Autoregressive Models for Matrix-Valued Time Series

Rong Chen, Han Xiao and Dan Yang

Rutgers University

Abstract

In finance, economics and many other fields, observations in a matrix form are often generated over time. For example, a set of key economic indicators are regularly reported in different countries every quarter. The observations at each quarter neatly form a matrix and are observed over many consecutive quarters. Dynamic transport networks with observations generated on the edges can be formed as a matrix observed over time. Although it is natural to turn the matrix observations into a long vector, and then use the standard vector time series models for analysis, it is often the case that the columns and rows of the matrix represent different types of structures that are closely interplayed. In this paper we follow the autoregressive structure for modeling time series and propose a novel matrix autoregressive model in a bilinear form that maintains and utilizes the matrix structure to achieve a greater dimensional reduction, as well as more interpretable results. Probabilistic properties of the models are investigated. Estimation procedures with their theoretical properties are presented and demonstrated with simulated and real examples.

KEYWORDS: Autoregressive; Bilinear; Economic Indicators; Kronecker Product; Multivariate Time Series; Matrix-valued Time Series; Nearest Kronecker Product Projection; Prediction;

1 Rong Chen is Professor at Department of Statistics, Rutgers University, Piscataway, NJ 08854. E-mail: rongchen@stat.rutgers.edu. Han Xiao is Associate Professor at Department of Statistics, Rutgers University, Piscataway, NJ 08854. E-mail: hxiao@stat.rutgers.edu. Dan Yang is Assistant Professor at Department of Statistics, Rutgers University, Piscataway, NJ 08854. E-mail: dyang@stat.rutgers.edu. Rong Chen is the corresponding author. Chen’s research is supported in part by National Science Foundation grants DMS-1503409, DMS-1737857 and IIS-1741390. Xiao’s research is supported in part by a research grant from NEC Labs America. Yang’s research is supported in part under NSF grant IIS-1741390.
1 Introduction

Multivariate time series is a classical area in time series analysis, and has been extensively studied in the literature (see Hannan 1970, Lütkepohl 2005, Tiao and Box 1981, Tsay 2014, for an overview). Recently there has been an emerging interest in modeling high dimensional time series. Roughly speaking these works fall into two major categories: (i) vector autoregressive modeling with regularization (Basu and Michailidis 2015, Davis et al. 2012, Guo et al. 2015, Han et al. 2015, 2016, Kock and Callot 2015, Nardi and Rinaldo 2011, Negahban and Wainwright 2011, Nicholson et al. 2015, Song and Bickel 2011 among others), and (ii) statistical or dynamic factor models (Bai and Ng 2002, Fan et al. 2013, Forni et al. 2005, Lam and Yao 2012, Lam et al. 2011, Wang et al. 2019 among others). In most of these studies, the multiple observations at each time point are treated as a vector.

Although it has been conventional to treat multiple observations as a vector, often the inter-relationship among the time series provides certain structure. For example, Hallin and Liška (2011) studied subpanel structures in multivariate time series, and Tsai and Tsay (2010) considered group constraints among the time series. When the time series are collected under the intersections of two classifications, they naturally form matrices.

Figure 1: Time series of four economic indicators from five countries.
In Figure 1, we plot four economic indicators from five countries, resulting in a $4 \times 5$ matrix observed at each time point. In this example, the rows and columns correspond to different classifications (economic indicators and countries). Univariate time series analysis would deal with individual time series separately (e.g., US interest rate, or UK GDP). Panel time series analysis deals with one row at a time (e.g., interest rates of the five countries), or one column at a time (e.g., all economic indicators of US). Obviously every time series is related to all other time series in the matrix and we wish to model them jointly. It is reasonable to assume that the same economic indicator from different countries (the rows) form a strong relationship, and at the same time economic indicators from the same country (the columns) also naturally move together closely. Hence there is a strong structure in the relationship among the time series. If the matrices are concatenated into vectors, the underlying structure is lost with significant impacts on the model complexity and interpretations.

There has been a recent interest in modeling time series with matrix-valued observations. For example, Wang et al. (2019) considered a factor model for matrix time series. In this paper we propose to model the matrix-valued time series under the autoregressive framework with a bilinear form. Specifically, in this model, the conditional mean of the matrix observation at time $t$ is obtained by multiplying the previous observed matrix at time $t - 1$ from both left and right by two autoregressive coefficient matrices. It can be extended to involve the previous $p$ observed matrices to form an order $p$ autoregressive model. If it involves $p$ previously observed matrices, we call it the matrix autoregressive model of order $p$, with the acronym MAR($p$). Compared with the traditional vector autoregressive models, our approach has two advantages: (i) it keeps the original matrix structure, and its two coefficient matrices have corresponding interpretations; and (ii) it reduces the number of parameters in the model significantly.

Similar bilinear models have been used in regression settings. For instance, Wang et al. (2018) considered the regression with matrix-valued covariates for low-dimensional data. Zhao and Leng (2014) studied the bilinear regression with sparse coefficient vectors under high-dimensional setting. Zhou et al. (2013) and Raskutti et al. (2015) mainly addressed the multi-linear regression with general tensor covariates.

Since the error term in the matrix AR model is also a matrix, its (internal) covariances form a 4-dimensional tensor. Here we also consider to exploit the matrix structure to reduce the dimen-
sionality of this covariance tensor, by separating the row-wise and column-wise dependencies of the error matrix.

In this paper we investigate some probabilistic properties of the proposed model. Several estimators for MAR(1) model are developed, with different computing algorithms, under different assumptions on the error covariance structure. Their asymptotic properties are investigated. We also compare the efficiencies of the estimators. In addition, the finite sample performances of the estimators are demonstrated through simulation studies. The matrix time series of four economic indicators from five countries, shown in Figure 1, is analyzed in detail.

The rest of the paper is organized as follows. We introduce the autoregressive model for matrix-valued time series in Section 2 along with some of its probabilistic properties. The estimation procedures are presented in Section 3. Statistical inferences and the asymptotic properties of the estimators will be considered in Section 4. Numerical studies are carried out in Section 5. Section 6 contains a short summary. All the proofs are collected in Section 7.

2 Autoregressive Model for Matrix-Valued Time Series

Consider a time series of length $T$, in which at each time $t$, a $m \times n$ matrix $X_t$ is observed. Here we use $X$ in boldface to emphasize the fact that it is a matrix. Let vec$(\cdot)$ be the vectorization of a matrix by stacking its columns. The traditional vector autoregressive model (VAR) of order 1 is directly applicable for vec$(X_t)$. That is,

$$\text{vec}(X_t) = \Phi \text{vec}(X_{t-1}) + \text{vec}(E_t). \tag{1}$$

It is immediately seen that the roles of rows and columns are mixed in the VAR model in (1). Using the example shown in Figure 1, the VAR model in (1) fails to recognize the strong connections within the columns (same country) and within the rows (same indicator). The (large) $mn \times mn$ coefficient matrix $\Phi$ does not have any assumed structure; and the model does not fully utilize the matrix structure, or any prior knowledge of the potential relationship among the time series. The coefficient matrix $\Phi$ is also very difficult to interpret.

To overcome the drawback of the direct VAR modeling that requires vectorization, and to take advantage of the original matrix structure, we propose the matrix autoregressive model (of order
1), denoted by MAR(1), in the form
\[ X_t = AX_{t-1}B' + E_t, \]  
where \( A = (a_{ij}) \) and \( B = (b_{ij}) \) are \( m \times m \) and \( n \times n \) autoregressive coefficient matrices, and \( E_t = (e_{t,ij}) \) is a \( m \times n \) matrix white noise. Clearly the model can be extended to an order \( p \) model in the form
\[ X_t = A_1X_{t-1}B'_1 + \cdots + A_pX_{t-p}B'_p + E_t. \]

We will defer the interpretations of \( A \) and \( B \) in (2) to Section 2.1.

Kronecker products can be used to obtain convenient representations of many linear matrix equations that will arise in the subsequent discussions. In fact, using the Kronecker product, the MAR(1) model in (2) can be represented in the form of a vector autoregressive model
\[ \text{vec}(X_t) = (B \otimes A)\text{vec}(X_{t-1}) + \text{vec}(E_t), \]
where \( \otimes \) denotes the matrix Kronecker product. A thorough discussion of the Kronecker product and its relationship with linear matrix equations can be found in Chapter 4 of Horn and Johnson (1994). In Proposition 6 of Section 7, we collect some basic properties of the Kronecker product.

The representation (3) means that the MAR(1) model can be viewed as a special case of the classical VAR(1) model in [1], with its autoregressive coefficient matrix given by a Kronecker product. On the other hand, comparing (1) and (3), we see that the MAR(1) requires \( m^2 + n^2 \) coefficients as the entries of \( A \) and \( B \), while a general VAR(1) needs \( m^2n^2 \) coefficients for \( \Phi \). Of course the latter can be much larger when both \( m \) and \( n \) are large.

There is an obvious identifiability issue with the MAR(1) model in (2), regarding the two coefficient matrices \( A \) and \( B \). The model remains unchanged if the two matrices \( A \) and \( B \) are divided and multiplied by the same nonzero constant respectively. To avoid ambiguity, we make the convention that \( A \) is normalized so its Frobenius norm is one. On the other hand, the uniqueness holds for the Kronecker product \( B \otimes A \).

The error matrix sequence \( \{ E_t \} \) is assumed to be a matrix white noise, i.e. there is no correlation between \( E_t \) and \( E_s \) as long as \( t \neq s \). But \( E_t \) is still allowed to have concurrent correlations among its own entries. As a matrix, its covariances form a 4-dimensional tensor, which is difficult to express. In the following we will discuss it in the form of \( \Sigma = \text{Cov}(\text{vec}(E_t)) \), a \( mn \times mn \) matrix.
In its simplest form, we may assume the entries of $E_t$ are independent so that $\text{Cov}(\text{vec}(E_t))$ is a diagonal matrix, and in a general form, we may allow its entries to have arbitrary correlations. We also consider a structured covariance matrix

\[
\text{Cov}(\text{vec}(E_t)) = \Sigma_c \otimes \Sigma_r,
\]

where $\Sigma_r$ and $\Sigma_c$ are $m \times m$ and $n \times n$ symmetric positive definite matrices. Under normality, this is equivalent to assuming $E_t = \Sigma_r^{1/2} Z_t \Sigma_c^{1/2}$, where all the entries of $Z_t$ are independent, and following the standard Normal distribution. Therefore, $\Sigma_r$ corresponds to row-wise covariances and $\Sigma_c$ introduces column-wise covariances.

**Remarks:** There are many possible extensions of the model. For example, the model can be extended to have multiple lag-one autoregressive terms. That is

\[
X_t = A_1 X_{t-1} B'_1 + \cdots + A_d X_{t-1} B'_d + E_t. \tag{5}
\]

This is still an order-1 autoregressive model, but with more parallel terms. In the stacked vector form, it corresponds to

\[
\text{vec}(X_t) = \left( \sum_{i=1}^d B_i \otimes A_i \right) \text{vec}(X_{t-1}) + \text{vec}(E_t).
\]

Such a structure provides more flexibility to capture the different interactions among rows and columns of the matrix time series, though it becomes more challenging, since there is obviously a more severe identifiability issue.

