Modelling Matrix Time Series via a Tensor CP-Decomposition

Jinyuan Chang [a] Jing He [a] Lin Yang [a] Qiwei Yao [b]

[a] Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China

[b] The London School of Economics and Political Science, London, WC2A 2AE, U.K.

16 December 2022

Abstract

We propose to model matrix time series based on a tensor 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. A key idea of the new procedure is to project a generalized eigenequation defined in terms of rank-reduced matrices to a lower-dimensional one with full-ranked matrices, to avoid the intricacy of the former of which the number of eigenvalues can be zero, finite and infinity. The asymptotic theory has been established under a general setting without the stationarity. It shows, for example, that all the component coefficient vectors in the CP-decomposition are estimated consistently with the different error rates, depending on the relative sizes between the dimensions of time series and the sample size. The proposed model and the estimation method are further 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 $p \cdot q$ 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 j-th variable on the i-th individual at time t. With 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 and Vader 2009). Then $y_{i,j,t}$ is the (i,j,t)-th element $\mathcal{Y}$. Conceptually we decompose $y_{i,j,t}$ 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) below). 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 modeling the dynamic behavour of the process.

The 'workhorse' method for CP-decompositions is the so-called alternative least squares (ALS) algorithm which is easy to understand and to 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 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 at. (2014), Colombo and Vlassis (2016), Sun at al. (2017), Sharan and Valiant (2017), Wang and Song (2017), Zhang and Xia (2018), and Han and Zhang (2021).

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 $\mathcal{E} \equiv \mathbf{0}$ in (1). In contrast, our new procedure eliminates the impact of the noise by incorporating the serial dependence into the estimation. More importantly, to overcome the intricacy in solving a generalized eigenequation defined by rank-reduced matrices (see Section 7.7 of Golub and Van Load (2013)), we project a high-dimensional generalized eigenequation to a lower-dimensional one in which the matrices are full-ranked, and the corresponding eigenvectors are uniquely determined upto the scaling and reflection indeterminacy.

Most existing literature on matrix time series is based on the factor modelling via the Tucker decomposition; see Chen and Chen (2019), Chen et al. (2020), and Wang et al. (2019). The key difference between our approach and a Tucker decomposition based approaches is two fold: 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 below. Secondly, a Tucker decomposition entails more conventional factor models, and, therefore, we only need to identify and to estimate factor loading spaces, for which the standard factor model methods (e.g. Lam and Yao (2012), and Chang et al. (2015)) are applicable. However for a CP-decomposition, we need to identify and to estimate the component coefficient vectors precisely. Therefore a radically new inference procedure is required. The other approaches for modelling matrix time series includes: the matrix-coefficient autoregressive models of Chen et al. (2021), the bilinear transformation segmentation method of Han et al. (2021a). Han et al. (2021b) models tensor time series also based on a CP-decomposition. But their approach is radically different from ours, as the CP-decomposition is estimated based on an iterative simultaneous orthogonalization algorithm with a warm-start initialization using the so-call composite principal component analysis for tensors; see Section 3 of Han et al. (2021b). Note that our estimation is an 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 procedure. The key idea of our approach is first elucidated for weakly stationary processes. The formal identification result and the estimation procedure are presented under a general setting without the stationarity condition. The asymptotic results, including the convergence rates for the estimated component vectors in the CP-decomposition, are presented in Section 5. Numerical illustration with both simulated and real data sets is given in Section 6. All the technical proofs are relegated to the Appendix.

For a positive integer m, write $[m] = \{1, \ldots, m\}$, and denote by $\mathbf{I}_m$ the $m \times m$ identity matrix. Denote by $I(\cdot)$ the indicator function. For an $m_1 \times m_2$ matrix $\mathbf{H} = (h_{i,j})_{m_1 \times m_2}$, let $\|\mathbf{H}\|_2$, $\text{rank}(\mathbf{H})$, $\sigma_{\min}(\mathbf{H})$ and $\text{vec}(\mathbf{H})$ 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-dimensional vector $\mathbf{H}$. 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 $\text{tr}(\mathbf{H})$ the trace of $\mathbf{H}$. For any $m \times r$ matrix $\mathbf{H}$, denote by $\mathcal{M}(\mathbf{H})$ the linear space spanned by the r columns of $\mathbf{H}$. Let $\otimes$ and $\circ$ denote the Kronecker product and the vector outer product, respectively.

2 Models

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 $\mathbf{a}_\ell = (a_{1,\ell}, \ldots, a_{p,\ell})^{\mathrm{T}}$ and $\mathbf{b}_\ell = (b_{1,\ell}, \ldots, b_{q,\ell})^{\mathrm{T}}$ are, respectively, $p \times 1$ and $q \times 1$ constant vectors, and $\mathbf{x}_\ell = (x_{1,\ell}, \ldots, x_{n,\ell})^{\mathrm{T}}$ is an $n \times 1$ random vectors, and $d \geq 1$ is an unknown integer. Furthermore we assume $d < \min(p, q)$. Put

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

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 times 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 $\mathcal{R}(\mathbf{A}) + \mathcal{R}(\mathbf{B}) + \mathcal{R}(\mathscr{X}) \geq 2d + 2$, where $\mathscr{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ and $\mathcal{R}(\cdot) = \max\{k : \text{any } k \text{ columns of the matrix are linear independent}\}$. Such requirement provides a sufficient condition for the uniqueness (Kolda and 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. The setting adopted by Dunlavy et al. (2011) is equivalent to (2) with $\mathcal{E} \equiv \mathbf{0}$.

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

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

where $\mathbf{X}_t = \mathrm{diag}(x_{t,1}, \ldots, x_{t,d})$ and $\boldsymbol{\varepsilon}_t$ 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 $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{B}) = d$. Furthermore, $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ for any t, $\mathbb{E}(\boldsymbol{\varepsilon}_t \otimes \boldsymbol{\varepsilon}_s) = \mathbf{0}$ for all $t \neq s$, and $\mathbb{E}(x_{t,\ell}\boldsymbol{\varepsilon}_s) = \mathbf{0}$ for any $\ell \in [d]$ and (t,s).

3 Identification and estimation

3.1 The key idea

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

$$\mathbf{Y}_t\mathbf{b}^\ell = x_{t,\ell}\mathbf{a}_\ell + \boldsymbol{\varepsilon}_t\mathbf{b}^\ell \,, \quad \ell \in [d] \,. \tag{5}$$

When $\boldsymbol{\varepsilon}_t \equiv \mathbf{0}$, 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 (i) to eliminate the impact of non-zero $\boldsymbol{\varepsilon}_t$, (ii) to increase the estimation efficiency by augmenting the information over time t, and (iii) to overcome the intricacy in solving a generalized eigenequation with rank-reduced matrices, as the number of the eigenvalues of such an equation may be 0, finite or infinite; see Section 7.7 of Golub and Van Load (2013). To highlight the key idea of our new approach, we proceed within this subsection with the assumption that $\{\mathbf{Y}_t\}$ and $\{\boldsymbol{\varepsilon}_t\}$ are both weakly stationary. This stationarity condition will be relaxed hereafter when we formally present the identification result, the estimation method, and the associated asymptotic theory.

As $\{\boldsymbol{\varepsilon}_t\}$ is uncorrelated with $\{x_{t,k}\}$ (see Condition 1 above), it is easy to eliminate the impact of $\boldsymbol{\varepsilon}_t$. For example, let ξ_t be a linear combination of $\mathbf{Y}_t$, and let $\boldsymbol{\Sigma}_k = \mathrm{Cov}(\xi_{t-k}, \mathbf{Y}_t)$ for any $k \geq 1$. It follows from (5) that $\boldsymbol{\Sigma}_2\mathbf{b}^\ell = \lambda_\ell\boldsymbol{\Sigma}_1\mathbf{b}^\ell$ with $\lambda_\ell = \mathrm{Cov}(x_{t,\ell}, \xi_{t-2})/\mathrm{Cov}(x_{t,\ell}, \xi_{t-1})$. Hence the rows of $\mathbf{B}^+$ are the eigenvectors of the generalized eigenequation

$$\boldsymbol{\Sigma}_1^{\mathrm{T}}\boldsymbol{\Sigma}_2\mathbf{b} = \lambda\boldsymbol{\Sigma}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1\mathbf{b} \,. \tag{6}$$

With $\min(p,q) > d$, both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ have the ranks not greater than d, as $\mathbf{\Sigma}_k = \mathbf{A}\mathrm{Cov}(\xi_{t-k}, \mathbf{X}_t)\mathbf{B}^{\mathrm{T}}$; see (4). Thus the generalized eigenequation (6) can have infinite number of eigenvalues and the minimum gap between the eigenvalues can be 0, which makes the identification of those 'correct' eigenvectors, both theoretically and empirically, extremely difficult (if not impossible). To avoid this intricacy, below we project the q-dimensional genealized eigen-problem (6) into a d-dimensional one in which the matrices are full-ranked.

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

$$\mathbf{M}_1 = \sum_{k=1}^{K} \mathbf{\Sigma}_k \mathbf{\Sigma}_k^{\mathrm{T}} \quad \text{and} \quad \mathbf{M}_2 = \sum_{k=1}^{K} \mathbf{\Sigma}_k^{\mathrm{T}} \mathbf{\Sigma}_k \,. \tag{7}$$

In practice, both p and q are much greater than d. It is reasonable to assume $\mathrm{rank}(\mathbf{M}_1) = \mathrm{rank}(\mathbf{M}_2) = d$. Perform the spectral decomposition:

$$\mathbf{M}_1 = \mathbf{P}\mathbf{\Lambda}_1\mathbf{P}^{\mathrm{T}} \quad \text{and} \quad \mathbf{M}_2 = \mathbf{Q}\mathbf{\Lambda}_2\mathbf{Q}^{\mathrm{T}} \,, \tag{8}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, $p \times d$ and $q \times d$ matrices, the columns of $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, the d orthonormal eigenvectors corresponding to the d non-zero eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$, and $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are the diagonal matrices with the corresponding eigenvalues as the diagonal elements. Recall $\mathbf{\Sigma}_k = \mathbf{A}\mathbf{G}_k\mathbf{B}^{\mathrm{T}}$, where $\mathbf{G}_k = \mathrm{Cov}(\xi_{t-k}, \mathbf{X}_t)$ is a diagonal matrix; see (4). Then

$$\mathbf{M}_1 = \mathbf{A}\bigg(\sum_{k=1}^{K} \mathbf{G}_k\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{G}_k\bigg)\mathbf{A}^{\mathrm{T}} \quad \text{and} \quad \mathbf{M}_2 = \mathbf{B}\bigg(\sum_{k=1}^{K} \mathbf{G}_k\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{G}_k\bigg)\mathbf{B}^{\mathrm{T}} \,.$$

This, together with (8), implies that

$$\mathbf{A} = \mathbf{P}\mathbf{U} \quad \text{and} \quad \mathbf{B} = \mathbf{Q}\mathbf{V} \,, \tag{9}$$

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 $\mathbf{Z}_t = \mathbf{P}^{\mathrm{T}}\mathbf{Y}_t\mathbf{Q}$. It follows from (4) and (9) that

$$\mathbf{Z}_t = \mathbf{U}\mathbf{X}_t\mathbf{V}^{\mathrm{T}} + \mathbf{\Delta}_t = \sum_{\ell=1}^{d} x_{t,\ell}\mathbf{u}_\ell\mathbf{v}_\ell^{\mathrm{T}} + \mathbf{\Delta}_t \,, \tag{10}$$

where $\mathbf{\Delta}_t = \mathbf{P}^{\mathrm{T}}\boldsymbol{\varepsilon}_t\mathbf{Q}$ is uncorrelated with $\mathbf{X}_s$, and $\mathbf{u}_\ell$ and $\mathbf{v}_\ell$ are, respectively, the ℓ-th column of

$\mathbf{U}$ and $\mathbf{V}$. Choose η_t to be a linear combination of $\mathbf{Z}_t$ such that $\mathbf{W}_k \equiv \mathrm{Cov}(\eta_{t-k}, \mathbf{Z}_t)$ is full-ranked for $k = 1, 2$. This condition is equivalent to $\mathrm{Cov}(\eta_{t-k}, x_{t,\ell}) \neq 0$ for $k = 1, 2$ and $\ell \in [d]$, which can be fulfilled easily. Then the same argument towards (6) implies that the rows of the $d \times d$ inverse matrix $\mathbf{V}^{-1}$ are the eigenvectors of the generalized eigen-equation

$$\mathbf{W}_1^{\mathrm{T}} \mathbf{W}_2 \mathbf{v} = \lambda \mathbf{W}_1^{\mathrm{T}} \mathbf{W}_1 \mathbf{v} \,, \tag{11}$$

which has exactly d eigenvectors, corresponding to the d eigenvalues $\lambda_\ell = \mathrm{Cov}(x_{t,\ell}, \eta_{t-2}) / \mathrm{Cov}(x_{t,\ell}, \eta_{t-1})$ for $\ell \in [d]$. Furthermore those d eigenvectors are unique upto the scaling and reflection indeterminacy if those d eigenvalues are distinct. To compute $\mathbf{V}$ from $\mathbf{V}^{-1}$, the lengths of the rows of $\mathbf{V}^{-1}$ must be correctly specified, which, unfortunately, cannot be determined by (11). Note that (10) implies $\mathbf{W}_1(\mathbf{V}^{-1})^{\mathrm{T}} = \mathbf{U}\mathrm{Cov}(\eta_{t-1}, \mathbf{X}_t)$ with the diagonal matrix $\mathrm{Cov}(\eta_{t-1}, \mathbf{X}_t)$. Recall $|\mathbf{u}_\ell|_2 = 1$ for any $\ell \in [d]$. Hence the columns of $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ can be identified as follows:

$$\mathbf{u}_\ell = \frac{\mathbf{W}_1 \mathbf{v}^\ell}{|\mathbf{W}_1 \mathbf{v}^\ell|_2} \,, \quad \ell \in [d] \,,$$

where $\mathbf{v}^1, \ldots, \mathbf{v}^d$ are the d eigenvectors of (11). By the symmetry, the columns of $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ are obtained as follows:

$$\mathbf{v}_\ell = \frac{\mathbf{W}_1^{\mathrm{T}} \mathbf{u}^\ell}{|\mathbf{W}_1^{\mathrm{T}} \mathbf{u}^\ell|_2} \,, \quad \ell \in [d] \,,$$

where $\mathbf{u}^1, \ldots, \mathbf{u}^d$ are the d eigenvectors of the generalized eigen-equation $\mathbf{W}_1 \mathbf{W}_2^{\mathrm{T}} \mathbf{u} = \lambda \mathbf{W}_1 \mathbf{W}_1^{\mathrm{T}} \mathbf{u}$ which shares the same eigenvalues as (11). With $\mathbf{U}$ and $\mathbf{V}$ specified above, $\mathbf{A}$ and $\mathbf{B}$ can be determined by (9).

