Journal of Econometrics 229 (2022) 201–217

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Projected estimation for large-dimensional matrix factor models

Long Yu [a], Yong He [b], Xinbing Kong [c,*], Xinsheng Zhang [a]

[a] Fudan University, Shanghai, 200433, China
[b] Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan, 250100, China
[c] Nanjing Audit University, Nanjing, 211815, China

ARTICLE INFO

Article history:
Received 3 April 2020
Received in revised form 5 April 2021
Accepted 7 April 2021
Available online 24 April 2021

JEL classification:
C23
C33
C38
C55

Keywords:
Matrix factor model
Vector factor model
Column covariance matrix
Row covariance matrix

ABSTRACT

In this study, we propose a projection estimation method for large-dimensional matrix factor models with cross-sectionally spiked eigenvalues. By projecting the observation matrix onto the row or column factor space, we simplify factor analysis for matrix series to that of a lower-dimensional tensor. This method also reduces the magnitudes of the idiosyncratic error components, thereby increasing the signal-to-noise ratio, because the projection matrix linearly filters the idiosyncratic error matrix. We theoretically prove that the projected estimators of the factor loading matrices achieve faster convergence rates than existing estimators under similar conditions. Asymptotic distributions of the projected estimators are also presented. A novel iterative procedure is given to specify the pair of row and column factor numbers. Extensive numerical studies verify the empirical performance of the projection method. Two real examples in finance and macroeconomics reveal factor patterns across rows and columns, which coincide with financial, economic, or geographical interpretations.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Time-series models with factor structures are widely used in fields such as finance, macroeconomics, and machine learning (e.g. Chamberlain and Rothschild (1983), Fama and French (1993), Stock and Watson (2002a) and Fan et al. (2015)). With the increasing complexity of data structure, the factor models for time series have experienced three stages: factor models for multivariate time series of fixed dimension (e.g. Ross (1977)), large-dimensional vector factor models (e.g. Forni et al. (2000), Bai and Ng (2002), Bai (2003), Ahn and Horenstein (2013), Fan et al. (2013), Kong et al. (2019) and He et al. (2020)), and matrix factor models (e.g. Wang et al. (2019) and Chen et al. (2020a)). The matrix factor model not only reveals the serial dynamics for a panel of variables, but also explores the spatial correlations among entries of the observation matrix in a parsimonious way.

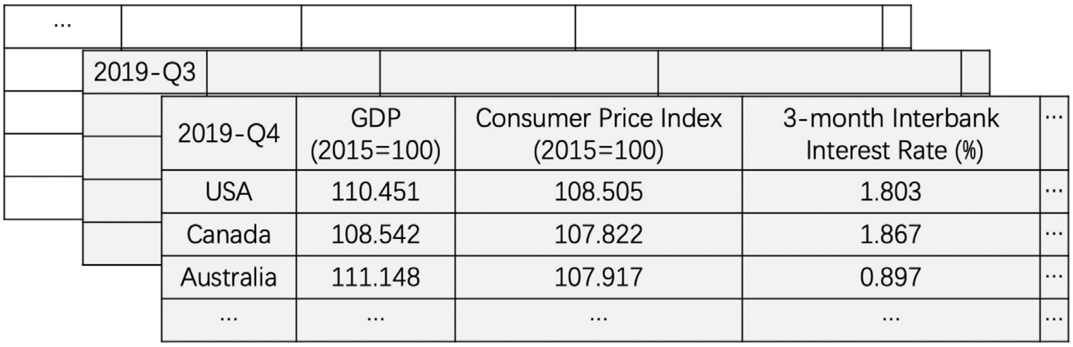
Wang et al. (2019) were the first to introduce a factor model for matrix time series. For ease of presentation, let $\mathbf{X}_t$ be a $p_1 \times p_2$ matrix of variables observed at t. Wang et al. (2019) factorized $\mathbf{X}_t$ as

$$(\mathbf{X}_t)_{p_1\times p_2} = (\mathbf{R})_{p_1\times k_1}(\mathbf{F}_t)_{k_1\times k_2}(\mathbf{C}^\top)_{k_2\times p_2} + (\mathbf{E}_t)_{p_1\times p_2}, \quad t = 1,\ldots,T, \tag{1.1}$$

where $\mathbf{R}$ is the $p_1 \times k_1$ row factor loading matrix exploiting the variations of $\mathbf{X}_t$ across the rows, $\mathbf{C}$ is the $p_2 \times k_2$ column factor loading matrix reflecting the differences across the columns of $\mathbf{X}_t$, $\mathbf{F}_t$ is the common factor matrix for all cells in

[*] Corresponding author.
E-mail addresses: fduyulong@163.com (L. Yu), heyong@sdu.edu.cn (Y. He), xinbingkong@126.com (X. Kong), xszhang@fudan.edu.cn (X. Zhang).

https://doi.org/10.1016/j.jeconom.2021.04.001
0304-4076/© 2021 Elsevier B.V. All rights reserved.



2019-Q4	GDP (2015=100)	Consumer Price Index (2015=100)	3-month Interbank Interest Rate (%)	···
USA	110.451	108.505	1.803	···
Canada	108.542	107.822	1.867	···
Australia	111.148	107.917	0.897	···
···	···	···	···	···

Fig. 1. A real example of matrix-variate observations consisting of macroeconomic variables for a number of countries.

$\mathbf{X}_t$, and $\mathbf{E}_t$ is the idiosyncratic component of a matrix form. Model (1.1) is particularly suited to modeling well-structured tables of macroeconomic indicators, financial characteristics, and frames of pictures. For example, Fig. 1 shows a time list of tables recording the macroeconomic variables across a number of countries. In this example, the dynamics of the panels might be driven by a much lower-dimensional matrix series of composite indices by taking the interrelationship among countries and macroeconomic variables into consideration. The cross-country (column-sectional) and cross-variable (row-sectional) exposures on these latent composite factors can be summarized in $\mathbf{R}$ and $\mathbf{C}$, respectively. In Wang et al. (2019), two interesting interpretations of model (1.1) on integrating the column and row interactions were explicitly illustrated. The two-step hierarchical interpretation shows that (1.1) reduces the number of parameters of the vector factor modeling by stacking the columns of $\mathbf{X}_t$ from $(p_1p_2+1)k_1k_2$ to $p_1k_1+p_2k_2+k_1k_2$, resulting in a much more parsimonious model. The empirical study in Section 5 shows that the accuracy in forecasting the inflation of the US and Canada and the GDP growth rate of Canada and Norway is improved with both the country and index factors augmented into the regression, and the parsimonious matrix factor structure always helps in the above prediction.

Model (1.1) was later extended to the constrained version by Chen et al. (2020b) and the threshold matrix factor model in Liu and Chen (2019). Chen and Chen (2020) applied model (1.1) to the dynamic transport network in application to international trade flow. Noticeably, all these works are along the line of Lam et al. (2011) and Lam and Yao (2012) by implementing an eigen-analysis of the auto-cross-covariance matrix, which relies heavily on the serial correlations of the factors. Other works follow the other line (e.g. Bai (2003) and Fan et al. (2013)) that the matrix factor series influences all the series and hence leads to spiked eigenvalues along the column and row dimensions. For example, Virta et al. (2017) constructed independent components from low-rank spiked observation matrices, but they assumed a noiseless model. Chen et al. (2020a) extended the approach used by Lettau and Pelger (2020) to the matrix factor model and proposed estimators of the factor loading matrices and factor matrices under model (1.1). Their estimation is mainly based on the eigen-decomposition of an aggregate of the sample mean matrix and the column (or row) covariance matrices, which are assumed to be pervasive along the two cross-sectional dimensions. However, their spectral method handles the rows or columns of the matrix individually, which does not take full advantage of the joint low rank structure across both the rows and columns. Therefore, we believe that the efficiency of the estimated row and column factor spaces can still be improved, with the hope to achieve faster convergence rates. A mathematically rigorous comparison of the convergence rates with Chen et al. (2020a) is provided in Section 3.

In this study, we follow the second line by assuming that the factors are pervasive along the two cross-sectional dimensions and adopt similar assumptions on model (1.1) as in Chen et al. (2020a). We propose a projection estimation method for (1.1). Ideally, if the normalized column loading matrix $\mathbf{C}/\sqrt{p_2}$ is known and orthonormal, then the transformed data matrix $\mathbf{X}_t\mathbf{C}/p_2 = \mathbf{R}\mathbf{F}_t + \mathbf{E}_t\mathbf{C}/p_2$ simply consists of k_2 linear combinations of the columns of $\mathbf{R}$ plus k_2 linear combinations of the columns of $\mathbf{E}_t$. This transformation amounts to projecting the columns of $\mathbf{X}_t$ onto the product space composed of the columns of $\mathbf{R}$ and the k_2 factors explaining the columns of $\mathbf{X}_t$, plus an error matrix $\mathbf{E}_t\mathbf{C}/p_2$. An advantage of the projection is that the original error matrix $\mathbf{E}_t$ is linearly filtered and the resulting entries are of order $O_p(p_2^{-1/2})$ under Assumptions C and D in Section 3. In summary, the projection simultaneously achieves dimension reduction and denoising, which is in the spirit of constructing principal portfolios in finance to reduce the idiosyncratic risk. As the k_2 columns of the transformed data matrix $\mathbf{X}_t\mathbf{C}/p_2$ all lay in the column space of $\mathbf{R}$ asymptotically under a mild condition, in a second step, a simple principal component analysis on the projected data matrix yields an estimator of $\mathbf{R}$. For the example illustrated in Fig. 1, the projection procedure amounts to first summarizing the k_2 factors behind the macroeconomic indicators (k_2 columns of $\mathbf{F}_t$), and then recovering $\mathbf{R}$ from the variations of $\mathbf{R}\mathbf{F}_t$ across countries. Through the above two steps, the low-rank structure along the two cross-sectional dimensions is exploited in a succeeding manner. The preceding argument is heuristic. In practice, $\mathbf{C}$ is unknown and has to be estimated. Indeed, the estimator of $\mathbf{C}$ proposed by Chen et al. (2020a) serves as a good projection matrix, and we describe it in Section 2. By simply applying the procedure to $\mathbf{X}_t^\top$, we can estimate $\mathbf{C}$ in the same manner.

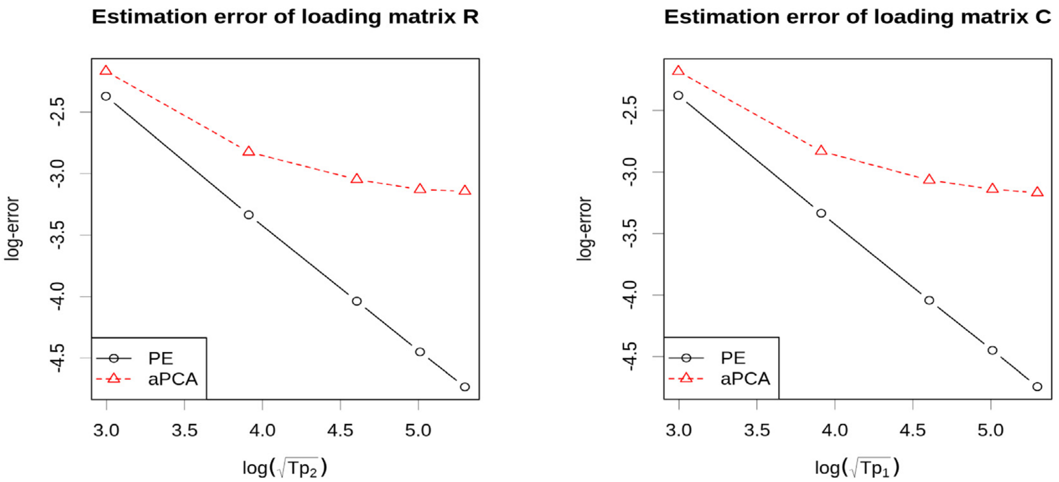


Fig. 2. Mean log error for estimating the loading matrices over 2000 replications. Left: for **R**, $p_1 = 20$, $T = p_2 \in \{20, 50, 100, 150, 200\}$. Right: for **C**, $p_2 = 20$, $T = p_1 \in \{20, 50, 100, 150, 200\}$. "PE": the proposed projected method. "aPCA": α-PCA with $\alpha = 0$.

In this study, we theoretically prove that our projection estimators improve the convergence rates of those in Chen et al. (2020a). Fig. 2 in the simulation study clearly illustrates the improvement. We also propose an iterative algorithm to consistently determine the pair of column and row factor numbers, which performs impressively well in the numerical studies.

