DSSR

Discussion Paper No. 108

Estimation of Weak Factor Models

Yoshimasa Uematsu and Takashi Yamagata

September, 2020

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Estimation of Weak Factor Models

Yoshimasa Uematsu* and Takashi Yamagata†

*Department of Economics and Management, Tohoku University
†Department of Economics and Related Studies, University of York
†Institute of Social Economic Research, Osaka University

September 2020 (3rd version)

Abstract

This paper investigates estimation of sparsity-induced weak factor (sWF) models, with large cross-sectional and time-series dimensions (N and T, respectively). It assumes that the kth largest eigenvalue of data covariance matrix grows proportionally to N^{α_k} with unknown exponents $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. Employing the same rotation of the principal component (PC) estimator, in the sWF models the growth rate α_k is linked to the degree of sparsity of kth factor loadings. This is much weaker than the typical assumption on the recent factor models, in which all the r largest eigenvalues diverge proportionally to N. We apply the SOFAR method of Uematsu et al. (2019) to estimate the sWF models and derive the estimation error bound. Importantly, our method yields consistent estimation of α_k's as well. A finite sample experiment shows that the performance of the new estimator uniformly dominates that of the PC estimator. We apply our method to forecasting bond yields and results demonstrate that our method outperforms that based on the PC. In another application we analyze S&P500 firm security returns and find that the first factor is consistently near strong while the others are indeed weak.

Keywords. Sparsity-induced weak factor models, (Adaptive) SOFAR estimator, Estimation error bound, Estimating diverging exponents, Interpreting factors, Group factor structure.

1 Introduction

The approximate factor model with large cross-sectional and time-series dimensions (N and T, respectively) has become an increasingly important tool for the analysis of psychology, finance, economics, and biology, among many others. In finance, the model is firstly introduced by Chamberlain and Rothschild (1983), then developed in the subsequent articles by Connor and Korajczyk (1986, 1993), Bai and Ng (2002), Bai (2003), Fan et al. (2008),

*Department of Economics and Management, Tohoku University, 27-1 Kawauchi, Aobaku, Sendai 980-8576, Japan (E-mail: yoshimasa.uematsu.e7@tohoku.ac.jp). He gratefully acknowledges the partial support of Grant-in-Aid for JSPS Overseas Research Fellow 29-60 and JSPS KAKENHI 19K13665.

†Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK and Institute of Social and Economic Research (ISER), Osaka University, Japan (E-mail: takashi.yamagata@york.ac.uk). He gratefully acknowledges the partial support of JSPS KAKENHI JP15H05728 and JP18K01545. The authors appreciate Kun Chen giving helpful suggestions and modification of the R package, rrpack.

Fan et al. (2011, 2013), among many others. In macroeconomics, Stock and Watson (2002) propose to extract a small number of factors from the large macroeconomic and financial series and use them to forecast a macroeconomic variable of interest. Ludvigson and Ng (2009) take a similar approach to forecast bond yields. See, for example, Fan et al. (2018) for an excellent review of the high-dimensional factor models and their applications.

1.1 Weak factor model, rotation, and sparsity

Suppose that a vector of zero-mean stationary time series $\mathbf{x}_t = (x_{t1}, \ldots, x_{tN})' \in \mathbb{R}^N$, $t = 1, \ldots, T$, is generated from the factor model

$$\mathbf{x}_t = \mathbf{B}^* \mathbf{f}_t^* + \mathbf{e}_t, \tag{1}$$

where $\mathbf{B}^* = (\mathbf{b}_1^*, \ldots, \mathbf{b}_N^*)' \in \mathbb{R}^{N \times r}$ with $\mathbf{b}_i^* \in \mathbb{R}^r$ is a matrix of deterministic factor loadings, $\mathbf{f}_t^* \in \mathbb{R}^r$ is a vector of zero-mean latent factors, and $\mathbf{e}_t \in \mathbb{R}^N$ is an idiosyncratic error vector. For a while suppose r is given. Let $\mathbf{\Sigma}_x = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t']$, $\mathbf{\Sigma}_f^* = \mathbb{E}[\mathbf{f}_t^* \mathbf{f}_t^{*\prime}]$, and $\mathbf{\Sigma}_e = \mathbb{E}[\mathbf{e}_t \mathbf{e}_t']$. Assuming uniform boundedness of $\lambda_k(\mathbf{\Sigma}_e)$ together with an exogeneity condition, we observe that

$$\lambda_k(\mathbf{\Sigma}_x) \asymp \lambda_k(\mathbf{B}^* \mathbf{\Sigma}_f^* \mathbf{B}^{*\prime}) \quad \text{for each} \quad k = 1, \ldots, r$$

and $\lambda_k(\mathbf{\Sigma}_x)$ are uniformly bounded for all $k = r+1, \ldots, N$.

In the studies on high-dimensional factor models, including Connor and Korajczyk (1986, 1993), Stock and Watson (2002), Bai and Ng (2002, 2006, 2013), Bai (2003) and Fan et al. (2018), it is typically assumed that all the r largest eigenvalues diverge proportional to N, namely, $\lambda_k(\mathbf{B}^* \mathbf{\Sigma}_f^* \mathbf{B}^{*\prime}) \asymp N$ for all $k = 1, \ldots, r$. We call the models with this condition the *strong factor (SF) models*. This SF assumption seems unduly restrictive, as it does not permit slower divergence rates than N nor different divergence rates among the r largest eigenvalues. The original *approximate factor model* proposed by Chamberlain and Rothschild (1983) is an important exception, which assumes that $\lambda_r(\mathbf{B}^* \mathbf{\Sigma}_f^* \mathbf{B}^{*\prime}) \to \infty$ as $N \to \infty$. Inspired by this condition, we will significantly relax the SF condition, and consider the structure,

$$\lambda_k(\mathbf{B}^* \mathbf{\Sigma}_f^* \mathbf{B}^{*\prime}) \asymp N_k := N^{\alpha_k} \text{ with } 0 < \alpha_k \leq 1 \text{ for each } k = 1, \ldots, r. \tag{2}$$

We call the factor models with (2) the *weak factor (WF) models* in this paper. The WF models allow different divergence rates of the signal eigenvalues, which can be slower than N. Our definition of the WF models is similar to the one in De Mol et al. (2008), but the readers are cautioned that the definition varies in the literature. For example, they assume non-diverging factors (i.e. $\alpha_r = 0$), which Chamberlain and Rothschild (1983) and we exclude; see Onatski (2012), Bryzgalova (2016), Lettau and Pelger (2020)). Chudik et al. (2011) categorize the factors according to the values of the exponents.

It is well-known that estimation of factor models, including (1), has an identification issue. To address it, we must impose r^2 restrictions on the model. Since the column and row spaces of $\mathbf{F}^* = (\mathbf{f}_1^*, ..., \mathbf{f}_T^*)'$ and $\mathbf{B}^{*\prime}$ are identical to those of $\mathbf{F}^* \mathbf{H}$ and $\mathbf{H}^{-1} \mathbf{B}^{*\prime}$, respectively, for any invertible matrix $\mathbf{H}$, we choose a specific (but frequently employed) rotation without loss of generality. That is, we put $\mathbf{f}_t^0 = \mathbf{H} \mathbf{f}_t^*$ and $\mathbf{B}^{0\prime} = \mathbf{H}^{-1} \mathbf{B}^{*\prime}$ with $\mathbf{\Sigma}_f = \mathbb{E}[\mathbf{f}_t^0 \mathbf{f}_t^{0\prime}] = \mathbf{I}_r$ and $\mathbf{B}^{0\prime} \mathbf{B}^0$ being a diagonal matrix. Then the model in (1) becomes

$$\mathbf{x}_t = \mathbf{B}^0 \mathbf{f}_t^0 + \mathbf{e}_t, \tag{3}$$

and is identifiable. Because the eigenvalues of (2) are invariant to any rotation, we have

$$N_k \asymp \lambda_k(\mathbf{B}^0\mathbf{B}^{0\prime}) = \lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \quad \text{for each} \quad k = 1, \ldots, r. \tag{4}$$

To estimate the identifiable model (3) with satisfying (4), we require the assumption that $\mathbf{B}^0$ is *sparse* such that (4) and diagonality of $\mathbf{B}^{0\prime}\mathbf{B}^0$ simultaneously hold.[1] It is called the *sparsity-induced weak factor (sWF) model*, and we investigate estimation of the sWF models hereafter. As the earlier discussion implies, the WF structure in (4) can be induced by *non-sparse* factor loadings. For instance, it is the case when a factor affects all the variables at similar strengths thinly,[2] but we do not consider this class in this paper.

1.2 Empirical evidence of the sWF models

A growing body of evidence in the literature supports the sWF models. First, influential empirical studies often find that the factors identified under the restrictions we impose are loaded on small subsets of the variables. Stock and Watson (2002) find that each of the extracted six factors from macroeconomic indicators are essentially loaded on the variables only in a few of the 14 categories; see figure 1 and discussions therein. By implementing a similar analysis, Ludvigson and Ng (2009) find a sharp contrast in intensity of correlation between each of the five factors and the measures of economic activity from which the factors are extracted, across the categories; see figures 1–5 and discussions therein. We will examine this feature by observing the sparse factor loadings; see Section 6.2. Note that Uematsu and Yamagata (2020) formally establish an inferential method for zeros in the loadings.

Another strand of empirical support for the sWF models comes from the literature on hierarchical (group) factor structures, which contain two types of factors, global and local factors. The factor loadings of the global factors are all non-zeros, whereas the local factors are associated with the loadings with nonzero elements only among specific cross-sectional groups. Ando and Bai (2017), Choi et al. (2018) provide empirical evidence for such a structure in financial and macroeconomic data sets. Importantly, the sWF model (3) nests the hierarchical factor model, to which the same identification restrictions have typically been imposed, and thus our method can be applied; see Section 5.3. In this context, Andreou et al. (2019) propose a test for the number of factors in the group factor models.

1.3 Contributions

Unlike the principal components (PC) estimator, our estimator for the sWF models requires the ℓ_1-norm regularization; see Section 3.1. Although the numerical optimization becomes much more complicated due to the imposition of both sparsity and orthogonality on the estimator, we can obtain a highly efficient estimator by employing the recently developed framework, the *sparse orthogonal factor regression (SOFAR)* of Uematsu et al. (2019). Hereafter the new estimator is called the *SOFAR estimator*.

As theoretical contributions, we will establish the estimation error bounds of the SOFAR and PC estimators as well as validating the method of Onatski (2010) for determining the number of factors for the sWF models. Perhaps surprisingly, our SOFAR estimator can

[1] Although sparsity of $\mathbf{B}^0$ is not generally rotation invariant, we can identify the r signal eigenvalues of model (1) as long as $\mathbf{B}^0$ is sparse. Also the sparse structure of $\mathbf{B}^0$ is row permutation invariant, or invariant to orderings of cross-section units; see Bai et al. (2016).

[2] For illustration, when the kth column vector of $\mathbf{B}^0$ is not sparse and composed of nonzero values of order $N^{\alpha_k - 1}$, it is easy to see that $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0)$ diverges proportionally to N^{α_k}.

be consistent for the sWF models with α_k less than $1/2$. We also propose the *adaptive* SOFAR estimator, which yields *factor selection consistency*. This property asymptotically guarantees the true support recovery of the sparse loadings. The assumptions we will make are in line with the literature of the approximate factor models. Thus the statistical theory substantially departs from that in Uematsu et al. (2019). In particular, the theoretical investigation of the adaptive SOFAR is completely new to the literature.

Importantly, the factor selection consistency enables us to consistently estimate each exponent α_k of the divergence rates. Recently estimation of the exponents has drawn great attention of empirical researchers since it is a useful measure of strength of the cross-sectional correlations. Assuming sparse loadings, Bailey et al. (2016, 2020) and Gao et al. (2020) propose methods that make use of cross-sectional averages of data for estimation and inference of the exponent, but they can only identify the largest divergence rate, α_1. This is essentially because they focus on estimation of the structural model (1). In contrast, our method can identify all the divergence rates because we impose the identification restrictions and focus on the rotated model (3).[3]

We implement extensive finite sample experiments in terms of the determining the number of factors and estimation accuracy of each parameter. Regarding estimation accuracy, we find that the SOFAR estimate uniformly dominates the PC estimate across all the designs we consider. We also conduct empirical analysis with a large data set of macroeconomic variables and S&P 500 monthly returns. In the first analysis, we compare the out-of-sample performance of forecasting bond yields using extracted factors from the macroeconomics variables via our method and the PC method. The statistical evidence suggests that our SOFAR method outperforms the PC method. In the second analysis, we illustrate usefulness of looking into sparse factor loadings to find properties of the extracted factors. The third analysis shows that the first factor in S&P 500 monthly returns is consistently near strong, while the second to fourth exponents vary over months between 0.90 and 0.65.

1.4 Related work

To our knowledge, this is the first study to propose a method that can estimate the WF models, separately identifying spans of $\mathbf{B}^*$ and $\mathbf{F}^*$, while taking the possibly different rates (2) into account. Recently some alternative approaches have been proposed. Freyaldenhoven (2020) advocates a two-step identification strategy; after obtaining the (non-sparse) PC estimator, it seeks a rotation that maximizes the number of zeros in the loading matrix. This approach can be seen as a complementary one to ours since it can reveal an alternative sparsity property with a different rotation. Another related recent work is Daniele et al. (2020). Extending Bai and Li (2012) and Bai and Liao (2017), they propose a method to estimate the idiosyncratic variance-covariance matrix for the sWF models, but it is questionable whether their strategy really works to separately identify the factors and factor loadings. [4]

There are some studies that consider WF models, but most of them have focused only on the case where all the divergence rates are identical. Such examples are seen in De Mol et al. (2008) and Lam et al. (2011); the former consider the Bayesian forecasts with the PC estimates for WF models, and the latter propose an efficient estimator for WF models with a specific correlation structure. Other related research includes Onatski (2012), Bryzgalova

[3] Bailey et al. (2020) permit estimation of multiple exponent for the models with observed factors.

[4] Daniele et al. (2020) counts the number of (*ex post*) zeros in the estimated loadings as a part of the r^2 identification restrictions.

(2016), and Lettau and Pelger (2020). They consider the properties of the PC estimator with the bounded maximum eigenvalue of $\mathbf{\Sigma}_x$, i.e., $\alpha_k = 0$ for all k in our WF specification.

We finally mention a large literature called *sparse principal component analysis* (sPCA), which introduces sparsity in the loadings of principal components by minimizing a penalized-regression-type criterion; see Zou et al. (2006), Shen and Huang (2008), among many others. The sPCA is related to but significantly different from ours in the following two points. First, it does not consider any factor model such as (3). Second, sPCA does not separately identify factors and loadings when $r > 1$. For example, the sPCA of Zou et al. (2006) can be interpreted that it estimates $\mathbf{B}\mathbf{f}_t$ as a predictor of $\mathbf{x}_t$, allowing sparsity in $\mathbf{B}$. However, they solve the problem imposing the $r(r+1)/2$ restrictions, $\mathbf{F}'\mathbf{F}/T = \mathbf{I}$, only. A similar comment applies to Shen and Huang (2008). We emphasize that this paper considers estimation of $\mathbf{F}^0$ and $\mathbf{B}^0$ in model (3) under relevant assumptions for economic and financial data, which requires very different mathematical proofs from those for sPCA. See Uematsu et al. (2019) for discussions on the relation between sPCA and SOFAR.

1.5 Organization and notational remarks

The rest of this paper is organized as follows. Section 2 formally defines the sWF models. Section 3 proposes the (adaptive) SOFAR estimator for the sWF models. Section 4 investigates the theoretical properties, including determination of the number of weak factors, the estimation error bounds of the SOFAR and PC estimators, and factor selection consistency. Section 5 confirms the validity of our method by Monte Carlo experiments. Section 6 gives three empirical illustrations. Section 7 concludes. All the proofs are collected in Supplementary Material.

For any matrix $\mathbf{M} = (m_{ti}) \in \mathbb{R}^{T\times N}$, we define the Frobenius norm, ℓ_2-induced (spectral) norm, entrywise ℓ_1-norm, and entrywise ℓ_∞-norm as $\|\mathbf{M}\|_{\mathrm{F}} = (\sum_{t,i} m_{ti}^2)^{1/2}$, $\|\mathbf{M}\|_2 = \lambda_1^{1/2}(\mathbf{M}'\mathbf{M})$, $\|\mathbf{M}\|_1 = \sum_{t,i}|m_{ti}|$, and $\|\mathbf{M}\|_{\max} = \max_{t,i}|m_{ti}|$, respectively, where $\lambda_i(\mathbf{S})$ refers to the ith largest eigenvalue of any symmetric matrix $\mathbf{S}$. We denote by $\mathbf{I}_N$ and $\mathbf{0}_{T\times N}$ the $N \times N$ identity matrix and $T \times N$ zero matrix, respectively. We use $\lesssim$ ($\gtrsim$) to represent $\leq$ ($\geq$) up to a positive constant factor. For any positive sequences a_n and b_n, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. For any positive values a and b, $a \vee b$ and $a \wedge b$ stand for $\max(a, b)$ and $\min(a, b)$, respectively. The indicator function is denoted by $1\{\cdot\}$.

2 Sparsity-Induced Weak Factor Models

Consider the factor model in (3) more precisely. Stacking the vectors vertically like $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)'$, $\mathbf{F}^0 = (\mathbf{f}_1^0, \ldots, \mathbf{f}_T^0)'$, and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_T)'$, we rewrite it as the matrix form

$$\mathbf{X} = \mathbf{F}^0\mathbf{B}^{0\prime} + \mathbf{E} = \mathbf{C}^0 + \mathbf{E}, \tag{5}$$

where $\mathbf{C}^0$ is called the matrix of common components. By the construction, the model satisfies the restrictions: $\mathbb{E}\,\mathbf{F}^{0\prime}\mathbf{F}^0/T = \mathbf{I}_r$ and $\mathbf{B}^{0\prime}\mathbf{B}^0$ is a diagonal matrix. Then the covariance matrix reduces to

$$\mathbf{\Sigma}_x = \mathbf{B}^0\mathbf{B}^{0\prime} + \mathbf{\Sigma}_e.$$

As discussed in Introduction, we consider *sparsity-induced* WF (sWF) models. Specifically, we assume sparse factor loadings $\mathbf{B}^0$ such that the sparsity of kth column (i.e., the number of nonzero elements in $\mathbf{b}_k^0 \in \mathbb{R}^N$) is $N_k := N^{\alpha_k}$ for $k \in \{1, \ldots, r\}$, where $1 \geq \alpha_1 \geq \cdots \geq \alpha_r > 0$

and exponents α_k's are unknown. Note that N_r must diverge since $\alpha_r > 0$ and $N \to \infty$. We may relax the *exact* sparseness by introducing the *approximate* sparse loadings; that is, $\mathbf{B}^0 = (b_{ik})$ such that $\sum_{i=1}^N |b_{ik}| \asymp N_k$. This does not necessarily require exact zeros in $\mathbf{B}^0$. However, we choose not to pursue this direction to avoid a complicated technical issue.

By the sparsity assumption and the diagonality of $\mathbf{B}^{0\prime}\mathbf{B}^0$, we can write

$$\mathbf{B}^{0\prime}\mathbf{B}^0 = \operatorname{diag}\left(d_1^2 N_1, \ldots, d_r^2 N_r\right)$$

with $d_1^2 N_1 \geq \cdots \geq d_r^2 N_r > 0$ for some positive constants $d_1, \ldots, d_r$. Then, under the assumption of uniform boundedness of $\lambda_j(\mathbf{\Sigma}_e)$, we have

$$\lambda_j(\mathbf{\Sigma}_x) \begin{cases} \asymp \lambda_j(\mathbf{B}^0\mathbf{B}^{0\prime}) = \lambda_j(\mathbf{B}^{0\prime}\mathbf{B}^0) = d_j^2 N_j & \text{for } \ j \in \{1, \ldots, r\}, \\ \text{is uniformly bounded} & \text{for } \ j \in \{r+1, \ldots, N\}. \end{cases}$$

Apparently, this specification fulfills the requirement of the WF structure (4).

For later use, we confirm the connection between $\mathbf{C}^0 = \mathbf{F}^0\mathbf{B}^{0\prime}$ and its singular value decomposition (SVD) $\mathbf{C}^0 = \mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}$. Here, $\mathbf{U}^0 \in \mathbb{R}^{T\times r}$ and $\mathbf{V}^0 \in \mathbb{R}^{N\times r}$ are respectively matrices of the (scaled) left- and sparse right-singular vectors of $\mathbf{C}^0$ that satisfy restrictions $\mathbf{U}^{0\prime}\mathbf{U}^0/T = \mathbf{I}_r$ and $\mathbf{V}^{0\prime}\mathbf{V}^0 = \mathbf{N}$ with $\mathbf{N} = \operatorname{diag}(N_1, \ldots, N_r)$, and $\mathbf{D}^0 = \operatorname{diag}(d_1, \ldots, d_r)$ is composed of the (scaled) singular values. In view of the restrictions on model (5), it is reasonable to set $\mathbf{F}^0 = \mathbf{U}^0$ and $\mathbf{B}^0 = \mathbf{V}^0\mathbf{D}^0$. This construction yields $\mathbf{F}^0\mathbf{B}^{0\prime} = \mathbf{C}^0$ and satisfies the restrictions.

3 Estimation

We propose our SOFAR estimator based on the SOFAR framework of Uematsu et al. (2019) for the WF models. In this section, we denote by $\hat{r}$ an estimate of the number of factors. The actual method of estimating r is introduced in Section 4.1.

3.1 SOFAR estimation

Once the sWF model is defined, it is natural to introduce a sparsity-inducing penalty term, such as the ℓ_1-norm of $\mathbf{B}$, to obtain a sparse estimate of $\mathbf{B}^0$ in the same fashion as the Lasso by Tibshirani (1996). The SOFAR estimator is defined as

$$(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}) = \underset{(\mathbf{F},\mathbf{B})\in\mathbb{R}^{T\times\hat{r}}\times\mathbb{R}^{N\times\hat{r}}}{\arg\min} \left\{ \frac{1}{2}\left\|\mathbf{X} - \mathbf{F}\mathbf{B}'\right\|_{\mathrm{F}}^2 + \eta\|\mathbf{B}\|_1 \right\} \tag{6}$$
$$\text{subject to } \mathbf{F}'\mathbf{F}/T = \mathbf{I}_{\hat{r}} \text{ and } \mathbf{B}'\mathbf{B} \text{ diagonal},$$

where $\hat{r}$ is the predetermined number of factors and $\eta > 0$ is a regularization coefficient. If $\eta = 0$ in (6), then the resulting estimator reduces to the PC estimator $(\widehat{\mathbf{F}}_{\mathsf{PC}}, \widehat{\mathbf{B}}_{\mathsf{PC}})$.

It is well-known that the PC estimator is easily obtained by the eigenvalue problem on $\mathbf{X}\mathbf{X}'$; specifically, for given $\hat{r}$, $\widehat{\mathbf{F}}_{\mathsf{PC}}$ is obtained as $T^{1/2}$ times the eigenvectors corresponding to the top $\hat{r}$ largest eigenvalues of $(NT)^{-1}\mathbf{X}\mathbf{X}'$ and $\widehat{\mathbf{B}}_{\mathsf{PC}} = \mathbf{X}'\widehat{\mathbf{F}}_{\mathsf{PC}}/T$. On the other hand, the SOFAR estimator is no longer computed by the eigenvalue problem. Even some algorithms used for the lasso, such as coordinate descent, cannot be directly applied to the problem due to the restrictions, sparsity and orthogonality (diagonality). In order to overcome this difficulty, we apply the SOFAR algorithm proposed by Uematsu et al. (2019) to solving (6). Roughly speaking, the algorithm provides estimates for the SVD of a coefficient matrix in a

multiple linear regression, with simultaneously exhibiting both low-rankness in the singular values matrix and sparsity in the singular vectors matrices. Recall the connection between $(\mathbf{F}, \mathbf{B})$ and $(\mathbf{U}, \mathbf{D}, \mathbf{V})$, which has been defined by the SVD of $\mathbf{C}$, in Section 2. Then for given $\hat{r}$, the SOFAR algorithm can solve (6) to get $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}) = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}\widehat{\mathbf{D}})$.

The algorithm to compute the SOFAR estimate is based on the *augmented Lagrangian method* coupled with the *block coordinate decent*, and is numerically stable. For detailed information on the algorithm, see Uematsu et al. (2019). The associated R package (rrpack) is available at https://cran.r-project.org/package=rrpack.

