A Two-Way Transformed Factor Model for Matrix-Variate Time Series

Article *in* Econometrics and Statistics · August 2021

DOI: 10.1016/j.ecosta.2021.08.008

Zhaoxing Gao
The London School of Economics and Political Science

ELSEVIER
Econometrics
and
Statistics
The official journal of
Computational and Financial Econometrics
Computational and Methodological Statistics
http://CFEnetwork.org
http://CMStatistics.org
http://www.elsevier.com/locate/ecosta

A Two-Way Transformed Factor Model for Matrix-Variate Time Series

Zhaoxing Gao, Ruey S. Tsay

PII: S2452-3062(21)00102-7
DOI: https://doi.org/10.1016/j.ecosta.2021.08.008
Reference: ECOSTA 265

To appear in: *Econometrics and Statistics*

Received date: 15 October 2020
Revised date: 17 August 2021
Accepted date: 17 August 2021

Please cite this article as: Zhaoxing Gao, Ruey S. Tsay, A Two-Way Transformed Factor Model for Matrix-Variate Time Series, *Econometrics and Statistics* (2021), doi: https://doi.org/10.1016/j.ecosta.2021.08.008

A Two-Way Transformed Factor Model for Matrix-Variate Time Series

Zhaoxing Gao[a], Ruey S. Tsay*[b]

[a] *Center for Data Science, Zhejiang University, Hangzhou, China*

[b] *Booth School of Business, University of Chicago, Chicago, USA*

Abstract

A new framework is proposed for modeling high-dimensional matrix-variate time series via a two-way transformation, where the transformed data consist of a matrix-variate factor process, which is dynamically dependent, and three other blocks of white noises. For a given $p_1 \times p_2$ matrix-variate time series, nonsingular transformations are sought to project the rows and columns onto another p_1 and p_2 directions according to the strength of the dynamical dependence of the series on their past values. Consequently, the data are nonsingular linear row and column transformations of dynamically dependent common factors and white noise idiosyncratic components. A common orthonormal projection method is proposed to estimate the front and back loading matrices of the matrix-variate factors. Under the setting that the largest eigenvalues of the covariance of the vectorized idiosyncratic term diverge for large p_1 and p_2, a two-way projected Principal Component Analysis is introduced to estimate the associated loading matrices of the idiosyncratic terms to mitigate such diverging noise effects. A new white-noise testing procedure is proposed to estimate the dimension of the factor matrix. Asymptotic properties of the proposed method are established for both fixed and diverging dimensions as the sample size increases to infinity. Simulated and real examples are used to assess the performance of the proposed method. Comparisons of the proposed method with some existing ones in the literature concerning the forecastability of the factors are studied and it is found that the proposed approach not only provides interpretable results, but also performs well in out-of-sample forecasting.

Keywords: Common factor, Eigen-analysis, Projected principal component analysis, Kronecker product, Diverging eigenvalues, High-dimensional white noise test.

1. Introduction

Modern scientific studies often collect data under combinations of multiple attributes. For example, neuroimaging experiments record brain activities at multiple spatial locations and various orientations

*Corresponding author.
E-mail addresses: zhaoxing_gao@zju.edu.cn (Z. Gao), ruey.tsay@chicagobooth.edu (R.S. Tsay). Supplementary materials for this article are available online.

under a variety of experimental stimuli. Studies of social networks record social links for a variety of settings from multiple initiators of social activity to multiple receivers of the activity. These data are naturally represented not as a list or table of numbers, but as a multi-indexed array, or tensor. Furthermore, such data are often collected repeatedly over time, and it is then natural to view them as tensor-valued time series. The matrix-variate time series is a sequence of second-order random tensors. For example, financial and economic studies often collect data on a number of economic indicators (e.g., growth rate of the gross domestic product, unemployment rate, etc.) from multiple countries every quarter. Therefore, it is important and interesting to develop appropriate statistical methods to analyze such matrix-variate time series. The most commonly used approach to modeling such data is to stack the matrix into a long vector and to apply the standard multivariate statistical methods. However, such an approach ignores the matrix structure of the data and often overlooks some important patterns embedded in the data. Werner et al. (2008) pointed out that after vectorizing matrices the resulting vectors have a Kronecker-product structure, and ignoring this structure means that a much larger number of parameters need to be estimated. Moreover, the dimension of a matrix-variate time series itself can become large in the current era of big data. Therefore, it is important to make use of the matrix structure and to find an effective way to reduce the number of parameters, especially when the dimension is high. There are some works on tensor time series, e.g., Rogers et al. (2013) and Surana et al. (2016), but these articles focus on data processing rather than on statistical properties or the high dimensional case.

In modeling vector time series, the available methods to reduce the number of parameters can be classified into two categories: regularization and dimension reduction. The former imposes some conditions on the structure of a vector autoregressive moving-average (VARMA) model, and the latter assumes a lower dimensional representation for the high-dimensional process. For the regularization methods, some special structures are often imposed on the VARMA model. For example, Chapter 4 of Tsay (2014) and the references therein discussed two different canonical structures. Davis et al. (2016) studied the VAR model with sparse coefficient matrices based on partial spectral coherence. The Lasso regularization has also been applied to VAR models, see Shojaie and Michailidis (2010), Song and Bickel (2011), and Han and Tsay (2020), among others. For dimension reduction, popular methods include the canonical correlation analysis (CCA) of Box and Tiao (1977), the principal component analysis (PCA) of Stock and Watson (2002), and the scalar component model of Tiao and Tsay (1989). The factor model approach can be found in Bai and Ng (2002), Stock and Watson (2005), Forni et al. (2000, 2005), Pan and Yao (2008), Lam et al. (2011), Lam and Yao (2012), Gao and Tsay (2019, 2021a,b,c), among others. However, none of the methods mentioned above can directly be used to model matrix-variate time series without vectorization. The matrix-variate time series are less well studied in the literature; Walden and Serroukh (2002) handled this type of data in signal and image processing, Wang et al. (2019) proposed a factor model for matrix-variate time series, which maintains and utilizes the matrix structure to achieve dimension reduction, and it was later extended to the tensor case in Han et al. (2020). With domain or prior knowledge of the series under study, Chen et al. (2020) studied the constrained matrix-variate factor models by imposing linear constraints on the loading matrices. However, the mechanism of the proposed matrix factor model deserves a further study and the bounded eigenvalue assumption of the covariance matrix of the vectorized idiosyncratic term is often violated in the high-dimensional setting, especially for the notable case of low signal-to-noise ratio commonly seen in finance and economics. See, for example, Black (1986). Recently, Wang et al. (2020) and Wang et al. (2021) considered tensor-variate time series modeling that focuses on autoregressive models.

The goal of this paper is to study the common dynamical dependence of matrix-variate time series from a new perspective. We first illustrate our basic idea below and propose our approach in Section 2. Let $\mathbf{Y}_t \in \mathbb{R}^{p_1 \times p_2}$ be an observable matrix-variate time series, which is weakly stationary with $E(\mathbf{Y}_t) = \mathbf{0}$. We postulate that there exist two full-rank matrices $\mathbf{T}_1 \in R^{p_1 \times p_1}$ and $\mathbf{T}_2 \in R^{p_2 \times p_2}$ such that $\mathbf{T}_1\mathbf{Y}_t\mathbf{T}_2'$ assumes the form

$$\mathbf{T}_1\mathbf{Y}_t\mathbf{T}_2' = \begin{bmatrix} \mathbf{F}_t & \mathbf{Z}_{12,t} \\ \mathbf{Z}_{21,t} & \mathbf{Z}_{22,t} \end{bmatrix}, \tag{1.1}$$

where $\mathbf{F}_t \in \mathbb{R}^{r_1 \times r_2}$ is a matrix-variate factor that captures the dynamical dependence of $\mathbf{Y}_t$, and $\mathbf{Z}_{12,t}$, $\mathbf{Z}_{21,t}$ and $\mathbf{Z}_{22,t}$ are matrix-variate idiosyncratic components, which are white noise processes. Equivalently, model (1.1) is to seek two nonsingular transformation matrices $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2) := \mathbf{T}_1^{-1}$ and $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2) := \mathbf{T}_2^{-1}$ with $\mathbf{L}_1 \in \mathbb{R}^{p_1 \times r_1}$ and $\mathbf{R}_1 \in \mathbb{R}^{p_2 \times r_2}$ such that $\mathbf{L}_1$ and $\mathbf{R}_1$ are the front and back loading matrices associated with the common factors. Although the factor process $\mathbf{F}_t$ is assumed to be dynamically dependent, model (1.1) is fundamentally different from that of Wang et al. (2019) in the sense that it maintains the number of innovation processes on both sides of (1.1) to be p_1p_2, if we further assume that $\mathbf{F}_t$ has an $r_1 \times r_2$ white noise innovation matrix. On the other hand, the data matrix $\mathbf{Y}_t$ in Wang et al. (2019) is generated by a latent factor process $\mathbf{F}_t$ with front and back loading matrices and a $p_1 \times p_2$ white noise innovation matrix, leading to the possibility of involving $r_1r_2 + p_1p_2$ innovations. Consequently, the covariance matrix of their estimated white noise process is singular in applications.

To see the rationale of model (1.1), let vec$(\cdot)$ be the conventional vectorization operator that converts a matrix to a vector by stacking columns of the matrix on top of each other. By the basic properties of Kronecker product, we rewrite the model in the following vector form:

$$\mathbf{y}_t := \text{vec}(\mathbf{Y}_t) = \mathbf{A}\begin{bmatrix} \mathbf{f}_t \\ \mathbf{z}_t \end{bmatrix}, \tag{1.2}$$

where $\mathbf{A} = [\mathbf{R}_1 \otimes \mathbf{L}_1, \mathbf{R}_1 \otimes \mathbf{L}_2, \mathbf{R}_2 \otimes \mathbf{L}_1, \mathbf{R}_2 \otimes \mathbf{L}_2] \in \mathbb{R}^{p_1p_2 \times p_1p_2}$, $\mathbf{f}_t = \text{vec}(\mathbf{F}_t) \in \mathbb{R}^{r_1r_2}$ and $\mathbf{z}_t = [\text{vec}(\mathbf{Z}_{21,t})', \text{vec}(\mathbf{Z}_{12,t})', \text{vec}(\mathbf{Z}_{22,t})']' \in \mathbb{R}^{p_1p_2 - r_1r_2}$. For identifiability, we assume that both $\mathbf{f}_t$ and $\mathbf{z}_t$ have zero mean and identity covariance matrices. This is a special case of the model considered in Gao and Tsay (2021b) for vector time series by assuming that the covariance of the vectorized data has a Kronecker structure. That is, we expect there exists a transformation matrix $\mathbf{A}^{-1}$ with a Kronecker structure such that $\mathbf{A}^{-1}\mathbf{y}_t = (\mathbf{f}_t', \mathbf{z}_t')'$, and this can be done via canonical correlation analysis between $\mathbf{y}_t$ and its lagged variables, and the components of the transformed vector $(\mathbf{f}_t', \mathbf{z}_t')'$ are contemporaneously uncorrelated with an identity covariance matrix. See the discussions in Gao and Tsay (2021b) and Tiao and Tsay (1989). The structure of $\mathbf{A}$ is different from that in Gao and Tsay (2021b) in order to preserve the structure of the matrix-valued data. Consequently, the main task of the proposed method is to estimate $\mathbf{L}_1$ and $\mathbf{R}_1$ and to recover the matrix factor $\mathbf{F}_t$.

To summarize, our goal is to propose a new statistical framework for modeling matrix-variate time series based on two-way linear transformations and the concept of factor models. We re-parametrize the model by leveraging the strengths of linear transformation matrices to separate the factors from the idiosyncratic components, and the resulting front and back loading matrices associated with the common factor and the idiosyncratic terms are all semi-orthogonal. To this end, our first step is to find common orthonormal projections for the row and column vectors, respectively, based on an eigen-analysis of certain matrices, and the top few projected coordinates form a matrix-variate common factor process. The rest of the projected coordinates form a matrix-variate white noise series. To recover the factor matrix, we introduce a two-way projected principal component analysis (PCA) to estimate the loading matrices associated with the idiosyncratic matrix; see Section 2 for details. In the presence of diverging noise components, the projected PCA can mitigate the effect of the idiosyncratic component in estimating the common factor matrix. Furthermore, we propose a diagonal-path selection method to estimate the order (or dimension) of the factor matrix based on a white noise testing procedure. This testing procedure is more powerful and provides more statistically interpretable results than the ratio-based method in Wang et al. (2019), which, in turn, follows that of Lam et al. (2011). Consequently, the extracted matrix-variate factors capture most of the dynamical dependence of the data and is useful if one is interested in out-of-sample forecasting of matrix-variate time series. An autoregressive type of model can then be used to model the low-dimensional common factor process. See, for example, the model in Chen et al. (2020). Asymptotic properties of the proposed method are established for both fixed and diverging dimensions as the sample size n goes to infinity. We use simulated and real examples to assess the performance of the proposed method in finite samples.

The rest of the paper is organized as follows. We introduce the proposed model and estimation methodology in Section 2. In Section 3, we study the theoretical properties of the proposed model and its associated estimates. Numerical illustrations with both simulated and real data sets are reported in Section 4. Section 5 provides some concluding remarks. All technical proofs are given in an online supplement. Throughout the article, we use the following notation. For a $p \times 1$ vector $\mathbf{u} = (u_1, ..., u_p)'$, $||\mathbf{u}||_2 = \|\mathbf{u}'\|_2 = (\sum_{i=1}^p u_i^2)^{1/2}$ is the Euclidean norm, and $\mathbf{I}_p$ denotes a $p \times p$ identity matrix. For a matrix $\mathbf{H} = (h_{ij})$, $|\mathbf{H}|_\infty = \max_{i,j} |h_{ij}|$, $\|\mathbf{H}\|_F = \sqrt{\sum_{i,j} h_{ij}^2}$ is the Frobenius norm, $\|\mathbf{H}\|_2 = \sqrt{\lambda_{\max}(\mathbf{H}'\mathbf{H})}$ is the operator norm, where $\lambda_{\max}(\cdot)$ denotes for the largest eigenvalue of a matrix, and $\|\mathbf{H}\|_{\min}$ is the square root of the minimum non-zero eigenvalue of $\mathbf{H}'\mathbf{H}$. The superscript

$'$ denotes the transpose of a vector or matrix. We also use the notation $a \asymp b$ to denote $a = O(b)$ and $b = O(a)$.