The remainder of this paper is organized as follows. Section 2 introduces the model setup and our projection approach. In Section 3, we present the technical assumptions and asymptotic results, including the convergence rates and limiting distributions. Section 4 is devoted to extensive numerical studies. Two real data examples are provided in Section 5. Section 6 concludes and discusses possible future works.

To end this section, we introduce some notations used throughout the study. For a matrix $\mathbf{X}_t$ observed at time t, $x_{t,ij}$ denotes its ijth entry, $\boldsymbol{x}_{t,i\cdot}$ ($\boldsymbol{x}_{t,\cdot j}$) denotes its ith row (jth column) and let $\mathrm{Vec}(\mathbf{X}_t)$ be the vector obtained by stacking the columns of $\mathbf{X}_t$. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ represent the spectral norm and Frobenius norm, respectively. $\|\mathbf{A}\|_{\max}$ is the maximum of $|\mathrm{A}_{ij}|$'s. $\lambda_j(\mathbf{A})$ is the jth eigenvalue of $\mathbf{A}$ if $\mathbf{A}$ is symmetric. The notations $\xrightarrow{p}$, $\xrightarrow{d}$ and $\xrightarrow{a.s.}$ represent convergence in probability, in distribution and almost surely, respectively. The o_p is for convergence to zero in probability and O_p is for stochastic boundedness. For two random series X_n and Y_n, $X_n \lesssim Y_n$ means that $X_n = O_p(Y_n)$, and $X_n \gtrsim Y_n$ means that $Y_n = O_p(X_n)$. The notation $X_n \asymp Y_n$ means that $X_n \lesssim Y_n$ and $X_n \gtrsim Y_n$. For two random vectors $\boldsymbol{X}, \boldsymbol{Y}$, $\boldsymbol{X} \perp \boldsymbol{Y}$ means that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. $[n]$ denotes the set $\{1, \ldots, n\}$. $\otimes$ denotes the Kronecker product. The constant c may not be identical in different lines.

2. Matrix factor model and projected estimators

2.1. Matrix factor model

The model (1.1) factorizes each matrix as a low-rank common component plus an idiosyncratic component, which can be regarded as an extension of the vector factor model to the matrix regime. It provides a new framework and interpretation for the analysis of 3D tensor data. The loading matrices **R** and **C** in model (1.1) are not separately identifiable. In the current paper, only the loading spaces are of interest, and thus we assume without loss of generality that

$$\|p_1^{-1}\mathbf{R}^\top\mathbf{R} - \mathbf{I}_{k_1}\| \to 0, \quad \text{and} \quad \|p_2^{-1}\mathbf{C}^\top\mathbf{C} - \mathbf{I}_{k_2}\| \to 0. \tag{2.1}$$

If this is not the case, then two matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ will always exist with orthogonal columns, such that

$$\mathbf{R} = \mathbf{Q}_1\mathbf{W}_1 \quad \text{and} \quad \mathbf{C} = \mathbf{Q}_2\mathbf{W}_2,$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are $k_1 \times k_1$ and $k_2 \times k_2$ full rank matrices, respectively. Therefore, **R** (or **C**) lies in the same column space as $\mathbf{Q}_1$ (or $\mathbf{Q}_2$), and $\mathbf{X}_t$ can be rewritten as

$$\mathbf{X}_t = (\sqrt{p_1}\mathbf{Q}_1)\widetilde{\mathbf{F}}_t(\sqrt{p_2}\mathbf{Q}_2)^\top + \mathbf{E}_t, \quad \text{where} \quad \widetilde{\mathbf{F}}_t = \frac{1}{\sqrt{p_1p_2}}\mathbf{W}_1\mathbf{F}_t\mathbf{W}_2^\top.$$

Then, $\mathbf{X}_t$ becomes a matrix factor model with row and column loading matrices satisfying (2.1). Assumption (2.1) is not only an identifiability condition for the factor loading spaces, but also a strong factor condition assuming pervasive factors along the row and column dimensions.

2.2. Projected estimation

In this section, we introduce our projection estimation approach. As a simple heuristic argument, $\mathbf{C}$ is assumed to be known and satisfies the orthogonal condition $\mathbf{C}^\top\mathbf{C}/p_2 = \mathbf{I}_{k_2}$. As stated in the introduction, we project the data matrix to a lower dimensional space by setting

$$\mathbf{Y}_t = \frac{1}{p_2}\mathbf{X}_t\mathbf{C} = \frac{1}{p_2}\mathbf{R}\mathbf{F}_t\mathbf{C}^\top\mathbf{C} + \frac{1}{p_2}\mathbf{E}_t\mathbf{C} := \mathbf{R}\mathbf{F}_t + \widetilde{\mathbf{E}}_t := \mathbf{R}(\boldsymbol{f}_{t,\cdot 1},\ldots,\boldsymbol{f}_{t,\cdot k_2}) + (\widetilde{\boldsymbol{e}}_{t,\cdot 1},\ldots,\widetilde{\boldsymbol{e}}_{t,\cdot k_2}). \tag{2.2}$$

After transformation, $\mathbf{Y}_t$ is a $p_1 \times k_2$ matrix-valued observation, lying in a much lower column space than $\mathbf{X}_t$. $\mathbf{F}_t$ and $\widetilde{\mathbf{E}}_t$ can be regarded as factors and errors for $\mathbf{Y}_t$. When $k_2 = 1$, it is exactly a vector factor model. As a result, the projection achieves dimension reduction of the error matrix and decrease of the noise levels. For the ith row of $\widetilde{\mathbf{E}}_t$, denoted as $\widetilde{\boldsymbol{e}}_{t,i\cdot}$, $\mathbb{E}\|\widetilde{\boldsymbol{e}}_{t,i\cdot}\|^2 \le cp_2^{-1}$ as long as the original errors $\{e_{t,ij}\}_{j=1}^{p_2}$ are weakly dependent column-wise. When p_2 is large enough, $\mathbf{Y}_t$ can be treated as a nearly noise-free factor model with $O(p_1)$ loading parameters to be estimated.

Given $\mathbf{Y}_t$, we define

$$\mathbf{M}_1 = \frac{1}{Tp_1}\sum_{t=1}^{T}\mathbf{Y}_t\mathbf{Y}_t^\top,$$

and then the row factor loading matrix $\mathbf{R}$ can be estimated by the leading k_1 eigenvectors of $\mathbf{M}_1$. Heuristically, under some mild conditions,

$$\mathbf{M}_1 = \frac{1}{Tp_1}\sum_{t=1}^{T}\sum_{j=1}^{k_2}\mathbf{y}_{t,\cdot j}\mathbf{y}_{t,\cdot j}^\top \approx \frac{1}{p_1}\mathbf{R}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{k_2}\boldsymbol{f}_{t,\cdot j}\boldsymbol{f}_{t,\cdot j}^\top\right)\mathbf{R}^\top + \frac{1}{Tp_1}\sum_{t=1}^{T}\sum_{j=1}^{k_2}\widetilde{\boldsymbol{e}}_{t,\cdot j}\widetilde{\boldsymbol{e}}_{t,\cdot j}^\top. \tag{2.3}$$

By looking at the k_2 columns of $\mathbf{Y}_t$ as observations within time unit t, (2.2) and (2.3) demonstrate that $\{\mathbf{Y}_t\}_{t=1}^T$ is effectively a vector factor model of length Tk_2 with asymptotically vanishing idiosyncratic entries.

One problem with the above ideal argument is that the projection matrix $\mathbf{C}$ is unavailable in practice. A natural solution is to replace it with a consistent initial estimator $\widehat{\mathbf{C}}$. The column factor loading matrix $\mathbf{C}$ can be similarly estimated by projecting $\mathbf{X}_t$ onto the space of $\mathbf{C}$ with transformation matrix $\mathbf{R}$ or its estimator $\widehat{\mathbf{R}}$. The choices of $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ will be discussed later. We summarize the projection procedure in Algorithm 1 by starting from $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$, which results in an estimated $\mathbf{M}_1$ denoted by $\widetilde{\mathbf{M}}_1$.

Algorithm 1 Projected method for estimating matrix factor spaces

```
Input: Data matrices $\{\mathbf{X}_t\}_{t\le T}$, the pair of row and column factor numbers $k_1$ and $k_2$
Output: Factor loading matrices $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{C}}$
1: obtain the initial estimators $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$;
2: project the data matrices to lower dimensions by defining $\widehat{\mathbf{Y}}_t = p_2^{-1}\mathbf{X}_t\widehat{\mathbf{C}}$ and $\widehat{\mathbf{Z}}_t = p_1^{-1}\mathbf{X}_t^\top\widehat{\mathbf{R}}$;
3: given $\widehat{\mathbf{Y}}_t$ and $\widehat{\mathbf{Z}}_t$, define $\widetilde{\mathbf{M}}_1 = (Tp_1)^{-1}\sum_{t=1}^T\widehat{\mathbf{Y}}_t\widehat{\mathbf{Y}}_t^\top$ and $\widetilde{\mathbf{M}}_2 = (Tp_2)^{-1}\sum_{t=1}^T\widehat{\mathbf{Z}}_t\widehat{\mathbf{Z}}_t^\top$, and estimate the loading spaces by the
   leading $k_i$ eigenvectors of $\widetilde{\mathbf{M}}_i$, denoted as $\widetilde{\mathbf{Q}}_i$, $i = 1, 2$;
4: the row and column loading matrices are finally given by $\widetilde{\mathbf{R}} = \sqrt{p_1}\widetilde{\mathbf{Q}}_1$ and $\widetilde{\mathbf{C}} = \sqrt{p_2}\widetilde{\mathbf{Q}}_2$.
```

The projection method can be implemented recursively by plugging in the newly estimated $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{C}}$ to replace $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ in **Step 2** and iterating **Steps 2-4**. Theoretical analysis of the recursive solution is challenging. The simulation results in Figure 2 of our supplementary material show that the projection estimators with a single iteration perform sufficiently well compared with the recursive method. Actually, with $T \asymp p_1 \asymp p_2$ and $\widehat{\mathbf{C}}$ (or $\widehat{\mathbf{R}}$) chosen suitably, we can prove that the projected estimator $\widetilde{\mathbf{R}}$ (or $\widetilde{\mathbf{C}}$) converges to $\mathbf{R}$ (or $\mathbf{C}$) after rotation at rate $O_p(1/\sqrt{Tp_2})$ (or $O_p(1/\sqrt{Tp_1})$) in terms of the averaged squared errors, which is the optimal rate even when the loading matrix $\mathbf{C}$ (or $\mathbf{R}$) is known in advance.

2.3. Initial projection matrices $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$

The columns of $\mathbf{X}_t$ in model (1.1) can be written in the form of a vector factor model as

$$\boldsymbol{x}_{t,\cdot j} = \mathbf{R}\mathbf{F}_t\boldsymbol{C}_{j\cdot}^\top + \boldsymbol{e}_{t,\cdot j} := \mathbf{R}\bar{\boldsymbol{f}}_{t,\cdot j} + \boldsymbol{e}_{t,\cdot j}, \quad t = 1,\ldots,T, \quad j = 1,\ldots,p_2, \tag{2.4}$$

where $\bar{\boldsymbol{f}}_{t,\cdot j} = \mathbf{F}_t\boldsymbol{C}_{j\cdot}^\top$. Therefore, to estimate $\mathbf{R}$, a natural approach is to regard each column as an individual vector observation and apply the conventional PCA method for vector time series. Specifically, we define the scaled column sample covariance matrix as

$$\widehat{\mathbf{M}}_1 = \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\sum_{j=1}^{p_2}\boldsymbol{x}_{t,\cdot j}\boldsymbol{x}_{t,\cdot j}^\top = \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t^\top.$$

When the columns of **C** are orthogonal and under other mild conditions, approximately

$$\begin{aligned}\widehat{\mathbf{M}}_1 &\approx \frac{1}{p_1}\mathbf{R}\left(\frac{1}{Tp_2}\sum_{t=1}^{T}\sum_{j=1}^{p_2}\bar{\boldsymbol{f}}_{t,\cdot j}\bar{\boldsymbol{f}}_{t,\cdot j}^{\top}\right)\mathbf{R}^{\top}+\frac{1}{p_1}\frac{1}{Tp_2}\sum_{t=1}^{T}\sum_{j=1}^{p_2}\boldsymbol{e}_{t,\cdot j}\boldsymbol{e}_{t,\cdot j}^{\top}\\ &=\frac{1}{p_1}\mathbf{R}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{k_2}\boldsymbol{f}_{t,\cdot j}\boldsymbol{f}_{t,\cdot j}^{\top}\right)\mathbf{R}^{\top}+\frac{1}{Tp_1}\left(\sum_{t=1}^{T}\sum_{j=1}^{p_2}\boldsymbol{e}_{t,\cdot j}p_2^{-1/2}\boldsymbol{e}_{t,\cdot j}^{\top}p_2^{-1/2}\right).\end{aligned} \tag{2.5}$$

The term $T^{-1}\sum_{t=1}^{T}\sum_{j=1}^{k_2}\boldsymbol{f}_{t,\cdot j}\boldsymbol{f}_{t,\cdot j}^{\top}$ typically converges to a symmetric positive definite matrix while the error terms are asymptotically negligible under certain conditions. Consequently, only the leading k_1 eigenvalues of $\widehat{\mathbf{M}}_1$ are spiky. Motivated by Davis–Kahan's sin(Θ) theorem (e.g. Davis and Kahan (1970) and Yu et al. (2015)), we find that the k_1 leading eigenvectors of $\widehat{\mathbf{M}}_1$ lie in the same column space of **R** asymptotically. Therefore, we propose to use the leading k_1 eigenvectors of $\widehat{\mathbf{M}}_1$ as an estimator of $\mathbf{Q}_1$, denoted as $\widehat{\mathbf{Q}}_1$. The row loading matrix is then estimated by $\widehat{\mathbf{R}}=\sqrt{p_1}\widehat{\mathbf{Q}}_1$. The column loading matrix **C** can be estimated in parallel steps applied to $\{\mathbf{X}_t^{\top}\}_{t\le T}$.