3.2 Adaptive SOFAR estimation

It is interesting to observe which factors truly contribute to x_{ti}. Expecting the true support recovery of $\mathbf{B}^0$, we introduce the *adaptive* SOFAR based on a similar principle of the adaptive lasso by Zou (2006). Let $\widehat{\mathbf{B}}^{\mathsf{ini}} = (\hat{b}_{ij}^{\mathsf{ini}})$ denote the first-stage initial estimator, such as the PC estimator. Then the (i,j)th element of the weighting matrix $\mathbf{W} = (w_{ij})$ is defined as $w_{ij} = 1/|\hat{b}_{ij}^{\mathsf{ini}}|$. The adaptive SOFAR estimator is defined as a minimizer of the second-stage weighted SOFAR problem:

$$(\widehat{\mathbf{F}}^{\mathsf{ada}}, \widehat{\mathbf{B}}^{\mathsf{ada}}) = \underset{(\mathbf{F},\mathbf{B})\in\mathbb{R}^{T\times\hat{r}}\times\mathbb{R}^{N\times\hat{r}}}{\arg\min} \left\{ \frac{1}{2} \left\| \mathbf{X} - \mathbf{F}\mathbf{B}' \right\|_{\mathrm{F}}^2 + \eta \|\mathbf{W} \circ \mathbf{B}\|_1 \right\} \tag{7}$$
$$\text{subject to } \mathbf{F}'\mathbf{F}/T = \mathbf{I}_{\hat{r}} \text{ and } \mathbf{B}'\mathbf{B} \text{ diagonal},$$

where $\mathbf{A} \circ \mathbf{B}$ represents the Hadamard product of two matrices, $\mathbf{A}$ and $\mathbf{B}$, of the same size.

Estimating exponents α_k's is of great interest to empirical research since, as discussed in Bailey et al. (2016), they are interpreted as the strength of the influence of the common factors and of the cross-sectional correlations. Recall that the kth column of $\mathbf{B}^0$, $\mathbf{b}_k^0$, has $N_k = N^{\alpha_k}$ nonzero entries. Similarly, let $\widehat{N}_k$ denote the number of nonzero elements in $\widehat{\mathbf{b}}_k^{\mathsf{ada}}$. As the lasso in a linear regression, we may expect that the adaptive SOFAR estimate $\widehat{\mathbf{B}}^{\mathsf{ada}}$ can successfully recover the true sparsity pattern of $\mathbf{B}^0$. If this is true, the estimators of exponents α_k's can naturally be obtained as $\hat{\alpha}_k = \log \widehat{N}_k / \log N$ by a simple algebraic formulation. In the next section, we will prove this estimator is actually consistent for α_k.

4 Theory

We first reveal the asymptotic behavior of the eigenvalues of $\mathbf{X}\mathbf{X}'$ for the sWF model in Section 4.1. This helps us to determine the number of factors. Next we derive the estimation error bound in Section 4.2. Furthermore, the asymptotic property of the adaptive SOFAR estimator is derived in Section 4.3.

For the sake of convenience, we assume the existence of some underlying sequence n that satisfies the principle that N and T are both functions of n and that they simultaneously diverge as $n \to \infty$ (i.e., $N = N(n) \to \infty$ and $T = T(n) \to \infty$ as $n \to \infty$). For example, we may simply suppose $n = N \wedge T \to \infty$. Furthermore, following Rigollet and Hütter (2017), we introduce a sub-Gaussian random variable: a random variable $X \in \mathbb{R}$ is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies $\mathbb{E}[\exp(sX)] \le \exp(\sigma^2 s^2/2)$ for all $s \in \mathbb{R}$. This is denoted by $X \sim \mathrm{subG}(\sigma^2)$. Define $L_n = (N \vee T)^\nu - 1$ for an arbitrary constant $\nu > 0$. Throughout the paper, including all the proofs in Appendix, ν is assumed to be fixed, and n is sufficiently large.

Assumption 1 (Latent factors)**.** The factor matrix $\mathbf{F}^0 = (\mathbf{f}_1^0, \ldots, \mathbf{f}_T^0)'$ is specified as the vector moving average process of order L_n (VMA(L_n)) such that

$$\mathbf{f}_t^0 = \sum_{\ell=0}^{L_n} \mathbf{\Psi}_\ell \boldsymbol{\zeta}_{t-\ell}, \qquad \lim_{n\to\infty} \sum_{\ell=0}^{L_n} \mathbf{\Psi}_\ell \mathbf{\Psi}_\ell' = \mathbf{I}_r,$$

where $\boldsymbol{\zeta}_t = (\zeta_{t1}, \ldots, \zeta_{tr})'$ with $\{\zeta_{tk}\}_{t,k}$ i.i.d. subG(σ_ζ^2) that has $\mathbb{E}\,\zeta_{tk}^2 = 1$, and where $\mathbf{\Psi}_0$ is a nonsingular, lower triangular matrix.

Assumption 2 (Factor loadings)**.** Each column $\mathbf{b}_k^0$ of $\mathbf{B}^0$ has the sparsity $N_k = N^{\alpha_k}$ with $0 < \alpha_r \le \cdots \le \alpha_1 \le 1$ and $\mathbf{B}^{0\prime}\mathbf{B}^0 = \mathrm{diag}\{d_1^2 N_1, \ldots, d_r^2 N_r\}$ with $0 < d_r N_r^{1/2} \le \cdots \le d_1 N_1^{1/2}$. For k such that $\alpha_k = \alpha_{k-1}$, it holds that $d_{k-1}^2 - d_k^2 \ge \delta^{1/2} d_{k-1}^2$ for some constant $\delta > 0$.

Assumption 3 (Idiosyncratic errors)**.** The error matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_T)'$ is specified as the VMA(L_n) such that

$$\mathbf{e}_t = \sum_{\ell=0}^{L_n} \mathbf{\Phi}_\ell \boldsymbol{\varepsilon}_{t-\ell}, \qquad \limsup_{n\to\infty} \sum_{\ell=0}^{L_n} \|\mathbf{\Phi}_\ell\|_2 < \infty,$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \ldots, \varepsilon_{tN})'$ with $\{\varepsilon_{ti}\}_{t,i}$ i.i.d. subG(σ_ε^2) and $\mathbf{\Phi}_0$ is a nonsingular, lower triangular matrix.

Assumption 4 (Parameter space)**.** The parameter space of $\mathbf{B}$ in optimization (6) is given by $\mathcal{B}(\widetilde{N}) = \{\mathbf{B} \in \mathbb{R}^{N\times r} : \|\mathbf{B}\|_0 \lesssim \widetilde{N}/2\}$ for $\widetilde{N} \in [N_1, N]$. (Define $\tilde{\alpha}$ to be such that $\widetilde{N} = N^{\tilde{\alpha}}$.)

These assumptions are quite different from those on the original SOFAR theory by Uematsu et al. (2019), which consider a multiple regression with deterministic regressors and i.i.d. errors. Assumptions 1 and 3 specify the stochastic processes $\{\mathbf{f}_t\}$ and $\{\mathbf{e}_t\}$, respectively, to be stationary VMA(L_n), where $L_n \asymp (N \vee T)^\nu$ diverges with an arbitrary fixed constant $\nu > 0$. This construction is regarded as the *asymptotic linear process*, which includes a wide range of multivariate weakly dependent processes. Assumption 2 is key to our analysis and provides the sWF models. The sparsity makes the divergence rate of $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0)$ possibly slower than N. Assumption 4 is required only when the parameter estimation is considered. If $\widetilde{N}$ is set to N, $\mathcal{B}(N)$ coincides with the whole space, $\mathbb{R}^{T\times r}$. Whereas, if $\widetilde{N}$ is set to N_1, $\mathcal{B}(N_1)$ becomes as sparse as $\mathbf{B}^0$. The PC estimator always requires optimization on $\mathcal{B}(N)$ since it cannot be sparse, but the SOFAR estimator can allow sparse $\mathcal{B}(\widetilde{N})$ with $\widetilde{N} \in [N_1, N)$ when the true loadings matrix is expected to be sparse. An important consequence of taking sparser space is that, as explained in Section 4.2, a wider class of the sWF models can be allowed in consistent estimation.

4.1 Determining the number of factors

Before investigating the properties of the estimator, we first observe the asymptotic behavior of the eigenvalues of $\mathbf{X}\mathbf{X}'$ under the sWF model. This result yields important information for determining the number of weak factors, r. Write $T = N^\tau$ for some constant $\tau > 0$ to understand the size of T relative to N. Recall that $N_k = N^{\alpha_k}$ for some $\alpha_k \in (0, 1]$.

Theorem 1. *Suppose that Assumptions 1–3 and condition*

$$\alpha_1 < 2\alpha_r \tag{8}$$

hold. Then for any finite integer $k_{\max} > r$*, the* j*th largest eigenvalue of* $(N \vee T)^{-1}\mathbf{X}\mathbf{X}'$*, denoted by* λ_k*, satisfies*

$$\lambda_k \begin{cases} \gtrsim \dfrac{N_k T}{N \vee T} & \text{for } k \in \{1, \ldots, r\}, \\ = O(1) & \text{for } k \in \{r+1, \ldots, k_{\max}\}, \end{cases}$$

with probability at least $1 - O((N \vee T)^{-\nu})$*. Divergence of* λ_r *is ensured by condition*

$$\alpha_r + \tau > 1. \tag{9}$$

Theorem 1 suggests the means of determining the number of weak factors. This presents a case in which the method of Onatski (2010) works. Namely, for $\delta > 0$, define

$$\hat{r}(\delta) = \max\{k = 1, \ldots, k_{\max} - 1 : \lambda_k - \lambda_{k+1} \geq \delta\}.$$

Then, the following important corollary is obtained.

Corollary 1. *Suppose that Assumptions 1–3 hold. If conditions (8) and (9) are true, then for any fixed positive constant* δ*, we have* $\hat{r}(\delta) \to r$ *with probability at least* $1 - O((N \vee T)^{-\nu})$*.*

In practice, δ should appropriately be predetermined. In fact, Onatski (2010) suggested the *edge distribution* (ED) method based on a calibration; see that paper for full details. If δ is appropriately chosen, $\hat{r}(\delta)$ will successfully detect the true number of factors r even when the biggest gap is observed not between λ_r and λ_{r+1} but among $\lambda_1, \ldots, \lambda_r$. Meanwhile, the method of Ahn and Horenstein (2013), which was designed for SF models, is likely to fail in detecting r in the WF models because it defines $\hat{r}$ as the point at which the largest gap is observed among $\lambda_1, \ldots, \lambda_{k_{\max}}$; this is not always the case for the WF models. In Section 5, we will check the validity of Onatski's ED estimator in our model through numerical simulations.

4.2 Estimation error bound

We suppose that the sWF model satisfies conditions (8) and (9) and that r is known in view of Corollary 1. Recall that $\widetilde{N} = N^{\tilde{\alpha}}$ (see Assumption 4), and introduce an additional condition that restricts the class of sWF models:

$$\alpha_1 + (\tilde{\alpha} \vee \tau)/2 < \alpha_r + \alpha_r \wedge \tau. \tag{10}$$

This condition is necessary to derive a nontrivial error bound. Note that condition (10) with any $\tilde{\alpha} \in [\alpha_1, 1]$ implies (8) because $\alpha_1 < \alpha_r + \alpha_r \wedge \tau - (\tilde{\alpha} \vee \tau)/2 < \alpha_r + \alpha_r \wedge \tau \leq 2\alpha_r$. For notational convenience, we put $K_n = \{N_1 \log^{1/2}(N \vee T)\}/\{N_r(N_r \wedge T)\}$.

Theorem 2 (SOFAR)**.** *Set* $\eta_n \asymp T^{1/2} \log^{1/2}(N \vee T)$ *in optimization (6). If Assumptions 1–4 and conditions (9) and (10) hold with any* $\widetilde{N} \in [N_1, N]$ *(i.e.,* $\tilde{\alpha} \in [\alpha_1, 1]$*), then the following error bounds hold with probability at least* $1 - O((N \vee T)^{-\nu})$*:*

$$T^{-1/2}\|\widehat{\mathbf{F}} - \mathbf{F}^0\|_{\mathrm{F}} \lesssim N_1^{1/2} K_n, \quad N^{-1/2}\|\widehat{\mathbf{B}} - \mathbf{B}^0\|_{\mathrm{F}} \lesssim \frac{N_1^{1/2} T^{1/2}}{N^{1/2}} K_n.$$

In particular, the upper bounds converge to zero.

The convergence rates do not depend on the choice of $\widetilde{N}$. Through condition (10), however, it provides a class of the sWF models that can consistently be estimated. In fact, the range of α_r restricted by (10) becomes the largest when $\widetilde{N} = N_1$ (i.e., $\tilde{\alpha} = \alpha_1$). This point is reconsidered in Remark 1 below in comparison with the PC estimator.

Theorem 3 (PC). *If Assumptions 1–4 and conditions (9) and (10) hold with $\widetilde{N} = N$ (i.e., $\tilde{\alpha} = 1$), then the following error bounds hold with probability at least $1 - O((N \vee T)^{-\nu})$:*

$$T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}} \lesssim N^{1/2}K_n, \quad N^{-1/2}\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\mathrm{F}} \lesssim T^{1/2}K_n.$$

In particular, the upper bounds converge to zero.

First, when the model has strong factors only (i.e., $N_r = N$), the convergence rates in the theorems correspond to that obtained from Bai (2003) up to the logarithmic factor. We also observe that the convergence rates of the SOFAR and the PC estimators become identical if $N_1 = N$. On the other hand, when the model has weak factors with $N_1 < N$, the SOFAR can take advantage of utilizing the sparsity and achieve a tighter upper bound. Therefore, the SOFAR estimator is likely to converge at least as fast as the PC estimator even when all the factors are strong. Of course a precise discussion requires a lower bound, but it is beyond the scope of this paper and left for a future study.

While the SOFAR can choose $\widetilde{N} = N_1$ as already mentioned, the PC necessarily selects $\widetilde{N} = N$ since it does not exploit sparse parameter spaces. In view of model restriction (10), this leads to the fact that the SOFAR can consistently estimate a wider class of the sWF models than the PC can. For a more detailed discussion, see the following remark.

Remark 1. We are interested in the class of sWF models that can consistently be estimated by the SOFAR and the PC, respectively. Condition (10) with $\widetilde{N} = N_1$ (i.e., $\tilde{\alpha} = \alpha_1$) naturally brings the largest class of the sWF models. In this case, the lower bound of α_r is $1/3$, which is achievable when $\alpha_1 = \alpha_r$ and $\tau = 2/3$. Likewise, the upper bound of the difference $\alpha_1 - \alpha_r$ is found to be $1/4$, which is attainable when $\tau \in (3/4, 1]$ and $\alpha_1 = 1$. Note that these results can be obtained not by PC but by SOFAR. Contrary to the case of $\widetilde{N} = N_1$, condition (10) with $\widetilde{N} = N$ restricts α_r to be strictly larger than $1/2$. This is more restrictive than the case of $\widetilde{N} = N_1$ though the upper bound of the difference is the same.

In sum, the SOFAR can consistently estimate the sWF models with exponents α_k's smaller than or equal to $1/2$ by supposing a sparse parameter space. The finite sample evidence in Section 5 shows that the SOFAR estimator seems quite robust to the violation of the restrictions on the region of $(\tau, \alpha_1, \alpha_r)$ discussed in Remark 1.

4.3 Factor selection consistency

We prove the *factor selection consistency*, which guarantees that the adaptive SOFAR recovers the true sparsity pattern of the loadings and correctly selects the relevant factors. As a corollary, we also establish consistency of the estimated exponents, $\hat{\alpha}_k$'s.

Before stating the theorem, define $\mathcal{S} = \mathrm{supp}(\mathbf{B}^0) \subset \{1, \ldots, N\} \times \{1, \ldots, r\}$, the index set of nonzero signals in $\mathbf{B}^0$. For any (sparse) matrix $\mathbf{A} = (a_{ik}) \in \mathbb{R}^{N \times r}$, we define $\mathbf{A}_{\mathcal{S}} = (a_{ik}1\{(i,k) \in \mathcal{S}\})$ and $\mathbf{a}_{\mathcal{S}} = \mathrm{vec}\,\mathbf{A}_{\mathcal{S}} \in \mathbb{R}^{rN}$. Write $\underline{b}_n^0 = \min_{(i,k)\in\mathcal{S}} |b_{ik}^0|$. Furthermore, introduce additional conditions:

$$\alpha_1 - \alpha_r < \tau/4, \tag{11}$$

$$1 \lesssim \frac{\eta_n/\underline{b}_n^0}{T^{1/2}\log^{1/2}(N \vee T)} \lesssim \frac{N_1(N_r \vee T)}{N_r(N_r \wedge T)}. \tag{12}$$

Condition (11) further restricts the model in terms of the maximum difference of α_1 and α_r when $\tau < 1$. However, the difference can be $1/4$, which is the same as the maximum value obtained by constraint (10) only, as long as $\tau = 1$. Condition (12) restricts the relation between η_n and $\underline{b}_n^0$.

Theorem 4 (Adaptive SOFAR)**.** *If Assumptions 1–3 and conditions (9)–(12) hold, then for the weighting matrix* $\mathbf{W}$ *constructed by any estimator* $\widehat{\mathbf{B}}^{\mathsf{ini}}$ *such that*

$$\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\|_{\max} \lesssim \underline{b}_n^0 \quad \textit{(with high probability)}, \tag{13}$$

the adaptive SOFAR estimator satisfies

$$T^{-1/2} \left\| \widehat{\mathbf{F}}^{\mathsf{ada}} - \mathbf{F}^0 \right\|_{\mathrm{F}} = O_p\left(N_1^{1/2} K_n\right), \tag{14}$$

$$N^{-1/2} \left\| \widehat{\mathbf{B}}_{\mathcal{S}}^{\mathsf{ada}} - \mathbf{B}_{\mathcal{S}}^0 \right\|_{\mathrm{F}} = O_p\left(\frac{N_1^{1/2} T^{1/2}}{N^{1/2}} K_n\right), \tag{15}$$

$$\mathbb{P}\left(\mathrm{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}}) = \mathcal{S}\right) \to 1. \tag{16}$$

If the PC estimator is used for the initial estimator, $\underline{b}_n^0 \gtrsim T^{-1/2} \log^{1/2}(N \vee T)$ is allowed in (13) (see Lemma 6 in Appendix). The rates of convergence (14) and (15) are identical to those in Theorem 2, and hence they converge to zero. Finally, we prove that $\hat{\alpha}_k = \log \widehat{N}_k / \log N$, which is defined in Section 3.2, is consistent for α_k because of (16).

Corollary 2. *If the model selection consistency in (16) holds, then we have*

$$\mathbb{P}\left(\hat{\alpha}_k = \alpha_k \textit{ for all } k = 1, \ldots, r\right) \to 1.$$

It is well-known that the adaptive Lasso can establish the asymptotic normality for the nonzero subvector of the estimator. Likewise, the asymptotic normality of the adaptive SOFAR estimator might be proved. However, we do not consider it due to the criticism by Leeb and Pötscher (e.g., Leeb and Pötscher (2008) and references therein). Instead, it is interesting to investigate inferential theory based on "debiasing" the SOFAR estimator in a manner similar to Javanmard and Montanari (2014). This direction is explored in Uematsu and Yamagata (2020).

5 Monte Carlo Experiments

We investigate thee Monte Carlo experiments. In this section, indexes i, t, and k run over $1, \ldots, N$, $1, \ldots, T$, and $1, \ldots, r$, respectively, unless otherwise noted. We consider the Data Generating Process (DGP), $x_{ti} = \sum_{k=1}^r b_{ik} f_{tk} + \sqrt{\theta} e_{ti}$. The factor loadings b_{ik} and factors f_{tk} are formed such that $N^{-1} \sum_{i=1}^N b_{ik} b_{i\ell} = 1\{k = \ell\}$ and $T^{-1} \sum_{t=1}^T f_{tk} f_{t\ell} = 1\{k = \ell\}$, by applying Gram–Schmidt orthonormalization to b_{ik}^* and f_{tk}^*, respectively, where $b_{ik}^* \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, N_k$ and $b_{ik}^* = 0$ for $i = N_k + 1, \ldots, N$, and $f_{tk}^* = \rho_{fk} f_{t-1,k}^* + v_{tk}$ with $v_{kt} \sim$ i.i.d.$N(0, 1 - \rho_{fk}^2)$ and $f_{0k}^* \sim$ i.i.d.$N(0,1)$. The idiosyncratic errors e_{ti} are generated by $e_{ti} = \rho_e e_{t-1,i} + \beta \varepsilon_{t,i-1} + \beta \varepsilon_{t,i+1} + \varepsilon_{ti}$, where $\varepsilon_{ti} \sim$ i.i.d.$N(0, \sigma_{\varepsilon,ti}^2)$ with $\sigma_{\varepsilon,ti}^2$ being set such that $\mathrm{Var}(e_{ti}) = 1$. The DGP is in line with the existing representative literature, such as Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013), among many others, but the main difference is that the absolute sums of the factor loadings over i are allowed to diverge proportionally to $N_k = N^{\alpha_k}$.

As the benchmark DGP, we set $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for all k, $\beta = 0.2$, and $\theta = 1$. We focus on the performance of the estimators for different values of exponents (α_1, α_2). In particular, we consider the combinations $(0.9, 0.9)$, $(0.8, 0.5)$[5] and $(0.5, 0.4)$. All the experimental results are based on 1,000 replications.

5.1 Determining the number of weak factors

Based on Corollary 1 and the discussion in Section 4.1, we confirm validity of $\hat{r}(\delta)$. As already explained, the estimator is the maximum value of j with which $\lambda_k - \lambda_{k+1}$ exceeds the threshold δ. Following the *ED* algorithm of Onatski (2010), we compute $\hat{\delta}$ by calibration.[6] The other competitor statistics include the *ER* (eigenvalue ratio) and *GR* (growth ratio) estimators of Ahn and Horenstein (2013). We also consider the information criteria IC_3 and BIC_3 proposed by Bai and Ng (2002). Note that these competitors are designed for SF models. Especially, the *ER* and *GR* just identify the maximum gap between the ordered eigenvalues. Hence, when the gap, $\lambda_k - \lambda_{k+1}$, is relatively large, these statistics will pick up k as the estimate of r even when $k < r$.

Table 1 reports the average of the estimated number of factors over the replications by the *ED*, *GR*, and BIC_3.[7] We set the maximum number of factors, $k_{\max}$, as five. As can be seen in Table 1, when α_1 and α_2 are both close to unity, all the methods perform well; see the case of exponents $(\alpha_1, \alpha_2) = (0.9, 0.9)$. However, the performance of *GR* and BIC_3 deteriorates when the gap of the values between α_1 and α_2 widens, or when both values α_1 and α_2 are further away from unity; e.g., see the cases when $(\alpha_1, \alpha_2) = (0.8, 0.5)$ and $(\alpha_1, \alpha_2) = (0.5, 0.4)$. In contrast, *ED* performs very well, and its estimation quality is very similar to that when both exponents are close to unity. Even under the most challenging set up $(\alpha_1, \alpha_2) = (0.5, 0.4)$, *ED* consistently estimates the number of factors for sufficiently large T and N.

We conclude that the finite sample evidence suggests that the *ED* method of Onatski (2010) provides a reliable estimation of the number of factors in sWF models, while the methods of *GR* and BIC_3 may not be as reliable as the *ED* in general.

5.2 Finite sample properties of the SOFAR estimator

We investigate the finite sample properties of our SOFAR estimator in comparison with the PC estimator. Here we treat the number of factors, r, as given. We report the results of the adaptive SOFAR estimator with regularization coefficient η determined by BIC, which we recommend to use.[8] For performance comparison purposes, we consider the ℓ_2-norm losses based on the scaled estimators: $L(\widehat{\mathbf{F}}) = \| \sum_{k=1}^{r} T^{-1/2}[\mathrm{abs}(\widehat{\mathbf{f}}_k) - \mathrm{abs}(\mathbf{f}_k^0)] \|_2$, $L(\widehat{\mathbf{B}}) = \| \sum_{k=1}^{r} N_k^{-1/2}[\mathrm{abs}(\widehat{\mathbf{b}}_k) - \mathrm{abs}(\mathbf{b}_k^0)] \|_2$, and $L(\widehat{\mathbf{C}}) = \| \sum_{k=1}^{r} T^{-1/2} N_k^{-1/2}[\widehat{\mathbf{C}}_k - \mathbf{C}_k^0] \|_{\mathrm{F}}$, where $\mathrm{abs}(\mathbf{a})$ takes elementwise absolute value of a real vector $\mathbf{a}$. Due to the scaling, the performance of the estimators can be comparable across different combinations of the values of N, T, and α_k's.