2. Models and Methodology

2.1. Setting

Let $\mathbf{Y}_t = [y_{ij,t}] = (\mathbf{y}_{1,t}, ..., \mathbf{y}_{p_2,t})$ be an observable $p_1 \times p_2$ matrix-variate time series with $\mathbf{y}_{j,t} = (y_{1j,t}, ..., y_{p_1j,t})' \in \mathbb{R}^{p_1}$ and $E(\mathbf{y}_{j,t}) = \mathbf{0}$, for $1 \le j \le p_2$. We assume $\mathbf{Y}_t$ admits a latent structure:

$$\mathbf{Y}_t = \mathbf{L} \begin{bmatrix} \mathbf{F}_t & \mathbf{Z}_{12,t} \\ \mathbf{Z}_{21,t} & \mathbf{Z}_{22,t} \end{bmatrix} \mathbf{R}' = \mathbf{L}_1\mathbf{F}_t\mathbf{R}_1' + \mathbf{L}_2\mathbf{Z}_{21,t}\mathbf{R}_1' + \mathbf{L}_1\mathbf{Z}_{12,t}\mathbf{R}_2' + \mathbf{L}_2\mathbf{Z}_{22,t}\mathbf{R}_2', \tag{2.1}$$

where $\mathbf{F}_t \in \mathbb{R}^{r_1 \times r_2}$ is a matrix-variate common factor process, $\mathbf{Z}_{12,t} \in \mathbb{R}^{r_1 \times v_2}$, $\mathbf{Z}_{21,t} \in \mathbb{R}^{v_1 \times r_2}$, and $\mathbf{Z}_{22,t} \in \mathbb{R}^{v_1 \times v_2}$ are matrix-variate idiosyncratic noise processes with $r_1 + v_1 = p_1$ and $r_2 + v_2 = p_2$, $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2) \in \mathbb{R}^{p_1 \times p_1}$ is the front loading matrix with $\mathbf{L}_1 \in \mathbb{R}^{p_1 \times r_1}$ and $\mathbf{L}_2 \in \mathbb{R}^{p_1 \times v_1}$, and $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2) \in \mathbb{R}^{p_2 \times p_2}$ is the back loading matrix with $\mathbf{R}_1 \in \mathbb{R}^{p_2 \times r_2}$ and $\mathbf{R}_2 \in \mathbb{R}^{p_2 \times v_2}$. We assume $\mathbf{L}$ and $\mathbf{R}$ are of full-rank so that $\mathbf{F}_t$, $\mathbf{Z}_{12,t}$, $\mathbf{Z}_{21,t}$, and $\mathbf{Z}_{22,t}$ can be viewed as transformed processes by applying the inverses of $\mathbf{L}$ and $\mathbf{R}'$, respectively, to the left and right of the data matrix $\mathbf{Y}_t$ as discussed in Section 1. Furthermore, letting $\mathbf{f}_t$ and $\mathbf{z}_t$ be the vectorized factor and idiosyncratic terms, we assume that $\mathrm{Cov}(\mathbf{f}_t) = \mathbf{I}_{r_1r_2}$ and $\mathrm{Cov}(\mathbf{z}_t) = \mathbf{I}_{p_1p_2 - r_1r_2}$. This assumption holds because one can adjust the scales of $\mathbf{L}$ and $\mathbf{R}$ accordingly. Therefore, the three noise terms are uncorrelated with each other and individually identified. Model (2.1) is general if one allows r_1, r_2, v_1 and v_2 to be zero, but for effective dimension reduction, r_1 and r_2 should be small and fixed positive integers. In addition, we assume that $\mathbf{f}_t$ and $\mathbf{z}_s$ are uncorrelated for any t and s. This is only for the simplicity in illustration, and it can be relaxed by imposing some dynamic dependence between $\mathbf{f}_t$ and $\mathbf{z}_s$. See Gao and Tsay (2021b) for details. Note that $\mathbf{L}$ and $\mathbf{R}$ are not uniquely identified because $c\mathbf{L}$ and $\mathbf{R}/c$, where $c \neq 0$, also hold for Equation (2.1).

To proceed, we further decompose $\mathbf{L}$ and $\mathbf{R}$ as follows:

$$\mathbf{L}_1 = \mathbf{A}_1\mathbf{W}_1, \quad \mathbf{L}_2 = \mathbf{A}_2\mathbf{W}_2, \quad \mathbf{R}_1 = \mathbf{P}_1\mathbf{G}_1, \text{ and } \mathbf{R}_2 = \mathbf{P}_2\mathbf{G}_2,$$

where $\mathbf{A}_i$ and $\mathbf{P}_i$ $(i = 1, 2)$ are semi-orthogonal matrices, i.e., $\mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_{r_i}$ and $\mathbf{P}_i'\mathbf{P}_i = \mathbf{I}_{v_i}$. This can be done via QR or singular value decomposition. Furthermore, let $\mathbf{X}_t = \mathbf{W}_1\mathbf{F}_t\mathbf{G}_1'$, $\mathbf{E}_{21,t} = \mathbf{W}_2\mathbf{Z}_{21,t}\mathbf{G}_1'$, $\mathbf{E}_{12,t} = \mathbf{W}_1\mathbf{Z}_{12,t}\mathbf{G}_2'$, and $\mathbf{E}_{22,t} = \mathbf{W}_2\mathbf{Z}_{22,t}\mathbf{G}_2'$, then model (2.1) can be rewritten as

$$\mathbf{Y}_t = \mathbf{A}_1\mathbf{X}_t\mathbf{P}_1' + \mathbf{A}_2\mathbf{E}_{21,t}\mathbf{P}_1' + \mathbf{A}_1\mathbf{E}_{12,t}\mathbf{P}_2' + \mathbf{A}_2\mathbf{E}_{22,t}\mathbf{P}_2'. \tag{2.2}$$

Even though $\mathbf{L}$ and $\mathbf{R}$ are of full rank, $\mathbf{A}_1$ $(\mathbf{P}_1)$ is not orthogonal to $\mathbf{A}_2$ $(\mathbf{P}_2)$ in general. Note that model (2.2) is still not identified since we can replace the triplets $(\mathbf{A}_1,\mathbf{X}_t,\mathbf{P}_1)$ by $(\mathbf{A}_1\mathbf{H}_1,\mathbf{H}_1'\mathbf{X}_t\mathbf{H}_2,$

$\mathbf{P}_1\mathbf{H}_2$) for any orthonormal matrices $\mathbf{H}_1 \in \mathbb{R}^{r_1\times r_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{r_2\times r_2}$ without altering the data generating process. The same issue exists for the idiosyncratic terms. Nevertheless the linear spaces spanned by the columns of $\mathbf{A}_i$ and $\mathbf{P}_i$, denoted respectively by $\mathcal{M}(\mathbf{A}_i)$ and $\mathcal{M}(\mathbf{P}_i)$, are uniquely defined and $\mathcal{M}(\mathbf{A}_i) = \mathcal{M}(\mathbf{L}_i)$ and $\mathcal{M}(\mathbf{P}_i) = \mathcal{M}(\mathbf{R}_i)$.

2.2. Common Orthonormal Projections

To illustrate our estimation method, we first introduce some notation. For $i = 1, 2$, let $\mathbf{B}_i$ and $\mathbf{Q}_i$ be the orthonormal complements of $\mathbf{A}_i$ and $\mathbf{P}_i$, respectively, i.e., $\mathbf{B}_i \in \mathbb{R}^{p_i\times r_i}$ and $\mathbf{Q}_i = \mathbb{R}^{p_i\times v_i}$ are semi-orthogonal matrices with $\mathbf{B}_i'\mathbf{A}_i = \mathbf{0}$ and $\mathbf{Q}_i'\mathbf{P}_i = \mathbf{0}$. Furthermore, denote $\boldsymbol{\ell}_{i,j}$, $\mathbf{r}_{i,j}$, $\mathbf{a}_{i,j}$, $\mathbf{b}_{i,j}$, $\mathbf{p}_{i,j}$ and $\mathbf{q}_{i,j}$ the j-th columns of $\mathbf{L}_i$, $\mathbf{R}_i$, $\mathbf{A}_i$, $\mathbf{B}_i$, $\mathbf{P}_i$ and $\mathbf{Q}_i$, respectively, where the range of j depends on the dimension of the corresponding matrix.

Let $\boldsymbol{\eta}_t = [\text{vec}(\mathbf{Y}_{t-1})', ..., \text{vec}(\mathbf{Y}_{t-k_0})']'$ be the vector of past k_0 lagged values of $\mathbf{Y}_t$, where $\text{vec}(\mathbf{Y}_t) = (\mathbf{y}_{1,t}', ..., \mathbf{y}_{p_2,t}')'$ and k_0 is a prescribed positive integer. Define $\boldsymbol{\Sigma}_{y,ij}(k) = \text{Cov}(\mathbf{y}_{i,t}, \mathbf{y}_{j,t-k})$. We seek the direction $\mathbf{a} \in \mathbb{R}^{p_1}$ that solves the following optimization problem:

$$\max_{\mathbf{a}\in\mathbb{R}^{p_1}} \sum_{i=1}^{p_2} \|\text{Cov}(\mathbf{a}'\mathbf{y}_{i,t}, \boldsymbol{\eta}_t)\|_2^2, \quad \text{subject to} \quad \mathbf{a}'\mathbf{a} = 1. \tag{2.3}$$

That is, we look for a common direction $\mathbf{a}$ with $\mathbf{a}'\mathbf{a} = 1$ such that it maximizes the sum of the covariance between $\mathbf{a}'\mathbf{y}_{i,t}$ and the past lagged variables, which characterize the dynamic dependence of the columns. Note that

$$\sum_{i=1}^{p_2} \|\text{Cov}(\mathbf{a}'\mathbf{y}_{i,t}, \boldsymbol{\eta}_t)\|_2^2 = \mathbf{a}' \left[\sum_{k=1}^{k_0}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2} \boldsymbol{\Sigma}_{y,ij}(k)\boldsymbol{\Sigma}_{y,ij}(k)'\right] \mathbf{a}.$$

Then, $\mathbf{a}$ is an eigenvector of the matrix

$$\mathbf{M}_1 = \sum_{k=1}^{k_0}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2} \boldsymbol{\Sigma}_{y,ij}(k)\boldsymbol{\Sigma}_{y,ij}(k)'. \tag{2.4}$$

On the other hand, under model (2.2), let $\mathbf{p}_{1,i\bullet}$ be the i-th row vector of $\mathbf{P}_1$ and define $\boldsymbol{\Sigma}_{xp,ij}(k) = \text{Cov}(\mathbf{X}_t\mathbf{p}_{1,i\bullet}', \mathbf{X}_{t-k}\mathbf{p}_{1,j\bullet}')$, for $k > 0$. Then

$$\boldsymbol{\Sigma}_{y,ij}(k) = \mathbf{A}_1\boldsymbol{\Sigma}_{xp,ij}(k)\mathbf{A}_1', \tag{2.5}$$

where $\mathbf{f}_t$ and $\mathbf{z}_s$ are uncorrelated for any t and s by assumption. Therefore,

$$\mathbf{M}_1 = \mathbf{A}_1 \left\{\sum_{k=1}^{k_0}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2} [\boldsymbol{\Sigma}_{xp,ij}(k)\boldsymbol{\Sigma}_{xp,ij}(k)']\right\} \mathbf{A}_1'. \tag{2.6}$$

We observe that $\mathbf{M}_1\mathbf{B}_1 = \mathbf{0}$, that is, the columns of $\mathbf{B}_1$ are the eigenvectors associated with the zero eigenvalues of $\mathbf{M}_1$, and the front factor loading space $\mathcal{M}(\mathbf{A}_1)$ is spanned by the eigenvectors corresponding to the r_1 non-zero eigenvalues of $\mathbf{M}_1$. Equivalently, the space spanned by the first r_1 solutions to the problem (2.3) are just the front factor loading space $\mathcal{M}(\mathbf{A}_1)$. Note that $\mathbf{M}_1$ in (2.6) is the same as equation (11) in Wang et al. (2019) as the factors are assumed to be dynamically dependent in both models. However, $\mathbf{M}_1$ in (2.6) is derived from a common orthogonal projection procedure, which provides a rational illustration for its use in factor modeling.

The r_2 orthonormal directions of the columns of $\mathbf{P}_1$ can be obtained by performing the same procedure on the the transpose of $\mathbf{Y}_t$. We can similarly construct $\mathbf{M}_2$ as $\mathbf{M}_1$ in (2.4) such that $\mathbf{M}_2\mathbf{Q}_1 = \mathbf{0}$, and therefore, $\mathcal{M}(\mathbf{P}_1)$ is the space spanned by the first r_2 non-zero eigenvectors of $\mathbf{M}_2$.

2.3. Two-Way Projected Principal Component Analysis

In this section, we introduce the idea of a 2-way projected PCA in order to recover the factor matrix $\mathbf{X}_t$ and, hence, $\mathbf{F}_t$. Using the notation in Section 2.2, it follows from model (2.2) that

$$\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1 = \mathbf{B}_1'\mathbf{A}_2\mathbf{E}_{22,t}\mathbf{P}_2'\mathbf{Q}_1, \tag{2.7}$$

which implies that $\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1$ is a matrix-variate white noise process and, hence, $\{\mathbf{b}_{1,i}'\mathbf{Y}_t\mathbf{q}_{1,j}|t = 0, \pm1, ...\}$ is a univariate white noise process, for all $1 \le i \le v_1$ and $1 \le j \le v_2$. Furthermore,

$$\mathbf{B}_2'\mathbf{Y}_t = \mathbf{B}_2'\mathbf{A}_1\mathbf{X}_t\mathbf{P}_1' + \mathbf{B}_2'\mathbf{A}_1\mathbf{E}_{12,t}\mathbf{P}_2' \text{ and } \mathbf{Y}_t\mathbf{Q}_2 = \mathbf{A}_1\mathbf{X}_t\mathbf{P}_1'\mathbf{Q}_2 + \mathbf{A}_2\mathbf{E}_{21,t}\mathbf{P}_1'\mathbf{Q}_2. \tag{2.8}$$

Therefore, $\mathbf{B}_2'\mathbf{Y}_t$ and $\mathbf{Y}_t\mathbf{Q}_2$ are uncorrelated with $\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1$ defined in (2.7). Let $\mathbf{\Omega}_{y_i} = \mathrm{Cov}(\mathbf{y}_{i,t}, \mathrm{vec}(\mathbf{Y}_t))$ and $\mathbf{\Omega}_{e_{22,i}p} = \mathrm{Cov}(\mathbf{E}_{22,t}\mathbf{p}_{2,i\bullet}', \mathrm{vec}(\mathbf{E}_{22,t}))$, where $\mathbf{p}_{2,i\bullet}$ is the i-th row vector of $\mathbf{P}_2$. It follows from (2.2) and (2.7) that

$$\mathrm{Cov}(\mathbf{y}_{i,t}, \mathrm{vec}(\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1)) = \mathbf{\Omega}_{y_i}(\mathbf{Q}_1 \otimes \mathbf{B}_1) = \mathbf{A}_2\mathbf{\Omega}_{e_{22,i}p}(\mathbf{P}_2'\mathbf{Q}_1 \otimes \mathbf{A}_2'\mathbf{B}_1). \tag{2.9}$$

Note that $\mathbf{B}_2'\mathbf{y}_{i,t}$ is uncorrelated with $\mathrm{vec}(\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1)$, for $1 \le i \le p_2$. Therefore, define

$$\mathbf{S}_1 := \sum_{i=1}^{p_2}[\mathbf{\Omega}_{y_i}(\mathbf{Q}_1 \otimes \mathbf{B}_1)][\mathbf{\Omega}_{y_i}(\mathbf{Q}_1 \otimes \mathbf{B}_1)]', \tag{2.10}$$

from which we can see that, via (2.9), $\mathbf{S}_1\mathbf{B}_2 = \mathbf{0}$. In addition, the rank of $\mathbf{S}_1 \in \mathbb{R}^{p_1 \times p_1}$ is v_1 so that $\mathbf{B}_2$ contains all the eigenvectors corresponding to the zero eigenvalues of $\mathbf{S}_1$. From the form of $\mathbf{S}_1$, we can see that $\mathbf{\Omega}_{y_i}(\mathbf{Q}_1 \otimes \mathbf{B}_1) = E[\mathbf{y}_t\mathrm{vec}(\mathbf{Y}_t)'(\mathbf{Q}_1 \otimes \mathbf{B}_1)]$, where the component $\mathrm{vec}(\mathbf{Y}_t)'(\mathbf{Q}_1 \otimes \mathbf{B}_1) = \mathrm{vec}(\mathbf{E}_{22,t})'(\mathbf{P}_2'\mathbf{Q}_1 \otimes \mathbf{A}_2'\mathbf{B}_1)$ only contains the information of the white noise. Yet we seek the direction $\mathbf{b}_{2,j} \in \mathbb{R}^{p_2}$ such that $\mathbf{b}_{2,j}'\mathbf{y}_{i,t}$ minimizes the covariance between the projected direction and the white noise processes and, consequently, $\mathbf{b}_{2,j}'\mathbf{y}_{i,t}$ should contain the information of the signal $\mathbf{F}_t$.

