Discussion Paper No. 1053

ESTIMATION OF WEAK FACTOR MODELS

Yoshimasa Uematsu
Takashi Yamagata

April 2019

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, 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

April 16, 2019

Abstract

In this paper, we propose a novel consistent estimation method for the approximate factor model of Chamberlain and Rothschild (1983), with large cross-sectional and time-series dimensions (N and T, respectively). Their model assumes that the r ($\ll N$) largest eigenvalues of data covariance matrix grow as N rises without specifying each diverging rate. This is weaker than the typical assumption on the recent factor models, in which all the r largest eigenvalues diverge proportionally to N, and is frequently referred to as the weak factor models. We extend the sparse orthogonal factor regression (SOFAR) proposed by Uematsu et al. (2019) to consider consistent estimation of the weak factors structure, where the k-th largest eigenvalue grows proportionally to N^{α_k} with some unknown exponents $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. Importantly, our method enables us to consistently estimate α_k as well. In our finite sample experiment, the performance of the new estimator uniformly dominates that of the principal component (PC) estimators in terms of mean absolute loss, and its superiority gets larger as the common components become weaker. We apply our method to analyze S&P500 firm security monthly returns from January 1984 to April 2018, and the results show that the first factor is consistently near strong, whilst the second to the fourth exponents vary over months between 0.90 and 0.65 and they cross. In another application, we consider out-of-sample performance of forecasting regressions for bond yield using extracted factors by our method and by the PC, and the forecasting performance test concludes that our method outperforms the PC method.

Keywords. Approximate factor models, Weak factors with sparse factor loadings, Determining the number of weak factors, Non-asymptotic error bound, Factor selection consistency, Firm security returns, Forecasting bond yields.

*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.

†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 on computation.

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 finance, macroeconomics and beyond. 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), 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. Fan et al. (2018) give an excellent review of the large dimensional factor models.

Suppose that a vector of zero-mean stationary time series $\mathbf{x}_t \in \mathbb{R}^N$, $t = 1, \ldots, T$, is generated from the factor model

$$\mathbf{x}_t = \mathbf{B}^0 \mathbf{f}_t^0 + \mathbf{e}_t, \tag{1}$$

where $\mathbf{B}^0 = (\mathbf{b}_1^0, \ldots, \mathbf{b}_N^0)' \in \mathbb{R}^{N \times r}$ with $\mathbf{b}_i^0 \in \mathbb{R}^r$ is a matrix of deterministic factor loadings, $\mathbf{f}_t^0 \in \mathbb{R}^r$ is a vector of probabilistic latent factors, and $\mathbf{e}_t \in \mathbb{R}^N$ is an idiosyncratic error vector. For a while suppose r is given. Imposing the familiar identification restrictions, $\mathbb{E}[\mathbf{f}_t^0 \mathbf{f}_t^{0\prime}] = \mathbf{I}_r$ and $\mathbf{B}^{0\prime}\mathbf{B}^0$ is a diagonal matrix with different elements, and assuming an exogeneity condition, we have

$$\mathbf{\Sigma}_x = \mathbf{B}^0 \mathbf{B}^{0\prime} + \mathbf{\Sigma}_e, \tag{2}$$

where $\mathbf{\Sigma}_x = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t']$ and $\mathbf{\Sigma}_e = \mathbb{E}[\mathbf{e}_t \mathbf{e}_t']$. If the eigenvalues of $\mathbf{\Sigma}_e$ are uniformly bounded, the asymptotic variation of $\mathbf{\Sigma}_x$ is fully controlled by loadings $\mathbf{B}^0$. Actually, we can observe that

$$\lambda_k(\mathbf{\Sigma}_x) \asymp \lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0)$$

for $k = 1, \ldots, r$ and $\lambda_k(\mathbf{\Sigma}_x)$ are uniformly bounded for $k = r+1, \ldots, N$. Most of the existing researches after Chamberlain and Rothschild (1983), including Connor and Korajczyk (1986, 1993), Stock and Watson (2002), Bai and Ng (2002, 2006, 2013), and Bai (2003), consider *strong factor models*, where it is assumed that all the r largest eigenvalues of $\mathbf{\Sigma}_x$ diverge proportional to N (i.e., $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N$ for all $k = 1, \ldots, r$) while the rest of the eigenvalues remain bounded, and employ the principal components (PC) estimator.

Actually, Chamberlain and Rothschild (1983) introduced the celebrated *approximate factor models*, where $\lambda_k(\mathbf{\Sigma}_x)$ diverges for each $k = 1, \ldots, r$ through assuming $\lambda_r(\mathbf{B}^{0\prime}\mathbf{B}^0) \to \infty$ as $N \to \infty$, and the rest of the eigenvalues remain bounded (due to the uniform boundedness of $\lambda_k(\mathbf{\Sigma}_e)$). This assumption is far less restrictive than that of the strong factor models, since the diverging rates of $\lambda_k(\mathbf{\Sigma}_x)$ can be less than or equal to N, which may be different for each $k = 1, \ldots, r$. Following De Mol et al. (2008), we call this model the *weak factor model* (WF model), in order to distinct from the strong factor (SF) model. To our knowledge, no method has been proposed which can estimate the WF models and each of the diverging rates of $\lambda_k(\mathbf{\Sigma}_x)$, jointly.

In this paper, we will fill this important gap in the literature of approximate factor models, by considering estimation of the WF models induced by *sparse* factor loadings. Specifically, $\mathbf{B}^0$ is supposed to contain many zeros such that $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N_k$, where $N_k = N^{\alpha_k}$ with some unknown exponent $0 < \alpha_k \leq 1$. Also note that our model nests the SF model when $\alpha_k = 1$ for all $k = 1, \ldots, r$. The sparse factor loadings are frequently observed in macroeconomic

and finance data. As an illustration, we have regressed each of 451 monthly security excess returns, which constitute the S&P500 index on December 2015, with 120 months observations back (among 500 securities) on the celebrated Fama and French (2015) five common factors, *Market, SMB, HML, RMW and CMA*, and an intercept. The numbers of securities, for which the *Market, SMB, HML, RMW or CMA* is significant at the 5% level t-test, are 446, 107, 126, 68 and 62, respectively.[1] Apart from the market factor, the common factors are *not* significantly different from zero for large portions of the securities. This evidence strongly suggests sparse factor loadings for the firm security returns and supports our approach[2].

In practice, the number of (weak) factors, r, is unknown, so that it has to be determined in prior to the model estimation. The widely used methods for determining the number of factors proposed by Connor and Korajczyk (1993), Bai and Ng (2002), Hallin and Liška (2007), Amengual and Watson (2007), Ahn and Horenstein (2013), Caner and Han (2014), among many others, assume the SF models. The only exceptions are Onatski (2010) and Kapetanios (2010), which permit to determine the number of weak factors. One of our contributions is to provide asymptotic justification of using the edge distribution (ED) estimator by Onatski (2010) for our WF models, and to report the experimental results. The evidence therein suggests that the ED estimator is reliable for the WF models while other estimators assuming SF models are in general unreliable.

Regarding the estimation of our WF models, we no longer rely on the eigenvalue problem but on the sparsity-inducing ℓ_1-norm regularization, unlike the PC estimator of the SF models. This problem requires numerical optimization, but is substantially complicated due to the imposition of both sparsity and orthogonality on the estimator. Despite this intrinsic difficulty, we propose a novel estimator of the WF models by employing the recently developed framework, the *sparse orthogonal factor regression* (SOFAR) of Uematsu et al. (2019). Hereafter the new estimator is called the WF-SOFAR estimator. As a theoretical contribution, we derive the non-asymptotic error bound for the WF-SOFAR estimator and its rate of convergence. Furthermore, we propose the *adaptive* WF-SOFAR estimator, which reduces the bias caused by the regularization and can enjoy *factor selection consistency*. This property asymptotically guarantees the true support recovery of the sparse loadings. It is remarkable that this enables us to consistently estimate each exponent α_k of the divergence rates as a corollary. Since, as discussed in Bailey et al. (2016), these are interpreted as the strength of the influence of the common factors and of the cross-sectional correlations, their estimation is of great interest to empirical research.[3] Perhaps surprisingly, our WF-SOFAR estimator can consistently estimate the WF models with α_k less than $1/2$. To our knowledge,

[1]Specifically, we run the time series regression $r_{ti} - r_{ft} = a_i + b_i\,(r_{mt} - r_{ft}) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{ti}$, where r_{ti} is the i-th security monthly return at the month t, r_{ft} is the one-month treasury bill rate, r_{mt} is the market return, SMB_t is the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks, HML_t is the difference between the returns on diversified portfolios of high and low B/M stocks, RMW_t is the difference between the returns on diversified portfolios of stocks with robust and weak profitability,and CMA_t is the difference between the returns on diversified portfolios of the stocks of low and high investment firms, which is called conservative and aggressive, and e_{ti} is the error term. Then we implement the t-tests for $b_i = 0$, $s_i = 0$, $h_i = 0$, $r_i = 0$ and $c_i = 0$, referring their absolute values to 1.96. The firm security return is computed as explained in Section 6.1, and other variables are obtained from the Kenneth R. French Data Library. See Fama and French (2015) for more details of the data and the regression.

[2]We repeated this exercise over the window months between September 1989 and April 2018, using the 1% and 5% significance level t-tests. The full results, which suggest sparse factor loadings, are summarized and reported in the online supplement.

[3]Bailey et al. (2016) propose a method for estimation and inference of the exponent of the largest divergence rate, α_1.

in the literature there have been no methods permitting consistent estimation of the factor model diverging at a rate slower than $N^{1/2}$. Note that the assumptions we will make are in line with the literature of the approximate factor models. Thus the statistical theory and the proofs we will explore are substantially different from those in Uematsu et al. (2019). In particular, the theoretical investigation of the adaptive estimator has not been considered in Uematsu et al. (2019) and is completely new to the literature.

We have to point out that the sparsity of the factor loadings is not necessarily rotation invariant. However, when the sparsity holds under the identification restrictions, each of the divergence rates of the r largest eigenvalues is invariant to the rotation of $\mathbf{B}^0$ and $\mathbf{f}_t^0$ with any invertible square matrix in $\mathbb{R}^{r\times r}$.[4] Finally, even when the model is not sparse at all, our WF-SOFAR estimator remains consistent to $\mathbf{f}_t^0$ and $\mathbf{B}^0$ because it coincide with the PC estimator, as the regularization coefficient tends to zero.

In order to evaluate the theoretical results of our WF-SOFAR estimator, we have implemented extensive finite sample experiments. In terms of the norm loss of estimators of factors, factor loadings and common components, the WF-SOFAR estimator uniformly dominates the PC estimator across all the designs we have considered. Perhaps surprisingly, the WF-SOFAR estimator of the common factors is uniformly more efficient than the PC estimator even in the most favorable experimental design of the PC, with the exponents of the divergence rates being 0.9. Also, the exponents of the divergence rates are correctly estimated, even when it is as small as 0.4, with sufficiently large sample size.

We consider two empirical applications. As the first empirical example, we apply our method to the S&P500 security excess returns, following Bailey et al. (2016). We estimate and plot the first four largest exponents of divergence rates using the 120 months estimation windows from January 1984 to April 2018. It is found that the first factor is near strong with the value of α_1 between 1.00 and 0.98, but the plots of the second to the fourth exponents vary over months between 0.90 and 0.65 and they also cross. A sharp rise of α_2 around the peak and burst of the dot-com bubble (March 2000) and a steep increase of α_3 at the time of the 2008 financial crisis draw our attention, which are in line with the well observed phenomenon that the correlation of the financial market tends to rise during the periods of market turmoils. As the second empirical example, we consider out-of-sample performance of forecasting regressions for bond yields using extracted factors via our method and the PC method, from a large number of macroeconomic (prediction) variables. We use analysis and data similar to that employed in Ludvigson and Ng (2009). For all the bond maturities considered, the forecasting performance test strongly rejects the null of the same forecasting performance, in favor of the alternative that our method outperforms the PC method.

We are not the first to consider the WF models. De Mol et al. (2008) consider the Bayesian forecasts for WF models and show that their rates of consistency are slower than the ones derived for the SF models based on the PC forecasts. Freyaldenhoven (2018) investigates the properties of the PC estimator in the WF models and proposes methods to estimate the number of common components diverging faster than a specific rate. Other related research includes Johnstone and Lu (2009), Onatski (2012) and Lettau and Pelger (2018), which 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.[5]

[4]Indeed, as can be seen in Section 6.1, in our empirical applications the sparse loadings assumption is appropriate under the identification restrictions.

[5]Bai and Ng (2017) focus on robust estimation of the factor models against outliers (in the sense of Candès et al. (2011)). For this purpose, they employ a ridge type estimator with ℓ_2 regularization, unlike our WF-SOFAR estimator which employs the ℓ_1 regularization. Bryzgalova (2016) considers estimation of

The rest of this paper is organized as follows. In Section 2, we formally define the WF models. Section 3 proposes the novel WF-SOFAR estimator of the WF models and considers its adaptive extension. Section 4 investigates the theoretical properties of the proposed estimators; specifically, we first consider determination of the number of weak factors and then derive the nonasymptotic error bound of the WF-SOFAR estimator and rate of convergence of the adaptive WF-SOFAR estimator. The error bound for the PC estimator in the WF models is also derived to compare with that of our estimator. Section 5 confirms the validity of our WF-SOFAR estimator in finite sample situations by Monte Carlo experiments. Section 6 gives empirical illustrations; the first example applies the WF-SOFAR to the firm security returns to uncover the market behavior and the second example reports forecasting performance of the bond yields by the WF-SOFAR. Finally, Section 7 concludes. The proofs for the main results are collected in Appendix A, and the related lemmas and their proofs as well as other supplementary materials are relegated to Appendices B-D in Online Supplements.

1.1 Notational remarks

For any matrix $\mathbf{M} = (m_{ti}) \in \mathbb{R}^{T\times N}$, we denote by $\|\mathbf{M}\|_{\mathrm{F}}$, $\|\mathbf{M}\|_2$, $\|\mathbf{M}\|_1$, and $\|\mathbf{M}\|_{\max}$ the Frobenius norm, ℓ_2-induced (spectral) norm, entrywise ℓ_1-norm, and entrywise ℓ_∞-norm, respectively. Specifically, they are defined by $\|\mathbf{M}\|_{\mathrm{F}} = (\sum_{t,i} m_{ti}^2)^{1/2}$, $\|\mathbf{M}\|_2 = \lambda_1^{1/2}(\mathbf{M}'\mathbf{M})$, $\|\mathbf{M}\|_1 = \sum_{t,i} |m_{ti}|$, and $\|\mathbf{M}\|_{\max} = \max_{t,i} |m_{ti}|$, where $\lambda_i(\mathbf{S})$ refers to the ith largest eigenvalue of any square matrix $\mathbf{S}$. We denote by $\mathbf{I}_N$ and $\mathbf{0}_{T\times N}$ the $N \times N$ identity matrix and $T \times N$ matrix with all the entries being zero, respectively. We use $\lesssim$ ($\gtrsim$) to represent $\leq$ ($\geq$) up to a positive constant factor. For any positive sequence a_n and b_n that converge to some points or diverge as $n \to \infty$, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Moreover, we denote by $a_n \sim b_n$ if $a_n/b_n \to 1$. We also use $X \sim \mu$ to denote that random variable X has distribution μ. 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 Weak Factor Models

Consider the factor model in (1). We investigate the case of large dimensional factor models, where the number of variables N and the number of observations T diverge at the same time. For the sake of convenience, we assume the existence of some underlying divergent 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 $n = N \wedge T \to \infty$. In Section 4, we also write $T = N^\tau$ for some constant $\tau > 0$ to understand the size of T relative to N. The number of factors r is unknown and to be determined in advance. 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 similarly rewrite model (1) as the matrix form

$$\mathbf{X} = \mathbf{F}^0\mathbf{B}^{0\prime} + \mathbf{E} = \mathbf{C}^0 + \mathbf{E}, \tag{3}$$

where $\mathbf{C}^0$ is called the matrix of common components. In order to identify $\mathbf{F}^0$ and $\mathbf{B}^0$ separately, r^2 restrictions are required. Throughout the paper, we impose the set of identifiability conditions called PC1 by Bai and Ng (2013): $\mathbf{F}^{0\prime}\mathbf{F}^0/T = \mathbf{I}_r$ and $\mathbf{B}^{0\prime}\mathbf{B}^0$ is a diagonal matrix

cross-sectional asset pricing models with non-diverging common components.

with distinct entries. We can see that the PC1 restriction makes the covariance matrix in (2) reduce to

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

As mentioned in the Introduction, Chamberlain and Rothschild (1983) consider approximate factor models in (3) allowing possibly different divergence rates of $\lambda_j(\mathbf{\Sigma}_x)$ for $j = 1, \ldots, r$ while $\lambda_{r+1}(\mathbf{\Sigma}_x)$ is bounded, which has recently been called the WF structure by De Mol et al. (2008). In this paper, we consider *sparsity-induced* WF models. Specifically, we assume *exactly 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 given by $N_k = N^{\alpha_k}$ for $k \in \{1, \ldots, r\}$, where $N \geq N_1 \geq \cdots \geq N_r$ (i.e., $1 \geq \alpha_1 \geq \cdots \geq \alpha_r > 0$) and α_k's are unknown. Note that N_r must diverge since $\alpha_r > 0$ and $N \to \infty$. We may theoretically consider the *weakly sparse* factor loadings; that is, $\mathbf{B}^0 = (b_{ik})$ such that $\sum_{i=1}^N |b_{ik}| \leq N_k$. This assumption does not necessarily require exact zeros in $\mathbf{B}^0$. However, we choose not to pursue this direction to avoid a complicated technical issue and leave it as a question for future work.

