

High-dimensional vector autoregressive time series modeling via tensor decomposition

Di Wang, Heng Lian, Yao Zheng and Guodong Li

*University of Hong Kong, City University of Hong Kong
and University of Connecticut*

September 17, 2019

Abstract

The classical vector autoregressive model is a fundamental tool for multivariate time series analysis. However, it involves too many parameters when the number of time series and lag order are even moderately large. This paper proposes to rearrange the coefficient matrices of the model into a tensor form such that the parameter space can be restricted in three directions simultaneously via tensor decomposition. The proposed method substantially expands the capacity of vector autoregressive modeling for a large number of time series. In contrast, the widely used reduced-rank regression method can restrict the parameter space in only one direction. Moreover, to handle high-dimensional time series, this paper considers imposing sparsity on factor matrices to improve the interpretability and estimation efficiency, which leads to a sparsity-inducing estimator. For the low-dimensional case, we derive asymptotic properties of the proposed least squares estimator and introduce an alternating least squares algorithm. For the high-dimensional case, we establish non-asymptotic properties of the sparsity-inducing estimator and propose an ADMM-based algorithm for regularized estimation. Simulation experiments and a real data example demonstrate the advantages of the proposed approach over various existing methods.

Keywords: High-dimensional time series; Reduced-rank regression; Regularization; Tucker decomposition; Variable selection.

1 Introduction

High-dimensional time series is one of the most common types of “big data” and can be found in many areas including meteorology, genomics, finance and economics (Hallin and Lippi, 2013). The classical vector autoregressive (VAR) model is fundamental to multivariate time series modeling and has recently been applied to the high-dimensional case under certain structural assumptions, e.g., the banded structure (Guo et al., 2016) and the network structure (Zhu et al., 2017). Consider the VAR model of the form (Lütkepohl, 2005; Tsay, 2010):

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A}_P \mathbf{y}_{t-P} + \boldsymbol{\epsilon}_t, \quad -P + 1 \leq t \leq T, \quad (1)$$

where $\{\mathbf{y}_t\}$ is the observed time series with $\mathbf{y}_t = (y_{1t}, \dots, y_{Nt})' \in \mathbb{R}^N$, $\{\boldsymbol{\epsilon}_t\}$ are independent and identically distributed (*i.i.d.*) innovations with $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \dots, \epsilon_{Nt})'$, $E(\boldsymbol{\epsilon}_t) = 0$ and $\text{var}(\boldsymbol{\epsilon}_t) < \infty$, \mathbf{A}_j s are $N \times N$ transition matrices of unknown parameters, and T is the sample size. It can be difficult to perform the estimation even when the dimension N is moderately large (De Mol et al., 2008; Carriero et al., 2011; Koop, 2013).

On the other hand, compared with model (1), the vector autoregressive moving average (VARMA) model usually performs better in practice since it can provide a more flexible autocorrelation structure (Athanasopoulos and Vahid, 2008; Chan et al., 2016). However, the VARMA model may have a serious identification problem (Chan et al., 2016; Wilms et al., 2017; Dias and Kapetanios, 2018), and its estimation is often unstable since the corresponding objective function involves a high-order polynomial. As a result, it is common in practice to employ a VAR model to approximate VARMA processes, and the order P may be very large in order to provide a better fit for the data (Ravenna, 2007). For example, to guarantee the approximation accuracy, we need to assume that $P \rightarrow \infty$ and $PT^{-1/3} \rightarrow 0$ as $T \rightarrow \infty$ for univariate and multivariate cases (Said and Dickey, 1984; Li et al., 2014). This makes the number of parameters in model (1), N^2P , much larger.

Therefore, to make inference on the VAR model for high-dimensional time series, it is necessary to restrict the parameter space of model (1) to a reasonable number of degrees of freedom. A direct method is to assume that the transition matrices \mathbf{A}_j s are sparse and apply sparsity-inducing regularized estimation, e.g., the ℓ_1 regularization (Lasso or

Dantzig estimator), for VAR models (Kock and Callot, 2015; Davis et al., 2016; Basu and Michailidis, 2015; Han et al., 2015; Wu and Wu, 2016). However, unlike the traditional linear regression, **time series data have non-negligible temporal and cross-sectional dependencies, which will seriously affect the accuracy of the regularized estimation.** More importantly, for the reason explained in Remark 1 in Section 2, the stationarity of the VAR model essentially entails that the average magnitude of parameters is bounded by $O(N^{-1/2})$. This makes the variable selection much more challenging and hence limits the popularity of sparsity-inducing regularized estimation for time series data.