Similarly, we can construct $\mathbf{S}_2$ such that $\mathbf{S}_2\mathbf{Q}_2 = \mathbf{0}$, and $\mathbf{Q}_2$ contains all the eigenvectors associated with the zero eigenvalues of $\mathbf{S}_2$ at the population level. Furthermore, if $\mathbf{A}_1$, $\mathbf{P}_1$, $\mathbf{B}_2$ and $\mathbf{Q}_2$ are known, it follows from (2.2) that

$$\mathbf{B}_2'\mathbf{Y}_t\mathbf{Q}_2 = \mathbf{B}_2'\mathbf{A}_1\mathbf{X}_t\mathbf{P}_1'\mathbf{Q}_2, \tag{2.11}$$

and, consequently,

$$\mathbf{X}_t = (\mathbf{B}_2'\mathbf{A}_1)^{-1}\mathbf{B}_2'\mathbf{Y}_t\mathbf{Q}_2(\mathbf{P}_1'\mathbf{Q}_2)^{-1}, \tag{2.12}$$

where $\mathbf{B}_2'\mathbf{A}_1 \in \mathbb{R}^{r_1\times r_1}$ and $\mathbf{P}_1'\mathbf{Q}_2 \in \mathbb{R}^{r_2\times r_2}$ are two invertible matrices. To see this, note that the $\mathbf{L}$ matrix is of full rank, thus there exist matrices $\mathbf{H}_1 \in \mathbb{R}^{r_1\times r_1}$ and $\mathbf{H}_2 \in \mathbb{R}^{v_1\times r_1}$ such that

$$\mathbf{B}_2 = \mathbf{L}_1\mathbf{H}_1 + \mathbf{L}_2\mathbf{H}_2 = \mathbf{A}_1\mathbf{W}_1\mathbf{H}_1 + \mathbf{A}_2\mathbf{W}_2\mathbf{H}_2.$$

Then,

$$\mathbf{I}_{r_1} = \mathbf{B}_2'\mathbf{B}_2 = \mathbf{B}_2'\mathbf{A}_1\mathbf{W}_1\mathbf{H}_1,$$

implying that $\text{rank}(\mathbf{B}_2'\mathbf{A}_1) = r_1$, which corresponds to full rank. The invertibility of $\mathbf{P}_1'\mathbf{Q}_2$ follows from a similar argument.

2.4. Estimation

In practice, given a sample $\{\mathbf{Y}_t : t = 1, ..., n\}$, the goal is to estimate $\mathbf{A}_1$ and $\mathbf{P}_1$ or equivalently $\mathcal{M}(\mathbf{A}_1)$ and $\mathcal{M}(\mathbf{P}_1)$, the dimension (r_1, r_2) of the factor matrix, and to recover the latent factor matrix process $\mathbf{X}_t$. To illustrate the main idea, we first assume that (r_1, r_2) is known. The estimation of the factor dimension is considered in the next subsection.

For the estimation of $\mathbf{A}_1$ and $\mathbf{P}_1$, we construct the sample version of $\mathbf{M}_1$ defined in (2.4) as follows:

$$\widehat{\mathbf{M}}_1 = \sum_{k=1}^{k_0}\sum_{i=1}^{p_2}\sum_{j=1}^{p_2}\widehat{\mathbf{\Sigma}}_{y,ij}(k)\widehat{\mathbf{\Sigma}}_{y,ij}(k)', \tag{2.13}$$

where

$$\widehat{\mathbf{\Sigma}}_{y,ij}(k) = \frac{1}{n}\sum_{t=k+1}^{n}(\mathbf{y}_{i,t} - \bar{\mathbf{y}}_i)(\mathbf{y}_{j,t-k}, -\bar{\mathbf{y}}_j)', \tag{2.14}$$

and $\bar{\mathbf{y}}_i = n^{-1}\sum_{t=1}^{n}\mathbf{y}_{i,t}$, which is essentially $\mathbf{0}$ if the data are centered. Then, $\mathcal{M}(\mathbf{A}_1)$ can be estimated by $\mathcal{M}(\widehat{\mathbf{A}}_1)$, where $\widehat{\mathbf{A}}_1 = (\widehat{\mathbf{a}}_{1,1}, ..., \widehat{\mathbf{a}}_{1,r_1})$ with $\widehat{\mathbf{a}}_{1,1}, ..., \widehat{\mathbf{a}}_{1,r_1}$ being the eigenvectors corresponding to the r_1 largest eigenvalues of $\widehat{\mathbf{M}}_1$. Consequently, the orthogonal space $\mathcal{M}(\widehat{\mathbf{B}}_1)$ can easily be obtained by $\widehat{\mathbf{B}}_1 = (\widehat{\mathbf{b}}_{1,1}, ..., \widehat{\mathbf{b}}_{1,v_1})$, where $\widehat{\mathbf{b}}_{1,1}, ..., \widehat{\mathbf{b}}_{1,v_1}$ are the eigenvectors corresponding to the v_1 smallest eigenvalues of $\widehat{\mathbf{M}}_1$.

Applying the same procedure to $\{\mathbf{Y}_t', t = 1, ..., n\}$, we can construct $\widehat{\mathbf{M}}_2$, and obtain the estimator $\widehat{\mathbf{P}}_1$ of $\mathbf{P}_1$. Once we have the estimators $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{P}}_1$, we consider methods for obtaining the estimators of $\mathbf{B}_2$ and $\mathbf{Q}_2$. The choices of the estimators $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{Q}}_2$ depend on the dimensions p_1 and p_2. We only discuss the cases when p_1 and p_2 are either both small or large, and the case when one of them is small can be solved by applying both methods jointly. Let

$$\widehat{\mathbf{S}}_1 = \sum_{i=1}^{p_2} [\widehat{\mathbf{\Omega}}_{y_i}(\widehat{\mathbf{Q}}_1 \otimes \widehat{\mathbf{B}}_1)][\widehat{\mathbf{\Omega}}_{y_i}(\widehat{\mathbf{Q}}_1 \otimes \widehat{\mathbf{B}}_1)]', \tag{2.15}$$

where $\widehat{\mathbf{\Omega}}_{y_i}$ is the sample estimator of $\mathbf{\Omega}_{y_i}$ defined in Section 2.3. When p_1 and p_2 are small, we perform an eigen-analysis on $\widehat{\mathbf{S}}_1$, and let $\widehat{\mathbf{B}}_2 = (\widehat{\mathbf{b}}_{2,1}, ..., \widehat{\mathbf{b}}_{2,r_1})$, where $\widehat{\mathbf{b}}_{2,1}, ..., \widehat{\mathbf{b}}_{2,r_1}$ are the eigenvectors of $\widehat{\mathbf{S}}_1$ corresponding to the r_1 smallest eigenvalues. We can similarly obtain $\widehat{\mathbf{Q}}_1$ based on the eigen-analysis on $\widehat{\mathbf{S}}_2$, which is calculated based on the transposed data $\mathbf{Y}_t'$.

When the dimensions p_1 and p_2 are relatively large, the aforementioned choices of $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{Q}}_2$ may fare poorly because the linear spaces spanned by the chosen eigenvectors are not consistent to the true ones in the high-dimensional case. Suppose that the elements $z_{ij,t}$ of $\mathbf{Z}_{22,t}$ are independent of each other, for $1 \leq i \leq v_1$ and $1 \leq j \leq v_2$. A reasonable assumption to make in the high-dimensional case is that the leading eigenvalues of the covariance matrix of the idiosyncratic component $\text{vec}(\mathbf{L}_2\mathbf{Z}_{22,t}\mathbf{R}_2')$, or equivalently $\text{vec}(\mathbf{A}_2\mathbf{E}_{22,t}\mathbf{P}_2')$, are diverging. Therefore, we assume that the largest singular values of $\mathbf{L}_2$ and $\mathbf{R}_2$ are diverging. See also Assumption 4 in Section 3. We can partition the singular vectors as $\mathbf{A}_2 = (\mathbf{A}_{21}, \mathbf{A}_{22})$ and $\mathbf{P}_2 = (\mathbf{P}_{21}, \mathbf{P}_{22})$ with $\mathbf{A}_{21} \in \mathbb{R}^{p_1 \times k_1}$ and $\mathbf{P}_{21} \in \mathbb{R}^{p_2 \times k_2}$, which correspond to the k_1 and k_2 diverging singular values of $\mathbf{L}_2$ and $\mathbf{R}_2$, respectively. Let $\mathbf{B}_2^* = (\mathbf{A}_{22}, \mathbf{B}_2) \in \mathbb{R}^{p_1 \times (p_1 - k_1)}$ and $\mathbf{Q}_2^* = (\mathbf{P}_{22}, \mathbf{Q}_2) \in \mathbb{R}^{p_2 \times (p_2 - k_2)}$. Under the assumption that the leading k_1 singular values of $\mathbf{L}_2$ and the leading k_2 of $\mathbf{R}_2$ are diverging, we can consistently estimate the spaces $\mathcal{M}(\mathbf{A}_{21})$ and $\mathcal{M}(\mathbf{Q}_{21})$ and, hence, their orthogonal parts $\mathcal{M}(\mathbf{B}_2^*)$ and $\mathcal{M}(\mathbf{Q}_2^*)$. From the above discussion, $\mathcal{M}(\mathbf{B}_2)$ and $\mathcal{M}(\mathbf{Q}_2)$ are sub-spaces of $\mathcal{M}(\mathbf{B}_2^*)$ and $\mathcal{M}(\mathbf{Q}_2^*)$, respectively. Once we have consistent estimators for $\mathbf{B}_2^*$ and $\mathbf{Q}_2^*$, denoted by $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{Q}}_2^*$, respectively, there exist semi-orthogonal matrices $\mathbf{\Xi}_1 \in \mathbb{R}^{(p_1 - k_1) \times r_1}$ and $\mathbf{\Xi}_2 \in \mathbb{R}^{(p_2 - k_2) \times r_2}$ such that $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^*\mathbf{\Xi}_1$ and $\widehat{\mathbf{Q}}_2 = \widehat{\mathbf{Q}}_2^*\mathbf{\Xi}_2$. In practice, it is not easy to find $\mathbf{\Xi}_1$ and $\mathbf{\Xi}_2$ such that $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{Q}}_2$ are consistent estimators of $\mathbf{B}_2$ and $\mathbf{Q}_2$. Nevertheless, any choices of $\mathbf{\Xi}_1$ and $\mathbf{\Xi}_2$ can mitigate the diverging effect of the leading eigenvalues of the noise covariance matrices since they are all orthogonal to $\widehat{\mathbf{A}}_{21}$ and $\widehat{\mathbf{P}}_{21}$, respectively. Therefore, we only need to guarantee the invertibilities of the the matrices $\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{P}}_1'\widehat{\mathbf{Q}}_2$ in order to recover the latent factors.

In applications, with the estimators $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{Q}}_2^*$, the columns of $\mathbf{\Xi}_1$ are chosen as the r_1 eigenvectors of $\widehat{\mathbf{B}}_2^{*\prime}\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\widehat{\mathbf{B}}_2^*$ corresponding the r_1 largest eigenvalues, and the columns of $\mathbf{\Xi}_2$ are the r_2 eigenvectors of $\widehat{\mathbf{Q}}_2^{*\prime}\widehat{\mathbf{P}}_1\widehat{\mathbf{P}}_1'\widehat{\mathbf{Q}}_2^*$ corresponding to the largest r_2 eigenvalues. These choices guarantee that both $\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{P}}_1'\widehat{\mathbf{Q}}_2$ behave well in practical calculations. Finally, we recover the latent factor matrix as

$$\widehat{\mathbf{X}}_t = (\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1)^{-1}\widehat{\mathbf{B}}_2'\mathbf{Y}_t\widehat{\mathbf{Q}}_2(\widehat{\mathbf{P}}_1'\widehat{\mathbf{Q}}_2)^{-1}. \tag{2.16}$$

With $\widehat{\mathbf{A}}_1$, $\widehat{\mathbf{P}}_1$ and the estimated factor process $\widehat{\mathbf{X}}_t$, we can make an h-step ahead prediction for the $\mathbf{Y}_t$ series using the formula $\widehat{\mathbf{Y}}_{n+h} = \widehat{\mathbf{A}}_1\widehat{\mathbf{X}}_{n+h}\widehat{\mathbf{P}}_1'$, where $\widehat{\mathbf{X}}_{n+h}$ is an h-step ahead forecast for $\mathbf{X}_t$ series based on the estimated past values $\widehat{\mathbf{X}}_1, ..., \widehat{\mathbf{X}}_n$. This can be done, for example, by fitting a matrix-autoregressive model to $\{\widehat{\mathbf{X}}_1, ..., \widehat{\mathbf{X}}_n\}$ as the one introduced in Chen et al. (2020).

2.5. Diagonal-Path Selections of the Order of Factor Matrix

The estimation of $\mathbf{A}_1$, $\mathbf{P}_1$, and $\mathbf{X}_t$ of the prior sections are based on given r_1 and r_2, which are unknown in practice. To the best of our knowledge, there is no efficient method available to estimate them in the literature. The most relevant one is the ratio-based method of Wang et al. (2019), but it can be shown that the method is not appropriate when the leading eigenvalues of the covariance of the idiosyncratic term are diverging. See the simulation results in Section 4. For the vector factor models, there are some methods available. See, for example, the information criterion in Bai and Ng (2002) and Bai (2003), the random matrix theory method in Onatski (2010), the ratio-based method in Lam and Yao (2012), the canonical correlation analysis in Gao and Tsay (2019), and the white noise testing approach in Gao and Tsay (2021b), among others. But those methods cannot be applied to the matrix-factor models directly.

In this section, we propose a diagonal-path method to search the dimension (r_1, r_2) by extending the approach of Gao and Tsay (2021b). The idea of our method follows from equation (2.7) that $\mathbf{B}_1'\mathbf{Y}_t\mathbf{Q}_1$ is a matrix-variate white noise process. Let $\widehat{\mathbf{\Gamma}}_1$ and $\widehat{\mathbf{\Gamma}}_2$ be the matrices of eigenvectors (in the decreasing order of the corresponding eigenvalues) of the sample matrix $\widehat{\mathbf{M}}_1$ in (2.13) and $\widehat{\mathbf{M}}_2$, respectively. Define $\widehat{\mathbf{W}}_t = \widehat{\mathbf{\Gamma}}_1{}'\mathbf{Y}_t\widehat{\mathbf{\Gamma}}_2$ and let $\widehat{\mathbf{W}}_t(i,j) \in \mathbb{R}^{(p_1-i+1)\times(p_2-j+1)}$ be the lower-right submatrix consisting of the i-th to the p_1-th rows and the j-th to the p_2-th columns of $\widehat{\mathbf{W}}_t$, and $\widehat{\mathbf{W}}_t^*(i,j) \in \mathbb{R}^{(i-1)\times(j-1)}$ be the upper-left submatrix of $\widehat{\mathbf{W}}_t$. Our test procedure searches the order (i,j) such that $\widehat{\mathbf{W}}_t^*(i,j)$ consists of all the factors and the remaining elements of $\widehat{\mathbf{W}}_t$ are white noises. The estimate of (r_1, r_2) is then $(i-1, j-1)$. The testing procedure is discussed below, and the test statistic used depends on the dimension p_1p_2.

