Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Modeling and Learning on High-Dimensional Matrix-Variate Sequences

Xu Zhang, Catherine C. Liu, Jianhua Guo, K. C. Yuen & A. H. Welsh

To cite this article: Xu Zhang, Catherine C. Liu, Jianhua Guo, K. C. Yuen & A. H. Welsh (24 May 2024): Modeling and Learning on High-Dimensional Matrix-Variate Sequences, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2344687

To link to this article: https://doi.org/10.1080/01621459.2024.2344687

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2024.2344687

Taylor & Francis
Taylor & Francis Group

Modeling and Learning on High-Dimensional Matrix-Variate Sequences

Xu Zhang[a,b], Catherine C. Liu[b], Jianhua Guo[c], K. C. Yuen[d], and A. H. Welsh[e]

[a]South China Normal University, Guangzhou, China; [b]The Hong Kong Polytechnic University, Hung Hom, HKSAR; [c]Beijing Technology and Business University, Beijing, China; [d]The University of Hong Kong, Hong Kong, HKSAR; [e]The Australian National University, Canberra, Australia

ABSTRACT

We propose a new matrix factor model, named RaDFaM, which is strictly derived from the general rank decomposition and assumes a high-dimensional vector factor model structure for each basis vector. RaDFaM contributes a novel class of low-rank latent structures that trade off between signal intensity and dimension reduction from a tensor subspace perspective. Based on the intrinsic separable covariance structure of RaDFaM, for a collection of matrix-valued observations, we derive a new class of PCA variants for estimating loading matrices, and sequentially the latent factor matrices. The peak signal-to-noise ratio of RaDFaM is proved to be superior in the category of PCA-type estimators. We also establish an asymptotic theory including the consistency, convergence rates, and asymptotic distributions for components in the signal part. Numerically, we demonstrate the performance of RaDFaM in applications such as matrix reconstruction, supervised learning, and clustering, on uncorrelated and correlated data, respectively. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received June 2023
Accepted April 2024

KEYWORDS
Image reconstruction; Matrix factor model; Peak signal-to-noise ratio; Rank decomposition; Separable covariance structure; Tensor subspace

1. Introduction

High-dimensional matrix objects arise in a broad range of applications, occurring as the slices of computed tomography (CT) or as magnetic resonance imaging (MRI) data in medical imaging, and as multinational macroeconomic indices data in economics, to name a few (Zhang 2017; Gupta and Nagar 2018; Fan et al. 2020). Low-rank approximation of a matrix-variate is critical for discovery based on matrix-valued observations. In this article, we aim to model a new class of latent low-rank signal, and establish the inferential methodology and theory under a collection of matrix-variates.

1.1. *Rank-Decomposition-Based Factor Modeling*

Recall that a matrix $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ of rank-l has the general rank decomposition expressible as $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_l) \in \mathbb{R}^{p_1\times l}$ and $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_l) \in \mathbb{R}^{p_2\times l}$ are *full column rank matrices*. Let span$(\mathbf{M})$ be the *column space* spanned by the columns of the placeholder matrix $\mathbf{M}$. Then $\{\mathbf{U}_i\}_{i=1}^{l}$ and $\{\mathbf{V}_i\}_{i=1}^{l}$ are *bases* of span$(\mathbf{U})$ and span$(\mathbf{V})$, respectively.

For a high-dimensional matrix-variate $\mathbf{X}$, where p_1 and p_2 may tend to infinity, by enduing all basis vectors $\mathbf{U}_i$ and $\mathbf{V}_i$ with the structure of the high-dimensional vector factor model, we postulate the following rank-decomposition-based matrix factor model (RaDFaM)

$$\begin{cases} \mathbf{X} = \mathbf{U}\mathbf{V}^\top = \sum_{i=1}^{l} \mathbf{U}_i\mathbf{V}_i^\top, & \text{(rank-}l\text{ decomposition)} \\ \mathbf{U}_i = \mathbf{R}\mathbf{A}_{i\cdot} + \boldsymbol{\xi}_i, \quad \mathbf{V}_i = \mathbf{C}\mathbf{B}_{i\cdot} + \boldsymbol{\eta}_i, & \text{(vector factor models)} \end{cases} \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{p_1\times k_1}$ $(k_1 \ll p_1)$ and $\mathbf{C} \in \mathbb{R}^{p_2\times k_2}$ $(k_2 \ll p_2)$ are called loading matrices, $\mathbf{A}_{i\cdot} \in \mathbb{R}^{k_1}$ and $\mathbf{B}_{i\cdot} \in \mathbb{R}^{k_2}$ are vectors of latent factors, and $\boldsymbol{\xi}_i \in \mathbb{R}^{p_1}$ and $\boldsymbol{\eta}_i \in \mathbb{R}^{p_2}$ are vectors of idiosyncratic errors.

We adopt matrix notation to express model (1) more concisely. Denote $\mathbf{A} = (\mathbf{A}_{1\cdot}, \dots, \mathbf{A}_{l\cdot})^\top \in \mathbb{R}^{l\times k_1}$, $\mathbf{B} = (\mathbf{B}_{1\cdot}, \dots, \mathbf{B}_{l\cdot})^\top \in \mathbb{R}^{l\times k_2}$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_l) \in \mathbb{R}^{p_1\times l}$, and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_l) \in \mathbb{R}^{p_2\times l}$. Let $\mathbf{G}_{i\cdot} \in \mathbb{R}^{n}$ and $\mathbf{G}_j \in \mathbb{R}^{m}$ be the ith row and the jth column of a placeholder matrix $\mathbf{G} \in \mathbb{R}^{m\times n}$. Denote $\mathbf{Z} = \mathbf{A}^\top\mathbf{B} \in \mathbb{R}^{k_1\times k_2}$, $\mathbf{E} = \boldsymbol{\eta}\mathbf{A} \in \mathbb{R}^{p_2\times k_1}$, $\mathbf{F} = \boldsymbol{\xi}\mathbf{B} \in \mathbb{R}^{p_1\times k_2}$, and $\mathbf{e} = \boldsymbol{\xi}\boldsymbol{\eta}^\top \in \mathbb{R}^{p_1\times p_2}$. Then in pure matrix notation, RaDFaM has the following equivalent expression

$$\mathbf{X} = \underbrace{\mathbf{R}\mathbf{Z}\mathbf{C}^\top + \mathbf{R}\mathbf{E}^\top + \mathbf{F}\mathbf{C}^\top}_{\text{low-rank signal}} + \underbrace{\mathbf{e}}_{\text{noise}} \equiv \mathbf{S} + \mathbf{e}. \tag{2}$$

Different insights of assignments of parts of the signal and noise in model (2) may lead to different low-rank approximations of the matrix-variate $\mathbf{X}$. Specifically,

$$\begin{aligned} \mathbf{X} &= \underbrace{\mathbf{R}\mathbf{Z}\mathbf{C}^\top}_{\text{latent structure}} + \underbrace{\mathbf{R}\mathbf{E}^\top + \mathbf{F}\mathbf{C}^\top + \mathbf{e}}_{\text{error}} \\ &\equiv \underbrace{\mathbf{R}\mathbf{Z}\mathbf{C}^\top}_{\text{low-rank signal}} + \text{noise}, \quad \text{(BiMFaM)} \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbf{X} &= \underbrace{\mathbf{R}\mathbf{E}^\top + \mathbf{F}\mathbf{C}^\top}_{\text{latent structure}} + \underbrace{\mathbf{R}\mathbf{Z}\mathbf{C}^\top + \mathbf{e}}_{\text{error}} \\ &\equiv \underbrace{\mathbf{R}\mathbf{E}^\top + \mathbf{F}\mathbf{C}^\top}_{\text{low-rank signal}} + \text{noise}. \quad \text{(2w-DFM)} \end{aligned} \tag{4}$$

CONTACT Catherine C. Liu catherine.chunling.liu@polyu.edu.hk The Hong Kong Polytechnic University, Hung Hom, HKSAR; Jianhua Guo jhguo@btbu.edu.cn Beijing Technology and Business University, Beijing, China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In fact, models (3) and (4) move specific quantities from the individual latent structure part into the error part, respectively. That is, if one exposes only the bilinear-form product involving the low-dimensional interactive latent factor ($\mathbf{Z}$) or the sum of the linear terms involving the two low-dimensional main latent factors (row-wise $\mathbf{F}$ and column-wise $\mathbf{E}$) as signals, respectively, then model (1) generates two classes of matrix factor models in the existing literature, the popular bilinear-form matrix factor model (BiMFaM) and a newly presented two-way dynamic factor model (2w-DFM), respectively (Wang, Liu, and Chen 2019; Yu et al. 2022; Chen and Fan 2023; Yuan et al. 2023).

1.2. Tensor Subspace and Dimension Reduction

In this section, we aim to elucidate advantages and disadvantages of RaDFaM from the tensor subspace perspective. Recall that a tensor space $\mathbb{R}^{p_1\times\cdots\times p_D}$ is conceptually consistent with a vector space, the operations of which satisfy the axioms of a real vector space (Shores 2018, sec. 3.1). A tensor subspace $\mathcal{S} \subset \mathbb{R}^{p_1\times\cdots\times p_D}$ includes the zero tensor and is closed under finite tensor addition and scalar multiplication; the dimension of $\mathcal{S}$, denoted as $\dim(\mathcal{S})$, is the number of elements in its basis (Zhou et al. 2021).

For different tensor decompositions, the deduced low-rank signal structures correspond to different tensor subspaces (Lu, Plataniotis, and Venetsanopoulos 2014; Wang, Aggarwal, and Aeron 2019; Zhang et al. 2022). For the case of $D = 2$ in our study, the signal in BiMFaM is essentially in the form of the second-order Tucker decomposition, and corresponds to the tensor subspace (Zare et al. 2018, sec. IV-B)

$$\mathcal{S}_{BiM} = \left\{ \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} \mathbf{Z}_{i_1 i_2}\left(\mathbf{R}_{i_1}\circ \mathbf{C}_{i_2}\right) | \mathbf{Z} = (\mathbf{Z}_{i_1 i_2}) \in \mathbb{R}^{k_1\times k_2} \right\},$$

where $\{\mathbf{R}_{i_1}\circ\mathbf{C}_{i_2} \in \mathbb{R}^{p_1\times p_2}\}$ represents the set of rank-1 matrices spanning $\mathcal{S}_{BiM}$ with $\circ$ being the vector outer product, and the interactive latent factor $\mathbf{Z}$ becomes the matrix of coefficients of the linear combination. Denote the vectorized $\mathcal{S}_{BiM}$ as $\widetilde{\mathcal{S}}_{BiM} = \{\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\mathbf{Z}_{i_1 i_2}(\mathbf{C}_{i_2}\otimes\mathbf{R}_{i_1}) | \mathbf{Z} = (\mathbf{Z}_{i_1 i_2}) \in \mathbb{R}^{k_1\times k_2}\}$, where $\mathbf{C}_{i_2}\otimes\mathbf{R}_{i_1} = \mathrm{vec}(\mathbf{R}_{i_1}\circ\mathbf{C}_{i_2})$, and $\otimes$ and $\mathrm{vec}(\cdot)$ are the Kronecker product and the vectorization operation, respectively. The isomorphism of $\mathcal{S}_{BiM}$ and $\widetilde{\mathcal{S}}_{BiM}$ yields their identical dimensionality (Nicholson 2020, sec. 7.3). It is easy to see that $\{\mathbf{C}_{i_2}\otimes\mathbf{R}_{i_1}\}$ is the basis of $\widetilde{\mathcal{S}}_{BiM}$ due to the column orthogonality of $\mathbf{R}$ and $\mathbf{C}$ (see Assumption 3). Hence,

$$\dim(\mathcal{S}_{BiM}) = \dim(\widetilde{\mathcal{S}}_{BiM}) = k_1k_2 \ll p_1p_2.$$

In the same spirit, some algebra relates the following tensor subspaces to RaDFaM and 2w-DFM:

$$\begin{aligned}\mathcal{S}_{RaD} = \Bigg\{ &\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} \mathbf{Z}_{i_1 i_2}\left(\mathbf{R}_{i_1}\circ\mathbf{C}_{i_2}\right) + \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{p_2} \mathbf{E}_{i_2 i_1}\left(\mathbf{R}_{i_1}\circ\mathbf{I}_{p_2,i_2}\right) \\ &+ \sum_{i_1=1}^{p_1}\sum_{i_2=1}^{k_2} \mathbf{F}_{i_1 i_2}\left(\mathbf{I}_{p_1,i_1}\circ\mathbf{C}_{i_2}\right) | \mathbf{Z} = (\mathbf{Z}_{i_1 i_2}) \in \mathbb{R}^{k_1\times k_2}, \\ &\mathbf{E} = (\mathbf{E}_{i_2 i_1}) \in \mathbb{R}^{p_2\times k_1}, \mathbf{F} = (\mathbf{F}_{i_1 i_2}) \in \mathbb{R}^{p_1\times k_2}\Bigg\},\end{aligned}$$

$$\begin{aligned}\mathcal{S}_{2w} = \Bigg\{ &\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{p_2} \mathbf{E}_{i_2 i_1}\left(\mathbf{R}_{i_1}\circ\mathbf{I}_{p_2,i_2}\right) \\ &+ \sum_{i_1=1}^{p_1}\sum_{i_2=1}^{k_2} \mathbf{F}_{i_1 i_2}\left(\mathbf{I}_{p_1,i_1}\circ\mathbf{C}_{i_2}\right) | \mathbf{E} \in \mathbb{R}^{p_2\times k_1}, \mathbf{F} \in \mathbb{R}^{p_1\times k_2}\Bigg\},\end{aligned}$$

where $\mathbf{I}_d$ is the $d\times d$ identity matrix. Note that $\mathrm{vec}(\mathbf{R}_{i_1}\circ\mathbf{I}_{p_2,i_2}) = \mathbf{I}_{p_2,i_2}\otimes\mathbf{R}_{i_1}$ and $\mathrm{vec}(\mathbf{I}_{p_1,i_1}\circ\mathbf{C}_{i_2}) = \mathbf{C}_{i_2}\otimes\mathbf{I}_{p_1,i_1}$. Then the sets of vectors that span the *vectorized isomorphic subspaces* under RaDFaM and 2w-DFM are $\{\mathbf{C}_{i_2}\otimes\mathbf{R}_{i_1}, \mathbf{I}_{p_2,j_2}\otimes\mathbf{R}_{i_1}, \mathbf{C}_{i_2}\otimes\mathbf{I}_{p_1,j_1}\}$ and $\{\mathbf{I}_{p_2,j_2}\otimes\mathbf{R}_{i_1}, \mathbf{C}_{i_2}\otimes\mathbf{I}_{p_1,j_1}\}$, respectively. Due to the fact that $\mathrm{span}(\mathbf{C}\otimes\mathbf{R}) \subset \{\mathrm{span}(\mathbf{I}_{p_2}\otimes\mathbf{R}) \cup \mathrm{span}(\mathbf{C}\otimes\mathbf{I}_{p_1})\}$, we have

$$\begin{aligned}\dim(\mathcal{S}_{RaD}) = \dim(\mathcal{S}_{2w}) &= p_1k_2 \\ &\quad + k_1p_2 - \dim\{\mathrm{span}(\mathbf{I}_{p_2}\otimes\mathbf{R}) \cap \mathrm{span}(\mathbf{C}\otimes\mathbf{I}_{p_1})\} \\ &\ll p_1p_2.\end{aligned}$$

Based on the above discussion of tensor subspaces and companion dimensions, we can see that: (i) RaDFaM possesses the strongest low-rank signal strength. Though RaDFaM reduces the dimension less than BiMFaM ($\dim(\mathcal{S}_{RaD}) > \dim(\mathcal{S}_{BiM})$), RaDFaM is able to extract additional mode-wise information ($\mathbf{E}$ and $\mathbf{F}$); though RaDFaM and 2w-DFM have the same tensor subspace dimension ($\dim(\mathcal{S}_{RaD}) = \dim(\mathcal{S}_{2w})$), RaDFaM is able to extract additional interaction information ($\mathbf{Z}$). ii) RaDFaM incurs no extra computational burden in the sense that the basis parameters $\mathbf{R}$ and $\mathbf{C}$ are identical for all three factor models.