In this paper we focus on MAR(1) model in all our discussions. Extensions will be investigated elsewhere.

### 2.1 Model interpretations

The MAR(1) model is not a straightforward model. A thorough discussion of the interpretations of its coefficient matrices is needed. Here we offer interpretations from three different angles.

First, in model (2), the left matrix $A$ reflects row-wise interactions, and the right matrix $B'$ introduces column-wise dependence, and therefore the conditional mean in (2) combines the row-wise and column-wise interactions. It is easier to see how the coefficient matrices $A$ and $B$ reflects the row and column structures by looking at a few special cases.
To isolate the effect from the bilinear form in (2), let us assume $A = I$. Then the model reduces to

$$X_t = X_{t-1}B' + E_t.$$  

Consider the example shown in Figure 1, using columns for countries and rows for economic indicators. The conditional expectation of the first column of $X_t$ is given by

$$
\begin{pmatrix}
\text{USA} \\
\text{DEU} \\
\text{CAN}
\end{pmatrix}
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_t
= b_{11}
\begin{pmatrix}
\text{USA} \\
\text{DEU} \\
\text{CAN}
\end{pmatrix}
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_{t-1}
+ b_{12}
\begin{pmatrix}
\text{USA} \\
\text{DEU} \\
\text{CAN}
\end{pmatrix}
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_{t-1}
+ \cdots + b_{1n}
\begin{pmatrix}
\text{USA} \\
\text{DEU} \\
\text{CAN}
\end{pmatrix}
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_{t-1},
$$

which means that at time $t$, the conditional expectation of an economic indicator of one country is a linear combination of the same indicator from all countries at $t - 1$, and this linear combination is the same for different indicators. Therefore, this model (and $B$) captures the column-wise interactions, i.e. interactions among the countries. However, the interactions are confined within each indicator. There are no interactions among the indicators.

On the other hand, if we let $B = I$ in model (2), then a similar interpretation can be obtained, where the matrix $A$ reflects the row-wise interactions, i.e. interactions among the economic indicators within each country. There are no interactions among the countries.

Second, we can interpret the model from a row-wise and column-wise VAR model point of view. For example, if $B = I$, then the model becomes

$$X_t = AX_{t-1} + E_t.$$  

In this case, each column of $X_t$ follows

$$X_{t,j} = AX_{t-1,j} + E_{t,j}, \quad j = 1, \ldots, n.$$  

That is, each column of $X_t$ follows the same VAR(1) model of dimension $m$. Specifically, for the first two columns in the example of Figure 1, the models are
\[
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_t = \begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_{t-1} + \begin{pmatrix}
e\text{Int} \\
e\text{GDP} \\
e\text{Prod} \\
e\text{CPI}
\end{pmatrix}_t
\]

\[
\begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_t = \begin{pmatrix}
\text{Int} \\
\text{GDP} \\
\text{Prod} \\
\text{CPI}
\end{pmatrix}_{t-1} + \begin{pmatrix}
e\text{Int} \\
e\text{GDP} \\
e\text{Prod} \\
e\text{CPI}
\end{pmatrix}_t
\]

In other words, for each country, its economic indicators follow a VAR(1) model (of its own past) of dimension \(m\); and different countries would follow the same VAR(1) model.

If \(A = I\), then

\[
X_t = X_{t-1}B' + E_t.
\]

In this case, each row of \(X_t\) (same indicator from different countries) would follow a VAR(1) model. And the coefficient matrices corresponding to different rows would be the same.

Obviously both models \(X_t = AX_{t-1} + E_t\) and \(X_t = X_{t-1}B' + E_t\) are too restrictive. It is difficult to reason that Germany’s economic indicators follow the same model as the US’s, and there is no interaction between Germany and US. There are two possible ways to add flexibility. One can assume an additive interaction structure to make the model as

\[
X_t = AX_{t-1} + X_{t-1}B' + E_t,
\]

which is essentially a special case of the multi-term model in \([5]\) with \(d = 2\) and \(B_1 = I\) and \(A_2 = I\). Or we can assume a one-term multiplicative interaction structure, which leads to MAR(1). Of course, one can also use

\[
X_t = A_1X_{t-1} + X_{t-1}B'_1 + A_2X_{t-1}B'_2 + E_t,
\]

similar to the model with main effects plus two-way interactions. In this paper we choose to work on MAR(1) in \([2]\).

The third way to interpret MAR(1) is through a defined hierarchical structure. Let \(Y_{t-1} = X_{t-1}B'\). It would be the prediction of \(X_t\) (or the conditional mean) if \(A = I\). Since each column of \(Y_{t-1}\) is based on the linear combination of all columns of \(X_{t-1}\) with no row (indicator) interaction, we can view each entry in \(Y_{t-1}\) as the **globally adjusted indicator**. For example, \(Y_{t-1,GDP,US}\) is
a linear combination of the GDPs of all countries at time $t-1$. Next, we consider $Z_{t-1} = AY_{t-1}$. This would be the prediction of $X_t$ if the model is $X_t = AY_{t-1} + E_t$. It replaces each entry (indicator) in $X_{t-1}$ by its corresponding globally adjusted indicator in $Y_{t-1}$. Each entry in $Z_{t-1}$ is a linear combination of the adjusted indicators from the same country. For example, $Z_{t-1,GDP,US}$ is a linear combination of $Y_{t-1,GDP,US}, Y_{t-1,INT,US}$ etc. It can be viewed as a second adjustment by other indicators (within the same country). Putting everything together, we have

$$X_t = Z_{t-1} + E_t = AY_{t-1} + E_t = AX_{t-1}B' + E_t.$$ 

Note that, if $X_t$ follows the MAR(1) in (2), each entry $x_{t,ij}$ in $X_t$ follows

$$x_{t,ij} = \sum_{k_1=1}^{m} \sum_{k_2=1}^{n} a_{ik_1}x_{t-1,k_1}b_{jk_2} + e_{t,ij}.$$ 

Hence $x_{t,ij}$ is controlled only by $i$-th row of $A$ and $j$-th column of $B'$. In the example of Figure 1, the $i$-th row of $A$ can be viewed as the coefficient corresponding to $i$-th indicator and the $j$-th column of $B'$ as the coefficient corresponding to the $j$-th country. Their values can be interpreted, as we will demonstrate in the real example in Section 4.

The error covariance structure $\text{Cov}(\text{vec}(E_t)) = \Sigma_c \otimes \Sigma_r$ in (4), which consists of all pairwise covariances $\text{Cov}(e_{t,ij}, e_{t,kl}) = \sigma_{c,ij}\sigma_{r,ik}$, has a similar interpretation. For example, if $\Sigma_r = I$, then $E_t = Z\Sigma_c^{1/2}$, which implies that the $\Sigma_c$ matrix captures the concurrent dependence of the columns of shocks in $E_t$. Note that each row of $E_t$ in this case is $E_{t,i} = Z_{t,i}\Sigma_c^{1/2}$. Hence $\text{Cov}(E_{t,i}) = \Sigma_c$ for all rows $i = 1, \ldots, m$, and therefore $\Sigma_c$ captures the covariance among the (column) elements in each row. In parallel, $\Sigma_r$ captures the concurrent dependence among the rows of shocks in $E_t$.

### 2.2 Probabilistic properties of MAR(1)

For any square matrix $C$, we use $\rho(C)$ to denote its spectral radius, which is defined as the maximum modulus of the (complex) eigenvalues of $C$. Since the MAR(1) model can be represented in the form (3), we see that (2) admits a causal and stationary solution if the spectral radius of $B \otimes A$, which is the product of the spectral radii of $A$ and $B$, is strictly less than 1. Hence we have

**Proposition 1.** If $\rho(A) \cdot \rho(B) < 1$, then the MAR(1) model (2) is stationary and causal.

The detailed proof of the proposition will be given in the appendix.
We remark that the property of being “stationary and causal” is referred to as “stable” in Lütkepohl (2005). Here we follow the terminology and definitions used in Brockwell and Davis (1991).

When the condition of Proposition 1 is fulfilled, the MAR(1) model in (2) has the following causal representation after vectorization:

$$\text{vec}(X_t) = \sum_{k=0}^{\infty} \left( B^k \otimes A^k \right) \text{vec}(E_{t-k}).$$

(6)

It follows that the autocovariance matrices of (2) is given by

$$\Gamma_k := \text{Cov}(\text{vec}(X_t), \text{vec}(X_{t-k})) = \sum_{l=0}^{\infty} \left( B^{k+l} \otimes A^{k+l} \right) \Sigma^{l} \left( B^l \otimes A^l \right)^\prime, \quad k \in \mathbb{Z},$$

(7)

where $\Sigma$ is the covariance matrix of vec($E_t$). The condition $\rho(A) \cdot \rho(B) < 1$ guarantees that the infinite matrix series is absolutely summable.

We discuss the impulse response function of the MAR(1) model, following the definition given in Section 2.10 of Tsay (2014) regarding the vector AR models. Let $\Sigma[\cdot, i]$ be the $i$-th column of $\Sigma$, and $\sigma_{ij}$ be the $(i, j)$-th entry of $\Sigma$. Following Tsay (2014), the effect of a unit standard deviation change (which is $\sigma_{m(j-1)+i,m(j-1)+i}$) in $e_{t,ij}$ on the future vec($X_{t+k}$) is given by $(B^k \otimes A^k) \Sigma[\cdot, m(j-1)+i].$

This type of impulse response function exhibits a special structure if $\Sigma$ has the Kronecker product form (4). Let $\Sigma_{r,i}$ and $\Sigma_{c,i}$ be the $i$-th columns of $\Sigma_r$ and $\Sigma_c$, respectively. Then the effect of a unit standard deviation change in $e_{t,ij}$ on the future vec($X_{t+k}$) is given by $(B^k \Sigma_{c,j}) \otimes (A^k \Sigma_{r,i}).$

Consider that case that a one-standard deviation shock occurs at $(1, 1)$ series. Let $f_i^r(k) = (B^k \Sigma_{c,1})[i], \text{ the } i\text{-th element of } B^k \Sigma_{c,1}$ and $f_j^c(k) = (A^k \Sigma_{r,1})[j], \text{ the } j\text{-th element of } A^k \Sigma_{r,1}.$ Then the impulse response function of $(i, j)$-th series at lag $k$ is

$$f_{i,j}(k) = f_i^r(k)f_j^c(k) = (A^k \Sigma_{r,1})[i] \cdot (B^k \Sigma_{c,1})[j].$$