3.2 Identification of A and B

We summarize the finding in Section 3.1 in Proposition 1 below. The notation is more cumbersome as we present the result for a more general process without the stationarity condition. But the key idea remains the same.

Let ξ_t be a linear combination of $\mathbf{Y}_t$ (e.g. the first principal component of $\mathrm{vec}(\mathbf{Y}_t)$). For any $k \geq 1$, define

$$\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\}\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}] \,, \tag{12}$$

where $\bar{\mathbf{Y}} = n^{-1}\sum_{t=1}^n \mathbf{Y}_t$, and $\bar{\xi} = n^{-1}\sum_{t=1}^n \xi_t$. For some prescribed integer $K \geq 1$, let

$$\mathbf{M}_1 = \sum_{k=1}^K \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^{\mathrm{T}} \text{ and } \mathbf{M}_2 = \sum_{k=1}^K \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Y},\xi}(k),$$

which reduces to (7) with the the stationarity condition. Under Condition 2 below, perform the spectral decomposition as in (8), and define $\mathbf{Z}_t = \mathbf{P}^{\mathrm{T}}\mathbf{Y}_t\mathbf{Q}$. Let $\eta_t = \mathbf{e}^{\mathrm{T}}\mathrm{vec}(\mathbf{Z}_t)$ be a linear combination of $\mathbf{Z}_t$ for some constant vector $\mathbf{e} \in \mathbb{R}^{d^2}$. Put

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

where $\bar{\mathbf{Z}} = n^{-1}\sum_{t=1}^n \mathbf{Z}_t$, and $\bar{\eta} = n^{-1}\sum_{t=1}^n \eta_t$. Any $\mathbf{e} \in \mathbb{R}^{d^2}$ such that the associated $d \times d$ matrix $\{\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\}^{-1}\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(2)$ has d distinct eigenvalues is valid for the identification of $\mathbf{U}$ and $\mathbf{V}$. See Proposition 1 below. Write $\mathbf{\Theta} = \mathbf{I}_p \otimes \{(\mathbf{Q} \otimes \mathbf{P})\mathbf{e}\}$ and $\mathbf{\Sigma}_{\mathring{\mathbf{Y}}}(k) = (n-k)^{-1}\sum_{t=k+1}^n \mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\} \otimes \mathrm{vec}\{\mathbf{Y}_{t-k} - \mathbb{E}(\bar{\mathbf{Y}})\}]$. Then $\mathbf{\Sigma}_{\mathbf{Z},\eta}(k)$ defined above can be reformulated as

$$\mathbf{\Sigma}_{\mathbf{Z},\eta}(k) = \mathbf{P}^{\mathrm{T}}\mathbf{\Theta}^{\mathrm{T}}\mathbf{\Sigma}_{\mathring{\mathbf{Y}}}(k)\mathbf{Q}. \tag{13}$$

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

Proposition 1. Let Conditions 1 and 2 hold, and the eigenvalues of the $d \times d$ matrix

$$\{\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\}^{-1}\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(2)$$

be distinct. Then $\mathbf{A}$ and $\mathbf{B}$ are uniquely defined as in (9) upto the scaling 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, respectively, as

$$\mathbf{u}_\ell = \frac{\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v}^\ell}{|\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v}^\ell|_2} \text{ and } \mathbf{v}_\ell = \frac{\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{u}^\ell}{|\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{u}^\ell|_2},$$

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

$$\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(2)\mathbf{v} = \lambda\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{v} \text{ and } \mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{\Sigma}_{\mathbf{Z},\eta}(2)^{\mathrm{T}}\mathbf{u} = \lambda\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)\mathbf{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\mathbf{u}. \tag{14}$$

Furthermore those two generalized eigenequations share the same d eigenvalues.

Remark 1. (i) It is important that the orders of the eigenvectors $\{\mathbf{v}^\ell\}$ and $\{\mathbf{u}^\ell\}$ are in accordance with the order of the d shared eigenvalues of the two equations in (14) such that the ℓ-th columns of $\mathbf{A}$ and $\mathbf{B}$ obtained in (9) are paired together in model (4).

(ii) It is easy to see from (14) that $\mathbf{v}^\ell$ and $\mathbf{u}^\ell$ are, respectively, the eigenvectors of matrices $\{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2)$ and $\{\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\}^{-1}\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(2)^{\mathrm{T}}$.

3.3 Estimation of d and $(\mathbf{A}, \mathbf{B})$

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

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

for $k \geq 1$. When $p \cdot q$ diverges faster than $n^{1/2}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)$ is no longer a consistent estimator for $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)$ in (12) under $\|\cdot\|_2$. In the spirit of Bickel and Levina (2008), we define the threshold estimators for $\mathbf{M}_1$ and $\mathbf{M}_2$ as follows:

$$\widehat{\mathbf{M}}_1 = \sum_{k=1}^{K} T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\}T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)^{\mathrm{T}}\} \text{ and } \widehat{\mathbf{M}}_2 = \sum_{k=1}^{K} T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)^{\mathrm{T}}\}T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\}, \quad (15)$$

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 $p, q \gg n$. When $\delta_1 = 0$, $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ are defined directly based on the sample covariance matrices $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)$ without truncation, which is appropriate when, for example, p and q are fixed.

To estimate the rank d, let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ be the eigenvalues of $\widehat{\mathbf{M}}_1$. Following Chang et al. (2015), we define

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

where $R = \lfloor \alpha \min(p, q)\rfloor$ for some prescribed constants $\alpha \in (0, 1)$ and $c_n \to 0^+$. In practice, we may set $\alpha = 0.5$. The convergence rate of c_n will be specified in Theorem 1 and Remark 3 in Section 4. Note that the true eigenvalues of $\mathbf{M}_1$ satisfy the condition $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$ (see Condition 2). Adding a small constant $c_n > 0$ in (16) is to avoid the ratio "0/0". Theorem 1 shows that $\hat{d}$ is consistent, i.e. $\mathbb{P}(\hat{d} \neq d) \to 0$ as $n \to \infty$.

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 of $\widehat{\mathbf{M}}_2$ corresponding to its $\hat{d}$ largest eigenvalues. Define

$$\widehat{\mathbf{Z}}_t = \widehat{\mathbf{P}}^{\mathrm{T}}\mathbf{Y}_t\widehat{\mathbf{Q}} \text{ and } \hat{\eta}_t = \mathbf{e}^{\mathrm{T}}\mathrm{vec}(\widehat{\mathbf{Z}}_t) \quad (17)$$

for some constant vector $\mathbf{e} \in \mathbb{R}^{\hat{d}^2}$ with bounded ℓ^2-norm. Based on (13), we put

$$\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(k) = \widehat{\mathbf{P}}^{\mathrm{T}}\widehat{\mathbf{\Theta}}^{\mathrm{T}}T_{\delta_2}\{\widehat{\mathbf{\Sigma}}_{\check{\mathbf{Y}}}(k)\}\widehat{\mathbf{Q}}\,, \quad k = 1, 2\,, \tag{18}$$

where $\delta_2 \geq 0$ is the threshold level, $\widehat{\mathbf{\Theta}} = \mathbf{I}_p \otimes \{(\widehat{\mathbf{Q}} \otimes \widehat{\mathbf{P}})\mathbf{e}\}$, and

$$\widehat{\mathbf{\Sigma}}_{\check{\mathbf{Y}}}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n}(\mathbf{Y}_t - \bar{\mathbf{Y}}) \otimes \mathrm{vec}(\mathbf{Y}_{t-k} - \bar{\mathbf{Y}})\,. \tag{19}$$

Now the estimators for $\mathbf{A}$ and $\mathbf{B}$ are defined as

$$\widehat{\mathbf{A}} = \widehat{\mathbf{P}}\widehat{\mathbf{U}} \quad \text{and} \quad \widehat{\mathbf{B}} = \widehat{\mathbf{Q}}\widehat{\mathbf{V}}\,, \tag{20}$$

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

$$\hat{\mathbf{u}}_\ell = \frac{\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)\hat{\mathbf{v}}^\ell}{|\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)\hat{\mathbf{v}}^\ell|_2}\,, \quad \hat{\mathbf{v}}_\ell = \frac{\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\hat{\mathbf{u}}^\ell}{|\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\hat{\mathbf{u}}^\ell|_2}\,, \quad \ell \in [\hat{d}]\,.$$

In the above expression, $\hat{\mathbf{v}}^1, \ldots, \hat{\mathbf{v}}^{\hat{d}}$ are the $\hat{d}$ eigenvectors of the $\hat{d} \times \hat{d}$ matrix $\{\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)\}^{-1}$ $\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(2)$, and $\hat{\mathbf{u}}^1, \ldots, \hat{\mathbf{u}}^{\hat{d}}$ are those of $\{\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\mathbf{\Sigma}}_{\mathbf{Z},\eta}(2)^{\mathrm{T}}$. Furthermore, the two sets of eigenvectors are arranged such that the corresponding eigenvalues are in the descending order in module. See Remark 1 above.

By (4), we have $\mathrm{vec}(\mathbf{Y}_t) = (\mathbf{B} \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}_t) + \mathrm{vec}(\boldsymbol{\varepsilon}_t)$. Write $\mathbf{H} = \mathbf{B} \otimes \mathbf{A}$ and $\mathcal{S} = \{(k-1)(d+1)+1 : k \in [d]\}$. Denote by $\mathbf{H}_{\cdot,\mathcal{S}}$ the submatrix of $\mathbf{H}$ including the columns of $\mathbf{H}$ indexed by $\mathcal{S}$. Since $\mathbf{X}_t = \mathrm{diag}(x_{t,1}, \ldots, x_{t,d})$, by omitting the zero components in $\mathrm{vec}(\mathbf{X}_t)$, we know $\mathrm{vec}(\mathbf{Y}_t) = \mathbf{H}_{\cdot,\mathcal{S}}(x_{t,1}, \ldots, x_{t,d})^{\mathrm{T}} + \mathrm{vec}(\boldsymbol{\varepsilon}_t)$. Define $\widehat{\mathbf{H}} = \widehat{\mathbf{B}} \otimes \widehat{\mathbf{A}}$. We can recover $\mathbf{X}_t$ by $\widehat{\mathbf{X}}_t = \mathrm{diag}(\hat{x}_{t,1}, \ldots, \hat{x}_{t,\hat{d}})$ with $(\hat{x}_{t,1}, \ldots, \hat{x}_{t,\hat{d}})^{\mathrm{T}} = \widehat{\mathbf{H}}^{+}_{\cdot,\widehat{\mathcal{S}}}\mathrm{vec}(\mathbf{Y}_t)$ and $\widehat{\mathcal{S}} = \{(k-1)(\hat{d}+1)+1 : k \in [\hat{d}]\}$.

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

$$\alpha(k) = \sup_r \sup_{A \in \mathcal{F}_{-\infty}^r, B \in \mathcal{F}_{r+k}^{\infty}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|\,, \quad k \geq 1\,, \tag{21}$$

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 3 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 $\max_{k\in[K]} \|\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 \le C_1$. (ii) Write $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \{\sigma^{(k)}_{y,\xi,i,j}\}_{p\times q}$. It holds that $\max_{i\in[p]} \sum_{j=1}^{q} |\sigma^{(k)}_{y,\xi,i,j}|^{\iota} \le s_1$ and $\max_{j\in[q]} \sum_{i=1}^{p} |\sigma^{(k)}_{y,\xi,i,j}|^{\iota} \le s_2$ 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 $\max_{i\in[p]} \max_{j\in[q]} \max_{t\in[n]} \mathbb{P}(|y_{i,j,t}| > x) \le C_2 \exp(-C_3 x^{r_1})$ and $\max_{t\in[n]} \mathbb{P}(|\xi_t| > x) \le C_2 \exp(-C_3 x^{r_1})$ 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 (21) satisfy $\alpha(k) \le C_4 \exp(-C_5 k^{r_2})$ for all $k \ge 1$.

Recall $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)$ is a $p \times q$ matrix. Condition 3(i) requires the singular values of $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)$ to be uniformly bounded away from infinity for any $k \in [K]$, which is a mild condition. Condition 3(ii) imposes sparsity on $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)$. Notice that $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \mathbf{A}\mathbf{G}_k\mathbf{B}^{\mathrm{T}}$ 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. Causal ARMA processes with continuous innovation distributions are α-mixing with exponentially decaying α-mixing coefficients, so are stationary Markov chains satisfying certain conditions; see Section 2.6.1 of Fan and Yao (2003). Stationary GARCH models with finite second moments and continuous innovation distributions are also α-mixing with exponentially decaying α-mixing coefficients; see Proposition 12 of Carrasco and Chen (2002). Under certain conditions, VAR processes, multivariate ARCH processes, and multivariate GARCH processes are all α-mixing with exponentially decaying α-mixing coefficients; see Hafner and Preminger (2009), Boussama et al. (2011) and Wong et al. (2020). 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 work when p, q diverge at some polynomial rate of n. See Remark 3(ii) below. Let

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

Theorem 1 below shows that the ratio-based estimator $\hat{d}$ defined in (16) 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 (16) 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 $c > 0$ depending only on r_1 and r_2 specified in Condition 4.

