Modelling matrix time series via a tensor CP-decomposition

Jinyuan Chang[1,2] , **Jing He[2]**, **Lin Yang[2] and Qiwei Yao[3]**

[1]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, Zhejiang, China
[2]Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan, China
[3]Department of Statistics, The London School of Economics and Political Science, London, UK

Address for correspondence: Jinyuan Chang, Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China. Email: changjinyuan@swufe.edu.cn

Abstract

We consider to model matrix time series based on a tensor canonical polyadic (CP)-decomposition. Instead of using an iterative algorithm which is the standard practice for estimating CP-decompositions, we propose a new and one-pass estimation procedure based on a generalized eigenanalysis constructed from the serial dependence structure of the underlying process. To overcome the intricacy of solving a rank-reduced generalized eigenequation, we propose a further refined approach which projects it into a lower-dimensional full-ranked eigenequation. This refined method can significantly improve the finite-sample performance. We show that all the component coefficient vectors in the CP-decomposition can be estimated consistently. The proposed model and the estimation method are also illustrated with both simulated and real data, showing effective dimension-reduction in modelling and forecasting matrix time series.

Keywords: dimension-reduction, generalized eigenanalysis, matrix time series, tensor CP-decomposition

1 Introduction

Let $\mathbf{Y}_t = (y_{i,j,t})$ be a $p \times q$ matrix time series, i.e., there are pq recorded values at each time t from, for example, p individuals and over q indices or variables, and $y_{i,j,t}$ is then the value of the jth variable on the ith individual at time t. Given available observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, the goal is to build a dynamic model for $\mathbf{Y}_t$ and to forecast the future values $\mathbf{Y}_{n+\ell}$ for $\ell \geq 1$. With moderately large p and q, any direct attempts based on the time series ARMA framework are unlikely to be successful due to overparametrization. We seek a low-dimensional structure via a tensor canonical polyadic (CP) decomposition. To this end, we denote by $\mathcal{Y}$ the $p \times q \times n$ tensor with $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ as its n frontal slices (Kolda & Bader, 2009). Then $y_{i,j,t}$ is the (i, j, t)th element of $\mathcal{Y}$. Conceptually, we decompose $\mathcal{Y}$ into two parts:

$$\mathcal{Y} = \mathcal{X} + \mathcal{E}, \tag{1}$$

where all the dynamic structure of $\mathcal{Y}$ is reflected by $\mathcal{X}$, and the frontal slices of $\mathcal{E} \equiv (\varepsilon_{i,j,t})$ are matrix white noise, i.e., $\mathrm{Cov}(\varepsilon_{i,j,t}, \varepsilon_{k,\ell,s}) = 0$ for any $t \neq s$. The key idea is to perform a CP-decomposition for $\mathcal{X}$, i.e., to express it as a sum of rank one tensors (see (2)). This effectively represents the dynamic structure of matrix process $\mathbf{Y}_t$ in terms of that of a vector process, and, hence, achieving an effective dimension-reduction in modelling the dynamic behaviour of the process.

The 'workhorse' method for CP-decompositions is the so-called alternative least squares (ALS) algorithm which is easy to understand and implement. See Section 3.4 of Kolda and Bader (2009) and the references therein. However, it has obvious drawbacks. For example, an ALS algorithm takes many iterations to converge. It is not guaranteed to converge to the global minimum even

for moderately large p, q, or n. Furthermore, it also depends sensitively on the selection of the initial values. Substantial effort has been made to improve the convergence and the performance of the ALS algorithm, including, among others, Anandkumar et al. (2014), Liu et al. (2014), Colombo and Vlassis (2016), Sun et al. (2017), Sharan and Valiant (2017), Wang and Song (2017), Zhang and Xia (2018) and Han and Zhang (in press).

We propose a new and one-pass estimation procedure in this paper. The new method is inspired by Sanchez and Kowalski (1990) which transforms a CP-decomposition into a generalized eigenanalysis problem. While Sanchez and Kolwalski's approach does not require iteration, it only works for the noise-free cases with $\mathcal{E} \equiv 0$ in (1). In contrast, our new procedure eliminates the impact of the noise by incorporating the serial dependence into the estimation. Furthermore to overcome the intricacy in solving a generalized eigenequation defined by rank-reduced matrices (see Section 7.7 of Golub & Van Load, 2013), we propose a new refined approach which projects the rank-reduced generalized eigenequation to a full-ranked lower-dimensional one which is, therefore, equivalent to a standard eigenequation. The numerical results in simulation also demonstrate the significant improvement in the finite-sample performance by this refined method.

Most existing literature on matrix time series is based on the factor modelling via the Tucker decomposition; see Chen and Chen (2019), Wang et al. (2019) and Chen et al. (2020). The key difference between our approach and the Tucker decomposition-based approaches is twofold. First, a Tucker decomposition represents a matrix process as a linear combination of a smaller matrix process while a CP-decomposition is more canonical in the sense that it represents a matrix process in terms of a vector process; see also the real data example in Section 5.2. Second, a Tucker decomposition entails more conventional factor models, and, therefore, we only need to identify and estimate the factor loading spaces, for which the standard factor model methods (e.g., Chang et al., 2015; Lam & Yao, 2012) are applicable. However, for a CP-decomposition, we need to identify and estimate the component coefficient vectors precisely. Therefore, a radically new inference procedure is required. The other approaches for modelling matrix time series include: the matrix-coefficient autoregressive models of Chen et al. (2021), and the bilinear transformation segmentation method of Han et al. (in press). Han et al. (2021) models tensor time series also based on a CP-decomposition. But their approach is radically different from ours, as they estimate the CP-decomposition based on an iterative simultaneous orthogonalization algorithm with a warm-start initialization using the so-called composite principal component analysis for tensors; see Section 3 of Han et al. (2021). Note that our estimation is a one-pass procedure, and no iterations are required.

The rest of the paper is organized as follows. The matrix time series model based on a CP-decomposition is presented in Section 2. Section 3 deals with the model identification and presents the newly proposed estimation procedures. The asymptotic results, including the convergence rates for the estimated component vectors in the CP-decomposition, are presented in Section 4. Numerical illustration with both simulated and real datasets is given in Section 5. All the technical proofs are relegated to the online supplementary material.

Notations. For a positive integer m, write $[m] = \{1, \ldots, m\}$, and denote by $\mathbf{I}_m$ the $m \times m$ identity matrix. Let $I(\cdot)$ be the indicator function. For an $m_1 \times m_2$ matrix $\mathbf{H} = (h_{i,j})_{m_1 \times m_2}$, let $\|\mathbf{H}\|_2$, rank$(\mathbf{H})$, $\sigma_{\min}(\mathbf{H})$, and vec$(\mathbf{H})$ be, respectively, its spectral norm, its rank, its smallest singular value, and a vector obtained by stacking together the columns of $\mathbf{H}$. Specifically, if $m_2 = 1$, we use $|\mathbf{H}|_2 = (\sum_{i=1}^{m_1} |h_{i,1}|^2)^{1/2}$ to denote the ℓ^2-norm of the $m_1 \times 1$ vector $\mathbf{H}$. Also, denote by $\mathbf{H}^\top$ and $\mathbf{H}^H$, respectively, the transpose and conjugate transpose of $\mathbf{H}$. When rank$(\mathbf{H}) = m_2$, denote by $\mathbf{H}^+$, an $m_2 \times m_1$ matrix, the Moore–Penrose inverse of $\mathbf{H}$ such that $\mathbf{H}^+\mathbf{H} = \mathbf{I}_{m_2}$. When $m_1 = m_2$, denote by det$(\mathbf{H})$ and tr$(\mathbf{H})$ the determinant and the trace of $\mathbf{H}$, respectively. Let $\otimes$ and $\circ$ denote the Kronecker product and the vector outer product, respectively. For any vector $\mathbf{h} = (h_1, \ldots, h_m)^\top$, we write Re$(\mathbf{h}) = \{Re(h_1), \ldots,$ Re$(h_m)\}^\top$ and Im$(\mathbf{h}) = \{$Im$(h_1), \ldots,$ Im$(h_m)\}^\top$, where Re(h_i) and Im(h_i) denote, respectively, the real part and the imaginary part of h_i.

2 Model

We impose a low-dimensional dynamic structure in model (1) as follows:

$$\mathcal{Y} = \sum_{\ell=1}^{d} \mathbf{a}_\ell \circ \mathbf{b}_\ell \circ \mathbf{x}_\ell + \mathcal{E}, \tag{2}$$

where $\mathbf{a}_\ell = (a_{1,\ell}, \ldots, a_{p,\ell})^\top$ and $\mathbf{b}_\ell = (b_{1,\ell}, \ldots, b_{q,\ell})^\top$ are, respectively, $p \times 1$ and $q \times 1$ constant vectors, $\mathbf{x}_\ell = (x_{1,\ell}, \ldots, x_{n,\ell})^\top$ is an $n \times 1$ random vector, and $1 \leq d < \min(p, q)$ is an unknown integer. Put

$$\mathbf{A} \equiv (a_{i,\ell})_{p \times d} = (\mathbf{a}_1, \ldots, \mathbf{a}_d) \quad \text{and} \quad \mathbf{B} \equiv (b_{j,\ell})_{q \times d} = (\mathbf{b}_1, \ldots, \mathbf{b}_d).$$

Then componentwisely (2) admits the representation

$$y_{i,j,t} = \sum_{\ell=1}^d a_{i,\ell} b_{j,\ell} x_{t,\ell} + \varepsilon_{i,j,t}. \tag{3}$$

Hence the dynamic structure in $\mathcal{Y}$ is entirely determined by that of the d time series $\mathbf{x}_1, \ldots, \mathbf{x}_d$. There is a clearly scaling indeterminacy in (2), as the triple $(\mathbf{a}_\ell, \mathbf{b}_\ell, \mathbf{x}_\ell)$ can be replaced by $(\alpha_\ell \mathbf{a}_\ell, \beta_\ell \mathbf{b}_\ell, \gamma_\ell \mathbf{x}_\ell)$ as long as $\alpha_\ell \beta_\ell \gamma_\ell = 1$. We assume that all $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are unit vectors (i.e., $|\mathbf{a}_\ell|_2 = |\mathbf{b}_\ell|_2 = 1$). Once $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are specified, $|\mathbf{x}_\ell|_2$ will be determined by (2) accordingly. Note that $\mathbf{a}_1, \ldots, \mathbf{a}_d$ (or $\mathbf{b}_1, \ldots, \mathbf{b}_d$) are not required to be orthogonal with each other.

Model (2) is resulted from applying the CP-decomposition to $\mathcal{X}$ in (1), where d is the order of the CP-decomposition. Note that this decomposition is unique upto the scaling and permutation indeterminacy if $\mathcal{R}(\mathbf{A}) + \mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathcal{X}) \geq 2d + 2$, where $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ and $\mathcal{R}(\cdot) = \max\{k : \text{any } k \text{ columns of the matrix} \cdot \text{are linear independent}\}$. Such requirement provides a sufficient condition for the uniqueness (Kolda & Bader, 2009, p. 467). See also Theorems 1.5 and 1.7 of Domanov and De Lathauwer (2014) for more refined results on the uniqueness of the CP-decomposition.

Though $y_{i,j,t}$ is a linear combination of $x_{t,1}, \ldots, x_{t,d}$ under (2), the factor representation of the model admits some special structure, i.e., the elements of the factor loading matrix are of the form of $a_{i,\ell} b_{j,\ell}$; see (3). In fact, we need to identify and estimate all the vectors in the first term on the RHS of (2) precisely (upto the permutation and scaling indeterminacy). Therefore, the conventional factor model estimation methods such as Lam and Yao (2012) and Chang et al. (2015) do not apply.

The frontal slice equation of (2) admits the form

$$\mathbf{Y}_t = \sum_{\ell=1}^d \mathbf{a}_\ell \circ \mathbf{b}_\ell x_{t,\ell} + \varepsilon_t = \sum_{\ell=1}^d x_{t,\ell} \mathbf{a}_\ell \mathbf{b}_\ell^\top + \varepsilon_t = \mathbf{A} \mathbf{X}_t \mathbf{B}^\top + \varepsilon_t, \tag{4}$$

where $\mathbf{X}_t = \mathrm{diag}(x_{t,1}, \ldots, x_{t,d})$ and ε_t denotes the $p \times q$ matrix with $\varepsilon_{i,j,t}$ as its (i, j)th element. We impose the following regularity condition on the model.

Condition 1 It holds that $\mathrm{rank}(\mathbf{A}) = d = \mathrm{rank}(\mathbf{B})$. Furthermore, $\mathbb{E}(\varepsilon_t) = 0$ for any t, $\mathbb{E}(\varepsilon_t \otimes \varepsilon_s) = 0$ for all $t \neq s$, and $\mathbb{E}(x_{t,\ell} \varepsilon_s) = 0$ for any $\ell \in [d]$ and $t \leq s$.

Remark 1 Write $\mathbf{f}_t = (x_{t,1}, \ldots, x_{t,d})^\top$. Model (4) is then equivalent to

$$\mathrm{vec}(\mathbf{Y}_t) = (\mathbf{b}_1 \otimes \mathbf{a}_1, \ldots, \mathbf{b}_d \otimes \mathbf{a}_d) \mathbf{f}_t + \mathrm{vec}(\varepsilon_t). \tag{5}$$

This may entice to consider a factor model for the vector process $\mathrm{vec}(\mathbf{Y}_t)$ directly:

$$\mathrm{vec}(\mathbf{Y}_t) = \mathbf{C} \tilde{\mathbf{f}}_t + \tilde{\varepsilon}_t, \tag{6}$$

where $\mathbf{C}$ is a $(pq) \times d$ loading matrix, $\tilde{\mathbf{f}}_t$ is a $d \times 1$ factor, and $\tilde{\varepsilon}_t$ is an error term. In comparison to (6), our model (4) has the following advantages: (a) The number of parameters to be estimated in (4) is $(p + q)d$ which is smaller than pqd, i.e., the number of parameters in (6), and (b) model (4) preserves the original column and row structures of the data while model (6) does not. More

precisely, the row and column variables of $\mathbf{Y}_t$ are typically of different nature. For example, the rows stand for p individuals and the columns stand for q indices. Note that (4) implies $\mathbf{Y}_{\cdot,j,t} = \sum_{\ell=1}^{d} b_{j,\ell}\mathbf{a}_\ell x_{t,\ell} + \varepsilon_{\cdot,j,t}$, i.e., the dynamic part of the jth column $\mathbf{Y}_{\cdot,j,t}$ of $\mathbf{Y}_t$ is a randomly weighted linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_d$. By the symmetry, the dynamic part of any row of $\mathbf{Y}_t$ is a randomly weighted linear combination of $\mathbf{b}_1, \ldots, \mathbf{b}_d$. In contrast model (6) treats the rows and the columns of $\mathbf{Y}_t$ on an equal footing; losing the original meaning and interpretation of the matrix process.

