not\supseteq\xi^\star:\ |\xi'\setminus\xi^\star|\le M_0 s}\ \sup_{\Theta_{4n}\cap\{\xi=\xi'\}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{4n}^{\xi'}).$$

On each partition $\Theta_{4n}\cap\{\xi=\xi'\}$ of Θ_{4n}, due to the restrictions $\|(\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta})-(\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star)\|>M_4\sigma^\star\sqrt{n}\epsilon_n$, and $\frac{\sigma^{\star 2}}{\sigma^2}>\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}$ of Θ_{4n},

$$\begin{aligned}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{4n}^{\xi'}) &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|(\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta})-(\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star)+\sigma(\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger})\boldsymbol{\varepsilon}\|<M_4\sigma^\star\sqrt{n}\epsilon_n/2\right)\\ &\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\|(\widehat{\mathbf{F}}\widehat{\mathbf{F}}^{\mathrm{T}}/n+\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger})\boldsymbol{\varepsilon}\|\ge\sqrt{\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}}\times M_4\sqrt{n}\epsilon_n/2\right)\\ &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\chi^2_{k+|\xi\cup\xi^\star|}\ge\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_4^2 n\epsilon_n^2/4\right)\\ &\le \mathbb{P}\left(\chi^2_{k+(1+M_0)s}\ge\frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_4^2 n\epsilon_n^2/4\right).\end{aligned}$$

This bound holds for any ξ' such that $|\xi'\setminus\xi^\star|\le M_0 s$ and any $(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})\in\Theta_{4n}\cap\{\xi=\xi'\}$. Applying Lemma B6(b) yields

$$\sup_{\Theta_{4n}}\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{4n})\lesssim e^{-(M_4^2/8-o)n\epsilon_n^2}.$$

□

Proof of Lemma B3(e). Claim that

$$\min_{\xi\not\supseteq\xi^\star:\ |\xi\setminus\xi^\star|\le M_0 s}\left\|\left(\widehat{\mathbf{U}}_{\xi\cup\xi^\star}\widehat{\mathbf{U}}_{\xi\cup\xi^\star}^{\dagger}-\widehat{\mathbf{U}}_{\xi}\widehat{\mathbf{U}}_{\xi}^{\dagger}\right)\widehat{\mathbf{U}}_{\xi^\star}\boldsymbol{\beta}_{\xi^\star}^\star\right\|\ge M_5\sigma^\star\sqrt{n}\epsilon_n.\tag{16}$$

Indeed, for any $\xi \not\supseteq \xi^\star$ with $|\xi \setminus \xi^\star| \le M_0 s$,

$$\left\| \left(\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star} \boldsymbol{\beta}^\star_{\xi^\star} \right\|^2 = \left\| \left(\mathbf{I} - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star} \boldsymbol{\beta}^\star_{\xi^\star} \right\|^2 = \left\| \left(\mathbf{I} - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star\setminus\xi} \boldsymbol{\beta}^\star_{\xi^\star\setminus\xi} \right\|^2$$
$$= \boldsymbol{\beta}^{\mathrm{T}}_{\xi^\star\setminus\xi} \widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star\setminus\xi} \left(\mathbf{I} - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star\setminus\xi} \boldsymbol{\beta}_{\xi^\star\setminus\xi}$$

Note that $\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star\setminus\xi} \left(\mathbf{I} - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star\setminus\xi}$ is the Schur complement of the principal submatrix $\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star\setminus\xi} \widehat{\mathbf{U}}_{\xi^\star\setminus\xi}$ in the matrix $\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}$. By Lemma B7,

$$\lambda_{\min} \left(\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star\setminus\xi} \left(\mathbf{I} - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \widehat{\mathbf{U}}_{\xi^\star\setminus\xi} \right) \ge \lambda_{\min} \left(\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star} \right) \ge n\kappa_0^2.$$

Putting the last two displays together proves (16). Under the null hypothesis, using (16), the fact that $\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \le \widehat{\mathbf{U}}_{\xi^\star} \widehat{\mathbf{U}}_{\xi^\star}^\dagger$ and Lemma B6(b),

$$\begin{aligned}
\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)} \phi_{5n} &= \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)} \left(\min_{\xi\not\supseteq\xi^\star:\ |\xi\setminus\xi^\star|\le M_0 s} \|(\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger)(\widehat{\mathbf{U}}_{\xi^\star} \boldsymbol{\beta}^\star_{\xi^\star} + \sigma^\star \boldsymbol{\varepsilon})\| \le M_5 \sigma^\star \sqrt{n} \epsilon_n / 2 \right) \\
&\le \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)} \left(\max_{\xi\not\supseteq\xi^\star:\ |\xi\setminus\xi^\star|\le M_0 s} \|(\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger)\boldsymbol{\varepsilon}\| \ge M_5 \sqrt{n} \epsilon_n / 2 \right) \\
&\le \widehat{\mathbb{P}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)} \left(\|\widehat{\mathbf{U}}_{\xi^\star} \widehat{\mathbf{U}}_{\xi^\star}^\dagger \boldsymbol{\varepsilon}\| \ge M_5 \sqrt{n} \epsilon_n / 2 \right) = \mathbb{P} \left(\chi^2_s \ge M_5^2 n \epsilon_n^2 / 4 \right) \lesssim e^{-(M_5^2/8 - o) n \epsilon_n^2}.
\end{aligned}$$

