Submitted to the Annals of Statistics

IDENTIFICATION AND ESTIMATION FOR MATRIX TIME SERIES CP-FACTOR MODELS

BY JINYUAN CHANG[1,*], YUE DU[1,†], GUANGLIN HUANG[1,‡], AND QIWEI YAO[2,§]

[1]*Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences,* [*]*changjinyuan@swufe.edu.cn;* [†]*122071400004@smail.swufe.edu.cn;* [‡]*huanggl@swufe.edu.cn*

[2]*Department of Statistics, London School of Economics and Political Science,* [§]*q.yao@lse.ac.uk*

We propose a new method for identifying and estimating the CP-factor models for matrix time series. Unlike the generalized eigenanalysis-based method of Chang et al. (2023) for which the convergence rates of the associated estimators may suffer from small eigengaps as the asymptotic theory is based on some matrix perturbation analysis, the proposed new method enjoys faster convergence rates which are free from any eigengaps. It achieves this by turning the problem into a joint diagonalization of several matrices whose elements are determined by a basis of a linear system, and by choosing the basis carefully to avoid near co-linearity (see Proposition 5 and Section 4.3). Furthermore, unlike Chang et al. (2023) which requires the two factor loading matrices to be full-ranked, the proposed new method can handle rank-deficient factor loading matrices. Illustration with both simulated and real matrix time series data shows the advantages of the proposed new method.

1. Introduction. The modern capacity for data collection has resulted in an abundance of time series data, with those in high-dimensional matrix format increasingly prevalent across diverse fields such as economics, finance, engineering, environmental sciences, medical research, network traffic monitoring, image processing and others. The demand of modeling and forecasting high-dimensional matrix time series brings the opportunities with challenges. Let $\mathbf{Y}_t = (y_{i,j,t})$ be a $p \times q$ matrix recorded at time t, where $y_{i,j,t}$ represents the value of, for example, the j-th variable on the i-th individual at time t. A popular approach to model $\mathbf{Y}_t$ in the existing literature is via the so-called Tucker decomposition, namely the matrix Tucker-factor model. See, for example, Wang, Liu and Chen (2019), Chen, Tsay and Chen (2020), Chen and Chen (2022), and Han et al. (2024a). It represents a high-dimensional matrix time series as a linear combination of a lower-dimensional matrix process. The Tucker decomposition can be viewed as a natural extension of the factor model for vector time series considered in Lam and Yao (2012) and Chang, Guo and Yao (2015). Similarly we can only identify the factor loading spaces (the linear spaces spanned by the columns of the factor loading matrices) in the matrix Tucker-factor model while the factor loading matrices themselves are not uniquely defined. Parallel to the approaches based on Tucker decomposition, Chang et al. (2023) and Han et al. (2024b) consider to model $\mathbf{Y}_t$ via the so-called canonical polyadic (CP) decomposition, namely the matrix CP-factor model. It provides a more comprehensive dimensionality reduction as the dynamic structure of a matrix time series is driven by a vector process rather than a matrix process. Furthermore the factor loading matrices in the matrix CP-factor model can be identified uniquely up to the column reflection and permutation indeterminacy under some regularity conditions.

MSC2020 subject classifications: Primary 62M10; secondary 62H25.

Keywords and phrases: CP-decomposition, dimension-reduction, matrix time series, non-orthogonal joint diagonalization.

The CP-factor model for matrix time series $\mathbf{Y}_t$ admits the form

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t\mathbf{B}^\top + \boldsymbol{\varepsilon}_t\,, \quad t \geq 1\,, \tag{1}$$

where $\mathbf{X}_t = \mathrm{diag}(\mathbf{x}_t)$ with $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,d})^\top$ being a $d \times 1$ time series, $\boldsymbol{\varepsilon}_t$ is a $p \times q$ matrix white noise, and $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ are, respectively, $p \times d$ and $q \times d$ constant matrices which are called factor loading matrices. See, for example, Chang et al. (2023). Without loss of generality, we assume $|\mathbf{a}_\ell|_2 = 1 = |\mathbf{b}_\ell|_2$ for each $\ell = 1, \ldots, d$. For matrix CP-factor model (1), we cannot observe $(\mathbf{A}, \mathbf{B}, \mathbf{X}_t, \boldsymbol{\varepsilon}_t)$ and only assume $1 \leq d < \min(p, q)$ is an unknown fixed integer. Based on the assumption $\mathrm{rank}(\mathbf{A}) = d = \mathrm{rank}(\mathbf{B})$, Chang et al. (2023) proposes a one-pass estimation procedure for $(d, \mathbf{A}, \mathbf{B})$ which identifies $(\mathbf{A}, \mathbf{B})$ uniquely up to the column reflection and permutation indeterminacy. In contrast to the standard alternating least squares method and its variations (Han and Zhang, 2022; Han et al., 2024b), the estimation procedure proposed in Chang et al. (2023) is based on solving some generalized eigenequations and requires no iterations. Note that the incoherence conditions imposed in Han and Zhang (2022) and Han et al. (2024b) also require both $\mathbf{A}$ and $\mathbf{B}$ to be full-ranked. In fact those conditions imply that both $\{\mathbf{a}_\ell\}_{\ell=1}^d$ and $\{\mathbf{b}_\ell\}_{\ell=1}^d$ are two sets of near-orthogonal vectors. We do not require such an incoherence condition in this paper.

In this paper, we investigate the identification issue of the CP-factor model (1) for matrix time series without imposing the condition $\mathrm{rank}(\mathbf{A}) = d = \mathrm{rank}(\mathbf{B})$. Let

$$\mathrm{rank}(\mathbf{A}) = d_1 \;\text{ and }\; \mathrm{rank}(\mathbf{B}) = d_2\,.$$

Then $1 \leq d_1, d_2 \leq d$. As the CP-decomposition for 3-way tensors often exhibits rank-deficient factor loading matrices (Kolda and Bader, 2009), i.e., in model (1) it may hold that $\max(d_1, d_2) < d$. We identify the condition under which $\mathbf{A}$ and $\mathbf{B}$ are uniquely identifiable up to the column reflection and permutation indeterminacy. Our setting allows all scenarios in terms of the relationships among d_1, d_2 and d.

The proposed new estimation procedure consists of several steps (see Section 4). The key idea is to transform the $p \times q$ matrix CP-factor model (1) to a $(d_1 d_2)$-vector factor model, and then to identify the columns of $\mathbf{A}$ and $\mathbf{B}$ by a joint diagonalization of several symmetric matrices whose elements are determined by a basis, and in fact any basis, of a linear system (see Proposition 5). Therefore, we can choose an appropriate basis to avoid near co-linearity such that our estimator enjoys faster convergence rate than those eigenanalysis-based estimators (see Section 4.3). Note that the convergence rates of the eigenanalysis-based estimators are derived based on some matrix perturbation analysis, and may suffer from the adverse impact of eigen-gap (i.e., the minimum pairwise gap among a set of eigenvalues). Our newly proposed estimator is free from this adversity. For example, the convergence rate of the estimator of Chang et al. (2023) can be formulated as the product of the rate of our new estimator and the inverse of an eigen-gap (See Remark 2). Note that the eigen-gap typically diminishes to 0 when p or/and q diverge to infinity.

The rest of the paper is organized as follows. Section 2 gives preliminaries of the matrix CP-factor model (1). A general identification strategy for the matrix CP-factor model is presented in Section 3. Section 4 provides a one-pass estimation procedure for $(d_1, d_2, d, \mathbf{A}, \mathbf{B})$. Section 5 gives a unified prediction approach for the matrix CP-factor model. We investigate the associated theoretical properties of the proposed method in Section 6. Numerical results with simulation studies and real data analysis are given in Section 7. The R-function `CP_MTS` for implementing our newly proposed method is available publicly in the `HDTSA` package (Chang et al., 2024). All technical proofs and some additional simulation studies are relegated in the supplementary material.

Notation. 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 $\mathcal{R}(\mathbf{H}) = \max\{k : \text{any } k \text{ columns of the matrix } \mathbf{H} \text{ are linearly independent}\}$, and denote by $\mathcal{M}(\mathbf{H})$ the linear space spanned by the columns of $\mathbf{H}$. Let $\|\mathbf{H}\|_2$, $\|\mathbf{H}\|_{\mathrm{F}}$, $\mathrm{rank}(\mathbf{H})$, $\lambda_i(\mathbf{H})$, and $\sigma_i(\mathbf{H})$ be, respectively, the spectral norm, Frobenius norm, rank, i-th largest eigenvalue, and i-th largest singular value of matrix $\mathbf{H}$. Specifically, if $m_2 = 1$, we use $|\mathbf{H}|_1 = \sum_{i=1}^{m_1} |h_{i,1}|$ and $|\mathbf{H}|_2 = (\sum_{i=1}^{m_1} h_{i,1}^2)^{1/2}$ to denote, respectively, the L_1-norm and L_2-norm of the m_1-dimensional vector $\mathbf{H}$. Also, denote by $\mathbf{H}^\top$ and $\mathbf{H}^+$, respectively, the transpose and the Moore-Penrose inverse of $\mathbf{H}$. The operator $\mathrm{diag}(\cdot)$ stacks a vector into a square diagonal matrix. Let $\otimes$ denote the Kronecker product, and $\odot$ denote the Khatri-Rao product such that $\check{\mathbf{H}} \odot \tilde{\mathbf{H}} = (\check{\mathbf{h}}_1 \otimes \tilde{\mathbf{h}}_1, \ldots, \check{\mathbf{h}}_m \otimes \tilde{\mathbf{h}}_m)$ for any matrices $\check{\mathbf{H}} = (\check{\mathbf{h}}_1, \ldots, \check{\mathbf{h}}_m)$ and $\tilde{\mathbf{H}} = (\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_m)$. Moreover, for any two sequences of positive numbers $\{\tau_k\}$ and $\{\tilde{\tau}_k\}$, we write $\tau_k \asymp \tilde{\tau}_k$ if $\tau_k/\tilde{\tau}_k = O(1)$ and $\tilde{\tau}_k/\tau_k = O(1)$ as $k \to \infty$, and write $\tau_k \ll \tilde{\tau}_k$ or $\tilde{\tau}_k \gg \tau_k$ if $\limsup_{k\to\infty} \tau_k/\tilde{\tau}_k = 0$. To simplify our presentation, for a matrix $\mathbf{H} = (h_{i,j})_{m_1\times m_2}$, we write $\vec{\mathbf{H}}$ or $\mathrm{vec}(\mathbf{H})$ as an (m_1m_2)-dimensional vector with the $\{(j-1)m_1 + i\}$-th element being $h_{i,j}$, and for a tensor $\mathcal{H} = (h_{i,j,k,l})_{m_1\times m_2\times m_3\times m_4}$, we write $\vec{\mathcal{H}}$ as an $(m_1m_2m_3m_4)$-dimensional vector with the $\{(i-1)m_2m_3m_4 + (j-1)m_3m_4 + (k-1)m_4 + l\}$-th element being $h_{i,j,k,l}$.

2. Preliminary. Recall that, in the matrix CP-factor model (1), $\mathbf{A}$ and $\mathbf{B}$ are, respectively, $p \times d$ and $q \times d$ matrices with $\mathrm{rank}(\mathbf{A}) = d_1$ and $\mathrm{rank}(\mathbf{B}) = d_2$, and $d_1, d_2 \in [d]$. Model (1) can be equivalently represented as

$$\vec{\mathbf{Y}}_t = (\mathbf{B} \odot \mathbf{A})\mathbf{x}_t + \vec{\boldsymbol{\varepsilon}}_t\,, \quad t \geq 1\,,$$

where $\mathbf{B} \odot \mathbf{A} = (\mathbf{b}_1 \otimes \mathbf{a}_1, \ldots, \mathbf{b}_d \otimes \mathbf{a}_d)$ and $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,d})^\top$. Condition 1(i) below holds naturally. If $\mathrm{rank}(\mathbf{B} \odot \mathbf{A}) = \tilde{d} < d$, the matrix $\mathbf{B} \odot \mathbf{A}$ has $\tilde{d}$ linearly independent columns that span its column space. Therefore, we can find $\{\mathbf{b}_{\ell_1} \otimes \mathbf{a}_{\ell_1}, \ldots, \mathbf{b}_{\ell_{\tilde{d}}} \otimes \mathbf{a}_{\ell_{\tilde{d}}}\}$ with some distinct $\ell_1, \ldots, \ell_{\tilde{d}} \in [d]$ such that they provide a basis for $\mathcal{M}(\mathbf{B} \odot \mathbf{A})$. The remaining columns of $\mathbf{B} \odot \mathbf{A}$ can be expressed as linear combinations of this set of basis vectors. Then $\mathbf{B} \odot \mathbf{A} = (\mathbf{b}_{\ell_1} \otimes \mathbf{a}_{\ell_1}, \ldots, \mathbf{b}_{\ell_{\tilde{d}}} \otimes \mathbf{a}_{\ell_{\tilde{d}}})\tilde{\mathbf{C}}$ for some $\tilde{d} \times d$ matrix $\tilde{\mathbf{C}}$. Since $(\mathbf{A}, \mathbf{B}, \mathbf{X}_t)$ are unobserved, we can reformulate $\vec{\mathbf{Y}}_t$ in a new form $\vec{\mathbf{Y}}_t = (\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})\tilde{\mathbf{x}}_t + \vec{\boldsymbol{\varepsilon}}_t$ with $\tilde{\mathbf{A}} = (\mathbf{a}_{\ell_1}, \ldots, \mathbf{a}_{\ell_{\tilde{d}}})$, $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_{\ell_1}, \ldots, \tilde{\mathbf{b}}_{\ell_{\tilde{d}}})$ and $\tilde{\mathbf{x}}_t = \tilde{\mathbf{C}}\mathbf{x}_t$. In this new form, the newly defined factor loading matrices $\tilde{\mathbf{A}} \in \mathbb{R}^{p\times\tilde{d}}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{q\times\tilde{d}}$ satisfy $\mathrm{rank}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = \tilde{d}$. On the other hand, since $\boldsymbol{\varepsilon}_t$ is a matrix white noise, Condition 1(ii) holds automatically.

CONDITION 1. (i) $\mathrm{rank}(\mathbf{B} \odot \mathbf{A}) = d$. (ii) $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ *for any* $t \geq 1$, $\mathbb{E}(\boldsymbol{\varepsilon}_t \otimes \boldsymbol{\varepsilon}_s) = \mathbf{0}$ *for all* $t \neq s$, *and* $\mathbb{E}(x_{t,\ell}\boldsymbol{\varepsilon}_s) = \mathbf{0}$ *for any* $\ell \in [d]$ *and* $t \leq s$.

When $d_1 = d_2 = d$, Chang et al. (2023) provides a one-pass estimator for $(\mathbf{A}, \mathbf{B})$ by solving some generalized eigenequations defined by the matrices

$$\boldsymbol{\Sigma}_{\mathbf{Y},\xi}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\}\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}]\,, \quad k \geq 1\,, \tag{2}$$

where $\bar{\mathbf{Y}} = n^{-1}\sum_{t=1}^{n} \mathbf{Y}_t$, ξ_t is a scalar defined as a linear combination of the elements of $\mathbf{Y}_t$, and $\bar{\xi} = n^{-1}\sum_{t=1}^{n} \xi_t$. For example, we can select ξ_t as the first principal component of $\vec{\mathbf{Y}}_t$.

Recall $\mathbf{A}^+$ and $\mathbf{B}^+$ are, respectively, the Moore-Penrose inverse of $\mathbf{A}$ and $\mathbf{B}$. The key requirement underlying the results of Chang et al. (2023) is $\mathbf{A}^+\mathbf{A} = \mathbf{B}^+\mathbf{B} = \mathbf{I}_d$, which only holds when $d_1 = d_2 = d$. Hence, the estimation method of Chang et al. (2023) is not applicable when $\min(d_1, d_2) < d$. Note that the CP-decomposition for 3-way tensors can

often exhibit rank-deficient factor loading matrices (Kolda and Bader, 2009), i.e., in model (1) it may hold that $\min(d_1, d_2) < d$ or even $\max(d_1, d_2) < d$. In this paper, we consider a new approach which identifies $(d, \mathbf{A}, \mathbf{B})$ without the condition $d_1 = d_2 = d$. Furthermore we propose a unified and more efficient one-pass estimation for $(\mathbf{A}, \mathbf{B})$ regardless they are rank-deficient or not.

3. Identification of $(\mathbf{A}, \mathbf{B})$. We need to identify in model (1) the order d and the factor loading pairs $(\mathbf{a}_1, \mathbf{b}_1), \ldots, (\mathbf{a}_d, \mathbf{b}_d)$. To carry out this task, we first introduce a reduced model for a $d_1 \times d_2$ matrix time series, and then identify d and the CP-factor loadings for the reduced model via (i) a factor model for a vector time series, and (ii) a non-orthogonal joint diagonalization of d symmetric matrices.

3.1. *A reduced model.* For a prescribed integer $K > 1$ and $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ specified in (2), define

$$\mathbf{M}_1 = \sum_{k=1}^{K} \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^\top \text{ and } \mathbf{M}_2 = \sum_{k=1}^{K} \mathbf{\Sigma}_{\mathbf{Y},\xi}(k)^\top\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\,. \tag{3}$$

Furthermore, due to $\mathbf{X}_t = \mathrm{diag}(\mathbf{x}_t)$ with $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,d})^\top$, we let

$$\mathbf{G}_k = \mathrm{diag}(\mathbf{g}_k) = \frac{1}{n-k}\sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{X}_t - \mathbb{E}(\bar{\mathbf{X}})\}\{\xi_{t-k} - \mathbb{E}(\bar{\xi})\}]\,, \quad k \in [K]\,,$$

where $\bar{\mathbf{X}} = n^{-1}\sum_{t=1}^{n}\mathbf{X}_t$. It follows from (1) and Condition 1 that

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

Let $\mathbf{G} = \sum_{k=1}^{K}\mathbf{g}_k\mathbf{g}_k^\top$. Proposition 1 shows that d_1 and d_2 can be identified, repsectively, by $\mathrm{rank}(\mathbf{M}_1)$ and $\mathrm{rank}(\mathbf{M}_2)$.

