Estimation of Sparsity-Induced 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*

October 2021

Abstract

This paper investigates estimation of sparsity-induced weak factor (sWF) models, with large cross-sectional and time-series dimensions (N and T, respectively). It assumes that the kth largest eigenvalue of a data covariance matrix grows proportionally to N^{α_k} with unknown exponents $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. Employing the same rotation of the principal components (PC) estimator, the growth rate α_k is linked to the degree of sparsity of kth factor loadings. This is much weaker than the typical assumption on the recent factor models, in which all the r largest eigenvalues diverge proportionally to N. We apply the method of *sparse orthogonal factor regression* (SOFAR) by Uematsu et al. (2019) to estimate the sWF models and derive the estimation error bound. Importantly, our method also yields consistent estimation of α_k. A finite sample experiment shows that the performance of the new estimator uniformly dominates that of the PC estimator. We apply our method to forecasting bond yields and the results demonstrate that our method outperforms that based on the PC. We also analyze S&P500 firm security returns and find that the first factor is consistently near strong while the others are weak.

[*] Correspondence: Yoshimasa Uematsu, Department of Economics and Management, Tohoku University, 27-1 Kawauchi, Aobaku, Sendai 980-8576, Japan (E-mail: yoshimasa.uematsu.e7@tohoku.ac.jp).

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

1 Introduction

The approximate factor model with large cross-sectional and time-series dimensions (N and T, respectively) has become an increasingly important tool for the analysis of finance, economics, psychology, and biology, among many other academic fields. In finance, the model was firstly introduced by Chamberlain and Rothschild (1983), then developed in 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 (2002a,b) propose to extract a small number of factors from the large macroeconomic and financial series and use them to forecast a macroeconomic variable of interest. Ludvigson and Ng (2009) take a similar approach to forecast bond yields. See, for example, Fan et al. (2018) for an excellent review of the high-dimensional factor models and their applications.

1.1 Weak factor model, rotation, and sparsity

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

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

where $\mathbf{B}^* = (\mathbf{b}_1^*, \ldots, \mathbf{b}_r^*) \in \mathbb{R}^{N \times r}$ with $\mathbf{b}_k^* \in \mathbb{R}^N$ is a matrix of deterministic factor loadings which has full column rank, $\mathbf{f}_t^* \in \mathbb{R}^r$ is a vector of zero-mean latent factors, and $\mathbf{e}_t \in \mathbb{R}^N$ is an idiosyncratic error vector independent of $\mathbf{f}_t^*$. For the present time, suppose r is given. Let $\mathbf{\Sigma}_x = \mathbb{E}[\mathbf{x}_t \mathbf{x}_t']$, $\mathbf{\Sigma}_f^* = \mathbb{E}[\mathbf{f}_t^* \mathbf{f}_t^{*\prime}]$, and $\mathbf{\Sigma}_e = \mathbb{E}[\mathbf{e}_t \mathbf{e}_t']$ with assuming all the eigenvalues of $\mathbf{\Sigma}_f^*$ and $\mathbf{\Sigma}_e$ are bounded away from zero and from above (uniformly in N). We then observe that $\lambda_k(\mathbf{\Sigma}_x) \asymp \lambda_k(\mathbf{B}^* \mathbf{\Sigma}_f^* \mathbf{B}^{*\prime})$ for each $k = 1, \ldots, r$ and $\lambda_{r+1}(\mathbf{\Sigma}_x) = O(1)$, where $\lambda_k(\cdot)$ denotes the kth largest eigenvalue.

In the studies on high-dimensional factor models which employ the principal components (PC) estimator, including Connor and Korajczyk (1986, 1993), Stock and Watson (2002a,b), Bai and Ng (2002, 2006, 2013), Bai (2003) and Fan et al. (2018), it is typically assumed that all the r largest eigenvalues diverge proportional to N, namely, $\lambda_k(\mathbf{B}^*\mathbf{\Sigma}_f^*\mathbf{B}^{*\prime}) \asymp N$ for all $k = 1, \ldots, r$. We call the models the *strong factor (SF) models*. This SF assumption seems unduly restrictive. The SF model does not permit slower divergence rates than N, nor different divergence rates among the r largest eigenvalues. The original *approximate factor model* proposed by Chamberlain and Rothschild (1983) is an important exception, which assumes that $\lambda_r(\mathbf{B}^*\mathbf{\Sigma}_f^*\mathbf{B}^{*\prime}) \to \infty$ as $N \to \infty$. Furthermore, the SF model implies a large gap between the values of $\lambda_r(\mathbf{\Sigma}_x)$ and $\lambda_{r+1}(\mathbf{\Sigma}_x)$, but it is often missing in representative financial and economic data sets; see the discussion in Fan et al. (2013) by Onatski.

In light of the above discussion, we will significantly relax the SF condition as follows:

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

We call the factor models with (2) the *weak factor (WF) models* in this paper. The WF models allow different divergence rates of the signal eigenvalues, which can be slower than N. Our definition of the WF models is similar to that in De Mol et al. (2008) and Onatski (2012, p.246), but the reader is cautioned that the definition varies in the literature. For example, Onatski (2012), Bryzgalova (2016), Lettau and Pelger (2020) assume non-diverging factors (i.e. $\alpha_r = 0$), which Chamberlain and Rothschild (1983) and ourselves exclude. Chudik et al. (2011) categorize the factors according to the values of the exponents.

It is well known that estimation of factor models, including (1), has an identification issue. To identify the factors and the loading matrix separately, we must impose r^2 (or more) restrictions on the model. To directly identify the true loading matrix $\mathbf{B}^*$, e.g. for studying a structural shock in macroeconomics, r^2 (or more) restrictions, together with cross-section ordering of x_{ti}, should be imposed. Such restrictions are typically exogenously informed by economic or financial theory; see the "named factor normalization" in Stock and Watson (2016), for example.[1)] However, such exogenous information is not always available. For many empirical studies, it is sufficient to identify the subspace spanned by the column vectors of $\mathbf{F}^*$. In these cases, the identification can be achieved by imposing an arbitrarily chosen set of r^2 restrictions, e.g. for mathematical convenience.

To discuss our identification restrictions further, observe that the column and row spaces of $\mathbf{F}^* = (\mathbf{f}_1^*, ..., \mathbf{f}_T^*)'$ and $\mathbf{B}^{*\prime}$ are identical to those of $\mathbf{F}^*\mathbf{Q}$ and $\mathbf{Q}^{-1}\mathbf{B}^{*\prime}$, respectively, for any invertible matrix $\mathbf{Q}$. We choose a specific rotation which is also used for the PC estimator. That is, we put $\mathbf{f}_t^0 = \mathbf{H}\mathbf{f}_t^*$ and $\mathbf{B}^{0\prime} = \mathbf{H}^{-1}\mathbf{B}^{*\prime}$ with $\mathbf{\Sigma}_f = \mathbb{E}[\mathbf{f}_t^0\mathbf{f}_t^{0\prime}] = \mathbf{I}_r$ and $\mathbf{B}^{0\prime}\mathbf{B}^0$ being a diagonal matrix. Then the model in (1) becomes

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

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

$$N_k \asymp \lambda_k(\mathbf{B}^0\mathbf{B}^{0\prime}) = \lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) = \mathbf{b}_k^{0\prime}\mathbf{b}_k^0 \;\text{ for each }\; k = 1, \dots, r, \tag{4}$$

where the last equality is due to the specific choice of the rotation matrix, $\mathbf{H}$. In our approach, we assume $\mathbf{B}^0$ is sparse and link the degree of sparsity in $\mathbf{b}_k^0$ to the divergence rate of $\lambda_k(\mathbf{\Sigma}_x)$, N_k.[2] This is called the *sparsity-induced weak factor (sWF) model*, and we investigate the estimation. As the earlier discussion implies, the WF structure in (4) can be induced by *non-sparse* factor loadings. For instance, this is the case when a factor affects all the variables at similar strengths thinly, but we do not consider this class in the paper.[3]

1.2 Empirical evidence of the sWF models

Influential empirical studies often give implicit yet strong evidence of sWF models under the restrictions we impose. Stock and Watson (2002b) and Ludvigson and Ng (2009) extract the PC factors from standardized N macroeconomic variables (x_{ti}). They run N time-series regressions of the variables on each of the extracted PC factors, then report N values of R^2s rather than the PC loadings; see figure 1 in Stock and Watson (2002b) and figures 1–5 in Ludvigson and Ng (2009). They find interesting local spikes in some R^2s while the rest are very close to zero. As an illustration, we have conducted a similar exercise by extracting two PC factors from the standardized 131 macroeconomic variables between January 1982 and December 2001, which is used in Ludvigson and Ng (2009).[4] Figure 1 reports the R^2s for the time-series regressions of the variables on the second PC factor. The values of R^2s for 79 variables out of 131 (60.3%) are less than 0.01. This factor has virtually zero explanatory power for these variables. Thus, it is to be regarded as a sWF model under the orthogonality restrictions. The corresponding PC loadings are illustrated in Figure 2. Note that such figures are not reported in Stock and Watson (2002b) and Ludvigson and Ng (2009). As can be seen, taking into account the standardization of x_{ti}, the absolute values of the loadings corresponding to the near-zero R^2s are disproportionately large. This strongly indicates that ℓ_1-norm regularization of the loadings will improve the estimation efficacy. Such estimators will be proposed and investigated in this paper, and in Section 6.2, Figure 11 will visualize that the sparse loading estimates can provide sharper and richer information about the factors than the R^2s and the PC loadings (e.g. the signs of the loadings), which are illustrated by Figures 1 and 2.

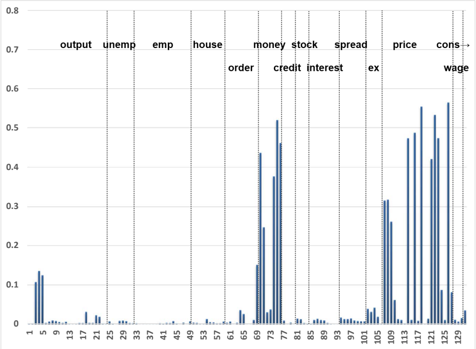

Figure 1: R^2 for regression of x_{it} on $\hat{f}_{t2}^{\mathsf{PC}}$

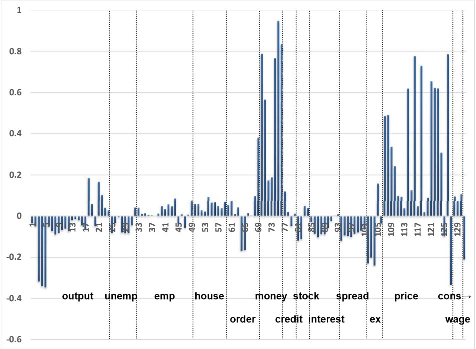

Figure 2: PC estimates $\hat{b}_{i2}^{\mathsf{PC}}$

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

Finally, the sparsity assumption is testable, and an inferential method is explored in the separate paper, Uematsu and Yamagata (2021). Therein, we find evidence of high degrees of sparsity in $\mathbf{B}^0$ with the FRED-MD data set discussed in McCracken and Ng (2016) and in the firm security excess returns of the S&P500 index.

1.3 Contributions

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

As theoretical contributions, we will establish the estimation error bounds of the SOFAR and PC estimators as well as validating the method of Onatski (2010) for determining the number of factors for the sWF models. Our results reveal that when the model is even slightly weaker than the strong one, the SOFAR utilizes the sparsity and can converge faster than the PC estimator. Even when all the factors are strong, the SOFAR estimator is likely to converge at least as fast as the PC estimator. We also propose the *adaptive* SOFAR estimator, which yields *factor selection consistency*. This property asymptotically guarantees the true support recovery of the sparse loadings. The assumptions we will make are in line with the literature of the approximate factor models with serially correlated factors as well as cross-sectionally and serially correlated errors. Thus the statistical theory substantially departs from that in Uematsu et al. (2019), which considers multivariate linear regression models with i.i.d. Gaussian errors.

Importantly, the factor selection consistency enables us to consistently estimate each exponent α_k of the divergence rates. Recently estimation of the exponents has drawn great attention among empirical researchers since it is a useful measure of the strength of the cross-sectional correlations. Assuming sparse loadings, Bailey et al. (2016, 2021) and Gao et al. (2020) propose methods that make use of cross-sectional averages of data for estimation and inference of the exponent, but they can only identify the largest divergence rate, α_1. This is essentially because they focus on estimation of the structural model (1). In contrast, our method can identify all the divergence rates because we impose the identification restrictions and focus on the rotated model (3).[5)] We implement extensive finite sample experiments in terms of determining the number of factors and estimation accuracy, and find that the SOFAR estimate uniformly dominates the PC estimate over all the designs we consider.

We also conduct empirical analysis with a large data set of macroeconomic variables and S&P 500 monthly returns. In the first analysis, we compare the out-of-sample performance of forecasting bond yields using extracted factors from the macroeconomic variables via our method and the PC method. The statistical evidence suggests that our SOFAR outperforms the PC. In the second analysis, we compare the PC and the SOFAR loading estimates to illustrate the usefulness of looking into sparse factor loadings (rather than the R^2s discussed above) for finding properties of the extracted factors. In the third analysis, applying our method to the residuals of the regressions of the S&P 500 monthly excess returns on the Fama and French (2015) five factors uncovers a hidden, very weak factor with the exponent 0.39. The factor affects only eight firm securities, all of which belong to the Technology Hardware & Equipment sector. The fourth analysis shows that the first factor in S&P 500 monthly returns is consistently near strong, while the second to fourth exponents vary over months between 0.90 and 0.65.

1.4 Related work

To our knowledge, this is the first study to propose a method that can estimate the WF models, separately identifying rotations of $\mathbf{B}^*$ and $\mathbf{F}^*$, while taking the possibly different rates (2) into account.[6)] There are some studies that consider WF models, but most of them have focused only on the case in which all the divergence rates are identical. Such examples are seen in De Mol et al. (2008) and Lam et al. (2011); the former consider the Bayesian forecasts with the PC estimates for WF models, and the latter propose an efficient estimator for WF models with a specific correlation structure. Another related recent work is Daniele et al. (2020) who extend Bai and Li (2012) and Bai and Liao (2017) to sparse factor models.

Given initial estimators of $\mathbf{B}^0$ and $\mathbf{f}_t^0$, some researchers consider identifying a specific rotation which achieves a desired criterion. Perhaps the most well-known criterion is the varimax rotation of Kaiser (1958), which finds the rotation that maximizes the sum of the variances of the squared loadings. Recently, Freyaldenhoven (2020) has proposed another criterion to discover a rotation which maximizes the sparsity in the loadings. Our estimator is complementary for them as it can add a potentially more efficient initial estimator.