We further let $f^c(k) = [f_1^c(k), \ldots, f_m^c(k)]'$, and $f^r(k) = [f_1^r(k), \ldots, f_n^r(k)]'$. (We use the boldfaced $f(\cdot)$ here to emphasis it is a vector of functions.) We can view $f^c(k)$ as the column response function and $f^r(k)$ as the row response function. Using the example shown in Figure 1 if there is a unit standard deviation shock in the interest rate of US (location $(1, 1)$ in the matrix), then the lag $k$ effect on the four economic indicators for $j$-th country is

$$[f_{1,j}(k), f_{2,j}(k), f_{3,j}(k), f_{4,j}(k)]' = [f_1^r(k), f_2^r(k), f_3^r(k), f_4^r(k)]' f_j^c(k) = f_j^c(k) \cdot f^r(k).$$
which has the following interpretations. The effect of the shock on four economic indicators in the same \(j\)-th country, a 4-dimensional vector \([f_{1,j}(k), f_{2,j}(k), f_{3,j}(k), f_{4,j}(k)]'\), is proportional to the vector \(f^r(k)\), for all countries \(1 \leq j \leq 5\). Hence the five (4-dimensional economic indicator) vectors corresponding to the five countries are parallel to each other and only differ by the multiplier \(f^c_j(k)\). This form of impulse response function implies that the economies of different countries have a co-movement as responses to a shock, but the impacts on different countries are of different scales. Similarly, we have

\[
[f_{i,1}(k), \ldots, f_{i,5}(k)] = f^r_i(k) \cdot [f^c_1(k), \ldots, f^c_5(k)] = f^r_i(k) \cdot [f^c(k)]', \quad 1 \leq i \leq 4.
\]

That is, the effect on five countries regarding each economic indicator, which is a 5-dimensional row vector, is proportional to \(f^c(k)\), and the four vectors corresponding to four indicators only differ by lengths \(f^r_i(k)\).

In general, for a shock occurs at location \((i,j)\), the lag-\(k\) row and column response functions are \(f^{c,j}(k) = B^{k} \Sigma_{c,j}\) and \(f^{r,i}(k) = A^{k} \Sigma_{r,i}\), respectively. In fact, the response function in matrix form in this case is given by the rank-one matrix:

\[
\text{IR}_{i,j}(k) = f^{r,i}(k)[f^{c,j}(k)]'.
\]

## 3 Estimation

### 3.1 Projection method

To estimate the coefficient matrices \(A\) and \(B\), our first approach is to view the MAR(1) model in (2) as the structured VAR(1) model in (3). We first obtain the maximum likelihood estimate or the least square estimate \(\hat{\Phi}\) of \(\Phi\) in (1) without the structure constraint, then we find the estimators by projecting \(\hat{\Phi}\) onto the space of Kronecker products under the Frobenius norm:

\[
(\hat{A}_1, \hat{B}_1) = \arg \min_{A, B} \| \hat{\Phi} - B \otimes A \|_F^2.
\]

This minimization problem is called the nearest Kronecker product (NKP) problem in matrix computation \(\text{[Van Loan, 2000; Van Loan and Pitsianis, 1993]}\). It turns out that an explicit solution exists, which can be obtained through a singular value decomposition (SVD) of a rearranged version of \(\hat{\Phi}\).
We note that the set of all entries in \( B \otimes A \) is exactly the same as the set of all entries in \( \text{vec}(A)\text{vec}(B)' \). The two matrices have the same set of elements, and only differ by the placement of the elements in the matrices. Define a re-arrangement operator \( \mathcal{G} : \mathbb{R}^{mn} \times \mathbb{R}^{mn} \rightarrow \mathbb{R}^{m^2} \times \mathbb{R}^{n^2} \) such that

\[
\mathcal{G}(B \otimes A) = \text{vec}(A)\text{vec}(B)'.
\]

It is easy to see that the operator is a linear operator such that \( \mathcal{G}(C_1 - C_2) = \mathcal{G}(C_1) - \mathcal{G}(C_2) \). We also note that the Frobenius norm of a matrix only depends on the elements in the matrix, but not the arrangement, hence \( ||\mathcal{G}(C)||_F = ||C||_F \). Then we have

\[
\min_{A,B} ||\hat{\Phi} - B \otimes A||_F^2 = \min_{A,B} ||\mathcal{G}(\hat{\Phi}) - \mathcal{G}(B \otimes A)||_F^2 = \min_{A,B} ||\mathcal{G}(\hat{\Phi}) - \text{vec}(A)\text{vec}(B)'||_F^2 = \min_{A,B} ||\hat{\Phi} - \text{vec}(A)\text{vec}(B)'||_F^2,
\]

where \( \hat{\Phi} = \mathcal{G}(\hat{\Phi}) \) is the re-arranged \( \hat{\Phi} \). It follows that the solution of (8) can be obtained through

\[
\text{vec}(\hat{A})\text{vec}(\hat{B})' = d_1 u_1 v_1',
\]

where \( d_1 \) is the largest singular value of \( \hat{\Phi} \), and \( u_1 \) and \( v_1 \) are the corresponding first left and right singular vectors, respectively. By converting the vectors into matrices, we obtain corresponding estimators of \( A \) and \( B \), denoted by \( \hat{A}_1 \) and \( \hat{B}_1 \), with the normalization that \( ||\hat{A}_1||_F = 1 \). We call them projection estimators, and will use the acronym PROJ for later references.

We illustrate the re-arrangement operation with a special case of \( m = n = 2 \). We first rearrange the entries of the Kronecker product \( B \otimes A \):

\[
\begin{bmatrix}
  b_{11}a_{11} & b_{11}a_{12} & b_{12}a_{11} & b_{12}a_{12} \\
  b_{11}a_{21} & b_{11}a_{22} & b_{12}a_{21} & b_{12}a_{22} \\
  b_{21}a_{11} & b_{21}a_{12} & b_{22}a_{11} & b_{22}a_{12} \\
  b_{21}a_{21} & b_{21}a_{22} & b_{22}a_{21} & b_{22}a_{22}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
  b_{11}a_{11} & b_{12}a_{11} & b_{12}a_{11} & b_{22}a_{11} \\
  b_{11}a_{21} & b_{12}a_{21} & b_{12}a_{21} & b_{22}a_{21} \\
  b_{11}a_{12} & b_{12}a_{12} & b_{12}a_{12} & b_{22}a_{12} \\
  b_{11}a_{22} & b_{12}a_{22} & b_{12}a_{22} & b_{22}a_{22}
\end{bmatrix}
\]

We then rearrange the entries of \( \hat{\Phi} \) in exactly the same way:

\[
\hat{\Phi} = \begin{bmatrix}
  \phi_{11} & \phi_{12} & \phi_{13} & \phi_{14} \\
  \phi_{21} & \phi_{22} & \phi_{23} & \phi_{24} \\
  \phi_{31} & \phi_{32} & \phi_{33} & \phi_{34} \\
  \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44}
\end{bmatrix}
\rightarrow
\begin{bmatrix}
  \phi_{11} & \phi_{31} & \phi_{13} & \phi_{33} \\
  \phi_{21} & \phi_{41} & \phi_{23} & \phi_{43} \\
  \phi_{12} & \phi_{32} & \phi_{14} & \phi_{34} \\
  \phi_{22} & \phi_{42} & \phi_{24} & \phi_{44}
\end{bmatrix} =: \tilde{\Phi}.
\]
By abuse of notation, we omit the hat on each individual $\hat{\phi}_{ij}$. Now it is clear that the NKP problem (8) is equivalent to $\min_{A,B} \| \tilde{\Phi} - \text{vec}(A)\text{vec}(B)' \|^2_F$.

Note that this procedure requires the estimation of the $mn \times mn$ coefficient matrix $\Phi$ first. This task is often formidable and inaccurate with moderately large $m$ and $n$ and a finite sample size. Hence the resulting projection estimator may not be accurate. However, it can serve as the initial value for a more elaborate iterative procedure.

### 3.2 Iterated least squares

If we assume the entries of $E_t$ are i.i.d. normal with mean zero and a constant variance, the maximum likelihood estimator, denoted by $\hat{A}_2$ and $\hat{B}_2$, is the solution of the least squares problem

$$
\min_{A,B} \sum_t \| X_t - AX_{t-1}B' \|^2_F.
$$

We refer to this estimator as LSE for the rest of this paper. If the error covariance matrix is arbitrary, the LSE is still an intuitive and reasonable estimator. To see the connection between the two estimators PROJ and LSE, we define

$$
\mathbf{Y} = [\text{vec}(X_2), \text{vec}(X_3), \ldots, \text{vec}(X_T)],
$$

$$
\mathbf{X} = [\text{vec}(X_1), \text{vec}(X_2), \ldots, \text{vec}(X_{T-1})].
$$

The minimization problem (9) can be rewritten as

$$
\min_{A,B} \| \mathbf{Y} - (B \otimes A)\mathbf{X} \|^2_F.
$$

Comparing (10) and (8), we see the problem (10) can be viewed as an inverse NKP problem. Unfortunately it does not have an explicit SVD solution (Van Loan, 2000).

There is another way to understand the minimization problem (9). Define

$$
\mathbf{Y}' = [X'_2, X'_3, \ldots, X'_T],
$$

$$
\mathbf{X}'_A = [X'_1A', X'_2A', \ldots, X'_{T-1}A'].
$$

With these notations, the least squares problem (9) is equivalent to

$$
\min_A \left\{ \min_B \| \mathbf{Y}' - \mathbf{X}'_A B \|^2_F \right\}.
$$

13
In other words, we aim to find the optimal $A$, so that the projection of the columns of $Y$ on the column space of $X_A$ is maximized.

Taking partial derivatives of (9) with respect to the entries of $A$ and $B$ respectively, we obtain

$$\sum_t AX_t B'BX_t' - \sum_t X_tBX_t' = 0$$

$$\sum_t BX_t'A'AX_t - \sum_t X_t'AX_t = 0.$$

(13)

The function $\sum_t \|X_t - AX_t B'\|^2_F$ is guaranteed to have at least one global minimum, so solutions to (13) are guaranteed to exist. On the other hand, if $\hat{A}$ and $\hat{B}$ solve the equations in (13), so are $\tilde{A} := \hat{A}/c$ and $\tilde{B} := \hat{B} \cdot c$, where $c$ is any nonzero constant. We should regard them as the same solution because they yield the same matrix product $\hat{A}X_t B' = \tilde{A}X_t B'$. Equivalently, we say that $(\hat{A}, \hat{B})$ and $(\tilde{A}, \tilde{B})$ are the same solution of (13), if $\hat{B} \otimes \hat{A} = \tilde{B} \otimes \tilde{A}$. With this convention, we argue that with probability one, the global minimum of (9) is unique. For this purpose, we need the following condition.

(Condition: R:) The innovations $E_t$ are independent and identically distributed, and absolutely continuous with respect to Lebesque measure.