We note that the above initial estimator is the α-PCA solution in Chen et al. (2020a) with $\alpha=0$. A comparison of (2.3) and (2.5) shows that $\mathbf{M}_1$ and $\widehat{\mathbf{M}}_1$ have approximately the same covariance matrix of the common components. However, $\widehat{\mathbf{M}}_1$ accumulates more error terms than $\mathbf{M}_1$, thus implying a higher signal-to-noise ratio for $\mathbf{Y}_t$ than that of $\mathbf{X}_t$. This explains the gain of efficiency of our projection estimation method over the initial estimator, or more generally α-PCA procedure. Other choices of initial estimates of **R** and **C** are admissible as long as two sufficient conditions (3.2) and (3.3) in the following section are fulfilled. For simplicity, we only demonstrate theoretically that the above initial estimators work.

3. Theoretical results

In this section, we present theoretical results on the convergence rates and asymptotic distributions of the projected estimators. The estimation of factors and common components is also considered. The numbers of factors are treated as given initially, and later we will propose an iterative algorithm to consistently estimate them.

3.1. Technical assumptions

The matrix factor models are specifically designed for 3D tensor data. The correlation structure for complex high-order tensor data makes the theoretical analysis challenging. Throughout this study, we make the following assumptions on the correlations across time, row, and column.

Assumption A (*Alpha Mixing*)**.** The vectorized factor series $\{\mathrm{Vec}(\mathbf{F}_t)\}$ and noise series $\{\mathrm{Vec}(\mathbf{E}_t)\}$ are α-mixing. A vector process $\{\boldsymbol{z}_t, t=0,\pm1,\pm2,\ldots\}$ is α-mixing if, for some $\gamma>2$, the mixing coefficients satisfy the condition that

$$\sum_{h=1}^{\infty}\alpha(h)^{1-2/\gamma}<\infty,$$

where $\alpha(h)=\sup_t\sup_{A\in\mathcal{F}_{-\infty}^{t},B\in\mathcal{F}_{t+h}^{\infty}}|P(A\cap B)-P(A)P(B)|$ and $\mathcal{F}_{\tau}^{s}$ is the σ-field generated by $\{\boldsymbol{z}_t:\tau\le t\le s\}$.

Assumption B (*Factor Matrix*)**.** The factor matrix satisfies $\mathbb{E}(\mathbf{F}_t)=\mathbf{0}$, $\mathbb{E}\|\mathbf{F}_t\|^4\le c<\infty$ for some constant $c>0$, and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t^{\top}\xrightarrow{p}\boldsymbol{\Sigma}_1 \text{ and } \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^{\top}\mathbf{F}_t\xrightarrow{p}\boldsymbol{\Sigma}_2, \tag{3.1}$$

where $\boldsymbol{\Sigma}_i$ is a $k_i\times k_i$ positive definite matrix with distinct eigenvalues and spectral decomposition $\boldsymbol{\Sigma}_i=\boldsymbol{\Gamma}_i\boldsymbol{\Lambda}_i\boldsymbol{\Gamma}_i^{\top}$, $i=1,2$. The factor numbers k_1 and k_2 are fixed as $\min\{T,p_1,p_2\}\to\infty$.

Assumption C (*Loading Matrix*)**.** Positive constants $\bar{r}$ and $\bar{c}$ exist such that $\|\mathbf{R}\|_{\max}\le\bar{r}$, $\|\mathbf{C}\|_{\max}\le\bar{c}$. As $\min\{p_1,p_2\}\to\infty$, $\|p_1^{-1}\mathbf{R}^{\top}\mathbf{R}-\mathbf{I}_{k_1}\|\to0$ and $\|p_2^{-1}\mathbf{C}^{\top}\mathbf{C}-\mathbf{I}_{k_2}\|\to0$.

The α-mixing condition in Assumption A allows weak temporal correlations for both the factors and noises. In Assumption B, the factor matrix is centered with bounded fourth moment. The condition (3.1) in Assumption B is easily fulfilled under the α-mixing assumption, by Corollary 16.2.4 in Athreya and Lahiri (2006). The eigenvalues of $\boldsymbol{\Sigma}_i$'s are assumed to be distinct such that the corresponding eigenvectors are identifiable. We assume strong factor conditions in Assumption C, which means that the row and column factors are pervasive along both dimensions. This result is an extension of the pervasive assumption in Stock and Watson (2002a) to the matrix regime. For identifiability, we assume $\|p_1^{-1}\mathbf{R}^{\top}\mathbf{R}-\mathbf{I}_{k_1}\|\to0$ and $\|p_2^{-1}\mathbf{C}^{\top}\mathbf{C}-\mathbf{I}_{k_2}\|\to0$ as $\min\{p_1,p_2\}\to\infty$. Assumptions A, B, and C are standard and common in the literature and similar assumptions are adopted by Chen et al. (2020a), except that the factor matrix is not centralized in their setting.

Assumption D (*Weak Correlation of Noise* $\mathbf{E}_t$ *Across Column, Row, and Time*)**.** A positive constant $c < \infty$ exists such that

1. $\mathbb{E}e_{t,ij} = 0$, $\mathbb{E}(e_{t,ij}^8) \le c$.
2. for any $t \in [T]$, $i \in [p_1]$, $j \in [p_2]$,

$$(1).\ \sum_{s=1}^{T}\sum_{l=1}^{p_1}\sum_{h=1}^{p_2}|\mathbb{E}e_{t,ij}e_{s,lh}| \le c, \quad (2).\ \sum_{l=1}^{p_1}\sum_{h=1}^{p_2}|\mathbb{E}e_{t,lj}e_{t,ih}| \le c.$$

3. for any $t \in [T]$, $i, l_1 \in [p_1]$, $j, h_1 \in [p_2]$,

$$(1).\ \sum_{s=1}^{T}\sum_{l_2=1}^{p_1}\sum_{h=1}^{p_2}\left|\mathrm{Cov}(e_{t,ij}e_{t,l_1j}, e_{s,ih}e_{s,l_2h})\right| \le c, \quad \sum_{s=1}^{T}\sum_{l=1}^{p_1}\sum_{h_2=1}^{p_2}\left|\mathrm{Cov}(e_{t,ij}e_{t,ih_1}, e_{s,lj}e_{s,lh_2})\right| \le c,$$

$$(2).\ \sum_{s=1}^{T}\sum_{l_2=1}^{p_1}\sum_{h_2=1}^{p_2}\left(\left|\mathrm{Cov}(e_{t,ij}e_{t,l_1h_1}, e_{s,ij}e_{s,l_2h_2})\right| + \left|\mathrm{Cov}(e_{t,l_1j}e_{t,ih_1}, e_{s,l_2j}e_{s,ih_2})\right|\right) \le c.$$

Assumption D is essentially an extension of Assumption C in Bai (2003) to the matrix regime. Similar conditions are adopted by Chen et al. (2020a). Assumption D.2 (1) allows weak correlation of the noises across time, row and column. It can be a sufficient condition to the Assumptions D.2, E and G.3 in Chen et al. (2020a). Assumption D.2 (2) further controls the column-wise and row-wise correlation of the noises. Assumption D.3 (1) is similar to the Assumption G.1 in Chen et al. (2020a), where they require that

$$\mathbb{E}\left\|\frac{1}{\sqrt{Tp_1p_2}}\sum_{t=1}^{T}\sum_{l=1}^{p_1}\sum_{j=1}^{p_2}\mathbf{R}_{l\cdot}\left(e_{t,ij}e_{t,lj} - \mathbb{E}e_{t,ij}e_{t,lj}\right)\right\|^2 \le c.$$

Therefore, the correlation of noises up to the second moment is controlled. Assumption D.3 (2) is similar to D.3 (1), but on different combinations of noise pairs. Suppose that the $\{e_{t,ij}\}$'s are located in a 3D space indexed by time, row and column, Assumption D is satisfied if $e_{t,ij} \perp e_{s,lh}$ as long as the index distance between them is larger than some bandwidth, or the correlation decays sufficiently fast as the distance increases.

Assumption E (*Weak Dependence Between Factor* F_t *and Noise* E_t)**.** There exists a constant $c > 0$, such that,

1. for any deterministic vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ satisfying $\|\boldsymbol{v}\| = 1$ and $\|\boldsymbol{w}\| = 1$,

$$\mathbb{E}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\mathbf{F}_t\boldsymbol{v}^\top\mathbf{E}_t\boldsymbol{w})\right\|^2 \le c;$$

2. for any $i, l_1 \in [p_1]$ and $j, h_1 \in [p_2]$,

$$(1).\left\|\sum_{h=1}^{p_2}\mathbb{E}(\bar{\boldsymbol{\zeta}}_{ij} \otimes \bar{\boldsymbol{\zeta}}_{ih})\right\|_{\max} \le c, \quad \left\|\sum_{l=1}^{p_1}\mathbb{E}(\bar{\boldsymbol{\zeta}}_{ij} \otimes \bar{\boldsymbol{\zeta}}_{lj})\right\|_{\max} \le c,$$

$$(2).\left\|\sum_{l=1}^{p_1}\sum_{h_2=1}^{p_2}\mathrm{Cov}(\bar{\boldsymbol{\zeta}}_{ij} \otimes \bar{\boldsymbol{\zeta}}_{ih_1}, \bar{\boldsymbol{\zeta}}_{lj} \otimes \bar{\boldsymbol{\zeta}}_{lh_2})\right\|_{\max} \le c, \left\|\sum_{l_2=1}^{p_1}\sum_{h=1}^{p_2}\mathrm{Cov}(\bar{\boldsymbol{\zeta}}_{ij} \otimes \bar{\boldsymbol{\zeta}}_{l_1j}, \bar{\boldsymbol{\zeta}}_{ih} \otimes \bar{\boldsymbol{\zeta}}_{l_2h})\right\|_{\max} \le c,$$

where $\bar{\boldsymbol{\zeta}}_{ij} = \mathrm{Vec}(\sum_{t=1}^{T}\mathbf{F}_te_{t,ij}/\sqrt{T})$.

Assumption E.1 is summarized from the Assumptions F and G.2 in Chen et al. (2020a). Indeed, we can view $\boldsymbol{v}^\top\mathbf{E}_t\boldsymbol{w}$ as a random variable with zero mean and bounded variance since the noise is weakly correlated across row and column. Therefore, Assumption E.1 simply implies that $\mathbb{E}(\mathbf{F}_t\boldsymbol{v}^\top\mathbf{E}_t\boldsymbol{w}) \approx \mathbf{0}$ and the temporal correlations of the series $\{\mathbf{F}_t\boldsymbol{v}^\top\mathbf{E}_t\boldsymbol{w}\}$ are also weak. Assumption E.2 controls higher-order correlations between the factor and noise series, where $\bar{\boldsymbol{\zeta}}_{ij}$ can be viewed as random vectors with fixed dimension and bounded marginal variances (under Assumption E.1). Assumption E is satisfied if the noise variables are independent across time and independent of the factor series, given Assumptions A to D.