Combining the sparsity assumption with the PC1 restriction, we then observe that there exist some constants $d_1 > \cdots > d_r > 0$ such that

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

Therefore, under the assumption of uniform boundedness of $\lambda_j(\mathbf{\Sigma}_e)$, it is not hard to see that

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

where the equality in the first line holds because $\lambda_j(\mathbf{B}^0\mathbf{B}^{0\prime}) = \lambda_j(\mathbf{B}^{0\prime}\mathbf{B}^0)$ for $j \in \{1, \ldots, r\}$. Apparently, this specification fulfills the requirement of the WF structure.

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 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 $\mathbf{N} = \mathrm{diag}(N_1, \ldots, N_r)$, and $\mathbf{D}^0 = \mathrm{diag}(d_1, \ldots, d_r) \in \mathbb{R}^{r\times r}$ is composed of the well-separated singular values $d_1 > \cdots > d_r > 0$. The boundedness assumption on d_1 makes the signal strength fully controllable through the sparsity of the loadings. In view of PC1 on model (3), 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 PC1.

3 Estimation

In this section, we first review the PC estimator of Bai and Ng (2002) and then propose our estimator based on the SOFAR framework of Uematsu et al. (2019) for the WF models. In what follows, we refer to the WF-SOFAR estimator. 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 Review of the PC estimation

In the literature of estimating model (3) with the strong factor assumption, the PC estimator has been widely used for macroeconomic and financial time series analysis. The estimator is

defined as

$$(\widehat{\mathbf{F}}_{\mathsf{PC}}, \widehat{\mathbf{B}}_{\mathsf{PC}}) = \underset{(\mathbf{F},\mathbf{B})\in\mathbb{R}^{T\times\hat{r}}\times\mathbb{R}^{N\times\hat{r}}}{\arg\min} \frac{1}{2}\left\|\mathbf{X}-\mathbf{F}\mathbf{B}'\right\|_{\mathrm{F}}^2 \tag{4}$$
$$\text{subject to } \mathbf{F}'\mathbf{F}/T = \mathbf{I}_{\hat{r}} \text{ and } \mathbf{B}'\mathbf{B} \text{ diagonal,}$$

where $\hat{r}$ is the predetermined number of factors. It is well-known that this optimization problem reduces to the eigenvalue problem on $\mathbf{X}\mathbf{X}'$; more 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 $\widehat{\mathbf{B}}_{\mathsf{PC}} = \mathbf{X}'\widehat{\mathbf{F}}_{\mathsf{PC}}/T$. So the PC estimator does not need a special numerical optimization and is very easy to compute. Note that the PC estimator requires the SF assumption for proofs of the asymptotic normality. Under the WF assumption, however, that result is unlikely to hold.

3.2 WF-SOFAR estimation

Once the WF structure is well-defined via the sparsity assumption on $\mathbf{B}^0$, it is natural to introduce a sparsity-inducing penalty term, such as the ℓ_1-norm of $\mathbf{B}$, to (4) to obtain a sparse estimate of $\mathbf{B}^0$ in the same fashion as the Lasso by Tibshirani (1996). In fact, the WF-SOFAR estimator is conceptually defined as

$$(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}) = \underset{(\mathbf{F},\mathbf{B})\in\mathbb{R}^{T\times\hat{r}}\times\mathbb{R}^{N\times\hat{r}}}{\arg\min} \left\{\frac{1}{2}\left\|\mathbf{X}-\mathbf{F}\mathbf{B}'\right\|_{\mathrm{F}}^2 + \eta_n\|\mathbf{B}\|_1\right\} \tag{5}$$
$$\text{subject to } \mathbf{F}'\mathbf{F}/T = \mathbf{I}_{\hat{r}} \text{ and } \mathbf{B}'\mathbf{B} \text{ diagonal.}$$

We should note that the estimator is no longer computed by the eigenvalue problem. Even some algorithms used for the lasso estimation, such as coordinate descent, cannot be directly applied to the problem due to the seemingly incompatible restrictions, sparsity and orthogonality (diagonality).

In order to overcome this difficulty, we apply the SOFAR algorithm proposed by Uematsu et al. (2019) to solving (5). Generally 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 $\mathbf{C}^0$, in Section 2. Then for given $\hat{r}$, the SOFAR algorithm can solve (5) to get $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}) = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}\widehat{\mathbf{D}})$.

The algorithm to compute the WF-SOFAR estimate is based on the *augmented Lagrangian method* coupled with the *block coordinate decent,* and is numerically stable. Regularization parameter η_n must be determined prior to implementing the optimization. To select an "optimal" η_n, we can rely on cross-validation (CV) or information criteria, such as AIC, BIC, and GIC; see Uematsu et al. (2019) for more information on the computational aspects.

3.3 Adaptive WF-SOFAR estimation

It is interesting to observe which factors truly contribute to $\mathbf{x}_t$ for each t. In general, the lasso estimator in a linear regression model tends to select more variables than necessary due to the bias caused by the regularization that equally penalizes the elements in the coefficient vector. To reduce the bias, Zou (2006) proposed the adaptive lasso. Here we introduce the adaptive WF-SOFAR based on a similar principle. Let $\widehat{\mathbf{B}}^{\mathsf{ini}} = (\hat{b}_{ij}^{\mathsf{ini}})$ denote the first-stage

initial estimator, such as the PC estimator, of $\mathbf{B}^0$. Then the (i,j)th element of the weighting matrix $\mathbf{W} = (w_{ij})$ is defined as $w_{ij} = 1/|\hat{b}_{ij}^{\mathsf{ini}}|$. In using this weight matrix, the adaptive WF-SOFAR estimator is defined as a minimizer of the second-stage weighted SOFAR problem:

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

where $\mathbf{A} \circ \mathbf{B}$ represents the Hadamard product of two matrices, $\mathbf{A}$ and $\mathbf{B}$, of the same size. Through this construction, small weight w_{ij} indicates the importance of b_{ij}.

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 WF-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 subsequent section, we will prove this estimator is actually consistent for α_k.

4 Theory

In this section, we investigate the theoretical properties of the (adaptive) WF-SOFAR estimators, which are introduced in the previous section. We first reveal the asymptotic behavior of the eigenvalues of $\mathbf{X}\mathbf{X}'$ under the weak factor structure in Section 4.1. This helps us to determine the number of weak factors. Next we derive the non-asymptotic estimation error bound for the WF-SOFAR estimator in Section 4.2. Furthermore, the asymptotic property of the adaptive WF-SOFAR estimator is derived in Section 4.3.

In order to achieve these results, we need technical assumptions on the model. Following Rigollet and Hütter (2017), we first introduce sub-Gaussian and sub-exponential random variables. 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 $\mathbb{E}[\exp(sX)] \leq \exp(\sigma^2 s^2/2)$ for all $s \in \mathbb{R}$. In this case, we write $X \sim \mathrm{subG}(\sigma^2)$. A random variable $Y \in \mathbb{R}$ is said to be *sub-exponential* with parameter γ if $\mathbb{E}[Y] = 0$ and its moment generating function satisfies $\mathbb{E}[\exp(sY)] \leq \exp(\gamma^2 s^2/2)$ for all $|s| \leq \gamma^{-1}$. In this case, we write $Y \sim \mathrm{subE}(\gamma)$. The tails of sub-Gaussian random variables decay at least as fast as the Gaussian tail. More specifically, $X \sim \mathrm{subG}(\sigma^2)$ has the tail probability $\mathbb{P}(|X| > x) \leq 2\exp\{-x^2/(2\sigma^2)\}$, which is a crucial property in the proofs. Another important consequence is that $\sum_{i=1}^n a_i X_i \sim \mathrm{subG}(\sigma^2 \sum_{i=1}^n a_i^2)$ holds for $X_i \sim$ i.i.d. $\mathrm{subG}(\sigma^2)$. Note that σ^2 is not the variance of $X \sim \mathrm{subG}(\sigma^2)$. In fact, it is known that $\mathbb{E}[X^2] \leq 4\sigma^2$. In developing our theory, we assume sub-Gaussianity on the latent factors and idiosyncratic errors, but the condition is sometimes argued to be restrictive. In particular, a fatter-tailed distribution is preferable when financial time series are considered. In such a case, we may alternatively suppose random variable X that has the characteristic function $\exp(-|s|^\xi)$ for some $\xi \in (0,2)$. This satisfies $\mathbb{P}(|X| > x) \lesssim c_\xi x^{-\xi}$ and, moreover, $\mathbb{P}(\sum_{i=1}^n a_i X_i) \sim \sum_{i=1}^n |a_i|^\xi c_\alpha x^{-\xi}$. For more information, see Fan et al. (2014).

Define $L_n = (N \vee T)^\nu - 1$ for an arbitrary positive number ν. Throughout the paper, including all the proofs in Appendix A, ν is assumed to be fixed. Now we are in position to introduce assumptions on the WF model.

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

$$\mathbf{f}_t^0 = \sum_{\ell=0}^{L_n} \mathbf{\Psi}_\ell \boldsymbol{\zeta}_{t-\ell}, \quad \lim_{n\to\infty} \sum_{\ell=0}^{L_n} \mathbf{\Psi}_\ell \mathbf{\Psi}_\ell' = \mathbf{I}_r, \quad \min_{\ell\geq 0} \|\mathbf{\Psi}_\ell\|_2 > 0,$$

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 $N_k = N^{\alpha_k}$ such that $0 < N_r \leq \cdots \leq N_1 \leq N$ (i.e., $0 < \alpha_r \leq \cdots \leq \alpha_1 \leq 1$). Furthermore, $\mathbf{B}^{0\prime}\mathbf{B}^0 = \mathrm{diag}\{d_1^2 N_1, \ldots, d_r^2 N_r\}$ such that $d_{k-1}^2 - d_k^2 \geq \delta^{1/2} d_{k-1}^2$ for $k \in \{2, \ldots, r\}$ for some positive constant δ.

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

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

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 $\widehat{\mathbf{B}}$ in optimization (5) is given by $\mathcal{B}(\widetilde{N}) = \{\mathbf{B} \in \mathbb{R}^{N\times r} : \|\mathbf{B}\|_0 \lesssim \widetilde{N}/2\}$ for $\widetilde{N} \in [N_1, N]$.

Assumptions 1 and 3 specify the stochastic processes $\{\mathbf{f}_t\}$ and $\{\mathbf{e}_t\}$, respectively, to be stationary VMA(L_n), where $L_n \sim (N \vee T)^\nu$ diverges with an arbitrary fixed positive constant ν. This construction is regarded as the *asymptotic linear process*. Thanks to this specification, we can consider a wide range of cross-sectional and time series dependent processes for modeling $\{\mathbf{f}_t^0\}$ and $\{\mathbf{e}_t\}$ as asymptotic approximations of the Wold representation. By Assumption 3, we can observe that $\lambda_1(\mathbb{E}\,\mathbf{e}_t\mathbf{e}_t') < \infty$. Assumption 2 is key to our analysis and provides the sparse structure of the factor loadings $\mathbf{B}^0$ that leads to the WF models. The sparsity makes the divergence rate of $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0)$ possibly slower than N. This can be called *weak pervasiveness* in contrast to the so-called pervasive condition of Fan et al. (2013) that assumes the SF model (i.e., $N_k = N$ for all $k \in \{1, \ldots, r\}$). Note that PC1 is satisfied under Assumptions 1 and 2; the summability condition, together with the strong law of large numbers, gives $\mathbf{F}^{0\prime}\mathbf{F}^0/T = \mathbf{I}_r(1 + o(1))$ a.s. and the relative eigengap condition entails the eigen-separation required in $\mathbf{B}^{0\prime}\mathbf{B}^0$.

Assumption 4 is used only when the parameter estimation is considered. Note that $\mathbf{B}^0$ is included in $\mathcal{B}(\widetilde{N})$ for any $\widetilde{N} \in [N_1, N]$ under Assumption 2. 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 the true parameter, $\mathbf{B}^0$. The PC estimator always requires optimization on $\mathcal{B}(N)$ since it cannot be sparse, but the WF-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 milder condition on the WF model can be allowed. This means that the WF-SOFAR can estimate the model with a much wider class of (α_1, α_r) which cannot be consistently estimated by the PC. This will be one of the distinctive features of using the WF-SOFAR relative to the PC.

Under these assumptions, we obtain the following lemma which is useful for deriving the subsequent main theorems.

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

Lemma 1 guarantees that the stochastic terms can be bounded by some deterministic sequences with high probability. As a result, in the event that the lemma occurs, we may deal with these stochastic terms as if they were deterministic sequences in the proofs. It is worth mentioning that the results are of independent interest in the literature of high-dimensional time series analysis.

4.1 Determining the number of weak factors

Before investigating the properties of the estimator, we first observe the asymptotic behavior of the eigenvalues of $\mathbf{X}\mathbf{X}'$ under the WF model. This result yields important information for determining the number of weak factors, r. Here we write λ_j to express $\lambda_j\left((N \vee T)^{-1}\mathbf{X}\mathbf{X}'\right)$ for notational convenience.

Theorem 1. *Suppose that Assumptions 1–3 hold. Then for any finite integer* $k_{\max} \geq r$ *and for each* $j \in \{1, \ldots, r\}$ *such that*

$$\frac{N_1^{1/2} \log^{1/2}(N \vee T)}{N_j} = o(1), \tag{7}$$

the j*th largest eigenvalue of* $(N \vee T)^{-1}\mathbf{X}\mathbf{X}'$*, denoted by* λ_j*, satisfies*

$$\lambda_j \begin{cases} \geq \dfrac{d_j^2 N_j T}{N \vee T}(1 + o(1)) & \text{for } \ j \in \{1, \ldots, r\}, \\ = O(1) & \text{for } \ j \in \{r+1, \ldots, k_{\max}\}, \end{cases}$$

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

It is worth discussing a condition that makes λ_r diverge for determining the number of factors. If λ_r is bounded, any asymptotic argument cannot distinguish it from λ_{r+1}, which removes the possibility of correctly detecting r. Recall that $N_j = N^{\alpha_j}$ for some $\alpha_j \in (0, 1]$. Then, condition (7) for $j = r$ is

$$\frac{N_1^{1/2} \log^{1/2}(N \vee T)}{N_r} = N^{\alpha_1/2 - \alpha_r} \log^{1/2}(N \vee T) = o(1), \tag{8}$$

which holds if $\alpha_1/2 < \alpha_r$. Under condition (8), Theorem 1 establishes

$$\lambda_r \gtrsim \frac{N_r T}{N \vee T} = \begin{cases} N_r T/N & \text{for } \ N > T, \\ N_r & \text{for } \ N \leq T, \end{cases}$$

with high probability. Namely, λ_r always diverges as long as $N \leq T$, but when $N > T$ (i.e., $1 > \tau$ because $T = N^\tau$), diverging λ_r requires

$$\frac{N}{N_r T} = N^{1-\alpha_r-\tau} = o(1). \tag{9}$$

This condition is rephrased as $\alpha_r > 1 - \tau$. When λ_r tends to infinity while $\lambda_{r+1}, \ldots, \lambda_{k_{\max}}$ cluster around a single point, Theorem 1 suggests the means of determining the number of weak factors, r. This presents a case in which the method of Onatski (2010) works. In that paper, the *edge distribution* (ED) estimator,

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

was proposed, where δ is a fixed positive constant. In short, the following important corollary of Theorem 1 is obtained.

Corollary 1. *Suppose that Assumptions 1–3 hold. If conditions* (8) *and* (9) *are true, then λ_r diverges. In this case, for a 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 method based on a calibration; see that paper for full details. As long as 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 that we will compare to Ahn and Horenshtein's *eigenvalue ratio* (ER) and *growth ratio* (GR) and Bai and Ng's IC_3 and BIC_3.