PROPOSITION 1. *Let Condition 1 hold and all the main diagonal elements of* $\mathbf{G}$ *are non-zero. The following two assertions hold.*

(i) *If* $\max\{\mathcal{R}(\mathbf{G}) + d_2, \mathcal{R}(\mathbf{B}^\top\mathbf{B}) + \mathrm{rank}(\mathbf{G})\} > d$, *then* $\mathrm{rank}(\mathbf{M}_1) = d_1$.
(ii) *If* $\max\{\mathcal{R}(\mathbf{G}) + d_1, \mathcal{R}(\mathbf{A}^\top\mathbf{A}) + \mathrm{rank}(\mathbf{G})\} > d$, *then* $\mathrm{rank}(\mathbf{M}_2) = d_2$.

The conditions required in Proposition 1 are mild. Notice that all the main diagonal elements of $\mathbf{G}$ are positive if all the components of some $\mathbf{g}_k$ are non-zero with $k \in [K]$, which implies $\mathcal{R}(\mathbf{G}) \geq 1$. For the scenario $d_1 = d_2 = d$, Proposition 1 holds automatically. When $\min(d_1, d_2) < d$, we suppose that $d_2 \leq d_1$ without loss of generality. For the scenario $d_2 < d_1 = d$, we only need to identify d_2. Proposition 1(ii) holds automatically in this scenario, which implies d_2 could be identified trivially. For the scenario $\max(d_1, d_2) < d$, by Condition 1(i), we know $\mathbf{a}_\ell \neq \mathbf{0}$ and $\mathbf{b}_\ell \neq \mathbf{0}$ for each $\ell \in [d]$, which implies $\mathcal{R}(\mathbf{A}^\top\mathbf{A}) \geq 1$ and $\mathcal{R}(\mathbf{B}^\top\mathbf{B}) \geq 1$. Proposition 2 proposes some sufficient conditions such that $\mathrm{rank}(\mathbf{G}) = d$, which make Proposition 1 hold automatically. Define

$$\mathbf{\Sigma}_{\mathbf{x}}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{x}_t - \mathbb{E}(\bar{\mathbf{x}})\}\{\mathbf{x}_{t-k} - \mathbb{E}(\bar{\mathbf{x}})\}^\top]\,, \quad k \in [K]\,,$$

where $\bar{\mathbf{x}} = n^{-1}\sum_{t=1}^{n}\mathbf{x}_t$. Write $\xi_t = \boldsymbol{\omega}^\top\vec{\mathbf{Y}}_t$ and $\mathbf{\Sigma}_{\mathbf{x},K} = \{\vec{\mathbf{\Sigma}}_{\mathbf{x}}(1), \ldots, \vec{\mathbf{\Sigma}}_{\mathbf{x}}(K)\} \in \mathbb{R}^{d^2\times K}$.

PROPOSITION 2. *Assume that* $\mathbb{E}(\mathbf{x}_t \otimes \vec{\boldsymbol{\varepsilon}}_{t-k}) = \mathbf{0}$ *for any* $k \in [K]$ *with* $K \geq d^2$. *If* $\boldsymbol{\omega}^\top(\mathbf{B} \odot \mathbf{A}) \neq \mathbf{0}$ *and* $\operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{x},K}) = d^2$, *then* $\operatorname{rank}(\mathbf{G}) = d$.

Due to $\mathbf{a}_\ell \neq \mathbf{0}$ and $\mathbf{b}_\ell \neq \mathbf{0}$ for each $\ell \in [d]$, the requirement $\boldsymbol{\omega}^\top(\mathbf{B} \odot \mathbf{A}) \neq \mathbf{0}$ is generally mild and can be satisfied by appropriately choosing a non-zero vector $\boldsymbol{\omega}$. If $\mathbf{x}_t$ satisfies $\operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{x},K}) = d^2$, and $\mathbf{x}_t$ and $\vec{\boldsymbol{\varepsilon}}_{t-k}$ are uncorrelated for $k \in [K]$, Proposition 2 shows that $\operatorname{rank}(\mathbf{G}) = d$. Combining with Proposition 1, it is reasonable to assume Condition 2, which ensures $\mathcal{M}(\mathbf{M}_1) = \mathcal{M}(\mathbf{A})$ and $\mathcal{M}(\mathbf{M}_2) = \mathcal{M}(\mathbf{B})$, i.e., the information on the loadings $\{\mathbf{a}_\ell\}_{\ell=1}^d$ and $\{\mathbf{b}_\ell\}_{\ell=1}^d$ is, respectively, kept in $\mathbf{M}_1$ and $\mathbf{M}_2$.

CONDITION 2. $\operatorname{rank}(\mathbf{M}_1) = d_1$ *and* $\operatorname{rank}(\mathbf{M}_2) = d_2$.

Now perform the spectral decomposition for $\mathbf{M}_1$ and $\mathbf{M}_2$:

$$\mathbf{M}_1 = \mathbf{P}\mathbf{D}_1\mathbf{P}^\top \text{ and } \mathbf{M}_2 = \mathbf{Q}\mathbf{D}_2\mathbf{Q}^\top, \tag{4}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are, respectively, $d_1 \times d_1$ and $d_2 \times d_2$ full-ranked diagonal matrices, $\mathbf{P}^\top\mathbf{P} = \mathbf{I}_{d_1}$ and $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_{d_2}$. As $\mathcal{M}(\mathbf{P}) = \mathcal{M}(\mathbf{M}_1) = \mathcal{M}(\mathbf{A})$ and $\mathcal{M}(\mathbf{Q}) = \mathcal{M}(\mathbf{M}_2) = \mathcal{M}(\mathbf{B})$, then

$$\mathbf{A} = \mathbf{P}\mathbf{U} \text{ and } \mathbf{B} = \mathbf{Q}\mathbf{V}, \tag{5}$$

where $\mathbf{U}$ and $\mathbf{V}$ are, respectively, $d_1 \times d$ and $d_2 \times d$ matrices with unit column vectors. Since $\mathbf{P}$ and $\mathbf{Q}$ are determined by the spectral decomposition (4), we only need to identify $(\mathbf{U}, \mathbf{V})$ in order to identify $(\mathbf{A}, \mathbf{B})$. When $d = 1$, we may take $\mathbf{a}_1 = \mathbf{A} = \mathbf{P}$ and $\mathbf{b}_1 = \mathbf{B} = \mathbf{Q}$. Therefore only the non-trivial case with $d \geq 2$ will be considered in the sequel.

Define a $d_1 \times d_2$ process $\mathbf{Z}_t = \mathbf{P}^\top\mathbf{Y}_t\mathbf{Q}$. It follows from (1) and (5) that

$$\mathbf{Z}_t = \mathbf{U}\mathbf{X}_t\mathbf{V}^\top + \boldsymbol{\Delta}_t, \quad t \geq 1, \tag{6}$$

where $\boldsymbol{\Delta}_t = \mathbf{P}^\top\boldsymbol{\varepsilon}_t\mathbf{Q}$ is a matrix white noise. This is a reduced form of the CP-factor model (1) for the matrix time series $\mathbf{Y}_t$. We will identify $(\mathbf{U}, \mathbf{V})$ based on this reduced model.

3.2. *A vector factor model.* Recall $\mathbf{X}_t = \operatorname{diag}(\mathbf{x}_t)$ with $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,d})^\top$. It follows from (6) that

$$\vec{\mathbf{Z}}_t = (\mathbf{V} \odot \mathbf{U})\mathbf{x}_t + \vec{\boldsymbol{\Delta}}_t, \quad t \geq 1. \tag{7}$$

This is the standard factor model for vector time series considered by Lam and Yao (2012) and Chang, Guo and Yao (2015). Note that $\mathbf{B} \odot \mathbf{A} = (\mathbf{Q} \otimes \mathbf{P})(\mathbf{V} \odot \mathbf{U})$, where $\mathbf{P} \in \mathbb{R}^{p \times d_1}$ and $\mathbf{Q} \in \mathbb{R}^{q \times d_2}$ with $\operatorname{rank}(\mathbf{P}) = d_1$ and $\operatorname{rank}(\mathbf{Q}) = d_2$. By Condition 1(i), we know the dimension of the factor loading space $\mathcal{M}(\mathbf{V} \odot \mathbf{U})$ in (7) is d, as $\operatorname{rank}(\mathbf{V} \odot \mathbf{U}) = \operatorname{rank}(\mathbf{B} \odot \mathbf{A}) = d$. Using the techniques developed in Chang, Guo and Yao (2015), we can identify d and $\mathcal{M}(\mathbf{V} \odot \mathbf{U})$ uniquely based on an eigenanalysis. More precisely, we can find a $(d_1 d_2) \times d$ matrix $\mathbf{W}$, with $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_d$, such that

$$\mathbf{V} \odot \mathbf{U} \equiv (\mathbf{v}_1 \otimes \mathbf{u}_1, \ldots, \mathbf{v}_d \otimes \mathbf{u}_d) = \mathbf{W}\boldsymbol{\Theta}, \tag{8}$$

where $\boldsymbol{\Theta}$ is an unknown $d \times d$ invertible matrix with unit column vectors. Since the d columns of $\mathbf{W}$ are the orthogonal basis of $\mathcal{M}(\mathbf{V} \odot \mathbf{U})$, we can select $\mathbf{W}$ in (8) as an arbitrary $(d_1 d_2) \times d$ matrix such that $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_d$ and $\mathcal{M}(\mathbf{W}) = \mathcal{M}(\mathbf{U} \odot \mathbf{V})$. In (8), different selections of $\mathbf{W}$ will lead to different $\boldsymbol{\Theta}$. As we will show in Section 3.3, for any given $\mathbf{W}$, the associated rotation matrix $\boldsymbol{\Theta}$ can be uniquely identified up to the column reflection and permutation indeterminacy. Write

$$\mathbf{C} \equiv (\vec{\mathbf{C}}_1, \ldots, \vec{\mathbf{C}}_d) = \mathbf{W}\boldsymbol{\Theta} \equiv (\vec{\mathbf{W}}_1, \ldots, \vec{\mathbf{W}}_d)\boldsymbol{\Theta}, \tag{9}$$

where $\mathbf{C}_\ell$ and $\mathbf{W}_\ell$ are $d_1 \times d_2$ matrices. We put the columns of both $\mathbf{C}$ and $\mathbf{W}$ in the form of vectorized $d_1 \times d_2$ matrices for some technical convenience which will be obvious soon. It follows from (8) and (9) that $\vec{\mathbf{C}}_\ell = \mathbf{v}_\ell \otimes \mathbf{u}_\ell$, which implies $\mathbf{u}_\ell \mathbf{v}_\ell^\top = \mathbf{C}_\ell$. Given $\mathbf{W}$ and its associated rotation matrix $\mathbf{\Theta}$, the $(d_1 d_2) \times d$ matrix $\mathbf{C}$ specified in (9) is uniquely identified, which can be used to identify $(\mathbf{U}, \mathbf{V})$. See Proposition 3 for details.

PROPOSITION 3. *Let Conditions 1 and 2 hold. Then matrices $\mathbf{C}_1, \ldots, \mathbf{C}_d$ specified in (9) are all of rank 1 with the nonzero singular value equal to 1, and $(\mathbf{u}_\ell, \mathbf{v}_\ell)$ are the unit singular vectors of $\mathbf{C}_\ell$ for each $\ell \in [d]$.*

3.3. *A non-orthogonal joint diagonalization.* For given $\mathbf{W}$ in (8), Proposition 3 implies that the task of identifying $(\mathbf{U}, \mathbf{V})$ boils down to identifying $\mathbf{\Theta}$ specified in (8) such that $\mathbf{C}_1, \ldots, \mathbf{C}_d$ defined in (9) satisfying $\mathrm{rank}(\mathbf{C}_\ell) = 1$ for each $\ell \in [d]$. By (9), it holds that

$$\mathbf{C}_\ell = \sum_{i=1}^d \theta_{i,\ell} \mathbf{W}_i \ \text{ and } \ \mathbf{W}_\ell = \sum_{i=1}^d \theta^{i,\ell} \mathbf{C}_i\,, \quad \ell \in [d]\,, \tag{10}$$

where $\theta_{i,j}$ and $\theta^{i,j}$ denote, respectively, the (i,j)-th elements of $\mathbf{\Theta}$ and $\mathbf{\Theta}^{-1}$.

For any two matrices $\mathbf{D} = (d_{i,j})$ and $\mathbf{F} = (f_{i,j})$ of the same size, define $\mathbf{\Psi}(\mathbf{D}, \mathbf{F})$ to be a 4-way tensor with the (i,j,k,ℓ)-th element $d_{i,k} f_{j,\ell} + d_{j,\ell} f_{i,k} - d_{i,\ell} f_{j,k} - d_{j,k} f_{i,\ell}$. By Theorem 2.1 of De Lathauwer (2006), for any matrix $\mathbf{D} \neq \mathbf{0}$, $\mathrm{rank}(\mathbf{D}) = 1$ if and only if $\mathbf{\Psi}(\mathbf{D}, \mathbf{D}) = \mathbf{0}$. Hence, by (10), for given $\mathbf{W}$ in (8), we know $\mathbf{\Theta} = (\theta_{i,j})$ is the solution of

$$\mathbf{0} = \mathbf{\Psi}(\mathbf{C}_\ell, \mathbf{C}_\ell) = \sum_{i,j=1}^d \theta_{i,\ell} \theta_{j,\ell} \mathbf{\Psi}(\mathbf{W}_i, \mathbf{W}_j)\,, \quad \ell \in [d]\,. \tag{11}$$

This is a set of quadratic equations. Consider a $(d_1^2 d_2^2) \times d(d+1)/2$ matrix

$$\mathbf{\Omega} = \big(\vec{\mathbf{\Psi}}(\mathbf{W}_1, \mathbf{W}_1), \ldots, \vec{\mathbf{\Psi}}(\mathbf{W}_1, \mathbf{W}_d), \vec{\mathbf{\Psi}}(\mathbf{W}_2, \mathbf{W}_2), \ldots, \vec{\mathbf{\Psi}}(\mathbf{W}_d, \mathbf{W}_d)\big)\,. \tag{12}$$

Proposition 4 is instrumental in solving those quadratic equations.

PROPOSITION 4. *Let $d \geq 2$ and Condition 2 hold. The following three assertions hold.*

(i) $\mathrm{rank}(\mathbf{\Omega}) \leq d(d-1)/2$.

(ii) $\mathrm{rank}(\mathbf{\Omega}) = d(d-1)/2$ *if and only if the* $d(d-1)/2$ *vectors* $\vec{\mathbf{\Psi}}(\mathbf{C}_1, \mathbf{C}_2), \ldots, \vec{\mathbf{\Psi}}(\mathbf{C}_1, \mathbf{C}_d)$, $\vec{\mathbf{\Psi}}(\mathbf{C}_2, \mathbf{C}_3), \ldots, \vec{\mathbf{\Psi}}(\mathbf{C}_{d-1}, \mathbf{C}_d)$ *are linearly independent.*

(iii) *Let* $\ker(\mathbf{\Omega}) = \{\mathbf{h} \in \mathbb{R}^{d(d+1)/2} : \mathbf{\Omega}\mathbf{h} = \mathbf{0}\}$. *Then* $\dim\{\ker(\mathbf{\Omega})\} = d$ *if and only if* $\mathrm{rank}(\mathbf{\Omega}) = d(d-1)/2$.

Now assume $\mathrm{rank}(\mathbf{\Omega}) = d(d-1)/2$. Let $\mathbf{h}_m = (h^m_{1,1}, \ldots, h^m_{1,d}, h^m_{2,2}, \ldots, h^m_{d,d})^\top$, $m \in [d]$, be a set of basis vectors of $\ker(\mathbf{\Omega})$. Recall $\mathbf{\Psi}(\mathbf{C}_\ell, \mathbf{C}_\ell) = \mathbf{0}$ for any $\ell \in [d]$. By (10), it holds that

$$\begin{aligned} \mathbf{0} &= \sum_{1 \leq i \leq j \leq d} h^m_{i,j} \mathbf{\Psi}(\mathbf{W}_i, \mathbf{W}_j) = \sum_{1 \leq i \leq j \leq d} h^m_{i,j} \sum_{k,\ell=1}^d \theta^{k,i} \theta^{\ell,j} \mathbf{\Psi}(\mathbf{C}_k, \mathbf{C}_\ell) \\ &= \sum_{1 \leq k < \ell \leq d} \mathbf{\Psi}(\mathbf{C}_k, \mathbf{C}_\ell) \sum_{1 \leq i \leq j \leq d} (\theta^{k,i} \theta^{\ell,j} + \theta^{k,j} \theta^{\ell,i}) h^m_{i,j}\,. \end{aligned}$$

By Proposition 4(ii), we have

$$\sum_{1 \leq i \leq j \leq d} (\theta^{k,i} \theta^{\ell,j} + \theta^{k,j} \theta^{\ell,i}) h^m_{i,j} = 0 \ \text{ for all } \ 1 \leq k < \ell \leq d\,. \tag{13}$$

Let $\mathbf{H}_m$ be a $d \times d$ matrix with the (i,i)-th element being $h_{i,i}^m$ for any i, and the (i,j)-th and (j,i)-th elements being $h_{i,j}^m/2$ for any $i<j$. Based on (13), we know $\mathbf{\Gamma}_m \equiv \mathbf{\Theta}^{-1}\mathbf{H}_m(\mathbf{\Theta}^{-1})^\top$ is a diagonal matrix, i.e., we can find $\mathbf{\Theta}^{-1}$ which diagonalizes jointly $\mathbf{H}_m = \mathbf{\Theta}\mathbf{\Gamma}_m\mathbf{\Theta}^\top$ for each $m \in [d]$.