To investigate the asymptotic properties of the estimator $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ in (20), 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 Remark 2 below.

Proposition 2. 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 $\|\widehat{\mathbf{P}}\mathbf{E}_1 - \mathbf{P}\|_2 = O_{\rm p}(\Pi_{1,n}) = \|\widehat{\mathbf{Q}}\mathbf{E}_2 - \mathbf{Q}\|_2$, provided that $\Pi_{1,n} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c > 0$ 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 non-zero 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 (non-zero) eigenvalues. Let $\tilde{\mathbf{e}} = (\mathbf{E}_2 \otimes \mathbf{E}_1)^{\rm T}\mathbf{e}$, with $\mathbf{e} \in \mathbb{R}^{d^2}$ involved in (17) for the definition of $\hat{\eta}_t = \mathbf{e}^{\rm T}{\rm vec}(\widehat{\mathbf{Z}}_t)$, and

$$\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k) = \mathbf{P}^{\rm T}\widetilde{\boldsymbol{\Theta}}^{\rm T}\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\mathbf{Q}\,, \quad k \geq 1\,,$$

where $\widetilde{\boldsymbol{\Theta}} = \mathbf{I}_p \otimes \{(\mathbf{Q} \otimes \mathbf{P})\tilde{\mathbf{e}}\}$, and $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)$ is specified in (13). As indicated in Lemma 2 in the Appendix, $\mathbf{E}_1^{\rm T}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(k)\mathbf{E}_2$ is consistent to $\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)$ under $\|\cdot\|_2$ rather than $\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(k)$ given in (13). In comparison to $\boldsymbol{\Sigma}_{\mathbf{Z},\eta}(k)$, we replace $\mathbf{e}$ by $\tilde{\mathbf{e}}$ in defining $\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)$. As we discussed above Condition 2, the selection of $\mathbf{e}$ for the identification of $\mathbf{U}$ and $\mathbf{V}$ is not unique. Define

$$\widetilde{\mathbf{S}}_1 = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\rm T}\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\,, \quad \widetilde{\mathbf{S}}_2 = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\rm T}\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(2)\,,$$
$$\widetilde{\mathbf{S}}_1^* = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\rm T}\,, \quad \widetilde{\mathbf{S}}_2^* = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(2)^{\rm T}\,.$$

Let $\tilde{\mu}_\ell = \tilde{c}_{2,\ell}\tilde{c}_{1,\ell}^{-1}$ with $\tilde{c}_{k,\ell} = (n-k)^{-1}\sum_{t=k+1}^n \tilde{\mathbf{e}}^{\rm T}\mathbb{E}[{\rm vec}\{\mathbf{Z}_{t-k} - \mathbb{E}(\bar{\mathbf{Z}})\}\{x_{t,\ell} - \mathbb{E}(\bar{x}_\ell)\}]$. Under Condition 5 below, Proposition 1 indicates 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

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

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

$$\widetilde{\mathbf{S}}_2\boldsymbol{\delta} = \tilde{\mu}_\ell \widetilde{\mathbf{S}}_1\boldsymbol{\delta} \quad \text{and} \quad \widetilde{\mathbf{S}}_2^*\boldsymbol{\delta} = \tilde{\mu}_\ell \widetilde{\mathbf{S}}_1^*\boldsymbol{\delta}\,. \tag{22}$$

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 $\max_{k\in\{1,2\}} \|\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\|_2 \le C_6$. (ii) Write $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k) = \{\sigma_{\mathring{y},r,s}^{(k)}\}_{p^2q\times q}$. It holds that $\max_{r\in[p^2q]}\sum_{s=1}^q |\sigma_{\mathring{y},r,s}^{(k)}|^\iota \le s_3$ and $\max_{s\in[q]}\sum_{r=1}^{p^2q} |\sigma_{\mathring{y},r,s}^{(k)}|^\iota \le s_4$ 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 (22) upto the scaling and permutation indeterminacy. Recall $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)$ is a $(p^2q)\times q$ matrix. Condition 6(i) requires the largest singular value of $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)$ is uniformly bounded away from infinity. Our technical proofs indeed allow $\max_{k\in\{1,2\}} \|\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\|_2$ to diverge with n. We impose Condition 6(i) just for simplifying the presentation. Condition 6(ii) imposes some sparsity requirement on $\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)$. Same as our discussion above for the validity of Condition 3(ii) imposed on the sparsity of $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)$, 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 (22) 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

$$\theta_\ell = \sigma_{\min}(\mathbf{R}_{\mathbf{v},\ell}^{\mathrm{T}}\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\mathbf{R}_{\mathbf{v},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}) \quad \text{and} \quad \theta_\ell^* = \sigma_{\min}\{\mathbf{R}_{\mathbf{u},\ell}^{\mathrm{T}}(\widetilde{\mathbf{S}}_1^*)^{-1}\widetilde{\mathbf{S}}_2^*\mathbf{R}_{\mathbf{u},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}\}\,, \tag{23}$$

the smallest singular values of $\mathbf{R}_{\mathbf{v},\ell}^{\mathrm{T}}\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\mathbf{R}_{\mathbf{v},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}$ and $\mathbf{R}_{\mathbf{u},\ell}^{\mathrm{T}}(\widetilde{\mathbf{S}}_1^*)^{-1}\widetilde{\mathbf{S}}_2^*\mathbf{R}_{\mathbf{u},\ell} - \tilde{\mu}_\ell\mathbf{I}_{d-1}$, 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 $\min_{j:\,j\ne\ell}|\tilde{\mu}_j - \tilde{\mu}_\ell|$. 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

$$\Pi_{2,n} = (s_3s_4)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}\,.$$

Theorem 2 indicates that the columns of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ defined in (20) are, respectively, consistent to those of $\mathbf{A}$ and $\mathbf{B}$ upto the scaling 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 $|\kappa_{1,\ell}\hat{\mathbf{a}}_{j_\ell} - \mathbf{a}_\ell|_2 = (1 + \theta_\ell^{-1}) \cdot O_{\rm p}(\Pi_{1,n} + \Pi_{2,n})$ and $|\kappa_{2,\ell}\hat{\mathbf{b}}_{j_\ell} - \mathbf{b}_\ell|_2 = \{1 + (\theta_\ell^*)^{-1}\} \cdot O_{\rm p}(\Pi_{1,n} + \Pi_{2,n})$ 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 $c > 0$ depending only on r_1 and r_2 specified in Condition 4. Here, the term $O_{\rm p}(\Pi_{1,n} + \Pi_{2,n})$ holds uniformly over $\ell \in [d]$.

For two $m \times r$ half orthogonal matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ satisfying $\mathbf{H}_1^{\rm T}\mathbf{H}_1 = \mathbf{H}_2^{\rm T}\mathbf{H}_2 = \mathbf{I}_r$, the distance between the two linear spaces $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$ is given by

$$\mathrm{dist}\{\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)\} = \sqrt{1 - \frac{1}{r}\mathrm{tr}(\mathbf{H}_1\mathbf{H}_1^{\rm T}\mathbf{H}_2\mathbf{H}_2^{\rm T})}\,.$$

For any unit ℓ^2-norm vectors $\hat{\mathbf{c}}$ and $\mathbf{c}$, we know $|\kappa\hat{\mathbf{c}} - \mathbf{c}|_2^2 \geq 1 - \mathrm{tr}(\hat{\mathbf{c}}\hat{\mathbf{c}}^{\rm T}\mathbf{c}\mathbf{c}^{\rm T})$ with any $\kappa \in \{-1, 1\}$. Hence, when $\hat{d} = d$, by Theorem 2, $\mathrm{dist}\{\mathcal{M}(\hat{\mathbf{a}}_{j_\ell}), \mathcal{M}(\mathbf{a}_\ell)\} = (1 + \theta_\ell^{-1}) \cdot O_{\rm p}(\Pi_{1,n} + \Pi_{2,n}) = o_{\rm p}(1)$ and $\mathrm{dist}\{\mathcal{M}(\hat{\mathbf{b}}_{j_\ell}), \mathcal{M}(\mathbf{b}_\ell)\} = \{1 + (\theta_\ell^*)^{-1}\} \cdot O_{\rm p}(\Pi_{1,n} + \Pi_{2,n}) = o_{\rm p}(1)$ for any $\ell \in [d]$.

Remark 2. For any matrices $\mathbf{H}_1 = \{\mathbf{h}_1^{(1)}, \ldots, \mathbf{h}_{d_1}^{(1)}\} \in \mathbb{R}^{m \times d_1}$ and $\mathbf{H}_2 = \{\mathbf{h}_1^{(2)}, \ldots, \mathbf{h}_{d_2}^{(2)}\} \in \mathbb{R}^{m \times d_2}$, we can measure the difference between $\mathbf{H}_1$ and $\mathbf{H}_2$ by

$$\rho(\mathbf{H}_1, \mathbf{H}_2) = \max_{\ell \in [d_1]} \min_{j \in [d_2]} \mathrm{dist}\{\mathcal{M}(\mathbf{h}_\ell^{(1)}), \mathcal{M}(\mathbf{h}_j^{(2)})\}\,.$$

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

Remark 3. (i) If p and q are fixed constants, we can select the threshold levels $\delta_1 = \delta_2 = 0$ in (15) and (18). 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 $\max_{i \in [p]}\max_{j \in [q]}\max_{t \in [n]}\mathbb{E}(|y_{i,j,t}|^{2\nu}) = O(1)$, $\max_{t \in [n]}\mathbb{E}(|\xi_t|^{2\nu}) = O(1)$, and $\sum_{k=1}^\infty\{\alpha(k)\}^{1-2/\nu} = O(1)$ for some constant $\nu > 2$. Under these conditions, using the Davydov inequality, we have

Theorem 1, Proposition 2 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 $\delta_1 = C_*(pq)^{1/\ell}n^{-1/2}$ and $\delta_2 = C_{**}(pq)^{2/\ell}n^{-1/2}$ in (15) and (18) for some sufficiently large constants $C_* > 0$ and $C_{**} > 0$, Theorem 1, Proposition 2 and Theorem 2 hold with $\Pi_{1,n}^* = (s_1s_2)^{1/2}\{(pq)^{1/\ell}n^{-1/2}\}^{1-\iota}$, $\Pi_{2,n}^* = (s_3s_4)^{1/2}\{(pq)^{2/\ell}n^{-1/2}\}^{1-\iota}$ 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 method by simulation based on model (2) or, equivalently, model (3). Recall $\mathbf{x}_\ell = (x_{1,\ell}, \ldots, x_{n,\ell})^{\mathrm{T}}$. In our simulation, the elements of the loading matrices $\mathbf{A}$ and $\mathbf{B}$ are drawn from the uniform distribution on $[-3, 3]$ independently with the restriction $\mathrm{rank}(\mathbf{A}) = d = \mathrm{rank}(\mathbf{B})$, and $\mathbf{x}_1, \ldots, \mathbf{x}_d$ are 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 $\boldsymbol{\mathcal{E}}$ in (2) are drawn from $\mathcal{N}(0,1)$ independently. We set $d \in \{1, 3, 6\}$, $n \in \{300, 600, 900\}$, and p, q taking values between 4 and 256. To specify ξ_t in our procedure, we perform the principal component analysis for the (pq)-dimensional time series $\mathrm{vec}(\mathbf{Y}_t)$, 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. Similarly, we select $\hat{\eta}_t$ in the same way with replacing $\mathbf{Y}_t$ by $\widehat{\mathbf{Z}}_t$. We only present the results for the cases with $p \geq q$. More numerical results can be found in the supplementary material.

First we estimate d by (16) with $c_n = 0$. Table 1 reports the relative frequency estimates of $\mathbb{P}(\hat{d} = d)$ with $K \in \{3, 5, 7\}$ and $\delta = 0$ in (15) based on 2000 repetitions. When $d = 1$, we observe $\hat{d} \equiv d$ in all the simulation replications. For $d > 1$, the relative frequency estimates of $\mathbb{P}(\hat{d} = d)$ increase as K, n, p and q increase in most of the cases. The results based on different selections of c_n are similar to those with $c_n = 0$. To evaluate the errors in estimating the columns of $\mathbf{A}$, we define the difference measure between $\mathbf{A}$ and $\widehat{\mathbf{A}}$:

$$L(\mathbf{A}, \widehat{\mathbf{A}}) = \begin{cases} \sqrt{1-(\mathbf{A}_*^{\mathrm{T}}\widehat{\mathbf{A}})^2}\,, & \text{if } \hat{d} = d = 1\,, \\ \dfrac{1}{2d(d-1)}\displaystyle\sum_{j=1}^{d}\bigg(\frac{\sum_{i=1}^d |r_{i,j}|}{\max_{i\in[d]}|r_{i,j}|} + \frac{\sum_{i=1}^d |r_{j,i}|}{\max_{i\in[d]}|r_{j,i}|} - 2\bigg)\,, & \text{if } \hat{d} = d \geq 2\,, \end{cases}$$

where $\mathbf{A}_*$ is the column-normalized $\mathbf{A}$ with each column being a unit vector, $r_{i,j}$ is the (i,j)-th element of $\mathbf{A}_*^+\widehat{\mathbf{A}}$. We take the convention that $(\max_{i\in[d]}|r_{i,j}|)^{-1}\sum_{i=1}^d |r_{i,j}| = d$ if $\max_{i\in[d]}|r_{i,j}| = 0$. Note that $L(\mathbf{A}, \widehat{\mathbf{A}})$ always takes value between 0 and 1. When $\hat{d} = d \geq 2$, $L(\mathbf{A}, \widehat{\mathbf{A}})$ is the divergence measure of Amari et al. (1996). Denote by $\mathbf{I}_d^*$ a $d \times d$ diagonal matrix with the main diagonal elements being either 1 or -1. When $\widehat{\mathbf{A}}$ is a 'perfect' estimate (i.e. free from errors), $\mathbf{A}_*^+\widehat{\mathbf{A}}$ will be a column-permutation of $\mathbf{I}_d^*$, and $L(\mathbf{A}, \widehat{\mathbf{A}}) = 0$. Analogously, we can also define the same measure between $\mathbf{B}$ and $\widehat{\mathbf{B}}$.

