Statistica Sinica Preprint No: SS-2024-0341	
Title	Regularized Estimation of High-Dimensional Matrix-Variate Autoregressive Models
Manuscript ID	SS-2024-0341
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0341
Complete List of Authors	Hangjin Jiang, Baining Shen, Yuzhou Li and Zhaoxing Gao
Corresponding Authors	Zhaoxing Gao
E-mails	mazxgao@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Regularized Estimation of High-Dimensional Matrix-Variate Autoregressive Models

Hangjin Jiang[1], Baining Shen[1], Yuzhou Li[1], and Zhaoxing Gao[2*]

[1]*Center for Data Science, Zhejiang University*

[2]*School of Mathematical Sciences, University of Electronic Science and Technology of China*

Abstract: Matrix-variate time series data are increasingly popular in economics, statistics, and environmental studies, among other fields. The bilinear autoregressive structure is a popular modeling approach for such data, as it reduces model complexity while capturing dynamic interactions between rows and columns. However, in high-dimensional settings, the conventional iterated least-squares method requires estimating a large number of parameters, which hampers interpretability and scalability. To address this challenge, we propose regularized estimation procedures designed for settings in which the autoregressive coefficient matrices exhibit banded or sparse structures. Specifically, we introduce a Bayesian Information Criterion (BIC)-based approach to estimate the bandwidth in the banded case, and employ the LASSO technique for enforcing sparsity in the coefficient matrices. We derive asymptotic properties for both methods as the dimensions diverge and the sample size $T \to \infty$. Simulations and real data examples demonstrate the effectiveness of our methods, comparing their forecasting performance against common autoregressive models in the literature.

Key words and phrases: Matrix Time Series, High-dimension, Iterated Least-Squares, Band, Lasso

1 Introduction

In recent years, with the development of advanced information technologies, modern data collection and storage capabilities have led to massive amounts of time series data. Multiple and high-dimensional time series are routinely observed in a wide range of applications, includ-

*Corresponding author: zhaoxing.gao@uestc.edu.cn (Z. Gao), School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, 611731 P.R. China.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

ing economics, finance, engineering, environmental sciences, medical research, and others. In the past decades, various multivariate time series modeling methods have been studied in the literature. See Tsay (2014) and the references therein for details. Recently, large tensor (or multi-dimensional array) time series data have become increasingly popular in the literature across various fields, including those mentioned. For example, a group of countries will report a set of economic indicators each quarter, forming a matrix-variate (2-dimensional array) time series, with each column representing a country and each row representing an economic indicator. To analyze large and high-dimensional datasets, dimension-reduction techniques have gained popularity for achieving efficient and effective analysis of high-dimensional time series data. Examples include the canonical correlation analysis (CCA) of Box and Tiao (1977) and Gao and Tsay (2019), principal component analysis (PCA) of Stock and Watson (2002), the scalar component model of Tiao and Tsay (1989), and the factor model approach 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), and Gao and Tsay (2021, 2022, 2023b), among others. However, all the techniques developed for vector time series cannot be directly applied to matrix-variate time series, and simple vectorization of the matrix data often results in a significant number of estimated parameters, losing the original data structure. Therefore, further analysis methods should be developed to model such complex and dynamic datasets.

Recently, several methods have been developed for analyzing matrix- and tensor-variate time series data, including factor models in Wang et al. (2019), Chen et al. (2020), Yu et al. (2022), Gao and Tsay (2023a), Han et al. (2024), and Han et al. (2024), as well as the bilin-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

ear matrix-variate autoregressive model in Chen et al. (2021) and its extension to tensors in Li and Xiao (2021). To the best of our knowledge, only the method in Chen et al. (2021) can be directly applied for out-of-sample forecasting, while others primarily focus on dimension reduction of the matrix data structures. Although Chen et al. (2021) introduced effective techniques for estimating autoregressive coefficient matrices and explored their asymptotic properties, these methods are applicable only to matrix-variate time series data with fixed and small dimensions. Given that large-dimensional matrix-variate data are increasingly common in applications, the traditional iterated least-squares methods presented in Chen et al. (2021) may not perform well, and the theoretical results may no longer hold. Therefore, new estimation methods must be considered in such contexts.

This paper represents an extension of the bilinear matrix-variate autoregressive model developed in Chen et al. (2021) and the spatio-temporal data framework in Hsu et al. (2021). We focus on the scenario where the dimensions of matrix-variate data are growing, thereby extending the approach in Chen et al. (2021) to high-dimensional contexts. To facilitate meaningful dimension reduction, we recognize that each observed data point interacts only with a limited number of others. For instance, spatio-temporal data points, such as $PM_{2.5}$ observations, may rely primarily on a few neighboring locations. More generally, each observation may dynamically depend on only a subset of other components. Our goal is to identify sparse autoregressive matrices that allow for further dimensional reduction while maintaining interpretability.

In this paper, we propose two regularized estimation methods to reduce the model's dimen-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

sions further. The first method assumes that the autoregressive coefficient matrices are banded, indicating that each observed data point interacts only with a limited number of neighboring points. We introduce a two-step estimation approach: the first step utilizes traditional iterated least-squares to obtain initial estimates, while the second step employs a banded iterated least-squares method. Additionally, we propose using the Bayesian Information Criterion (BIC) to estimate the bandwidths of the coefficient matrices. The second method is similar but assumes that the autoregressive matrices are sparse, applying the LASSO technique for estimation. We derive the asymptotic properties of the proposed methods for diverging dimensions of the matrix-variate data as the sample size $T \to \infty$. Both simulated and real examples are used to evaluate the performance of our methods in finite samples, comparing them with commonly used techniques in the literature regarding the forecasting ability of autoregressive models.

This paper presents multiple contributions. First, the methods introduced in Chen et al. (2021) are applicable only to matrix-variate time series data with fixed and relatively small dimensions. We extend this model to a high-dimensional environment, offering a broader perspective on matrix-autoregressive models that is increasingly relevant for practitioners as such data become more common in applications. Second, coefficients obtained from traditional least-squares methods can be challenging to interpret due to the large number of parameters associated with higher dimensions. Our approaches, utilizing banded and general sparse structures, address this issue by facilitating meaningful dimensional reductions. The banded approach is particularly well-suited for analyzing spatio-temporal data, as the matrix structure corresponds to the locations of observations, making it reasonable to assume that each data

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

point depends dynamically on only a few neighboring points. The effectiveness of the banded structure has been demonstrated in Gao et al. (2019) and the references therein across various applications. Finally, we provide rigorous theoretical analysis, deriving the asymptotic properties of our proposed methods under these circumstances, thereby contributing to the theoretical foundation of this field.

The rest of the paper is organized as follows. We introduce the model and proposed estimation methodology in Section 2 and study the theoretical properties of the proposed model and its associated estimates in Section 3. Numerical studies with both simulated and real data sets are given in Section 4, and Section 5 provides some concluding remarks. All technical proofs are given in an online Supplementary Material. 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, $\|\mathbf{u}\|_\infty = \max_{1\le i\le p} |u_i|$ is the ℓ_∞-norm, and $\mathbf{I}_p$ denotes a $p \times p$ identity matrix. For a matrix $\mathbf{H} = (h_{ij})$, $\|\mathbf{H}\|_1 = \max_j \sum_i |h_{ij}|$, $\|\mathbf{H}\|_\infty = \max_i \sum_j |h_{ij}|$, $\|\mathbf{H}\|_F = \sqrt{\sum_{i,j} h_{ij}^2}$ 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 $b = O(a)$. $\mathbb{E}X$ denotes the expectation of random variable X, and $\mathbf{i}_k$ be the unit vector with the k-th element equal to 1. Finally, C is a constant having different values in different contexts.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

2 Model and Methodology

2.1 Setting

Let $\mathbf{Y}_t \in \mathbb{R}^{p_1 \times p_2}$ be an observable $p_1 \times p_2$ matrix-variate time series, we consider the matrix-variate autoregressive model of order $d \geq 1$ (MAR(d)) introduced by Chen et al. (2021) as follows:

$$\mathbf{Y}_t = \mathbf{A}_1 \mathbf{Y}_{t-1} \mathbf{B}_1' + ... + \mathbf{A}_d \mathbf{Y}_{t-d} \mathbf{B}_d' + \mathbf{E}_t, \tag{2.1}$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are the coefficient matrices, and $\mathbf{E}_t$ is a white noise term. By a similar argument as that in traditional AR models, we may let $\mathbf{A} = [\mathbf{A}_1, ..., \mathbf{A}_d]$, $\mathbf{B} = [\mathbf{B}_1, ..., \mathbf{B}_d]$, and $\mathbf{G}_t = \text{diag}(\mathbf{Y}_{t-1}, ..., \mathbf{Y}_{t-d})$, the regression part in Model (2.1) can be written as $\mathbf{A}\mathbf{G}_t\mathbf{B}'$. Thus, without loss of generality, we only consider the case when $d = 1$, i.e., study the following MAR(1) model:

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1}\mathbf{B}' + \mathbf{E}_t, \tag{2.2}$$

where $\mathbf{Y}_t \in \mathbb{R}^{p_1 \times p_2}$, $\mathbf{A} \in \mathbb{R}^{p_1 \times p_1}$, and $\mathbf{B} \in \mathbb{R}^{p_2 \times p_2}$ are the coefficient matrices, and $\mathbf{E}_t \in \mathbb{R}^{p_1 \times p_2}$ is the white noise term.

As discussed in Chen et al. (2021), the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are not uniquely defined due to the identification issue. For example, $(\mathbf{A}, \mathbf{B})$ can be replaced by $(c\mathbf{A}, \mathbf{B}/c)$ for some constant $c \neq 0$ without altering the equation in (2.2). Therefore, some identification conditions are required to impose on the coefficients. There are several ways to achieve this, for instance, we may assume $\|\mathbf{A}\|_F = 1$ and $\text{sign}(\text{tr}(\mathbf{A})) = 1$ as that in Hsu et al. (2021).

Chen et al. (2021) proposed three methods to estimate the coefficient matrices when the

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

dimensions p_1 and p_2 are fixed, which are (1) the Projection method, (2) Iterated least squares approach, and (3) Maximum likelihood estimation (MLE). Asymptotic properties of the estimators are also established therein. However, both their methods and asymptotic theory are derived under finite and fixed dimensions.