Under the alternative hypothesis, write $\phi_{5n} = \max_{\xi'\not\supseteq\xi^\star:\ |\xi'\setminus\xi^\star|\le M_0 s} \phi_{5n}^{\xi'}$ with

$$\phi_{5n}^{\xi'} = 1 \left\{ \left\| \left(\widehat{\mathbf{U}}_{\xi'\cup\xi^\star} \widehat{\mathbf{U}}_{\xi'\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_{\xi'} \widehat{\mathbf{U}}_{\xi'}^\dagger \right) \mathbf{Y} \right\| \le M_5 \sigma^\star \sqrt{n} \epsilon_n / 2 \right\},$$

then, by Lemma B5,

$$\sup_{\Theta_{5n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} (1 - \phi_{5n}) \le \max_{\xi'\not\supseteq\xi^\star:\ |\xi'\setminus\xi^\star|\le M_0 s} \sup_{\Theta_{5n}\cap\{\xi=\xi'\}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} (1 - \phi_{5n}^{\xi'}).$$

On each partition $\Theta_{5n} \cap \{\xi = \xi'\}$ of Θ_{5n}, using the restriction $\frac{\sigma^{\star 2}}{\sigma^2} > \frac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n}$ of Θ_{5n} and the fact that $\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \le \widehat{\mathbf{U}}_{\xi^\star} \widehat{\mathbf{U}}_{\xi^\star}^\dagger$ again,

$$\begin{aligned}
\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} (1 - \phi_{5n}^{\xi'}) &= \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} \left(\|(\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger)\sigma\boldsymbol{\varepsilon}\| > M_5 \sigma^\star \sqrt{n} \epsilon_n / 2 \right) \\
&\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} \left(\left\| \left(\widehat{\mathbf{U}}_{\xi\cup\xi^\star} \widehat{\mathbf{U}}_{\xi\cup\xi^\star}^\dagger - \widehat{\mathbf{U}}_\xi \widehat{\mathbf{U}}_\xi^\dagger \right) \boldsymbol{\varepsilon} \right\| > \sqrt{\frac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n}} \times M_5 \sqrt{n} \epsilon_n / 2 \right) \\
&\le \widehat{\mathbb{P}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} \left(\left\| \widehat{\mathbf{U}}_{\xi^\star} \widehat{\mathbf{U}}_{\xi^\star}^\dagger \boldsymbol{\varepsilon} \right\| > \sqrt{\frac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n}} \times M_5 \sqrt{n} \epsilon_n / 2 \right) \\
&= \mathbb{P} \left(\chi^2_s > \frac{1 - M_1\epsilon_n}{1 + M_1\epsilon_n} \times M_5^2 n \epsilon_n^2 / 4 \right).
\end{aligned}$$

This bound holds for any $\xi' \not\supseteq \xi^\star$ such that $|\xi' \setminus \xi^\star| \le M_0 s$ and any $(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) \in \Theta_{5n} \cap \{\xi = \xi'\}$. Applying Lemma B6(b) yields

$$\sup_{\Theta_{5n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})} (1 - \phi_{5n}) \lesssim e^{-(M_5^2/8 - o) n \epsilon_n^2}.$$

□

Proof of Lemma B3(f). Since $\phi_{6n} \le \phi_{3n}$,

$$\widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{6n} \le \widehat{\mathbb{E}}_{(\sigma^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)}\phi_{3n} \lesssim e^{-(M_3^2/8-M_0-o)s\log p}.$$

Under the alternative hypothesis, write $\phi_{6n} = \max_{\xi'\supseteq\xi^\star:\ |\xi'\setminus\xi^\star|\le M_0 s} \phi_{6n}^{\xi'}$ with

$$\phi_{6n}^{\xi'} = 1\left\{\|\widehat{\mathbf{U}}_{\xi'}^\dagger \mathbf{Y} - \boldsymbol{\beta}_{\xi'}^\star\| \ge M_3\sigma^\star\epsilon_n/2\kappa_0\right\},$$

then, by Lemma B5,

$$\sup_{\Theta_{6n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{6n}) \le \max_{\xi'\supseteq\xi^\star:\ |\xi'\setminus\xi^\star|\le M_0 s} \sup_{\Theta_{6n}\cap\{\xi=\xi'\}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{6n}^{\xi'}).$$

On each partition $\Theta_{6n}\cap\{\xi=\xi'\}$ of Θ_{6n}, note that $\phi_{6n}^{\xi'} = \phi_{3n}^{\xi'}$ and reuse results in the proof of part (b).

$$\widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{6n}^{\xi'}) = \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{3n}^{\xi'}) \le \mathbb{P}\left(\chi^2_{(1+M_0)s} \ge \frac{1-M_1\epsilon_n}{1+M_1\epsilon_n}\times M_3^2 n\epsilon_n^2/4\right).$$

This bound holds for any $\xi'\supseteq\xi^\star$ such that $|\xi'\setminus\xi^\star| \le M_0 s$ and any $(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta}) \in \Theta_{6n}\cap\{\xi=\xi'\}$. Applying Lemma B6(b) yields

$$\sup_{\Theta_{6n}} \widehat{\mathbb{E}}_{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta})}(1-\phi_{6n}) \lesssim e^{-(M_3^2/8-o)n\epsilon_n^2}.$$