3 Methodology

3.1 Direct estimation for **A**, **B**, and *d*

Without loss of generality, we assume $q \leq p$ in this section, as **A** and **B** are on the equal footing in model (2); see also (3). Then both the identification and the estimation of **A**, **B**, and d essentially reduce to solving a generalized eigenequation defined by two rank-reduced $q \times q$ matrices.

3.1.1 Identification

Let $\mathbf{B}^+ \equiv (\mathbf{b}^1, \ldots, \mathbf{b}^d)^\top$ be the Moore–Penrose inverse of **B**, i.e., $\mathbf{b}_k^\top \mathbf{b}^\ell = I(k = \ell)$ for any $k, \ell \in [d]$. Hence it follows from (4) that

$$\mathbf{Y}_t \mathbf{b}^\ell = x_{t,\ell}\mathbf{a}_\ell + \varepsilon_t \mathbf{b}^\ell, \quad \ell \in [d]. \tag{7}$$

When $\varepsilon_t \equiv 0$, this leads to $\mathbf{Y}_t \mathbf{b}^\ell = \lambda \mathbf{Y}_{t+1}\mathbf{b}^\ell$ with $\lambda = x_{t,\ell}/x_{t+1,\ell}$. Thus, $\mathbf{b}^\ell$ can be obtained from solving this generalized eigenequation. This is essentially the idea of Sanchez and Kowalski (1990). We proceed differently from this point onwards in order (a) to eliminate the impact of nonzero ε_t, (b) to increase the estimation efficiency by augmenting the information over time t, and (c) to improve the estimation performance in solving a generalized eigenequation with rank-reduced matrices.

Let ξ_t be a linear combination of $\mathbf{Y}_t$. For any $k \geq 1$ and $t \geq k+1$, we define $\boldsymbol{\Xi}_{t,k} = \mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\}\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}]$ with $\bar{\mathbf{Y}} = n^{-1}\sum_{t=1}^{n}\mathbf{Y}_t$ and $\bar{\xi} = n^{-1}\sum_{t=1}^{n}\xi_t$. Let

$$\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n}\boldsymbol{\Xi}_{t,k} \tag{8}$$

for any $k \geq 1$. Furthermore, we write $\lambda_{t,k,\ell} = \mathbb{E}[\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}\{x_{t,\ell} - \mathbb{E}(\bar{x}_{\cdot,\ell})\}]$ with $\bar{x}_{\cdot,\ell} = n^{-1}\sum_{t=1}^{n}x_{t,\ell}$ for any $k \geq 1$, $t \geq k+1$, and $\ell \in [d]$. By (7) and Condition 1, it holds that $\boldsymbol{\Xi}_{t,k}\mathbf{b}^\ell = \lambda_{t,k,\ell}\mathbf{a}_\ell$, which implies

$$\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\mathbf{b}^\ell = \left(\frac{1}{n-k}\sum_{t=k+1}^{n}\lambda_{t,k,\ell}\right)\mathbf{a}_\ell, \quad \ell \in [d]. \tag{9}$$

Then, we have

$$\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(2)\mathbf{b}^\ell = \tilde{\lambda}_\ell \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(1)\mathbf{b}^\ell \text{ with } \tilde{\lambda}_\ell = \frac{(n-1)\sum_{t=3}^{n}\lambda_{t,2,\ell}}{(n-2)\sum_{t=2}^{n}\lambda_{t,1,\ell}}. \tag{10}$$

Write

$$\mathbf{K}_{1,q} = \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(1)^\top \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(1) \quad \text{and} \quad \mathbf{K}_{2,q} = \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(1)^\top \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(2).$$

Hence the rows of $\mathbf{B}^+ = (\mathbf{b}^1, \ldots, \mathbf{b}^d)^\top$ are the eigenvectors of the generalized eigenequation

$$\mathbf{K}_{2,q}\mathbf{b} = \lambda\mathbf{K}_{1,q}\mathbf{b}. \tag{11}$$

This is a generalized eigenequation defined by rank-reduced matrices $\mathbf{K}_{1,q}$ and $\mathbf{K}_{2,q}$. In general, the number of eigenvalues of a generalized eigenequation defined by rank-reduced matrices may be 0, finite or infinite; see Section 7.7 of Golub and Van Load (2013). However, since $\mathbf{K}_{1,q}$ is positive definite with rank d, (11) admits exactly d eigenvalues. To verify this statement, recall $\min(p, q) > d$ and $\mathbf{\Sigma}_{\mathbf{Y},\xi}(1) = \mathbf{A}\mathbf{\Lambda}\mathbf{B}^\top$ for some $d \times d$ diagonal matrix $\mathbf{\Lambda}$. If the elements in the main diagonal of $\mathbf{\Lambda}$ are nonzero, together with Condition 1, we know $\mathrm{rank}(\mathbf{K}_{1,q}) = d$. Hence $\mathbf{K}_{1,q} = \mathbf{\Gamma}\mathbf{C}\mathbf{\Gamma}^\top$, where $\mathbf{\Gamma}$ is a $q \times q$ orthogonal matrix, and $\mathbf{C} = \mathrm{diag}(c_1, \ldots, c_d, 0, \ldots, 0)$ with $c_1 \geq \cdots \geq c_d > 0$. Then the characteristic equation of the generalized eigenequation (11) is

$$0 = \det(\mathbf{K}_{2,q} - \lambda\mathbf{K}_{1,q}) = \det^2(\mathbf{\Gamma})\det(\mathbf{\Gamma}^\top\mathbf{K}_{2,q}\mathbf{\Gamma} - \lambda\mathbf{C}).$$

The RHS of the above equation is a polynomial in λ of order d, which, therefore, has d roots.

Let $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_d$ specified in (10) be distinct. Then the rows of $\mathbf{B}^+$ can be identified by (11) upto the scaling and permutation indeterminacy. However, to specify $\mathbf{B}^+$ completely, both the length and direction of each row need to be determined precisely, which is beyond what can be learned from (11). Nevertheless the eigenvectors of (11) can identify the columns of $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ based on the following identity:

$$\mathbf{a}_\ell = \frac{\mathbf{\Sigma}_{\mathbf{Y},\xi}(1)\mathbf{b}^\ell}{|\mathbf{\Sigma}_{\mathbf{Y},\xi}(1)\mathbf{b}^\ell|_2}, \quad \ell \in [d], \tag{12}$$

which is implied by (9). For $\mathbf{A}$ specified above, let $\mathbf{A}^+ = (\mathbf{a}^1, \ldots, \mathbf{a}^d)^\top$ be its Moore–Penrose inverse. By the symmetry, the columns of $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ are uniquely identified by

$$\mathbf{b}_\ell = \frac{\mathbf{\Sigma}_{\mathbf{Y},\xi}(1)^\top\mathbf{a}^\ell}{|\mathbf{\Sigma}_{\mathbf{Y},\xi}(1)^\top\mathbf{a}^\ell|_2}, \quad \ell \in [d]. \tag{13}$$

3.1.2 Estimation

With the available observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, we define

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n}(\mathbf{Y}_t - \bar{\mathbf{Y}})(\xi_{t-k} - \bar{\xi}) \tag{14}$$

for $k \geq 1$. When $pq \gg n$, $\widehat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)$ is no longer a consistent estimator for $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ under the spectral norm $\|\cdot\|_2$. In the spirit of Bickel and Levina (2008), we select $\widehat{\mathbf{\Sigma}}_k$ defined as follows for the estimate of $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$:

$$\widehat{\mathbf{\Sigma}}_k = T_{\delta_1}\{\widehat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)\}, \tag{15}$$

where $T_{\delta_1}(\cdot)$ is a threshold operator $T_{\delta_1}(\mathbf{W}) = \{w_{i,j}I(|w_{i,j}| \geq \delta_1)\}_{m_1 \times m_2}$ for any matrix $\mathbf{W} = (w_{i,j})_{m_1 \times m_2}$ with the threshold level $\delta_1 \geq 0$. We choose $\delta_1 > 0$ when $pq \gg n$. When $\delta_1 = 0$, we have $\widehat{\mathbf{\Sigma}}_k = \widehat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)$, which is appropriate when, for example, p and q are fixed constants. Then $\widehat{\mathbf{K}}_{1,q} = \widehat{\mathbf{\Sigma}}_1^\top\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{K}}_{2,q} = \widehat{\mathbf{\Sigma}}_1^\top\widehat{\mathbf{\Sigma}}_2$ provide the estimates of $\mathbf{K}_{1,q}$ and $\mathbf{K}_{2,q}$, respectively.

Let $\hat{\lambda}_1(\widehat{\mathbf{K}}_{1,q}) \geq \cdots \geq \hat{\lambda}_q(\widehat{\mathbf{K}}_{1,q}) \geq 0$ be the eigenvalues of $\widehat{\mathbf{K}}_{1,q}$. Since $\mathrm{rank}(\mathbf{K}_{1,q}) = d$, following Chang et al. (2015), we can estimate d as

$$\hat{d} = \arg\min_{j\in[R]} \frac{\hat{\lambda}_{j+1}(\widehat{\mathbf{K}}_{1,q}) + c_n}{\hat{\lambda}_j(\widehat{\mathbf{K}}_{1,q}) + c_n}, \tag{16}$$

where $R = \lfloor \alpha\min(p, q)\rfloor$ for a prescribed constant $\alpha \in (0, 1)$, and $c_n \to 0^+$ as $n \to \infty$. In practice, we may set $\alpha = 0.5$. Note that the true eigenvalues of $\mathbf{K}_{1,q}$ satisfy the condition $\lambda_1(\mathbf{K}_{1,q}) \geq \cdots \geq \lambda_d(\mathbf{K}_{1,q}) > 0 = \lambda_{d+1}(\mathbf{K}_{1,q}) = \cdots = \lambda_q(\mathbf{K}_{1,q})$. Adding a small constant $c_n > 0$ in (16) is to avoid the ratio '0/0'. Under some regularity conditions, $\hat{d}$ defined in (16) is a consistent estimate for d in the sense that $\mathbb{P}(\hat{d} = d) \to 1$ as $n \to \infty$.

Applying the spectral decomposition to $\widehat{\mathbf{K}}_{1,q}$, we have $\widehat{\mathbf{K}}_{1,q} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{C}}\widehat{\mathbf{\Gamma}}^{\top}$, where $\widehat{\mathbf{\Gamma}} = (\hat{\mathbf{\gamma}}_1, \ldots, \hat{\mathbf{\gamma}}_q)$ is a $q \times q$ orthogonal matrix, and $\widehat{\mathbf{C}} = \mathrm{diag}(\hat{c}_1, \ldots, \hat{c}_q)$ with $\hat{c}_1 \geq \cdots \geq \hat{c}_q \geq 0$. For $\hat{d}$ specified in (16), we define

$$\widetilde{\mathbf{K}}_{1,q} = \sum_{j=1}^{\hat{d}} \hat{c}_j \hat{\mathbf{\gamma}}_j \hat{\mathbf{\gamma}}_j^{\top}, \tag{17}$$

which is a truncated version of $\widehat{\mathbf{K}}_{1,q}$. Then $\mathrm{rank}(\widetilde{\mathbf{K}}_{1,q}) = \hat{d}$. Let $\hat{\mathbf{b}}^1, \ldots, \hat{\mathbf{b}}^{\hat{d}}$ be the eigenvectors of the generalized eigenequation

$$\widehat{\mathbf{K}}_{2,q}\mathbf{b} = \lambda\widetilde{\mathbf{K}}_{1,q}\mathbf{b}, \tag{18}$$

which is a sample version of (11). We can use the function `geigen` in the R-package `geigen` to solve (18). Then the columns of $\mathbf{A}$ can be estimated as

$$\hat{\mathbf{a}}_{\ell} = \frac{\widehat{\mathbf{\Sigma}}_1\hat{\mathbf{b}}^{\ell}}{|\widehat{\mathbf{\Sigma}}_1\hat{\mathbf{b}}^{\ell}|_2}, \quad \ell \in [\hat{d}]. \tag{19}$$

Let $\widehat{\mathbf{A}}^+ = (\hat{\mathbf{a}}^1, \ldots, \hat{\mathbf{a}}^{\hat{d}})^{\top}$ be the Moore–Penrose inverse of $\widehat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_{\hat{d}})$. Then the columns of $\mathbf{B}$ can be estimated as

$$\hat{\mathbf{b}}_{\ell} = \frac{\widehat{\mathbf{\Sigma}}_1^{\top}\hat{\mathbf{a}}^{\ell}}{|\widehat{\mathbf{\Sigma}}_1^{\top}\hat{\mathbf{a}}^{\ell}|_2}, \quad \ell \in [\hat{d}]. \tag{20}$$

The truncation of $\widehat{\mathbf{K}}_{1,q}$ given in (17) is necessary here for estimating the rows of $\mathbf{B}^+$. Note that $\widehat{\mathbf{K}}_{1,q}$ is a $q \times q$ matrix with $q > d$. Since $\mathrm{rank}(\widehat{\mathbf{K}}_{1,q})$ may be larger than $\hat{d}$ in finite samples, the generalized eigenequation $\widehat{\mathbf{K}}_{2,q}\mathbf{b} = \lambda\widehat{\mathbf{K}}_{1,q}\mathbf{b}$ may have more than $\hat{d}$ eigenvectors. Since we do not know which eigenvalues are associated with our required eigenvectors, it will be extremely difficult (if not impossible) for us to pick out $\hat{\mathbf{b}}^1, \ldots, \hat{\mathbf{b}}^{\hat{d}}$ from all the eigenvectors of $\widehat{\mathbf{K}}_{2,q}\mathbf{b} = \lambda\widehat{\mathbf{K}}_{1,q}\mathbf{b}$.