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

1.5 Organization and notational remarks

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

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

2 Sparsity-Induced Weak Factor Models

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

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

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

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

As discussed in our Introduction, we consider *sparsity-induced* WF (sWF) models. Specifically, we assume sparse factor loadings $\mathbf{B}^0$ such that the sparsity of kth column (i.e., the number of non-zero elements in $\mathbf{b}_k^0 \in \mathbb{R}^N$) is $N_k := N^{\alpha_k}$ for $k \in \{1, \ldots, r\}$, where $1 \geq \alpha_1 \geq \cdots \geq \alpha_r > 0$ and exponents α_k is unknown. Note that N_r must diverge since $\alpha_r > 0$ and $N \to \infty$. We may relax the *exact* sparseness by introducing the *approximate* sparse loadings; that is, $\mathbf{B}^0 = (b_{ik})$ such that $\sum_{i=1}^N |b_{ik}| \asymp N_k$. This does not necessarily require exact zeros in $\mathbf{B}^0$. However, we choose not to pursue this direction to avoid a complicated technical issue.

By the sparsity assumption and the diagonality of $\mathbf{B}^{0\prime}\mathbf{B}^0$, there exist some positive constants $d_1, \ldots, d_r$ such that

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

and $d_1^2 N_1 \geq \cdots \geq d_r^2 N_r > 0$. Then, under the assumption $\max_N \lambda_1(\mathbf{\Sigma}_e) < \infty$, we have

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

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

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

3 Estimation

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

3.1 SOFAR estimation

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

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

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

It is well known that the PC estimator is easily obtained by the eigenvalue problem on $\mathbf{X}\mathbf{X}'$; specifically, for given $\hat{r}$, $\widehat{\mathbf{F}}_{\mathsf{PC}}$ is obtained as $T^{1/2}$ times the eigenvectors corresponding to the top $\hat{r}$ largest eigenvalues of $\mathbf{X}\mathbf{X}'$ and $\widehat{\mathbf{B}}_{\mathsf{PC}} = \mathbf{X}'\widehat{\mathbf{F}}_{\mathsf{PC}}/T$. On the other hand, the SOFAR estimator is no longer computed by the eigenvalue problem. Even some algorithms used for the lasso, such as coordinate descent, cannot be directly applied to the problem due to the restrictions, sparsity and orthogonality (diagonality). In order to overcome this difficulty, we apply the SOFAR algorithm proposed by Uematsu et al. (2019) to solve (6). Roughly speaking, the algorithm provides estimates for the SVD of a coefficient matrix in a multiple linear regression, while simultaneously exhibiting both low-rankness in the singular values matrix and sparsity in the singular vectors matrices. Recall the connection between $(\mathbf{F}, \mathbf{B})$ and $(\mathbf{U}, \mathbf{D}, \mathbf{V})$, which has been defined by the SVD of $\mathbf{C}$, in Section 2. Then for given $\hat{r}$, the SOFAR algorithm can solve (6) to get $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}) = (\widehat{\mathbf{U}}, \widehat{\mathbf{V}}\widehat{\mathbf{D}})$.

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

3.2 Adaptive SOFAR estimation

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

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

$$\text{subject to } \mathbf{F}'\mathbf{F}/T = \mathbf{I}_{\hat{r}} \text{ and } \mathbf{B}'\mathbf{B} \text{ diagonal,}$$

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

Estimating exponents α_k, $k = 1, ..., r$, is of great interest to empirical research since, as discussed in Bailey et al. (2016, 2021), they are interpreted as the strength of the influence of the common factors and of the cross-sectional correlations. Recall that the kth column of $\mathbf{B}^0$, $\mathbf{b}_k^0$, has $N_k = N^{\alpha_k}$ non-zero entries. Similarly, let $\widehat{N}_k$ denote the number of non-zero elements in $\widehat{\mathbf{b}}_k^{\mathsf{ada}}$. As the lasso in a linear regression, we may expect that the adaptive SOFAR estimate $\widehat{\mathbf{B}}^{\mathsf{ada}}$ can successfully recover the true sparsity pattern of $\mathbf{B}^0$. If this is true, the estimators of exponents α_k can naturally be obtained as $\hat{\alpha}_k = \log \widehat{N}_k / \log N$ by a simple algebraic formulation. In Section 4.3, we will prove this estimator is consistent for α_k.

4 Theory

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

For the sake of convenience, define $n = N \wedge T$. Then we have $N = N(n) \to \infty$ and $T = T(n) \to \infty$ as $n \to \infty$. Furthermore, following Vershynin (2018), we introduce a *sub-Gaussian* random variable: a random variable $Y \in \mathbb{R}$ is said to be sub-Gaussian and denoted as $Y \sim \mathrm{subG}$ if there exists some constant $c > 0$ such that $\mathbb{P}(|Y| \geq y) \leq 2\exp(-y^2/c)$ for all $y \geq 0$. Throughout this paper, including all the proofs in Supplementary Material, $\nu > 0$ is a fixed large constant, and n is sufficiently large.

Write $T = N^{\tau}$ for some constant $\tau > 0$ to understand the size of T relative to N. Recall that $N_k = N^{\alpha_k}$ for some $\alpha_k \in (0, 1]$.

Assumption 1 (Latent factors)**.** The factor matrix $\mathbf{F}^0 = (\mathbf{f}_1^0, \dots, \mathbf{f}_T^0)'$ is specified as the vector linear process $\mathbf{f}_t^0 = \sum_{\ell=0}^{\infty} \mathbf{\Psi}_\ell \boldsymbol{\zeta}_{t-\ell}$, where $\boldsymbol{\zeta}_t = (\zeta_{t1}, \dots, \zeta_{tr})'$ with $\{\zeta_{tk}\}_{t,k}$ are i.i.d. subG that has $\mathbb{E}\,\zeta_{tk}^2 = 1$ and $\sum_{\ell=0}^{\infty} \mathbf{\Psi}_\ell \mathbf{\Psi}_\ell' = \mathbf{I}_r$. Moreover, there exist constants $C_f > 0$ and $\ell_f \in \mathbb{N}$ such that $\|\mathbf{\Psi}_\ell\|_2 \leq C_f \ell^{-(\nu+2)}$ for all $\ell \geq \ell_f$.

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

Assumption 3 (Idiosyncratic errors)**.** The error matrix $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_T)'$ is independent of $\mathbf{F}^0$ and is specified as the vector linear process $\mathbf{e}_t = \sum_{\ell=0}^{\infty} \mathbf{\Phi}_\ell \boldsymbol{\varepsilon}_{t-\ell}$, where $\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \dots, \varepsilon_{tN})'$ with $\{\varepsilon_{ti}\}_{t,i}$ are i.i.d. subG and $\mathbf{\Phi}_0$ is a nonsingular, lower triangular matrix. Moreover, there exist constants $C_e > 0$ and $\ell_e \in \mathbb{N}$ such that $\|\mathbf{\Phi}_\ell\|_2 \leq C_e \ell^{-(\nu+2)}$ for all $\ell \geq \ell_e$.

Assumptions 1 and 3 specify the stochastic processes $\{\mathbf{f}_t^0\}$ and $\{\mathbf{e}_t\}$, respectively, to be the stationary vector linear processes that satisfy the summability condition $\sum_{\ell=0}^{\infty}(\|\mathbf{\Psi}_\ell\|_2 + \|\mathbf{\Phi}_\ell\|_2) < \infty$. The decaying rates are at most polynomial, which includes a wide range of multivariate weakly dependent processes (Chen and Wu, 2018). Under this condition we can achieve the concentration inequalities in Lemma 1 in Supplementary Material, which play a crucial role in deriving the error bounds. Assumption 2 is key to our analysis and provides the sWF models. The sparsity in $\mathbf{B}^0$ makes the divergence rate of $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0)$ possibly slower than N for each $k = 1, \dots, r$.

4.1 Determining the number of factors

Before investigating the properties of the estimator, we first observe the asymptotic behavior of the eigenvalues of $\mathbf{XX}'$ under the sWF model. This result yields important information for determining the number of weak factors, r.

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

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

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

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

with probability at least $1 - O((N \vee T)^{-\nu})$. In particular, λ_r diverges if the following condition is true:

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

Theorem 1 suggests the means of determining the number of weak factors. Condition (8) is needed for a technical reason. Theorem 1 presents a case in which the method of Onatski

(2010) works. Namely, for $\delta > 0$, define

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

Then, the following important corollary is obtained.

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

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

4.2 Estimation error bound

Here we introduce the condition that restricts the class of sWF models:

$$\alpha_1 + (1 \vee \tau)/2 < 3\alpha_r/2 + \tau/2. \tag{10}$$

This condition is used to derive a nontrivial error bound in the following theorems. Without this condition, it is not guaranteed that the error bound is positive and meaningful. To understand the role of this condition more precisely, see Section C and Remark 2 in Supplementary Material.

It is easy to observe that condition (10) implies both conditions (8) and (9). Thus we suppose r is known in view of Corollary 1 provided that condition (10) is true. Define

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

Theorem 2 (SOFAR). *Set $\eta_n \asymp T^{1/2}\log^{1/2}(N \vee T)$ in optimization* (6)*. If Assumptions 1–3 and condition* (10) *hold, then the following error bounds hold with probability at least $1 - O((N \vee T)^{-\nu})$:*

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

Theorem 3 (PC). *If Assumptions 1–3 and condition* (10) *hold, then the following error bounds hold with probability at least $1 - O((N \vee T)^{-\nu})$:*

$$T^{-1/2}\|\widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0\|_{\mathrm{F}} \lesssim N_1^{1/2}K_n(1+\gamma_n), \quad N_1^{-1/2}\|\widehat{\mathbf{B}}_{\mathsf{PC}} - \mathbf{B}^0\|_{\mathrm{F}} \lesssim T^{1/2}K_n(1+\gamma_n).$$

All the bounds share K_n while γ_n only appears in the PC bounds. First, when the model contains strong factors only (i.e., $N_r = N$), the convergence rates in Theorems 2 and 3 reduce to that obtained by Bai and Ng (2013) up to the logarithmic factor. In fact, $K_n = (N \vee T)^{-1}\log^{1/2}(N \vee T)$ and $\gamma_n \leq 1$ in this case. The extra (but low) cost $\log^{1/2}(N \vee T)$ is incurred in using the union bound under the subG tail assumptions, which leads to an effective treatment of the sparsity; see the proof of Theorem 2 with Lemma 1 in Supplementary Material.

We also observe that the convergence rates of the SOFAR and the PC estimators become identical if $\gamma_n = O(1)$, which occurs when $N_1 = N$, or T is larger than N, for instance. On the other hand, when the model has weak factors with $N_1 < N$ and relatively smaller T, the SOFAR can take advantage of utilizing the sparsity and achieve a sharper upper bound. Therefore, the SOFAR estimator is likely to converge at least as fast as the PC estimator even when all the factors are strong. Of course a precise discussion requires a lower bound, but this is beyond the scope of this paper and left for a future study.

Remark 1. We are interested in the class of sWF models that can consistently be estimated by the SOFAR and PC, respectively. As for the SOFAR, the lower bound of α_r is $1/3$, which is achievable when $\alpha_1 = \alpha_r$ and $\tau = 2/3$. Likewise, the upper bound of the difference $\alpha_1 - \alpha_r$ is found to be $1/4$, which is attainable when $\tau \in (3/4, 1]$ and $\alpha_1 = 1$. Contrarily, the PC restricts α_r to be strictly larger than $1/2$, though the upper bound of the difference is the same.

In sum, the SOFAR can consistently estimate the sWF models with exponents α_k smaller than 1/2. The SOFAR bound is sharp enough to allow weaker factors compared with the PC, since the SOFAR can manage the sparsity well thanks to the ℓ_1-regularization. The finite sample evidence in Section 5 shows that the SOFAR estimator seems quite robust to the violation of the restrictions on the region of $(\tau, \alpha_1, \alpha_r)$ discussed in Remark 1.

4.3 Factor selection consistency

We prove the *factor selection consistency*, which guarantees that the adaptive SOFAR recovers the true sparsity pattern of the loadings and correctly selects the relevant factors. As a corollary, we also establish the consistency of the estimated exponents, $\hat{\alpha}_k$, $k = 1, ..., r$. Before stating the theorem, define $\mathcal{S} = \text{supp}(\mathbf{B}^0) \subset \{1, \ldots, N\} \times \{1, \ldots, r\}$, the index set of non-zero signals in $\mathbf{B}^0$. For any matrix $\mathbf{A} = (a_{ik}) \in \mathbb{R}^{N \times r}$, define $\mathbf{A}_{\mathcal{S}} = (a_{ik} 1\{(i,k) \in \mathcal{S}\})$. Write $\underline{b}_n^0 = \min_{(i,k)\in\mathcal{S}} |b_{ik}^0|$.

Introduce additional conditions for the factor selection consistency:

$$\alpha_1 < \tau \wedge (4\alpha_r/3), \tag{11}$$

$$3\alpha_1/2 + \tau - 2\alpha_r < \beta \leq \alpha_1 + 3\tau/2 - 2\alpha_r, \tag{12}$$

where β is such that $\eta_n/\underline{b}_n^0 \asymp N^\beta \log^{1/2}(N \vee T)$.

Condition (11) consists of the two inequalities, $\alpha_1 < \tau$ and $\alpha_1 < 4\alpha_r/3$, which are technically important in the proof. The first condition requires sufficiently large T relative to the largest signal strength N_1. The second condition further restricts the model; it implies condition (10) if $\tau \geq 1$. Condition (12) determines the relation between η_n and $\underline{b}_n^0$. The interval is not empty when $\alpha_1 < \tau$. If $\underline{b}_n^0$ is assumed to be a (small) fixed constant, (12) becomes a condition for η_n only. Note that even decaying $\underline{b}_n^0$ is allowed by an appropriate choice of η_n.

Theorem 4 (Adaptive SOFAR)**.** *If Assumptions 1–3 and conditions* (10)–(12) *hold, then for the weighting matrix* $\mathbf{W}$ *constructed by any estimator* $\widehat{\mathbf{B}}^{\mathsf{ini}}$ *such that*

$$\mathbb{P}\left(\|\widehat{\mathbf{B}}^{\mathsf{ini}} - \mathbf{B}^0\|_{\max} \lesssim \underline{b}_n^0\right) \to 1, \tag{13}$$

the adaptive SOFAR estimator satisfies

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

$$\mathbb{P}\left(\mathrm{supp}(\widehat{\mathbf{B}}^{\mathsf{ada}})=\mathcal{S}\right)\to 1.\tag{15}$$

The rates of convergence (14) are identical to those in Theorem 2. Thus the most interesting property of the adaptive SOFAR is the factor selection consistency given by (15). If the PC estimator is used for the initial estimator, $\underline{b}_n^0 \gtrsim N_1^{1/2}K_n(1+\gamma_n)\log^{1/2}(N\vee T)$ is allowed in (13) (see Lemma 6 in Supplementary Material). This lower bound can shrink to zero in many cases. Finally, we prove that the estimated exponent $\hat{\alpha}_k=\log\widehat{N}_k/\log N$, defined in Section 3.2, is consistent for α_k because of (15).

Corollary 2. *If the model selection consistency in* (15) *holds, then we have*

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

It is well-known that the adaptive Lasso can establish the asymptotic normality for the non-zero subvector of the estimator. Likewise, the asymptotic normality of the adaptive SOFAR estimator might be proved. However, we do not consider it, due to the criticism raised by Leeb and Pötscher (e.g., Leeb and Pötscher (2008) and references therein). Instead, it is interesting to investigate inferential theory based on "debiasing" the SOFAR estimator in a manner similar to Javanmard and Montanari (2014). This direction is explored in Uematsu and Yamagata (2021).