In this paper, we consider the estimation of the coefficient matrices in high-dimensional scenarios, i.e., $p_1, p_2 \to \infty$ as $T \to \infty$. It is widely known that traditional methods usually fail when the dimensions are growing as the sample size increases, and one of the main reasons is that there will be much more parameters to be estimated. Therefore, some interpretable structures are often imposed in a high-dimensional framework. Here, we present the estimation methods under two different cases: 1) the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are banded ones, and 2) $\mathbf{A}$ and $\mathbf{B}$ are sparse, where the banded coefficients matrices are often used in spatio-temporal data when the observed value of one location only depends on those of a few neighborhoods. On the other hand, in the second scenario, we assume only a small proportion of elements in $\mathbf{A}$ and $\mathbf{B}$ are non-zero, which serves as a general sparsity condition. Our goal is to estimate the coefficients $\mathbf{A}$ and $\mathbf{B}$ under such conditions in a high-dimensional framework. We will discuss these two scenarios in the following sections.

2.2 Estimation with the Banded Case

In this section, we consider the scenario that the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are banded ones with bandwidths $k_1 > 0$ and $k_2 > 0$, respectively. That is, we assume that

$$a_{i,j} = 0, b_{k,l} = 0 \text{ for all } |i-j| > k_1, |k-l| > k_2, \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

where $\mathbf{A} = (a_{i,j})_{i,j=1}^{p_1}$, $\mathbf{B} = (b_{k,l})_{k,l=1}^{p_2}$, and k_1 and k_2 are unknown bandwidths. Our goal is to estimate the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$, and their corresponding bandwidth parameters k_1 and k_2.

We first assume that k_1 and k_2 are known, and we will propose a BIC approach to consistently estimate them in subsection 2.2.2 below. For the estimation of the coefficient matrices with banded structures, the procedure can be carried out in a similar way as the iterated least-squares method in Chen et al. (2021). Specifically, we first obtain initial estimators of $\mathbf{A}$ and $\mathbf{B}$ by the iterated least squares method proposed in Chen et al. (2021). Then we start with the initial estimators, and perform another iterated least squares method to estimate the banded coefficient matrices. For example, we may estimate $\mathbf{A}$ and its bandwidths k_1 when the latest estimator for $\mathbf{B}$ is given, and then estimate $\mathbf{B}$ and its bandwidths k_2 by fixing the latest estimator for $\mathbf{A}$. We repeat this procedure and the algorithm stops when the estimators converge. A description of the algorithm is outlined in **Algorithm 1**. Details on Steps 2(a) and 2(b) in **Algorithm 1** are given in the following subsections.

Algorithm 1 Estimating algorithm for banded case

1. We use the iterated least-squares method in Chen et al. (2021) to obtain the estimators $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$ for $\mathbf{A}$ and $\mathbf{B}$, respectively. Denote the initial estimators as $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_0$ and $\widehat{\mathbf{A}}^{(0)} = \widehat{\mathbf{A}}_0$.
2. For the i-th iteration ($i = 1, 2, \cdots$),
   (a) Fix the estimator $\widehat{\mathbf{B}}^{(i-1)}$ of $\mathbf{B}$, the estimator $\widehat{\mathbf{A}}^{(i)}$ of $\mathbf{A}$ is obtained by applying the least-squares method to Model (2.2). The estimator of the unknown bandwidth k_1, denoted by $\widehat{k}_1^{(i)}$, is obtained based on a BIC given in section 2.2.2 below.
   (b) Fix the estimator $\widehat{\mathbf{A}}^{(i)}$ of $\mathbf{A}$, we estimate $\widehat{\mathbf{B}}^{(i)}$ and $\widehat{k}_2^{(i)}$ using the same procedure as that in (a).
   (c) The iteration stops if the convergence criterion is satisfied, otherwise we go to the next iteration and repeat Steps 2(a)–2(b).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Note that either $\widehat{\mathbf{B}}^{(i-1)}$ in Step 2(a) or $\widehat{\mathbf{A}}^{(i)}$ in Step 2(b) needs to be normalized according to the identification conditions mentioned in Section 2.1. The initial estimator can be either $\widehat{\mathbf{B}}^{(0)}$ or $\widehat{\mathbf{A}}^{(0)}$ which does not influence the properties of the final estimators. For the convergence conditions, there are several useful ones that we can adopt in Step 2(c) of **Algorithm 1**. For example, we may take the following two convergence criteria:

$$\|\widehat{\mathbf{A}}^{(i)} - \widehat{\mathbf{A}}^{(i-1)}\|_F \leq \eta \text{ and } \|\widehat{\mathbf{B}}^{(i)} - \widehat{\mathbf{B}}^{(i-1)}\|_F \leq \eta,$$

or

$$\|\widehat{\mathbf{B}}^{(i)} \otimes \widehat{\mathbf{A}}^{(i)} - \widehat{\mathbf{B}}^{(i-1)} \otimes \widehat{\mathbf{A}}^{(i-1)}\|_F \leq \eta,$$

where $\eta > 0$ is a prescribed small constant. In practice, we may choose $\eta = 10^{-6}$, and simulation results in Section 4 suggest that our algorithm works well in finite samples.

2.2.1 Iterated Least-Squares Estimation

Given $\mathbf{B} = \widehat{\mathbf{B}}^{(i-1)}$, in order to obtain the estimator of $\mathbf{A}$ and its bandwidth k_1, we write $\widehat{\mathbf{Q}}_t = \mathbf{Y}_t\widehat{\mathbf{B}}^{(i-1)\prime}$ and $\mathbf{Q}_t = \mathbf{Y}_t\mathbf{B}'$, and model (2.2) can be written as

$$\mathbf{Y}_t = \mathbf{A}\widehat{\mathbf{Q}}_{t-1} + \mathbf{F}_{1t}, \mathbf{F}_{1t} = \mathbf{A}(\mathbf{Q}_{t-1} - \widehat{\mathbf{Q}}_{t-1}) + \mathbf{E}_t. \tag{2.4}$$

Let $\mathbf{A}' = [\mathbf{a}_1, \cdots, \mathbf{a}_{p_1}]$, we have, $\mathbf{Y}_t'\mathbf{i}_j = \widehat{\mathbf{Q}}_{t-1}'\mathbf{a}_j + \mathbf{F}_{1t}'\mathbf{i}_j, j = 1, 2, \cdots, p_1,$ where $\mathbf{i}_j$ denotes the j-th canonical basis (unit) vector in $\mathbb{R}^{p_1}$, with 1 in the j-th entry and 0 elsewhere.

Assuming the bandwidth of $\mathbf{A}$ is k, then there are $\tau_j(k)$ non-zero elements in its j-th row

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

$\mathbf{a}_j$ of $\mathbf{A}$, where

$$\tau_j(k) = \begin{cases} k+j & \text{if} \quad j \le k+1, \\ 2k+1 & \text{if} \quad k+1 < j \le p_1 - k, \\ p_1 + k - j + 1 & \text{if} \quad p_1 - k < j \le p_1. \end{cases}$$

Let $\boldsymbol{\beta}_{j,k}$ be the $\tau_j(k) \times 1$ vector obtained by stacking non-zero elements in $\mathbf{a}_j$, and $\mathbf{X}^k_{j,t-1}$ be the corresponding $\tau_j(k)$ columns of $\widehat{\mathbf{Q}}'_{t-1}$. Denote $\mathbf{v}_j = [\mathbf{i}'_j\mathbf{Y}_2, \cdots, \mathbf{i}'_j\mathbf{Y}_{T+1}]' \in \mathbb{R}^{Tp_2 \times 1}$, $\mathbf{X}_{j,k} = [\mathbf{X}^k_{j,1}, \cdots, \mathbf{X}^k_{j,T}] \in \mathbb{R}^{Tp_2 \times \tau_j(k)}$, and $\boldsymbol{f}_j = [\mathbf{i}'_j\mathbf{F}_{12}, \cdots, \mathbf{i}'_j\mathbf{F}_{1,T+1}]'$, we have

$$\mathbf{v}_j = \mathbf{X}_{j,k}\boldsymbol{\beta}_{j,k} + \boldsymbol{f}_j. \tag{2.5}$$

Now, it follows from (2.5) that the least-squares estimator of $\boldsymbol{\beta}_{j,k}$ is denoted by $\widehat{\boldsymbol{\beta}}_{j,k} = (\mathbf{X}'_{j,k}\mathbf{X}_{j,k})^{-1}\mathbf{X}'_{j,k}\mathbf{v}_j$, and the corresponding residual sum of squares can be written as

$$\text{RSS}_j(k, \mathbf{a}_j) = \mathbf{v}'_j(\mathbf{I} - \mathbf{H}_{\mathbf{X}_{j,k}})\mathbf{v}_j, \tag{2.6}$$

where $\mathbf{H}_{\mathbf{X}_{j,k}} = \mathbf{X}_{j,k}(\mathbf{X}'_{j,k}\mathbf{X}_{j,k})^{-1}\mathbf{X}'_{j,k}$ is a hat matrix, which is a function of the unknown bandwidth k.

Next, we consider the estimation of $\mathbf{B}$ given $\widehat{\mathbf{A}}^{(i)}$. Similar to the technique used in (2.4), let $\widehat{\mathbf{R}}_t = \widehat{\mathbf{A}}^{(i)}\mathbf{Y}_t$ and $\mathbf{R}_t = \mathbf{A}\mathbf{Y}_t$, model (2.2) can be written as

$$\mathbf{Y}_t = \widehat{\mathbf{R}}_{t-1}\mathbf{B}' + \mathbf{F}_{2t}, \mathbf{F}_{2t} = (\mathbf{R}_{t-1} - \widehat{\mathbf{R}}_{t-1})\mathbf{B}' + \mathbf{E}_t. \tag{2.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Let $\mathbf{B}' = [\mathbf{b}_1, \cdots, \mathbf{b}_{p_2}]$, we have $\mathbf{Y}_t\mathbf{i}_j = \widehat{\mathbf{R}}_{t-1}\mathbf{b}_j + \mathbf{F}_{2t}\mathbf{i}_j, j = 1, 2, \cdots, p_2$.