Our extensive simulation indicates the estimation errors for the columns of $\mathbf{A}$ and $\mathbf{B}$ with different $K \in \{3, 5, 7\}$ are almost the same. We only present the results for $\mathbf{A}$ with $K = 3$ in Figure 1, and report the results for $\mathbf{B}$ and for $\mathbf{A}$ with $K \in \{5, 7\}$ in the supplementary material. For each selected (p, q, d), the estimation error decreases as the sample size n increases. When $d > 1$, the estimation error decreases when p and/or q increases. However, such phenomenon is not true for $d = 1$ in general and can be only observed in the cases with $p = q$.

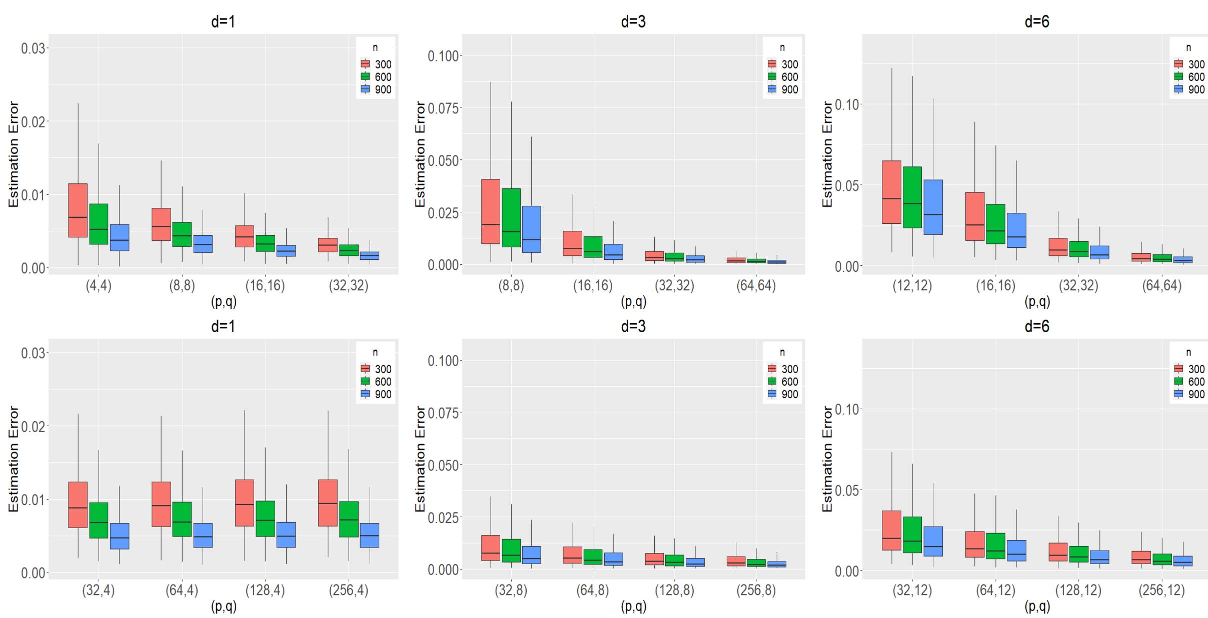

Figure 1: The boxplots of $L(\mathbf{A}, \widehat{\mathbf{A}})$ with $K = 3$ based on 2000 repetitions.

Table 1: *Relative frequency estimates of* $\mathbb{P}(\hat{d} = d)$ *based on* 2000 *repetitions.*

	n	(p, q)	$K = 3$	$K = 5$	$K = 7$	(p, q)	$K = 3$	$K = 5$	$K = 7$
$d = 1$	300	(4, 4)	100.00	100.00	100.00	(32, 4)	100.00	100.00	100.00
	600		100.00	100.00	100.00		100.00	100.00	100.00
	900		100.00	100.00	100.00		100.00	100.00	100.00
	300	(8, 8)	100.00	100.00	100.00	(64, 4)	100.00	100.00	100.00
	600		100.00	100.00	100.00		100.00	100.00	100.00
	900		100.00	100.00	100.00		100.00	100.00	100.00
	300	(16, 16)	100.00	100.00	100.00	(128, 4)	100.00	100.00	100.00
	600		100.00	100.00	100.00		100.00	100.00	100.00
	900		100.00	100.00	100.00		100.00	100.00	100.00
	300	(32, 32)	100.00	100.00	100.00	(256, 4)	100.00	100.00	100.00
	600		100.00	100.00	100.00		100.00	100.00	100.00
	900		100.00	100.00	100.00		100.00	100.00	100.00
$d = 3$	300	(8, 8)	78.50	79.65	79.35	(32, 8)	88.45	89.95	91.15
	600		81.45	81.10	80.90		91.35	92.80	93.20
	900		87.65	87.15	86.75		94.05	94.75	95.35
	300	(16, 16)	89.75	91.45	92.40	(64, 8)	88.60	91.55	92.80
	600		93.00	93.50	94.15		92.70	94.20	95.35
	900		94.40	94.30	94.90		96.15	96.80	97.10
	300	(32, 32)	94.55	96.55	96.95	(128, 8)	91.70	93.95	95.10
	600		94.70	95.90	96.80		95.20	96.15	96.75
	900		97.20	97.90	98.30		97.20	97.20	97.60
	300	(64, 64)	95.40	96.75	97.90	(256, 8)	92.35	94.35	95.60
	600		96.90	98.00	98.75		95.05	96.50	96.80
	900		98.15	98.50	98.95		97.50	98.15	98.20
$d = 6$	300	(12, 12)	73.35	78.15	81.85	(32, 12)	84.80	90.65	94.35
	600		76.15	80.00	82.45		89.15	93.45	95.05
	900		82.20	85.10	86.50		91.15	94.30	95.75
	300	(16, 16)	81.70	87.05	89.40	(64, 12)	88.60	93.80	96.15
	600		84.65	87.45	90.65		91.30	96.10	97.25
	900		87.30	90.50	93.20		93.35	96.15	97.40
	300	(32, 32)	91.15	95.10	96.55	(128, 12)	90.80	95.40	96.90
	600		92.75	96.05	97.60		92.45	95.95	98.20
	900		92.85	96.05	98.10		94.90	97.20	98.25
	300	(64, 64)	95.00	98.25	99.05	(256, 12)	90.30	95.25	97.65
	600		95.40	98.35	99.25		92.80	96.70	98.30
	900		96.40	98.40	99.45		95.70	97.70	98.95

5.2 A real data analysis

In this section, we analyze the monthly 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 (book to 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 the 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 i-th level of size and j-th 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 2 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 2 correspond to the ten levels of size and the columns correspond to the ten 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.

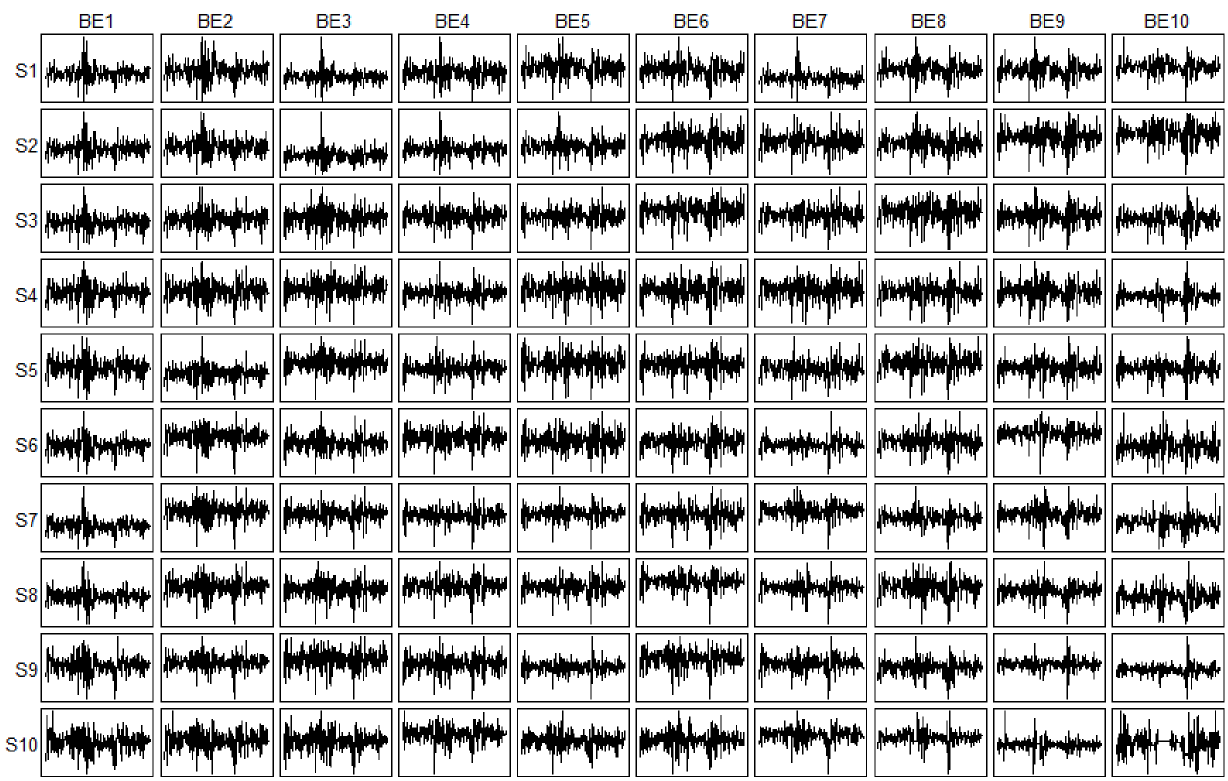

Figure 2: The plots of the return series of the portfolios formed on different levels of size (by rows) and book to equity ratio (by columns). The horizontal axis represents time and the vertical axis represents the monthly returns. The ranges of the vertical values are not the same.

We apply our model (4) to fit the standardized matrix time series $\{\mathbf{Y}_t\}_{t=1}^{336}$; leading to $\hat{d} \equiv 1$

with $K = 3, 5$ or 7. See (16). In the sequel, we only present the results with $K = 5$. The results based on $K \in \{3, 7\}$ are almost identical and thus omitted here. Based on (20), we obtain $\widehat{\mathbf{A}} = (0.46, 0.36, 0.33, 0.34, 0.29, 0.24, 0.31, 0.30, 0.26, 0.21)^{\mathrm{T}}$ and $\widehat{\mathbf{B}} = (0.27, 0.35, 0.29, 0.31, 0.34, 0.31, 0.30, 0.26, 0.35, 0.36)^{\mathrm{T}}$. Following the arguments in the end of Section 3.3, we can recover the latent time series $\{\hat{x}_{t,1}\}_{t=1}^{336}$. Figure 3 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 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 following methods:

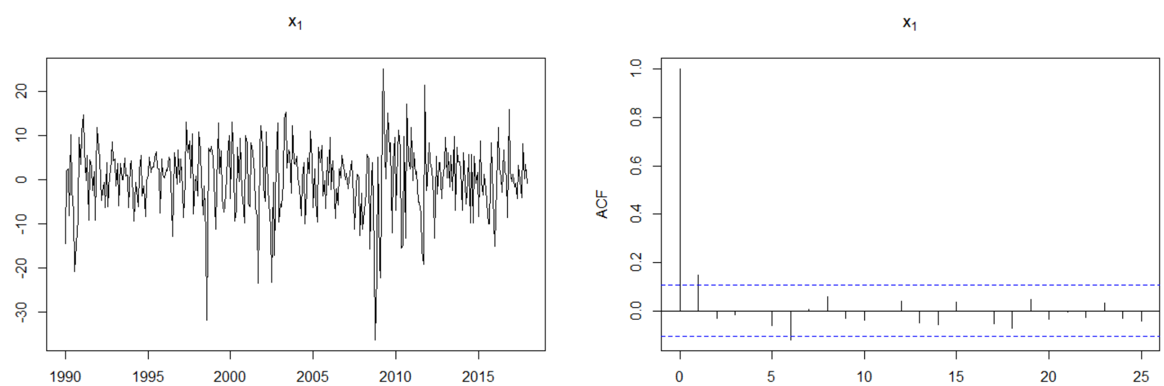

Figure 3: The plots of the latent time series $\{\hat{x}_{t,1}\}_{t=1}^{336}$ and its ACF.

- (UniARMA) For each of 100 component time series $\{y_{i,j,t}\}_{t=1}^{336}$, we fit an ARMA model specified by AIC; leading to the estimation for 135 coefficient parameters in the total 100 models.
- (SVAR) Fit a sparse VAR(ℓ) model to $\{\mathrm{vec}(\mathbf{Y}_t)\}_{t=1}^{336}$ using the R-function `sparseVAR` in the R-package `bigtime`. The AIC selects $\ell = 27$, and there are 270000 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 of Chang et al. (2018) to the 100 dimensional time series $\{\mathrm{vec}(\mathbf{Y}_t)\}_{t=1}^{336}$ 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

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	0.9914	1.0021	0.9923	0.9895	0.9985	**0.9613**
MAE	0.7433	0.7501	0.7436	0.7417	0.7443	**0.7235**
# Parameters	1	93	1	135	270000	200
time (seconds)	**0.2493**	6.3387	1.2064	7.0821	1723.5310	1.8343

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 $\{\mathbf{Y}_t\}_{t=1}^{336}$. 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.