4.2 Non-asymptotic error bound for the WF-SOFAR estimator

Hereafter, we suppose that the WF model satisfies conditions (8) and (9) and that r is known in view of Corollary 1. We first derive the non-asymptotic error bound of our WF-SOFAR estimator.

Theorem 2. *Suppose that Assumptions 1–4 and conditions* (8) *and* (9) *hold. Then for*

$$\eta_n \asymp T^{1/2} \log^{1/2}(N \vee T)$$

in optimization (5)*, the WF-SOFAR estimator satisfies the non-asymptotic error bound*

$$\|\widehat{\mathbf{F}} - \mathbf{F}^0\|_{\mathrm{F}} + \|\widehat{\mathbf{B}} - \mathbf{B}^0\|_{\mathrm{F}} \lesssim \frac{\kappa_n^{-1} N_1^{1/2} T^{1/2} \log^{1/2}(N \vee T)}{1 - \kappa_n^{-1} (\widetilde{N} \vee T)^{1/2} \log^{1/2}(N \vee T)}$$

with probability at least $1 - O((N \vee T)^{-\nu})$, where $\kappa_n = N_r(N_r \wedge T)/N_1$.

Theorem 2 yields the non-asymptotic error bound that holds with high probability for any combination of N and T. It is noteworthy that Theorem 2 is not a direct consequence of the error bounds achieved by Uematsu et al. (2019). In fact, Theorem 2 does not require the assumptions of sparse factors and minimum strength of the signals that are necessary in

Uematsu et al. (2019). Moreover, we assume the stochastic factors with serial dependence and the approximate factor structure in the error term by Assumptions 1 and 3, respectively, which greatly extends the original theory of Uematsu et al. (2019).

In order to enjoy consistency, κ_n must diverge relatively faster than $(\widetilde{N} \vee T)^{1/2} \log^{1/2}(N \vee T)$, where $\widetilde{N}$ is chosen from N_1 to N according to the parameter space of Assumption 4. That is, we need the condition

$$\gamma_n(\widetilde{N}) := \frac{N_1(\widetilde{N} \vee T)^{1/2} \log^{1/2}(N \vee T)}{N_r(N_r \wedge T)} = o(1) \tag{10}$$

for assigned $\widetilde{N} \in [N_1, N]$. With condition (10), we can obtain the rate of convergence for the WF-SOFAR estimator.

Corollary 2. *Suppose that Assumptions 1–4 and conditions* (9) *and* (10) *hold. Then the non-asymptotic error bound of the WF-SOFAR estimator in Theorem 2 reduces to*

$$T^{-1/2}\|\widehat{\mathbf{F}} - \mathbf{F}^0\|_{\mathrm{F}} = O_p\left(\frac{N_1^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}\right), \tag{11}$$

$$N^{-1/2}\|\widehat{\mathbf{B}} - \mathbf{B}^0\|_{\mathrm{F}} = O_p\left(\frac{N_1^{1/2}}{N^{1/2}}\frac{T^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}\right). \tag{12}$$

Observe that under condition (10), rates (11) and (12) always converge to zero. It is also easily seen that the rates do not depend on $\widetilde{N}$ because $(\widetilde{N} \vee T)^{1/2}$ in the denominator and that in $\gamma_n(\widetilde{N})$ cancel out. Note that (10) with any $\widetilde{N} \in [N_1, N]$ implies (8) because $\alpha_1 < \alpha_r + \alpha_r \wedge \tau - (\alpha_1 \vee \tau)/2 < \alpha_r + \alpha_r \wedge \tau \leq 2\alpha_r$.

Remark 1. Condition (10), together with condition (9), restricts the class of WF models in response to $\widetilde{N} \in [N_1, N]$ through limiting the region of $(\tau, \alpha_1, \alpha_r)$, where τ is such that $T = N^\tau$. To understand the meaning of condition (10), we consider the "sparsest" parameter space for $\widehat{\mathbf{B}}$. Condition (10) with $\widetilde{N} = N_1$, which is equivalently written as $\alpha_1 + (\alpha_1 \vee \tau)/2 < \alpha_r + \alpha_r \wedge \tau$, naturally brings the largest class of the WF models. Let us consider the region of $(\tau, \alpha_1, \alpha_r)$ restricted by this condition in turn. First we find that $\tau \in (1/2, 2]$. The upper bound of the difference in values of α_1 and α_r is found to be $1/4$, which is attainable when $\tau \in (3/4, 1]$ and $\alpha_1 = 1$. Likewise, it turns out that the lower bound of α_r is $1/3$, which is achievable when $\alpha_1 = \alpha_r$ and $\tau = 2/3$. In general, the sparser the factor loadings, the smaller value of τ is required. This is because the information on common factors can be collected only from the cross-section units with non-zero factor loadings. Hence larger cross-section units relative to T are required for the models with weaker factor structures. 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 WF-SOFAR can consistently estimate the WF 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 WF-SOFAR estimator seems quite robust to the violation of the restrictions on the region of $(\tau, \alpha_1, \alpha_r)$ discussed in this remark.

It is interesting to compare the results of the WF-SOFAR estimator obtained in Theorem 2 and Corollary 2 with those of the PC estimator for the WF models. Notice that the PC estimator must postulate the non-sparse parameter space (i.e., Assumption with $\widetilde{N} = N$).

Theorem 3. *Suppose that Assumptions 1–4 with $\widetilde{N} = N$ and conditions* (9) *and* (10) *hold. Then the PC estimator satisfies the non-asymptotic error bound*

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

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

We can find the difference from Theorem 2 in the error bound; both N_1 and $\widetilde{N}$ are replaced by N. In particular, thanks to the change in the denominator, condition (10) needs to hold with $\widetilde{N} = N$ to achieve consistency of the PC estimator. As discussed in Remark 1, this leads to a more restrictive class of WF models which can consistently be estimated.

Corollary 3. *Suppose that Assumptions 1–4 with $\widetilde{N} = N$ and conditions* (9) *and* (10) *hold. Then the non-asymptotic error bound of the PC estimator in Theorem 3 reduces to*

$$T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}} = O_p\left(\frac{N^{1/2}\gamma_n(N)}{(N \vee T)^{1/2}}\right), \tag{13}$$

$$N^{-1/2}\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\mathrm{F}} = O_p\left(\frac{T^{1/2}\gamma_n(N)}{(N \vee T)^{1/2}}\right). \tag{14}$$

Remark 2. The potential advantage of the WF-SOFAR estimator over the PC estimator for the WF models is twofold. First, since the PC estimation does not exploit sparse parameter space (i.e. $\widetilde{N} = N$ in Assumption 4), the PC cannot consistently estimate the WF models with α_r smaller than or equal to $1/2$, unlike the WF-SOFAR estimator (see Remark 1). Second, the rate of convergence of the PC estimator given by Corollary 3 deteriorates in comparison to that of the WF-SOFAR estimator given by Corollary 2 due to the replacement of N_1 to N in the numerators. This is because the PC estimator cannot take advantage of utilizing the sparsity in the model, but the WF-SOFAR can. The finite sample evidence in Section 5 strongly supports this observation. Finally, when the model has strong factors only, namely $N_1 = N_r = N$, the convergence rates of the WF-SOFAR and the PC estimators become identical when $\widetilde{N} = N$. Therefore, the WF-SOFAR estimator is likely to converge as fast as the PC estimator even when all the factors are strong.

4.3 Factor selection consistency of the adaptive WF-SOFAR estimator

We have derived the error bound and rate of convergence for the WF-SOFAR estimator. We next investigate the asymptotic property of the adaptive WF-SOFAR estimator introduced in Section 3.3. Specifically, we prove the *factor selection consistency*, which guarantees that the adaptive WF-SOFAR can recover the true sparsity pattern of the loadings and correctly select the relevant factors. As a corollary of the factor selection consistency, we finally establish the consistency of the estimated exponents of the divergence rates.

Before stating the theorem, define the index set of nonzero signals in $\mathbf{B}^0$ as

$$\mathcal{S} = \mathrm{supp}(\mathbf{B}^0) \subset \{1, \ldots, N\} \times \{1, \ldots, r\}.$$

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 $\mathbf{a}_{\mathcal{S}} = \mathrm{vec}\, \mathbf{A}_{\mathcal{S}} \in \mathbb{R}^{rN}$. We further write $\underline{b}_n^0 = \min_{(i,k)\in\mathcal{S}} |b_{ik}^0|$ and introduce condition

$$\frac{N_1^2}{N_r^2 T^{1/2}} = N^{2\alpha_1 - 2\alpha_r - \tau/2} = o(1). \tag{15}$$

Condition (15) further restricts the region of $(\tau, \alpha_1, \alpha_r)$ 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 case constrained only by (10), as long as $\tau = 1$. The lower bound of α_r can also achieve $1/3$ even if (15) is additionally supposed.

Theorem 4. *Suppose that Assumptions 1–3 and conditions* (9), (10) *and* (15) *hold and that the weighting matrix* $\mathbf{W}$ *is constructed by an estimator* $\widehat{\mathbf{B}}^{\mathsf{ini}}$ *such that*

$$\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\|_{\max} \lesssim \frac{N^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}} \tag{16}$$

with probability at least $1 - O((N \vee T)^{-\nu})$. *If* $\underline{b}_n^0$ *and* η_n *satisfy the conditions*

$$\underline{b}_n^0 \gtrsim \frac{N_r(N_r \wedge T)\eta_n}{N_1(N_r \vee T)T^{1/2}\log^{1/2}(N \vee T)} \vee \frac{N^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}, \tag{17}$$

$$\eta_n \asymp \frac{N^{1/2}T^{1/2}}{(\widetilde{N} \vee T)^{1/2}}, \tag{18}$$

then the adaptive WF-SOFAR estimator satisfies

$$T^{-1/2}\left\|\widehat{\mathbf{F}}^{\mathsf{ada}} - \mathbf{F}^0\right\|_{\mathrm{F}} = O_p\left(\frac{N_1^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}\right), \tag{19}$$

$$N^{-1/2}\left\|\widehat{\mathbf{B}}_{\mathcal{S}}^{\mathsf{ada}} - \mathbf{B}_{\mathcal{S}}^0\right\|_{\mathrm{F}} = O_p\left(\frac{N_1^{1/2}}{N^{1/2}}\frac{T^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}\right), \tag{20}$$

$$\mathbb{P}\left(\mathrm{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}}) = \mathcal{S}\right) \to 1. \tag{21}$$

As described in Section 3.1, we can use the PC estimator as the the initial estimator. Lemma 6 in Appendix A reveals that the PC estimator satisfies condition (16). Condition (17) restricts the magnitude of the minimum signal $\underline{b}_n^0 = \min_{(i,k)\in\mathcal{S}}|b_{ik}^0|$ from below. In the literature of sparse estimation, this kind of condition is said to be the *beta-min* condition and is frequently supposed to achieve variable selection consistency (e.g., Fan and Lv, 2011). Notice that the lower bound always converges to zero under condition (10) even in the worst case. As a result, condition (17) can asymptotically allow a wide class of the WF models. The rates of convergence (19) and (20) in Theorem 4 are identical to (11) and (12) in Corollary 2, respectively, hence they converge to zero.

Next, we prove that $\hat{\alpha}_k = \log \widehat{N}_k / \log N$, which is defined in Section 3.3, is consistent to α_k. Due to the factor selection consistency of (21) in Theorem 4, we immediately reach to the following corollary.

Corollary 4. *If the model selection consistency in* (21) *holds, then we have*

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

In the literature on the adaptive Lasso and penalized regressions with folded-concave penalties, such as the SCAD by Fan and Li (2001) or the MCP by Zhang (2014), the asymptotic normality can be established for the nonzero subvector of the estimator as well as the variable selection consistency; this is known as the *oracle property*. It was thought to be useful for statistical inference for the estimated model, but has been criticized recently by

Leeb and Pötscher (2005, 2006, 2008) and Pötscher and Leeb (2009), among others, on the basis that inferences based on the property lack uniformity over sequences of models that include even minor deviations from the beta-min condition. The same criticism could apply to the WF models and the adaptive WF-SOFAR estimator, and hence we do not consider inference based on the adaptive WF-SOFAR estimator in the present paper.

Instead of the adaptive estimation and the *oracle property,* a number of methods have been proposed for inference in high-dimensional linear regressions. Especially, the method called *debiasing* (*desparsification*) by Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014) has gained popularity. This framework tries directly to correct the Lasso estimator to remove the bias using the Karush-Kuhn-Tucker (KKT) conditions which the Lasso must satisfy. Such debiased Lasso estimators for time series models have been proposed by, for instance, Kock and Tang (2019). However, the question of how to desparsify the WF-SOFAR estimator for the WF models is nontrivial. We leave the problem as a future challenge.

5 Monte Carlo Experiments

In this section we investigate the finite sample behavior of estimators of the number of factors, and the behavior of the proposed WF-SOFAR estimators by means of Monte Carlo experiments.

In this section, indexes i, t, and k run over $1,\dots,N$, $1,\dots,T$, and $1,\dots,r$, respectively, unless otherwise noted. Denote by $N_k = \lfloor N^{\alpha_k} \rfloor$, where $\lfloor\cdot\rfloor$ is the floor function, with $0 < \alpha_k \leq 1$ for each k. We consider the following Data Generating Process (DGP):

$$x_{ti} = \sum_{k=1}^{r} b_{ik} f_{tk} + \sqrt{\theta} e_{ti}. \tag{22}$$

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 $i = 1,\dots,N_k$ and $b^*_{ik} = 0$ for $i = N_k+1,\dots,N$, and

$$f^*_{tk} = \rho_{fk} f^*_{t-1,k} + v_{tk}$$

with $v_{kt} \sim$ i.i.d.$N(0, 1-\rho^2_{fk})$ 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.$N(0,\sigma^2_{\varepsilon,ti})$ with $\sigma^2_{\varepsilon,ti}$ being set such that $\mathrm{Var}(e_{ti}) = 1$.

The above DGP is in line with that considered in the existing representative literature of approximate factor models, such as Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013), among many others. The main difference in our DGP from the literature is that the absolute sums of the factor loadings over i are allowed to diverge proportionally to N_k.[6]

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

[6]Ahn and Horenstein (2013) allow the number of pairs of cross-correlated errors, e_{ti}, to rise with N. The approximate factor models (of Chamberlain and Rothschild (1983)) are not in line with such a design.

particular, we consider the combinations $(0.9, 0.9)$, $(0.9, 0.8)$, $(0.9, 0.5)$, $(0.8, 0.8)$, $(0.8, 0.5)$, and more challenging cases, $(0.8, 0.4)$, $(0.5, 0.5)$ and $(0.5, 0.4)$. All the experimental results are based on 1,000 replications.

5.1 Determining the number of weak factors

As in Section 4.1, we write λ_k to express $\lambda_k((N \vee T)^{-1}\mathbf{X}\mathbf{X}')$. As discussed in that section, λ_r will diverge while λ_{r+1} will be bounded. Due to this, as summarized in Corollary 1, we can consider a class of estimators

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

where $k_{\max}$ $(\geq r)$ is given. The estimator is the maximum value of k with which $\lambda_k - \lambda_{k+1}$ exceeds the threshold δ. The question is how to choose the threshold value δ. We employ ED (edge distribution) algorithm to calibrate δ, which is proposed by Onatski (2010; p.1008). The algorithm exploits the square root shape of the edge of the eigenvalue distribution.[7] We denote the associated estimator of r as

$$\hat{r}_{\mathsf{ED}} = \hat{r}(\hat{\delta}) \tag{23}$$

where $\hat{\delta}$ is the calibrated δ.

We consider other competitor statistics to determine the number of factors. The ER (eigenvalue ratio) and GR (growth ratio) estimators of Ahn and Horenstein (2013) are closely related to Onatski (2010) in that they utilize the eigenvalues. Under the classical approximate factor model, the probability limit of $N^{-1}\sum_{i=1}^{N}\left(b_{ki}^0\right)^2$ is a bounded positive constant for all k, while the cross correlation among the idiosyncratic errors is typically bounded in N. Therefore, it is sufficient to distinguish a single big "jump" in the value of several largest eigenvalues. We define the ER and GR estimators as the maximizers of the following functions of the eigenvalues:

$$\hat{r}_{\mathsf{ER}} = \max_{1 \leq k \leq k_{\max}} ER(k), \quad ER(k) = \frac{\lambda_k}{\lambda_{k+1}}, \tag{24}$$

$$\hat{r}_{\mathsf{GR}} = \max_{1 \leq k \leq k_{\max}} GR(k), \quad GR(k) = \frac{\log\left[V(k-1)/V(k)\right]}{\log\left[V(k)/V(k+1)\right]} \tag{25}$$

with $V(k) = \sum_{\ell=k+1}^{N \wedge T - k_{\max} - 1} \lambda_\ell$. Note that, as these estimators just identify the maximum gap between the ordered eigenvalues, when the gap of divergence rate of factors, $\alpha_k - \alpha_{k+1}$, is relatively large, these statistics might pick up k as the estimator of r, even when $k < r$.