Assuming the bandwidth of $\mathbf{B}$ is k, then there are $\tau_j(k)$ non-zero elements in $\mathbf{b}_j$, where

$$\tau_j(k) = \begin{cases} k + j & \text{if} \quad j \leq k + 1, \\ 2k + 1 & \text{if} \quad k + 1 < j \leq p_2 - k, \\ p_2 + k - j + 1 & \text{if} \quad p_2 - k < j \leq p_2. \end{cases}$$

Let $\boldsymbol{\gamma}_{j,k}$ be the $\tau_j(k) \times 1$ vector obtained by stacking non-zero elements in $\mathbf{b}_j$, and $\mathbf{G}^k_{j,t-1}$ be the corresponding $\tau_j(k)$ columns of $\widehat{\mathbf{R}}_{t-1}$. Denote $\mathbf{w}_j = [\mathbf{i}_j'\mathbf{Y}_2', \cdots, \mathbf{i}_j'\mathbf{Y}_{T+1}']'$, $\mathbf{G}_{j,k} = [\mathbf{G}^k_{j,1}; \cdots; \mathbf{G}^k_{j,T}]$, and $\boldsymbol{r}_j = [\mathbf{i}_j'\mathbf{F}_{22}', \cdots, \mathbf{i}_j'\mathbf{F}_{2,T+1}']'$, we have

$$\mathbf{w}_j = \mathbf{G}_{j,k}\boldsymbol{\gamma}_{j,k} + \boldsymbol{r}_j, \tag{2.8}$$

and obtain the least-squares estimator of $\boldsymbol{\gamma}_{j,k}$ as $\widehat{\boldsymbol{\gamma}}_{j,k} = (\mathbf{G}_{j,k}'\mathbf{G}_{j,k})^{-1}\mathbf{G}_{j,k}'\mathbf{w}_j$. The corresponding residual sum of squares is given by

$$\text{RSS}_j(k, \mathbf{b}_j) = \mathbf{w}_j'(\mathbf{I} - \mathbf{H}_{\mathbf{G}_{j,k}})\mathbf{w}_j, \tag{2.9}$$

where $\mathbf{H}_{\mathbf{G}_{j,k}} = \mathbf{G}_{j,k}(\mathbf{G}_{j,k}'\mathbf{G}_{j,k})^{-1}\mathbf{G}_{j,k}'$ is a hat matrix, which is also a function of the unknown bandwidth k as that in (2.6).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

2.2.2 Determining the bandwidth

As discussed in Section 2.2.1, the estimation of the unknown coefficients depends on the bandwidth parameters k_1 and k_2, which are unknown in practice. In this section, we propose a Bayesian information criterion (BIC) to determine the unknown bandwidth of $\mathbf{A}$. We first consider the estimation of the bandwidth k_1 of $\mathbf{A}$. For each prescribed $k_1 \geq 1$, we may obtain the least-squares estimator of $\widehat{\boldsymbol{\beta}}_{j,k}$ from Model (2.5) as well as the residual-sum of squares in (2.6). For $j = 1, ..., p_1$, we define

$$\text{BIC}_j(k) = \log \text{RSS}_j(k, \mathbf{a}_j) + \frac{C_{M_2}}{M_2} \tau_j(k) \log(p_1 \vee M_2),$$

where $M_2 = p_2 T$ and $C_{M_2} = \log\log(M_2)$. The bandwidth of the j-th row of $\mathbf{A}$ estimated from $(\mathbf{X}_{j,k}, \mathbf{v}_j)$ is given by

$$\widehat{k}_{1,j} = \arg \min_{1 \leq k \leq K} \text{BIC}_j(k),$$

where K is a prescribed upper bound of k which may be taken as $\lceil T^{1/2} \rceil$. Finally, the estimated bandwidth of $\mathbf{A}$ in the i-th iteration of **Algorithm 1** is given by $\widehat{k}_1^{(i)} = \max_{1 \leq j \leq p_1} \widehat{k}_{1,j}$, and the estimator for $\mathbf{A}$ in the i-th iteration is denoted by $\widehat{\mathbf{A}}^{(i)} = [\widehat{\mathbf{a}}_1, \widehat{\mathbf{a}}_2, \cdots, \widehat{\mathbf{a}}_{p_1}]'$, where the estimator $\widehat{\mathbf{a}}_j$ of $\mathbf{a}_j$ is obtained by replacing the corresponding non-zero elements in $\mathbf{a}_j$ by $\widehat{\boldsymbol{\beta}}_{j,\widehat{k}_j}$.

Similarly, for the estimation of the bandwidth parameter k_2, we can define the following BIC criterion

$$\text{BIC}_j(k) = \log \text{RSS}_j(k, \mathbf{b}_j) + \frac{C_{M_1}}{M_1} \tau_j(k) \log(p_2 \vee M_1),$$

where $M_1 = p_1 T$, and $C_{M_1} = \log\log(M_1)$. The bandwidth of the j-th row of $\mathbf{B}$ can be

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

estimated by

$$\widehat{k}_{2,j} = \arg \min_{1 \le k \le K} \text{BIC}_j(k),$$

where K is a prescribed upper bound of k which may be taken as $\lceil T^{1/2} \rceil$. Finally, the bandwidth of $\mathbf{B}$, is estimated as $\widehat{k}_2^{(i)} = \max_{1 \le j \le p_2} \widehat{k}_{2,j}$ and $\mathbf{B}$ is estimated as $\widehat{\mathbf{B}}^{(i)} = [\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \cdots, \widehat{\mathbf{b}}_{p_2}]'$, where estimator $\widehat{\mathbf{b}}_j$ of $\mathbf{b}_j$ is obtained by replacing corresponding non-zero elements in $\mathbf{b}_j$ by $\widehat{\boldsymbol{\beta}}_{j,\widehat{k}_j}$, which is similar as that in estimating k_1.

In practice, the upper bound $K > 0$ in the BICs defined above is a prescribed integer. Our numerical results show that the procedure is insensitive to the choice of K so long as $K > k_1$ and $K > k_2$. In practice, we may take K to be $\min([T^{1/2}], [p_1^{1/2}], [p_2^{1/2}])$ or choose K by checking the curvature of $\text{BIC}_i(k)$ directly.

2.3 Estimation with Sparse Coefficient Matrices

In this section, we consider the estimation of the coefficient matrices in model (2.2) under the scenario that the coefficients are sparse, i.e., we assume that $\mathbf{A}$ and $\mathbf{B}$ are sparse in the sense that only a few elements within are nonzero. Note that we may apply the properties of the Kronecker product to model (2.2), and rewrite the model in the following two ways:

$$\text{vec}(\mathbf{Y}_t) = ((\mathbf{B}\mathbf{Y}'_{t-1}) \otimes \mathbf{I}_{p_1})\text{vec}(\mathbf{A}) + \text{vec}(\mathbf{E}_t), \tag{2.10}$$

and

$$\text{vec}(\mathbf{Y}'_t) = ((\mathbf{A}\mathbf{Y}_{t-1}) \otimes \mathbf{I}_{p_2})\text{vec}(\mathbf{B}) + \text{vec}(\mathbf{E}'_t), \tag{2.11}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

where vec$(\cdot)$ is the vectorization operator that stacks all the columns of a matrix into a vector in order.

Let $\mathbf{y}_t = \text{vec}(\mathbf{Y}_t)$ and $\mathbf{y}_t^* = \text{vec}(\mathbf{Y}_t')$, it is possible to estimate $\mathbf{A}$ when $\mathbf{B}$ is known in (2.10) and to estiamte $\mathbf{B}$ when $\mathbf{A}$ is known in (2.11). For any consistent estimators $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{A}}$ for $\mathbf{B}$ and $\mathbf{A}$, respectively, we may define $\mathbf{Z}_{t-1} = (\mathbf{B}\mathbf{Y}_{t-1}') \otimes \mathbf{I}_{p_1}$ and $\widehat{\mathbf{Z}}_{t-1} = (\widehat{\mathbf{B}}\mathbf{Y}_{t-1}') \otimes \mathbf{I}_{p_1}$. Similarly, we may also define $\mathbf{Z}_{t-1}^* = (\mathbf{A}\mathbf{Y}_{t-1}) \otimes \mathbf{I}_{p_2}$ and $\widehat{\mathbf{Z}}_{t-1}^* = (\widehat{\mathbf{A}}\mathbf{Y}_{t-1}) \otimes \mathbf{I}_{p_2}$. In view of the spare structures of the coefficient matrices, we may adopt some penalized method such as the Lasso to obtain the estimators.

Specifically, similar to the approach in the banded case in Section 2.2, we first obtain the initial estimators of $\mathbf{A}$ and $\mathbf{B}$ by applying the alternative least squares method proposed by Chen et al. (2021). Starting with these initial estimators, we may apply the Lasso technique to estimate $\mathbf{A}$ by replacing $\mathbf{B}$ with its latest estimator, and then obtain the Lasso estimate of $\mathbf{B}$ by replacing $\mathbf{A}$ with its latest estimator. For example, denote the estimate for $\mathbf{B}$ as $\widehat{\mathbf{B}}^{(i-1)}$ in the i-th iteration, we solve the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} = \arg \min_{\boldsymbol{\alpha} \in R^{p_1^2}} \left\{ \frac{1}{T} \sum_{t=2}^{T} \|\mathbf{y}_t - \widehat{\mathbf{Z}}_{t-1}\boldsymbol{\alpha}\|_2^2 + \lambda_{1,T}\|\boldsymbol{\alpha}\|_1 \right\}, \tag{2.12}$$

where $\widehat{\mathbf{B}}$ is equal to $\widehat{\mathbf{B}}^{(i-1)}$ in $\widehat{\mathbf{Z}}_{t-1}$ and $\lambda_{1,T} > 0$ is a tuning parameter. Then the estimator $\widehat{\mathbf{A}}^{(i)}$ is obtained by reverting the $\widehat{\boldsymbol{\alpha}}$ to a $p_1 \times p_1$ matrix according to the way vec$(\cdot)$ is performed. Similarly, $\widehat{\boldsymbol{\beta}}$ is obtained by solving the following optimization problem:

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in R^{p_2^2}} \left\{ \frac{1}{T} \sum_{t=2}^{T} \|\mathbf{y}_t^* - \widehat{\mathbf{Z}}_{t-1}^*\boldsymbol{\beta}\|_2^2 + \lambda_{2,T}\|\boldsymbol{\beta}\|_1 \right\}, \tag{2.13}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

where $\widehat{\mathbf{A}}$ is equal to $\widehat{\mathbf{A}}^{(i)}$ in $\widehat{\mathbf{Z}}^{*}_{t-1}$. Then, the estimator $\widehat{\mathbf{B}}^{(i)}$ is obtained by reverting the $\widehat{\boldsymbol{\beta}}$ to a $p_2 \times p_2$ matrix as before. We can repeat this procedure until convergence. The estimation procedure is summarized in **Algorithm 2**. The convergence criteria are similar to those in **Algorithm 1**, and we do not repeat them to save space.