If the dimension p_1p_2 is small, implying that $\mathbf{Y}_t$ is a low dimensional matrix, we recommend using the well-known Ljung-Box statistic $Q_s(m)$ for multivariate time series, where s and m denote the dimension of the vector and the number of lags used. See, for example, Hosking (1980) and Tsay (2014). Specifically, we first search the minimum of r_1 and r_2 along the diagonal of $\widehat{\mathbf{W}}_t$. Consider the null hypothesis

$$H_0(l) : \text{vec}(\widehat{\mathbf{W}}_t(l,l)) \text{ is a vector white noise,}$$

with type-I error α. $H_0(l)$ is rejected if $Q_{d_l}(m) \geq \chi^2_{d_l^2 m, 1-\alpha}$, where $d_l = (p_1 - l + 1)(p_2 - l + 1)$ is the dimension of $\text{vec}(\widehat{\mathbf{W}}_t(l,l))$ and $\chi^2_{d_l^2 m, 1-\alpha}$ is the $(1-\alpha)$-th quantile of a chi-squared distribution with $d_l^2 m$ degrees of freedom. We start with $l = 1$. If $H_0(1)$ is rejected, we increase l by 1 and repeat the testing procedure until we cannot reject $H_0(l)$, and denote the resulting order as l^*. Two situations can happen. If $l^* = \min(p_1, p_2)$ and we still reject $H_0(l^*)$, we fix one dimension (say p_1 when $p_1 = l^*$), and test whether $\text{vec}(\widehat{\mathbf{W}}_t(p_1, p_1 + j))$ is white noise or not by starting with $j = 1$ until we cannot reject H_0. If $l^* < \min(p_1, p_2)$, then we perform a back testing to determine the maximum order of the factor matrix. That is, we first test whether $\text{vec}(\widehat{\mathbf{W}}_t(l^* - 1 + i, l^* - 1))$ is a vector white noise starting with $i = 1$. Increase i by 1 and repeat the testing procedure until we cannot reject H_0 at $i = i^*$. Second, we test whether $\text{vec}(\widehat{\mathbf{W}}_t(l^* + i^* - 2, l^* - 1 + j))$ is a vector white noise starting with $j = 1$. Increase j by 1 and repeat the testing procedure until we reject H_0 at $j = j^*$. Then, we have $\widehat{r}_1 = l^* + i^* - 2$ and $\widehat{r}_2 = l^* + j^* - 2$. Finally, $\widehat{\mathbf{\Gamma}}_1 = [\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1]$ and $\widehat{\mathbf{\Gamma}}_2 = [\widehat{\mathbf{P}}_1, \widehat{\mathbf{Q}}_1]$, where $\widehat{\mathbf{A}}_1 \in \mathbb{R}^{p_1 \times \widehat{r}_1}$ and $\widehat{\mathbf{P}}_1 \in \mathbb{R}^{p_2 \times \widehat{r}_2}$.

For large p_1 and/or p_2, we use the same testing procedure, but replace the $Q_s(m)$ test statistics by high-dimensional white noise (HDWN) tests. This is so, because $Q_s(m)$ is no longer adequate. Instead, we consider two HDWN test statistics in this paper. The first test statistic is introduced by Chang et al. (2017) and makes use of the maximum absolute auto- and cross-correlations of the component series. Specifically, let $\widehat{\mathbf{\Gamma}}_w(k) = [\widehat{\rho}_{ij}(k)]_{1 \le i,j \le d_l}$ be the lage-k sample auto-correlation matrix of $\text{vec}(\widehat{\mathbf{W}}_t(l,l))$, the test statistic T_n is defined as

$$T_n = \max_{1 \le k \le m} \max_{1 \le i,j \le d_l} n^{1/2} |\widehat{\rho}_{ij}(k)|,$$

and its limiting distribution under $H_0(l)$ can be approximated by that of the L_∞-norm of a normal random vector, which can be simulated by a bootstrapping algorithm. The second HDWN test statistic is developed by Tsay (2020). Let $\widehat{\mathbf{\Gamma}}_{w,k} = [\widehat{\Gamma}_{w,k}(i,j)]_{1 \le i,j \le d_l}$ be the lag-k sample rank auto-correlation matrix of an orthogonalized vector of $\text{vec}(\widehat{\mathbf{W}}_t(l,l))$, where the orthogonalization can be done via PCA if $d_l < n$ and we only vectorize the top-left $\min(p_1, \sqrt{\varepsilon n})$-by-$\min(p_2, \sqrt{\varepsilon n})$ principal submatrix of $\widehat{\mathbf{W}}_t(l,l)$ if $d_l \ge n$; See Gao and Tsay (2021b) for details. The test statistic is defined as

$$T(m) = \max\{\sqrt{n} |\widehat{\Gamma}_{w,k}(i,j)| : 1 \le i,j \le d_l, 1 \le k \le m\},$$

and its limiting distribution under $H_0(l)$ is a function of the standard Gumbel distribution via the extreme value theory. The critical values and the rejection regions of the test statistics are available in closed form and can be found in Tsay (2020) or Section 2.3 in Gao and Tsay (2021b).

3. Theoretical Properties

We present first the asymptotic theory for the estimation method of Section 2, assuming that r_1 and r_2 are fixed. The consistency of the white noise tests in determining r_1 and r_2 of the matrix factor is shown thereafter. Traditional asymptotic properties are established under the setting that the sample size n tends to ∞, but everything else is fixed. Modern time series analysis encounters the situation that the number of time series $p_1 p_2$ can be as large as, or even larger than, the sample size n. We deal with these two situations separately in Sections 3.1 and 3.2 below.

3.1. Asymptotics When $n \to \infty$, But p_1 and p_2 Are Fixed

Consider the asymptotic properties of the estimates when p_1 and p_2 are fixed, but $n \to \infty$. These properties reflect the behavior of our proposed estimates when n is large and the dimensions p_1 and p_2 are relatively small. We begin with some assumptions.

Assumption 1. *The process $\{\text{vec}(\mathbf{Y}_t), \text{vec}(\mathbf{F}_t)\}$ is α-mixing with the mixing coefficient satisfying the condition $\sum_{k=1}^{\infty} \alpha_p(k)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, where*

$$\alpha_p(k) = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i, B \in \mathcal{F}_{i+k}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{(\text{vec}(\mathbf{Y}_t), \text{vec}(\mathbf{F}_t)) : i \le t \le j\}$.

Assumption 2. *For any $i = 1, ..., r_1 r_2$ and $1 \le j \le p_1 p_2 - r_1 r_2$, $E|f_{i,t}|^{2\gamma} < C_1$ and $E|z_{j,t}|^{2\gamma} < C_2$, where $f_{i,t}$ and $z_{j,t}$ are the i-th and j-th element of $\mathbf{f}_t$ and $\mathbf{z}_t$, respectively, C_1 and C_2 are positive constants, and γ is given in Assumption 1.*

Assumption 1 is standard for dependent random processes. See Gao et al. (2019) for a theoretical justification for VAR series. The conditions in Assumption 2 imply that $E|y_{ij,t}|^{2\gamma} < C$ under the setting that p_1 and p_2 are fixed. We adopt the discrepancy measure used by Pan and Yao (2008): for two $p \times r$ semi-orthogonal matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ satisfying the condition $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{H}_2'\mathbf{H}_2 = \mathbf{I}_r$, the difference between the two linear spaces $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$ is measured by

$$D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}\mathbf{H}_2) = \sqrt{1 - \frac{1}{r}\text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')}. \tag{3.1}$$

Note that $D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}\mathbf{H}_2) \in [0, 1]$. It is equal to 0 if and only if $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and to 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. By Lemma A1(i) in Pan and Yao (2008), $D(\cdot, \cdot)$ is a well-defined distance measure on some quotient space of matrices. On the other hand, as only $\mathbf{H}_i\mathbf{H}_i'$ is uniquely defined, we may also adopt the measure

$$\rho(\mathbf{H}_1, \mathbf{H}_2) = \|\mathbf{H}_1\mathbf{H}_1' - \mathbf{H}_2\mathbf{H}_2'\|_F, \tag{3.2}$$

which is the Frobenius norm of the difference between the projection matrices of two spaces and is also a well-defined distance between linear subspaces. In addition, if we denote the singular values of $\mathbf{H}_1'\mathbf{H}_2$ by $\{\sigma_i\}_{i=1}^r$, in descending order, then the principal angles between $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$, $\mathbf{\Theta}(\mathbf{H}_1, \mathbf{H}_2) = \text{diag}(\theta_1, ..., \theta_r)$, are defined as $\text{diag}\{\cos^{-1}(\sigma_1), ..., \cos^{-1}(\sigma_r)\}$; see, for example, Theorem I.5.5 of Stewart and Sun (1990). The squared Frobenius norm of the so-called $\sin \mathbf{\Theta}$ distance, defined as

$$\|\sin \mathbf{\Theta}(\mathbf{H}_1, \mathbf{H}_2)\|_F^2 := \sum_{i=1}^r \sin^2(\theta_i), \tag{3.3}$$

can also be used to measure the distance between two linear spaces. In fact, if r is finite, the distances in (3.1)–(3.3) are equivalent, because

$$\rho^2(\mathbf{H}_1,\mathbf{H}_2) =\|\mathbf{H}_1\mathbf{H}_1'\|_F^2 + \|\mathbf{H}_2\mathbf{H}_2'\|_F^2 - 2\text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2') = 2r\{D(\mathcal{M}(\mathbf{H}_1),\mathcal{M}(\mathbf{H}_2))\}^2$$
$$=2r - 2\|\mathbf{H}_1'\mathbf{H}_2\|_F^2 = 2\sum_{i=1}^r(1-\sigma_i^2) = 2\sum_{i=1}^r \sin^2(\theta_i) = 2\|\sin\mathbf{\Theta}(\mathbf{H}_1,\mathbf{H}_2)\|_F^2. \qquad (3.4)$$

Therefore, we shall only use the distance in (3.1) to present our theoretical results. The following theorem establishes the consistency of the estimated loading matrices $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{P}}_1$, their orthonormal complements $\widehat{\mathbf{B}}_1$ and $\widehat{\mathbf{Q}}_1$, the matrices $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{Q}}_2$, and the extracted common factor $\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}_t\widehat{\mathbf{P}}_1'$.

Theorem 1. *Suppose Assumptions 1 and 2 hold and (r_1,r_2) are known and fixed. Then, for fixed p_1 and p_2,*

$$D(\mathcal{M}(\widehat{\mathbf{A}}_1),\mathcal{M}(\mathbf{A}_1)) = O_p(n^{-1/2}),\quad D(\mathcal{M}(\widehat{\mathbf{B}}_1),\mathcal{M}(\mathbf{B}_1)) = O_p(n^{-1/2}),$$
$$D(\mathcal{M}(\widehat{\mathbf{P}}_1),\mathcal{M}(\mathbf{P}_1)) = O_p(n^{-1/2}),\quad D(\mathcal{M}(\widehat{\mathbf{Q}}_1),\mathcal{M}(\mathbf{Q}_1)) = O_p(n^{-1/2})$$

and

$$D(\mathcal{M}(\widehat{\mathbf{B}}_2),\mathcal{M}(\mathbf{B}_2)) = O_p(n^{-1/2}),\quad D(\mathcal{M}(\widehat{\mathbf{Q}}_2),\mathcal{M}(\mathbf{Q}_2)) = O_p(n^{-1/2}),$$

as $n\to\infty$. Furthermore,

$$\|\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}_t\widehat{\mathbf{P}}_1' - \mathbf{A}_1\mathbf{X}_t\mathbf{P}_1'\|_2 = O_p(n^{-1/2}).$$

From Theorem 1, as expected, the convergence rates of all estimates are standard at $\sqrt{n}$, which is commonly seen in the traditional statistical theory. If the largest r_1 and r_2 eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$ are distinct, then $\mathbf{A}_1$ and $\mathbf{P}_1$ are uniquely defined up to a change of signs in columns. Note that the consistency in estimating the linear spaces $\mathcal{M}(\mathbf{B}_1)$ and $\mathcal{M}(\mathbf{B}_2)$ is more meaningful than that in estimating $\mathbf{B}_1$ and $\mathbf{B}_2$, because their columns correspond to the zero eigenvalues of $\mathbf{M}_1$ and $\mathbf{S}_1$, respectively, and, as such, they cannot be uniquely characterized.

3.2. Asymptotics When $n, p_1, p_2 \to \infty$

Turn to the case of high-dimensional matrices. For vectorized variables, it is well known that if the dimension p_1p_2 diverges faster than $n^{1/2}$, then their sample covariance matrix is no longer a consistent estimate of their population one. On the other hand, if $p_1p_2 = o(n^{1/2})$, it is still possible to consistently estimate the factor loading matrix and the number of common factors. See Gao and Tsay (2019) for details. Thus, without any additional assumptions on the underlying structure of the time series under study, p_1p_2 can only be as large as $o(n^{1/2})$. With recent developments in high-dimensional PCA, under the assumption that a small number of the eigenvalues of the covariance matrix are diverging, which are called the spiked ones, and all the remaining eigenvalues are bounded, the eigenspace can still be consistently estimated using the matrix perturbation theory; see, for example, Stewart and Sun (1990) and Fan et al. (2013). Therefore, to deal with the case of large p_1p_2, we impose some conditions on the transformation matrices $\mathbf{L}$ and $\mathbf{R}$ of Equation (2.1) and the cross dependence of $\mathbf{Y}_t$, which allow the factors to be either pervasive or slightly weak.

Assumption 3. *(i)* $\mathbf{L}_1 = (\boldsymbol{\ell}_1, ..., \boldsymbol{\ell}_{r_1})$ *and* $\mathbf{R}_1 = (\mathbf{r}_1, ..., \mathbf{r}_{r_2})$ *such that* $\|\boldsymbol{\ell}_i\|_2^2 \asymp p_1^{1-\delta_1}$, $\|\mathbf{r}_j\|_2^2 \asymp p_2^{1-\delta_1}$, *for* $i = 1, ..., r_1$, $j = 1, ..., r_2$, *and* $\delta_1 \in [0, 1)$; *(ii) For each* $i = 1, ..., r_1$, $j = 1, ..., r_2$ *and* δ_1 *given in (i),* $\min_{\theta_k \in \mathbb{R}, k \neq i} \|\boldsymbol{\ell}_i - \sum_{1 \leq k \leq r_1, k \neq i} \theta_k \boldsymbol{\ell}_k\|_2^2 \asymp p_1^{1-\delta_1}$ *and* $\min_{\theta_k \in \mathbb{R}, k \neq j} \|\mathbf{r}_j - \sum_{1 \leq k \leq r_2, k \neq j} \theta_k \mathbf{r}_k\|_2^2 \asymp p_2^{1-\delta_1}$.

Assumption 4. *(i)* $\mathbf{L}_2$ *and* $\mathbf{R}_2$ *admit a singular value decomposition* $\mathbf{L}_2 = \mathbf{A}_2 \mathbf{D}_2 \mathbf{U}_2'$ *and* $\mathbf{R}_2 = \mathbf{P}_2 \boldsymbol{\Lambda}_2 \mathbf{V}_2'$, *where* $\mathbf{A}_2 \in \mathbb{R}^{p_1 \times v_1}$ *and* $\mathbf{P}_2 \in \mathbb{R}^{p_2 \times v_2}$ *are given in Equation (2.2),* $\mathbf{D}_2 = diag(d_1, ..., d_{v_1})$ *and* $\mathbf{U}_2 \in \mathbb{R}^{v_1 \times v_1}$ *satisfying* $\mathbf{U}_2'\mathbf{U}_2 = \mathbf{I}_{v_1}$, $\boldsymbol{\Lambda}_2 = diag(\gamma_1, ..., \gamma_{v_2})$, $\mathbf{V}_2 \in \mathbb{R}^{v_2 \times v_2}$ *satisfying* $\mathbf{V}_2'\mathbf{V}_2 = \mathbf{I}_{v_2}$; *(ii) There exist finite integers* $1 \leq k_1 < v_1$ *and* $1 \leq k_2 < v_2$ *such that* $d_1 \asymp ... \asymp d_{k_1} \asymp p_1^{(1-\delta_2)/2}$ *and* $\gamma_1 \asymp ... \asymp \gamma_{k_2} \asymp p_2^{(1-\delta_2)/2}$ *for some* $\delta_2 \in [0, 1)$ *and* $d_{k_1+1} \asymp ... \asymp d_{v_1} \asymp 1 \asymp \gamma_{k_2+1} \asymp ... \asymp \gamma_{v_2}$.