The information criteria proposed by Bai and Ng (2002) are widely used to determine the number of factors. Among these criteria, IC_3 and BIC_3 are of relevance for the models in which we are interested. IC_3 is appropriate for large N and T panel data and also recommended by Bai and Ng (2002). BIC_3 is recommended by Bai to include in the simulation exercises of Ahn and Horenstein (2013), particularly for cases in which idiosyncratic errors are cross-sectionally correlated. We define the associated estimators

$$\hat{r}_{\mathsf{IC}_3} = \min_{1 \leq k \leq k_{\max}} IC_3(k), \quad \hat{r}_{\mathsf{BIC}_3} = \min_{1 \leq k \leq k_{\max}} BIC_3(k),$$

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

where

$$IC_3(k) = \log v(k) + \frac{k \log(N \wedge T)}{N \wedge T},$$
$$BIC_3(k) = v(k) + \frac{k v(k_{\max})(N + T - k) \log(NT)}{NT}, \tag{26}$$

with

$$v(k) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{ti} - \sum_{\ell=1}^{k} \hat{b}_{\mathsf{PC}i\ell} \hat{f}_{\mathsf{PC}t\ell} \right).$$

Observe that computation of the information criteria requires estimation of the factor models, which is not necessary for $\hat{r}_{\mathsf{ED}}$, $\hat{r}_{\mathsf{ER}}$, and $\hat{r}_{\mathsf{GR}}$.

5.1.1 Results

Table 1 reports the average of the estimated number of factors over the replications by the ED of Onatski (2010), GR of Ahn and Horenstein (2013), and BIC_3 of Bai and Ng (2002), which are defined by (23), (25) and (26), respectively.[8] We employ the benchmark DGP of equation (22) with $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for $k = 1, \ldots, r$, $\beta = 0.2$, $\theta = 1$ and the sets of the values of exponents (α_1, α_2), as described above. We set the maximum number of factors, $k_{\max}$, as five. All the combinations of $N, T = 100, 200, 500, 1000$ are considered and all the results are based on 1,000 replications. As can been seen in Table 1, when α_1 and α_2 are both close to unity, all the methods perform very well, picking up the true number of factors with very high probability. Indeed, in the case of exponents $(\alpha_1, \alpha_2) = (0.9, 0.9), (0.9, 0.8), (0.8, 0.8)$, GR and BIC_3 choose the correct number of factors for all the replications, whilst ED very slightly tends to overestimate the number of factors.

However, the performance of GR and BIC3 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. For example, when $(\alpha_1, \alpha_2) = (0.9, 0.5)$, for $N = T = 100, 200, 500, 1000$, the averages of the estimated numbers of factors by GR are 1.15, 1.10, 1.05 and 1.00, respectively. It converges to one as the sample size increases, despite the fact that the true number of factors is two. With the same sets of exponents, for $N = T = 100, 200, 500, 1000$, the averages of the estimated number of factors by BIC_3 are 1.30, 1.36, 1.53 and 1.42, which apparently does not tend towards the true number of factors as the sample size increases. In contrast, ED performs very well, and its estimation quality is very similar to that when both exponents are close to unity. Similar observations for ED, GR and BIC_3 apply to the cases in which $(\alpha_1, \alpha_2) = (0.8, 0.5), (0.8, 0.4)$, however, for the latter case ED tends to underestimate the number of factors when $T < 500$ and $N < 200$. When $(\alpha_1, \alpha_2) = (0.5, 0.5)$, $(0.5, 0.4)$, as the values of α_1 and α_2 are similar, GR might be expected to pick up the right number of factors, but it tends to significantly underestimate r. For example, when $(\alpha_1, \alpha_2) = (0.5, 0.4)$, for $N = T = 1000$, the average of the estimated number of factors is 1.59. BIC_3 preforms even worse, giving the values between 1.00 and 1.23 for all the combinations of N and T. Under this challenging set up, ED consistently estimates the number of factors for sufficiently large T and N. For example, when $(\alpha_1, \alpha_2) = (0.5, 0.4)$, for $N = T = 500, 1000$, the averages of the estimated numbers of factors are 2.01 and 2.02.

[8] 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.

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 approximate factor models, whilst the methods proposed by Ahn and Horenstein (2013) and Bai and Ng (2002) may not be as reliable as the ED, in general.

Table 1: Average of the chosen number of factors for WF models by edge distribution algorithm (ED), Growth Ratio (GR), and BIC_3 methods: $r = 2$, $k_{\max} = 5$

	ED				GR				BIC3			
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
$(\alpha_1, \alpha_2) = (0.9, 0.8)$												
100	2.04	2.03	2.02	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
200	2.04	2.03	2.03	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
500	2.03	2.03	2.02	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1000	2.04	2.03	2.02	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
$(\alpha_1, \alpha_2) = (0.9, 0.5)$												
100	1.96	1.95	1.95	1.90	1.15	1.05	1.00	1.00	1.30	1.17	1.02	1.00
200	2.02	2.02	2.03	2.02	1.23	1.10	1.01	1.00	1.40	1.36	1.12	1.01
500	2.04	2.03	2.02	2.02	1.41	1.22	1.05	1.00	1.41	1.51	1.53	1.42
1000	2.02	2.03	2.02	2.02	1.35	1.22	1.05	1.00	1.43	1.51	1.53	1.42
$(\alpha_1, \alpha_2) = (0.8, 0.8)$												
100	2.05	2.03	2.02	2.01	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
200	2.05	2.03	2.03	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
500	2.03	2.03	2.02	2.02	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
1000	2.03	2.03	2.02	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
$(\alpha_1, \alpha_2) = (0.8, 0.4)$												
100	1.70	1.66	1.52	1.35	1.09	1.03	1.00	1.00	1.09	1.02	1.00	1.00
200	1.86	1.90	1.89	1.86	1.12	1.05	1.00	1.00	1.10	1.06	1.00	1.00
500	2.01	2.00	2.01	2.02	1.21	1.11	1.01	1.00	1.10	1.13	1.05	1.01
1000	1.94	2.00	2.01	2.02	1.18	1.11	1.01	1.00	1.12	1.13	1.05	1.01
$(\alpha_1, \alpha_2) = (0.5, 0.5)$												
100	1.80	1.84	1.79	1.72	1.64	1.62	1.65	1.62	1.08	1.01	1.00	1.00
200	1.98	2.02	2.02	2.02	1.69	1.80	1.83	1.87	1.11	1.09	1.01	1.00
500	2.03	2.03	2.02	2.02	1.87	1.91	1.95	1.98	1.14	1.21	1.23	1.14
1000	2.02	2.03	2.02	2.02	1.78	1.91	1.95	1.98	1.14	1.21	1.23	1.14
$(\alpha_1, \alpha_2) = (0.5, 0.4)$												
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

Notes: The DGP is $x_{ti} = \sum_{k=1}^{r} b_{ik}^0 f_{tk}^0 + \sqrt{\theta} e_{ti}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $b_{ik}^* \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, \lfloor N^{\alpha_k} \rfloor = N_k$ and $b_{ik}^* = 0$ for $i = N_k{+}1, \ldots, N$, with $0 < \alpha_k \leq 1$, $f_{tk}^* = \rho_{fk} f_{k,t-1}^* + v_{tk}$ for $t = 1, ..., T$, $v_{tk} \sim$ i.i.d.$N(0, 1 - \rho_{fk}^2)$ with $f_{0k}^* \sim$ i.i.d.$N(0,1)$ for $k = 1, \ldots, r$, $e_{ti} = \rho_e e_{t-1,i} + \beta \varepsilon_{t,i-1} + \beta \varepsilon_{t,i+1} + \varepsilon_{ti}$,

$\varepsilon_{ti} \sim$ i.i.d.$N(0, \sigma^2_{\varepsilon,ti})$, $\sigma^2_{\varepsilon,ti}$ is set so that $\text{Var}(e_{ti}) = 1$. Once b^*_{ik} and f^*_{tk} are generated, we apply Gram–Schmidt orthonormalization and form b^0_{ik} and f^0_{tk}, where $N^{-1}\sum_{i=1}^N b^0_{ik}b^0_{i\ell} = 1\{k = \ell\}$ and $T^{-1}\sum_{t=1}^T f^0_{tk}f^0_{t\ell} = 1\{k = \ell\}$. We set $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for $k = 1, \ldots, r$, $\beta = 0.2$, $\theta = 1$. The reported values are the average of estimated number of factors over the replications by the edge distribution algorithm (ED) of Onatski (2010), growth ratio (GR) proposed by Ahn and Horenstein (2013), and BIC_3 of Bai and Ng (2002), which are defined by (23), (25) and (26), respectively. All the results are based on 1,000 replications.

5.2 Finite sample properties of the WF-SOFAR estimator

In this subsection we investigate the finite sample properties of our WF-SOFAR estimator, and compare these with those of the PC estimator where appropriate. Here we treat the number of factors, r, as given, in order to distinguish the analysis from the quality of methods for choosing the number of factors, which is investigated above. We employ the benchmark DGP of equation (22) with $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for $k = 1, \ldots, r$, $\beta = 0.2$, $\theta = 1$. Initially we consider the exponents $(\alpha_1, \alpha_2) = (0.9, 0.9)$, $(0.9, 0.8)$, $(0.9, 0.5)$, $(0.8, 0.8)$, $(0.8, 0.5)$ and all the combinations of $N, T = 100, 200, 500, 1000$.[9] Then, we investigate more challenging cases, $(0.8, 0.4)$, $(0.5, 0.5)$ and $(0.5, 0.4)$. In these cases, we consider large sample sizes, $N, T = 500, 1000$, only. All the results are based on 1000 replications.

We report the results of the adaptive WF-SOFAR estimator with BIC, which we recommend to use.[10] We consider the PC estimator as defined in Section 3.1.

For performance comparison purposes, we consider the following ℓ_2-norm losses based on the kth scaled estimators:

$$
\begin{aligned}
L(\widehat{\mathbf{F}}) &= \left\| \sum_{k=1}^{r} T^{-1/2} \left[\text{abs}(\widehat{\mathbf{f}}_k) - \text{abs}(\mathbf{f}^0_k) \right] \right\|_2, \\
L(\widehat{\mathbf{B}}) &= \left\| \sum_{k=1}^{r} N_k^{-1/2} \left[\text{abs}(\widehat{\mathbf{b}}_k) - \text{abs}(\mathbf{b}^0_k) \right] \right\|_2, \\
L(\widehat{\mathbf{C}}) &= \left\| \sum_{k=1}^{r} T^{-1/2} N_k^{-1/2} \left[\widehat{\mathbf{C}}_k - \mathbf{C}^0_k \right] \right\|_{\text{F}},
\end{aligned}
\tag{27}
$$

where 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. Observe that these norm losses are not sensitive to the sign indeterminacy of the estimators (i.e. $\mathbf{f}^0_k\mathbf{b}^{0\prime}_k = \left(-\mathbf{f}^0_k\right)\left(-\mathbf{b}^{0\prime}_k\right)$) and the change of the order of the factor components (e.g., for $r = 2$, the estimated first factor can be of the true second factor).

5.2.1 Results

Table 2 reports the averages and standard deviations (s.d.) of the estimates of α_1 and α_2 based on Corollary 4, and the average of the norm losses (multiplied by 100) of the scaled estimated factors, factor loadings, and common components by the WF-SOFAR (WS in the tables) and PC estimators over the replications. The experimental design is essentially

[9] Note that when the values of α_1 are 0.9 and 0.8, the associated lowest bounds of α_r implied by condition (10) are 0.675 and 0.6, respectively.

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

determined by the two values of the exponents, (α_1, α_2). In this table, we consider $(\alpha_1, \alpha_2) = (0.9, 0.9)$, $(0.9, 0.8)$, $(0.9, 0.5)$, $(0.8, 0.8)$, $(0.8, 0.5)$. Each panel of the table has two column blocks for $T = 100$ and 200 (and $T = 500$ and 1000), and each column block for given T contains four row blocks for $N = 100, 200, 500, 1000$.

First, let us focus on the WF-SOFAR estimates of (α_1, α_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, when $(\alpha_1, \alpha_2) = (0.9, 0.5)$, for $N = T = 100, 200, 500, 1000$, the averages of $(\hat{\alpha}_1, \hat{\alpha}_2)$ are $(0.85, 0.54)$, $(0.87, 0.53)$, $(0.88, 0.52)$, $(0.89, 0.52)$ and the standard deviations (rounded to two decimal places) are $(0.02, 0.07)$, $(0.01, 0.04)$, $(0.00, 0.03)$, $(0.00, 0.02)$. This precision is remarkable since given α_1 is 0.9, the value of the lower bound of α_r implied by condition (10) is 0.675, which is much larger than the actual value considered here, 0.5. Similar comments apply to the results for the combinations $(\alpha_1, \alpha_2) = (0.8, 0.5)$.

Now we turn our attention to the performance of the WF-SOFAR and PC estimates. In terms of the norm loss in (27), the WF-SOFAR uniformly beats the PC across all the designs we have considered. Perhaps surprisingly, the WF-SOFAR estimate of the factors is much more accurate than the PC estimator even in the most favorable experimental design to the PC, with $(\alpha_1, \alpha_2) = (0.9, 0.9)$. For example, for $N = 100, 200, 500, 1000$ with $T = 100$, the average squared norm losses (scaled by 100) of the WF-SOFAR and PC factor estimates, $\{\text{WS}, \text{PC}\}$, are $\{6.2, 11.6\}$, $\{4.6, 10.1\}$, $\{3.5, 9.3\}$, $\{2.8, 9.0\}$, respectively. In terms of their ratios, the WF-SOFAR factor estimates become more accurate than the PC factor estimates as N rises with given T.

As expected, the accuracy of the WF-SOFAR estimator of factor loadings is uniformly superior to that of the PC estimator. This gap in accuracy becomes wider when the exponents are further from unity. For example, when $(\alpha_1, \alpha_2) = (0.9, 0.5)$, for $N = 100, 200, 500, 100$ with $T = 500$, the average squared norm losses (scaled by 100) for the WF-SOFAR and PC, $\{\text{WS}, \text{PC}\}$, are $\{2.1, 6.9\}$, $\{1.7, 8.2\}$, $\{1.6, 11.3\}$, $\{1.6, 14.8\}$, respectively. The norm loss of the WF-SOFAR stabilizes while that of the PC fast rises. However, when T grows with given N, the accuracy of both the WF-SOFAR and the PC factor loadings estimates improves (but the former is always more accurate than the latter). For example, when $(\alpha_1, \alpha_2) = (0.9, 0.5)$, for $T = 100, 200, 500, 1000$ with $N = 500$, the average squared norm losses (scaled by 100) of the WF-SOFAR and PC factor loadings estimates, $\{\text{WS}, \text{PC}\}$, are $\{10.3, 63.2\}$, $\{4.5, 29.5\}$, $\{1.6, 11.3\}$, $\{0.8, 5.7\}$, respectively.

Consequently, the accuracy of the WF-SOFAR estimator of common component is uniformly superior to that of the PC estimator. For example, when $(\alpha_1, \alpha_2) = (0.9, 0.5)$, $T = N = 100, 200, 500, 1000$, the average squared norm losses (scaled by 100) of the WF-SOFAR and PC factor component estimates, $\{\text{WS}, \text{PC}\}$, are $\{19.5, 47.6\}$, $\{11.5, 30.0\}$, $\{6.2, 16.5\}$, $\{3.9, 10.5\}$, respectively.

Table 3 reports the same information as Table 2, but for more challenging models with $(\alpha_1, \alpha_2) = (0.8, 0.4)$, $(0.5, 0.5)$ and $(0.5, 0.4)$. As the WF-SOFAR estimation naturally requires a larger sample size for these cases, we consider the combinations for $N, T = 500, 1000$ only. As can be seen in the table, remarkably, even when one of the exponent is 0.4, our WF-SOFAR method provides sufficiently accurate estimates of α_1 and α_2 as well as far superior estimates of factors, factor loadings and common components to the PC method.

To summarize, the WF-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 1/2 or below, the WF-SOFAR estimator is sufficiently precise in terms of norm

loss but requires a larger sample size. Significantly, even when the gap between the largest exponent and the smallest exponent is larger than the condition (10) implies (the latter must be greater than half of the former to ensure the divergence of the common components though), the WF-SOFAR estimator is sufficiently accurate in terms of norm loss 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 WF-SOFAR estimator in terms of norm loss.