Based on $\hat{\mathbf{a}}_{\ell}$ and $\hat{\mathbf{b}}_{\ell}$ specified in (19) and (20), we define

$$\widehat{\mathbf{H}} = (\hat{\mathbf{b}}_1 \otimes \hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{b}}_{\hat{d}} \otimes \hat{\mathbf{a}}_{\hat{d}}).$$

By (5), we can recover $\mathbf{X}_t$ by $\widehat{\mathbf{X}}_t = \mathrm{diag}(\hat{x}_{t,1}, \ldots, \hat{x}_{t,\hat{d}})$ with

$$(\hat{x}_{t,1}, \ldots, \hat{x}_{t,\hat{d}})^{\top} = \widehat{\mathbf{H}}^+\mathrm{vec}(\mathbf{Y}_t).$$

We need to point out that the eigenvalues of the generalized eigenequation (18) are not necessary to be real. Proposition 1 shows that its complex eigenvalues always occur in complex conjugate pairs.

Proposition 1 Assume the eigenvalues of the generalized eigenequation (18) are distinct. If $\lambda_* \in \mathbb{C}$ is a complex eigenvalue of (18) such that $\widehat{\mathbf{K}}_{2,q}\hat{\mathbf{b}}^\ell = \lambda_* \widetilde{\mathbf{K}}_{1,q}\hat{\mathbf{b}}^\ell$ for some $\ell \in [\hat{d}]$, then $\overline{\lambda_*}$, the complex conjugate of λ_*, is also a complex eigenvalue of (18). More specifically, there exists some $\tilde{\ell} \in [\hat{d}]$ and a constant $\kappa \in \{-1, 1\}$ satisfying $\widehat{\mathbf{K}}_{2,q}\hat{\mathbf{b}}^{\tilde{\ell}} = \overline{\lambda_*}\widetilde{\mathbf{K}}_{1,q}\hat{\mathbf{b}}^{\tilde{\ell}}$, $\hat{\mathbf{a}}_{\tilde{\ell}} = \kappa \overline{\hat{\mathbf{a}}_\ell}$, $\hat{\mathbf{b}}_{\tilde{\ell}} = \kappa \overline{\hat{\mathbf{b}}_\ell}$, and $\hat{x}_{t,\tilde{\ell}} = \overline{\hat{x}_{t,\ell}}$, where $\overline{\hat{\mathbf{a}}_\ell}$, $\overline{\hat{\mathbf{b}}_\ell}$, and $\overline{\hat{x}_{t,\ell}}$ are the complex conjugate of $\hat{\mathbf{a}}_\ell$, $\hat{\mathbf{b}}_\ell$, and $\hat{x}_{t,\ell}$, respectively.

Assume the generalized eigenequation (18) has s real eigenvalues and $\hat{d} - s$ complex eigenvalues. Since the complex eigenvalues always occur in complex conjugate pairs, $\hat{d} - s$ is an even integer. Write $\hat{d} - s = 2m$. Let $\lambda_1, \overline{\lambda_1}, \ldots, \lambda_m, \overline{\lambda_m}$ be the $\hat{d} - s$ complex eigenvalues, where $\overline{\lambda_1}, \ldots, \overline{\lambda_m}$ are the complex conjugate of $\lambda_1, \ldots, \lambda_m$, respectively. For each $j \in [m]$, there exist $\ell_j, \tilde{\ell}_j \in [\hat{d}]$ such that the eigenvectors associated with λ_j and $\overline{\lambda_j}$ are, respectively, $\hat{\mathbf{b}}^{\ell_j}$ and $\hat{\mathbf{b}}^{\tilde{\ell}_j}$. By Proposition 1, there exists $(\kappa_1, \ldots, \kappa_m) \in \{-1, 1\}^m$ such that $\hat{\mathbf{a}}_{\tilde{\ell}_j} = \kappa_j \overline{\hat{\mathbf{a}}_{\ell_j}}$, $\hat{\mathbf{b}}_{\tilde{\ell}_j} = \kappa_j \overline{\hat{\mathbf{b}}_{\ell_j}}$, and $\hat{x}_{t,\tilde{\ell}_j} = \overline{\hat{x}_{t,\ell_j}}$ for each $j \in [m]$. Then

$$\sum_{j=1}^m (\hat{x}_{t,\ell_j}\hat{\mathbf{a}}_{\ell_j}\hat{\mathbf{b}}_{\ell_j}^\top + \hat{x}_{t,\tilde{\ell}_j}\hat{\mathbf{a}}_{\tilde{\ell}_j}\hat{\mathbf{b}}_{\tilde{\ell}_j}^\top) \in \mathbb{R}^{p \times q}.$$

Write $\{k_1, \ldots, k_s\} = [\hat{d}] \setminus \{\ell_1, \tilde{\ell}_1, \ldots, \ell_m, \tilde{\ell}_m\}$. Then $\hat{\mathbf{b}}^{k_1}, \ldots, \hat{\mathbf{b}}^{k_s}$ are the eigenvectors of the generalized eigenequation (18) associated with the s real eigenvalues. Hence, $\hat{\mathbf{a}}_{k_j} \in \mathbb{R}^p$, $\hat{\mathbf{b}}_{k_j} \in \mathbb{R}^q$, and $\hat{x}_{t,k_j} \in \mathbb{R}$ for each $j \in [s]$. To do prediction of $\mathbf{Y}_t$ based on (4), we only need to model $\hat{d}$ univariate time series $\{\hat{x}_{t,k_1}\}, \ldots, \{\hat{x}_{t,k_s}\}, \{\mathrm{Re}(\hat{x}_{t,\ell_1})\}, \{\mathrm{Im}(\hat{x}_{t,\ell_1})\}, \ldots, \{\mathrm{Re}(\hat{x}_{t,\ell_m})\}, \{\mathrm{Im}(\hat{x}_{t,\ell_m})\}$.

Remark 2 Solving a generalized eigenequation defined by rank-reduced matrices could be a complex computational task. See Section 7.7 of Golub and Van Load (2013). In principle, we can also estimate $\mathbf{A}^+$ first; leading to the estimate for $\mathbf{B}$ and then that for $\mathbf{A}$. Technically this boils down to solving a generalized eigenequation defined by two $p \times p$ rank-reduced matrices, which is computationally more expensive and less stable by using the R-function geigen when $p > q$; often leading to, for example, more than $\hat{d}$ eigenvalues/vectors.

3.2 A refined estimation procedure

To overcome the complication in solving a rank-reduced generalized eigenequation, which plays the key role in the method proposed in Section 3.1, we propose a refinement which reduces the q-dimensional rank-reduced generalized eigenequation to a d-dimensional full-ranked one. Therefore, effectively the new refined method only requires to solve a d-dimensional eigenequation.

Simulation results in Section 5 indicate that this new procedure outperforms the direct estimation, proposed in Section 3.1.2, uniformly over various settings.

3.2.1 Identification

For a prescribed integer $K \geq 1$, define

$$\mathbf{M}_1 = \sum_{k=1}^K \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^\top \quad \text{and} \quad \mathbf{M}_2 = \sum_{k=1}^K \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^\top\mathbf{\Sigma}_{\mathbf{Y},\xi}(k) \tag{21}$$

with $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ defined as (8). Recall $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k) = \mathbf{A}\mathbf{G}_k\mathbf{B}^\top$, where $\mathbf{G}_k = (n-k)^{-1}\sum_{t=k+1}^n \mathbb{E}[\{\mathbf{X}_t - \mathbb{E}(\bar{\mathbf{X}})\}\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}]$ is a $d \times d$ diagonal matrix with $\bar{\mathbf{X}} = n^{-1}\sum_{t=1}^n \mathbf{X}_t$ and $\bar{\xi} = n^{-1}\sum_{t=1}^n \xi_t$; see (4). Then

$$\mathbf{M}_1 = \mathbf{A}\left(\sum_{k=1}^{K} \mathbf{G}_k \mathbf{B}^\top \mathbf{B} \mathbf{G}_k\right)\mathbf{A}^\top \quad \text{and} \quad \mathbf{M}_2 = \mathbf{B}\left(\sum_{k=1}^{K} \mathbf{G}_k \mathbf{A}^\top \mathbf{A} \mathbf{G}_k\right)\mathbf{B}^\top. \tag{22}$$

Since both p and q are much greater than d in practice, it is reasonable to impose the following assumption.

> **Condition 2**   It holds that $\mathrm{rank}(\mathbf{M}_1) = d = \mathrm{rank}(\mathbf{M}_2)$. Furthermore, the nonzero eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$ are uniformly bounded away from zero.

> **Remark 3**   Write $\mathbf{G} = (\mathbf{G}_1^\top, \ldots, \mathbf{G}_K^\top)^\top$. Then $\mathbf{M}_1 = \mathbf{A}\mathbf{G}^\top(\mathbf{I}_K \otimes \mathbf{B})^\top(\mathbf{I}_K \otimes \mathbf{B})\mathbf{G}\mathbf{A}^\top$, which implies $\mathrm{rank}(\mathbf{M}_1) = \mathrm{rank}\{(\mathbf{I}_K \otimes \mathbf{B})\mathbf{G}\mathbf{A}^\top\}$. Notice that each $\mathbf{G}_k$ is a $d \times d$ diagonal matrix and $(\mathbf{I}_K \otimes \mathbf{B})\mathbf{G}\mathbf{A}^\top = (\mathbf{A}\mathbf{G}_1\mathbf{B}^\top, \ldots, \mathbf{A}\mathbf{G}_K\mathbf{B}^\top)^\top$. It holds that $d \geq \mathrm{rank}(\mathbf{G}) \geq \mathrm{rank}\{(\mathbf{I}_K \otimes \mathbf{B})\mathbf{G}\mathbf{A}^\top\} \geq \max_{k \in [K]} \mathrm{rank}(\mathbf{A}\mathbf{G}_k\mathbf{B}^\top)$. Since $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \mathbf{A}\mathbf{G}_k\mathbf{B}^\top$, $\mathrm{rank}(\mathbf{M}_1) = d$ provided that there exists some $k \in [K]$ such that $\mathrm{rank}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} = d$. By the same argument, $\mathrm{rank}(\mathbf{M}_2) = d$ provided that there exists some $k \in [K]$ such that $\mathrm{rank}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} = d$. Since $\mathrm{rank}(\mathbf{A}) = d = \mathrm{rank}(\mathbf{B})$ (see Condition 1), $\mathrm{rank}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} = \mathrm{rank}(\mathbf{G}_k)$. Consequently, $\mathrm{rank}(\mathbf{M}_1) = d = \mathrm{rank}(\mathbf{M}_2)$ if the elements in the main diagonal of some $\mathbf{G}_k$ are nonzero.

Perform the spectral decomposition:

$$\mathbf{M}_1 = \mathbf{P}\boldsymbol{\Lambda}_1\mathbf{P}^\top \quad \text{and} \quad \mathbf{M}_2 = \mathbf{Q}\boldsymbol{\Lambda}_2\mathbf{Q}^\top,$$

where the columns of $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, the d orthonormal eigenvectors corresponding to the d nonzero eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$, $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are the diagonal matrices with the corresponding eigenvalues as the diagonal elements. This, together with (22), implies that

$$\mathbf{A} = \mathbf{P}\mathbf{U} \quad \text{and} \quad \mathbf{B} = \mathbf{Q}\mathbf{V}, \tag{23}$$

where $\mathbf{U}$ and $\mathbf{V}$ are two $d \times d$ invertible matrices. Furthermore all the columns of $\mathbf{U}$ and $\mathbf{V}$ are unit vectors, which is implied by the assumption that all $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are unit vectors.

To identify $\mathbf{A}$ and $\mathbf{B}$, we only need to identify $\mathbf{U}$ and $\mathbf{V}$, which can be solved from a generalized eigenequation with two full-ranked matrices. To this end, define $d \times d$ matrix process $\mathbf{Z}_t = \mathbf{P}^\top\mathbf{Y}_t\mathbf{Q}$. It follows from (4) and (23) that

$$\mathbf{Z}_t = \mathbf{U}\mathbf{X}_t\mathbf{V}^\top + \boldsymbol{\Delta}_t = \sum_{\ell=1}^{d} x_{t,\ell}\mathbf{u}_\ell\mathbf{v}_\ell^\top + \boldsymbol{\Delta}_t,$$

where $\boldsymbol{\Delta}_t = \mathbf{P}^\top\boldsymbol{\varepsilon}_t\mathbf{Q}$ is uncorrelated with $\{\mathbf{X}_s\}_{s \leq t}$, $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ are, respectively, the ℓth column of $\mathbf{U}$ and $\mathbf{V}$. Choose η_t to be a linear combination of $\mathbf{Z}_t$ such that

$$\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{Z}_t - \mathbb{E}(\bar{\mathbf{Z}})\}\{\eta_{t-k} - \mathbb{E}(\bar{\eta})\}] \tag{24}$$

is full-ranked for $k = 1, 2$, where $\bar{\mathbf{Z}} = n^{-1}\sum_{t=1}^{n}\mathbf{Z}_t$, and $\bar{\eta} = n^{-1}\sum_{t=1}^{n}\eta_t$. Then the same argument towards (11) implies that the rows of the $d \times d$ inverse matrix $\mathbf{V}^{-1} = (\mathbf{v}^1, \ldots, \mathbf{v}^d)^\top$ are the eigenvectors of the generalized eigenequation

$$\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2)\mathbf{v} = \lambda\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v}, \tag{25}$$

which has exactly d eigenvectors. Furthermore those d eigenvectors are unique upto the scaling

indeterminacy if the d eigenvalues associated with (25) are distinct. Parallel to (12) and (13), the columns of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ can be identified as follows:

$$\mathbf{u}_\ell = \frac{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v}^\ell}{|\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v}^\ell|_2} \quad \text{and} \quad \mathbf{v}_\ell = \frac{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\mathbf{u}^\ell}{|\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\mathbf{u}^\ell|_2} \tag{26}$$

for each $\ell \in [d]$, where $(\mathbf{u}^1, \ldots, \mathbf{u}^d)^\top$ is the inverse of $\mathbf{U}$. With $\mathbf{U}$ and $\mathbf{V}$ specified above, $\mathbf{A}$ and $\mathbf{B}$ can be determined by (23). Write

$$\mathbf{J}_1 = \{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2). \tag{27}$$

Proposition 2   Let Conditions 1 and 2 hold, and the eigenvalues of the $d \times d$ matrix $\mathbf{J}_1$ specified in (27) be distinct. Then $\mathbf{A}$ and $\mathbf{B}$ are uniquely defined as in (23) upto the reflection and permutation indeterminacy, where the columns of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ are defined as (26).