While UniARMA, SVAR and MAR model $\mathbf{Y}_t$ or $\text{vec}(\mathbf{Y}_t)$ 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:

$$\text{RMSE} = \left\{ \frac{1}{33600} \sum_{t=1}^{336} \sum_{i=1}^{10} \sum_{j=1}^{10} (\hat{y}_{i,j,t} - y_{i,j,t})^2 \right\}^{1/2}, \quad \text{MAE} = \frac{1}{33600} \sum_{t=1}^{336} \sum_{i=1}^{10} \sum_{j=1}^{10} |\hat{y}_{i,j,t} - y_{i,j,t}| \,.$$

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 two 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 obtained 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 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 two-step ahead forecasts are obtained by plug-in the one-step forecasts into the models. The post-sample forecasting performance is evaluated by the rRMSE and rMAE defined as

$$\text{rRMSE} = \left[\frac{1}{2400} \sum_{s=1}^{24} \sum_{i=1}^{10} \sum_{j=1}^{10} \{\hat{y}_{i,j,312+s}^{(s)} - y_{i,j,312+s}\}^2 \right]^{1/2},$$

$$\text{rMAE} = \frac{1}{2400} \sum_{s=1}^{24} \sum_{i=1}^{10} \sum_{j=1}^{10} \left|\hat{y}_{i,j,312+s}^{(s)} - y_{i,j,312+s}\right| .$$

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

	Proposed	TS-PCA	FAC	UniARMA	SVAR	MAR
			$h = 1$			
rRMSE	**0.7678**	0.7795	0.7703	0.7724	0.7690	0.8067
rMAE	**0.5609**	0.5755	0.5643	0.5652	0.5614	0.5948
			$h = 2$			
rRMSE	0.7067	0.7163	0.7055	0.7083	0.7076	**0.7043**
rMAE	0.5179	0.5245	0.5162	0.5198	0.5180	**0.5144**

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 2nd smallest in the two-step ahead forecast for which only MAR has slightly smaller rRMSE and rMAE.

Appendix

Throughout the Appendix, we use $C \in (0, \infty)$ to denote a generic finite constant that does not depend on (n, p, q, d), and may be different in different uses. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if $\limsup_{n\to\infty} a_n/b_n \le c_0$ for some constant $c_0 > 0$.

A Proofs of Theorem 1 and Proposition 2

Recall $\Pi_{1,n} = (s_1 s_2)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}$. To construct Theorem 1 and Proposition 2, we need the following lemma whose proof is given in Section C.1.

Lemma 1. Under Conditions 1, 3 and 4, if the threshold level $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constant $C_* > 0$, we have $\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 = O_{\rm p}(\Pi_{1,n}) = \|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2$, provided that $\Pi_{1,n} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c > 0$ depending only on r_1 and r_2 specified in Condition 4.

A.1 Proof of Theorem 1

Denote by $\hat\lambda_1 \ge \cdots \ge \hat\lambda_p$ and $\lambda_1 \ge \cdots \ge \lambda_p$, respectively, the eigenvalues of $\widehat{\mathbf{M}}_1$ and $\mathbf{M}_1$. By Lemma 1, we have $\max_{i\in[p]} |\hat\lambda_i - \lambda_i| \le \|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 = o_{\rm p}(c_n)$ for some $c_n = o(1)$ satisfying $\Pi_{1,n} = o(c_n)$. By Condition 2, we know $\lambda_d \ge C$, and $\lambda_{d+1} = \cdots = \lambda_p = 0$. Hence, $(\hat\lambda_{d+1}+c_n)/(\hat\lambda_d+c_n) \xrightarrow{\mathbb{P}} 0$, $(\hat\lambda_{j+1}+c_n)/(\hat\lambda_j+c_n) \xrightarrow{\mathbb{P}} \lambda_{j+1}/\lambda_j > 0$ for any $j < d$, and $(\hat\lambda_{j+1}+c_n)/(\hat\lambda_j+c_n) \xrightarrow{\mathbb{P}} 1$ for any $j > d$, which implies $(\hat\lambda_{d+1}+c_n)/(\hat\lambda_d+c_n) = \min_{j\in[R]}(\hat\lambda_{j+1}+c_n)/(\hat\lambda_j+c_n)$ with probability approaching one. This indicates that $\mathbb{P}(\hat d = d) \to 1$ as $n \to \infty$. We complete the proof of Theorem 1. $\square$

A.2 Proof of Proposition 2

Denote by $\lambda_{1,1} \geq \cdots \geq \lambda_{1,p}$ and $\lambda_{2,1} \geq \cdots \geq \lambda_{2,q}$, respectively, the eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$. It follows from Condition 2 that $\lambda_{1,1} \geq \cdots \geq \lambda_{1,d} > 0 = \lambda_{1,d+1} = \cdots = \lambda_{1,p}$ and $\lambda_{2,1} \geq \cdots \geq \lambda_{2,d} > 0 = \lambda_{2,d+1} = \cdots = \lambda_{2,q}$. Notice that $\lambda_{1,d}$ and $\lambda_{2,d}$ are uniformly bounded away from zero. Lemma 1 of Chang et al. (2018) implies that $\|\widehat{\mathbf{P}}\mathbf{E}_1 - \mathbf{P}\|_2 \leq 8\lambda_{1,d}^{-1}\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 \leq C\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2$ and $\|\widehat{\mathbf{Q}}\mathbf{E}_2 - \mathbf{Q}\|_2 \leq 8\lambda_{2,d}^{-1}\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2 \leq C\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2$, where $\mathbf{E}_1$ and $\mathbf{E}_2$ are two orthogonal matrices. Together with Lemma 1, we complete the proof of Proposition 2. $\square$

B Proof of Theorem 2

Recall $\Pi_{1,n} = (s_1 s_2)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}$ and $\Pi_{2,n} = (s_3 s_4)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}$. Write $\Pi_n = \Pi_{1,n} + \Pi_{2,n}$. For $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(k)$ defined as (18), we write $\widehat{\mathbf{S}}_1 = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)$, $\widehat{\mathbf{S}}_2 = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(2)$, $\widehat{\mathbf{S}}_1^* = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}$ and $\widehat{\mathbf{S}}_2^* = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(2)^{\mathrm{T}}$. Recall $\widetilde{\mathbf{S}}_1 = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)$, $\widetilde{\mathbf{S}}_2 = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(2)$, $\widetilde{\mathbf{S}}_1^* = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\mathrm{T}}$ and $\widetilde{\mathbf{S}}_2^* = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(2)^{\mathrm{T}}$. To construct Theorem 2, we need the following lemmas. The proofs of Lemmas 2 and 3 are given in Sections C.2 and C.3, respectively. Lemma 4 is Corollary 7.2.6 of Golub and Van Load (2013).

Lemma 2. Under Conditions 1–4, and 6, if 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 $C_{**} > 0$, we have $\max_{k\in\{1,2\}} \|\mathbf{E}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(k)\mathbf{E}_2 - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)\|_2 = O_{\mathrm{p}}(\Pi_n)$ for $(\mathbf{E}_1, \mathbf{E}_2)$ specified in Proposition 2, provided that $\Pi_n = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c > 0$ depending only on r_1 and r_2 specified in Condition 4.

Lemma 3. Under conditions of Lemma 2, we have $\|\mathbf{E}_2^{\mathrm{T}}\widehat{\mathbf{S}}_1\mathbf{E}_2 - \widetilde{\mathbf{S}}_1\|_2 = O_{\mathrm{p}}(\Pi_n) = \|\mathbf{E}_2^{\mathrm{T}}\widehat{\mathbf{S}}_2\mathbf{E}_2 - \widetilde{\mathbf{S}}_2\|_2$ and $\|\mathbf{E}_1^{\mathrm{T}}\widehat{\mathbf{S}}_1^*\mathbf{E}_1 - \widetilde{\mathbf{S}}_1^*\|_2 = O_{\mathrm{p}}(\Pi_n) = \|\mathbf{E}_1^{\mathrm{T}}\widehat{\mathbf{S}}_2^*\mathbf{E}_1 - \widetilde{\mathbf{S}}_2^*\|_2$ for $(\mathbf{E}_1, \mathbf{E}_2)$ specified in Proposition 2.

Lemma 4. Suppose $\mathbf{W}, \boldsymbol{\Delta} \in \mathbb{C}^{d\times d}$ and that $\mathbf{R} = (\mathbf{r}_1, \mathbf{R}_2) \in \mathbb{C}^{d\times d}$ is unitary with $\mathbf{r}_1 \in \mathbb{C}^d$. Assume

$$\mathbf{R}^{\mathrm{H}}\mathbf{W}\mathbf{R} = \begin{pmatrix} \lambda & \mathbf{v}^{\mathrm{H}} \\ \mathbf{0} & \mathbf{D}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{R}^{\mathrm{H}}\boldsymbol{\Delta}\mathbf{R} = \begin{pmatrix} \epsilon & \boldsymbol{\gamma}^{\mathrm{H}} \\ \boldsymbol{\delta} & \boldsymbol{\Delta}_{22} \end{pmatrix},$$

where $^{\mathrm{H}}$ denotes the conjugate transpose. Let θ be the smallest singular value of $\mathbf{D}_{22} - \lambda\mathbf{I}_{d-1}$, and denote by $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm of $\cdot$. If $\theta > 0$ and $5\|\boldsymbol{\Delta}\|_{\mathrm{F}}(1 + 5\theta^{-1}|\mathbf{v}|_2) \leq \theta$, then there exists $\mathbf{u} \in \mathbb{C}^{d-1}$ with $|\mathbf{u}|_2 \leq 4\theta^{-1}|\boldsymbol{\delta}|_2$ such that $\tilde{\mathbf{r}}_1 = (\mathbf{r}_1 + \mathbf{R}_2\mathbf{u})/\sqrt{1 + \mathbf{u}^{\mathrm{H}}\mathbf{u}}$ is a unit ℓ^2-norm eigenvector for $\mathbf{W} + \boldsymbol{\Delta}$. Moreover, $1 - (\mathbf{r}_1^{\mathrm{H}}\tilde{\mathbf{r}}_1)^2 \leq 16\theta^{-2}|\boldsymbol{\delta}|_2^2$.

Now we begin to prove Theorem 2. Let $\{\hat{\boldsymbol{\beta}}_{\mathbf{v},1}, \ldots, \hat{\boldsymbol{\beta}}_{\mathbf{v},d}\}$ and $\{\boldsymbol{\beta}_{\mathbf{v},1}, \ldots, \boldsymbol{\beta}_{\mathbf{v},d}\}$ be, respectively, the eigenvectors of $\widehat{\mathbf{S}}_1^{-1}\widehat{\mathbf{S}}_2$ and $\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2$ with unit ℓ^2-norm, i.e., $\widehat{\mathbf{S}}_1^{-1}\widehat{\mathbf{S}}_2\hat{\boldsymbol{\beta}}_{\mathbf{v},\ell} = \hat{\mu}_\ell\hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}$ and $\widetilde{\mathbf{S}}_1^{-1}\widetilde{\mathbf{S}}_2\boldsymbol{\beta}_{\mathbf{v},\ell} =$

$\tilde{\mu}_\ell \boldsymbol{\beta}_{\mathbf{v},\ell}$ for any $\ell \in [d]$. Write $\hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell} = \mathbf{E}_2^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}$ with $\mathbf{E}_2$ specified in Proposition 2. Then $\hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell}$ is the eigenvector of $\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \widehat{\mathbf{S}}_2 \mathbf{E}_2$ associated with eigenvalue $\hat{\mu}_\ell$. Under Condition 5(i), applying Lemma 4 with $\mathbf{W} = \widetilde{\mathbf{S}}_1^{-1} \widetilde{\mathbf{S}}_2$, $\mathbf{W} + \boldsymbol{\Delta} = \mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \widehat{\mathbf{S}}_2 \mathbf{E}_2$ and $\mathbf{r}_1 = \boldsymbol{\beta}_{\mathbf{v},\ell}$, it holds that $|\kappa^\ell \hat{\boldsymbol{\beta}}^*_{\mathbf{v},j_\ell} - \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 \lesssim \theta_\ell^{-1} \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \widehat{\mathbf{S}}_2 \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1} \widetilde{\mathbf{S}}_2\|_2$ for any $\ell \in [d]$ provided that $5\|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \widehat{\mathbf{S}}_2 \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1} \widetilde{\mathbf{S}}_2\|_{\mathrm{F}}(1 + 5\theta_\ell^{-1} \|\widetilde{\mathbf{S}}_1^{-1} \widetilde{\mathbf{S}}_2\|_2) \le \theta_\ell$, where $\kappa^\ell \in \{-1, 1\}$, $(j_1, \ldots, j_d)$ is a permutation of $(1, \ldots, d)$, and θ_ℓ is given in (23). Without loss of generality, we assume $(j_1, \ldots, j_d) = (1, \ldots, d)$. For any $\ell \in [d]$, let

$$\Phi_\ell = \bigg| \frac{\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}}{|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2} - \frac{\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}}{|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2} \bigg|_2$$

with $\mathbf{E}_1$ specified in Proposition 2. In this sequel, we will specify the convergence rate of Φ_ℓ.

