| Statistica Sinica Preprint No: SS-2023-0219 | |
|---|---|
| **Title** | Sparse Factor Model for High Dimensional Time Series |
| **Manuscript ID** | SS-2023-0219 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202023.0219 |
| **Complete List of Authors** | Xiaoran Wu,<br>Baojun Dou and<br>Rongmao Zhang |
| **Corresponding Authors** | Rongmao Zhang |
| **E-mails** | rmzhang@zju.edu.cn |

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219

Statistica Sinica

# SPARSE FACTOR MODEL FOR HIGH DIMENSIONAL TIME SERIES

Xiaoran Wu[1], Baojun Dou[2] and Rongmao Zhang[3]

*Zhejiang University*[1], *City University of Hong Kong*[2]

*and Zhejiang Gongshang University*[3]

*Abstract:* Factor models have been extensively employed in high dimensional time series. However, little is known for the case with the sparse loading matrix. This paper introduces a sparse factor model with an easy-to-implement estimation method, aiming to enhance interpretability and relax the constraints on the dimension $p$ of the time series. In particular, it is shown that under weak conditions, the loading space could be consistently estimated with a convergence rate related to the sparseness of each column in the loading matrix and the eigenvalues used to recover the latent factor and loading matrix. In addition, a randomized sequential test is introduced to determine the number of sparse factors. Simulations and real data analysis on sea surface air pressure and stock portfolios are also provided to illustrate the performance of the proposed method.

*Key words and phrases:* High dimensional time series, $\alpha$-mixing, Orthogonal projection, Sparse factor model.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



## 1. Introduction

Modeling high dimensional time series is of great interest and importance in a wide range of fields including signal process, medical research and financial analysis. For interpretability and simplicity, it is commonly assumed that the high dimensional data is driven by a low dimensional latent factor model. Regarding high dimensional factor models that utilize PCA estimator, a strong-factor framework is usually employed for theoretical consistency, which assumes that the leading eigenvalues driven by common factors are proportional to the dimension $p$, or the loading matrix is dense with each element being non-vanishing, see Fan et al. (2013) and Lam et al. (2011). Otherwise the consistency of the estimator is compromised. For example, Baik and Silverstein (2006), Paul (2007) and Nadler (2008) show that when the largest eigenvalue has a finite upper bound, the leading sample principal eigenvector is asymptotically orthogonal to the leading population principal eigenvector almost surely.

Through the strong factor benefits from a large dimension $p$, namely the "blessing of dimensionality" (see the convergence rate in Lam et al. (2011) for example), it suffers less interpretability as the dimension $p$ increases, since the loading matrix has non-vanishing elements in all the $p$ coordinates, see Pelger and Xiong (2022). To facilitate interpretability in

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



high dimensional series, we introduce the concept of sparse factor, which indicates that the factor does not contribute to all the $p$ series. It can be regarded as a special case of factor rotation, which is commonly used to find more interpretable factors. Due to the non-uniqueness of factors and loadings, it is possible to rotate principle components to get sparse loadings with some coordinates equal to or close to zero. A special type of sparse factor is the group factor structure, which is also an empirical motivation for our work, see Ando and Bai (2016), Chang et al. (2018), and Zhang et al. (2023). In group structure, the factors are associated with the loadings with nonzero elements only within specific groups. Ando and Bai (2016) provide empirical evidence for such a structure in finance.

The concept corresponding to the strong factor is the weak factor, which has some overlap with the sparse factor we propose. The weak factor has multiple definitions, one of which is that the eigenvalues driven by latent factors are $o(p)$, see Bai and Ng (2023). Others use the norm for each column of the loading matrix to define, see Chudik et al. (2011) and Lam et al. (2011). For example, Lam et al. (2011) define weak factor by the $L_2$ norm with $o(p^{1/2})$. In this sense, weak factors are attributed to two reasons, one is the magnitude of most elements in the loading matrix increases slowly (but none of elements is zero), the other is the loading matrix has a lot of

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



zero elements, or both. We use $L_1$ norm of each column to define sparse factors. We also show that the induced eigenvalues is $o(p)$ and related to our sparse index, which shares some similarities with the weak factors. Besides, weak factors are often found in financial and spatial data, but none of these papers metioned above considers incorporating such sparse or weak information in estimating the factor model, which also motivates our current study.

For high dimensional factor models, there are typically two methods for estimating, one uses the sample covariance matrix to recover the loading matrix, see Bai and Ng (2002), Bai (2003) and subsequent papers based on them. The other uses a symmetric and non negative-definite matrix constructed by auto-cross covariance over different time lags, see Lam et al. (2011), Lam and Yao (2012), Chang et al. (2015), Wang et al. (2019), Chang et al. (2023), Chang et al. (2024), among many others. The latter assumes that the idiosyncratic errors are white noise but the factor processes are serially dependent, indicating that the factors drive dynamics of most time series. To take advantage of the dynamic information of the time series, we carry on the construction of auto-cross covariance in Lam et al. (2011) to recover the loading space, and simultaneously add a sparse constraint on the eigenvectors.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



The main purpose of this paper is to propose an efficient method to estimate a factor model with sparse loadings for high dimensional time series. To implement the estimation procedure, we propose a new algorithm by the means of divide-and-conquer, which is efficient in computation and easy to implement. It iteratively alternates between two subtasks: constrained rank-one variance maximization and orthogonal projection. The former subtask aims to find the leading pseudo eigen vector with constraints; the latter projects the constructed matrix onto the orthogonal complementary space of the estimated leading vector in the former subtask, aiming to eliminate the influence of it. The rank-one variance maximization is proposed to obtain sparse loading in one direction, the orthogonal projection helps to obtain the orthogonal factor loadings. The divide-and-conquer method keeps the optimization convex in each single direction and reduces the computation complexities in high dimensional series therefore.

To the best of our knowledge, this paper is the first to propose a sparse factor model based on auto-cross covariance, allowing the sparsity varies for factors. The empirical studies on sea surface air pressure and stock portfolios show that the sparsity enhances the interpretability and efficiency of the factor model in high dimensional time series. Last but not least, we derive the convergence rate of the sparse factor loading under weak

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



assumptions, where we only assume that the tail of time series follows a polynomial order, while many literature require it to be subgaussian in high dimensional settings. Besides, it is shown that when the factors are sufficiently sparse, our convergence rate is better than that of Lam et al. (2011).

Regarding the topic of sparse factor model, we mention three of the latest relevant literature. Pelger and Xiong (2022) shrink the PCA factor weights and set many of them to zero to attain sparse and more interpretable factors. The so-called proximate factor is inconsistent but performs well in terms of the generalized correlation. Uematsu and Yamagata (2023) propose a special sparsity-induced weak factor, which constrains both $L_0$ and $L_1$ norm. For theoretical research, Bai and Ng (2023) consider the inference of weaker loadings in terms of the limit of the loading matrix multiplied by its transpose in the PCA-based approaches.

Notably, a large body of literature with the theme of Sparse PCA are also ralated to our work. Zou et al. (2006) present iterations between the singular value decomposition and the elastic net regression step. Witten et al. (2009) propose a penalized matrix decomposition for sample covariance matrix. Johnstone and Lu (2009) apply the classical PCA along with thresholding into a selected subset of variables with the larger sample vari-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



ances. Ma (2013) adds an additional thresholding step to the usual orthogonal iteration steps to seek sparse basis vectors for the subspace. In theoretical study, Vu and Lei (2012) investigate minimax rates for the estimator of the first principle sparse vector. Cai et al. (2013) consider the minimax optimality and adaptive estimation of the principal subspace. But this category of literature does not consider any factor structures, and their theoretical consistency is mainly obtained under the assumption of i.i.d. normality, which is divergent to our work.

The rest of the paper is organized as follows. The sparse factor model and the estimation methods are introduced in Section 2. The asymptotic theories are investigated in Section 3. Simulation results are reported in Section 4. The analyses of real data on both sea surface air pressure and stock portfolios are provided in Section 5. Conclusions and discussions are in Section 6. All mathematical proofs and some simulated results are relegated to the supplementary material.

Throughout this paper, we always use the following notation. $\|\mathbf{u}\|_1 = (\sum_{i=1}^{p} |u_i|)$ is the $L_1$ norm of a $p$-dimensional vector $\mathbf{u} = (u_1, \ldots, u_p)^T$, $\|\mathbf{u}\|_2 = (\sum_{i=1}^{p} u_i^2)^{1/2}$ is the Euclidean norm, and $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. For a matrix $\mathbf{H} = (h_{ij})$, $\|\mathbf{H}\|_F = \sqrt{trace\,(\mathbf{H}^T\mathbf{H})}$ is the Frobenius norm, The superscript $^T$ denotes the transpose of a vector or matrix.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



Finally, we use the notation $a \asymp b$ to denote $a = O(b)$ and $b = O(a)$.

## 2. Models and the Estimation

### 2.1 Sparse Factor Model

Let $\mathbf{y_t}$ be a $p \times 1$ observation from vector time series process at time $t$. Let $n$ denote the sample size and $p$ be the number of the series. $\mathbf{y_t}$ is said to have a factor structure if it has the representation as

$$\mathbf{y_t} = \Theta \mathbf{x_t} + \epsilon_t, t = 1, 2, ..., n, \tag{2.1}$$

where $\mathbf{x_t} = (x_{1,t}, ..., x_{r,t})^T$ is a $r \times 1$ latent process with unknown $r \ll p$, $\Theta = (\theta_1, ..., \theta_r)$ is a $p \times r$ unknown constant matrix. $\epsilon_t \sim WN(\mu_\epsilon, \Sigma_\epsilon)$ is a vector white-noise process. An effective dimension-reduction is achieved in the sense that the serial dependence of $\mathbf{y_t}$ is driven by that of a much lower-dimensional process $\mathbf{x_t}$. We refer to $\mathbf{x_t}$ a factor process and $\Theta$ a loading matrix.