Remark 4   By the symmetry, we also know that $\mathbf{u}^1, \ldots, \mathbf{u}^d$ are the d eigenvectors of the generalized eigenequation $\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2)^\top\mathbf{u} = \lambda\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\mathbf{u}$. Write

$$\mathbf{J}_2 = \{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^\top\}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2)^\top.$$

It holds that $\mathbf{v}^\ell$ and $\mathbf{u}^\ell$ are, respectively, the eigenvectors of the $d \times d$ matrices $\mathbf{J}_1$ and $\mathbf{J}_2$ associated with the same eigenvalue.

3.2.2 Estimation

Let $\eta_t = \mathbf{w}^\top\text{vec}(\mathbf{Z}_t)$ be a linear combination of $\mathbf{Z}_t$ for some constant vector $\mathbf{w} \in \mathbb{R}^{d^2}$. Any $\mathbf{w} \in \mathbb{R}^{d^2}$ such that the associated $d \times d$ matrix $\mathbf{J}_1$ specified in (27) has d distinct eigenvalues is valid for the identification of $\mathbf{U}$ and $\mathbf{V}$. See Proposition 2 for details. Write $\boldsymbol{\Theta} = \mathbf{I}_p \otimes \{(\mathbf{Q} \otimes \mathbf{P})\mathbf{w}\}$ and $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k) = (n - k)^{-1}\sum_{t=k+1}^{n}\mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\} \otimes \text{vec}\{\mathbf{Y}_{t-k} - \mathbb{E}(\bar{\mathbf{Y}})\}]$. Then $\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(k)$ defined as (24) can be reformulated as

$$\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(k) = \mathbf{P}^\top\boldsymbol{\Theta}^\top\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\mathbf{Q}. \tag{28}$$

For $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)$ defined as (14), we define the threshold estimators for $\mathbf{M}_1$ and $\mathbf{M}_2$ given in (21) as follows:

$$\widehat{\mathbf{M}}_1 = \sum_{k=1}^{K}\widehat{\boldsymbol{\Sigma}}_k\widehat{\boldsymbol{\Sigma}}_k^\top \quad \text{and} \quad \widehat{\mathbf{M}}_2 = \sum_{k=1}^{K}\widehat{\boldsymbol{\Sigma}}_k^\top\widehat{\boldsymbol{\Sigma}}_k, \tag{29}$$

where $\widehat{\boldsymbol{\Sigma}}_k$ is defined as (15). Let $\hat{\lambda}_1(\widehat{\mathbf{M}}_1) \geq \cdots \geq \hat{\lambda}_p(\widehat{\mathbf{M}}_1) \geq 0$ be the eigenvalues of the $p \times p$ matrix $\widehat{\mathbf{M}}_1$. Recall $\text{rank}(\mathbf{M}_1) = d$. Analogous to (16), we can also estimate d as

$$\hat{d} = \arg\min_{j \in [R]} \frac{\hat{\lambda}_{j+1}(\widehat{\mathbf{M}}_1) + c_n}{\hat{\lambda}_j(\widehat{\mathbf{M}}_1) + c_n}, \tag{30}$$

where R and c_n are same as those in (16). The convergence rate of c_n will be specified in Theorem 1 and Remark 7 in Section 4. Theorem 1 shows that $\hat{d}$ is consistent, i.e., $\mathbb{P}(\hat{d} \neq d) \to 0$ as $n \to \infty$.

Remark 5   Analogously, we can also estimate d by replacing $\widehat{\mathbf{M}}_1$ in (30) by $\widehat{\mathbf{M}}_2$. Recall $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ are, respectively, $p \times p$ and $q \times q$ matrices. For $K = 1$, since the nonzero eigenvalues of $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ are identical, such replacement will lead to a same estimate for d as that by (30). For $K > 1$, although the estimates based on $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ are both consistent, their finite-sample performance is a little bit different. More specifically, simulation results show that (a) the estimate based on $\widehat{\mathbf{M}}_1$ has higher probability of correctly estimating d when $p > q$, (b) the estimate based on $\widehat{\mathbf{M}}_2$ has higher probability of correctly estimating d when $q > p$, and (c) the estimates based on $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ are almost identical when $p = q$. See the online supplementary material, Tables S1–S3 for details. We suggest to estimate d based on $\widehat{\mathbf{M}}_1$ when $p \geq q$, and based on $\widehat{\mathbf{M}}_2$ when $p < q$.

Now let $\widehat{\mathbf{P}}$ be the $p \times \hat{d}$ matrix of which the columns are the $\hat{d}$ orthonormal eigenvectors of $\widehat{\mathbf{M}}_1$ corresponding to its $\hat{d}$ largest eigenvalues, and $\widehat{\mathbf{Q}}$ be the $q \times \hat{d}$ matrix of which the columns are the $\hat{d}$ orthonormal eigenvectors of $\widehat{\mathbf{M}}_2$ corresponding to its $\hat{d}$ largest eigenvalues. Define

$$\widehat{\mathbf{Z}}_t = \widehat{\mathbf{P}}^\top \mathbf{Y}_t \widehat{\mathbf{Q}} \quad \text{and} \quad \hat{\eta}_t = \mathbf{w}^\top \mathrm{vec}(\widehat{\mathbf{Z}}_t) \tag{31}$$

for some constant vector $\mathbf{w} \in \mathbb{R}^{\hat{d}^2}$ with bounded ℓ^2-norm. Based on (28), we put

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(k) = \widehat{\mathbf{P}}^\top \widehat{\boldsymbol{\Theta}}^\top T_{\delta_2}\{\widehat{\boldsymbol{\Sigma}}_{\check{\mathbf{Y}}}(k)\}\widehat{\mathbf{Q}}, \tag{32}$$

where $T_{\delta_2}(\cdot)$ is a threshold operator with the threshold level $\delta_2 \geq 0$, $\widehat{\boldsymbol{\Theta}} = \mathbf{I}_p \otimes \{(\widehat{\mathbf{Q}} \otimes \widehat{\mathbf{P}})\mathbf{w}\}$, and

$$\widehat{\boldsymbol{\Sigma}}_{\check{\mathbf{Y}}}(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} (\mathbf{Y}_t - \bar{\mathbf{Y}}) \otimes \mathrm{vec}(\mathbf{Y}_{t-k} - \bar{\mathbf{Y}}). \tag{33}$$

Write $\widehat{\mathbf{J}}_1 = \{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(2)$ and let $\hat{\mathbf{v}}^1, \ldots, \hat{\mathbf{v}}^{\hat{d}}$ be the $\hat{d}$ eigenvectors of the $\hat{d} \times \hat{d}$ matrix $\widehat{\mathbf{J}}_1$. Now the estimators for $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\widehat{\mathbf{A}} = \widehat{\mathbf{P}}\widehat{\mathbf{U}} \quad \text{and} \quad \widehat{\mathbf{B}} = \widehat{\mathbf{Q}}\widehat{\mathbf{V}}, \tag{34}$$

where $\widehat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{\hat{d}})$ and $\widehat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_{\hat{d}})$ with

$$\hat{\mathbf{u}}_\ell = \frac{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\hat{\mathbf{v}}^\ell}{|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\hat{\mathbf{v}}^\ell|_2} \quad \text{and} \quad \hat{\mathbf{v}}_\ell = \frac{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^\top \hat{\mathbf{u}}^\ell}{|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^\top \hat{\mathbf{u}}^\ell|_2}. \tag{35}$$

In the above expression, $(\hat{\mathbf{u}}^1, \ldots, \hat{\mathbf{u}}^{\hat{d}})^\top$ is the inverse of $\widehat{\mathbf{U}}$.

Remark 6   Our above-presented estimation procedure essentially estimates $\mathbf{V}^{-1}$, $\mathbf{U}$, $\mathbf{U}^{-1}$, and $\mathbf{V}$ sequentially. Parallel to Remark 2 in Section 3.1, we can also consider estimating $\mathbf{U}^{-1}$ first. Remark 4 indicates that the d rows of $\mathbf{U}^{-1}$ are the d eigenvectors of $\mathbf{J}_2$. Since $\widehat{\mathbf{J}}_1$ and $\widehat{\mathbf{J}}_2 = \{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^\top\}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(2)^\top$ are full-ranked, the difference between these two solutions are negligible, which is confirmed by the simulation not reported here.

4 Asymptotic properties

As we do not impose the stationarity on $\{\mathbf{Y}_t\}$, we use the concept of 'α-mixing' to characterize the serial dependence of $\{\mathbf{Y}_t\}$ with the α-mixing coefficients defined as

$$\alpha(k) = \sup_r \sup_{A \in \mathcal{F}_{-\infty}^r, B \in \mathcal{F}_{r+k}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \quad k \geq 1, \tag{36}$$

where $\mathcal{F}_r^s$ is the σ-field generated by $\{\mathbf{Y}_t : r \leq t \leq s\}$. To simplify our presentation, we first present the theoretical results for the most challenging scenario with $p, q \gg n$ in Theorems 1 and 2, and then give the associated results in Remark 7 for the cases with fixed (p, q) or (p, q) diverging at some polynomial rate of n. We need the following regularity conditions.

Condition 3   (i) There exists a universal constant $C_1 > 0$ such that $\max_{k \in [K]} \|\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 \leq C_1$. (ii) Write $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k) = \{\sigma_{y,\xi,i,j}^{(k)}\}_{p \times q}$. It holds that $\max_{i \in [p]} \sum_{j=1}^q |\sigma_{y,\xi,i,j}^{(k)}|^\iota \leq s_1$ and $\max_{j \in [q]} \sum_{i=1}^p |\sigma_{y,\xi,i,j}^{(k)}|^\iota \leq s_2$ for some universal constant $\iota \in [0, 1)$, where s_1 and s_2 may, respectively, diverge together with p and q.

Condition 4   (i) There exist some universal constants $C_2 > 0$, $C_3 > 0$, and $r_1 \in (0, 2]$ such that $\max_{i \in [p]} \max_{j \in [q]} \max_{t \in [n]} \mathbb{P}(|y_{i,j,t}| > x) \leq C_2 \exp(-C_3 x^{r_1})$ and $\max_{t \in [n]} \mathbb{P}(|\xi_t| > x) \leq C_2 \exp(-C_3 x^{r_1})$ for any $x > 0$. (ii) There exist some universal constants $C_4 > 0$, $C_5 > 0$, and $r_2 \in (0, 1]$ such that the mixing coefficients $\alpha(k)$ given in (36) satisfy $\alpha(k) \leq C_4 \exp(-C_5 k^{r_2})$ for all $k \geq 1$.

Recall $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ is a $p \times q$ matrix. Condition 3(i) requires the singular values of $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ to be uniformly bounded away from infinity for any $k \in [K]$. Our technical proofs indeed allow $\max_{k \in [K]} \|\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\|_2$ to diverge with n. We impose Condition 3(i) just for simplifying the presentation. Condition 3(ii) imposes some sparsity on $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$. Notice that $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k) = \mathbf{A}\mathbf{G}_k\mathbf{B}^\top$ for some $d \times d$ diagonal matrix $\mathbf{G}_k$. Under some sparsity condition on $\mathbf{A}$ and $\mathbf{B}$, applying the technique used to derive Lemma 5 of Chang et al. (2018), we can show that Condition 3(ii) holds for certain (s_1, s_2). Condition 4 is a common assumption in the literature on ultrahigh-dimensional data analysis, which ensures exponential-type upper bounds for the tail probabilities of the statistics concerned when $p, q \gg n$. See Chang et al. (2021) and reference therein. The α-mixing assumption in Condition 4(ii) is mild. See the discussion below Equation (3) and Assumption 1 in Chang et al. (in press) for the widely used time series models which satisfy Condition 4(ii). If we only require $\max_{i \in [p]} \max_{j \in [q]} \max_{t \in [n]} \mathbb{P}(|y_{i,j,t}| > x) = O\{x^{-2(l+\tau)}\}$ for any $x > 0$, $\max_{t \in [n]} \mathbb{P}(|\xi_t| > x) = O\{x^{-2(l+\tau)}\}$ for any $x > 0$ and $\alpha(k) = O\{k^{-(l-1)(l+\tau)/\tau}\}$ as $k \to \infty$ with two constants $l > 2$ and $\tau > 0$, we can apply Fuk–Nagaev-type inequalities to construct the upper bounds for the tail probabilities of the statistics concerned for which our procedure still works when p and q diverge at some polynomial rate of n. See Remark 7(ii). Let

$$\Pi_{1,n} = (s_1 s_2)^{1/2} \{n^{-1}\log(pq)\}^{(1-\iota)/2}.$$

Theorem 1 shows that the ratio-based estimator $\hat{d}$ defined in (30) is consistent.