1.3. Separable Covariance Structure

Recall that a matrix-variate is said to have a separable covariance structure if the variance-covariance matrix of its vectorization is the Kronecker product of column-wise and row-wise variance-covariance matrices (Dawid 1981; Gupta and Nagar 2018). Indeed, the rank-l matrix-variate $\mathbf{X}$ under RaDFaM possesses a separable covariance structure, where its vectorization is expressible as

$$\mathrm{vec}(\mathbf{X}) = \sum_{i=1}^{l}\mathbf{V}_i\otimes\mathbf{U}_i.$$

Let Ψ_G be the variance-covariance matrix of the rows or columns of the placeholder matrix $\mathbf{G}$. Under regularity assumptions of zero mean, mutual uncorrelatedness of $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$, and uncorrelatedness among rows or columns of the factor matrix $\mathbf{A}$ ($\mathbf{B}$) and the error matrix $\boldsymbol{\xi}$ ($\boldsymbol{\eta}$), by some algebra, one has $\mathrm{cov}(\mathbf{V}_i) = \mathbf{C}\Psi_B\mathbf{C}^\top + \Psi_\eta$, $\mathrm{cov}(\mathbf{U}_i) = \mathbf{R}\Psi_A\mathbf{R}^\top + \Psi_\xi$; and

$$\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\} = l\{\mathrm{cov}(\mathbf{V}_i)\otimes\mathrm{cov}(\mathbf{U}_i)\} \equiv l(\Sigma_C\otimes\Sigma_R), \quad (5)$$

where Σ_R and Σ_C are called row-wise and column-wise variance-covariance matrices of $\mathbf{X}$ under RaDFaM, respectively, and their elements control the correlation of different rows and columns in $\mathbf{X}$, respectively.

Separable covariance structure brings convenience for inference based on covariance decompositions (Hoff 2011; Fosdick and Hoff 2014; Fan, Li, and Liao 2021). The proposed RaDFaM enjoys this property, which guarantees our PCA-type estimation of the loading matrices is doable. Thus, later we name our proposed approach of PCA-type estimation as sPCA.

1.4. Related Work and Organization

The bilinear-form low-rank structure has been extensively researched in engineering and computer vision from the algorithmic perspective (Yang et al. 2004; Ye 2005; Liu et al. 2010; Ahmadi and Rezghi 2020, among others). The corresponding BiMFaM was first explored for matrix-variate time series (Wang, Liu, and Chen 2019), then extended to uncorrelated or weakly correlated observations (Chen and Fan 2023), and its convergence rates were enhanced by the projection technique (Yu et al. 2022); their estimation approaches were PCA based. Recently, 2w-DFM was presented and a refined quasi maximum likelihood estimation (Q-MLE) enhanced the PCA-type initial estimation (Yuan et al. 2023); a similar latent structure was separately investigated from the spiked covariance perspective (Tang, Yuan, and Zhang 2023). Our proposed RaDFaM with sPCA estimation shows superior signal intensity both theoretically and numerically in the category of PCA-type estimation approaches, albeit with some sacrifice of theoretical convergence rates; and sPCA has a much lower computational cost than Q-MLE, though they are comparable in reconstruction accuracy.

The remainder of this article is organized as follows. Section 2 develops PCA-type estimation based on the separable covariance structure, and states the theoretical evidence for the strong signal intensity of RaDFaM. Section 3 establishes the asymptotic theory, and Section 4 presents simulation studies evaluating the finite sample performance. Section 5 conducts real data analysis on uncorrelated and correlated matrix-variate observations, respectively, followed by concluding remarks in Section 6. All the technical proofs are given in the online supplement.

2. Estimation for Matrix-Valued Observations

Let $\mathbf{X}_t \in \mathbb{R}^{p_1 \times p_2}$, $t \in [T] \equiv \{1, \dots, T\}$, be a sequence of high-dimensional matrix-valued observations that might be weakly correlated or uncorrelated, where the mode-wise dimensions p_1 and p_2, and the number of observations T, can all tend to infinity. The matrix-valued observations under RaDFaM are expressible as

$$\mathbf{X}_t = \mathbf{S}_t + \mathbf{e}_t, \;\; \mathbf{S}_t = \mathbf{R}\mathbf{Z}_t\mathbf{C}^\top + \mathbf{R}\mathbf{E}_t^\top + \mathbf{F}_t\mathbf{C}^\top. \tag{6}$$

In this section, under model+data in (6), we propose a two-step estimation of the components in the signal part, including so-called sPCA estimators (based on the property of separable covariance structure) for the loading matrices in Section 2.1 and least squares estimators of the factor matrices in Section 2.2. Furthermore, we show the superiority of matrix reconstruction using RaDFaM compared to BiMFaM and 2w-DFM based on PCA-type estimators in Section 2.3. We consider the estimation of factor numbers in Section 2.4.

2.1. Separable PCA Estimators for Loading Matrices

It is known that estimators of loading matrices $\mathbf{R}$ and $\mathbf{C}$ are uniquely determined up to a rotation (Wang, Liu, and Chen 2019; Chen and Fan 2023). Thus, we restrict the loading matrices by setting $p_1^{-1}\mathbf{R}^\top\mathbf{R} = \mathbf{I}_{k_1}$ and $p_2^{-1}\mathbf{C}^\top\mathbf{C} = \mathbf{I}_{k_2}$, known as the strong factor assumption in the literature of factor models (Stock and Watson 2002; Lam and Yao 2011). In Assumption 3 of Section 3, we use the strict limit form of the strong factor assumption.

The spectral or PCA-type methods that are most commonly used for estimating loading matrices are based on the eigen-decomposition or singular value decomposition of some moment statistics (Chen and Fan 2023, Remark 2). One of the rationales behind these methods is that, under the strong factor assumption, top eigenspaces (spaces spanned by the top eigenvectors) of population moment matrices are consistent estimators of the column spaces of loading matrices (Davis and Kahan 1970).

Nonetheless, spectral methods based on second moments may not work under model (6) for matrix-valued observations. Take the row-wise second moment for instance

$$\mathbb{E}\left(\mathbf{X}_t\mathbf{X}_t^\top\right) = p_2\mathbf{R}\mathbb{E}\left(\mathbf{Z}_t\mathbf{Z}_t^\top\right)\mathbf{R}^\top + \left\{\mathbf{R}\mathbb{E}\left(\mathbf{E}_t^\top\mathbf{E}_t\right)\mathbf{R}^\top + p_2\mathbb{E}\left(\mathbf{F}_t\mathbf{F}_t^\top\right)\right\} + \mathbb{E}\left(\mathbf{e}_t\mathbf{e}_t^\top\right),$$

where we assume $\mathbf{Z}_t$, $\mathbf{E}_t$, $\mathbf{F}_t$, and $\mathbf{e}_t$ are uncorrelated. Compared to BiMFaM (Chen and Fan 2023, sec. 2.2), the two terms in braces on the right-hand side, which are exactly the main terms in $\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top)$ under 2w-DFM, are not negligible. Fortunately, from (5), we have the following result.

Proposition 1. (Row-wise second moment) Under RaDFaM, based on the separable covariance structure in (5), we have

$$\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top) = \{l\text{tr}(\Sigma_C)\}\Sigma_R.$$

Proposition 1 shows that $\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top)$ and Σ_R are equivalent up to a constant $l\text{tr}(\Sigma_C)$, so have identical top eigenspaces, that is

$$\text{span}[\text{eig}\{\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top), k_1\}] = \text{span}\{\text{eig}(\Sigma_R, k_1)\}, \tag{7}$$

where $\text{eig}(\mathbf{M}, r)$ denotes a matrix with columns given by the top r eigenvectors of the placeholder matrix $\mathbf{M}$. Recall that $\Sigma_R = \mathbf{R}\Psi_A\mathbf{R}^\top + \Psi_\xi$. Using the general pervasiveness assumption and applying the Davis-Kahan Theorem (Davis and Kahan 1970; Fan et al. 2020), we can show the approximate equivalence between the top eigenspace of Σ_R and the column space of $\mathbf{R}$

$$\frac{1}{p_1}\left\|\text{eig}(\Sigma_R, k_1)\text{eig}^\top(\Sigma_R, k_1) - \mathbf{R}\mathbf{R}^\top\right\|_F = o(1), \tag{8}$$

where $\|\cdot\|_F$ is the Frobenius norm. Hence, based on (7) and (8), we have the approximate equivalence of the top eigenspace of $\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top)$ and the column space of $\mathbf{R}$

$$\text{span}[\text{eig}\{\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top), k_1\}] \approx \text{span}(\mathbf{R}), \tag{9}$$

and we can derive the sPCA estimator of $\mathbf{R}$ as

$$\widehat{\mathbf{R}} = \sqrt{p_1}\text{eig}(\widehat{\mathbf{M}}_1, k_1) \;\text{ with }\; \widehat{\mathbf{M}}_1 = \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\mathbf{X}_t\mathbf{X}_t^\top. \tag{10}$$

By a similar argument, the sPCA estimator of $\mathbf{C}$ is

$$\widehat{\mathbf{C}} = \sqrt{p_2}\text{eig}(\widehat{\mathbf{M}}_2, k_2) \;\text{ with }\; \widehat{\mathbf{M}}_2 = \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\mathbf{X}_t^\top\mathbf{X}_t. \tag{11}$$

2.2. Least Squares Estimators for Factor Matrices

Given $\mathbf{R}$ and $\mathbf{C}$, estimators of the factor matrices $\mathbf{Z}_t$, $\mathbf{E}_t$, and $\mathbf{F}_t$ can be obtained by minimizing the matrix reconstruction error

$$\min_{\substack{p_1^{-1}\mathbf{R}^\top\mathbf{R}=\mathbf{I}_{k_1}, p_2^{-1}\mathbf{C}^\top\mathbf{C}=\mathbf{I}_{k_2} \\ \{\mathbf{Z}_t\}_{t=1}^T, \{\mathbf{E}_t\}_{t=1}^T, \{\mathbf{F}_t\}_{t=1}^T}} \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\left\|\mathbf{X}_t - \mathbf{R}\mathbf{Z}_t\mathbf{C}^\top - \mathbf{R}\mathbf{E}_t^\top - \mathbf{F}_t\mathbf{C}^\top\right\|_F^2.$$

Using the fact that $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top)$, we expand the objective function, and then take partial derivatives with respect to $\mathbf{Z}_t$, $\mathbf{E}_t$ and $\mathbf{F}_t$, respectively, to obtain the normal equations. After some algebra, we obtain the system of equations

$$\begin{cases} \mathbf{Z}_t = \dfrac{\mathbf{R}^\top\mathbf{X}_t\mathbf{C}}{p_1p_2} - \dfrac{\mathbf{E}_t^\top\mathbf{C}}{p_2} - \dfrac{\mathbf{R}^\top\mathbf{F}_t}{p_1}, \\ (\mathbf{I}_{p_2} - \dfrac{\mathbf{C}\mathbf{C}^\top}{p_2})(\mathbf{E}_t - \dfrac{\mathbf{X}_t^\top\mathbf{R}}{p_1}) = \mathbf{0}_{p_2\times k_1}, \\ (\mathbf{I}_{p_1} - \dfrac{\mathbf{R}\mathbf{R}^\top}{p_1})(\mathbf{F}_t - \dfrac{\mathbf{X}_t\mathbf{C}}{p_2}) = \mathbf{0}_{p_1\times k_2}, \end{cases}$$

where $\mathbf{0}_{p\times q}$ is the $p \times q$ matrix of zeros. For any linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A}^\dagger\mathbf{b}$ is the one of minimum-norm among the least-squares solutions, where $\mathbf{A}^\dagger$ represents the pseudoinverse of $\mathbf{A}$ (Ben-Israel and Greville 2003, Corollary 3, chap. 3). Note that $\mathbf{I}_{p_1} - p_1^{-1}\mathbf{R}\mathbf{R}^\top$ is an orthogonal projection matrix, and its pseudoinverse is itself. Hence, for the linear equations $(\mathbf{I}_{p_1} - p_1^{-1}\mathbf{R}\mathbf{R}^\top)(\mathbf{F}_t - p_2^{-1}\mathbf{X}_t\mathbf{C}) = \mathbf{0}$, the minimum-norm solutions of the columns of $\mathbf{F}_t - p_2^{-1}\mathbf{X}_t\mathbf{C}$ are the columns of $(\mathbf{I}_{p_1} - p_1^{-1}\mathbf{R}\mathbf{R}^\top)\mathbf{0}$ $(= \mathbf{0})$. That is, $\mathbf{F}_t - p_2^{-1}\mathbf{X}_t\mathbf{C} = \mathbf{0}$. Similarly, $\mathbf{E}_t - p_1^{-1}\mathbf{X}_t^\top\mathbf{R} = \mathbf{0}$. Then we have $\mathbf{E}_t = p_1^{-1}\mathbf{X}_t^\top\mathbf{R}$ and $\mathbf{F}_t = p_2^{-1}\mathbf{X}_t\mathbf{C}$. Substituting these expressions into the first equation of the system, we have $\mathbf{Z}_t = -(p_1p_2)^{-1}\mathbf{R}^\top\mathbf{X}_t\mathbf{C}$. Using $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ in (10) and (11), the factor matrices can be estimated by

$$\widehat{\mathbf{E}}_t = \frac{\mathbf{X}_t^\top\widehat{\mathbf{R}}}{p_1}, \quad \widehat{\mathbf{F}}_t = \frac{\mathbf{X}_t\widehat{\mathbf{C}}}{p_2}, \quad \widehat{\mathbf{Z}}_t = -\frac{\widehat{\mathbf{R}}^\top\mathbf{X}_t\widehat{\mathbf{C}}}{p_1p_2}. \tag{12}$$

Accordingly, the estimated signal part of RaDFaM is

$$\widehat{\mathbf{S}}_t = -\frac{\widehat{\mathbf{R}}\widehat{\mathbf{R}}^\top\mathbf{X}_t\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top}{p_1p_2} + \frac{\widehat{\mathbf{R}}\widehat{\mathbf{R}}^\top\mathbf{X}_t}{p_1} + \frac{\mathbf{X}_t\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top}{p_2}. \tag{13}$$

Hereafter, we refer to the two-step estimation procedure which is summarized in Algorithm 1 as sPCA for ease of reference.