It is also clear from (13) that the diagonal property is independent of the norms of row vectors of $\mathbf{\Theta}^{-1}$. Hence all the columns of $\mathbf{\Theta}$ can be set as unit vectors. Proposition 5 shows that $\mathbf{\Theta}$ is invariant with respect to the choice of the basis vectors for $\ker(\mathbf{\Omega})$. The available algorithms for this joint diagonalization include the joint approximate diagonalization of Pham and Cardoso (2001), the fast Frobenius diagonalization of Ziehe et al. (2004), and the quadratic diagonalization of Vollgraf and Obermayer (2006).

PROPOSITION 5. *Let Condition 2 hold. For a given $(d_1d_2)\times d$ matrix $\mathbf{W}$ in (8) such that $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_d$ and $\mathcal{M}(\mathbf{W}) = \mathcal{M}(\mathbf{U}\odot\mathbf{V})$, if $\mathbf{\Omega}$ defined in (12) satisfies $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$, then $\mathbf{\Theta}$ in (8) can be uniquely identified by the non-orthogonal joint diagonalization $\mathbf{H}_m = \mathbf{\Theta}\mathbf{\Gamma}_m\mathbf{\Theta}^\top$, $m \in [d]$, up to the column reflection and permutation indeterminacy, and $\mathbf{\Theta}$ is invariant with respect to the choice of the basis vectors of $\ker(\mathbf{\Omega})$.*

Proposition 6 provides a sufficient condition under which $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$. Such sufficient condition holds automatically when $d_1 = d_2 = d$, as then $\mathcal{R}(\mathbf{A}) = \mathcal{R}(\mathbf{B}) = d$. When $d_1 \neq d_2$, we assume $d_2 < d_1$ without loss of generality. For the scenario $d_2 < d_1 = d$, since $\mathcal{R}(\mathbf{A}) = d$, Proposition 6 indicates that $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$ if $\mathcal{R}(\mathbf{B}) \geq 2$. Actually, the requirement $\mathcal{R}(\mathbf{B}) \geq 2$ is necessary for the identification of $(\mathbf{A},\mathbf{B})$ when $d_2 < d_1 = d$. Recall $\vec{\mathbf{Y}}_t = (\mathbf{B}\odot\mathbf{A})\mathbf{x}_t + \vec{\boldsymbol{\varepsilon}}_t$. If $\mathcal{R}(\mathbf{B}) = 1$, since $|\mathbf{b}_\ell|_2 = 1$ for each $\ell \in [d]$, we can assume $\mathbf{b}_2 = \mathbf{b}_1$ without loss of generality. Let $\tilde{\mathbf{B}} = \mathbf{B}$ and $\tilde{\mathbf{A}} = (\tilde{\mathbf{a}}_1,\ldots,\tilde{\mathbf{a}}_d)$, where $\tilde{\mathbf{a}}_\ell = \mathbf{a}_\ell$ for any $\ell \geq 2$, and $\tilde{\mathbf{a}}_1 = c_1\mathbf{a}_1 + c_2\mathbf{a}_2$ for some nonzero constants c_1, c_2 such that $|\tilde{\mathbf{a}}_1|_2 = 1$. Select $\mathbf{\Xi} = (\xi_{i,j})$ with $\xi_{1,1} = c_1$, $\xi_{2,1} = c_2$, $\xi_{i,i} = 1$ for any $2 \leq i \leq d$, and $\xi_{i,j} = 0$ otherwise. Then $\mathbf{Y}_t$ can be also formulated by another matrix CP-factor model $\vec{\mathbf{Y}}_t = (\tilde{\mathbf{B}}\odot\tilde{\mathbf{A}})\mathbf{\Xi}^{-1}\mathbf{x}_t + \vec{\boldsymbol{\varepsilon}}_t$.

PROPOSITION 6. *Let $d \geq 2$, and Conditions 1 and 2 hold. Then $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$ provided that $\mathcal{R}(\mathbf{A}) + d_2 \geq d+2$ and $\mathcal{R}(\mathbf{B}) + d_1 \geq d+2$.*

By Propositions 3 and 5, if $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$, then $\mathbf{U}$ and $\mathbf{V}$ specified in (5) can be uniquely defined up to the column reflection and permutation indeterminacy, which implies $\mathbf{A}$ and $\mathbf{B}$ can be uniquely defined up to the column reflection and permutation indeterminacy. For $d \geq 2$, Proposition 7 shows that the requirement $\operatorname{rank}(\mathbf{\Omega}) = d(d-1)/2$ is necessary for identifying $(\mathbf{A},\mathbf{B})$, and it is impossible to obtain the consistent estimators for $(\mathbf{A},\mathbf{B})$ without such requirement.

PROPOSITION 7. *Let $d \geq 2$. Consider the following parameter space for the matrix CP-factor model (1):*

$$\mathcal{U} = \big\{(\mathbf{A},\mathbf{B}) : \mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_d) \text{ and } \mathbf{B} = (\mathbf{b}_1,\ldots,\mathbf{b}_d) \text{ with } |\mathbf{a}_\ell|_2 = 1 = |\mathbf{b}_\ell|_2 \text{ for each } \ell \in [d], \text{ and } \operatorname{rank}(\mathbf{\Omega}) < d(d-1)/2 \text{ with } \mathbf{\Omega} \text{ defined as (12)}\big\}.$$

Write $\mathcal{G} = \{(\breve{\mathbf{A}},\breve{\mathbf{B}}) : \breve{\mathbf{A}} = (\breve{\mathbf{a}}_1,\ldots,\breve{\mathbf{a}}_d) \in \mathbb{R}^{p\times d}, \breve{\mathbf{B}} = (\breve{\mathbf{b}}_1,\ldots,\breve{\mathbf{b}}_d) \in \mathbb{R}^{q\times d}\}$ for the class of all measurable estimators of $(\mathbf{A},\mathbf{B})$ based on the data $\{\mathbf{Y}_t\}_{t=1}^n$. Under Conditions 1 and 2, it holds that

$$\inf_{(\breve{\mathbf{A}},\breve{\mathbf{B}})\in\mathcal{G}} \sup_{(\mathbf{A},\mathbf{B})\in\mathcal{U}} \mathbb{P}\left[\max\{\mathscr{D}(\breve{\mathbf{A}},\mathbf{A}), \mathscr{D}(\breve{\mathbf{B}},\mathbf{B})\} \geq \frac{1}{8}\right] \geq \frac{1}{2},$$

where $\mathscr{D}(\breve{\mathbf{A}},\mathbf{A}) = \max_{\ell\in[d]}|\breve{\mathbf{a}}_\ell - \mathbf{a}_\ell|_2$ and $\mathscr{D}(\breve{\mathbf{B}},\mathbf{B}) = \max_{\ell\in[d]}|\breve{\mathbf{b}}_\ell - \mathbf{b}_\ell|_2$.

4. Estimation. Based on Section 3, we can estimate d and $(\mathbf{a}_\ell, \mathbf{b}_\ell)$ for $\ell \in [d]$ via the following five steps:

Step 1. Based on (4), we can obtain the estimates for d_1, d_2, $\mathbf{P}$ and $\mathbf{Q}$, denoted by $\hat{d}_1$, $\hat{d}_2$, $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, respectively.

Step 2. Based on (7), we can obtain the estimates for d and $\mathbf{W}$ (the orthogonal basis of $\mathcal{M}(\mathbf{V} \odot \mathbf{U})$) with replacing $\mathbf{Z}_t$ by $\hat{\mathbf{Z}}_t = \hat{\mathbf{P}}^\top \mathbf{Y}_t \hat{\mathbf{Q}}$. Denote by $\hat{d}$ and $\hat{\mathbf{W}}$ the associated estimators.

Step 3. With replacing $\mathbf{W}$ involved in (8) by $\hat{\mathbf{W}}$, we can use the joint diagonalization algorithm mentioned in Section 3.3 to obtain $\hat{\mathbf{\Theta}}$, the estimate of $\mathbf{\Theta}$ involved in (8).

Step 4. Let $\hat{\mathbf{C}} = (\mathrm{vec}(\hat{\mathbf{C}}_1), \ldots, \mathrm{vec}(\hat{\mathbf{C}}_{\hat{d}})) = \hat{\mathbf{W}}\hat{\mathbf{\Theta}}$. For each $\ell \in [\hat{d}]$, we select $\hat{\mathbf{u}}_\ell$ and $\hat{\mathbf{v}}_\ell$, respectively, as the unit eigenvectors corresponding to the largest eigenvalues of $\hat{\mathbf{C}}_\ell \hat{\mathbf{C}}_\ell^\top$ and $\hat{\mathbf{C}}_\ell^\top \hat{\mathbf{C}}_\ell$. Based on Proposition 3, we can estimate $(\mathbf{u}_\ell, \mathbf{v}_\ell)$ by $(\hat{\mathbf{u}}_\ell, \hat{\mathbf{v}}_\ell)$ for each $\ell \in [\hat{d}]$.

Step 5. Based on (5), we can estimate $\mathbf{A}$ and $\mathbf{B}$, respectively, by $\hat{\mathbf{A}} = \hat{\mathbf{P}}(\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{\hat{d}})$ and $\hat{\mathbf{B}} = \hat{\mathbf{Q}}(\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_{\hat{d}})$.

Steps 4 and 5 are straightforward. More details of Steps 1–3 are given, respectively, in Sections 4.1–4.3. Especially Step 3 involves a further rotation to improve the convergence rate of the estimation. All the estimation is based on observations $\{\mathbf{Y}_t\}_{t=1}^n$.

4.1. *Estimating d_1, d_2, $\mathbf{P}$ and $\mathbf{Q}$.* Let ξ_t be a prescribed linear combination of $\mathbf{Y}_t$ (e.g. the first principal component of $\vec{\mathbf{Y}}_t$), and $K > 1$ be a prescribed integer. Based on (3), we put

$$\hat{\mathbf{M}}_1 = \sum_{k=1}^{K} T_{\delta_1}\{\hat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)\} T_{\delta_1}\{\hat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)^\top\}\,,$$

$$\hat{\mathbf{M}}_2 = \sum_{k=1}^{K} T_{\delta_1}\{\hat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)^\top\} T_{\delta_1}\{\hat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k)\}\,, \tag{14}$$

where $T_{\delta_1}(\cdot)$ is a truncation operator with the threshold level $\delta_1 \geq 0$, i.e., $T_{\delta_1}(\mathbf{S}) = (s_{i,j} I(|s_{i,j}| \geq \delta_1))$ for any matrix $\mathbf{S} = (s_{i,j})$, and

$$\hat{\mathbf{\Sigma}}_{\mathbf{Y},\xi}(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} (\mathbf{Y}_t - \bar{\mathbf{Y}})(\xi_{t-k} - \bar{\xi})\,, \quad k \in [K]\,. \tag{15}$$

We set $\delta_1 > 0$ in (14) when $pq \geq n$. Note that $\hat{\mathbf{M}}_1$ is a $p \times p$ matrix, and $\hat{\mathbf{M}}_2$ is a $q \times q$ matrix. By Condition 2, we can estimate d_1 and d_2 by the eigenvalue-ratio based method (Chang, Guo and Yao, 2015) as follows:

$$\hat{d}_1 = \arg\min_{j \in [p]} \frac{\lambda_{j+1}(\hat{\mathbf{M}}_1) + c_{1,n}}{\lambda_j(\hat{\mathbf{M}}_1) + c_{1,n}} \quad \text{and} \quad \hat{d}_2 = \arg\min_{j \in [q]} \frac{\lambda_{j+1}(\hat{\mathbf{M}}_2) + c_{2,n}}{\lambda_j(\hat{\mathbf{M}}_2) + c_{2,n}} \tag{16}$$

for some $c_{1,n}, c_{2,n} \to 0^+$ as $n \to \infty$. The proposed eigenvalue-ratio based method here is an extension of that in Lam and Yao (2012). Adding $c_{1,n}$ and $c_{2,n}$ is to avoid the technical difficulties associated with handling potential "0/0" cases and can lead to consistent estimates for d_1 and d_2. See Theorem 1 in Section 6 for details. In contrast, the eigenvalue-ratio based method proposed in Lam and Yao (2012) without adding $c_{1,n}$ and $c_{2,n}$ only ensures that the numbers of factors are not underestimated, without providing consistency.

Perform the spectral decomposition for the non-negative definite matrices $\hat{\mathbf{M}}_1$ and $\hat{\mathbf{M}}_2$. Let $\hat{\mathbf{P}}$ be the $p \times \hat{d}_1$ matrix of which the columns are the $\hat{d}_1$ orthonormal eigenvectors of $\hat{\mathbf{M}}_1$ corresponding to its $\hat{d}_1$ largest eigenvalues, and $\hat{\mathbf{Q}}$ be the $q \times \hat{d}_2$ matrix of which the columns

are the $\hat{d}_2$ orthonormal eigenvectors of $\hat{\mathbf{M}}_2$ corresponding to its $\hat{d}_2$ largest eigenvalues. Now we are ready to reduce the original $p \times q$ process $\mathbf{Y}_t$ to the $\hat{d}_1 \times \hat{d}_2$ process

$$\hat{\mathbf{Z}}_t = \hat{\mathbf{P}}^\top \mathbf{Y}_t \hat{\mathbf{Q}}, \quad t \geq 1.$$

4.2. *Estimating d and $\mathbf{W}_1, \ldots, \mathbf{W}_d$.* Based on (7) and (8), we can reformulate (7) as $\vec{\mathbf{Z}}_t = \mathbf{W}\mathbf{x}_t^* + \vec{\boldsymbol{\Delta}}_t$ with $\mathbf{x}_t^* = \boldsymbol{\Theta}\mathbf{x}_t$. Hence, we can estimate a factor loading matrix $\mathbf{W} = (\vec{\mathbf{W}}_1, \ldots, \vec{\mathbf{W}}_d)$ based on the method proposed in Lam, Yao and Bathia (2011), Lam and Yao (2012) and Chang, Guo and Yao (2015). To do this, we put

$$\hat{\mathbf{M}} = \sum_{k=1}^{\tilde{K}} \hat{\boldsymbol{\Sigma}}_{\vec{\mathbf{Z}}}(k) \hat{\boldsymbol{\Sigma}}_{\vec{\mathbf{Z}}}(k)^\top \tag{17}$$

with a prescribed integer $\tilde{K} \geq 1$ and

$$\hat{\boldsymbol{\Sigma}}_{\vec{\mathbf{Z}}}(k) = (\hat{\mathbf{Q}}^\top \otimes \hat{\mathbf{P}}^\top) T_{\delta_2}\{\hat{\boldsymbol{\Sigma}}_{\vec{\mathbf{Y}}}(k)\} (\hat{\mathbf{Q}} \otimes \hat{\mathbf{P}}), \quad k \in [\tilde{K}], \tag{18}$$

where $T_{\delta_2}(\cdot)$ is a truncation operator with the threshold level $\delta_2 \geq 0$, and

$$\hat{\boldsymbol{\Sigma}}_{\vec{\mathbf{Y}}}(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} (\vec{\mathbf{Y}}_t - \bar{\vec{\mathbf{Y}}})(\vec{\mathbf{Y}}_{t-k} - \bar{\vec{\mathbf{Y}}})^\top, \quad k \in [\tilde{K}],$$

with $\bar{\vec{\mathbf{Y}}} = n^{-1} \sum_{t=1}^{n} \vec{\mathbf{Y}}_t$. Analogous to (16), we can estimate d as

$$\hat{d} = \left\{ \arg \min_{j \in [\hat{d}_1 \hat{d}_2]} \frac{\lambda_{j+1}(\hat{\mathbf{M}}) + c_{3,n}}{\lambda_j(\hat{\mathbf{M}}) + c_{3,n}} \right\} I(\hat{d}_1 \hat{d}_2 \geq 2) + I(\hat{d}_1 \hat{d}_2 = 1) \tag{19}$$

for some $c_{3,n} \to 0^+$ as $n \to \infty$. Furthermore we let $\hat{\mathbf{W}} \equiv (\mathrm{vec}(\hat{\mathbf{W}}_1), \ldots, \mathrm{vec}(\hat{\mathbf{W}}_{\hat{d}}))$ be the $(\hat{d}_1 \hat{d}_2) \times \hat{d}$ matrix of which the columns are the $\hat{d}$ orthonormal eigenvectors of $\hat{\mathbf{M}}$ corresponding to its largest $\hat{d}$ eigenvalues.

REMARK 1. We can also consider an alternative two-stage procedure to estimate d and $\mathbf{W}$ in Step 2. Notice that $\vec{\mathbf{Y}}_t = (\mathbf{B} \odot \mathbf{A})\mathbf{x}_t + \vec{\boldsymbol{\varepsilon}}_t$ for $t \geq 1$. We can firstly obtain the estimates of d and $\mathbf{T}$ (the orthogonal basis of $\mathcal{M}(\mathbf{B} \odot \mathbf{A})$), denoted by $\hat{d}$ and $\hat{\mathbf{T}}$, based on the method proposed in Lam, Yao and Bathia (2011), Lam and Yao (2012) and Chang, Guo and Yao (2015). Recall $\mathbf{V} \odot \mathbf{U} = (\mathbf{Q} \otimes \mathbf{P})^\top (\mathbf{B} \odot \mathbf{A})$ and $\mathbf{W}$ is an orthogonal basis of $\mathcal{M}(\mathbf{V} \odot \mathbf{U})$. Based on $(\hat{\mathbf{P}}, \hat{\mathbf{Q}})$, the estimates of $\mathbf{P}$ and $\mathbf{Q}$ obtained in Step 1, we can then estimate $\mathbf{W}$ by $(\hat{\mathbf{Q}} \otimes \hat{\mathbf{P}})^\top \hat{\mathbf{T}}$. Figure S1 in the supplementary material shows that, although this alternative two-stage approach yields estimation errors nearly identical to those of our proposed method, it is considerably more computationally expensive when p and q are large. This is because the first stage of this alternative approach requires an eigen-decomposition of a $(pq) \times (pq)$ matrix defined based on the sample auto-covariance matrices of $\{\vec{\mathbf{Y}}_t\}_{t=1}^n$, whereas Step 2 of our proposed method only involves an eigen-decomposition of a $(\hat{d}_1 \hat{d}_2) \times (\hat{d}_1 \hat{d}_2)$ matrix. This indicates that our proposed method can significantly reduce the computational complexity, especially in high-dimensional settings.