Since none of the elements on the pair $(\Theta, \mathbf{x_t})$ are observable, the model remains unchanged if we replace $(\Theta, \mathbf{x_t})$ by $(\Theta\mathbf{H}, \mathbf{H^{-1}x_t})$ for any $r \times r$ invertible matrix $\mathbf{H}$. However the factor loading space is uniquely defined, which is the linear space spanned by the columns of $\Theta$ and denoted by

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



$\mathcal{M}(\Theta)$. Note that $\mathcal{M}(\Theta) = \mathcal{M}(\Theta\mathbf{H})$ for any invertible $\mathbf{H}$. Without loss of generality, we first assume the loading matrix $\Theta$ to be column-orthogonal, then we decompose $\Theta$ by $\Theta = \mathbf{QR}$ where $\mathbf{Q}$ is unit-orthogonal and $\mathbf{R}$ is upper triangular, and replace $(\Theta, \mathbf{x_t})$ by $(\mathbf{Q}, \mathbf{Rx_t})$. So in the following, we assume that $\Theta^T\Theta = \mathbf{I}_r$, where $\mathbf{I}_r$ is the $r \times r$ identity matrix.

The non uniqueness of $(\Theta, \mathbf{x_t})$ makes it possible to add some constraints to get a specific $\Theta$ that behaves well both theoretically and practically. Now recall the factor strength $\delta_0$ defined in Lam et al. (2011), which is for $\Theta = (\theta_1 \cdots \theta_r)$, $\|\theta_i\|_2^2 \asymp p^{\delta_0}$, for $i = 1, \cdots, r$ and $0 \leq \delta_0 \leq 1$. When $\delta_0 = 1$, the corresponding factor is named strong factor since it includes the case where all the elements of $\theta_i$ are $O(1)$. When $\delta_0 < 1$, the factors are weak factors. According to Theorem 1 of Lam et al. (2011), the convergence rate of the estimated loadings is slower in the presence of weak factors. In this situation, adding a constraint to the norm of $\Theta$ will improve the estimation. In this paper, we purpose a sparse index to character the sparsity of factors, that is

$$\|\theta_i\|_1 \asymp p^{\delta_i}, for \quad i = 1, \cdots, r \quad and \quad 0 \leq \delta_i \leq 1/2. \tag{2.2}$$

Notably, our configuration differs from the factor strength mentioned above, but it is similar to the definition of the semi-weak factor in Chudik et al. (2011). Since we assume $\|\theta_i\|_2 = 1$ and $\|\theta_i\|_1 \leq \sqrt{p}\,\|\theta_i\|_2 = \sqrt{p}$, $\delta_i$ should

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



be no more than 1/2, otherwise the $L_1$-penalty does not work.

Our goal is to estimate the $p \times r$ sparse factor loading $\Theta$, or more precisely the sparse factor loading space $\mathcal{M}(\Theta)$. Once an estimator $\widehat{\Theta}$ is obtained, we can estimate the factor process as $\widehat{\mathbf{x}}_t = (\widehat{\Theta}^T\widehat{\Theta})^{-1}\widehat{\Theta}^T\mathbf{y}_t$. and the resulting residuals are $\widehat{\varepsilon}_t = \left(\mathbf{I}_d - \widehat{\Theta}\widehat{\Theta}^T\right)\mathbf{y}_t$.

In this paper, we focus on the estimation of $\Theta$ so the number of factors $r$ is first assumed to be fixed and known. Then we adopt the randomized sequential procedure in Trapani (2018) to determine the factor number $r$. Other useful methods on the determination of $r$ include the information criterion, see Bai and Ng (2002) and Hallin and Liška (2007), the ratio-based method, see Lam and Yao (2012), and the hypothesis testing, see Pan and Yao (2008), Onatski (2009) and Onatski (2010).

## 2.2 The Estimation

We are considering the stationary case. We introduce some notations now.

$$\Sigma_y(k) = \mathrm{Cov}\left(\mathbf{y}_{t+k}, \mathbf{y}_t\right), \quad \Sigma_x(k) = \mathrm{Cov}\left(\mathbf{x}_{t+k}, \mathbf{x}_t\right),$$

$$\Sigma_{x\epsilon}(k) = \mathrm{Cov}\left(\mathbf{x}_{t+k}, \varepsilon_t\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



For a prescribed integer $k_0 \geq 1$,

$$\mathbf{M} = \sum_{k=1}^{k_0} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)^T,$$

**M** makes full use of the information in different time lags, and simultaneously remains nonnegative and symmetric, just like the covariance matrix. For the sparse factor model we propose, the following assumptions are required.

**Condition 1.** *(factor)* $\mathbf{x_t}$ *is weak stationary and* $\Sigma_{\mathbf{x}}(k)$ *is full-ranked for* $k = 0, 1, \cdots, k_0$, *where* $k_0 \geq 1$ *is a small positive integer.*

**Condition 2.** *(loading)* $\Theta = (\theta_1 \cdots \theta_r)$ *is column-orthogonal and sparse in the sense that* $\|\theta_i\|_2 = 1, \|\theta_i\|_1 \asymp p^{\delta_i}, i = 1, \cdots, r, \quad 0 \leq \delta_i \leq 1/2$. $\delta_i$ *is the sparse index.*

**Condition 3.** *(noise) (i)*$\epsilon_t \sim WN(\mu_\epsilon, \Sigma_\epsilon)$ *with the elements of* $\Sigma_\epsilon$ *bounded as* $(n, p) \to \infty$; *(ii)* $\Sigma_{\mathrm{x},\epsilon}(k) = \mathrm{Cov}\,(\mathbf{x}_{t+k}, \varepsilon_t)$ *has elements of order* $O(1)$; *(iii) For* $k > 0$, $\mathrm{Cov}\,(\mathbf{x}_t, \varepsilon_{t+k}) = 0$.

**Condition 4.** *The first* $r$ *eigenvalues of* $M$ *satisfy* $\lambda_1 > \cdots > \lambda_r$.

Condition 1 is commonly assumed in factor model, if $\Sigma_x$ is not full-ranked, we need to reduce the number of factors to eliminate the redundant

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



ones. Condition 2 is the sparse condition. Condition 3 is to ensure that the construction of $\mathbf{M}$ is effective to offset the impact of noise. Condition 4 assumes that the $r$ nonzero eigenvalues are distinct from each other so that we can distinguish $r$ factors. In this paper, we allow $\lambda_i (1 \leq i \leq r)$ to be diverging with $p$ and the entries of $\theta_i$ to depend on $\lambda_i$. Since we scale the loading matrix $\Theta$ such that $\Theta^T\Theta = \mathbf{I}_r$, which means that if the entries of $\theta_i$ is scaled by $\sqrt{\lambda_i}$, we multiple the variances of the factors with $\lambda_i$. This ensures that each component of $\mathbf{y}_t$ has a constant variance and is stationary.

Let $\widehat{\mathbf{M}}$ be the sample version of $\mathbf{M}$, that is

$$\widehat{\mathbf{M}} = \sum_{k=1}^{k_0} \widehat{\mathbf{\Sigma}}_{\mathbf{y}}(k) \widehat{\mathbf{\Sigma}}_{\mathbf{y}}(k)^T, \tag{2.3}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} (\mathbf{y}_{t+k} - \overline{\mathbf{y}}) (\mathbf{y}_t - \overline{\mathbf{y}})^T$, and $\overline{\mathbf{y}} = n^{-1} \sum_{t=1}^{n} \mathbf{y}_t$.

Under (2.2), we estimate $\Theta$ by imposing constraints on sparsity and column-orthogonality, which is equivalent to solving the following optimization problem:

$$\widehat{\Theta} = \arg\max_{\Theta} \operatorname{tr}\left(\Theta^T \widehat{\mathbf{M}} \Theta\right) \quad \text{subject to } \|\theta_i\|_1 \leq p^{\delta_i} \text{ and } \Theta^T\Theta = I_r, \tag{2.4}$$

where $\theta_i$ is the $i$-th column of $\Theta$. In $L_1$-penalty the coefficients are not required to be zeros, but their absolute magnitude must decay at a relatively

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



rapid rate. This kind of soft sparsity remains convex and is more realistic for many applications. (2.4) is similar to the SCoTLASS procedure, see Jolliffe et al. (2003). Such kind of problem is not convex and of high computational cost therefore. We finesse this problem by iteratively alternating between two subtasks: constrained rank-one variance maximization and orthogonal projection. In the $i$-th round we get a sparse vector $\widehat{\theta}_i$ by solving the following optimization problem

$$\widehat{\theta}_i = \arg\max_{\theta} \theta^T \widehat{\mathbf{M}} \theta, \text{ subject to } \|\theta\|_2 = 1, \|\theta\|_1 \leq p^{\delta_i}. \tag{2.5}$$

We adopt the Penalized Matrix Decomposition Analysis (PMA) for sparse PCA in Witten et al. (2009) to deal with (2.5), named the rank one maximization. Then we restrict $\widehat{\theta}_{i+1}$ to the orthogonal complementary space of $\widehat{\theta}_i$, where the projection operator is expressed as $I_p - \widehat{\theta}_i \widehat{\theta}_i^T$. Then, in the $(i+1)$-th round, the optimization problem is converted to

$$\begin{aligned} \widehat{\theta}_{i+1} = \arg\max_{\theta} \theta^T (I_p - \widehat{\theta}_i \widehat{\theta}_i^T) \widehat{\mathbf{M}} (I_p - \widehat{\theta}_i \widehat{\theta}_i^T) \theta, \\ \text{subject to } \|(I_p - \widehat{\theta}_i \widehat{\theta}_i^T)\theta\|_2 = 1, \|(I_p - \widehat{\theta}_i \widehat{\theta}_i^T)\theta\|_1 \leq p^{\delta_{i+1}}. \end{aligned} \tag{2.6}$$

Taking $(I_p - \widehat{\theta}_i \widehat{\theta}_i^T) \widehat{\mathbf{M}} (I_p - \widehat{\theta}_i \widehat{\theta}_i^T)$ as a whole, such a treatment is as the matrix deflation which modifies the matrix $\widehat{\mathbf{M}}$ to eliminate the influence of a given

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



vector $\widehat{\theta}_i$ (see White (1958) and Mackey (2008)). The complete steps of our method are shown in Algorithm 1.