□

Proof of Lemma B4. Define

$$A_n^\star = A_n^\star(\eta_1,\eta_2) = \left\{(\sigma,\boldsymbol{\alpha},\boldsymbol{\beta}):\ \begin{array}{l}\sigma^2/\sigma^{\star 2}\in[1,1+\eta_1\epsilon_n^2],\\ \xi=\xi^\star,\\ |\alpha_j-\alpha_j^\star|\le\sigma^\star\eta_2\epsilon_n/\sqrt{k}, j=1,\dots,k\\ |\beta_j-\beta_j^\star|\le\tau_j\sigma^\star\eta_2\epsilon_n/s, j\in\xi^\star,\end{array}\right\}$$

Evidently,

$$\int \frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}d\pi \ge \pi(A_n^\star)\inf_{A_n^\star}\frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}.$$

Suppose at this moment we have shown that

$$\boldsymbol{\varepsilon}^{\mathrm{T}}(\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\dagger+\widehat{\mathbf{U}}_{\xi^\star}\widehat{\mathbf{U}}_{\xi^\star}^\dagger)\boldsymbol{\varepsilon} < 3C_4' n\epsilon_n^2 \implies \inf_{A_n^\star}\frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})} > e^{-C_4''s\log p}. \tag{17}$$

with $C_4'' = (1+\kappa_1^2/\kappa_0^2)\eta_2^2/2 + \sqrt{3C_4'}(1+\kappa_1/\kappa_0)\eta_2+\eta_1$, and that for any constant $o>0$

$$\pi(A_n^\star) \gtrsim e^{-(2+o)s\log p}. \tag{18}$$

Then, if $C_4 - C_4'' > 2$ and n is sufficiently large

$$\begin{aligned}\int \frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})}d\pi \le e^{-C_4 s\log p} &\implies \inf_{A_n^\star}\frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})} \le e^{-C_4''s\log p}\\ &\implies \boldsymbol{\varepsilon}^{\mathrm{T}}(\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\dagger+\widehat{\mathbf{U}}_{\xi^\star}\widehat{\mathbf{U}}_{\xi^\star}^\dagger)\boldsymbol{\varepsilon} \ge 3C_4' s\log p\end{aligned}$$

The last event happens with probability $\mathbb{P}(\chi^2_{k+s} \geq 3C_4' s \log p) \lesssim e^{-C_4' s \log p}$, due to Lemma B6. For given constants $C_4 > 2$, $C_4' > 0$, choosing sufficiently small η_1, η_2 such that $C_4'' < C_4 - 2$ would complete the proof.

Turn to prove claims (17) and (18). For claim (17), write

$$\begin{aligned}
-\log \frac{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}\boldsymbol{\beta},\sigma^2\mathbf{I})}{\mathcal{N}(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star+\widehat{\mathbf{U}}\boldsymbol{\beta}^\star,\sigma^{\star 2}\mathbf{I})} &= \|\mathbf{Y}-\widehat{\mathbf{F}}\boldsymbol{\alpha}-\widehat{\mathbf{U}}\boldsymbol{\beta}\|^2/2\sigma^2 - \|\mathbf{Y}-\widehat{\mathbf{F}}\boldsymbol{\alpha}^\star-\widehat{\mathbf{U}}\boldsymbol{\beta}^\star\|^2/2\sigma^{\star 2} + n\log(\sigma^2/\sigma^{\star 2}) \\
&= \|\sigma^\star\boldsymbol{\varepsilon} + \widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha}) + \widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})\|^2/2\sigma^2 - \|\boldsymbol{\varepsilon}\|^2/2 + n\log(\sigma^2/\sigma^{\star 2}) \\
&\leq \|\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})\|^2/2\sigma^2 + \|\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})\|^2/2\sigma^2 \\
&\quad + \sigma^\star\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})/\sigma^2 + \sigma^\star\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})/\sigma^2 + \eta_1 n\epsilon_n^2.
\end{aligned}$$

Note that $\tau_j^{-1} = \|\widehat{\mathbf{U}}_j\|/\sqrt{n} \geq \kappa_0$ for $j \in \xi^\star$, and $\|\widehat{\mathbf{U}}_{\xi^\star}\| \leq \|\widehat{\mathbf{U}}_{\xi^\star}\|_{\mathrm{F}} \leq \kappa_1\sqrt{s}$. Each term is the last display can be bounded as follows.

$$\begin{aligned}
\|\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})\|^2/2\sigma^2 &\leq \lambda_{\max}\left(\widehat{\mathbf{F}}^{\mathrm{T}}\widehat{\mathbf{F}}\right)\|\boldsymbol{\alpha}^\star-\boldsymbol{\alpha}\|^2/2\sigma^2 \leq \eta_2^2 n\epsilon_n^2/2 \\
\|\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})\|^2/2\sigma^2 &\leq \lambda_{\max}\left(\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star}\widehat{\mathbf{U}}_{\xi^\star}\right)\|\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star}\|^2/2\sigma^2 \leq \eta_2^2\kappa_1^2 n\epsilon_n^2/2\kappa_0^2 \\
\sigma^\star\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})/\sigma^2 &= \boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\dagger\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})/\sigma^\star \times \sigma^{\star 2}/\sigma^2 \\
&\leq \|\widehat{\mathbf{F}}^\dagger\widehat{\mathbf{F}}^{\mathrm{T}}\boldsymbol{\varepsilon}\| \times \|\widehat{\mathbf{F}}(\boldsymbol{\alpha}^\star-\boldsymbol{\alpha})/\sigma^\star\| \times 1 \\
&\leq \sqrt{3C_4'}\sqrt{n}\epsilon_n \times \eta_2\sqrt{n}\epsilon_n = \sqrt{3C_4'}\eta_2 n\epsilon_n^2 \\
\sigma^\star\boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})/\sigma^2 &= \boldsymbol{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{U}}_{\xi^\star}\widehat{\mathbf{U}}^\dagger_{\xi^\star}\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star-\boldsymbol{\beta})/\sigma^\star \times (\sigma^{\star 2}/\sigma^2) \\
&\leq \|\widehat{\mathbf{U}}^\dagger_{\xi^\star}\widehat{\mathbf{U}}^{\mathrm{T}}_{\xi^\star}\boldsymbol{\varepsilon}\| \times \|\widehat{\mathbf{U}}_{\xi^\star}(\boldsymbol{\beta}^\star_{\xi^\star}-\boldsymbol{\beta}_{\xi^\star})/\sigma^\star\| \times 1 \\
&\leq \sqrt{3C_4'}\sqrt{n}\epsilon_n \times \eta_2\kappa_1\sqrt{n}\epsilon_n/\kappa_0 = \sqrt{3C_4'}\eta_2\kappa_1 n\epsilon_n^2/\kappa_0.
\end{aligned}$$