4.3. *Estimating $\boldsymbol{\Theta}$ via joint diagonalization.* For 4-way tensor $\boldsymbol{\Psi}(\cdot, \cdot)$ defined in Section 3.3, we define a $(\hat{d}_1^2 \hat{d}_2^2) \times \hat{d}(\hat{d}+1)/2$ matrix $\hat{\boldsymbol{\Omega}}$ as follows:

$$\hat{\boldsymbol{\Omega}} = \left(\vec{\boldsymbol{\Psi}}(\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_1), \ldots, \vec{\boldsymbol{\Psi}}(\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_{\hat{d}}), \vec{\boldsymbol{\Psi}}(\hat{\mathbf{W}}_2, \hat{\mathbf{W}}_2), \ldots, \vec{\boldsymbol{\Psi}}(\hat{\mathbf{W}}_{\hat{d}}, \hat{\mathbf{W}}_{\hat{d}}) \right), \tag{20}$$

which is an estimate of $\mathbf{\Omega}$ defined as in (12). Let

$$\tilde{\mathbf{h}}_m = (\tilde{h}^m_{1,1}, \ldots, \tilde{h}^m_{1,\hat{d}}, \tilde{h}^m_{2,2}, \ldots, \tilde{h}^m_{\hat{d},\hat{d}})^\top, \quad m \in [\hat{d}], \tag{21}$$

be the right-singular vectors of $\hat{\mathbf{\Omega}}$ corresponding to the $\hat{d}$ smallest singular values. Such selected $\{\tilde{\mathbf{h}}_m\}_{m=1}^{\hat{d}}$ provides the estimate for a basis of $\ker(\mathbf{\Omega})$. By Proposition 5, an estimator for $\mathbf{\Theta}$ can be obtained by the joint diagonalization of $\tilde{\mathbf{H}}_1, \ldots, \tilde{\mathbf{H}}_{\hat{d}}$, which are constructed in the same manner as $\mathbf{H}_m$ with $\mathbf{h}_m$ replaced by $\tilde{\mathbf{h}}_m$. See the statement below (13).

Though $\mathbf{\Theta}$ can be uniquely identified by any set of basis $\{\mathbf{h}_m\}_{m=1}^d$ of $\ker(\mathbf{\Omega})$ (see Proposition 5), the accuracy of its estimator depends on the choice of $\{\mathbf{h}_m\}_{m=1}^d$ sensitively. Motivated by Proposition 8 at the end of this section, a good choice is to rotate the basis vectors $\{\tilde{\mathbf{h}}_m\}_{m=1}^{\hat{d}}$ in (21) first. More specifically, let $(\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_{\hat{d}}) = (\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_{\hat{d}})\hat{\mathbf{\Pi}}$ with

$$\hat{\mathbf{\Pi}} = \{2(\hat{\mathbf{\Upsilon}}_0^\top \hat{\mathbf{\Upsilon}}_2)(\hat{\mathbf{\Upsilon}}_1^\top \hat{\mathbf{\Upsilon}}_2 + \hat{\mathbf{\Upsilon}}_2^\top \hat{\mathbf{\Upsilon}}_1)^{-1}(\hat{\mathbf{\Upsilon}}_2^\top \hat{\mathbf{\Upsilon}}_0)\}^{-1/2},$$

where

$$\hat{\mathbf{\Upsilon}}_0 = \big(\mathrm{vec}(\tilde{\mathbf{H}}_1), \ldots, \mathrm{vec}(\tilde{\mathbf{H}}_{\hat{d}})\big), \quad \hat{\mathbf{\Upsilon}}_1 = \big(\mathrm{vec}(\tilde{\mathbf{H}}^{-1}\tilde{\mathbf{H}}_1), \ldots, \mathrm{vec}(\tilde{\mathbf{H}}^{-1}\tilde{\mathbf{H}}_{\hat{d}})\big),$$

$$\hat{\mathbf{\Upsilon}}_2 = \big(\mathrm{vec}(\tilde{\mathbf{H}}_1\tilde{\mathbf{H}}^{-1}), \ldots, \mathrm{vec}(\tilde{\mathbf{H}}_{\hat{d}}\tilde{\mathbf{H}}^{-1})\big), \quad \tilde{\mathbf{H}} = \sum_{m=1}^{\hat{d}} \phi_m \tilde{\mathbf{H}}_m$$

for some $\hat{d}$-dimensional vector $(\phi_1, \ldots, \phi_{\hat{d}})^\top \neq \mathbf{0}$ such that $\tilde{\mathbf{H}}$ is invertible. Section 7.1 specifies how to select $(\phi_1, \ldots, \phi_{\hat{d}})^\top$ in practice. Define $\hat{\mathbf{H}}_1, \ldots, \hat{\mathbf{H}}_{\hat{d}}$ in the same manner as $\mathbf{H}_m$ but with replacing $\mathbf{h}_m$ by $\hat{\mathbf{h}}_m$. Utilizing the fast Frobenius diagonalization algorithm introduced by Ziehe et al. (2004), we can obtain the non-orthogonal joint diagonalizer $\mathbf{\Phi}$ for $\hat{\mathbf{H}}_1, \ldots, \hat{\mathbf{H}}_{\hat{d}}$ such that all the columns of $\mathbf{\Phi}^{-1}$ are unit vectors. Then, $\mathbf{\Theta}$ involved in (8) can be estimated by $\hat{\mathbf{\Theta}} = \mathbf{\Phi}^{-1}$.

Now we give some illustrations on how the set of basis $\{\mathbf{h}_m\}_{m=1}^d$ of $\ker(\mathbf{\Omega})$ used to identify $\mathbf{\Theta}$ affects the convergence rate of the associated estimator of $\mathbf{\Theta}$ based on the non-orthogonal joint diagonalization. Recall

$$\mathbf{H}_m = \mathbf{\Theta}\,\mathrm{diag}(\gamma_{1,m}, \ldots, \gamma_{d,m})\,\mathbf{\Theta}^\top, \quad m \in [d].$$

By Theorem 3 of Afsari (2008), the convergence rate of the estimator for $\mathbf{\Theta}$ based on the fast Frobenius diagonalization algorithm is bounded by

$$\frac{\eta(\mathbf{h}_1, \ldots, \mathbf{h}_d)}{1 - \rho^2(\mathbf{h}_1, \ldots, \mathbf{h}_d)} \times K_n(d, \mathbf{\Theta}),$$

where $K_n(d, \mathbf{\Theta})$ is a universal quantity only depending on $(n, d, \mathbf{\Theta})$, and

$$\rho(\mathbf{h}_1, \ldots, \mathbf{h}_d) = \max_{k,\ell \in [d]:\, k \neq \ell} \frac{|\sum_{m=1}^d \gamma_{k,m}\gamma_{\ell,m}|}{(\sum_{m=1}^d \gamma_{k,m}^2)^{1/2}(\sum_{m=1}^d \gamma_{\ell,m}^2)^{1/2}},$$

$$\eta(\mathbf{h}_1, \ldots, \mathbf{h}_d) = \max_{k,\ell \in [d]:\, k \neq \ell} \left(\frac{1}{\sum_{m=1}^d \gamma_{\ell,m}^2} + \frac{1}{\sum_{m=1}^d \gamma_{k,m}^2}\right). \tag{22}$$

Ideally we should choose $\{\mathbf{h}_m\}_{m=1}^d$ such that $\rho(\mathbf{h}_1, \ldots, \mathbf{h}_d) = 0$ and $\eta(\mathbf{h}_1, \ldots, \mathbf{h}_d)$ as small as possible. For any given set of basis $\{\mathbf{h}_m\}_{m=1}^d$ of $\ker(\mathbf{\Omega})$, Proposition 8 indicates that we should replace $\{\mathbf{h}_m\}_{m=1}^d$ by its rotation $\{\mathbf{h}^*_m\}_{m=1}^d$ such that $(\mathbf{h}^*_1, \ldots, \mathbf{h}^*_d) = (\mathbf{h}_1, \ldots, \mathbf{h}_d)\mathbf{\Pi}$ with

$$\mathbf{\Pi} = \{2(\mathbf{\Upsilon}_0^\top \mathbf{\Upsilon}_2)(\mathbf{\Upsilon}_1^\top \mathbf{\Upsilon}_2 + \mathbf{\Upsilon}_2^\top \mathbf{\Upsilon}_1)^{-1}(\mathbf{\Upsilon}_2^\top \mathbf{\Upsilon}_0)\}^{-1/2},$$

where

$$\boldsymbol{\Upsilon}_0 = \big(\mathrm{vec}(\mathbf{H}_1), \ldots, \mathrm{vec}(\mathbf{H}_d)\big)\,, \quad \boldsymbol{\Upsilon}_1 = \big(\mathrm{vec}(\mathbf{H}^{-1}\mathbf{H}_1), \ldots, \mathrm{vec}(\mathbf{H}^{-1}\mathbf{H}_d)\big)\,,$$

$$\boldsymbol{\Upsilon}_2 = \big(\mathrm{vec}(\mathbf{H}_1\mathbf{H}^{-1}), \ldots, \mathrm{vec}(\mathbf{H}_d\mathbf{H}^{-1})\big)\,, \quad \mathbf{H} = \sum_{m=1}^{d} \phi_m \mathbf{H}_m \tag{23}$$

for some d-dimensional vector $(\phi_1, \ldots, \phi_d)^\top \neq \mathbf{0}$ such that $\mathbf{H}$ is invertible.

PROPOSITION 8. *$\rho(\mathbf{h}_1^*, \ldots, \mathbf{h}_d^*) = 0$ and $\eta(\mathbf{h}_1^*, \ldots, \mathbf{h}_d^*) = 2$.*

5. Prediction. Given observations $\{\mathbf{Y}_t\}_{t=1}^n$, we can also use the matrix CP-factor model (1) to forecast the future values $\mathbf{Y}_{n+h}$ for $h \geq 1$. More specifically, we can predict $\mathbf{Y}_{n+h}$ by recovering the latent process $\{\mathbf{X}_t\}_{t=1}^n$. Let $\hat{\mathbf{L}} = \hat{\mathbf{B}} \odot \hat{\mathbf{A}}$ with $\hat{\mathbf{A}} \in \mathbb{R}^{p\times\hat{d}}$ and $\hat{\mathbf{B}} \in \mathbb{R}^{q\times\hat{d}}$ being, respectively, the estimates of the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ in the matrix CP-factor model (1). If $\mathrm{rank}(\hat{\mathbf{L}}) = \hat{d}$, we can recover $\mathbf{X}_t$ by $\hat{\mathbf{X}}_t = \mathrm{diag}(\hat{\mathbf{x}}_t)$ with $\hat{\mathbf{x}}_t = \hat{\mathbf{L}}^+\mathbf{Y}_t = (\hat{x}_{t,1}, \ldots, \hat{x}_{t,\hat{d}})^\top$. In order to predict $\mathbf{Y}_{n+h}$, we only need to fit a $\hat{d}$-dimensional multivariate time series model for $\{\hat{\mathbf{x}}_t\}_{t=1}^n$. Then we can predict $\mathbf{Y}_{n+h}$ by $\hat{\mathbf{Y}}_{n+h} = \hat{\mathbf{A}}\tilde{\hat{\mathbf{X}}}_{n+h}\hat{\mathbf{B}}^\top$ with $\tilde{\hat{\mathbf{X}}}_{n+h} = \mathrm{diag}(\tilde{\hat{\mathbf{x}}}_{n+h})$, where $\tilde{\hat{\mathbf{x}}}_{n+h}$ is the h-step ahead forecast of $\hat{\mathbf{x}}_{n+h}$ based on the fitted model for $\{\hat{\mathbf{x}}_t\}_{t=1}^n$. Chang et al. (2023) uses this idea to predict $\mathbf{Y}_{n+h}$ under the assumption $d_1 = d_2 = d$ based on the CP-refined estimate considered there for $(\mathbf{A}, \mathbf{B})$. Since the CP-refined estimate (Chang et al., 2023) does not work if the assumption $d_1 = d_2 = d$ is not satisfied, we cannot select $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as the CP-refined estimate to recover $\mathbf{X}_t$ in these cases. When $d_1 = d_2 = d$ is not satisfied, if the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ can be uniquely identified, we can select $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as our newly proposed estimate specified in Section 4, and use the same idea to predict $\mathbf{Y}_{n+h}$.

As shown in Proposition 7, if $\mathrm{rank}(\boldsymbol{\Omega}) < d(d-1)/2$, the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ cannot be uniquely identified, which implies that we cannot recover $\mathbf{X}_t$ successfully. Hence, above mentioned strategy for predicting $\mathbf{Y}_{n+h}$ does not always work. A natural question is that whether we can propose a unified prediction procedure for $\mathbf{Y}_{n+h}$ based on the matrix CP-factor model (1) without any assumption on the relationship among d_1, d_2 and d. By (5) and (6), we have

$$\mathbf{Y}_t = \mathbf{P}\mathbf{Z}_t\mathbf{Q}^\top + \underbrace{\boldsymbol{\varepsilon}_t - \mathbf{P}\mathbf{P}^\top\boldsymbol{\varepsilon}_t\mathbf{Q}\mathbf{Q}^\top}_{\text{white noise}} \tag{24}$$

with $\mathbf{Z}_t = \mathbf{P}^\top\mathbf{Y}_t\mathbf{Q}$. In order to predict $\mathbf{Y}_{n+h}$, we only need to predict $\mathbf{Z}_{n+h}$. For $(\hat{\mathbf{P}}, \hat{\mathbf{Q}}, \hat{\mathbf{W}})$ specified in Sections 4.1 and 4.2, we define

$$\hat{\mathbf{x}}_t^* = \hat{\mathbf{W}}^\top\mathrm{vec}(\hat{\mathbf{P}}^\top\mathbf{Y}_t\hat{\mathbf{Q}})\,, \quad t \in [n]\,.$$

Proposition 9 in Section 6 indicates that such defined $\hat{d}$-dimensional vector $\hat{\mathbf{x}}_t^*$ provides a recovery of $\mathbf{E}_3\mathbf{x}_t^*$, where $\mathbf{x}_t^* = \boldsymbol{\Theta}\mathbf{x}_t$, and $\mathbf{E}_3$ is an orthogonal matrix specified in Proposition 9. Hence, we can fit a $\hat{d}$-dimensional vector time series model for $\{\hat{\mathbf{x}}_t^*\}_{t=1}^n$. Let $\tilde{\hat{\mathbf{x}}}_{n+h}^*$ be the h-step ahead forecast of $\hat{\mathbf{x}}_{n+h}^*$. By (7) and Proposition 9, we know $\hat{\mathbf{W}}\tilde{\hat{\mathbf{x}}}_{n+h}^*$ provides a prediction of $(\mathbf{E}_2 \otimes \mathbf{E}_1)\vec{\mathbf{Z}}_{n+h}$, where $\mathbf{E}_1$ and $\mathbf{E}_2$ are two orthogonal matrices specified in Proposition 9. Let $\hat{\mathbf{Z}}_{n+h}$ satisfy $\mathrm{vec}(\hat{\mathbf{Z}}_{n+h}) = \hat{\mathbf{W}}\tilde{\hat{\mathbf{x}}}_{n+h}^*$. Applying Proposition 9 again, by (24), we know $\hat{\mathbf{Y}}_{n+h} = \hat{\mathbf{P}}\hat{\mathbf{Z}}_{n+h}\hat{\mathbf{Q}}^\top$ provides a prediction of $\mathbf{Y}_{n+h}$. This new prediction idea only depends on the calculation of three matrices $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ and $\hat{\mathbf{W}}$. As we have discussed in Sections 4.1 and 4.2, determining $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ and $\hat{\mathbf{W}}$ only involves the spectral decomposition

of $\hat{\mathbf{M}}_1$, $\hat{\mathbf{M}}_2$ and $\hat{\mathbf{M}}$, respectively, which does not require any additional assumption on the relationship among d_1, d_2 and d. Hence, our newly proposed prediction strategy provides a unified prediction procedure for $\mathbf{Y}_{n+h}$ based on the matrix CP-factor model (1) regardless of the relationship among d_1, d_2 and d. When the linear dynamic structure is concerned for the latent process $\mathbf{X}_t$, our numerical studies in Section 7.2 indicate that if the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ can be uniquely identified, the finite-sample performance of our newly proposed prediction method is almost identical to the prediction idea considered in Chang et al. (2023) with selecting $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as our proposed estimate of $(\mathbf{A}, \mathbf{B})$ specified in Section 4. However, if the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ cannot be uniquely identified, our newly proposed prediction method outperforms that of Chang et al. (2023).