If Condition (R) is fulfilled, it holds that with probability one, the solutions of (13) have full ranks, and they have no zero entries. Let us restrict our discussion to this event of probability one. Without loss of generality, we fix the first entry of $A$ at $a_{11} = 1$. Let us use $Z$ to denote the set of entries of $A$ and $B$:

$$Z := \{a_{ij}, b_{kl} : 1 \leq i, j \leq m, (i,j) \neq (1,1), 1 \leq k, l \leq n\}.$$ 

The matrix equations in (13) involves $m^2 + n^2$ individual equations, and each equation takes the form $f(Z) = 0$, where $f(Z)$ is a multivariate polynomial in the polynomial ring $\mathbb{C}[Z]$ over the complex field $\mathbb{C}$. The collection $V$ of all solutions of (13) is thus an affine variety in the space $\mathbb{C}^{m^2+n^2-1}$. By computing a Groebner basis for the ideal generated by the polynomials in (13), we see that $V$ is a finite set (see for example, Theorem 6, page 251 of Cox et al., 2015). Equivalently, the equations in (13) have a finite number of solutions.

Now consider the uniqueness of the global minimum. Assume $A_1$ and $A_2$ are two nonzero $m \times m$ matrices such that $A_1 \neq c \cdot A_2$ for any constant $c \in \mathbb{R}$. Define $X_1$ and $X_2$ as in (11) with $A_1$ and $A_2$ respectively. We further define $X(c) = cX_1 + (1 - c)X_2$. The projection of $Y$ on the
column space of $\mathcal{X}(c)$ is given by $\mathcal{X}(c)[\mathcal{X}(c)']^{-1}\mathcal{X}(c)'\mathcal{Y}$, and its Frobenius norm

$$\text{trace}\left\{\mathcal{Y}'\mathcal{X}(c)[\mathcal{X}(c)']^{-1}\mathcal{X}(c)'\mathcal{Y}\right\}$$  \hspace{1cm} (14)

is a rational function of $c \in \mathbb{R}$ with random coefficients determined by $X_1, X_2, \ldots, X_T$. With probability one, this rational function takes distinct values at different local minima. Combining this fact and the preceding argument, we see that with probability one, the least squares problem (9) has an unique global minimum, and finitely many local minima.

To solve (9), we iteratively update the two matrices $\hat{A}$ and $\hat{B}$, by updating one of them in the least squares (9) while holding the other one fixed, starting with some initial $A$ and $B$. By (13) the iteration of updating $B$ given $A$ is

$$B \leftarrow \left(\sum_t X_t'AX_{t-1}\right)\left(\sum_t X_t'A'AX_{t-1}\right)^{-1},$$

and similarly by (13) the iteration of updating $A$ given $B$ is

$$A \leftarrow \left(\sum_t X_tBX_{t-1}'\right)\left(\sum_t X_{t-1}B'BX_{t-1}'\right)^{-1}.$$  \hspace{1cm} (13)

We denote these estimators by $\hat{A}_2$ and $\hat{B}_2$, with the name least squares estimators, and the acronym LSE.

The iterative least squares may converge to a local minimum. In practice, we suggest to use the PROJ estimators $\hat{A}_1$ and $\hat{B}_1$ as the starting values of the iterations. On the other hand, by permuting the entries of the corresponding matrices (Van Loan 2000), the problem (9) can be rewritten as a problem of best rank-one approximation under a linear transform, which in turn can be viewed as a generalized SVD problem. The variable projection methods discussed in Golub and Pereyra (1973) and Kaufman (1975) may also be applicable here.

3.3 MLE under a structured covariance tensor

When the covariance matrix of the error matrix $E_t$ assumes the structure in (4), it can be utilized to improve the efficiency of the estimators. The log likelihood under normality can be written as

$$-m(T-1)\log|\Sigma_c| - n(T-1)\log|\Sigma_r| - \sum_t \text{tr}\left(\Sigma_r^{-1}(X_t - AX_{t-1}B')\Sigma_c^{-1}(X_t - AX_{t-1}B')'\right).$$
Four matrix parameters $A, B, \Sigma_r, \Sigma_c$ are involved in the log likelihood function. The gradient condition at the MLE is given by

$$A \sum_t X_{t-1}' B \Sigma_c^{-1} B X_{t-1}' - \sum_t X_t \Sigma_c^{-1} B X_{t-1}' = 0$$

$$B \sum_t X_{t-1}' A \Sigma_r^{-1} A X_{t-1}' - \sum_t X_t \Sigma_r^{-1} A X_{t-1}' = 0$$

$m(T-1)\Sigma_c - \sum_t (X_t - AX_{t-1}') \Sigma_r^{-1} (X_t - AX_{t-1}') = 0$

$n(T-1)\Sigma_r - \sum_t (X_t - AX_{t-1}') \Sigma_c^{-1} (X_t - AX_{t-1}')' = 0$.

To find the MLE, we iteratively update one, while keeping the other three fixed. These iterations are given by

$$A \leftarrow \left( \sum_t X_t \Sigma_c^{-1} B X_{t-1}' \right)^{-1} \left( \sum_t X_{t-1} B \Sigma_c^{-1} B X_{t-1}' \right)$$

$$B \leftarrow \left( \sum_t X_t' \Sigma_r^{-1} A X_{t-1} \right)^{-1} \left( \sum_t X_{t-1}' A \Sigma_r^{-1} A X_{t-1} \right)$$

$$\Sigma_c \leftarrow \frac{\sum_t R_t' \Sigma_r^{-1} R_t}{m(T-1)}$$, where $R_t = X_t - AX_{t-1}'$

$$\Sigma_r \leftarrow \frac{\sum_t R_t \Sigma_c^{-1} R_t'}{n(T-1)}$$, where $R_t = X_t - AX_{t-1}'$

The MLE for $A$ and $B$ under the covariance structure (4) will be denoted by $\hat{A}_3$ and $\hat{B}_3$, with an acronym MLEs, where the “s” stands for the special structure (4) of the error covariance matrix. Due to the unidentifiability of the pairs of $A, B$ and $\Sigma_c, \Sigma_r$, to make sure the numerical computation is stable, after looping through $A, B, \Sigma_c$ and $\Sigma_r$ in each iteration, we renormalize so that both $||A||_F = 1$ and $||\Sigma_r||_F = 1$.

4 Asymptotics and Efficiency

Due to the identifiability issue regarding $A$ and $B$, we make the convention that $||A||_F = 1$, and the three estimators ($\hat{A}_i, \hat{B}_i$), $1 \leq i \leq 3$ are rescaled so that $||\hat{A}_i||_F = 1$. Since the Kronecker product $B \otimes A$ is unique, we also state the asymptotic distributions of the estimated Kronecker product $B \otimes A$, in addition to that of $A$ and $B$.

We first present the central limit theorem for the projection estimators $\hat{A}_1$ and $\hat{B}_1$. Following standard theory of multivariate ARMA models [Dunsmuir and Hannan 1976, Hannan 1970], the
conditions of Theorem 2 guarantees that \( \hat{\Phi} \) converges to a multivariate normal distribution:

\[
\sqrt{T} \begin{pmatrix} \text{vec}(\hat{\Phi} - B \otimes A) \\ \text{vec}(\hat{\Phi} - B' \otimes A') \end{pmatrix} \Rightarrow N(0, \Gamma_0^{-1} \otimes \Sigma),
\]

where \( \Sigma \) is the covariance matrix of vec\((E_t)\), and \( \Gamma_0 \) is given in (7). Let \( \tilde{\Phi} = \mathcal{G}(\hat{\Phi}) \) be the rearranged version of \( \hat{\Phi} \), and \( \Xi_1 \) be the asymptotic covariance matrix of vec\((\tilde{\Phi})\). The matrix \( \Xi_1 \) is obtained by rearranging the entries of \( \Gamma_0^{-1} \otimes \Sigma \), and can be expressed using permutation matrices and Kronecker products, but we omit the explicit formula here.

**Theorem 2.** Set \( \alpha_1 := \text{vec}(A)/\|\text{vec}(A)\|, \beta_1 := \text{vec}(B)/\|\text{vec}(B)\|, \) and

\[
V_0 := \begin{pmatrix} \|\beta\|^{-1}[\beta_1' \otimes (I - \alpha_1\alpha_1')] \\ I \otimes \alpha_1 \end{pmatrix},
\]

\[
V_1 := (\beta_1\beta_1') \otimes I + I \otimes (\alpha_1\alpha_1') - (\beta_1\beta_1') \otimes (\alpha_1\alpha_1').
\]

Assume that \( E_1, \ldots, E_T \) are iid with mean zero and finite second moments. Also assume the stationarity condition \( \rho(A) \cdot \rho(B) < 1 \), and \( A, B \) and \( \Sigma \) are nonsingular. It holds that

\[
\sqrt{T} \begin{pmatrix} \text{vec}(\hat{A}_1 - A) \\ \text{vec}(\hat{B}_1 - B) \end{pmatrix} \Rightarrow N(0, V_0\Xi_1 V_0'),
\]

and

\[
\sqrt{T} \begin{pmatrix} \text{vec}(\hat{B}_1) \otimes \text{vec}(\hat{A}_1) - \text{vec}(B) \otimes \text{vec}(A) \end{pmatrix} \Rightarrow N(0, V_1\Xi_1 V_1').
\]

The proof of the theorem is in Section 7.

Now we consider the least squares estimators \((\hat{A}_2, \hat{B}_2)\). Let \( \alpha := \text{vec}(A), \beta := \text{vec}(B'), \gamma' := (\alpha', 0) \) be a vector in \( \mathbb{R}^{m^2+n^2} \). Note that \( \beta \) should not be confused with \( \beta_1 \) defined in Theorem 2, which is related to the vectorization of \( B \). Recall that \( \Sigma \) is the covariance matrix of vec\((E_t)\). We have the following result for \( \hat{A}_2 \) and \( \hat{B}_2 \).

**Theorem 3.** Define \( W_t' := [(BX_t') \otimes I : I \otimes (AX_t)] \), and \( H := \mathbb{E}(W_t W_t') + \gamma' \gamma' \). Let \( \Xi_2 := H^{-1}\mathbb{E}(W_t \Sigma W_t')H^{-1} \), and \( V := [\beta \otimes I, I \otimes \alpha] \). In addition to the conditions of Theorem 2, we also assume (R). It holds that

\[
\sqrt{T} \begin{pmatrix} \text{vec}(\hat{A}_2 - A) \\ \text{vec}(\hat{B}_2 - B) \end{pmatrix} \Rightarrow N(0, \Xi_2);
\]

and equivalently,

\[
\sqrt{T} \begin{pmatrix} \text{vec}(\hat{B}_2') \otimes \text{vec}(\hat{A}_2) - \text{vec}(B') \otimes \text{vec}(A) \end{pmatrix} \Rightarrow N(0, V\Xi_2 V').
\]
The proof of the theorem is in Section 7.