[5] When $\alpha_1 = 0.8$, the smallest value of α_r implied by condition (10) is 0.6, which is much larger than 0.5.

[6] We have found no experimental results on the finite sample performance of the ED estimator with the WF models apart from ours.

[7] To save the space, we do not report the results for ER and IC_3 since the performance of ER is very similar to that of GR, and the performance of IC_3 is mostly outperformed by BIC_3. These results are available upon request from the authors.

[8] We examined all the combinations of SOFAR and adaptive SOFAR with AIC, cross-validation, BIC and GIC. The results of which are available upon request from the authors.

Table 1: Average of the chosen number of factors for sWF models by ED, GR, and BIC_3

	ED				GR				BIC_3			
T, N	100	200	500	1000	100	200	500	1000	100	200	500	1000
$(\alpha_1, \alpha_2) = (0.9, 0.9)$												
100	2.05	2.04	2.02	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
200	2.04	2.04	2.03	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
500	2.04	2.04	2.03	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1000	2.02	2.04	2.03	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
$(\alpha_1, \alpha_2) = (0.8, 0.5)$												
100	1.96	1.96	1.95	1.90	1.30	1.18	1.04	1.00	1.30	1.17	1.02	1.00
200	2.02	2.02	2.03	2.02	1.40	1.30	1.09	1.01	1.39	1.36	1.12	1.01
500	2.03	2.03	2.02	2.02	1.61	1.45	1.24	1.10	1.41	1.51	1.53	1.42
1000	2.02	2.03	2.02	2.02	1.52	1.45	1.24	1.10	1.43	1.51	1.53	1.42
$(\alpha_1, \alpha_2) = (0.5, 0.4)$												
100	1.54	1.52	1.36	1.14	1.50	1.47	1.39	1.33	1.03	1.00	1.00	1.00
200	1.83	1.88	1.89	1.86	1.52	1.53	1.50	1.39	1.03	1.02	1.00	1.00
500	2.00	2.00	2.01	2.02	1.67	1.64	1.65	1.59	1.03	1.05	1.02	1.01
1000	1.92	2.00	2.01	2.02	1.60	1.64	1.65	1.59	1.04	1.05	1.02	1.01

Table 2 reports the averages and standard deviations (s.d.) of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ based on Corollary 2, and the average of the norm losses of the scaled estimated factors, factor loadings, and common components by the SOFAR (SO in the tables) and PC estimators over the replications. First, focus on $(\hat{\alpha}_1, \hat{\alpha}_2)$. In a nutshell, they are sufficiently accurate but tend to slightly underestimate when α_k is closer to one and overestimate when it is around 0.5. The precision improves as T and N increase. For example, see the results when $(\alpha_1, \alpha_2) = (0.8, 0.5)$. Now we turn to the performance of the SOFAR and PC estimates. In terms of the norm loss given above, the SOFAR uniformly beats the PC across all the designs. Perhaps surprisingly, the SOFAR estimate of the factors is much more accurate than the PC even in the most favorable experimental design to the PC, with $(\alpha_1, \alpha_2) = (0.9, 0.9)$. As expected, moreover, the accuracy of the SOFAR estimates of factor loadings is uniformly superior to that of the PC estimates. This gap in accuracy becomes wider when the exponents are further from unity. Consequently, the accuracy of the SOFAR estimator of common component is uniformly superior to that of the PC estimator.

Table 3 reports the same information as Table 2, but for more challenging models with $(0.5, 0.4)$. Remarkably, even when one of the exponent is 0.4, our SOFAR method provides sufficiently accurate estimates of α_1 and α_2 as well as far superior estimates of the factors, factor loadings, and common components to the PC method.

To summarize, the SOFAR estimator performs very well when the exponents are close to unity, thus, signal of common components is high, even with a smaller sample size. When the signal of common components is weak, namely when the value(s) of exponent(s) are around 0.5 or below, the SOFAR estimator is sufficiently precise in terms of norm loss, but requires a larger sample size. Significantly, even when the gap between α_1 and α_2 is larger than that condition (10) implies, the SOFAR estimator is sufficiently accurate, and its accuracy improves as the sample size rises. Conversely, the PC estimator fails to improve the performance when N rises due to its inability to identify zero elements in sparse loadings, and consequently the PC estimator is uniformly superseded by the SOFAR estimator.

Table 2: Performance of the SOFAR (SO) and PC estimators for approximate factor models with two factor components with $(\alpha_1, \alpha_2) = (0.9, 0.9), (0.8, 0.5)$

	T=100				**T=200**				**T=500**				**T=1000**			
Design (α_1, α_2)	$(0.9, 0.9)$		$(0.8, 0.5)$		$(0.9, 0.9)$		$(0.8, 0.5)$		$(0.9, 0.9)$		$(0.8, 0.5)$		$(0.9, 0.9)$		$(0.8, 0.5)$	
N=100	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.86	0.02	0.75	0.03	0.87	0.01	0.76	0.02	0.88	0.01	0.78	0.02	0.89	0.01	0.78	0.01
$\hat{\alpha}_2$	0.85	0.02	0.52	0.07	0.86	0.02	0.52	0.06	0.88	0.01	0.51	0.05	0.88	0.01	0.51	0.04
	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	6.2	11.6	13.8	21.8	5.1	7.8	13.1	17.0	4.2	5.3	12.8	14.4	3.9	4.5	12.3	13.1
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	9.0	9.9	10.4	38.2	4.7	5.5	4.8	19.2	2.2	2.6	2.1	8.2	1.4	1.6	1.2	4.4
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	8.2	14.5	20.9	50.6	5.6	8.7	16.5	31.2	4.1	5.5	14.4	20.5	3.6	4.3	13.6	16.7
N=200	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.86	0.01	0.75	0.02	0.87	0.01	0.76	0.01	0.88	0.01	0.78	0.01	0.89	0.01	0.78	0.01
$\hat{\alpha}_2$	0.86	0.01	0.52	0.05	0.87	0.01	0.51	0.04	0.88	0.01	0.50	0.03	0.89	0.01	0.50	0.03
	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	4.6	10.1	10.4	19.5	3.5	6.4	9.2	13.6	2.8	4.1	8.8	10.5	2.5	3.1	8.7	9.5
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	9.1	10.4	10.0	50.0	4.7	5.7	4.5	24.2	2.2	2.8	1.8	9.7	1.4	1.6	1.0	5.0
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	6.8	13.1	16.4	56.8	4.1	7.5	12.1	31.6	2.6	4.0	10.1	17.8	2.1	2.9	9.5	13.3
N=500	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.87	0.01	0.75	0.01	0.88	0.01	0.77	0.01	0.88	0.00	0.78	0.01	0.89	0.00	0.78	0.01
$\hat{\alpha}_2$	0.86	0.01	0.52	0.04	0.87	0.01	0.51	0.03	0.88	0.00	0.51	0.02	0.89	0.00	0.50	0.02
	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	3.5	9.3	7.0	18.8	2.3	5.6	6.0	11.2	1.8	3.2	5.5	7.3	1.5	2.2	5.3	6.2
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	9.4	11.2	10.8	74.8	4.5	6.0	4.6	35.0	2.2	3.0	1.6	13.5	1.3	1.7	0.8	6.7
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	6.1	12.7	13.4	76.0	3.3	6.9	8.9	37.6	1.7	3.2	6.5	17.4	1.2	2.0	5.9	11.3
N=1000	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.87	0.01	0.76	0.01	0.88	0.00	0.77	0.01	0.89	0.00	0.78	0.00	0.89	0.00	0.79	0.00
$\hat{\alpha}_2$	0.86	0.01	0.53	0.03	0.87	0.00	0.51	0.03	0.88	0.00	0.51	0.02	0.89	0.00	0.51	0.02
	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	2.8	9.0	5.2	20.1	1.9	5.4	4.3	10.6	1.4	2.9	3.8	5.7	1.2	2.0	3.6	4.5
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	9.4	12.0	11.5	101.8	4.7	6.5	4.8	46.8	2.1	3.1	1.7	17.5	1.3	1.9	0.8	8.6
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	6.0	12.7	12.3	99.6	3.0	6.8	7.2	46.3	1.4	2.9	4.8	19.0	0.9	1.7	4.1	11.0

Table 3: Performance of the SOFAR (SO) and PC estimators for approximate factor models with two factor components with $(\alpha_1, \alpha_2) = (0.5, 0.4)$

	T=500		**T=1000**	
Design (α_1, α_2)	$(0.5, 0.4)$		$(0.5, 0.4)$	
N=500	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.47	0.03	0.47	0.03
$\hat{\alpha}_2$	0.41	0.04	0.40	0.04
	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	13.4	17.9	13.1	15.2
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	4.6	48.3	2.9	24.4
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	17.3	48.6	16.0	31.1

	T=500		**T=1000**	
Design (α_1, α_2)	$(0.5, 0.4)$		$(0.5, 0.4)$	
N=1000	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.48	0.02	0.48	0.02
$\hat{\alpha}_2$	0.40	0.03	0.40	0.03
	SO	PC	SO	PC
$L_F^2(\hat{\mathbf{F}})_{\times 100}$	9.7	15.2	9.5	12.0
$L_F^2(\hat{\mathbf{\Lambda}})_{\times 100}$	3.7	65.6	2.3	32.2
$L_F^2(\hat{\mathbf{C}})_{\times 100}$	13.0	57.4	12.0	32.9

5.3 A hierarchical factor structure

Recently estimation of a hierarchical factor structure or a multi-level factor structure has been gaining serious interest in the literature. Ando and Bai (2017) and Choi et al. (2018) consider two types of factors, called global and local factors. The global factors have the loadings with non-zero values for all the cross-section units, whereas the local factors have the non-zero loadings among the cross-section units of some specific groups. They propose sequential procedures to identify the global and local factors separately.[9] In fact, the sWF model nests the hierarchical factor structure, and hence our SOFAR method can be readily applied. In contrast to the existing approaches, given the total number of global and local factor, our approach permits us to consistently estimate the hierarchical model in one go. Furthermore, our method can identify "near global" (or "near local") factors as the strongest, which influence many but not all the variables; see Section 6.2 for the evidence of such factors. The near global factors may not be distinguished from the global (or strictly strong) factors by the aforementioned existing methods.

For illustration, We generate the data of four factors models as above. The first factor is global, i.e., $b_{i1} \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, N$. The other three factors are local, i.e., b_{i2} is drawn from $N(0,1)$ for the first third, b_{i3} for the second third, and b_{i4} for the last third of cross section units while the rests are zero. We obtained a simulated data with $N = 450$ and $T = 120$, and estimated the model given $r = 4$ by the PC and SOFAR. To visualize the quality of the factor loadings, we provide heat maps of three $N \times N$ matrices, $\sum_{k=1}^4 \omega_k \mathbf{b}_k^0 \mathbf{b}_k^{0\prime}$, $\sum_{k=1}^4 \omega_k \widehat{\mathbf{b}}_k \widehat{\mathbf{b}}_k'$ and $\sum_{k=1}^4 \omega_k \widehat{\mathbf{b}}_{\mathsf{PC},k} \widehat{\mathbf{b}}_{\mathsf{PC},k}'$, which are reported in Figures 1-3, respectively. To clarify the difference between the global factor loadings and local ones, which overlaps in the heat maps, we use the weight $\omega_1 = 1/8$ and $\omega_2 = \omega_3 = \omega_4 = 1$. As is clear, the SOFAR estimator successfully recover the hierarchical pattern while the PC estimator fails.

6 Empirical Applications

We provide three empirical applications. Section 6.1 conducts forecasting bond yields with comparing the SOFAR and PC. Section 6.2 investigates interpreting the extracted factors in Section 6.1. Section 6.3 considers estimation of the exponents based on the adaptive SOFAR with a large number of stock returns.

[9] Andreou et al. (2019) propose a similar sequential method to estimate the number of global and local factors separately.

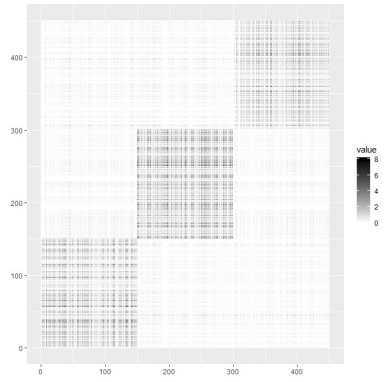

Figure 1: True factor loadings

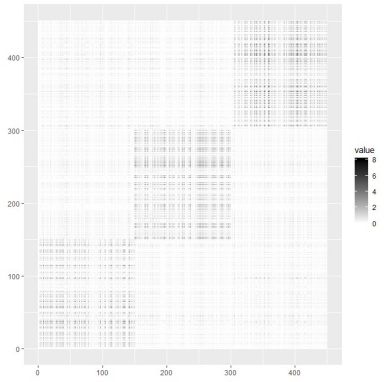

Figure 2: SOFAR estimate

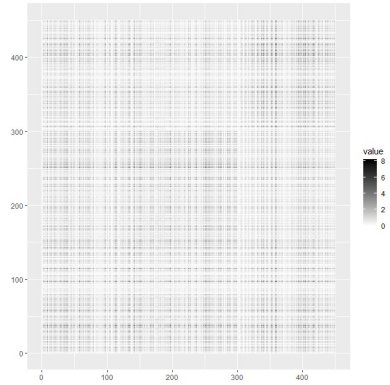

Figure 3: PC estimate

6.1 Forecasting bond yields

We consider out-of-sample forecasting of bond yields using extracted factors via our SOFAR and the PC, from a large number of macroeconomic variables in line with Ludvigson and Ng (2009). We use the same data set provided by Sydney Ludvigson's web page.[10] Specifically, the data consists of the continuously compounded (log) annual excess returns on an n-year discount bond at month t, $y_t^{(n)}$, and a balanced panel of $i = 1, \ldots, 131$ monthly macroeconomic series at month t, x_{ti}, spanning the period from January 1964 to December 2003. We consider the maturities $n = 2, 3, 4, 5$. For more details of data, see Section 3 of Ludvigson and Ng (2009).

We conduct one-year-ahead out of sample forecast comparisons. In order to minimize possible adverse effects of structural breaks, we set the rolling window at 252 months. The forecast comparison procedure is explained below. For the $\mathtt{T}$th month rolling window and maturity n, we extract factors $\{\hat{f}_{tk}\}_{k=1}^{\hat{r}_\mathtt{T}}$ from standardized x_{ti} via our SOFAR and the PC, $i = 1, \ldots, N = 131$, $t = \mathtt{T}, \ldots, T_\mathtt{T} - 12$, where t denotes months from January 1964 to December 2003, $\mathtt{T}$ and $T_\mathtt{T}$ denote the start and end months of the $\mathtt{T}$th rolling window, respectively. Observe that r is estimated for each window according to Section 4.1, where the estimates vary from one to six over the forecast windows. Then, run the predictive regression

$$y_{t+12}^{(n)} = \tilde{\beta}_0^{(n)} + \sum_{k=1}^{\hat{r}_\mathtt{T}} \tilde{\beta}_k^{(n)} \hat{f}_{tk} + \tilde{\varepsilon}_t^{(n)}, \quad t = \mathtt{T}, \ldots, T_\mathtt{T} - 12, \quad n = 2, 3, 4, 5$$

and obtain the forecast error $\hat{\varepsilon}_{T_\mathtt{T}+12|T_\mathtt{T}}^{(n)} = y_{T_\mathtt{T}+12}^{(n)} - \hat{y}_{T_\mathtt{T}+12|T_\mathtt{T}}^{(n)}$, with $\hat{y}_{T_\mathtt{T}+12|T_\mathtt{T}}^{(n)} = \tilde{\beta}_0^{(n)} + \sum_{k=1}^{\hat{r}_\mathtt{T}} \tilde{\beta}_k^{(n)} \hat{f}_{T_\mathtt{T}k}$. This produces $H = 217$ forecast errors.

In Table 4, we report the mean absolute deviation of the forecast errors, $MAE^{(n)} = H^{-1} \sum_{s=1}^{H} \left| \hat{\epsilon}_{s|s-1}^{(n)} \right|$, and the DM forecasting performance test statistics of Diebold and Mariano (1995) with associated p-values, based on the MAEs. As can be seen, the MAEs of the SOFAR are smaller than those of the PC for all the maturities. The DM test strongly rejects the null of the same forecasting performance for all the maturities, in favor of the alternative that our method outperforms the PC. The average values of exponents over the windows are $\{\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6\} = \{0.92, 0.82, 0.87, 0.78, 0.77, 0.74\}$, which suggests that

[10] https://www.sydneyludvigson.com/s/RFS2009-u1e1.xls

even the (first) strongest factor is not strictly strong ($\widehat{N}_1 = 89$). As is evidenced in the previous section, the accuracy of our estimator is much higher than the PC estimator under such situations, and the better forecasting performance may not be too surprising in this empirical exercise.

Table 4: Mean absolute forecast errors and DM forecast comparison test

	SOFAR	PC	DM Statistic	[p-value]
$y_{t+12}^{(2)}$	1.164	1.191	-3.58	[0.0003]
$y_{t+12}^{(3)}$	2.304	2.354	-3.54	[0.0004]
$y_{t+12}^{(4)}$	3.354	3.429	-3.73	[0.0002]
$y_{t+12}^{(5)}$	4.197	4.278	-3.20	[0.0014]

Notes: For the computation of the long-run variance for the DM test statistic, the window is chosen by the Schwert criterion with the maximum lag of 14.

6.2 Interpreting the factors

Since no statistical methods will recover the structural or true factors $\mathbf{F}^*$ and factor loadings $\mathbf{B}^*$ in model (1), it is irrelevant to discuss their detailed properties based on the consistent estimates of their rotations, $\mathbf{F}^0$ and $\mathbf{B}^0$ in sWF model (3). Nonetheless, it is certainly useful to look into the properties of $(\mathbf{F}^0, \mathbf{B}^0)$ or its consistent estimate $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$. As discussed in Ludvigson and Ng (2007, 2009), when the loadings are not sparse, all the variables x_{ti} are subject to the factors, and any economic labeling, such as "output" and/or "unemployment," to a factor can be irrelevant. For this reason, to illustrate the characterization of the factors, empirical studies based on the PC estimate typically report the R^2 statistic of the time-series regression of $(x_{ti})_t$ on each factor $(\hat{f}_{tk}^{\mathsf{PC}})_t$ for k for each i; see figure 1 of Stock and Watson (2002) and figures 1-5 in Ludvigson and Ng (2009).

Importantly, our SOFAR estimates $\widehat{\mathbf{B}}$ of the sparse loadings $\mathbf{B}^0$ can provide more information on the individual factors than the PC estimates because $b_{ik}^0 = 0$ literally means no influence of f_{tk}^0 on x_{ti}. Therefore, together with the orthogonality of the factors, the information about the association of a factor to the variables and its strength is contained in the corresponding loadings. In addition, the sign of a non-zero loading reveals whether the associated variable responds in the same or opposite direction to the other variables in terms of the corresponding factor.

For illustration, we investigate a set of extracted factors from the 131 macroeconomic variables used in Section 6.1 in more detail. In particular, we estimate the model using the variables between January 1982 and December 2001. Two factors (i.e., $\widehat{r} = 2$) are extracted by the PC and SOFAR methods (adaptive, BIC). The exponents are $\{\hat{\alpha}_1, \hat{\alpha}_2\} = \{0.91, 0.71\}$. Figure 4 displays the R^2 of the regressions of the 131 individual time series on the first PC factor over the period. These R^2 are plotted as bar charts, and the variables are ordered as described in the aforementioned data file. Figure 5 displays the PC estimates of the loadings on the first factor. Comparing Figures 4 and 5 reveals that the variables 70–83 and 101–131 except 78 and 113 have little association in terms of R^2, whereas the magnitude of the corresponding loadings are not as small as R^2. Figures 6 and 7 report corresponding results of the adaptive SOFAR to Figures 4 and 5. The patters of R^2 and loadings of SOFAR and PC are very similar. The striking difference is, however, that for the variables 70–83 and

101–131 except 78 and 113, the R^2s for the regressions with the SOFAR factor in Figure 6 are as close to zeros as those of PC (actually the former are slightly closer to zeros), and the associated loadings of SOFAR in Figure 7 are (rightly) zeros. In addition, comparing Figures 6 and 7, the magnitude of R^2 is largely similar to that of loadings. These contrasts in the PC and SOFAR estimation results are more pronounced for the second factor, which are reported in Figures 8–11. In summary, unlike the PC, the SOFAR loadings contain sharper information on which variables are associated with which factor. Among the variables with nonzero loadings, the value of the SOFAR loadings can provide information on strength and direction of the influence of the factor relative to the other variables.

With this encouraging result, we investigate properties of each empirical factor, making use of the information contained in SOFAR loadings. Based on the description of each of the 131 variables in the aforementioned data file, we categorize the 131 variables as follows: 1-24 *Output*; 25-32 *Unemployment*; 33-49 *Employment*; 50-59 *Housing*; 60-69 *Orders*; 70-76 *Money Supply*; 77-80 *Credits*; 81-84 *Stock Prices*; 85-93 *Interest Rate*; 94-101 *Spreads*; 102-106 *Exchange Rates*; 107-127 *Prices*; 128-130 *Wages*; 131 *Consumer Expectation*. From Figure 7 it is easily seen that the first SOFAR factor is *exclusively* loaded on *Output*, *Unemployment*, *Employment*, *Housing*, *Orders*, *Interest Rates* and *Spreads* with a few exceptions only. Observe that the signs of the loadings on the unemployment variables are different from those on the employment variables, as expected. Figure 11 reveals that the second SOFAR factor is exclusively loaded on *Money Supply*, *Exchange Rates* and *Prices*, with scarce exceptions.

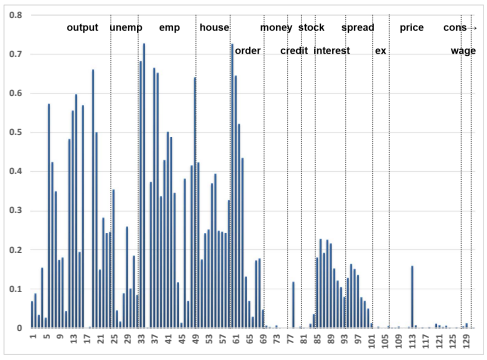

Figure 4: R^2 for regression of x_{it} on $\hat{f}_{t1}^{\mathsf{PC}}$

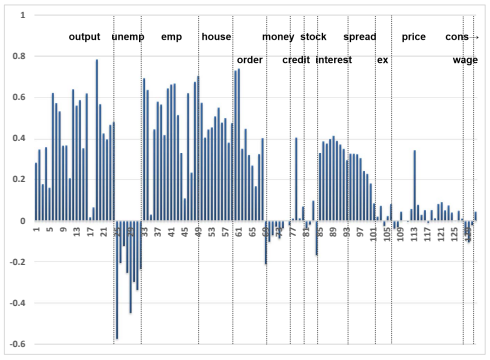

Figure 5: PC estimates $\hat{b}_{i1}^{\mathsf{PC}}$

6.3 Estimating exponents with stock returns

We estimate the sWF model using excess returns of components of the Standard & Poor's 500 Stock Index (S&P 500). In particular, we obtain the 500 securities each month over the period from January 1984 to April 2018 from Datastream. The monthly excess return of security i for month t is computed as $r_{e,ti} = 100 \times (P_{ti} - P_{t-1,i})/P_{t-1,i} + DY_{ti}/12 - r_{ft}$, where P_{ti} is the end-of-the-month price, DY_{ti} is the percent per annum dividend yield, and r_{ft} is the one-month US treasury bill rate chosen as the risk-free rate.[11] We standardize the obtained excess returns and denote them as $r^*_{e,ti}$.

For each window month, $\mathtt{T}$ = September 1998 to April 2018, we chose securities that contain the data extending 120 months back ($T = 120$) from $\mathtt{T}$. This gives the different

[11] This is obtained from Ken French's data library web page.