5 Monte Carlo Experiments

In this section, we conduct thee Monte Carlo experiments. Indexes i, t, and k run over $1,\ldots,N$, $1,\ldots,T$, and $1,\ldots,r$, respectively, unless otherwise noted. We consider the Data Generating Process (DGP), $x_{ti}=\sum_{k=1}^{r}b_{ik}^0 f_{tk}^0+\sqrt{\theta}e_{ti}$. The factor loadings b_{ik}^0 and factors f_{tk}^0 are formed such that $N^{-1}\sum_{i=1}^N b_{ik}^0 b_{i\ell}^0=1\{k=\ell\}$ and $T^{-1}\sum_{t=1}^T f_{tk}^0 f_{t\ell}^0=1\{k=\ell\}$, by applying Gram–Schmidt orthonormalization to b_{ik}^* and f_{tk}^*, respectively, where $b_{ik}^*\sim$ i.i.d.$N(0,1)$ for $i=1,\ldots,N_k$ and $b_{ik}^*=0$ for $i=N_k+1,\ldots,N$, and $f_{tk}^*=\rho_{fk}f_{t-1,k}^*+v_{tk}$ with

$v_{kt} \sim$ i.i.d.$N(0, 1-\rho_{fk}^2)$ and $f_{0k}^* \sim$ i.i.d.$N(0,1)$. The idiosyncratic errors e_{ti} are generated by $e_{ti} = \rho_e e_{t-1,i} + \beta\varepsilon_{t,i-1} + \beta\varepsilon_{t,i+1} + \varepsilon_{ti}$, where $\varepsilon_{ti} \sim$ i.i.d.$N(0, \sigma_{\varepsilon,ti}^2)$ with $\sigma_{\varepsilon,ti}^2$ being set such that $\mathrm{Var}(e_{ti}) = 1$. The DGP is in line with the existing representative literature, such as Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013), among many others, but the main difference is that the absolute sums of the factor loadings over i are allowed to diverge proportionally to $N_k = N^{\alpha_k}$.

As the benchmark DGP, we set $r = 2$, $\rho_{fk} = \rho_e = 0.5$ for all k, $\beta = 0.2$, and $\theta = 1$. We focus on the performance of the estimators for different values of exponents (α_1, α_2). In particular, we consider the combinations $(0.9, 0.9)$, $(0.8, 0.5)$[7] and $(0.5, 0.4)$. All the experimental results are based on 1,000 replications.

5.1 Determining the number of weak factors

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

Table 1 reports the average of the estimated number of factors over the replications by the *ED*, *GR*, and BIC_3.[8] We set the maximum number of factors, $k_{\max}$, as five. As can be seen in Table 1, when α_1 and α_2 are both close to unity, all the methods perform well; see the case of exponents $(\alpha_1, \alpha_2) = (0.9, 0.9)$. However, the performance of *GR* and BIC_3 deteriorates when the gap of the values between α_1 and α_2 widens, or when both values α_1 and α_2 are further away from unity; e.g., see the cases when $(\alpha_1, \alpha_2) = (0.8, 0.5)$ and $(\alpha_1, \alpha_2) = (0.5, 0.4)$. In contrast, *ED* performs very well, and its estimation quality is very similar to that when both exponents are close to unity. Even under the most challenging set up $(\alpha_1, \alpha_2) = (0.5, 0.4)$, *ED* consistently estimates the number of factors for sufficiently large T and N.

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

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

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

5.2 Finite sample properties of the SOFAR estimator

Here we investigate the finite sample properties of our SOFAR estimator in comparison with the PC estimator. Here we treat the number of factors, r, as given. We report the results of the adaptive SOFAR estimator with regularization coefficient η determined by BIC, which we recommend to use.[9)] For performance comparison purposes, we consider the ℓ_2-norm losses based on the scaled estimators: $L(\widehat{\mathbf{F}}) = \| \sum_{k=1}^{r} T^{-1/2}[\mathrm{abs}(\widehat{\mathbf{f}}_k) - \mathrm{abs}(\mathbf{f}_k^0)]\|_2$, $L(\widehat{\mathbf{B}}) = \| \sum_{k=1}^{r} N_k^{-1/2}[\mathrm{abs}(\widehat{\mathbf{b}}_k) - \mathrm{abs}(\mathbf{b}_k^0)]\|_2$, and $L(\widehat{\mathbf{C}}) = \| \sum_{k=1}^{r} T^{-1/2} N_k^{-1/2}[\widehat{\mathbf{C}}_k - \mathbf{C}_k^0]\|_{\mathrm{F}}$, where $\mathrm{abs}(\mathbf{a})$ takes elementwise absolute value of a real vector $\mathbf{a}$. Due to the scaling, the performance of the estimators can be comparable across different combinations of the values of N, T, and α_k's.

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

Table 3 reports the same information as Table 2, but for more challenging models $(\alpha_1, \alpha_2) = (0.5, 0.4)$. Remarkably, even when one of the exponents is 0.4, our SOFAR method provides sufficiently accurate estimates of α_1 and α_2 as well as estimates of the factors, factor loadings, and common components that are far superior to the PC method.

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

5.3 A hierarchical factor structure

Estimation of a hierarchical factor structure or a multi-level factor structure has recently been gaining serious interest in the literature. Ando and Bai (2017) and Choi et al. (2018) consider two types of factors, called global and local factors. The global factors have loadings with non-zero values for all the cross-section units, whereas the local factors have non-zero loadings among the cross-section units of some specific groups. They propose sequential

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

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

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

	T=500		**T=1000**			**T=500**		**T=1000**	
Design (α_1, α_2)	$(0.5, 0.4)$		$(0.5, 0.4)$		Design (α_1, α_2)	$(0.5, 0.4)$		$(0.5, 0.4)$	
N=500	mean	s.d.	mean	s.d.	**N=1000**	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.47	0.03	0.47	0.03	$\hat{\alpha}_1$	0.48	0.02	0.48	0.02
$\hat{\alpha}_2$	0.41	0.04	0.40	0.04	$\hat{\alpha}_2$	0.40	0.03	0.40	0.03
	SO	PC	SO	PC		SO	PC	SO	PC
$L^2(\hat{\mathbf{F}})_{\times 100}$	13.4	17.9	13.1	15.2	$L^2(\hat{\mathbf{F}})_{\times 100}$	9.7	15.2	9.5	12.0
$L^2(\hat{\mathbf{B}})_{\times 100}$	4.6	48.3	2.9	24.4	$L^2(\hat{\mathbf{B}})_{\times 100}$	3.7	65.6	2.3	32.2
$L^2(\hat{\mathbf{C}})_{\times 100}$	17.3	48.6	16.0	31.1	$L^2(\hat{\mathbf{C}})_{\times 100}$	13.0	57.4	12.0	32.9

procedures to identify the global and local factors separately.[10)] In fact, the sWF model nests the hierarchical factor structure, and hence our SOFAR method can be readily applied. In contrast to the existing approaches, given the total number of global and local factors, our approach permits us to consistently estimate the hierarchical model in one go. Furthermore, our method can identify "near global" (or "near local") factors as the strongest, which influence many but not all the variables; see Section 6.2 for the evidence of such factors. The aforementioned existing methods may not distinguish between the near global factors and the global (or strictly strong) factors.

As an illustration, we generate the data of a four-factor model as above. The first factor is global, i.e., $b_{i1} \sim$ i.i.d.$N(0,1)$ for $i = 1, \ldots, N$. The other three factors are local, i.e., b_{i2} is drawn from $N(0,1)$ for the first third, b_{i3} for the second third, and b_{i4} for the last third of cross section units while the rest are zero. We obtained simulated data with $N = 450$ and $T = 120$, and estimated the model given $r = 4$ by the PC and SOFAR. To visualize the quality of the factor loadings, we provide heat maps of three $N \times N$ matrices, $\sum_{k=1}^{4} \omega_k \mathbf{b}_k^0 \mathbf{b}_k^{0\prime}$, $\sum_{k=1}^{4} \omega_k \widehat{\mathbf{b}}_k \widehat{\mathbf{b}}_k'$ and $\sum_{k=1}^{4} \omega_k \widehat{\mathbf{b}}_{\mathsf{PC},k} \widehat{\mathbf{b}}_{\mathsf{PC},k}'$, which are reported in Figures 3-5, respectively. To clarify the difference between the global factor loadings and local ones, which overlap in the heat maps, we use the weight $\omega_1 = 1/8$ and $\omega_2 = \omega_3 = \omega_4 = 1$. As is clear, the SOFAR estimator successfully recovers the hierarchical pattern, while the PC estimator fails.

6 Empirical Applications

Here we provide four empirical applications. Section 6.1 compares the out-of-sample forecast performance of the predictive regressions with the SOFAR and the PC factors. Section 6.2

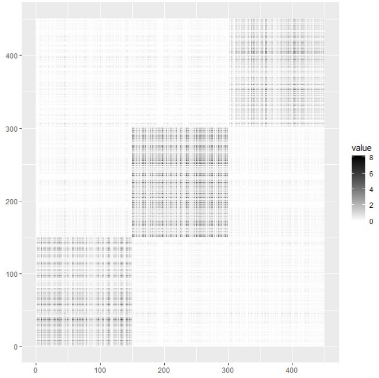

Figure 3: True factor loadings

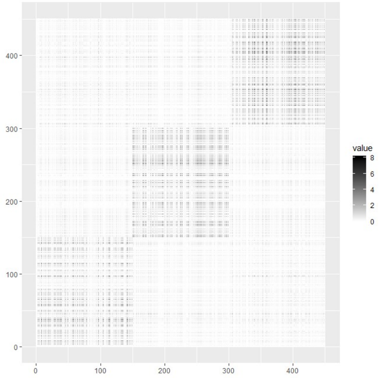

Figure 4: SOFAR estimate

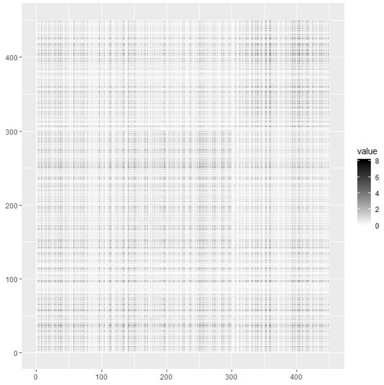

Figure 5: PC estimate

investigates the properties of the extracted SOFAR factors in Section 6.1 by looking into the sparsity pattern and the signs of the loadings. In section 6.3 we search the residuals of the Fama and French (2015) five-factor regressions for omitted weak factors. Section 6.4 considers estimation of the exponents using a large number of excess stock returns.

6.1 Forecasting bond yields

We consider out-of-sample forecasting of bond yields using extracted factors via our SOFAR and the PC, from a large number of macroeconomic variables in line with Ludvigson and Ng (2009). We use the same data set provided by Sydney Ludvigson's web page. 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 $i = 1, \dots, 131$ monthly macroeconomic series at month t, x_{ti}, spanning the period from January 1964 to December 2003. We consider the maturities $n = 2, 3, 4, 5$. For more details, see Section 3 of Ludvigson and Ng (2009).

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

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

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

In Table 4, we report the mean absolute deviation of the forecast errors, $MAE^{(n)} = H^{-1} \sum_{s=1}^{H} \left| \hat{\epsilon}_{s|s-12}^{(n)} \right|$ and the mean squared forecast errors, $MSE^{(n)} = H^{-1} \sum_{s=1}^{H} (\hat{\epsilon}_{s|s-12}^{(n)})^2$, $n = 2, 3, 4, 5$, for our estimation method (WF-SF) and principal component method (PC), and Diebold-Mariano forecasting performance test statistics with associated p-values, based on the MAEs and the MSEs. As can be seen, the MAEs and the MSEs of WF-SF are smaller than those of the PC for all the maturities. For all the maturities, the Diebold-Mariano forecasting performance test strongly rejects the null of the same forecasting performance, in favor of the alternative that our method outperforms the PC method. The average values of alphas 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 a strictly strong factor. As is evidenced in the previous section, the accuracy of our estimator is much higher than the PC estimator under such situations, and the better forecasting performance may not be too surprising in this empirical exercise.

Table 4: Mean absolute forecast errors, mean squared forecast errors, and Diebold-Mariano (DM) forecast comparison test results

	Mean Absolute Forecast Errors			Mean Squared Forecast Errors		
	SO	*PC*	DM statistic [p-value]	*SO*	*PC*	DM statistic [p-value]
$y_{t+12}^{(2)}$	1.16	1.19	-3.58 [0.000]	2.31	2.42	-4.14 [0.000]
$y_{t+12}^{(3)}$	2.30	2.35	-3.54 [0.000]	9.16	9.53	-4.13 [0.000]
$y_{t+12}^{(4)}$	3.35	3.43	-3.73 [0.000]	19.57	20.34	-4.33 [0.000]
$y_{t+12}^{(5)}$	4.20	4.28	-3.20 [0.001]	30.54	31.66	-4.08 [0.000]

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

6.2 Interpreting the factors by analyzing the sparse loadings

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

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

As an illustration, we investigate a set of extracted factors from the 131 macroeconomic variables used in Section 6.1 in more detail. In particular, we estimate the model using the variables between January 1982 and December 2001. Two factors (i.e., $\widehat{r} = 2$) are extracted by the PC and SOFAR methods (adaptive, BIC). The exponents are $\{\hat{\alpha}_1, \hat{\alpha}_2\} = \{0.91, 0.71\}$.

Figures 6 and 8 display the R^2s of the regressions of the 131 individual time series on the first PC factor and the first SOFAR factor over the period, respectively. These R^2s are plotted as bar charts, and the variables are ordered as described in the aforementioned data file. As can be seen from the figures, the patterns of R^2s of the PC and SOFAR factors are very similar. The variables 70–83 and 101–131, except 78 and 113, have little association in terms of R^2. The striking difference in the PC and SOFAR results is found in their loading estimates, which are reported in Figures 7 and 9, respectively. The SOFAR loadings associated with the near-zero R^2s are (rightly) zeros, whereas the magnitude of the corresponding PC loadings are not as small as R^2s. These contrasts in the PC and SOFAR estimation results are more pronounced for the second factor. The PC results are as reported in Figures 1 and 2 in Section 1.2, and the SOFAR results are reported in Figures 10 and 11.

In summary, the SOFAR loadings contain sharper information on which variables are associated with which factor than is the case with the PC. Among the variables with non-zero loadings, the value of the SOFAR loadings can provide information on the strength and direction of the influence of the factor, relative to the other variables.