Theorem 1   Let Conditions 1–4 hold and the threshold level $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constant $C_* > 0$. For any c_n in (30) satisfying $\Pi_{1,n} \ll c_n \ll 1$, it holds that $\mathbb{P}(\hat{d} = d) \to 1$ as $n \to \infty$, provided that $\Pi_{1,n} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c \in (0, 1)$ depending only on r_1 and r_2 specified in Condition 4.

To investigate the asymptotic properties of the estimator $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ given in (34), we first assume $\hat{d} = d$. Due to the consistency of $\hat{d}$ presented in Theorem 1, we can prove, using the same arguments below Theorem 2.4 of Chang et al. (2015), that the same results still hold without the assumption $\hat{d} = d$. See our discussion below Theorem 2.

Proposition 3   Let Conditions 1–4 hold and the threshold level $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constant $C_* > 0$. If $\hat{d} = d$, there exist some orthogonal matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ such that $\|\widehat{\mathbf{P}}\mathbf{E}_1 - \mathbf{P}\|_2 = O_p(\Pi_{1,n}) = \|\widehat{\mathbf{Q}}\mathbf{E}_2 - \mathbf{Q}\|_2$,

provided that $\Pi_{1,n} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c \in (0, 1)$ depending only on r_1 and r_2 specified in Condition 4.

Recall the columns of $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, the d orthonormal eigenvectors corresponding to the d nonzero eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$. The presence of $\mathbf{E}_1$ and $\mathbf{E}_2$ accounts for the indeterminacy of those eigenvectors due to reflections and/or possible tied (nonzero) eigenvalues. Let $\tilde{\mathbf{w}} = (\mathbf{E}_2 \otimes \mathbf{E}_1)^\top \mathbf{w}$, with $\mathbf{w} \in \mathbb{R}^{d^2}$ involved in (31) for the definition of $\hat{\eta}_t = \mathbf{w}^\top \mathrm{vec}(\widehat{\mathbf{Z}}_t)$, and define

$$\Sigma_{Z,\tilde{\eta}}(k) = \mathbf{P}^\top \widetilde{\Theta}^\top \Sigma_{\overset{\circ}{Y}}(k) \mathbf{Q},$$

where $\widetilde{\Theta} = \mathbf{I}_p \otimes \{(\mathbf{Q} \otimes \mathbf{P})\tilde{\mathbf{w}}\}$, and $\Sigma_{\overset{\circ}{Y}}(k)$ is specified in (28). As indicated in the online supplementary material, Lemma 2, $\mathbf{E}_1^\top \widehat{\Sigma}_{Z,\eta}(k)\mathbf{E}_2$ is consistent to $\Sigma_{Z,\tilde{\eta}}(k)$ under the spectral norm $\|\cdot\|_2$ rather than $\Sigma_{Z,\eta}(k)$ given in (28). In comparison to $\Sigma_{Z,\eta}(k)$, we replace $\mathbf{w}$ by $\tilde{\mathbf{w}}$ in defining $\Sigma_{Z,\tilde{\eta}}(k)$. As we discussed in the beginning of Section 3.2.2, the selection of $\mathbf{w}$ for the identification of $\mathbf{U}$ and $\mathbf{V}$ is not unique. Define

$$\widetilde{\mathbf{S}}_1 = \Sigma_{Z,\tilde{\eta}}(1)^\top \Sigma_{Z,\tilde{\eta}}(1), \quad \widetilde{\mathbf{S}}_2 = \Sigma_{Z,\tilde{\eta}}(1)^\top \Sigma_{Z,\tilde{\eta}}(2),$$

$$\widetilde{\mathbf{S}}_1^* = \Sigma_{Z,\tilde{\eta}}(1) \Sigma_{Z,\tilde{\eta}}(1)^\top, \quad \widetilde{\mathbf{S}}_2^* = \Sigma_{Z,\tilde{\eta}}(1) \Sigma_{Z,\tilde{\eta}}(2)^\top.$$

Let $\tilde{\mu}_\ell = \tilde{c}_{2,\ell} \tilde{c}_{1,\ell}^{-1}$ with $\tilde{c}_{k,\ell} = (n-k)^{-1} \sum_{t=k+1}^n \tilde{\mathbf{w}}^\top \mathbb{E}[\mathrm{vec}\{\mathbf{Z}_{t-k} - \mathbb{E}(\bar{\mathbf{Z}})\}\{x_{t,\ell} - \mathbb{E}(\bar{x}_{\cdot,\ell})\}]$. Under Condition 5, parallel to Proposition 2 in Section 3.2, we have that the columns of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ can be also defined, respectively, as

$$\mathbf{u}_\ell = \frac{\Sigma_{Z,\tilde{\eta}}(1)\mathbf{v}^\ell}{|\Sigma_{Z,\tilde{\eta}}(1)\mathbf{v}^\ell|_2} \quad \text{and} \quad \mathbf{v}_\ell = \frac{\Sigma_{Z,\tilde{\eta}}(1)^\top \mathbf{u}^\ell}{|\Sigma_{Z,\tilde{\eta}}(1)^\top \mathbf{u}^\ell|_2},$$

with $\mathbf{v}^\ell$ and $\mathbf{u}^\ell$ being, respectively, the eigenvectors of the generalized eigenequations

$$\widetilde{\mathbf{S}}_2 \boldsymbol{\delta} = \tilde{\mu}_\ell \widetilde{\mathbf{S}}_1 \boldsymbol{\delta} \quad \text{and} \quad \widetilde{\mathbf{S}}_2^* \boldsymbol{\delta} = \tilde{\mu}_\ell \widetilde{\mathbf{S}}_1^* \boldsymbol{\delta}. \tag{37}$$

The following conditions are needed in our theoretical analysis.

Condition 5   (i) All the values $\tilde{\mu}_1, \ldots, \tilde{\mu}_d$ are finite and distinct. (ii) The eigenvalues of $\widetilde{\mathbf{S}}_1$ are uniformly bounded away from zero.

Condition 6   (i) There exists a universal constant $C_6 > 0$ such that $\max_{k \in \{1,2\}} \|\Sigma_{\overset{\circ}{Y}}(k)\|_2 \leq C_6$.
(ii) Write $\Sigma_{\overset{\circ}{Y}}(k) = \{\sigma_{\overset{\circ}{y},r,s}^{(k)}\}_{(p^2 q) \times q}$. It holds that $\max_{r \in [p^2 q]} \sum_{s=1}^q |\sigma_{\overset{\circ}{y},r,s}^{(k)}|^\iota \leq s_3$ and $\max_{s \in [q]} \sum_{r=1}^{p^2 q} |\sigma_{\overset{\circ}{y},r,s}^{(k)}|^\iota \leq s_4$ for some universal constant ι specified in Condition 3(ii), where s_3 and s_4 may, respectively, diverge together with p and q.

Under Condition 5, $\mathbf{v}^\ell$ and $\mathbf{u}^\ell$ can be uniquely identified by the generalized eigenequations (37) upto the scaling and permutation indeterminacy. Recall $\Sigma_{\overset{\circ}{Y}}(k)$ is a $(p^2 q) \times q$ matrix. Condition 6(i) requires the largest singular value of $\Sigma_{\overset{\circ}{Y}}(k)$ is uniformly bounded away from infinity. Our technical proofs indeed allow $\max_{k \in \{1,2\}} \|\Sigma_{\overset{\circ}{Y}}(k)\|_2$ to diverge with n. We impose Condition 6(i) just for simplifying the presentation. Condition 6(ii) imposes some sparsity requirement on $\Sigma_{\overset{\circ}{Y}}(k)$. Same as our discussion above for the validity of Condition 3(ii) imposed on the sparsity of $\Sigma_{Y,\xi}(k)$, Condition 6(ii) holds automatically for certain (s_3, s_4) under some sparsity condition imposed on the loading matrices $\mathbf{A}$ and $\mathbf{B}$.

Let $\boldsymbol{\beta}_{\mathbf{v},\ell}$ and $\boldsymbol{\beta}_{\mathbf{u},\ell}$ be the eigenvectors with unit ℓ^2-norm of the generalized eigenequations (37) associated with $\tilde{\mu}_\ell$, i.e., $\widetilde{\mathbf{S}}_2\boldsymbol{\beta}_{\mathbf{v},\ell} = \tilde{\mu}_\ell\widetilde{\mathbf{S}}_1\boldsymbol{\beta}_{\mathbf{v},\ell}$ and $\widetilde{\mathbf{S}}_2^*\boldsymbol{\beta}_{\mathbf{u},\ell} = \tilde{\mu}_\ell\widetilde{\mathbf{S}}_1^*\boldsymbol{\beta}_{\mathbf{u},\ell}$. By Condition 5(ii), we know $\widetilde{\mathbf{S}}_1$ and $\widetilde{\mathbf{S}}_1^*$ are two invertible symmetric matrices. Hence, $\boldsymbol{\beta}_{\mathbf{v},\ell}$ and $\boldsymbol{\beta}_{\mathbf{u},\ell}$ are, respectively, also the eigenvectors of the eigenequations $\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\boldsymbol{\delta} = \tilde{\mu}_\ell\boldsymbol{\delta}$ and $(\widetilde{\mathbf{S}}_1^*)^{-1}\widetilde{\mathbf{S}}_2^*\boldsymbol{\delta} = \tilde{\mu}_\ell\boldsymbol{\delta}$. For given $\boldsymbol{\beta}_{\mathbf{v},\ell}$ and $\boldsymbol{\beta}_{\mathbf{u},\ell}$, there exist two $d \times (d-1)$ matrices $\mathbf{R}_{\mathbf{v},\ell}$ and $\mathbf{R}_{\mathbf{u},\ell}$ such that $(\boldsymbol{\beta}_{\mathbf{v},\ell}, \mathbf{R}_{\mathbf{v},\ell})$ and $(\boldsymbol{\beta}_{\mathbf{u},\ell}, \mathbf{R}_{\mathbf{u},\ell})$ are two orthogonal matrices. For any $\ell \in [d]$, define

$$\theta_\ell = \sigma_{\min}(\mathbf{R}_{\mathbf{v},\ell}^\top\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\mathbf{R}_{\mathbf{v},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}) \quad \text{and} \quad \theta_\ell^* = \sigma_{\min}\{\mathbf{R}_{\mathbf{u},\ell}^\top(\widetilde{\mathbf{S}}_1^*)^{-1}\widetilde{\mathbf{S}}_2^*\mathbf{R}_{\mathbf{u},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}\}, \tag{38}$$

the smallest singular values of $\mathbf{R}_{\mathbf{v},\ell}^\top\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\mathbf{R}_{\mathbf{v},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}$ and $\mathbf{R}_{\mathbf{u},\ell}^\top(\widetilde{\mathbf{S}}_1^*)^{-1}\widetilde{\mathbf{S}}_2^*\mathbf{R}_{\mathbf{u},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}$, respectively. Under Condition 5(i), we know $\min_{\ell\in[d]}\theta_\ell > 0$ and $\min_{\ell\in[d]}\theta_\ell^* > 0$. Such defined θ_ℓ and θ_ℓ^* can be viewed as the extension of the concept 'eigen-gap' in symmetric matrices to non-symmetric matrices. If $\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2$ is a symmetric matrix, such defined θ_ℓ is actually the eigen-gap $\min_{j:j\neq\ell}|\tilde{\mu}_j - \tilde{\mu}_\ell|$. Write $\widehat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_{\hat{d}})$ and $\widehat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_{\hat{d}})$. Define

$$\Pi_{2,n} = (s_3 s_4)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}.$$

Theorem 2 indicates that the columns of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ defined in (34) are, respectively, consistent to those of $\mathbf{A}$ and $\mathbf{B}$ upto the reflection and permutation indeterminacy.

Theorem 2 Let Conditions 1–6 hold and the threshold levels $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ and $\delta_2 = C_{**}\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constants $C_* > 0$ and $C_{**} > 0$. If $\hat{d} = d$, there exists a permutation of $(1, \ldots, d)$, denoted by $(j_1, \ldots, j_d)$, such that $|\kappa_{1,\ell}\hat{\mathbf{a}}_{j_\ell} - \mathbf{a}_\ell|_2 = (1 + \theta_\ell^{-1})\cdot O_p(\Pi_{1,n} + \Pi_{2,n})$ and $|\kappa_{2,\ell}\hat{\mathbf{b}}_{j_\ell} - \mathbf{b}_\ell|_2 = \{1 + (\theta_\ell^*)^{-1}\}\cdot O_p(\Pi_{1,n} + \Pi_{2,n})$ for any $\ell \in [d]$ with some $\kappa_{1,\ell}, \kappa_{2,\ell} \in \{-1, 1\}$, provided that $(\Pi_{1,n} + \Pi_{2,n})\max\{1, d^{1/2}\theta_\ell^{-1}, d^{1/2}(\theta_\ell^*)^{-1}, d^{1/2}\theta_\ell^{-2}, d^{1/2}(\theta_\ell^*)^{-2}\} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c \in (0, 1)$ depending only on r_1 and r_2 specified in Condition 4. Furthermore, it also holds that $1 - |\hat{\mathbf{a}}_{j_\ell}^H\mathbf{a}_\ell|^2 = (1 + \theta_\ell^{-1})^2\cdot O_p(\Pi_{1,n}^2 + \Pi_{2,n}^2)$ and $1 - |\hat{\mathbf{b}}_{j_\ell}^H\mathbf{b}_\ell|^2 = \{1 + (\theta_\ell^*)^{-1}\}^2\cdot O_p(\Pi_{1,n}^2 + \Pi_{2,n}^2)$ for any $\ell \in [d]$. Here, the terms $O_p(\Pi_{1,n} + \Pi_{2,n})$ and $O_p(\Pi_{1,n}^2 + \Pi_{2,n}^2)$ hold uniformly over $\ell \in [d]$.