Recall $\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k) = \mathbf{P}^{\mathrm{T}} \widetilde{\boldsymbol{\Theta}}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k) \mathbf{Q}$, where $\widetilde{\boldsymbol{\Theta}} = \mathbf{I}_p \otimes \{(\mathbf{Q} \otimes \mathbf{P})(\mathbf{E}_2 \otimes \mathbf{E}_1)^{\mathrm{T}} \mathbf{e}\}$. Note that $\|\widetilde{\boldsymbol{\Theta}}\|_2 = |\mathbf{e}|_2 = O(1)$. By Condition 6(i), $\|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)\|_2 \le \|\mathbf{P}\|_2 \|\widetilde{\boldsymbol{\Theta}}\|_2 \|\mathbf{Q}\|_2 \|\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\|_2 \le C$ for any $k = 1, 2$, which implies $\|\widetilde{\mathbf{S}}_2\|_2 \le \|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2 \|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(2)\|_2 \le C$. Recall $\widetilde{\mathbf{S}}_1$ and $\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1 \mathbf{E}_2$ are two symmetric matrices. Denote by $\lambda_1 \ge \cdots \ge \lambda_d$ and $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_d$, respectively, the eigenvalues of $\widetilde{\mathbf{S}}_1$ and $\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1 \mathbf{E}_2$. By Condition 5(ii), we have $\lambda_d > 0$ is uniformly bounded away from zero. By Lemma 3, $|\hat{\lambda}_d - \lambda_d| \le \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1 \mathbf{E}_2 - \widetilde{\mathbf{S}}_1\|_2 = o_{\mathrm{p}}(1)$, which implies $\|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1}\|_2 \le \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \mathbf{E}_2\|_2 \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1 \mathbf{E}_2 - \widetilde{\mathbf{S}}_1\|_2 \|\widetilde{\mathbf{S}}_1^{-1}\|_2 = O_{\mathrm{p}}(\Pi_n)$. By Triangle inequality and Lemma 3, $\|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \widehat{\mathbf{S}}_2 \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1} \widetilde{\mathbf{S}}_2\|_2 \le \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1}\|_2 \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_2 \mathbf{E}_2 - \widetilde{\mathbf{S}}_2\|_2 + \|\widetilde{\mathbf{S}}_1^{-1}\|_2 \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_2 \mathbf{E}_2 - \widetilde{\mathbf{S}}_2\|_2 + \|\mathbf{E}_2^{\mathrm{T}} \widehat{\mathbf{S}}_1^{-1} \mathbf{E}_2 - \widetilde{\mathbf{S}}_1^{-1}\|_2 \|\widetilde{\mathbf{S}}_2\|_2 = O_{\mathrm{p}}(\Pi_n)$, which implies $|\kappa^\ell \hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell} - \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 = \theta_\ell^{-1} \cdot O_{\mathrm{p}}(\Pi_n)$ for any $\ell \in [d]$. Note that $\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell} = \kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \mathbf{E}_2 \hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell}$. By Lemma 2 and Triangle inequality, it holds that $|\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell} - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 \le \|\mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \mathbf{E}_2 - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2 |\kappa^\ell \hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell}|_2 + \|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2 |\kappa^\ell \hat{\boldsymbol{\beta}}^*_{\mathbf{v},\ell} - \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 = (1 + \theta_\ell^{-1}) \cdot O_{\mathrm{p}}(\Pi_n)$ for any $\ell \in [d]$. Then $|\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2 \ge |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 - o_{\mathrm{p}}(1)$. By Condition 5(ii), $|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2^2 = \boldsymbol{\beta}_{\mathbf{v},\ell}^{\mathrm{T}} \widetilde{\mathbf{S}}_1 \boldsymbol{\beta}_{\mathbf{v},\ell} \ge \lambda_d > C$, which implies $|\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2 \ge C$ with probability approaching one. Due to $|\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2 = |\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2$, by Triangle inequality, $||\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2^{-1} - |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2^{-1}| \le |\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2^{-1} |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2^{-1} |\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell} - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 = (1 + \theta_\ell^{-1}) \cdot O_{\mathrm{p}}(\Pi_n)$ for any $\ell \in [d]$. Hence, $\Phi_\ell \le |\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2^{-1} |\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell} - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 + ||\kappa^\ell \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2^{-1} - |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2^{-1}| |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2 = (1 + \theta_\ell^{-1}) \cdot O_{\mathrm{p}}(\Pi_n)$.

Write $\widehat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_d)$ and $\widetilde{\mathbf{P}} = \widehat{\mathbf{P}} \mathbf{E}_1$. For any $\ell \in [d]$, $\kappa^\ell \hat{\mathbf{a}}_\ell = \kappa^\ell \widehat{\mathbf{P}} \hat{\mathbf{u}}_\ell = |\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}|_2^{-1} \kappa^\ell \widetilde{\mathbf{P}} \mathbf{E}_1^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1) \hat{\boldsymbol{\beta}}_{\mathbf{v},\ell}$. Recall $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ with $\mathbf{a}_\ell = \mathbf{P} \mathbf{u}_\ell = |\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}|_2^{-1} \mathbf{P} \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1) \boldsymbol{\beta}_{\mathbf{v},\ell}$. By Proposition 2 and Triangle inequality, we have $|\kappa^\ell \hat{\mathbf{a}}_\ell - \mathbf{a}_\ell|_2 \le \|\widetilde{\mathbf{P}} - \mathbf{P}\|_2 + \Phi_\ell = (1 + \theta_\ell^{-1}) \cdot O_{\mathrm{p}}(\Pi_n)$ for any $\ell \in [d]$. Write $\widehat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_d)$. Analogously, we can also prove $|\kappa_*^\ell \hat{\mathbf{b}}_\ell - \mathbf{b}_\ell|_2 = \{1 + (\theta_\ell^*)^{-1}\} \cdot O_{\mathrm{p}}(\Pi_n)$ for any $\ell \in [d]$, where $\kappa_*^\ell \in \{-1, 1\}$. We complete the proof of Theorem 2. $\square$

C Proofs of auxiliary lemmas

C.1 Proof of Lemma 1

We first show that, with $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constant $C_* > 0$,

$$\|T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\} - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 = O_{\rm p}(\Pi_{1,n}) \tag{24}$$

for any $k \in [K]$, provided that $\log(pq) = o(n^c)$ for some constant $c > 0$.

To simplify the notation, we write $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k) = (\hat{\sigma}_{i,j}^{(k)})_{p\times q}$ and $\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = (\sigma_{i,j}^{(k)})_{p\times q}$. Write $n_k = n - k$ and $\bar{y}_{i,j} = n^{-1}\sum_{t=1}^n y_{i,j,t}$. Then we have

$$\begin{aligned}\hat{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)} &= \frac{1}{n_k}\sum_{t=k+1}^n \{y_{i,j,t}\xi_{t-k} - \mathbb{E}(y_{i,j,t}\xi_{t-k})\} - \bigg\{\frac{\bar{y}_{i,j}}{n_k}\sum_{t=k+1}^n \xi_{t-k} - \frac{\mathbb{E}(\bar{y}_{i,j})}{n_k}\sum_{t=k+1}^n \mathbb{E}(\xi_{t-k})\bigg\} \\ &\quad - \bigg\{\frac{\bar{\xi}}{n_k}\sum_{t=k+1}^n y_{i,j,t} - \frac{\mathbb{E}(\bar{\xi})}{n_k}\sum_{t=k+1}^n \mathbb{E}(y_{i,j,t})\bigg\} + \{\bar{y}_{i,j}\bar{\xi} - \mathbb{E}(\bar{y}_{i,j})\mathbb{E}(\bar{\xi})\} \\ &=: \mathrm{I}_{i,j}(1) + \mathrm{I}_{i,j}(2) + \mathrm{I}_{i,j}(3) + \mathrm{I}_{i,j}(4)\,.\end{aligned}$$

Under Condition 4(i), applying Lemma 2 of Chang et al. (2013), we have $\mathbb{P}\{|y_{i,j,t}\xi_{t-k} - \mathbb{E}(y_{i,j,t}\xi_{t-k})| > x\} \lesssim \exp(-Cx^{r_1/2})$ for any $x > 0$. Write $|\cdot|_+ = \max(\cdot, 0)$. Notice that $\{y_{i,j,t}\xi_{t-k} - \mathbb{E}(y_{i,j,t}\xi_{t-k})\}_{t=k+1}^n$ is an α-mixing sequence with α-mixing coefficients $\{\alpha(|m-k|_+)\}_{m\geq 1}$, where $\alpha(\cdot)$ is given in (21) in Section 4. Together with Condition 4(ii), Lemma 7 of Chang et al. (2021) implies $\mathbb{P}\{|\mathrm{I}_{i,j}(1)| \geq x\} \lesssim \exp(-Cnx^2) + \exp(-Cn^{\tilde{r}}x^{\tilde{r}})$ for any $x \in (0,1)$, where $\tilde{r}^{-1} = 1 + 2r_1^{-1} + r_2^{-1}$. Since $\mathbb{E}(y_{i,j,t}) = O(1)$ and $\mathbb{E}(\xi_t) = O(1)$ for any $t \in [n]$, applying Lemma 7 of Chang et al. (2021) again, $\mathbb{P}\{|\mathrm{I}_{i,j}(2) + \mathrm{I}_{i,j}(3) + \mathrm{I}_{i,j}(4)| \geq x\} \lesssim \exp(-Cnx^2) + \exp(-Cn^{\check{r}}x^{\check{r}})$ for any $x \in (0,1)$, where $\check{r}^{-1} = 2 + |r_1^{-1} - 1|_+ + r_2^{-1}$. Then $Z := \max_{i\in[p]}\max_{j\in[q]}|\hat{\sigma}_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| = O_{\rm p}[\{n^{-1}\log(pq)\}^{1/2}]$, provided that $\log(pq) = o(n^c)$ for some constant $c > 0$ depending only on r_1 and r_2.

By Triangle inequality, $\|T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\} - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 \leq \|T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 + \|T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\} - T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\}\|_2$, where $T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} = (\sigma_{i,j}^{(k)} I\{|\sigma_{i,j}^{(k)}| \geq \delta_1\})_{p\times q}$. On the one hand, $\|T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2^2 \leq [\max_{i\in[p]}\sum_{j=1}^q |\sigma_{i,j}^{(k)}| I\{|\sigma_{i,j}^{(k)}| < \delta_1\}][\max_{j\in[q]}\sum_{i=1}^p |\sigma_{i,j}^{(k)}| I\{|\sigma_{i,j}^{(k)}| < \delta_1\}]$. By Condition 3(ii), we have $\sum_{j=1}^q |\sigma_{i,j}^{(k)}| I\{|\sigma_{i,j}^{(k)}| < \delta_1\} \leq \delta_1^{1-\iota}s_1$ and $\sum_{i=1}^p |\sigma_{i,j}^{(k)}| I\{|\sigma_{i,j}^{(k)}| < \delta_1\} \leq \delta_1^{1-\iota}s_2$, which implies that $\|T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\} - \boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 \leq \delta_1^{1-\iota}(s_1s_2)^{1/2}$. On the other hand, we have

$$\begin{aligned}\|T_{\delta_1}\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y},\xi}(k)\} - T_{\delta_1}\{\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k)\}\|_2^2 &\leq \bigg[\max_{i\in[p]}\sum_{j=1}^q \big|\hat{\sigma}_{i,j}^{(k)} I\{|\hat{\sigma}_{i,j}^{(k)}| \geq \delta_1\} - \sigma_{i,j}^{(k)} I\{|\sigma_{i,j}^{(k)}| \geq \delta_1\}\big|\bigg] \\ &\quad \times \bigg[\max_{j\in[q]}\sum_{i=1}^p \big|\hat{\sigma}_{i,j}^{(k)} I\{|\hat{\sigma}_{i,j}^{(k)}| \geq \delta_1\} - \sigma_{i,j}^{(k)} I\{|\sigma_{i,j}^{(k)}| \geq \delta_1\}\big|\bigg]\end{aligned}$$

$$=: \mathrm{I}(1) \times \mathrm{I}(2)\,. \tag{25}$$

By Triangle inequality, $\mathrm{I}(1) \le \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| \ge \delta_1\} + \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| < \delta_1\} + \max_{i\in[p]} \sum_{j=1}^q |\sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| < \delta_1, |\sigma_{i,j}^{(k)}| \ge \delta_1\} =: \mathrm{I}(1,1) + \mathrm{I}(1,2) + \mathrm{I}(1,3)$. Recall $Z = \max_{i\in[p]} \max_{j\in[q]} |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}|$. By Condition 3(ii), $\mathrm{I}(1,1) \le Z \times \max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| \ge \delta_1\} \le Z\delta_1^{-\iota} s_1$. Applying Triangle inequality and Condition 3(ii) again, we have $\mathrm{I}(1,2) \le \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| < \delta_1\} + \max_{i\in[p]} \sum_{j=1}^q |\sigma_{i,j}^{(k)}| I\{|\sigma_{i,j}^{(k)}| < \delta_1\} \le \delta_1^{1-\iota} s_1 + \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| < \delta_1\}$. Taking $\theta \in (0,1)$, by Triangle inequality and Condition 3(ii), we have $\max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| < \delta_1\} \le \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{y,\xi,i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, |\sigma_{i,j}^{(k)}| \le \theta\delta_1\} + \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| \ge \delta_1, \theta\delta_1 < |\sigma_{i,j}^{(k)}| < \delta_1\} \le Z \times \max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\} + Z\theta^{-\iota}\delta_1^{-\iota} s_1$. Thus, $\mathrm{I}(1,2) \le Z \times \max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\} + Z\theta^{-\iota}\delta_1^{-\iota} s_1 + \delta_1^{1-\iota} s_1$. Meanwhile, by Triangle inequality and Condition 3(ii), we have $\mathrm{I}(1,3) \le \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| < \delta_1, |\sigma_{i,j}^{(k)}| \ge \delta_1\} + \max_{i\in[p]} \sum_{j=1}^q |\hat\sigma_{i,j}^{(k)}| I\{|\hat\sigma_{i,j}^{(k)}| < \delta_1, |\sigma_{i,j}^{(k)}| \ge \delta_1\} \le Z\delta_1^{-\iota} s_1 + \delta_1^{1-\iota} s_1$. Hence, $\mathrm{I}(1) \lesssim Z\delta_1^{-\iota} s_1 + \delta_1^{1-\iota} s_1 + Z\theta^{-\iota}\delta_1^{-\iota} s_1 + Z \times \max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\}$. Selecting $\delta_1 = C_*\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large $C_* > 0$, by Markov inequality, we have $\mathbb{P}[\max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\} > \lambda] \le \lambda^{-1} \sum_{i=1}^p \sum_{j=1}^q \mathbb{P}\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\} \le C\lambda^{-1}$ for any $\lambda > 1$, provided that $\log(pq) = o(n^c)$ for some constant $c > 0$ depending only on r_1 and r_2, which implies $\max_{i\in[p]} \sum_{j=1}^q I\{|\hat\sigma_{i,j}^{(k)} - \sigma_{i,j}^{(k)}| \ge (1-\theta)\delta_1\} = O_{\mathrm{p}}(1)$. Recall $Z = O_{\mathrm{p}}[\{n^{-1}\log(pq)\}^{1/2}]$. Therefore, $\mathrm{I}(1) = O_{\mathrm{p}}[s_1\{n^{-1}\log(pq)\}^{(1-\iota)/2}]$. Analogously, we also have $\mathrm{I}(2) = O_{\mathrm{p}}[s_2\{n^{-1}\log(pq)\}^{(1-\iota)/2}]$. By (25), we have $\|T_{\delta_1}\{\widehat{\boldsymbol\Sigma}_{\mathbf{Y},\xi}(k)\} - T_{\delta_1}\{\boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\}\|_2 = O_{\mathrm{p}}(\Pi_{1,n})$ for any $k \in [K]$. Together with $\|T_{\delta_1}\{\boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\} - \boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\|_2 \le \delta_1^{1-\iota}(s_1 s_2)^{1/2} = O(\Pi_{1,n})$, we complete the proof of (24).