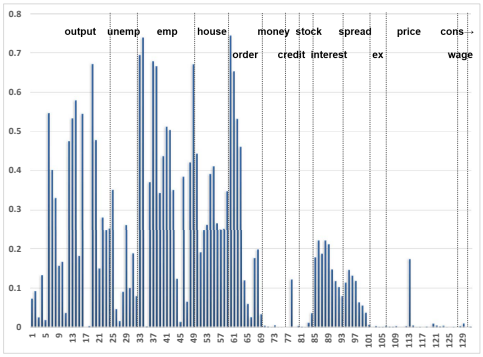

Figure 6: R^2 for regression of x_{it} on $\hat{f}_{t1}^{\mathsf{ada}}$

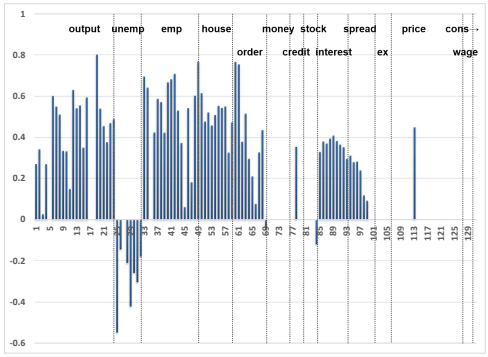

Figure 7: adaptive SOFAR estimates $\hat{b}_{i1}^{\mathsf{ada}}$

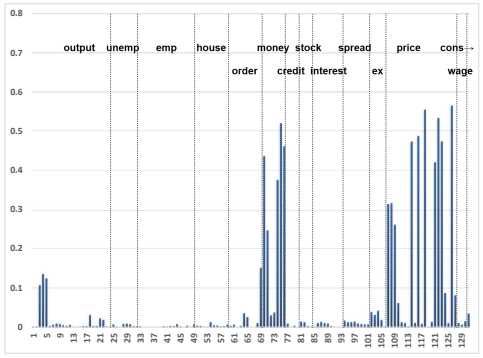

Figure 8: R^2 for regression of x_{it} on $\hat{f}_{t2}^{\mathsf{PC}}$

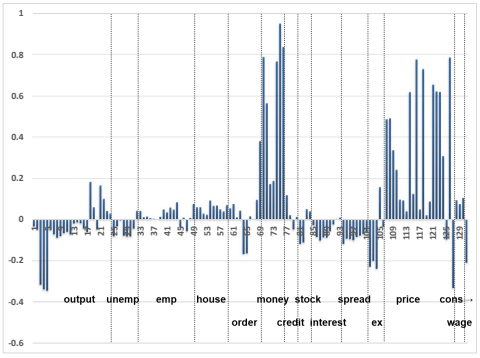

Figure 9: PC estimates $\hat{b}_{i2}^{\mathsf{PC}}$

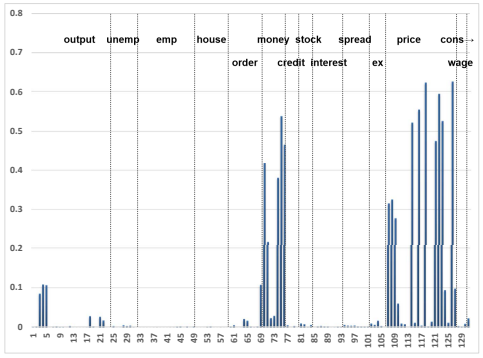

Figure 10: R^2 for regression of x_{it} on $\hat{f}_{t2}^{\mathsf{ada}}$

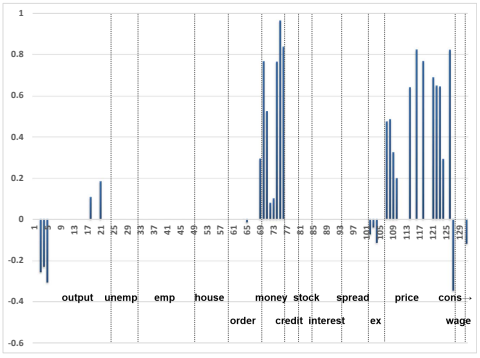

Figure 11: adaptive SOFAR estimates $\hat{b}_{i2}^{\mathsf{ada}}$

number of securities for each window $\mathtt{T}$ ($N_{\mathtt{T}}$). The average number of securities over the estimation windows is 443 ($\bar{N} = 443$). As will be shown below, three or four factors are estimated over the windows. We identify the factors and signs of the factors and factor loadings, given the estimates of the initial window month, $\mathtt{T}$ = September 1989, based on the correlation coefficients between the factors at $\mathtt{T}$ and the appropriately lagged $\mathtt{T}$.[12]

[12]For example, define $(T-1)$-dimensional vector of ℓth factor of $\mathtt{T}$ as $\widehat{\mathbf{f}}_{\ell\mathtt{T}} = (\hat{f}_{\ell\mathtt{T},1}, \hat{f}_{\ell\mathtt{T},2}, \ldots, \hat{f}_{\ell\mathtt{T},\mathtt{T}-1})'$ and that of $\mathtt{T}-\mathtt{1}$ as $\widehat{\mathbf{f}}_{\ell\mathtt{T}-1} = (\hat{f}_{\ell\mathtt{T}-1,2}, \hat{f}_{\ell\mathtt{T}-1,2}, \ldots, \hat{f}_{\ell\mathtt{T}-1,\mathtt{T}})'$, $\ell = 1, \ldots, r$. For $\widehat{\mathbf{f}}_{\ell\mathtt{T}}$, if $\max_{1\le k\le r} |\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{k\mathtt{T}-1})| =$

We report $\hat{\alpha}_k$, $k = 1, 2, 3, 4$, of the stock return covariance matrix, which are associated with the four factors. Observe that, as discussed earlier, the estimated exponents are invariant to the rotation of the estimated common components. Figure 12 plots $\hat{\alpha}_k$ over the estimation window months, $\mathtt{T}$ = September 1989 to April 2018. Apart from the first factor, which is always strong, the strengths of the signals vary over the months and can become quite weak. These strongly imply a potentially substantial efficiency gain in estimation of the approximate factor models through our SOFAR over the PC. It is also interesting that the orders in terms of values of the exponents, α_2, α_3, and α_4, change over the period.

In line with the well-observed phenomenon that the correlation among the securities in the financial market rises during periods of turmoil, sharp rises of exponents in some months can be observed. For example, α_2 goes up sharply around February 2000 then rises gradually. This period corresponds to the peak of the dot-com bubble and its burst on March 2000 (the main contributor to the factor loadings of the second factor is Technology industry, see Appendix C.1). Similarly, a sharp rise of α_3 is observed from July 2008 to April 2009. This period coincides with the 2008 financial crisis. In just ten months, it goes up by 0.12, from 0.74 to 0.86 (one of the main contributors to the factor loadings of the third factor is the Financial industry, see Appendix C.1).

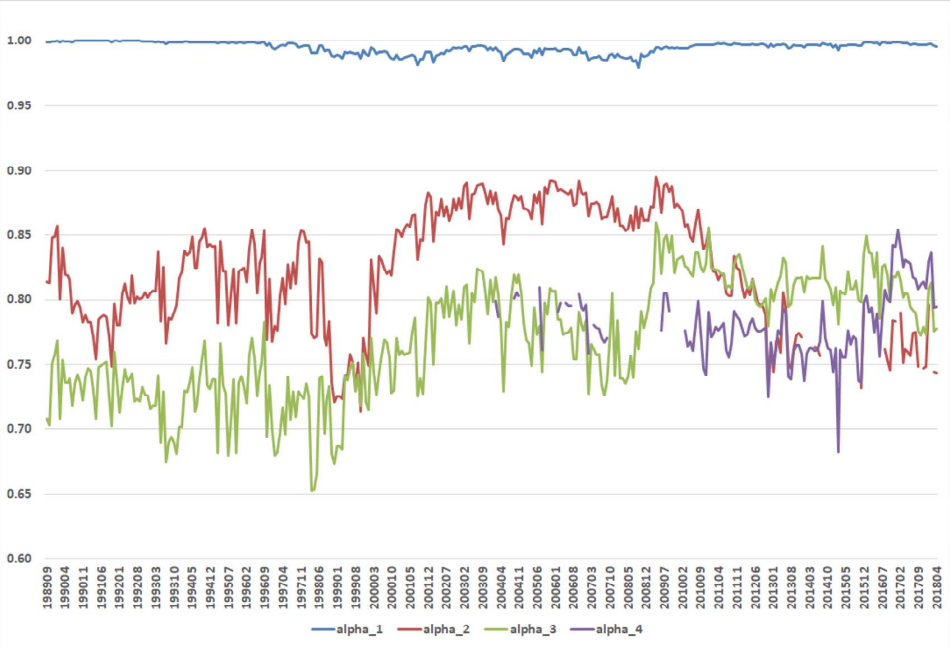

Figure 12: Plot of the estimated α_k's from September 1989 to April 2018.

7 Conclusion

This paper has considered estimation of the sparsity-induced weak factor (sWF) models in a high-dimensional setting. We suppose sparse factor loadings $\mathbf{B}^0$ that lead to the WF structure, $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N^{\alpha_k}$ with $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. This model is much less restrictive than the widely employed strong factor (SF) model in the literature, in which $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N$ for $k = 1, \ldots, r$. The SOFAR estimator and its adaptive version enable us

$|\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{2\mathtt{T}-1})|$ and $\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{2\mathtt{T}-1}) < 0$, say, $\widehat{\mathbf{f}}_{2\mathtt{T}} \equiv -\widehat{\mathbf{f}}_{\ell,\mathtt{T}}$ and $\widehat{\mathbf{b}}_{i2\mathtt{T}} \equiv -\widehat{\mathbf{b}}_{i\ell\mathtt{T}}$.

to consistently estimate the sWF models, separately identifying $\mathbf{B}^0$ and $\mathbf{F}^0$. As theoretical contributions, we have established the estimation error bound of the SOFAR estimators, the factor selection consistency of the adaptive SOFAR estimator, and consistent estimation of each exponent α_k as well as validating the method of Onatski (2010) for determining the number of weak factors. All the theoretical results are supported by the Monte Carlo experiments, and three empirical examples demonstrate practical usefulness of our estimator in comparison to the principal component (PC) estimator.

The proposed method has large potential applicability and many direction to extend. The hierarchical factor model, which contains global and local factors, are recently considered by Ando and Bai (2017), Choi et al. (2018) and Andreou et al. (2019). Our sWF model nests the hierarchical factor model, and hence the SOFAR method can be applied to readily estimate such models. It is of interest to estimate the stock returns covariance matrix for optimal portfolio allocation and portfolio risk assessment. This can be achieved by consistently estimating the covariance matrix of idiosyncratic errors, in line with Fan et al. (2008) and Fan et al. (2011), which is an interesting extension of this paper. Having provided the consistent estimation in this paper, the statistical inference for the sWF models is an important research agenda. This is considered in Uematsu and Yamagata (2020). Yet another possible extension of interest is the estimation of panel data models with interactive effects, which is considered by Pesaran (2006) and Bai (2009), among others: $y_{ti} = \mathbf{x}_{ti}'\boldsymbol{\beta} + u_{ti},\ u_{ti} = \mathbf{f}_t'\mathbf{b}_i + \varepsilon_{ti}$. For the PC based estimators, such as Bai (2009), u_{ti} is typically assumed to be a SF model and estimated by PC, given an initial estimator of $\boldsymbol{\beta}$. The SOFAR estimation, instead of the PC, would potentially improve the precision of the estimates of $\boldsymbol{\beta}$.

References

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

Ando, T. and J. Bai (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association 112*, 1182–1198.

Andreou, E., P. Gagliardini, E. Ghysels, and M. Rubin (2019). Inference in group factor models with an application to mixed frequency data. *Econometrica, to appear*.

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

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*, 1229–1279.

Bai, J. and K. Li (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics 40*, 436–465.

Bai, J., K. Li, and L. Lu (2016). Estimation and inference of FAVAR models. *Journal of Business & Economic Statistics 34*, 620–641.

Bai, J. and Y. Liao (2017). Inferences in panel data with interactive effects using large covariance matrices. *Journal of Econometrics 200*, 59–78.

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

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

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics 176*, 18–29.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2016). Exponent of cross-sectional dependence: Estimation and inference. *Journal of Applied Econometrics 31*, 929–960.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2020). Measurement of factor strength: Theory and practice. *mimeo*.

Bryzgalova, S. (2016). Spurious factors in linear asset pricing models. *mimeo*.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica 51*, 1281–1304.

Choi, I., D. Kim, Y. J. Kim, and N.-S. Kwark (2018). A multilevel factor model: Identification, asymptotic theory and applications. *Journal of Applied Econometrics 33*, 355–377.

Chudik, A., , H. Pesaran, and E. Tosetti (2011). Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal 14*, C45–C90.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*, 373–394.

Connor, G. and R. A. Korajczyk (1993). A test for the number of factors in an approximate factor modela test for the number of factors in an approximate factor model. *Journal of Finance 48*, 1263–1291.

Daniele, M., W. Pohlmeier, and A. Zagidullina (2020). Sparse approximate factor estimation for high-dimensional covariance matrices. *arXiv:1906.05545v1*.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*, 253–263.

Fan, J., Y. Fan, and E. Barut (2014). Adaptive robust variable selection. *Annals of Statistics 42*, 324–351.

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

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

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

Fan, J., K. Wang, Y. Zhong, and Z. Zhu (2018). Robust high-dimensional factor models with applications to statistical machine learning. *arXiv:1808.03889v1*.

Fan, Y., J. Lv, M. Sharifvaghefi, and Y. Uematsu (2019). IPAD: stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association, to appear*.

Freyaldenhoven, S. (2020). Identification through sparsity in factor models. *Federal Reserve Bank of Philadelphia, WP20-25*.

Gao, J., G. Pan, Y. Yan, and B. Zhang (2020). Estimation of cross-sectional dependence in large panels. *arXiv:1904.06843v1*.

Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research 15*, 2869–2909.

Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series. *Biometrika 98*, 901–918.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics 142*, 201–211.

Lettau, M. and M. Pelger (2020). Estimating latent asset-pricing factors. *Journal of Econometrics 218*, 1–31.

Ludvigson, C. S. and S. Ng (2007). Empirical risk-return relation: A factor analysis approach. *Journal of Financial Econometrics 83*, 171–222.

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

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics 92*, 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*, 244–258.

Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*, 967–1012.

Rigollet, P. and J.-C. Hütter (2017). *High Dimensional Statistics*. Massachusetts Institute of Technology, MIT Open CourseWare.

Rudelson, M. and R. Vershynin (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability 18*, 1–9.

Shen, H. and J. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis 99*, 1015–1034.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 30*, 147–162.

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

Uematsu, Y., Y. Fan, K. Chen, J. Lv, and W. Lin (2019). SOFAR: large-scale association network learning. *IEEE Transactions on Information Theory 65*, 4929–4939.

Uematsu, Y. and S. Tanaka (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *Econometrics Journal 22*, 34–56.

Uematsu, Y. and T. Yamagata (2020). Inference in weak factor models. *Available at SSRN: https://ssrn.com/abstract=3556275*.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Practice*, pp. 210–268. Cambridge University Press.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*, 265–286.

Supplementary Material for

Estimation of Weak Factor Models

Yoshimasa Uematsu* and Takashi Yamagata†

*Department of Economics and Management, Tohoku University
†Department of Economics and Related Studies, University of York
†Institute of Social Economic Research, Osaka University

A Proofs of the Main Results

A.1 Proof of Theorem 1

Proof. We denote by $\mathbf{M}_{k:\ell} \in \mathbb{R}^{T\times(\ell-k+1)}$ a submatrix of $\mathbf{M}$ constructed by its kth to ℓth columns. Following Ahn and Horenstein (2013), we evaluate the eigenvalues of $\mathbf{XX}'$ with recalling notation based on the SVD rather than $\mathbf{F}^0$ and $\mathbf{B}^0$. We define $\mathbf{P} = \mathbf{V}^0\mathbf{N}^{-1}\mathbf{V}^{0\prime}$, $\mathbf{Q} = \mathbf{I}_N - \mathbf{P}$, and $\mathbf{U}^* = \mathbf{U}^0 + \mathbf{E}\mathbf{V}^0\mathbf{N}^{-1}(\mathbf{D}^0)^{-1}$. Then, we can write $\mathbf{XX}' = \mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime} + \mathbf{EQE}'$ since $\mathbf{V}^{0\prime}\mathbf{V}^0 = \mathbf{N} = \mathrm{diag}(N_1,\dots,N_r)$ by the definition. We also define $\mathbf{W}_{1:k}$ as the matrix of k eigenvectors corresponding to the first k largest eigenvalues of $\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}$.

We first evaluate the r largest eigenvalues of $\mathbf{XX}'$. Because $\lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) = d_k^2 N_k T$, it is sufficient to show that for any $k \in \{1,\dots,r\}$,

$$\lambda_k(\mathbf{XX}') = \lambda_k(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + O(N\vee T), \tag{A.1}$$

$$\lambda_k(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) = \lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) + O\left(TN_1^{1/2}\log^{1/2}(N\vee T) + N\vee T\right). \tag{A.2}$$

Then (A.1) and (A.2) lead to

$$\begin{aligned}\lambda_k(\mathbf{XX}') &= \lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) + O(TN_1^{1/2}\log^{1/2}(N\vee T) + N\vee T)\\ &= d_k^2 N_k T + O\left(TN_1^{1/2}\log^{1/2}(N\vee T) + N\vee T\right),\end{aligned}$$

which gives the desired result under condition (8). We show (A.1). Lemma A.5 of Ahn and Horenstein (2013) yields the upper bound

$$\begin{aligned}&\sum_{j=1}^{k}\lambda_j(\mathbf{XX}') = \sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime} + \mathbf{EQE}')\\ &\le \sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + k\lambda_1(\mathbf{EQE}' + \mathbf{EPE}')\\ &= \sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + k\lambda_1(\mathbf{EE}') \lesssim \sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + T\vee N,\end{aligned}$$

where the last inequality follows from Lemma 1(a), with probability at least $1-O((N\vee T)^{-\nu})$.

Moreover, the lower bound is given by

$$\begin{aligned}\sum_{j=1}^{k}\lambda_j(\mathbf{X}\mathbf{X}') &\geq T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{X}\mathbf{X}'\mathbf{W}_{1:k})\\ &= T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}\mathbf{W}_{1:k}) + T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{E}\mathbf{Q}\mathbf{E}'\mathbf{W}_{1:k})\\ &\geq \sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}).\end{aligned}$$

Hence, these two inequalities imply (A.1). Next, we verify (A.2). By the construction of $\mathbf{U}^*$, the upper bound is

$$\begin{aligned}\sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) &= T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}\mathbf{W}_{1:k})\\ &\quad + 2T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\mathbf{W}_{1:k}) + T^{-1}\operatorname{tr}(\mathbf{W}_{1:k}'\mathbf{E}\mathbf{P}\mathbf{E}'\mathbf{W}_{1:k})\\ &\lesssim \sum_{j=1}^{k}\lambda_j(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) + TN_1^{1/2}\log^{1/2}(N\vee T) + N\vee T,\end{aligned}$$

where the last inequality holds by Lemma 3 with probability at least $1-O((N\vee T)^{-\nu})$. Similarly, the lower bound is

$$\sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) \gtrsim \sum_{j=1}^{k}\lambda_j(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) - TN_1^{1/2}\log^{1/2}(N\vee T).$$

Hence, these two inequalities imply (A.2).

Finally, we consider the lower and upper bounds of $\lambda_{r+j}(\mathbf{X}\mathbf{X}')$ for $j=1,\ldots,k_{\max}$. Because $\lambda_{r+j}(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime})=0$ for all $j\geq 1$, Lemma 3 entails

$$\lambda_{r+j}(\mathbf{X}\mathbf{X}') \leq \lambda_{r+j}(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + \lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}') = \lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}') \lesssim T\vee N$$

with probability at least $1-O((N\vee T)^{-\nu})$. This completes the proof. □

A.2 Proof of Theorem 2

Proof. The optimality of the SOFAR estimator implies

$$2^{-1}\|\mathbf{X}-\widehat{\mathbf{F}}\widehat{\mathbf{B}}'\|_{\mathrm{F}}^2 + \eta_n\|\widehat{\mathbf{B}}\|_1 \leq 2^{-1}\|\mathbf{X}-\mathbf{F}^0\mathbf{B}^{0\prime}\|_{\mathrm{F}}^2 + \eta_n\|\mathbf{B}^0\|_1.$$

By plugging model (5) and letting $\boldsymbol{\Delta}=\widehat{\mathbf{F}}\widehat{\mathbf{B}}'-\mathbf{F}^0\mathbf{B}^{0\prime}$, this is equivalently written as

$$2^{-1}\|\mathbf{E}-\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \eta_n\|\widehat{\mathbf{B}}\|_1 \leq 2^{-1}\|\mathbf{E}\|_{\mathrm{F}}^2 + \eta_n\|\mathbf{B}^0\|_1.$$

Define $\boldsymbol{\Delta}^f=\widehat{\mathbf{F}}-\mathbf{F}^0$ and $\boldsymbol{\Delta}^b=\widehat{\mathbf{B}}-\mathbf{B}^0$. Expanding the first term and using decomposition $\boldsymbol{\Delta}=\boldsymbol{\Delta}^f\mathbf{B}^{0\prime}+\boldsymbol{\Delta}^f\boldsymbol{\Delta}^{b\prime}+\mathbf{F}^0\boldsymbol{\Delta}^{b\prime}$ lead to

$$\begin{aligned}(1/2)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 &\leq \operatorname{tr}\mathbf{E}\boldsymbol{\Delta}' + \eta_n\left(\|\mathbf{B}^0\|_1-\|\widehat{\mathbf{B}}\|_1\right)\\ &\leq \left|\operatorname{tr}\mathbf{E}\mathbf{B}^0\boldsymbol{\Delta}^{f\prime}\right| + \left|\operatorname{tr}\mathbf{E}\boldsymbol{\Delta}^b\boldsymbol{\Delta}^{f\prime}\right| + \left|\operatorname{tr}\boldsymbol{\Delta}^b\mathbf{F}^{0\prime}\mathbf{E}\right| + \eta_n\left(\|\mathbf{B}^0\|_1-\|\widehat{\mathbf{B}}\|_1\right). \qquad (A.3)\end{aligned}$$

We bound the traces in (A.3). By applying Hölder's inequality and using properties of the norms, the first term is bounded as

$$\left|\operatorname{tr} \mathbf{E}\mathbf{B}^0\boldsymbol{\Delta}^{f\prime}\right| \leq \|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}^f\|_1 \leq (rT)^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}}.$$

Similarly, the second and third terms of (A.3) are bounded as

$$\begin{aligned}\left|\operatorname{tr} \mathbf{E}\boldsymbol{\Delta}^b\boldsymbol{\Delta}^{f\prime}\right| + \left|\operatorname{tr} \boldsymbol{\Delta}^b\mathbf{F}^{0\prime}\mathbf{E}\right| &\leq \|\mathbf{E}\boldsymbol{\Delta}^b\|_2\|\boldsymbol{\Delta}^f\|_* + \|\boldsymbol{\Delta}^b\|_1\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} \\ &\leq r^{1/2}\|\mathbf{E}\boldsymbol{\Delta}^b\|_2\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + \|\boldsymbol{\Delta}^b\|_1\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max}.\end{aligned}$$

From these inequalities, the upper bound of (A.3) becomes

$$\begin{aligned}(1/2)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leq (rT)^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + r^{1/2}\|\mathbf{E}\boldsymbol{\Delta}^b\|_2\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ + \|\boldsymbol{\Delta}^b\|_1\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} + \eta_n\left(\|\mathbf{B}^0\|_1 - \|\widehat{\mathbf{B}}\|_1\right).\end{aligned} \tag{A.4}$$

From Lemmas 1 and 4, there exist some positive constants c_1–c_3 such that the event

$$\begin{aligned}\mathcal{E} = &\left\{\|\mathbf{E}\boldsymbol{\Delta}^b\|_2 \leq c_1\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\right\} \\ &\cap\left\{\|\mathbf{E}\mathbf{B}^0\|_{\max} \leq c_2 N_1^{1/2}\log^{1/2}(N\vee T)\right\}\cap\left\{\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} \leq c_3 T^{1/2}\log^{1/2}(N\vee T)\right\}\end{aligned}$$

occurs with probability at least $1 - O((N\vee T)^{-\nu})$ for any fixed constant $\nu > 0$. Set the regularization parameter to be $\eta_n = 2c_3T^{1/2}\log^{1/2}(N\vee T)$. Then on event $\mathcal{E}$, we have $\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} \leq \eta_n/2$, and (A.4) is further bounded as