Putting these bounds together proves claim (17).

For claim (18), note that $\sigma^\star$, $\|\boldsymbol{\alpha}^\star\|$, $\|\boldsymbol{\beta}^\star\|$, and $\tau_j^{-1} = \|\widehat{\mathbf{U}}_j\|/\sqrt{n} \leq \kappa_1$ for $j \in \xi^\star$ are assumed to be of constant order. For all $(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) \in A_n^\star(\eta_1, \eta_2)$, find constants c_1, c_2 such that

$$\begin{aligned}
|\alpha_j| &\leq |\alpha_j^\star| + \eta_2\sigma^\star\epsilon_n/\sqrt{k} \leq c_1, \quad j = 1, \ldots, k \\
|\beta_j/\tau_j| &\leq |\beta_j^\star/\tau_j| + \eta_2\sigma^\star\epsilon_n/s \leq c_2, \quad j \in \xi^\star.
\end{aligned}$$

Then

$$\begin{aligned}
\pi(A_n^\star(\eta_1,\eta_2)) &= \left(\frac{s_0}{p}\right)^s \int_{\sigma^{\star 2}}^{\sigma^{\star 2}(1+\eta_1\epsilon_n^2)} g(\sigma^2)d\sigma^2 \times \prod_{j=1}^k \int_{\alpha_j^\star-\eta_2\sigma^\star\epsilon_n/\sqrt{k}}^{\alpha_j^\star+\eta_2\sigma^\star\epsilon_n/\sqrt{k}} h(\alpha_j)d\alpha_j \\
&\quad \times \prod_{j\in\xi^\star} \int_{\beta_j^\star/\tau_j-\eta_2\sigma^\star\epsilon_n/s}^{\beta_j^\star/\tau_j+\eta_2\sigma_\star\epsilon_n/s} h(\beta_j/\tau_j)d(\beta_j/\tau_j) \\
&\geq \left(\frac{s_0}{p}\right)^s \times \sigma^{\star 2}\eta_1\epsilon_n^2 g(\sigma^{\star 2})/2 \times \left(\frac{2\eta_2\sigma^\star\epsilon_n}{\sqrt{k}} \inf_{|z|\leq c_1} h_1(z)\right)^k \times \left(\frac{2\eta_2\sigma^\star\epsilon_n}{s} \inf_{|z|\leq c_2} h_2(z)\right)^s \\
&= c_3 c_4^s \times s^{(2+k+s)/2-s} \times (\log p)^{(2+k+s)/2} \times n^{-(2+k+s)/2} \times p^{-s} \gtrsim p^{-(2+o)s}
\end{aligned}$$

with $c_3 = 2^{k-1}\eta_1\eta_2^k\sigma^{\star(2+k)}g(\sigma^{\star 2})[\inf_{|z|\leq c_1} h_1(z)]^k$, $c_4 = 2s_0\eta_2\sigma^\star \inf_{|z|\leq c_2} h_2(z)$. □

C Gibbs Sampler

For the prior (8), we set g as the inverse-gamma density function with shape $a_0 = 1$ and scale $b_0 = 1$, $h_1(z) = \mathcal{N}(z|0, \sigma_1^2)$ and $h_2(z) = \mathcal{N}(z|0, \sigma_2^2)$. A Gibbs sampler is implemented to learn the pseudo-posterior distribution $\widehat{\pi}(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}|\widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})$ given by (9). This Gibbs sampler converges towards the pseudo-posterior joint distribution of $(\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by iterating the following steps: (1) draw ξ given $\boldsymbol{\alpha}$ and σ^2, (2) draw $\boldsymbol{\beta}$ given ξ, $\boldsymbol{\alpha}$ and σ^2, (3) draw $\boldsymbol{\alpha}$ given $\xi, \boldsymbol{\beta}$ and σ^2, (4) draw σ^2 given $\xi, \boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For simplicity, implementation details with $\tau_j = 1$ for $j = 1, \ldots, p$ are presented.