Algorithm 2 Estimating algorithm for sparse case

1. Obtain $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$ by the method in Chen et al. (2021), denoted as $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_0$ and $\widehat{\mathbf{A}}^{(0)} = \widehat{\mathbf{A}}_0$, respectively.
2. For the i-th iteration ($i = 1, 2, ...$),
   (a) Fix $\mathbf{B} = \widehat{\mathbf{B}}^{(i-1)}$, apply Lasso to (2.12) and obtain $\widehat{\mathbf{A}}^{(i)}$,
   (b) Fix $\mathbf{A} = \widehat{\mathbf{A}}^{(i)}$, apply Lasso to (2.13) and obtain $\widehat{\mathbf{B}}^{(i)}$,
   (c) The iteration stops if the convergence criterion is satisfied, otherwise we go to the next iteration and repeat Steps 2(a)–2(b).

3 Theoretical Properties

In this section, we establish the asymptotic properties of the estimators proposed in Section 2. We begin by outlining the regular conditions necessary for the theoretical proofs, followed by the statement of the asymptotic theorems. All proofs for the theorems are provided in an online Supplementary Material.

3.1 Regular Conditions

We introduce some notations first. A process $\text{vec}(\mathbf{Y}_t)$ is α-mixing if

$$\alpha_p(k) = \sup_i \sup_{A \in \mathcal{F}^i_{-\infty}, B \in \mathcal{F}^{\infty}_{i+k}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \to 0,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

where $\mathcal{F}_l^k$ is the σ-field generated by $\{\text{vec}(\mathbf{Y}_t) : l \le t \le k\}$. For $i = 1, 2$, define $C_{i,\alpha}(S_i) = \{\Delta \in R^{p_i^2} : \|\Delta_{S_i^c}\|_1 \le \alpha\|\Delta_{S_i}\|_1\}$, where S_i is a subset of $\{1, 2, ..., p_i^2\}$ and Δ_{S_i} is the vector of Δ restricted on the positions of S_i and the other elements on indexes of S_i^c are zero. Now, we introduce some assumptions for Model (2.2).

A1. For $\mathbf{Y}_t = \{y_{ijt}\}$, we assume

   a. The process $\text{vec}(\mathbf{Y}_t)$ is α-mixing with the mixing coefficient satisfying the condition $\alpha_p(k) \le \exp(-ck^{\gamma_1})$ for some $\gamma_1 > 0$.

   b. $\sup_{i,j,t} \mathbb{P}(|y_{ij,t}| > s) \le \exp(1 - s^{\gamma_2})$ for $s > 0$ and some $\gamma_2 > 0$.

A2. The innovations $\{\mathbf{E}_t = (e_{ijt})\}$ are independent and identically distributed (i.i.d.) with mean 0, and

   a. $\frac{1}{p_1}\sum_{i=1}^{p_1} \mathbb{E}e_{ijt}^2 \to \sigma_1^2$ for $j = 1, 2, \cdots, p_2$ and $\frac{1}{p_2}\sum_{j=1}^{p_2} \mathbb{E}e_{ijt}^2 \to \sigma_2^2$, for $i = 1, 2, \cdots, p_1$.

   b. $\sup_{i,j,t} \mathbb{P}(|e_{ij,t}| > s) \le \exp(1 - s^{\gamma_3})$ for $s > 0$ and some $\gamma_3 > 0$.

A3. $\rho(\mathbf{A})\rho(\mathbf{B}) < 1$, where $\rho(\mathbf{A})$ and $\rho(\mathbf{B})$ are the spectral radii of $\mathbf{A}$ and $\mathbf{B}$, respectively.

A4. For any two sub-columns of $\mathbf{Q}_t'$ (or $\mathbf{R}_t$), denoted by $\mathbf{W}_t$ and $\mathbf{U}_t$, and $\mathbf{W}_t \ne \mathbf{U}_t$, let $p = p_2$ (or p_1), $\Sigma_{\mathbf{U}} = p^{-1}\mathbb{E}\mathbf{U}_t'\mathbf{U}_t$, $\Sigma_{\mathbf{WU}} = p^{-1}\mathbb{E}\mathbf{W}_t'\mathbf{U}_t$, and $\Sigma_{\mathbf{W}} = p^{-1}\mathbb{E}\mathbf{W}_t'\mathbf{W}_t$, there exists some positive constants $\lambda_1 \le \lambda_2$, such that $\lambda_1 \le \lambda_{\min}(\Sigma_{\mathbf{U}}) \le \lambda_{\max}(\Sigma_{\mathbf{U}}) \le \lambda_2$, and $\lambda_1 \le \lambda_{\min}(\Sigma_{\mathbf{W}} - \Sigma_{\mathbf{WU}}\Sigma_{\mathbf{U}}^{-1}\Sigma_{\mathbf{WU}}') \le \lambda_{\max}(\Sigma_{\mathbf{W}} - \Sigma_{\mathbf{WU}}\Sigma_{\mathbf{U}}^{-1}\Sigma_{\mathbf{WU}}') \le \lambda_2$.

A5. For the banded matrix $\mathbf{A}$, $|a_{i,i-k_1}|$ or $|a_{i,i+k_1}|$, $i = 1, 2, \cdots, p_1$, is greater than $\{C_{M_2}k_1M_2^{-1}\log(p_1 \vee M_2)\}^{1/2}$; Similarly, for the banded matrix $\mathbf{B}$, $|b_{i,i-k_2}|$ or $|b_{i,i+k_2}|$, $i = 1, 2, \cdots, p_2$, is greater than $\{C_{M_1}k_2M_1^{-1}\log(p_2 \vee M_1)\}^{1/2}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

A6. For matrices $\mathbf{A}$ and $\mathbf{B}$, $a_{i,j}$ and $b_{i,j}$ are bounded uniformly, and $\|\mathbf{A}^k\|_2 \leq \delta^k$ and $\|\mathbf{B}^k\|_2 \leq \delta^k$ for $k \geq 2$, where $\delta \in (0,1)$ is independent of p_1, p_2, and k.

A7. Let S_0 be a subset of $\{1,2,...,p_1^2\}$ with cardinality s_0 consisting of the indexes of the non-zero components in $\boldsymbol{\alpha} = \text{vec}(\mathbf{A})$, and S_0^c be its complement. Let $\mathbf{Z}_t = (\mathbf{B}\mathbf{Y}_t') \otimes \mathbf{I}_{p_1}$, (a) when p_1 is finite, there exists a constant $C_2 > 0$ such that $\lambda_{min}\{\mathbb{E}(\mathbf{Z}_t\mathbf{Z}_t')\} > C_2$; (b) when p_1 is diverging, the matrix $\mathbf{Z} := (\mathbf{Z}_1, ..., \mathbf{Z}_T)'$ satisfies the restricted eigenvalues condition, $\frac{1}{T}\|\mathbf{Z}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$, for all $\Delta \in C_{1,3}(S_0)$. Similar assumptions also hold for $\mathbf{Z}_t^*$ defined in Section 2.3.

A8. For matrices $\mathbf{A}$ and $\mathbf{B}$, $\lambda_1 < \lambda_{\min}\{p_2^{-1}\mathbb{E}[(\mathbf{A}\mathbf{Y}_t) \otimes (\mathbf{Y}_t\mathbf{B}')]\} < \lambda_{\max}\{p_2^{-1}\mathbb{E}[(\mathbf{A}\mathbf{Y}_t) \otimes (\mathbf{Y}_t\mathbf{B}')]\} < \lambda_2$.

Conditions A1(a-b) are standard for econometric time series models. Condition A2(a) ensures the row and column variances of $\mathbf{E}_t$ exits, and condition A2(b) is used to bound a new time series built on $\mathbf{E}_t$ and $\mathbf{Y}_t$. Condition A3 ensures that model (2.2) is stationary and causal, as shown in proposition 1 in Chen et al. (2021). Conditions A4-A7 are imposed to prove the consistency of the estimated bandwidth by BIC in Section 2.2.2. Condition A5 ensures that the bandwidth is asymptotically identifiable, since both $\{C_{M_2}k_1M_2^{-1}\log(p_1 \vee M_2)\}^{1/2}$ and $\{C_{M_1}k_2M_1^{-1}\log(p_2 \vee M_1)\}^{1/2}$ is the minimum magnitude of a non-zero coefficient to be identifiable, see, e.g. Gao et al. (2019). Condition A7(a) indicates that the regressors have a non-singular covariance and the least-squares estimators are well defined when p_1 is finite. Condition A7(b) is the well-known restricted-eigenvalue condition in Lasso regressions; see Chapter 6 in Bühlmann and Van De Geer (2011). The condition in Condition A7(b) can also

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

be replaced by a more general Restricted Strong Convexity condition that is commonly used in high-dimensional regularized estimation problems. See Chapter 9 of Wainwright (2019) for details.