Algorithm 1 sPCA

```
1: Input: matrix observations $\{\mathbf{X}_t\}_{t=1}^T$, factor numbers $k_1$ and $k_2$.
2: Estimate loading matrices by (10) and (11).
3: Estimate factor matrices and the signal part by (12) and (13) for $t \in [T]$.
4: Output: $\widehat{\mathbf{R}}$, $\widehat{\mathbf{C}}$, $\{\widehat{\mathbf{Z}}_t\}_{t=1}^T$, $\{\widehat{\mathbf{F}}_t\}_{t=1}^T$, $\{\widehat{\mathbf{E}}_t\}_{t=1}^T$, and $\{\widehat{\mathbf{S}}_t\}_{t=1}^T$.
```

2.3. Reconstruction Error Comparison

A grayscale digital image can be represented by a matrix in which each element represents a pixel value (i.e., brightness) ranging from 0 to 255 (Suetens 2017). We compare the matrix reconstruction performance of RaDFaM, BiMFaM, and 2w-DFM in terms of the peak signal-to-noise ratio (PSNR), which is a standard metric to measure the quality of reconstructed images compared with the originals. The PSNR is defined as

$$\mathrm{PSNR} = 10\log_{10}\left\{\frac{\|\mathbf{X}\|_{max}^2}{\|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2/(p_1p_2)}\right\},$$

where $\mathbf{X} = (\mathbf{X}_{ij}) \in \mathbb{R}^{p_1\times p_2}$ is an original image, $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_{ij}) \in \mathbb{R}^{p_1\times p_2}$ is a reconstructed image, and $\|\mathbf{X}\|_{\max}$ is the maximum of the absolute pixel values of $\mathbf{X}$; the denominator inside the logarithm is the mean square error (MSE) between $\mathbf{X}$ and $\widehat{\mathbf{X}}$, that is $\mathrm{MSE} = (p_1p_2)^{-1}\|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2$, the average of the square of pixel differences of the two images, that is, $(p_1p_2)^{-1}\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}(\mathbf{X}_{ij} - \widehat{\mathbf{X}}_{ij})^2$. Therefore, the more the reconstructed image resembles the original image, the smaller MSE will be and the larger PSNR will be (Salomon 2004). The PSNR is preferred to the MSE because it is dimensionless and the logarithm in the PSNR reduces the sensitivity to small variations in the reconstructed image.

Before conducting the comparison, we show that $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ ((10) and (11)) can also act as PCA-type estimators for the loading matrices $\mathbf{R}$ and $\mathbf{C}$ under BiMFaM and 2w-DFM. For BiMFaM, $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ are exactly α-PCA estimators with $\alpha = 0$ in Chen and Fan (2023) or the initial estimators in Yu et al. (2022). The corresponding estimators of the latent factor matrix $\mathbf{Z}_t$ and the signal part are straightforwardly

$$\widetilde{\mathbf{Z}}_t = \frac{\widehat{\mathbf{R}}^\top\mathbf{X}_t\widehat{\mathbf{C}}}{p_1p_2} \quad \text{and} \quad \widetilde{\mathbf{S}}_t = \frac{\widehat{\mathbf{R}}\widehat{\mathbf{R}}^\top\mathbf{X}_t\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top}{p_1p_2}. \tag{14}$$

For 2w-DFM, let Ψ_E and Ψ_F be the variance-covariance matrices of the rows of $\mathbf{E}_t$ and $\mathbf{F}_t$, respectively, and σ^2 be the average of the element-wise variances of the noise matrix. Based on the working variance-covariance matrix (Σ) of $\mathrm{vec}(\mathbf{X}_t)$ (Yuan et al. 2023, sec. 3.1), one may show $\mathbb{E}(\mathbf{X}_t\mathbf{X}_t^\top) = p_2\mathbf{R}\Psi_E\mathbf{R}^\top + p_2\mathrm{tr}(\Psi_F)\mathbf{I}_{p_1} + p_2\sigma^2\mathbf{I}_{p_1}$; the order of eigenvalues of $p_2\mathrm{tr}(\Psi_F)\mathbf{I}_{p_1} + p_2\sigma^2\mathbf{I}_{p_1}$ is smaller than that of $p_2\mathbf{R}\Psi_E\mathbf{R}^\top$. Then, under the strong factor assumption, (9) holds and yields $\widehat{\mathbf{R}}$, as does its column-wise counterpart, yielding $\widehat{\mathbf{C}}$. Analogously to the minimization of reconstruction error in Section 2.2, the estimators of the latent factor matrices and the signal part are

$$\check{\mathbf{F}}_t = \frac{\mathbf{X}_t\widehat{\mathbf{C}}}{p_2}, \quad \check{\mathbf{E}}_t = \frac{\mathbf{X}_t^\top\widehat{\mathbf{R}}}{p_1}, \quad \check{\mathbf{S}}_t = \frac{\widehat{\mathbf{R}}\widehat{\mathbf{R}}^\top\mathbf{X}_t}{p_1} + \frac{\mathbf{X}_t\widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top}{p_2}. \tag{15}$$

For the tth image or matrix, $\widehat{\mathbf{S}}_t$, $\widetilde{\mathbf{S}}_t$, and $\check{\mathbf{S}}_t$ in (13)–(15), which are the specific estimates of $\mathbf{X}_t$ under unified PCA estimates $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$, will yield their levels of the PSNR or the MSE under the three matrix factor models. The following proposition shows that RaDFaM is superior to BiMFaM and 2w-DFM, respectively, in terms of the PSNR or the MSE.

Proposition 2. For each matrix observation $\mathbf{X}_t$, $t \in [T]$, we have

$$\mathrm{PSNR}_{t,RaD} \geq \mathrm{PSNR}_{t,BiM} \text{ and } \mathrm{PSNR}_{t,RaD} \geq \mathrm{PSNR}_{t,2w};$$

or equivalently,

$$\mathrm{MSE}_{t,RaD} \leq \mathrm{MSE}_{t,BiM} \text{ and } \mathrm{MSE}_{t,RaD} \leq \mathrm{MSE}_{t,2w}.$$

In addition, our numerical studies show that, even based on other PCA-type estimators, that is, autoPCA (Wang, Liu, and Chen 2019) and proPCA (Yu et al. 2022) for BiMFaM, and StepApp for 2w-DFM (Yuan et al. 2023), RaDFaM still has better reconstruction performance.

2.4. Factor Number Estimation

The above subsections all assume known k_1 and k_2. When k_1 and k_2 are unknown, we need to determine them before estimation. We adopt general ratio-type estimators for factor numbers. Let $\widehat{\lambda}_1(\widehat{\mathbf{M}}_1) \geq \widehat{\lambda}_2(\widehat{\mathbf{M}}_1) \geq \cdots \geq \widehat{\lambda}_{p_1}(\widehat{\mathbf{M}}_1) \geq 0$ denote the ordered eigenvalues of $\widehat{\mathbf{M}}_1$ and let $k_{\max}$ be a given upper bound for k_1. The number of row-wise factors can be estimated by

$$\widehat{k}_1 = \arg\max_{1 \leq j \leq k_{\max}} \frac{\widehat{\lambda}_j(\widehat{\mathbf{M}}_1)}{\widehat{\lambda}_{j+1}(\widehat{\mathbf{M}}_1)}, \tag{16}$$

and the column-wise $\widehat{k}_2$ can be defined similarly. This ratio-type estimator is widely used in factor models including the vector case and the matrix case (Lam and Yao 2012; Yu et al. 2022). We will prove its consistency in the next section.

3. Asymptotic Properties

In this section, we state necessary assumptions in Section 3.1 and provide asymptotic properties including the rate of convergence for estimators of loading matrices and the signal parts, the asymptotic distributions of estimators of the loading matrices, and the consistency of estimators of factor numbers, in Section 3.2.

3.1. Assumptions

Assumption 1. α-mixing: the vectorized factor processes $\{\mathrm{vec}(\mathbf{Z}_t)\}$, $\{\mathrm{vec}(\mathbf{F}_t)\}$, $\{\mathrm{vec}(\mathbf{E}_t)\}$, and the noise process $\{\mathrm{vec}(\mathbf{e}_t)\}$ are α-mixing.

A vector process $\{\boldsymbol{u}_t\}$ is α-mixing, if $\sum_{h=1}^{\infty} \alpha(h)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, where $\alpha(h) = \sup_i \sup_{A \in \mathcal{C}_{-\infty}^{i}, B \in \mathcal{C}_{i+h}^{\infty}} |P(A \cap B) - P(A)P(B)|$ with $\mathcal{C}_i^j$ the σ-field generated by $\left\{\boldsymbol{u}_t : i \leq t \leq j\right\}$. The α-mixing condition means that variables that are sufficiently far apart are asymptotic independent (Francq and Zakoian 2019, Appendix A.3).

Assumption 2. Common factors: assume that k_1 and k_2 are fixed. For any $t \in [T], i \in [p_1], j \in [p_2], u \in [k_1]$ and $v \in [k_2]$, there exists a positive constant m such that $\mathbb{E}(\mathbf{Z}_{t,uv}^4) \leq m$, $\mathbb{E}(\mathbf{F}_{t,iv}^8) \leq m$, $\mathbb{E}(\mathbf{E}_{t,ju}^8) \leq m$ and

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_t \mathbf{Z}_t^\top \xrightarrow{a.s.} \Sigma_{Z1}, \quad \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_t^\top \mathbf{Z}_t \xrightarrow{a.s.} \Sigma_{Z2},$$

$$\frac{1}{Tp_1}\sum_{t=1}^{T} \mathbf{F}_t^\top \mathbf{F}_t \xrightarrow{a.s.} \Sigma_F, \quad \frac{1}{Tp_2}\sum_{t=1}^{T} \mathbf{E}_t^\top \mathbf{E}_t \xrightarrow{a.s.} \Sigma_E,$$

where $\Sigma_{Z1} \in \mathbb{R}^{k_1 \times k_1}$, $\Sigma_{Z2} \in \mathbb{R}^{k_2 \times k_2}$, $\Sigma_F \in \mathbb{R}^{k_2 \times k_2}$ and $\Sigma_E \in \mathbb{R}^{k_1 \times k_1}$ are positive definite matrices. Let $\Sigma_1 = \Sigma_E + \Sigma_{Z1}$ and $\Sigma_2 = \Sigma_F + \Sigma_{Z2}$. Assume that Σ_1 and Σ_2 have spectral decompositions $\Sigma_1 = \Gamma_1 \Lambda_1 \Gamma_1^\top$ and $\Sigma_2 = \Gamma_2 \Lambda_2 \Gamma_2^\top$, where the diagonal elements of Λ_1 and Λ_2 are distinct and arranged in decreasing order.

Assumption 2 requires the interactive factor matrix $\mathbf{Z}_t$ to have bounded fourth moments, and the row-wise and column-wise factor matrices $\mathbf{F}_t$ and $\mathbf{E}_t$ to have bounded eighth moments. It also requires the second sample moments of the factor matrices to converge to positive definite matrices, as can be derived under Assumption 1; see Chapter 16 in Athreya and Lahiri (2006) and Appendix A.3. in Francq and Zakoian (2019) for more details. The assumption of distinct and decreasing eigenvalues of Σ_1 and Σ_2 results in unique eigen-decompositions and identifiable eigenvectors. Assumption 2 is an extension of the assumption in Yu et al. (2022) with extra row-wise and column-wise factors.

Assumption 3. Loading matrices: there exist positive constants $\bar{r}$ and $\bar{c}$ such that $\|\mathbf{R}\|_{max} \leq \bar{r}$ and $\|\mathbf{C}\|_{max} \leq \bar{c}$. As $\min\left\{p_1, p_2\right\} \to \infty$, we have $\|p_1^{-1}\mathbf{R}^\top\mathbf{R} - \mathbf{I}_{k_1}\| \to 0$ and $\|p_2^{-1}\mathbf{C}^\top\mathbf{C} - \mathbf{I}_{k_2}\| \to 0$, where $\|\cdot\|$ is the L_2-norm of a matrix.

Assumption 3 is a model identification condition. It guarantees that our model belongs to the strong factor regime, which means that the factors are pervasively shared by elements of observations, and the signal component has spiked eigenvalues relative to the noise component (Stock and Watson 2002; Lam and Yao 2011; Fan, Liao, and Mincheva 2013; Chen, Fan, and Wang 2020).

Assumption 4. Correlation: for any $s, t, t_1 \in [T], i, i_1, i_2 \in [p_1]$, $j, j_1, j_2 \in [p_2], u_1, u_2, u_3, u_4 \in [k_1]$ and $v_1, v_2, v_3, v_4 \in [k_2]$, there exists a positive constant m such that

(4.1) $\sum_{t_2=1}^{T} |\mathbb{E}(\mathbf{Z}_{t_1,u_1v_1}\mathbf{Z}_{t_2,u_2v_2})| \leq m.$

(4.2) $\mathbb{E}(\mathbf{e}_{t,ij}^8) \leq m;\ \sum_{t_2=1}^{T}\sum_{i_2=1}^{p_1}\sum_{j_2=1}^{p_2} |\mathbb{E}(\mathbf{e}_{t_1,i_1j_1}\mathbf{e}_{t_2,i_2j_2})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{i_3,i_4=1}^{p_1}\sum_{j_3,j_4=1}^{p_2} |\mathrm{cov}(\mathbf{e}_{t_1,i_1j_1}\mathbf{e}_{t_1,i_2j_2}, \mathbf{e}_{t_2,i_3j_3}\mathbf{e}_{t_2,i_4j_4})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{i_3,i_4=1}^{p_1}\sum_{j_3,j_4=1}^{p_2} |\mathrm{cov}(\mathbf{e}_{s,i_1j_1}\mathbf{e}_{t_1,i_2j_2}, \mathbf{e}_{s,i_3j_3}\mathbf{e}_{t_2,i_4j_4})| \leq m.$

(4.3) $\sum_{t_2=1}^{T}\sum_{i_2=1}^{p_1} |\mathbb{E}(\mathbf{F}_{t_1,i_1v_1}\mathbf{F}_{t_2,i_2v_2})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{i_3,i_4=1}^{p_1} |\mathrm{cov}(\mathbf{F}_{t_1,i_1v_1}\mathbf{F}_{t_1,i_2v_2}, \mathbf{F}_{t_2,i_3v_3}\mathbf{F}_{t_2,i_4v_4})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{i_3,i_4=1}^{p_1} |\mathrm{cov}(\mathbf{F}_{s,i_1v_1}\mathbf{F}_{t_1,i_2v_2}, \mathbf{F}_{s,i_3v_3}\mathbf{F}_{t_2,i_4v_4})| \leq m.$

(4.4) $\sum_{t_2=1}^{T}\sum_{j_2=1}^{p_2} |\mathbb{E}(\mathbf{E}_{t_1,j_1u_1}\mathbf{E}_{t_2,j_2u_2})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{j_3,j_4=1}^{p_2} |\mathrm{cov}(\mathbf{E}_{t_1,j_1u_1}\mathbf{E}_{t_1,j_2u_2}, \mathbf{E}_{t_2,j_3u_3}\mathbf{E}_{t_2,j_4u_4})| \leq m;$

$\sum_{t_2=1}^{T}\sum_{j_3,j_4=1}^{p_2} |\mathrm{cov}(\mathbf{E}_{s,j_1u_1}\mathbf{E}_{t_1,j_2u_2}, \mathbf{E}_{s,j_3u_3}\mathbf{E}_{t_2,j_4u_4})| \leq m.$

Assumption 4 focuses on the cross-sectional and temporal correlation of the noise matrices $\mathbf{e}_t$ and the factor matrices $\mathbf{Z}_t$, $\mathbf{F}_t$, and $\mathbf{E}_t$. Assumption 4.1 requires the temporal correlation of $\mathbf{Z}_t$ to be weak, but no constraint is put on the cross-sectional correlation $\mathbf{Z}_t$ due to the finiteness of k_1 and k_2. Similar to Assumption 4.1, Assumptions 4.2–4.4 are made for $\mathbf{e}_t$, $\mathbf{F}_t$, and $\mathbf{E}_t$, respectively. Taking Assumption 4.2 of the matrix noise as an example, it requires a bounded eighth moment for each element, the cross-sectional and temporal correlation to be weak, and the cross-sectional and temporal correlation of $\mathbf{e}_t \circ$

$\mathbf{e}_t \in \mathbb{R}^{p_1\times p_2\times p_1\times p_2}$ and $\mathbf{e}_s \circ \mathbf{e}_t \in \mathbb{R}^{p_1\times p_2\times p_1\times p_2}$ to be weak. In summary, we assume the correlations in the noise part and the factors are weak, and the independent matrix observations satisfy Assumption 4.