6 Empirical Applications

In this section we provide two empirical applications. In the first subsection, the WF-SOFAR is applied to firm security returns to analyse changes in the presence of systematic risks in the market over the decades. In the second subsection we compare the forecasting performance of predictive regressions based on the factors extracted by the WF-SOFAR and WF-PC.

6.1 Firm security returns

In this subsection we apply our method to estimate approximate factor models using excess returns of firm securities which are used to compute the Standard & Poor's 500 (S&P 500) index of large cap U.S. equities market. In particular, we obtain the 500 securities that constitute the S&P 500 index each month over the period from January 1984 to April 2018 from Datastream. The monthly return of security i for month t is computed as $r_{ti} = 100 \times (P_{ti} - P_{t-1,i})/P_{t-1,i} + DY_{ti}/12$, where P_{ti} is the end-of-the-month price of the security and DY_{ti} is the per cent per annum dividend yield on the security. The one-month US treasury bill rate is chosen as the risk-free rate (r_{ft}), which is obtained from Ken French's data library web page. The excess return is defined as $r_{e,ti} = r_{ti} - r_{ft}$.

Following the literature, we estimate the factor model for the standardized excess return, $r^*_{e,ti}$. In view of the experimental results shown earlier, we report the results based on the adaptive WF-SOFAR with BIC. For each window month, $\mathtt{T}$ = Sept 1998,...,April 2018, we chose securities that contain the data extending 120 months back ($T = 120$) from $\mathtt{T}$. This gives the different number of securities for each window $\mathtt{T}$ ($N_{\mathtt{T}}$). The average number of securities over the estimation windows is 443 ($\bar{N} = 443$). In this exercise, we set the maximum number of factors as four. 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 $\mathtt{T}$.[11]

We report the estimates of the exponents of the eigenvalues of the security return covariance matrix, α_ℓ, $\ell = 1, 2, 3, 4$, 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. Table 4 reports the summary statistics of estimated α_ℓ's and the portion of non-zero factors, $N_{\ell\mathtt{T}}/N_{\mathtt{T}}$, $\ell = 1, 2, 3, 4$, and Figure 1 plots the estimated α_ℓ, $\ell = 1, 2, 3, 4$ over the estimation window months, $\mathtt{T}$ = September 1989,...,April 2018.

In turn we discuss the trajectories of the estimated exponents in some details by referring to Table 4 and Figure 1. The first factor does seem to be almost always "strong," in that

[11] 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 $\max_{1\le k\le r} |\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{k\mathtt{T}-1})| = |\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{2\mathtt{T}-1})|$ 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 $\widehat{\mathbf{b}}_{i2\mathtt{T}} \equiv -\widehat{\mathbf{b}}_{i\ell\mathtt{T}}$.

Table 2: Performance of the WF-SOFAR (WS) and PC estimators for approximate factor models with two factor components with $\alpha_1 \geq 0.8$ and $\alpha_2 \geq 0.5$

	T=100										**T=200**									
Design (α_1, α_2)	$(0.9, 0.9)$		$(0.9, 0.8)$		$(0.9, 0.5)$		$(0.8, 0.8)$		$(0.8, 0.5)$		$(0.9, 0.9)$		$(0.9, 0.8)$		$(0.9, 0.5)$		$(0.8, 0.8)$		$(0.8, 0.5)$	
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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.86	0.02	0.84	0.03	0.85	0.02	0.76	0.03	0.75	0.03	0.87	0.01	0.86	0.03	0.86	0.02	0.77	0.02	0.76	0.02
$\hat{\alpha}_2$	0.85	0.02	0.78	0.04	0.54	0.07	0.75	0.03	0.52	0.07	0.86	0.02	0.79	0.03	0.53	0.06	0.77	0.02	0.52	0.06
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	6.2	11.6	6.3	11.9	13.0	21.0	8.2	13.4	13.8	21.8	5.1	7.8	5.3	8.1	12.4	16.3	7.1	9.6	13.1	17.0
$L^2(\widehat{\mathbf{B}})_{\times 100}$	9.0	9.9	8.9	12.0	9.7	31.7	9.9	15.4	10.4	38.2	4.7	5.5	4.7	6.7	4.6	16.1	5.3	8.3	4.8	19.2
$L^2(\widehat{\mathbf{C}})_{\times 100}$	8.2	14.5	10.9	18.4	19.5	47.6	10.5	19.8	20.9	50.6	5.6	8.7	7.7	11.5	15.5	29.4	7.6	12.3	16.5	31.2
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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.86	0.01	0.85	0.02	0.85	0.01	0.77	0.02	0.75	0.02	0.87	0.01	0.87	0.02	0.87	0.01	0.77	0.01	0.76	0.01
$\hat{\alpha}_2$	0.86	0.01	0.77	0.03	0.54	0.05	0.76	0.02	0.52	0.05	0.87	0.01	0.78	0.02	0.53	0.04	0.77	0.02	0.51	0.04
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	4.6	10.1	4.2	9.9	9.9	19.0	5.9	11.5	10.4	19.5	3.5	6.4	3.2	6.1	8.6	13.0	4.8	7.7	9.2	13.6
$L^2(\widehat{\mathbf{B}})_{\times 100}$	9.1	10.4	8.6	12.4	9.8	41.9	10.2	17.4	10.0	50.0	4.7	5.7	4.0	6.3	4.4	20.3	5.3	9.3	4.5	24.2
$L^2(\widehat{\mathbf{C}})_{\times 100}$	6.8	13.1	8.7	16.8	15.9	54.1	8.6	18.7	16.4	56.8	4.1	7.5	5.4	9.7	11.5	29.9	5.6	11.0	12.1	31.6
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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.87	0.01	0.86	0.02	0.86	0.01	0.77	0.01	0.75	0.01	0.88	0.01	0.87	0.01	0.87	0.01	0.78	0.01	0.77	0.01
$\hat{\alpha}_2$	0.86	0.01	0.77	0.02	0.55	0.04	0.76	0.01	0.52	0.04	0.87	0.01	0.78	0.01	0.53	0.03	0.77	0.01	0.51	0.03
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	3.5	9.3	3.0	8.6	6.8	18.5	4.3	10.1	7.0	18.8	2.3	5.6	1.9	4.9	5.7	10.8	3.3	6.4	6.0	11.2
$L^2(\widehat{\mathbf{B}})_{\times 100}$	9.4	11.2	9.0	13.6	10.3	63.2	10.6	20.3	10.8	74.8	4.5	6.0	3.9	6.6	4.5	29.5	5.4	10.7	4.6	35.0
$L^2(\widehat{\mathbf{C}})_{\times 100}$	6.1	12.7	7.5	16.6	13.1	72.9	7.5	19.1	13.4	76.0	3.3	6.9	3.9	8.8	8.5	35.9	4.2	10.3	8.9	37.6
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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.87	0.01	0.86	0.01	0.86	0.01	0.77	0.01	0.76	0.01	0.88	0.00	0.87	0.00	0.87	0.00	0.78	0.01	0.77	0.01
$\hat{\alpha}_2$	0.86	0.01	0.77	0.01	0.56	0.03	0.76	0.01	0.53	0.03	0.87	0.00	0.78	0.01	0.54	0.03	0.77	0.01	0.51	0.03
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	2.8	9.0	2.2	7.8	5.1	20.0	3.4	9.4	5.2	20.1	1.9	5.4	1.5	4.5	4.2	10.4	2.4	5.8	4.3	10.6
$L^2(\widehat{\mathbf{B}})_{\times 100}$	9.4	12.0	8.8	14.4	10.9	86.7	11.0	23.1	11.5	101.8	4.7	6.5	4.0	7.3	4.6	39.6	5.3	12.0	4.8	46.8
$L^2(\widehat{\mathbf{C}})_{\times 100}$	6.0	12.7	6.7	16.6	12.2	96.0	7.2	20.1	12.3	99.6	3.0	6.8	3.4	8.8	7.0	44.4	3.8	10.7	7.2	46.3

Notes: The DGP is $x_{ti} = \sum_{k=1}^{r} b_{ik}^0 f_{tk}^0 + \sqrt{\theta} e_{ti}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $b_{ik}^* \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, \lfloor N^{\alpha_k} \rfloor = N_k$ and $b_{ik}^* = 0$ for $i = N_k + 1, \ldots, N$, with $0 < \alpha_k \leq 1$, $f_{tk}^* = \rho_{fk} f_{k,t-1}^* + v_{tk}$ for $t = 1, ..., T$, $v_{tk} \sim$ i.i.d.$N(0, 1 - \rho_{fk}^2)$ with $f_{0k}^* \sim$ i.i.d.$N(0,1)$ for $k = 1, \ldots, r$, $e_{ti} = \rho_e e_{t-1,i} + \beta \varepsilon_{t,i-1} + \beta \varepsilon_{t,i+1} + \varepsilon_{ti}$, $\varepsilon_{ti} \sim$ i.i.d.$N(0, \sigma_{\varepsilon,ti}^2)$, $\sigma_{\varepsilon,ti}^2$ is set so that $\text{Var}(e_{ti}) = 1$. Once b_{ik}^* and f_{tk}^* are generated, we apply Gram–Schmidt orthonormalization and form b_{ik}^0 and f_{tk}^0, where $N^{-1} \sum_{i=1}^{N} b_{ik}^0 b_{i\ell}^0 = 1\{k = \ell\}$ and $T^{-1} \sum_{t=1}^{T} f_{tk}^0 f_{t\ell}^0 = 1\{k = \ell\}$. We set $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for $k = 1, \ldots, r$, $\beta = 0.2$, $\theta = 1$. The rows $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in the table report the averages and standard deviations (s.d.) of the WF-SOFAR estimates of α_1 and α_2, and the rows $L^2(\widehat{\mathbf{F}})_{\times 100}$, $L^2(\widehat{\mathbf{B}})_{\times 100}$ and $L^2(\widehat{\mathbf{C}})_{\times 100}$ report the averages of the norm losses (multiplied by 100) of the scaled estimated factors, factor loadings and common components by the WF-SOFAR and PC over the replications. All the results are based on 1000 replications.

Table 2 (Continued):

	T=500										**T=1000**									
Design (α_1, α_2)	$(0.9, 0.9)$		$(0.9, 0.8)$		$(0.9, 0.5)$		$(0.8, 0.8)$		$(0.8, 0.5)$		$(0.9, 0.9)$		$(0.9, 0.8)$		$(0.9, 0.5)$		$(0.8, 0.8)$		$(0.8, 0.5)$	
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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.88	0.01	0.87	0.02	0.87	0.01	0.78	0.02	0.78	0.02	0.89	0.01	0.88	0.02	0.88	0.01	0.79	0.01	0.78	0.01
$\hat{\alpha}_2$	0.88	0.01	0.79	0.03	0.52	0.05	0.78	0.02	0.51	0.05	0.88	0.01	0.80	0.03	0.52	0.04	0.78	0.01	0.51	0.04
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	4.2	5.3	4.5	5.7	11.9	13.6	6.1	7.2	12.8	14.4	3.9	4.5	4.3	4.9	11.5	12.4	5.7	6.2	12.3	13.1
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.2	2.6	2.1	3.0	2.1	6.9	2.4	3.8	2.1	8.2	1.4	1.6	1.3	1.8	1.2	3.7	1.5	2.2	1.2	4.4
$L^2(\widehat{\mathbf{C}})_{\times 100}$	4.1	5.5	5.9	7.5	13.4	19.2	6.1	8.0	14.4	20.5	3.6	4.3	5.3	6.1	12.7	15.6	5.5	6.5	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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.88	0.01	0.88	0.01	0.88	0.01	0.78	0.01	0.78	0.01	0.89	0.01	0.89	0.01	0.88	0.01	0.79	0.01	0.78	0.01
$\hat{\alpha}_2$	0.88	0.01	0.79	0.02	0.52	0.04	0.78	0.01	0.50	0.03	0.89	0.01	0.79	0.01	0.51	0.03	0.78	0.01	0.50	0.03
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	2.8	4.1	2.6	3.8	8.2	9.9	4.0	5.1	8.8	10.5	2.5	3.1	2.4	3.0	8.1	8.9	3.7	4.3	8.7	9.5
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.2	2.8	1.6	2.6	1.7	8.2	2.3	4.1	1.8	9.7	1.4	1.6	0.7	1.3	1.0	4.2	1.5	2.4	1.0	5.0
$L^2(\widehat{\mathbf{C}})_{\times 100}$	2.6	4.0	3.5	5.3	9.5	16.8	3.9	6.1	10.1	17.8	2.1	2.9	2.9	3.8	8.9	12.5	3.4	4.5	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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.88	0.00	0.88	0.00	0.88	0.00	0.78	0.01	0.78	0.01	0.89	0.00	0.89	0.00	0.89	0.00	0.79	0.01	0.78	0.01
$\hat{\alpha}_2$	0.88	0.00	0.79	0.01	0.52	0.03	0.78	0.01	0.51	0.02	0.89	0.00	0.79	0.01	0.51	0.02	0.79	0.01	0.50	0.02
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	1.8	3.2	1.4	2.5	5.2	7.0	2.5	3.9	5.5	7.3	1.5	2.2	1.2	1.8	5.0	5.9	2.2	2.9	5.3	6.2
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.2	3.0	1.4	2.6	1.6	11.3	2.4	4.9	1.6	13.5	1.3	1.7	0.7	1.3	0.8	5.7	1.5	2.7	0.8	6.7
$L^2(\widehat{\mathbf{C}})_{\times 100}$	1.7	3.2	2.0	4.1	6.2	16.5	2.4	4.9	6.5	17.4	1.2	2.0	1.5	2.6	5.6	10.7	1.9	3.2	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.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.89	0.00	0.88	0.00	0.88	0.00	0.79	0.00	0.78	0.00	0.89	0.00	0.89	0.00	0.89	0.00	0.79	0.00	0.79	0.00
$\hat{\alpha}_2$	0.88	0.00	0.79	0.01	0.53	0.02	0.78	0.00	0.51	0.02	0.89	0.00	0.79	0.00	0.52	0.02	0.79	0.00	0.51	0.02
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	1.4	2.9	1.0	2.1	3.6	5.5	1.9	3.3	3.8	5.7	1.2	2.0	0.7	1.3	3.4	4.3	1.5	2.2	3.6	4.5
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.1	3.1	1.4	2.9	1.6	14.8	2.4	5.3	1.7	17.5	1.3	1.9	0.6	1.5	0.7	7.3	1.3	2.9	0.8	8.6
$L^2(\widehat{\mathbf{C}})_{\times 100}$	1.4	2.9	1.6	3.8	4.6	18.1	1.9	4.7	4.8	19.0	0.9	1.7	1.1	2.2	3.9	10.5	1.3	2.7	4.1	11.0

Table 3: Performance of the WF-SOFAR (WS) and PC estimators for approximate factor models with two factor components with $(\alpha_1, \alpha_2) = (0.8, 0.4)$, $(0.5, 0.5)$, $(0.5, 0.4)$

	T=500						**T=1000**					
Design (α_1, α_2)	$(0.8, 0.4)$		$(0.5, 0.5)$		$(0.5, 0.4)$		$(0.8, 0.4)$		$(0.5, 0.5)$		$(0.5, 0.4)$	
N=500	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.78	0.01	0.48	0.02	0.47	0.03	0.78	0.01	0.48	0.02	0.47	0.03
$\hat{\alpha}_2$	0.44	0.04	0.48	0.02	0.41	0.04	0.43	0.03	0.48	0.02	0.40	0.04
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	9.3	13.4	10.5	12.9	13.4	17.9	9.2	11.1	10.0	11.0	13.1	15.2
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.0	25.9	3.2	30.6	4.6	48.3	1.1	12.8	1.9	15.4	2.9	24.4
$L^2(\widehat{\mathbf{C}})_{\times 100}$	10.9	34.3	10.6	30.0	17.3	48.6	10.0	21.3	10.0	19.4	16.0	31.1
N=1000	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.78	0.00	0.49	0.01	0.48	0.02	0.79	0.00	0.49	0.01	0.48	0.02
$\hat{\alpha}_2$	0.44	0.03	0.48	0.02	0.40	0.03	0.43	0.03	0.49	0.01	0.40	0.03
	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC	WS	PC
$L^2(\widehat{\mathbf{F}})_{\times 100}$	6.8	11.7	7.5	10.1	9.7	15.2	6.7	8.9	7.2	8.3	9.5	12.0
$L^2(\widehat{\mathbf{B}})_{\times 100}$	2.0	36.1	3.1	40.4	3.7	65.6	1.0	17.4	1.8	20.0	2.3	32.2
$L^2(\widehat{\mathbf{C}})_{\times 100}$	8.3	41.5	7.8	33.3	13.0	57.4	7.5	23.1	7.0	19.4	12.0	32.9

Notes: See the notes to Table 2.

the divergence rate N_1 is very close to N. As reported in Table 4, 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. Actually, the values of the factor loadings to this factor have the same sign, which strongly suggests that this is the market factor. Apart from the first factor, which is always strong, the strengths of the common components vary over the months and can become quite weak. For example, for the second to the fourth factors, the maximum portion of nonzero factor loadings is between 44.5% and 59.5%, whilst the minimum portion is merely 12.0% to 17.6%. Furthermore, the divergence rates are very different over the factors. For example, for the window month of March 1998, $\{\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3\} = \{0.991, 0.774, 0.653\}$, and the corresponding numbers of non-zero factors are 425, 113 and 54 out of 450 securities. These strongly imply a potentially substantial efficiency gain in estimation of the approximate factor models through our WF-SOFAR method over the PC method.