3.2 Asymptotic properties

In this section, we study the theoretical properties of the proposed method, i.e., the convergence of **Algorithm 1** and **Algorithm 2** in Section 2. Since we take estimators, $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$, from the iterated least squares in Chen et al. (2021) as our initials in **Algorithm 1** and **Algorithm 2**, we first study the convergence rate of $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$. Let Σ be the covariance matrix of $\text{vec}(\mathbf{E}_t)$, and define $\boldsymbol{H} = \mathbb{E}\left(\boldsymbol{W}_t\boldsymbol{W}_t'\right) + \boldsymbol{\gamma}\boldsymbol{\gamma}'$, where $\boldsymbol{W}_t' = [(\mathbf{B}\mathbf{Y}_t') \otimes \mathbf{I} : \mathbf{I} \otimes (\mathbf{A}\mathbf{Y}_t)] \in \mathbb{R}^{p_1p_2\times(p_1^2+p_2^2)}$ and $\boldsymbol{\gamma} = (\text{vec}(\mathbf{A})', \mathbf{0})' \in \mathbb{R}^{p_1^2+p_2^2}$. Let $p = \max\{p_1, p_2\}$, we have following result for $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$.

Proposition 1. *Let conditions A1-A3 hold. If* $\mathbf{A}$, $\mathbf{B}$, Σ *are nonsingular,* $\lambda_{\min}(\mathbf{H}) \geq \lambda_h > 0$, *and* $T, p_1, p_2 \to \infty$, *then,*

$$\|\widehat{\mathbf{A}}_0 - \mathbf{A}\|_F^2 + \|\widehat{\mathbf{B}}_0 - \mathbf{B}\|_F^2 = O_p(\frac{p_1^2p_2}{T} + \frac{p_2^2p_1}{T}).$$

Remark 3.1. (i) When the dimensions of $\mathbf{Y}_t$, p_1 and p_2, are fixed, the convergence rate of $\|\widehat{\mathbf{A}}_0 - \mathbf{A}\|_F^2$ and $\|\widehat{\mathbf{B}}_0 - \mathbf{B}\|_F^2$ are both of order $O_p(1/T)$. This recovers the low-dimensional case. (ii) According to this Proposition, we have $\|\widehat{\mathbf{A}}_0 - \mathbf{A}\|_F^2 + \|\widehat{\mathbf{B}}_0 - \mathbf{B}\|_F^2 \to 0$ if $p_1^2p_2/T \to 0$ and $p_2^2p_1/T \to 0$, which is ensured if $pp_1p_2/T \to 0$. (iii) The convergence rate stated in Theorem 3 of Li and Xiao (2021) is slightly faster than that in Proposition 1. However, the detailed proof is not fully transparent and appears technically involved. We leave the task

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

of refining the convergence rates of the initial estimators under high-dimensional settings to future research.

Next, we show the convergence rate of estimators for $\mathbf{A}$ and $\mathbf{B}$ in **Algorithm 1** by assuming the bandwidth k_1 and k_2 are known and fixed. Theorem 3.1 below shows that estimators from **Algorithm 1** are consistent when $p^3/T \to 0$.

Theorem 3.1. *Assume conditions A1-A6 and A8 hold. Let* $\widehat{\mathbf{B}}^{(i-1)}$ *be the latest estimator of* $\mathbf{B}$ *in* ***Algorithm 1*** *with* $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_0$ *and* $i \geq 1$. *We have, for* $i > 1$,

$$\|\widehat{\mathbf{A}}^{(i)} - \mathbf{A}\|_F^2 = O_p(p_1 p_2 p/T) \text{ and } \|\widehat{\mathbf{B}}^{(i)} - \mathbf{B}\|_F^2 = O_p(p_1 p_2 p/T).$$

Remark 3.2. (i) In this theorem, **Algorithm 1** is assumed to begin estimating $\mathbf{A}$ by fixing $\mathbf{B}$ at its most recent estimate. However, the same conclusion will hold if we reverse the estimation order. (ii) In the first iteration, we take $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_0$, by Proposition 1, we have $\|\widehat{\mathbf{B}}_0 - \mathbf{B}\|_F^2 \to 0$ when $pp_1p_2/T \to 0$. The condition $pp_1p_2/T \to 0$ also ensures that the error of $\mathbf{A}^{(i)}$ is primarily influenced by the error arising from approximating $\mathbf{B}$ with its latest estimate.

Theorem 3.1 is based on the assumption that the bandwidths are known, which is often not the case in real-world problems. However, if consistent estimators for these two unknown bandwidths exist, Theorem 3.1 remains valid. We will now demonstrate the consistency of the estimated bandwidths in **Algorithm 1**.

Theorem 3.2. *Assume conditions A1-A6 hold. Let* $\widehat{\mathbf{B}}^{(i-1)}$ *be the latest estimator of* $\mathbf{B}$ *in*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Algorithm 1 *with* $\widehat{\mathbf{B}}^{(0)} = \widehat{\mathbf{B}}_0$. *For* $i > 1$, *if* $p^5/T \to 0$, *then*

$$P(\widehat{k}_1^{(i)} = k_1) \to 1, \text{ and } P(\widehat{k}_2^{(i)} = k_2) \to 1, \text{ as } T, p_1, p_2 \to \infty.$$

Remark 3.3. (i) In Theorem 3.1, k_1 and k_2 are assumed to be fixed, since model (2.2) is useful only when k_1 and k_2 are small and finite. However, Theorem 3.1 still holds when k_1 and k_2 diverges to ∞ along with T, p_1, p_2 so long as $k_1 = o\{C_{M_2}^{-1} M_2 / \log(p_1 \vee M_2)\}$, and $k_2 = o\{C_{M_1}^{-1} M_1 / \log(p_2 \vee M_1)\}$. See the proof of Theorem 3.1 in Supplementary Material.
(ii) In this theorem, **Algorithm 1** is assumed to begin by estimating $\mathbf{A}$ while holding $\mathbf{B}$ fixed at its most recent estimate. However, the same conclusion holds if we reverse the estimation order.

Next, we examine the convergence rate of **Algorithm2** for estimating $\mathbf{A}$ and $\mathbf{B}$ in the sparse case. Similar to **Algorithm1**, **Algorithm 2** also begins with $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{B}}_0$, which are derived from the iterated least squares method in Chen et al. (2021).

Theorem 3.3. *Assume conditions A1-A3 and A6-A7 hold. Let* $\widehat{\mathbf{B}}^{(i-1)}$ *be the latest estimator of* $\mathbf{B}$ *in* ***Algorithm 2*** *with* $\widehat{\mathbf{B}}^{0} = \widehat{\mathbf{B}}_0$ *and* $i \geq 1$, *we have,*

$$\|\widehat{\mathbf{A}}^{(i)} - \mathbf{A}\|_2 = O_p(\sqrt{\frac{p_1 p_2^2 s_0}{T}}) \text{ and } \|\widehat{\mathbf{B}}^{(i)} - \mathbf{B}\|_2 = O_p(\sqrt{\frac{p_1^2 p_2 s_0}{T}}),$$

where $s_0 = |S_0|$.

Remark 3.4. (i) Similar to the Frobenius norm results in the banded case, Theorem 3.3 establishes that the estimated sparse coefficient matrices are consistent in the ℓ_2 norm, provided that

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

$pp_1p_2/T \to 0$ and the sparsity level satisfies $0 < s_0 < \infty$. In fact, the convergence rates are roughly the same as those in Theorem 3.1, since we treat the bandwidth parameter as finite and absorb it into the upper bound. (ii) The convergence rates in Theorem 3.1 and Theorem 3.3 depend on the rates of the initial estimators. As suggested by one of the reviewers, these rates can potentially be improved by using better initial estimates. One suggested approach is to begin with banded or sparse initial estimators in Algorithm 1 and Algorithm 2, respectively—similar to the strategy in Yang et al. (2016)—since the effective number of parameters in such structured matrices is smaller, which may lead to faster convergence. However, the ALS setting presents a different challenge. In this case, it is not straightforward to separate a proportion of the parameters in the matrices for individual convergence rate analysis. As a result, we must still rely on the convergence rates of the full initial estimators, as described in Proposition 1. Simulation results (see Table S9 of the Supplement) suggest that, in the banded case, there is no notable difference in estimation error between starting with the full initial estimator and starting with a banded initial estimator—both lead to nearly identical performance.

4 Numerical Results

In this section, we examine the finite sample properties of the proposed method and provide real data examples to evaluate its forecasting performance compared to the alternative least-squares (ALSE) method proposed by Chen et al. (2021). In Section 4.1, we conduct Monte Carlo experiments to demonstrate the convergence of the proposed methods in estimating the coefficient matrices under two scenarios, alongside comparisons with the estimators obtained

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

using the ALSE method. In Section 4.2, we apply our method to two real data examples.

4.1 Simulation

This section outlines our methodology for evaluating the finite sample properties of the proposed methods through Monte-Carlo experiments. The observed data matrix $\mathbf{Y}_t$ are simulated from model (2.2) under different conditions for matrices $\mathbf{A}$ and $\mathbf{B}$, and each entry in the white noise $\mathbf{E}_t$ is generated from the standard normal distribution, with $\text{Cov}(\text{vec}(\mathbf{E}_t)) = \mathbf{I}_{p_1 p_2}$. Through these simulations, we aim to demonstrate the convergence of our proposed methods in comparison with the ALSE method, as the sample size increases. Furthermore, we examine the accuracy of our proposed methods in estimating unknown bandwidths under the banded cases. The impact of penalty parameters on the estimation results is also investigated under the sparse cases. All of the results are obtained by conducting 100 independent replications.

To study the convergence of the estimators for matrices $\mathbf{A}$ and $\mathbf{B}$, some identification conditions are required to impose on the coefficients. In this experiment, we assume $\|\mathbf{A}\|_F = 1$ and $\text{sign}(\text{tr}(\mathbf{A})) = 1$. The convergence criteria of the iterations in our algorithms are specified as $\|\widehat{\mathbf{A}}^{(i+1)} - \widehat{\mathbf{A}}^{(i)}\|_F < 10^{-6}$ and $\|\widehat{\mathbf{B}}^{(i+1)} - \widehat{\mathbf{B}}^{(i)}\|_F < 10^{-6}$.