Assumption 5. *(i) For any* $1 \leq l_1 \leq v_1$, $1 \leq l_2 \leq v_2$, $\mathbf{h} \in \mathbb{R}^{l_1 l_2}$, $\mathbf{U} \in \mathbb{R}^{v_1 \times l_1}$ *and* $\mathbf{V} \in \mathbb{R}^{v_2 \times l_2}$ *with* $\|\mathbf{h}\|_2 = c < \infty$, $\mathbf{U}'\mathbf{U} = \mathbf{I}_{l_1}$ *and* $\mathbf{V}'\mathbf{V} = \mathbf{I}_{l_2}$, *we assume* $E|\mathrm{h}'\mathrm{vec}(\mathbf{U}'\mathbf{Z}_{22,t}\mathbf{V})|^{2\gamma} < \infty$; *(ii)* $\sigma_{\min}(\boldsymbol{\Xi}_1'\mathbf{B}_2^{*\prime}\mathbf{A}_1) \geq C_3$ *and* $\sigma_{\min}(\boldsymbol{\Xi}_2'\mathbf{Q}_2^{*\prime}\mathbf{P}_1) \geq C_4$ *for some constants* $C_3, C_4 > 0$ *and some semi-orthogonal matrices* $\boldsymbol{\Xi}_1 \in \mathbb{R}^{(p_1 - v_1) \times r_1}$ *and* $\boldsymbol{\Xi}_2 \in \mathbb{R}^{(p_2 - v_2) \times r_2}$ *satisfying* $\boldsymbol{\Xi}_1'\boldsymbol{\Xi}_1 = \mathbf{I}_{r_1}$ *and* $\boldsymbol{\Xi}_2'\boldsymbol{\Xi}_2 = \mathbf{I}_{r_2}$, *where* $\sigma_{\min}$ *denotes the minimum non-zero singular value of a matrix.*

The quantity δ_1 of Assumption 3 is used to quantify the strength of the factors. If $\delta_1 = 0$, the corresponding factors are called strong factors or pervasive factors, which are used in Bai and Ng (2002) and Fan et al. (2013), because it includes the case where each element of $\boldsymbol{\ell}_i$ and $\mathbf{r}_j$ is $O(1)$ under which the eigenvalues of the covariance of $\mathrm{vec}(\mathbf{X}_t)$ in (2.2) are of order $p_1 p_2$. If $\delta_1 > 0$, the corresponding factors are weak factors and the smaller the δ_1 is, the stronger the factors are. One advantage of using index δ_1 is to link the convergence rates of the estimated factors explicitly to the strength of the factors. This assumption is slightly different from Condition 4 in Wang et al. (2019), which actually imposes two different strengths ς_1 and ς_2 on the front and back loading matrices, respectively. Due to the non-uniqueness of the loading matrices, we can always choose δ_1 such that $(p_1 p_2)^{(1-\delta_1)/2} \asymp p_1^{(1-\varsigma_1)/2} p_2^{(1-\varsigma_2)/2}$. Hence Assumption 4 ensures that all common factor components in $\mathbf{F}_t$ are of equal strength δ_1. There are many sufficient conditions for Assumption 4 to hold. See the discussion of Assumption 5 in Gao and Tsay (2021b). Assumption 5(i) is mild and includes the standard normal distribution as a special case. Assumption 5(ii) is reasonable since $\mathbf{B}_2$ is a subspace of $\mathbf{B}_2^*$, $\widehat{\mathbf{Q}}_2$ is a subspace of $\widehat{\mathbf{Q}}_2^*$, and the discussion in Section 2.3 implies that that $\boldsymbol{\Xi}_1'\mathbf{B}_2^{*\prime}\mathbf{A}_1$ and $\boldsymbol{\Xi}_2'\mathbf{Q}_2^{*\prime}\mathbf{P}_1$ are invertible. The choices of $\widehat{\boldsymbol{\Xi}}_1$ and $\widehat{\boldsymbol{\Xi}}_2$, and hence $\widehat{\mathbf{B}}_2 = \widehat{\mathbf{B}}_2^*\widehat{\boldsymbol{\Xi}}_1$ and $\widehat{\mathbf{Q}}_2 = \widehat{\mathbf{Q}}_2^*\widehat{\boldsymbol{\Xi}}_2$ will be discussed later.

If p_1 and p_2 are large, it is not possible to consistently estimate $\mathbf{B}_2$ (also $\mathbf{Q}_2$) or even $\mathcal{M}(\mathbf{B}_2)$ (also $\mathcal{M}(\mathbf{Q}_2)$). Instead, we will estimate $\mathbf{B}_2^* = (\mathbf{A}_{22}, \mathbf{B}_2)$ or equivalently $\mathcal{M}(\mathbf{B}_2^*)$, which is the subspace spanned by the eigenvectors associated with the $p_1 - k_1$ smallest eigenvalues of $\mathbf{S}_1$. Assume $\widehat{\mathbf{B}}_2^*$ consists of the eigenvectors corresponding to the smallest $p - k_1$ eigenvalues of $\widehat{\mathbf{S}}_1$. Under some conditions, we can show that $\mathcal{M}(\widehat{\mathbf{B}}_2^*)$ is consistent to $\mathcal{M}(\mathbf{B}_2^*)$. This is also the case in the literature on high-dimensional PCA with i.i.d. data. See, for example, Shen et al. (2016) and the references therein. Therefore, the choice of $\widehat{\mathbf{B}}_2$ should be a subspace of $\widehat{\mathbf{B}}_2^*$, and we discuss it before Theorem 3 below.

Theorem 2. *Suppose Assumptions 1–5 hold and r_1 and r_2 are known and fixed. As $n\to\infty$, if $p_1^{\delta_1}p_2^{\delta_1}n^{-1/2}=o(1)$, then*

$$\|D(\mathcal{M}(\widehat{\mathbf{A}}_1),\mathcal{M}(\mathbf{A}_1))\|_2=O_p(p_1^{\delta_1}p_2^{\delta_1}n^{-1/2})\ \textit{and}\ \|D(\mathcal{M}(\widehat{\mathbf{P}}_1),\mathcal{M}(\mathbf{P}_1))\|_2=O_p(p_1^{\delta_1}p_2^{\delta_1}n^{-1/2}),$$

and the above results also hold for $\|D(\mathcal{M}(\widehat{\mathbf{B}}_1),\mathcal{M}(\mathbf{B}_1))\|_2$ and $\|D(\mathcal{M}(\widehat{\mathbf{Q}}_1),\mathcal{M}(\mathbf{Q}_1))\|_2$. Furthermore,

$$\|D(\mathcal{M}(\widehat{\mathbf{B}}_2^*),\mathcal{M}(\mathbf{B}_2^*))\|_2=O_p(p_1^{\delta_2}p_2^{3\delta_2/2}n^{-1/2}+p_1^{\delta_1}p_2^{\delta_1+\delta_2}n^{-1/2}),$$

and

$$\|D(\mathcal{M}(\widehat{\mathbf{Q}}_2^*),\mathcal{M}(\mathbf{Q}_2^*))\|_2=O_p(p_1^{3\delta_2/2}p_2^{\delta_2}n^{-1/2}+p_1^{\delta_1+\delta_2}p_2^{\delta_1}n^{-1/2}).$$

Remark 1. *(i) For the consistencies of $\widehat{\mathbf{A}}_1$ and $\widehat{\mathbf{P}}_1$, we require $p_1p_2=o(n^{\frac{1}{2\delta_1}})$. When the strength $\delta_1\in[0,1/2]$, the range of the total dimensions p_1p_2 can be greater than $\sqrt{n}$. When $\delta_1=0$, i.e., all the factors are strong or pervasive, we can achieve the standard convergence rate $\sqrt{n}$, because the strong factors are easier to estimate.*
(ii) The conditions for the consistencies of $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{Q}}_2^*$ are slightly stronger since they depend on the estimation error in the first step. Specifically, we require $p_1^{\delta_2}p_2^{3\delta_2/2}n^{-1/2}=o(1)$, $p_1^{\delta_1}p_2^{\delta_1+\delta_2}n^{-1/2}=o(1)$, $p_1^{3\delta_2/2}p_2^{\delta_2}n^{-1/2}=o(1)$ and $p_1^{\delta_1+\delta_2}p_2^{\delta_1}n^{-1/2}=o(1)$. To give a better illustration, we assume $p_1\asymp p_2\asymp p$, then we have $p^2=o(n^{\frac{1}{2\delta_1}})$ for the consistency of $\widehat{\mathbf{A}}_1$ (also $\widehat{\mathbf{P}}_1$), and $p^2=o(\min\{n^{\frac{2}{5\delta_2}},n^{\frac{1}{2\delta_1+\delta_2}}\})$ for that of $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{Q}}_2^*$, which is slightly stronger than the former.
(iii) When r_1 and r_2 increase slowly with p_1, p_2, and n, by the proof of Theorem 2 or Lemma 1 in Gao and Tsay (2021b), we may re-establish the asymptotic bound of $\|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_2$ such that

$$\|D[\mathcal{M}(\widehat{\mathbf{A}}_1),\mathcal{M}(\mathbf{A}_1)]\|_2\le C\|\widehat{\mathbf{M}}_1-\mathbf{M}_1\|_2/\lambda_{1,r_1}(\mathbf{M}_1),$$

and we may still obtain consistent results by including some exponents of r_1 and/or r_2 in the asymptotic bounds of Theorem 2, but the argument for the matrix-variate case is lengthy and we do not pursue the issue here to save space.

Equipped with $\widehat{\mathbf{B}}_2^*$ and $\widehat{\mathbf{Q}}_2^*$, we suggest to choose $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{Q}}_2$ as $\widehat{\mathbf{B}}_2=\widehat{\mathbf{B}}_2^*\widehat{\boldsymbol{\Xi}}_1$ and $\widehat{\mathbf{Q}}_2=\widehat{\mathbf{Q}}_2^*\widehat{\boldsymbol{\Xi}}_2$, where $\widehat{\boldsymbol{\Xi}}_1=(\widehat{\boldsymbol{\xi}}_{1,1},..,\widehat{\boldsymbol{\xi}}_{1,r_1})\in\mathbb{R}^{(p_1-k_1)\times r_1}$ and $\widehat{\boldsymbol{\Xi}}_2=(\widehat{\boldsymbol{\xi}}_{2,1},..,\widehat{\boldsymbol{\xi}}_{2,r_2})\in\mathbb{R}^{(p_2-k_2)\times r_2}$, where $\widehat{\boldsymbol{\xi}}_{1,i}$ is the vector associated with the i-th largest eigenvalues of $\widehat{\mathbf{B}}_2^{*\prime}\widehat{\mathbf{A}}_1\widehat{\mathbf{A}}_1'\widehat{\mathbf{B}}_2^*$ and $\widehat{\boldsymbol{\xi}}_{2,j}$ is the vector associated with the j-th largest eigenvalues of $\widehat{\mathbf{Q}}_2^{*\prime}\widehat{\mathbf{P}}_1\widehat{\mathbf{P}}_1'\widehat{\mathbf{Q}}_2^*$. These choices can guarantee that the matrices $(\widehat{\mathbf{B}}_2'\widehat{\mathbf{A}}_1)^{-1}$ and $(\widehat{\mathbf{Q}}_2'\widehat{\mathbf{P}}_1)^{-1}$ behave well in recovering the factor $\widehat{\mathbf{X}}_t$. Furthermore, they can still eliminate the diverging part of the noise covariance matrix and give prominent convergence rate, as shown in Theorem 3. There are many ways to choose the numbers of components k_1 and k_2 in Assumption 4 so long as $p_1-k_1>r_1$ and $p_2-k_2>r_2$. We discuss the choices of k_1 and k_2 in Remark 2 below. The following theorem states the convergence rate of the extracted common factors.

Theorem 3. *Under the assumptions in Theorem 2, we have*

$$\begin{aligned}(p_1p_2)^{-1/2}\|\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}_t\widehat{\mathbf{P}}_1' - \mathbf{A}_1\mathbf{X}_t\mathbf{P}_1'\|_2 =& O_p\Big(p_1^{-\delta_1/2}p_2^{-\delta_1/2}(\|D(\mathcal{M}(\widehat{\mathbf{A}}_1),\mathcal{M}\mathbf{A}_1))\|_2\\ &+\|D(\mathcal{M}(\widehat{\mathbf{P}}_1),\mathcal{M}(\mathbf{P}_1))\|_2) + p_1^{-\delta_2/2}\|D(\mathcal{M}(\widehat{\mathbf{B}}_2^*),\mathcal{M}(\mathbf{B}_2^*))\|_2\\ &+p_2^{-\delta_2/2}\|D(\mathcal{M}(\widehat{\mathbf{Q}}_2^*),\mathcal{M}(\mathbf{Q}_2))\|_2 + p_1^{-1/2}p_2^{-1/2}\Big).\end{aligned}$$

Remark 2. *(i) A similar result is given in Theorem 3 of Lam et al. (2011) and Theorem 5 of Gao and Tsay (2021b), which deal with the approximate factor model and a structured factor model, respectively. When $\delta_1 = \delta_2 = 0$, i.e. the factors and the noise terms are all strong, the convergence rate in Theorem 3 is $O_p((p_1p_2)^{-1/2} + n^{-1/2})$, which is the optimal rate specified in Theorem 3 of Bai (2003) when dealing with the traditional approximate factor models.*
(ii) It is a common issue to select the number of principal components in the literature and there are many possible approaches available. Since it is impossible to eliminate all the noise effects in recovering the factors and we only need to guarantee that the diverging part of the noises is removed for large p_1, we may select k_1 in a range of possible values. In practice, let $\widehat{\mu}_{1,1} \geq ... \geq \widehat{\mu}_{1,p_1}$ be the sample eigenvalues of $\widehat{\mathbf{S}}_1$ and define $\widehat{k}_{1,L}$ as

$$\widehat{k}_{1,L} = \arg\min_{1\leq j\leq k_{1,U}}\{\widehat{\mu}_{1,j+1}/\widehat{\mu}_{1,j}\}, \tag{3.5}$$

and $\widehat{k}_{1,U}$ is a pre-specified integer. We suggest $\widehat{k}_{1,U} = \min\{\sqrt{p_1}, \sqrt{n}, p_1 - \widehat{r}_1, 5\}$. Then the estimator $\widehat{k}_1$ for k_1 can assume some value between $\widehat{k}_{1,L}$ and $\widehat{k}_{1,U}$. We can select $\widehat{k}_2$ in a similar manner.

Next, we study the consistency of the white noise tests described in Section 2. In fact, the consistency conditions depend on the test statistic used. We only consider the two test statistics T_n and $T(m)$ of Section 2.5. and present the consistency when p_1 and p_2 are large since the case of small p_1 and p_2 is trivial. For any random vector $\mathbf{x}_t$ to be sub-Gaussian we mean there exists a constant $C > 0$ such that $P(|\mathbf{v}'(\mathbf{x}_t - E\mathbf{x}_t)| > x) \leq C\exp(-Cx^2)$ for any constant vector $\|\mathbf{v}\|_2 = 1$. We need an additional assumption.