With the additional assumption (4) on the covariance structure of \( E_t \), we have a similar result. Recall that \( W'_t = [(BX'_t) \otimes I, I \otimes (AX_t)] \), and \( \gamma' = (\alpha', 0')' \). Let \( \hat{H} := E(W_t \Sigma^{-1} W'_t) + \gamma \gamma' \). Define \( \Xi_3 := \hat{H}^{-1} E(W_t \Sigma^{-1} W'_t) \hat{H}^{-1} \). Note that the covariance matrix \( \Sigma \) takes the form \( \Sigma = \Sigma_c \otimes \Sigma_r \).

The MLEs \( \hat{A}_3 \) and \( \hat{B}_3 \) under the assumption (4) have the following joint limiting distribution.

**Theorem 4.** Under the same conditions of Theorem 3, it holds that

\[
\sqrt{T} \begin{pmatrix}
\text{vec}(\hat{A}_3 - A) \\
\text{vec}(\hat{B}'_3 - B')
\end{pmatrix} \Rightarrow N(0, \Xi_3);
\]

and equivalently,

\[
\sqrt{T} \begin{pmatrix}
\text{vec}(\hat{B}'_3) \otimes \text{vec}(\hat{A}_3) - \text{vec}(B') \otimes \text{vec}(A)
\end{pmatrix} \Rightarrow N(0, V\Xi_3 V').
\]

The proof of the theorem is in Section 7.

We remark that in these three versions of the central limit theorem, there may be zero diagonal entries in the asymptotic covariance matrix. For example, in Theorem 3 there may be a zero on the diagonal of the matrix \( V\Xi_2 V' \). It happens when the corresponding true values of the entries \( a_{i1j1} \) and \( b_{i2j2} \) are both zero; and in this situation, the product estimator \( \hat{a}_{i1j1} \hat{b}_{i2j2} \) has a convergence rate of \( 1/T \) instead of \( 1/\sqrt{T} \).

We now compare the efficiencies of the LSE and the MLEs, when the covariance matrix of \( \text{vec}(E_t) \) has the Kronecker product structure (4). In Theorems 3 and 4, both asymptotic covariance matrices take the form \( V\Xi_i V', \ i = 2, 3 \). The following corollary asserts that the MLEs is asymptotically more efficient under the structured covariance matrix (4).

**Corollary 5.** Under the same conditions of Theorem 3 and assume that the covariance matrix of \( \text{vec}(E_t) \) takes the form (4), then it holds that

\[ \Xi_2 \geq \Xi_3. \]

The proof of the Corollary is in Section 7.

Here the matrix relationship \( \geq \) means that the difference of the two matrices is positive semi-definite. Consequently, we see that when the covariance structure is correctly specified by (4), the MLEs \( \hat{A}_3 \) and \( \hat{B}_3 \) are more efficient than the LSE \( \hat{A}_2 \) and \( \hat{B}_2 \) asymptotically. A comparison of the
efficiencies of the projection estimators and least squares estimators can also be made, where the least squares estimators are more efficient. However, we omit this result here, because in practice we suggest to use either LSE or MLEs, and only use PROJ as initial values for the other two estimation methods.

5 Numerical Results

5.1 Simulations

Figure 2: Comparison of four estimators, PROJ, LSE, MLEs, and VAR, under Setting I. The three rows correspond to \((m, n) = (3, 2), (6, 4)\) and \((9, 6)\) respectively, and the four columns \(T = 100, 200, 400\) and \(5000\) respectively.

In this section, we compare the performances of the aforementioned estimators and the stacked VAR(1) estimator under different settings for various choices of the matrix dimensions \(m\) and \(n\), as well as the length of the time series \(T\).

Specifically, for given dimensions \(m\) and \(n\), the observed data \(X_t\) are simulated according
Figure 3: Comparison of four estimators, PROJ, LSE, MLEs, and VAR, under Setting II. The three rows correspond to $(m,n) = (3,2), (6,4)$ and $(9,6)$ respectively, and the four columns $T = 100, 200, 400$ and $5000$ respectively.

In what follows, we perform four experiments: the first three experiments demonstrate the finite-sample comparisons under three settings of the covariance structure of the innovation matrix $E_t$ respectively, and the fourth one compares the asymptotic properties of all estimators when $T \to \infty$ under these three settings.

- **Setting I**: The covariance matrix $\text{Cov}(\text{vec}(E_t))$ is set to $\Sigma = I$.

- **Setting II**: The covariance matrix $\text{Cov}(\text{vec}(E_t)) = \Sigma$ is randomly generated according to $\text{Cov}(\text{vec}(E_t)) = \Lambda Q Q'$, where the eigenvalues in the diagonal matrix $\Lambda$ are the absolute...
values of i.i.d. standard normal random variates, and the eigenvector matrix $Q$ is a random orthonormal matrix.

- Setting III: The covariance matrix $\text{Cov}(\text{vec}(E_t))$ takes the Kronecker product form (4), where $\Sigma_c$ and $\Sigma_r$ are generated similarly as the $\Sigma$ in Setting II.

In addition to the three estimators (PROJ, LSE, and MLEs) discussed in Section 3, we also include the estimator under the stacked VAR(1) model in (1) as a benchmark for comparison. For each configuration, we repeat the simulation 100 times, and show a box plot of

$$\log(\|\hat{B} \otimes \hat{A} - B \otimes A\|_F^2).$$

Figures 2 to 4 show the simulation results under three settings respectively for relatively small sample sizes. For each of these three figures, the dimensions $m, n$ increase from top to bottom,
Figure 5: Comparison of the asymptotic efficiencies of three estimators, PROJ, LSE, and MLEs, under three settings. The left column shows the average error over 100 repetitions for $\|\hat{B} \otimes \hat{A} - B \otimes A\|_F$ and the right for $T\|\hat{B} \otimes \hat{A} - B \otimes A\|_F$.

taking values in $(m, n) = (3, 2), (6, 4)$ and $(9, 6)$. The sample size $T$ increases from left to right at $T = 100, 200, 400$ and 5000, respectively. One common finding from these three figures is that all three estimators, PROJ, LSE, and MLEs, obtained under the MAR(1) model in (2) outperform the stacked VAR estimator.

In the first experiment under Setting I, Figure [2] shows that LSE is the best estimator when the covariance matrix is indeed a diagonal matrix. This is intuitive since LSE is the maximum likelihood estimator under this setting. The very close second best is the MLEs, which is comparable with LSE throughout different combinations of $m, n, T$ and only performs slightly worse when $m, n$ are large and $T$ is small. This is expected since MLEs has to estimate the additional row and column covariance matrices of sizes $m \times m$ and $n \times n$ when it is not necessary. Both LSE and MLEs are superior over the PROJ, especially when the sample size is small and the dimensions are large as seen in the lower left corner of the figure. As sample size increases, the advantage becomes less obvious.
In the second experiment under Setting II, the overall pattern of Figure 3 is similar to that in Figure 2. The VAR estimator performs the worst; PROJ is the second worst, and LSE and MLEs are very much similar. Note that under Setting II, the covariance structure is arbitrary and does not follow the Kronecker structure, which is the underlying assumption for the MLEs. The LSE does not assume any covariance structure in the estimation process. Hence one would expect MLEs, obtained under the wrong assumption, should perform worse than LSE. However, the simulation results show that they perform similarly.

In the third experiment under Setting III, Figure 4 shows that MLEs dominates LSE for any choice of $m$, $n$, $T$, as Corollary 5 predicts. LSE in turn prevails PROJ, which in turn always leads the stacked VAR estimator.

In the fourth experiment, we compare the asymptotic efficiencies of PROJ, LSE, and MLEs by letting the length of the time series $T$ go to infinity. The main purpose of this experiment is to obtain qualitative understanding of the asymptotic covariances of different estimators. Although Theorems 2 to 4 provide the theoretical form of the asymptotic covariances and Corollary 5 ascertains the relative magnitude of the errors from LSE and MLEs under Setting III, we have little concrete insight on the relative performances of the three estimators under other settings. For this purpose, we fix the dimensions $(m, n) = (3, 2)$ for all three settings in this experiment. Figure 5 shows the results. The left three panels show the average estimation errors over 100 repetitions of $\| \hat{B} \otimes \hat{A} - B \otimes A \|_F^2$ for different $T$. The right panels shows $T \| \hat{B} \otimes \hat{A} - B \otimes A \|_F^2$ as a function of $T$. The three rows correspond to the three settings respectively. In each of the six panels, the solid line, dashed line, and dotted line correspond to PROJ, LSE, and MLEs respectively. The three figures on the left show the decreasing trend of all three estimators as $T$ grows. The three figures on the right magnify the differences of the three estimators, with a clear ordering. PROJ estimator clearly has the lowest efficiency. Under Setting I (top panels), LSE and MLEs performs similarly, as both are maximum likelihood estimators under the setting. MLEs estimates total 4 more parameters in $\Sigma_r$ and $\Sigma_c$. The bottom panels in the figure show that MLEs is more efficient than LSE under Setting III, which is expected by Corollary 5. However, under Setting II (the middle panels), it is interesting to observe that MLEs outperforms LSE slightly but consistently, although MLEs is obtained under a wrong model assumption. This is probably due to the use of a regularized covariance structure (4), which is beneficial because of the dimension reduction from 21 parameters in an arbitrary $\Sigma$.
to $\Sigma_c \otimes \Sigma_r$. Note that the performance also depends on how close the Kronecker product approximation is to the arbitrary (random) covariance matrix we used in the simulation.

### 5.2 Economic indicator from five countries

We now revisit the example shown in Figure 1. The data consists of quarterly observations of four economic indicators: interest rate, GDP growth, Production growth and consumer price index, from five countries: Canada, France, Germany, United Kingdom and United States. It ranges from 1990 to 2016. The data was obtained from Organisation for Economic Co-operation and Development (OECD) at [https://data.oecd.org/](https://data.oecd.org/). Before fitting the autoregressive models, we adjusted the seasonality of CPI by subtracting the sample quarterly means. All series are normalized so that the combined variance of each indicator (each row) is 1.

MAR(1) model was estimated using the three estimation methods. We also fitted a stacked VAR(1) model, and univariate AR(1) and AR(2) models for each individual time series. The residual sum of squares of each model and the sum of squares of the (normalized) original data are listed in Table 1. The MAR(1) estimated using the least squares method has the smallest residual sum of squares, among all models and methods, except the VAR(1) model. Note that MAR(1) model uses $16 + 25 - 1 = 40$ parameters in the two coefficient matrices, comparing to 20 and 40 parameters in fitting 20 univariate AR(1) and AR(2) models to each series, respectively. The VAR(1) model has total 400 parameters in the AR coefficient matrix. The large number of parameters results in a small residual sum of squares. It is deemed to be overfitting as we will show later in out-sample rolling forecasting performance evaluation.

| MAR(1)PROJ | MAR(1)LSE | MAR(1)MLEs | VAR(1) | iAR(1) | iAR(2) | original |
|------------|-----------|------------|--------|--------|--------|----------|
| 1572       | 1468      | 1487       | 1121   | 1724   | 1663   | 2116     |

Table 1: Residual sum of squares of MAR(1) model using three different estimators and the stacked VAR(1) estimator; and the total residual sum of squares of fitting univariate AR(1) and AR(2) to each individual time series; and the total sum of squares of the original (normalized) data.