(1) For the conditional distribution of ξ, write

$$\widehat{\pi}(\xi = \emptyset|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) = \left(1 - \frac{s_0}{p}\right)^p \times (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\|\mathbf{Y} - \widehat{\mathbf{F}}\boldsymbol{\alpha}\|^2}{2\sigma^2}\right),$$

and, for $\xi \neq \emptyset$,

$$\widehat{\pi}(\xi, \boldsymbol{\beta}_\xi|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) \propto \left(\frac{s_0}{p - s_0}\right)^{|\xi|} \exp\left(-\frac{\|\mathbf{Y} - \widehat{\mathbf{F}}\boldsymbol{\alpha} - \widehat{\mathbf{U}}_\xi\boldsymbol{\beta}_\xi\|^2}{2\sigma^2}\right)(2\pi\sigma_1^2)^{-|\xi|/2} \exp\left(-\frac{\|\boldsymbol{\beta}_\xi\|^2}{2\sigma_1^2}\right).$$

It follows that

$$\begin{aligned}\frac{\widehat{\pi}(\xi|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})}{\widehat{\pi}(\emptyset|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})} &= \frac{\int \widehat{\pi}(\xi, \boldsymbol{\beta}_\xi|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) d\boldsymbol{\beta}_\xi}{\widehat{\pi}(\emptyset|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})} \\ &= \left(\frac{s_0}{p - s_0}\right)^{|\xi|} \sigma^{|\xi|}\sigma_1^{-|\xi|} \det(\mathbf{S}_\xi)^{1/2} \exp\left(\frac{\mathbf{Y}^{\mathrm{T}}\widehat{\mathbf{U}}_\omega\mathbf{S}_\omega\widehat{\mathbf{U}}_\omega^{\mathrm{T}}\mathbf{Y}}{2\sigma^2}\right),\end{aligned} \tag{19}$$

where $\mathbf{S}_\xi = (\widehat{\mathbf{U}}_\xi^{\mathrm{T}}\widehat{\mathbf{U}}_\xi + \sigma^2\sigma_1^{-2}\mathbf{I})^{-1}$. However, it is computationally prohibitive to directly sample from this conditional distribution, as ξ takes 2^p possible values. As a remedy, we flip $Z_j = 1\{j \in \xi\}$ in random scans with probability

$$\widehat{\pi}(Z_j = 1|\{Z_{j'}\}_{1 \leq j' \neq j \leq p}, \sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y}) = \left[1 + \frac{\widehat{\pi}(\xi = \omega|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})}{\widehat{\pi}(\xi = \omega \cup \{j\}|\sigma^2, \boldsymbol{\alpha}, \widehat{\mathbf{F}}, \widehat{\mathbf{U}}, \mathbf{Y})}\right]^{-1},$$

where $\omega = \{j' \neq j : Z_j' = 1\}$. For $\omega \neq \emptyset$, the Bayes factor between models ω and $\omega \cup \{j\}$ is given by

$$\frac{p - s_0}{s_0} \times \frac{\sigma_1}{\sigma} \times \left[\frac{\det(\mathbf{S}_\omega)}{\det(\mathbf{S}_{\omega\cup\{j\}})}\right]^{1/2} \exp\left(\frac{\mathbf{Y}^{\mathrm{T}}\widehat{\mathbf{U}}_\omega\mathbf{S}_\omega\widehat{\mathbf{U}}_\omega^{\mathrm{T}}\mathbf{Y}}{2\sigma^2} - \frac{\mathbf{Y}^{\mathrm{T}}\widehat{\mathbf{U}}_{\omega\cup\{j\}}\mathbf{S}_{\omega\cup\{j\}}\widehat{\mathbf{U}}_{\omega\cup\{j\}}^{\mathrm{T}}\mathbf{Y}}{2\sigma^2}\right),$$

where

$$\frac{\det(\mathbf{S}_\omega)}{\det(\mathbf{S}_{\omega\cup\{j\}})} = \frac{\det(\widehat{\mathbf{U}}_{\omega\cup\{j\}}^{\mathrm{T}}\widehat{\mathbf{U}}_{\omega\cup\{j\}} + \sigma^2\sigma_1^{-2}\mathbf{I})}{\det(\widehat{\mathbf{U}}_\omega^{\mathrm{T}}\widehat{\mathbf{U}}_\omega + \sigma^2\sigma_1^{-2}\mathbf{I})} = \left(\widehat{\mathbf{U}}_j^{\mathrm{T}}\widehat{\mathbf{U}}_j + \sigma^2\sigma_1^{-2}\right) - \widehat{\mathbf{U}}_j^{\mathrm{T}}\widehat{\mathbf{U}}_\omega\mathbf{S}_\omega\widehat{\mathbf{U}}_\omega^{\mathrm{T}}\widehat{\mathbf{U}}_j,$$

due to the property of the Schur complement. For $\omega = \emptyset$, the Bayes factor between models $\emptyset$ and $\{j\}$ is given by

$$\frac{p - s_0}{s_0} \times \frac{\sigma_1}{\sigma} \times \left(\widehat{\mathbf{U}}_j^{\mathrm{T}}\widehat{\mathbf{U}}_j + \sigma^2\sigma_1^{-2}\right)^{1/2} \times \exp\left(-\frac{\mathbf{Y}^{\mathrm{T}}\widehat{\mathbf{U}}_{\{j\}}\mathbf{S}_{\{j\}}\widehat{\mathbf{U}}_{\{j\}}^{\mathrm{T}}\mathbf{Y}}{2\sigma^2}\right),$$

In experiments, we find that just one random scan suffices for the proposed method to perform well.