Another important approach to reducing the dimensionality of model (1) arises naturally from the reduced-rank regression (Yuan et al., 2007; Negahban and Wainwright, 2011; Chen et al., 2013; Raskutti et al., 2019). The VAR model in (1) can be rewritten as

$$\mathbf{y}_t = \mathbf{A}^{(C)} \mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad (2)$$

where $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-P})'$, and $\mathbf{A}^{(C)} = (\mathbf{A}_1, \dots, \mathbf{A}_P)$ is assumed to have a low rank (Velu et al., 1986; Velu and Reinsel, 2013). Based on the reduced-rank VAR model in (2), Carriero et al. (2011) considered a Bayesian method to predict large macroeconomic data, and both the number of variables N and the sample size T diverge to infinity. However, unlike the reduced-rank regression, we may have alternative ways to define the low-rankness of parameter matrices \mathbf{A}_j s with $P > 1$. Specifically, the rank of $\mathbf{A}^{(C)}$ is the dimension of the column space of \mathbf{A}_j s. Denote $\mathbf{A}^{(R)} = (\mathbf{A}'_1, \mathbf{A}'_2, \dots, \mathbf{A}'_P)$ and $\mathbf{A}^{(L)} = (\text{vec}(\mathbf{A}_1), \text{vec}(\mathbf{A}_2), \dots, \text{vec}(\mathbf{A}_P))'$, where $\text{vec}(\mathbf{A}_j)$ is the vectorization of \mathbf{A}_j . **The ranks of $\mathbf{A}^{(R)}$ and $\mathbf{A}^{(L)}$ are then the dimensions of the row space and vectorized matrix space of \mathbf{A}_j s, respectively.** The three dimensions are different in general, and the corresponding low-rank structures have different physical interpretations; see Section 2 for details. Similarly to model (2) above, Reinsel (1983) proposed an autoregressive index model, where the low-rank assumption was imposed on $\mathbf{A}^{(R)}$. Moreover, the transition matrices \mathbf{A}_j s may have a low-rank structure along different lags, i.e. $\mathbf{A}^{(L)}$ has a low rank. In fact, the VARMA model can be treated as a parsimonious formulation for VAR models, since it restricts the degrees of freedom on transition matrices over different lags (Tsay, 2010).

It is noteworthy that imposing the low-rank assumption on any one of $\mathbf{A}^{(C)}$, $\mathbf{A}^{(R)}$ and $\mathbf{A}^{(L)}$ leads to a different physical interpretation as it amounts to reducing the dimensionality along one of the three different directions. This inspires us to rearrange the transition matrices \mathbf{A}_j s into a tensor, and interestingly, the corresponding mode-1, -2 and -3 matricizations of the tensor happen to be $\mathbf{A}^{(C)}$, $\mathbf{A}^{(R)}$ and $\mathbf{A}^{(L)}$, respectively; see Kolda and Bader (2009) and Section 2. By adopting the standard Tucker decomposition for the transition tensor, different low-rank structures can be assumed simultaneously along the three directions, and hence the parameter space of the VAR model can be efficiently restricted. We call the resulting model the multilinear low-rank VAR model, since the Tucker ranks are also called multilinear ranks.

Furthermore, another important contribution of this paper is to introduce a sparse decomposition for the transition tensor as a more efficient approach to modeling much higher dimensional time series. In the literature, sparsity-inducing regularization has been widely considered in reduced-rank regression to improve interpretability and efficiency. For example, Chen and Huang (2012) and Bunea et al. (2012) considered row-wise sparsity in singular value decomposition, where zero rows imply irrelevance of the corresponding predictors to the responses; Lian et al. (2015) proposed to directly restrict the rank of the coefficient matrix with entry-wise sparsity, which however does not lead to a sparse decomposition; Chen et al. (2012) obtained a sparse singular value decomposition of the coefficient matrix by slightly relaxing the strict orthogonality; and Uematsu et al. (2019) achieved the sparsity and strict orthogonality simultaneously. Note that as in Uematsu et al. (2019), our proposed method is able to keep the strict orthogonality of the factor matrices in the tensor decomposition.