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \lesssim (N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + (\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ + \eta_n\left(\|\boldsymbol{\Delta}^b\|_1 + 2\|\mathbf{B}^0\|_1 - 2\|\widehat{\mathbf{B}}\|_1\right).\end{aligned} \tag{A.5}$$

We then focus on the last parenthesis of (A.5). Define index set $\mathcal{S} = \{(i,k) : b_{ik}^0 \neq 0\}$, the support of $\mathbf{B}^0$. Note that $|\mathcal{S}| = \sum_{k=1}^r N_k \leq rN_1$. The last parenthesis of (A.5) is rewritten and bounded as

$$\begin{aligned}&\|\boldsymbol{\Delta}^b\|_1 + 2\|\mathbf{B}^0\|_1 - 2\|\widehat{\mathbf{B}}\|_1 = \|\boldsymbol{\Delta}^b_{\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}^b_{\mathcal{S}^c}\|_1 + 2\|\mathbf{B}^0_{\mathcal{S}}\|_1 - 2\|\widehat{\mathbf{B}}_{\mathcal{S}}\|_1 - 2\|\widehat{\mathbf{B}}_{\mathcal{S}^c}\|_1 \\ &\quad\leq \|\boldsymbol{\Delta}^b_{\mathcal{S}}\|_1 + \|\boldsymbol{\Delta}^b_{\mathcal{S}^c}\|_1 + 2\|\mathbf{B}^0_{\mathcal{S}}\|_1 - 2\left(\|\mathbf{B}^0_{\mathcal{S}}\|_1 - \|\boldsymbol{\Delta}^b_{\mathcal{S}}\|_1\right) - 2\|\widehat{\mathbf{B}}_{\mathcal{S}^c}\|_1 \\ &\quad= 3\|\boldsymbol{\Delta}^b_{\mathcal{S}}\|_1 - \|\widehat{\mathbf{B}}_{\mathcal{S}^c}\|_1 \leq 3(rN_1)^{1/2}\|\boldsymbol{\Delta}^b_{\mathcal{S}}\|_{\mathrm{F}} \leq 3(rN_1)^{1/2}\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}.\end{aligned}$$

Therefore, the upper bound of (A.5) is given by

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \lesssim (N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ + (\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + N_1^{1/2}\eta_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}.\end{aligned} \tag{A.6}$$

Meanwhile, Lemma 5 establishes the lower bound of (A.6). Consequently, we obtain

$$\begin{aligned}\kappa_n\left(\|\boldsymbol{\Delta}^f\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}^b\|_{\mathrm{F}}^2\right) &\lesssim (N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ &\quad + (\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + N_1^{1/2}\eta_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}} \\ &=: \alpha_n\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + \mu_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + \beta_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}} \\ &\leq \alpha_n\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + \mu_n\left(\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}^f\|_{\mathrm{F}}^2\right) + \beta_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}},\end{aligned}$$

where $\kappa_n = N_r(N_r \wedge T)/N_1$, $\alpha_n = (N_1 T)^{1/2}\log^{1/2}(N \vee T)$, $\mu_n = (\widetilde{N} \vee T)^{1/2}\log^{1/2}(N \vee T)$, and $\beta_n = N_1^{1/2}\eta_n$. By condition (10), we have

$$\|\mathbf{\Delta}^f\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}^b\|_{\mathrm{F}}^2 \leq \frac{(\alpha_n/\kappa_n)\|\mathbf{\Delta}^f\|_{\mathrm{F}} + (\beta_n/\kappa_n)\|\mathbf{\Delta}^b\|_{\mathrm{F}}}{1-\mu_n/\kappa_n}.$$

Rearranging this inequality gives

$$\|\mathbf{\Delta}^f\|_{\mathrm{F}} + \|\mathbf{\Delta}^b\|_{\mathrm{F}} \leq \frac{3}{2}\left(\frac{\alpha_n/\kappa_n + \beta_n/\kappa_n}{1-\mu_n/\kappa_n}\right).$$

Finally, since $\eta_n = 2c_3 T^{1/2}\log^{1/2}(N \vee T)$, we observe that

$$\alpha_n + \beta_n = (N_1 T)^{1/2}\log^{1/2}(N \vee T) + N_1^{1/2}\eta_n \lesssim (N_1 T)^{1/2}\log^{1/2}(N \vee T).$$

This completes the proof. □

A.3 Proof of Theorem 3

Proof. Following the proof of Theorem 2, we derive the bound. From (A.4) with putting $\eta_n = 0$, we have

$$\begin{aligned}&(1/2)\|\mathbf{\Delta}_{\mathsf{PC}}\|_{\mathrm{F}}^2 \\ &\quad\lesssim T^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + \|\mathbf{E}\mathbf{\Delta}_{\mathsf{PC}}^b\|_2\|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + N^{1/2}\|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max}. \qquad \text{(A.7)}\end{aligned}$$

Lemmas 1 and 4 states that the event

$$\begin{aligned}\mathcal{E} = &\left\{\|\mathbf{E}\mathbf{\Delta}_{\mathsf{PC}}^b\|_2 \lesssim \|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}(N \vee T)^{1/2}\log^{1/2}(N \vee T)\right\} \\ &\cap \left\{\|\mathbf{E}\mathbf{B}^0\|_{\max} \lesssim N_1^{1/2}\log^{1/2}(N \vee T)\right\} \cap \left\{\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} \lesssim T^{1/2}\log^{1/2}(N \vee T)\right\}\end{aligned}$$

occurs with probability at least $1 - O((N \vee T)^{-\nu})$ for any fixed constant $\nu > 0$. On event $\mathcal{E}$ together with Lemma 5, (A.7) becomes

$$\kappa_n\left(\|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}^2\right) \lesssim \alpha_n\|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + \mu_n\left(\|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}^2\right) + \beta_n\|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}},$$

where

$$\begin{aligned}&\kappa_n = \frac{N_r(N_r \wedge T)}{N_1}, \qquad \mu_n = (N \vee T)^{1/2}\log^{1/2}(N \vee T) \\ &\alpha_n = (N_1 T)^{1/2}\log^{1/2}(N \vee T), \qquad \beta_n = (NT)^{1/2}\log^{1/2}(N \vee T).\end{aligned}$$

The desired result is obtained by rearranging this inequality as in the proof of Theorem 2. In fact, we have

$$\|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + \|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}} \lesssim \frac{3}{2}\left(\frac{\alpha_n/\kappa_n + \beta_n/\kappa_n}{1-\mu_n/\kappa_n}\right).$$

Finally, we observe that

$$\alpha_n + \beta_n = (N_1 T)^{1/2}\log^{1/2}(N \vee T) + (NT)^{1/2}\log^{1/2}(N \vee T) \lesssim (NT)^{1/2}\log^{1/2}(N \vee T).$$

This completes the proof of Theorem 3. □

A.4 Proof of Theorem 4

Proof. Throughout this proof, we omit the superscript of the adaptive estimators $(\widehat{\mathbf{F}}^{\mathsf{ada}}, \widehat{\mathbf{B}}^{\mathsf{ada}})$ and simply write them as $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$. Recall $\mathcal{S} = \mathrm{supp}(\mathbf{B}^0)$, which is a subset of $\{1, \ldots, N\} \times \{1, \ldots, r\}$. For any matrix $\mathbf{B} = (b_{ik}) \in \mathbb{R}^{N \times r}$, define $\mathbf{B}_{\mathcal{S}} \in \mathbb{R}^{N \times r}$ as the matrix whose (i,k)th element is $b_{ik} 1\{(i,k) \in \mathcal{S}\}$. Similarly, define $\mathbf{B}_{\mathcal{S}^c} \in \mathbb{R}^{N \times r}$ whose (i,k)th element is $b_{ik} 1\{(i,k) \in \mathcal{S}^c\}$. By the definition, note that $\mathbf{B}^0_{\mathcal{S}} = \mathbf{B}^0$ and $\mathbf{B}^0_{\mathcal{S}^c} = \mathbf{0}$. Recall that the objective function for obtaining the adaptive SOFAR estimator is given by

$$Q_n(\mathbf{F}, \mathbf{B}) := \frac{1}{2} \left\| \mathbf{X} - \mathbf{F}\mathbf{B}' \right\|_{\mathrm{F}}^2 + \eta_n \|\mathbf{W} \circ \mathbf{B}\|_1 \tag{A.8}$$

subject to $\mathbf{F}'\mathbf{F}/T = \mathbf{I}_r$ and $\mathbf{B}'\mathbf{B}$ being diagonal. The strategy of this proof consists of two steps. In the first step, we show that the *oracle estimator* $(\widehat{\mathbf{F}}^o, \widehat{\mathbf{B}}^o_{\mathcal{S}})$, which is defined as a minimizer of $Q_n(\mathbf{F}, \mathbf{B}_{\mathcal{S}})$, is consistent to $(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}})$ with some rate of convergence. In the second step, we prove that the oracle estimator is indeed a minimizer of the unrestricted problem, $\min Q_n(\mathbf{F}, \mathbf{B})$ over $\mathbb{R}^{T \times r} \times \mathbb{R}^{N \times r}$.

(First step) We derive the rate of convergence of the oracle estimator. To this end, it suffices to show that as $n \to \infty$, there exists a (large) constant $C > 0$ such that

$$\mathbb{P}\left(\inf_{\|\mathbf{U}\|_{\mathrm{F}} = C,\, \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} = C} Q_n(\mathbf{F}^0 + r_n \mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}}) > Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \right) \to 1, \tag{A.9}$$

where $\mathbf{U} \in \mathbb{R}^{T \times r}$ and $\mathbf{V} \in \mathbb{R}^{N \times r}$ are deterministic matrices, and

$$r_n = \frac{N_1 (N_1 T)^{1/2} \log^{1/2}(N \vee T)}{N_r (N_r \wedge T)}.$$

This implies that the oracle estimator $(\widehat{\mathbf{F}}^o, \widehat{\mathbf{B}}^o_{\mathcal{S}})$ lies in the ball

$$\left\{ (\mathbf{F}, \mathbf{B}_{\mathcal{S}}) \in \mathbb{R}^{T \times r} \times \mathbb{R}^{N \times r} : \|\mathbf{F} - \mathbf{F}^0\|_{\mathrm{F}} \le C r_n, \|\mathbf{B}_{\mathcal{S}} - \mathbf{B}^0_{\mathcal{S}}\|_{\mathrm{F}} \le C r_n \right\}$$

with high probability, which gives the desired rate of convergence. In this proof, write $\ell_n = \log(N \vee T)$ for notational simplicity.

To show (A.9), we first have

$$\begin{aligned}
&Q_n(\mathbf{F}^0 + r_n \mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}}) - Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \\
&\quad = 2^{-1} \|\mathbf{X} - (\mathbf{F}^0 + r_n \mathbf{U})(\mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}})'\|_{\mathrm{F}}^2 - 2^{-1} \|\mathbf{X} - \mathbf{F}^0 \mathbf{B}^{0\prime}_{\mathcal{S}}\|_{\mathrm{F}}^2 \\
&\qquad\quad + \eta_n \|\mathbf{W} \circ (\mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}})\|_1 - \eta_n \|\mathbf{W} \circ \mathbf{B}^0_{\mathcal{S}}\|_1 \\
&\quad \ge -\mathrm{tr}(r_n \mathbf{E}'\mathbf{F}^0 \mathbf{V}'_{\mathcal{S}} + r_n \mathbf{E}'\mathbf{U}\mathbf{B}^{0\prime}_{\mathcal{S}} + r_n^2 \mathbf{E}'\mathbf{U}\mathbf{V}'_{\mathcal{S}}) \\
&\qquad\quad + 2^{-1} \|r_n \mathbf{F}^0 \mathbf{V}'_{\mathcal{S}} + r_n \mathbf{U}\mathbf{B}^{0\prime}_{\mathcal{S}} + r_n^2 \mathbf{U}\mathbf{V}'_{\mathcal{S}}\|_{\mathrm{F}}^2 - r_n \eta_n \|\mathbf{W}_{\mathcal{S}} \circ \mathbf{V}_{\mathcal{S}}\|_1 \\
&\quad =: (I) + (II) + (III).
\end{aligned} \tag{A.10}$$

By Lemma 7 (a)–(c), we bound (I) as

$$\begin{aligned}
|(I)| &\le r_n \left| \mathrm{tr}\, \mathbf{V}'_{\mathcal{S}} \mathbf{E}' \mathbf{F}^0 \right| + r_n \left| \mathrm{tr}\, \mathbf{B}^{0\prime}_{\mathcal{S}} \mathbf{E}' \mathbf{U} \right| + r_n^2 \left| \mathrm{tr}\, \mathbf{V}'_{\mathcal{S}} \mathbf{E}' \mathbf{U} \right| \\
&\lesssim r_n \left(T^{1/2} \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} + N_1^{1/2} \|\mathbf{U}\|_{\mathrm{F}} \right) \ell_n^{1/2} + r_n^2 \|\mathbf{U}\|_{\mathrm{F}} \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} \ell_n^{1/2}.
\end{aligned}$$

Next, we bound (II) from below as

$$\begin{aligned}(II) &= 2^{-1}\|r_n\mathbf{F}^0\mathbf{V}'_{\mathcal{S}} + r_n\mathbf{U}\mathbf{B}^{0\prime}_{\mathcal{S}} + r_n^2\mathbf{U}\mathbf{V}'_{\mathcal{S}}\|_{\mathrm{F}}^2 \\ &\geq 2^{-1}\|r_n\mathbf{U}\mathbf{B}^{0\prime}_{\mathcal{S}}\|_{\mathrm{F}}^2 + 2^{-1}\|r_n\mathbf{F}^0\mathbf{V}'_{\mathcal{S}}\|_{\mathrm{F}}^2 - r_n^3\left|\operatorname{tr}\mathbf{V}_{\mathcal{S}}\mathbf{U}'\mathbf{F}^0\mathbf{V}'_{\mathcal{S}}\right| - r_n^3\left|\operatorname{tr}\mathbf{B}^0_{\mathcal{S}}\mathbf{U}'\mathbf{U}\mathbf{V}'_{\mathcal{S}}\right| - r_n^2\left|\operatorname{tr}\mathbf{B}^0_{\mathcal{S}}\mathbf{U}'\mathbf{F}^0\mathbf{V}'_{\mathcal{S}}\right| \\ &= (i) + (ii) + (iii) + (iv) + (v).\end{aligned}$$

In view of the Rayleigh quotient, (i) and (ii) are further bounded from below as

$$\begin{aligned}(i) + (ii) &= 2^{-1}\|\mathbf{U}\mathbf{B}^{0\prime}\|_{\mathrm{F}}^2 + 2^{-1}\|\mathbf{F}^0\mathbf{V}'_{\mathcal{S}}\|_{\mathrm{F}}^2 \\ &= 2^{-1}r_n^2\|(\mathbf{I}_T \otimes \mathbf{B}^0)\operatorname{vec}(\mathbf{U}')\|_2^2 + 2^{-1}r_n^2\|(\mathbf{I}_N \otimes \mathbf{F}^0)\operatorname{vec}(\mathbf{V}'_{\mathcal{S}})\|_2^2 \\ &\gtrsim r_n^2\left\{\min_{\mathbf{u}\in\mathbb{R}^{Tr}\setminus\{\mathbf{0}\}}\left(\frac{\|(\mathbf{I}_T \otimes \mathbf{B}^0)\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2}\right)\|\mathbf{U}\|_{\mathrm{F}}^2 + \min_{\mathbf{v}\in\mathbb{R}^{Nr}\setminus\{\mathbf{0}\}}\left(\frac{\|(\mathbf{I}_N \otimes \mathbf{F}^0)\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2^2}\right)\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right\} \\ &\gtrsim r_n^2\left(N_r\|\mathbf{U}\|_{\mathrm{F}}^2 + T\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right).\end{aligned}$$

Meanwhile, by Lemma 7 (d)–(f), $|(iii) + (iv) + (v)|$ is bounded from above as

$$|(iii) + (iv) + (v)| \lesssim r_n^3\left(\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\ell_n^{1/2} + N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\right) + r_n^2N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2}.$$

Combining these bounds of (i)–(v) yields

$$\begin{aligned}(II) &\gtrsim (i) + (ii) - |(iii) + (iv) + (v)| \gtrsim r_n^2\left(N_r\|\mathbf{U}\|_{\mathrm{F}}^2 + T\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right) \\ &\quad - r_n^3\left(\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\ell_n^{1/2} + N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\right) - r_n^2N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2}.\end{aligned}$$

We then consider (III) in (A.10). Lemma 8 yields

$$|(III)| = r_n\eta_n\|\mathbf{W}_{\mathcal{S}} \circ \mathbf{V}_{\mathcal{S}}\|_1 \leq r_n\eta_n\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} \lesssim N_1^{1/2}r_n(\eta_n/\underline{b}_n^0)\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}},$$

where $\underline{b}_n^0 = \min_{(i,k)\in\mathcal{S}}|b_{ik}^0|$, with high probability.

Putting together the pieces obtained so far with (A.10), we have

$$\begin{aligned}&\inf_{\|\mathbf{U}\|_{\mathrm{F}}=C,\,\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}=C} Q_n(\mathbf{F}^0 + r_n\mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n\mathbf{V}_{\mathcal{S}}) - Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \\ &\gtrsim \inf_{\|\mathbf{U}\|_{\mathrm{F}}=C,\,\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}=C}\{(II) - |(I)| - |(III)|\} \\ &\gtrsim \inf_{\|\mathbf{U}\|_{\mathrm{F}}=C,\,\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}=C}\Big\{r_n^2\left(N_r\|\mathbf{U}\|_{\mathrm{F}}^2 + T\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right) \\ &- r_n^3\left(\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\ell_n^{1/2} + N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\right) - r_n^2N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2} \\ &- r_n\left(T^{1/2}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2} + N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\ell_n^{1/2}\right) - r_n^2\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2} - N_1^{1/2}r_n(\eta_n/\underline{b}_n^0)\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\Big\} \\ &\asymp r_n^2\left(N_r + T - N_1^{1/2}\ell_n^{1/2}\right)C^2 - r_n^3N_1^{1/2}C^3 - r_n\left(T^{1/2}\ell_n^{1/2} + N_1^{1/2}\ell_n^{1/2} + N_1^{1/2}(\eta_n/\underline{b}_n^0)\right)C.\end{aligned}$$

By condition (8), which is implied by (10), and the fact that $r_n \geq N_r^{1/2}T^{1/2}/(N_r \vee T) \geq 1$, we have

$$\begin{aligned}&\inf_{\|\mathbf{U}\|_{\mathrm{F}}=C,\,\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}=C} Q_n(\mathbf{F}^0 + r_n\mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n\mathbf{V}_{\mathcal{S}}) - Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \\ &\gtrsim r_n^2\left(N_r \vee T\right)C^2 - r_n^3N_1^{1/2}C^3 - r_nN_1^{1/2}(\eta_n/\underline{b}_n^0)C. \qquad \text{(A.11)}\end{aligned}$$

Furthermore, in (A.11), the first term dominates the second as the ratio, $r_n^3 N_1^{1/2}/\{r_n^2(N_r \vee T)\} = N_1^2/\{N_r^2 T^{1/2}\}\ell_n^{1/2}$, converges to zero by condition (11). Also, the first term dominates the third in (A.11) by the upper bound of conditions (12) as long as $C > 0$ is taken to be large enough. In consequence, the lower bound (A.11) tends to positive for such $C > 0$ and (A.9) holds.

(Second step) Set $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^o$ and $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}^o_{\mathcal{S}}$. If the estimator $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$ is indeed a minimizer of the unrestricted problem, $\min Q_n(\mathbf{F}, \mathbf{B})$ over $\mathbb{R}^{T\times r} \times \mathbb{R}^{N\times r}$, the proof completes. Note that $\operatorname{supp} \widehat{\mathbf{B}} = \mathcal{S}$ by the construction. Taking the same strategy as in Fan et al. (2014), we check the optimality of $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$. By a simple calculation, the (sub-)gradients of Q_n with respect to $\mathbf{F}$ and $\mathbf{B}$ are given by

$$\nabla_{\mathbf{F}} Q_n(\mathbf{F}, \mathbf{B}) = \mathbf{F}\mathbf{B}'\mathbf{B} - \mathbf{X}\mathbf{B}, \quad \nabla_{\mathbf{B}} Q_n(\mathbf{F}, \mathbf{B}) = \mathbf{B}\mathbf{F}'\mathbf{F} - \mathbf{X}'\mathbf{F} + \eta_n \mathbf{T},$$

where the (i,k)th element of $\mathbf{T} \in \mathbb{R}^{N\times r}$ is defined as

$$t_{ik} \begin{cases} = w_{ik} \operatorname{sgn}(b_{ik}) & \text{for} \quad b_{ik} \neq 0, \\ \in w_{ik}[-1, 1] & \text{for} \quad b_{ik} = 0. \end{cases}$$

Then $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$ is a strict minimizer of (7) if the following conditions hold:

$$\widehat{\mathbf{F}}\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{X}\widehat{\mathbf{B}} = \mathbf{0}_{T\times r}, \tag{A.12}$$

$$T\widehat{\mathbf{B}}_{\mathcal{S}} - (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}} + \eta_n \mathbf{W}_{\mathcal{S}} \circ \operatorname{sgn} \widehat{\mathbf{B}}_{\mathcal{S}} = \mathbf{0}_{N\times r}, \tag{A.13}$$

$$\left\| \mathbf{W}^{-}_{\mathcal{S}^c} \circ \left\{ T\widehat{\mathbf{B}}_{\mathcal{S}^c} - (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c} \right\} \right\|_{\max} < \eta_n, \tag{A.14}$$

where $\widehat{\mathbf{F}}'\widehat{\mathbf{F}} = T\mathbf{I}_r$ has been used, and $\mathbf{W}^- \in \mathbb{R}^{N\times r}$ is the matrix with its (i,k)th elements given by $1/w_{ik}$. Since $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}_{\mathcal{S}})$ is a minimizer of $Q_n(\mathbf{F}, \mathbf{B}_{\mathcal{S}})$, it satisfies the Karush–Kuhn–Tucker (KKT) conditions. Therefore, we only need to check condition (A.14), which is verified by Lemma 9. This completes the proof of Theorem 4. □

A.5 Proof of Corollary 2

Proof. Recall that $\hat{\alpha}_j = \log \widehat{N}_j / \log N$ with $\widehat{N}_j = |\operatorname{supp}(\widehat{\mathbf{b}}_j^{\mathsf{ada}})|$ and $\alpha_j = \log N_j / \log N$ by the definition. Because $\{\operatorname{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}}) = \operatorname{supp}(\mathbf{B}^0)\} \subset \{\widehat{N}_j = N_j \text{ for all } j = 1, \ldots, r\}$, we have

$$\begin{aligned} &\mathbb{P}\left(\hat{\alpha}_j = \alpha_j \text{ for all } j = 1, \ldots, r\right) \\ &= \mathbb{P}\left(\widehat{N}_j = N_j \text{ for all } j = 1, \ldots, r\right) \geq \mathbb{P}\left(\operatorname{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}}) = \operatorname{supp}(\mathbf{B}^0)\right). \end{aligned}$$

The last probability tends to one by the factor selection consistency. This completes the proof of Corollary 2. □

B Related Lemmas and the Proofs

Lemma 1. *Suppose that Assumptions 1–3 hold. Then the following inequalities simultaneously hold with probability at least* $1 - O\left((N \vee T)^{-\nu}\right)$*:*

(a) $\|\mathbf{E}\|_2 \lesssim (N \vee T)^{1/2}$,

(b) $\|\mathbf{E}\mathbf{B}^0\|_{\max} \lesssim N_1^{1/2} \log^{1/2}(N \vee T)$,

(c) $\|\mathbf{E}'\mathbf{F}^0\|_{\max} \lesssim T^{1/2}\log^{1/2}(N \vee T)$,

(d) $\max_{i\in\{1,\dots,N\}}\left|\sum_{t=1}^{T}\left(e_{ti}^2 - \mathbb{E}\, e_{ti}^2\right)\right| \lesssim T^{1/2}\log^{1/2}(N \vee T)$.