(2) For the conditional distribution of $\boldsymbol{\beta}$,

$$\widehat{\pi}(\boldsymbol{\beta}_\xi|\sigma^2,\boldsymbol{\alpha},\xi,\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{Y}) \sim \mathcal{N}\left(\mathbf{S}_\xi\widehat{\mathbf{U}}_\xi^{\mathrm{T}}\mathbf{Y},\sigma^2\mathbf{S}_\xi\right), \quad \boldsymbol{\beta}_{\xi^c}\equiv 0.$$

(3) For the conditional distribution of $\boldsymbol{\alpha}$,

$$\widehat{\pi}(\boldsymbol{\alpha}|\sigma^2,\boldsymbol{\beta},\xi,\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{Y}) \sim \mathcal{N}\left(\frac{\widehat{\mathbf{F}}^{\mathrm{T}}\mathbf{Y}}{n+\sigma^2/\sigma_2^2},\frac{\sigma^2}{n+\sigma^2/\sigma_2^2}\right).$$

(4) For the conditional distribution of σ^2,

$$\begin{aligned}\widehat{\pi}(\sigma^2|\boldsymbol{\alpha},\boldsymbol{\beta},\xi,\widehat{\mathbf{F}},\widehat{\mathbf{U}},\mathbf{Y}) &\propto g(\sigma^2|a_0,b_0)\mathcal{N}\left(\mathbf{Y}|\widehat{\mathbf{F}}\boldsymbol{\alpha}+\widehat{\mathbf{U}}_\xi\boldsymbol{\beta}_\xi,\sigma^2\mathbf{I}\right)\\ &\propto g\left(\sigma^2\middle|a_0+\frac{n}{2},b_0+\frac{\|\mathbf{Y}-\widehat{\mathbf{F}}\boldsymbol{\alpha}-\widehat{\mathbf{U}}_\xi\boldsymbol{\beta}_\xi\|^2}{2}\right).\end{aligned}$$

The overall time complexity of the factor-adjusted Bayesian method is $O(np^2)+O(Tps^3)$, where T is the number of iterations of the posterior computation algorithm. As suggested by Yang et al. (2016), $T=O(s^2\log p)$ may suffice for the posterior sampler to converge. Below are details of the time complexity analysis. The truncated singular value decomposition algorithms can compute $\widehat{\mathbf{F}},\widehat{\mathbf{U}}$ with time complexity $O(npk)$ (Allen-Zhu and Li, 2016). Computing $\widehat{\mathbf{F}}^{\mathrm{T}}\mathbf{Y}$, $\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{Y}$ and $\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}}$ takes $O(np^2)$ flops. Given $\widehat{\mathbf{F}}^{\mathrm{T}}\mathbf{Y}$, $\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{Y}$ and $\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}}$, each iteration of the posterior computation algorithm takes $O(ps^3)$ flops (per random scan) to sample from the conditional distribution of ξ, because computing the conditional probability ratio between models $\xi=\omega$ and $\xi=\omega\cup\{j\}$ for each flip update takes $O(|\omega|^3)=O(s^3)$ flops, and each random scan consists of p flip updates. Each iteration also takes $O(s^3)$, $O(1)$, $O(ns)$ flops to sample from the conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ^2, respectively.

D Response

We would like to thank two reviewers for their comments. We have revised and improved the manuscript to address their concerns.

To Reviewer 1:

On Interpretation of Model Setup

Reviewer 1: My first major concern is about the interpretation of the model setup. Instead of focusing on the conventional regression model

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\sigma\boldsymbol{\varepsilon}, \qquad (1) \text{ in the updated manuscript}$$

the paper considers the factor-adjusted regression model

$$\mathbf{Y}=\mathbf{F}\boldsymbol{\alpha}+\mathbf{U}\boldsymbol{\beta}+\sigma\boldsymbol{\varepsilon}, \qquad (4) \text{ in the updated manuscript.}$$

...I believe model (4) is equivalent to the factor-augmented high-dimensional linear regression model

$$\mathbf{Y}=\mathbf{F}\boldsymbol{\alpha}'+\mathbf{X}\boldsymbol{\beta}+\sigma\boldsymbol{\varepsilon},, \qquad (6) \text{ in the updated manuscript.}$$

I wonder if the paper can provide some discussions along this line, so that we may better understand the implications of the model. I also feel model (6) might be a more intuitive representation than model (4) in the sense that (6) clearly separates the roles played by $\mathbf{X}$ *and* $\mathbf{F}$.

We provide more discussions on the differences between models (1),(4) and (6) in the introduction section. We agree with you that model (6) is equivalent to model (4) up to reparametrization $\boldsymbol{\alpha}' = \boldsymbol{\alpha} - \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$. Model (6) has been studied by (Kneip and Sarda, 2011). However, for models (1) and (6), the sparse regression methods need to impose the weak correlation condition on $\mathbf{X}$. Bayesian sparse regression methods need the sparse eigenvalue (SE) condition, a specific type of weak correlation condition. The sparse eigenvalue of $\mathbf{X}$ could be much smaller than the sparse eigenvalue of $\mathbf{U}$. Section 2 in the updated manuscript makes this intuition precise as

$$\frac{\mathrm{SE}(\mathbf{X})}{\mathrm{SE}(\mathbf{U})} \leq \max_{j=1}^{p} \frac{\|\mathbf{U}_j\|^2}{\|\mathbf{X}_j\|^2} \times \mathrm{R}(\mathbf{U}), \quad \text{with } \mathrm{R}(\mathbf{U}) \asymp 1.$$

Note that the total variation of $\mathbf{X}_j$ mainly consists of two parts $\|\mathbf{X}_j\|^2 \approx \|\mathbf{F}\boldsymbol{b}_j\|^2 + \|\mathbf{U}_j\|^2$. When the strong correlation part $\mathbf{F}\boldsymbol{b}_j$ accounts for a large portion of the total variation, the sparse eigenvalue of $\mathbf{X}$ is small, causing incorrect estimation and slow convergence speed of the Bayesian method. Experimental results also verify this intuition (see Figs. 2 and 4 in the updated manuscript).