Our work is also related to the fast-growing literature on tensor regression; see, e.g., Zhou et al. (2013), Li et al. (2018), Li and Zhang (2017), Sun and Li (2017) and Raskutti et al. (2019). Whereas most of the existing work focuses on tensor-valued predictors or responses, we employ tensor decomposition as a novel approach to the dimensionality reduction of vector-valued time series models. To summarize, the proposed methodology has the following three attractive features:

- (a) The proposed methods reduce the dimensionality of the VAR model in three direc-

tions of the transition tensor, where each direction can have a different low-rank structure, and hence allow us to handle much higher dimensional data than the reduced-rank VAR model in (2).

- (b) As shown in Section 2, for VARMA processes, the transition tensor of the corresponding VAR(∞) representation possesses a low-rank structure over different lags under certain conditions. Thus we may expect that the performance of the proposed model is comparable to that of VARMA models.
- (c) The proposed high-dimensional method facilitates the interpretation, and improves the estimation efficiency, of the high-dimensional VAR model through the three sparse factor matrices in the tensor decomposition. An ADMM-based algorithm is also proposed to separate the ℓ_1 regularization and orthogonality constraints.

The rest of the paper is organized as follows. Section 2 introduces the multilinear low-rank VAR model defined via tensor decomposition and discusses its relationship with the VARMA model. Section 3 derives asymptotic properties of the least squares estimator for the proposed model and introduces an alternating least squares algorithm. For the high-dimensional setup, the sparse higher-order reduced-rank estimation is proposed in Section 4 which takes into account both the orthogonality and sparsity. The non-asymptotic properties are established, and an ADMM-based algorithm is also developed. Numerical studies are presented in Section 5, and a short discussion is given in Section 6. All technical proofs are relegated to the Appendix.

2 Multilinear low-rank vector autoregression

2.1 Tensor decomposition

Tensors, a.k.a. multidimensional arrays, are natural higher-order extensions of matrices. A multidimensional array $\mathcal{X} \in \mathbb{R}^{p_1 \times \dots \times p_K}$ is called a K th-order tensor, and the order of a tensor is known as the dimension, way or mode; we refer readers to Kolda and Bader (2009) for a detailed review on tensor notations and operations. This paper will focus on third-order tensors.

Throughout the paper, we denote vectors by small boldface letters $\mathbf{y}, \mathbf{x}, \dots$, matrices by capital letters $\mathbf{Y}, \mathbf{X}, \dots$, and tensors by Euler script capital letters $\mathcal{Y}, \mathcal{X}, \dots$. For a vector \mathbf{x} , denote by $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ its ℓ_1 and ℓ_2 norms, respectively. For a matrix \mathbf{X} , denote by $\|\mathbf{X}\|_F$, $\|\mathbf{X}\|_1$, $\|\mathbf{X}\|_0$, $\text{vec}(\mathbf{X})$, \mathbf{X}' and $\sigma_j(\mathbf{X})$ its Frobenius norm, vectorized ℓ_1 norm (i.e. $\|\mathbf{X}\|_1 = \|\text{vec}(\mathbf{X})\|_1$), ℓ_0 “norm”, vectorization, transpose and the j -th largest singular value, respectively. For two symmetric matrices \mathbf{X} and \mathbf{Y} , we write $\mathbf{X} \leq \mathbf{Y}$ if $\mathbf{Y} - \mathbf{X}$ is positive semidefinite. Furthermore, for a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, let $\|\mathcal{X}\|_F = \left(\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} \mathcal{X}_{ijk}^2 \right)^{1/2}$ and $\|\mathcal{X}\|_0 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} 1(\mathcal{X}_{ijk} \neq 0)$ be its Frobenius norm and ℓ_0 “norm”, respectively.