3.2. Asymptotics on projection estimators

We first present the following conditions on the convergence rates of the initial estimators $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ to guarantee that the projection procedure works.

(Sufficient Condition) There exist $k_1 \times k_1$ matrices $\widehat{\mathbf{H}}_1$ satisfying $\widehat{\mathbf{H}}_1\widehat{\mathbf{H}}_1^\top \xrightarrow{p} \mathbf{I}_{k_1}$ and

$$(a).\ \frac{1}{p_1}\|\widehat{\mathbf{R}} - \mathbf{R}\widehat{\mathbf{H}}_1\|_F^2 = O_p(w_1), \quad (b).\ \frac{1}{p_2}\left\|\frac{1}{Tp_1}\sum_{s=1}^{T}\mathbf{E}_s^\top(\widehat{\mathbf{R}} - \mathbf{R}\widehat{\mathbf{H}}_1)\mathbf{F}_s\right\|_F^2 = O_p(w_2), \tag{3.2}$$

where $w_1, w_2 \to 0$ as T, p_1 and p_2 go to infinity simultaneously. There exist $k_2 \times k_2$ matrices $\widehat{\mathbf{H}}_2$ satisfying $\widehat{\mathbf{H}}_2\widehat{\mathbf{H}}_2^\top \xrightarrow{p} \mathbf{I}_{k_2}$ and

$$\text{(a). } \frac{1}{p_2}\|\widehat{\mathbf{C}} - \mathbf{C}\widehat{\mathbf{H}}_2\|_F^2 = O_p(m_1), \quad \text{(b). } \frac{1}{p_1}\left\| \frac{1}{Tp_2}\sum_{s=1}^T \mathbf{E}_s(\widehat{\mathbf{C}} - \mathbf{C}\widehat{\mathbf{H}}_2)\mathbf{F}_s^\top \right\|_F^2 = O_p(m_2), \tag{3.3}$$

where $m_1, m_2 \to 0$ as T, p_1 and p_2 go to infinity simultaneously.

Now, we state a theorem on the convergence rates of our projection estimators.

Theorem 3.1 (*Consistency of the Projected Estimators*)**.** *Under Assumptions A to E and sufficient conditions (3.2) and (3.3), there exist matrices $\widetilde{\mathbf{H}}_1$ and $\widetilde{\mathbf{H}}_2$, satisfying $\widetilde{\mathbf{H}}_1^\top\widetilde{\mathbf{H}}_1/p_1 \xrightarrow{p} \mathbf{I}_{k_1}$ and $\widetilde{\mathbf{H}}_2^\top\widetilde{\mathbf{H}}_2/p_2 \xrightarrow{p} \mathbf{I}_{k_2}$, such that*

$$\frac{1}{p_1}\|\widetilde{\mathbf{R}} - \mathbf{R}\widetilde{\mathbf{H}}_1\|_F^2 = O_p(\widetilde{w}_1), \quad \|\widetilde{\boldsymbol{R}}_{i\cdot} - \boldsymbol{R}_{i\cdot}\widetilde{\mathbf{H}}_1\|^2 = O_p(\widetilde{w}_1), \quad i = 1, \ldots, p_1,$$
$$\frac{1}{p_2}\|\widetilde{\mathbf{C}} - \mathbf{C}\widetilde{\mathbf{H}}_2\|_F^2 = O_p(\widetilde{m}_1), \quad \|\widetilde{\mathbf{C}}_{j\cdot} - \mathbf{C}_{j\cdot}\widetilde{\mathbf{H}}_2\|^2 = O_p(\widetilde{m}_1), \quad j = 1, \ldots, p_2,$$

as T, p_1 and p_2 go to infinity simultaneously, where $\boldsymbol{R}_{i\cdot}$, $\widetilde{\boldsymbol{R}}_{i\cdot}$, $\mathbf{C}_{j\cdot}$ and $\widetilde{\mathbf{C}}_{j\cdot}$ are the ith /jth row of $\mathbf{R}$, $\widetilde{\mathbf{R}}$, $\mathbf{C}$ and $\widetilde{\mathbf{C}}$, respectively, and

$$\widetilde{w}_1 = \frac{1}{Tp_2} + \frac{1}{p_1^2p_2^2} + m_1^2\left(\frac{1}{p_1^2} + \frac{1}{Tp_1}\right) + m_2, \quad \widetilde{m}_1 = \frac{1}{Tp_1} + \frac{1}{p_1^2p_2^2} + w_1^2\left(\frac{1}{p_2^2} + \frac{1}{Tp_2}\right) + w_2.$$

The estimation error bounds for $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{C}}$ show clear dependence on the accuracy of the initial estimates. Actually, Theorem 3.3 verifies that the proposed initial estimates in Section 2.3 satisfy the sufficient conditions (3.2) and (3.3) with

$$w_1 = \frac{1}{p_1^2} + \frac{1}{Tp_2}, \quad w_2 = \frac{1}{Tp_1^2} + \frac{1}{T^2p_2^2}, \quad m_1 = \frac{1}{p_2^2} + \frac{1}{Tp_1}, \quad m_2 = \frac{1}{Tp_2^2} + \frac{1}{T^2p_1^2}. \tag{3.4}$$

Corollary 3.1. *Under Assumptions A–E and conditions (3.2)–(3.4), it holds in Theorem 3.1,*

$$\widetilde{w}_1 = O\left(\frac{1}{Tp_2} + \frac{1}{p_1^2p_2^2} + \frac{1}{T^2p_1^2}\right), \quad \widetilde{m}_1 = O\left(\frac{1}{Tp_1} + \frac{1}{p_1^2p_2^2} + \frac{1}{T^2p_2^2}\right).$$

The convergence rates for the estimators of $\mathbf{R}$ and $\mathbf{C}$ in Chen et al. (2020a) are $O_p\{(Tp_2)^{-1} + p_1^{-1}\}$ and $O_p\{(Tp_1)^{-1} + p_2^{-1}\}$, respectively. Corollary 3.1 demonstrates that our projected estimators of the row and column factor spaces perform no worse than (Chen et al., 2020a)'s estimator, and achieve faster convergence rates than Chen et al. (2020a)'s estimators, when $p_1 = o(Tp_2)$ for estimating the row factor loading matrix $\mathbf{R}$ and $p_2 = o(Tp_1)$ for estimating the column factor loading matrix $\mathbf{C}$.

If we take each column (or row) as individual observation and look at $\{\mathbf{X}_t\}$ as a vector time series of length Tp_2 (or Tp_1) and dimension p_1 (or p_2) as in (2.4), the theorems in Bai (2003) indicate a convergence rate of $(Tp_2)^{-1} + p_1^{-2}$ for estimating $\mathbf{R}$ (or $(Tp_1)^{-1} + p_2^{-2}$ for estimating $\mathbf{C}$). Indeed, under our assumptions, we can improve the results in Chen et al. (2020a) and show that the α-PCA estimator with $\alpha = 0$, i.e. the initial estimator, achieves the rates conceivable from Bai (2003). Recall that, as (2.2)–(2.3) show, our first-step projected matrix series $\{\mathbf{Y}_t\}$ can be interpreted as a series of p_1 dimensional vectors of length Tk_2 with asymptotically negligible error entries. A comparison of (2.2)–(2.3) with (2.4)–(2.5) demonstrates that the projection estimators benefit from a smaller noise level in the sense of the spectral norm. Indeed, the proof in the supplementary material shows that the idiosyncratic risk components for $\mathbf{M}_1$ and $\widehat{\mathbf{M}}_1$ are of orders $1/\sqrt{Tp_2} + 1/(Tp_1) + 1/(p_1p_2)$ and $1/\sqrt{Tp_2} + 1/p_1$, respectively.

To further study the entry-wise asymptotic distributions of the estimated loadings, we need the following assumptions.

Assumption F. For $i \leq p_1$,

$$\frac{1}{\sqrt{Tp_2}}\sum_{t=1}^T \mathbf{F}_t\mathbf{C}^\top\boldsymbol{e}_{t,i\cdot} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{1i}), \quad \text{where} \quad \mathbf{V}_{1i} = \lim_{T,p_1,p_2\to\infty}\frac{1}{Tp_2}\sum_{t=1}^T \mathbb{E}\mathbf{F}_t\mathbf{C}^\top\mathrm{Cov}(\boldsymbol{e}_{t,i\cdot})\mathbf{C}\mathbf{F}_t^\top.$$

For $j \leq p_2$,

$$\frac{1}{\sqrt{Tp_1}}\sum_{t=1}^T \mathbf{F}_t^\top\mathbf{R}^\top\boldsymbol{e}_{t,\cdot j} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_{2j}), \quad \text{where} \quad \mathbf{V}_{2j} = \lim_{T,p_1,p_2\to\infty}\frac{1}{Tp_1}\sum_{t=1}^T \mathbb{E}\mathbf{F}_t^\top\mathbf{R}^\top\mathrm{Cov}(\boldsymbol{e}_{t,\cdot j})\mathbf{R}\mathbf{F}_t.$$

$\mathbf{V}_{1i}$ and $\mathbf{V}_{2j}$ are positive definite matrices whose eigenvalues are bounded away from 0 and infinity.

Assumption F can be verified by the martingale central limit theorem. It is easily fulfilled under the proposed α-mixing condition and weak correlation assumptions. One can refer to Chapter 16 of Athreya and Lahiri (2006) for more details. Similar assumptions are imposed in Bai (2003) and Chen et al. (2020a). The following Theorem 3.2 presents the asymptotic distributions of the projected estimators of the loadings.

Theorem 3.2 (*Asymptotic Normality of Projection Estimators*)**.** *Under Assumptions A to F, if the initial estimators* $\widehat{\mathbf{R}}$ *and* $\widehat{\mathbf{C}}$ *are proposed as in Section 2.3,*

1. *for* $i \le p_1$,

$$\begin{cases} \sqrt{Tp_2}(\widetilde{\boldsymbol{R}}_i - \widetilde{\mathbf{H}}_1^\top \boldsymbol{R}_i) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Gamma}_1^\top \mathbf{V}_{1i}\boldsymbol{\Gamma}_1\boldsymbol{\Lambda}_1^{-1}), & \text{if} \quad Tp_2 = o(\min\{T^2p_1^2, p_2^2p_1^2\}), \\ \widetilde{\boldsymbol{R}}_i - \widetilde{\mathbf{H}}_1^\top \boldsymbol{R}_i = O_p\Big(\dfrac{1}{Tp_1} + \dfrac{1}{p_2p_1}\Big), & \text{if} \quad Tp_2 \gtrsim \min\{T^2p_1^2, p_2^2p_1^2\}; \end{cases}$$

2. *for* $j \le p_2$,

$$\begin{cases} \sqrt{Tp_1}(\widetilde{\boldsymbol{C}}_j - \widetilde{\mathbf{H}}_2^\top \boldsymbol{C}_j) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Gamma}_2^\top \mathbf{V}_{2j}\boldsymbol{\Gamma}_2\boldsymbol{\Lambda}_2^{-1}), & \text{if} \quad Tp_1 = o(\min\{T^2p_2^2, p_1^2p_2^2\}), \\ \widetilde{\boldsymbol{C}}_j - \widetilde{\mathbf{H}}_2^\top \boldsymbol{C}_j = O_p\Big(\dfrac{1}{Tp_2} + \dfrac{1}{p_1p_2}\Big), & \text{if} \quad Tp_1 \gtrsim \min\{T^2p_2^2, p_1^2p_2^2\}. \end{cases}$$

3.3. Theorems on initial estimators

As claimed, the initial estimators are α-PCA solutions in Chen et al. (2020a) with $\alpha = 0$. However, based on the argument below Corollary 3.1, the convergence rate in their paper is slower than the expected one by a factor of p_1^{-1} or p_2^{-1}. Under our assumptions, an improved rate is accessible and summarized in the following theorem.

Theorem 3.3. *Under Assumptions A to E, (3.4) holds for the initial estimators.*

For the initial estimator $\widehat{\mathbf{R}}$, the rate w_1 matches the typical rate $O_p(T^{-1} + p_1^{-2})$ of the vector factor model when $p_2 = 1$, as shown by Theorem 2 of Bai (2003). The initial estimators are also asymptotically normally distributed as shown in the next theorem.