**Algorithm 1** Orthogonal Projection Method for Sparse Factor Model

**Input:**
The auto-cross sample covariance matrix $\widehat{\mathbf{M}}$;
the number of factors $r$; the sparse index $\boldsymbol{\delta} = (\delta_1, ..., \delta_r)$.

**Output:**
Estimated sparse loadings matrix $\widehat{\Theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_r)$.

1: Initialize $i = 1$
2: **repeat**
3: **(Rank one maximization)**
(i) Initialize $\widehat{\theta}_i$ to have $L_2$-norm 1.
(ii) Iterate until convergence: $\widehat{\theta}_i \leftarrow \frac{S(\widehat{\mathbf{M}}\widehat{\theta}_i, \Delta)}{\|S(\widehat{\mathbf{M}}\widehat{\theta}_i, \Delta)\|_2}$,
where $\Delta = 0$ if $\|\widehat{\theta}_i\|_1 \leqslant p^{\delta_i}$; otherwise, $\Delta$ is chosen such that $\|\widehat{\theta}_i\|_1 = p^{\delta_i}$.
4: Return $\widehat{\theta}_i$
5: **(Orthogonal projection)**
Update $\widehat{\mathbf{M}}$ by $\widehat{\mathbf{M}} \leftarrow (I_p - \widehat{\theta}_i\widehat{\theta}_i^T)\widehat{\mathbf{M}}(I_p - \widehat{\theta}_i\widehat{\theta}_i^T)$.
Update $i \leftarrow i + 1$.
6: **until** $i = r + 1$

**Remark 1.** *(i) $S$ is the soft thresholding operator, which is, $S(a, \Delta) = \mathrm{sgn}(a)(|a| - \Delta)_+$, where $s > 0$ is a constant and $x_+$ is defined to equal $x$ if $x > 0$ and 0 if $x \leqslant 0$.*

*(ii) The algorithm is not sensitive to the initial estimator of $\widehat{\theta}_i$. When $p = o(n)$, it can be generated from the classical PCA, adopting the leading eigen vector of matrix $\widehat{\mathbf{M}}$. When $p > n$, it can be estimated from the Elastic net (Zou et al. (2006)), diagonal thresholding (Johnstone and Lu (2009))*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



*or other similar methods.*

*(iii) The algorithmic solution* $\widehat{\Theta}$ *from Algorithm 1 converges to the theoretical solution of* (2.4)*. This is based on two facts: the rank one maximization is convex, so the extremum is on the boundary according to KKT conditions; the loading space after orthogonal projection is a consistent estimator of the original one without projection, see Theorem 2 and Remark 3 of Zhang et al. (2023) for theoretical support.*

Algorithm 1 requires the number of factors $r$ and the sparse index $\boldsymbol{\delta}$ as inputs, but they are unknown in practice. We explore the estimation of $r$ and $\boldsymbol{\delta}$ in the following sections.

## 2.3 Determine the Number of Sparse Factors

In strong factor models, it is usually assumed that the first $r$ eigenvalues are $O_p(p)$, and $\lambda_{r+1}$ to $\lambda_p$ remain finite, see Fan et al. (2013). Many existing methods take advantage of the different convergence rates between the first $r$ eigenvalues and the others to determine the number of factors, for example, the ratio-based method in Lam and Yao (2012) and Ahn and Horenstein (2013) and the difference-based method in Onatski (2010). However, the methods for strong factor models are likely to fail in the presence of sparse

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



factors, since

$$\lambda_i = \theta_i^T \mathbf{M} \theta_i = \sum_k \sum_l \theta_{ik}\theta_{il}\sigma_{kl} \leq \max_{k,l} |\sigma_{kl}| \sum_k |\theta_{ik}| \sum_l |\theta_{il}| = O_p(p^{2\delta_i}), \quad (2.7)$$

where $0 \leq \delta_i \leq 1/2$ with each element of $\mathbf{M} = (\sigma_{kl})_{p\times p}$ be finite. Take the ratio-based method as an example, if both strong and sparse factors exist and the eigenvalues of them are $O_p(p)$ and $O_p(p^{1/3})$ respectively, then the biggest ratio of $\lambda_i/\lambda_{i+1}$ is not in $i = r$ but among $i < r$.

Under sparse factors, the first $r$ eigenvalues still diverge to infinity with $p \to \infty$ and others remain finite. We make use of this gap to separate $\lambda_1, \cdots, \lambda_r$ from $\lambda_{r+1}, \cdots, \lambda_p$ to obtain the estimated $\widehat{r}$. Specifically, let $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ be the eigenvalues of auto-cross covariance $\mathbf{M}$ in (2.3) in decreasing order. Define the null and the alternative as

$$H_0^{(i)} : \lambda_i = O(p^\alpha) \quad \text{v.s.} \quad H_A^{(i)} : \lambda_i = O(1),$$

Where $0 < \alpha \leq 1$. So $H_0^{(i)}$ should be accepted for $1 \leq i \leq r$, and be rejected from $i = r + 1$. Therefore, for each $\lambda_i$, we run the test in sequence until the $(i + 1)$-th hypothesis $H_0^{(i+1)}$ is rejected, then we can obtain $\widehat{r} = i$. But given that $\lambda_i = O(p^\alpha)$ under the null, we cannot use it directly, so we introduce the randomized sequential test in Trapani (2018). Specifically,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



for each $i$, we generate a series of artificial samples $\eta^{(i)}$ with distribution $N(0, \phi^{(i)})$ where $\phi^{(i)} \equiv \exp\left\{p^{-\gamma}(\widehat{\lambda}_i/\overline{\lambda})\right\}$, $\gamma$ is a tuning constant and $\overline{\lambda} = \sum_{i=1}^{p} \widehat{\lambda}_i/p$. Therefore, under the null, the sequence $\zeta_r^{(i)}(u) \equiv I\left(\eta^{(i)} \leq u\right)$ follows a Bernoulli distribution with $E\left\{\zeta_k^{(i)}(u)\right\} = \frac{1}{2}$ with $u$ extracted from a distribution $F(u)$ with support $\Omega \subset R\backslash\{0\}$. And under the alternative, for any $u \neq 0, E\left\{\zeta_r^{(i)}(u)\right\} \neq \frac{1}{2}$. For full details about the test statistics, one can refer to Trapani (2018).

**Remark 2.** *According to Theorem 1 of Lam and Yao (2012),* $|\widehat{\lambda}_i - \lambda_i| = O_p(p^2 n^{-1})$ *for* $i = r+1, \cdots, p$. *So on the construction of* $\phi^{(i)}$, *we also take* $\gamma \equiv 1 - \frac{1}{2}\frac{\ln n}{\ln p}$ *to make sure* $p^{-\gamma}\widehat{\lambda}_i$ *remains finite under the alternative when* $p > \sqrt{n}$.

In contrast to other methods for estimating the number of factors, this sequential test works under $p \ll n$, $p \asymp n$, and $p \gg n$. Theorem 4 shows that it produces a consistent estimator of $r$. In Section 4 we compare the sequence test with several other methods. Besides, when strong and sparse factors exist simultaneously, one can also use the two-step estimation in Lam and Yao (2012) or the local maximums of ratio in Zhang et al. (2023) to determine $\widehat{r}$. And the information criterion in Bai and Ng (2002) remains useful.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



### 2.4 Estimation of the sparse index

The sparse index reflects the strength of the factor, which is of great interest to empirical research. Recall that $p^{\delta_i}$ is the $L_1$ norm of $\theta_i$. If our proposed estimator $\widehat{\Theta}$ is consistent and recovers the loading matrix effectively, we can simply and directly estimate $\delta_i$ as $\widehat{\delta}_i = \log \|\widehat{\theta}_i\|_1 / \log p$. Meanwhile, $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_r)$ is an input of Algorithm 1. Without any prior information, we use cross-validation to determine the initial value of $\delta_i$, then run Algorithm 1 and use the output $\widehat{\theta}_i$ to obtain $\widehat{\delta}_i$, and iterate once again through Algorithm 1 with the estimated $\widehat{\delta}_i$ to obtain the final estimate of $\widehat{\theta}_i$, see Algorithm 2.

**Algorithm 2** Estimation for Sparse Factor Model

1: Determine the input $r$ by sequential test in Section 2.3
2: Use cross-validation to determine the input $\boldsymbol{\delta}$ and run Algorithm 1 to obtain $\widehat{\Theta}$.
3: Compute $\widehat{\delta}_i = \log \|\widehat{\theta}_i\|_1 / \log p$.
4: Run Algorithm 1 again with $\widehat{\boldsymbol{\delta}}$ to obtain the final $\widehat{\Theta}$.

## 3. Asymptotic theory

In this section, we consider the asymptotic properties of the estimator derived by (2.4) under $p$ varying with $n$. To this end, we supplement some technical conditions on the sparse factor model (2.1).

**Condition 5.** *As* $u \to \infty$, *it holds that* $\sup_t \max_{1 \le i \le p} P(|y_{i,t}| > u) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



$O\left\{u^{-2(l+\tau)}\right\}$ *for some constants* $l > 2$ *and* $\tau > 0$.

**Condition 6.** *The weak stationary process* $(\mathbf{x}_t, \epsilon_t)$ *is* $\alpha$*-mixing, that is, its mixing coefficients* $\alpha_{k,p} \to 0$ *as* $k \to \infty$, *where* $\alpha_{k,p} = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i, B \in \mathcal{F}_{i+k}^\infty} |P(A \cap B) - P(A)P(B)|$, *and* $\mathcal{F}_i^j$ *is the* $\sigma$*-field generated by* $\{(\mathbf{x}_t, \epsilon_t) : i \le t \le j\}$. *And* $\alpha_{k,p}$ *satisfies the condition* $\sup_{p \ge 1} \alpha_{k,p} = O\left\{k^{-(l-1)(l+\tau)/\tau}\right\}$, *as* $k \to \infty$, *where* $l$ *and* $\tau$ *are given in Condition 5.*

Conditions 5 and 6 ensure the Fuk-Nagaev type inequalities for $\alpha$-mixing processes, similarly to Chang et al. (2018). Our conditions are weaker than those in many papers where $\mathbf{y}_t$ is often required to be sub-gaussian, see Uematsu and Yamagata (2023) for example, but we only assume the tail decays at a polynomial order.