For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, its mode-1 matricization $\mathcal{X}_{(1)}$ is defined as the p_1 -by- $(p_2 p_3)$ matrix whose $\{i, (k-1)p_3 + j\}$ -th entry is \mathcal{X}_{ijk} , for $1 \leq i \leq p_1, 1 \leq j \leq p_2$ and $1 \leq k \leq p_3$, and $\mathcal{X}_{(1)}$ contains all mode-1 fibers $\{(\mathcal{X}_{[:,i_2,i_3]}) \in \mathbb{R}^{p_1} : 1 \leq i_2 \leq p_2, 1 \leq i_3 \leq p_3\}$. The mode-2 and mode-3 matricizations can be defined similarly. The matricization of tensors helps to link the concepts and properties of matrices to those of tensors. The mode-1 multiplication \times_1 of a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ and a matrix $\mathbf{Y} \in \mathbb{R}^{q_1 \times p_1}$ is defined as

$$\mathcal{X} \times_1 \mathbf{Y} = \left(\sum_{i=1}^{p_1} \mathcal{X}_{ijk} \mathbf{Y}_{si} \right)_{1 \leq s \leq q_1, 1 \leq j \leq p_2, 1 \leq k \leq p_3}.$$

Multiplications \times_2 and \times_3 can be defined similarly.

Unlike matrices, there is no universal definition of the rank for tensors. In this paper, we consider the multilinear ranks (r_1, r_2, r_3) of a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, where

$$r_1 = \text{rank}_1(\mathcal{X}) := \text{rank}(\mathcal{X}_{(1)}) = \dim(\text{span}\{\mathcal{X}_{[:,i_2,i_3]} \in \mathbb{R}^{p_1} : 1 \leq i_2 \leq p_2, 1 \leq i_3 \leq p_3\}),$$

and r_2 and r_3 are the ranks of $\mathcal{X}_{(2)}$ and $\mathcal{X}_{(3)}$, respectively. Note that r_1, r_2 and r_3 are analogous to the row rank and column rank of a matrix, but these three ranks are not necessarily equal. The multilinear ranks are also known as Tucker ranks, as they are closely related to the Tucker decomposition.

For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, if $\text{rank}_j(\mathcal{X}) = r_j$ for $1 \leq j \leq 3$, then there exists a Tucker decomposition (Tucker, 1966; De Lathauwer et al., 2000),

$$\mathcal{X} = \mathcal{Y} \times_1 \mathbf{Y}_1 \times_2 \mathbf{Y}_2 \times_3 \mathbf{Y}_3,$$

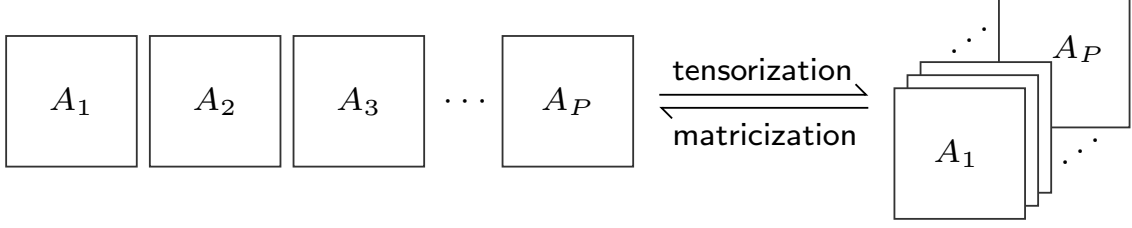


Figure 1: Rearranging P transition matrices of a VAR model into a tensor.

where $\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, $\mathbf{Y}_j \in \mathbb{R}^{p_j \times r_j}$ with $1 \leq j \leq 3$ are factor matrices, and the above decomposition can also be denoted by $\mathcal{X} = \llbracket \mathcal{Y}; \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3 \rrbracket$.

There is another definition of the rank for tensors which is related to the CP decomposition. For the tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, if there exist $\mathbf{x}_i^{(s)} \in \mathbb{R}^{p_i}$ with $1 \leq i \leq 3$ and $1 \leq s \leq r$ such that $\mathcal{X} = \sum_{s=1}^r \mathbf{x}_1^{(s)} \circ \mathbf{x}_2^{(s)} \circ \mathbf{x}_3^{(s)}$, its rank is then defined as $\text{rank}(\mathcal{X}) = r$, where \circ is the vector outer product. For second-order tensors, i.e. matrices, $\text{rank}(\mathcal{X}) = \text{rank}_1(\mathcal{X}) = \text{rank}_2(\mathcal{X})$, while we may have $\max_{1 \leq j \leq K} \text{rank}_j(\mathcal{X}) \leq \text{rank}(\mathcal{X})$ for general K th-order tensors with $K \geq 3$. This paper will concentrate on the Tucker decomposition and Tucker ranks for reasons to be explained in the next subsection.