4.1.1 Banded case

This section presents a comprehensive analysis to investigate the performance of **Algorithm 1**. We will study the convergence of the estimators under the scenarios that we start with either $\widehat{\mathbf{A}}^{(0)}$ or $\widehat{\mathbf{B}}^{(0)}$ as initial estimates. Additionally, we examine the accuracy of the bandwidth estimation and the algorithm's convergence as the sample size T increases.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

For each configuration of (p_1, p_2, k_1, k_2, T), we generate $\mathbf{Y}_t$ according to model (2.2). Specifically, for given dimensions p_1, p_2, and bandwidths k_1 and k_2, the observed data $\mathbf{Y}_t$ are simulated according to model (2.2), where the entries of $\mathbf{A}$ and $\mathbf{B}$ are generated as follows: (1) for entries of $\mathbf{A}$, $\{a_{i,j} : |i-j| \leq k_1\}$ are generated independently from $U[-1,1]$, and other elements are zero. We re-scale $\mathbf{A}$ such that $\|\mathbf{A}\|_F = 1$ and $\text{sign}(\text{tr}(\mathbf{A})) = 1$; (2) for entries of $\mathbf{B}$, $\{b_{i,j} : |i-j| \leq k_2\}$ are generated independently from $U[-1,1]$, and other elements are zero. We re-scale $\mathbf{B}$ so that $\rho = \rho(\mathbf{A})\rho(\mathbf{B}) = 0.5$. The white noise $\mathbf{E}_t$ are generated from standard normal distribution with $\text{Cov}(\text{vec}(\mathbf{E}_t)) = \mathbf{I}_{p_1 p_2}$.

Firstly, we examine the convergence of **Algorithm 1** is insensitive to the choice of the initial estimators. Note that there are two iteration orders that may occur in **Algorithm 1**. We may first estimate $\mathbf{A}$ for given initial estimator $\widehat{\mathbf{B}}^{(0)}$, or estimate $\mathbf{B}$ for given $\widehat{\mathbf{A}}^{(0)}$. We denote the estimated coefficient matrices via these two procedures by $(\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1)$ and $(\widehat{\mathbf{A}}_2, \widehat{\mathbf{B}}_2)$, respectively. The dimensions are set as $(p_1, p_2) = (6,4), (8,5)$ and $(9,6)$ with bandwidths $(k_1, k_2) = (2,1)$ for each (p_1, p_2). The mean, median, and maximum of $\log_{10}(\|\widehat{\mathbf{A}}_1 - \widehat{\mathbf{A}}_2\|_F)$ and $\log_{10}(\|\widehat{\mathbf{B}}_1 - \widehat{\mathbf{B}}_2\|_F)$ are reported in Table S1. From Table S1 we see that the convergence of the estimators is insensitive to the choice of the initial estimators that we use in **Algorithm 1**. On the other hand, the reported errors in Table S1 are all less than -6, which is in accordance with the convergence criteria where the upper bound η is chosen as 10^{-6} in Section 2.

Second, we show the accuracy of **Algorithm 1** in estimating the unknown bandwidths k_1 and k_2. The empirical frequencies of the events $\{\widehat{k}_1 = k_1\}$ and $\{\widehat{k}_2 = k_2\}$ are reported in Table 1, where we set $(k_1, k_2) = (1,1)$ and $(2,1)$ and the sample size $T = 100, 200, 400$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

800. For a fixed sample size T and the bandwidths (k_1, k_2), the dimensions (p_1, p_2) are set to $(6, 4)$, $(8, 5)$, and $(9, 6)$. Results in Table 1 show that the accuracy of estimated $\widehat{k}_1$ and $\widehat{k}_2$ is pretty satisfactory, and it increases with sample size T for each (p_1, p_2, k_1, k_2) in most cases.

Table 1: Accuracy of **Algorithm 1** in estimating unknown bandwidths under different settings, where E_1 and E_2 represent the empirical frequencies of the events $\{\widehat{k}_1 = k_1\}$ and $\{\widehat{k}_2 = k_2\}$, respectively.

	$T = 100$		$T = 200$		$T = 400$		$T = 800$	
(p_1, p_2)	E_1	E_2	E_1	E_2	E_1	E_2	E_1	E_2
$(k_1, k_2) = (1, 1)$								
$(6, 4)$	100	100	100	100	100	100	100	100
$(8, 5)$	100	99	100	99	100	100	100	100
$(9, 6)$	98	99	100	99	100	100	100	100
$(k_1, k_2) = (2, 1)$								
$(6, 4)$	98	100	100	100	100	100	100	100
$(8, 5)$	99	100	100	100	100	100	100	100
$(9, 6)$	99	100	100	100	100	100	100	100

Next, we show the convergence pattern of **Algorithm 1** under different configurations of (p_1, p_2, k_1, k_2) as the sample size T increases. We also compare the estimation accuracy with the ALSE method in Chen et al. (2021). The estimation errors for $\mathbf{A}$ and $\mathbf{B}$, denoted by $\log(\|\widehat{\mathbf{A}} - \mathbf{A}\|_F)$ and $\log(\|\widehat{\mathbf{B}} - \mathbf{B}\|_F)$, respectively, are reported in Table S2, where $(p_1, p_2) = (6, 4)$ and $(9, 6)$, the sample size $T = 200, 500, 1000, 2000$, and the bandwidths $(k_1, k_2) = (2, 1)$. For each setting, we consider two scenarios that $\rho(\mathbf{A})\rho(\mathbf{B}) = 0.5$ and 0.8 to show the results are consistent for different strengths of the coefficient matrices. From Table S2, we see that estimation errors obtained by the proposed method and the ALSE all decrease as the sample size increase for each configuration, which is in line with our theoretical results. On the other hand, we also see that the estimation error obtained by our proposed method is smaller than that by the ALSE, implying that our estimation procedure is more accurate than

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

the ALSE.

Furthermore, we define $\mathbf{S} := \mathbf{B} \otimes \mathbf{A}$ and $\widehat{\mathbf{S}} := \widehat{\mathbf{B}} \otimes \widehat{\mathbf{A}}$, using the distance $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$ to evaluate the overall performance of our proposed method. For simplicity, we set the dimensions $(p_1, p_2) = (6, 4)$ and $(9, 6)$, and the bandwidths (k_1, k_2) to $(1, 1)$ and $(2, 1)$ for each (p_1, p_2). We fix $\rho(\mathbf{A})\rho(\mathbf{B}) = 0.5$ in this experiment, and the box plots of the estimated errors $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$ are shown in Figure S1. It is clear that both methods converge under these settings, and our proposed **Algorithm 1** performs better than the ALSE method in terms of estimation errors, which aligns with our theoretical results. Similar results are obtained from simulations conducted in higher-dimensional settings (Figure S2, Table 2).

Table 2: The average estimation errors of the coefficient matrices by **Algorithm 1** and ALSE.

		Algorithm 1				ALSE			
(p_1, p_2)	ρ	T = 200	500	1000	2000	T = 200	500	1000	2000
					$\log(\|\widehat{\mathbf{A}} - \mathbf{A}\|_F)$				
$(12, 15)$	0.5	-1.712	-2.378	-2.779	-3.09	-1.631	-2.111	-2.455	-2.809
$(12, 15)$	0.8	-2.658	-3.122	-3.48	-3.808	-2.262	-2.733	-3.091	-3.421
$(20, 20)$	0.5	-2.295	-2.828	-3.178	-3.534	-1.788	-2.25	-2.602	-2.951
$(20, 20)$	0.8	-2.761	-3.211	-3.563	-3.901	-2.158	-2.624	-2.974	-3.314
					$\log(\|\widehat{\mathbf{B}} - \mathbf{B}\|_F)$				
$(12, 15)$	0.5	-0.427	-0.857	-1.266	-1.594	0.04	-0.456	-0.813	-1.155
$(12, 15)$	0.8	-0.681	-1.112	-1.4	-1.677	-0.052	-0.515	-0.865	-1.209
$(20, 20)$	0.5	-0.274	-0.866	-1.267	-1.609	0.314	-0.169	-0.523	-0.874
$(20, 20)$	0.8	-0.392	-0.957	-1.287	-1.597	0.229	-0.239	-0.59	-0.938
					$\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$				
$(12, 15)$	0.5	0.001	-0.547	-0.956	-1.277	0.285	-0.207	-0.559	-0.906
$(12, 15)$	0.8	-0.35	-0.799	-1.118	-1.416	0.178	-0.289	-0.642	-0.981
$(20, 20)$	0.5	0.1	-0.46	-0.833	-1.184	0.647	0.173	-0.181	-0.531
$(20, 20)$	0.8	-0.05	-0.558	-0.901	-1.225	0.56	0.091	-0.259	-0.604

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

4.1.2 Sparse case

This section evaluates the performance of **Algorithm 2** in estimating sparse coefficient matrices. First, we assess its ability to recover the non-zero elements of matrices $\mathbf{A}$ and $\mathbf{B}$. The tuning parameters $\lambda_{1,T}$ and $\lambda_{2,T}$ in equations (2.12) and (2.13) significantly influence the sparsity of $\mathbf{A}$ and $\mathbf{B}$, respectively. It is important to note that if these parameters are adjusted in each iteration of **Algorithm 2**, the algorithm will not converge, as the objective function changes with the tuning parameters. Therefore, we select the tuning parameters in the first iteration and keep them fixed for subsequent iterations. In practice, tuning parameters are usually chosen through cross-validation (CV). The R package glmnet offers two options: (1) sdCV, which selects $\lambda_{1,T}$ ($\lambda_{2,T}$) as the largest value of λ such that the corresponding CV error is within 1 standard error of the minimum, and (2) mCV, which selects $\lambda_{1,T}$ ($\lambda_{2,T}$) as the value of λ that minimizes the CV error.