**Theorem 1.** *Let Conditions 1-6 hold and the eigenvalues of* $\widehat{\mathbf{M}}$ *be distinct. Denote* $\widehat{\theta}_1$ *as the leading vector of the theoretical solution to (2.4). Then*

$$\left\|\widehat{\theta}_1\widehat{\theta}_1^T - \theta_1\theta_1^T\right\|_F^2 = O_p\left(\frac{\sqrt{\lambda_1}p^{\delta_1}}{\lambda_1 - \lambda_2}\sqrt{\frac{p\log p}{n}}\right).$$

Theorem 2 is an extension to Theorem 1.

**Theorem 2.** *Let Conditions 1-6 hold, the eigenvalues of* $\widehat{\mathbf{M}}$ *be distinct and the sparse indices be the same, that is* $\delta_1 = \cdots = \delta_r = \delta$. *Denote* $\widehat{\Theta}$ *as the*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



*theoretical solution to* (2.4). *Then*

$$\left\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\right\|_F^2 = O_p\left(\frac{\sqrt{\lambda_1}p^\delta}{\lambda_r}\sqrt{\frac{p\log p}{n}}\right).$$

**Corollary 1.** *With the condition* $\lambda_i = O_p(p^{2\delta})$ *in* (2.7) *and* $\lambda_i - \lambda_{i-1} = O_p(p^{2\delta})$ *for* $1 \le i \le r$. *Theorem 2 can be simplified to* $\left\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\right\|_F^2 = O_p\left(\sqrt{\frac{p\log p}{n}}\right)$.

**Remark 3.** *(i) The F-norm* $\left\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\right\|_F$ *measures the distance between* $\mathcal{M}(\Theta)$ *and* $\mathcal{M}(\widehat{\Theta})$ *and is uniquely defined. It is similar to the commonly used statistic which measures the distance between two spaces in other literature, which is* $\mathcal{D}(\boldsymbol{O}_1, \boldsymbol{O}_2) = \left(1 - \frac{1}{\max(q_1,q_2)}\operatorname{tr}(\boldsymbol{O}_1\boldsymbol{O}_1'\boldsymbol{O}_2\boldsymbol{O}_2')\right)^{1/2}$, *where two orthogonal matrices* $\boldsymbol{O}_1$ *and* $\boldsymbol{O}_2$ *are of sizes* $p \times q_1$ *and* $p \times q_2$, *since*

$$\begin{aligned}\left\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\right\|_F^2 &= tr(\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T)(\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T)\\ &= (r+\widehat{r})\left\{1 - \frac{2}{r+\widehat{r}}tr(\widehat{\Theta}\widehat{\Theta}^T\Theta\Theta^T)\right\},\end{aligned}$$

*by* $tr(\Theta\Theta^T) = r$ *and* $tr(\widehat{\Theta}\widehat{\Theta}^T) = \widehat{r}$. *Then* $\left\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\right\|_F$ *is between 0 and* $\sqrt{r+\widehat{r}}$. *It is equal to 0 if the column spaces of* $\widehat{\Theta}$ *and* $\Theta$ *are the same and equal to* $\sqrt{r+\widehat{r}}$ *if they are orthogonal. When* $r = 1$, *it is equivalent to both the Euclidean distance between* $\theta_1$ *and* $\widehat{\theta}_1$, *and the magnitude of the sine of the angle between* $\theta_1$ *and* $\widehat{\theta}_1$, *see Vu and Lei (2012) for details.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



*(ii) When $p$ is fixed, $\lambda_1$ and $\lambda_2$ are finite. It follows from Theorem 1 that the convergence rate is the traditional root-n rate for fixed $p$, which is consistent with the Proposition 1 of Lam and Yao (2012). And when $p$ is fixed, the needed condition for mixing coefficients in Condition 6 is simplified to $\sum_{k=1}^{\infty} \alpha_{k,p}^{1-2/\gamma} < \infty$, where $\gamma > 2$ is a constant.*

*(iii) Theorem 2 shows that the sparser the factors are, the faster the convergence rate is, since we introduce the sparse penalty. In Lam et al. (2011), the convergence rate is $p^{1-\delta_0}/\sqrt{n}$ with $\|\theta_i\|_2^2 \asymp p^{\delta_0}$ and $0 \leq \delta_0 \leq 1$. It indicates that when $\delta_0 < 1/2$, namely the factor is weak enough in Lam et al. (2011), our sparse estimator has a faster convergence rate.*

**Theorem 3.** *If all the eigenvalues of $\Sigma_\epsilon$ are uniformly bounded from infinity, it holds that*

$$p^{-\frac{1}{2}} \left\| \widehat{\Theta}\widehat{\mathbf{x}}_t - \Theta \mathbf{x}_t \right\|_2 = O_p\left( p^{-\frac{1}{2}} \left\| \widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T \right\|_F + p^{-\frac{1}{2}} \right).$$

Theorem 3 specifies the convergence rate for the estimated factors.

**Theorem 4.** *Let Conditions 1–6 hold, and define the level of each individual test as $\alpha = \alpha(n,p)$. As $min(n,p) \to \infty$. If $\alpha(n,p) \to 0$, it holds that $P(\widehat{r} = r) = 1$ a.s. conditionally on the sample.*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



# 4. Numerical Results

## 4.1 Simulation

To illustrate the asymptotic properties in Section 3 above, we report some simulation results. We design two examples with different constructions for spare loadings as well as the factors, and one example with strong factors in the appendix. We measure the estimation error by the Frobenius norm discussed in Theorem 2 of Section 3, that is $\|\Theta\Theta^T - \widehat{\Theta}\widehat{\Theta}^T\|_F$. We also calculate the root-mean-square error (RMSE) given by RMSE $= \left(\frac{1}{np}\sum_{t=1}^{n}\left\|\widehat{\Theta}\widehat{\mathbf{x}}_t - \Theta\mathbf{x}_t\right\|^2\right)^{1/2}$.

**Example 1.** We consider the model with three sparse factors $\mathbf{y_t} = \Theta\mathbf{x_t} + \epsilon_t$, where $\epsilon_{t,i} \sim i.i.d.\ N(0,1)$. For each column of $\Theta$, we generate the first $h = [\ p^{\delta_*}\ ]$ ($\delta_* = 0.3$ and $0.5$) elements randomly from $N(0,1)$ and set the rest to zero. We perform SVD decomposition on the nonzero part of $\Theta$ for column orthogonality. Note that here we use $L_0$-constraint, differently from the sparse index $\delta$ defined in (2.2), but it can also generate sparse factors as desired. We adopt the generation scheme for intuition and convenience. The sparse index $\delta_i$ for each example is shown in the Table 2. We generate factors $x_i$ independently from AR(1) process $x_{i,t} = \eta_i x_{i,t-1} + e_{i,t}$, for $i = 1, 2, 3$, with coefficient $\eta_i$ equal to 0.8, 0.6 and 0.4

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



respectively. The noise term $e_{i,t}$ is independently sampled from $N(0,1)$ for all $i$ and $t$.

We calculate the results with $k_0 = 3$ in the definition of $\widehat{\mathbf{M}}$ and the true number of factors $r = 3$. We compare our method (labeled with "SFM") with other similar methods including the one in Lam et al. (2011) (denoted as "Eigen"), which performs eigen analysis on $\mathbf{M}$ in (2.3) without sparse constraints and PCA in Bai and Ng (2002), Sparse PCA (denoted as "SPCA") in Zou et al. (2006), the Sparse Orthogonal Factor Regression (denoted as "SOFAR") in Uematsu and Yamagata (2023) and Proximate PCA (denoted as "PPCA") in Pelger and Xiong (2022). We present the results for $p = 50, 100, 200, 500, 1000$, with $n$ selected to be (i) less than $p$, (ii) comparable to $p$, (iii) large than $p$, and conduct the simulation 200 times for each parameter pair. Figure 1 presents the results with $\delta_* = 0.3$, which distinctly indicates that basically the method with the sparse constraint is better than the method without it in the presence of sparse factors. And our SFM model performs better than others in most cases. Compared to low-dimensional cases, the performance of PPCA is not competitive in high dimensional ones, which may due to the fact that the PPCA method does not guarantee pointwise convergence. Besides, SOFAR and our SFM perform well in high dimensional situations, but the F-norm of SFM is

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



slightly better. The plot with $\delta_* = 0.5$ is shown in the appendix.