Due to $\widehat{\mathbf{M}}_1 = \sum_{k=1}^K T_{\delta_1}\{\widehat{\boldsymbol\Sigma}_{\mathbf{Y},\xi}(k)\} T_{\delta_1}\{\widehat{\boldsymbol\Sigma}_{\mathbf{Y},\xi}(k)^{\mathrm{T}}\}$ and $\mathbf{M}_1 = \sum_{k=1}^K \boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\boldsymbol\Sigma_{\mathbf{Y},\xi}(k)^{\mathrm{T}}$, by Triangle inequality, (24) and Condition 3(i), $\|\widehat{\mathbf{M}}_1 - \mathbf{M}_1\|_2 \le \sum_{k=1}^K \|T_{\delta_1}\{\widehat{\boldsymbol\Sigma}_{\mathbf{Y},\xi}(k)\} - \boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\|_2^2 + 2\sum_{k=1}^K \|\boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\|_2 \|T_{\delta_1}\{\widehat{\boldsymbol\Sigma}_{\mathbf{Y},\xi}(k)\} - \boldsymbol\Sigma_{\mathbf{Y},\xi}(k)\|_2 = O_{\mathrm{p}}(\Pi_{1,n})$, provided that $\Pi_{1,n} = o(1)$ and $\log(pq) = o(n^c)$ for some constant $c > 0$ depending only on r_1 and r_2. Analogously, we also have $\|\widehat{\mathbf{M}}_2 - \mathbf{M}_2\|_2 = O_{\mathrm{p}}(\Pi_{1,n})$. □

C.2 Proof of Lemma 2

By the same arguments for (24), if $\delta_2 = C_{**}\{n^{-1}\log(pq)\}^{1/2}$ for some sufficiently large constant $C_{**} > 0$, we have $\|T_{\delta_2}\{\widehat{\boldsymbol\Sigma}_{\check{\mathbf{Y}}}(k)\} - \boldsymbol\Sigma_{\mathring{\mathbf{Y}}}(k)\|_2 = O_{\mathrm{p}}(\Pi_{2,n})$. Write $\widetilde{\mathbf{P}} = \widehat{\mathbf{P}}\mathbf{E}_1$ and $\widetilde{\mathbf{Q}} = \widehat{\mathbf{Q}}\mathbf{E}_2$ for $(\mathbf{E}_1, \mathbf{E}_2)$ specified in Proposition 2. Then $\mathbf{E}_1^{\mathrm{T}}\widehat{\boldsymbol\Sigma}_{\mathbf{Z},\eta}(k)\mathbf{E}_2 = \widetilde{\mathbf{P}}^{\mathrm{T}}\widehat{\boldsymbol\Theta}^{\mathrm{T}} T_{\delta_2}\{\widehat{\boldsymbol\Sigma}_{\check{\mathbf{Y}}}(k)\}\widetilde{\mathbf{Q}}$. Since $\|\widehat{\boldsymbol\Theta} - \widetilde{\boldsymbol\Theta}\|_2 = |\{\widehat{\mathbf{Q}} \otimes \widehat{\mathbf{P}} - (\mathbf{Q}\mathbf{E}_2^{\mathrm{T}}) \otimes (\mathbf{P}\mathbf{E}_1^{\mathrm{T}})\}\mathbf{e}|_2 \lesssim \|\widehat{\mathbf{Q}} \otimes \widehat{\mathbf{P}} - (\mathbf{Q}\mathbf{E}_2^{\mathrm{T}}) \otimes (\mathbf{P}\mathbf{E}_1^{\mathrm{T}})\|_2 \le \|\widehat{\mathbf{Q}}\mathbf{E}_2 - \mathbf{Q}\|_2 + \|\widehat{\mathbf{P}}\mathbf{E}_1 - \mathbf{P}\|_2$, by Proposition 2, we have $\|\widehat{\boldsymbol\Theta} - \widetilde{\boldsymbol\Theta}\|_2 = O_{\mathrm{p}}(\Pi_{1,n})$. Note that $\|\widehat{\boldsymbol\Theta}\|_2 = \|\widetilde{\boldsymbol\Theta}\|_2 = |\mathbf{e}|_2 = O(1)$ and

$\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k) = \mathbf{P}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\mathbf{Q}$. Together with Condition 6(i), we have $\max_{k\in\{1,2\}}\|\mathbf{E}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(k)\mathbf{E}_2 - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)\|_2 = O_{\mathrm{p}}(\Pi_n)$ with $\Pi_n = \Pi_{1,n} + \Pi_{2,n}$. □

C.3 Proof of Lemma 3

Recall $\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k) = \mathbf{P}^{\mathrm{T}}\widetilde{\boldsymbol{\Theta}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathring{\mathbf{Y}}}(k)\mathbf{Q}$ with $\widetilde{\boldsymbol{\Theta}} = \mathbf{I}_p \otimes \{(\mathbf{Q}\otimes\mathbf{P})(\mathbf{E}_2\otimes\mathbf{E}_1)^{\mathrm{T}}\mathbf{e}\}$, and $\|\widetilde{\boldsymbol{\Theta}}\|_2 = |\mathbf{e}|_2 = O(1)$. By Condition 6(i), we have $\max_{k\in\{1,2\}}\|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(k)\|_2 \le C$. Note that $\widetilde{\mathbf{S}}_1 = \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)$ and $\widehat{\mathbf{S}}_1 = \widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)$. By Triangle inequality and Lemma 2, $\|\mathbf{E}_2^{\mathrm{T}}\widehat{\mathbf{S}}_1\mathbf{E}_2 - \widetilde{\mathbf{S}}_1\|_2 \le \|\mathbf{E}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\mathbf{E}_2 - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2^2 + 2\|\boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2\|\mathbf{E}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Z},\eta}(1)\mathbf{E}_2 - \boldsymbol{\Sigma}_{\mathbf{Z},\tilde{\eta}}(1)\|_2 = O_{\mathrm{p}}(\Pi_n)$. Analogously, we can also construct other results. □

References

Amari, S., Cichochi, A. and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems, 8* edi. by D.S. Touretzky, M.C. Mozer and M.E. Hasselma. MIT Press.

Anandkumar, A., Ge, R., Hsu, D. and Janzamin, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. `arXiv:1402.5180`.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577–2604.

Boussama, F., Fuchs, F. and Stelzer, R. (2011). Stationary and geometric ergodicity of BEKK multivariate GARCH models. *Stochastic Processes and Their Applications*, **121**, 2331–2360.

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

Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, **18**, 17–39.

Chang, J., Chen, X. and Wu, M. (2021). Central limit theorems for high dimentional dependent data. `arXiv:2104.12929`.

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

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

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, **41**, 2123–2148.

Chen, E. Y. and Chen, R. (2019). Modeling dynamic transport network with matrix factor models: with an application to international trade flow. `arXiv:1901.00769`.

Chen, E. Y., Tsay, R. S. and Chen, R. (2020). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, **115**, 775–793.

Chen, R., Xiao, H. and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, **222**, 539–560.

Chen, R., Yang, D. and Zhang, C.-H. (2021). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, to appear. DOI: 10.1080/01621459.2021.1912757.

Domanov, I. and De Lathauwer, L. (2014). Canonical polyadic decomposition of third-order tensors: reduction to generalized eigenvalue decomposition. *SIAM Journal of Matrix Analysis and Applications*, **35**, 636–660.

Dunlavy, D. M., Kolda, T. G. and Acar, E. (2011). Temporal link prediction using matrix and tensor factorizations. *ACM Trans. Knowledge Discovery from Data*, **5**-2, Artile 10.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Golub, G. H. and Van Load, C. F. (2013). *Matrix Computations* (4th Edition). Johns Hopkins University Press, Baltimore.

Hafner, C. M. and Preminger, A. (2009). On asymptotic theory for multivariate GARCH models. *Journal of Multivariate Analysis*, **100**, 2044–2054.

Han, Y., Chen, R., Zhang, C.-H. and Yao, Q. (2021a). Simultaneous decorrelation of matrix time series. `arXiv:2103.09411`.

Han, Y. and Zhang, C.-H. (2021). Tensor principal component analysis in high dimensional CP models. `arXiv:2108.04428`.

Han, Y., Zhang, C.-H. and Chen, R. (2021b). CP factor model for dynamic tensors. `arXiv:2110.15517`.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, **51**, 455–500.

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

Linnemann, H. (1966). *An Econometric Study of International Trade Flows*, **234**. North-Holland Publishing Company Amsterdam.

Liu, Y., Shang, F., Fan, W., Cheng, J. and Cheng, H. (2014). Generalized higher-order orthogonal iteration for tensor decomposition and completion. In *Advances in Neural Information Processing Systems*, 1763–1771.

Sanchez, E. and Kowalski, B. R. (1990). Tensorial resolution: a direct trilinear decomposition. *Journal of Chemometrics*, **4**, 29–45.

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

Sun, W. W., Lu, J., Liu, H. Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society* **B**, **79**, 899–916.

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

Wang, M. and Song, Y. (2017). Tensor decompositions via two-mode higher-order svd (hosvd). *International Conference on Artificial Intelligence and Statistics, PMLR* **54**, 614–622.

Wong, K. C., Li, Z. and Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, **48**, 1124–1142.

Zhang, A. and Xia, D. (2018). Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, **64**, 7311–7338.

Supplementary Material for "Modelling Matrix Time Series via a Tensor CP-Decomposition" by Jinyuan Chang, Jing He, Lin Yang and Qiwei Yao

Table 4: *The averages and standard deviations (in parentheses) of the estimation errors for* $\widehat{\mathbf{A}}$ *and* $\widehat{\mathbf{B}}$ *when* $d = 1$ *based on* 2000 *repetitions. All numbers in the table are the true numbers multiplied by* 100 *for ease of presentation.*

		$K = 3$		$K = 5$		$K = 7$	
(p, q)	n	$L(\mathbf{A}, \widehat{\mathbf{A}})$	$L(\mathbf{B}, \widehat{\mathbf{B}})$	$L(\mathbf{A}, \widehat{\mathbf{A}})$	$L(\mathbf{B}, \widehat{\mathbf{B}})$	$L(\mathbf{A}, \widehat{\mathbf{A}})$	$L(\mathbf{B}, \widehat{\mathbf{B}})$
(4, 4)	300	0.91(0.82)	0.91(0.84)	0.96(0.83)	0.95(0.85)	0.98(0.84)	0.97(0.87)
	600	0.69(0.65)	0.70(0.66)	0.73(0.67)	0.73(0.67)	0.75(0.68)	0.75(0.67)
	900	0.48(0.45)	0.48(0.42)	0.51(0.47)	0.50(0.43)	0.52(0.47)	0.52(0.44)
(8, 8)	300	0.63(0.35)	0.63(0.36)	0.66(0.36)	0.66(0.37)	0.67(0.36)	0.67(0.37)
	600	0.49(0.27)	0.48(0.27)	0.51(0.28)	0.51(0.28)	0.52(0.28)	0.52(0.28)
	900	0.35(0.19)	0.34(0.18)	0.36(0.20)	0.36(0.19)	0.37(0.20)	0.37(0.19)
(16, 16)	300	0.45(0.22)	0.45(0.21)	0.47(0.22)	0.47(0.22)	0.48(0.22)	0.48(0.22)
	600	0.34(0.16)	0.34(0.16)	0.36(0.16)	0.36(0.16)	0.37(0.16)	0.37(0.16)
	900	0.24(0.11)	0.24(0.11)	0.25(0.11)	0.25(0.11)	0.26(0.11)	0.26(0.11)
(32, 32)	300	0.32(0.14)	0.32(0.14)	0.34(0.14)	0.34(0.14)	0.34(0.14)	0.34(0.14)
	600	0.24(0.10)	0.24(0.10)	0.25(0.11)	0.26(0.11)	0.26(0.11)	0.26(0.11)
	900	0.17(0.07)	0.17(0.07)	0.18(0.07)	0.18(0.08)	0.18(0.07)	0.19(0.08)
(32, 4)	300	1.02(0.72)	0.30(0.25)	1.07(0.73)	0.31(0.25)	1.09(0.72)	0.31(0.25)
	600	0.78(0.52)	0.22(0.19)	0.82(0.52)	0.24(0.19)	0.83(0.52)	0.24(0.19)
	900	0.54(0.36)	0.15(0.12)	0.57(0.37)	0.16(0.13)	0.58(0.37)	0.17(0.13)
(64, 4)	300	1.02(0.75)	0.21(0.18)	1.07(0.76)	0.22(0.19)	1.09(0.75)	0.22(0.18)
	600	0.78(0.55)	0.16(0.15)	0.82(0.55)	0.17(0.16)	0.84(0.55)	0.17(0.16)
	900	0.55(0.42)	0.11(0.09)	0.58(0.42)	0.11(0.09)	0.59(0.42)	0.12(0.09)
(128, 4)	300	1.03(0.65)	0.15(0.12)	1.08(0.66)	0.16(0.13)	1.10(0.66)	0.16(0.12)
	600	0.79(0.49)	0.11(0.08)	0.83(0.50)	0.12(0.08)	0.85(0.50)	0.12(0.09)
	900	0.55(0.34)	0.08(0.06)	0.58(0.35)	0.08(0.07)	0.60(0.35)	0.08(0.07)
(256, 4)	300	1.04(0.62)	0.10(0.07)	1.08(0.63)	0.11(0.08)	1.11(0.63)	0.11(0.08)
	600	0.79(0.48)	0.08(0.06)	0.83(0.49)	0.08(0.06)	0.85(0.49)	0.08(0.06)
	900	0.55(0.33)	0.05(0.04)	0.58(0.34)	0.06(0.04)	0.59(0.34)	0.06(0.04)
(4, 32)	300	0.30(0.27)	1.00(0.69)	0.31(0.27)	1.05(0.70)	0.32(0.28)	1.08(0.71)
	600	0.22(0.19)	0.77(0.51)	0.23(0.19)	0.80(0.51)	0.24(0.20)	0.82(0.52)
	900	0.16(0.13)	0.54(0.36)	0.17(0.14)	0.57(0.38)	0.17(0.14)	0.58(0.37)
(4, 64)	300	0.21(0.21)	1.03(0.79)	0.22(0.22)	1.08(0.80)	0.22(0.21)	1.10(0.80)
	600	0.16(0.15)	0.79(0.56)	0.16(0.15)	0.82(0.57)	0.17(0.15)	0.84(0.58)
	900	0.11(0.10)	0.55(0.42)	0.12(0.10)	0.58(0.42)	0.12(0.11)	0.60(0.43)
(4, 128)	300	0.14(0.11)	1.01(0.58)	0.15(0.12)	1.06(0.59)	0.15(0.12)	1.09(0.59)
	600	0.11(0.08)	0.78(0.42)	0.12(0.08)	0.82(0.43)	0.12(0.08)	0.84(0.43)
	900	0.08(0.05)	0.54(0.29)	0.08(0.05)	0.57(0.30)	0.08(0.06)	0.59(0.30)
(4, 256)	300	0.10(0.09)	1.03(0.75)	0.11(0.09)	1.08(0.76)	0.11(0.09)	1.10(0.75)
	600	0.08(0.08)	0.79(0.61)	0.08(0.08)	0.83(0.61)	0.09(0.08)	0.85(0.61)
	900	0.06(0.05)	0.55(0.40)	0.06(0.05)	0.58(0.41)	0.06(0.06)	0.59(0.41)