Assumption 5. Central limit theorems: for $i \in [p_1]$, $j \in [p_2]$, $s, t \in [T]$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{Z}_t\mathbf{F}_{t,i\cdot} \xrightarrow{d} N(\mathbf{0},\mathbf{V}_{1i}),\quad \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{Z}_t^\top\mathbf{E}_{t,j\cdot} \xrightarrow{d} N(\mathbf{0},\mathbf{V}_{2j}),$$

where $\mathbf{V}_{1i} = \lim_{T\to\infty} T^{-1}\sum_{s,t}\mathbb{E}(\mathbf{Z}_t\mathbf{F}_{t,i\cdot}\mathbf{F}_{s,i\cdot}^\top\mathbf{Z}_s^\top)$ and $\mathbf{V}_{2j} = \lim_{T\to\infty} T^{-1}\sum_{s,t}\mathbb{E}(\mathbf{Z}_t^\top\mathbf{E}_{t,j\cdot}\mathbf{E}_{s,j\cdot}^\top\mathbf{Z}_s^\top)$.

Assumption 5 is useful when deriving the asymptotic distributions of the estimated loading matrices. Under Assumptions 1-4, Assumption 5 can be derived by applying the central limit theorem for α-mixing processes (Athreya and Lahiri 2006; Francq and Zakoian 2019). Assumption 5 is different from the corresponding condition for BiMFaM in that the main terms differ due to the incorporation of row-wise and column-wise factors.

3.2. Asymptotic Properties

Let $\widehat{\Lambda}_1$ and $\widehat{\Lambda}_2$ be the diagonal matrices with elements being the top k_1 and k_2 eigenvalues of $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$, respectively. Based on $\widehat{\mathbf{M}}_1\widehat{\mathbf{R}} = \widehat{\mathbf{R}}\widehat{\Lambda}_1$ and $\widehat{\mathbf{M}}_2\widehat{\mathbf{C}} = \widehat{\mathbf{C}}\widehat{\Lambda}_2$, one may define the following asymptotic orthogonal matrices (see Appendix D in the online supplement for more details),

$$\mathbf{H}_1 = \frac{1}{Tp_1}\sum_{t=1}^{T}\mathbf{Z}_t\mathbf{Z}_t^\top\mathbf{R}^\top\widehat{\mathbf{R}}\widehat{\Lambda}_1^{-1} + \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\mathbf{E}_t^\top\mathbf{E}_t\mathbf{R}^\top\widehat{\mathbf{R}}\widehat{\Lambda}_1^{-1},$$
$$\mathbf{H}_2 = \frac{1}{Tp_2}\sum_{t=1}^{T}\mathbf{Z}_t^\top\mathbf{Z}_t\mathbf{C}^\top\widehat{\mathbf{C}}\widehat{\Lambda}_2^{-1} + \frac{1}{Tp_1p_2}\sum_{t=1}^{T}\mathbf{F}_t^\top\mathbf{F}_t\mathbf{C}^\top\widehat{\mathbf{C}}\widehat{\Lambda}_2^{-1}.$$

We first present the following convergence rates of $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ in the Frobenius norm.

Theorem 1. Suppose that T, p_1, and p_2 tend to infinity, and that k_1 and k_2 are fixed. If Assumptions 1–4 hold, then there exist matrices $\mathbf{H}_1$ and $\mathbf{H}_2$, satisfying $\mathbf{H}_1^\top\mathbf{H}_1 \xrightarrow{p} \mathbf{I}_{k_1}$ and $\mathbf{H}_2^\top\mathbf{H}_2 \xrightarrow{p} \mathbf{I}_{k_2}$, such that

$$\frac{1}{p_1}||\widehat{\mathbf{R}} - \mathbf{R}\mathbf{H}_1||_F^2 = O_p(\frac{1}{T} + \frac{1}{p_1^2}),$$
$$\frac{1}{p_2}||\widehat{\mathbf{C}} - \mathbf{C}\mathbf{H}_2||_F^2 = O_p(\frac{1}{T} + \frac{1}{p_2^2}).$$

Our convergence rates for $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ are slower than those for PCA-type estimators under BiMFaM, though they achieve those in the high-dimensional vector factor model (Bai 2003); the convergence rates of proPCA (Yu et al. 2022) are so far the fastest in the class of PCA-based methods. This result reflects our strategy of sacrificing the theoretical rate of convergence to enhance the signal strength.

We next derive the asymptotic distributions of the estimated loading matrices. As the row sizes of $\mathbf{R}$ and $\mathbf{C}$ tend to infinity, we establish the row-wise asymptotic distribution in the following Theorem 2.

Theorem 2. Suppose that T, p_1, and p_2 tend to infinity, and that k_1 and k_2 are fixed. If Assumptions 1–5 hold, then

1. for $i \in [p_1]$, we have

$$\begin{cases}\sqrt{T}(\widehat{\mathbf{R}}_{i\cdot} - \mathbf{H}_1^\top\mathbf{R}_{i\cdot}) \xrightarrow{d} N(\mathbf{0}, \Lambda_1^{-1}\Gamma_1^\top\mathbf{V}_{1i}\Gamma_1\Lambda_1^{-1}), & T = o(p_1^2),\\ \widehat{\mathbf{R}}_{i\cdot} - \mathbf{H}_1^\top\mathbf{R}_{i\cdot} = O_p(\frac{1}{p_1}), & p_1^2 = O(T);\end{cases}$$

2. for $j \in [p_2]$, we have

$$\begin{cases}\sqrt{T}(\widehat{\mathbf{C}}_{j\cdot} - \mathbf{H}_2^\top\mathbf{C}_{j\cdot}) \xrightarrow{d} N(\mathbf{0}, \Lambda_2^{-1}\Gamma_2^\top\mathbf{V}_{2j}\Gamma_2\Lambda_2^{-1}), & T = o(p_2^2),\\ \widehat{\mathbf{C}}_{j\cdot} - \mathbf{H}_2^\top\mathbf{C}_{j\cdot} = O_p(\frac{1}{p_2}), & p_2^2 = O(T).\end{cases}$$

Theorem 2 shows that the loading matrix estimator can be asymptotically normal when p_1 and p_2 are sufficiently large. Even if the central limit theorem does not hold, the rows of estimated loading matrices are still consistent with no restriction on the limiting relationship between T and $\{p_1, p_2\}$.

We now turn to the convergence rate of the estimated signal component $\widehat{\mathbf{S}}_t$ in (13). As the size of $\widehat{\mathbf{S}}_t$ tends to infinity, we will prove its element-wise consistency.

Theorem 3. Suppose that T, p_1, and p_2 tend to infinity, and that k_1 and k_2 are fixed. If Assumptions 1–4 hold, then we have

$$|\widehat{\mathbf{S}}_{t,ij} - \mathbf{S}_{t,ij}| = O_p(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p_1}} + \frac{1}{\sqrt{p_2}}),$$
$$\text{for any } i \in [p_1] \text{ and } j \in [p_2].$$

Similar to the convergence rates of the estimators of the loading matrices, the convergence rate of the estimator of the signal component in RaDFaM is slower than that in BiMFaM. However, as shown in Proposition 2, RaDFaM has better reconstruction performance.

Theorems 1–3 are proved with fixed factor numbers k_1 and k_2. If the factor numbers are unknown, we show that the ratio-type estimator (16) provides consistent estimators.

Theorem 4. Suppose that T, p_1, and p_2 tend to infinity, and k_{max} is not smaller than $\max\{k_1, k_2\}$. If Assumptions 1–4 hold, then we have

$$P(\widehat{k}_1 \neq k_1) \to 0 \quad \text{and} \quad P(\widehat{k}_2 \neq k_2) \to 0.$$

Theorem 4 shows that with large enough $k_{\max}$, the factor numbers can be estimated consistently. The proof follows arguments similar to those used in the proof of Lemma 1 given in the online supplement. As $\widehat{\mathbf{R}}$ and $\widehat{\mathbf{C}}$ of sPCA are identical to the estimator of $\mathbf{R}$ and $\mathbf{C}$ in α-PCA with $\alpha = 0$, they can use the same ratio-type estimators.

4. Simulation Studies

In this section, we present simulations to demonstrate the finite sample performance of the proposed RaDFaM. The simulation settings for six different scenarios and measures of performance are introduced in Section 4.1. The simulation results assessing the performance of loading matrix estimators, signal part estimators and matrix reconstruction, the computation time, the estimation error of factor numbers, and the asymptotic normality of loading matrix estimators are reported in Section 4.2.

4.1. Simulation Settings and Performance Measures

We first introduce the simulation settings. The matrix observations are generated with factor numbers $(k_1, k_2) = (3, 3)$, and 6 sets of (T, p_1, p_2), which are $(20, 50, 100)$, $(20, 100, 100)$, $(50, 20, 50)$, $(50, 100, 100)$, $(100, 20, 50)$, and $(100, 50, 100)$, respectively. The loading matrix $\mathbf{R}$ is taken as $\sqrt{p_1}$ times the matrix of the top k_1 left singular vectors of the SVD decomposition of a $p_1 \times k_1$-dimensional matrix with independent standard normal elements, and the loading matrix $\mathbf{C}$ is generated similarly. Then $\mathbf{Z}_t$, $\mathbf{E}_t$, and $\mathbf{F}_t$ are generated from the following vector auto-regression models

$$\begin{aligned}
\text{vec}(\mathbf{Z}_t) &= \phi \text{vec}(\mathbf{Z}_{t-1}) + \sqrt{1-\phi^2}\mathbf{u}_t,\\
\text{vec}(\mathbf{F}_t) &= \psi \text{vec}(\mathbf{F}_{t-1}) + \sqrt{1-\psi^2}\boldsymbol{\xi}_t,\\
\text{vec}(\mathbf{E}_t) &= \gamma\, vec(\mathbf{E}_{t-1}) + \sqrt{1-\gamma^2}\boldsymbol{\eta}_t,
\end{aligned}$$

where $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{I}_{k_1k_2})$, $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \mathbf{I}_{p_1k_2})$, $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{I}_{p_2k_1})$, and the entries of the initial values of $\text{vec}(\mathbf{Z}_t)$, $\text{vec}(\mathbf{F}_t)$ and $\text{vec}(\mathbf{E}_t)$ are standard normal. The coefficients ϕ, ψ and γ control the temporal correlation of the factor matrices. We generate $\mathbf{e}_t \sim MN(\mathbf{0}; \Omega_e, \Delta_e)$, where Ω_e and Δ_e both have 1's on the diagonal and constant off-diagonal elements $1/p_1$ and $1/p_2$, respectively. We consider six scenarios shown in Table 1: the first three are the uncorrelated cases, and the last three are the correlated cases

We compare sPCA under RaDFaM; autoPCA (Wang, Liu, and Chen 2019), α-PCA with $\alpha = 0$ (Chen and Fan 2023), and proPCA (Yu et al. 2022) under BiMFaM; 2w-PCA (estimators in (15)), and Step-App and Q-MLE, the initial estimators and quasi-likelihood estimators in Yuan et al. (2023), respectively, under 2w-DFM.

Next, we introduce the measures of performance. For the loading matrices, we use the following column space distance to characterize the performance of factor loading estimators

$$\begin{aligned}
\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R}) &= \frac{1}{p_1}\left\| \widehat{\mathbf{R}}\widehat{\mathbf{R}}^\top - \mathbf{R}\mathbf{R}^\top \right\|_2,\\
\mathcal{D}(\widehat{\mathbf{C}}, \mathbf{C}) &= \frac{1}{p_2}\left\| \widehat{\mathbf{C}}\widehat{\mathbf{C}}^\top - \mathbf{C}\mathbf{C}^\top \right\|_2.
\end{aligned}$$

For the signal part, we measure the distance between the estimator and the true signal using

$$\mathcal{D}_{signal} = \sqrt{\frac{1}{Tp_1p_2}\sum_{t=1}^{T} \|\widehat{\mathbf{S}}_t - \mathbf{S}_t\|_F^2}.$$

Table 1. Simulation scenarios.

		ϕ	ψ	γ	Generating Model
Uncorrelated	Scenario I	0	0	0	RaDFaM
	Scenario II	0	–	–	BiMFaM
	Scenario III	–	0	0	2w-DFM
Correlated	Scenario IV	0.6	0.8	0.8	RaDFaM
	Scenario V	0.6	–	–	BiMFaM
	Scenario VI	–	0.8	0.8	2w-DFM

For assessing the reconstruction performance, we use

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\text{MSE}_t}, \quad \overline{\text{PSNR}} = \frac{1}{T}\sum_{t=1}^{T}\text{PSNR}_t,$$

where the MSE_t and the PSNR_t are defined in Section 2.3. For the estimated factor numbers, we assess the accuracy by

$$\text{acc} = \mathbf{I}(\widehat{k}_1 = k_1, \widehat{k}_2 = k_2) * 100\%.$$

4.2. Simulation Results

First, we focus on the performance of row-wise loading matrix estimators and the similar results for the column-wise counterpart are left in Appendix H of the online supplement. Boxplots of $\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R})$ for the seven comparison methods over 100 replications are reported in Figure 1. Note that α-PCA, 2w-PCA, and sPCA provide identical estimators regardless of the generating model, so they have identical results under all scenarios. For Scenarios I and IV, autoPCA performs the worst, proPCA performs poorly, and Step-App and Q-MLE perform the best, followed by α-PCA/2w-PCA/sPCA; for Scenarios II and V, Q-MLE fails most (abnormal upright bars in Subfigures b and e) and performs poorly when it works, Step-App performs worst, and other methods have similar results; for Scenarios III and VI, proPCA performs the worst, autoPCA performs poorly for correlated observations, and Step-App and Q-MLE perform the best, followed by α-PCA/2w-PCA/sPCA.