2.2 Multilinear low-rank vector autoregression

For the VAR model in (1), we can rearrange its transition matrices into a tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times P}$; see Figure 1 for an illustration. Denote by $\mathcal{A}_{(j)}$ the mode- j matricization of \mathcal{A} , where $1 \leq j \leq 3$. It can be verified that $\mathcal{A}_{(1)} = (\mathbf{A}_1, \dots, \mathbf{A}_P)$, $\mathcal{A}_{(2)} = (\mathbf{A}'_1, \mathbf{A}'_2, \dots, \mathbf{A}'_P)$ and $\mathcal{A}_{(3)} = (\text{vec}(\mathbf{A}_1), \text{vec}(\mathbf{A}_2), \dots, \text{vec}(\mathbf{A}_P))'$, which correspond to the column, row and vectorized matrix spaces of \mathbf{A}_j s, respectively.

If the transition tensor \mathcal{A} has multilinear low ranks (r_1, r_2, r_3) , i.e. $\text{rank}(\mathcal{A}_{(j)}) = r_j$ for $1 \leq j \leq 3$, then there exists a Tucker decomposition, $\mathcal{A} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ or $\mathcal{A} = \llbracket \mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \rrbracket$, where $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, and $\mathbf{U}_1 \in \mathbb{R}^{N \times r_1}$, $\mathbf{U}_2 \in \mathbb{R}^{N \times r_2}$ and $\mathbf{U}_3 \in \mathbb{R}^{P \times r_3}$ are factor matrices. As a result, model (1) can be written as

$$\mathbf{y}_t = (\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} \mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad (3)$$

where $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \dots, \mathbf{y}'_{t-P})'$. For simplicity, we call model (3) the multilinear low-rank

VAR model.

Assumption 1. *All roots of the matrix polynomial $\mathcal{A}(z) = \mathbf{I}_N - \mathbf{A}_1 z - \dots - \mathbf{A}_P z^P$, $z \in \mathbb{C}$, are outside the unit circle, where \mathbb{C} is the set of complex numbers.*

Assumption 1 is the sufficient and necessary condition for the existence of a unique strictly stationary solution to model (1). When $P = 1$, Assumption 1 is equivalent to the spectral radius of \mathbf{A}_1 being strictly less than one.

Remark 1. *Suppose that the entries of \mathbf{A}_1 are i.i.d. random variables with mean zero and variance σ^2 , i.e. they are equally important. Then the spectral radius of \mathbf{A}_1 will converge to $\sqrt{N}\sigma$ in probability as $N \rightarrow \infty$ (Bai, 1997).*

Note that the Tucker decomposition in (3) is not unique since $\llbracket \mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \rrbracket = \llbracket \mathcal{G} \times_1 \mathbf{O}_1 \times_2 \mathbf{O}_2 \times_3 \mathbf{O}_3; \mathbf{U}_1 \mathbf{O}_1^{-1}, \mathbf{U}_2 \mathbf{O}_2^{-1}, \mathbf{U}_3 \mathbf{O}_3^{-1} \rrbracket$ for any nonsingular matrices $\mathbf{O}_1 \in \mathbb{R}^{r_1 \times r_1}$, $\mathbf{O}_2 \in \mathbb{R}^{r_2 \times r_2}$ and $\mathbf{O}_3 \in \mathbb{R}^{r_3 \times r_3}$. Hence, we consider a special Tucker decomposition: the higher-order singular value decomposition (HOSVD); see De Lathauwer et al. (2000). Specifically, we let \mathbf{U}_j be a tall matrix consisting of the top r_j left singular vectors of $\mathcal{A}_{(j)}$ for each $1 \leq j \leq 3$, where (r_1, r_2, r_3) are the multilinear ranks of the tensor \mathcal{A} . Let the core tensor $\mathcal{G} = \mathcal{A} \times_1 \mathbf{U}'_1 \times_2 \mathbf{U}'_2 \times_3 \mathbf{U}'_3$. Then \mathcal{G} has the following *all-orthogonal* property: for each $1 \leq j \leq 3$, the rows of $\mathcal{G}_{(j)}$ are pairwise orthogonal. If the following condition further holds, then all components in the HOSVD are uniquely defined.