Tables 2 and 3 show the estimated parameters and their corresponding standard errors (in the parentheses) of $A$ and $B$ using the least squares method. Due to ambiguity between the two matrices, the left matrix is scaled so that its Frobenius norm is one. On the right of the table we
also indicate the positively significant, negatively significant and insignificant parameters (at 10% level) using symbols (+, −, 0), respectively.

|       | Int  | GDP    | Prod   | CPI   | Int  | GDP    | Prod   | CPI   |
|-------|------|--------|--------|-------|------|--------|--------|-------|
| Int   | 0.272| -0.164 | -0.1977| -0.108| +    | 0      | −      | −     |
|       | (0.074)| (0.11) | (0.12) | (0.062)|      |        |        |       |
| GDP   | 0.304| 0.447  | 0.3933 | 0.031 | +    | +      | +      | 0     |
|       | (0.079)| (0.11) | (0.13) | (0.058)|      |        |        |       |
| Prod  | 0.066| 0.320  | 0.4294 | 0.036 | 0    | +      | +      | 0     |
|       | (0.078)| (0.11) | (0.13) | (0.063)|      |        |        |       |
| CPI   | 0.018| -0.036 | 0.0093 | 0.327 | 0    | 0      | 0      | +     |
|       | (0.091)| (0.13) | (0.15) | (0.080)|      |        |        |       |

Table 2: Estimated left coefficient matrix $A$ of MAR(1) using LS method. Standard errors are shown in the parentheses. The right panel indicates the positively significant, negatively significant and insignificant parameters at 10% level using symbols (+, −, 0), respectively.

The left coefficient matrix shows an interesting pattern. For example, the first row in Table 2 shows the influence on the current interest rate from the past quarter’s economic indicators. The influence from the past GDP growth, Production growth and CPI are all negative, meaning that strong past GDP growth and Production growth actually push the current interest rate lower. Higher CPI in the past is also negatively related to the current interest rate. The second row in Table 2 shows that the influence on the current GDP growth from the past quarter’s economic indicators. They are all positive, except the insignificant CPI influence. The last row Table 2 shows that the past economic indicators do not have significant influence on the current CPI, except its own past.

Table 3 shows the estimated $B$. Its effect should be considered in the view of $B X_t'$. It is seen that the influence of US’s last quarter’s indicators on the current quarter’s indicators of all countries (shown by the first column in $\hat{B}$) are very significantly positive and all larger than those of all other countries. This is intuitively correct as US is the world’s largest economy. Although it is understandable that Canada has a relatively small influence on other countries (shown by the last column), it is surprising to see that Germany has almost no influence (shown by the second column).
|     | USA  | DEU  | FRA  | GBR  | CAN  | USA  | DEU  | FRA  | GBR  | CAN  |
|-----|------|------|------|------|------|------|------|------|------|------|
| USA |  0.75 | -0.132 |  0.159 |  0.46 | -0.057 |  +  |  −  |  0  |  +  |  0   |
|     | (0.084) | (0.056) | (0.098) | (0.085) | (0.067) |     |     |     |     |     |
| DEU |  0.45  |  0.179 |  0.678 |  0.36  | -0.387 |  +  |  +  |  +  |  +  |  −   |
|     | (0.128) | (0.100) | (0.140) | (0.123) | (0.095) |     |     |     |     |     |
| FRA |  0.36  |  0.073 |  0.292 |  0.25  |  0.041 |  +  |  0  |  +  |  +  |  0   |
|     | (0.082) | (0.063) | (0.099) | (0.079) | (0.072) |     |     |     |     |     |
| GBR |  0.38  | -0.098 |  0.099 |  0.69  |  0.069 |  +  |  0  |  0  |  +  |  0   |
|     | (0.087) | (0.058) | (0.096) | (0.086) | (0.070) |     |     |     |     |     |
| CAN |  0.51  | -0.083 |  0.054 |  0.63  |  0.308 |  +  |  0  |  0  |  +  |  +   |
|     | (0.098) | (0.067) | (0.114) | (0.096) | (0.086) |     |     |     |     |     |

Table 3: Estimated right coefficient matrix $\mathbf{B}$ of MAR(1) using LS method. Standard errors are shown in parentheses. The right panel indicates the positively significant, negatively significant and insignificant parameters at 10% level using symbols (+, −, 0), respectively.

Most of the large coefficients are positive, showing positive influences among the countries.

Figures 6, 7 and 8 shows the impulse response functions with one standard deviation shock on US interest rate, US GDP and US CPI, respectively, using that given in Section 2.2. The dotted horizontal lines mark the values (0.1, 0, −0.1) and the dotted vertical lines marks the time (5, 10, 15). It can be seen from Figure 6 that a US interest rate shock would be responded strongly by the interest rates in other countries in the same pattern, and the impact lasts about 10 quarters. It is surprising to see that GDP of all countries responds positively to the interest rate shocks. However, the interest rate itself responds negatively, except at the first time period. This is because the positive GDP responses produce future negative responses to the interest rate (see Figure 7). CPI does not have much response to interest rate shocks.

From Figure 7 it is seen that GDP growth and Production growth of all countries respond positively to US GDP shock, while interest rates of all countries respond negatively, though the impact lasts only about a year. Again, CPI does not respond at all.

On the other hand, Figure 8 shows that a shock on US CPI generates positive responses from CPI, GDP growth and Production growth, and negative responses from the interest rate.
Figure 6: Impulse response function of MAR(1) model with a unit variance shock on US interest rate.

Figure 9 shows the residual plots of the MAR(1) estimated using LS method. There are some outliers. As an illustration of the MAR(1) model, in this analysis we do not try to do any adjustment. In Figure 10 we plot the autocorrelation function (ACF) of the 20 residual series, after fitting the MAR(1) model using the least squares method. Figure 11 shows the ACF plots of the 20 original series. It is seen that the MAR(1) model is able to capture the serial correlations in the 20 time series simultaneously, and lead to relatively clean ACF plots of the residuals.

We also obtain out-sample rolling forecast performances of the MAR(1) model as well as univariate AR(1) and AR(2) models for comparison. Specifically, starting from the last quarter of 2011 ($t = 87$) to the end of the series (the last quarter of 2016, $t = 107$), we fit the corresponding models using all available data at time $t - 1$ and obtained the one step ahead prediction $\hat{X}_{t-1}(1)$ for $X_t$ at time $t$. Sum of prediction error squares $||\hat{X}_{t-1}(1) - X_t||_F^2$ of all methods are shown in Table 4. It seems that MAR(1) with least squares estimation performs about 3% worse than the individual AR(1) models. It is commonly observed in multivariate time series that the joint model often performs worse than individual AR models in prediction. Figure 12 shows the difference between the sum of squares of prediction error (or all countries and all indicators) for each quarter between the MAR(1) model and the individual AR(1) model. It is seen that although MAR(1) model performs quite poorly in three out of the 20 quarters, it performs better in the later three
6 Conclusion

We proposed an autoregressive model for matrix-valued time series in a bilinear form. It respects the original matrix structure, and provides a much more parsimonious model, comparing with the direct VAR approach. Several interpretations of the model, along with possible extensions are provided. Different estimation methods are studied under different covariance structures of the error matrix. Asymptotic distributions of the various estimators are established, which facilitate the statistical inferences.

Table 4: Sum of out-of-sample prediction error squares of MAR(1) model using three different estimators and the stacked VAR(1) estimator, and the total sum of out-of-sample prediction error squares of fitting univariate AR(1) and AR(2) to each individual time series.

| MAR(1)projection | MAR(1)LS | MAR(1)MLE | iAR(1) | iAR(2) | VAR(1) |
|------------------|---------|----------|--------|--------|--------|
| 146.89           | 141.82  | 136.69   | 136.00 | 135.72 | 296.62 |

Figure 7: Impulse response function of MAR(1) model with a unit variance shock on US GDP rate.
Figure 8: Impulse response function of MAR(1) model with a unit variance shock on US CPI rate.

On the other hand, when the matrix observation has large dimensions itself, our model still involves a large number of parameters, although much less than that of the corresponding stacked VAR model. Note that it is natural to have relatively large total number of parameters, due to the large number of time series involved. For example, fitting univariate AR(2) models to the \( mn \) time series individually would require total \( 2mn \) AR coefficients, while MAR(1) involves \( m^2 + n^2 - 1 \) AR coefficients. When \( m \sim n \), they use about the same number of parameters. Also, the structured covariance model involves \( m(m+1)/2 + n(n+1)/2 - 1 \) parameters while individual AR models involve \( mn \) variance parameters, without considering any correlation among the series. Of course, regularized estimation approach can be used for MAR(1) model, potentially shrinking some of the insignificant parameters in the coefficient matrices to zero, as we have done in a rather ad hoc way in Tables 2 and 3 in the real example.

For very large dimensional matrix time series, Wang et al. (2018) proposed a factor model in a bilinear form. The MAR(1) can be used to model the factor matrix in that of Wang et al. (2018) to build a dynamic factor model in matrix form.

It is also of interest and importance to validate the assumption that the conditional expectation of \( \mathbf{X}_t \) is given by \( \mathbf{A} \mathbf{X}_{t-1} \mathbf{B}' \) in model (2), by a testing procedure. We will also consider MAR models of order larger than one in the future. Furthermore, the idea of MAR can be applied for volatility modeling (Bollerslev 1986; Engle 1982) as well.
7 Proofs

We collect the proofs of Proposition 1, Theorem 2, Theorem 3, Theorem 4 and Corollary 5 in this section.

We begin by listing some basic properties of the Kronecker product and its relationship with linear matrix equations. Let $M_{m,n}$ be the set of all $m \times n$ matrices over the field of complex numbers $\mathbb{C}$. The Kronecker product of $C = (c_{ij}) \in M_{m,n}$, and $D = (d_{ij}) \in M_{p,q}$, denoted by $C \otimes D$, is
defined to be the block matrix

\[ C \otimes D = \begin{pmatrix}
  c_{11}D & \cdots & c_{1n}D \\
  \vdots & \ddots & \vdots \\
  c_{m1}D & \cdots & c_{mn}D
\end{pmatrix} \in M_{mpnq}. \]

In the following proposition, we list some facts regarding the Kronecker product, which are used in this section at various places without specific references. Proofs of these facts can be found in Chapter 4 of [Horn and Johnson (1994)].

**Proposition 6.** Let \( C \in M_{m,n} \), \( D \in M_{p,q} \), \( F \in M_{n,k} \), \( G \in M_{q,l} \) and \( Z \in M_{n,p} \).

(i) \((C \otimes D)' = C' \otimes D'\).

(ii) If both \( C \) and \( D \) are invertible square matrices, then \( C \otimes D \) is also invertible, and \((C \otimes D)^{-1} = C^{-1} \otimes D^{-1}\).
(iii) \((C \otimes D)(F \otimes G) = (CF) \otimes (DG)\).