With this encouraging result, we investigate the properties of each empirical factor, making use of the information contained in the SOFAR loadings. Based on the description of each of the 131 variables in the aforementioned data file, we categorize the 131 variables as follows: 1-24 *Output*; 25-32 *Unemployment*; 33-49 *Employment*; 50-59 *Housing*; 60-69 *Orders*; 70-76 *Money Supply*; 77-80 *Credits*; 81-84 *Stock Prices*; 85-93 *Interest Rate*; 94-101 *Spreads*; 102-106 *Exchange Rates*; 107-127 *Prices*; 128-130 *Wages*; 131 *Consumer Expectation.* From Figure 9 it is easily seen that the first SOFAR factor is almost *exclusively* loaded on *Output*, *Unemployment*, *Employment*, *Housing*, *Orders*, *Interest Rates*, and *Spreads*, with a few exceptions only. Observe that the signs of the loadings on the unemployment variables are different from those on the employment variables, as expected. Figure 11 reveals that the second SOFAR factor is exclusively loaded on *Money Supply*, *Exchange Rates*, and *Prices*, with scarce exceptions.

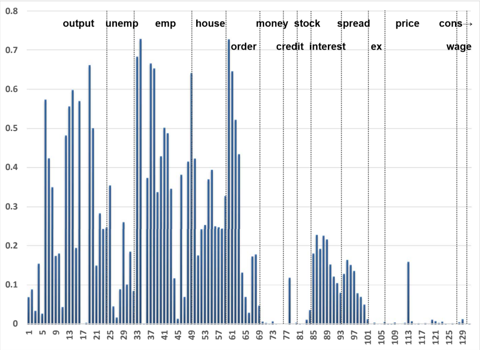

Figure 6: R^2 for regression of x_{it} on $\hat{f}_{t1}^{\mathsf{PC}}$

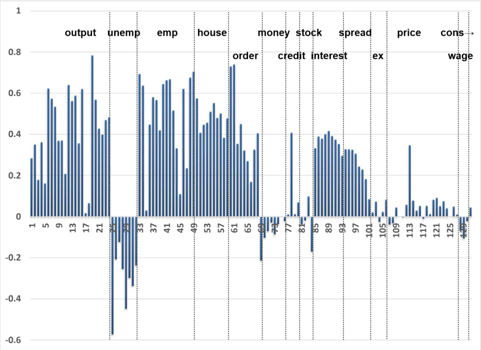

Figure 7: PC estimates $\hat{b}_{i1}^{\mathsf{PC}}$

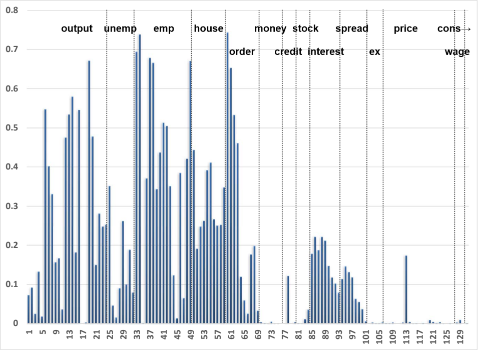

Figure 8: R^2 for regression of x_{it} on $\hat{f}_{t1}^{\text{ada}}$

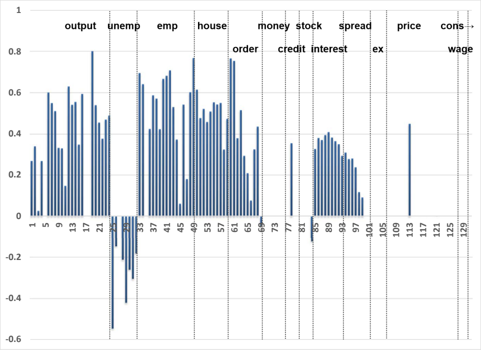

Figure 9: adaptive SOFAR estimates $\hat{b}_{i1}^{\text{ada}}$

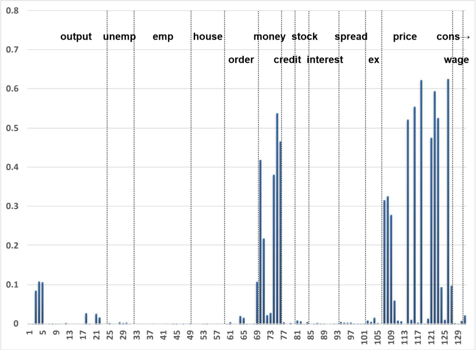

Figure 10: R^2 for regression of x_{it} on $\hat{f}_{t2}^{\text{ada}}$

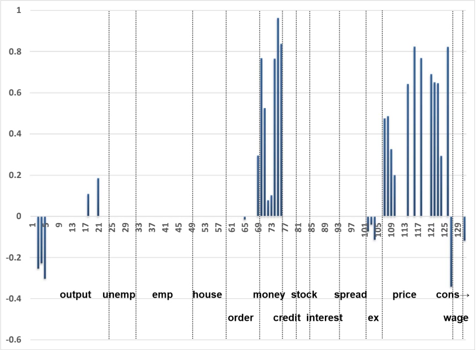

Figure 11: adaptive SOFAR estimates $\hat{b}_{i2}^{\text{ada}}$

6.3 Are there any omitted weak factors in the Fama-French five factor regression model?

In this subsection, we examine the regression residuals of the celebrated Fama and French (2015) five-factor model in order to check whether any weak common factors have been left-out. We consider monthly security excess returns, which constitute the S&P500 index on April 2018, with 500 months os observations back, leaving 194 securities. The firm security return is computed as explained in Section 6.4, 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. Specifically, we run the time series regression $r_{ti} - r_{ft} = a_i + b_i\left(r_{mt} - r_{ft}\right) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{ti}$, 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 13 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.

We have applied our method to the obtained residual, $\hat{e}_{ti}$, say. The ED estimates that there is one factor, and the adaptive SOFAR picks up eight non-zero factor loadings ($\hat{\alpha} = 0.39$), all of which have the same sign. Interestingly, all the associated firms belong to the same industrial category, Technology Hardware & Equipment: Advanced Micro Devices, Analog Devices, Applied Materials, Intel, Texas Instruments, Western Digital, Skyworks Solutions and Xerox Holdings. This result could not be found using the conventional PC method.

6.4 Estimating exponents with stock returns

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

For each window month, $\mathtt{T}$ = September 1998 to April 2018, we chose securities that contain the data extending 120 months back ($T = 120$) from $\mathtt{T}$. This gives the different number of securities for each window $\mathtt{T}$ ($N_{\mathtt{T}}$). The average number of securities over the estimation windows is 443 ($\bar{N} = 443$). As will be shown below, three or four factors are estimated over the windows. We identify the factors and signs of the factors and factor loadings, given the estimates of the initial window month, $\mathtt{T}$ = September 1989, based on the correlation coefficients between the factors at $\mathtt{T}$ and the appropriately lagged $\mathtt{T}$.[12)]

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

In line with the well-observed phenomenon that the correlation among the securities in the financial market rises during periods of turmoil, sharp rises of exponents in some months can be observed. For example, α_2 goes up sharply around February 2000 then rises gradually. This period corresponds to the peak of the dot-com bubble and its burst on March 2000 (the main contributor to the factor loadings of the second factor is the Technology industry, see Section F.1 in Supplementary Material). 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 Section F.1 in Supplementary Material).

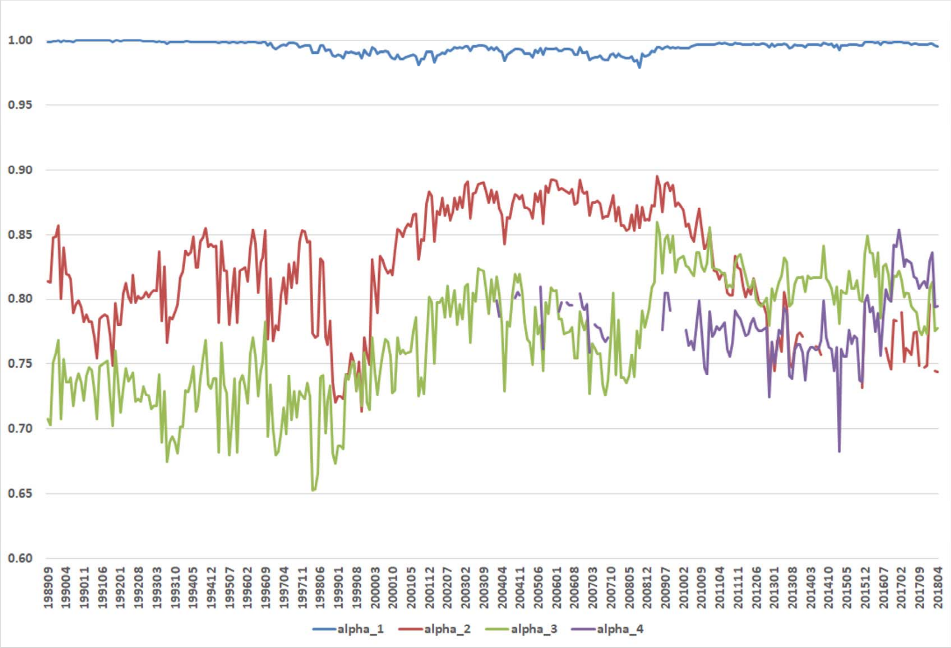

Figure 12: Plot of the estimated α_k from September 1989 to April 2018.

7 Conclusion

This paper has considered estimation of the sparsity-induced weak factor (sWF) models in a high-dimensional setting. We suppose sparse factor loadings $\mathbf{B}^0$ that lead to the WF structure, $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N^{\alpha_k}$ with $0 < \alpha_k \leq 1$ for $k = 1, \ldots, r$. This model is much less restrictive than the widely employed strong factor (SF) model in the literature, in which $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) \asymp N$ for $k = 1, \ldots, r$. The SOFAR estimator and its adaptive version enable us to consistently estimate the sWF models, separately identifying $\mathbf{B}^0$ and $\mathbf{F}^0$. As theoretical contributions, we have established the estimation error bound of the SOFAR estimators, the factor selection consistency of the adaptive SOFAR estimator, and consistent estimation of each exponent α_k, as well as validating the method of Onatski (2010) for determining the number of weak factors. All the theoretical results are supported by the Monte Carlo experiments, and four empirical examples demonstrate the practical usefulness of our estimator in comparison to the principal components (PC) estimator.

The proposed method has large potential applicability and many direction to extend. The hierarchical factor model, which contains global and local factors, are recently considered by Ando and Bai (2017), Choi et al. (2018) and Andreou et al. (2019). Our sWF model nests the hierarchical factor model, and hence the SOFAR method can be applied to readily estimate such models. It is of interest to estimate the stock returns covariance matrix for optimal portfolio allocation and portfolio risk assessment. This can be achieved by consistently estimating the covariance matrix of idiosyncratic errors, in line with Fan et al. (2008) and Fan et al. (2011), which is an interesting extension of this paper. Having provided the consistent estimation in this paper, the statistical inference for the sWF models becomes an important research agenda. This is considered in Uematsu and Yamagata (2021). Yet another possible extension of interest is the estimation of panel data models with interactive effects, which is considered by Pesaran (2006) and Bai (2009), among others: $y_{ti} = \mathbf{x}_{ti}'\boldsymbol{\beta} + u_{ti}$, $u_{ti} = \mathbf{f}_t'\mathbf{b}_i + \varepsilon_{ti}$. For the PC based estimators, such as Bai (2009), u_{ti} is typically assumed to be a SF model and estimated by PC, given an initial estimator of $\boldsymbol{\beta}$. The SOFAR estimation, instead of the PC, would potentially improve the precision of the estimates of $\boldsymbol{\beta}$.

Notes

[1] In the finance literature, see also Kleibergen (2009).

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

[3] As an illustration, when $\mathbf{b}_k^0$ is not sparse and composed of non-zero values of order $N^{(\alpha_k-1)/2}$, it is easy to see that $\lambda_k(\mathbf{B}^{0\prime}\mathbf{B}^0) = \mathbf{b}_k^{0\prime}\mathbf{b}_k^0$ diverges proportionally to N^{α_k}. Connor and Korajczyk (2019) consider such a structure with observed factors.

[4] https://www.sydneyludvigson.com/data-and-appendixes

[5] Bailey et al. (2021) mainly consider the estimation and inference of the divergence rates of structural loadings $\mathbf{B}^*$ in (1) when $\mathbf{F}^*$ is observed.

[6] Very recently, Freyaldenhoven (2021) has proposed a method to estimate the number of common components, the divergence rates of which exceed a preassigned threshold in the sWF model.

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

[8] To save 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.

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

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

[11] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[12] For example, define $(T-1)$-dimensional vector of ℓth factor of $\mathtt{T}$ as $\widehat{\mathbf{f}}_{\ell\mathtt{T}} = (\hat{f}_{\ell\mathtt{T},1}, \hat{f}_{\ell\mathtt{T},2}, \ldots, \hat{f}_{\ell\mathtt{T},\mathtt{T}-1})'$ and that of $\mathtt{T}-\mathtt{1}$ as $\widehat{\mathbf{f}}_{\ell\mathtt{T}-1} = (\hat{f}_{\ell\mathtt{T}-1,2}, \hat{f}_{\ell\mathtt{T}-1,2}, \ldots, \hat{f}_{\ell\mathtt{T}-1,\mathtt{T}})'$, $\ell = 1, \ldots, r$. For $\widehat{\mathbf{f}}_{\ell\mathtt{T}}$, if $\max_{1\le k\le r}|\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{k\mathtt{T}-1})| = |\mathrm{corr}(\widehat{\mathbf{f}}_{\ell\mathtt{T}}, \widehat{\mathbf{f}}_{2\mathtt{T}-1})|$ 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 $\widehat{\mathbf{b}}_{i2\mathtt{T}} \equiv -\widehat{\mathbf{b}}_{i\ell\mathtt{T}}$.

Supplementary Materials

Supplementary Material consists of six sections. Section A contains the proofs of the main results. Section B contains some lemmas and their proofs. Section C contains some details on derivation of the estimation error bound, which relates to the proof of Theorems 2 and 3. Section D contains a brief review of SOFAR. Section E contains additional experimental results. Section F contains additional estimation results.

Acknowledgments

The authors thank Christian Hansen (the editor), the associate editor, and two anonymous referees for their valuable comments and suggestions. We appreciate Kun Chen giving helpful suggestions and modification of the R package, `rrpack`. We are grateful to Greg Connor, Jianqing Fan, Yingying Fan, Simon Freyaldenhoven, Kazuhiko Hayakawa, Oliver Linton, George Kapetanios, Robert Korajczyk, Jinchi Lv, Alexey Onatskiy, Hashem Pesaran and participants to NBER-NFS times-series conference 2019 for their helpful comments on the earlier version of this paper, entitled "Estimation of weak factor models."

Funding

This work was supported by JSPS KAKENHI Grant Numbers 18K01545, 19K13665, 20H01484, and 20H05631.

References

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

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

Andreou, E., P. Gagliardini, E. Ghysels, and M. Rubin (2019). Inference in group factor models with an application to mixed frequency data. *Econometrica 87*, 1267–1305.

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

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

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

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

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

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

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

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

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

Bailey, N., G. Kapetanios, and M. H. Pesaran (2021). Measurement of factor strength: Theory and practice. *Journal of Applied Econometrics 36*, 587–613.

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

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

Chen, L. and W. B. Wu (2018). Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research 18*, 1–46.