Table 5: *The averages and standard deviations (in parentheses) of the estimation errors for* $\widehat{\mathbf{A}}$ *and* $\widehat{\mathbf{B}}$ *when* $d = 3$ *based on* 2000 *repetitions. All numbers in the table are the true numbers multiplied by* 100 *for ease of presentation.*

		$K=3$		$K=5$		$K=7$	
(p,q)	n	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$
(8, 8)	300	4.02(5.74)	4.03(5.61)	4.12(5.84)	4.11(5.62)	4.25(6.07)	4.18(5.71)
	600	3.58(5.44)	3.61(5.48)	3.37(4.98)	3.44(5.10)	3.33(4.86)	3.45(5.06)
	900	2.96(5.04)	2.94(5.00)	2.98(5.17)	2.98(5.09)	2.89(4.94)	2.92(5.01)
(16, 16)	300	1.77(3.36)	1.82(3.48)	1.71(3.31)	1.77(3.44)	1.67(3.08)	1.72(3.21)
	600	1.59(3.43)	1.56(3.33)	1.62(3.43)	1.58(3.33)	1.60(3.40)	1.56(3.27)
	900	1.26(3.00)	1.26(2.85)	1.25(2.95)	1.23(2.75)	1.22(2.82)	1.20(2.66)
(32, 32)	300	0.82(2.15)	0.83(2.26)	0.86(2.25)	0.86(2.41)	0.88(2.33)	0.88(2.43)
	600	0.79(2.46)	0.77(2.32)	0.80(2.50)	0.80(2.43)	0.83(2.60)	0.82(2.49)
	900	0.63(2.13)	0.60(1.96)	0.65(2.24)	0.63(2.17)	0.65(2.18)	0.62(2.06)
(64, 64)	300	0.44(1.34)	0.43(1.32)	0.44(1.37)	0.43(1.36)	0.48(1.53)	0.47(1.51)
	600	0.45(1.97)	0.45(2.11)	0.46(2.06)	0.45(2.17)	0.47(2.09)	0.47(2.22)
	900	0.32(1.39)	0.32(1.29)	0.35(1.60)	0.35(1.46)	0.32(1.52)	0.32(1.39)
(32, 8)	300	1.89(3.61)	1.84(3.52)	1.94(3.69)	1.90(3.62)	1.91(3.63)	1.87(3.53)
	600	1.80(3.79)	1.82(3.89)	1.81(3.80)	1.81(3.85)	1.83(3.82)	1.81(3.79)
	900	1.27(2.66)	1.25(2.41)	1.34(2.88)	1.31(2.64)	1.38(2.94)	1.34(2.69)
(64, 8)	300	1.36(3.07)	1.34(2.90)	1.42(3.28)	1.43(3.24)	1.42(3.11)	1.41(3.01)
	600	1.12(2.59)	1.14(2.67)	1.10(2.54)	1.12(2.65)	1.10(2.55)	1.12(2.67)
	900	0.97(2.52)	0.99(2.46)	0.98(2.64)	0.98(2.51)	0.97(2.58)	0.96(2.42)
(128, 8)	300	1.06(2.90)	1.06(2.87)	1.06(2.90)	1.06(2.90)	1.10(3.02)	1.10(3.03)
	600	1.00(2.75)	1.00(2.70)	0.98(2.69)	0.98(2.68)	1.01(2.84)	1.01(2.76)
	900	0.67(1.86)	0.66(1.83)	0.69(1.98)	0.69(2.00)	0.66(1.76)	0.66(1.82)
(256, 8)	300	0.78(2.13)	0.78(2.11)	0.80(2.28)	0.81(2.30)	0.81(2.26)	0.82(2.29)
	600	0.64(1.89)	0.65(1.90)	0.66(1.99)	0.65(1.95)	0.71(2.17)	0.69(2.09)
	900	0.49(1.41)	0.50(1.49)	0.49(1.36)	0.49(1.37)	0.47(1.34)	0.48(1.36)
(8, 32)	300	1.90(3.51)	1.87(3.52)	1.87(3.39)	1.83(3.36)	1.97(3.56)	1.91(3.48)
	600	1.77(3.51)	1.76(3.53)	1.82(3.76)	1.81(3.70)	1.76(3.54)	1.76(3.50)
	900	1.43(3.42)	1.40(3.25)	1.46(3.45)	1.47(3.47)	1.48(3.54)	1.49(3.54)
(8, 64)	300	1.37(3.00)	1.39(2.95)	1.35(3.01)	1.33(2.91)	1.40(3.06)	1.39(2.97)
	600	1.19(2.78)	1.19(2.74)	1.20(2.79)	1.20(2.78)	1.20(2.87)	1.22(2.91)
	900	1.00(2.64)	1.01(2.62)	1.02(2.75)	1.02(2.69)	0.99(2.66)	1.00(2.65)
(8, 128)	300	1.09(2.93)	1.07(2.89)	1.05(2.73)	1.03(2.67)	1.08(2.76)	1.05(2.69)
	600	0.83(2.23)	0.83(2.32)	0.82(2.09)	0.84(2.21)	0.85(2.20)	0.87(2.27)
	900	0.69(2.06)	0.68(1.98)	0.69(2.12)	0.69(2.04)	0.73(2.33)	0.73(2.28)
(8, 256)	300	0.74(2.01)	0.77(2.06)	0.74(1.97)	0.76(2.05)	0.74(1.95)	0.76(2.00)
	600	0.74(2.28)	0.74(2.48)	0.71(2.22)	0.72(2.38)	0.73(2.32)	0.74(2.44)
	900	0.49(1.72)	0.49(1.68)	0.51(1.76)	0.50(1.73)	0.51(1.77)	0.51(1.75)

Table 6: *The averages and standard deviations (in parentheses) of the estimation errors for $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ when $d = 6$ based on 2000 repetitions. All numbers in the table are the true numbers multiplied by 100 for ease of presentation.*

		$K=3$		$K=5$		$K=7$	
(p,q)	n	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$	$L(\mathbf{A},\widehat{\mathbf{A}})$	$L(\mathbf{B},\widehat{\mathbf{B}})$
(12, 12)	300	5.00(3.29)	5.02(3.26)	5.03(3.30)	5.00(3.21)	5.00(3.35)	4.99(3.30)
	600	4.64(3.14)	4.65(3.15)	4.61(3.06)	4.64(3.10)	4.60(3.11)	4.59(3.10)
	900	4.00(2.79)	4.02(2.84)	3.91(2.77)	3.92(2.85)	3.99(2.85)	3.97(2.88)
(16, 16)	300	3.32(2.38)	3.33(2.41)	3.34(2.48)	3.33(2.47)	3.37(2.57)	3.39(2.56)
	600	2.96(2.36)	2.97(2.36)	2.96(2.39)	2.96(2.38)	2.97(2.29)	2.96(2.28)
	900	2.56(2.12)	2.56(2.13)	2.60(2.27)	2.58(2.21)	2.63(2.20)	2.60(2.19)
(32, 32)	300	1.55(1.60)	1.55(1.67)	1.56(1.59)	1.54(1.59)	1.59(1.58)	1.57(1.60)
	600	1.37(1.50)	1.36(1.47)	1.37(1.51)	1.36(1.49)	1.39(1.52)	1.39(1.52)
	900	1.10(1.22)	1.09(1.21)	1.16(1.30)	1.17(1.31)	1.17(1.34)	1.15(1.31)
(64, 64)	300	0.76(0.99)	0.76(0.98)	0.76(0.98)	0.76(0.95)	0.76(0.97)	0.76(0.97)
	600	0.67(0.89)	0.66(0.89)	0.66(0.90)	0.66(0.88)	0.67(0.87)	0.67(0.89)
	900	0.55(0.82)	0.56(0.85)	0.57(0.86)	0.57(0.86)	0.57(0.84)	0.58(0.85)
(32, 12)	300	2.81(2.27)	2.82(2.23)	2.79(2.21)	2.80(2.22)	2.93(2.33)	2.94(2.33)
	600	2.54(2.09)	2.54(2.10)	2.52(2.18)	2.51(2.16)	2.52(2.13)	2.53(2.16)
	900	2.19(2.01)	2.15(1.93)	2.18(1.99)	2.15(1.96)	2.21(2.04)	2.20(2.03)
(64, 12)	300	2.02(1.94)	2.03(1.92)	1.98(1.76)	1.96(1.75)	1.99(1.79)	1.98(1.77)
	600	1.87(1.85)	1.89(1.89)	1.84(1.82)	1.85(1.81)	1.82(1.75)	1.83(1.74)
	900	1.56(1.54)	1.54(1.52)	1.49(1.48)	1.49(1.50)	1.53(1.59)	1.53(1.59)
(128, 12)	300	1.51(1.57)	1.51(1.57)	1.48(1.47)	1.49(1.50)	1.48(1.50)	1.48(1.55)
	600	1.34(1.44)	1.34(1.42)	1.35(1.52)	1.33(1.44)	1.36(1.47)	1.37(1.48)
	900	1.14(1.35)	1.15(1.36)	1.08(1.23)	1.08(1.25)	1.11(1.33)	1.12(1.31)
(256, 12)	300	1.10(1.23)	1.11(1.26)	1.09(1.27)	1.10(1.29)	1.09(1.29)	1.09(1.29)
	600	0.97(1.17)	0.99(1.19)	0.96(1.11)	0.95(1.13)	0.96(1.14)	0.97(1.18)
	900	0.84(1.09)	0.85(1.10)	0.86(1.14)	0.86(1.14)	0.84(1.13)	0.84(1.12)
(12, 32)	300	2.72(2.24)	2.72(2.22)	2.74(2.20)	2.74(2.19)	2.79(2.21)	2.81(2.22)
	600	2.51(2.08)	2.52(2.09)	2.55(2.12)	2.54(2.12)	2.52(2.10)	2.52(2.07)
	900	2.23(1.99)	2.22(1.97)	2.27(2.03)	2.27(2.03)	2.20(1.93)	2.21(1.96)
(12, 64)	300	1.99(1.69)	2.01(1.76)	2.03(1.76)	2.03(1.80)	2.02(1.81)	2.01(1.79)
	600	1.83(1.75)	1.82(1.75)	1.85(1.70)	1.83(1.69)	1.86(1.77)	1.86(1.76)
	900	1.52(1.53)	1.52(1.52)	1.55(1.58)	1.54(1.56)	1.56(1.62)	1.56(1.59)
(12, 128)	300	1.49(1.52)	1.50(1.62)	1.49(1.50)	1.50(1.60)	1.53(1.57)	1.52(1.63)
	600	1.37(1.49)	1.36(1.46)	1.33(1.47)	1.32(1.44)	1.34(1.45)	1.33(1.43)
	900	1.13(1.33)	1.14(1.35)	1.07(1.21)	1.09(1.23)	1.06(1.24)	1.09(1.29)
(12, 256)	300	1.15(1.31)	1.16(1.32)	1.17(1.35)	1.16(1.33)	1.13(1.34)	1.13(1.33)
	600	1.03(1.22)	1.02(1.25)	1.00(1.18)	0.99(1.15)	1.02(1.23)	1.03(1.23)
	900	0.81(1.04)	0.80(1.05)	0.81(1.06)	0.81(1.07)	0.81(1.07)	0.82(1.08)