6. 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

$$\alpha(k) = \sup_{r} \sup_{A \in \mathcal{F}_{-\infty}^{r}, B \in \mathcal{F}_{r+k}^{\infty}} |\mathbb{P}(AB) - \mathbb{P}(A)\mathbb{P}(B)|, \quad k \geq 1, \tag{25}$$

where $\mathcal{F}_r^s$ is the σ-filed generated by $\{\mathbf{Y}_t : r \leq t \leq s\}$. Write

$$\mathbf{\Sigma}_{\vec{\mathbf{Y}}}(k) = \frac{1}{n-k} \sum_{t=k+1}^{n} \mathbb{E}[\{\vec{\mathbf{Y}}_t - \mathbb{E}(\bar{\vec{\mathbf{Y}}})\}\{\vec{\mathbf{Y}}_{t-k} - \mathbb{E}(\bar{\vec{\mathbf{Y}}})\}^\top], \quad k \geq 1,$$

where $\bar{\vec{\mathbf{Y}}} = n^{-1}\sum_{t=1}^{n} \vec{\mathbf{Y}}_t$. Define $\mathbf{M} = \sum_{k=1}^{\tilde{K}} \mathbf{\Sigma}_{\vec{\mathbf{Z}}}(k)\mathbf{\Sigma}_{\vec{\mathbf{Z}}}(k)^\top$ with $\tilde{K}$ given in (17) and

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

where $\bar{\vec{\mathbf{Z}}} = n^{-1}\sum_{t=1}^{n} \vec{\mathbf{Z}}_t$. Following the arguments in Chang, Guo and Yao (2015), we can identify d as $d = \text{rank}(\mathbf{M})$, and select $\vec{\mathbf{W}}_1, \ldots, \vec{\mathbf{W}}_d$ involved in (9) as the d orthonormal eigenvectors of $\mathbf{M}$ corresponding to the d non-zero eigenvalues $\lambda_1(\mathbf{M}) \geq \cdots \geq \lambda_d(\mathbf{M}) > 0$, i.e., $\vec{\mathbf{W}}_\ell$ is the eigenvector associated with the eigenvalue $\lambda_\ell(\mathbf{M})$ for $\ell \in [d]$. We need the following regularity conditions in our theoretical analysis.

CONDITION 3. (i) *The nonzero singular values of* $\mathbf{B} \odot \mathbf{A}$ *are uniformly bounded away from zero.* (ii) *The nonzero eigenvalues of* $\mathbf{M}_1$, $\mathbf{M}_2$ *and* $\mathbf{M}$ *are uniformly bounded away from zero.*

CONDITION 4. (i) *There exist some universal constants* $K_1 > 0$, $K_2 > 0$ *and* $r_1 \in (0, 2]$ *such that* $\mathbb{P}(|y_{i,j,t}| > x) \leq K_1 \exp(-K_2 x^{r_1})$ *and* $\mathbb{P}(|\xi_t| > x) \leq K_1 \exp(-K_2 x^{r_1})$ *for any* $x > 0$, $i \in [p]$, $j \in [q]$ *and* $t \in [n]$. (ii) *There exist some universal constants* $K_3 > 0$, $K_4 > 0$ *and* $r_2 \in (0, 1]$ *such that the* α*-mixing coefficients* $\alpha(k)$ *defined as in* (25) *satisfy* $\alpha(k) \leq K_3 \exp(-K_4 k^{r_2})$ *for* $k \geq 1$.

CONDITION 5. (i) *There exists a universal constant* $K_5 > 0$ *such that* $\|\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\|_2 \leq K_5$ *for any* $k \in [K]$, *and* $\|\mathbf{\Sigma}_{\vec{\mathbf{Y}}}(k)\|_2 \leq K_5$ *for any* $k \in [\tilde{K}]$. (ii) *Write* $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k) = (\sigma^{(k)}_{y,\xi,i,j})_{p\times q}$ *and* $\mathbf{\Sigma}_{\vec{\mathbf{Y}}}(k) = (\sigma^{(k)}_{i,j})_{pq\times pq}$. *There exists a universal constant* $\iota \in [0, 1)$ *such that* $\sum_{j_1=1}^{q} |\sigma^{(k)}_{y,\xi,i_1,j_1}|^\iota \leq s_1$, $\sum_{i_1=1}^{p} |\sigma^{(k)}_{y,\xi,i_1,j_1}|^\iota \leq s_2$, $\sum_{j_2=1}^{pq} |\sigma^{(k)}_{i_2,j_2}|^\iota \leq s_3$ *and* $\sum_{i_2=1}^{pq} |\sigma^{(k)}_{i_2,j_2}|^\iota \leq s_4$ *for any* $i_1 \in [p]$, $j_1 \in [q]$ *and* $i_2, j_2 \in [pq]$, *where* s_1, s_2, s_3 *and* s_4 *may, respectively, diverge together with* p *and* q.

Condition 3 is used to simplify the presentation for the results. Our technical proofs indeed allow the nonzero singular values of $\mathbf{B} \odot \mathbf{A}$, and the nonzero eigenvalues of $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}$ decay to zero as p and/or q grow to infinity. Condition 4 is also used in Chang et al. (2023), which is a common assumption in the literature on ultrahigh-dimensional data analysis. See Chang et al. (2023) for the discussion of their validity. We impose Condition 5(i) just for simplifying the presentation. Our technical proofs indeed allow $\max_{k\in[K]} \|\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)\|_2$ and $\max_{k\in[K]} \|\mathbf{\Sigma}_{\vec{\mathbf{Y}}}(k)\|_2$ to diverge as p and/or q grow to infinity. Condition 5(ii) imposes some sparsity requirement on $\mathbf{\Sigma}_{\mathbf{Y},\xi}(k)$ and $\mathbf{\Sigma}_{\vec{\mathbf{Y}}}(k)$. Under some sparsity condition on $\mathbf{A}$ and $\mathbf{B}$, applying the technique used to derive Lemma 5 of Chang, Guo and Yao (2018), we can show that Condition 5(ii) holds for certain (s_1, s_2, s_3, s_4). Let

$$\Pi_{1,n} = (s_1 s_2)^{1/2} \left\{ \frac{\log(pq)}{n} \right\}^{(1-\iota)/2} \quad \text{and} \quad \Pi_{2,n} = (s_3 s_4)^{1/2} \left\{ \frac{\log(pq)}{n} \right\}^{(1-\iota)/2}. \tag{26}$$

Theorem 1 shows that the eigenvalue-ratio based estimators $\hat{d}_1$, $\hat{d}_2$ and $\hat{d}$ provide consistent estimates for d_1, d_2 and d, respectively.

THEOREM 1. *Let Conditions 1–5 hold. Select the threshold levels in (14) and (18) as*

$$\delta_1 = \breve{C}\sqrt{\frac{\log(pq)}{n}} \ \text{ and } \ \delta_2 = \tilde{C}\sqrt{\frac{\log(pq)}{n}}$$

for some sufficiently large constants $\breve{C}, \tilde{C} > 0$. For any $(c_{1,n}, c_{2,n}, c_{3,n})$ given in (16) and (19) satisfying $\Pi_{1,n} \ll c_{1,n}, c_{2,n} \ll 1$ and $\max(\Pi_{1,n}, \Pi_{2,n}) \ll c_{3,n} \ll 1$, it holds that

$$\mathbb{P}(\hat{d}_1 = d_1) \to 1\,, \ \mathbb{P}(\hat{d}_2 = d_2) \to 1 \ \text{ and } \ \mathbb{P}(\hat{d} = d) \to 1$$

as $n \to \infty$, provided that $\Pi_{1,n} + \Pi_{2,n} \ll 1$ and $\log(pq) \ll n^c$ for some constant $c \in (0,1)$ depending only on r_1 and r_2.

Proposition 9 states the asymptotic performance of $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ and $\hat{\mathbf{W}}$.

PROPOSITION 9. *Let Conditions 1–5 hold. Select the threshold levels in (14) and (18) as*

$$\delta_1 = \breve{C}\sqrt{\frac{\log(pq)}{n}} \ \text{ and } \ \delta_2 = \tilde{C}\sqrt{\frac{\log(pq)}{n}}$$

for some sufficiently large constants $\breve{C}, \tilde{C} > 0$. Assume that $\Pi_{1,n} + \Pi_{2,n} \ll 1$ and $\log(pq) \ll n^c$ for some constant $c \in (0,1)$ depending only on r_1 and r_2. If $(\hat{d}_1, \hat{d}_2) = (d_1, d_2)$, there exist some orthogonal matrices $\mathbf{E}_1 \in \mathbb{R}^{d_1 \times d_1}$ and $\mathbf{E}_2 \in \mathbb{R}^{d_2 \times d_2}$ such that

$$\|\hat{\mathbf{P}}\mathbf{E}_1 - \mathbf{P}\|_2 = O_{\mathrm{p}}(\Pi_{1,n}) = \|\hat{\mathbf{Q}}\mathbf{E}_2 - \mathbf{Q}\|_2\,.$$

Furthermore, if $(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)$, there exists an orthogonal matrix $\mathbf{E}_3 \in \mathbb{R}^{d \times d}$ such that

$$\|(\mathbf{E}_2 \otimes \mathbf{E}_1)^\top \hat{\mathbf{W}}\mathbf{E}_3 - \mathbf{W}\|_2 = O_{\mathrm{p}}(\Pi_{1,n} + \Pi_{2,n})\,.$$

If the nonzero eigenvalues of $\mathbf{M}$ are distinct, $\mathbf{E}_3$ will be a diagonal matrix with its diagonal elements being 1 or -1. For the trivial case $d = 1$, if $(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)$, we have $\mathbf{E}_1 = \pm 1$ and $\mathbf{E}_2 = \pm 1$. Following the discussion below (5), it holds in this trivial case that $|\hat{\mathbf{A}}\mathbf{E}_1 - \mathbf{A}|_2 = O_{\mathrm{p}}(\Pi_{1,n}) = |\hat{\mathbf{B}}\mathbf{E}_2 - \mathbf{B}|_2$ provided that $\Pi_{1,n} \ll 1$ and $\log(pq) \ll n^c$ for some constant $c \in (0,1)$ depending only on r_1 and r_2. Note that $\mathbb{P}\{(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)\} \to 1$ as $n \to \infty$.

Hence, in the trivial case $d = 1$, $(\mathbf{A}, \mathbf{B})$ can be consistently estimated up to the reflection indeterminacy. For the non-trivial case $d \geq 2$, the convergence rates of the estimation errors for $\mathbf{A}$ and $\mathbf{B}$ will be shown in Theorem 2.

To present Theorem 2, we need to introduce some notation first. For $(\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3)$ specified in Proposition 9, let $\breve{\mathbf{W}} = (\mathbf{E}_2 \otimes \mathbf{E}_1)\mathbf{W}\mathbf{E}_3^\top$. Define $\breve{\mathbf{\Omega}}$ in the same manner as $\mathbf{\Omega}$ given in (12) but with replacing $\mathbf{W}$ by $\breve{\mathbf{W}}$. Following the discussions of Propositions 5–7, the requirement $\mathrm{rank}(\mathbf{\Omega}) = d(d-1)/2$ is crucial for the identification of $(\mathbf{A}, \mathbf{B})$ when $d \geq 2$. As shown in Section D.3 in the supplementary material for the proof of Lemma 4, we know $\mathrm{rank}(\breve{\mathbf{\Omega}}) = \mathrm{rank}(\mathbf{\Omega})$. By Proposition 4(i), it holds that $\mathrm{rank}(\mathbf{\Omega}) = d(d-1)/2$ if and only if $\lambda_{d(d-1)/2}(\breve{\mathbf{\Omega}}^\top \breve{\mathbf{\Omega}}) > 0$. Note that $\breve{\mathbf{\Omega}}^\top \breve{\mathbf{\Omega}}$ is a $\{d(d+1)/2\} \times \{d(d+1)/2\}$ matrix. We require the following mild condition in our theoretical analysis.

CONDITION 6. *$\lambda_{d(d-1)/2}(\breve{\mathbf{\Omega}}^\top \breve{\mathbf{\Omega}})$ is uniformly bounded away from zero.*

Write $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_{\hat{d}})$ and $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_{\hat{d}})$, where $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are specified in Section 4. Recall $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$. Theorem 2 indicates that the columns of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are, respectively, consistent to those of $\mathbf{A}$ and $\mathbf{B}$ up to the reflection and permutation indeterminacy.

THEOREM 2. *Let $d \geq 2$ and Conditions 1–6 hold. Select the threshold levels in (14) and (18) as*

$$\delta_1 = \breve{C}\sqrt{\frac{\log(pq)}{n}} \quad \text{and} \quad \delta_2 = \tilde{C}\sqrt{\frac{\log(pq)}{n}}$$

for some sufficiently large constants $\breve{C}, \tilde{C} > 0$. If $(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)$, there exists a permutation of $(1, \ldots, d)$, denoted by $(j_1, \ldots, j_d)$, such that

$$\max_{\ell \in [d]} |\kappa_{1,\ell}\hat{\mathbf{a}}_{j_\ell} - \mathbf{a}_\ell|_2 = O_{\mathrm{p}}(\Pi_{1,n} + \Pi_{2,n}) = \max_{\ell \in [d]} |\kappa_{2,\ell}\hat{\mathbf{b}}_{j_\ell} - \mathbf{b}_\ell|_2$$

with some $\kappa_{1,\ell}, \kappa_{2,\ell} \in \{1, -1\}$, provided that $\Pi_{1,n} + \Pi_{2,n} \ll 1$ and $\log(pq) \ll n^c$ for some constant $c \in (0, 1)$ depending only on r_1 and r_2.

REMARK 2. The convergence rates of the estimates for $\mathbf{a}_\ell$ and $\mathbf{b}_\ell$ suggested in Chang et al. (2023) are, respectively, $(1 + \vartheta_\ell^{-1}) \cdot O_{\mathrm{p}}(\tilde{\Pi}_{1,n} + \tilde{\Pi}_{2,n})$ and $\{1 + (\vartheta_\ell^*)^{-1}\} \cdot O_{\mathrm{p}}(\tilde{\Pi}_{1,n} + \tilde{\Pi}_{2,n})$, where ϑ_ℓ and ϑ_ℓ^* are the eigen-gaps defined as in Equation (38) of Chang et al. (2023), $\tilde{\Pi}_{1,n} = \Pi_{1,n}$, and $\tilde{\Pi}_{2,n} = (\tilde{s}_3\tilde{s}_4)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}$. Here, $\tilde{s}_3$ and $\tilde{s}_4$ control the sparsity of the matrix

$$\mathbf{\Sigma}_{\mathring{\mathbf{Y}}}(k) = \frac{1}{n-k}\sum_{t=k+1}^{n} \mathbb{E}[\{\mathbf{Y}_t - \mathbb{E}(\bar{\mathbf{Y}})\} \otimes \mathrm{vec}\{\mathbf{Y}_{t-k} - \mathbb{E}(\bar{\mathbf{Y}})\}] =: \big(\sigma^{(k)}_{\mathring{y},r,s}\big)_{(p^2q)\times q}$$

in the sense that $\sum_{s=1}^{q} |\sigma^{(k)}_{\mathring{y},r,s}|^\iota \leq \tilde{s}_3$ and $\sum_{r=1}^{p^2q} |\sigma^{(k)}_{\mathring{y},r,s}|^\iota \leq \tilde{s}_4$ for any $r \in [p^2q]$ and $s \in [q]$. Recall $\Pi_{2,n} = (s_3 s_4)^{1/2}\{n^{-1}\log(pq)\}^{(1-\iota)/2}$ with (s_3, s_4) specified in Condition 5(ii). By direct calculation, we have $s_3 \leq p\tilde{s}_3$ and $\tilde{s}_4 \leq ps_4$. Under some mild conditions, it holds that $s_3 s_4 \asymp \tilde{s}_3\tilde{s}_4$, which implies $\tilde{\Pi}_{2,n} \asymp \Pi_{2,n}$. Hence, if ϑ_ℓ and ϑ_ℓ^* are uniformly bounded away from zero, Theorem 2 indicates that our new estimators share the same convergence rates of those proposed in Chang et al. (2023). If $\vartheta_\ell \to 0$ or $\vartheta_\ell^* \to 0$, our new estimators will have faster convergence rates than the estimators considered in Chang et al. (2023).