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 D). 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 D).

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 particular, from September 1989, α_2 is larger than α_3 most of the time until December 2010, then α_3 is almost always larger than α_2. Since the estimate of α_4 first appeared in February 2004, it was mostly smaller than other exponents. It is estimated every month from March 2010 onward, seemingly becoming more and more strong toward the latest month, April 2018. After the sharp one-off drop in February 2015,[12] α_4 rises to become the highest next to the first factor from November 2016 onward.

Table 4: Summary statistics of the estimated exponents of factor loadings, $\alpha_{\ell\mathtt{T}}$, and the portion of non-zero factor loadings, $N_{\ell\mathtt{T}}/N_{\mathtt{T}}$, $\ell = 1, 2, 3, 4$, from September 1989 to April 2018.

	Exponents of Loadings				Portion of Non-zero Loadings			
	α_1	α_2	α_3	α_4	N_1/N	N_2/N	N_3/N	N_4/N
mean	0.995	0.824	0.770	0.781	97.1%	36.2%	26.2%	27.3%
s.d.	0.004	0.046	0.045	0.028	2.4%	9.6%	7.0%	5.0%
max	1.000	0.895	0.860	0.854	100.0%	59.5%	44.5%	49.7%
min	0.979	0.713	0.653	0.665	88.2%	17.6%	12.0%	13.1%

Notes: The estimated α_ℓ, $\ell = 1, 2, 3, 4$ for the each month of 120 months window, $\mathtt{T}$ = September 1989,...,April 2018.

6.2 Forecasting bond yields

In this subsection we consider out-of-sample performance of forecasting regressions for bond yields using extracted factors via our WF-SOFAR method and the PC method, from a large

[12]This coincides with the period at bottom of the biggest sharp fall of oil price between 2014-2015.

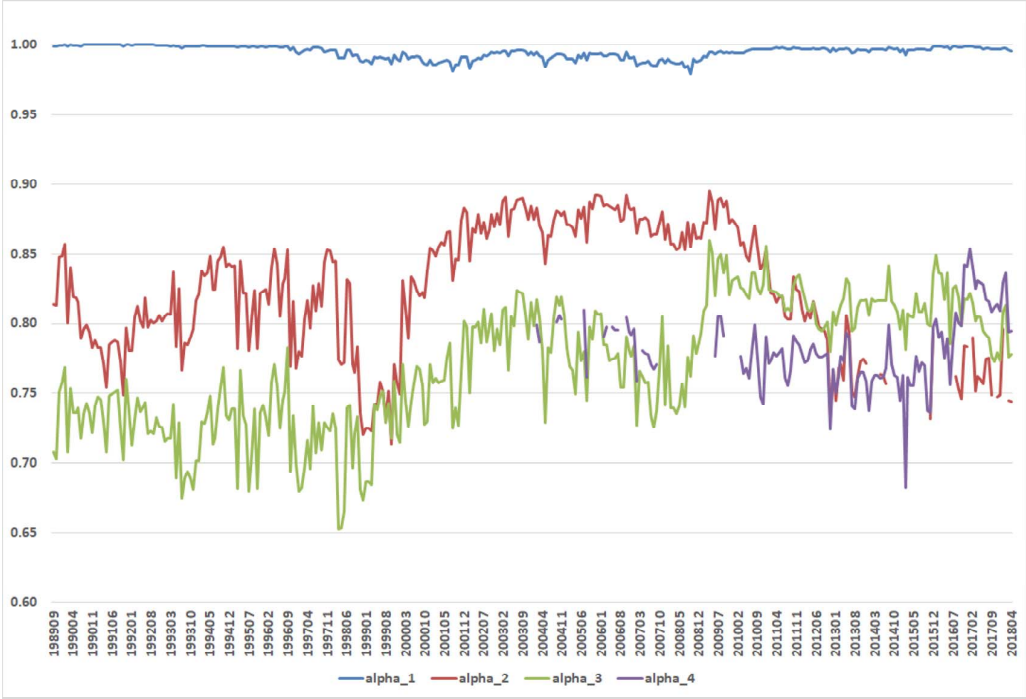

Figure 1: Plot of the estimated α_k's from September 1989 to April 2018. The estimated α_ℓ, $\ell = 1, 2, 3, 4$ for the each month of 120 months window, $\mathtt{T}$ = September 1989,...,April 2018.

number of macroeconomic (prediction) variables in line with the analysis of Ludvigson and Ng (2009). We use the same data set which is provided by Ludvigson and Ng.[13] 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 $i = 1, \ldots, 132$ monthly macroeconomic series at month t, x_{ti}, spanning the period from January 1964 to December 2003. We consider the maturities $n = 2, 3, 4, 5$.[14]

We consider one-year-ahead out of sample forecast comparisons of continuously compounded annual log excess bond returns, $y_{t+12}^{(n)}$. 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 x_{ti} via our WF-SOFAR and the PC, $i = 1, \ldots, N = 132$, $t = \mathtt{T}, \ldots, T_\mathtt{T} - 12$, 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 the number of factors is estimated for each estimation window to avoid using "future" information.[15] Then, run the predictive regression

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

[13]The data file is obtained from Sydney Ludvigson's web page: https://www.sydneyludvigson.com/s/RFS2009-u1e1.xls

[14]For more details of the data, see Section 3 of Ludvigson and Ng (2009).

[15]In another experiment, we estimated the number of factors using a whole sample period and implemented a similar exercise. The forecast based on our estimator uniformly outperformed that based on the PC estimator.

and obtain the forecast error

$$\hat{\varepsilon}^{(n)}_{T_{\mathrm{T}}+12|T_{\mathrm{T}}} = y^{(n)}_{T_{\mathrm{T}}+12} - \hat{y}^{(n)}_{T_{\mathrm{T}}+12|T_{\mathrm{T}}},$$

with $\hat{y}^{(n)}_{T_{\mathrm{T}}+12|T_{\mathrm{T}}} = \tilde{\beta}^{(n)}_0 + \sum_{k=1}^{\hat{r}_{\mathrm{T}}} \tilde{\beta}^{(n)}_k \hat{f}_{T_{\mathrm{T}}k}$. This produces $H = 217$ forecast errors. To estimate the number of factors for each window, we set the maximum number of factors to nine and use the ED estimator. The estimated number of factors varies from one to six over the forecast windows. In Table 1, we report the mean absolute deviation of the forecast errors, $MAE^{(n)} = H^{-1} \sum_{s=1}^{H} \left| \hat{\epsilon}^{(n)}_{s|s-1} \right|$ for $n = 2, 3, 4, 5$, for our WF-SOFAR and the PC, and Diebold-Mariano forecasting performance test statistics with associated p-values, based on the MAEs.[16] As can be seen, the MAEs of the WF-SF are smaller than those of the PC for all the maturities. The Diebold-Mariano forecasting performance test strongly rejects the null of the same forecasting performance for all the maturities, in favour of the alternative that our method ourperforms the PC method. The average values of exponents over the windows are $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6\} = \{0.92, 0.82, 0.87, 0.78, 0.77, 0.74\}$, which suggests that even the (first) strongest factor component is not strictly strong. 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 5: Mean absolute forecast errors and Diebold-Mariano forecast comparison test result

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

Notes: For the computation of the long-run variance for the Diebold-Mariano test statistic of Diebold and Mariano (1995), the window is chosen by the Schwert criterion with the maximum lag of 14.

7 Conclusion

This paper considers the determination of number of factors and efficient estimation of the large dimensional approximate factor models of Chamberlain and Rothschild (1983)), called weak factor (WF) models, in which the r largest eigenvalues of a data covariance matrix grow as N rises with possibly different diverging rates N^{α_k}, $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. We consider the WF structure induced by *sparse* factor loadings $\mathbf{B}^0$ that leads to $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N^{\alpha_k}$. The proposed *weak factor sparse orthogonal factor regression* (WF-SOFAR) estimator and its adaptive version enable us to consistently estimate the WF models. As theoretical contributions, the non-asymptotic error bound and the rate of convergence for the WF-SOFAR estimators are derived. For the adaptive WF-SOFAR estimator, we have established the factor selection consistency and consistent estimation of each exponent α_k of the r divergence rates. The proposed estimator is expected to be more efficient than

[16] We computed the Diebold-Mariano test statistic based on mean squared errors, and the null was more strongly rejected.

the widely used principal component (PC) estimator. In line with our theoretical results, the Monte Carlo experiments show that the ED estimator of Onatski (2010) for the number of factors r is reliable while other widely used estimators such as Bai and Ng (2002) and Ahn and Horenstein (2013) can be unreliable. Furthermore, the evidence demonstrates that the WF-SOFAR estimator uniformly dominates the PC estimator in terms of the norm loss, including under the experimental design most favorable to the PC estimator, in which the exponents of the divergence rates are 0.9.

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, Kim, Kim, and Kwark (Choi et al.) consider factor models with two types of factors, global factors and local factors. The factor loadings of global factors are non-zero values for all the cross-section units, whereas the local factors have non-zero loadings among the cross-section units of specific cross sectional groups. Ando and Bai (2017) and Choi, Kim, Kim, and Kwark (Choi et al.) propose sequential procedures to identify the global and local factors separately. In fact, the WF structure nests the hierarchical factor structure and hence our WF-SOFAR method can be applied to readily estimate such models. In contrast to existing approaches, given the total number of global and local factors, our approach permits us to consistently estimate the number of local groups, the number of global and local factors and its memberships in one go. For further information and additional simulation results, see Appendix C.

Having provided the consistency result of the WF-SOFAR estimator in this paper, the statistical inference for the high-dimensional WF models is an important research agenda. The application of the method named *debiasing* (desparsification) by Javanmard and Montanari (2014) and van de Geer et al. (2014) to the WF-SOFAR estimator seems the most promising pathway. This methodology will correct the Lasso estimator to remove the bias based on the Karush-Kuhn-Tucker (KKT) conditions. However, it is nontrivial to establish how the WF-SOFAR estimator can be debiased, and this is our future challenge.

In this paper we have focused on the estimation of the common factors and the exponents of the divergence rates of the r largest eigenvalues. It is of interest to estimate the stock return 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.

Another possible extension of interest is to consider the estimation of panel data models with unobservable multiple interactive effects. Pesaran (2006) and Bai (2009), among others, develop the estimation methods of the panel data model:

$$y_{ti} = \mathbf{x}_{ti}'\boldsymbol{\beta} + u_{ti}, \qquad u_{ti} = \mathbf{f}_t'\mathbf{b}_i + \varepsilon_{ti}.$$

For the PC based estimators, such as Bai (2009), u_{it} is typically assumed to have the strong factor structure (i.e., $\sum_{i=1}^{N} \mathbf{b}_i\mathbf{b}_i'/N$ tends to a fixed matrix), which may not hold in practice, and the WF structure seems more appropriate. The iterative procedure proposed by Bai (2009) based on the WF-SOFAR estimation of $\mathbf{f}_t'\mathbf{b}_i$, instead of the PC estimation, 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.

Amengual, D. and M. Watson (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business & Economic Statistics 25*, 91–96.

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.

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 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.

Bai, J. and S. Ng (2017). Principal components and regularized estimation of factor models. *arXiv:1708.08137v2*.

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

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

Candès, E. J., X. Li, Y. Ma, and J. Wright (2011). Robust principal component analysis? *Journal of the ACM 58*, 11:1–11:37.

Caner, M. and X. Han (2014). Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators. *Journal of Business & Economic Statistics 32*, 359–374.

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. A multilevel factor model: Identification, asymptotic theory and applications. *Journal of Applied Econometrics 33*, 355–377.

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.

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.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*, 1–22.

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

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. and J. Lv (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory 57*, 5467–5484.

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

Freyaldenhoven, S. (2018). A generalized factor model with local factors. *mimeo*.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*, 603–617.

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

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

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business & Economic Statistics 28*, 397–409.

Kock, A. B. and H. Tang (2019). Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory 35*, 295–359.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*, 21–59.

Leeb, H. and B. M. Pötscher (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics 34*, 2554–2591.

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 (2018). Estimating latent asset-pricing factors. *NBER Working Paper 24618*.

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.

Pötscher, B. M. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis 100*, 2065–2082.

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.

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

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, to appear*.

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

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 42*, 1166–1202.

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.

Zhang, C.-H. (2014). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics 38*, 894–942.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B 76*, 217–242.

Appendix

A Proofs of the Main Results

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 $\widetilde{\mathbf{E}}_\ell = (\boldsymbol{\varepsilon}_{1-\ell}, \ldots, \boldsymbol{\varepsilon}_{T-\ell})' \in \mathbb{R}^{T\times N}$. 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

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

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

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

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 $1-\exp\left(-|O(N\vee T)|\right)$.

(b) By the definition, the (t,k)th element of $\mathbf{EB}^0$ 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

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

Since $\{\varepsilon_{t-\ell,i}\tilde{b}_{\ell k,i}\}_{i=1}^N$ is a sequence of i.i.d. 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

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

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

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

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

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

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

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

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

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

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 $\mathbf{M} \in \mathbb{R}^{m\times m}$,

$$\mathbb{P}\left(\left|\boldsymbol{\xi}'\mathbf{M}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{M}\boldsymbol{\xi}\right| > u\right) \le 2\exp\left\{-c\min\left(\frac{u^2}{K^4\|\mathbf{M}\|_{\mathrm{F}}^2}, \frac{u}{K^2\|\mathbf{M}\|_2}\right)\right\}, \tag{A.1}$$

where c and K are positive constants such that $\sup_{k\ge1} k^{-1/2}(\mathbb{E}\,|\varepsilon|^k)^{1/k} \le K$. In our setting, we can take $K = 3\sigma_\varepsilon^2$ (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

$$\begin{aligned}\max_i\left|T^{-1}\sum_{t=1}^T\left(e_{ti}^2 - \mathbb{E}\,e_{ti}^2\right)\right| &= \max_i\left|T^{-1}\sum_{t=1}^T\sum_{\ell=0}^{L_n}\left(\boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}_{\ell,i}'\boldsymbol{\varepsilon}_{t-\ell} - \mathbb{E}\,\boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\phi}_{\ell,i}\boldsymbol{\phi}_{\ell,i}'\boldsymbol{\varepsilon}_{t-\ell}\right)\right| \\ &\le T^{-1}\sum_{\ell=0}^{L_n}\max_i\left|\tilde{\boldsymbol{\varepsilon}}_\ell'\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell - \mathbb{E}\,\tilde{\boldsymbol{\varepsilon}}_\ell'\mathbf{A}_{\ell i}\tilde{\boldsymbol{\varepsilon}}_\ell\right|,\end{aligned}$$

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.1) with the union bound gives

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

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

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

with probability at least

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

This completes all the proofs. □

Proof of Theorem 1. 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 $\mathbf{V}^{0\prime}\mathbf{V}^0 = \mathbf{N} = \mathrm{diag}(N_1,\ldots,N_r)$ by the definition. We also define $\mathbf{W}_{1:k}$ to be 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 $\lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) = d_k^2N_kT$, it is sufficient to show that for any $k \in \{1,\ldots,r\}$,

$$\lambda_k(\mathbf{XX}') = \lambda_k(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + O(N\vee T), \tag{A.2}$$

$$\lambda_k(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) = \lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) + O\left(TN_1^{1/2}\log^{1/2}(N\vee T) + N\vee T\right). \tag{A.3}$$

Then (A.2) and (A.3) lead to the desired result,

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

We show (A.2). Lemma A.5 of Ahn and Horenstein (2013) yields the upper bound

$$\begin{aligned}\sum_{j=1}^k \lambda_j(\mathbf{XX}') &= \sum_{j=1}^k \lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime} + \mathbf{EQE}')\\ &\le \sum_{j=1}^k \lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + k\lambda_1(\mathbf{EQE}')\\ &\le \sum_{j=1}^k \lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + k\lambda_1(\mathbf{EQE}' + \mathbf{EPE}')\\ &= \sum_{j=1}^k \lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) + k\lambda_1(\mathbf{EE}').\end{aligned}$$

Moreover, the lower bound is given by

$$\begin{aligned}\sum_{j=1}^k \lambda_j(\mathbf{XX}') &\ge T^{-1}\,\mathrm{tr}(\mathbf{W}_{1:k}'\mathbf{XX}'\mathbf{W}_{1:k})\\ &= T^{-1}\,\mathrm{tr}(\mathbf{W}_{1:k}'\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}\mathbf{W}_{1:k}) + T^{-1}\,\mathrm{tr}(\mathbf{W}_{1:k}'\mathbf{EQE}'\mathbf{W}_{1:k})\\ &\ge \sum_{j=1}^k \lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}).\end{aligned}$$

Since $\lambda_1(\mathbf{EE}') = \|\mathbf{E}\|_2^2 \lesssim T\vee N$ with probability at least $1 - O((N\vee T)^{-\nu})$ by Lemma 1(a), these two inequalities imply (A.2).