Theorem 3.4 (*Asymptotic Normality of the Initial Estimators*)**.** *Under Assumptions A to F, as* $T, p_1, p_2 \to \infty$,

1. *for* $i \le p_1$,

$$\begin{cases} \sqrt{Tp_2}(\widehat{\boldsymbol{R}}_i - \widehat{\mathbf{H}}_1^\top \boldsymbol{R}_i) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\Gamma}_1^\top \mathbf{V}_{1i}\boldsymbol{\Gamma}_1\boldsymbol{\Lambda}_1^{-1}), & \text{if} \quad Tp_2 = o(p_1^2), \\ \widehat{\boldsymbol{R}}_i - \widehat{\mathbf{H}}_1^\top \boldsymbol{R}_i = O_p(p_1^{-1}), & \text{if} \quad Tp_2 \gtrsim p_1^2; \end{cases}$$

2. *for* $j \le p_2$,

$$\begin{cases} \sqrt{Tp_1}(\widehat{\boldsymbol{C}}_j - \widehat{\mathbf{H}}_2^\top \boldsymbol{C}_j) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_2^{-1}\boldsymbol{\Gamma}_2^\top \mathbf{V}_{2j}\boldsymbol{\Gamma}_2\boldsymbol{\Lambda}_2^{-1}), & \text{if} \quad Tp_1 = o(p_2^2), \\ \widehat{\boldsymbol{C}}_j - \widehat{\mathbf{H}}_2^\top \boldsymbol{C}_j = O_p(p_2^{-1}), & \text{if} \quad Tp_1 \gtrsim p_2^2, \end{cases}$$

where $\widehat{\boldsymbol{R}}_i$ *and* $\widehat{\boldsymbol{C}}_j$ *are the ith and jth row vectors of* $\widehat{\mathbf{R}}$ *and* $\widehat{\mathbf{C}}$*, respectively.*

Compared with Theorem 3.2, $\widehat{\mathbf{R}}$ and $\widetilde{\mathbf{R}}$ share the same asymptotic covariance matrix when p_1 is sufficiently large. However, the normality of $\widehat{\mathbf{R}}$ requires a more stringent condition that $p_1^2 \gg Tp_2$, while for the projected estimator we only require $p_1^2 \gg \max\{p_2/T, T/p_2\}$. A similar conclusion holds for $\widehat{\mathbf{C}}$ and $\widetilde{\mathbf{C}}$.

3.4. Estimating factor matrix and common components

As long as the loading matrices are determined, the factor matrix $\mathbf{F}_t$ can be estimated easily by $\widetilde{\mathbf{F}}_t = \widetilde{\mathbf{R}}^\top \mathbf{X}_t \widetilde{\mathbf{C}}/(p_1p_2)$. The common component matrix is then given by $\widetilde{\mathbf{S}}_t = \widetilde{\mathbf{R}}\widetilde{\mathbf{F}}_t\widetilde{\mathbf{C}}^\top$. The next theorem provides the consistency of the estimated factors and common components.

Theorem 3.5. *Under Assumptions A to E, as* $\min\{T, p_1, p_2\} \to \infty$*, for any* $t \in [T]$, $i \in [p_1]$ *and* $j \in [p_2]$,

$$(1).\|\widetilde{\mathbf{F}}_t - \widetilde{\mathbf{H}}_1^{-1}\mathbf{F}_t\widetilde{\mathbf{H}}_2^{-1}\| = O_p\left(\frac{1}{\sqrt{T} \times \min\{p_1, p_2\}} + \frac{1}{\sqrt{p_1p_2}}\right),$$

$$(2).|\widetilde{\mathrm{S}}_{t,ij} - \mathrm{S}_{t,ij}| = O_p\left(\frac{1}{\sqrt{Tp_1}} + \frac{1}{\sqrt{Tp_2}} + \frac{1}{\sqrt{p_1p_2}}\right).$$

Remark 6. The convergence rates in Theorem 3.5 are the same as those in Chen et al. (2020a) when $p_1 \asymp p_2$, although the estimated loadings by the projection method are generally more accurate. The reason is that the estimation error of $\widetilde{\mathbf{F}}_t$ mainly comes from the error term $(p_1p_2)^{-1}\mathbf{R}^\top\mathbf{E}_t\mathbf{C}$. Even if the loadings $\mathbf{R}$ and $\mathbf{C}$ are known, the best convergence rate for estimating $\mathbf{F}_t$ is still of the rate $(p_1p_2)^{-1/2}$ under the spectral norm. This error further affects the estimation of the common components. One can easily verify the asymptotic normality of $\widetilde{\mathbf{F}}_t$ by imposing certain conditions on $(p_1p_2)^{-1}\mathbf{R}^\top\mathbf{E}_t\mathbf{C}$.

3.5. Determining the pair of row and column factor numbers k_1 and k_2

The dimensions k_1 and k_2 of the common factor matrix need to be determined before the procedures can be applied. In this study, we specify the numbers of row and column factors by borrowing the eigenvalue-ratio statistics discussed in Lam and Yao (2012) and Ahn and Horenstein (2013). In detail, $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ are selected as the initial projection matrices, and then k_1 is estimated by

$$\widehat{k}_1 = \arg\max_{j \le k_{\max}} \frac{\lambda_j(\widetilde{\mathbf{M}}_1)}{\lambda_{j+1}(\widetilde{\mathbf{M}}_1)}, \tag{3.5}$$

where $k_{\max}$ is a predetermined upper bound for k_1. Chen et al. (2020a) proposed a similar criterion using $\widehat{\mathbf{M}}_1$. We use $\widetilde{\mathbf{M}}_1$ rather than $\widehat{\mathbf{M}}_1$ because $\widetilde{\mathbf{M}}_1$ is usually more accurate for approximating the column covariance matrix of the common components.

When the common factors are sufficiently strong, the leading k_1 eigenvalues of $\widetilde{\mathbf{M}}_1$ are well separated from the others. Thus, the eigenvalue ratios in Eq. (3.5) are asymptotically maximized exactly at $j = k_1$. To avoid vanishing denominators, we can add an asymptotically negligible term, such as $c\delta$ for some small constant c and $\delta = \max\{1/\sqrt{Tp_2}, 1/\sqrt{Tp_1}, p_1^{-1}\}$, to the denominator of Eq. (3.5). One problem to calculate $\widetilde{\mathbf{M}}_1$ is that $\widehat{\mathbf{C}}$ must be predetermined, which means k_2 must be given first. Empirically, both k_1 and k_2 are unknown. To address this difficulty, we suggest using the following iterative Algorithm 2 to determine the paired numbers of factors.

Algorithm 2 Iterative algorithm to specify numbers of factors

```
Input: Data matrices $\{\mathbf{X}_t\}_{t \le T}$, maximum number $k_{\max}$, maximum iterative step $m$
Output: Numbers of row and column factors $\widehat{k}_1$ and $\widehat{k}_2$
1: initialization: $\widehat{k}_1^{(0)} = k_{\max}$, $\widehat{k}_2^{(0)} = k_{\max}$;
2: for $t = 1, \ldots, m$, given $\widehat{k}_2^{(t-1)}$, estimate $\widehat{\mathbf{C}}^{(t)}$ by the initial estimator, and calculate $\widetilde{\mathbf{M}}_1^{(t)}$ using $\widehat{\mathbf{C}}^{(t)}$, then $\widehat{k}_1^{(t)}$ is given by equation (3.5);
3: given $\widehat{k}_1^{(t)}$, estimate $\widehat{\mathbf{R}}^{(t)}$ by the initial estimator, and calculate $\widetilde{\mathbf{M}}_2^{(t)}$ using $\widehat{\mathbf{R}}^{(t)}$, then $\widehat{k}_2^{(t)}$ is given by a parallel "ER" approach by replacing $\widetilde{\mathbf{M}}_1$ with $\widetilde{\mathbf{M}}_2^{(t)}$ in equation (3.5);
4: repeat Steps 2 and 3 until $\widehat{k}_1^{(t)} = \widehat{k}_1^{(t-1)}$ and $\widehat{k}_2^{(t)} = \widehat{k}_2^{(t-1)}$, or reach the maximum iterative step.
```

Remark 7. The term $c\delta$ only works as a lower bound of the denominator in our technical proofs. As one reviewer pointed out, adding such a term may affect the finite sample performance. We compared the empirical performances of the iterative algorithm with $c = 0$, $c = 10^{-4}$ and $c = 1$ in the simulation study. The numerical results are not much sensitive to the term $c\delta$. In practice, we suggest setting c sufficiently small in case of underestimation.

The consistency of the iterative algorithm is guaranteed by the following theorem.

Theorem 3.8 (*Specifying the Numbers of Row and Column Factors*)**.** *Under Assumptions A to E, when* $\min\{k_1, k_2\} > 0$, $\min\{T, p_1, p_2\} \to \infty$ *and* $k_{\max}$ *is a predetermined constant no smaller than* $\max\{k_1, k_2\}$, *if* $\widehat{k}_2^{(t-1)} \in [k_2, k_{\max}]$ *for some* t *in the iterative Algorithm 2, then*

$$Pr(\widehat{k}_1^{(t)} = k_1) \to 1;$$

and if $\widehat{k}_1^{(t)} \in [k_1, k_{\max}]$ *for some* t *in Algorithm 2, then*

$$Pr(\widehat{k}_2^{(t)} = k_2) \to 1.$$

Theorem 3.8 indicates that as long as we start with some $k_1^{(0)}$ and $k_2^{(0)}$ larger than the true k_1 and k_2, the iterative algorithm can consistently estimate the numbers of factors. The algorithm is computationally very fast because it has a large probability to stop within finite steps.

The finite sample performance of the eigenvalue-ratio method usually depends on the maximized ratio at the true number of factors. A larger ratio implies a better separation of the spiked eigenvalues, which leads to better estimation of the number of factors. The maximized ratio of the above algorithm is shown to be $\min\{\sqrt{Tp_2}, \sqrt{Tp_1}, p_1\}$ for k_1 and $\min\{\sqrt{Tp_1}, \sqrt{Tp_2}, p_2\}$ for k_2 in the proof. However, if we vectorize the data matrices and apply the eigenvalue-ratio approach in Ahn and Horenstein (2013), the maximized ratio for the total number of factors will be of the rate

Table 1
Averaged estimation errors and standard errors (in parentheses) of $\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R})$ and $\mathcal{D}(\widehat{\mathbf{C}}, \mathbf{C})$ for Settings A and B (effects of T, p_1, p_2), over 2000 replications. "PE": proposed projected method. "(a)PCA": α-PCA with $\alpha = a$.

Evaluation	T	p_1	p_2	PE	(-1)PCA	(0)PCA	(1)PCA
$\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R})$	20	20	20	**0.0933(0.0158)**	0.1173(0.0329)	0.1144(0.0319)	0.1179(0.0324)
	50		50	**0.0356(0.0052)**	0.0597(0.0224)	0.0594(0.0223)	0.0598(0.0223)
	100		100	**0.0176(0.0026)**	0.0475(0.0213)	0.0475(0.0213)	0.0475(0.0213)
	150		150	**0.0117(0.0016)**	0.0438(0.0200)	0.0437(0.0199)	0.0438(0.0199)
	200		200	**0.0088(0.0013)**	0.0432(0.0244)	0.0432(0.0244)	0.0432(0.0243)
$\mathcal{D}(\widehat{\mathbf{C}}, \mathbf{C})$	20	20	20	**0.0927(0.0157)**	0.1154(0.0304)	0.1126(0.0297)	0.1162(0.0305)
	50	50		**0.0356(0.0052)**	0.0591(0.0209)	0.0590(0.0211)	0.0596(0.0213)
	100	100		**0.0175(0.0025)**	0.0466(0.0194)	0.0466(0.0194)	0.0467(0.0194)
	150	150		**0.0117(0.0016)**	0.0433(0.0202)	0.0433(0.0201)	0.0433(0.0201)
	200	200		**0.0087(0.0012)**	0.0421(0.0205)	0.0421(0.0205)	0.0421(0.0205)

$\min\{\sqrt{T}, p_1p_2\}$. Therefore, when $T = o(\min\{p_1^2, p_2^2\})$, our iterative algorithm may perform better than the standard eigenvalue-ratio method for the vectorized model, which is checked in the simulation studies. However, the eigenvalue-ratio method for the vectorized model may perform better when $\min\{p_1^2, p_2^2\} = o(T)$, which is conceivable as we have a pair of factor numbers to be estimated. The estimation of k_2 brings an error to $\widehat{k}_1$ in the iterative algorithm. Actually, we show in the proof that when k_2 is given, the maximized eigenvalue-ratio for estimating k_1 is of the order $\min\{\sqrt{Tp_2}, Tp_1, p_1p_2\}$, which is even better than that of the vectorized model. As $\widehat{k}_1^{(t)}$ or $\widehat{k}_2^{(t)}$ has large probability to be exactly k_1 or k_2 after a few iterations, the iterative algorithm performs impressively well empirically.