Second, for the performance of the signal parts estimators and matrix reconstruction, boxplots of the $\mathcal{D}_{\text{signal}}$, the RMSE and the $\overline{\text{PSNR}}$ from 100 replications are shown in Figures 2–4. For Scenarios I and IV, 2w-PCA and Step-App perform the worst, autoPCA, α-PCA, and proPCA perform poorly, and sPCA and Q-MLE perform the best; for Scenarios II and V, Q-MLE fails at most cases, 2w-PCA performs worst, Step-App performs poorly, and other methods have similar results; for Scenarios III and VI, autoPCA, α-PCA, and proPCA perform the worst, 2w-PCA and Step-App perform poorly, and sPCA and Q-MLE perform the best.

Third, to evaluate the computational time required for different methods, boxplots of the run time in seconds from 100 replications are shown in Figure 5. It is evident that the Q-MLE incurs the highest computational time cost, particularly for large matrices and sample sizes. Following Q-MLE, autoPCA exhibits a relatively higher time cost, while the remaining methods show similar time requirements.

Fourth, the results of the estimation error of factor numbers are shown in Table 2, from which it can be seen that α-PCA/2w-PCA/sPCA estimate the factor numbers almost correctly under all scenarios, autoPCA does not work under RaDFaM and 2w-DFM, proPCA does not work under 2w-DFM, and Step-App and Q-MLE do not work under BiMFaM.

Finally, we illustrate the asymptotic normality of the sPCA estimator $\widehat{\mathbf{R}}$ under Scenario I with $(T, p_1, p_2) = (100, 150, 150)$. In this case, the asymptotic variance-covariance matrix of $\widehat{\mathbf{R}}_{i\cdot}$ is $k_2/(k_2+1)^2\mathbf{I}_{k_1}$, so we present the histogram and QQ plot of the third element of $(k_2+1)\sqrt{T/k_2}(\widehat{\mathbf{R}}_{1\cdot} - \mathbf{H}_1^\top\mathbf{R}_{1\cdot})$ over 1000 replications in Figure 6. The smooth curve in the left plot is the

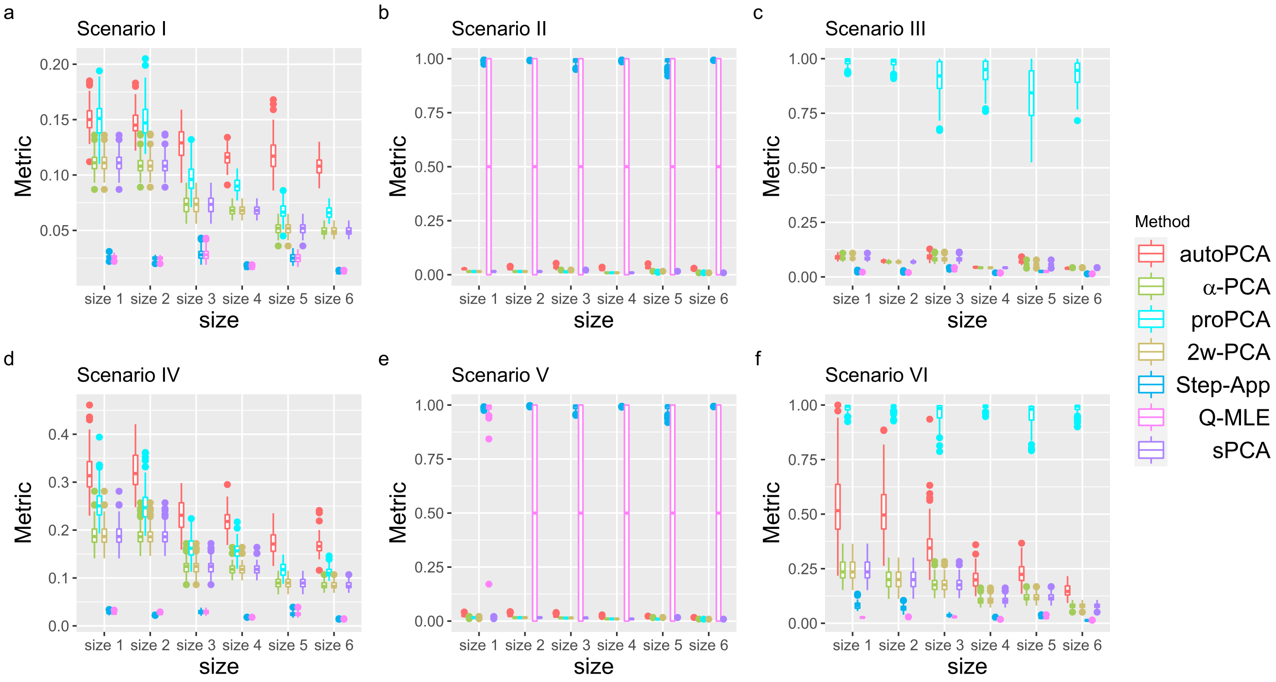


Figure 1. Boxplots of $\mathcal{D}(\widehat{\mathbf{R}}, \mathbf{R})$ for seven estimation methods under six size settings in six scenarios.

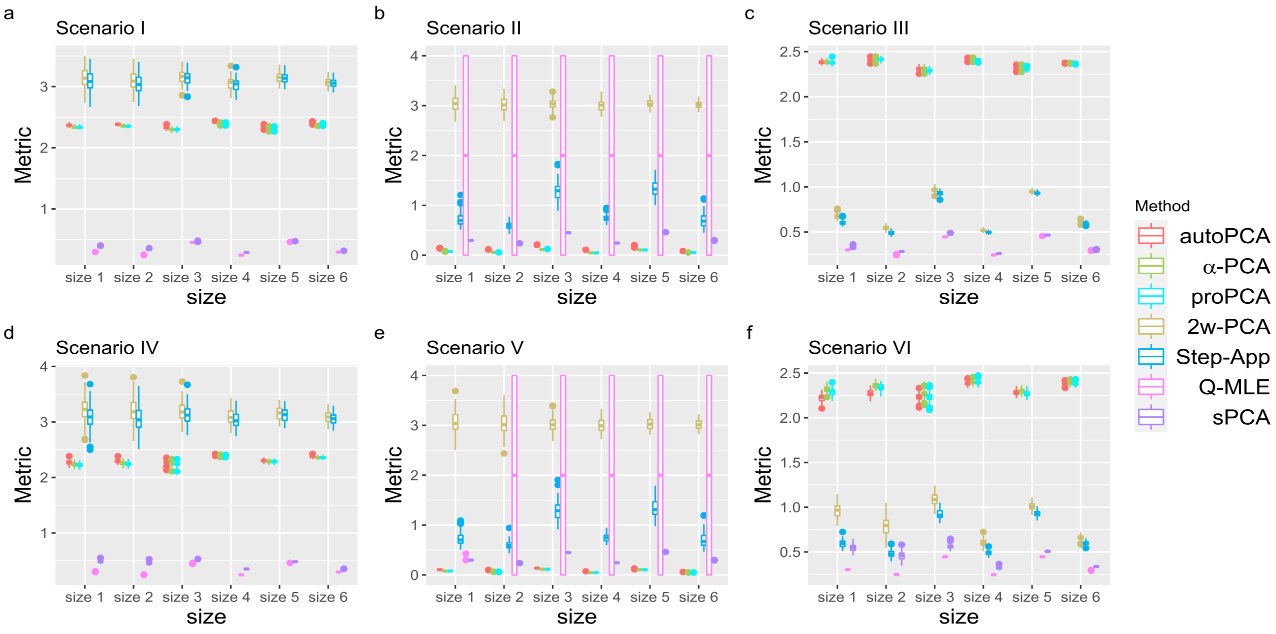


Figure 2. Boxplots of $\mathcal{D}_{\text{signal}}$ for seven estimation methods under six size settings in six scenarios.

probability density function of the standard normal distribution, and the QQ plot supports the normal approximation (the *p*-value of the Shapiro-Wilk test for normality is 0.4243).

In summary, based on PCA-type estimation, sPCA demonstrates excellent performance in estimating loading matrices and factor numbers, and outperforms other methods in terms of matrix reconstruction, even when dealing with data generated from BiMFaM and 2w-DFM. Compared to Q-MLE, which refines Step-App, sPCA's performance in estimating loading matrices, factor numbers, and matrix reconstruction is comparable with much less computational time, and is robust for dealing with data generated from BiMFaM when Q-MLE usually fails.

5. Real Data Analysis

5.1. Uncorrelated Data: CT Images for COVID-19

In this section, we apply the proposed approach to a real open-source chest CT dataset called COVID-CT, which was one of the largest publicly available COVID-19 CT scan dataset in the early pandemic (Yang et al. 2020; Zhang et al. 2023). It contains 349 COVID-19 positive CT scans manually selected from the embedded figures out of 760 preprints on medRxiv2 and bioRxiv3, and 397 negative CT scans from four open-access databases. Due to figure format problems, we finally have 317 positive scans and 397 negative scans. As discussed in Section 2.3, grayscale image data can be regarded as matrix data,

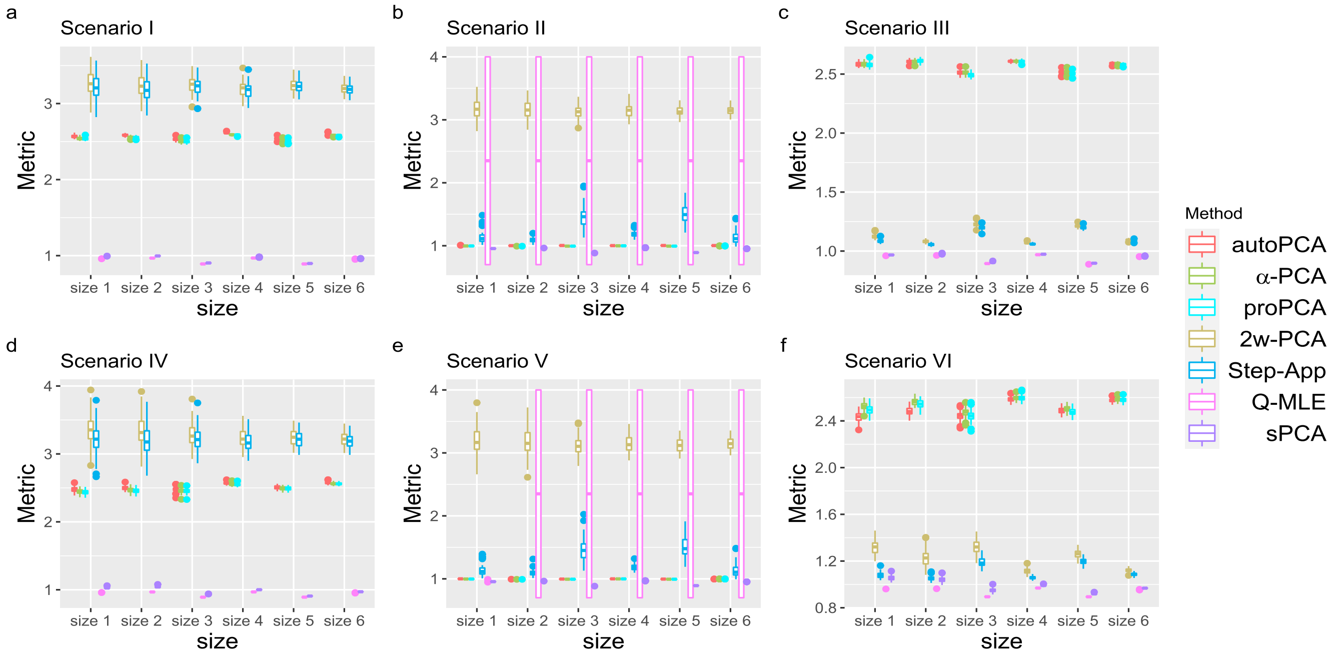


Figure 3. Boxplots of the RMSE for seven estimation methods under six size settings in six scenarios.

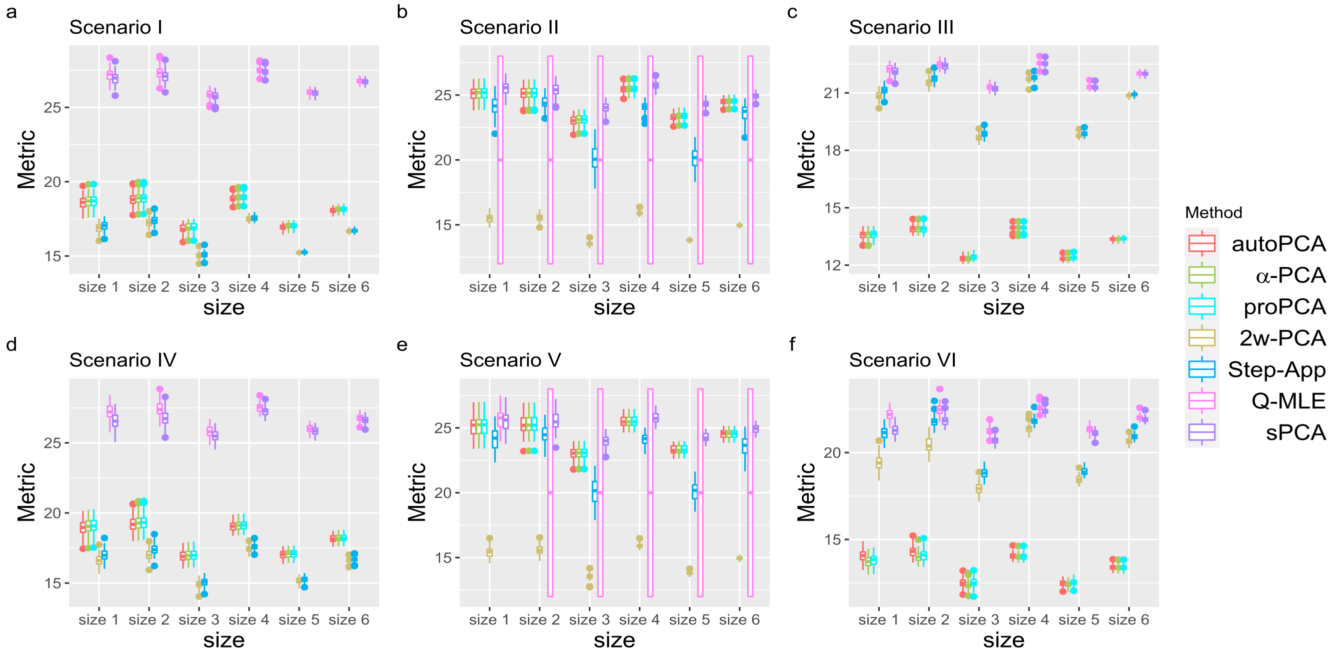


Figure 4. Boxplots of the $\overline{\text{PSNR}}$ for seven estimation methods under six size settings in six scenarios.

we have $T = 714$, $p_1 = 150$ and $p_2 = 150$ by resizing each image to 150 × 150.