Next, we verify (A.3). 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 $1-O((N\vee T)^{-\nu})$. Consider the lower bound. We have

$$\begin{aligned}\sum_{j=1}^{k}\lambda_j(\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}) &\geq T^{-1}\operatorname{tr}(\mathbf{U}^0_{1:k}\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}\mathbf{U}^0_{1:k}) \\ &\gtrsim \sum_{j=1}^{k}\lambda_j(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) - TN_1^{1/2}\log^{1/2}(N\vee T).\end{aligned}$$

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

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

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

Proof of Theorem 2. The optimality of the WF-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 (3) and letting $\boldsymbol{\Delta}=\widehat{\mathbf{F}}\widehat{\mathbf{B}}'-\mathbf{F}^0\mathbf{B}^{0\prime}$, this inequality 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

$$\begin{aligned}(1/2)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 &\leq \operatorname{tr}\mathbf{E}\boldsymbol{\Delta}'+\eta_n\left(\|\mathbf{B}^0\|_1-\|\widehat{\mathbf{B}}\|_1\right) \\ &\leq \left|\operatorname{tr}\mathbf{E}\mathbf{B}^0\boldsymbol{\Delta}^{f\prime}\right|+\left|\operatorname{tr}\mathbf{E}\boldsymbol{\Delta}^b\boldsymbol{\Delta}^{f\prime}\right|+\left|\operatorname{tr}\boldsymbol{\Delta}^b\mathbf{F}^{0\prime}\mathbf{E}\right|+\eta_n\left(\|\mathbf{B}^0\|_1-\|\widehat{\mathbf{B}}\|_1\right). \qquad \text{(A.4)}\end{aligned}$$

We bound the traces in (A.4). 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.4) 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.4) becomes

$$\begin{aligned}(1/2)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leq{}& (rT)^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + r^{1/2}\|\mathbf{E}\boldsymbol{\Delta}^b\|_2\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ &+ \|\boldsymbol{\Delta}^b\|_1\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} + \eta_n\left(\|\mathbf{B}^0\|_1 - \|\widehat{\mathbf{B}}\|_1\right). \qquad \text{(A.5)}\end{aligned}$$

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.5) is further bounded as

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \lesssim{}& (N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + (\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ &+ \eta_n\left(\|\boldsymbol{\Delta}^b\|_1 + 2\|\mathbf{B}^0\|_1 - 2\|\widehat{\mathbf{B}}\|_1\right). \qquad \text{(A.6)}\end{aligned}$$

We then focus on the last parenthesis of (A.6). 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.6) is rewritten and bounded as

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

Therefore, the upper bound of (A.6) is given by

$$\begin{aligned}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \lesssim{}& (N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} \\ &+ (\widetilde{N}\vee T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + N_1^{1/2}\eta_n\|\boldsymbol{\Delta}^b\|_{\mathrm{F}}. \qquad \text{(A.7)}\end{aligned}$$

Meanwhile, Lemma 5 establishes the lower bound of (A.7). Consequently, we obtain

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

where

$$\kappa_n = \frac{N_r(N_r \wedge T)}{N_1},$$
$$\alpha_n = (N_1 T)^{1/2} \log^{1/2}(N \vee T), \quad \mu_n = (\widetilde{N} \vee T)^{1/2} \log^{1/2}(N \vee T). \quad \beta_n = N_1^{1/2} \eta_n.$$

Then we have

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

The desired result is obtained by rearranging this inequality. In fact, we have

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

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

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

This completes the proof. □

Proof of Theorem 3. Following the proof of Theorem 2, we derive the bound. From (A.5) with putting $\eta_n = 0$, we have

$$\begin{aligned}&(1/2)\|\boldsymbol{\Delta}_{\mathsf{PC}}\|_{\mathrm{F}}^2 \\ &\quad\lesssim T^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + \|\mathbf{E}\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_2\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}} + N^{1/2}\|\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max}. \qquad \text{(A.8)}\end{aligned}$$

Lemmas 1 and 4 states that the event

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

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.8) becomes

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

where

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

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

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

Finally, we observe that

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

This completes the proof of Theorem 3. □

Proof of Theorem 4. 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 WF-SOFAR estimator is given by

$$Q_n(\mathbf{F}, \mathbf{B}) := \frac{1}{2} \left\| \mathbf{X} - \mathbf{F}\mathbf{B}' \right\|_{\mathrm{F}}^2 + \eta_n \|\mathbf{W} \circ \mathbf{B}\|_1 \tag{A.9}$$

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 $Q_n(\mathbf{F}, \mathbf{B}_{\mathcal{S}})$ (i.e., a minimizer of the correctly zero-restricted minimization problem), 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

$$\mathbb{P}\left(\inf_{\|\mathbf{U}\|_{\mathrm{F}} = C,\, \|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} = C} Q_n(\mathbf{F}^0 + r_n \mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}}) > Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \right) \to 1, \tag{A.10}$$

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

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

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.

To show (A.10), we have

$$\begin{aligned}
&Q_n(\mathbf{F}^0 + r_n \mathbf{U}, \mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}}) - Q_n(\mathbf{F}^0, \mathbf{B}^0_{\mathcal{S}}) \\
&\quad = 2^{-1} \|\mathbf{X} - (\mathbf{F}^0 + r_n \mathbf{U})(\mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}})'\|_{\mathrm{F}}^2 - 2^{-1} \|\mathbf{X} - \mathbf{F}^0 \mathbf{B}^{0\prime}_{\mathcal{S}}\|_{\mathrm{F}}^2 \\
&\qquad + \eta_n \|\mathbf{W} \circ (\mathbf{B}^0_{\mathcal{S}} + r_n \mathbf{V}_{\mathcal{S}})\|_1 - \eta_n \|\mathbf{W} \circ \mathbf{B}^0_{\mathcal{S}}\|_1 \\
&\quad \ge \mathrm{tr}\, \mathbf{E}' (r_n \mathbf{F}^0 \mathbf{V}'_{\mathcal{S}} + r_n \mathbf{U} \mathbf{B}^{0\prime}_{\mathcal{S}} + r_n^2 \mathbf{U} \mathbf{V}'_{\mathcal{S}}) \\
&\qquad + 2^{-1} \|r_n \mathbf{F}^0 \mathbf{V}'_{\mathcal{S}} + r_n \mathbf{U} \mathbf{B}^{0\prime}_{\mathcal{S}} + r_n^2 \mathbf{U} \mathbf{V}'_{\mathcal{S}}\|_{\mathrm{F}}^2 - r_n \eta_n \|\mathbf{W}_{\mathcal{S}} \circ \mathbf{V}_{\mathcal{S}}\|_1 \\
&\quad = (I) + (II) + (III).
\end{aligned} \tag{A.11}$$

By Lemma 7 (a)–(c), we bound (I) as

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

Next, we bound (II) from below as

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

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

$$\begin{aligned}
|(iii) + (iv) + (v)| \lesssim r_n^3 \Big(&\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2 \log^{1/2}(N \vee T) + N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} \Big) \\
&+ r_n^2 N_1^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} \log^{1/2}(N \vee T).
\end{aligned}$$

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

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

We then consider (III) in (A.11). Lemma 8 yields

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

Putting together the pieces obtained so far with (A.11), we have

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

By condition (8) and the fact that

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

we have

$$\begin{aligned}
&\inf_{\|\mathbf{U}\|_{\mathrm{F}}=C,\,\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}=C} Q_n(\mathbf{F}^0+r_n\mathbf{U},\mathbf{B}^0_{\mathcal{S}}+r_n\mathbf{V}_{\mathcal{S}})-Q_n(\mathbf{F}^0,\mathbf{B}^0_{\mathcal{S}})\\
&\quad\gtrsim r_n^2\left(N_r\vee T\right)C^2-r_n^3N_1^{1/2}C^3-r_nN_1^{1/2}(\eta_n/\underline{b}_n^0)C. \qquad\text{(A.12)}
\end{aligned}$$

Furthermore, in (A.12), the first term asymptotically dominates the second term because the ratio,

$$\frac{r_n^3N_1^{1/2}}{r_n^2(N_r\vee T)}=\frac{N_1^2}{N_r^2T^{1/2}},$$

converges to zero by condition (15). Finally, we compare the first term with the third term in (A.12). Observe that conditions (17) (first bound) and (18) give

$$\underline{b}_n^0\gtrsim\frac{N_1^{1/2}\eta_n}{r_n(N_r\vee T)}.$$

Therefore, the first term in (A.12) becomes strictly larger than the third term in absolute value as long as $C>0$ is taken to be large enough. This means that the lower bound tends to positive for such $C>0$ and (A.10) 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

$$\begin{aligned}
&\nabla_{\mathbf{F}}Q_n(\mathbf{F},\mathbf{B})=\mathbf{F}\mathbf{B}'\mathbf{B}-\mathbf{X}\mathbf{B},\\
&\nabla_{\mathbf{B}}Q_n(\mathbf{F},\mathbf{B})=\mathbf{B}\mathbf{F}'\mathbf{F}-\mathbf{X}'\mathbf{F}+\eta_n\mathbf{T},
\end{aligned}$$

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

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

Then $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}})$ is a strict minimizer of (6) if the following conditions hold:

$$\widehat{\mathbf{F}}\widehat{\mathbf{B}}'\widehat{\mathbf{B}} - \mathbf{X}\widehat{\mathbf{B}} = \mathbf{0}_{T\times r}, \tag{A.13}$$

$$T\widehat{\mathbf{B}}_{\mathcal{S}} - (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}} + \eta_n \mathbf{W}_{\mathcal{S}} \circ \mathrm{sgn}\,\widehat{\mathbf{B}}_{\mathcal{S}} = \mathbf{0}_{N\times r}, \tag{A.14}$$

$$\left\| \mathbf{W}^{-}_{\mathcal{S}^c} \circ \left\{ T\widehat{\mathbf{B}}_{\mathcal{S}^c} - (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c} \right\} \right\|_{\max} < \eta_n, \tag{A.15}$$

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.15). Condition (A.15) is indeed verified by Lemma 9. This completes the proof of Theorem 4. □

Proof of Corollary 4. Recall that $\hat{\alpha}_j = \log \widehat{N}_j / \log N$ with $\widehat{N}_j = |\,\mathrm{supp}(\widehat{\mathbf{b}}_j^{\mathsf{ada}})|$ and $\alpha_j = \log N_j / \log N$ by the definition. Because $\{\mathrm{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}}) = \mathrm{supp}(\mathbf{B}^0)\} \subset \{\widehat{N}_j = N_j \text{ for all } j = 1,\ldots,r\}$, we have

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

The last probability tends to one by the factor selection consistency. This completes the proof of Corollary 4. □

B Related Lemmas and their 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) $X_iX_j \sim$ *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), but is repeated here for completeness. (a) Since X_i is subG(α_i^2), we obtain $\mathbb{E}\,|X|^k \leq (2\alpha_i^2)^{k/2}k\Gamma(k/2)$; see Rigollet and Hütter (2017), for instance. Then we see from the dominated convergence theorem and

independence that

$$\begin{aligned}
\mathbb{E}\exp(sX_iX_j) &= 1+\sum_{k=2}^{\infty}\frac{s^k\,\mathbb{E}(X_iX_j)^k}{k!} \le 1+\sum_{k=2}^{\infty}\frac{s^k\,\mathbb{E}\,|X_i|^k\,\mathbb{E}\,|X_j|^k}{k!} \\
&\le 1+\sum_{k=2}^{\infty}\frac{s^k(2\alpha_i\alpha_j)^k k^2\Gamma(k/2)^2}{k!} \le 1+\sum_{k=2}^{\infty}\frac{s^k(2\alpha_i\alpha_j)^k k^2(k/2)^k}{k!} \\
&= 1+\sum_{k=2}^{\infty}\frac{s^k(\alpha_i\alpha_j)^k k^{k+2}}{k!} \le 1+\sum_{k=2}^{\infty}(2e\alpha_i\alpha_j s)^k \\
&= 1+(2e\alpha_i\alpha_j s)^2\sum_{k=0}^{\infty}(2e\alpha_i\alpha_j s)^k,
\end{aligned}$$

where we have used $\Gamma(k/2) \le (k/2)^{k/2}$ and $k^{k+2} \le (2\pi)^{-1/2}k!e^k k^{3/2} \le k!(2e)^k$. Therefore, for any $|s| \le (4e\alpha_i\alpha_j)^{-1}$, it holds that

$$\mathbb{E}\exp(sX_iX_j) \le 1+8(e\alpha_i\alpha_j)^2s^2 \le \exp((4e\alpha_i\alpha_j)^2s^2/2).$$

This means that the product X_iX_j is $\mathrm{subE}(4e\alpha_i\alpha_j)$.

(b) By the definition of subG, we have

$$\begin{aligned}
\mathbb{E}\exp\left(s\sum_{i=1}^{n}\phi_iX_i\right) &= \prod_{i=1}^{n}\mathbb{E}\exp\left(s\phi_iX_i\right) \\
&\le \prod_{i=1}^{n}\exp\left(s^2\phi_i^2\alpha_i^2/2\right) = \exp\left(s^2\sum_{i=1}^{n}\phi_i^2\alpha_i^2/2\right),
\end{aligned}$$

which yields the result.

(c) First note that $\psi_iY_i \sim \mathrm{subE}(\psi_i\gamma_i, |\psi_i|\gamma_i)$ because $\mathbb{E}\exp(s\psi_iY_i) \le \exp(s^2\psi_i^2\gamma_i^2/2)$ holds for all $|s| \le (|\psi_i|\gamma_i)^{-1}$. Thus, we can see that

$$\begin{aligned}
\mathbb{E}\exp\left(s\sum_{i=1}^{n}\psi_iY_i\right) &= \prod_{i=1}^{n}\mathbb{E}\exp\left(s\psi_iY_i\right) \\
&\le \prod_{i=1}^{n}\exp\left(s^2\psi_i^2\gamma_i^2/2\right) = \exp\left(s^2\sum_{i=1}^{n}\psi_i^2\gamma_i^2/2\right),
\end{aligned}$$

where the inequality holds for all $|s| \le (\max_i|\psi_i|\gamma_i)^{-1}$. This gives the result by the definition of subE, and completes all the proofs. □

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

(*a*) $T^{-1}\left|\mathrm{tr}\,\mathbf{H}'\mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}\mathbf{E}'\mathbf{H}\right| \lesssim TN_1^{1/2}\log^{1/2}(N\vee T)$,

(*b*) $T^{-1}\,\mathrm{tr}\,\mathbf{H}'\mathbf{E}\mathbf{P}\mathbf{E}'\mathbf{H} \lesssim N\vee T$,

(*c*) $\lambda_1(\mathbf{E}\mathbf{Q}\mathbf{E}') \lesssim T\vee N$,

(*d*) $T^{-1}\,\mathrm{tr}(\mathbf{H}'\mathbf{E}\mathbf{Q}\mathbf{E}'\mathbf{H}) \lesssim T\vee N$.