Proof of Lemma 1. (a) The tth row of $\mathbf{E}$, $\mathbf{e}_t \in \mathbb{R}^N$, is specified as $\mathbf{e}_t = \sum_{\ell=0}^{L}\mathbf{\Phi}_\ell \boldsymbol{\varepsilon}_{t-\ell}$, where $\boldsymbol{\varepsilon}_t \in \mathbb{R}^N$ is composed of i.i.d. subG(σ_ε^2) by Assumption 3. We also define $\widetilde{\mathbf{E}}_\ell = (\boldsymbol{\varepsilon}_{1-\ell},\dots,\boldsymbol{\varepsilon}_{T-\ell})' \in \mathbb{R}^{T\times N}$. Then, we can write $\mathbf{E} = \sum_{\ell=0}^{L_n}\widetilde{\mathbf{E}}_\ell\mathbf{\Phi}_\ell'$, so that the spectral norm is bounded as

$$\|\mathbf{E}\|_2 \le \sum_{\ell=0}^{L_n}\|\widetilde{\mathbf{E}}_\ell\|_2\|\mathbf{\Phi}_\ell\|_2 \le \max_{\ell\in\{0,\dots,L_n\}}\|\widetilde{\mathbf{E}}_\ell\|_2\sum_{\ell=0}^{\infty}\|\mathbf{\Phi}_\ell\|_2.$$

By Assumption 3, the last infinite sum is bounded from above. Because of the union bound and sub-Gaussianity (see Section 4 and Theorem 5.39 of Vershynin 2012), there is a positive constant M such that

$$\begin{aligned}\mathbb{P}\left(\max_{\ell\in\{0,\dots,L_n\}}\left\|(N\vee T)^{-1/2}\widetilde{\mathbf{E}}_\ell\right\|_2 > M\right) &\le L_n \max_{\ell\in\{0,\dots,L_n\}}\mathbb{P}\left(\left\|(N\vee T)^{-1/2}\widetilde{\mathbf{E}}_\ell\right\|_2 > M\right)\\ &\le 2(N\vee T)^{\nu}\exp\left(-c_1|N\vee T|\right) \le \exp\left(-c_2|N\vee T|\right)\end{aligned}$$

for some constants $c_1, c_2 > 0$, where the last inequality holds since ν is a fixed positive constant. Thus, $\|(N\vee T)^{-1/2}\mathbf{E}\|_2$ is bounded by a constant with probability at least $1 - \exp\left(-|c_2(N\vee T)|\right)$.

(b) By the definition, the (t,k)th element of $\mathbf{E}\mathbf{B}^0$ is given by $\mathbf{e}_t'\mathbf{b}_k^0 = \sum_{\ell=0}^{L_n}\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{\Phi}_\ell'\mathbf{b}_k^0$. Let $\tilde{b}_{\ell k,i}$ denote the ith element of $\mathbf{\Phi}_\ell'\mathbf{b}_k^0$. Then, we have

$$\begin{aligned}\|\mathbf{E}\mathbf{B}^0\|_{\max} &= \max_{t\in\{1,\dots,T\},k\in\{1,\dots,r\}}\left|\sum_{\ell=0}^{L_n}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right|\\ &\le \sum_{\ell=0}^{L_n}\max_{t\in\{1,\dots,T\},k\in\{1,\dots,r\}}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2\\ &\le \max_{t\in\{1,\dots,T\},k\in\{1,\dots,r\},\ell\in\{0,\dots,L_n\}}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right|\max_k\sum_{\ell=0}^{L_n}\|\mathbf{\Phi}_\ell\|_2\|\mathbf{b}_k\|_2\\ &\lesssim N_1^{1/2}\max_{t,k,\ell}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right|\sum_{\ell=0}^{\infty}\|\mathbf{\Phi}_\ell\|_2.\end{aligned}$$

Since $\{\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\}_{i=1}^N$ is a sequence of independent subG$(\sigma_\varepsilon^2\tilde{b}_{\ell k,i}^2)$ for each t,k,ℓ, we can further see that $\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^N\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i} \sim$ subG(σ_ε^2) by Lemma 2(b). Thus, the union bound yields

$$\begin{aligned}&\mathbb{P}\left(\max_{t,k,\ell}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right| > x\right)\\ &\le rT(L_n+1)\max_{t,k,\ell}\mathbb{P}\left(\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right| > x\right) \le 2r(N\vee T)^{\nu+1}\exp\left(-\frac{x^2}{2\sigma_\varepsilon^2}\right).\end{aligned}$$

Setting $x = \left(2\sigma_\varepsilon^2(2\nu+1)\log(N\vee T)\right)^{1/2}$ leads to

$$\max_{t,k,\ell}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^{N}\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right| \le \left(2\sigma_\varepsilon^2(2\nu+1)\log(N\vee T)\right)^{1/2},$$

which holds with probability at least $1 - O\left((N \vee T)^{-\nu}\right)$. This together with the first inequality achieves the result.

(c) Let $\widetilde{\mathbf{Z}}_\ell = (\boldsymbol{\zeta}_{1-\ell}, \ldots, \boldsymbol{\zeta}_{T-\ell})' \in \mathbb{R}^{T \times r}$. Then, by Assumptions 1 and 3, we can write $\mathbf{E}'\mathbf{F} = \sum_{\ell,m=0}^{L} \boldsymbol{\Phi}_\ell \widetilde{\mathbf{E}}'_\ell \widetilde{\mathbf{Z}}_m \boldsymbol{\Psi}'_m$. By the triangle inequality and property of matrix norms, we observe that

$$\begin{aligned}
\|\mathbf{E}'\mathbf{F}\|_{\max} &\leq \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\Phi}_\ell \widetilde{\mathbf{E}}'_\ell \widetilde{\mathbf{Z}}_m \boldsymbol{\Psi}'_m\|_{\max} \leq r^{1/2} \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\Psi}_m\|_2 \max_{i \in \{1,\ldots,N\}, k \in \{1,\ldots,r\}} \left|\boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k}\right| \\
&\leq r^{1/2} \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\Psi}_m\|_2 \max_{i,k} \left| \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| \max_i \|\boldsymbol{\phi}_{\ell,i}\|_2 \\
&\leq r^{1/2} \max_{\ell,m,i,k} \left| \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\Psi}_m\|_2 \max_i \|\boldsymbol{\phi}_{\ell,i}\|_2 \\
&\leq r^{1/2} \max_{\ell,m,i,k} \left| \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| \sum_{m=0}^{\infty} \|\boldsymbol{\Psi}_m\|_2 \sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2,
\end{aligned}$$

where $\boldsymbol{\phi}'_{\ell,i}$ and $\boldsymbol{\zeta}_{m,k}$ are the ith row vector of $\boldsymbol{\Phi}_\ell$ and kth column vector of $\widetilde{\mathbf{Z}}_m$, respectively. We can see that for each i and ℓ, the row vector

$$\boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell = \left(\sum_{j=1}^{N} \phi_{\ell,ij} \varepsilon_{1-\ell,j}, \ldots, \sum_{j=1}^{N} \phi_{\ell,ij} \varepsilon_{T-\ell,j} \right)$$

is composed of independent $\mathrm{subG}(\sigma_\varepsilon^2 \|\boldsymbol{\phi}_{\ell,i}\|_2^2)$. Since $\boldsymbol{\zeta}_{m,k} = (\zeta_{1-m,k}, \ldots, \zeta_{T-m,k})'$ consists of i.i.d. $\mathrm{subG}(\sigma_\zeta^2)$, Lemma 2(a) entails that

$$\|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} = \sum_{t=1}^{T} \left(\|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \sum_{j=1}^{N} \phi_{\ell,ij} \varepsilon_{t-\ell,j} \right) \zeta_{t-m,k}$$

is the sum of i.i.d. $\mathrm{subE}(4e\sigma_\varepsilon\sigma_\zeta)$. Therefore, the union bound and Bernstein's inequality for the sum of sub-exponential random variables give

$$\begin{aligned}
&\mathbb{P}\left(\max_{\ell,m,i,k} \left| T^{-1} \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| > x \right) \\
&\quad \leq rN(L_n+1)^2 \max_{\ell,m,i,k} \mathbb{P}\left(\left| T^{-1} \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| > x \right) \\
&\quad \leq 2r(N \vee T)^{2\nu+1} \exp\left\{ -\frac{T}{2} \left(\frac{x^2}{16e^2\sigma_\varepsilon^2\sigma_\zeta^2} \wedge \frac{x}{4e\sigma_\varepsilon\sigma_\zeta} \right) \right\}
\end{aligned}$$

for all $x > 0$. Putting $x = \left(32e^2\sigma_\varepsilon^2\sigma_\zeta^2(3\nu+1)T^{-1}\log(N \vee T)\right)^{1/2}$ gives

$$\max_{\ell,m,i,k} \left| \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1} \boldsymbol{\phi}'_{\ell,i} \widetilde{\mathbf{E}}'_\ell \boldsymbol{\zeta}_{m,k} \right| \leq \left(32e^2\sigma_\varepsilon^2\sigma_\zeta^2(3\nu+1)T\log(N \vee T)\right)^{1/2},$$

which holds with probability at least $1 - O\left((N \vee T)^{-\nu}\right)$. Combining this with the first bound yields the result.

(d) To obtain the result, we apply the Hanson–Wright inequality in Rudelson and Vershynin (2013). Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)' \in \mathbb{R}^m$ denote a random vector of m independent copies of $\varepsilon \sim \mathrm{subG}(\sigma_\varepsilon^2)$. Then the inequality states that for any (nonrandom) matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$,

$$\mathbb{P}\left(\left|\boldsymbol{\xi}'\mathbf{M}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{M}\boldsymbol{\xi}\right| > u\right) \leq 2\exp\left\{-c\min\left(\frac{u^2}{K^4\|\mathbf{M}\|_{\mathrm{F}}^2}, \frac{u}{K^2\|\mathbf{M}\|_2}\right)\right\}, \tag{A.15}$$

where c and K are positive constants such that $\sup_{k\geq 1} k^{-1/2}(\mathbb{E}\,|\varepsilon|^k)^{1/k} \leq K$. In our setting, we can take $K = 3\sigma_\varepsilon$ (e.g., Rigollet and Hütter (2017), Lemma 1.4).

Let $\boldsymbol{\phi}'_{\ell,i}$ denote the ith row vector of $\boldsymbol{\Phi}_\ell$. Then we have

$$\begin{aligned}\max_i \left|T^{-1}\sum_{t=1}^{T}\left(e_{ti}^2 - \mathbb{E}\,e_{ti}^2\right)\right| = \max_i \Bigg|T^{-1}\sum_{t=1}^{T}\sum_{\ell=0}^{L_n}\left(\boldsymbol{\varepsilon}'_{t-\ell}\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell} - \mathbb{E}\,\boldsymbol{\varepsilon}'_{t-\ell}\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\right)\\ + T^{-1}\sum_{t=1}^{T}\sum_{\ell,m=0,\ell\neq m}^{L_n}\boldsymbol{\varepsilon}'_{t-\ell}\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{m,i}\boldsymbol{\varepsilon}_{t-m}\Bigg|.\end{aligned}$$

The first term (sum of the diagonal elements) is bounded as

$$\begin{aligned}&\max_i \left|T^{-1}\sum_{t=1}^{T}\sum_{\ell=0}^{L_n}\left(\boldsymbol{\varepsilon}'_{t-\ell}\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell} - \mathbb{E}\,\boldsymbol{\varepsilon}'_{t-\ell}\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\right)\right|\\ &\leq T^{-1}\sum_{\ell=0}^{L_n}\max_i\left|\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell - \mathbb{E}\,\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell\right|,\end{aligned}$$

where $\tilde{\boldsymbol{\varepsilon}}_\ell = (\boldsymbol{\varepsilon}'_{1-\ell}, \ldots, \boldsymbol{\varepsilon}'_{T-\ell})' \in \mathbb{R}^{NT}$ and $\mathbf{A}_{\ell i} = \mathrm{diag}(\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}, \ldots, \boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}) \in \mathbb{R}^{NT\times NT}$. For any $\ell \in \{0, \ldots, L\}$ and $u > 0$, the Hanson–Wright inequality in (A.15) with the union bound gives

$$\begin{aligned}\mathbb{P}\left(\max_i\left|\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell - \mathbb{E}\,\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell\right| > u\right) &\leq N\max_i\mathbb{P}\left(\left|\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell - \mathbb{E}\,\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell\right| > u\right)\\ &\leq 2N\exp\left(-c\frac{u^2}{K^4\max_i\|\mathbf{A}_{\ell i}\|_{\mathrm{F}}^2}\right)\end{aligned}$$

Setting $u = ((\nu+1)/c)^{1/2}K^2\max_i\|\mathbf{A}_{\ell i}\|_{\mathrm{F}}\log^{1/2}(N\vee T)$ yields

$$\begin{aligned}&T^{-1}\sum_{\ell=0}^{L_n}\max_i\left|\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell - \mathbb{E}\,\tilde{\boldsymbol{\varepsilon}}'_\ell\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell\right| \leq K^2T^{-1}\log^{1/2}(N\vee T)\sum_{\ell=0}^{L_n}\max_i\|\mathbf{A}_{\ell i}\|_{\mathrm{F}}\\ &\lesssim T^{-1/2}\log^{1/2}(N\vee T)\sum_{\ell=0}^{L_n}\max_i\|\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\|_{\mathrm{F}} = T^{-1/2}\log^{1/2}(N\vee T)\sum_{\ell=0}^{\infty}\max_i\|\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}'_{\ell,i}\|_2\\ &\lesssim T^{-1/2}\log^{1/2}(N\vee T)\end{aligned}$$

with probability at least

$$1 - 2N\exp\left(-(\nu+1)\log(N\vee T)\right) = 1 - O((N\vee T)^{-\nu}).$$

The second term (sum of the off-diagonal elements) is bounded in the same way, and we omit it. For detail, see the proof of Lemma 7 in Fan et al. (2019). This completes all the proofs. □

Lemma 2. *Assume $X_i \sim$ ind. subG(α_i^2) and $Y_i \sim$ ind. subE(γ_i). Then, for any deterministic sequences (ϕ_i) and (ψ_i), the following statements are true:*

(a) $X_i X_j \sim$ subE$(4e\alpha_i\alpha_j)$ *for* $i \neq j$.

(b) $\sum_{i=1}^n \phi_i X_i \sim$ subG$(\sum_{i=1}^n \phi_i^2\alpha_i^2)$.

(c) $\sum_{i=1}^n \psi_i Y_i \sim$ subE$((\sum_{i=1}^n \psi_i^2\gamma_i^2)^{1/2}, \max_i |\psi_i|\gamma_i)$.

Proof. This proof was achieved in Uematsu and Tanaka (2019). □

Lemma 3. *Suppose the same conditions as Theorem 1. Then, for any $\mathbf{H} \in \mathbb{R}^{T\times k}$ $(k \leq r)$ such that $\mathbf{H}'\mathbf{H} = T\mathbf{I}_k$, the following inequalities simultaneously hold with probability at least $1 - O((N \vee T)^{-\nu})$:*

$$
\begin{aligned}
&(a)\ \ T^{-1}\left|\operatorname{tr}\mathbf{H}'\mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\mathbf{H}\right| \lesssim TN_1^{1/2}\log^{1/2}(N\vee T),\\
&(b)\ \ T^{-1}\operatorname{tr}\mathbf{H}'\mathbf{E}\mathbf{P}\mathbf{E}'\mathbf{H} \lesssim N\vee T,\\
&(c)\ \ \lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}') \lesssim T\vee N,\\
&(d)\ \ T^{-1}\operatorname{tr}(\mathbf{H}'\mathbf{E}\mathbf{Q}\mathbf{E}'\mathbf{H}) \lesssim T\vee N.
\end{aligned}
$$

Proof. Recall the notation based on the SVD of $\mathbf{C}^0$: $\mathbf{U}^0 = \mathbf{F}^0$ and $\mathbf{V}^0\mathbf{D}^0 = \mathbf{B}^0$. We derive the results on the event that Lemma 1 hold, which occurs with probability at least $1 - O((N \vee T)^{-\nu})$. Prove (a). Low rankness of each matrix and Lemma 1(b) give

$$
\begin{aligned}
\left|\operatorname{tr}\mathbf{H}'\mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\mathbf{H}\right| &\leq \|\mathbf{H}\mathbf{H}'\|_{\mathrm{F}}\|\mathbf{U}^0\|_{\mathrm{F}}\|\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\|_{\mathrm{F}} \lesssim \|\mathbf{H}\mathbf{H}'\|_{\mathrm{F}}\|\mathbf{U}^0\|_{\mathrm{F}}\|\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\|_2\\
&\lesssim TT^{1/2}T^{1/2}\|\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\|_{\max} \lesssim T^2N_1^{1/2}\log^{1/2}(N\vee T).
\end{aligned}
$$

Prove (b). Since the rank of $\mathbf{P}$ is at most r, Lemma 1(a) gives

$$
\operatorname{tr}\mathbf{H}'\mathbf{E}\mathbf{P}\mathbf{E}'\mathbf{H} \lesssim \|\mathbf{H}\mathbf{H}'\|_{\mathrm{F}}\|\mathbf{E}\mathbf{P}\mathbf{E}'\|_2 \leq T\|\mathbf{E}\|_2^2\|\mathbf{P}\|_2 \lesssim T(N\vee T).
$$

Prove (c). By the argument of the proof of Lemma A.8 in Ahn and Horenstein (2013) and Lemma 1(a), the bound

$$
\lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}') \leq \lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}' + \mathbf{E}\mathbf{P}\mathbf{E}') = \lambda_1(\mathbf{E}\mathbf{E}') = \|\mathbf{E}\|_2^2 \lesssim T\vee N.
$$

Prove (d). From the triangle inequality and result (c), we have

$$
\operatorname{tr}(\mathbf{H}'\mathbf{E}\mathbf{Q}\mathbf{E}'\mathbf{H}) \lesssim \|\mathbf{H}\mathbf{H}'\|_{\mathrm{F}}\|\mathbf{E}\mathbf{Q}\mathbf{E}'\|_2 \leq \|\mathbf{H}\mathbf{H}'\|_{\mathrm{F}}(\|\mathbf{E}\mathbf{E}'\|_2 + \|\mathbf{E}\mathbf{P}\mathbf{E}'\|_2) \lesssim T(T\vee N).
$$

This completes all the proofs of (a)–(d). □

Lemma 4. *Suppose the same conditions as Theorem 2. Then we have*

$$
\|\mathbf{E}\boldsymbol{\Delta}^b\|_2 \lesssim \|\boldsymbol{\Delta}^b\|_{\mathrm{F}}(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)
$$

with probability at least $1 - O((N \vee T)^{-\nu})$.

Proof. In the upper bound of (A.3), we consider a tighter bound of the second trace. The second trace in the upper bound of(A.3) is bounded as

$$\left|\operatorname{tr} \mathbf{E}\boldsymbol{\Delta}^b \boldsymbol{\Delta}^{f\prime}\right| \leq \|\mathbf{E}\boldsymbol{\Delta}^b\|_2 \|\boldsymbol{\Delta}^f\|_*.$$

Because $\widehat{\mathbf{B}}$ and $\mathbf{B}^0$ lie in the set $\mathcal{B}(\widetilde{N}) = \{\mathbf{B} \in \mathbb{R}^{N\times r} : \|\mathbf{B}\|_0 \lesssim \widetilde{N}/2\}$ for $\widetilde{N} \in [N_1, N]$ by Assumption 4, we have

$$\|\boldsymbol{\Delta}^b\|_0 \leq \|\widehat{\mathbf{B}}\|_0 + \|\mathbf{B}^0\|_0 \lesssim \widetilde{N}/2 + \widetilde{N}/2 \leq \widetilde{N}.$$

Define a set of sparse vectors $\mathcal{V}(\mathcal{A}) = \{\mathbf{v} \in \mathbb{R}^N \backslash \{\mathbf{0}\} : \|\mathbf{v}\|_0 = |\mathcal{A}|\}$ with $\mathcal{A} \subset \{1, \ldots, N\}$. Then, by the definition of the spectral norm, we have

$$\begin{aligned}
\|\mathbf{E}\boldsymbol{\Delta}^b\|_2^2 &= \max_{\mathbf{u}\in\mathbb{R}^r\backslash\{\mathbf{0}\}} \frac{\mathbf{u}'\boldsymbol{\Delta}^{b\prime}\mathbf{E}'\mathbf{E}\boldsymbol{\Delta}^b\mathbf{u}}{\mathbf{u}'\mathbf{u}} \leq \max_{\mathbf{u}\in\mathbb{R}^r\backslash\{\mathbf{0}\}} \frac{\mathbf{u}'\boldsymbol{\Delta}^{b\prime}\mathbf{E}'\mathbf{E}\boldsymbol{\Delta}^b\mathbf{u}}{\mathbf{u}'\boldsymbol{\Delta}^{b\prime}\boldsymbol{\Delta}^b\mathbf{u}} \max_{\mathbf{u}\in\mathbb{R}^r\backslash\{\mathbf{0}\}} \frac{\mathbf{u}'\boldsymbol{\Delta}^{b\prime}\boldsymbol{\Delta}^b\mathbf{u}}{\mathbf{u}'\mathbf{u}} \\
&\leq \max_{|\mathcal{A}|\lesssim\widetilde{N}} \max_{\mathbf{v}\in\mathcal{V}(\mathcal{A})} \frac{\mathbf{v}'\mathbf{E}'\mathbf{E}\mathbf{v}}{\mathbf{v}'\mathbf{v}} \|\boldsymbol{\Delta}^b\|_2^2 = \max_{|\mathcal{A}|\lesssim\widetilde{N}} \max_{\mathbf{v}_\mathcal{A}\in\mathbb{R}^{|\mathcal{A}|}} \frac{\mathbf{v}_\mathcal{A}'\mathbf{E}_\mathcal{A}'\mathbf{E}_\mathcal{A}\mathbf{v}_\mathcal{A}}{\mathbf{v}_\mathcal{A}'\mathbf{v}_\mathcal{A}} \|\boldsymbol{\Delta}^b\|_2^2 \\
&\leq \max_{|\mathcal{A}|\lesssim\widetilde{N}} \|\mathbf{E}_\mathcal{A}\|_2^2 \|\boldsymbol{\Delta}^b\|_2^2 \leq \max_{|\mathcal{A}|\lesssim\widetilde{N}} \max_{\ell\in\{1,\ldots,L_n\}} \|\widetilde{\mathbf{E}}_{\mathcal{A},\ell}\|_2^2 \left(\sum_{\ell=0}^{L_n} \|\boldsymbol{\Phi}_\ell\|_2\right)^2 \|\boldsymbol{\Delta}^b\|_2^2
\end{aligned}$$

where $\mathbf{v}_\mathcal{A} \in \mathbb{R}^{|\mathcal{A}|}$ consists of elements $\{v_i : i \in \mathcal{A}\}$ and $\mathbf{E}_\mathcal{A} \in \mathbb{R}^{T\times|\mathcal{A}|}$ is composed of the corresponding columns. Note that the second inequality holds since $\|\boldsymbol{\Delta}^b\mathbf{u}\|_0 \lesssim \widetilde{N}$, and in the last inequality $\widetilde{\mathbf{E}}_{\mathcal{A},\ell}$ is defined in the proof of Lemma 1. We also observe that $\sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2 < \infty$ by Assumption 3. By Theorem 5.39 of Vershynin (2012) with the union bound, for some positive constants c_1 and c_2 such that $c_1 < c_2$ and C, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{|\mathcal{A}|\lesssim\widetilde{N}} \max_{\ell\in\{0,\ldots,L_n\}} \|\widetilde{\mathbf{E}}_{\mathcal{A},\ell}\|_2 > C(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\right) \\
&\leq \binom{N}{c_1\widetilde{N}} (L_n+1) \max_{|\mathcal{A}|\lesssim\widetilde{N}} \max_{\ell\in\{1,\ldots,L_n\}} \mathbb{P}\left(\|\widetilde{\mathbf{E}}_{\mathcal{A},\ell}\|_2 > C(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\right) \\
&\lesssim N^{c_1\widetilde{N}}(N\vee T)^\nu \exp\left\{-c_2(\widetilde{N}\vee T)\log(N\vee T)\right\} \\
&= O\left((N\vee T)^{-\widetilde{N}\vee T}\right) = O\left((N\vee T)^{-\nu}\right).
\end{aligned}$$

Thus, we have with probability at least $1 - O((N\vee T)^{-\nu})$,

$$\|\mathbf{E}\boldsymbol{\Delta}^b\|_2 \lesssim \|\boldsymbol{\Delta}^b\|_2(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T) \leq \|\boldsymbol{\Delta}^b\|_\mathrm{F}(\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T),$$

giving the desired bound. □

Lemma 5. *Suppose the same conditions as Theorem 2. Then we have*

$$\|\boldsymbol{\Delta}\|_\mathrm{F}^2 \gtrsim \kappa_n \left(\|\widehat{\mathbf{F}} - \mathbf{F}^0\|_\mathrm{F}^2 + \|\widehat{\mathbf{B}} - \mathbf{B}^0\|_\mathrm{F}^2\right),$$

where $\kappa_n = N_r(N_r \wedge T)/N_1$.