Assumption 6. vec($\mathbf{F}_t$), vec($\mathbf{Z}_{12,t}$), vec($\mathbf{Z}_{21,t}$), *and* vec($\mathbf{Z}_{22,t}$) *are sub-Gaussian random vectors.*

Theorem 4. *Assume Assumptions 1–6 hold.*
(i) If $p_1p_2 = o\left\{\min\left(n^{\frac{2}{1+3\delta_1}}, n^{\frac{1}{1+2\delta_1-\delta_2}}\right)\right\}$, then the test statistic T_n can consistently estimate r_1 and r_2, i.e. $P(\widehat{r}_1 = r_1, \widehat{r}_2 = r_2) \to 1$ as $n \to \infty$.
(ii) If $p_1^{1+\delta_1-\delta_2/2}p_2^{1+\delta_1-\delta_2/2}n^{-1/2}\sqrt{\log(np_1p_2)} = o(1)$, then the test statistic $T(m)$ can consistently estimate r_1 and r_2.

With the estimator $\widehat{r}_1$, we may estimate $\mathbf{A}_1$ by $\widehat{\mathbf{A}}_1 = (\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_{\widehat{r}_1})$, where $\widehat{\mathbf{a}}_1, ..., \widehat{\mathbf{a}}_{\widehat{r}_1}$ are the orthonormal eigenvectors of $\widehat{\mathbf{M}}_1$, defined in (2.13), corresponding to the $\widehat{r}_1$ largest eigenvalues. In addition, we may also replace r_1 by $\widehat{r}_1$ in the entire methodology described in Section 2. We can define $\widehat{\mathbf{P}}_1$ in a similar way.

4. Numerical Properties

4.1. Simulation

We illustrate the finite-sample properties of the proposed methodology under different choices of p_1 and p_2. Because the actual dimension p_1p_2 can easily go to hundreds for even relatively small p_1 and p_2, we focus on the high-dimensional case, which is of interest. As the dimensions of $\widehat{\mathbf{A}}_1$ and $\mathbf{A}_1$ are not necessarily the same, and $\mathbf{L}_1$ is not an orthogonal matrix in general, we first extend the discrepancy measure in Equation (3.1) to a more general form below. Let $\mathbf{H}_i$ be a $p \times h_i$ matrix with $\text{rank}(\mathbf{H}_i) = h_i$, and $\mathbf{P}_i = \mathbf{H}_i(\mathbf{H}_i'\mathbf{H}_i)^{-1}\mathbf{H}_i'$, for $i = 1, 2$. Define

$$\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \sqrt{1 - \frac{1}{\max(h_1, h_2)}\text{tr}(\mathbf{P}_1\mathbf{P}_2)}. \tag{4.1}$$

Then $\bar{D} \in [0, 1]$. Furthermore, $\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = 0$ if and only if either $\mathcal{M}(\mathbf{H}_1) \subset \mathcal{M}(\mathbf{H}_2)$ or $\mathcal{M}(\mathbf{H}_2) \subset \mathcal{M}(\mathbf{H}_1)$, and it is 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. When $h_1 = h_2 = h$ and $\mathbf{H}_i'\mathbf{H}_i = \mathbf{I}_r$, $\bar{D}(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2))$ reduces to that in Equation (3.1). We only present the simulation results for $k_0 = 2$ in Equation (2.13) to save space because other choices of k_0 produce similar patterns.

Example 1. Consider model (2.1) with common factors satisfying

$$\mathbf{F}_t = \mathbf{\Phi}\mathbf{F}_{t-1}\mathbf{\Psi}' + \mathbf{N}_t,$$

where $\mathbf{N}_t$ is a matrix-variate white noise process with independent entries, $\mathbf{\Phi} \in \mathbb{R}^{r_1 \times r_1}$ and $\mathbf{\Psi} \in \mathbb{R}^{r_2 \times r_2}$ are two diagonal coefficient matrices. We set the true dimension of the matrix factor to $(r_1, r_2) = (2, 3)$, the orders of the diverging noise components $(k_1, k_2) = (1, 2)$ as defined in Assumption 4, the dimensions $(p_1, p_2) = (7, 7)$, $(10, 15)$, $(20, 20)$ and $(20, 30)$, and the sample sizes are $n = 300$, 500, 1000, 1500, 3000. We consider three scenarios for δ_1 and δ_2: $(\delta_1, \delta_2) = (0, 0.9)$, $(0.2, 0.8)$ and $(0.5, 0.5)$. We obtain similar results for other settings, but omit the details to save space. For each scenario mentioned above, the elements of $\mathbf{L}$ and $\mathbf{R}$ are drawn independently from $U(-2, 2)$, and then we divide $\mathbf{L}_1$ ($\mathbf{R}_1$) by $p_1^{\delta_1/2}$ ($p_2^{\delta_1/2}$), the first k_1 (k_2) columns of $\mathbf{L}_2$ ($\mathbf{R}_2$) by $p_1^{\delta_2/2}$ ($p_2^{\delta_2/2}$) and the remaining $v_1 - k_1$ ($v_2 - k_2$) columns by p_1 (p_2) to satisfy Assumptions 3 and 4. $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are diagonal matrices with their diagonal elements drawn independently from $U(0.5, 0.9)$, $\text{vec}(\mathbf{Z}_{12,t}) \sim N(0, \mathbf{I}_{r_1v_2})$, $\text{vec}(\mathbf{Z}_{21,t}) \sim N(0, \mathbf{I}_{v_1r_2})$, $\text{vec}(\mathbf{Z}_{22,t}) \sim N(0, \mathbf{I}_{v_1v_2})$, $\text{vec}(\mathbf{N}_t) \sim N(0, \mathbf{I}_{r_1r_2})$. We use 500 replications in each experiment.

We first study the performance of estimating the dimension of the matrix-variate factors. For simplicity, we only report the results of the test statistic $T(m)$ with $m = 10$, defined in Section 2, and the results for the other test are similar. When $p_1p_2 > n$, we only keep the upper $\varepsilon\sqrt{n}$ row- and column-transformed series of $\widehat{\mathbf{\Gamma}}_1'\mathbf{Y}_t\widehat{\mathbf{\Gamma}}_2$ with $\varepsilon = 0.9$ in the testing. Similar results are obtained for other choices of ε, but we do not report them here. The testing results are given in Table 1. From

Table 1: Empirical probabilities $P(\widehat{r}_2 = r_1, \widehat{r}_2 = r_2)$ for Example 1 with $(r_1, r_2) = (2, 3)$ and $(k_1, k_2) = (1, 2)$, where (p_1, p_2) and n are the dimension and the sample size, respectively. δ_1 and δ_2 are the strength parameters of the factors and the errors, respectively. 500 iterations are used.

			n				
(δ_1, δ_2)	(p_1, p_2)	p_1p_2	300	500	1000	1500	3000
(0,0.9)	$(7, 7)$	49	0.956	0.982	0.984	0.980	0.976
	$(10, 15)$	150	0.930	0.988	0.964	0.984	0.978
	$(20, 20)$	400	0.818	0.976	0.962	0.970	0.968
	$(20, 30)$	600	0.834	0.986	0.976	0.964	0.972
(0.2,0.8)	$(7, 7)$	49	0.848	0.992	0.988	0.972	0.986
	$(10, 15)$	150	0.882	0.982	0.974	0.978	0.984
	$(20, 20)$	400	0.742	0.964	0.972	0.982	0.972
	$(20, 30)$	600	0.816	0.994	0.976	0.968	0.966
(0.5,0.5)	$(7, 7)$	49	0.104	0.438	0.950	0.974	0.972
	$(10, 15)$	150	0.304	0.710	0.946	0.974	0.980
	$(20, 20)$	400	0.028	0.074	0.334	0.696	0.980
	$(20, 30)$	600	0.020	0.080	0.296	0.636	0.938

the table, we see that for each setting of (δ_1, δ_2) and fixed (p_1, p_2), the performance of the white noise test improves as the sample size increases. The performance is also quite satisfactory for moderately large p_1p_2 when the factor strength is stronger than that of the noises. When $(\delta_1, \delta_2) = (0.5, 0.5)$, we see that the test statistic fares poorly for small sample sizes when the dimension is high, which is understandable since, in this case, the factors and the noises have the same level of strength while the diverging noise effect is much more prominent by Equation (2.1). However, the performance of the test statistic improves significantly when the sample size increases.

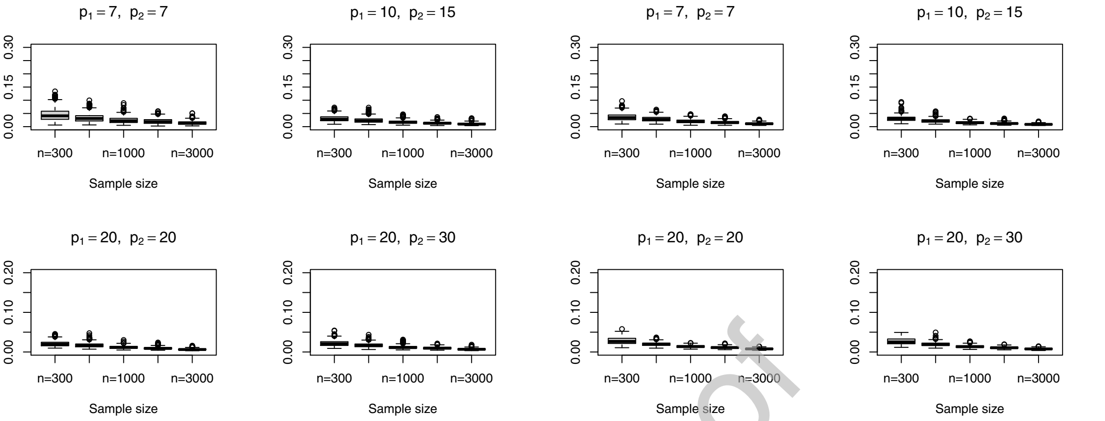
Next, we study the accuracy of the estimated loading matrices. The boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$ and $\bar{D}(\mathcal{M}(\widehat{\mathbf{P}}_1), \mathcal{M}(\mathbf{R}_1))$ are shown in Figures 1(a) and (b), respectively. From Figure 1, we see that the estimation accuracy of the loading matrix improves as the sample size increases even for moderately large p_1p_2, which is in line with our asymptotic theory. We then study the estimation accuracy of the estimated factor process by

$$D(\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}\widehat{\mathbf{P}}_1', \mathbf{L}_1\mathbf{F}\mathbf{R}_1') = \frac{1}{n\sqrt{p_1p_2}} \sum_{t=1}^{n} \|\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}_t\widehat{\mathbf{P}}_1' - \mathbf{L}_1\mathbf{F}_t\mathbf{R}_1'\|_2. \tag{4.2}$$

The results are shown in Figure 2, from which we see that, for fixed (p_1, p_2), the estimation accuracy also improves as the sample size increases. This result is consistent with our Theorem 3 of Section 3.

To see the advantages of the proposed method, we compare it with that of Wang et al. (2019) (denoted by WLC) in selecting the dimension of the matrix-variate factors. For the ratio-based method in WLC, let $\widehat{\lambda}_{i,1}, ..., \widehat{\lambda}_{i,p_i}$ be the eigenvalues (in decreasing order) of $\widehat{\mathbf{M}}_i$, for $i = 1, 2$, define

$$\widehat{r}_i = \arg \min_{1 \le j \le p_i/2} \{\widehat{\lambda}_{i,j+1}/\widehat{\lambda}_{i,j}\}, \ i = 1, 2. \tag{4.3}$$



(a)

(b)

Figure 1: (a) Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{A}}_1), \mathcal{M}(\mathbf{L}_1))$; (b) Boxplots of $\bar{D}(\mathcal{M}(\widehat{\mathbf{P}}_1), \mathcal{M}(\mathbf{R}_1))$. We set $(r_1, r_2) = (2, 3)$, $(k_1, k_2) = (1, 2)$, and $(\delta_1, \delta_2) = (0, 0.9)$ in Example 1. The sample sizes used are $300, 500, 1000, 1500, 3000$, respectively, and the number of iterations is 500.

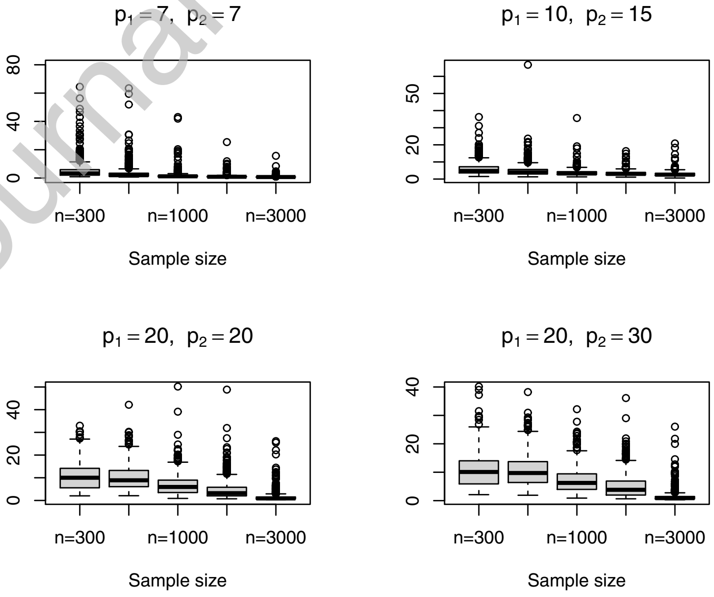
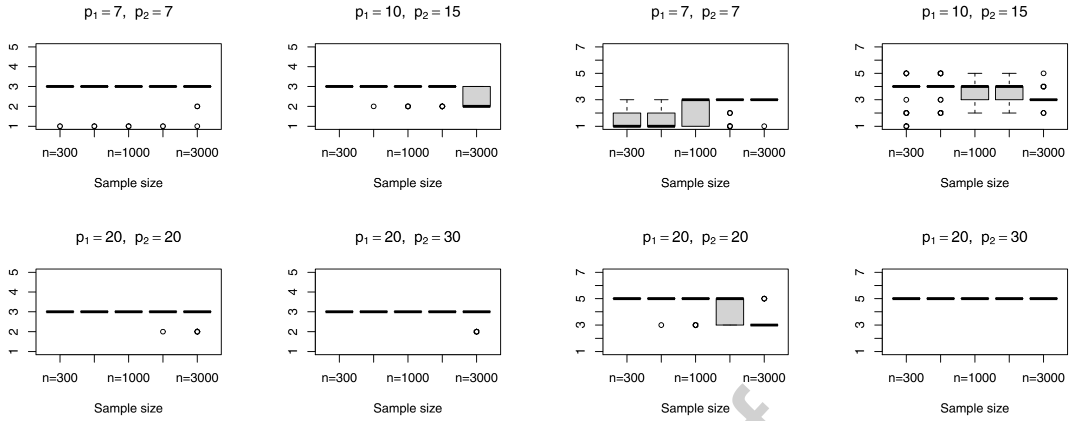


Figure 2: Boxplots of $D(\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}\widehat{\mathbf{P}}_1, \mathbf{L}_1\mathbf{F}\mathbf{R}_1')$ defined in (4.2) when $(r_1, r_2) = (2, 3)$, $(k_1, k_2) = (1, 2)$, and $(\delta_1, \delta_2) = (0, 0.9)$ in Example 1. The sample sizes used are $300, 500, 1000, 1500, 3000$, respectively, and the number of iterations is 500.



(a) (b)

Figure 3: (a) Boxplots of $\widehat{r}_1$ by the ratio-based method of Wang et al. (2019); (b) Boxplots of $\widehat{r}_2$ by the ratio-based method of Wang et al. (2019). We set $(r_1, r_2) = (2, 3)$, $(k_1, k_2) = (1, 2)$, and $(\delta_1, \delta_2) = (0.5, 0.5)$ in Example 1. The sample sizes used are $300, 500, 1000, 1500, 3000$, respectively and the number of iterations is 500.