Assumption 2. *For each $1 \leq j \leq 3$, the singular values of $\mathcal{A}_{(j)}$ are distinct, and the first nonzero element in each column of \mathbf{U}_j is positive.*

Since $(\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} = \mathbf{U}_1 \mathcal{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)'$ and \mathbf{U}_1 is orthonormal, model (3) implies that

$$\mathbf{U}'_1 \mathbf{y}_t = \mathcal{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)' \mathbf{x}_t + \mathbf{U}'_1 \boldsymbol{\epsilon}_t = \mathcal{G}_{(1)} \text{vec}(\mathbf{U}'_2 \mathbf{X}_t \mathbf{U}_3) + \mathbf{U}'_1 \boldsymbol{\epsilon}_t,$$

where $\mathbf{X}_t = (\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-P})$, and \otimes denotes the Kronecker product. Then, we may view $\mathbf{U}'_1 \mathbf{y}_t$ at the left side of the above model equation as r_1 factors across the N variables of the response \mathbf{y}_t . Similarly, for the bilinear form $\mathbf{U}'_2 \mathbf{X}_t \mathbf{U}_3$, we may view $\mathbf{U}'_2 \mathbf{X}_t = (\mathbf{U}'_2 \mathbf{y}_{t-1}, \dots, \mathbf{U}'_2 \mathbf{y}_{t-P})$ as r_2 factors across the N variables of the predictors

\mathbf{y}_{t-j} s, and $\mathbf{U}'_3 \mathbf{X}'_t = (\mathbf{U}'_3 \mathbf{x}_{1t}^{(L)}, \dots, \mathbf{U}'_3 \mathbf{x}_{Nt}^{(L)})$ as r_3 factors across the P time lags, where $\mathbf{x}_{jt}^{(L)} = (y_{j,t-1}, \dots, y_{j,t-P})'$ with $1 \leq j \leq N$; note that similar formulations can be found in matrix variate regressions such as Zhao and Leng (2014) and Ding and Cook (2018). Thus, we call r_1 , r_2 and r_3 the (cross-sectional) response, (cross-sectional) predictor and temporal ranks, respectively.

Due to the HOSVD, the proposed multilinear low-rank VAR model in (3) has only $r_1 r_2 r_3 + (N - r_1)r_1 + (N - r_2)r_2 + (P - r_3)r_3$ parameters, i.e. the dimension increases linearly in N and P ; see Zhang (2019). By contrast, model (1) has $N^2 P$ parameters, while the reduced-rank VAR model in (2) has $(NP + N - r_1)r_1$ parameters, where $r_1 = \text{rank}(\mathcal{A}_{(1)})$.

One may also consider the CP decomposition for the transition tensor \mathcal{A} . The CP decomposition can be treated as a special case of the Tucker decomposition, where the factor matrices have the same number of columns and the core tensor is superdiagonal. This will result in a model with even fewer parameters. However, the assumption that the ranks of \mathcal{A} are equal in all directions may be too restrictive in practice.

2.3 Relationship with VARMA processes

This subsection discusses the relationship between the proposed multilinear low-rank model in (3) and VARMA processes.

We first consider the vector moving average (VMA) process of order one:

$$\mathbf{y}_t = \boldsymbol{\epsilon}_t - \boldsymbol{\Theta} \boldsymbol{\epsilon}_{t-1}, \quad (4)$$

where $\{\boldsymbol{\epsilon}_t\}$ are defined as in (3). Suppose that the above process is invertible, i.e. the spectral radius of $\boldsymbol{\Theta}$ is strictly less than one. Then it has the following VAR(∞) representation,

$$\mathbf{y}_t = \boldsymbol{\epsilon}_t + \boldsymbol{\Phi}_1 \mathbf{y}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{y}_{t-2} + \boldsymbol{\Phi}_3 \mathbf{y}_{t-3} + \dots,$$

where $\boldsymbol{\Phi}_j = -\boldsymbol{\Theta}^j$ for $j \geq 1$, and the corresponding transition tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times \infty}$ satisfies that $\mathcal{A}_{(1)} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \boldsymbol{\Phi}_3, \dots)$.