Image reconstruction RaDFaM, BiMFaM, and 2w-DFM can be used to extract low-dimensional features from the matrix-valued observations, and the estimated low-rank signal parts can be regarded as the reconstruction of the original image. We conduct a comparison of the reconstruction performance of the seven methods used in simulation studies. The factor numbers k_1 and k_2 are set to be equal and take on the values of 5, 15, 25, 35, and 45. Furthermore, we also compare with the autoencoder (AE) methods: i) AE-1 and AE-5 with 1 and 5 hidden layers are used, and the number of neurons are set to be k_1 and $(256, 128, k_1, 128, 256)$, respectively (Hinton and Salakhutdinov 2006; Fan, Ma, and Zhong 2021). We use the whole sample as the batch, and carry out 500 epochs; ii) a convolutional autoencoder (CAE) of 8 hidden layers is used to make full use of the matrix structural information, where the number of neurons in the central fully connected bottleneck layer is set to be k_1 (Guo et al. 2017). We take 32 as the batch size, and carry out 50 epochs.

The results of the RMSE and the $\overline{\text{PSNR}}$ for all methods are shown in Tables 3 and 4, respectively (bold values are the best ones). We can see that: (i) except 2w-PCA, Step-App and AE

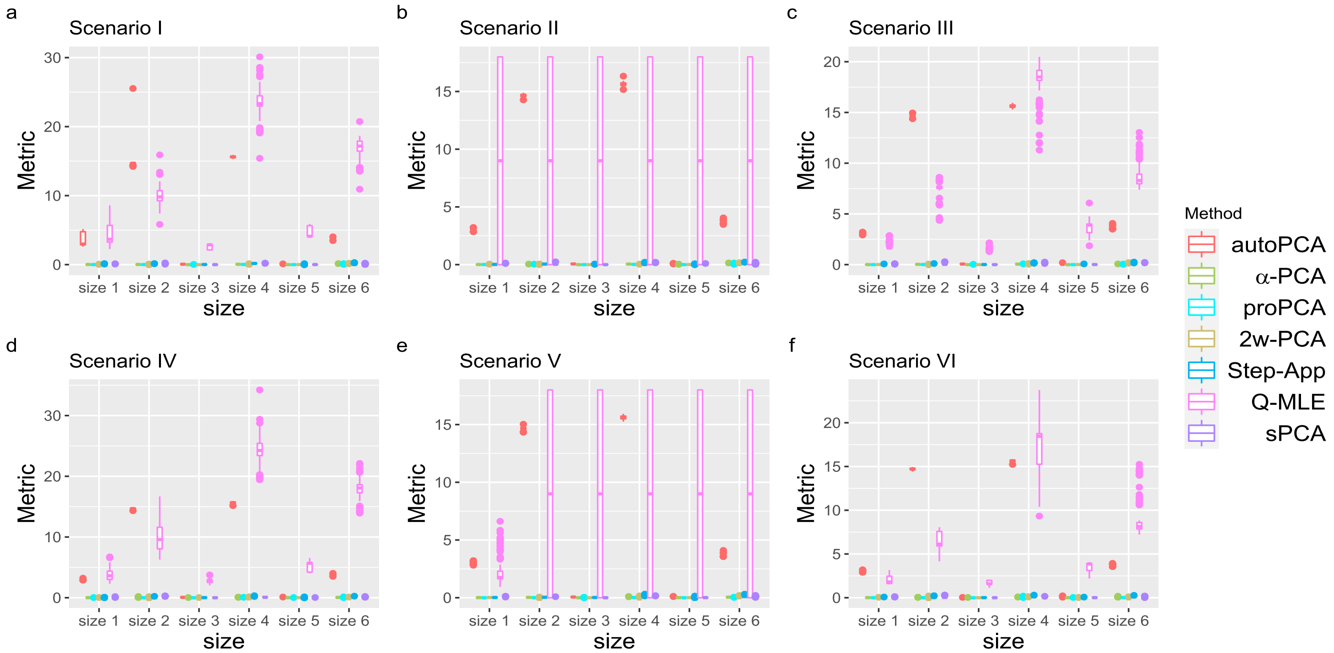


Figure 5. Boxplots of the computational time for seven estimation methods under six size settings in six scenarios.

Table 2. The result for estimated factor numbers.

	Scenario I						Scenario II						Scenario III					
Methods	size 1	size 2	size 3	size 4	size 5	size 6	size 1	size 2	size 3	size 4	size 5	size 6	size 1	size 2	size 3	size 4	size 5	size 6
autoPCA	2	10	6	99	61	100	100	100	100	100	100	100	0	0	0	0	0	0
α-PCA	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
proPCA	100	100	100	100	100	100	100	100	100	100	100	100	2	1	13	2	18	18
2w-PCA	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Step-App	100	100	100	100	100	100	0	0	0	0	0	0	100	100	100	100	100	100
Q-MLE	100	100	100	100	100	100	–	–	–	–	–	–	100	100	100	100	100	100
sPCA	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
	Scenario IV						Scenario V						Scenario VI					
Methods	size 1	size 2	size 3	size 4	size 5	size 6	size 1	size 2	size 3	size 4	size 5	size 6	size 1	size 2	size 3	size 4	size 5	size 6
autoPCA	0	0	0	0	0	0	100	100	100	100	100	100	33	57	1	77	15	4
α-PCA	100	100	100	100	100	100	100	100	100	100	100	100	94	100	100	100	100	100
proPCA	99	99	100	100	100	100	100	100	100	100	100	100	4	6	0	2	4	1
2w-PCA	100	100	100	100	100	100	100	100	100	100	100	100	94	100	100	100	100	100
Step-App	100	100	100	100	100	100	0	0	0	0	0	0	100	100	100	100	100	100
Q-MLE	100	100	100	100	100	100	0	–	–	–	–	–	100	100	100	100	100	100
sPCA	100	100	100	100	100	100	100	100	100	100	100	100	94	100	100	100	100	100

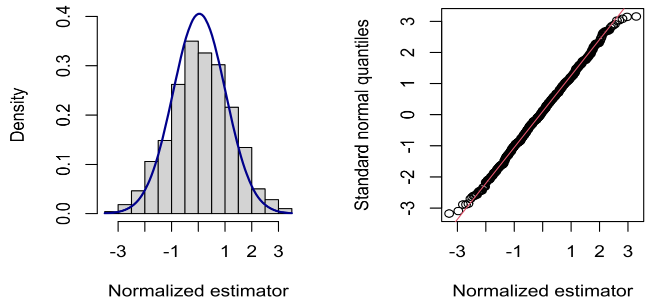
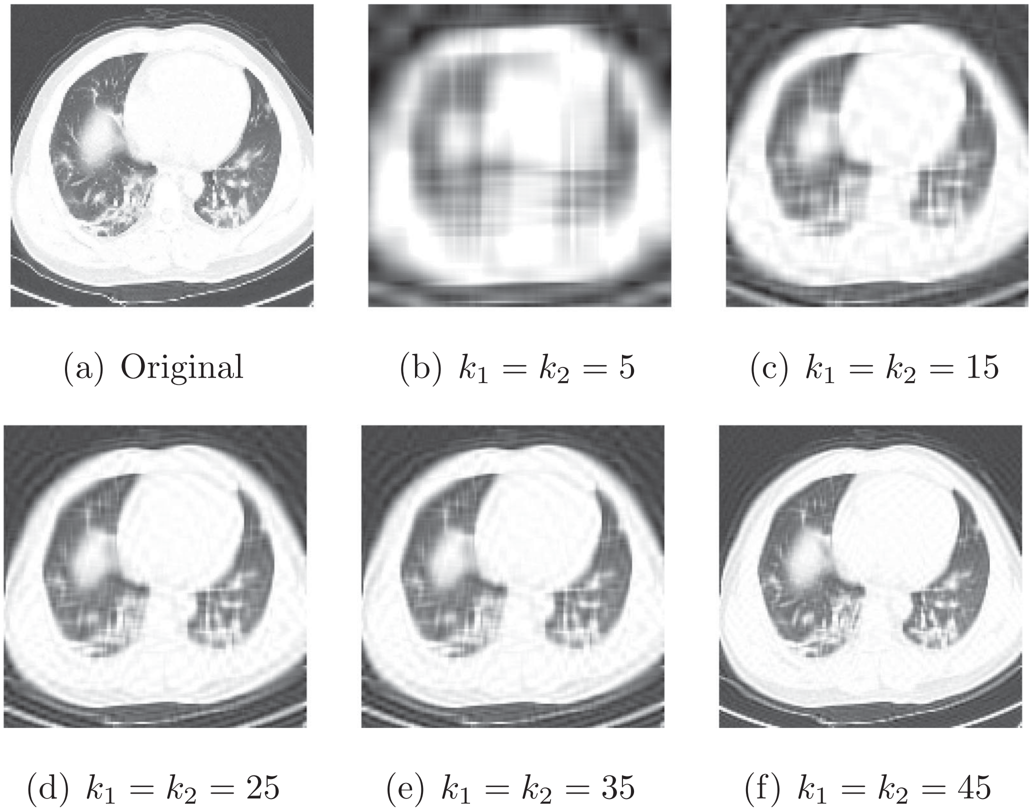


Figure 6. Histogram (with superimposed standard normal density) and QQ-plot of the normalized estimator $\widehat{\mathbf{R}}_{13}$.

Table 3. RMSEs of different estimation methods for different choices of (k_1, k_2) applied to the COVID-CT data.

Model	Method/ (k_1, k_2)	$(5, 5)$	$(15, 15)$	$(25, 25)$	$(35, 35)$	$(45, 45)$
BiMFaM	autoPCA	0.2009	0.1347	0.1097	0.0949	0.0836
	α-PCA	0.1907	0.1238	0.0975	0.0814	0.0698
	proPCA	0.1899	0.1230	0.0969	0.0810	0.0695
2w-DFM	2w-PCA	0.6623	0.6710	0.6732	0.6744	0.6752
	Step-App	0.6370	0.6558	0.6567	0.6580	0.6594
	Q-MLE	**0.1169**	**0.0708**	**0.0528**	0.0488	**0.0368**
RaDFaM	sPCA	0.1246	0.0782	0.0579	**0.0456**	0.0369
	AE-1	0.3403	0.3403	0.3403	0.3404	0.3404
AE	AE-5	0.3403	0.3403	0.3403	0.3403	0.3403
	CAE	0.3383	0.3383	0.3383	0.3384	0.3384

methods, the performance of other methods improves as the number of factors increases; see the reconstructed results of sPCA for different factor numbers in Figure 7 for visualization; (ii) the relatively small sample size ($T = 714$) makes the performance of AE methods worse than the statistical methods; (iii) sPCA outperforms other PCA-type estimations; (iv) sPCA uses much less computational time than Q-MLE and has comparable


(a) Original

(b) $k_1 = k_2 = 5$

(c) $k_1 = k_2 = 15$

(d) $k_1 = k_2 = 25$

(e) $k_1 = k_2 = 35$

(f) $k_1 = k_2 = 45$

Figure 7. Reconstructed images for different factor numbers under RaDFaM.

Table 4. $\overline{\text{PSNR}}$s of different estimation methods for different choices of (k_1, k_2) applied to the COVID-CT data.

Model	Method/ (k_1, k_2)	$(5, 5)$	$(15, 15)$	$(25, 25)$	$(35, 35)$	$(45, 45)$
BiMFaM	autoPCA	14.1498	17.6611	19.4595	20.7566	21.9110
	α-PCA	14.6213	18.3889	20.5233	22.1957	23.6577
	proPCA	14.6597	18.4334	20.5756	22.2437	23.6945
2w-DFM	2w-PCA	3.9152	3.7891	3.7549	3.7362	3.7242
	Step-App	4.2524	4.0008	3.9877	3.9683	3.9478
	Q-MLE	**18.8588**	**23.3783**	**26.1569**	26.9265	29.7442
RaDFaM	sPCA	18.3016	22.4788	25.3164	**27.6773**	**29.8249**
	AE-1	9.5018	9.5010	9.5004	9.4994	9.4976
AE	AE-5	9.5023	9.5022	9.5019	9.5020	9.5020
	CAE	9.5534	9.5537	9.5518	9.5516	9.5499

Table 5. Computational time (s) of different estimation methods for different choices of (k_1, k_2) applied to the COVID-CT data.

Model	Method/ (k_1, k_2)	$(5, 5)$	$(15, 15)$	$(25, 25)$	$(35, 35)$	$(45, 45)$
BiMFaM	autoPCA	287.64	291.38	288.88	289.68	293.33
	α-PCA	3.33	3.27	3.23	3.47	3.66
	proPCA	0.47	0.80	1.27	2.73	2.35
2w-DFM	2w-PCA	3.88	4.39	4.82	4.83	5.50
	Step-App	6.42	8.67	10.68	15.14	18.45
	Q-MLE	435.47	3374.95	1247.70	1723.92	3894.55
RaDFaM	sPCA	4.99	4.75	6.17	5.88	9.01

performance; refer to the computational time (in seconds) in Table 5.

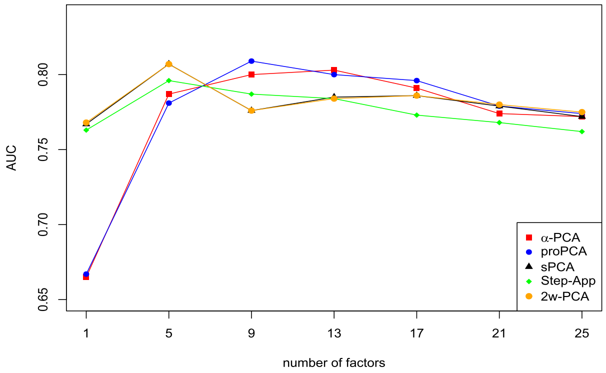
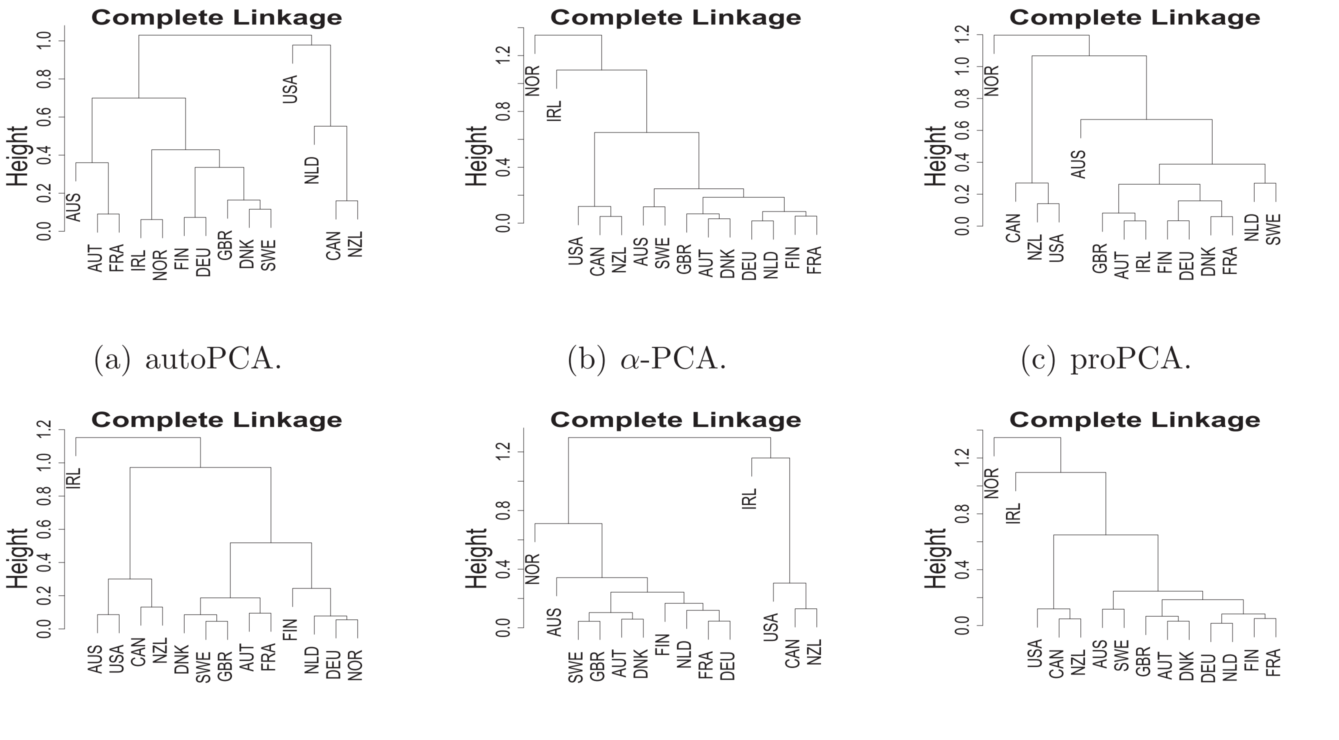


Figure 8. Average AUC values for different choice of (k_1, k_2) for COVID-CT data.