Alternatively, tuning parameters can also be selected based on variable selection stability, as discussed in Meinshausen and Buhlmann (2010) and Sun et al. (2013). The key idea is to choose tuning parameters that ensure stability in the variable selection process of the penalized regression model. We employ the Kappa Selection Criterion (KSC) proposed by Sun et al. (2013) to select $\lambda_{1,T}$ and $\lambda_{2,T}$. Here, variable selection stability is defined as the expected value of Cohen's kappa coefficient (Cohen, 1960) between active sets obtained from two independent and identical datasets. For instance, consider problem (2.12). Given $\lambda_{1,T} = \lambda$, KSC first estimates the variable selection stability $S(\lambda)$ by randomly partitioning the samples $(\mathbf{y}_t, \widehat{\mathbf{Z}}_{t-1}) : t = 2, \cdots, T$ into two subsets, repeating this process B times. Then, $\lambda_{1,T}$ is cho-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

sen as $\widehat{\lambda}_{1,T} = \min \lambda : \frac{S(\lambda)}{\max_{\lambda'} S(\lambda')} \geq 1 - \alpha_T$, where we set $B = 50$ and $\alpha_T = 0.4$. In summary, there are three methods for selecting tuning parameters in **Algorithm 2**, and their impact on the algorithm's performance is discussed in the following sections.

In our simulations, the coefficient matrices $\mathbf{A} \in \mathbb{R}^{p_1 \times p_1}$ and $\mathbf{B} \in \mathbb{R}^{p_2 \times p_2}$ are generated as follows: for a given dimension p_1, let r_1 be the proportion of nonzero entries in $\mathbf{A}$. For each row of $\mathbf{A}$, $\lfloor \frac{p_1}{2} \rfloor$ entries are generated from $U[1, 2]$, and the remaining $p_1 - \lfloor \frac{p_1}{2} \rfloor$ entries are generated from $U[-2, -1]$. Next, $p_1 - \lfloor r_1 p_1 \rfloor$ entries are set to zero, and the elements are randomly rearranged to form one row of $\mathbf{A}$. Finally, we rescale $\mathbf{A}$ so that $|\mathbf{A}|_F = 1$. The procedure for generating $\mathbf{B}$ follows the same steps as for $\mathbf{A}$, except that $\mathbf{B}$ is rescaled to satisfy $\rho(\mathbf{A})\rho(\mathbf{B}) = 0.9$.

To measure the accuracy of **Algorithm 2** in recovering non-zero elements, we define the following sets for $\mathbf{A} = (a_{i,j})$ and $\widehat{\mathbf{A}} = (\widehat{a}_{i,j})$:

$$S_1 = \{(i,j)|a_{i,j} = 0, \widehat{a}_{i,j} = 0\}, \qquad S_2 = \{(i,j)|a_{i,j} = 0, \widehat{a}_{i,j} \neq 0\},$$
$$S_3 = \{(i,j)|a_{i,j} \neq 0, \widehat{a}_{i,j} \neq 0\}, \qquad S_4 = \{(i,j)|a_{i,j} \neq 0, \widehat{a}_{i,j} = 0\}.$$

The recovery accuracy for non-zero elements in $\mathbf{A}$ is then defined as

$$\text{cr}(\widehat{\mathbf{A}}) = \frac{|S_1| + |S_3|}{p_1 \times p_1}, \tag{4.1}$$

which represents the proportion of correctly estimated zero and non-zero entries in $\widehat{\mathbf{A}}$ relative to $\mathbf{A}$. The recovery accuracy for non-zero elements in $\mathbf{B} = (b_{i,j})$ is defined similarly.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Table S3 reports cr($\widehat{\mathbf{A}}$) and cr($\widehat{\mathbf{B}}$) under different settings and tuning methods from 100 independent replications. **Algorithm 2** shows varying performance depending on the tuning method. Specifically, tuning with sdCV results in the highest accuracy for recovering non-zero elements but also the largest error in $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$. In contrast, tuning with mCV yields the best accuracy for $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$ but the worst recovery accuracy. Finally, tuning with KSC provides a balanced performance, achieving comparable results in both recovery accuracy and the error in $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$, making it a well-rounded choice.

Next, we compare our estimators with those obtained by the ALSE method from Chen et al. (2021) in terms of estimation errors. Figure S3 presents the box plot of $\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$ from 100 independent replications. Detailed results on the estimation errors of $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, and $\widehat{\mathbf{S}}$ using the KSC tuning method are reported in Table S4, while results for the sdCV and mCV tuning methods are shown in Table S5 and Table S6, respectively. Figure S3 and Table S4 indicate that the estimators produced by **Algorithm 2** generally outperform those from the ALSE method, as the Lasso solutions yield smaller estimation errors in most cases. This suggests that the proposed procedure generates more accurate estimators. Additionally, consistent with previous findings, **Algorithm 2** tuned with KSC shows comparable performance to that tuned with mCV, and significantly outperforms the sdCV-tuned version, as seen in Figure S3, Table S4, Table S5, and Table S6. It also performs better than ALSE (Figure S3). Similar conclusions are drawn from simulations conducted in higher-dimensional settings (Figure S4, Table 3, Tables S7- S8).

In summary, **Algorithm 2** tuned by KSC demonstrates satisfactory performance in both

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

recovery accuracy and estimation error, outperforming ALSE with significantly sparser coefficient matrices that are easier to interpret. Therefore, we recommend it for real data applications. From this point forward, we will refer to **Algorithm 2** as **Algorithm 2** tuned by KSC.

Table 3: The estimation errors of the estimators obtained by ALSE and **Algorithm 2** with tuning parameters method KSC.

		Algorithm 2 tunned by KSC				ALSE			
(p_1, p_2)	ρ	T = 100	500	1000	2000	T = 100	500	1000	2000
				$\log(\|\widehat{\mathbf{A}} - \mathbf{A}\|_F)$					
$(12, 15)$	0.5	-2.315	-2.843	-3.218	-3.579	-1.976	-2.449	-2.799	-3.147
$(12, 15)$	0.9	-3.165	-3.648	-4.003	-4.352	-2.763	-3.232	-3.578	-3.924
$(20, 20)$	0.5	-2.349	-2.869	-3.249	-3.617	-1.963	-2.43	-2.775	-3.123
$(20, 20)$	0.9	-3.197	-3.69	-4.059	-4.412	-2.749	-3.218	-3.564	-3.91
				$\log(\|\widehat{\mathbf{B}} - \mathbf{B}\|_F)$					
$(12, 15)$	0.5	0.09	-0.36	-0.706	-1.051	-0.003	-0.48	-0.825	-1.169
$(12, 15)$	0.9	-0.096	-0.556	-0.902	-1.256	-0.217	-0.689	-1.033	-1.373
$(20, 20)$	0.5	0.391	-0.066	-0.409	-0.752	0.281	-0.195	-0.548	-0.895
$(20, 20)$	0.9	0.267	-0.186	-0.526	-0.869	0.078	-0.386	-0.738	-1.084
				$\log(\|\widehat{\mathbf{S}} - \mathbf{S}\|_F)$					
$(12, 15)$	0.5	0.192	-0.265	-0.613	-0.959	0.246	-0.231	-0.578	-0.924
$(12, 15)$	0.9	0.002	-0.459	-0.806	-1.157	0.037	-0.434	-0.779	-1.122
$(20, 20)$	0.5	0.527	0.061	-0.287	-0.633	0.621	0.148	-0.201	-0.549
$(20, 20)$	0.9	0.381	-0.077	-0.422	-0.766	0.419	-0.048	-0.397	-0.743

4.2 Real Data Examples

In this section, we apply the proposed regularized estimation methods to two real-world examples. In the first example, we utilize three iterative algorithms: the ALSE method from Chen et al. (2021), **Algorithm 1**, and **Algorithm 2** to estimate the coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ in Model (2.1). The **Algorithm 2** is tuned by KSC since its interpretability and lower estimation error, see section 4.1. We then examine the out-of-sample forecasting errors produced by the MAR(1) model using parameters estimated by the three approaches. The empirical

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

findings demonstrate that our proposed algorithms achieve smaller out-of-sample forecast error with a high degree of sparsity in modeling the matrix-variate data, resulting in a significant reduction in model parameters. In the second example, we compared the performance of vector auto-regressive model, ALSE from Chen et al. (2021), reduced rank MAR (Xiao et al. (2022)), Dynamic matrix factor models (Yu et al., (2024)) and our proposed **Algorithm 1** and **Algorithm 2** on financial data. The out-of-sample rolling forecast results demonstrate that our methods consistently outperform the competing approaches.

4.2.1 Wind Speed Data

In this example, we apply our methodology to a wind speed dataset consisting of the east–west component of the wind speed vector over a region between latitudes $14°$S and $16°$N and longitudes $145°$E and $175°$E in the western Pacific Ocean. The data records the average wind speed every 6 hours on a 17×17 grid (covering 289 locations) from November 1992 to February 1993, resulting in $T = 480$ and $p_1 = p_2 = 17$. Previous studies by Hsu et al. (2012) and Hsu et al. (2021) have shown evidence of non-stationary spatial dependence, while indicating temporal stationarity with positive temporal correlations.

The observed data was divided into training data $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{400}\}$ and validation data $\{\mathbf{Y}_{401}, \ldots, \mathbf{Y}_{480}\}$. We apply the three iterative estimation methods (ALSE, **Algorithm 1**, and **Algorithm 2**) to obtain the estimated coefficients $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ in the model (2.2). To evaluate the performances of these methods, we calculate the average prediction mean-squared error

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

(PMSE).

$$\text{PMSE} = \frac{1}{289 \times 80} \sum_{t=400}^{479} \|\widehat{\mathbf{Y}}_{t+1} - \mathbf{Y}_{t+1}\|_F,$$

and the average prediction mean-absolute error (PMAE)

$$\text{PMAE} = \frac{1}{289 \times 80} \sum_{t=400}^{479} \|\widehat{\mathbf{Y}}_{t+1} - \mathbf{Y}_{t+1}\|_1,$$

on the validation data, where $\widehat{\mathbf{Y}}_{t+1} = \widehat{\mathbf{A}}\mathbf{Y}_t\widehat{\mathbf{B}}'$. Furthermore, the sparsity of the coefficient matrices estimated by these methods, in terms of the proportions of zero entries in each matrix and the number of iteration steps, is reported in Table 4. From Table 4, we see that our proposed **Algorithm 1** under the banded case and **Algorithm 2** under the sparse case perform better than the ALSE method in terms of both PMSE and PMAE. Moreover, the degree of sparsity of the coefficient matrices estimated by our methods is much higher than that by the ALSE method, which implies that our methods greatly simplify the model. In general, **Algorithm 2** with $\lambda_1 = \lambda_2 = 0.1$ in the Lasso estimation performed best among the three methods. The bandwidths of $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ chosen by the proposed BIC are 4 and 5, respectively. The heat maps of the $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ estimated by the three methods are shown in Figure 1, which clearly illustrate the sparsity of the parameters estimated by our methods.