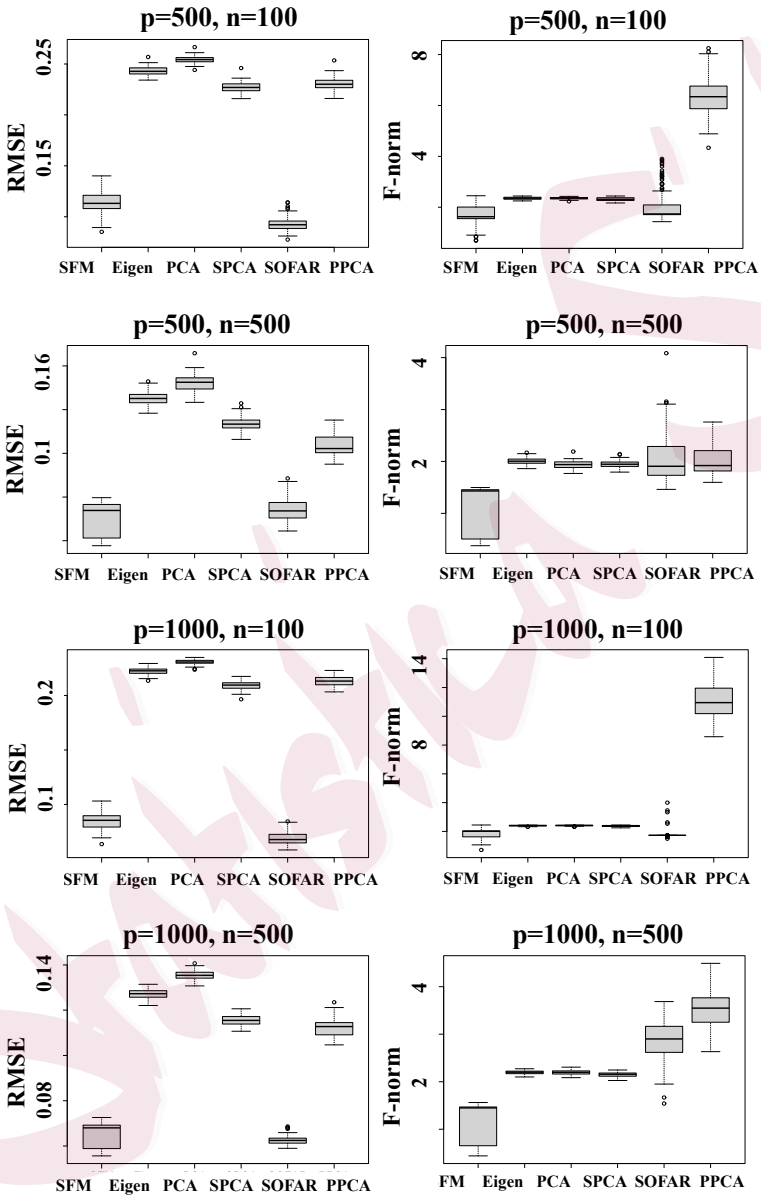



Figure 1: Boxplots for RMSE and F-norm of different methods. The plots are based on the simulated data set of $\delta_* = 0.3$ in Example 1 of Section 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



Table 1: The average number of zeros in the estimated loading matrix. The data is generated from Example 1 with $\delta^* = 0.5$.

| | | | Method | | | | | |
|---|---|---|---|---|---|---|---|---|
| p | n | #{0} | SFM | Eigen | PCA | SPCA | SOFAR | PPCA |
| 50 | 100 | 43 | 38.7 | 0 | 0 | 42.5 | 48.66 | 0 |
| | 200 | 43 | 35.18 | 0 | 0 | 47.62 | 47.1 | 0 |
| | 500 | 43 | 37.46 | 0 | 0 | 48.8 | 46.6 | 0 |
| 100 | 100 | 90 | 83.38 | 0 | 0 | 72.34 | 96.56 | 0 |
| | 200 | 90 | 82.86 | 0 | 0 | 91.98 | 97.8 | 0 |
| | 500 | 90 | 78.66 | 0 | 0 | 96.56 | 89.7 | 0 |
| 200 | 100 | 186 | 171.76 | 0 | 0 | 132.2 | 195.26 | 0 |
| | 200 | 186 | 173.96 | 0 | 0 | 159.5 | 195.56 | 0 |
| | 500 | 186 | 172.2 | 0 | 0 | 194.86 | 191.2 | 0 |
| 500 | 100 | 478 | 462.26 | 0 | 0 | 403.26 | 491.84 | 0 |
| | 200 | 478 | 452.54 | 0 | 0 | 353.86 | 486.76 | 0 |
| | 500 | 478 | 448.2 | 0 | 0 | 438.06 | 492.9 | 0 |

To better observe the effects of different methods, Table 1 demonstrates the average number of 0 in each column of the load matrix obtained by the different methods, where #{0} denotes the true the number of zeros in each set of parameters. The method without sparsity, including "Eigen" and "PCA", naturally has no 0 elements. The PPCA method performs a factorial regression after thresholding, so there are also no 0 elements. Our method slightly underestimates the number of 0 elements due to the use of the $L_1$-norm penalty, and the SOFAR method shows a slight overestimation.

**Example 2.** In this example we construct the sparse factor loading matrix $\Theta$ by diagonal blocks, that is $\Theta = diag(\Theta_1, \Theta_2, \cdots, \Theta_d)$. For convenience, we still set up a three-factor model. Let $d = 3$ and each block to be a vector with dimensions of $0.4p$, $0.3p$, $0.3p$ correspondingly. All elements of the blocks are i.i.d generated from $N(0, 1)$. Table 2 shows the sparse

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



index of each factor.

Table 2: Sparse indices for three factors in our simulation of the Example 1 and Example 2

| | Exaple 1 | | | | | | Example 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\delta_* = 0.3$ | | | $\delta_* = 0.5$ | | | | | |
| p | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_1$ | $\delta_2$ | $\delta_3$ |
| 50 | 0.12 | 0.12 | 0.13 | 0.21 | 0.22 | 0.21 | 0.31 | 0.28 | 0.25 |
| 100 | 0.09 | 0.08 | 0.11 | 0.19 | 0.20 | 0.23 | 0.34 | 0.31 | 0.33 |
| 200 | 0.12 | 0.11 | 0.12 | 0.22 | 0.21 | 0.20 | 0.36 | 0.34 | 0.34 |
| 500 | 0.11 | 0.12 | 0.12 | 0.21 | 0.21 | 0.21 | 0.37 | 0.36 | 0.36 |
| 1000 | 0.13 | 0.13 | 0.11 | 0.21 | 0.22 | 0.21 | 0.40 | 0.38 | 0.38 |

We generate a moving average factors $\mathbf{x}_t = (x_{1,t}, x_{2,t}, x_{3,t})$, defined by $x_{1,t} = \omega_t$, $x_{2,t} = \omega_{t-1}$, $x_{3,t} = \omega_{t-2}$, where $\omega_t = 0.2z_{t-1} + z_t$, and $z_t$ are independent $N(0,1)$ random variables.

We calculate the mean and standard deviation of the F-norm, RMSE and forecast error (FE) for different methods. The results are reported in the Supplementary Material. Next, we estimate the number of factors. We adopt the factor loadings in Example 2 and consider two scenarios for the factors $\mathbf{x}_t$. In Scenario I, $\mathbf{x}_t$ is the same as the one in Example 1, which consists of three independent AR(1) process with coefficient 0.4, 0.6 and 0.8 respectively and independent $N(0,1)$ innovations. In Scenario II, $\mathbf{x}_t$ is the moving average series in Example 2. We compare the sequential test with the ratio-base method in Lam and Yao (2012), which is $\widehat{r} = \arg\max_k \widehat{\lambda}_k/\widehat{\lambda}_{k+1}$, and the BIC-type information criterion of Bai and Ng (2002) given by $\widehat{r} = \arg\min_k \left\{ \log\left(p^{-1}T^{-1}\sum_{j=1}^{p} \|\widehat{\epsilon}_j\|_2^2\right) + k\left(\frac{n+p}{np}\right)\log\left(\frac{np}{n+p}\right)\right\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



We denote the sequential test, the ratio-based method and the information criterion method as "SeqTest, Ratio, IC" relatively, and report the performance of different methods in Table 3.

Table 3 shows that the ratio-based method underperforms with sparse factors, especially in the $p > n$ cases. The conclusion aligns with the discussion in Lam and Yao (2012). In Scenario I, the ratio-based method and IC fail to capture the signal of three factors, but the sequential test outperforms. In Scenario II, the three methods perform almost equally, but the computational cost of IC is much higher than that of the sequential test and ratio-based method.

### 4.2 Choose of tuning parameter

In this section we discuss the choice of tuning parameter $s_i = p^{\delta_i}$ (or $\delta_i$ equally). First we range $s$ from 1 to $\sqrt{p}$ and examine the finite-sample performance in different tuning parameters. We take both RMSE and $F$-norm $\|\widehat{\Theta}\widehat{\Theta}^T - \Theta\Theta^T\|_F$ as criteria. Figure 2 shows the results in different combinations of $(n, p, \delta_*)$. It is shown that the error decreases first and then increases with $s$ grows up, and reaches the optimum near the true value of $s$ under both RMSE and the F-norm. The monotonicity in both sides of truth value indicates the validity of our estimation. In practice,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



Table 3: $Pr(r = \widehat{r})$ and mean$(\widehat{r})$ of different method for determine the number of factors. The true factor number is $r = 3$.

| | | SeqTest | | Ratio | | IC | |
|---|---|---|---|---|---|---|---|
| p | n | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ |
| 100 | 100 | 0.60 | 2.58 | 0.30 | 2.30 | 0.16 | 5.99 |
| | 200 | 0.34 | 2.31 | 0.18 | 2.18 | 0.00 | 7.29 |
| | 500 | 0.03 | 2.02 | 0.02 | 2.02 | 0.02 | 5.86 |
| | 1000 | 0.04 | 2.04 | 0.05 | 2.05 | 0.00 | 6.78 |
| | 1500 | 0.08 | 2.03 | 0.00 | 2.00 | 0.00 | 7.99 |
| 200 | 100 | 0.80 | 2.79 | 0.24 | 2.24 | 0.01 | 6.96 |
| | 200 | 0.47 | 2.43 | 0.15 | 2.15 | 0.00 | 7.90 |
| | 500 | 0.61 | 2.59 | 0.05 | 2.05 | 0.00 | 8.76 |
| | 1000 | 0.16 | 2.11 | 0.02 | 2.02 | 0.00 | 9.29 |
| | 1500 | 0.21 | 2.21 | 0.01 | 2.01 | 0.00 | 9.46 |
| 500 | 100 | 0.76 | 3.22 | 0.20 | 2.44 | 0.19 | 4.70 |
| | 200 | 0.96 | 2.93 | 0.14 | 2.26 | 0.05 | 5.08 |
| | 500 | 0.84 | 2.80 | 0.05 | 2.05 | 0.00 | 6.98 |
| | 1000 | 0.53 | 2.50 | 00.00 | 2.00 | 0.00 | 8.15 |
| | 1500 | 0.35 | 2.33 | 0.01 | 2.01 | 0.00 | 7.89 |
| | 2000 | 0.44 | 2.43 | 0.03 | 2.03 | 0.00 | 8.83 |
| 1000 | 100 | 0.24 | 3.68 | 0.22 | 2.88 | 0.02 | 1.91 |
| | 200 | 0.44 | 3.47 | 0.10 | 2.34 | 0.07 | 4.43 |
| | 500 | 0.95 | 2.89 | 0.04 | 2.08 | 0.00 | 9.57 |
| | 500 | 0.95 | 2.89 | 0.04 | 2.08 | 0.00 | 9.57 |
| | 1000 | 0.96 | 2.93 | 0.03 | 2.03 | 0.00 | 9.85 |
| | 1500 | 0.92 | 2.90 | 0.00 | 2.00 | 0.00 | 9.98 |
| | 2000 | 0.76 | 2.72 | 0.00 | 2.00 | 0.00 | 10 |