(iv) \(\text{vec}(CZD) = (D' \otimes C)\text{vec}(Z)\).

(v) \(\text{rank}(C \otimes D) = \text{rank}(D \otimes C) = \text{rank}(C) \cdot \text{rank}(D)\).

(vi) Let \(C \in M_{m,m}\) and \(D \in M_{n,n}\). Let \(\{\lambda_1, \lambda_2, \ldots, \lambda_m\}\) be eigenvalues of \(C\) (including multiplicities), and \(\{\eta_1, \eta_2, \ldots, \eta_n\}\) be eigenvalues of \(D\). The \(mn\) eigenvalues (including multiplicities) of \(C \otimes D\) are \(\{\lambda_i \eta_j : 1 \leq i \leq m, 1 \leq j \leq n\}\).

**Proof of Proposition**

It is known that the VAR(1) model in (1) admits a stationary and causal solution if the spectral radius of the coefficient matrix \(\Phi\) is strictly less than 1 (See for example, §11.3 of Brockwell and Davis, 1991). By Proposition 6~(vi), all the eigenvalues of \(B \otimes A\) are of the form \(\lambda_i \eta_j\), where \(\lambda_i\) and \(\eta_j\) are the eigenvalues of \(A\) and \(B\) respectively. As a consequence, \(\rho(B \otimes A) = \rho(A) \cdot \rho(B)\). Since the MAR(1) model in (2) can be represented as a VAR model as given by (3), the proposition then follows.

\[\square\]
Figure 12: The difference of the sum of prediction error squares between the MAR(1) model and individual AR(1) model at each quarter.

Proof of Theorem 3. Let $\alpha = \text{vec}(A)$ and $\beta = \text{vec}(B)$. Note that the convention that $\|A\|_F = 1$ is equivalent to $\|\alpha\| = 1$. The gradient condition of the NKP problem (8) is given by

$$\hat{\alpha} \beta' \hat{\beta} - \tilde{\Phi} \hat{\beta} = 0 \quad (15)$$

$$\hat{\beta} \alpha' \hat{\alpha} - \tilde{\Phi}' \hat{\alpha} = 0.$$

Recall that we require $\|A_1\|_F = 1$, so we similarly also require that the solution of the NKP problem satisfies $\|\hat{\alpha}\| = 1$. Since both $\|\alpha\| = 1$ and $\|\hat{\alpha}\| = 1$ it follows that $(\hat{\alpha} - \alpha)' \alpha = o_p(T^{-1/2})$. Replacing $\tilde{\Phi}$ by $\alpha \beta' + (\hat{\Phi} - \alpha \beta')$ in the gradient conditions, we have

$$(\hat{\alpha} - \alpha) \beta' \beta + \alpha (\hat{\beta} - \beta)' \beta = (\hat{\Phi} - \alpha \beta') \beta + o_p(T^{-1/2})$$

$$\hat{\beta} - \beta = (\hat{\Phi} - \alpha \beta')' \alpha + o_p(T^{-1/2}).$$

(16)
It follows that
\[
\begin{pmatrix}
\text{vec}(\hat{A}_1 - A) \\
\text{vec}(\hat{B}_1 - B)
\end{pmatrix} = V_0 \text{vec}(\hat{\Phi} - \alpha\beta') + o_P(T^{-1/2}),
\]
(17)
and
\[
\hat{\alpha}'\hat{\beta}' - \alpha\beta' = (\hat{\Phi} - \alpha\beta')\beta_1\hat{\beta}_1' + \alpha_1\alpha'_1(\hat{\Phi} - \alpha\beta') - \alpha_1\alpha'_1(\hat{\Phi} - \alpha\beta')\beta_1\hat{\beta}_1' + o_P(T^{-1/2}).
\]
(18)
The first central limit theorem stated in Theorem 2 is an immediate consequence of (17). Taking vectorization on both sides of (18), we have
\[
\hat{\alpha} \otimes \hat{\beta} - \alpha \otimes \beta = V_1 \text{vec}(\tilde{\Phi} - \alpha\beta') + o_P(T^{-1/2}),
\]
and the second central limit theorem follows.

To prove Theorem 2, we first state and prove the following lemma.

**Lemma 7.** Consider the VAR(1) representation of (3), and let \(\Phi = \mathbf{B} \otimes \mathbf{A}\). Assume the conditions of Theorem 2. Then for any sequence \(\{c_T\}\) such that \(c_T \to \infty\),
\[
P\inf_{\sqrt{T}/\|\hat{\Phi} - \Phi\|_{F} \geq c_T} \sum_{t=2}^{T} \left\| \text{vec}(X_t) - \tilde{\Phi} \text{vec}(X_{t-1}) \right\|^2 \leq \sum_{t=2}^{T} \left\| \text{vec}(E_t) \right\|^2 \to 0.
\]
(19)
Proof. First of all, by the ergodic theorem
\[
\frac{1}{T} \sum_{t=2}^{T} \text{vec}(X_{t-1})\text{vec}(X_{t-1})' \to \Gamma_0 \quad \text{a.s.}
\]
It follows that for any constant \(c > 0\),
\[
\sup_{\sqrt{T}/\|\hat{\Phi} - \Phi\|_{F} \leq c} \left| \frac{1}{T} \sum_{t=2}^{T} \text{tr} \left[ (\tilde{\Phi} - \Phi) \text{vec}(X_{t-1})\text{vec}(X_{t-1})'(\tilde{\Phi} - \Phi)' \right] \right| = 0 \quad \text{a.s.}
\]
(20)
In (20), the superme is taken over \(\tilde{\Phi}\). As a consequence of (20), there exists a sequence \(\{c'_T\}\) such that \(c'_T \to \infty, c'_T \leq c_T\), and
\[
\sup_{\sqrt{T}/\|\hat{\Phi} - \Phi\|_{F} \leq c'_T} \left| \frac{1}{T} \sum_{t=2}^{T} \text{tr} \left[ (\tilde{\Phi} - \Phi) \text{vec}(X_{t-1})\text{vec}(X_{t-1})'(\tilde{\Phi} - \Phi)' \right] \right| = 0 \quad \text{in probability.}
\]
(21)
Now we write
\[
\sum_{t=2}^{T} \| \text{vec}(X_t) - \Phi \text{vec}(X_{t-1}) \|^2 - \sum_{t=2}^{T} \| \text{vec}(E_t) \|^2 = -2 \sum_{t=2}^{T} \text{tr} \left[ (\Phi - \Phi) \text{vec}(X_{t-1}) \text{vec}(E_t)' \right] + \sum_{t=2}^{T} \left[ (\Phi - \Phi) \text{vec}(X_{t-1}) \text{vec}(X_{t-1})' (\Phi - \Phi)' \right].
\] (22)

On the boundary set $\sqrt{T} \| \bar{\Phi} - \Phi \|_F = c'_T$, by calculating the variance, we know that
\[
\sum_{t=2}^{T} \text{tr} \left[ (\Phi - \Phi) \text{vec}(X_{t-1}) \text{vec}(E_t)' \right] = O_P(c'_T).
\] (23)

On the other hand, on the boundary set $\sqrt{T} \| \bar{\Phi} - \Phi \|_F = c'_T$,
\[
T \cdot \text{tr} \left[ (\Phi - \Phi) \Gamma_0 (\Phi - \Phi)' \right] \geq \lambda(\Gamma_0)(c'_T)^2,
\] (24)
where $\lambda(\Gamma_0)$ is the minimum eigenvalue of $\Gamma_0$, which is strictly positive under the condition that $A$, $B$ and $\Sigma$ are nonsingular. Combining (22)∼(24), and with fact that $c'_T \to \infty$, we have
\[
P \left[ \inf_{\sqrt{T} \| \bar{\Phi} - \Phi \|_F = c'_T} \sum_{t=2}^{T} \| \text{vec}(X_t) - \Phi \text{vec}(X_{t-1}) \|^2 \leq \sum_{t=2}^{T} \| \text{vec}(E_t) \|^2 \right] \to 0.
\] (25)

Observe that $\sum_{t=2}^{T} \| \text{vec}(X_t) - \Phi \text{vec}(X_{t-1}) \|^2$ is a convex function of $\bar{\Phi}$, so (19) is implied by (25) and the convexity.

Now we are ready to give the proof of Theorem 3.

Proof of Theorem 3. Let $\mathcal{S} = \{ C : C \text{ is a } m \times m \text{ matrix, and } \| C \|_F = 1 \}$. Let $\{ c_T \}$ be any sequence such that $c_T \to \infty$, and $c_T/\sqrt{T} \to 0$. By the conditions that $A$ and $B$ are nonsingular, and $A \in \mathcal{S}$, it can be show that if $\hat{A}$ and $\hat{B}$ are such that $\hat{A} \in \mathcal{S}$, and $T \| \hat{A} - A \|^2_F + T \| \hat{B} - B \|^2_F \geq c'_T$, then $\| \hat{B} \otimes A - B \otimes \hat{A} \|_F \geq c \cdot c_T$, where $c$ is a constant determined by $A$ and $B$. By Lemma 7, we have
\[
P \left[ \min_{T \| \hat{A} - A \|^2_F + T \| \hat{B} - B \|^2_F \geq c'_T} \sum_{t=2}^{T} \| \text{vec}(X_t) - (\hat{B} \otimes \hat{A}) \text{vec}(X_{t-1}) \|^2 \leq \sum_{t=2}^{T} \| \text{vec}(E_t) \|^2 \right] \to 0,
\]
with the implicit requirement that $\hat{A} \in \mathcal{S}$. It follows that
\[
P \left[ T \| \hat{A} - A \|^2_F + T \| \hat{B} - B \|^2_F \leq c'_T \right] \to 1,
\] (26)
also with the implicit requirement that $\hat{A} \in \mathbb{S}$. Since (26) holds for any sequence $\{c_T\}$ such that $c_T \to \infty$, and $c_T/\sqrt{T} \to 0$, we have

$$\hat{A} = A + O_P(T^{-1/2}), \quad \text{and} \quad \hat{B} = B + O_P(T^{-1/2}).$$

We now repeat the gradient condition (13) here:

$$\sum_t \Delta_2 X_{t-1} \hat{B}_{2}^{T} B_{2} X_{t-1}^T - \sum_t X_t \hat{B}_{2} X_{t-1} = 0$$

$$\sum_t B_{2} X_{t-1}^T \Delta_2 A_{2} X_{t-1} - \sum_t X_t \hat{A}_{2} X_{t-1} = 0. \quad (27)$$

Replacing each $X_t$ by $AX_{t-1}B' + E_t$ in (27), we have

$$\sum_t (\Delta_2 - A) X_{t-1} B' B X_{t-1} + \sum_t A X_{t-1} (\hat{B}_2 - B)' B X_{t-1} = \sum_t E_t B X_{t-1} + o_P(\sqrt{T})$$

$$\sum_t X_{t-1}^T A' (\hat{A}_2 - A) X_{t-1} B' + \sum_t X_{t-1}^T A' AX_{t-1} (\hat{B}_2 - B)' = \sum_t X_{t-1}^T A' E_t + o_P(\sqrt{T}).$$