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

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

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

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

Connor, G. and R. A. Korajczyk (2019). Semi-strong factors in asset returns. *Economics Department Working Paper Series, National University of Ireland - Maynooth*.

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

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

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

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

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

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

Freyaldenhoven, S. (2021). Factor models with local factors - determining the number of relevant factors. *Journal of Econometrics, forthcoming*.

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

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

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*, 187–200.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics 149*(2), 149–173.

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

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

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

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

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

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

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.

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

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

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

Stock, J. H. and M. W. Watson (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, pp. 415–525. Northholland.

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

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

Uematsu, Y. and T. Yamagata (2021). Inference in sparsity-induced weak factor models. *Available at SSRN: https://ssrn.com/abstract=3556275*.

Vershynin, R. (2018). In *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press.

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

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

Supplementary Material for

Estimation of Sparsity-Induced Weak Factor Models

YOSHIMASA UEMATSU* and TAKASHI YAMAGATA†

Department of Economics and Management, Tohoku University

†*Department of Economics and Related Studies, University of York*

†*Institute of Social Economic Research, Osaka University*

A Proofs of the Main Results

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

A.1 Proof of Theorem 1

Proof. Following Ahn and Horenstein (2013), we evaluate the eigenvalues of $\mathbf{X}\mathbf{X}'$. Recall notation of the SVD, $\mathbf{C}^0 = \mathbf{U}^0\mathbf{D}\mathbf{V}^{0\prime}$. 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 $\mathbf{X}\mathbf{X}' = \mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime} + \mathbf{E}\mathbf{Q}\mathbf{E}'$ since $\mathbf{V}^{0\prime}\mathbf{V}^0 = \mathbf{N} = \mathrm{diag}(N_1, \ldots, N_r)$ by the definition. We also define $\mathbf{W}_{1:k}$ as the matrix of k eigenvectors corresponding to the first k largest eigenvalues of $\mathbf{U}^*\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{*\prime}$.

We first evaluate the r largest eigenvalues of $\mathbf{XX}'$. Because

$$\lambda_k(\mathbf{U}^0\mathbf{D}^0\mathbf{N}\mathbf{D}^0\mathbf{U}^{0\prime}) = d_k^2 N_k T \quad \text{for} \quad k \in \{1, \ldots, r\}, \tag{A.1}$$

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.1)–(A.3) lead to the desired result under condition (8).

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

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

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

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

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

Next, we verify (A.3). By the construction of $\mathbf{U}^*$, the upper bound is

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

where the last inequality holds by Lemma 2 with probability at least $1 - O((N\vee T)^{-\nu})$. Similarly, the lower bound is

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

Hence, these two inequalities imply (A.3).

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 2 entails

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

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

A.2 Proof of Theorem 2

Proof. The optimality of the SOFAR estimator implies

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

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

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

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

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

Bound the three trace terms in (A.4). By applying Hölder's inequality and the property of norms, they are bounded as

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

where $\|\cdot\|_*$ denotes the nuclear (Schatten-1) norm. From these inequalities, the upper bound of (A.4) becomes

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

From Lemma 1, there exist some positive constants c_1–c_3 such that the following event occurs with probability at least $1 - O((N \vee T)^{-\nu})$ for any fixed (large) constant $\nu > 0$:

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

Set the regularization parameter to be $\eta_n = 2c_3 T^{1/2} \log^{1/2}(N \vee T)$. Then on event $\mathcal{E}$, we have $\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max} \le \eta_n/2$, and thus (A.5) is further bounded as

$$
\begin{aligned}
\|\mathbf{\Delta}\|_{\mathrm{F}}^2 \lesssim (N_1 T)^{1/2} \log^{1/2}(N \vee T) \|\mathbf{\Delta}^f\|_{\mathrm{F}} + (N \vee T)^{1/2} \|\mathbf{\Delta}^f\|_{\mathrm{F}} \|\mathbf{\Delta}^b\|_{\mathrm{F}} \\
+ \eta_n \left(\|\mathbf{\Delta}^b\|_1 + 2\|\mathbf{B}^0\|_1 - 2\|\widehat{\mathbf{B}}\|_1 \right). \qquad (A.6)
\end{aligned}
$$

Define index set $\mathcal{S} = \{(i,k) : b^0_{ik} \neq 0\}$, 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

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

Meanwhile, Lemma 3 establishes the lower bound of (A.6). Consequently, combining these upper and lower bounds yields

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

Using (A.7), we can derive the upper bound

$$\begin{aligned}
\|\boldsymbol{\Delta}^f\|_{\mathrm{F}} + \|\boldsymbol{\Delta}^b\|_{\mathrm{F}} &\lesssim \left(\frac{T^{1/2}}{N_r^{1/2}} + \frac{N_r^{1/2}}{T^{1/2}}\right)\frac{N_1^{3/2}}{N_r^{3/2}}\log^{1/2}(N \vee T) \\
&\asymp \frac{N_1^{3/2}T^{1/2}}{N_r(N_r \wedge T)}\log^{1/2}(N \vee T).
\end{aligned}$$

(See Section C for the derivation.) This completes the proof. □

A.3 Proof of Theorem 3

Proof. Following the proof of Theorem 2, we derive the bound. From (A.5) with putting $\eta_n = 0$ and using the bound $\|\boldsymbol{\Delta}^b_{\mathsf{PC}}\|_1 \leq r^{1/2}N^{1/2}\|\boldsymbol{\Delta}^b_{\mathsf{PC}}\|_{\mathrm{F}}$, we have

$$\begin{aligned}
(1/2)\|\boldsymbol{\Delta}_{\mathsf{PC}}\|_{\mathrm{F}}^2 &\leq r^{1/2}T^{1/2}\|\mathbf{E}\mathbf{B}^0\|_{\max}\|\boldsymbol{\Delta}^f_{\mathsf{PC}}\|_{\mathrm{F}} \\
&\quad + \|\mathbf{E}\|_2\|\boldsymbol{\Delta}^b_{\mathsf{PC}}\|_{\mathrm{F}}\|\boldsymbol{\Delta}^f_{\mathsf{PC}}\|_{\mathrm{F}} + r^{1/2}N^{1/2}\|\boldsymbol{\Delta}^b_{\mathsf{PC}}\|_{\mathrm{F}}\|\mathbf{F}^{0\prime}\mathbf{E}\|_{\max}.
\end{aligned}$$

In the same way as the proof of Theorem 2, we obtain

$$\frac{N_r^2}{N_1}\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}^2+\frac{TN_r}{N_1}\|\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}^2\lesssim(N_1T)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}$$
$$+(N\vee T)^{1/2}\|\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}+(NT)^{1/2}\log^{1/2}(N\vee T)\|\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}.$$

From this, we can derive the upper bound

$$\begin{aligned}\|\boldsymbol{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}+\|\boldsymbol{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}&\lesssim\left\{\frac{T^{1/2}}{N_r^{1/2}}+\frac{N^{1/2}}{N_1^{1/2}}\left(1+\frac{N_r^{1/2}}{T^{1/2}}\right)\right\}\frac{N_1^{3/2}}{N_r^{3/2}}\log^{1/2}(N\vee T)\\&\asymp\frac{N_1^{3/2}T^{1/2}}{N_r(N_r\wedge T)}(1+\gamma_n)\log^{1/2}(N\vee T),\end{aligned}$$

where $\gamma_n=N^{1/2}(N_r\wedge T)^{1/2}/(N_1T)^{1/2}$. (See Section C for the derivation.) This completes the proof. □

A.4 Proof of Theorem 4

Proof. Throughout this proof, we omit the superscript of the adaptive estimators $(\widehat{\mathbf{F}}^{\mathsf{ada}},\widehat{\mathbf{B}}^{\mathsf{ada}})$ and simply write them as $(\widehat{\mathbf{F}},\widehat{\mathbf{B}})$. Recall $\mathcal{S}=\mathrm{supp}(\mathbf{B}^0)$, which is a subset of $\{1,\dots,N\}\times\{1,\dots,r\}$. 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 $\mathbf{B}_{\mathcal{S}}^0=\mathbf{B}^0$ and $\mathbf{B}_{\mathcal{S}^c}^0=\mathbf{0}$. Recall that the objective function for obtaining the adaptive SOFAR estimator is given by

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

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

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

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

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

$$r_n = \frac{N_1(N_1T)^{1/2}\log^{1/2}(N \vee T)}{N_r(N_r \wedge T)}.$$

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

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

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

To show (A.9), we first have

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

By Hölder's inequality, we bound (I) as

$$\begin{aligned}
|(I)| &\le r_n\left|\operatorname{tr}\mathbf{V}'_{\mathcal{S}}\mathbf{E}'\mathbf{F}^0\right| + r_n\left|\operatorname{tr}\mathbf{B}^{0\,\prime}_{\mathcal{S}}\mathbf{E}'\mathbf{U}\right| + r_n^2\left|\operatorname{tr}\mathbf{V}'_{\mathcal{S}}\mathbf{E}'\mathbf{U}\right| \\
&\lesssim r_n\|\mathbf{V}_{\mathcal{S}}\|_1\|\mathbf{E}'\mathbf{F}^0\|_{\max} + r_n\|\mathbf{B}^{0\,\prime}_{\mathcal{S}}\mathbf{E}'\|_{\max}\|\mathbf{U}\|_1 + r_n^2\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\|\mathbf{E}_{\mathcal{S}}\|_2\|\mathbf{U}\|_{\mathrm{F}} \\
&\lesssim r_nN_1^{1/2}T^{1/2}\left(\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}} + \|\mathbf{U}\|_{\mathrm{F}}\right)\ell_n^{1/2} + r_n^2(N_1 \vee T)^{1/2}\|\mathbf{U}\|_{\mathrm{F}}\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}\ell_n^{1/2}.
\end{aligned}$$

Next, we bound (II) from below as

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

In view of the Rayleigh quotient, (IIa) is further bounded from below as

$$\begin{aligned}
(IIa) &= 2^{-1}r_n^2\left(\operatorname{tr}\mathbf{B}_{\mathcal{S}}^{0\prime}\mathbf{B}_{\mathcal{S}}^0\mathbf{U}'\mathbf{U} + \operatorname{tr}\mathbf{F}^{0\prime}\mathbf{F}^0\mathbf{V}_{\mathcal{S}}'\mathbf{V}_{\mathcal{S}} + r_n^2\operatorname{tr}\mathbf{U}'\mathbf{U}\mathbf{V}_{\mathcal{S}}'\mathbf{V}_{\mathcal{S}}\right) \\
&\geq 2^{-1}r_n^2\left(\lambda_{\min}(\mathbf{B}_{\mathcal{S}}^{0\prime}\mathbf{B}_{\mathcal{S}}^0)\|\mathbf{U}\|_{\mathrm{F}}^2 + \lambda_{\min}(\mathbf{F}^{0\prime}\mathbf{F}^0)\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right) \\
&\gtrsim r_n^2\left(N_r\|\mathbf{U}\|_{\mathrm{F}}^2 + T\|\mathbf{V}_{\mathcal{S}}\|_{\mathrm{F}}^2\right).
\end{aligned}$$

Meanwhile, $|(IIb)|$ is bounded from above as

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

Combining (IIa) and (IIb) yields

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

We then consider (III) in (A.10). Lemma 4 yields

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

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

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

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

Under conditions $\alpha_1 < \tau$ and $3\alpha_1 < 4\alpha_r$ implied by (11), a simple calculation reveals that $(i)/|(ii)| \to \infty$ and $(i)/|(iii)| \to \infty$. Furthermore, by the upper bound of condition (12), we find that $(i) \asymp |(ii)|$ and hence $(i) + (iv)$ tends to positive if $C > 0$ is taken to be large enough. In consequence, the lower bound (A.11) is positive for such $C > 0$ and (A.9) holds.

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

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

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

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

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

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

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

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

where $\widehat{\mathbf{F}}'\widehat{\mathbf{F}} = T\mathbf{I}_r$ has been used, and $\mathbf{W}^{-} \in \mathbb{R}^{N\times r}$ is the matrix with its (i,k)th elements given by $1/w_{ik}$. Since $(\widehat{\mathbf{F}}, \widehat{\mathbf{B}}_{\mathcal{S}})$ is a minimizer of $Q_n(\mathbf{F}, \mathbf{B}_{\mathcal{S}})$, it satisfies the Karush–Kuhn–Tucker (KKT) conditions. Therefore, we only need to check condition (A.14), which is verified by Lemma 5. This completes the proof of Theorem 4. □

A.5 Proof of Corollary 2

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

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

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

B Related Lemmas and the Proofs

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

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

(b) $\|\mathbf{E}\mathbf{B}^0\|_{\max} \lesssim \|\mathbf{b}_1^0\|_2 \log^{1/2}(N \vee T)$,

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

Proof. Throughout the proof, set $L_n = N \vee T$. Prove (a). The tth row of $\mathbf{E}$, $\mathbf{e}_t' \in \mathbb{R}^N$, is specified as $\mathbf{e}_t = \sum_{\ell=0}^{\infty} \mathbf{\Phi}_\ell \boldsymbol{\varepsilon}_{t-\ell}$, 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 $\mathbf{E} = (\sum_{\ell=0}^{L_n-1} + \sum_{\ell=L_n}^{\infty})\widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell'$ and hence,

$$\|\mathbf{E}\|_2 \le \left\| \sum_{\ell=0}^{L_n-1} \widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell' \right\|_2 + \left\| \sum_{\ell=L_n}^{\infty} \widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell' \right\|_2 .$$

Consider the first term. By the submultiplicativity of the spectral norm, we have

$$\left\| \sum_{\ell=0}^{L_n-1} \widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell' \right\|_2 \le \sum_{\ell=0}^{L_n-1} \|\widetilde{\mathbf{E}}_\ell\|_2 \|\mathbf{\Phi}_\ell\|_2 \le \max_{\ell\in\{0,\ldots,L_n-1\}} \|\widetilde{\mathbf{E}}_\ell\|_2 \sum_{\ell=0}^{L_n-1} \|\mathbf{\Phi}_\ell\|_2.$$

By Assumption 3, the last sum is bounded as

$$\sum_{\ell=0}^{L_n-1} \|\mathbf{\Phi}_\ell\|_2 \le \sum_{\ell=0}^{\infty} \|\mathbf{\Phi}_\ell\|_2 \le C_e \ell_e + C_e \sum_{\ell=\ell_e}^{\infty} \ell^{-(\nu+2)} < \infty.$$

Because of the union bound and the inequality for the largest singular value of a sub-Gaussian matrix (Vershynin, 2018, Theorem 4.4.5), there are positive constants M and C_1 such that

$$\begin{aligned}
&\mathbb{P}\left(\max_{\ell\in\{0,\ldots,L_n-1\}} \left\| (N\vee T)^{-1/2} \widetilde{\mathbf{E}}_\ell \right\|_2 > M \right) \\
&\le L_n \max_{\ell\in\{0,\ldots,L_n-1\}} \mathbb{P}\left(\left\| (N\vee T)^{-1/2} \widetilde{\mathbf{E}}_\ell \right\|_2 > M \right) \\
&\le 2(N\vee T) \exp\{-C_1(N\vee T)\} \lesssim (N\vee T)^{-\nu}.
\end{aligned}$$

Thus we conclude $\left\| \sum_{\ell=0}^{L_n-1} \widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell' \right\|_2 \lesssim (N\vee T)^{1/2}$ with probability at least $1 - O((N\vee T)^{-\nu})$.