4. Simulation studies

4.1. Data generation

In this section, we investigate the finite sample performance of the proposed projection procedure. The observed data matrices are generated according to model (1.1). In detail, we set $k_1 = k_2 = 3$, draw the entries of $\mathbf{R}$ and $\mathbf{C}$ independently from uniform distribution $\mathcal{U}(-1, 1)$, and let

$$\begin{aligned}\text{Vec}(\mathbf{F}_t) =& \phi \times \text{Vec}(\mathbf{F}_{t-1}) + \sqrt{1-\phi^2} \times \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_{k_1 \times k_2}),\\ \text{Vec}(\mathbf{E}_t) =& \psi \times \text{Vec}(\mathbf{E}_{t-1}) + \sqrt{1-\psi^2} \times \text{Vec}(\mathbf{U}_t), \quad \mathbf{U}_t \overset{i.i.d.}{\sim} \mathcal{MN}(\mathbf{0}, \mathbf{U}_E, \mathbf{V}_E),\end{aligned} \tag{4.1}$$

where $\mathbf{U}_t$ is from a matrix-normal distribution, i.e., $\text{Vec}(\mathbf{U}_t) \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{V}_E \otimes \mathbf{U}_E)$. $\mathbf{U}_E$ and $\mathbf{V}_E$ are matrices with ones on the diagonal, and the off-diagonal entries are $1/p_1$ and $1/p_2$, respectively. Thus, by setting ϕ and ψ unequal to zero, the simulated factors are temporally correlated, and the idiosyncratic noises are both temporally and cross-sectionally correlated. The pair of factor numbers is assumed to be known except in Section 4.5, where we investigate the empirical performance of Algorithm 2 to estimate the numbers of factors. All the simulation results hereafter are based on 2000 replications.

4.2. Verifying the convergence rates for loading spaces

We first compare the performances of our **P**rojected **E**stimation (PE) method with those of the α-PCA method by Chen et al. (2020a) in terms of estimating the loadings. We consider two settings, where Setting A is for estimating the row factor loading matrix $\mathbf{R}$ while Setting B is designed for estimating the column loading matrix $\mathbf{C}$.

Setting A: $p_1 = 20$, $T = p_2 \in \{20, 50, 100, 150, 200\}$, $\phi = \psi = 0.1$.

Setting B: $p_2 = 20$, $T = p_1 \in \{20, 50, 100, 150, 200\}$, $\phi = \psi = 0.1$.

In view of identifiability, we evaluate the performances by the distance between the estimated loading space and true loading space. That is,

$$\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R}) = \left(1 - \frac{1}{k_1}\text{tr}(\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^\top \mathbf{Q}\mathbf{Q}^\top)\right)^{1/2},$$

where $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$ are the left singular-vector matrices of the true loading $\mathbf{R}$ and its estimator $\widehat{\mathbf{R}}$, respectively. $\mathcal{D}(\widehat{\mathbf{C}}, \mathbf{C})$ is defined similarly. Here, we abuse the notations $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ to mean any estimators of $\mathbf{R}$ and $\mathbf{C}$. The distance $\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R})$ is always between 0 and 1. When the corresponding matrices lie in the same space, they are equal to 0. If the two spaces are orthogonal, they are equal to 1. Once Assumptions A–E are satisfied, the squared distances would converge to 0 with the same rates as in Corollary 3.1.

Table 1 shows the averaged estimation errors with standard errors in parentheses under Settings A and B. We take $\alpha = -1, 0, 1$ for the α-PCA as in Chen et al. (2020a). All the methods benefit from large dimensions, and PE always shows

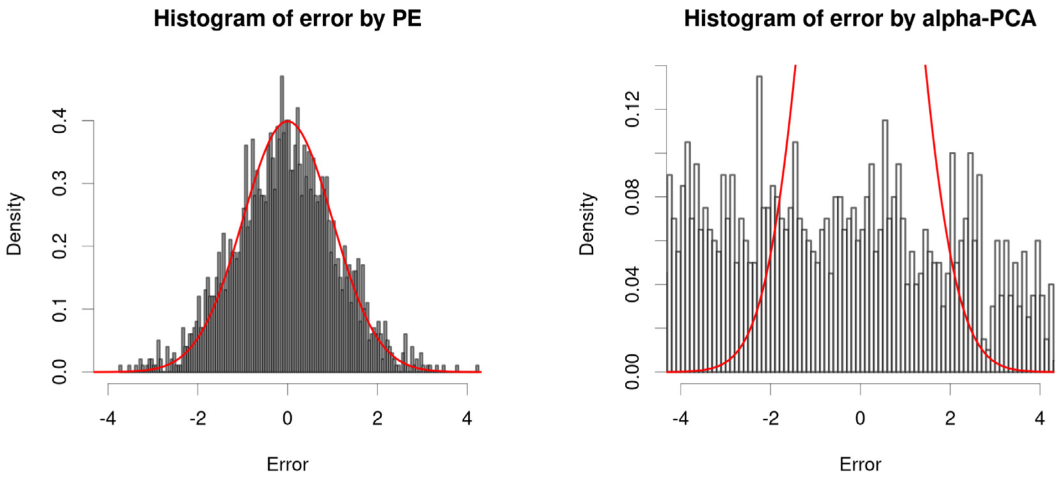


Fig. 3. Histograms of estimation errors for R_{11} after normalization over 2000 replications. $T = p_2 = 200$, $p_1 = 20$. left: PE. right: α-PCA with $\alpha = 0$. The red real line plots the probability density function of standard normal distribution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lowest estimation errors and standard errors. Fig. 2 plots the averaged log errors of the PE and α-PCA with $\alpha = 0$, which reflects different convergence rates of the estimators by PE and α-PCA. The left panel shows that the log error of the PE method for estimating $\mathbf{R}$ is almost linear to $\log(\sqrt{Tp_2})$ with slope -1, which matches the rate in Corollary 3.1. However, for the α-PCA method, the log error first decreases with growing $\log(\sqrt{Tp_2})$ but later tends to be invariant. This result is conceivable as the convergence rate of $\widehat{\mathbf{R}}$ by α-PCA depends only on p_1 when T and p_2 are sufficiently large. A similar conclusion can be drawn for the column factor loading matrix $\mathbf{C}$ from the right panel of Fig. 2. We conclude that the projected method leads to more accurate estimation of the loading spaces compared with α-PCA.

4.3. Verifying the asymptotic normality

In this section, we check the asymptotic normality of $\widetilde{\mathbf{R}}$ and verify the asymptotic variances in Theorem 3.2 by numerical studies. For data generation, we normalize $\mathbf{R}$ as $\sqrt{p_1}$ times its left singular-vector matrix such that the identification condition $\mathbf{R}^\top\mathbf{R}/p_1 = \mathbf{I}_{k_1}$ is satisfied. The column loading matrix $\mathbf{C}$ is normalized similarly. Let

$$\mathrm{Vec}(\mathbf{F}_t) \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad \text{with } \mathbf{D} = \mathrm{diag}(1.5, 1, 0.5, 1.5, 1, 0.5, 1.5, 1, 0.5),$$

so that the eigenvalues of $\boldsymbol{\Sigma}_1$ in Assumption B are distinct. The errors $\mathbf{E}_t$ are generated according to Eq. (4.1) with $\psi = 0$. Thus, both $\{\mathbf{F}_t, 1 \le t \le T\}$ and $\{\mathbf{E}_t, 1 \le t \le T\}$ are independent across time, which simplifies the calculation of the asymptotic covariance matrix. Actually, under the above setting, as $\min\{T, p_1, p_2\} \to \infty$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1.5 \end{pmatrix}, \sqrt{Tp_2}(\widetilde{\boldsymbol{R}}_{i\cdot} - \widetilde{\mathbf{H}}_1^\top \boldsymbol{R}_{i\cdot}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\widetilde{\mathbf{R}}}), \text{ where } \boldsymbol{\Sigma}_{\widetilde{\mathbf{R}}} := \frac{\mathrm{tr}(\mathbf{C}^\top \mathbf{V}_E \mathbf{C})}{3p_2}\boldsymbol{\Sigma}_1^{-1}.$$

We set $k_1 = k_2 = 3$, $p_1 = 20$, and $T = p_2 = 200$.

Fig. 3 shows the histograms of the first coordinates of $\sqrt{Tp_2}\boldsymbol{\Sigma}_{\widetilde{\mathbf{R}}}^{-1/2}(\widetilde{\boldsymbol{R}}_1 - \widetilde{\mathbf{H}}_1^\top \boldsymbol{R}_1)$ and $\sqrt{Tp_2}\boldsymbol{\Sigma}_{\widetilde{\mathbf{R}}}^{-1/2}(\widehat{\boldsymbol{R}}_1 - \widehat{\mathbf{H}}_1^\top \boldsymbol{R}_1)$. The asymptotic covariance matrices of the initial estimators and the projected estimators are the same theoretically, although the rotational matrices are not identical. Fig. 3 shows that the projected estimator is quite normally distributed, but the estimator by α-PCA deviates far from normality when $p_1 = 20$. This result is expected because the condition $Tp_2 = o(p_1^2)$ in Theorem 3.4 is not yet met, but the much looser condition for our projected estimators in Theorem 3.2 is already satisfied. When we increase p_1 to 400, both estimators show normality with the same standard deviations, as illustrated in Figure 1 in the supplementary material.

4.4. Estimation error for common components

In this subsection, we investigate the empirical performances of the PE and α-PCA methods in terms of estimating the common components under Setting A. We evaluate the performance of different methods by the mean squared error, i.e.,

$$\mathrm{MSE} = \frac{1}{Tp_1p_2}\sum_{t=1}^{T} \|\widehat{\mathbf{S}}_t - \mathbf{S}_t\|_F^2.$$

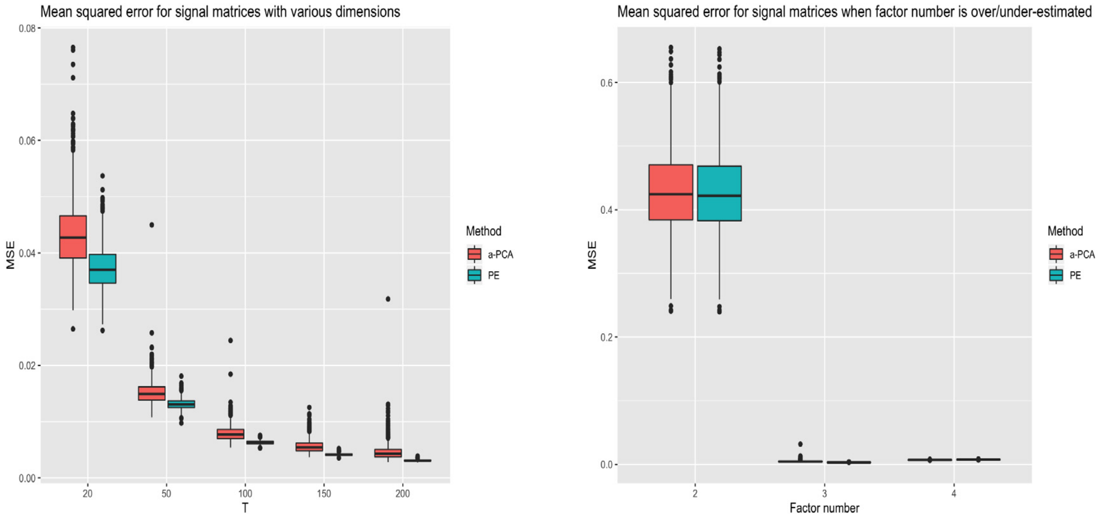


Fig. 4. Boxplots of MSEs for common components under Setting A over 2000 replications. Left: The true numbers of factors are given while $T = p_2$ grows. Right: The effects if we use less $(\widehat{k}_1 = \widehat{k}_2 = 2)$ or more $(\widehat{k}_1 = \widehat{k}_2 = 4)$ factors in the estimation, $T = p_2 = 200$. "PE": proposed projected method. "aPCA": α-PCA with $\alpha = 0$.

Table 2
The frequencies of exact estimation and underestimation (in parentheses) of the numbers of factors under Setting A over 2000 replications. $p_1 = 20$, $T = p_2$. "IterER(c)": the iterative algorithm with constant c in the denominator. "(a)PCA-ER": the α-PCA-ER in Chen et al. (2020a) with $\alpha = a$. "VER": eigenvalue ratio estimation of the vectorized model.