(a) Scenario I: MA factors

| | | SeqTest | | Ratio | | IC | |
|---|---|---|---|---|---|---|---|
| p | n | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ | $Pr(r=\widehat{r})$ | mean$(\widehat{r})$ |
| 50 | 50 | 0.39 | 1.89 | 0.80 | 2.68 | 0.84 | 3.17 |
| | 100 | 0.37 | 1.82 | 0.58 | 2.40 | 0.95 | 3.05 |
| | 200 | 0.59 | 2.18 | 0.96 | 2.91 | 0.46 | 3.54 |
| | 500 | 0.98 | 2.96 | 0.98 | 2.98 | 0.58 | 3.42 |
| | 1000 | 1.00 | 3.00 | 0.97 | 2.94 | 0.02 | 3.98 |
| | 2000 | 1.00 | 3.00 | 0.93 | 2.85 | 0.00 | 4.00 |
| 100 | 50 | 0.27 | 1.67 | 0.77 | 2.98 | 0.89 | 3.10 |
| | 100 | 0.47 | 2.01 | 0.89 | 2.75 | 0.42 | 3.58 |
| | 200 | 0.76 | 2.54 | 0.99 | 2.97 | 1.00 | 3.00 |
| | 500 | 1.00 | 3.00 | 0.96 | 2.94 | 1.00 | 3.00 |
| | 1000 | 1.00 | 3.00 | 0.95 | 2.89 | 1.00 | 3.00 |
| | 1500 | 1.00 | 3.00 | 0.98 | 2.94 | 1.00 | 3.00 |
| 200 | 50 | 0.28 | 1.63 | 0.32 | 3.23 | 1.00 | 3.00 |
| | 100 | 0.40 | 1.89 | 0.76 | 2.78 | 1.00 | 3.00 |
| | 200 | 0.68 | 2.36 | 0.96 | 2.91 | 0.93 | 3.07 |
| | 500 | 0.86 | 2.72 | 0.98 | 2.96 | 1.00 | 3.00 |
| | 1000 | 1.00 | 3.00 | 0.97 | 2.94 | 1.00 | 3.00 |
| | 1500 | 1.00 | 3.00 | 0.96 | 2.90 | 1.00 | 3.00 |
| 500 | 200 | 0.65 | 2.31 | 0.20 | 3.91 | 1.00 | 3.00 |
| | 500 | 0.98 | 2.96 | 0.94 | 2.89 | 1.00 | 3.00 |
| | 1000 | 1.00 | 3.00 | 0.96 | 2.94 | 1.00 | 3.00 |
| | 1500 | 1.00 | 3.00 | 0.98 | 2.95 | 1.00 | 3.00 |

(b) Scenario II: AR factors

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



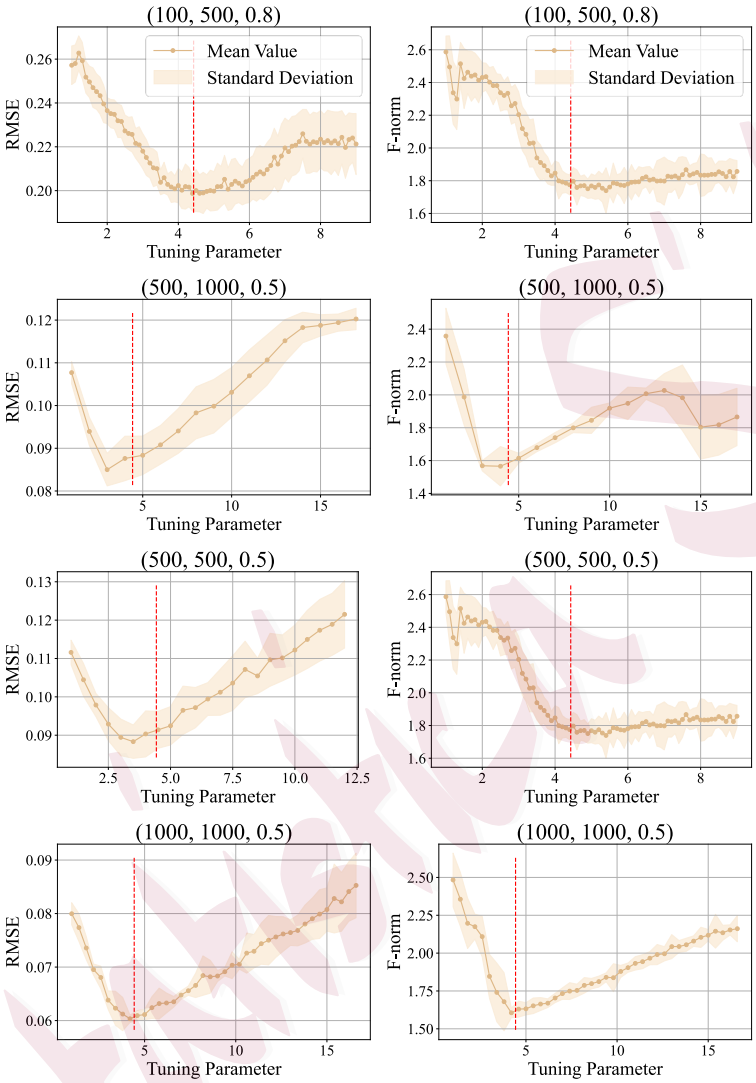



Figure 2: Errors as functions of tuning parameter $s$ under different $(p, n, \delta^*)$. The red line is the true value of $s$. The plot is based on the simulated data set in Example 1 of Section 4.

both RMSE and F-norm are unknown. We use Algorithm 2 to choose the tuning parameter and estimate $\boldsymbol{\delta}$. The mean and standard deviation of $\widehat{\delta}_1$ of Example 2 are presented in Table 4. Generally, they are sufficiently

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



accurate, but there is a slight tendency to overestimate. The precision also increases as $n$ increases.

Table 4: Performance for the estimation of sparse index $\delta_1$. The data is generated from Example 2 of Section 4.

| | p=50, $\delta_1$=0.31 | | p=100, $\delta_1$=0.34 | | p=200, $\delta_1$=0.36 | | p=500, $\delta_1$=0.37 | |
|---|---|---|---|---|---|---|---|---|
| n | mean($\widehat{\delta}$) | sd $\times$100 | mean($\widehat{\delta}$) | sd $\times$100 | mean($\widehat{\delta}$) | sd $\times$100 | mean($\widehat{\delta}$) | sd $\times$100 |
| 50 | 0.32 | 0.00 | 0.35 | 0.00 | 0.37 | 0.00 | 0.39 | 0.00 |
| 100 | 0.32 | 0.46 | 0.35 | 0.07 | 0.37 | 0.04 | 0.39 | 0.00 |
| 200 | 0.32 | 0.00 | 0.35 | 0.15 | 0.37 | 0.40 | 0.39 | 0.00 |
| 500 | 0.32 | 0.60 | 0.34 | 1.18 | 0.37 | 0.26 | 0.39 | 0.36 |
| 1000 | 0.32 | 0.92 | 0.35 | 0.07 | 0.37 | 0.36 | 0.39 | 0.19 |
| 1500 | 0.30 | 2.21 | 0.34 | 0.44 | 0.37 | 0.38 | 0.38 | 0.44 |

## 5. Real Data Analysis

This section analyzes two real datasets in both geography and finance to demonstrate our sparse factor model and newly proposed estimation method.

### 5.1 Real Data Example 1: Average Sea Surface Air Pressure.

We analyze the records of monthly average sea surface air pressure (in Pascal) from January 1958 to December 2001 (i.e., 528 months in total) over a $22 \times 39$ grid in the North Atlantic Ocean. Let $P_t(u, v)$ denote the air pressure in the $t$-th month at the location $(u, v)$, where $u = 1, \ldots, 22, v = 1, \ldots, 39$ and $t = 1, \ldots, 528$. We first subtract each data point by the monthly mean over the 44 years at its location: $\frac{1}{44}\sum_{i=1}^{44} P_{12(i-1)+j}(u, v)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



where $j = 1, \ldots, 12$, representing the 12 different months over a year. We then line up the new data over $22 \times 39 = 858$ grid points as a vector $\mathbf{y}_t$, so that $\mathbf{y}_t$ is a multi-variate time series with $p = 858$ dimensions and $n = 528$ observations. Different from Lam and Yao (2012), this is a situation with $p > n$.

To fit the sparse factor model to $\mathbf{y}_t$, we need to determine the number of factors first. We run the sequential test in Section 2.3 and the ratio-based method in Lam and Yao (2012). Figure 3 reports the p-value of the sequential test and the value of ratio $\widehat{\lambda}_i/\widehat{\lambda}_{i-1}$. In the sequential test, there is an obvious gap between the first 5 p-values and the latter ones. The red line is $\alpha = 0.05$, and we reject the null until $i > 5$ and $\widehat{r} = 5$ therefore. For the ratio-based method, the largest ratio occurs at $i = 1$, but Figure 3 also shows that the first local maximum of the ratio occurs in $i = 5$ (the red vertical line in the figure), this is due to the differences in strength among the common factors, see Remark 3 of Zhang et al. (2023). In other words, there may exist other four sparse factors in addition. The conclusion is in agreement with the result by the sequential test, indicating that the method we used is valid.

We present the Forecast Error (FE) for different methods in Table 5, where $FE = p^{-1/2}\|\widehat{\mathbf{y}}_T^{(1)} - \mathbf{y}_T\|_2$, and $\widehat{\mathbf{y}}_T^{(1)} = \widehat{\Theta}\widehat{\mathbf{x}}_T^{(1)}$. $\widehat{\mathbf{x}}_t^{(1)}$ is the one-step

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



predictor for $\mathbf{x}_T$ based on $\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_{T-1}$, on which a VAR process is assumed. We also select the initial value of tuning parameters on Algorithm 2 using the principle of optimizing FE.