REMARK 3. The model considered in Han et al. (2024b) for order 2 tensor is in the same form as our CP-factor model (1). Therefore, the two estimation procedures proposed in Han et al. (2024b), the composite PCA (denoted by cPCA) and the High-Order Projection Estimators (denoted by HOPE), can also be used to estimate the loading matrices $\mathbf{A}$ and $\mathbf{B}$ in our CP-factor model (1), where cPCA is a one-pass estimation and HOPE is an iterative refinement initialized at the cPCA solution. Han et al. (2024b) assumes each latent factor $x_{t,\ell} = w_\ell f_{t,\ell}$ where $\{f_{t,\ell}\}_{t\geq 1}$ is stationary with $\mathbb{E}(f_{t,\ell}^2) = 1$, and w_ℓ represents the signal strength. Under the model setting of Han et al. (2024b), the latent factor process $\{x_{t,\ell}\}_{t\geq 1}$ is stationary for each $\ell \in [d]$. Moreover, Han et al. (2024b) also assumes $\mathbb{E}(f_{t-h,\ell_1} f_{t,\ell_2}) = 0$ for all $\ell_1 \neq \ell_2$ and $h \geq 1$, which implies $\mathbb{E}(x_{t-h,\ell_1} x_{t,\ell_2}) = 0$ for all $\ell_1 \neq \ell_2$ and $h \geq 1$. However, these assumptions imposed on the latent factors are not necessary in our proposed method. Write $\delta = \|(\mathbf{B}\odot\mathbf{A})^\top(\mathbf{B}\odot\mathbf{A}) - \mathbf{I}_d\|_2$, $\psi_\ell = w_\ell^2 \mathbb{E}(f_{t-h,\ell} f_{t,\ell})$ with some fixed lag $h \geq 1$, and $\psi_* = \min_{\ell\in[d+1]}(\psi_{\ell-1} - \psi_\ell)$ with $\psi_0 = \infty$ and $\psi_{d+1} = 0$. To simplify the comparison between the theoretical results of Han et al. (2024b) and our proposed method, we ignore the permutation indeterminacy among the estimators. Theorem 1 of Han et al. (2024b) shows that the cPCA estimators $\hat{\mathbf{a}}_1^{\text{cpca}}, \ldots, \hat{\mathbf{a}}_d^{\text{cpca}}, \hat{\mathbf{b}}_1^{\text{cpca}}, \ldots, \hat{\mathbf{b}}_d^{\text{cpca}}$ satisfy

$$\max_{\ell\in[d]}\{1-(\mathbf{a}_\ell^\top \hat{\mathbf{a}}_\ell^{\text{cpca}})^2\}^{1/2} + \max_{\ell\in[d]}\{1-(\mathbf{b}_\ell^\top \hat{\mathbf{b}}_\ell^{\text{cpca}})^2\}^{1/2}$$
$$\lesssim \left(1+\frac{2\psi_1}{\psi_*}\right)\delta + \psi_*^{-1}\left\{\max_{\ell\in[d]} w_\ell^2\sqrt{\frac{\log n}{n}} + \left(1+\max_{\ell\in[d]} w_\ell\right)\sqrt{\frac{pq}{n}}\right\}$$

with probability at least $1-(nd)^{-C_1} - e^{-pq}$, where C_1 is a positive constant. Theorem 2 of Han et al. (2024b) shows that, after a sufficient number of iterations, the HOPE estimators $\hat{\mathbf{a}}_1^{\text{iso}}, \ldots, \hat{\mathbf{a}}_d^{\text{iso}}, \hat{\mathbf{b}}_1^{\text{iso}}, \ldots, \hat{\mathbf{b}}_d^{\text{iso}}$ satisfy

$$\max_{\ell\in[d]}\{1-(\mathbf{a}_\ell^\top \hat{\mathbf{a}}_\ell^{\text{iso}})^2\}^{1/2} + \max_{\ell\in[d]}\{1-(\mathbf{b}_\ell^\top \hat{\mathbf{b}}_\ell^{\text{iso}})^2\}^{1/2} \lesssim (\psi_d^{-1}+\psi_d^{-1/2})\sqrt{\frac{\max(p,q)}{n}}$$

with probability at least $1-(nd)^{-C_2} - e^{-p} - e^{-q}$, provided that the cPCA estimators satisfy certain convergence rates, where C_2 is a positive constant. For our proposed estimators $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_d, \hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_d$, due to $1-(\hat{\mathbf{a}}_\ell^\top \mathbf{a}_\ell)^2 \leq |\kappa_{1,\ell}\hat{\mathbf{a}}_\ell - \mathbf{a}_\ell|_2^2$ and $1-(\hat{\mathbf{b}}_\ell^\top \mathbf{b}_\ell)^2 \leq |\kappa_{2,\ell}\hat{\mathbf{b}}_\ell - \mathbf{b}_\ell|_2^2$ for $\kappa_{1,\ell}, \kappa_{2,\ell} \in \{1,-1\}$, then

$$\max_{\ell\in[d]}\{1-(\mathbf{a}_\ell^\top \hat{\mathbf{a}}_\ell)^2\}^{1/2} + \max_{\ell\in[d]}\{1-(\mathbf{b}_\ell^\top \hat{\mathbf{b}}_\ell)^2\}^{1/2} \lesssim \Pi_{1,n} + \Pi_{2,n}$$

with probability approaching one, where $\Pi_{1,n}$ and $\Pi_{2,n}$ are specified in (26). Hence, the two estimation procedures proposed in Han et al. (2024b) can only work for $pq \ll n$, while our proposed method allows $p, q \gg n$. More importantly, in order to obtain the consistency of the cPCA estimators, we need to require $\mathbf{B}\odot\mathbf{A}$ to be very close to an orthonormal matrix ($\delta \to 0$ as $n\to\infty$). However, such requirement may be too restrictive in practice. The larger δ is, or the smaller ψ_* is, the worse convergence rate of the cPCA estimators will be. Since the HOPE estimators are obtained through an iterative refinement method initialized with the cPCA estimators, the HOPE estimators will perform poorly if the cPCA estimators have large estimation errors. However, the convergence rate of our proposed method does not depend on these quantities.

Theorem 2 requires $(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)$. By Theorem 1, we have $\mathbb{P}\{(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)\} \to 1$ as $n\to\infty$. Hence, such requirement is reasonable in our theoretical analysis. More generally, without assuming $(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)$, we can consider to measure the difference between $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ and $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_{\hat{d}})$ by

$$\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) = \max_{\ell\in[d]} \min_{j\in[\hat{d}]}(1-|\hat{\mathbf{a}}_j^\top \mathbf{a}_\ell|^2)\,. \tag{27}$$

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

$$\varpi^2(\mathbf{B}, \hat{\mathbf{B}}) = \max_{\ell \in [d]} \min_{j \in [\hat{d}]} (1 - |\hat{\mathbf{b}}_j^\top \mathbf{b}_\ell|^2) . \tag{28}$$

Consider the event $\mathcal{G} = \{(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)\}$. Due to $|\hat{\mathbf{a}}_j|_2 = 1 = |\mathbf{a}_\ell|_2$ and $|\kappa_{1,\ell}\hat{\mathbf{a}}_{j_\ell} - \mathbf{a}_\ell|_2^2 \geq 2 - 2|\hat{\mathbf{a}}_{j_\ell}^\top \mathbf{a}_\ell|$ for any $\kappa_{1,\ell} \in \{1, -1\}$, restricted on $\mathcal{G}$, Theorem 2 indicates that $1 - |\hat{\mathbf{a}}_{j_\ell}^\top \mathbf{a}_\ell|^2 \leq 2(1 - |\hat{\mathbf{a}}_{j_\ell}^\top \mathbf{a}_\ell|) = O_{\mathrm{p}}(\Pi_{1,n}^2 + \Pi_{2,n}^2)$ provided that $\Pi_{1,n} + \Pi_{2,n} \ll 1$ and $\log(pq) \ll n^c$ for some constant $c \in (0, 1)$ depending only on r_1 and r_2. Hence, restricted on $\mathcal{G}$, for any $\epsilon > 0$, there exists some constant $C_\epsilon > 0$ such that $\mathbb{P}\{\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) > C_\epsilon(\Pi_{1,n}^2 + \Pi_{2,n}^2) \,|\, \mathcal{G}\} \leq \epsilon$. Together with Theorem 1, we have

$$\begin{aligned}
&\mathbb{P}\{\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) > C_\epsilon(\Pi_{1,n}^2 + \Pi_{2,n}^2)\} \\
&\quad \leq \mathbb{P}\{\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) > C_\epsilon(\Pi_{1,n}^2 + \Pi_{2,n}^2) \,|\, \mathcal{G}\}\, \mathbb{P}(\mathcal{G}) + \mathbb{P}(\mathcal{G}^{\mathrm{c}}) \\
&\quad \leq \mathbb{P}\{\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) > C_\epsilon(\Pi_{1,n}^2 + \Pi_{2,n}^2) \,|\, \mathcal{G}\} + \mathbb{P}(\hat{d}_1 \neq d_1) + \mathbb{P}(\hat{d}_2 \neq d_2) + \mathbb{P}(\hat{d} \neq d) \\
&\quad \leq \epsilon + o(1) \to \epsilon
\end{aligned}$$

as $n \to \infty$, which implies $\varpi^2(\mathbf{A}, \hat{\mathbf{A}}) = O_{\mathrm{p}}(\Pi_{1,n}^2 + \Pi_{2,n}^2)$. Also, we can show $\varpi^2(\mathbf{B}, \hat{\mathbf{B}}) = O_{\mathrm{p}}(\Pi_{1,n}^2 + \Pi_{2,n}^2)$.

7. Numerical studies. In this section, we will evaluate the finite-sample performance of our proposed method by simulation and real data analysis. The simulation setup is given in Section 7.1, and the analysis of the simulation results is presented in Section 7.2. The real data analysis is given in Section 7.3.

7.1. *Setting up.* Let $\mathbf{A}^\dagger \equiv (a_{i,j}^\dagger)_{p \times d}$ and $\mathbf{B}^\dagger \equiv (b_{i,j}^\dagger)_{q \times d}$ with the elements drawn from the uniform distribution on $[-3, 3]$ independently satisfying $\mathrm{rank}(\mathbf{A}^\dagger) = d = \mathrm{rank}(\mathbf{B}^\dagger)$. Define $\mathbf{P} \in \mathbb{R}^{p \times d_1}$ and $\mathbf{Q} \in \mathbb{R}^{q \times d_2}$ such that the columns of $\mathbf{P}$ and $\mathbf{Q}$ are, respectively, the d_1 and d_2 left-singular vectors corresponding to the d_1 and d_2 largest singular values of $\mathbf{A}^\dagger$ and $\mathbf{B}^\dagger$. Let $\mathbf{U}^* = \mathbf{P}^\top \mathbf{A}^\dagger = (\mathbf{u}_1^*, \ldots, \mathbf{u}_d^*)$ and $\mathbf{V}^* = \mathbf{Q}^\top \mathbf{B}^\dagger = (\mathbf{v}_1^*, \ldots, \mathbf{v}_d^*)$. Derive $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ with $\mathbf{u}_j = \mathbf{u}_j^*/|\mathbf{u}_j^*|_2$ and $\mathbf{v}_j = \mathbf{v}_j^*/|\mathbf{v}_j^*|_2$ for any $j \in [d]$. Write $\mathbf{x}_j^* = (x_{1,j}^*, \ldots, x_{n,j}^*)^\top$ and let $\mathbf{x}_1^*, \ldots, \mathbf{x}_d^*$ be d independent AR(1) processes with independent $\mathcal{N}(0, 1)$ innovations, and the autoregressive coefficients drawn from the uniform distribution on $[-0.95, -0.6] \cup [0.6, 0.95]$. Let $\mathbf{X}_t = \mathrm{diag}(x_{t,1}, \ldots, x_{t,d})$ with $x_{t,j} = x_{t,j}^*|\mathbf{v}_j^*|_2|\mathbf{u}_j^*|_2$ for each $t \in [n]$. The elements of the error term $\boldsymbol{\varepsilon}_t$ are drawn from $\mathcal{N}(0, 1)$ independently. Finally, we generate $\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t\mathbf{B}^\top + \boldsymbol{\varepsilon}_t$ for any $t \in [n]$ with $\mathbf{A} = \mathbf{P}\mathbf{U}$ and $\mathbf{B} = \mathbf{Q}\mathbf{V}$. We set $n \in \{300, 600, 900\}$, $d \in \{3, 5, 7\}$ and p, q taking values between 10 and 160. We consider three different scenarios for (d, d_1, d_2):

(R1) Let $d_1 = d_2 = d$. In this scenario, $\mathbf{A}$ and $\mathbf{B}$ are full rank.
(R2) Let $d_1 = d - 1$ and $d_2 = d$. In this scenario, only $\mathbf{B}$ is full rank.
(R3) Let $d_1 = d_2 = d - 1$. In this scenario, both $\mathbf{A}$ and $\mathbf{B}$ are not full rank.

We follow Chang et al. (2023) to specify ξ_t involved in (15). Let $\mathbf{Y} = (\vec{\mathbf{Y}}_1, \ldots, \vec{\mathbf{Y}}_n)^\top$. Perform the principal component analysis for $\mathbf{Y}$ 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. Let $\hat{\sigma}_0^2 = (npq)^{-1}\|\mathbf{Y}\|_{\mathrm{F}}^2$. We set $\delta_1 = \delta_2 = \hat{\sigma}_0\{n^{-1}\log(pq)\}^{1/2}$ in (14) and (18), and set $c_{1,n} = c_{2,n} = c_{3,n} = \hat{\sigma}_0 n^{-1}$ in (16) and (19). We also choose $K = 20$ and $\tilde{K} = 10$ with K and $\tilde{K}$ given in (14) and (17), respectively. Here, using a relatively large value for K is to ensure that $\mathbf{M}_1$ and $\mathbf{M}_2$ defined in (3) satisfy $\mathrm{rank}(\mathbf{M}_1) = d_1$ and $\mathrm{rank}(\mathbf{M}_2) = d_2$. These

two requirements are essential for our proposed method. See Propositions 1 and 2. As shown in (17), $\tilde{K}$ is the number of lags used in the methods of Lam, Yao and Bathia (2011), Lam and Yao (2012) and Chang, Guo and Yao (2015) to estimate the linear space spanned by the columns of the factor loading matrix in the standard factor model. In practice, a small $\tilde{K}$ (i.e., $1 \le \tilde{K} \le 10$) is enough and the estimation results are generally robust to the specific choice of $\tilde{K}$. See our sensitivity analysis with respect to the tuning parameters K and $\tilde{K}$ in Figures S2–S5 of the supplementary material for more details. As mentioned in Section 4.3, we need to select an appropriate constant vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{\hat{d}})^\top$ to ensure that $\tilde{\mathbf{H}} = \sum_{i=1}^{\hat{d}} \phi_i \tilde{\mathbf{H}}_i$ is an invertible matrix with $\tilde{\mathbf{H}}_i$ defined below (21). Let $\phi_i = I\{\sigma_{\hat{d}}(\tilde{\mathbf{H}}_i) = \max_{j\in[\hat{d}]} \sigma_{\hat{d}}(\tilde{\mathbf{H}}_j) > 0\}$ for any $i \in [\hat{d}]$. If $|\boldsymbol{\phi}|_1 = 0$, we randomly generate a unit vector $\boldsymbol{\phi}$ such that $\sigma_{\hat{d}}(\tilde{\mathbf{H}}) > 0$. If $|\boldsymbol{\phi}|_1 \ge 1$, we arbitrarily keep one non-zero element in $\boldsymbol{\phi}$ and set all other elements to zero. The simulation results show that our proposed procedure based on such selected $\boldsymbol{\phi}$ exhibits good finite-sample performance. We also compare our proposed method with the refined method (denoted by CP-refined) introduced by Chang et al. (2023), and the cPCA and the HOPE methods proposed by Han et al. (2024b) with the recommended tuning parameter $h = 1$ therein. All simulations are implemented in R. Our proposed method is available in R-package `HDTSA`, which is implemented by calling the R-function `CP_MTS` with setting `method = 'CP.Unified'`. The CP-refined method of Chang et al. (2023) can also be implemented by calling the R-function `CP_MTS` with setting `method = 'CP.Refined'`. All simulation results are based on 2000 replications.

7.2. *Simulation results* . We first consider the finite-sample performance of the estimation $(\hat{d}_1, \hat{d}_2, \hat{d})$ given in (16) and (19). Note that the CP-refined method of Chang et al. (2023) is developed under the assumption $d_1 = d_2 = d$. To fairly compare our proposed method and the CP-refined method, we compare the relative frequency estimate of $\mathbb{P}_d := \mathbb{P}(\hat{d} = d)$ with $\hat{d}$ specified in (19), and the relative frequency estimate of $\mathbb{P}_c := \mathbb{P}(\hat{d} = d)$ with $\hat{d}$ estimated by the CP-refined method. Note that $\mathbb{P}_{1,2,d} := \mathbb{P}\{(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)\} \le \mathbb{P}_d$ with $(\hat{d}_1, \hat{d}_2, \hat{d})$ estimated by our proposed method. Table 1 indicates that (i) our proposed method outperforms the CP-refined method across Scenarios R1–R3, and (ii) (d_1, d_2, d) can be consistently estimated by our proposed method. To conserve space, we omit the results for $p < q$ in Scenarios R1 and R3, as the symmetry in the data-generating process leads to results that are nearly identical to those obtained when $p > q$.

Figure 1 reports the averages of the estimation errors $\varpi^2(\mathbf{A}, \hat{\mathbf{A}})$ and $\varpi^2(\mathbf{B}, \hat{\mathbf{B}})$ defined in (27) and (28) based on 2000 repetitions across different scenarios. Our proposed method consistently outperforms all competing methods, except in Scenario R1 with $p > q$, where it performs comparably to the HOPE method in estimating $\mathbf{B}$. In contrast, the estimation errors of the CP-refined method are very large in Scenarios R2 and R3, which indicates that the CP-refined method does not work for the matrix CP-factor model (1) with rank-deficient factor loading matrices $\mathbf{A}$ and $\mathbf{B}$. Also, in Scenarios R2 and R3, the HOPE method offers no notable improvement over the cPCA method and even underperforms the cPCA method in some settings, suggesting that the iterative method HOPE is ineffective when the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ are rank-deficient. Additionally, all methods lose efficiency when $(d, d_1, d_2) = (3, 2, 2)$ since $(\mathbf{A}, \mathbf{B})$ cannot be identified uniquely, but our proposed method still yields the smallest estimation errors. The averages and standard deviations of the estimation errors $\varpi^2(\mathbf{A}, \hat{\mathbf{A}})$ and $\varpi^2(\mathbf{B}, \hat{\mathbf{B}})$ based on 2000 repetitions are summarized in Tables S1–S4 in the supplementary material.