For $(j_1, \ldots, j_d)$ specified in Theorem 2, Proposition 1 in Section 3.1 shows that $\hat{\mathbf{a}}_{j_\ell}$ and $\hat{\mathbf{b}}_{j_\ell}$ may not be real vectors for some $\ell \in [d]$ although $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ are real vectors for all $\ell \in [d]$. When $\hat{d} = d$, we can measure the difference between $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\widehat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_d)$ by $\max_{\ell\in[d]}(1 - |\hat{\mathbf{a}}_{j_\ell}^H\mathbf{a}_\ell|^2)$ with $(j_1, \ldots, j_d)$ specified in Theorem 2. In finite samples, $\hat{d}$ may not be exactly equal to d. In general scenario without assuming $\hat{d} = d$, we consider to measure the difference between $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\widehat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_{\hat{d}})$ by

$$\rho^2(\mathbf{A}, \widehat{\mathbf{A}}) = \max_{\ell\in[d]}\min_{j\in[\hat{d}]}(1 - |\hat{\mathbf{a}}_j^H\mathbf{a}_\ell|^2). \tag{39}$$

Analogously, we can measure the difference between $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ and $\widehat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_{\hat{d}})$ by

$$\rho^2(\mathbf{B}, \widehat{\mathbf{B}}) = \max_{\ell\in[d]}\min_{j\in[\hat{d}]}(1 - |\hat{\mathbf{b}}_j^H\mathbf{b}_\ell|^2). \tag{40}$$

When $\hat{d} = d$, Theorem 2 yields that $\rho^2(\mathbf{A}, \widehat{\mathbf{A}}) = \{1 + (\min_{\ell\in[d]}\theta_\ell)^{-1}\}^2\cdot O_p(\Pi_{1,n}^2 + \Pi_{2,n}^2)$ and

$\rho^2(\mathbf{B}, \widehat{\mathbf{B}}) = \{1 + (\min_{\ell \in [d]} \theta_\ell^*)^{-1}\}^2 \cdot O_p(\Pi_{1,n}^2 + \Pi_{2,n}^2)$. Write $\varphi_n = \{1 + (\min_{\ell \in [d]} \theta_\ell)^{-1}\}^2 (\Pi_{1,n}^2 + \Pi_{2,n}^2)$. For any $\varepsilon > 0$, there exists some constant $C_\varepsilon > 0$ such that $\mathbb{P}\{\rho^2(\mathbf{A}, \widehat{\mathbf{A}}) > C_\varepsilon \varphi_n \mid \hat{d} = d\} \leq \varepsilon$. Together with Theorem 1, we have $\mathbb{P}\{\rho^2(\mathbf{A}, \widehat{\mathbf{A}}) > C_\varepsilon \varphi_n\} \leq \mathbb{P}\{\rho^2(\mathbf{A}, \widehat{\mathbf{A}}) > C_\varepsilon \varphi_n \mid \hat{d} = d\}\mathbb{P}(\hat{d} = d) + \mathbb{P}(\hat{d} \neq d) \leq \varepsilon + o(1) \to \varepsilon$, which implies $\{1 + (\min_{\ell \in [d]} \theta_\ell)^{-1}\}^2 (\Pi_{1,n}^2 + \Pi_{2,n}^2)$, the convergence rate of $\rho^2(\mathbf{A}, \widehat{\mathbf{A}})$ conditional on $\hat{d} = d$, is also the convergence rate of $\rho^2(\mathbf{A}, \widehat{\mathbf{A}})$. Identically, we also know $\{1 + (\min_{\ell \in [d]} \theta_\ell^*)^{-1}\}^2 (\Pi_{1,n}^2 + \Pi_{2,n}^2)$ is the convergence rate of $\rho^2(\mathbf{B}, \widehat{\mathbf{B}})$.

Remark 7 (i) If p and q are fixed constants, we can select the threshold levels $\delta_1 = \delta_2 = 0$ in (29) and (32). In this scenario, Conditions 3 and 6 hold automatically with $\iota = 0$ and (s_1, s_2, s_3, s_4) being some fixed constants, and Condition 4 can be replaced by the weaker requirements that $\max_{i \in [p]} \max_{j \in [q]} \max_{t \in [n]} \mathbb{E}(|y_{i,j,t}|^{2v}) = O(1)$, $\max_{t \in [n]} \mathbb{E}(|\xi_t|^{2v}) = O(1)$, and $\sum_{k=1}^{\infty} \{\alpha(k)\}^{1-2/v} = O(1)$ for some constant $v > 2$. Under these conditions, using the Davydov inequality, we have Theorem 1, Proposition 3, and Theorem 2 hold with $\Pi_{1,n}^* = \Pi_{2,n}^* = n^{-1/2}$ and $\Pi_{1,n}^* \ll c_n \ll 1$, provided that $(\Pi_{1,n}^* + \Pi_{2,n}^*) \max\{1, \theta_\ell^{-2}, (\theta_\ell^*)^{-2}\} = o(1)$.

(ii) If p and q diverge at some polynomial rate of n, we can replace Condition 4 by the weaker requirements $\max_{i \in [p]} \max_{j \in [q]} \max_{t \in [n]} \mathbb{P}(|y_{i,j,t}| > x) = O\{x^{-2(l+\tau)}\}$ for any $x > 0$, $\max_{t \in [n]} \mathbb{P}(|\xi_t| > x) = O\{x^{-2(l+\tau)}\}$ for any $x > 0$, and $\alpha(k) = O\{k^{-(l-1)(l+\tau)/\tau}\}$ as $k \to \infty$ with some constants $l > 2$ and $\tau > 0$. Under these conditions, if the threshold levels $\delta_1 = C_*(pq)^{1/l} n^{-1/2}$ and $\delta_2 = C_{**}(pq)^{2/l} n^{-1/2}$ in (29) and (32) for some sufficiently large constants $C_* > 0$ and $C_{**} > 0$, Theorem 1, Proposition 3, and Theorem 2 hold with $\Pi_{1,n}^* = (s_1 s_2)^{1/2}\{(pq)^{1/l} n^{-1/2}\}^{1-\iota}$, $\Pi_{2,n}^* = (s_3 s_4)^{1/2}\{(pq)^{2/l} n^{-1/2}\}^{1-\iota}$ and $\Pi_{1,n}^* \ll c_n \ll 1$, provided that $(\Pi_{1,n}^* + \Pi_{2,n}^*) \max\{1, d^{1/2}\theta_\ell^{-1}, d^{1/2}(\theta_\ell^*)^{-1}, d^{1/2}\theta_\ell^{-2}, d^{1/2}(\theta_\ell^*)^{-2}\} = o(1)$.

5 Numerical studies

5.1 Simulation

We illustrate the finite-sample performance of the proposed methods by simulation based on model (2). Let $\mathbf{A}^* \equiv (a_{i,\ell}^*)_{p \times d} = (\mathbf{a}_1^*, \ldots, \mathbf{a}_d^*)$ and $\mathbf{B}^* \equiv (b_{j,\ell}^*)_{q \times d} = (\mathbf{b}_1^*, \ldots, \mathbf{b}_d^*)$ with the elements drawn from the uniform distribution on $[-3, 3]$ independently satisfying the restriction $\operatorname{rank}(\mathbf{A}^*) = d = \operatorname{rank}(\mathbf{B}^*)$. Write $\tilde{\mathbf{x}}_\ell = (\tilde{x}_{1,\ell}, \ldots, \tilde{x}_{n,\ell})^\top$ and let $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_d$ be independent AR(1) processes with independent $\mathcal{N}(0, 1)$ innovations, and the autoregressive coefficients drawn from the uniform distribution on $[-0.95, -0.6] \cup [0.6, 0.95]$. The elements of the error term $\mathcal{E}$ in (2) are drawn from $\mathcal{N}(0, 1)$ independently. Then, we generate the tensor $\mathcal{Y} = \sum_{\ell=1}^{d} \mathbf{a}_\ell^* \circ \mathbf{b}_\ell^* \circ \tilde{\mathbf{x}}_\ell + \mathcal{E}$. Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ with $\mathbf{a}_\ell = \mathbf{a}_\ell^*/|\mathbf{a}_\ell^*|_2$ and $\mathbf{b}_\ell = \mathbf{b}_\ell^*/|\mathbf{b}_\ell^*|_2$. Equivalently, we have $\mathcal{Y} = \sum_{\ell=1}^{d} \mathbf{a}_\ell \circ \mathbf{b}_\ell \circ \mathbf{x}_\ell + \mathcal{E}$, where $\mathbf{x}_\ell = |\mathbf{a}_\ell^*|_2 |\mathbf{b}_\ell^*|_2 \tilde{\mathbf{x}}_\ell$. We set $d \in \{1, 3, 6\}$, $n \in \{300, 600, 900\}$, and p, q taking values between 4 and 256. We consider the following two choices for ξ_t:

- (PCA) Let $\mathbf{Y} = \{\operatorname{vec}(\mathbf{Y}_1), \ldots, \operatorname{vec}(\mathbf{Y}_n)\}^\top$. Perform the principal component analysis for $\mathbf{Y}$ using the R-function `prcomp` in the R-package `stats`, and select ξ_t as the average of the first m principal components corresponding to the eigenvalues which count for at least 99% of the total variations.
- (Random weighting) Generate a (pq)-dimensional vector $\mathbf{h}$ with its components randomly from the uniform distribution on $[0, 1]$, and normalize $\mathbf{h}$ as a unit vector, which is denoted by $\mathbf{h}_0$. Then define $\xi_t = \mathbf{h}_0^\top \operatorname{vec}(\mathbf{Y}_t)$.

For the refined method, $\hat{\eta}_t$ is specified in the same manners with $\mathbf{Y}_t$ replaced by $\widehat{\mathbf{Z}}_t$. We only present the results for the cases with $p \geq q$. More simulation results with $p < q$ can be found in the online supplementary material.

We first consider the finite-sample performance of the estimation for d by (16) of the direct estimation and by (30) of the refined method. We set $\delta_1 = 0$ and $c_n = 0$ in (16) and (30). Table 1 reports the relative frequency estimates of $\mathbb{P}(\hat{d} = d)$ based on 2000 repetitions with ξ_t determined by PCA. When $d = 1$, we observe $\hat{d} \equiv d$ for both the direct and refined methods in all the simulation replications. For $d > 1$, the relative frequency estimates of $\mathbb{P}(\hat{d} = d)$ based on both the direct and refined methods increase as n, p, and q grow in most of the cases. The refined method works uniformly better than the direct method except $(p, q, d, n) = (8, 8, 3, 600)$, $(8, 8, 3, 900)$, and $(16, 16, 3, 900)$, and their performances in these three cases are similar. As d increases, the improvement from using the refined method also increases. Also, the refined method with larger K has better performance in most of the cases. As shown in the proof of Theorem 1 in the online supplementary material, the consistency of $\hat{d}$ depends on the convergence rate of $\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2$. Recall $\widehat{\mathbf{M}}_1 = \sum_{k=1}^{K} T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\} T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)^{\top}\}$ and $\mathbf{M}_1 = \sum_{k=1}^{K} \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)^{\top}$. The proof of Lemma 1 in the online supplementary material indicates that the convergence rate of $|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k) - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)|_{\infty}$ plays a key role in deriving the convergence rate of $\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2$. If K is a fixed constant, $\max_{k \in [K]} |\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k) - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)|_{\infty} = O_p[\{n^{-1} \log (pq)\}^{1/2}]$. If K diverges with n, K will appear in the convergence rate of $\max_{k \in [K]} |\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k) - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)|_{\infty}$. Then the convergence rate of $\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2$ with diverging K will be slower than that with fixed K. Hence, we cannot select K as large as possible since too large K may lead to a bad estimate $\hat{d}$. We suggest to restrict $K \leq 10$ in practice. In the online supplementary material, Table S4 reports the results using randomly weighted ξ_t; showing the similar patterns as those in Table 1. Note that using PCA-based ξ_t produces uniformly more accurate estimates than using randomly weighted ξ_t.

Tables S5–S7 in the online supplementary material present the averages and standard deviations of the estimation errors $\rho^2(\mathbf{A}, \widehat{\mathbf{A}})$ and $\rho^2(\mathbf{B}, \widehat{\mathbf{B}})$ defined in (39) and (40) based on 2,000 repetitions. To highlight the key information, Figure 1 plots the results of the direct method and the refined method with $K = 3$. It shows that (a) the refined method outperforms the direct method uniformly when $d > 1$, (b) two methods perform about the same in some cases when $d = 1$, and (c) the PCA-based ξ_t performs better than the randomly weighted ξ_t. Figure 2 summarizes the performance of the refined method with $K \in \{3, 5, 7\}$ and ξ_t determined by the PCA method. We can find that (a) the refined method performs about the same for $K \in \{3, 5, 7\}$ when $d = 1$ and (b) the refined method with larger K in general has slightly better performance when $d > 1$, mainly because larger K is more likely to lead to more accurate estimate of d, see Table 1.

5.2 A real data analysis

In this section, we analyse the monthly average value weighted returns of the 100 portfolios from January 1990 to December 2017. The portfolios include all NYSE, AMEX, and NASDAQ stocks, which are constructed by the intersections of 10 levels of size (market equity) and 10 levels of the book equity to market equity ratio (BE). The data were downloaded from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. Although this website provides monthly return data from July 1926 to June 2021, there are many missing values in the early years. We restrict the time period from January 1990 to December 2017 to avoid the large numbers of missing data and large fluctuations. The data can be represented as a 10×10 matrix $\mathbf{Y}_t = (y_{i,j,t})$ for $t = 1, \ldots, 336$ (i.e., $p = q = 10$, $n = 336$), where $y_{i,j,t}$ is the return of the portfolio at the ith level of size and jth level of the BE-ratio at time t. We impute the missing values by the weighted averages of the three previous months, i.e., set $y_{i,j,t} = 0.5y_{i,j,t-1} + 0.3y_{i,j,t-2} + 0.2y_{i,j,t-3}$ for missing $y_{i,j,t}$.