Figures 3(a)-(b) present the boxplots of $\widehat{r}_1$ and $\widehat{r}_2$, respectively. We see, from the plots, that the estimated number of factors $\widehat{r}_i$ tends to be the sum of the number of common factors r_i and the number of spiked components of the noises k_i in most scenarios. The result indicates that the ratio-based method of Wang et al. (2019) may fail to identify the correct dimension of the matrix-variate factor process with dynamical dependence if the covariance of the noises has diverging eigenvalues, whereas the proposed white noise test continues to work well, as shown in Table 1.

Finally, we compare our method with the one of Wang et al. (2019) in recovering the common factors since a key difference between the two methods is that we allow some of the eigenvalues of the noise covariance to diverge. We denote our method by GT and the results are reported in Table 2 for $(r_1, r_2) = (2, 3)$, $(k_1, k_2) = (1, 2)$, and $(\delta_1, \delta_2) = (0.5, 0.5)$. From the table, we see that the estimation error of our method is much smaller than that based on the method of WLC. This is understandable because the ratio-based method tends to overestimate the dimension of the common factors. In addition, for a given (p_1, p_2), the estimation error of our method tends to decrease as the sample size increases, which is in agreement with our asymptotic theory. Overall, under the assumption that the noise effect is prominent, the proposed method outperforms the existing one in the literature.

4.2. Real Data Analysis

Example 2. In this example, we use the Fama-French return series to illustrate applications of the proposed method. The data contain monthly returns of 100 portfolios, structured in a 10 by 10 matrix according to ten levels of market capitalization (Size, in rows, from small to large) and ten levels of investment (Inv, in columns, from low to high). Both size and investment are factors for average stock returns considered in Fama and French (2015). The return series spans from July 1963

Table 2: The $D(\widehat{\mathbf{A}}_1\widehat{\mathbf{X}}\widehat{\mathbf{P}}_1', \mathbf{L}_1\mathbf{F}\mathbf{R}_1')$ defined in (4.2) when $(r_1, r_2) = (2, 3)$, $(k_1, k_2) = (1, 2)$, and $(\delta_1, \delta_2) = (0.5, 0.5)$ in Example 1. The sample sizes used are $n = 300, 500, 1000, 1500, 3000$. Standard errors are given in the parentheses and 500 iterations are used. GT denotes the proposed method and WLC is the one in Wang et al. (2019).

		n				
(p_1, p_2)	Method	300	500	1000	1500	3000
$(7, 7)$	GT	0.862(0.199)	0.726(0.494)	0.460(0.164)	0.447(0.320)	0.416(0.096)
	WLC	1.178(0.036)	1.178(0.029)	1.182(0.026)	1.183(0.025)	1.179(0.047)
$(10, 15)$	GT	0.652(0.354)	0.403(0.185)	0.290(0.310)	0.254(0.208)	0.229(0.132)
	WLC	0.891(0.060)	0.886(0.066)	0.862(0.084)	0.783(0.148)	0.549(0.165)
$(20, 20)$	GT	0.530(0.167)	0.437(0.108)	0.301(0.147)	0.191(0.126)	0.103(0.043)
	WLC	0.696(0.011)	0.695(0.010)	0.686(0.036)	0.648(0.072)	0.485(0.117)
$(20, 30)$	GT	0.485(0.181)	0.394(0.102)	0.278(0.122)	0.181(0.124)	0.098(0.069)
	WLC	0.662(0.010)	0.663(0.008)	0.662(0.005)	0.662(0.005)	0.651(0.044)

to December 2019 and consists of 678 monthly observations for each individual process. Therefore, the series forms a $10 \times 10 \times 678$ tensor-valued data set. The data and relevant information are available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

Following Sharpe (1964) and Fama and French (2015), we adjust each of the return series by subtracting the corresponding risk-free asset returns, which are also available from the above website. The missing values were imputed by a simple exponential smoothing method. Time plots of the adjusted 10×10 series are shown in Figure 4 with $p_1 = p_2 = 10$ and $n = 678$.

We first applied the method of Section 2.4 with $k_0 = 2$ to the data and found that the test statistic $T(10) = 4.85$ for testing the null hypothesis $H_0(2)$ defined in Section 2.5. This statistic exceeds the critical value 4.81 based on the limiting Gumbel distribution in Tsay (2020) with $\alpha = 0.05$. But the null hypothesis is not rejected if we increase the order in either the column or row direction of the testing procedure. Therefore, $\widehat{r}_1 = 2$ and $\widehat{r}_2 = 2$, implying that a 2×2 matrix-variate latent factor process is detected. The estimated front and back loading matrices after being multiplied by 30 are reported in Table 3, which has several implications. First, for Size, it seems that the 10 rows of the portfolios can be divided into two or three groups. The one with the smallest size (corresponding to S1) depends on both the first and the second factors heavier than the others, the second smallest size portfolio depends more on the first row factor and less on the second one, and the 3rd to the 10th size portfolios have similar dependence on both the first and the second rows of the matrix-variate factors. Second, for Investment, all the portfolios have similar dependence on the first column of the factor matrix, and the dependence on the second columns seems to have three groups; the lowest investment portfolio (corresponding to Inv1) seems to depend heavily on the second row of the factors, the 5th to the 8th and the 10-th investment portfolios have similar dependence on the first and the second rows of the factors, whereas the 2nd to the 4th and the 9th investment portfolios depend more on the first column of the factors. In addition, the signs of the first coefficients of the size loading and the investment loading are the same, which implies that each return series shares a co-movement with respect to the $[1, 1]$-factor series. This is understandable since we can treat this common factor as representing the market factor of the capital asset pricing model (CAPM) of Sharpe (1964). The

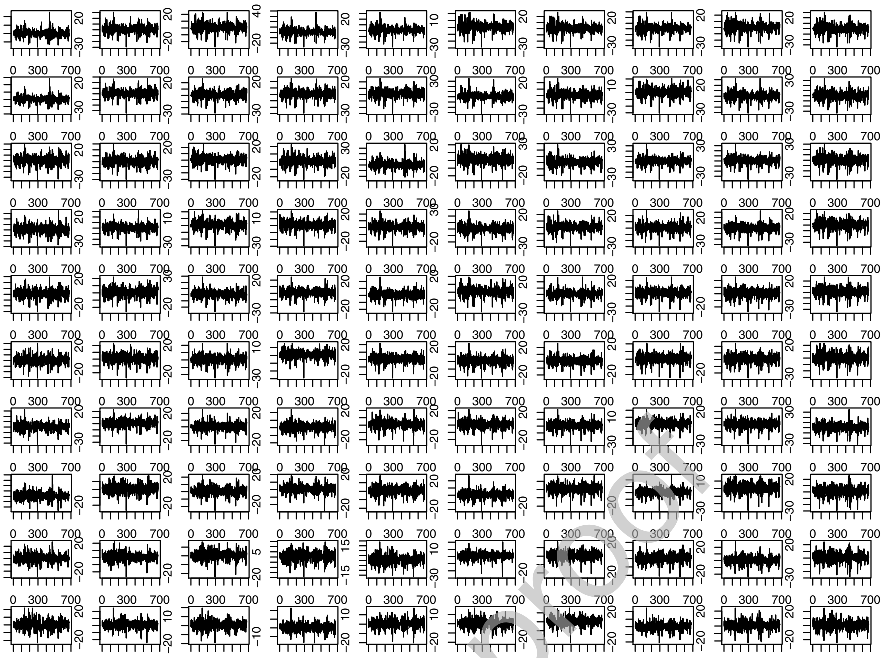


Figure 4: Time series plots of Fama-French 10 by 10 monthly excess return series based on Size and Investment from July 1963 to December 2019.

Table 3: Fama-French series: Size and Investment (Inv) loading matrices after being multiplied by 30. The two-dimensional loading vectors are ordered via sizes (S1–S10) and Investment (Inv1–Inv10) from small to large and from low to high, respectively.

Size Factor	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Row 1	-18	-12	-10	-9	-8	-7	-7	-6	-5	-3
Row 2	21	1	-6	-6	-6	-10	-7	-8	-8	-7

Inv Factor	Inv1	Inv2	Inv3	Inv4	Inv5	Inv6	Inv7	Inv8	Inv9	Inv10
Column 1	-11	-10	-9	-8	-8	-8	-9	-9	-10	-12
Column 2	27	-3	-1	1	-6	-6	-5	-5	1	-7

product of the first coefficients of the size loading and the investment loading can be treated as a market beta, even though the starting point of our approach is different from that of CAPM. The usefulness of the detected market factor, however, deserves a further investigation.

To obtain the extracted factors, by the two-way projected PCA of Sections 2.3 and 2.4, we first examine the eigenvalues of the sample covariance matrices $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$. From Figure 5, we see that the largest eigenvalues of $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$ are much larger than the others. Therefore, we choose $\widehat{k}_1 = \widehat{k}_2 = 1$, and the recovered matrix-variate factors are shown in Figure 6(a) as well as their corresponding spectrum densities in Figure 6(b). From Figure 6, we see that there are three series with non-trivial spectra, which are all dynamically dependent and they capture most of the dynamic information of the data, and the $[1, 2]$-factor appears to be not serially correlated by itself because its sample spectrum is flat. However, this does not imply that the $[1, 2]$-factor captures no dynamic information in the

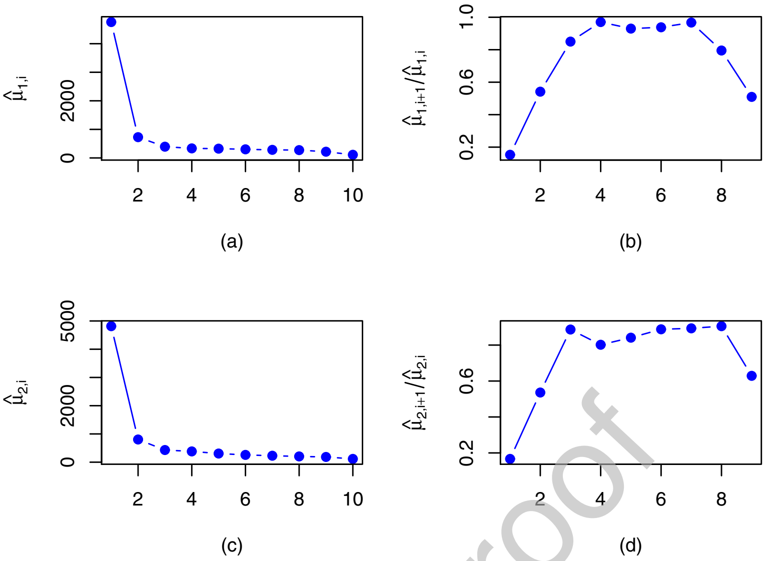


Figure 5: (a) The 10 eigenvalues of $\widehat{\mathbf{S}}_1$; (b) Plot of the ratios of consecutive eigenvalues of $\widehat{\mathbf{S}}_1$; (c) The 10 eigenvalues of $\widehat{\mathbf{S}}_2$; (d) Plot of the ratios of consecutive eigenvalues of $\widehat{\mathbf{S}}_2$

detected matrix-variate common factors. For example, the lag-1 cross-correlation between the $[1,2]$-factor and the $[2,2]$-factor is 0.08. If we test for the zero lag-1 cross correlation between these two series using the long-run covariance matrix calculated by the method in Andrews (1991), the p-value is 0.038, implying that the two factors are lag-1 cross-correlated at the 5% level. Therefore, the detected 2-by-2 matrix-variate common factor process does not violate the assumptions of the proposed model.

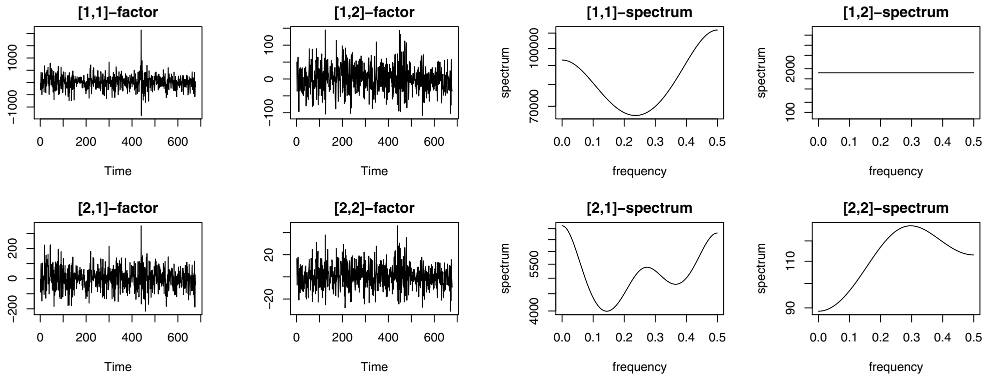
Next we examine and compare the forecasting performance of the extracted factors via the proposed method (denoted by GT) and those by Wang et al. (2019) (denoted by WLC). We estimate the models using the data in the time span $[1,\tau]$ with $\tau = 558, ..., 678-h$ for the h-step ahead forecasts, i.e., we use returns of the last ten years for out-of-sample forecasts. For the method of Wang et al. (2019), the estimated dimension of the matrix-variate factor is $(\widehat{r}_1, \widehat{r}_2) = (1,1)$. For simplicity, we employ a simple AR(1) model for each detected common factor to produce forecasts. We also fit a scalar AR(1) (denoted by SAR) model to each individual return series as a benchmark approach in the out-of-sample forecasting comparison. The following two criteria are used to measure the forecast errors:

$$\mathrm{FE}_F(h) = \frac{1}{120-h+1}\sum_{\tau=558}^{678-h}\frac{1}{\sqrt{p_1p_2}}\|\widehat{\mathbf{Y}}_{\tau+h}-\mathbf{Y}_{\tau+h}\|_F, \tag{4.4}$$

and

$$\mathrm{FE}_2(h) = \frac{1}{120-h+1}\sum_{\tau=558}^{678-h}\frac{1}{\sqrt{p_1p_2}}\|\widehat{\mathbf{Y}}_{\tau+h}-\mathbf{Y}_{\tau+h}\|_2, \tag{4.5}$$

where $p_1 = p_2 = 10$. Table 4 reports the 1-step to 4-step ahead forecast errors of Equations (4.4)



(a)

(b)

Figure 6: (a) The time series plots of the extracted 2 × 2 common factors; (b) the corresponding spectrum densities of the factor processes

.

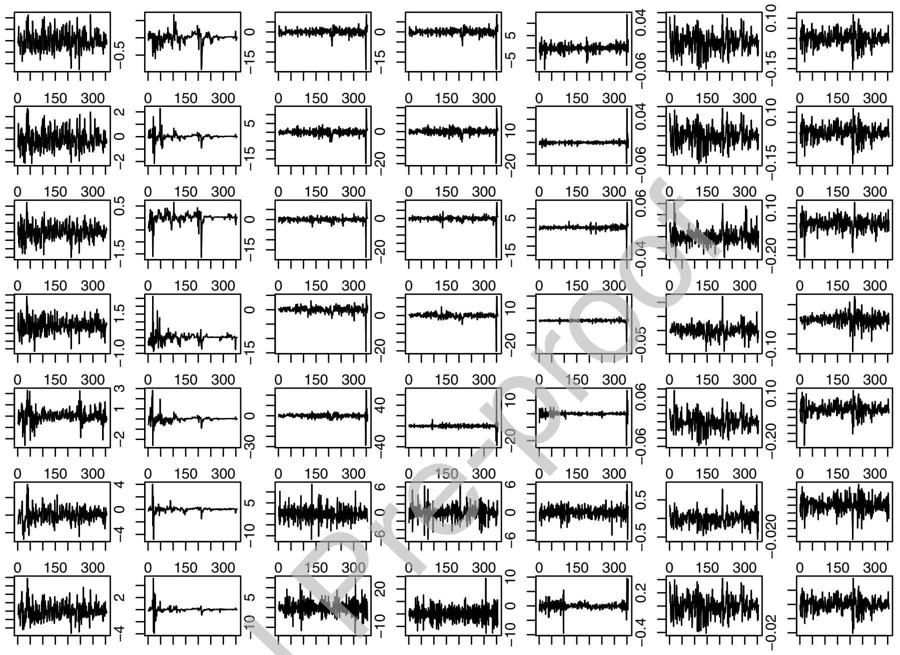
and (4.5) for the three methods GT, WLC, and SAR. The smallest forecast error of each step is shown in boldface. From the table, we see that our proposed method is capable of producing accurate forecasts and the associated forecast errors based on the extracted factors by our method are smaller than those based on the factor extracted by WLC or the benchmark approach SAR. The difference in forecasting errors between the three methods used in Table 4 is small. But it is generally not easy to produce accurate forecasts in asset returns and the improvements by our proposed method could have important implications to practitioners, especially over the ten-year horizon.