Table 6. RMSEs under different choices of (k_1, k_2) for the multinational macroeconomic indices data.

Model	Method /(k_1, k_2)	$(2, 2)$	$(3, 4)$	$(5, 6)$	$(8, 8)$
BiMFaM	autoPCA	0.3232	0.2736	0.2272	0.1743
	α-PCA	0.3075	0.2596	0.2142	0.1576
	proPCA	0.3103	0.2608	0.2068	0.1578
2w-DFM	2w-PCA	0.2886	0.3029	0.3149	0.3380
	Step-App	0.2440	0.2677	0.3009	0.3263
	Q-MLE	0.2443	–	0.2941	–
RaDFaM	sPCA	**0.1979**	**0.1480**	**0.0880**	**0.0429**

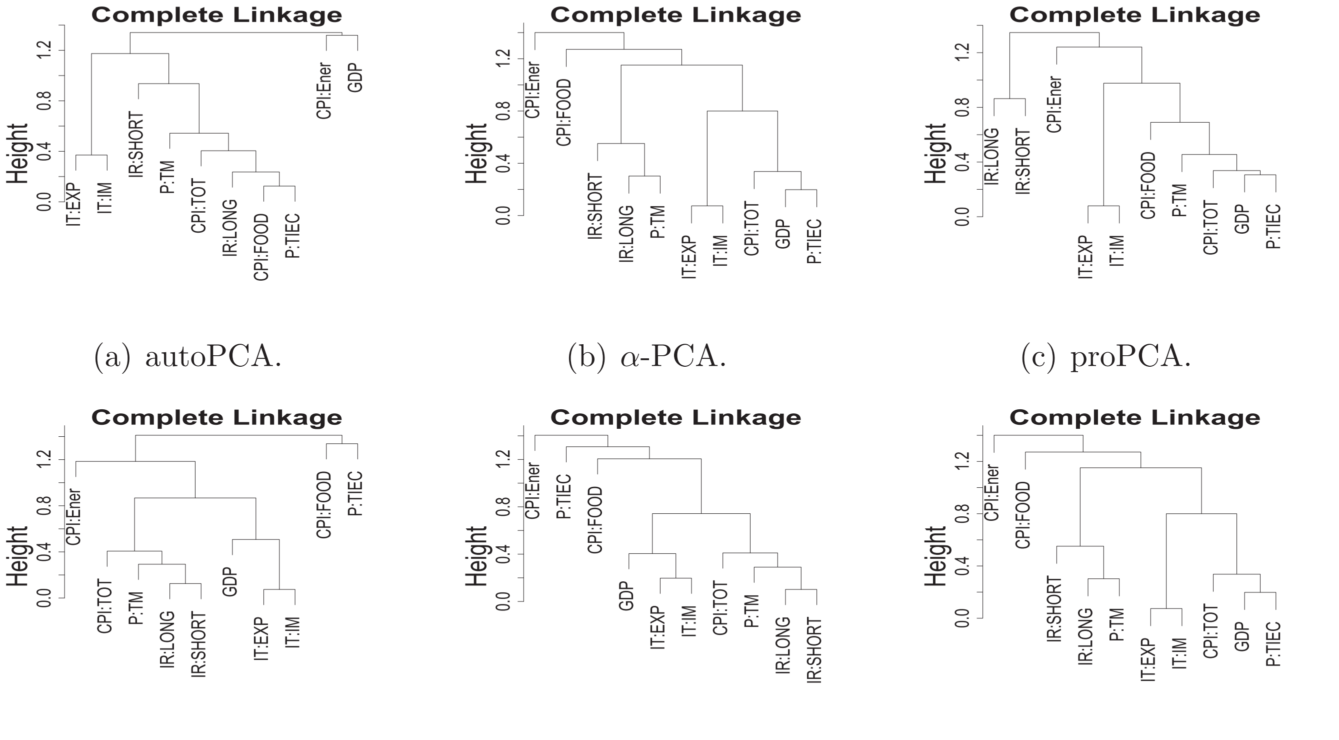
Image classification: RaDFaM versus BiMFaM versus 2w-DFM Once unsupervised learning on high-dimensional matrix-variate objects is built up under a specific latent structure, one may conduct supervised learning based on the extracted low-dimensional features. Take RaDFaM on the aforementioned CT scanning dataset $\{(Y_t, \mathbf{X}_t), t \in [714]\}$ for instance, where each CT scan $\mathbf{X}_t$ is a biomedical image biomarker to diagnose the disease label $Y_t \in \{0, 1\}$ with 0 being negative and 1 otherwise. Let $\langle \mathbf{A}, \mathbf{B} \rangle = \text{vec}^\top(\mathbf{A})\text{vec}(\mathbf{B})$ be the matrix inner product for the placeholder matrices $\mathbf{A}$ and $\mathbf{B}$. Then



(a) autoPCA. (b) α-PCA. (c) proPCA.

(d) Step-App. (e) Q-MLE. (f) sPCA.

Figure 9. Hierarchical clustering results from the estimated row-wise loading matrix for different methods of estimation applied to the multinational, macroeconomic indices data.



(a) autoPCA. (b) α-PCA. (c) proPCA.

(d) Step-App. (e) Q-MLE. (f) sPCA.

Figure 10. Hierarchical clustering results from the estimated column-wise loading matrix for different methods of estimation applied to the multinational, macroeconomic indices data.

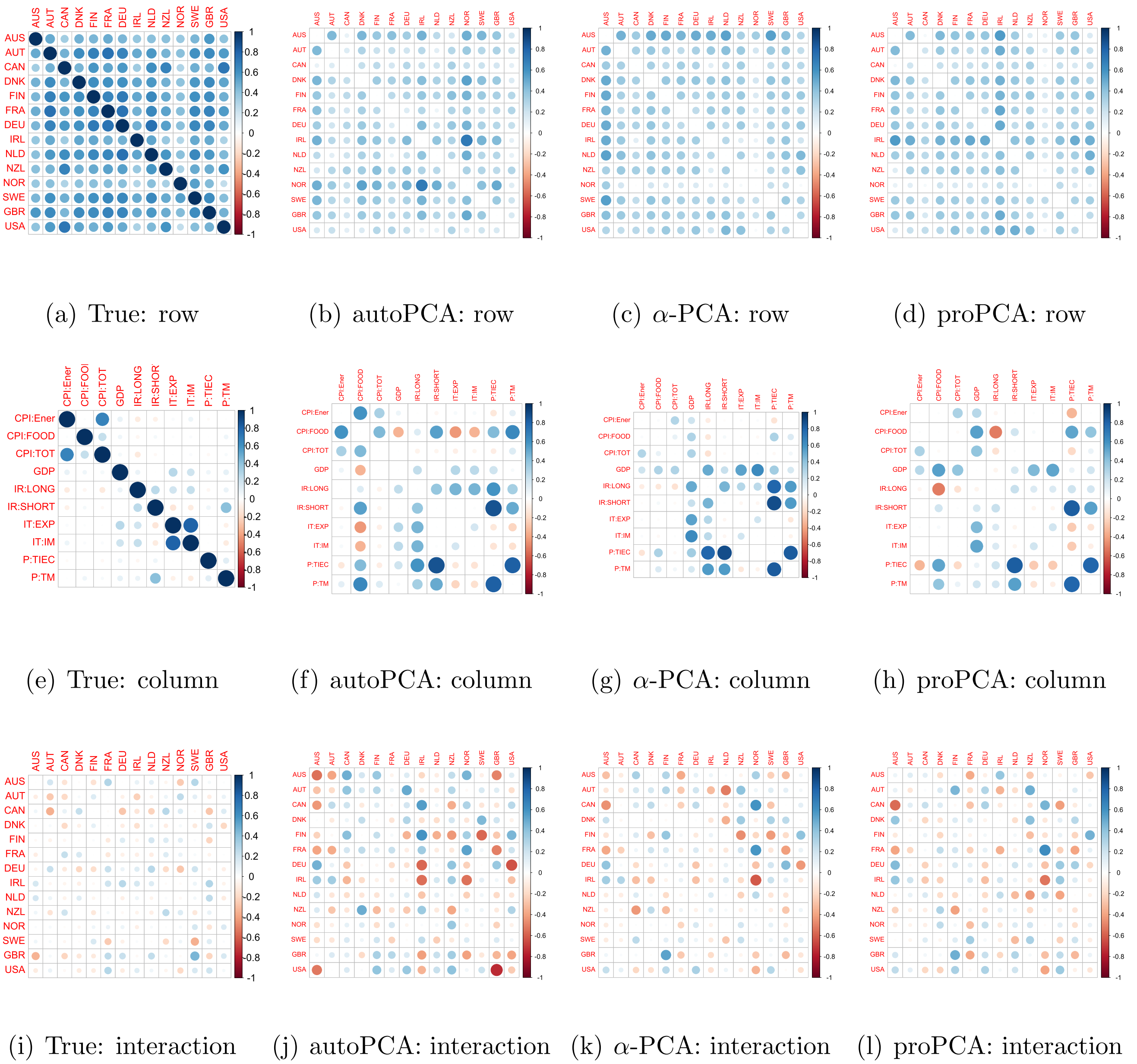
a two-step latent logistic regression model LLR_{RaD} can be fitted,

$$\begin{cases} \operatorname{logit}P(Y_t = 1|\mathbf{Z}_t, \mathbf{F}_t, \mathbf{E}_t) = \gamma + \langle \mathbf{B}_1, \mathbf{Z}_t\rangle + \langle \mathbf{B}_2, \mathbf{F}_t\rangle + \langle \mathbf{B}_3, \mathbf{E}_t\rangle, \\ \mathbf{X}_t = \mathbf{R}\mathbf{Z}_t\mathbf{C}^\top + \mathbf{R}\mathbf{E}_t^\top + \mathbf{F}_t\mathbf{C}^\top + \mathbf{e}_t, \end{cases}$$

where γ is the intercept, and $\mathbf{B}_1 \in \mathbb{R}^{k_1\times k_2}$, $\mathbf{B}_2 \in \mathbb{R}^{p_1\times k_2}$, and $\mathbf{B}_3 \in \mathbb{R}^{p_2\times k_1}$ are the regression coefficients for the latent-factor regressors $\mathbf{Z}_t$, $\mathbf{F}_t$, and $\mathbf{E}_t$, respectively. Analogous latent models on the same CT scan dataset are denoted as LLR_{BiM} and LLR_{2w} with the latent factor regressor $\mathbf{Z}_t$ and latent factor regressors $\mathbf{E}_t$, $\mathbf{F}_t$, based on the unsupervised learning of the matrix factor models BiMFaM and 2w-DFM, respectively.

We randomly split samples into a training set (80%) and a testing set (20%). In the training or fitting process, we estimate loadings and latent factors by sPCA (LLR_{RaD}), α-PCA and proPCA (LLR_{BiM}), and 2w-PCA and Step-App (LLR_{2w}), respectively, followed by estimating the matrix regression coefficients (considering the computational time cost, autoPCA and Q-MLE are not used). In the testing or prediction process, we obtain the new regressors by the factor score equations (such as (12), (14), and (15)) based on the estimated loadings, followed by predicting labels. The procedure is randomly repeated for 100 times.

The factor numbers k_1 and k_2 are set to be equal, and take values of 1, 5, 9, 13, 17, 21, and 25, respectively. Note that LASSO type thresholding is applied to sPCA, 2w-PCA, and Step-App due to the high dimensionality of vec($\mathbf{F}_t$) and vec($\mathbf{E}_t$), and it is also used for α-PCA and proPCA when the factor numbers exceed 5 due to the increasing dimensionality of vec($\mathbf{Z}_t$) with the growing factor numbers.



(a) True: row (b) autoPCA: row (c) α-PCA: row (d) proPCA: row

(e) True: column (f) autoPCA: column (g) α-PCA: column (h) proPCA: column

(i) True: interaction (j) autoPCA: interaction (k) α-PCA: interaction (l) proPCA: interaction

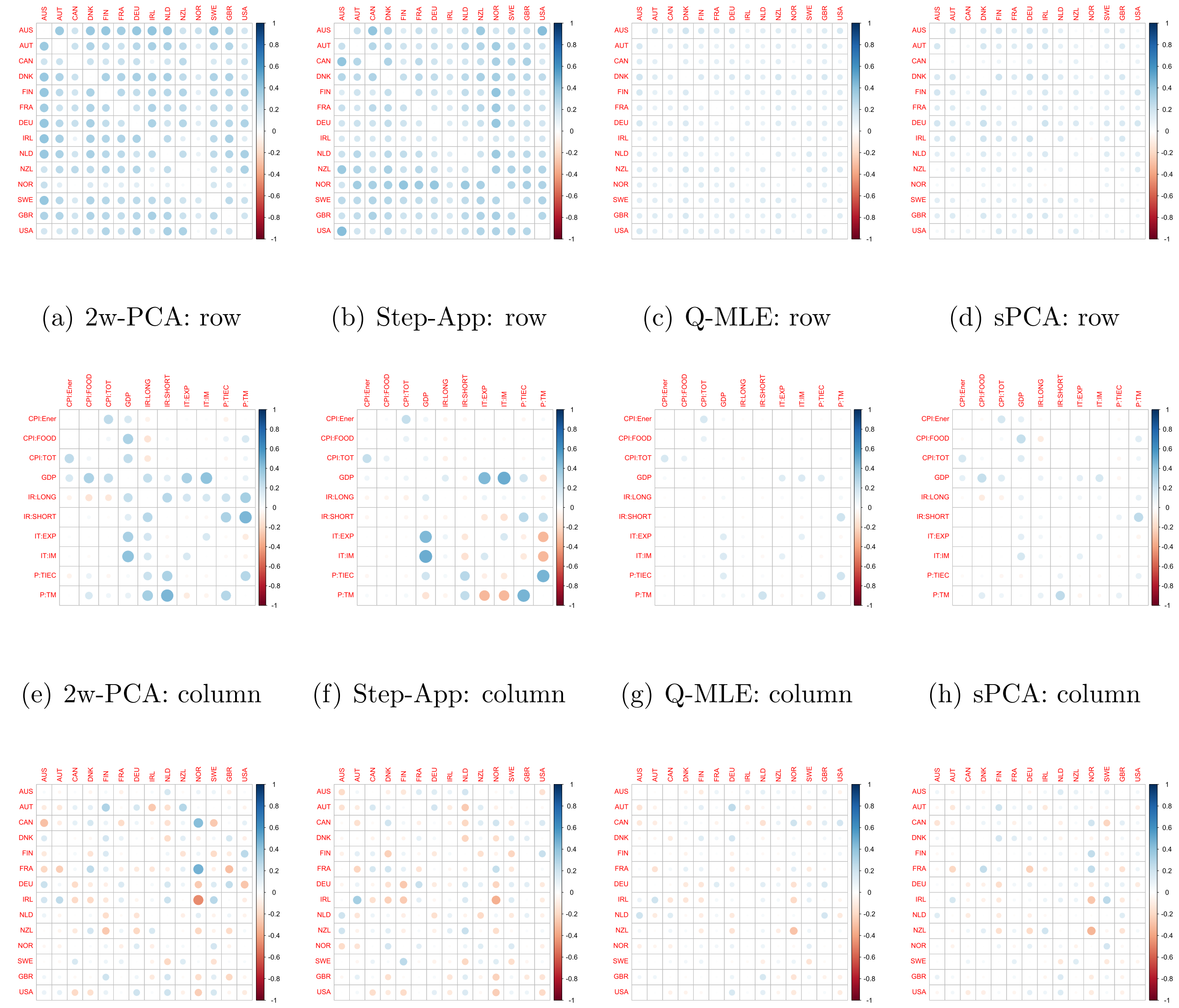
Figure 11. Observed correlations (the first column) and differences between estimated and observed correlations for the multinational, macroeconomic indices data under BiMFaM (the second to the fourth columns).