Next, we evaluate the finite-sample performance of our proposed prediction method introduced in Section 5. We generate a sequence $\{\mathbf{Y}_t\}_{t=1}^{n+m+1}$ defined in Section 7.1 with $m = 20$. For any $s \in [m]$, we apply our proposed prediction method to the data $\{\mathbf{Y}_t\}_{t=s}^{n+s-1}$ and then,

respectively, obtain the one-step forecast of $\mathbf{Y}_{n+s}$ (denoted by $\hat{\mathbf{Y}}^{(1)}_{n+s}$) and the two-step forecast of $\mathbf{Y}_{n+s+1}$ (denoted by $\hat{\mathbf{Y}}^{(2)}_{n+s+1}$). We also consider the prediction method introduced in Chang et al. (2023) to obtain the one-step ahead forecast of $\mathbf{Y}_{n+s}$ and the two-step ahead forecast of $\mathbf{Y}_{n+s+1}$ using $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ estimated from the data $\{\mathbf{Y}_t\}_{t=s}^{n+s-1}$ for each $s \in [m]$, where $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ can be selected as either (i) our proposed estimate of $(\mathbf{A}, \mathbf{B})$ specified in Section 4, or (ii) the CP-refined estimate of $(\mathbf{A}, \mathbf{B})$ given in Chang et al. (2023). Here, for the obtained univariate time series, we fit it by an autoregressive (AR) model with the order determined by the Akaike information criterion (AIC). For the obtained multivariate time series, we fit it by a vector autoregressive (VAR) model with the order determined by the AIC. Based on 2000 repetitions, Figure 2 plots the averages of the one-step ahead

$$\text{RMSE} := \frac{1}{m\sqrt{pq}} \sum_{s=1}^{m} \|\hat{\mathbf{Y}}^{(1)}_{n+s} - \mathbf{Y}_{n+s}\|_{\text{F}}.$$

It can be observed that (i) in all cases, the finite-sample performance of our newly proposed prediction method is better than the prediction method introduced in Chang et al. (2023) with selecting $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as the CP-refined estimate, (ii) in the cases expect $(d, d_1, d_2) = (3, 2, 2)$, the averages of the one-step ahead RMSE of our newly proposed prediction method are almost identical to those of the prediction method introduced in Chang et al. (2023) with selecting $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as our proposed estimate, and (iii) in the case $(d, d_1, d_2) = (3, 2, 2)$, our newly proposed prediction method outperforms the prediction method introduced in Chang et al. (2023) with selecting $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as our proposed estimate. Note that the factor loading matrices $\mathbf{A}$ and $\mathbf{B}$ cannot be uniquely identified in the case $(d, d_1, d_2) = (3, 2, 2)$. Hence, we can conclude that (i) when $(\mathbf{A}, \mathbf{B})$ can be uniquely identified, the prediction method introduced in Chang et al. (2023) with selecting $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ as our proposed estimate works quite well, which has almost identical performance as our newly proposed prediction method; and (ii) our newly proposed prediction method works very well regardless of whether $(\mathbf{A}, \mathbf{B})$ can be uniquely identified or not. The results of two-step ahead forecasting are similar to that of one-step ahead forecasting. See Figure S6 in the supplementary material for details.

7.3. *Real data analysis.* In this section, we illustrate the proposed method for the matrix CP-factor model (1) by using the Fama-French 10×10 return series. We collect the monthly returns from January 1964 to December 2021, which contains 69600 observations for total 696 months. The data are downloaded from `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. The portfolios are formed by the intersections of 10 levels of size, denoted by $(\text{S}_1, \ldots, \text{S}_{10})$, and 10 levels of the book equity to market equity ratio (BE), denoted by $(\text{BE}_1, \ldots, \text{BE}_{10})$. The data contain a small number of missing values in the early years and we transform them to zeros. Since all the 100 series are clearly related to the overall market condition, following Wang, Liu and Chen (2019), we decide to remove the influence of market effects before empirical analysis. Two filtering approaches are considered: (i) (CAPM filtering) fitting a standard CAPM model (Fama and MacBeth, 1973) to each of the series to remove the market effect, (ii) (Demean filtering) subtracting the corresponding monthly excess market return from each of the series. The market return data are obtained from the same website above. Based on each filtering approach, we finally obtain 100 market-adjusted return series. The 100 market-adjusted return series can be represented as a 10×10 matrix time series $\mathbf{Y}_t = (y_{i,j,t})$ for $t \in [696]$ (i.e., $p = q = 10$, $n = 696$), where $y_{i,j,t}$ is the market-adjusted return at the i-th level of size S_i and the j-th level of the BE-ratio BE_j at time t. Figure 3 shows the time series plots of the market-adjusted return series $\{y_{i,j,t}\}_{t=1}^{n}$ based on the CAPM filtering for $i, j \in [10]$. The rows in Figure 3 correspond to the ten levels of size and the columns correspond to the ten levels of the BE-ratio. All series are stationary because they reject the null hypothesis of Augmented Dickey-Fuller test at 5% significance level.

TABLE 1

Relative frequency estimates of $\mathbb{P}_{1,2,d} = \mathbb{P}\{(\hat{d}_1, \hat{d}_2, \hat{d}) = (d_1, d_2, d)\}$ and $\mathbb{P}_d = \mathbb{P}(\hat{d} = d)$ with $(\hat{d}_1, \hat{d}_2, \hat{d})$ estimated by our proposed method, and the relative frequency estimate of $\mathbb{P}_c = \mathbb{P}(\hat{d} = d)$ with $\hat{d}$ estimated by the CP-refined method of Chang et al. (2023) in Scenarios R1–R3. All numbers reported below are multiplied by 100.

d	n	R1								R2									R3					
		$p=q$				$p>q$				$p=q$			$p>q$			$p<q$			$p=q$			$p>q$		
		(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_d$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_d$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_c$	(p,q)	$\mathbb{P}_{1,2,d}$	$\mathbb{P}_c$
3	300		96.59	97.04	94.58		95.11	96.74	95.52		85.38	77.02		77.51	0.00		86.57	79.05		88.86	0.00		87.44	0.00
	600	(20,20)	97.35	97.70	96.15	(40,10)	95.16	97.02	96.01	(20,20)	86.36	79.23	(40,10)	76.70	0.00	(10,40)	87.91	81.21	(20,20)	89.86	0.00	(40,10)	89.09	0.00
	900		97.65	98.00	97.25		96.41	97.88	97.07		84.23	77.54		78.55	0.00		89.29	82.90		90.97	0.00		89.32	0.00
	300		98.75	98.80	97.90		94.99	97.04	96.07		90.88	83.56		78.11	0.00		91.00	83.13		92.83	0.00		90.02	0.00
	600	(40,40)	99.50	99.55	99.05	(80,10)	96.56	97.67	97.06	(40,40)	90.74	85.34	(80,10)	78.31	0.00	(10,80)	90.85	84.11	(40,40)	93.68	0.00	(80,10)	90.51	0.00
	900		99.80	99.80	99.40		96.97	98.69	97.88		91.17	85.51		76.49	0.00		90.19	84.41		93.52	0.00		91.54	0.00
	300		99.75	99.80	99.35		96.55	98.12	97.11		94.31	88.58		80.50	0.00		92.07	84.92		95.74	0.00		90.92	0.00
	600	(80,80)	100.00	100.00	99.50	(160,10)	97.78	99.04	98.38	(80,80)	94.34	89.89	(160,10)	79.97	0.00	(10,160)	91.85	86.37	(80,80)	95.08	0.00	(160,10)	91.35	0.00
	900		100.00	100.00	99.55		97.84	99.14	98.49		94.87	90.66		77.69	0.00		91.17	84.95		95.77	0.00		92.01	0.00
5	300		97.49	98.04	94.88		94.14	98.41	95.42		97.25	91.30		89.70	0.00		97.12	93.58		97.99	0.00		96.91	0.00
	600	(20,20)	98.20	98.55	96.29	(40,10)	94.69	98.62	97.24	(20,20)	97.19	93.32	(40,10)	89.86	0.00	(10,40)	97.36	94.37	(20,20)	98.64	0.00	(40,10)	97.68	0.00
	900		98.10	98.55	96.14		95.63	98.78	97.51		96.40	91.79		90.33	0.00		97.26	95.39		98.49	0.00		97.53	0.00
	300		99.60	99.60	99.05		95.24	99.02	97.15		98.99	96.88		91.13	0.00		97.93	94.61		99.40	0.00		97.66	0.00
	600	(40,40)	99.45	99.55	99.10	(80,10)	96.12	99.23	98.16	(40,40)	99.35	97.39	(80,10)	91.48	0.00	(10,80)	98.14	95.98	(40,40)	99.55	0.00	(80,10)	97.93	0.00
	900		99.70	99.70	99.35		96.89	99.49	98.73		99.05	97.39		91.32	0.00		98.54	96.48		99.55	0.00		98.08	0.00
	300		99.90	99.90	99.70		96.26	99.18	98.05		99.75	98.50		92.28	0.00		98.29	96.28		99.80	0.00		98.49	0.00
	600	(80,80)	99.95	99.95	99.65	(160,10)	96.84	99.23	98.52	(80,80)	99.70	98.85	(160,10)	91.21	0.00	(10,160)	98.59	96.83	(80,80)	100.00	0.00	(160,10)	98.04	0.00
	900		99.95	99.95	99.70		97.01	99.70	99.24		99.85	99.00		90.56	0.00		99.30	97.79		99.95	0.00		97.63	0.00
7	300		97.94	98.60	95.84		84.63	98.69	96.64		98.53	94.99		83.97	0.00		97.68	96.36		99.00	0.00		98.23	0.00
	600	(20,20)	98.59	99.39	97.33	(40,10)	87.07	99.00	97.84	(20,20)	99.19	95.76	(40,10)	84.77	0.00	(10,40)	98.24	96.78	(20,20)	99.70	0.00	(40,10)	98.47	0.00
	900		98.85	99.35	97.25		89.27	98.95	98.06		98.74	96.38		85.96	0.00		98.23	96.87		99.40	0.00		97.93	0.00
	300		99.85	99.85	99.30		85.49	98.85	97.39		99.90	98.60		83.71	0.00		98.13	97.73		100.00	0.00		98.53	0.00
	600	(40,40)	99.90	99.90	99.50	(80,10)	89.37	99.32	98.27	(40,40)	99.80	99.05	(80,10)	84.98	0.00	(10,80)	98.39	97.99	(40,40)	99.90	0.00	(80,10)	98.13	0.00
	900		99.85	99.90	99.30		89.15	99.53	98.81		99.75	98.65		85.63	0.00		99.15	98.49		99.95	0.00		98.79	0.00
	300		100.00	100.00	99.70		87.47	99.53	99.06		99.95	99.70		84.41	0.00		99.15	98.24		100.00	0.00		98.83	0.00
	600	(80,80)	99.95	99.95	99.85	(160,10)	89.06	99.74	99.11	(80,80)	99.90	99.65	(160,10)	86.38	0.00	(10,160)	99.25	98.79	(80,80)	100.00	0.00	(160,10)	98.48	0.00
	900		100.00	100.00	99.80		92.05	100.00	99.38		100.00	99.70		87.72	0.00		99.35	99.04		100.00	0.00		98.79	0.00

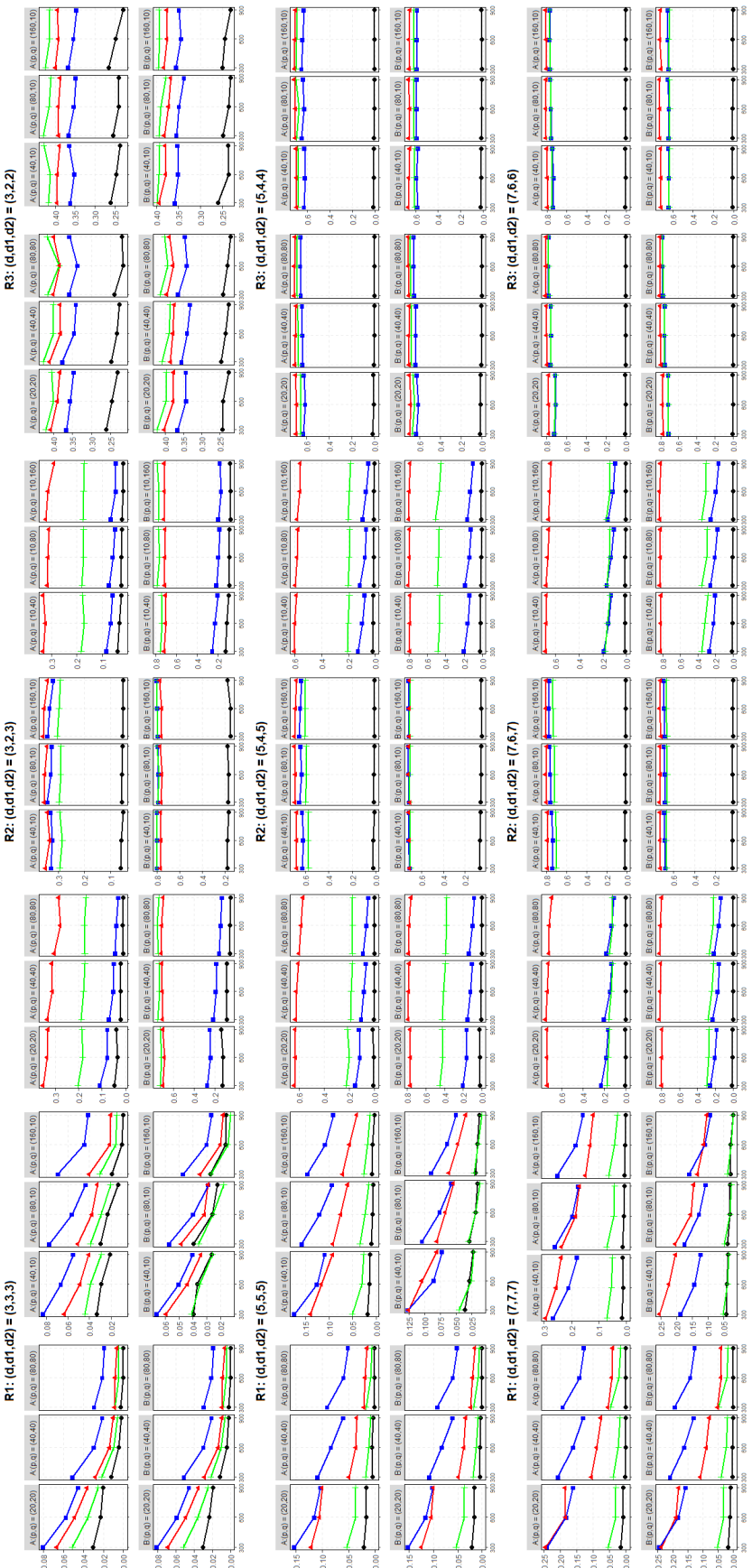


FIG 1. *The lineplots for the averages of estimation errors* $\varpi^2(\mathbf{A}, \hat{\mathbf{A}})$ *and* $\varpi^2(\mathbf{B}, \hat{\mathbf{B}})$ *based on 2000 repetitions in Scenarios R1–R3. The legend is defined as follows:* (i) *our proposed method* ($-\bullet-$), (ii) *the CP-refined method of Chang et al. (2023)* ($-\blacktriangle-$), (iii) *the cPCA of Han et al. (2024b)* ($-\blacksquare-$), *and* (iv) *the HOPE of Han et al. (2024b)* ($-+-$).

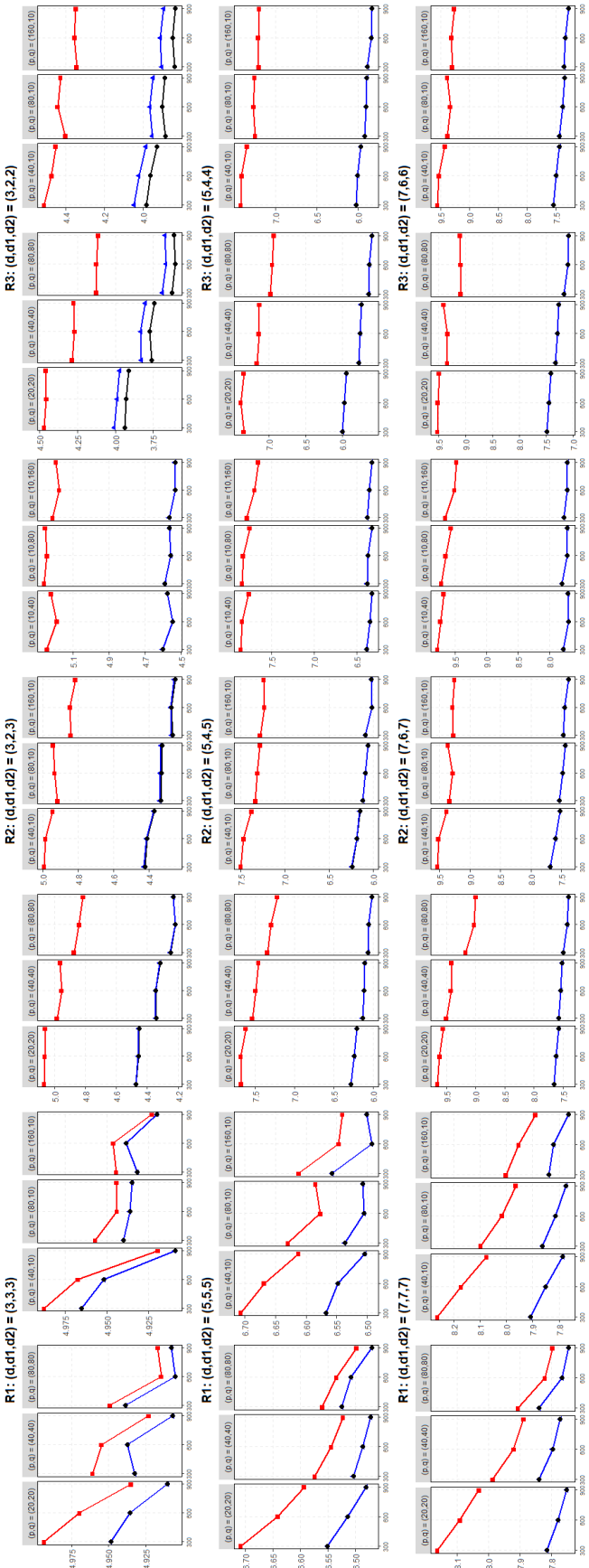


FIG 2. *The lineplots for the averages of one-step ahead forecast RMSE based on 2000 repetitions in Scenarios R1–R3. The legend is defined as follows:* (i) *our proposed prediction method* (− ● −), (ii) *the prediction method introduced in Chang et al. (2023) with* $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ *selected as our proposed estimate* (−▲−), (iii) *the prediction method introduced in Chang et al. (2023) with* $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ *selected as the CP-refined estimate* (−■−).