We feel that model (4) is a more better representation than model (6). Each covariate $\mathbf{X}_j$ has two parts $\mathbf{F}\boldsymbol{b}_j$ and $\mathbf{U}_j$. In model (4), the strong correlation parts $\mathbf{F}\boldsymbol{b}_j$'s of all covariates contribute to the response $\mathbf{Y}$ aggregately, while idiosyncratic components $\{\mathbf{U_j} : j \in \xi^\star\}$ of a small number of covariates have specific effects on $\mathbf{Y}$. In model (6), the effects of common factors $\mathbf{F}$ and strong correlation parts $\mathbf{F}\boldsymbol{b}_j$'s of $\mathbf{X}_j$'s are not clearly separated.

Adjusted Sparseness Assumption as Motivation

Reviewer 1: "My second major concern is about the motivation part. If the authors agree that models (1) and (4) are equivalent, then using sparsity of model (1) as a motivation seems not appropriate, given that a sparse $\boldsymbol{\beta}$ *in model (4) is the same as a sparse* $\boldsymbol{\beta}$ *in model (1).*

We change the motivation part of the proposed method in the introduction section. The adjustment of the sparseness assumption is a consequence of the adjustment of the weak correlation condition. We agree with you that a sparse $\boldsymbol{\beta}$ in model (4) is the same as a sparse $\boldsymbol{\beta}$ in model (1). But, the meaning of the sparseness of $\boldsymbol{\beta}$ has been changed. In (1), a nonzero β_j means an overall effect of $\mathbf{X}_j$ (sum of the effects of $\mathbf{F}\boldsymbol{b}_j$ and $\mathbf{U}_j$). In (1), a nonzero β_j means the specific effect of $\mathbf{X}_j$ (or $\mathbf{U}_j$), excluding $\mathbf{F}\boldsymbol{b}_j$.

On Constraint $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$

Reviewer 1: If the authors do not agree with the equivalence between models (1) and (4), I would like to know the following.

(a) The paper argues that model (4) covers model (1) as a special case by restricting the side constraint that $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$. *I wonder how to interpret* $\boldsymbol{\beta}$ *and* $\boldsymbol{\alpha}$ *in (4) when the constraint does not hold.*

(b) The paper assumes that $\boldsymbol{\beta}$ *is sparse. How is this assumption different from the sparsity assumption for model (1)?*

(c) In general, how to evaluate a particular covariate's impact on Y in model (4) when $\boldsymbol{\alpha} \neq \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$?

(d) Is it possible to test $\boldsymbol{\alpha} \neq \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ within the Bayesian framework? I think using the representation (6), a test for $\boldsymbol{\alpha} \neq \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ is equivalent to a test for $\alpha' = 0$ in (1), which looks more straightforward."

Model (1) is equivalent to

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \quad \boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}, \quad \text{(3) in the updated manuscript.}$$

We think model (4) is more general than model (1) or equivalently (3), as the former drops the constraint $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$.

(a) When the constraint is removed, covariates $\mathbf{X}_j$'s do not directly contribute to $\mathbf{Y}$. They are outcomes of some underlying factor model. The common factors and idiosyncratic components contribute to $\mathbf{Y}$, with coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

(b) As we have discussed, a sparse $\boldsymbol{\beta}$ in model (4) is the same as a sparse $\boldsymbol{\beta}$ in model (1). But, the meaning of the sparseness of $\boldsymbol{\beta}$ has been changed. In (1), a nonzero β_j means an overall effect of $\mathbf{X}_j$ (sum of the effects of $\mathbf{F}\boldsymbol{b}_j$ and $\mathbf{U}_j$). In (1), a nonzero β_j means the specific effect of $\mathbf{X}_j$ (or $\mathbf{U}_j$), excluding $\mathbf{F}\boldsymbol{b}_j$.

(c) In model (4) with $\boldsymbol{\alpha} \neq \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$, each covariate $\mathbf{X}_j$ does not directly contribute to $\mathbf{Y}$.

(d) This is an interesting question we had not thought about before. We can test $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ in the Bayesian framework by looking into the posterior distribution of $\boldsymbol{\alpha} - \widehat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\beta}$, since we have sample of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ from the pseudo posterior distribution and the estimate $\widehat{\mathbf{B}} \approx \mathbf{B}$ from PCA. We will leave it to future research.

On Estimated Latent Variables and Rate-optimality of Sparse Regression

Review 1: "The paper shows in Section 3 that the pseudo posterior distribution (9) achieves the best rate Bayesian methods can achieve with observed $[\mathbf{F}, \mathbf{U}]$. Can the authors provide more discussion why conditioning on $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ does not affect the convergence rate? In general, what are the implications for inference if the posterior is conditional on $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$ instead of on $[\mathbf{F}, \mathbf{U}]$?