Next show that with the same order of the probability the second term has a smaller upper bound than the first term. By the Markov inequality and the submultiplicativity of the spectral norm, it holds that for all $x \ge 0$,

$$\mathbb{P}\left(\left\| \sum_{\ell=L_n}^{\infty} \widetilde{\mathbf{E}}_\ell \mathbf{\Phi}_\ell' \right\|_2 > x \right) \le \sum_{\ell=L_n}^{\infty} \|\mathbf{\Phi}_\ell\|_2 \, \mathbb{E} \, \|\widetilde{\mathbf{E}}_\ell\|_2 / x.$$

By Vershynin (2018, Chapter 4.4.2), we have $\max_\ell \mathbb{E}\, \|\widetilde{\mathbf{E}}_\ell\|_2 \lesssim (N\vee T)^{1/2}$. Assumption 3 with a standard estimate of remainder gives

$$\sum_{\ell=L_n}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2 \lesssim \sum_{\ell=L_n}^{\infty} \ell^{-(\nu+2)} \le \int_{L_n-1}^{\infty} x^{-(\nu+2)} dx \lesssim (L_n-1)^{-(\nu+1)} \asymp (N \vee T)^{-(\nu+1)}.$$

Thus by setting x to be a positive large constant, the second term is bounded by a positive constant with probability at least $1 - O((N \vee T)^{-\nu})$. This completes the proof of (a).

Prove (b). By the definition, for $t \in \{1, \dots, T\}$ and $k \in \{1, \dots, r\}$, the (t,k)th element of $\mathbf{E}\mathbf{B}^0$ is given by $\mathbf{e}_t'\mathbf{b}_k^0 = \sum_{\ell=0}^{\infty} \boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0$. As the proof of (a), the summation is decomposed into two parts, $\sum_{\ell=0}^{\infty} = \sum_{\ell=0}^{L_n-1} + \sum_{\ell=L_n}^{\infty}$.

Consider the first summation. Let $\mathbf{a}_{k\ell} = \|\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\|_2^{-1}\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0$. We have

$$\begin{aligned}\max_{t,k}\left|\sum_{\ell=0}^{L_n-1} \boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right| &\le \max_{t,k}\sum_{\ell=0}^{L_n-1} \left|\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{a}_{k\ell}\right| \|\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\|_2 \\ &\le \max_{t,k}\max_{\ell\in\{0,\dots,L_n-1\}} \left|\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{a}_{k\ell}\right| \|\mathbf{b}_k^0\|_2 \sum_{\ell=0}^{L_n-1}\|\boldsymbol{\Phi}_\ell\|_2,\end{aligned}$$

where $\sum_{\ell=0}^{\infty}\|\boldsymbol{\Phi}_\ell\|_2 < \infty$ by Assumption 3 and the same argument as in the proof of (a). Note that $\|\mathbf{a}_{k\ell}\|_2 = 1$ for all k and ℓ. Therefore, by the union bound and the general Hoeffding inequality (Vershynin, 2018, Theorem 2.6.3), there exists a constant $C_2 > 0$ such that

$$\mathbb{P}\left(\max_{t,k,\ell}\left|\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{a}_{k\ell}\right| > x\right) \le rTL_n \max_{t,k,\ell}\mathbb{P}\left(\left|\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{a}_{k\ell}\right| > x\right) \lesssim (N \vee T)^2 \exp\left(-x^2/C_2\right).$$

Setting $x = C_2^{1/2}(\nu+2)^{1/2}\log^{1/2}(N \vee T)$ yields the bound

$$\max_{t,k,\ell}\left|\boldsymbol{\varepsilon}_{t-\ell}'\mathbf{a}_{k\ell}\right| \le C_2^{1/2}(\nu+2)^{1/2}\log^{1/2}(N \vee T),$$

which holds with probability at least $1 - O((N \vee T)^{-\nu})$. With this probability, combining the obtained inequalities gives the bound for the first summation

$$\max_{t,k}\left|\sum_{\ell=0}^{L_n-1} \boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right| \lesssim \max_k \|\mathbf{b}_k^0\|_2 \log^{1/2}(N \vee T).$$

Next consider the second summation in the same way as the proof of (a). The Markov inequality and Hölder's inequality entail

$$\mathbb{P}\left(\max_{t,k}\left|\sum_{\ell=L_n}^{\infty}\boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right|>x\right)\leq\mathbb{E}\max_{t,k}\left|\sum_{\ell=L_n}^{\infty}\boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right|/x$$
$$\leq\sum_{\ell=L_n}^{\infty}\mathbb{E}\max_t\|\boldsymbol{\varepsilon}_{t-\ell}\|_\infty\max_k\left\|\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right\|_1/x$$
$$\lesssim\mathbb{E}\max_t\|\boldsymbol{\varepsilon}_t\|_\infty\max_k\left\|\mathbf{b}_k^0\right\|_1\sum_{\ell=L_n}^{\infty}\left\|\boldsymbol{\Phi}_\ell'\right\|_1/x.$$

The sub-Gaussian property implies $\mathbb{E}\max_t\|\boldsymbol{\varepsilon}_t\|_\infty\lesssim\log^{1/2}(N\vee T)$. By Assumption 2, we also obtain $\max_k\left\|\mathbf{b}_k^0\right\|_1\leq N_1^{1/2}\max_k\left\|\mathbf{b}_k^0\right\|_2$. Under Assumption 3, we have

$$\sum_{\ell=L_n}^{\infty}\left\|\boldsymbol{\Phi}_\ell'\right\|_1=\sum_{\ell=L_n}^{\infty}\|\boldsymbol{\Phi}_\ell\|_\infty\leq N^{1/2}\sum_{\ell=L_n}^{\infty}\|\boldsymbol{\Phi}_\ell\|_2$$
$$\leq C_eN^{1/2}\sum_{\ell=L_n}^{\infty}\ell^{-(\nu+2)}\leq C_eN^{1/2}\int_{L_n-1}^{\infty}x^{-(\nu+2)}dx\lesssim N^{1/2}(L_n-1)^{-(\nu+1)}.$$

Thus setting $x=\max_k\|\mathbf{b}_k^0\|_2\log^{1/2}(N\vee T)$ yields

$$\mathbb{P}\left(\max_{t,k}\left|\sum_{\ell=L_n}^{\infty}\boldsymbol{\varepsilon}_{t-\ell}'\boldsymbol{\Phi}_\ell'\mathbf{b}_k^0\right|\gtrsim\max_k\|\mathbf{b}_k^0\|_2\log^{1/2}(N\vee T)\right)$$
$$\lesssim N/(L_n-1)^{\nu+1}=O((N\vee T)^{-\nu}).$$

Combining the results gives the proof of (b).

(c) Let $\boldsymbol{\phi}_{\ell,i}'$ and $\boldsymbol{\psi}_{m,k}'$ denote the ith and kth row vectors of $\boldsymbol{\Phi}_\ell$ and $\boldsymbol{\Psi}_m$, respectively. Note that $\max_k\|\boldsymbol{\psi}_{m,k}\|_1\leq r^{1/2}\|\boldsymbol{\Psi}_m\|_2$ and $\max_i\|\boldsymbol{\phi}_{\ell,i}\|_2\leq\|\boldsymbol{\Phi}_\ell\|_2$ by definition of the spectral norm. Thus we have

$$
\begin{aligned}
\|\mathbf{E}'\mathbf{F}\|_{\max} = \max_{i\in\{1,\dots,N\}} \max_{k\in\{1,\dots,r\}} &\left| \sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty}\sum_{t=1}^{T} \boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\boldsymbol{\zeta}'_{t-m}\boldsymbol{\psi}_{m,k}\right| \\
&\le \max_{i,k}\sum_{\ell}\sum_{m}\left|\sum_{t}\boldsymbol{\phi}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\boldsymbol{\zeta}'_{t-m}\boldsymbol{\psi}'_{m,k}\right| \\
&\le \max_{i,k}\sum_{\ell}\sum_{m}\left\|\sum_{t}\mathbf{w}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\boldsymbol{\zeta}'_{t-m}\right\|_{\infty}\|\boldsymbol{\phi}_{\ell,i}\|_2\|\boldsymbol{\psi}_{m,k}\|_1 \\
&\le r^{1/2}\max_{i}\sum_{\ell}\sum_{m}\left\|\sum_{t}\boldsymbol{\xi}_{t,\ell mi}\right\|_{\infty}\|\boldsymbol{\Psi}_m\|_2\|\boldsymbol{\Phi}_\ell\|_2.
\end{aligned}
$$

where $\mathbf{w}_{\ell,i} := \|\boldsymbol{\phi}_{\ell,i}\|_2^{-1}\boldsymbol{\phi}_{\ell,i}$ and $\boldsymbol{\xi}_{t,\ell mi} := \mathbf{w}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}\boldsymbol{\zeta}'_{t-m}$. Since $\{\varepsilon_{ti}\}$ are i.i.d. subG random variables (Assumption 3), there exists some constant $C_2 > 0$ such that

$$
\mathbb{P}\left(|\mathbf{w}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}| > x\right) \le 2\exp\left\{-x^2/(C_2\|\mathbf{w}_{\ell,i}\|_2^2)\right\} = 2\exp\left(-x^2/C_2\right).
$$

This implies that $\mathbf{w}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}$ is subG. Furthermore, because $\{\zeta_{tk}\}$ are i.i.d. subG random variables (Assumption 1) independent of $\mathbf{w}'_{\ell,i}\boldsymbol{\varepsilon}_{t-\ell}$, each element of $\{\boldsymbol{\xi}_{t,\ell mi}\}_t$ is a sequence of i.i.d. sub-exponential (subE) random variables. (See Vershynin (2018) for a detailed discussion on subG and subE random variables.)

The sum is divided as follows:

$$
\begin{aligned}
\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty} &= \left(\sum_{\ell=0}^{L_n-1} + \sum_{\ell=L_n}^{\infty}\right)\left(\sum_{m=0}^{L_n-1} + \sum_{m=L_n}^{\infty}\right) \\
&= \sum_{\ell=0}^{L_n-1}\sum_{m=0}^{L_n-1} + \sum_{\ell=0}^{L_n-1}\sum_{m=L_n}^{\infty} + \sum_{\ell=L_n}^{\infty}\sum_{m=0}^{L_n-1} + \sum_{\ell=L_n}^{\infty}\sum_{m=L_n}^{\infty}.
\end{aligned}
$$

Consider the first sum. Note that

$$
\begin{aligned}
&\max_{i}\sum_{\ell=0}^{L_n-1}\sum_{m=0}^{L_n-1}\left\|\sum_{t}\boldsymbol{\xi}_{t,\ell mi}\right\|_{\infty}\|\boldsymbol{\Psi}_m\|_2\|\boldsymbol{\Phi}_\ell\|_2 \\
&\le \max_{i}\max_{\ell,m}\left\|\sum_{t}\boldsymbol{\xi}_{t,\ell mi}\right\|_{\infty}\sum_{\ell=0}^{L_n-1}\sum_{m=0}^{L_n-1}\|\boldsymbol{\Phi}_\ell\|_2\|\boldsymbol{\Psi}_m\|_2,
\end{aligned}
$$

where $\sum_{\ell=0}^{\infty}\sum_{m=0}^{\infty}\|\boldsymbol{\Phi}_\ell\|_2\|\boldsymbol{\Psi}_m\|_2$ is bounded by Assumptions 1 and 3. By the Bernstein inequality for a sum of subE random variables (Vershynin, 2018, Theorem 2.8.1) together with the union bound, for any $0 \leq x \lesssim T^{1/2}$, there exists a constant $C_3 > 0$ such that

$$\mathbb{P}\left(\max_i \max_{\ell,m} \left\| \sum_t \boldsymbol{\xi}_{t,\ell mi} \right\|_\infty > xT^{1/2}\right) \leq 2NL_n^2 r \max_{\ell,m,i} \exp\left(-x^2/C_3\right).$$

Taking $x = C_3^{1/2}(\nu+3)^{1/2}\log^{1/2}(N \vee T)$ establishes the upper bound,

$$\max_i \max_{\ell,m} \left\| \sum_t \boldsymbol{\xi}_{t,\ell mi} \right\|_\infty \lesssim T^{1/2}\log^{1/2}(N \vee T),$$

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

Next consider the second summation. In a similar manner to the proof of (b), we have

$$\begin{aligned}
&\mathbb{P}\left(\max_i \sum_{\ell=0}^{L_n-1} \sum_{m=L_n}^{\infty} \left\| \sum_t \boldsymbol{\xi}_{t,\ell mi} \right\|_\infty \|\boldsymbol{\Phi}_\ell\|_2 \|\boldsymbol{\Psi}_m\|_2 > x\right) \\
&\leq \sum_{\ell=0}^{L_n-1} \sum_{m=L_n}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2 \|\boldsymbol{\Psi}_m\|_2 \sum_t \mathbb{E} \max_i \left\| \boldsymbol{\xi}_{t,\ell mi} \right\|_\infty / x \\
&\leq Tr\log(N) \sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2 \sum_{m=L_n}^{\infty} \|\boldsymbol{\Psi}_m\|_2 / x.
\end{aligned}$$

By the sub-Gaussianity of Assumptions 1 and 3, we have $\sum_{\ell=0}^{\infty} \|\boldsymbol{\Phi}_\ell\|_2 < \infty$ and

$$\sum_{m=L_n}^{\infty} \|\boldsymbol{\Psi}_m\|_2 \lesssim \sum_{m=L_n}^{\infty} m^{-(\nu+2)} \leq (N \vee T)^{-(\nu+1)}.$$

Therefore, taking $x \asymp \log(N \vee T)$ entails that the second sum is $O(\log(N \vee T))$, which holds with probability at least $1 - O((N \vee T)^{-\nu})$. As for the remaining two summations, we can achieve the same results under Assumptions 1 and 3. This completes the proof of (c). □

Lemma 2. *Suppose the same conditions as Theorem 1. Then, for any $\mathbf{H} \in \mathbb{R}^{T \times k}$ $(k \leq r)$ such that $\mathbf{H}'\mathbf{H} = T\mathbf{I}_k$, the following inequalities simultaneously hold with probability at least*

$1-O((N\vee T)^{-\nu})$:

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

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

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

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

Proof. Recall $\mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime}=\mathbf{C}^0=\mathbf{F}^0\mathbf{B}^{0\prime}$. 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

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

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 3. *Suppose the same conditions as Theorem 2. Then for any $\eta_n\geq 0$ we have*

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\gtrsim\frac{N_r^2}{N_1}\|\widehat{\mathbf{F}}-\mathbf{F}^0\|_{\mathrm{F}}^2+\frac{TN_r}{N_1}\|\widehat{\mathbf{B}}-\mathbf{B}^0\|_{\mathrm{F}}^2.$$