We standardize each of the 100 component time series $\{y_{i,j,t}\}_{t=1}^{n}$ so that they have mean zero and unit variance. To economize the notation, we still use $y_{i,j,t}$ to denote the standardized data. Figure 3 shows the plots of the standardized return series $\{y_{i,j,t}\}_{t=1}^{n}$, for $i, j = 1, \ldots, 10$. The rows in Figure 3 correspond to the 10 levels of size and the columns correspond to the 10 levels of the BE-ratio. Notice that the ranges of the vertical values are not the same, and the figures are not directly comparable. All the 100 return series appear to be stationary. The ACF (autocorrelation functions) plots of these 100 time series indicate that most series have significant ACF at the first lag, and all series do not show any seasonal patterns. The cross correlations between different time series are mostly significant at time lags 0 and 1.
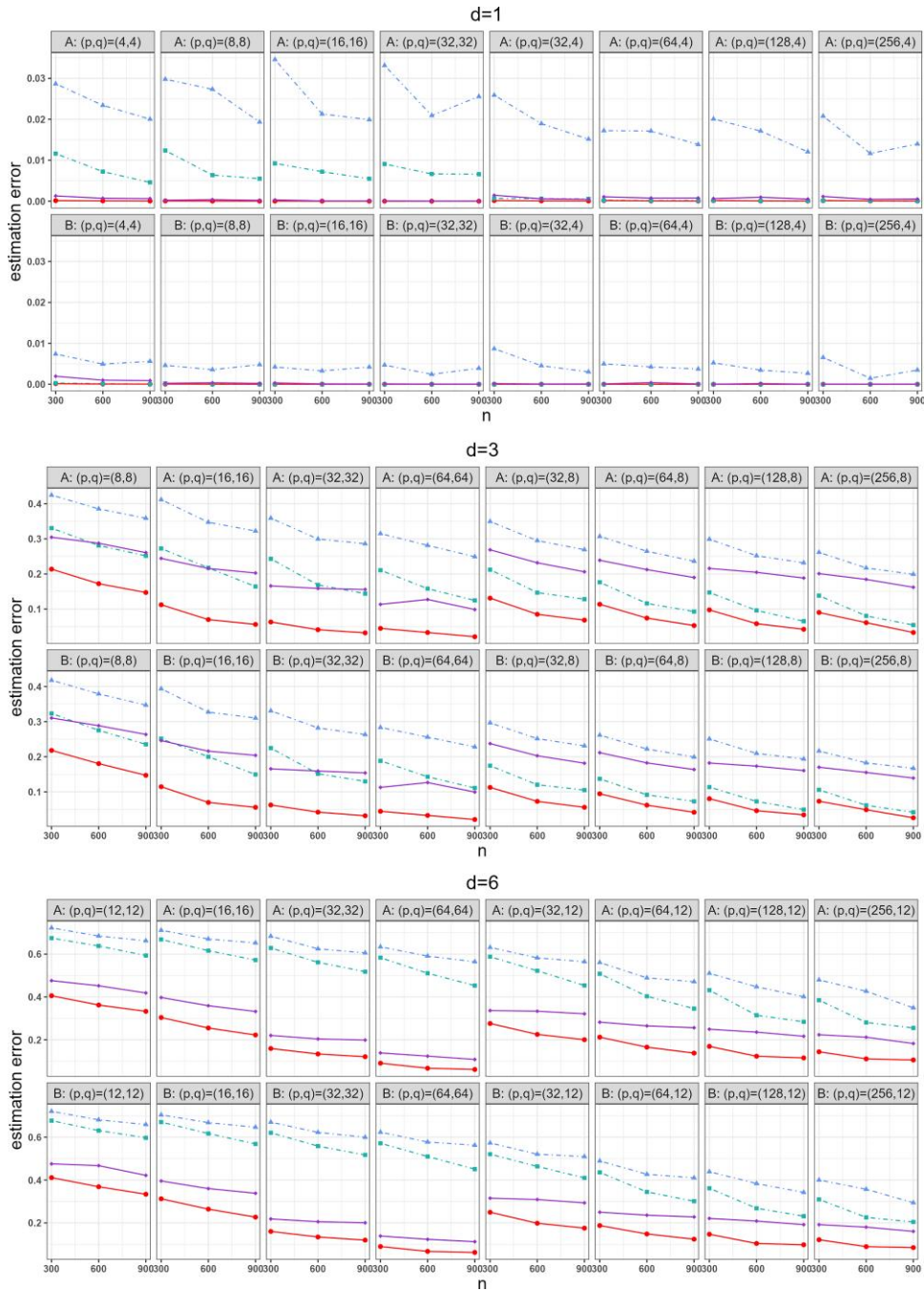
Figure 1. The lineplots for the averages of $\rho^2(\mathbf{A}, \widehat{\mathbf{A}})$ and $\rho^2(\mathbf{B}, \widehat{\mathbf{B}})$ based on 2,000 repetitions. The legend is defined as follows: (i) the direct method with PCA-based ξ_t ($- \cdot -x25A0 - \cdot -$); (ii) the direct method with randomly weighted ξ_t ($- \cdot -\blacktriangle - \cdot -$); (iii) the refined method ($K = 3$) with PCA-based ξ_t ($\!-\!\bullet\!-\!$); and (iv) the refined method ($K = 3$) with randomly weighted ξ_t ($\!-\!\blacklozenge\!-\!$).

We apply our model (4) to fit the standardized matrix time series $\{\mathbf{Y}_t\}_{t=1}^{336}$ using the refined estimation method with PCA-based ξ_t; leading to $\hat{d} \equiv 1$ with $K = 3$, 5, or 7. See (30). In the sequel, we only present the results with $K = 5$. The results based on $K \in \{3, 7\}$ are similar and thus omitted here. Based on (34), we obtain $\widehat{\mathbf{A}} = (0.44, 0.34, 0.32, 0.32, 0.29, 0.25, 0.30, 0.30, 0.33, 0.23)^\top$
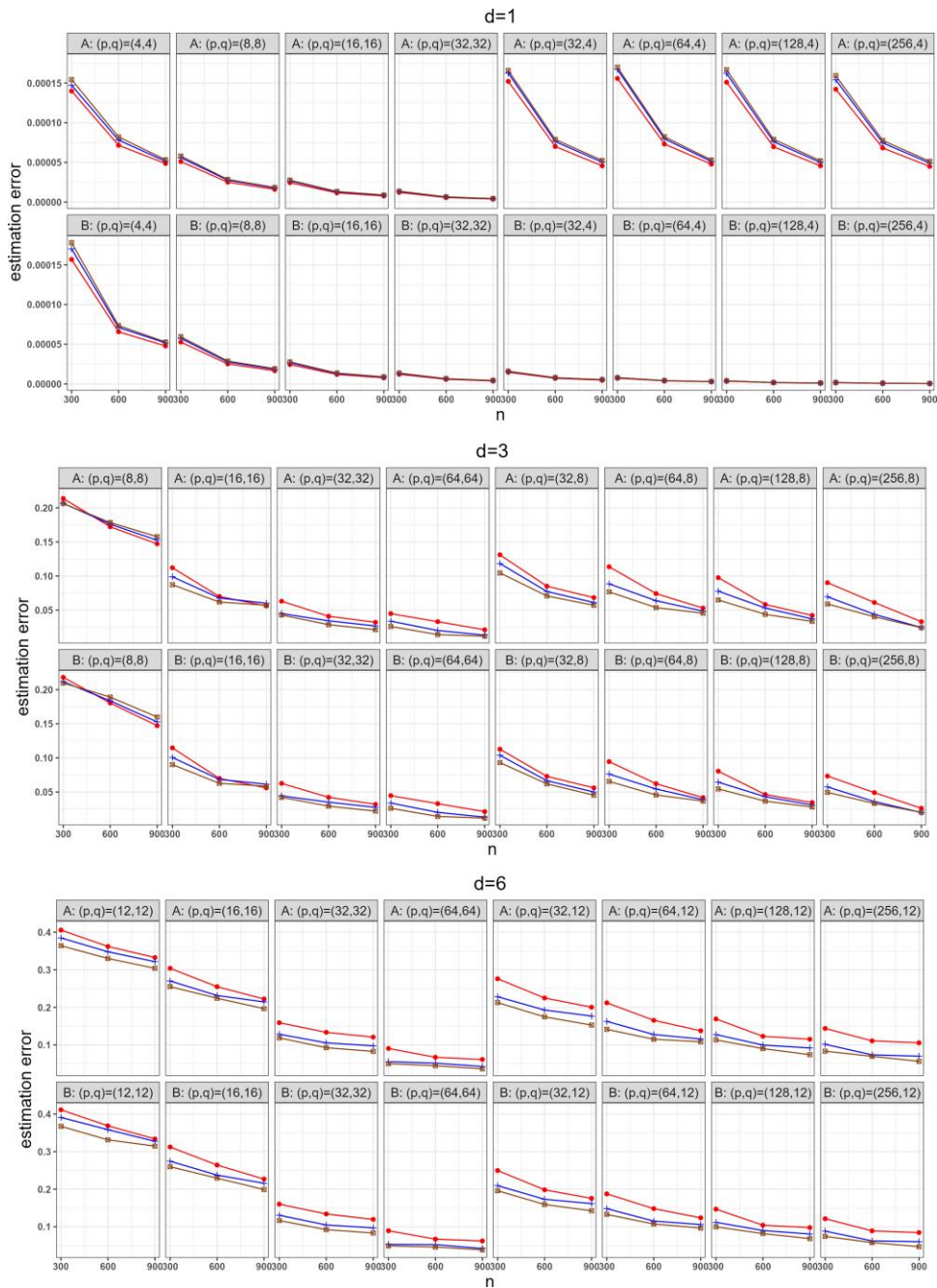
Figure 2. The lineplots for the averages of $\rho^2(\mathbf{A}, \widehat{\mathbf{A}})$ and $\rho^2(\mathbf{B}, \widehat{\mathbf{B}})$ based on 2,000 repetitions, where $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ is estimated by the refined method ($K = 3, 5, 7$) with PCA-based ξ_t. The legend is defined as follows: (i) $K = 3$ (—•—); (ii) $K = 5$ (—+—); and (iii) $K = 7$ (—⊠—).

and $\widehat{\mathbf{B}} = (0.20, 0.26, 0.27, 0.29, 0.35, 0.33, 0.34, 0.31, 0.37, 0.39)^{\top}$. Following the arguments above Proposition 1 in Section 3.1, we can recover the latent time series $\{\hat{x}_{t,1}\}_{t=1}^{336}$. Figure 4 displays the plots of time series $\{\hat{x}_{t,1}\}_{t=1}^{336}$ and its ACF, which shows that the autocorrelations of $\{\hat{x}_{t,1}\}_{t=1}^{336}$ is significant at the first lag that is consistent to the ACF patterns of $\mathbf{Y}_t$. The Akaike information criterion (AIC) suggests to fit $\{\hat{x}_{t,1}\}_{t=1}^{336}$ by an AR(1) model. Hence, to model this 10×10 matrix time

Table 1. Relative frequency estimates of $\mathbb{P}(\hat{d} = d)$ based on 2,000 repetitions with PCA-based ξ_t, where the direct estimate and the refined estimate are given in (16) and (30), respectively

|			Refined			Direct			Refined			Direct
	(**p**, **q**)	**n**	**K = 3**	**K = 5**	**K = 7**		(**p**, **q**)	**n**	**K = 3**	**K = 5**	**K = 7**	
$d = 1$	(4, 4)	300	100.00	100.00	100.00	100.00	(32, 4)	300	100.00	100.00	100.00	100.00
		600	100.00	100.00	100.00	100.00		600	100.00	100.00	100.00	100.00
		900	100.00	100.00	100.00	100.00		900	100.00	100.00	100.00	100.00
	(8, 8)	300	100.00	100.00	100.00	100.00	(64, 4)	300	100.00	100.00	100.00	100.00
		600	100.00	100.00	100.00	100.00		600	100.00	100.00	100.00	100.00
		900	100.00	100.00	100.00	100.00		900	100.00	100.00	100.00	100.00
	(16, 16)	300	100.00	100.00	100.00	100.00	(128, 4)	300	100.00	100.00	100.00	100.00
		600	100.00	100.00	100.00	100.00		600	100.00	100.00	100.00	100.00
		900	100.00	100.00	100.00	100.00		900	100.00	100.00	100.00	100.00
	(32, 32)	300	100.00	100.00	100.00	100.00	(256, 4)	300	100.00	100.00	100.00	100.00
		600	100.00	100.00	100.00	100.00		600	100.00	100.00	100.00	100.00
		900	100.00	100.00	100.00	100.00		900	100.00	100.00	100.00	100.00
$d = 3$	(8, 8)	300	78.85	79.85	80.35	78.75	(32, 8)	300	88.65	90.20	91.55	85.65
		600	82.45	82.15	81.65	83.75		600	92.95	93.65	94.50	92.05
		900	85.55	85.05	84.50	86.55		900	94.45	95.25	96.00	93.20
	(16, 16)	300	89.45	90.95	92.10	88.00	(64, 8)	300	89.85	92.45	93.70	87.70
		600	93.85	94.35	95.05	92.35		600	93.55	94.60	95.75	92.75
		900	94.95	94.75	95.10	94.80		900	95.65	96.05	96.50	94.40
	(32, 32)	300	94.30	96.25	96.45	91.25	(128, 8)	300	91.40	93.55	94.90	88.85
		600	96.20	96.95	97.60	94.90		600	95.05	95.65	96.55	93.60
		900	97.20	97.80	98.35	96.20		900	96.45	97.05	97.35	95.95
	(64, 64)	300	95.80	96.95	97.80	92.10	(256, 8)	300	91.90	94.00	95.05	88.85
		600	96.95	98.25	98.90	95.15		600	94.65	96.30	96.65	93.50
		900	98.15	98.90	99.10	97.40		900	97.25	98.00	98.10	96.40
$d = 6$	(12, 12)	300	73.35	78.15	81.80	60.70	(32, 12)	300	85.25	90.85	94.55	66.45
		600	77.85	81.50	84.85	68.00		600	89.55	93.75	95.00	76.70
		900	80.15	82.90	85.10	73.05		900	90.65	93.50	95.75	81.20
	(16, 16)	300	81.50	87.00	89.05	66.30	(64, 12)	300	88.35	93.55	95.50	69.25
		600	85.35	89.60	91.40	74.55		600	92.00	95.70	97.00	80.25
		900	88.45	90.90	93.20	79.40		900	93.75	96.50	97.70	85.95
	(32, 32)	300	90.65	94.90	96.40	76.65	(128, 12)	300	90.80	95.10	96.80	72.05
		600	92.50	96.40	97.80	81.50		600	94.05	96.45	97.60	83.30
		900	93.40	96.00	97.55	85.40		900	94.60	96.85	98.40	85.65
	(64, 64)	300	94.30	98.35	99.20	79.30	(256, 12)	300	90.85	95.40	97.65	71.95
		600	96.15	98.20	99.10	85.70		600	93.90	97.40	98.25	81.95
		900	96.30	98.35	99.10	89.15		900	93.85	97.35	98.75	84.50

series $\mathbf{Y}_t$, our method essentially only needs to estimate one parameter in an AR(1) model. We also consider to fit the matrix time series $\mathbf{Y}_t$ by the following methods:

- (UniARMA) For each of 100 component time series $\{y_{i,j,t}\}_{t=1}^{336}$, we fit an ARMA model specified by the AIC; leading to the estimation for 135 coefficient parameters in the total 100 models.
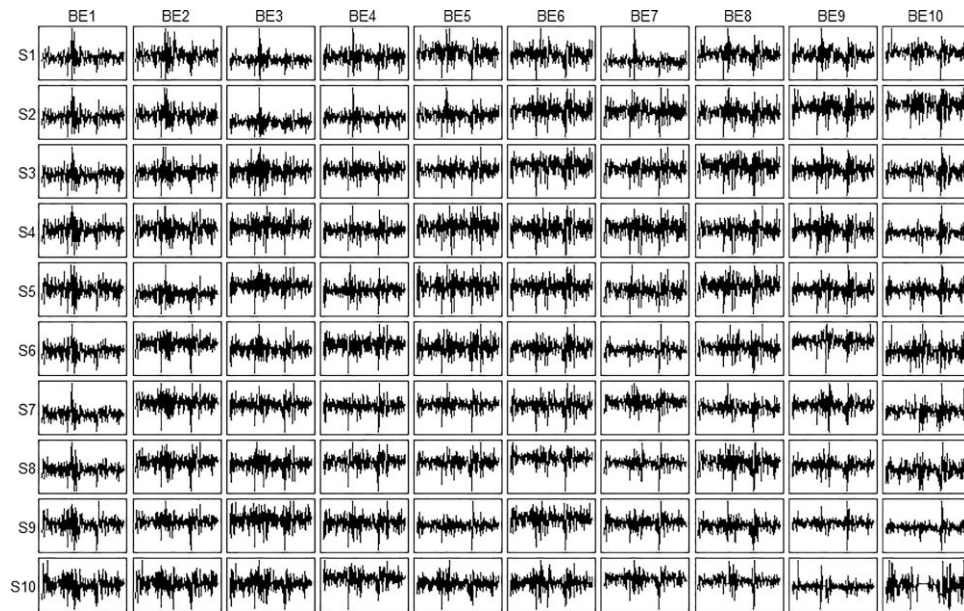
Figure 3. The plots of the return series of the portfolios formed on different levels of size (by rows) and book equity to market equity ratio (by columns). The horizontal axis represents time and the vertical axis represents the monthly returns. The ranges of the vertical values are not the same.
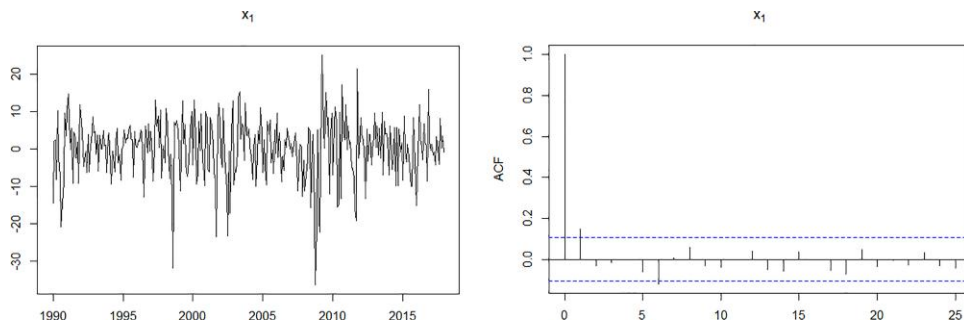
Figure 4. The plots of the latent time series $\{\hat{x}_{t,1}\}_{t=1}^{336}$ and its autocorrelation functions.

- (SVAR) Fit a sparse VAR(ℓ) model to $\{\text{vec}(\mathbf{Y}_t)\}_{t=1}^{336}$ using the R-function sparseVAR in the R-package `bigtime` with the standard lasso penalization and the optimal sparsity parameter selected by the time series cross validation procedure. The programme selects $\ell = 27$ automatically based on the time series length, and there are 270,000 parameters to be estimated.
- (MAR) Fit $\{\mathbf{Y}_t\}_{t=1}^{336}$ by the matrix-AR(1) of Chen et al. (2021), which involves 200 parameters.
- (TS-PCA) Apply the principle component analysis for time series suggested in Chang et al. (2018) to the 100-dimensional time series $\{\text{vec}(\mathbf{Y}_t)\}_{t=1}^{336}$ using the R-package HDTSA, leading to 98 univariate time series and one two-dimensional time series. For the obtained univariate time series, we fit it by an ARMA model with the order determined by the AIC. For the obtained two-dimensional time series, we fit it by an VAR model with the order determined by the AIC. There are in total 93 parameters in the models.
- (FAC) Apply the factor model of Wang et al. (2019) to matrix time series $\{\mathbf{Y}_t\}_{t=1}^{336}$ with the predetermined parameter $h_0 = 1$ as suggested in the real data analysis part of their paper. Based on their method, we find there is only one factor. We fit the latent factor series by an AR(1) model specified by the AIC which only needs to estimate one parameter.

Table 2. Fitting errors for the monthly data from year 1990 to 2017. The computational time is conducted on the Windows platform with Intel(R) Core(TM) i7-8550U CPU at 1.99 GHz

	Proposed	TS-PCA	FAC	UniARMA	SVAR	MAR
RMSE	0.9913	0.9935	0.9923	0.9895	0.9985	0.9613
MAE	0.7432	0.7456	0.7436	0.7417	0.7444	0.7235
#Parameters	1	93	1	135	270000	200
time (s)	0.3172	6.4618	0.6596	6.7335	1689.1860	1.8470

Table 3. One-step and two-step ahead forecasting errors for the monthly readings in the last 2 years 2016 and 2017

	Proposed	TS-PCA	FAC	UniARMA	SVAR	MAR
	One-step forecast					
rRMSE	0.7678	0.7802	0.7701	0.7724	0.7690	0.8067
rMAE	0.5609	0.5696	0.5649	0.5652	0.5614	0.5948
	Two-step forecast					
rRMSE	0.7668	0.7526	0.7683	0.7707	0.7693	0.7728
rMAE	0.5590	0.5512	0.5610	0.5638	0.5616	0.5627

While UniARMA, SVAR and MAR model $\mathbf{Y}_t$ or vec($\mathbf{Y}_t$) directly, our proposed method, TS-PCA and FAC seek dimension reduction first and then model the resulting low-dimensional time series. Both RMSE and MAE, defined as below, of the fitted models are listed in Table 2:

$$\text{RMSE} = \left\{ \frac{1}{33{,}600} \sum_{t=1}^{336} \sum_{i=1}^{10} \sum_{j=1}^{10} (\hat{y}_{i,j,t} - y_{i,j,t})^2 \right\}^{1/2}, \quad \text{MAE} = \frac{1}{33{,}600} \sum_{t=1}^{336} \sum_{i=1}^{10} \sum_{j=1}^{10} |\hat{y}_{i,j,t} - y_{i,j,t}|.$$

Among the three dimension-reduction methods, our proposed method has the smallest RMSE and MAE, while MAR achieves the overall minimum RMSE and MAE.

We also evaluate the post-sample forecasting performance of these methods by performing the one-step and two-step ahead rolling forecasts for the 24 monthly readings in the last 2 years (i.e., 2016 and 2017). For each $s = 1, \ldots, 24$, we use our proposed method and the other five methods to fit $\{\mathbf{Y}_t\}_{t=s}^{311+s}$ and then obtain the one-step forecast of $\mathbf{Y}_{312+s}$ denoted by $\widehat{\mathbf{Y}}_{312+s} = \{\hat{y}_{i,j,312+s}^{(s)}\}_{10 \times 10}$. For the two-step ahead forecast, we fit $\{\mathbf{Y}_t\}_{t=s}^{310+s}$ by the six methods, and the two-step ahead forecast $\widehat{\mathbf{Y}}_{312+s}$ can be obtained by plugging-in the one-step forecast into the models. For our proposed method, TS-PCA and FAC, if the dimension of the obtained latent time series is larger than 1 we fit it by a VAR model with the order determined by the AIC, otherwise, we fit it by an ARMA model with the order determined by the AIC. The post-sample forecasting performance is evaluated by the rRMSE and rMAE defined as

$$\text{rRMSE} = \left[\frac{1}{2{,}400} \sum_{s=1}^{24} \sum_{i=1}^{10} \sum_{j=1}^{10} \{\hat{y}_{i,j,312+s}^{(s)} - y_{i,j,312+s}\}^2 \right]^{1/2},$$

$$\text{rMAE} = \frac{1}{2{,}400} \sum_{s=1}^{24} \sum_{i=1}^{10} \sum_{j=1}^{10} |\hat{y}_{i,j,312+s}^{(s)} - y_{i,j,312+s}|.$$

Table 3 summarizes the post-sample forecasting rRMSE and rMAE. The newly proposed method, in spite of its simplicity, exhibits the promising post-sample forecasting performance, as its rRMSE

and rMAE are the smallest in one-step ahead forecasting among all the methods concerned, and are the second smallest in the two-step ahead forecast for which only TS-PCA has smaller rRMSE and rMAE.

Acknowledgments

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

Conflict of interest: None declared.

Funding

References

Anandkumar A., Ge R., Hsu D., & Janzamin M. (2014). 'Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates', arXiv, arXiv:1402.5180, preprint: not peer reviewed.

Bickel P. J., & Levina E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, *36*(6), 2577–2604. https://doi.org/10.1214/08-AOS600

Chang J., Chen X., & Wu M. (2021). 'Central limit theorems for high dimentional dependent data', arXiv, arXiv:2104.12929, preprint: not peer reviewed.

Chang J., Guo B., & Yao Q. (2015). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. *Journal of Econometrics*, *189*(2), 297–312. https://doi.org/10.1016/j.jeconom.2015.03.024

Chang J., Guo B., & Yao Q. (2018). Principal component analysis for second-order stationary vector time series. *The Annals of Statistics*, *46*(5), 2094–2124. https://doi.org/10.1214/17-AOS1613

Chang J., Hu Q., Liu C., & Tang C. Y. (in press). Optimal covariance matrix estimation for high-dimensional noise in high-frequency data. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2022.06.010

Chen E. Y., & Chen R. (2019). 'Modeling dynamic transport network with matrix factor models: With an application to international trade flow', arXiv, arXiv:1901.00769, preprint: not peer reviewed.

Chen E. Y., Tsay R. S., & Chen R. (2020). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, *115*(530), 775–793. https://doi.org/10.1080/01621459.2019.1584899

Chen R., Xiao H., & Yang D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, *222*(1), 539–560. https://doi.org/10.1016/j.jeconom.2020.07.015

Colombo N., & Vlassis N. (2016). Tensor decomposition via joint matrix schur decomposition. In *International Conference on Machine Learning* (Vol. 48, pp. 2820–2828). PMLR.

Domanov I., & De Lathauwer L. (2014). Canonical polyadic decomposition of third-order tensors: reduction to generalized eigenvalue decomposition. *SIAM Journal on Matrix Analysis and Applications*, *35*(2), 636–660. https://doi.org/10.1137/130916084

Golub G. H., & Van Load C. F. (2013). *Matrix computations* (4th ed.). Johns Hopkins University Press.

Han Y., Chen R., Zhang C.-H., & Yao Q. (in press). Simultaneous decorrelation of matrix time series. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2022.2151448

Han Y., Zhang C.-H., & Chen R. (2021). 'CP factor model for dynamic tensors', arXiv, arXiv:2110.15517, preprint: not peer reviewed.

Han Y., & Zhang C.-H. (in press). Tensor principal component analysis in high dimensional CP models. *IEEE Transcations on Information Theory*. https://doi.org/10.1109/TIT.2022.3203972

Kolda T. G., & Bader B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500. https://doi.org/10.1137/07070111X

Lam C., & Yao Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, *40*(2), 694–726. https://doi.org/10.1214/12-AOS970

Liu Y., Shang F., Fan W., Cheng J., & Cheng H. (2014). Generalized higher-order orthogonal iteration for tensor decomposition and completion. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*.

Sanchez E., & Kowalski B. R. (1990). Tensorial resolution: A direct trilinear decomposition. *Journal of Chemometrics*, *4*(1), 29–45. https://doi.org/10.1002/cem.1180040105

Sharan V., & Valiant G. (2017). Orthogonalized ALS: A theoretically principled tensor decomposition algorithm for practical use. In *International Conference on Machine Learning* (Vol. 70, pp. 3095–3104). PMLR.

Sun W. W., Lu J., Liu H., & Cheng G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *79*(3), 899–916. https://doi.org/10.1111/rssb.12190

Wang D., Liu X., & Chen R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, *208*(1), 231–248. https://doi.org/10.1016/j.jeconom.2018.09.013

Wang M., & Song Y. (2017). Tensor decompositions via two-mode higher-order SVD (HOSVD). In *International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 614–622). PMLR.

Zhang A., & Xia D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, *64*(11), 7311–7338. https://doi.org/10.1109/TIT.2018.2841377