Proof. Recall the notation based on the SVD of $\mathbf{C}^0$ and $\widehat{\mathbf{C}}$: $\mathbf{U}^0 = \mathbf{F}^0$, $\mathbf{V}^0\mathbf{D}^0 = \mathbf{B}^0$, $\widehat{\mathbf{U}} = \widehat{\mathbf{F}}$, and $\widehat{\mathbf{V}}\widehat{\mathbf{D}} = \widehat{\mathbf{B}}$. To establish the statement, we derive the following two inequalities:

$$(a) \quad \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \gtrsim \frac{N_r^2}{N_1}\|\widehat{\mathbf{U}} - \mathbf{U}^0\|_{\mathrm{F}}^2,$$
$$(b) \quad \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \gtrsim \frac{TN_r}{N_1}\|\widehat{\mathbf{D}}\widehat{\mathbf{V}}' - \mathbf{D}^0\mathbf{V}^{0\prime}\|_{\mathrm{F}}^2.$$

Using them, we can immediately obtain the result.

First we prove (a). We define matrices: $\widehat{\mathbf{U}}_* = T^{-1/2}\widehat{\mathbf{U}}$, $\widehat{\mathbf{D}}_* = \widehat{\mathbf{D}}\widehat{\mathbf{N}}^{1/2}$, $\widehat{\mathbf{V}}_* = \widehat{\mathbf{V}}\widehat{\mathbf{N}}^{-1/2}$, $\mathbf{U}_*^0 = T^{-1/2}\mathbf{U}^0$, $\mathbf{D}_*^0 = \mathbf{D}^0\mathbf{N}^{1/2}$, and $\mathbf{V}_*^0 = \mathbf{V}^0\mathbf{N}^{-1/2}$, where $\widehat{\mathbf{N}}$ is any p.d. diagonal matrix. Then, we can see that

$$T^{-1/2}\boldsymbol{\Delta} = \widehat{\mathbf{U}}_*\widehat{\mathbf{D}}_*\widehat{\mathbf{V}}_*' - \mathbf{U}_*^0\mathbf{D}_*^0\mathbf{V}_*^{0\prime} =: \boldsymbol{\Delta}_*.$$

For this expression, we can apply the proof of Lemma 3 in Uematsu et al. (2019). That is, under Assumptions 1 and 2, we have

$$\begin{aligned}\|\widehat{\mathbf{U}}_* - \mathbf{U}_*^0\|_{\mathrm{F}}^2 &= \sum_{k=1}^r \|\widehat{\mathbf{u}}_{*k} - \mathbf{u}_{*k}^0\|_2^2 \lesssim d_{*1}^2\|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \sum_{k=1}^r \frac{1}{\delta d_{*k}^4} \\ &= d_1^2 N_1 \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \sum_{k=1}^r \frac{1}{\delta d_k^4 N_k^2} \lesssim \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \frac{N_1}{N_r^2}.\end{aligned}$$

Rewriting this inequality with the original scaling gives result (a).

Next, we prove (b). We begin with rewriting $\boldsymbol{\Delta}_*$ as

$$\widehat{\mathbf{U}}_*(\widehat{\mathbf{D}}_*\widehat{\mathbf{V}}_*' - \mathbf{D}_*^0\mathbf{V}_*^{0\prime}) = \boldsymbol{\Delta}_* - (\widehat{\mathbf{U}}_* - \mathbf{U}_*^0)\mathbf{D}_*^0\mathbf{V}_*^{0\prime}.$$

The triangle inequality and unitary property of the Frobenius norm entail that

$$\|\widehat{\mathbf{D}}_*\widehat{\mathbf{V}}_*' - \mathbf{D}_*^0\mathbf{V}_*^{0\prime}\|_{\mathrm{F}} \le \|\boldsymbol{\Delta}_*\|_{\mathrm{F}} + \|(\widehat{\mathbf{U}}_* - \mathbf{U}_*^0)\mathbf{D}_*^0\|_{\mathrm{F}}.$$

We can bound the second term of the upper bound as in the proof of (a). That is, we have

$$\begin{aligned}\|(\widehat{\mathbf{U}}_* - \mathbf{U}_*^0)\mathbf{D}_*^0\|_{\mathrm{F}}^2 &\le \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 (cd_{*1}^2/\delta) \sum_{k=1}^r d_{*k}^{-2} \\ &= \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 (cd_1^2 N_1/\delta) \sum_{k=1}^r (d_k N_k^{1/2})^{-2} \lesssim \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \frac{N_1}{N_r}.\end{aligned}$$

Combining these inequalities gives

$$\begin{aligned}\|\widehat{\mathbf{D}}_*\widehat{\mathbf{V}}_*' - \mathbf{D}_*^0\mathbf{V}_*^{0\prime}\|_{\mathrm{F}}^2 &\le 2\|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 + 2\|(\widehat{\mathbf{U}}_* - \mathbf{U}_*^0)\mathbf{D}_*^0\|_{\mathrm{F}}^2 \\ &\lesssim \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \frac{N_1}{N_r} = T^{-1}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \left(1 + \frac{N_1}{N_r}\right).\end{aligned}$$

Noting that the left-hand side is equal to $\|\widehat{\mathbf{D}}\widehat{\mathbf{V}}' - \mathbf{D}^0\mathbf{V}^{0\prime}\|_{\mathrm{F}}^2$, we obtain

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 &\gtrsim T\left(1 + \frac{N_1}{N_r}\right)^{-1} \|\widehat{\mathbf{D}}\widehat{\mathbf{V}}' - \mathbf{D}^0\mathbf{V}^{0\prime}\|_{\mathrm{F}}^2 \\ &= \frac{TN_r}{N_1 + N_r}\|\widehat{\mathbf{D}}\widehat{\mathbf{V}}' - \mathbf{D}^0\mathbf{V}^{0\prime}\|_{\mathrm{F}}^2 \gtrsim \frac{TN_r}{N_1}\|\widehat{\mathbf{D}}\widehat{\mathbf{V}}' - \mathbf{D}^0\mathbf{V}^{0\prime}\|_{\mathrm{F}}^2.\end{aligned}$$

This completes the proof. □

Lemma 6. *Suppose that Assumptions 1–4 with $\widetilde{N} = N$ and conditions (9) and (10) hold. Then we have*

$$\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\max} \lesssim T^{-1/2} \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$.

Proof. Let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0$. Define

$$\mathcal{F} = \left\{ \mathbf{\Delta} = (\delta_{tk}) \in \mathbb{R}^{T \times r} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq C r_n^{\mathsf{PC}} \right\} \quad \text{with} \quad r_n^{\mathsf{PC}} = \frac{N_1 (NT)^{1/2} \log^{1/2}(N \vee T)}{N_r (N_r \wedge T)},$$

where C is some positive constant introduced in the proof of Theorem 4. By the definition of the PC estimator under PC1 restriction, we have

$$\begin{aligned}
\widehat{\mathbf{B}}_{\mathsf{PC}} &= T^{-1} \mathbf{X}' \widehat{\mathbf{F}}_{\mathsf{PC}} = T^{-1} (\mathbf{B}^0 \mathbf{F}^{0\prime} + \mathbf{E}') \widehat{\mathbf{F}}_{\mathsf{PC}} \\
&= T^{-1} (\mathbf{B}^0 \mathbf{F}^{0\prime} + \mathbf{E}') \mathbf{F}^0 + T^{-1} (\mathbf{B}^0 \mathbf{F}^{0\prime} + \mathbf{E}') \widehat{\mathbf{\Delta}} \\
&= \mathbf{B}^0 + T^{-1} \mathbf{E}' \mathbf{F}^0 + T^{-1} \mathbf{B}^0 \mathbf{F}^{0\prime} \widehat{\mathbf{\Delta}} + T^{-1} \mathbf{E}' \widehat{\mathbf{\Delta}}.
\end{aligned}$$

Then the triangle inequality implies that

$$\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\max} \leq T^{-1} \|\mathbf{E}' \mathbf{F}^0\|_{\max} + T^{-1} \|\mathbf{B}^0 \mathbf{F}^{0\prime} \widehat{\mathbf{\Delta}}\|_{\max} + T^{-1} \|\mathbf{E}' \widehat{\mathbf{\Delta}}\|_{\max}. \tag{A.16}$$

From Lemma 1(c), the first term of (A.16) is bounded by $T^{-1/2} \log^{1/2}(N \vee T)$ (up to a positive constant factor) with probability at least $1 - O((N \vee T)^{-\nu})$. We then consider the remaining two terms. For any $\mathbf{\Delta} \in \mathbb{R}^{T \times r}$, we have

$$\left\| \mathbf{B}^0 \mathbf{F}^{0\prime} \mathbf{\Delta} \right\|_{\max} \leq r \left\| \mathbf{B}^0 \right\|_{\max} \left\| \mathbf{F}^{0\prime} \mathbf{\Delta} \right\|_{\max} \lesssim \left\| \mathbf{F}^{0\prime} \mathbf{\Delta} \right\|_{\max} \leq \max_k \sum_\ell \left\| \mathbf{\Psi}_\ell \sum_t \boldsymbol{\zeta}_{t-\ell} \delta_{tk} \right\|_{\max}.$$

By Lemma 2 with Assumption 1, we have $z_{\ell,jk} := \sum_t \zeta_{t-\ell,j} \delta_{tk} \sim \mathrm{subG}(\sigma_\zeta^2 \|\boldsymbol{\delta}_k\|_2^2)$ for each fixed δ_{tk}, j, and ℓ. By the independence of $z_{\ell,jk}$ across j and Lemma 2 again, we have $\sum_j \psi_{\ell,ij} z_{\ell,jk} \sim \mathrm{subG}(\sigma_\zeta^2 \|\boldsymbol{\delta}_k\|_2^2 \|\mathbf{\Psi}_{\ell,i\cdot}\|_2^2)$ for each i, k, and ℓ. Therefore, for any fixed $\mathbf{\Delta}$ and ℓ, the subG tail inequality with the union bound entails that

$$\max_k \left\| \mathbf{\Psi}_\ell \sum_t \boldsymbol{\zeta}_{t-\ell} \delta_{tk} \right\|_{\max} \lesssim \max_i \|\mathbf{\Psi}_{\ell,i\cdot}\|_2 \|\mathbf{\Delta}\|_{\mathrm{F}} \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$. Because $\max_i \|\mathbf{\Psi}_{\ell,i\cdot}\|_2 \leq \|\mathbf{\Psi}_\ell\|_2$ by the definition of the spectral norm, we have

$$\sup_{\mathbf{\Delta} \in \mathcal{F}} \left\| \mathbf{F}^{0\prime} \mathbf{\Delta} \right\|_{\max} \leq C \sum_\ell \|\mathbf{\Psi}_\ell\|_2 r_n^{\mathsf{PC}} \log^{1/2}(N \vee T) \lesssim r_n^{\mathsf{PC}} \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$. Moreover, by the same argument as above with Assumption 3, we have

$$\sup_{\mathbf{\Delta} \in \mathcal{F}} \left\| \mathbf{E}' \mathbf{\Delta} \right\|_{\max} \lesssim r_n^{\mathsf{PC}} \log^{1/2}(N \vee T) = \frac{N_1 N^{1/2} T \log^{1/2}(N \vee T)}{N_r (N_r \wedge T)} \cdot T^{-1/2} \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$. Consequently, by Theorem 3 with condition (10), the bound in (A.16) becomes

$$\begin{aligned}\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\max} &\lesssim T^{-1/2}\log^{1/2}(N \vee T) + \frac{N_1 N^{1/2}\log^{1/2}(N \vee T)}{N_r(N_r \wedge T)} \cdot T^{-1/2}\log^{1/2}(N \vee T) \\ &= T^{-1/2}\log^{1/2}(N \vee T) + o(1)T^{-1/2}\log^{1/2}(N \vee T)\end{aligned}$$

with probability at least $1 - O((N \vee T)^{-\nu})$. This completes the proof of Lemma 6. □

Lemma 7. *Suppose the same conditions as Theorem 4. Then, for any deterministic matrices* $\mathbf{U} = (u_{tk}) \in \mathbb{R}^{T\times r}$ *and* $\mathbf{V} = (v_{ik}) \in \mathbb{R}^{N\times r}$*, the following inequalities simultaneously hold with probability at least* $1 - O((N \vee T)^{-\nu})$*:*

(a) $\left|\operatorname{tr}\mathbf{E}\mathbf{B}^0\mathbf{U}'\right| \lesssim N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\log^{1/2}(N \vee T)$,

(b) $\left|\operatorname{tr}\mathbf{E}'\mathbf{F}^0\mathbf{V}_{\mathcal{S}}'\right| \lesssim T^{1/2}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\log^{1/2}(N \vee T)$,

(c) $\left|\operatorname{tr}\mathbf{V}_{\mathcal{S}}'\mathbf{E}'\mathbf{U}\right| \lesssim \|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\log^{1/2}(N \vee T)$,

(d) $\left|\operatorname{tr}\mathbf{V}_{\mathcal{S}}\mathbf{U}'\mathbf{F}^0\mathbf{V}_{\mathcal{S}}'\right| \lesssim \|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\log^{1/2}(N \vee T)$,

(e) $\left|\operatorname{tr}\mathbf{B}^0\mathbf{U}'\mathbf{U}\mathbf{V}_{\mathcal{S}}'\right| \lesssim N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}$,

(f) $\left|\operatorname{tr}\mathbf{B}^0\mathbf{U}'\mathbf{F}^0\mathbf{V}_{\mathcal{S}}'\right| \lesssim N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\log^{1/2}(N \vee T)$.

Proof. Recall that $\mathbf{V}_{\mathcal{S}} \in \mathbb{R}^{N\times r}$ is defined as the matrix whose (i,k)th element is $v_{ik}1\{(i,k) \in \mathcal{S}\}$, where $\mathcal{S} = \operatorname{supp}(\mathbf{B}^0)$; see the proof of Theorem 4.

(a) First note that the (t,k)th element of $\mathbf{E}\mathbf{B}^0$ is given by $\mathbf{e}_t'\mathbf{b}_k^0$. We observe that

$$\left|\operatorname{tr}\mathbf{E}\mathbf{B}^0\mathbf{U}'\right| = \left|\operatorname{vec}(\mathbf{E}\mathbf{B}^0)'\mathbf{u}\right| \le r\max_k\left|\sum_{t=1}^T \mathbf{e}_t'\mathbf{b}_k^0 u_{tk}\right|,$$

where we have written as $\mathbf{u} = \operatorname{vec}(\mathbf{U})$. From Assumption 3, recall that $\mathbf{e}_t = \sum_{\ell=0}^L \mathbf{\Phi}_\ell\boldsymbol{\varepsilon}_{t-\ell}$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1},\ldots,\varepsilon_{tN})'$ with $\{\varepsilon_{ti}\}_{t,i} \sim$ i.i.d. subG(σ_ε^2). Let $\tilde{b}_{\ell k,i}$ denote the ith element of $\mathbf{\Phi}_\ell'\mathbf{b}_k^0$ as in the proof of Lemma 1(b). Then, we have

$$\begin{aligned}\max_k\left|\sum_{t=1}^T \mathbf{e}_t'\mathbf{b}_k^0 u_{tk}\right| &= \max_k\left|\sum_{t=1}^T\sum_{\ell=0}^L\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}u_{tk}\right| \\ &\le \sum_{\ell=0}^L\max_k\left|\sum_{t=1}^T\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}u_{tk}\right| \le \sum_{\ell=0}^L\max_k\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{t=1}^T\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}u_{tk}\right|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2 \\ &\le \max_{k,\ell}\left|\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{t=1}^T\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}u_{tk}\right|\max_k\|\mathbf{b}_k\|_2\sum_{\ell=0}^\infty\|\mathbf{\Phi}_\ell\|_2 \\ &\lesssim N_1^{1/2}\max_{k,\ell}\left|\sum_{t=1}^T u_{tk}\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\right|.\end{aligned}$$

Since $\{\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\}_i$ is a sequence of independent subG$(\sigma_\varepsilon^2\tilde{b}_{\ell k,i}^2)$ for each t,k,ℓ, we can see that $\{\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\}_t \sim$ indep. subG(σ_ε^2) by Lemma 2. Moreover, Lemma 2 gives

$$Z_{k\ell} := \sum_{t=1}^T u_{tk}\|\mathbf{\Phi}_\ell'\mathbf{b}_k\|_2^{-1}\sum_{i=1}^N \varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i} \sim \text{subG}(\sigma_\varepsilon^2\|\mathbf{u}_k\|_2^2).$$

Therefore, the subG tail inequality and the union bound entail

$$\mathbb{P}\left(\max_{k,\ell}|Z_{k\ell}|>x\right)\le r(L+1)\max_{k,\ell}\mathbb{P}\left(|Z_{k\ell}|>x\right)$$
$$\le 2r(N\vee T)^{\nu}\exp\left(-\frac{x^2}{2\sigma_\varepsilon^2\max_k\|\mathbf{u}_k\|_2^2}\right)\le 2r(N\vee T)^{\nu}\exp\left(-\frac{x^2}{2\sigma_\varepsilon^2\|\mathbf{U}\|_{\mathrm{F}}^2}\right).$$

Setting $x^2=4\sigma_\varepsilon^2\|\mathbf{U}\|_{\mathrm{F}}^2\nu\log(N\vee T)$ leads to getting the bound

$$\max_{k,\ell}|Z_{k\ell}|\le 2\sigma_\varepsilon\|\mathbf{U}\|_{\mathrm{F}}\nu^{1/2}\log^{1/2}(N\vee T)$$

with probability at least $1-O((N\vee T)^{-\nu})$. Thus the desired upper bound

$$\left|\operatorname{tr}\mathbf{E}\mathbf{B}^0\mathbf{U}'\right|\lesssim N_1^{1/2}\log^{1/2}(N\vee T)\|\mathbf{U}\|_{\mathrm{F}}$$

holds with probability at least $1-O((N\vee T)^{-\nu})$.

(b) As in the proof of Lemma 1, we write $\widetilde{\mathbf{E}}_\ell=(\boldsymbol{\varepsilon}_{1-\ell},\ldots,\boldsymbol{\varepsilon}_{T-\ell})'\in\mathbb{R}^{T\times N}$ and $\widetilde{\mathbf{Z}}_\ell=(\boldsymbol{\zeta}_{1-\ell},\ldots,\boldsymbol{\zeta}_{T-\ell})'\in\mathbb{R}^{T\times r}$. Then we can write $\mathbf{E}'\mathbf{F}=\sum_{\ell,m=0}^{L_n}\boldsymbol{\Phi}_\ell\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\boldsymbol{\Psi}_m'$ under Assumptions 1 and 3. By the same way as in (a), we have

$$\left|\operatorname{tr}\mathbf{E}'\mathbf{F}^0\mathbf{V}_{\mathcal{S}}'\right|=\left|\sum_{(i,k)\in\mathcal{S}}\sum_{\ell,m=0}^{L_n}\boldsymbol{\phi}_{\ell,i}'\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\boldsymbol{\psi}_{m,k}v_{ik}\right|\le\sum_{\ell,m=0}^{L_n}\left|\sum_{(i,k)\in\mathcal{S}}\boldsymbol{\phi}_{\ell,i}'\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\boldsymbol{\psi}_{m,k}v_{ik}\right|$$
$$=\sum_{\ell,m=0}^{L_n}\left|\sum_{(i,k)\in\mathcal{S}}v_{ik}\operatorname{tr}\boldsymbol{\psi}_{m,k}\boldsymbol{\phi}_{\ell,i}'\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\right|=\sum_{\ell,m=0}^{L_n}\left|\operatorname{tr}\boldsymbol{\Theta}_{\ell m}\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\right|,$$

where $\boldsymbol{\Theta}_{\ell m}:=\sum_{(i,k)\in\mathcal{S}}v_{ik}\boldsymbol{\psi}_{m,k}\boldsymbol{\phi}_{\ell,i}'$ with its (h,j)th component given by $\theta_{\ell m,hj}$ for $h=1,\ldots,r$ and $j=1,\ldots,N$. Recall that $\widetilde{\mathbf{E}}_\ell'=(\boldsymbol{\varepsilon}_{1-\ell},\ldots,\boldsymbol{\varepsilon}_{T-\ell})$ and $\widetilde{\mathbf{Z}}_m'=(\boldsymbol{\zeta}_{1-m},\ldots,\boldsymbol{\zeta}_{T-m})$ from the proof of Lemma 1. Then we have

$$\sum_{\ell,m=0}^{L_n}\left|\operatorname{tr}\boldsymbol{\Theta}_{\ell m}\widetilde{\mathbf{E}}_\ell'\widetilde{\mathbf{Z}}_m\right|=\sum_{\ell,m=0}^{L_n}\left|\sum_{h=1}^{r}\sum_{t=1}^{T}\left(\sum_{j=1}^{N}\theta_{\ell m,hj}\varepsilon_{t-\ell,j}\right)\zeta_{t-m,h}\right|$$
$$\le r\max_h\sum_{\ell,m=0}^{L_n}\left|\sum_{t=1}^{T}\left(\|\boldsymbol{\theta}_{\ell m,h}\|_2^{-1}\sum_{j=1}^{N}\theta_{\ell m,hj}\varepsilon_{t-\ell,j}\right)\zeta_{t-m,h}\right|\|\boldsymbol{\theta}_{\ell m,h}\|_2$$
$$\lesssim\max_{h,\ell,m}\left|\sum_{t=1}^{T}\left(\|\boldsymbol{\theta}_{\ell m,h}\|_2^{-1}\sum_{j=1}^{N}\theta_{\ell m,hj}\varepsilon_{t-\ell,j}\right)\zeta_{t-m,h}\right|\sum_{\ell,m=0}^{L_n}\|\boldsymbol{\theta}_{\ell m,h}\|_2$$
$$\lesssim\max_{h,\ell,m}\left|\sum_{t=1}^{T}\left(\|\boldsymbol{\theta}_{\ell m,h}\|_2^{-1}\sum_{j=1}^{N}\theta_{\ell m,hj}\varepsilon_{t-\ell,j}\right)\zeta_{t-m,h}\right|\max_h\sum_{\ell,m=0}^{L_n}\|\boldsymbol{\theta}_{\ell m,h}\|_2,,$$

where $\boldsymbol{\theta}_{\ell m,h}'$ is the hth row vector of $\boldsymbol{\Theta}_{\ell m}$. By the same reason as in the proof of Lemma 1(c), Lemma 2 entails that the inside of the absolute value is the sum of i.i.d. subE$(4e\sigma_\varepsilon\sigma_\zeta)$ random variables. Thus, the same bound in that proof can be used. Thus, applying the union bound, we obtain with probability at least $1-O((N\vee T)^{-\nu})$,

$$\max_{h,\ell,m}\left|\sum_{t=1}^{T}\left(\|\boldsymbol{\theta}_{\ell m,h}\|_2^{-1}\sum_{j=1}^{N}\theta_{\ell m,hj}\varepsilon_{t-\ell,j}\right)\zeta_{t-m,h}\right|\le(96e^2\sigma_\varepsilon^2\sigma_\zeta^2\nu T\log(N\vee T))^{1/2}.$$

Finally, we evaluate $\max_h \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\theta}_{\ell m,h}\|_2$. By the construction, we have

$$\max_h \sum_{\ell,m=0}^{L_n} \|\boldsymbol{\theta}_{\ell m,h}\|_2 = \max_h \sum_{\ell,m=0}^{L_n} \left(\sum_{j=1}^{N} \left(\sum_{(i,k)\in\mathcal{S}} v_{ik}\psi_{m,hk}\phi_{\ell,ij} \right)^2 \right)^{1/2}$$
$$\le \max_h \sum_{\ell,m=0}^{L_n} \left(\sum_{k=1}^{r} \psi_{m,hk}^2 \sum_{i,j=1}^{N} \phi_{\ell,ij}^2 \right)^{1/2} \|\mathbf{v}_{\mathcal{S}}\|_2 \le \sum_{m=0}^{\infty} \|\boldsymbol{\Psi}_m\|_2 \sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_{\mathrm{F}} \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}.$$

Thus the desired upper bound holds with probability at least $1 - O((N \vee T)^{-\nu})$.