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, it suffices to prove the following two inequalities:

$$(a) \quad \|\mathbf{\Delta}\|_{\mathrm{F}}^2 \gtrsim \frac{N_r^2}{N_1}\|\widehat{\mathbf{U}} - \mathbf{U}^0\|_{\mathrm{F}}^2,$$

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

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

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

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

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

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

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

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

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

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

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

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

Because $N_1/N_r \geq 1$, combining these inequalities gives

$$\|\widehat{\mathbf{D}}_* \widehat{\mathbf{V}}_*' - \mathbf{D}_*^0 \mathbf{V}_*^{0\prime}\|_{\mathrm{F}}^2 \leq 2\|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 + 2\|(\widehat{\mathbf{U}}_* - \mathbf{U}_*^0)\mathbf{D}_*^0\|_{\mathrm{F}}^2 \lesssim \|\boldsymbol{\Delta}_*\|_{\mathrm{F}}^2 \frac{N_1}{N_r} = \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \frac{N_1}{TN_r}.$$

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 (b). This completes the proof. □

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

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

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

$$\mathbb{P}\left(\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}} > x\right) \leq \mathbb{P}\left(\|\mathbf{W}_{\mathcal{S}}\|_{\mathrm{F}} > x \mid \hat{\underline{b}}_n > \underline{b}_n^0/2\right) + \mathbb{P}\left(\hat{\underline{b}}_n \leq \underline{b}_n^0/2\right). \tag{A.15}$$

Set $x = 2(rN_1)^{1/2}/\underline{b}_n^0$ in (A.15). Then the first probability in the upper bound of (A.15) 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 (13), the second probability of the upper bound of (A.15) is bounded as

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

These two bounds together with (A.15) imply the result. □

Lemma 5. *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 $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{F}} - \mathbf{F}^0$. Then we have

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

where the last inequality follows by condition (13). By the property of norms and Lemma 1, we further obtain

$$\begin{aligned}
&\|\mathbf{B}^0\mathbf{F}^{0\prime}\widehat{\boldsymbol{\Delta}}\|_{\max} + \|\mathbf{E}'\widehat{\boldsymbol{\Delta}}\|_{\max} + \|\mathbf{E}'\mathbf{F}^0\|_{\max} \\
&\le T^{1/2} r \|\mathbf{B}^0\|_{\max}\|\mathbf{F}^0\|_{\max}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} + T^{1/2}\|\mathbf{E}\|_{\max}\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}} + \|\mathbf{E}'\mathbf{F}^0\|_{\max} \\
&\lesssim T^{1/2} r_n \log^{1/2}(N \vee T) + T^{1/2} r_n \log^{1/2}(N \vee T) + T^{1/2}\log^{1/2}(N \vee T)
\end{aligned}$$

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

$$\eta_n^{-1}\left\|\mathbf{W}^{-}_{\mathcal{S}^c} \circ (\mathbf{X}'\widehat{\mathbf{F}})_{\mathcal{S}^c}\right\|_{\max} \lesssim (\underline{b}^0_n/\eta_n)(1 \vee r_n) T^{1/2}\log^{1/2}(N \vee T) \tag{A.16}$$

with probability at least $1 - O((N \vee T)^{-\nu})$. The desired strict inequality is then obtained eventually by the lower bound of condition (12). This completes the proof. □

Lemma 6. *Suppose that Assumptions 1–3 and condition* (10) *hold. Then we have*

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

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

Proof. Let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{F}}_{\mathsf{PC}} - \mathbf{F}^0$ in this proof. By the definition of the PC estimator under PC1 restriction, we have

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

Then by Lemma 1 and the proof of Lemma 5, we have

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

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

C Derivation of the Estimation Error Bound

We explain the derivation of the estimation error upper bounds in the proofs of Theorems 2 and 3. Let $x = \|\mathbf{\Delta}^f\|_{\mathrm{F}}$ and $y = \|\mathbf{\Delta}^b\|_{\mathrm{F}}$ for Theorem 2 (SOFAR). Multiplying both sides of (A.7) by N_1/N_r^2 yields

$$x^2 + a_n y^2 \leq b_n x + c_n xy + d_n y, \tag{A.17}$$

where

$$\begin{aligned}
&a_n = T/N_r, \quad b_n = M(N_1^{3/2}T^{1/2}/N_r^2)\log^{1/2}(N \vee T), \\
&c_n = M(N_1/N_r^2)(N \vee T)^{1/2}, \quad d_n = b_n,
\end{aligned}$$

with M being some positive constant. For the proof of Theorem 3 (PC), let $x = \|\mathbf{\Delta}_{\mathsf{PC}}^f\|_{\mathrm{F}}$ and $y = \|\mathbf{\Delta}_{\mathsf{PC}}^b\|_{\mathrm{F}}$ with setting

$$d_n = M(N_1 N^{1/2}T^{1/2}/N_r^2)\log^{1/2}(N \vee T) \geq b_n.$$

We treat the two cases simultaneously.

We are ready to find a maximum value of $x+y$, where (x,y) satisfies inequality (A.17). Set $x=-y+m$ for some positive value m, and plug this into (A.17). Then we have

$$(-y+m)^2+a_ny^2 \leq b_n(-y+m)+c_n(-y+m)y+d_ny,$$

which is equivalently written as

$$f(y) := (a_n+c_n+1)y^2+(b_n-d_n-c_nm-2m)y+m^2-b_nm \leq 0.$$

We derive the region of m that satisfies the inequality. Because a_n+c_n+1 is positive and $f(y)=0$ must have at least a solution, it is required that

$$D(m) := \{b_n-d_n-(c_n+2)m\}^2-4(a_n+c_n+1)(m^2-b_nm) \geq 0.$$

Collecting the terms gives

$$\begin{aligned} D(m) &= -\{4(a_n+c_n+1)-(c_n+2)^2\}m^2 \\ &\quad +\{4b_n(a_n+c_n+1)+2(d_n-b_n)(c_n+2)\}m+(b_n-d_n)^2 \\ &= -(4a_n-c_n^2)m^2+2(2a_nb_n+b_nc_n+c_nd_n+2d_n)m+(b_n-d_n)^2. \end{aligned}$$

Derive the domain of m such that $D(m)\geq 0$. Under condition (10),

$$2\alpha_1+1\vee\tau < 3\alpha_r+\tau,$$

it holds that $a_n/c_n^2 \to \infty$. Thus it can be assumed that $4a_n > c_n^2$ for sufficiently large n. Note that $D(m)$ becomes concave by this condition. The two solutions to $D(m)=0$ are then computed as

$$\begin{aligned} m_1 &= \frac{2(2a_nb_n+b_nc_n+c_nd_n+2d_n)+\sqrt{4(2a_nb_n+b_nc_n+c_nd_n+2d_n)^2+4(4a_n-c_n^2)(b_n-d_n)^2}}{2(4a_n-c_n^2)}, \\ m_2 &= \frac{2(2a_nb_n+b_nc_n+c_nd_n+2d_n)-\sqrt{4(2a_nb_n+b_nc_n+c_nd_n+2d_n)^2+4(4a_n-c_n^2)(b_n-d_n)^2}}{2(4a_n-c_n^2)}. \end{aligned}$$

Hence, we obtain $0\vee m_2 \leq m \leq m_1$. As for m_2, we can observe $m_2 \leq 0$ since $4a_n > c_n^2$

with an easy algebra. Thus it is sufficient to evaluate m_1. By a simple calculation, we have $m_1 \in [\underline{m}, \overline{m}]$, where

$$\begin{aligned}\underline{m} &= \frac{2(2a_nb_n + b_nc_n + c_nd_n + 2d_n) + \sqrt{4(4a_n - c_n^2)(b_n - d_n)^2}}{2(4a_n - c_n^2)} \\ &= \frac{2(2a_nb_n + b_nc_n + c_nd_n + 2d_n) + (d_n - b_n)\sqrt{4(4a_n - c_n^2)}}{2(4a_n - c_n^2)}\end{aligned}$$

and

$$\begin{aligned}\overline{m} &= \frac{2(2a_nb_n + b_nc_n + c_nd_n + 2d_n) + \sqrt{4(2a_nb_n + b_nc_n + c_nd_n + 2d_n)^2} + \sqrt{4(4a_n - c_n^2)(b_n - d_n)^2}}{2(4a_n - c_n^2)} \\ &= \frac{4(2a_nb_n + b_nc_n + c_nd_n + 2d_n) + (d_n - b_n)\sqrt{4(4a_n - c_n^2)}}{2(4a_n - c_n^2)}.\end{aligned}$$

Since they are the same up to a positive constant factor, we obtain the sharp bound

$$\begin{aligned}m_1 &\asymp \frac{2(2a_nb_n + b_nc_n + c_nd_n + 2d_n) + (d_n - b_n)\sqrt{4(4a_n - c_n^2)}}{2(4a_n - c_n^2)} \\ &= \frac{2a_nb_n + (b_n + d_n)c_n + 2d_n}{4a_n - c_n^2} + \frac{d_n - b_n}{\sqrt{4a_n - c_n^2}}.\end{aligned}$$

Recall $a_n/c_n^2 \to \infty$. For sufficiently large n, we consequently obtain

$$\begin{aligned}m \leq m_1 &\lesssim \frac{a_nb_n + d_nc_n + d_n}{a_n} + \frac{d_n}{\sqrt{a_n}} \\ &\lesssim b_n + \frac{d_n}{\sqrt{a_n}} + \frac{d_n}{a_n} = \left(\sqrt{a_n}b_n + d_n + \frac{d_n}{\sqrt{a_n}}\right)\frac{1}{\sqrt{a_n}}.\end{aligned}$$

The SOFAR upper bound is obtained by putting $d_n = b_n$:

$$\begin{aligned}m \lesssim \left(\sqrt{a_n} + 1 + \frac{1}{\sqrt{a_n}}\right)\frac{b_n}{\sqrt{a_n}} &\asymp \left(\frac{T^{1/2}}{N_r^{1/2}} + \frac{N_r^{1/2}}{T^{1/2}}\right)\frac{N_1^{3/2}}{N_r^{3/2}}\log^{1/2}(N \vee T) \\ &\asymp \frac{N_1^{3/2}T^{1/2}}{N_r(N_r \wedge T)}\log^{1/2}(N \vee T).\end{aligned}$$

The PC upper bound is obtained by recalling $d_n \geq b_n$ and $d_n/b_n = N^{1/2}/N_1^{1/2}$:

$$\begin{aligned}
m &\lesssim \left(\sqrt{a_n} + \frac{d_n}{b_n} + \frac{d_n}{\sqrt{a_n} b_n}\right) \frac{b_n}{\sqrt{a_n}} \\
&\asymp \left\{\frac{T^{1/2}}{N_r^{1/2}} + \frac{N^{1/2}}{N_1^{1/2}}\left(1 + \frac{N_r^{1/2}}{T^{1/2}}\right)\right\} \frac{N_1^{3/2}}{N_r^{3/2}} \log^{1/2}(N \vee T) \\
&= \left(\frac{T^{1/2}}{N_r^{1/2}} + \frac{N_r^{1/2}}{T^{1/2}}\right) \frac{N_1^{3/2}}{N_r^{3/2}} \log^{1/2}(N \vee T) + \left\{\frac{N^{1/2}}{N_1^{1/2}}\left(1 + \frac{N_r^{1/2}}{T^{1/2}}\right) - \frac{N_r^{1/2}}{T^{1/2}}\right\} \frac{N_1^{3/2}}{N_r^{3/2}} \log^{1/2}(N \vee T).
\end{aligned}$$

The first term is the same as the SOFAR bound. The second term is further transformed as

$$\begin{aligned}
&\left\{\frac{N^{1/2}}{N_1^{1/2}}\left(1 + \frac{N_r^{1/2}}{T^{1/2}}\right) - \frac{N_r^{1/2}}{T^{1/2}}\right\} \frac{N_1^{3/2}}{N_r^{3/2}} \log^{1/2}(N \vee T) \\
&= \frac{N^{1/2} N_1}{N_r^{3/2}} \log^{1/2}(N \vee T) + \left(\frac{N^{1/2}}{N_1^{1/2}} - 1\right) \frac{N_1^{3/2}}{T^{1/2} N_r} \log^{1/2}(N \vee T) \\
&= \frac{N^{1/2} N_1}{N_r^{3/2}} \log^{1/2}(N \vee T) + (1 + O(1)) \frac{N^{1/2} N_1}{T^{1/2} N_r} \log^{1/2}(N \vee T) \\
&\asymp \frac{N^{1/2} N_1}{N_r (N_r \wedge T)^{1/2}} \log^{1/2}(N \vee T).
\end{aligned}$$

Combining the terms gives the PC upper bound.

Remark 2. If $4a_n^2 \leq c_n^2$, then the solutions to $D(m) = 0$ are negative or complex. This does not lead to achieving a meaningful upper bound. Condition (9) ensures $4a_n^2 > c_n^2$ eventually, under which we successfully derive the upper bounds in the theorems.

D A Brief Review of SOFAR

Uematsu et al. (2019) have proposed the estimation method called the SOFAR for general high-dimensional multivariate regression models. We briefly review the framework here. Consider estimation of the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^0 + \mathbf{E},$$

where $\mathbf{Y}$ is a $T \times q$ response matrix, $\mathbf{X}$ is a $T \times p$ input matrix, $\mathbf{E}$ is a $T \times q$ error matrix, and $\mathbf{C}^0$ is the $p \times q$ coefficient matrix. To manage the high dimensionality $(p, q \to \infty)$ efficiently, the coefficient matrix $\mathbf{C}^0$ is supposed to have the sparse and low-rank singular value decomposition (SVD) structure:

$$\mathbf{C}^0 = \mathbf{U}^0\mathbf{D}^0\mathbf{V}^{0\prime},$$

where $\mathbf{U}^0$ and $\mathbf{V}^0$ are sparse and orthogonal matrices of the left- and right-singular vectors, respectively, with $\mathbf{U}^{0\prime}\mathbf{U}^0 = \mathbf{I}$ and $\mathbf{V}^{0\prime}\mathbf{V}^0 = \mathbf{I}$ and $\mathbf{D}^0$ is the low-rank diagonal matrix of the singular values.

The SOFAR estimator $(\widehat{\mathbf{U}}, \widehat{\mathbf{D}}, \widehat{\mathbf{V}})$ is defined as a minimizer of the problem

$$\begin{aligned}&\text{minimize } \left\|\mathbf{Y} - \mathbf{XUDV}'\right\|_{\mathrm{F}}^2 + \lambda_a \left\|\mathbf{UD}\right\|_1 + \lambda_b \left\|\mathbf{VD}\right\|_1 + \lambda_d \left\|\mathbf{D}\right\|_1,\\ &\text{subject to } \mathbf{U}'\mathbf{U} = \mathbf{I},\ \mathbf{V}'\mathbf{V} = \mathbf{I},\ \mathbf{D} \text{ diagonal.}\end{aligned}$$

As for the statistical theory of the SOFAR estimator, Uematsu et al. (2019) establish the estimation error bound like Theorem 2 under the assumptions of nonrandom regressors $\mathbf{X}$ and coefficient $\mathbf{C}^0$ with Gaussian errors $\mathbf{E}$ that has i.i.d. rows. Regarding the numerical contribution, they provide the SOFAR optimization algorithm with a convergence property based on the augmented Lagrangian method and block coordinate descent, which is available in R package `rrpack`.