Table 4: The 1-step to 4-step ahead out-of-sample forecast errors of various methods for Example 2. GT denotes the proposed method, WLC denotes the forecasting errors based on the extracted factor by the method in Wang et al. (2019), and SAR denotes a scalar AR model to each individual return series. Boldface numbers denote the smallest error for a given forecast horizon.

	$\mathrm{FE}_F(h)$			$\mathrm{FE}_2(h)$		
Step-h	GT	WLC	SAR	GT	WLC	SAR
1	**4.51**	4.61	4.60	**4.04**	4.14	4.14
2	**4.47**	4.51	4.60	**3.98**	4.04	4.14
3	**4.48**	4.51	4.60	**4.00**	4.04	4.14
4	**4.47**	4.49	4.57	**3.98**	4.01	4.11

Example 3: In this application, we apply the proposed method to a multinational macroeconomic data set. We consider seven monthly macroeconomic indices of seven countries. The indices used are: (1) Long-Term Government Bond Yields: 10-year: Main (Including Benchmark) (LTBY); (2) 3-Month or 90-day Rates and Yields: Interbank Rates (IBR); (3) Production of Total Industry (PTI); (4) Production in Total Manufacturing (PTM); (5) Total Retail Trade (TRT); (6) National Currency to US Dollar Exchange Rate: Average of Daily Rates (CER); (7) US Dollar to National Currency Spot Exchange Rate (CSER), and the countries are: (1) Germany (DE); (2) France (FR); (3) United Kingdom (GB); (4) Canada (CA); (5) Italy (IT); (6) Norway (NO); and (7) Denmark (DK). The data

set is obtained from the Economic database of U.S. Federal Reserve Bank of St. Louis, and it spans from March 1, 1991 to October 1, 2020 consisting of 356 monthly observations. We selected this time span, because there are missing values in other time intervals. To remove any possible trends in the series, we take the first difference of the 7 × 7 matrix-variate data. The resulting series are shown in Figure 7 with $p_1 = p_2 = 7$ and $n = 355$.



(a)

Figure 7: Time series plots of the 7 × 7 first-differenced macroeconomic variables from March 1, 1991 to October 1, 2020. Rows (countries):(1) Germany (DE); (2) France (FR); (3) United Kingdom (GB); (4) Canada (CA); (5) Italy (IT); (6) Norway (NO); and (7) Denmark (DK). Columns (indices):(1) Long-Term Government Bond Yields: 10-year: Main (Including Benchmark) (LTBY); (2) 3-Month or 90-day Rates and Yields: Interbank Rates (IBR); (3) Production of Total Industry (PTI); (4) Production in Total Manufacturing (PTM); (5) Total Retail Trade (TRT); (6) National Currency to US Dollar Exchange Rate: Average of Daily Rates (CER); (7) US Dollar to National Currency Spot Exchange Rate (CSER).

Following the proposed modeling procedure with the same chosen parameters as those of Example 2, we applied the white-noise tests to identify the dimension of the matrix-variate factors and found that $(\widehat{r}_1, \widehat{r}_2) = (4, 4)$, which implies that a 4 × 4 matrix-variate latent factor process is detected. The estimated front and back loading matrices after being multiplied by 20 are reported in Table 5. From the table, we have several observations. For the country factors, we see that they can roughly be divided into two or three groups. The indices of DE, FR, GB, and CA depend less on the first and the second factors, but more on the third and the fourth ones. The indices of IT and DK depend more on the first two factors, but less on the last two factors. Finally, the indices of NO have a similar dependence on all the four factors. For the index factors, we see that they can also roughly be divided into two or three groups. The LTBY, IBR, and CER of all 7 countries have weak dependence on the first three column factors, but have heavier dependence on the fourth one, while the PTI, PTM, and

Table 5: Multinational macroeconomic indices: Country and index loading matrices after being multiplied by 20.

Country Factor	DE	FR	GB	CA	IT	NO	DK
Row 1	1	-1	2	1	-15	0	-13
Row 2	3	3	4	4	-12	1	15
Row 3	-6	-12	-8	-11	-5	-1	4
Row 4	11	10	-6	11	-1	-4	0

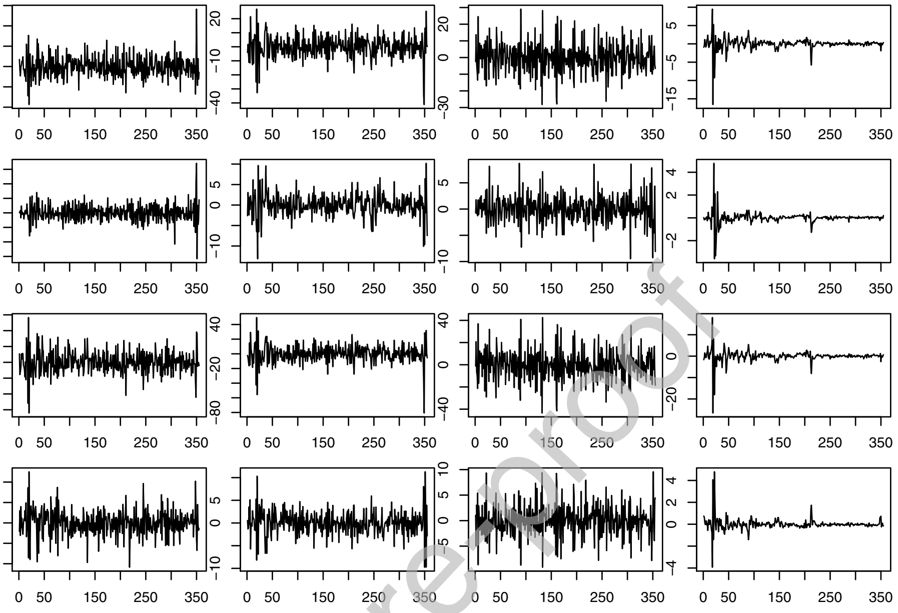
Index Factor	LTBY	IBR	PTI	PTM	TRT	CER	CSER
Column 1	0	0	-4	-19	-3	0	0
Column 2	0	0	-8	5	-18	0	0
Column 3	0	0	18	-3	-8	0	0
Column 4	6	19	0	0	0	-2	0

TRT have heavier dependence on the first three factors but weak dependence on the last one. The CSER of all countries have little dependence on all the factors.

Next, we apply the two-way projected PCA to obtain the extracted factors. Similarly to that of Example 2, we first examine the eigenvalues of the sample covariance matrices $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$, and find that the first eigenvalue of $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$ are much larger than the others, and therefore, we choose $\widehat{k}_1 = \widehat{k}_2 = 1$. The recovered matrix-variate factors are shown in Figure 8, which can be used for out-of-sample forecasting or to study the dynamic dependence of the original macroeconomic processes. In this particular instance, the proposed method reduces the dimensions from 49 to 16, making it easier to pursue any further investigation of the data.

Similarly to that in Example 2, we also examine and compare the forecasting performance of the extracted factors via the proposed method (denoted by GT) and those by Wang et al. (2019) (denoted by WLC). For the method of WLC, the estimated dimension of the matrix-variate factors is $(\widehat{r}_1, \widehat{r}_2) = (3, 3)$. We use the observed data of the last 20 months to do out-of-sample testing and the fitted models for the factors are the same as those in Example 2. We use the forecast errors in (4.4)–(4.5), where $p_1 = p_2 = 7$, the number of rolling-windows is 20, and the summation is summing from $\tau = 335$ to $355 - h$. From Table 6, we see that our proposed method is capable of producing accurate forecasts and the associated forecast errors based on the extracted factors are smaller than those based on factors extracted by WLC, indicating that our proposed method provides another useful approach to practitioners who are interested in out-of-sample forecasting.

In summary, for the monthly excess return series and the multinational macroeconomic indices considered, the proposed method not only produced interpretable factors, but also improved on out-of-sample forecasts. We like to emphasize that the proposed method is different from the traditional factor model analysis, especially those based on the conventional principal component analysis, which focuses on the variance decomposition rather than the dynamical dependence of the series. The proposed model explores the underlying structure of the data via a two-way transformation. Finally, one can further improve the out-of-sample forecasting if he/she adopts some regularization method to



(a)

Figure 8: The time series plots of the extracted 4 × 4 common factors.

Table 6: The 1-step to 4-step ahead out-of-sample forecast errors of various methods for Example 3. GT denotes the proposed method, and WLC denotes the forecasting errors based on the extracted factor by the method in Wang et al. (2019). Boldface numbers denote the smallest error for a given forecast horizon.

	$\mathrm{FE}_F(h)$			$\mathrm{FE}_2(h)$	
Step h	GT	WLC		GT	WLC
1	**3.37**	3.42		**3.08**	3.17
2	**3.44**	3.68		**3.14**	3.34
3	**3.55**	3.70		**3.25**	3.41
4	**3.70**	3.82		**3.37**	3.51

increase the accuracy in parameter estimation, especially when the dimension (r_1, r_2) of the common factors is high.

5. Discussion and Concluding Remarks

This paper proposed a new approach to analyze high-dimensional, dynamically dependent matrix-variate data in the presence of prominent noise effect. The proposed approach is an extension of that for vector time series in Tiao and Tsay (1989) and Gao and Tsay (2021b). The approach not only can reduce the dimension of the matrix-variate data, but also preserves the matrix structure to mitigate information loss or to reduce the number of parameters. The proposed approach is easy to implement for high-dimensional matrix-variate time series data and the empirical results obtained show that it can effectively extract the number of common factors from complex data. In addition, the extracted common factors could be useful in out-of-sample predictions.

It is worth mentioning that the Kronecker-product structure of the covariance matrix of a vectorized matrix-variate time series is used to reduce the number of parameters in the proposed model. In practice, this assumption can be inferred by a statistical test, which is referred to as a separability test of the variance-covariance structure in the literature; see an overview in Dutilleul (2018) and the references therein. However, most of the likelihood-ratio or Rao's score tests mentioned therein require that the data follow a multivariate normal distribution and the dimension of the series is not high. It is of interest to develop a proper test statistic to verify the separability of the variance-covariance matrix for high-dimensional time-series without the normality assumption. We leave details of such a test for future research.

Supplementary Material

The supplementary material contains all technical proofs of the theorems in Section 3.

Acknowledgments

We are grateful to the Editor, Associate Editor and the anonymous referees for their insightful comments and suggestions that have substantially improved the presentation and quality of the paper. This research is supported in part by the Booth School of Business, University of Chicago.

References

Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.

Bai J. (2003) Inferential theory for factor models of large dimensions. *Econometrica*, **71(1)**, 135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

Black, F. (1986). Noise. *The Journal of Finance*, **41(3)**, 528–543.

Box, G. E. P. and Tiao, G. C. (1977). A canonical analysis of multiple time series. *Biometrika*, **64**, 355–365.

Chang, J., Yao, Q. and Zhou, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika*, **104(1)**, 111–127.

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(530)**, 775–793.

Chen, R., Xiao, H., and Yang, D. (2020). Autoregressive models for matrix-valued time series. *Journal of Econometrics* (forthcoming).

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, **25(4)**, 1077–1096.

Dutilleul, P. (2018). Estimation and testing for separable variancecovariance structures. *Wiley Interdisciplinary Reviews: Computational Statistics*, **10(4)**, e1432.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, **116(1)**, 1–22.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society*, Series B, **75(4)**, 603–680.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000). Reference cycles: the NBER methodology revisited (No. 2400). Centre for Economic Policy Research.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, **100(471)**, 830–840.

Gao, Z., Ma, Y., Wang, H. and Yao, Q. (2019). Banded spatio-temporal autoregressions. *Journal of Econometrics,* **208(1)**, 211–230.

Gao, Z. and Tsay, R. S. (2019). A structural-factor approach for modeling high-dimensional time series and space-time data. *Journal of Time Series Analysis*, **40**, 343–362.

Gao, Z. and Tsay, R. S. (2021a). Modeling high-dimensional unit-root time series. *International Journal of Forecasting* (forthcoming).

Gao, Z. and Tsay, R. S. (2021b). Modeling high-dimensional time series: a factor model with dynamically dependent factors and diverging eigenvalues. *Journal of the American Statistical Association* (forthcoming).

Gao, Z. and Tsay, R. S. (2021c). Divide-and-conquer: a distributed hierarchical factor approach to modeling large-scale time series data. *arXiv preprint arXiv:2103.14626.*

Han, Y., Chen, R., Yang, D., and Zhang, C. H. (2020). Tensor factor model estimation by iterative projection. *arXiv preprint arXiv:2006.02611.*

Han, Y. and Tsay, R. S. (2020). High-dimensional linear regression for dependent data with applications to now-casting. *Statistica Sinica*, **30**, 1797–1827.

Hosking, J. R. (1980). The multivariate portmanteau statistic. *Journal of the American Statistical Association*, **75(371)**, 602–608.

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

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

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, **92(4)**, 1004–1016.

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95(2)**, 365–379.

Rogers, M., Li, L., and Russell, S. (2013). Multilinear dynamic systems for tensor time series. Conference for Neural Information Processing Systems, https://papers.nips.cc/paper/5117-multilinear-dynamical-systems-for-tensor-time-series.pdf.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, **19(3)**, 425–442.

Shen, D., Shen, H. and Marron, J. S. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research*, **17(150)**, 1–34.

Shojaie, A. and Michailidis, G. (2010). Discovering graphical Granger causality using the truncated lasso penalty. *Bioinformatics*, **26**, 517–523.

Song, S. and Bickel, P. J. (2011). Large vector auto regressions. Available at *arXiv:1106.3519.*

Stewart, G. W., and Sun, J. (1990). *Matrix Perturbation Theory.* Academic Press.

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.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. NBER Working Paper 11467.

Surana, A. Patterson, G., and Rajapakse, I. (2016). Dynamic tensor time series modeling and analysis. 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, 2016, pp. 1637-1642, doi: 10.1109/CDC.2016.7798500.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society,* **B51**, 157–213.

Tsay, R. S. (2014). *Multivariate Time Series Analysis.* Wiley, Hoboken, NJ.

Tsay, R. S. (2020). Testing for serial correlations in high-dimensional time series via extreme value theory. *Journal of Econometrics*, **216**, 106-117.

Walden, A. and Serroukh, A. (2002). Wavelet analysis of matrix-valued time series. *Proceedings: Mathematical, Physical and Engineering Sciences*, **458(2017)**, 157-179.

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

Wang, D., Zheng, Y., and Li, G. (2021). High-dimensional low-rank tensor autoregressive time series modeling. *arXiv: 2101.04276.*

Wang, D., Zheng, Y., Lian, H., and Li, G. (2021). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association* (forthcoming).

Werner, K., Jansson, M., and Stoica, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, **56(2)**, 478–491.