Proposition 1. *Suppose that the spectral radius of $\boldsymbol{\Theta}$ is strictly less than one, and $\text{rank}(\boldsymbol{\Theta}) = r$. Then the transition tensor \mathcal{A} of model (4) has multilinear ranks at most (r, r, r) .*

Fitting model (4) by directly applying the least squares method is quite challenging since a $2(T-1)$ th order polynomial with respect to Θ will be involved in the objective function, which may lead to unstable performance of the optimization. The common practice is to approximate the process by a VAR(P) model with a sufficiently large P . Note that the derivation of asymptotic properties for univariate and multivariate cases requires $P \rightarrow \infty$ and $PT^{-1/3} \rightarrow 0$ as $T \rightarrow \infty$ (Said and Dickey, 1984; Li et al., 2014). This condition would entail a VAR model with a large number of parameters, whereas the VMA(1) model in (4) has only $(2N-r)r$ parameters. For the proposed multilinear low-rank model, it follows from Proposition 1 that the number of parameters is only $r^3 + 2r(N-r) + r(P-r)$, which is comparable to that of the VMA(1) model as long as $P = O(N)$.

We next consider the VARMA(1, 1) process

$$\mathbf{y}_t = \Psi \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t - \Theta \boldsymbol{\epsilon}_{t-1}. \quad (5)$$

If the spectral radii of Ψ and Θ are both strictly less than one, then the process is stationary and invertible, with the VAR(∞) representation $\mathbf{y}_t = \boldsymbol{\epsilon}_t + \sum_{j=1}^{\infty} \Phi_j \mathbf{y}_{t-j}$, where the transition tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times \infty}$ satisfies $\mathcal{A}_{(1)} = (\Phi_1, \Phi_2, \Phi_3, \dots)$, and $\Phi_j = -\Theta^{j-1}(\Theta - \Psi)$ for $j \geq 1$.

Proposition 2. *Suppose that the spectral radii of Ψ and Θ are both strictly less than one, with $\text{rank}(\Theta) = r$ and $\text{rank}(\Psi) = s$. Then the transition tensor \mathcal{A} of model (5) has multilinear ranks at most $(r+s, r+s, r+1)$.*

The VARMA model suffers from a serious identification problem: its uniqueness cannot be guaranteed unless additional parameter constraints, such as the echelon form or the final equations form, are imposed (Lütkepohl, 2005). The estimation of the resulting constrained VARMA model involves complicated nonlinear optimization which may be even more challenging for large-scale data. Therefore, compared to VARMA modeling, the VAR(P) approximation is a more viable approach. This further motivates the proposed multilinear low-rank VAR modeling methodology.

Remark 2. *For the VARMA process of general orders, it can be shown that its VAR(∞) representation exists under certain regularity conditions, and the corresponding transi-*

tion tensor also has a multilinear low-rank structure. However, the assumptions and corresponding low-rank structures have complicated forms and hence are omitted here.

3 Low-dimensional time series modeling

3.1 Multilinear low-rank least squares estimation

For the multilinear low-rank VAR model in (3) with ranks (r_1, r_2, r_3) , the multilinear low-rank (MLR) least squares estimator can be defined as

$$\hat{\mathcal{A}}_{\text{MLR}} \equiv [\hat{\mathcal{G}}; \hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2, \hat{\mathbf{U}}_3] = \arg \min \sum_{t=1}^T \|\mathbf{y}_t - (\mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} \mathbf{x}_t\|_2^2. \quad (6)$$

This section studies asymptotic properties of $\hat{\mathcal{A}}_{\text{MLR}}$ with both N and P being fixed.

Let $\boldsymbol{\phi} = (\text{vec}(\mathcal{G}_{(1)})', \text{vec}(\mathbf{U}_1)', \text{vec}(\mathbf{U}_2)', \text{vec}(\mathbf{U}_3)')'$ be the true value of the vectorized HOSVD components and $\hat{\boldsymbol{\phi}}_{\text{MLR}} = (\text{vec}(\hat{\mathcal{G}}_{(1)})', \text{vec}(\hat{\mathbf{U}}_1)', \text{vec}(\hat{\mathbf{U}}_2)', \text{vec}(\hat{\mathbf{U}}_3)')'$ be the corresponding estimator. Let $\mathbf{h}(\boldsymbol{\phi}) = \text{vec}(\mathcal{A}_{(1)}) = \text{vec}(\mathbf{U}_1 \mathcal{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)')$ be a function of $\boldsymbol{\phi}$. Let $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \text{var}(\boldsymbol{\epsilon}_t)$, $\boldsymbol{\Gamma}_j = \text{cov}(\mathbf{y}_{t+j}, \mathbf{y}_t)$ with $j \geq 0$,