Taking vectorization on both sides, we have

$$\begin{pmatrix} \sum_t X_{t-1} B' B X_{t-1} \otimes I \\ \sum_t B X_{t-1} \otimes (X_{t-1} A') \otimes I \otimes (\sum_t A' A X_{t-1}) \end{pmatrix} \begin{pmatrix} \text{vec}(\Delta_2 - A) \\ \text{vec}(\hat{B}_2 - B') \end{pmatrix} = \sum_t \text{vec}(E_t) + o_P(\sqrt{T}),$$

which can be rewritten as

$$\begin{pmatrix} \sum_t W_{t-1} W_{t-1}' \end{pmatrix} \begin{pmatrix} \text{vec}(\Delta_2 - A) \\ \text{vec}(\hat{B}_2 - B') \end{pmatrix} = \sum_t W_{t-1} \text{vec}(E_t) + o_P(\sqrt{T}). \quad (28)$$

By the ergodic theorem as $X_t$ is strictly stationary with i.i.d. innovations under the conditions, we have

$$\frac{1}{T} \sum_t W_{t-1} W_{t-1}' \to \mathbb{E}(W_t W_t'), \quad \text{a.s.}$$

Observe that $\mathbb{E}(W_t W_t')$ is not a full rank matrix, because $\mathbb{E}(W_t W_t') (\alpha' - \beta')' = 0$. On the other hand, since we require $\|A\|_F = 1$ and $\|\Delta_2\|_F = 1$, it holds that $\alpha' (\text{vec}(\Delta_2) - \alpha) = o_P(T^{-1/2})$. and consequently

$$H \begin{pmatrix} \text{vec}(\Delta_2 - A) \\ \text{vec}(\hat{B}_2 - B') \end{pmatrix} = \frac{1}{T} \sum_t W_{t-1} \text{vec}(E_t) + o_P(T^{-1/2}),$$

36
$H := E(W_tW'_t) + \gamma\gamma'$. By martingale central limit theorem (Hall and Heyde, 1980)

$$\sum_t W_{t-1}vec(E_t) \Rightarrow N(0, E(W_t\Sigma W'_t)).$$

It follows that

$$\sqrt{T} \begin{pmatrix} vec(\hat{A}_2 - A) \\ vec(\hat{B}'_2 - B') \end{pmatrix} \Rightarrow N(0, \Xi_2),$$

where $\Xi_2 := H^{-1}E(W_t\Sigma W'_t)H^{-1}$. Furthermore, noting that

$$vec(\hat{B}'_2) \otimes vec(\hat{A}_2) - vec(B') \otimes vec(A) = V \begin{pmatrix} vec(\hat{A}_2 - A) \\ vec(\hat{B}'_2 - B') \end{pmatrix} + o_P(T^{-1/2}),$$

where $V := [\beta \otimes I, I \otimes \alpha]$, the statement about $B \otimes A$ in the theorem follows. The proof is complete.

**Proof of Theorem 4.** Following similar arguments of (Dunsmuir and Hannan, 1976; Hannan, 1970), we know that $\hat{A}_3 = A + O_P(T^{-1/2})$ and $\hat{B}_3 = B + O_P(T^{-1/2})$, and $\hat{\Sigma}_e$ and $\hat{\Sigma}_c$ are consistent estimators of $\Sigma_e$ and $\Sigma_c$ respectively. The rest of the proof is very similar with that of Theorem 3, and we omit the details here.

**Proof of Corollary 5.** From Theorem 3 and Theorem 4, we have that

$$\Xi_2 = [E(W_tW'_t) + \gamma\gamma']^{-1}[E(W_t\Sigma W'_t)]E(W_tW'_t) + \gamma\gamma']^{-1},$$

$$\Xi_3 = [E(W_t\Sigma^{-1}W'_t) + \gamma\gamma']^{-1}[E(W_t\Sigma^{-1}W'_t)]E(W_t\Sigma^{-1}W'_t) + \gamma\gamma']^{-1}.$$

Let $\Sigma = Q\text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_{mn}\}Q'$ be the spectral decomposition of $\Sigma$, and use $u_1, u_2, \ldots, u_{mn}$ to denote the $mn$ columns of the matrix $W_tQ$. We further let $\Theta_i = E(u_iu'_i)$, and note that $\Theta_i$ is a symmetric positive semi-definite matrix. Then the preceding equation becomes

$$\Xi_2 = \left(\sum_{i=1}^{mn} \Theta_i + \gamma\gamma'\right)^{-1} \left(\sum_{i=1}^{mn} \lambda_i \Theta_i\right) \left(\sum_{i=1}^{mn} \Theta_i + \gamma\gamma'\right)^{-1},$$

$$\Xi_3 = \left(\sum_{i=1}^{mn} \lambda_i^{-1} \Theta_i + \gamma\gamma'\right)^{-1} \left(\sum_{i=1}^{mn} \lambda_i^{-1} \Theta_i\right) \left(\sum_{i=1}^{mn} \lambda_i^{-1} \Theta_i + \gamma\gamma'\right)^{-1}.$$
Since \( W'_t \gamma = 0 \), it holds that \( \Theta_i \gamma = 0 \) for all \( 1 \leq i \leq mn \). It follows that
\[
\Xi_2 + \gamma \gamma' = \left( \sum_{i=1}^{mn} \Theta_i + \gamma \gamma' \right)^{-1} \left( \sum_{i=1}^{mn} \lambda_i \Theta_i + \gamma \gamma' \right) \left( \sum_{i=1}^{mn} \Theta_i + \gamma \gamma' \right)^{-1}
\]
\[
\Xi_3 + \gamma \gamma' = \left( \sum_{i=1}^{mn} \lambda_i^{-1} \Theta_i + \gamma \gamma' \right)^{-1}.
\]

To simplify the long equations, we let \( \Theta_{mn+1} = \gamma \gamma' \), \( \lambda_{mn+1} = 1 \), and make the convention that all the sums over \( i \) runs from \( i = 1 \) to \( i = mn + 1 \). The equation (29) becomes
\[
\Xi_2 + \gamma \gamma' = \left( \sum_i \Theta_i \right)^{-1} \left( \sum_i \lambda_i \Theta_i \right) \left( \sum_i \Theta_i \right)^{-1},
\]
\[
\Xi_3 + \gamma \gamma' = \left( \sum_i \lambda_i^{-1} \Theta_i \right)^{-1}.
\]

From (30), we see that in order to show that \( \Xi_2 \geq \Xi_3 \), it suffices to show that
\[
\sum_i \lambda_i^{-1} \Theta_i - \left( \sum_i \Theta_i \right) \left( \sum_i \lambda_i \Theta_i \right)^{-1} \left( \sum_i \Theta_i \right)
\]
is positive semi-definite. For this purpose, we construct the matrix
\[
\begin{pmatrix}
\sum_i \lambda_i^{-1} \Theta_i & \sum_i \Theta_i \\
\sum_i \Theta_i & \sum_i \lambda_i \Theta_i
\end{pmatrix} = \sum_i \begin{pmatrix}
\lambda_i^{-1} \Theta_i & \Theta_i \\
\Theta_i & \lambda_i \Theta_i
\end{pmatrix}.
\]

Since \( \Theta_i \) is positive semi-definite, each term on the right hand side of (32) is also positive semi-definite, and so is the sum in (32). If we view the matrix in (32) as a covariance matrix, then the matrix in (31) is the conditional covariance matrix of the first half given the second half, so it is positive semi-definite, and the proof is complete. \( \square \)
References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.*, 43(4):1535–1567.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

Cox, D. A., Little, J., and O’Shea, D. (2015). *Ideals, varieties, and algorithms*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition. An introduction to computational algebraic geometry and commutative algebra.

Davis, R. A., Zang, P., and Zheng, T. (2012). Sparse Vector Autoregressive Modeling. *ArXiv e-prints*.

Deistler, M., Dunsmuir, W., and Hannan, E. J. (1978). Vector linear time series models: corrections and extensions. *Adv. in Appl. Probab.*, 10(2):360–372.

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models. *Advances in Appl. Probability*, 8(2):339–364.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):603–680.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *J. Amer. Statist. Assoc.*, 100(471):830–840.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432.
Guo, S., Wang, Y., and Yao, Q. (2015). High Dimensional and Banded Vector Autoregressions. *ArXiv e-prints.*

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. Probability and Mathematical Statistics.

Hallin, M. and Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics,* 163(1):29–41.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res.*, 16:3115–3150.

Han, F., Xu, S., and Liu, H. (2016). Rate-optimal estimation of high dimensional time series. Technical report, Washington University, Department of Statistics.

Hannan, E. J. (1970). *Multiple time series.* John Wiley and Sons, Inc., New York-London-Sydney.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis.* Cambridge University Press, Cambridge. Corrected reprint of the 1991 original.

Kaufman, L. (1975). A variable projection method for solving separable nonlinear least squares problems. *BIT Numerical Mathematics,* 15(1):49–57.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics,* 186(2):325–344.

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

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika,* 98(4):901–918.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis.* Springer-Verlag, Berlin.

Nardi, Y. and Rinaldo, A. (2011). Autoregressive process modeling via the Lasso procedure. *J. Multivariate Anal.,* 102(3):528–549.
Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097.

Nicholson, W., Matteson, D., and Bien, J. (2015). VARX-L: Structured Regularization for Large Vector Autoregressions with Exogenous Variables. *ArXiv e-prints*.

Raskutti, G., Yuan, M., and Chen, H. (2015). Convex Regularization for High-Dimensional Multi-Response Tensor Regression. Technical report.

Song, S. and Bickel, P. J. (2011). Large Vector Auto Regressions. *ArXiv e-prints*.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications. *J. Amer. Statist. Assoc.*, 76(376):802–816.

Tsai, H. and Tsay, R. S. (2010). Constrained factor models. *J. Amer. Statist. Assoc.*, 105(492):1593–1605. Supplementary materials available online.

Tsay, R. S. (2014). *Multivariate time series analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ.

Van Loan, C. (2000). The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics*, 123(1):85 – 100. Numerical Analysis 2000. Vol. III: Linear Algebra.

Van Loan, C. and Pitsianis, N. (1993). Approximation with kronecker products. In Moonen, M. and Golub, G., editors, *Linear Algebra for Large Scale and Real Time Applications*, pages 293–314. Kluwer Publications, Dordrecht.

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., Yang, D., Shen, H., and Zhu, H. (2018). On scalar-on-matrix bilinear regression analysis. Technical report.

Zhao, J. and Leng, C. (2014). Structured lasso for regression with matrix covariates. *Statist. Sinica*, 24:799–814.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor Regression with Applications in Neuroimaging Data Analysis. *J. Amer. Statist. Assoc.*, 108(502):540–552.