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.4), we consider a tighter bound of the second trace. The second trace in the upper bound of(A.4) 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 constants c_1 and c_2 such that $c_1 < c_2$ and C, we have

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

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

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

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

$$\kappa_n = \frac{N_r(N_r\wedge T)}{N_1}.$$

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:

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

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

$$\begin{aligned}\|\widehat{\mathbf{U}}_* - \mathbf{U}_*^0\|_{\mathrm{F}}^2 &= \sum_{k=1}^{r} \|\widehat{\mathbf{u}}_{*k} - \mathbf{u}_{*k}^0\|_2^2 \leq \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 (cd_{*1}^2/\delta) \sum_{k=1}^{r} d_{*k}^{-4} \\ &= cd_1^2 N_1 \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \sum_{k=1}^{r} \frac{1}{\delta d_k^4 N_k^2} \\ &= \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 (cd_1^2 N_1/\delta) \sum_{k=1}^{r} (d_k N_k^{1/2})^{-4} \lesssim \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \frac{N_1}{N_r^2}.\end{aligned}$$

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

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

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

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

Combining these inequalities gives

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

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*

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

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

Proof. By the definition of the PC estimator under PC1 restriction, we have

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

which together with the triangle inequality implies that

$$\begin{aligned}&\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\max} \\ &\quad\le T^{-1}\|\mathbf{E}'\mathbf{F}^0\|_{\max} + T^{-1}\|\mathbf{B}^0\mathbf{F}^{0\prime}(\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0)\|_{\max} + T^{-1}\|\mathbf{E}'(\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0)\|_{\max}. \qquad \text{(A.16)}\end{aligned}$$

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. The second term of (A.16) is evaluated as

$$\begin{aligned}T^{-1}\|\mathbf{B}^0\mathbf{F}^{0\prime}(\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0)\|_{\max} &\le rT^{-1}\|\mathbf{B}^0\|_{\max}\|\mathbf{F}^{0\prime}(\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0)\|_{\max} \\ &\lesssim T^{-1}\|\mathbf{F}^0\|_{\mathrm{F}}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}} \\ &\lesssim T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}}.\end{aligned}$$

By the Cauchy-Schwarz inequality, the third term of (A.16) is evaluated as

$$\begin{aligned}T^{-1}\|\mathbf{E}'(\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0)\|_{\max} &= T^{-1}\max_{i,k}\left|\sum_t e_{ti}(\hat{f}_{tk} - f^0_{tk})\right| \\ &\le \max_i\left(T^{-1}\sum_t e_{ti}^2\right)^{1/2}\max_k\left(T^{-1}\sum_t(\hat{f}_{tk} - f^0_{tk})^2\right)^{1/2} \\ &\le \max_i\left\{\left|T^{-1}\sum_t\left(e_{ti}^2 - \mathbb{E}\,e_{ti}^2\right)\right|^{1/2} + \left|T^{-1}\sum_t \mathbb{E}\,e_{ti}^2\right|^{1/2}\right\}T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}} \\ &\lesssim T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}},\end{aligned}$$

where the last inequality follows from Lemma 1(d) and

$$\begin{aligned}\|\,\mathbb{E}\,\mathbf{e}_t\mathbf{e}_t'\|_{\max} &\le \sum_{\ell=0}^{L_n}\|\mathbf{\Phi}_\ell\,\mathbb{E}\,\boldsymbol{\varepsilon}_{t-\ell}\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{\Phi}_\ell'\|_{\max} \\ &\lesssim \sum_{\ell=0}^{L_n}\|\mathbf{\Phi}_\ell\mathbf{\Phi}_\ell'\|_{\max} \lesssim \sum_{\ell=0}^{\infty}\|\mathbf{\Phi}_\ell\|_2^2 < \infty.\end{aligned}$$

Consequently, by Theorem 3 with condition (10), the bound in (A.16) becomes

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

with probability at least $1 - O((N \vee T)^{-\nu})$. In this upper bound, the second term dominates the first term because

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

Thus, we have

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

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

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

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

$\boldsymbol{\Phi}_\ell' \mathbf{b}_k^0$ as in the proof of Lemma 1(b). Then, we have

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

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

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

Therefore, the subG tail inequality and the union bound entail

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

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

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

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

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

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

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

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 $h = 1,\dots,r$ and $j = 1,\dots,N$. Recall that $\widetilde{\mathbf{E}}'_\ell = (\boldsymbol{\varepsilon}_{1-\ell},\dots,\boldsymbol{\varepsilon}_{T-\ell})$ and $\widetilde{\mathbf{Z}}'_m = (\boldsymbol{\zeta}_{1-m},\dots,\boldsymbol{\zeta}_{T-m})$ from the proof of Lemma 1. Then we have

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

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

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

Thus the desired upper bound

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

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

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

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

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

which yields the result.

(d) By the property of norms, we obtain

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

By Assumption 1, the last stochastic part is evaluated as

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

where $\{\zeta_{tm}\}_{t,m} \sim$ i.i.d. subG(σ_ζ^2) and $\sum_{\ell=0}^\infty\|\mathbf{\Psi}_\ell\|_2$ is bounded. By Lemma 2(b), we have $\sum_{t=1}^T \zeta_{t-\ell,m}u_{tj} \sim$ 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

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

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

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

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

(e) We observe that

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

which gives the proof.

(f) By the property of norms, we obtain

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

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*

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

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

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

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

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

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

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

$$\begin{aligned}\mathbb{P}\left(\hat{\underline{b}}_n \leq \underline{b}_n^0/2\right) &\leq \mathbb{P}\left(\|\widehat{\mathbf{B}}_{\mathsf{ini}} - \mathbf{B}^0\|_{\max} \geq \underline{b}_n^0/2\right) \\ &\leq \mathbb{P}\left(\|\widehat{\mathbf{B}}_{\mathsf{ini}} - \mathbf{B}^0\|_{\max} \gtrsim \frac{N^{1/2}\gamma_n(\widetilde{N})}{(\widetilde{N} \vee T)^{1/2}}\right) = O((N \vee T)^{-\nu}).\end{aligned}$$

Consequently, we obtain

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

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

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

$$\mathcal{F} = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{T \times r} : \|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq C r_n\right\}$$

where C is some positive constant introduced in the proof of Theorem 4 and

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

Then we have

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

We evaluate the three terms in the parenthesis. Lemma 1(c) gives $\|\mathbf{E}'\mathbf{F}^0\|_{\max} \lesssim T^{1/2} \log^{1/2}(N \vee T)$. We next observe that for any $\boldsymbol{\Delta}$,

$$\left\|\mathbf{B}^0\mathbf{F}^{0\prime}\boldsymbol{\Delta}\right\|_{\max} \leq r \left\|\mathbf{B}^0\right\|_{\max} \left\|\mathbf{F}^{0\prime}\boldsymbol{\Delta}\right\|_{\max} \lesssim \left\|\mathbf{F}^{0\prime}\boldsymbol{\Delta}\right\|_{\max}.$$

The upper bound is further bounded as

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

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 $\sum_j \psi_{\ell,ij} z_{\ell,jk} \sim \mathrm{subG}(\sigma_\zeta^2 \|\boldsymbol{\delta}_k\|_2^2 \|\boldsymbol{\Psi}_{\ell,i\cdot}\|_2^2)$ for each i, k, and ℓ. Therefore, for any fixed $\boldsymbol{\Delta}$ and ℓ, the subG tail inequality with the union bound entails that

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

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

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

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

$$\sup_{\mathbf{\Delta} \in \mathcal{F}} \left\| \mathbf{E}' \mathbf{\Delta} \right\|_{\max} \lesssim r_n \log^{1/2}(N \vee T)$$

with probability at least $1 - O((N \vee T)^{-\nu})$. Combining the obtained bounds yields

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

We can obtain $T^{1/2} + r_n = O(T^{1/2})$ because condition (10) entails

$$\frac{r_n}{T^{1/2}} = \frac{N_1^{3/2}}{N_r(N_r \wedge T)} = o\left(\frac{N_1^{3/2}}{N_1 (N \vee T)^{1/2}} \right) = o\left(\frac{N_1^{1/2}}{(N \vee T)^{1/2}} \right) = o(1).$$

Hence, by Lemma 6 and condition (18), we have

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

with probability at least $1 - O((N \vee T)^{-\nu})$. The last value is $o(1)$ since $\gamma_n(\widetilde{N})$ tends to zero polynomially. This asymptotically guarantees the desired strict inequality. □

C Additional Simulation Result: Global and Local Factors

In empirical data analysis, often the factor models are applied to capture the dependence withing a sector or group, even hierarchical structure. For example, Ando and Bai (2017) consider two types of factors: global factors of which factor loadings have non-zero values for all the cross-section units, whilst the local factors of which loadings of specific cross sectional group are non-zero, and zero otherwise. Estimation of this class of hierarchical factor structure is of great interest in practice, and some estimation methods have been proposed.

Our WF-SOFAR estimator is easily applicable to estimate the model. To see how useful our estimator, we implemented a small experiment. We generate the data of four factors models, $x_{ti} = \sum_{k=1}^{r} b_{ik} f_{tk} + e_{ti}$, where $f_{tk} = \rho_{fk} f_{t-1,k} + v_{tk}$ with $v_{tk} \sim$ i.i.d.$N(0, 1 - \rho_{fk}^2)$ and $f_{1k} \sim$ i.i.d.$N(0,1)$; $e_{ti} = \rho_e e_{t-1,i} + \beta \varepsilon_{t,i-1} + \beta \varepsilon_{t,i+1} + \varepsilon_{ti} \varepsilon_{ti} \sim$ i.i.d. $N(0, \sigma_{\varepsilon,ti}^2)$ and $\sigma_{\varepsilon,ti}^2$ is set so that $\mathrm{Var}(e_{ti}) = 1$ for $i = 1, \ldots, N$, $t = 1, \ldots, T$. We set $r = 4$. The first factor is a global factor, i.e., $b_{i1} \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, N$. The other three factors are local ones, 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 factor model given $r = 4$ by the PC and WF-SOFAR. To visualize the quality of the factor loadings, we provide heat maps of three $N \times N$ matrices, $\sum_{k=1}^{4} \omega_k \, \mathrm{abs}(\mathbf{b}_k^0 \mathbf{b}_k^{0\prime})$, $\sum_{k=1}^{4} \omega_k \, \mathrm{abs}(\widehat{\mathbf{b}}_k \widehat{\mathbf{b}}_k')$ and $\sum_{k=1}^{4} \omega_k \, \mathrm{abs}(\widehat{\mathbf{b}}_{\mathsf{PC},k} \widehat{\mathbf{b}}_{\mathsf{PC},k}')$. To clarify the difference between the global factor loadings and local ones, which overlaps, we use the weight $\omega_1 = 1/8$ and $\omega_2 = \omega_3 = \omega_4 = 1$.

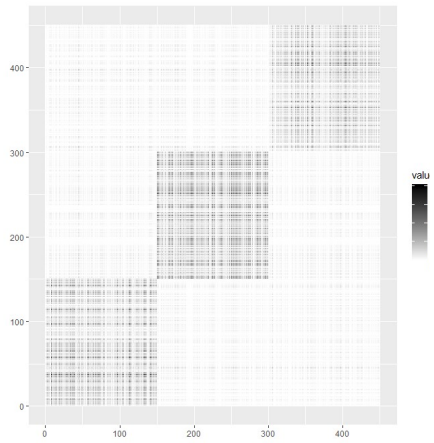

Figure 2: True factor loadings

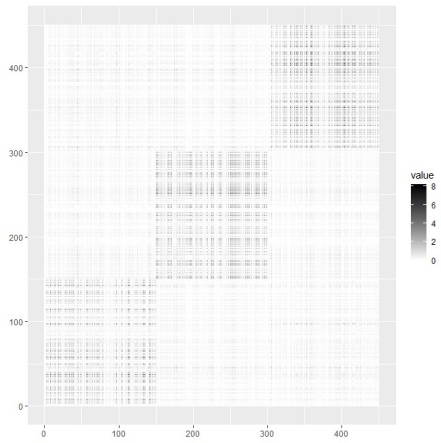

Figure 3: WF-SOFAR estimate

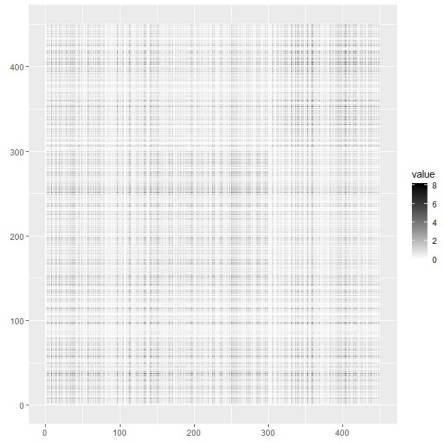

Figure 4: PC estimate

D Additional Results of Empirical Example 1: Firm Security Returns

In addition to reporting the divergence rates, 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)[17]: (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, whilst *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

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

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* 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,[18] α_4 rises to become the highest next to the first factor from November 2016 onward.

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

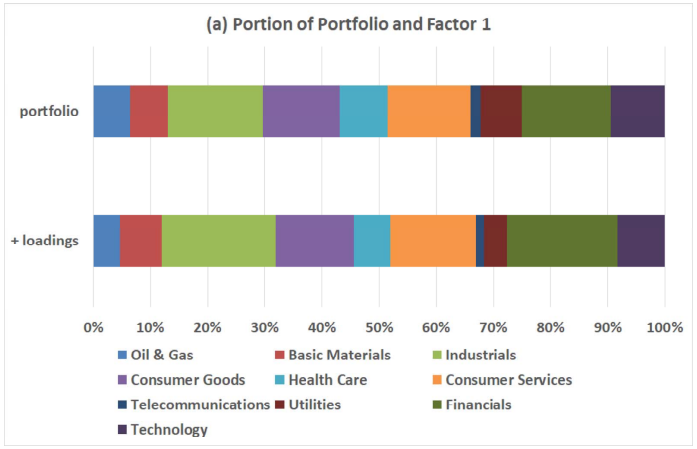

Figure 5: the portion of the industrial sectors in S&P500 and in the positive/negative 1st factor loadings

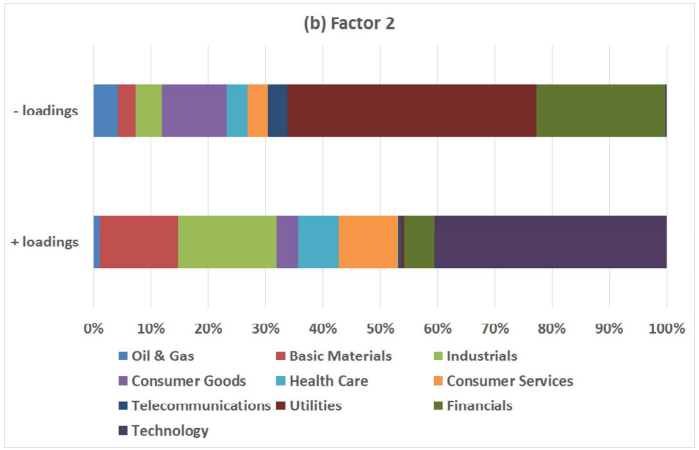

Figure 6: the portion of the industrial sectors in the positive/negative second factor loadings

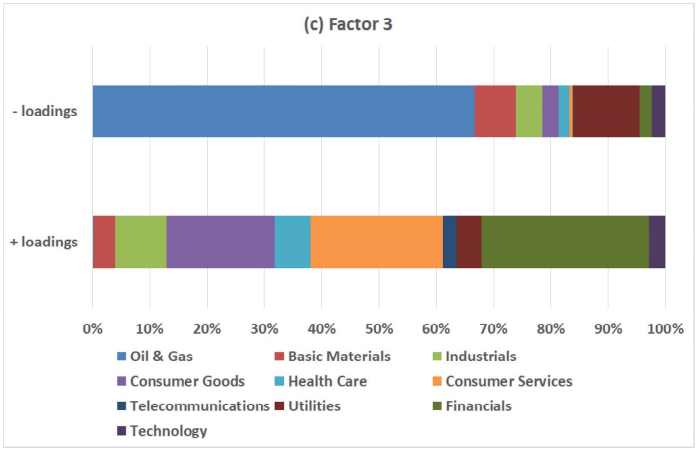

Figure 7: the portion of the industrial sectors in the positive/negative third factor loadings

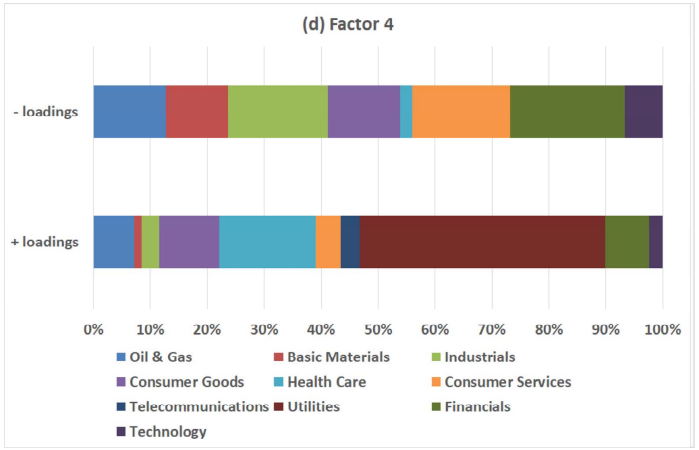

Figure 8: the portion of the industrial sectors in the positive/negative fourth factor loadings