Table 5: Forecast Error (FE) for different methods in Real Data Example 1

| Method | SFM | PCA | SPCA | SOFAR | PPCA |
|---|---|---|---|---|---|
| FE | **611.63** | 637.57 | 649.31 | 617.91 | 635.22 |

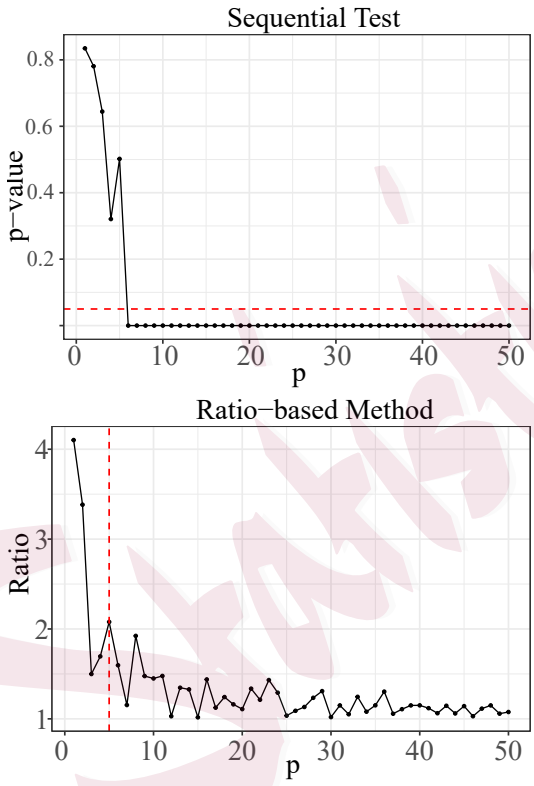



Figure 3: The estimate of factor numbers

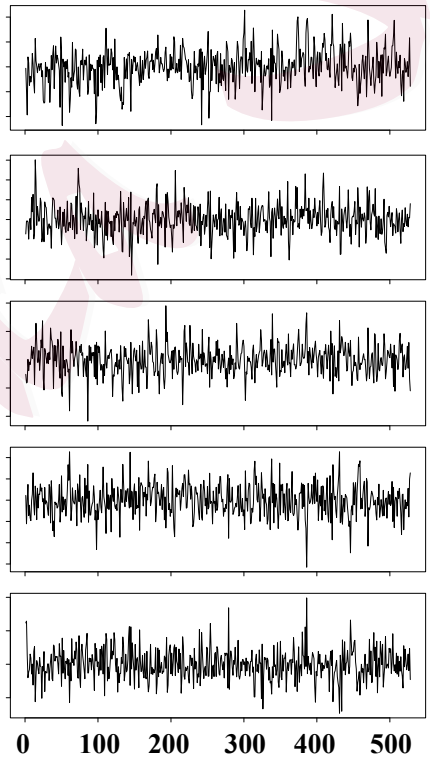



Figure 4: Time series plot of estimated factors

Table 5 shows the validity of our SFM model. We present the time series plots for the five estimated factors in Figure 4. The five selected factors

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



explain 96.84% of the total variance. Figure 5 presents the factor loading surfaces, revealing some regional patterns. For instance, the first factor is relatively strong, with positive effect in the north and negative effect in the middle. The second factor is mainly driving force in the central north. The third factor affects the dynamics of the middle east and west in opposite directions. The fifth factor is so sparse that it has no force on all regions except for the north.

**Remark 4.** *We test the stationarity of the five estimated factors. Both the ADF test and the ACF plot indicate that the factors are stationary. Therefore this data satisfies Condition 1.*

### 5.2 Real Data Example 2: Stock Portfolios.

In this example, we model the returns of different assets, which are the stock portfolio constructed in different ways. We apply our method on two datasets (DS). The portfolios are formed according to size (Market Equity) and momentum (prior (2-12) return) in DS 1, while according to size and long-term reversal (prior (13-60) return) in DS 2. Specifically, in DS 1, all stocks are ranked into quintiles (i.e., five groups) based on their size. Within each quintile, stocks are further ranked into quintiles based on their momentum, resulting 25 distinct portfolios. In each dataset, the portfolios

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



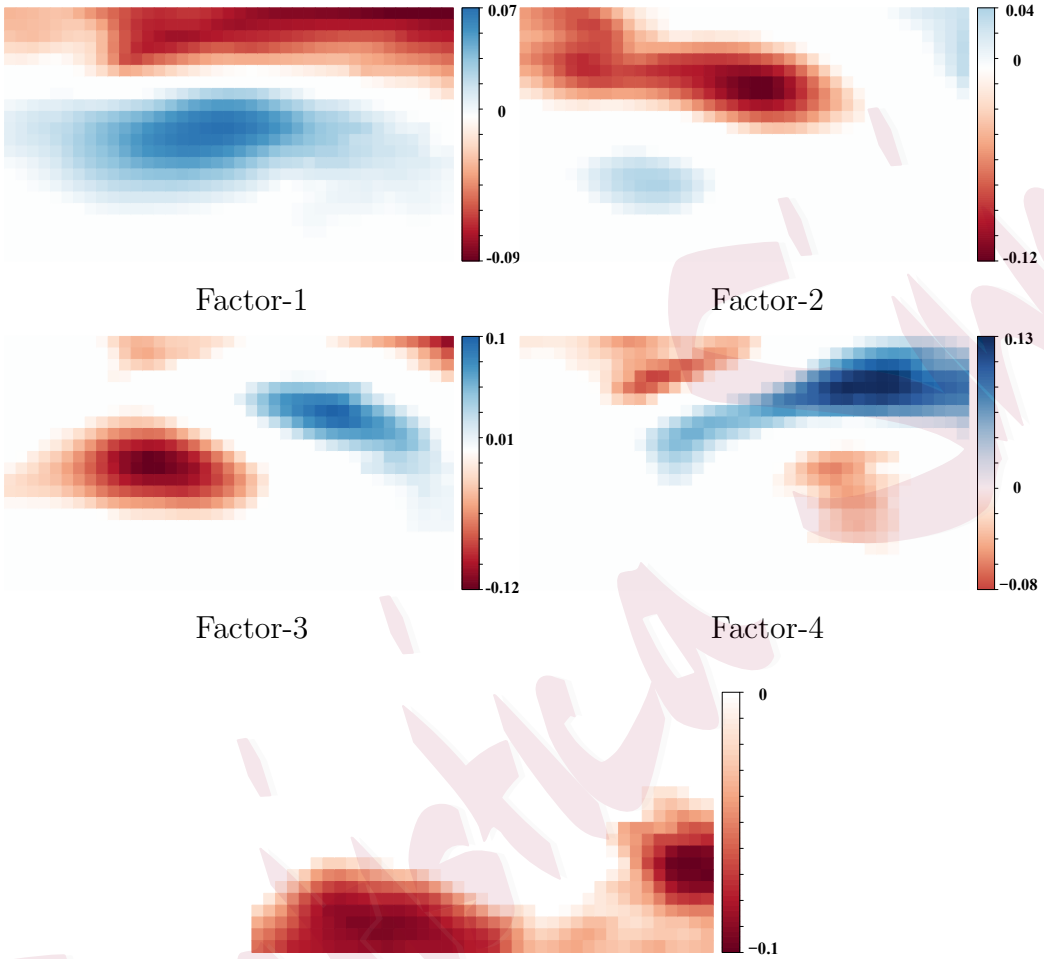



Figure 5: Factor loading surface of the first to fifth factors

monthly constructed include NYSE, AMEX, and NASDAQ stocks with prior return data. For further details regarding data construction, please refer to Kenneth R. French's website.

Instead of directly applying our sparse factor model on the data, we

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



model the asset returns in two parts: one is the known factors part, the other is the latent sparse factor part, represented as $\mathbf{y}_t = D\mathbf{z}_t + \eta_t = D\mathbf{z}_t + \Theta\mathbf{x}_t + \epsilon_t$, where $\mathbf{z}_t$ denotes the known factors and $D$ signifies the regression coefficient matrix. The sparse factor model is represented by $\eta_t$.

We construct $\mathbf{y}_t$ in two different examples. In Example 1, $\mathbf{y}_t s$ are monthly returns of 25 portfolios sorted in DS1 from January, 2008 to February, 2023, resulting in a dimension of $p = 25$ and a sample size of $n = 182$. The known factor $\mathbf{z}_t$ is designed as the market factor, aligning with the Capital Asset Pricing Model (CAPM). In Example 2, we merge the portfolios sorted by DS1 with those sorted by DS2 from January 2001 to February 2023, leading to a $\mathbf{y}_t$ with $p = 50$ and $n = 266$. In this case, $\mathbf{z}_t$ is designed as the Fama-French three factors (FF3).

In each example, we first estimate $D$ from regressing $\mathbf{y}_t$ on $\mathbf{z}_t$ asset by asset (see Chang et al. (2015)) and obtain $\widehat{\eta}_t = \mathbf{y}_t - \widehat{D}\mathbf{z}_t$. We then fit our sparse factor model (SFM) on $\widehat{\eta}_t$ to derive $\widehat{\Theta}$, $\widehat{\mathbf{x}}_t$. The sequential test is employed too. In both examples, the estimated factor number $\widehat{r} = 2$. The p-value of the test is provided in Figure 6.

Figure 7 displays the heat maps of the combined loading matrix $(\widehat{D}, \widehat{\Theta})$, which reveals some noticeable characteristics. In the CAPM+SFM model, the market factor exerts a strong positive influence on all portfolios, while

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



the latent sparse factors have no or negative impacts on these. In the FF3+SFM model, after taking out the effects of three known factors, the remaining factors become quite sparse.