T	IterER(0)	IterER(10^{-4})	IterER(1)	(-1)PCA-ER	(0)PCA-ER	(1)PCA-ER	VER
20	**0.9875(0.0125)**	**0.9875(0.0125)**	**0.9875(0.0125)**	0.6185(0.3790)	0.6300(0.3675)	0.6195(0.3795)	0.0010(0.9990)
50	**1(0)**	**1(0)**	**1(0)**	0.8895(0.1085)	0.8910(0.1070)	0.8900(0.1080)	0.7655(0.2345)
100	**1(0)**	**1(0)**	**1(0)**	0.9015(0.0945)	0.9020(0.0940)	0.9000(0.0965)	0.9895(0.0105)
150	**1(0)**	**1(0)**	**1(0)**	0.9105(0.0865)	0.9090(0.0870)	0.9070(0.0885)	0.9985(0.0015)
200	**1(0)**	**1(0)**	**1(0)**	0.8985(0.0980)	0.9005(0.0955)	0.8995(0.0960)	0.9965(0.0035)

We also investigate the effects when we underestimate/overestimate the numbers of factors.

Fig. 4 shows the boxplots of the MSEs by PE and α-PCA with $\alpha = 0$ under Setting A over 2000 replications. We do not show the results for $\alpha = \pm 1$ because the choice of α has a minimal effect on the results under this setting. The left panel of Fig. 4 shows the boxplots of MSEs for various T by the PE and α-PCA estimates with the true numbers of factors. Both methods perform better as the dimension $T = p_2$ grows, and the PE always leads to lower MSEs and smaller deviations. The right panel corresponds to the case when we overestimate/underestimate the numbers of factors with $T = p_2 = 200$. The effect of overestimation is negligible while underestimating the numbers of factors results in intolerable MSEs, which is consistent with the findings in vector factor models. Detailed numerical results are reported in Table 1 of the supplementary material.

4.5. Estimating the numbers of factors

As shown in Fig. 4, accurate specification of the numbers of factors is critical to the matrix factor model. In this subsection, we compare the empirical performances of the Vectorized Eigenvalue-Ratio (VER) criterion in Ahn and Horenstein (2013), α-PCA based ER method (α-PCA-ER) in Chen et al. (2020a), and the proposed iterative method in Algorithm 2 (IterER) in terms of estimating the numbers of factors.

Table 2 presents the frequencies of exact estimation and underestimation over 2000 replications under Setting A. For the proposed IterER method, we try $c = 0, 10^{-4}, 1$ in Eq. (3.5). For the VER criterion in Ahn and Horenstein (2013), we first vectorize the matrix observations and regard the true number of factors as $k_1 \times k_2$. We set $k_{\max} = 8$ for IterER and α-PCA-ER while $k_{\max} = 64$ for VER. We find that the IterER has the highest accuracy and lowest underestimation risk even with small T, while the others only work when T is large. Under this setting, the constant c and α have minimal effect on the results.

As one reviewer pointed out, the specification criterion of the numbers of factors deserves more concern when the mean of the data matrix is not zero. If the data are not demeaned suitably, this may enlarge the first eigenvalue of the corresponding matrix and result in underestimation of the numbers of factors. We propose to first demean the vectorized data either by subtracting the sample mean or adopting the double-demeaned strategy by Ahn and Horenstein (2013), and then structure the demeaned vectors into matrices. In the supplementary material, we have designed two settings

Table 3
Loading matrices for Fama–French data set, after varimax rotation and scaling by 30. “PE“ stands for the projected estimator, “ACCE” stands for the approach used by Wang et al. (2019), while α-PCA represents the method in Chen et al. (2020a) with $\alpha = 0$.

Size											
Method	Factor	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
PE	1	−16	−15	−12	−10	−8	−5	−3	−1	4	7
	2	−6	−1	3	5	8	11	12	13	15	10
ACCE	1	−12	−14	−12	−13	−10	−6	−3	−1	4	9
	2	−1	−1	−1	2	5	10	11	18	15	11
α-PCA	1	−14	−14	−13	−11	−9	−7	−4	−2	3	7
	2	−4	−2	1	3	6	9	12	13	16	14
Book-to-Equity											
Method	Factor	BE1	BE2	BE3	BE4	BE5	BE6	BE7	BE8	BE9	BE10
PE	1	6	1	−4	−7	−10	−11	−12	−12	−12	−10
	2	20	17	11	8	4	2	0	−1	−1	0
ACCE	1	6	−1	−4	−8	−8	−9	−10	−13	−15	−12
	2	21	15	11	6	5	2	1	−2	−3	1
α-PCA	1	6	2	−4	−7	−10	−11	−12	−13	−12	−11
	2	19	18	12	8	4	2	0	−1	−1	−1

with nonzero mean to investigate the empirical performances of different methods. Both demean strategies work well and the proposed iterative algorithm outperforms others, see Tables 1 and 2 in the supplementary material.

5. Real data analysis

5.1. Fama–French 10 × 10 portfolios

For ease of comparison, in our first real example we use the same dataset as that used by Wang et al. (2019). This dataset consists of monthly returns of 100 portfolios, structured into a 10 × 10 matrix according to 10 levels of market capital size (S1–S10) and 10 levels of book-to-equity ratio (BE1-BE10). The monthly returns from January 1964 to December 2019 are collected, covering 672 months. Detailed information can be found on the website http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Following Wang et al. (2019), we adjusted the return series by first subtracting the corresponding monthly market excess returns and then standardizing each of the series. We imputed the missing values by the factor-model-based method introduced in Xiong and Pelger (2019). The augmented Dickey–Fuller test rejects the null hypotheses for all the series, which indicates stationarity. With the preprocessed monthly returns, our iterative eigenvalue-ratio method suggests that $k_1 = 2$ while $k_2 = 1$. For better illustration, we take $k_1 = k_2 = 2$. The estimated front and back loading matrices after varimax rotation and scaling are reported in Table 3.

From the table, we observe that the PE, α-PCA and Auto-Cross-Correlation Estimation (ACCE) method by Wang et al. (2019) lead to very similar estimated loadings. From the perspective of size, the small size portfolios load heavily on the first factor while the large size portfolios load mainly on the second factor. From the perspective of book-to-equity, the small BE portfolios load heavily on the second factor while the large BE portfolios load mainly on the first factor.

To further compare these methods, we use a rolling-validation procedure as in Wang et al. (2019). For each year t from 1996 to 2019, we repeatedly use the n (bandwidth) years observations before t to fit the matrix-variate factor model and estimate the two loading matrices. The loadings are then used to estimate the factors and corresponding residuals of the 12 months in the current year. Specifically, let $\mathbf{Y}_t^i$ and $\widehat{\mathbf{Y}}_t^i$ be the observed and estimated price matrix of month i in year t, denote $\bar{\mathbf{Y}}_t$ as the mean price matrix, and further define

$$\text{MSE}_t = \frac{1}{12 \times 10 \times 10} \sum_{i=1}^{12} \|\widehat{\mathbf{Y}}_t^i - \mathbf{Y}_t^i\|_F^2, \quad \rho_t = \frac{\sum_{i=1}^{12} \|\widehat{\mathbf{Y}}_t^i - \mathbf{Y}_t^i\|_F^2}{\sum_{i=1}^{12} \|\mathbf{Y}_t^i - \bar{\mathbf{Y}}_t\|_F^2},$$

as the mean squared pricing error and unexplained proportion of total variances, respectively. During the rolling-validation procedure, the variation of loading space is measured by $v_t := \mathcal{D}(\widehat{\mathbf{C}}_t \otimes \widehat{\mathbf{R}}_t, \widehat{\mathbf{C}}_{t-1} \otimes \widehat{\mathbf{R}}_{t-1})$. The matrix factor model (1.1) can be written in vector form with $(\mathbf{C} \otimes \mathbf{R})$ being the loading matrix.

Table 4 reports the means of MSE, ρ and v by PE, ACCE, α-PCA and a conventional PCA estimation applied to the vectorized data. Diverse combinations of bandwidth n and numbers of factors ($k_1 = k_2 = k$) are compared. On the one hand, the pricing errors of PE, α-PCA, and the vector model are very close especially for large n and k, but lower than the ACCE method. On the other hand, in terms of estimating the loading space, PE always performs much more stably compared with the other two methods. Financial data are usually heavily-tailed with outliers, so the more robust PE method is preferred to control transaction costs and reduce risks.

Table 4
Rolling validation for the Fama–French portfolios. $12n$ is the sample size of the training set. $k_1 = k_2 = k$ is the number of factors. $\overline{MSE}$, $\bar{\rho}$, $\bar{v}$ are the mean pricing error, mean unexplained proportion of total variances and mean variation of the estimated loading space. "PE", "ACCE" and "α-PCA" are the same as in Table 3, while "Vec" is the PCA applied to the vectorized data.

n	k	$\overline{MSE}$				$\bar{\rho}$				$\bar{v}$			
		PE	ACCE	α-PCA	Vec	PE	ACCE	α-PCA	Vec	PE	ACCE	α-PCA	Vec
5	1	0.870	0.885	**0.862**	0.910	0.802	0.828	**0.796**	0.840	**0.176**	0.303	0.241	0.252
10	1	**0.855**	0.880	0.860	0.939	**0.784**	0.815	0.791	0.863	**0.085**	0.165	0.203	0.182
15	1	**0.853**	0.884	0.860	0.894	**0.782**	0.812	0.792	0.814	**0.064**	0.152	0.234	0.131
5	2	0.596	0.667	0.601	**0.588**	0.625	0.673	0.628	**0.623**	**0.239**	0.460	0.350	0.398
10	2	0.601	0.655	0.611	**0.592**	0.628	0.668	0.636	**0.622**	**0.092**	0.257	0.261	0.208
15	2	0.603	0.637	0.612	**0.587**	0.626	0.652	0.630	**0.614**	**0.057**	0.189	0.173	0.188
5	3	**0.522**	0.564	0.529	0.532	**0.550**	0.590	0.556	0.564	**0.286**	0.497	0.432	0.477
10	3	**0.519**	0.573	0.526	0.528	**0.548**	0.594	0.555	0.561	**0.114**	0.303	0.353	0.300
15	3	**0.517**	0.560	0.522	0.523	**0.545**	0.582	0.544	0.554	**0.084**	0.299	0.308	0.254

Table 5
Row (countries) loading matrices by PE, ACCE and α-PCA ($\alpha = 0$) for multinational macroeconomic index dataset, varimax rotated and multiplied by 10.

Method	Factor	AUS	NZL	USA	CAN	NOR	DEU	FRA	GBR
	1	0	1	−7	−6	3	−3	−2	−1
PE	2	2	−2	1	−1	−7	−2	−5	−5
	3	8	6	0	−1	0	2	−1	1
	1	2	−2	1	−2	−6	0	−6	−5
ACCE	2	7	5	0	0	0	5	0	0
	3	0	−2	8	4	−2	1	1	2
	1	−1	1	-7	−5	3	−3	−2	−1
α-PCA	2	1	−1	0	−1	−7	−2	−5	−4
	3	−7	−7	0	1	0	−1	1	−1

5.2. Multinational macroeconomic indices

In the second real example, we analyze a multinational macroeconomic index dataset collected from OECD. A similar dataset is studied in Chen et al. (2020a). It contains 10 macroeconomic indices across 8 countries over 130 quarters from 1988-Q1 to 2020-Q2. The countries are the United States, the United Kingdom, Canada, France, Germany, Norway, Australia and New Zealand. The indices are from 4 major groups, namely consumer price, interest rate, production, and international trade. Logarithm transform and difference operators are applied to each of the series according to Chen et al. (2020a) so that the α-mixing assumption is satisfied. Detailed description can be found in our supplementary material. We further standardize each of the transformed series to avoid the effects of non-zero mean or disparate variances.

The first step is to determine the numbers of row and column factors. The proposed iterative algorithm suggests taking $k_1 = 1$ and $k_2 = 5$ for the 8×10 matrix-valued observations. For better illustration, we take $k_1 = 3$ and $k_2 = 4$ such that the row and column factors can explain nearly 75% variances of the matrices $\widetilde{\mathbf{M}}_1$ and $\widetilde{\mathbf{M}}_2$. The estimated loading matrices are reported in Tables 5 and 6. For the row factors, Table 5 shows that they are closely related to the geographical location. The neighboring countries tend to load similarly on the factors. The 8 countries (excluding Germany) naturally divide into 3 groups, the Oceania, North American and European. For the column factors, the macroeconomic indices are divided into 4 groups (consumer price, interest rate, production and international trade) in Table 6, which coincides with the economic interpretation.