We evaluate the post-sample forecasting performance of our proposed method introduced in Section 5 by performing the one-step and two-step ahead rolling forecasts for the 240 monthly readings in the last twenty years (2002–2021). To do this, we first use the data $\{\mathbf{Y}_t\}_{t=1}^{456}$ to determine the rank parameters (d, d_1, d_2). With the tuning parameters selected as those in Section 7.1, our proposed method obtains $(\hat{d}, \hat{d}_1, \hat{d}_2) = (2, 2, 1)$, which aligns with the conventional scree plots of $\hat{\mathbf{M}}_1$ and $\hat{\mathbf{M}}_2$ given in Figures 4(a) and 4(b), respectively. We adopt $(\hat{d}, \hat{d}_1, \hat{d}_2) = (2, 2, 1)$ in the rolling forecasts. For each $s \in [240]$, we apply our proposed prediction method to the data $\{\mathbf{Y}_t\}_{t=s}^{455+s}$ and then obtain the one-step forecast of $\mathbf{Y}_{456+s}$, denoted by $\hat{\mathbf{Y}}_{456+s}^{(1)} = (\hat{y}_{i,j,456+s}^{(1)})$. For the two-step ahead forecast, we apply our proposed prediction method to the data $\{\mathbf{Y}_t\}_{t=s}^{454+s}$, and the two-step ahead forecast $\hat{\mathbf{Y}}_{456+s}^{(2)} = (\hat{y}_{i,j,456+s}^{(2)})$ can be obtained by plug-in the one-step forecast into the fitted model. More specifically, for each $s \in [240]$, we fit the obtained 2-dimensional time series by a VAR model with the order determined by the AIC. For comparison, we can also fit $\{\mathbf{Y}_t\}_{t=s}^{455+s}$ and $\{\mathbf{Y}_t\}_{t=s}^{454+s}$ by the following methods and obtain the associated one-step and two-step ahead forecasts:

- (CP-refined) The CP-refined method of Chang et al. (2023) with the pre-determined parameter $K = 10$ therein. The associated rank in this method is estimated as $\hat{d} = 1$ based on $\{\mathbf{Y}_t\}_{t=1}^{456}$ and then fixed in the rolling forecasts. Motivated by the scree plot in Figure 4(c), we also consider $\hat{d} = 2$ as an alternative. For $\hat{d} = 1$, we fit the obtained univariate time series by an AR model with the order determined by the AIC. For $\hat{d} = 2$, we fit the obtained 2-dimensional time series by a VAR model with the order determined by the AIC. The methods with $\hat{d} = 1$ and $\hat{d} = 2$ are referred to as CP-refined(1) and CP-refined(2), respectively.
- (cPCA, HOPE) The composite PCA and High-Order Projection Estimators in Han et al. (2024b) with the recommended tuning parameter $h = 1$ therein. Following the same rank specification strategy as in the CP-refined method, we consider both $\hat{r} = 1$ and $\hat{r} = 2$ for the associated rank in these two methods. For $\hat{r} = 1$, we fit the obtained univariate time series by an AR model with the order determined by the AIC. For $\hat{r} = 2$, we fit the obtained 2-dimensional time series by a VAR model with the order determined by the AIC. The methods with $\hat{r} = 1$ are referred to as cPCA(1) and HOPE(1), while those with $\hat{r} = 2$ are denoted as cPCA(2) and HOPE(2).
- (FAC) The matrix Tucker-factor model with the FAC method proposed by Wang, Liu and Chen (2019) with the pre-determined parameter $h_0 = 1$ as suggested therein. The associated ranks in this model are estimated as $(\hat{k}_1, \hat{k}_2) = (1, 1)$ by the ratio estimators suggested therein based on $\{\mathbf{Y}_t\}_{t=1}^{456}$, and are fixed in the rolling forecasts. Motivated by the scree plots in Figures 4(d) and 4(e), we also consider an alternative setting with $(\hat{k}_1, \hat{k}_2) = (2, 1)$. For $(\hat{k}_1, \hat{k}_2) = (1, 1)$, we fit the obtained univariate time series by an AR model with the order determined by the AIC. For $(\hat{k}_1, \hat{k}_2) = (2, 1)$, we fit the obtained 2-dimensional time series by a VAR model with the order determined by the AIC. The methods with $(\hat{k}_1, \hat{k}_2) = (1, 1)$ and $(\hat{k}_1, \hat{k}_2) = (2, 1)$ are referred to as FAC(1,1) and FAC(2,1), respectively.
- (TOPUP, TIPUP) The Time series Outer-Product Unfolding Procedure and the Time series Inner-Product Unfolding Procedure proposed by Han et al. (2024a) for the matrix Tucker-factor model. The associated ranks in this model are estimated as $(\hat{k}_1, \hat{k}_2) = (2, 2)$ by the information criterion considered in Han, Chen and Zhang (2022) based on $\{\mathbf{Y}_t\}_{t=1}^{456}$, and are fixed in the rolling forecasts. We fit the obtained 4-dimensional time series by a VAR model with the order determined by the AIC. The methods are implemented using the R package `tensorTS`.
- (MAR) The matrix-AR(1) model of Chen, Xiao and Yang (2021).

- (TS-PCA) Apply the principal component analysis for time series proposed by Chang, Guo and Yao (2018) to the 100-dimensional time series $\{\vec{\mathbf{Y}}_t\}_{t=s}^{455+s}$ and $\{\vec{\mathbf{Y}}_t\}_{t=s}^{454+s}$, respectively, to obtain the associated one-step and two-step ahead forecasts. The method is implemented using the R package `HDTSA`. For the obtained univariate time series, we fit it by an AR model with the order determined by the AIC. For the obtained multivariate time series, we fit it by a VAR model with the order determined by the AIC.
- (UniAR) Fit each of 100 component time series by an AR model with the order determined by the AIC.

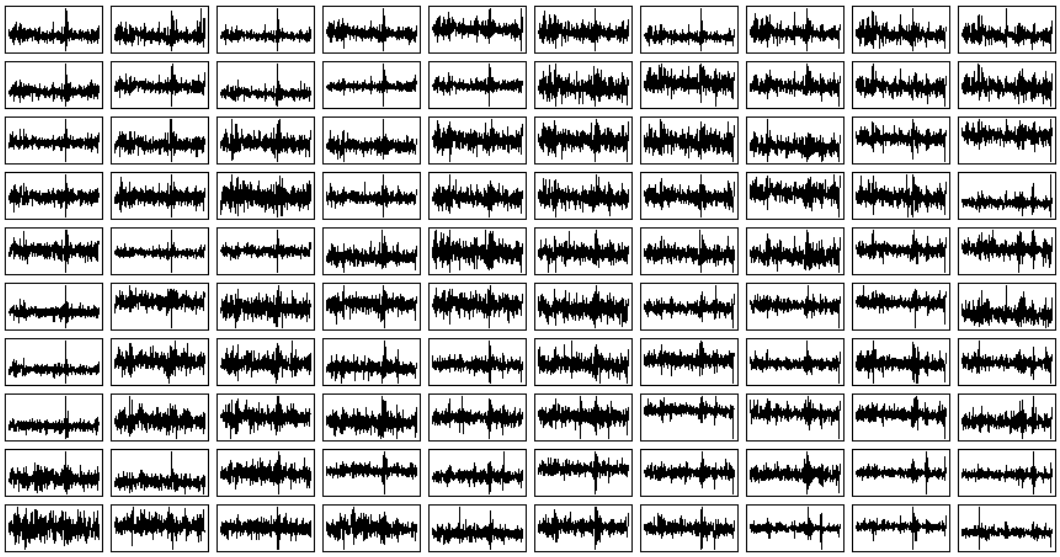


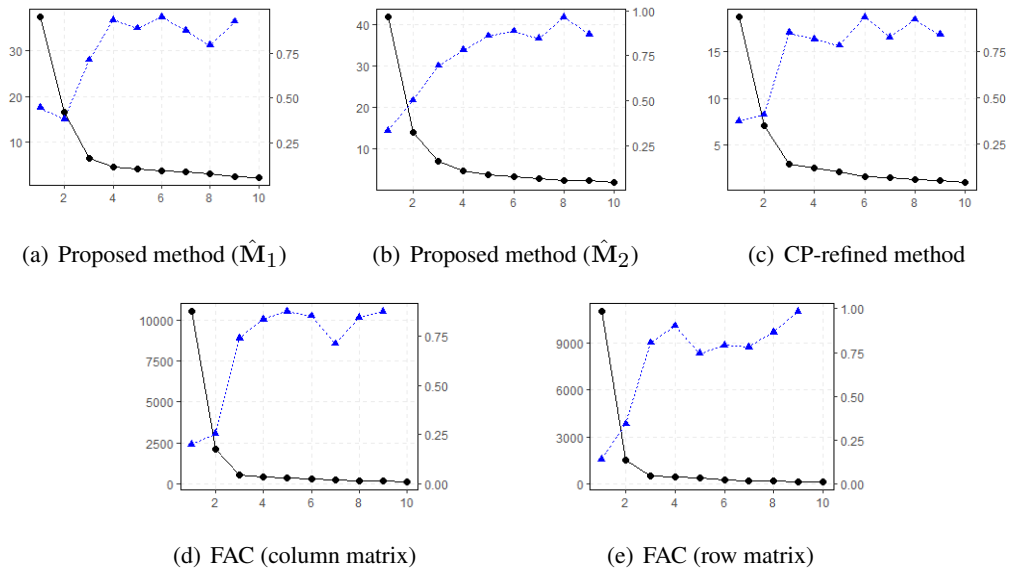
FIG 3. *The time series plots of 100 market-adjusted returns formed on different levels of size (by rows) and book equity to market equity ratio (by columns). The horizontal axis represents time and the vertical axis represents the monthly returns.*

For each $s \in [240]$, the one-step ahead forecasting performance is evaluated by the rRMSE(s) and rMAE(s) defined as

$$\text{rRMSE}(s) = \left\{ \frac{1}{100} \sum_{i=1}^{10} \sum_{j=1}^{10} |\hat{y}_{i,j,456+s}^{(1)} - y_{i,j,456+s}|^2 \right\}^{1/2},$$

$$\text{rMAE}(s) = \frac{1}{100} \sum_{i=1}^{10} \sum_{j=1}^{10} |\hat{y}_{i,j,456+s}^{(1)} - y_{i,j,456+s}|.$$

For the two-step ahead forecast, we can evaluate it by the associated rRMSE(s) and rMAE(s) analogously. Table 2 reports the averages of $\{\text{rRMSE}(s)\}_{s=1}^{240}$ and $\{\text{rMAE}(s)\}_{s=1}^{240}$, denoted by rRMSE and rMAE, respectively. The standard deviations of $\{\text{rRMSE}(s)\}_{s=1}^{240}$ and $\{\text{rMAE}(s)\}_{s=1}^{240}$ are reported in parentheses. As shown in Table 2, under CAPM filtering (Panel A), our proposed method achieves the lowest rRMSE and rMAE for both one- and two-step ahead forecasts, outperforming all competing methods. Under Demean filtering (Panel B), although the HOPE(2) achieves the lowest rRMSE and rMAE, our proposed method performs comparably and yields smaller standard deviations than the HOPE(2). Overall, the results show that our proposed method delivers robust and accurate forecasts across different market-adjustment schemes, often outperforming alternatives in both accuracy and stability.



(a) Proposed method ($\hat{\mathbf{M}}_1$) (b) Proposed method ($\hat{\mathbf{M}}_2$) (c) CP-refined method

(d) FAC (column matrix) (e) FAC (row matrix)

FIG 4. *Scree plots for our proposed method, the CP-refined method and the FAC based on* $\{\mathbf{Y}_t\}_{t=1}^{456}$. *The black solid line represents the eigenvalues, while the blue dashed line indicates the ratios of adjacent eigenvalues.*

Acknowledgments. The authors are grateful to the Editor, an Associate Editor and three referees for their helpful suggestions. The authors also thank Yuefeng Han for sharing code for implementing the methods proposed in Han et al. (2024b).

Funding. J. Chang, Y. Du and G. Huang were supported in part by the National Natural Science Foundation of China (Grant nos. 72125008 and 72495122).

Q. Yao was supported in part by the U.K. Engineering and Physical Sciences Research Council (Grant nos. EP/V007556/1 and EP/X002195/1).

SUPPLEMENTARY MATERIAL

Supplement to "Identification and Estimation for Matrix Time Series CP-factor Models".
This supplement contains additional simulation studies and all technical proofs.

REFERENCES

AFSARI, B. (2008). Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications* **30** 1148–1171.

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., HE, J., YANG, L. and YAO, Q. (2023). Modelling matrix time series via a tensor CP-decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85** 127–148.

CHANG, J., HE, J., LIN, C. and YAO, Q. (2024). HDTSA: An R package for high-dimensional time series analysis. *arXiv:2412.17341*.

CHEN, E. Y. and CHEN, R. (2022). Modeling dynamic transport network with matrix factor models: an application to international trade flow. *Journal of Data Science* **21** 490–507.

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.

DE LATHAUWER, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications* **28** 642–666.

FAMA, E. F. and MACBETH, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* **81** 607–636.

HAN, Y., CHEN, R. and ZHANG, C.-H. (2022). Rank determination in tensor factor model. *Electronic Journal of Statistics* **16** 1726–1803.

HAN, Y. and ZHANG, C. (2022). Tensor principal component analysis in high dimensional CP models. *IEEE Transactions on Information Theory* **69** 1147–1167.

HAN, Y., CHEN, R., YANG, D. and ZHANG, C.-H. (2024a). Tensor factor model estimation by iterative projection. *The Annals of Statistics* **52** 2641–2667.

HAN, Y., YANG, D., ZHANG, C. and CHEN, R. (2024b). CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 1384–1413.

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

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

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

PHAM, D. T. and CARDOSO, J. F. (2001). Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing* **49** 1837–1848.

VOLLGRAF, R. and OBERMAYER, K. (2006). Quadratic optimization for simultaneous matrix diagonalization. *IEEE Transactions on Signal Processing* **54** 3270–3278.

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

ZIEHE, A., LASKOV, P., NOLTE, G. and MÜLLER, K. R. (2004). A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *Journal of Machine Learning Research* **5** 777–800.

TABLE 2

The averages and standard deviations (in parentheses) of one-step and two-step ahead forecasting errors of the market-adjusted returns from January, 2002 to December, 2021. Panel A and Panel B represent the market-adjusted returns obtained by demean filtering and CAPM filtering, respectively.

	Proposed	CP-refined(1)	CP-refined(2)	cPCA(1)	cPCA(2)	HOPE(1)	HOPE(2)	FAC(1,1)	FAC(2,1)	TOPUP	TIPUP	MAR	TS-PCA	UniAR
Panel A: CAPM filtering														
one-step ahead forecast														
rRMSE	**3.4302**	3.4485	3.4408	3.4402	3.4361	3.4423	3.4373	3.4482	3.4610	3.4597	3.4669	3.4669	3.4710	3.4895
	(**1.5062**)	(1.5223)	(1.4992)	(1.5254)	(1.5154)	(1.5299)	(1.5163)	(1.5226)	(1.5271)	(1.5277)	(1.5303)	(1.5364)	(1.5028)	(1.5099)
rMAE	**2.6218**	2.6424	2.6325	2.6340	2.6294	2.6340	2.6307	2.6433	2.6541	2.6534	2.6566	2.6600	2.6579	2.6711
	(**1.0522**)	(1.0746)	(1.0468)	(1.0712)	(1.0561)	(1.0755)	(1.0569)	(1.0751)	(1.0791)	(1.0763)	(1.0764)	(1.0849)	(1.0578)	(1.0601)
two-step ahead forecast														
rRMSE	**3.4297**	3.4488	3.4378	3.4436	3.4362	3.4447	3.4370	3.4505	3.4627	3.4602	3.4641	3.4401	3.4626	3.4873
	(**1.5003**)	(1.5238)	(1.4972)	(1.5432)	(1.5178)	(1.5477)	(1.5161)	(1.5254)	(1.5291)	(1.5303)	(1.5331)	(1.5162)	(1.4947)	(1.5111)
rMAE	**2.6241**	2.6444	2.6317	2.6378	2.6327	2.6374	2.6328	2.6456	2.6558	2.6538	2.6552	2.6353	2.6545	2.6705
	(**1.0526**)	(1.0767)	(1.0494)	(1.0896)	(1.0684)	(1.0942)	(1.0668)	(1.0774)	(1.0805)	(1.0789)	(1.0823)	(1.0694)	(1.0575)	(1.0640)
Panel B: Demean filtering														
one-step ahead forecast														
rRMSE	3.4905	3.5146	3.4947	3.5293	3.4987	3.5255	**3.4862**	3.5057	3.5165	3.5268	3.5303	3.5154	3.5244	3.5470
	(1.5698)	(1.5829)	(1.5725)	(1.5883)	(1.5878)	(1.5876)	(**1.5824**)	(1.5952)	(1.5884)	(1.5826)	(1.5891)	(1.6093)	(1.6005)	(1.5789)
rMAE	2.6676	2.6910	2.6718	2.7047	2.6741	2.7008	**2.6622**	2.6834	2.6923	2.7022	2.7036	2.6923	2.6999	2.7143
	(1.1133)	(1.1288)	(1.1250)	(1.1333)	(1.1273)	(1.1327)	(**1.1182**)	(1.1402)	(1.1365)	(1.1283)	(1.1319)	(1.1597)	(1.1586)	(1.1250)
two-step ahead forecast														
rRMSE	3.4869	3.5142	3.4912	3.5239	3.4920	3.5208	**3.4700**	3.5018	3.5105	3.5269	3.5294	3.4959	3.5124	3.5413
	(1.5685)	(1.5863)	(1.5721)	(1.5926)	(1.5939)	(1.5943)	(**1.5954**)	(1.5954)	(1.5894)	(1.5899)	(1.5961)	(1.6057)	(1.5881)	(1.5817)
rMAE	2.6674	2.6922	2.6703	2.7013	2.6721	2.6982	**2.6511**	2.6805	2.6885	2.7036	2.7038	2.6765	2.6919	2.7130
	(1.1170)	(1.1358)	(1.1237)	(1.1408)	(1.1391)	(1.1422)	(**1.1376**)	(1.1403)	(1.1368)	(1.1367)	(1.1406)	(1.1539)	(1.1483)	(1.1323)