We add a paradigm Fig. 1 to illustrate the idea to prove Theorem 2. Conditioning on $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$, the properties of Bayesian sparse regression are established in the probability space of the pseudo data generating process $\mathbf{Y} = \widehat{\mathbf{F}}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$. Then the properties are translated back to the probability space of the pseudo data generating process $\mathbf{Y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$. In the probability space of the pseudo data generating process, the error arising from the Bayesian sparse regression method is determined by the strength of the sparse eigenvalue condition, measured by a constant M_0. The deviation between two probability spaces is controlled in terms of the ℓ_2 distance between their conditional means

$$\|(\widehat{\mathbf{F}}\mathbf{H}\boldsymbol{\alpha}^\star + \widehat{\mathbf{U}}\boldsymbol{\beta}^\star) - (\mathbf{F}\boldsymbol{\alpha}^\star + \mathbf{U}\boldsymbol{\beta}^\star)\| \leq L_5\sigma^\star\sqrt{n}\epsilon_n.$$

When $M_0 - 2 > L_5^2$, the error due to the estimation $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}] \approx [\mathbf{F}, \mathbf{U}]$ in the factor model is relatively small compared to that arising from the Bayesian sparse regression, and therefore the estimation of the factor model does not change the order of the error rate of the Bayesian method, but do change the constant factor of the error rate. If an inaccurate estimation of latent variables leads to a large L_5, a stronger sparse eigenvalue condition with larger M_0 is needed by the Bayesian sparse regression method.

On Assumption 1 regarding $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$

Review 1: "Can the authors provide more discussions on Assumption 1 regarding $[\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$? Are there any examples of DGP so that Assumption 1 holds? When will Assumption 1 be violated?

We add two concrete examples Examples 1 and 2 in which Assumption 1 holds. Roughly speaking, when $[\mathbf{F}, \mathbf{U}]$ contain subgaussian entries, Assumption 1 holds. It might be violated when entries of $[\mathbf{F}, \mathbf{U}]$ have heavy tails. In that case, we need to use some robust covariance matrix estimator in place of the sample covariance matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ in the PCA procedure.

On Theorem 1

Review 1: "Can the authors provide more discussions about Theorem 1, in particular, how do Assumptions 1 to 3 imply Assumption 5? What is the relationship between the assumptions in this paper and those in (Bai and Ng, 2002; Bai, 2003).

We add Section 4.3 to discuss the similarity and difference of Theorem 1 to (Bai and Ng, 2002). Theorem 1 can be viewed as a non-asymptotic version of (Bai and Ng, 2002). In general, the non-asymptotic analysis is more quantitative than the asymptotic analysis and more suitable for high-dimensional statistics.

On Table 2 in Simulation Experiment Section

Review 1: "In simulation experiment section's Table 2, I do not think a direct comparison with the generic Bayes or generic lasso is fair as models (1) and (4) are not the same models when $\boldsymbol{\alpha} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$. A more appropriate comparison should be with the Bayesian analysis of model (6), or maybe other similar models."

We redesign the experiments by setting $\boldsymbol{\alpha}^\star = \mathbf{B}^{\mathrm{T}}\boldsymbol{\beta}$ for a fair comparison between factor-adjusted methods and routine methods.

To Reviewer 2:

On Iteration Number of Gibbs Sampler

Reviewer 2: "Based on my personal experience, when the model is complex, the Bayesian Gibbs sampling algorithm is very difficult to converge. Hence, how to show the convergence of MCMC drawings is still an important concern for the applied Bayesian readers. However, in page 13, you said that 'we iterate a Gibbs sampler $T = 20$ times and drop the first $T/2 = 10$ iterations as the burn-in period.' I am very confusing about this sentence. Is it enough for convergence? Could you check this claim? However, from your proof from your appendix, the number of drawing requires an order with $O(tpns^2)$ in page 48. May I suggest that in the real data analysis in section 6, could you give some graphical tools or test statistics to show that the burn-in length is enough to achieve the convergence?"

We add Fig. 4 to show the fast convergence of the proposed Bayesian method in the setup of $n = 200, p = 500$ (larger than $n = 120$, $p = 131$ in the real dataset of U.S. bond risk premia). $T = 20$ iterations are indeed enough.

We revise the time complexity analysis in the appendix. The overall time complexity of the factor-adjusted Bayesian method is $O(np^2) + O(Tps^3)$, where $O(np^2)$ is for multiplication of large matrices and T is the number of iterations of the posterior computation algorithm. Details are given. As suggested by Yang et al. (2016), $T = O(s^2 \log p)$ may suffice for the posterior sampler to converge

in case that the covariates are weakly correlated. Fortunately, the proposed method has remove the strong correlation parts from covariates. The routine Bayesian method with strongly correlated covariates does encounter the slow convergence issue.

On Information-based Model Selection Criteria

Reviewer 2: "In Bayesian literature, as to model selection, Bayes factor or DIC, which are Bayesian version of BIC or AIC respectively, are very popular criterion for model selection. Hence, could you give a remark or discussion why these popular criteria cannot be used? After all, many Bayesian readers are familiar with the use of Bayes factor or DIC. I think that this kind of discussion can strengthen your motivation of your paper about why we need develop a new approach."

AIC, BIC and DIC are popular criteria for classical model selection problems. However, in the high-dimensional regression models considered in this paper, there are 2^p possible models and $n \prec p \prec e^n$. It is computationally prohibitive to perform these criteria. A short and good notes on this topic is Section 2.4 in Philippe Rigollet and Jan-Christian Hütter's lecture notes on high-dimensional statistics http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf. To extend these criteria to the high-dimensional regime, a remedy to restrict the focus on models of size at most Cs (Kim et al., 2012, JMLR, consistent model selection criteria on high dimensions). Still these criteria need to consider p^{Cs} possible models. In contrast, the Bayesian sparse regression method needs $O(s^2 \log p)$ iterations, and each iteration visits p possible models (Yang et al., 2016).

The presented paper is mainly motivated from the field of the high-dimensional linear regression, which has grew apart from information-based criteria in the last decade.