Table 7 reveals that the known factors and the latent sparse factors within the same model are uncorrelated, signifying that SFM can capture additional information. Specifically, in the CAPM+SFM model, the two sparse factors are correlated with HML and SMB, indicating that the latent factors identified share similarities with the well-established factors. In asset pricing field, there exist skeptical concerns on if HML and SMB are true risk factors. By employing our model, however, we can discover similar or more influential factors, although their practical implications remain to be elucidated. Furthermore, Table 8 illustrates that our CAPM+SFM model outperforms the FF3 model in terms of sum of residuals, specifically under the same premise of utilizing three factors, thereby highlighting the superiority of our methodology.

Table 6: Forecast Error (FE) for different methods in Real Data Example 2

| | Example 1: CAPM+ | | | Example 2: FF3+ | | |
|---|---|---|---|---|---|---|
| Method | SFM | SOFAR | PPCA | SFM | SOFAR | PPCA |
| FE | 7.15 | 7.50 | 7.51 | 2.35 | 2.36 | 2.36 |

The goodness of fit for our model compared to others is assessed by scatter plots of $\overline{\mathbf{y}}_t$ against the mean of $\widehat{\mathbf{y}}_t$, as depicted in Figure 8. With the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



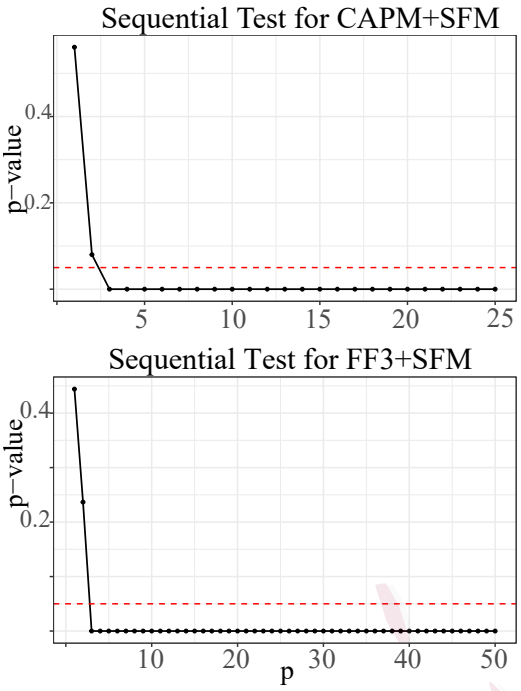



Figure 6: Estimation for the Factor Number

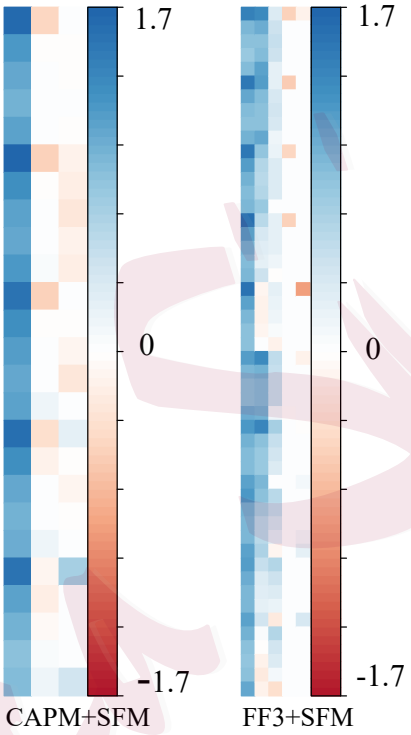



Figure 7: Loading Surface for Two Examples

Table 7: Correlation between the Sparse Factors and the Fama-French Factors

| | | Fama French Factors | | |
|---|---|---|---|---|
| | | Mkt-RF | HML | SMB |
| CAPM+SFM | SF-1 | -0.015 | -0.333 | -0.429 |
| | SF-2 | 0.025 | -0.200 | -0.607 |
| FF3+SFM | SF-1 | -0.146 | -0.008 | -0.003 |
| | SF-2 | -0.003 | -0.017 | -0.009 |

Note: SF-1 refers to the first sparse factor in the model.

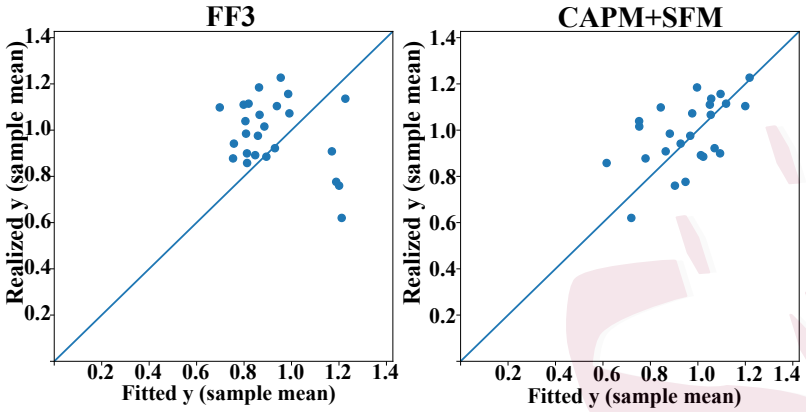
Table 8: Sum of Residuals from Different Models

| Example 1 | | | Example 2 | |
|---|---|---|---|---|
| CAPM | CAMP+SFM | FF3 | FF3 | FF3+SFM |
| 1.730 | 0.518 | 1.594 | 4.236 | 3.301 |

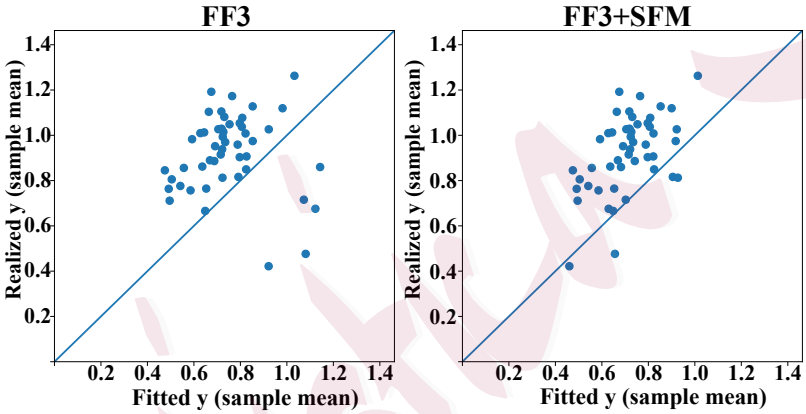
SFM incorporated, the scatter points are observed to be closer to the $y = x$ line, indicating an enhanced fitting performance of the model. It should be

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219






(a) Example 1




(b) Example 2

Figure 8: Scatter Plots for $\overline{\mathbf{y}}_t$ against the mean of $\widehat{\mathbf{y}}_t$

noted that though comparing the fitting error of FF3 with FF3+SFM is not fair as the latter has more factors, we aim to illustrate that after taking out the effects of known factors, if there are remaining effects (both serially and cross sectionally), we can always apply SFM to further model them.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



## 6. Conclusions and discussions

Factor model has been extensively studied in high dimensional time series. However, relatively little is known for the case where some factors take effect only on part of the series, which is the so-called sparse factor model. In this paper, we propose a sparse factor model with an algorithm. We estimate the sparse loading space and factors by auto-cross covariance over different time lags and $L_1$-penalty, and determine the factor number based on a randomized procedure. It is shown that the loading space and the factor are estimated consistently, and a sparser loading matrix leads to a faster convergence rate.

In this paper, we only consider $\mathbf{y}_t = \Theta\mathbf{x_t} + \epsilon_t$, but if the loading matrix is extremely sparse, some components of $\mathbf{y}_t$ may be driven totally by noise and the model would become less interpretable. To address this issue, one can add exogenous variables in the sparse model, allowing the latent factors and exogenous variables to jointly account for the observations, see Ando and Bai (2016), Bai (2009) and our Real Example 2 in Section 5. After detrending the exogenous variables term, we can then apply the proposed method to estimate the factors and loading space.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



## Supplementary Materials

The supplementary materials contains some technical lemmas, the proof of Theorem 1-4 of the main article, and some detailed tables and figures of simulation results which are discussed in the main paper.

## Acknowledgments

We would like to thank the Co-Editor, Associate Editor and two anonymous referees for their critical comments and thoughtful suggestions, which led to a much improved version of this paper. This research was supported in part by grants from NSFC, China (Nos. 12171427, U21A20426) and National Key R&D Program of China (2024YFA1013502).

## References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, 235(2):1893–1916.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408.

Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

Chang, J., Chen, C., Qiao, X., and Yao, Q. (2024). An autocovariance-based learning framework for high-dimensional functional time series. *Journal of Econometrics*, 239(2):105385.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



Chang, J., Guo, B., and Yao, Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, 189(2):297–312.

Chang, J., Guo, B., and Yao, Q. (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics*, 46(5):2094–2124.

Chang, J., He, J., Yang, L., and Yao, Q. (2023). Modelling matrix time series via a tensor cp-decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):127–148.

Chudik, A., Pesaran, M. H., and Tosetti, E. (2011). Weak and strong cross section dependence and estimation of large panels. *The Econometrics Journal*, 14(1):C45–C90.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801.

Mackey, L. (2008). Deflation methods for sparse pca. In *Advances in Neural Information Processing Systems*, volume 21.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77(5):1447–1479.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642.

Pelger, M. and Xiong, R. (2022). Interpretable sparse proximate factors for large dimensions. *Journal of Business & Economic Statistics*, 40(4):1642–1664.

Trapani, L. (2018). A randomized sequential procedure to determine the number of factors. *Journal of the American Statistical Association*, 113(523):1341–1349.

Uematsu, Y. and Yamagata, T. (2023). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):213–227.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0219



Vu, V. and Lei, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248.

White, P. A. (1958). The computation of eigenvalues and eigenvectors of a matrix. *Journal of the Society for Industrial and Applied Mathematics*, 6(4):393–437.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Zhang, B., Pan, G., Yao, Q., and Zhou, W. (2023). Factor modeling for clustering high-dimensional time series. *Journal of the American Statistical Association*, 0(0):1–12.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Xiaoran Wu

E-mail: xiaoran@zju.edu.cn

Baojun Dou

E-mail: baojun.dou@cityu.edu.hk

Rongmao Zhang

E-mail: rmzhang@zju.edu.cn