(c) We observe that

$$\left|\operatorname{tr} \mathbf{V}_{\mathcal{S}}'\mathbf{E}'\mathbf{U}\right| = \left| \sum_{k=1}^{r} \sum_{t=1}^{T} \mathbf{v}_k' \mathbf{e}_t u_{tk} \right| \le \sum_{k=1}^{r} \sum_{\ell=0}^{L} \left| \sum_{t=1}^{T} \mathbf{v}_k' \boldsymbol{\Phi}_\ell \boldsymbol{\varepsilon}_{t-\ell} u_{tk} \right|.$$

By Assumption 3 and Lemma 2, we have $(\mathbf{v}_k'\boldsymbol{\Phi}_\ell\boldsymbol{\varepsilon}_{t-\ell})_t \sim$ indep. $\mathrm{subG}(\sigma_\varepsilon^2\|\mathbf{v}_k'\boldsymbol{\Phi}_\ell\|_2^2)$ for each k and ℓ. Thus, by Lemma 2 again, we further have $\sum_{t=1}^{T} \mathbf{v}_k'\boldsymbol{\Phi}_\ell\boldsymbol{\varepsilon}_{t-\ell}u_{tk} \sim \mathrm{subG}(\sigma_\varepsilon^2\|\mathbf{v}_k'\boldsymbol{\Phi}_\ell\|_2^2\|\mathbf{u}_k\|_2^2)$ for each k and ℓ. Therefore, the subG tail probability gives

$$\left| \sum_{t=1}^{T} \mathbf{v}_k'\boldsymbol{\Phi}_\ell\boldsymbol{\varepsilon}_{t-\ell}u_{tk} \right| \lesssim \|\mathbf{v}_k'\boldsymbol{\Phi}_\ell\|_2\|\mathbf{u}_k\|_2 \log^{1/2}(N \vee T) \le \|\boldsymbol{\Phi}_\ell\|_2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{U}\|_{\mathrm{F}} \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$. Consequently, we have

$$\left|\operatorname{tr} \mathbf{V}_{\mathcal{S}}'\mathbf{E}'\mathbf{U}\right| \lesssim \sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{U}\|_{\mathrm{F}} \log^{1/2}(N \vee T) \lesssim \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{U}\|_{\mathrm{F}} \log^{1/2}(N \vee T),$$

which yields the result.

(d) By the property of norms, we obtain

$$\left|\operatorname{tr} \mathbf{V}_{\mathcal{S}}'\mathbf{V}_{\mathcal{S}}\mathbf{U}'\mathbf{F}^0\right| \le \|\mathbf{V}_{\mathcal{S}}'\mathbf{V}_{\mathcal{S}}\|_*\|\mathbf{U}'\mathbf{F}^0\|_2$$
$$\le r^{3/2}\|\mathbf{V}_{\mathcal{S}}'\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{U}'\mathbf{F}^0\|_{\max} \lesssim \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2 \max_{j,k} \left| \sum_{t=1}^{T} u_{tj} f_{tk}^0 \right|.$$

By Assumption 1, the last stochastic part is evaluated as

$$\max_{j,k} \left| \sum_{t=1}^{T} u_{tk} f_{tk}^0 \right| = \max_{j,k} \left| \sum_{\ell=0}^{L_n} \sum_{m=1}^{r} \psi_{\ell,km} \sum_{t=1}^{T} u_{tj}\zeta_{t-\ell,m} \right|$$
$$\le r \max_{k,m} \sum_{\ell=0}^{L_n} |\psi_{\ell,km}| \max_{j,m} \left| \sum_{t=1}^{T} \zeta_{t-\ell,m}u_{tj} \right| \le r \max_{j,m,\ell} \left| \sum_{t=1}^{T} \zeta_{t-\ell,m}u_{tj} \right| \max_{k,m} \sum_{\ell=0}^{L_n} |\psi_{\ell,km}|$$
$$\lesssim \max_{j,m,\ell} \left| \sum_{t=1}^{T} \zeta_{t-\ell,m}u_{tj} \right| \sum_{\ell=0}^{\infty} \|\boldsymbol{\Psi}_\ell\|_2,$$

where $\{\zeta_{tm}\}_{t,m} \sim$ i.i.d. $\mathrm{subG}(\sigma_\zeta^2)$ and $\sum_{\ell=0}^{\infty} \|\boldsymbol{\Psi}_\ell\|_2$ is bounded. By Lemma 2(b), we have $\sum_{t=1}^{T} \zeta_{t-\ell,m}u_{tj} \sim \mathrm{subG}(\sigma_\zeta^2\|\mathbf{u}_j\|_2^2)$ for any j, m, ℓ. Thus, the subG tail inequality together

with the union bound establishes that

$$\mathbb{P}\left(\max_{j,m,\ell}\left|\sum_{t=1}^{T}\zeta_{t-\ell,m}u_{tj}\right|>x\right)\le r^2(L_n+1)\max_{j,m,\ell}\mathbb{P}\left(\left|\sum_{t=1}^{T}\zeta_{t-\ell,m}u_{tj}\right|>x\right)$$
$$\lesssim (N\vee T)^{\nu}\exp\left(-\frac{x^2}{2\sigma_\zeta^2\max_j\|\mathbf{u}_j\|_2^2}\right).$$

Setting $x=2\nu^{1/2}\sigma_\zeta\max_j\|\mathbf{u}_j\|_2\log^{1/2}(N\vee T)$ yields

$$\max_{j,m,\ell}\left|\sum_{t=1}^{T}\zeta_{t-\ell,m}u_{tj}\right|\le 2\sigma_\zeta\max_j\|\mathbf{u}_j\|_2\log^{1/2}(N\vee T)\lesssim\|\mathbf{U}\|_{\mathrm{F}}\log^{1/2}(N\vee T)$$

with probability at least $1-O((N\vee T)^{-\nu})$. This together with the first inequality yields the result.

(e) We observe that

$$\left|\operatorname{tr}\mathbf{B}^0\mathbf{U}'\mathbf{U}\mathbf{V}_{\mathcal{S}}'\right|\le\|\mathbf{V}_{\mathcal{S}}'\mathbf{B}^0\|_{\mathrm{F}}\|\mathbf{U}'\mathbf{U}\|_{\mathrm{F}}\lesssim N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}},$$

which gives the proof.

(f) By the property of norms, we obtain

$$\left|\operatorname{tr}\mathbf{V}_{\mathcal{S}}'\mathbf{B}^0\mathbf{U}'\mathbf{F}^0\right|\le\|\mathbf{V}_{\mathcal{S}}'\mathbf{B}^0\|_*\|\mathbf{U}'\mathbf{F}^0\|_2$$
$$\le r^{3/2}\|\mathbf{V}_{\mathcal{S}}'\mathbf{B}^0\|_{\mathrm{F}}\|\mathbf{U}'\mathbf{F}^0\|_{\max}\lesssim N_1^{1/2}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\max_{j,k}\left|\sum_{t=1}^{T}u_{tj}f_{tk}^0\right|.$$

Thus by the same argument as the proof of (d), we conclude that the stochastic part is bounded by $\|\mathbf{U}\|_{\mathrm{F}}\log^{1/2}(N\vee T)$, which occurs with probability at least $1-O((N\vee T)^{-\nu})$. This completes the proofs of (a)–(f). □

Lemma 8. *Suppose the same conditions as Theorem 4. Then we have with high probability*

$$\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}}\le\frac{2(rN_1)^{1/2}}{\underline{b}_n^0}.$$

Proof. Let $\underline{b}_n^0=\min_{(i,k)\in\mathcal{S}}|b_{ik}^0|$ and $\hat{\underline{b}}_n=\min_{(i,k)\in\mathcal{S}}|\hat{\underline{b}}_{ik}^{\mathsf{ini}}|$. For any $x>0$, we have

$$\mathbb{P}\left(\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}}>x\right)\le\mathbb{P}\left(\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}}>x\mid\hat{\underline{b}}_n>\underline{b}_n^0/2\right)+\mathbb{P}\left(\hat{\underline{b}}_n\le\underline{b}_n^0/2\right).\tag{A.17}$$

With setting $x=2(rN_1)^{1/2}/\underline{b}_n^0$, we verify that the upper bound of (A.17) tends to zero. The first probability of the upper bound is bounded as

$$\mathbb{P}\left(\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}}>\frac{2(rN_1)^{1/2}}{\underline{b}_n^0}\mid\hat{\underline{b}}_n>\underline{b}_n^0/2\right)\le\mathbb{P}\left(\frac{rN_1}{\hat{\underline{b}}_n^2}>\frac{4rN_1}{(\underline{b}_n^0)^2}\mid\hat{\underline{b}}_n>\underline{b}_n^0/2\right)$$
$$\le\mathbb{P}\left(\frac{2}{\hat{\underline{b}}_n\underline{b}_n^0}>\frac{4}{(\underline{b}_n^0)^2}\mid\hat{\underline{b}}_n>\underline{b}_n^0/2\right)=\mathbb{P}\left(\underline{b}_n^0/2>\hat{\underline{b}}_n\mid\hat{\underline{b}}_n>\underline{b}_n^0/2\right)=0.$$

By condition (12) and Lemma 6, the second probability of the upper bound of (A.17) is bounded as

$$\mathbb{P}\left(\hat{\underline{b}}_n\le\underline{b}_n^0/2\right)\le\mathbb{P}\left(\|\widehat{\mathbf{B}}_{\mathsf{ini}}-\mathbf{B}^0\|_{\max}\ge\underline{b}_n^0/2\right)=o(1).$$

These two bounds together with (A.17) imply the result. □

Lemma 9. *Suppose the same conditions as Theorem 4. Then we have*

$$\left\|\mathbf{W}^{-}_{\mathcal{S}^c} \circ (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} < \eta_n$$

with probability at least $1 - O((N \vee T)^{-\nu})$.

Proof. Let $\boldsymbol{\Delta} = (\delta_{tk}) = \mathbf{F} - \mathbf{F}^0$ and $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{F}} - \mathbf{F}^0$. Define

$$\mathcal{F} = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{T\times r} : \|\boldsymbol{\Delta}\|_{\mathrm{F}} \le C r_n\right\} \quad \text{with} \quad r_n = \frac{N_1(N_1T)^{1/2}\log^{1/2}(N\vee T)}{N_r(N_r \wedge T)},$$

where C is some positive constant introduced in the proof of Theorem 4. Then we have

$$\begin{aligned}
\left\|\mathbf{W}^{-}_{\mathcal{S}^c} \circ (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} &\le \left\|\mathbf{W}^{-}_{\mathcal{S}^c}\right\|_{\max} \left\|(\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} \\
&= \left\|\widehat{\mathbf{B}}^{\mathsf{ini}}_{\mathcal{S}^c}\right\|_{\max} \left\|(\mathbf{B}^0\mathbf{F}^{0\prime}\widehat{\boldsymbol{\Delta}})_{\mathcal{S}^c} + (\mathbf{E}'\widehat{\boldsymbol{\Delta}})_{\mathcal{S}^c} + (\mathbf{E}'\mathbf{F}^0)_{\mathcal{S}^c}\right\|_{\max} \\
&\le \left\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\right\|_{\max} \left(\sup_{\boldsymbol{\Delta}\in\mathcal{F}} \left\|(\mathbf{B}^0\mathbf{F}^{0\prime}\boldsymbol{\Delta})_{\mathcal{S}^c}\right\|_{\max} + \sup_{\boldsymbol{\Delta}\in\mathcal{F}} \left\|(\mathbf{E}'\boldsymbol{\Delta})_{\mathcal{S}^c}\right\|_{\max} + \left\|(\mathbf{E}'\mathbf{F}^0)_{\mathcal{S}^c}\right\|_{\max}\right) \\
&\le \left\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\right\|_{\max} \left(\sup_{\boldsymbol{\Delta}\in\mathcal{F}} \left\|\mathbf{B}^0\mathbf{F}^{0\prime}\boldsymbol{\Delta}\right\|_{\max} + \sup_{\boldsymbol{\Delta}\in\mathcal{F}} \left\|\mathbf{E}'\boldsymbol{\Delta}\right\|_{\max} + \left\|\mathbf{E}'\mathbf{F}^0\right\|_{\max}\right).
\end{aligned}$$

Therefore, by the same argument as the proof of Lemma 6, we observe that

$$\eta_n^{-1}\left\|\mathbf{W}^{-}_{\mathcal{S}^c} \circ (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} \lesssim \eta_n^{-1}\left\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\right\|_{\max}\left(T^{1/2} + r_n\right)\log^{1/2}(N\vee T),$$

where $T^{1/2} + r_n = T^{1/2}(1 + o(1))$ by condition (10). Lemma 6 and condition (13) yield

$$\begin{aligned}
\eta_n^{-1}\left\|\mathbf{W}^{-}_{\mathcal{S}^c} \circ (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} &\lesssim \eta_n^{-1}\left\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\right\|_{\max} T^{1/2}\log^{1/2}(N\vee T) \\
&\lesssim (2\eta_n)^{-1}\underline{b}_n^0 T^{1/2}\log^{1/2}(N\vee T)
\end{aligned}$$

with high probability. By the lower bound of condition (12) with taking sufficiently large positive constant factor in η_n, the desired strict inequality is obtained. □

C Additional Estimation Results

C.1 Estimating Exponents with Stock Returns

In addition to reporting the divergence rates in Section 6.3, we summarize the estimates of the factor loadings, focusing on analysis of the contributions of industrial sectors to the non-zero factor loadings. Such contributions can be regarded as measures of sensitivities of industrial sectors to the factor. Also we look into the signs of the factor loadings. Notice that the firm securities with negative loadings react to the factor in the opposite direction to those with positive loadings. Therefore, given the systematic risk factor, the different sign of the factor loadings could be interpreted as the different investment positions, for example, being long and short. Note that our analyses on the measures of sensitivities of industrial sectors and the signs of the factor loadings are conditional on the identification restrictions on the factors and factor loadings.

For the above purposes, all the firms are categorized to one of the ten industrial sectors based on Industry Classification Benchmark (ICB)[13]: (i) *Oil & Gas*; (ii) *Basic Materials*; (iii) *Industrials*; (iv) *Consumer Goods*; (v) *Health Care*; (vi) *Consumer Services*; (vii) *Telecommunications*; (viii) *Utilities*; (ix) *Financials*; (x) *Technology*. Then, for a given factor, the factor loadings are grouped into the negatives and the positives. For each group, the portion of the sum of the absolute value of the factor loadings which belong to each industrial sector is computed and reported. Specifically, we compute the following statistics for factor ℓ and industry s for given estimation window:

$$T^-_{b_\ell,s} = \frac{\sum_{i=1}^N \hat{b}_{i\ell} 1\{\hat{b}_{i\ell} < 0\} 1\{i \in s\}}{\sum_{i=1}^N \hat{b}_{i\ell} 1\{\hat{b}_{i\ell} < 0\}}, \quad T^+_{b_\ell,s} = \frac{\sum_{i=1}^N \hat{b}_{i\ell} 1\{\hat{b}_{i\ell} > 0\} 1\{i \in s\}}{\sum_{i=1}^N \hat{b}_{i\ell} 1\{\hat{b}_{i\ell} > 0\}}$$

where $\hat{b}_{i\ell}$ is the estimated factor loading of ith firm security, and $1\{A\}$ is the indicator function which takes unity if A is true and zero otherwise. We regard the portion $T^-_{b_\ell,s}$ and $T^+_{b_\ell,s}$ as the statistical measure of the negative and positive sensitivities of the sth industry to the ℓth factor. The average of the portion of the industrial sectors in S&P500 and the average of $T^-_{b_\ell,s}$ and $T^+_{b_\ell,s}$ for the four factors over the estimation windows τ =Sept 1998,...,April 2018, are reported in Figure SP2.

Figure SP2(a) shows the portion of the industrial sectors to which the securities consists of S&P500 belong, and the measure $T^+_{b_1,s}$ for the first factor. All the loadings to the first factor have the same sign (and it is chosen to be positive), which strongly suggests that this is the market factor. As one might expect, the 'beta' (the factor loading) of defensive industries, *Oil&Gas, Health Care, Telecoms* and *Utilities* is relatively small. The 'beta' of cyclical industries such as *Industrials*, *Financials* and *Basic Materials*, is noticeably high. The averages of the measures of negative and positive industrial contributions to the second factor loadings are reported in Figure SP2(b). It shows that *Utility* and *Financials* account for around 43% and 23% of negative loadings, respectively, while *Technology*, *Industrials* and *Basic Materials* share 40%, 17% and 14% of positive loadings, respectively. The averages of $T^-_{b_\ell,s}$ and $T^+_{b_\ell,s}$ for the third factor are reported in Figure SP2(c). It is clear that this is the *Oil&Gas* factor, which share the 67% of the negative loadings. *Financials*, *Consumer Services* and *Consumer Goods* share 29%, 23% and 19% of positive loadings, which means that these industrial sectors move opposite direction to the *Oil&Gas* with respect to the third factor. In view of Figure SP2(d), the dominating industry of the fourth factor is *Utility*, which share 43% of positive loading, together with *Health Care* with 17% of the share. No dominant industry is found for negative loadings, which are equally shared by cyclical industries.

In turn we discuss each factors in more details by analyzing Table SP1, Figures SP1 and SP2. The first factor does seem to be almost always "strong," in that the absolute sum of factor loadings is proportional to N. As reported in Table SP1, the average of α_1 over the month windows is 0.995 and standard deviation is very small (0.004) with the minimum value of 0.979. Also as is shown later, all the values of the factor loadings to this factor have the same sign, which strongly suggests that this is the market factor. Now we turn our attention to the rest of the factors. The divergence rates for the rest of the common components, α_2, α_2 and α_4, exhibit very different trajectory over the months, and their orders in terms of value change (i.e., their plots cross).

Let us see the trajectory of α_2. From Figure SP2(b), under our identification condition, the second factor can be understood of *Utility* and *Financials* versus *Technology*, *Industrials*

[13]Refer to FTSE Russell for more details about ICB.

and *Basic Materials*. In Figure SP1 it is seen that α_2 moves around 0.80 until October 1998, but from this month it sharply goes down and stay below 0.75 to October 1999. Then it sharply goes up to achieve 0.83 in February 2000. Indeed, this period corresponds to the turbulence of *Basic Material* stock index during 1998-2003, the fall of *Industrials* stock index around 2001-2 and the dot com bubble towards the peak in 2000. Since then, during most of the 2000s, α_2 goes above 0.85. After achieving the peak of 0.895 in April 2009, it steadily decreases and stabilizes around 0.75 from November 2012 onward, during which often this factor is not estimated but the fourth factor is.

Now let us analyze the move of α_3. From Figure SP2(c), under our identification condition, the third factor can be understood of *Oil&Gas* versus *Financials*, *Consumer Services* and *Consumer Goods*. According to Table SP1, α_3 has the lowest average. In Figure SP1, it looks co-moving with α_2, around 0.1 below, between September 1989 and July 2008. The exceptions are the periods from 1991 to 1992 and from 1999 to 2000, during which α_3 and α_2 are very close. A sharp rise of α_3 is observed from July 2008 to April 2009. This period coincides with the 2008 financial crisis. In just ten months, it goes up by 0.12, from 0.74 to 0.86. This can be interpreted that the *Oil&Gas* industry was sharply affected by the crisis. α_3 exceeds α_2 in December 2010, and this change of the order remains to the latest data point, April 2018.

Now let us analyze the move of α_4. From Figure SP2(d), under our identification condition, the fourth factor can be understood of *Utility* and *Health Care* versus cyclical industries. As shown in Figure SP1, the first estimate of the fourth factor appears in February 2004, with the value of α_4 being 0.80. Since its appearance, often it is not estimated but it is from March 2010 onward, seemingly becoming more and more stronger toward the latest month, April 2018. Since its first appearance, the value of α_4 is mostly between 0.75 and 0.80. After the sharp one off drop in February 2015,[14] α_4 rises to become the highest next to the first factor from November 2016 onward.

[14]This coincides with the period at bottom of the biggest sharp fall in oil price between 2014–2015.

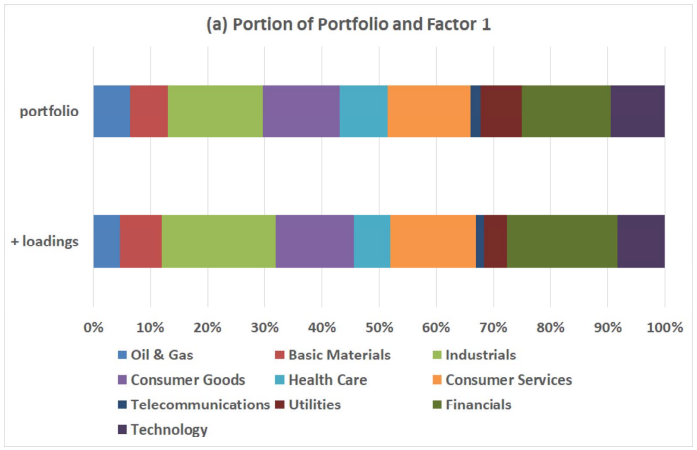

Figure 13: the portion of the industrial sectors in S&P500 and in the positive/negative 1st factor loadings

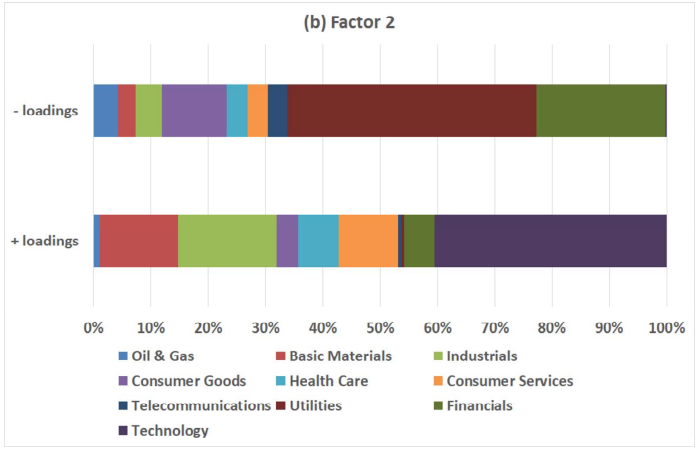

Figure 14: the portion of the industrial sectors in the positive/negative second factor loadings

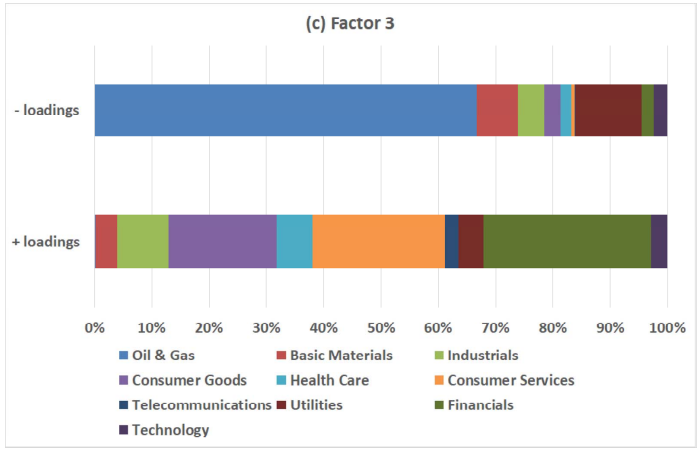

Figure 15: the portion of the industrial sectors in the positive/negative third factor loadings

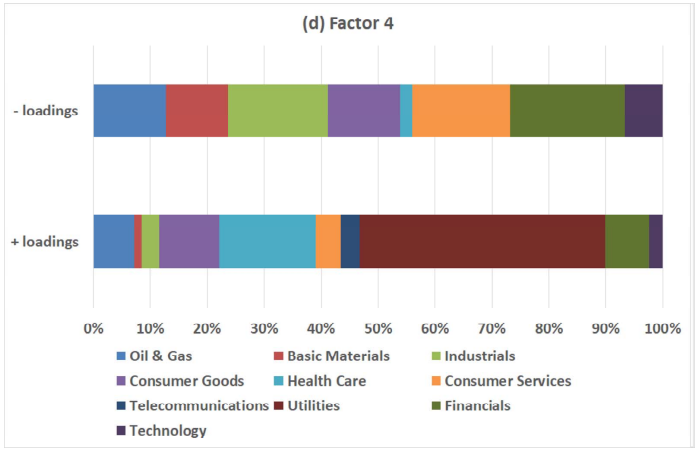

Figure 16: the portion of the industrial sectors in the positive/negative fourth factor loadings