Table 4: The performance of the three different methods on wind speed data.

Method	PMSE	PMAE	Sparsity of $\widehat{\mathbf{A}}$	Sparsity of $\widehat{\mathbf{B}}$	Iteration step
ALSE	0.16612	0.20007	0	0	45
Algorithm 1	0.16563	0.19983	0.6851	0.7578	8
Algorithm 2	0.16341	0.19665	0.7647	0.8651	6

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

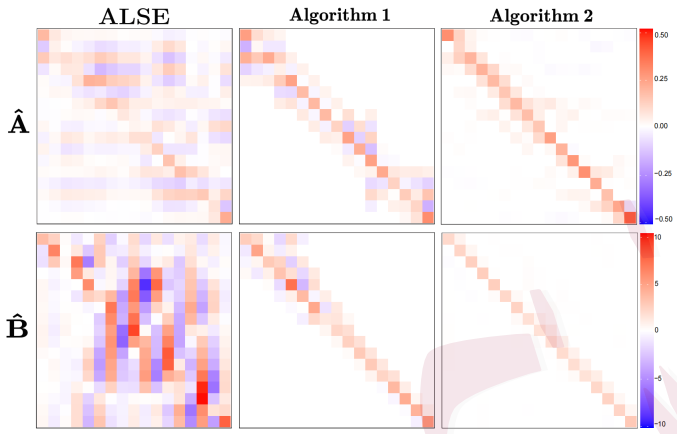

Figure 1: The coefficient matrices estimated by ALSE, **Algorithm 1**, and **Algorithm 2** from U-wind dataset. The first row shows results of $\widehat{\mathbf{A}}$, and the second row shows those of $\widehat{\mathbf{B}}$ obtained by different methods.

4.2.2 Economic Indicator Data

The data in this example consists of quarterly observations of 10 economic indicators for 10 countries. The 10 indicators are the total consumer price index (CPI, growth from the previous period), long-term interest rate (LTIR, first-order difference), short-term interest rate (STIR, first-order difference), total industrial production index (IPI, first-order log difference), manufacturing industrial production index (MIPI, first-order log difference), GDP growth same period previous year (GDPpy, percentage change), GDP growth previous period (GDPpp, percentage change), total exports (EXP, first-order log difference), total imports (IMP, first-order log difference) and unemployment rate forecast (UR, first-order difference). These indicators are sourced from 10 countries: Italy (ITA), Spain (ESP), France (FRA), Germany (DEU), the United Kingdom (GBR), the United States (USA), Canada (CAN), Korea (KOR), Australia (AUS) and Japan (JPN). The dataset spans from the first quarter of 1990 to the fourth quarter of 2019, resulting in a 10×10 matrix-valued time series with a time length of $T = 120$. The data

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

can be obtained from the Organisation for Economic Co-operation and Development (OECD) at https://data.oecd.org/. Figure S5 displays the original data, where rows and columns represent different economic indicators and countries, respectively. To remove seasonal effects, we adjusted each indicator by subtracting its sample quarterly mean.

We compare the out-of-sample rolling forecast performances of the VAR model, MAR model estimated by ALSE, our regularized MAR model, reduced rank MAR model (rrMAR, Xiao et al. (2022)), and the dynamic matrix factor models (DMFM, Yu et al., (2024)). The rolling forecasts are conducted from $t = 100$ to $t = 119$. For each time point, we fit the model using the data $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t$ to obtain the estimated coefficient matrices $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$. We then compute the predictive value $\widehat{\mathbf{Y}}_{t+1} = \widehat{\mathbf{A}}\mathbf{Y}_t\widehat{\mathbf{B}}'$, as well as the 1-norm predictive error $\|\widehat{\mathbf{Y}}_{t+1} - \mathbf{Y}_{t+1}\|_1$ and the F-norm predictive error $\|\widehat{\mathbf{Y}}_{t+1} - \mathbf{Y}_{t+1}\|_F$. The averages of the 1-norm and F-norm errors from $t = 100$ to $t = 119$ for the six methods are reported in Table 5. Notably, both our Lasso iterative method and the banded iterative method outperform the other methods.

Table 5: Out-of-sample rolling forecast performance of the six methods—VAR, MAR estimated by ALSE, **Algorithm 1**, **Algorithm 2**, rrMAR, and DMFM—on economic indicator data.

	VAR	ALSE	rrMAR	DMFM	**Algorithm 1**	**Algorithm 2**
1-norm	7.674	2.534	3.583	2.398	2.221	2.311
F-norm	6.709	2.409	4.140	2.361	2.283	2.338

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

5 Conclusion

In this paper, we studied statistical estimators for high-dimensional matrix-valued autoregressive models under two different settings: when the parameter matrix is banded or sparse. We established the asymptotic properties of these estimators. Both simulations and real data analyses demonstrate the advantages of our new methods over existing ones. The proposed method can be treated as another option in the toolbox for modeling high-dimensional matrix-variate time series and the dynamic models can be useful to practitioners who are interested in out-of-sample forecasting.

Supplementary Material

The online Supplementary Material provides additional simulation results and proofs of the theoretical results.

Acknowledgments

We thank the Editor, Associate Editor, and anonymous referees for their constructive comments and valuable suggestions, which have greatly improved the presentation and quality of this article. Z.G. acknowledges partial support from the National Natural Science Foundation of China (NSFC) under Grant Nos. 12201558, 72573029, and U23A2064, and from the Tianfu Emei Youth Talent Project of Sichuan Province. H.J. acknowledges partial support from the High-level Talent Special Support Program of Zhejiang Province and the National Natural

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Science Foundation of China (No.12531013).

References

Ahn, S. C., and, Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, **81(3)**, 1203–1227.

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.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science and Business Media.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Chen, R., Xiao, H., and Yang, D. (2021). Autoregressive models for matrix-valued time series. *Journal of Econometrics*, **222(1)**, 539–560.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37-46.

Ding, S. and Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B*, **80(2)**, 387–408.

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. *url: https://ideas.repec.org/p/cpr/ceprdp/2400.html*.

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. (2020). Segmenting high-dimensional matrix-valued time series via sequential transformations. *arXiv:2002.03382*.

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. (2021). Modeling high-dimensional unit-root time series. *International Journal of Forecasting*, **37(4)**, 1535–1555.

Gao, Z. and Tsay, R. S. (2022). Modeling high-dimensional time series: a factor model with dynamically dependent factors and diverging eigenvalues. *Journal of the American Statistical Association*, **117(539)**, 1398-1414.

Gao, Z. and Tsay, R. S. (2023a). A two-way transformed factor model for matrix-variate time series. *Econometrics and Statistics*, **27**, 83–101.

Gao, Z. and Tsay, R. S. (2023b). Divide-and-conquer: a distributed hierarchical factor approach to modeling large-scale time series data. *Journal of the American Statistical Association*, **118(544)**, 2698–2711.

Guo, S., Wang, Y. and Yao, Q. (2016). High dimensional and banded vector autoregression. *Biometrika,* **103(4)**, 889–903.

Han, Y., Chen, R., Yang, D., and Zhang, C. H. (2024). Tensor factor model estimation by iterative projection. *The Annals of Statistics*, **52(6)**, 2641–2667.

Han, Y., Yang, D., Zhang, C. H., and Chen, R. (2024). CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **86(5)**, 1383–1413.

Hauglustaine, D. A., and Ehhalt, D. H. (2002). A three-dimensional model of molecular hydrogen in the troposphere. *Journal of Geophysical Research: Atmospheres*, **107(D17)**, ACH-4.

Hsu, N. J., Huang, H. C., and Tsay, R. S. (2021). Matrix autoregressive spatio-temporal mod-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

els. *Journal of Computational and Graphical Statistics*, **30(4)**, 1143–1155.

Hsu, N. J., Chang, Y. M., and Huang, H. C. (2012). A group lasso approach for non-stationary spatial–temporal covariance estimation. *Environmetrics*, **23**, 12-23.

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

Hung, H., Wu, P., Tu, I., and Huang, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, **99(3)**, 569–583.

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.

Li Z. and Xiao H. (2021). Multi-linear tensor autoregressive models. *arXiv:2110.00928*.

Lozano, A. C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., and Abe, N. (2009). Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 587–596.

Meinshausen N. and Buhlmann P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, **72**, 414-473.

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

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

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.

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. *doi: 10.3386/w11467*

Sun, W., Wang, J. H., and Fang, Y. X. (2013). Consistent Selection of Tuning Parameters via Variable Selection Stability. *Journal of Machine Learning Research*, **71(14)**, 3419-3440.

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.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press.

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.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0341

Wang, D., Zheng, Y., and Li, G. (2024). High-dimensional low-rank tensor autoregressive time series modeling. *Journal of Econometrics*, **238(1)**, 105544.

Wang, D., Zheng, Y., Lian, H., and Li, G. (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, **117(539)**, 1338–1356.

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.

Xiao, H., Han, Y., Chen, R., and Liu, C. (2022). Reduced rank autoregressive models for matrix time series. Working paper, available at `https://yuefenghan.github.io/papers/Reduced_Rank_MAR.pdf`.

Yang, D, Ma, Z., and Buja, A. (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. *Journal of Machine Learning Research*, **17(92)**, 1–27.

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, **61(1-3)**, 167–191.

Yu, L., He, Y., Kong, X., and Zhang, X. (2022). Projected estimation for large-dimensional matrix factor models. *Journal of Econometrics,* **229(1)**, 201–217.

Yu, R., Chen, R., Xiao, H., and Han, Y. (2024). Dynamic matrix factor models for high dimensional time series. *arXiv: 2407.05624.*

[1]Center for Data Science, Zhejiang University. E-mail: {jianghj,12335035,12235025}@zju.edu.cn.

[2]School of Mathematical Sciences, University of Electronic Science and Technology of China. E-mail: zhaoxing.gao@uestc.edu.cn.