The SOFAR procedure is readily applicable to estimation of the sWF models. If we set $\mathbf{X} = \mathbf{I}_N$ in the model, it reduces to the sWF model with the (non-sparse) factors $\mathbf{F}^0 = \mathbf{U}^0$ and the (sparse) factor loadings $\mathbf{B}^0 = \mathbf{V}^0\mathbf{D}^0$ as in Section 2. The SOFAR estimator for the sWF model is obtained by setting $\lambda_a = 0$ and $\lambda_d = 0$.

E Additional experimental results

E.1 Orthogonality restrictions in $\mathbf{F}^0$ and $\mathbf{B}^0$ are violated

In the experiments summarised in Table 2, the data generating process (DGP) is given by $x_{ti} = \sum_{k=1}^r f_{tk}^0 b_{ik}^0 + e_{ti}$, where the factor loadings b_{ik}^0 and factors f_{tk}^0 are formed such that $N^{-1}\sum_{i=1}^N b_{ik}^0 b_{i\ell}^0 = 1\{k = \ell\}$ and $T^{-1}\sum_{t=1}^T f_{tk}^0 f_{t\ell}^0 = 1\{k = \ell\}$, by applying Gram–Schmidt orthonormalization to b_{ik}^* and f_{tk}^*, respectively, where $b_{ik}^* \sim$i.i.d.$N(0,1)$ for $i = 1, \ldots, N_k$ and $b_{ik}^* = 0$ for $i = N_k + 1, \ldots, N$, and $f_{tk}^* = \rho_{fk} f_{t-1,k}^* + v_{tk}$ with $v_{kt} \sim$i.i.d.$N(0, 1 - \rho_{fk}^2)$

and f^*_{0k} $\sim$i.i.d.$N(0,1)$. In order to investigate the effect of the violation of the orthogonalisations, we examine the performance of estimators for the DGP $x_{ti} = \sum_{k=1}^{r} f^*_{tk} b^*_{ik} + e_{ti}$. Except for this change, the experimental design is identical to that for Table 2. The results are summarised in Table 5. As can be seen, the results are qualitatively very similar to those reported in Table 2: The adaptive SOFAR estimator accurately estimates the exponents (α_1,α_2) and mostly dominates the PC estimator. Here the PC estimator for the loadings slightly outperforms the SOFAR estimator when the factors are very close to strong ((α_1,α_2)=(0.9,0.9)).

Table 5: Performance of the SOFAR (SO) and PC estimators for approximate factor models with two factor components with $(\alpha_1, \alpha_2) = (0.9, 0.9), (0.8, 0.8), (0.8, 0.5)$ and $(0.5, 0.4)$, when the orthogonality restrictions in $\mathbf{F}^0$ and $\mathbf{B}^0$ are violated

Design (α_1,α_2)	(0.9,0.9)		(0.8,0.8)		(0.8,0.5)		(0.5,0.4)*	
	$N = T = 100$							
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.86	0.02	0.77	0.02	0.75	0.04	0.48	0.05
$\hat{\alpha}_2$	0.85	0.02	0.75	0.03	0.56	0.15	0.45	0.11
	SO	PC	SO	PC	SO	PC	SO	PC
$L^2(\hat{\mathbf{F}})_{\times 100}$	11.62	16.13	13.18	18.02	16.12	23.64	13.59	19.36
$L^2(\hat{\mathbf{B}})_{\times 100}$	14.63	14.22	15.43	19.73	16.25	43.75	15.74	29.08
$L^2(\hat{\mathbf{C}})_{\times 100}$	8.16	14.52	10.53	20.09	25.83	53.50	10.50	25.02
	$N = T = 200$							
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
$\hat{\alpha}_1$	0.88	0.01	0.77	0.01	0.77	0.01	0.48	0.03
$\hat{\alpha}_2$	0.87	0.01	0.77	0.02	0.55	0.07	0.45	0.06
	SO	PC	SO	PC	SO	PC	SO	PC
$L^2(\hat{\mathbf{F}})_{\times 100}$	8.93	11.15	9.90	12.21	10.78	14.90	10.01	13.14
$L^2(\hat{\mathbf{B}})_{\times 100}$	10.52	10.39	11.04	13.86	8.14	26.96	12.64	22.80
$L^2(\hat{\mathbf{C}})_{\times 100}$	4.15	7.41	5.62	10.77	15.14	33.27	8.40	17.96

Note: * For this model, $\theta = 0.2$.

F Additional Estimation Results

F.1 Estimating exponents with stock returns

In addition to reporting the divergence rates in Section 6.4, 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. We also 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 in terms of 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 into one of the ten industrial sectors based on the Industry Classification Benchmark (ICB)[13]: (i) *Oil & Gas*; (ii) *Basic Materials*; (iii) *Industrials*; (iv) *Consumer Goods*; (v) *Health Care*; (vi) *Consumer Services*; (vii) *Telecommunications*; (viii) *Utilities*; (ix) *Financials*; (x) *Technology.* Then, for a given factor, the factor loadings are grouped into the negatives and the positives. For each group, the portion of the sum of the absolute value of the factor loadings which belong to each industrial sector is computed and reported. Specifically, we compute the following statistics for factor ℓ and industry s for given estimation window:

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

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

Figure SP2(a) shows the portion of the industrial sectors to which the securities consists of S&P500 belong, and the measure $T^{+}_{b_1,s}$ for the first factor. All the loadings to the first factor have the same sign (and it is chosen to be positive), which strongly suggests that this is the market factor. As one might expect, the 'beta' (the factor loading) of defensive industries, *Oil&Gas, Health Care, Telecoms*, and *Utilities* is relatively small. The 'beta' of cyclical industries such as *Industrials*, *Financials*, and *Basic Materials*, is noticeably high.

The averages of the measures of negative and positive industrial contributions to the second factor loadings are reported in Figure SP2(b). It shows that *Utility* and *Financials* account for around 43% and 23% of negative loadings, respectively, while *Technology*, *Industrials* and *Basic Materials* share 40%, 17%, and 14% of positive loadings, respectively. The averages of $T^{-}_{b_\ell,s}$ and $T^{+}_{b_\ell,s}$ for the third factor are reported in Figure SP2(c). It is clear that this is the *Oil&Gas* factor, which share the 67% of the negative loadings. *Financials*, *Consumer Services*, and *Consumer Goods* share 29%, 23%, and 19% of positive loadings, which means that these industrial sectors move in the 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 shares 43% of the positive loading, together with *Health Care* with 17% of the share. No dominant industry is found for negative loadings, which are equally shared by cyclical industries.

In turn, we discuss each factor in more detail by analyzing Table SP1, and 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 the 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_3 , and α_4, exhibit very different trajectories over the months, and their orders in terms of value change (i.e., their plots cross).

Let us look at the trajectory of α_2. From Figure SP2(b), under our identification condition, we can understand the second factor 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 declines and stays below 0.75 to October 1999. Then it rises sharply to achieve 0.83 in February 2000. Indeed, this period corresponds to the turbulence of the *Basic Material* stock index during 1998-2003, the fall of the *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 often estimated but the fourth factor is.

Now let us analyze the movement of α_3. From Figure SP2(c), under our identification condition, we can understand the third factor 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. We can therefore surmise 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 up to the latest data point, April 2018.

Now let us analyze the movement of α_4. From Figure SP2(d), under our identification condition, we can understand the fourth factor 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, it is often not estimated, but it is from March 2010 onward, seemingly becoming increasingly strong toward the latest month, April 2018. Since its first appearance, the value of α_4 is mostly between 0.75 and 0.80. After the sharp one-off drop in February 2015,[14)] α_4 rises to become the highest next to the first factor from November 2016 onward.

G Some Examples of Rotation

Inspired by a suggestion from the AE and referee, we illustrate when the sparsity of $\mathbf{B}^*$ is preserved. We find that sparsity is rotation variant, which implies that the degree of sparsity of $\mathbf{B}^0$ can rise after the rotation of sparse or dense $\mathbf{B}^*$. Of course we can easily give an example in which $\mathbf{B}^0$ is less sparse than $\mathbf{B}^*$; see Freyaldenhoven (2020).

Example 1: $\mathbf{B}^*$ is sparse and $\mathbf{B}^0$ is more sparse. Suppose $r = 2$ and consider a $N \times r$

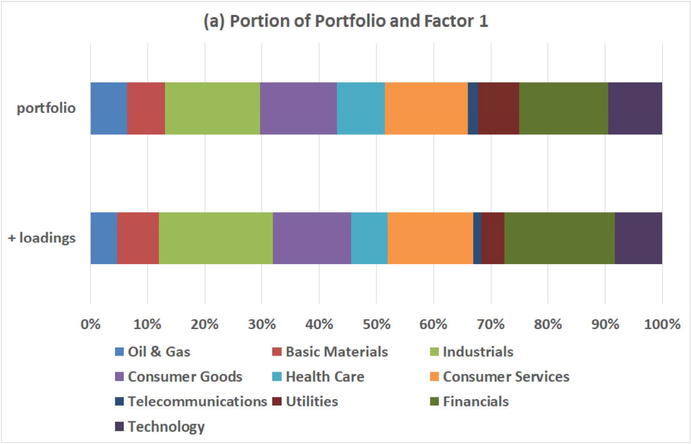

Figure 13: the portion of the industrial sectors in S&P500 and in the positive/negative 1st factor loadings

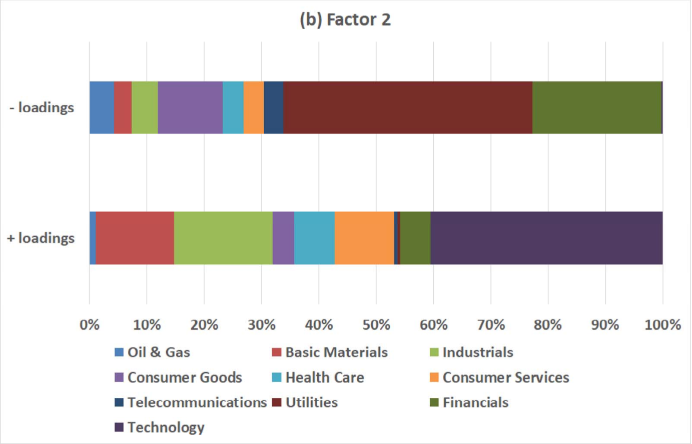

Figure 14: the portion of the industrial sectors in the positive/negative second factor loadings

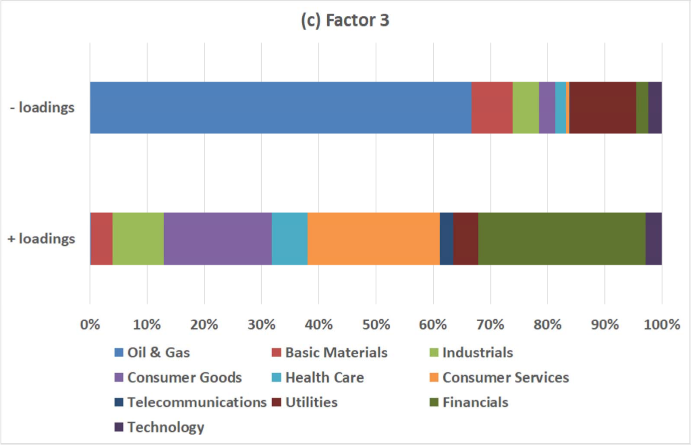

Figure 15: the portion of the industrial sectors in the positive/negative third factor loadings

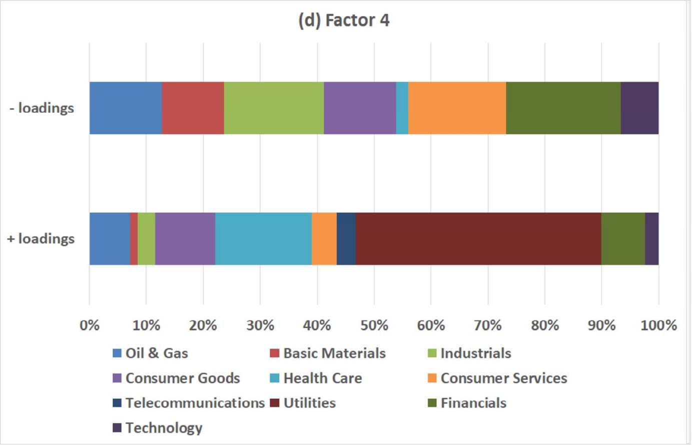

Figure 16: the portion of the industrial sectors in the positive/negative fourth factor loadings

matrix

$$\mathbf{B}^* = \begin{pmatrix} \mathbf{b}_1^* & \mathbf{b}_1^* \\ \mathbf{b}_2^* & \mathbf{0} \end{pmatrix},$$

where $\mathbf{b}_1^*$ and $\mathbf{b}_2^*$ are $N_1 \times 1$ and $N_2 \times 1$ vectors with non-zero elements such that $\mathbf{b}_1^* \neq \mathbf{b}_2^*$ and $N = N_1 + N_2$. Consider the $r \times r$ rotation matrix

$$\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{H}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

It is easily seen that

$$\mathbf{B}^*\mathbf{H}^{-1} = \mathbf{B}^0 = \begin{pmatrix} \mathbf{b}_1^* & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_2^* \end{pmatrix},$$

which is sparser than $\mathbf{B}^*$. Note that $\mathbf{B}^{0\prime}\mathbf{B}^0 = \mathrm{diag}(\mathbf{b}_1^{*\prime}\mathbf{b}_1^*, \mathbf{b}_2^{*\prime}\mathbf{b}_2^*)$.

Example 2: $\mathbf{B}^*$ is dense but $\mathbf{B}^0$ is sparse. Suppose $r = 2$ and consider a $N \times r$ dense matrix

$$\mathbf{B}^* = \begin{pmatrix} a\mathbf{b}_1^* & \mathbf{b}_1^* \\ a\mathbf{b}_1^* & -a\mathbf{b}_1^* \end{pmatrix},$$

where $a \neq 0, -1$ and $\mathbf{b}_1^*$ is an $N/2 \times 1$ vector with non-zero elements. Consider the $r \times r$ rotation matrix

$$\mathbf{H} = \frac{1}{1+a}\begin{pmatrix} a & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{H}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & -a \end{pmatrix}.$$

It is easily seen that

$$\mathbf{B}^*\mathbf{H}^{-1} = \mathbf{B}^0 = \begin{pmatrix} (a+1)\mathbf{b}_1^* & \mathbf{0} \\ \mathbf{0} & a(a+1)\mathbf{b}_1^* \end{pmatrix},$$

which is sparse. Note that $\mathbf{B}^{0\prime}\mathbf{B}^0 = \mathbf{b}_1^{*\prime}\mathbf{b}_1^* \, \mathrm{diag}((a+1)^2, a^2(a+1)^2)$.