A rolling procedure is applied to the multinational macroeconomic index data set to check the validation errors of different methods. Table 6 in the supplementary material shows that the PE method leads to most stable estimates, although the PCA method for the vectorized data produces smallest validation errors in this example, which is possibly due to deeper complexity in the sense of more parameters for the vector factor model.

At last, we evaluate the practical utility of different methods by a rolling prediction procedure. At each quarter t, denote y_t as the preprocessed stationary series. First, we set y_t as the change of inflation (second-order difference of the log level of the total consumer price index–CPI:Tot) of a selected country at time t, $\boldsymbol{x}_t$ the vector of all the other 9 indices of the selected country at time t, and $\mathbf{Z}_t$ the 8×10 matrix observation for all indices and all countries at time t. We predict y_{t+1} using the following Auto-Regression (AR) model (Model 1) and Factor-Augmented-Auto-Regression (FAAR) models (Models 2–4), which are similar to the Diffusion Index Forecasts in Stock and Watson (2002b).

Model 1 $y_{t+1} = a + by_t + \epsilon_{t+1}$,

Model 2 $y_{t+1} = a + by_t + \boldsymbol{\beta}^\top \boldsymbol{f}_{1t} + \epsilon_{t+1}$, where $\boldsymbol{f}_{1t}$'s are estimated from the vector factor model with observations $\{\boldsymbol{x}_t\}$.

Table 6
Column (indices) loading matrices by PE, ACCE, and α-PCA ($\alpha = 0$) for multinational macroeconomic index dataset, varimax rotated and multiplied by 10.

Method	Factor	CPI:Tot	CPI:Ener	CPI:NFNE	IR:3-Mon	IR:Long	P:TIEC	P:TM	GDP	IT:Ex	IT:Im
PE	1	1	−2	3	1	−1	6	7	2	−1	0
	2	6	7	3	−1	1	0	0	0	0	0
	3	0	0	−1	−6	−8	0	0	1	0	−1
	4	1	−2	3	0	0	−1	0	−5	−6	−5
ACCE	1	0	0	0	0	1	−7	−7	0	0	0
	2	1	0	0	−5	−4	1	−1	−2	−4	−6
	3	−4	2	−9	0	0	2	−1	2	0	0
	4	6	7	0	−1	3	0	0	2	1	−1
α-PCA	1	0	−1	1	1	−1	7	6	4	0	0
	2	7	5	5	−1	1	0	1	0	0	0
	3	0	0	0	−7	−7	1	0	−1	1	0
	4	0	2	−2	0	0	0	0	2	7	6

Table 7
MAPEs for inflation (at annual rate) for different countries with different models. $k_1 = 3, k_2 = 4$.

Country	Model 1	Model 2	Model 3	Model 4: PE	Model 4: ACCE	Model 4: α-PCA
AUS	**1.588**	1.626	**1.599**	**1.589**	**1.585**	**1.598**
NZL	**1.887**	1.909	2.051	2.004	1.929	**1.918**
USA	2.835	2.584	2.471	2.353	**2.247**	2.378
CAN	2.461	2.235	1.965	**1.852**	**1.827**	1.902
NOR	**1.936**	2.261	2.387	2.155	2.444	2.313
DEU	1.875	1.957	**1.728**	**1.733**	1.804	**1.74**0
FRA	1.806	1.832	**1.241**	1.407	1.382	1.433
GBR	1.578	1.502	**1.198**	1.324	1.335	1.349

Model 3 $y_{t+1} = a+by_t+\boldsymbol{\beta}^\top \boldsymbol{f}_{2t}+\epsilon_{t+1}$, where $\boldsymbol{f}_{2t}$'s are estimated from the vector factor model with observations $\{\mathrm{Vec}(\mathbf{Z}_t)\}$.

Model 4 $y_{t+1} = a + by_t + \boldsymbol{\beta}^\top \mathrm{Vec}(\mathbf{F}_t) + \epsilon_{t+1}$, where $\mathbf{F}_t$'s are estimated from the matrix factor model with observations $\{\mathbf{Z}_t\}$, by the PE, ACCE, and α-PCA, respectively.

Model 1 is a simple auto-regression model. In Model 2, we add common index factors of the selected country into the auto-regression model. In Model 3 and Model 4, both index and country factors are taken into account. Model 4 differs from Model 3 in that Model 4 assumes the more parsimonious matrix factor structure while Model 3 vectorizes the matrix series and assumes the vector factor structure. To avoid possible over-fitting in prediction, we use the least absolute shrinkage and selection operator (LASSO) invented in Tibshirani (1996) to select factors and estimate the coefficients for Models 2-4.

For each quarter t from 2008-Q1 to 2020-Q2, we use the 80 neighboring observations before t to train the models and predict y_{t+1} (denoted as $\widehat{y}_{t+1}$). As y_t was standardized in preprocessing, we transformed the predicted y_{t+1} to match the change of inflation rate by multiplying the standard deviation and adding back the sample mean. For simplicity of notation, we still denote the transformed predictor as $\widehat{y}_{t+1}$. The inflation I_{t+1} is then predicted by integrating $\widehat{y}_{t+1}$ and I_t, i.e., $\widehat{I_{t+1}} = \widehat{y}_{t+1} + I_t$. In Model 2 and Model 3, the factor numbers before model selection are set as k_2 and $k_1 \times k_2$, respectively.

We fix $(k_1, k_2) = (3, 4)$ to save space and similar conclusions can be drawn for other pairs of (k_1, k_2), see for example Tables 7–8 in the supplementary material. Table 7 reports the mean absolute prediction errors (MAPEs) for the annualized inflation rates. We observe from Table 7 that the simple AR model does the best for Norway and New Zealand in predicting the inflation rate, and the index and country factors act as noise in Models 2–4. For the largest Oceania country, Australia, Model 4 with ACCE does slightly better, but all models perform comparably well. For large American and European countries, USA, Canada, France, Great Britain and Germany, both the country and index factors do reduce the prediction errors substantially, while the more parsimonious matrix factor structure further improves the prediction for two American countries. It looks that more connected countries' inflation rates rely more on the country factors.

Next, we consider the rolling-prediction of the GDP growth rates (first-order difference of the log level of GDP) for all countries. The conclusion almost seems to reverse for predicting the GDP growth rates of different countries compared with that of inflation. The prediction errors for the GDP growth rates of different countries are displayed in Table 8. It demonstrates that for the strong manufacturing American and European countries, USA, Germany, France and Great Britain, simple auto-regressions suffice to predict the GDP growth rates well. That being said, for these countries, the GDP growth rates depend mainly on their own past. For the remaining countries, the country and index factors contribute in predicting the GDP growth rates, while for Canada and Norway, the matrix factor structure further improves the prediction. The above real data analysis also demonstrates that the factor augmentation may not always work in prediction

Table 8
MAPEs for the growth rate of GDP (at annual rate) for different countries by different models. $k_1 = 3, k_2 = 4$.

Country	Model 1	Model 2	Model 3	Model 4: PE	Model 4: ACCE	Model 4: α-PCA
AUS	1.890	**1.871**	**1.867**	**1.873**	1.912	**1.871**
NZL	2.921	2.912	**2.839**	2.927	3.037	3.036
USA	**2.432**	**2.468**	2.566	2.582	2.507	2.608
CAN	2.828	2.858	2.636	**2.550**	2.687	2.654
NOR	3.779	3.898	3.738	3.901	**3.674**	3.896
DEU	**3.297**	3.341	3.917	3.448	3.584	3.523
FRA	**2.678**	2.839	2.854	2.808	2.745	2.870
GBR	**3.252**	3.323	3.329	3.352	3.425	3.384

for all countries or any index. This naturally raises the question of whether the matrix series supports the factor structure, especially the parsimonious matrix factor structure. We leave it to the future work and further discussions can be found in the Conclusion-and-Discussion section later.

6. Conclusions and discussions

The current study focuses on the estimation of matrix factor models. We start with the column or row sample covariances instead of the auto-cross covariances for the estimation of the front and back loading matrices. A projected approach is proposed to improve the estimation accuracy. Statistical convergence rates and asymptotic distributions of the estimated loadings are provided under mild conditions. An iterative approach is introduced to determine the numbers of factors. Thorough numerical studies and real examples show the advantages of the projected method over existing methods. The matrix factor models can be further extended to analyze high-order tensor data, such as video streaming, which are widely applied in recommender systems. This subject will be addressed in a future study. We are also interested in incorporating the matrix factor structure into estimating large-dimensional covariance matrices or detecting structure breaks.

As the empirical studies show, the matrix factor structure may not improve the prediction of multinational macroeconomic variables. This raises the question of whether the matrix factor structure is true for the time series, and hence statistical tests are needed to check the matrix factor model. More precisely, the matrix factor model amounts to setting $\text{Vec}(\mathbf{X}_t) = (\mathbf{C}\otimes\mathbf{R})\text{Vec}(\mathbf{F}_t)+\text{Vec}(\mathbf{E}_t)$, where $\otimes$ is the Kronecker product operator. Under the framework of high-dimensional vector factor model, the model checking of the matrix factor structure is equivalent to testing for the Kronecker constraints on the loading parameters. We leave it to our future research works.

Acknowledgments

We thank Prof. Bing-Yi Jing with the Hong Kong University of Science and Technology for very kind help in revising the paper. We also thank the editor Serena Ng, the associate editor and two anonymous referees for constructive comments and suggestions which improve the quality of the paper a lot. He's work is supported by NSF China (11801316), NSF of Shandong Province, China (ZR2019QA002) and the Fundamental Research Fund of Shandong University, China. Kong's work is partially supported by NSF China (71971118 and 11831008) and the WRJH-QNBJ Project and Qinglan Project of Jiangsu Province, China. Zhang's work is partially supported by NSF China (11971116).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2021.04.001. The technical proofs of the main results, extra empirical studies, and details of the data sets are included in the Supplementary Material.

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81 (3), 1203–1227.
Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer Science & Business Media.
Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71 (1), 135–171.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.
Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica 51 (5), 1281–1304.
Chen, E.Y., Chen, R., 2020. Modeling dynamic transport network with matrix factor models: with an application to international trade flow.. arXiv:1901.00769.
Chen, E.Y., Fan, J., Li, E., 2020a. Statistical inference for high-dimensional matrix-variate factor model.. arXiv:2001.01890.
Chen, E.Y., Tsay, R.S., Chen, R., 2020b. Constrained factor models for high-dimensional matrix-variate time series. J. Amer. Statist. Assoc. 115 (530), 775–793.
Davis, C., Kahan, W.M., 1970. The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal. 7 (1), 1–46.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econom. 33 (1), 3–56.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 603–680.

Fan, J., Liao, Y., Shi, X., 2015. Risks of large portfolios. J. Econometrics 186 (2), 367–387.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. 82 (4), 540–554.

He, Y., Kong, X., Yu, L., Zhang, X., 2020. Large-dimensional factor analysis without moment constraints. J. Bus. Econ. Statist. http://dx.doi.org/10.1080/07350015.2020.1811101, (in press).

Kong, X., Wang, J., Xing, J., Xu, C., Ying, C., 2019. Factor and idiosyncratic empirical processes. J. Amer. Statist. Assoc. 114 (527), 1138–1146.

Lam, C., Yao, Q., 2012. Factor modeling for high-dimensional time series: inference for the number of factors. Ann. Statist. 40 (2), 694–726.

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

Lettau, M., Pelger, M., 2020. Factors that fit the time series and cross-section of stock returns. Rev. Financ. Stud. 33 (5), 2274–2325.

Liu, X., Chen, E., 2019. Helping effects against curse of dimensionality in threshold factor models for matrix time series.. arXiv:1904.07383.

Ross, S.A., 1977. The capital asset pricing model (CAPM), short-sale restrictions and related issues. J. Finance 32 (1), 177–183.

Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97 (460), 1167–1179.

Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20 (2), 147–162.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Virta, J., Li, B., Nordhausen, K., Oja, H., 2017. Independent component analysis for tensor-valued data. J. Multivariate Anal. 162, 172–192.

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

Xiong, R., Pelger, M., 2019. Large dimensional latent factor modeling with missing observations and applications to causal inference.. arXiv:1910.08273.

Yu, Y., Wang, T., Samworth, R.J., 2015. A useful variant of the davis-kahan theorem for statisticians. Biometrika 102 (2), 315–323.