Figure 8 is a scatterplot of the average AUC values against different factor numbers. We may observe that, (i) the performance of sPCA (black) and 2w-PCA (orange) are almost identical. This may be due to their identical estimations of $\mathbf{E}_t$ and $\mathbf{F}_t$ in equations (12) and (15), and the shrinkage function of LASSO that eliminates the interactive effect $\mathbf{Z}_t$ automatically; (ii) the number of factors may affect the prediction accuracy of the latent logistic regression models, particular for the supervised learning derived based on BiMFaM.

5.2. Correlated Data: Multinational Macroeconomic Indices

We analyze the quarterly multinational macroeconomic indices data collected from the Organization of Economic Cooperation and Development (OECD) website containing 10 quarterly macroeconomic indices from 14 countries for 74 quarters from 2000.Q1 to 2018.Q2. Samples of each quarter can be treated as a matrix-variate with rows, columns, and each element representing different countries, macroeconomic indices, and the value of the index of the corresponding country, respectively.

The 10 indices consist of four groups: the production group (P:TIEC, P:TM, GDP), the consumer price group (CPI:FOOD, CPI:ENER, CPI:TOT), the money market group (IR:Long, IR:SHORT), and the international trade group (IT:EXP, IT:IM). The countries and corresponding abbreviations are given in Appendix I of the online supplement. After transformations of each univariate time series along the index mode as in Chen and Fan (2023) (see Table 10 in their supplement), we center the matrix-variate series to eliminate mean effects, and obtain observations with $T = 72$, $p_1 = 14$ and $p_2 = 10$.



(a) 2w-PCA: row (b) Step-App: row (c) Q-MLE: row (d) sPCA: row

(e) 2w-PCA: column (f) Step-App: column (g) Q-MLE: column (h) sPCA: column

(i) 2w-PCA: interaction (j) Step-App:interaction (k) Q-MLE: interaction (l) sPCA: interaction

Figure 12. Differences between estimated and observed correlations for the multinational, macroeconomic indices data under 2w-DFM (the first to the third columns) and RaDFaM (the fourth column).

Clustering We choose factor numbers $k_1 = 3$ and $k_2 = 4$. For interpretability, the varimax rotation is applied to loading matrix estimates, and then the rotated loadings are multiplied by 30 and truncated to integer values. The hierarchical clustering results based on the rotated row-wise and column-wise loading matrix estimates are shown in Figures 9 and 10 (2w-PCA is omitted to save space), respectively. For row-wise loadings (countries), the results of α-PCA, 2w-PCA, and sPCA are identical, but are different from those of autoPCA, proPCA, Step-App, and Q-MLE. All methods cluster the United States of America (USA), Canada (CAN) and New Zealand (NZL) into one group. For column-wise loadings (indices), proPCA gives the closest clustering to the true index groups, followed closely by α-PCA, 2w-PCA, and sPCA. The tables of final loading matrix estimates are given in Appendix I of the online supplement.

Reconstruction comparison in terms of correlation effects To compare the reconstruction performance of BiMFaM, 2w-DFM, and RaDFaM, we calculate the difference between the row-wise, column-wise, and interactive correlation from the estimated signal part and the observations. The results in Figures 11–12 show that the recovered row-wise, column-wise, and interactive correlation for sPCA and Q-MLE are closer (much lighter color) to the observed ones (the first column in Figure 11). This result provides evidence that mode-wise latent factors substantially influence on the reconstruction performance.

Reconstruction comparison in terms of the RMSE We use 10-fold cross-validation to evaluate the difference between the observed and estimated matrices in terms of the RMSE for different values of factor numbers. The results in Table 6 demonstrate that sPCA performs the best; Q-MLE fails for the second and the fourth factor numbers, which may be attributed to the small sample size and matrix dimensions (bold values are the best ones).

6. Discussion and Future Work

The Bayesian networks of the three models provided in Appendix A of the online supplement show visually their modeling strategies. Another insight may be that using the block diagonal approach under BiMFaM, which may result in a special form of RaDFaM (He et al. 2023).

We end this article with a suggestion for future work. Since, as we illustrated in Section 1.2, BiMFaM has a signal part in the form of the second-order Tucker decomposition, it can be extended naturally to the higher-order Tucker tensor factor model (Chen, Yang, and Zhang 2022; Zhang et al. 2022). Similarly, we can rewrite expression (1) of RaDFaM as

$$\sum_{i=1}^{l} \mathbf{U}_i \mathbf{V}_i^\top = \sum_{i=1}^{l} \mathbf{U}_i \circ \mathbf{V}_i,$$

which is a special second-order CANDECOMP/PARAFAC (CP) decomposition (Kolda and Bader 2009, sec. 3). Due to the strong signal strength of RaDFaM, we may extend the hierarchical spirit to model a new tensor decomposition, and hence an induced tensor factor model. Specifically, for any Dth-order tensor $\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_D}$, we postulate the model

$$\begin{cases} \mathcal{X} = \sum_{i=1}^{l} \mathbf{U}_{1,i} \circ \cdots \circ \mathbf{U}_{D,i}, \\ \mathbf{U}_{d,i} = \mathbf{R}_d \mathbf{A}_{d,i\cdot} + \boldsymbol{\xi}_{d,i}, \ d \in [D]. \end{cases}$$

One may consider developing PCA-type estimators or building EM-variant algorithms to find the maximum likelihood estimators.

Supplementary Materials

The illustration of the Bayesian networks of the three models is provided in Appendix A. The proofs of Propositions 1 and 2 are provided in Appendices B and C, respectively. The proofs of Theorems 1–4 with the necessary lemmas are provided in Appendices D–G, respectively. Additional simulation results and real data analysis are provided in Appendices H–I, respectively.

Acknowledgments

The authors are grateful to the Co-editor, the Associate Editor, and the two anonymous referees for their constructive and helpful comments and suggestions. The authors thank Dr. Zhen Pang for the discussion and coding in the early stage of this work.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Xu Zhang's research was partially supported by the National Natural Science Foundation of China (grant no.12301338) and the Postdoc Fellowship of CAS AMSS-PolyU Joint Laboratory of Applied Mathematics. Catherine C. Liu's research was partially supported by General Research Fund 15301123, RGC, HKSAR, and PolyU Research Grant P0045497. Jianhua Guo's research was partially supported by the National Key Research and Development Program of China (grant no.2020YFA0714100). K. C. Yuen's research was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU17306220). A. H. Welsh's research was partially supported by the Australian Research Council Discovery Project DP230101908.

References

Ahmadi, S., and Rezghi, M. (2020), "Generalized Low-Rank Approximation of Matrices based on Multiple Transformation Pairs," *Pattern Recognition*, 108, 107545. [3]

Athreya, K. B., and Lahiri, S. N. (2006), *Measure Theory and Probability Theory* (1st ed.), New York: Springer. [5,6]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [6]

Ben-Israel, A., and Greville, T. N. (2003), *Generalized Inverses: Theory and Applications* (2nd ed.), Springer. [4]

Chen, E. Y., and Fan, J. (2023), "Statistical Inference for High-Dimensional Matrix-Variate Factor Models," *Journal of the American Statistical Association*, 118, 1038–1055. [2,3,4,7,14]

Chen, R., Yang, D., and Zhang, C.-H. (2022), "Factor Models for High-Dimensional Tensor Time Series," *Journal of the American Statistical Association*, 117, 94–116. [15]

Chen, Z., Fan, J., and Wang, D. (2020), "High-Dimensional Factor Model and its Applications to Statistical Machine Learning," *Science China Mathematics*, 50, 447–490. [5]

Davis, C., and Kahan, W. M. (1970), "The Rotation of Eigenvectors by a Perturbation. III," *SIAM Journal on Numerical Analysis*, 7, 1–46. [3]

Dawid, A. P. (1981), "Some Matrix-Variate Distribution Theory: Notational Considerations and a Bayesian Application," *Biometrika*, 68, 265–274. [2]

Fan, J., Li, K., and Liao, Y. (2021), "Recent Developments in Factor Models and Applications in Econometric Learning," *Annual Review of Financial Economics*, 13, 401–430. [2]

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical Foundations of Data Science* (1st ed.), Boca Raton, FL: Chapman and Hall/CRC. [1,3]

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [5]

Fan, J., Ma, C., and Zhong, Y. (2021), "A Selective Overview of Deep Learning," *Statistical Science*, 36, 264–290. [9]

Fosdick, B. K., and Hoff, P. D. (2014), "Separable Factor Analysis with Applications to Mortality Data," *The Annals of Applied Statistics*, 8, 120–147. [2]

Francq, C., and Zakoian, J.-M. (2019), *GARCH Models: Structure, Statistical Inference and Financial Applications* (2nd ed.), Hoboken, NJ: Wiley. [5,6]

Guo, X., Liu, X., Zhu, E., and Yin, J. (2017), "Deep Clustering with Convolutional Autoencoders," in *Neural Information Processing*, pp. 373–382, Cham: Springer. [9]

Gupta, A. K., and Nagar, D. K. (2018), *Matrix Variate Distributions* (1st ed.), Boca Raton, FL: CRC Press. [1,2]

He, Y., Kong, X., Trapani, L., and Yu, L. (2023), "One-Way or Two-Way Factor Model for Matrix Sequences?" *Journal of Econometrics*, 235, 1981–2004. [15]

Hinton, G. E., and Salakhutdinov, R. R. (2006), "Reducing the Dimensionality of Data with Neural Networks," *Science*, 313, 504–507. [9]

Hoff, P. D. (2011), "Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data," *Bayesian Analysis*, 6, 179–196. [2]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [15]

Lam, C., and Yao, Q. (2011), "Estimation of Latent Factors for High-Dimensional Time Series," *Biometrika*, 98, 901–918. [3,5]

——— (2012), "Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors," *The Annals of Statistics*, 40, 694–726. [5]

Liu, J., Chen, S., Zhou, Z.-H., and Tan, X. (2010), "Generalized Low-Rank Approximations of Matrices Revisited," *IEEE Transactions on Neural Networks*, 21, 621–632. [3]

Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. (2014), *Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data* (1st ed.), New York: Chapman and Hall/CRC. [2]

Nicholson, W. K. (2020), *Linear Algebra with Applications*, Calgary: Lyryx Learning Inc. [2]

Salomon, D. (2004), *Data Compression: the Complete Reference* (4th ed.), London: Springer. [4]

Shores, T. S. (2018), *Applied Linear Algebra and Matrix Analysis* (2nd ed.), Cham: Springer. [2]

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [3,5]

Suetens, P. (2017), *Fundamentals of Medical Imaging* (3rd ed.), Cambridge: Cambridge University Press. [4]

Tang, R., Yuan, M., and Zhang, A. R. (2023), "Mode-Wise Principal Subspace Pursuit and Matrix Spiked Covariance Model," arXiv preprint arXiv:2307.00575. [3]

Wang, D., Liu, X., and Chen, R. (2019), "Factor Models for Matrix-Valued High-Dimensional Time Series," *Journal of Econometrics*, 208, 231–248. [2,3,4,7]

Wang, W., Aggarwal, V., and Aeron, S. (2019), "Principal Component Analysis with Tensor Train Subspace," *Pattern Recognition Letters*, 122, 86–91. [2]

Yang, J., Zhang, D., Frangi, A. F., and Yang, J.-y. (2004), "Two-Dimensional PCA: A New Approach to Appearance-based Face Representation and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 131–137. [3]

Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., and Xie, P. (2020), "COVID-CT-dataset: A CT Image Dataset about COVID-19," arXiv preprint arXiv:2003.13865. [8]

Ye, J. (2005), "Generalized Low Rank Approximations of Matrices," *Machine Learning*, 61, 167–191. [3]

Yu, L., He, Y., Kong, X., and Zhang, X. (2022), "Projected Estimation for Large-Dimensional Matrix Factor Models," *Journal of Econometrics*, 229, 201–217. [2,3,4,5,6,7]

Yuan, C., Gao, Z., He, X., Huang, W., and Guo, J. (2023), "Two-Way Dynamic Factor Models for High-Dimensional Matrix-Valued Time Series," *Journal of the Royal Statistical Society*, Series B, 85, 1517–1537. [2,3,4,7]

Zare, A., Ozdemir, A., Iwen, M. A., and Aviyente, S. (2018), "Extension of PCA to Higher Order Data Structures: An Introduction to Tensors, Tensor Decompositions, and Tensor PCA," *Proceedings of the IEEE*, 106, 1341–1358. [2]

Zhang, X. (2017), *Matrix Analysis and Applications* (1st ed.), Cambridge: Cambridge University Press. [1]

Zhang, X., Li, G., Liu, C. C., and Guo, J. (2022), "Tucker Tensor Factor Models: Matricization and Mode-Wise PCA Estimation," arXiv preprint arXiv:2206.02508v2. [2,15]

Zhang, Y., Zhang, X., Zhang, H., Liu, A., and Liu, C. C. (2023), "Low-Rank Latent Matrix-Factor Prediction Modeling for Generalized High-Dimensional Matrix-Variate Regression," *Statistics in Medicine*, 42, 3616–3635. [8]

Zhou, P., Lu, C., Feng, J., Lin, Z., and Yan, S. (2021), "Tensor Low-Rank Representation for Data Recovery and Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 1718–1732. [2]