$$\boldsymbol{\Gamma}^* = \begin{bmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{P-1} \\ \boldsymbol{\Gamma}'_1 & \boldsymbol{\Gamma}_0 & \cdots & \boldsymbol{\Gamma}_{P-2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}'_{P-1} & \boldsymbol{\Gamma}'_{P-2} & \cdots & \boldsymbol{\Gamma}_0 \end{bmatrix},$$

and $\mathbf{J} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes \boldsymbol{\Gamma}^*$. Denote

$$\mathbf{H} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\phi}} = \left((\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1), [(\mathbf{U}_3 \otimes \mathbf{U}_2) \mathcal{G}'_{(1)}] \otimes \mathbf{I}_N, \mathbf{T}_{21} \{[(\mathbf{U}_1 \otimes \mathbf{U}_3) \mathcal{G}'_{(2)}] \otimes \mathbf{I}_N\}, \right. \\ \left. \mathbf{T}_{31} \{[(\mathbf{U}_1 \otimes \mathbf{U}_2) \mathcal{G}'_{(3)}] \otimes \mathbf{I}_P\} \right),$$

where \mathbf{T}_{ij} is an $(N^2 P) \times (N^2 P)$ permutation matrix such that $\text{vec}(\mathcal{A}_{(j)}) = \mathbf{T}_{ij} \text{vec}(\mathcal{A}_{(i)})$ with $1 \leq i, j \leq 3$.

Theorem 1. Suppose that the time series $\{\mathbf{y}_t\}$ is generated by model (3) with $E\|\boldsymbol{\epsilon}_t\|^4 < \infty$, and Assumption 1 holds. Then,

$$\sqrt{T} \{\mathbf{h}(\hat{\boldsymbol{\phi}}_{\text{MLR}}) - \mathbf{h}(\boldsymbol{\phi})\} \rightarrow N(0, \boldsymbol{\Sigma}_{\text{MLR}})$$

in distribution as $T \rightarrow \infty$, where $\Sigma_{\text{MLR}} = \mathbf{H}(\mathbf{H}'\mathbf{J}\mathbf{H})^\dagger\mathbf{H}'$, and \dagger denotes the Moore-Penrose inverse.

The proof of Theorem 1 relies on the technique for deriving asymptotic distributions of overparameterized models in Shapiro (1986); see Appendix A.1 for details. Under Assumption 2, we can further establish the asymptotic properties of $\hat{\mathbf{U}}_i$ and $\hat{\mathbf{G}}$ as follows.

Corollary 1. *Suppose that the conditions of Theorem 1 and Assumption 2 hold. Then $\sqrt{T}\{\text{vec}(\hat{\mathbf{G}}) - \text{vec}(\mathbf{G})\}$, $\sqrt{T}\{\text{vec}(\hat{\mathbf{U}}_1) - \text{vec}(\mathbf{U}_1)\}$, $\sqrt{T}\{\text{vec}(\hat{\mathbf{U}}_2) - \text{vec}(\mathbf{U}_2)\}$, and $\sqrt{T}\{\text{vec}(\hat{\mathbf{U}}_3) - \text{vec}(\mathbf{U}_3)\}$ converge to normal distributions as $T \rightarrow \infty$.*

The next theorem shows that the proposed estimator $\hat{\mathbf{A}}_{\text{MLR}}$ is asymptotically more efficient than the ordinary least squares (OLS) estimator

$$\hat{\mathbf{A}}_{\text{OLS}} = \arg \min_{\mathbf{B} \in \mathbb{R}^{N \times NP}} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{B}\mathbf{x}_t\|_2^2$$

for the full VAR model in (1) and the reduced-rank regression (RRR) estimator

$$\hat{\mathbf{A}}_{\text{RRR}} = \arg \min_{\mathbf{B} \in \mathbb{R}^{N \times NP}, \text{rank}(\mathbf{B})=r_1} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{B}\mathbf{x}_t\|_2^2$$

for the reduced-rank VAR model in (2), where r_1 is the rank of $\mathbf{A}_{(1)}$. Denote by $\hat{\mathbf{A}}_{\text{OLS}}$ and $\hat{\mathbf{A}}_{\text{RRR}}$ the transition tensors formed by $\hat{\mathbf{A}}_{\text{OLS}}$ and $\hat{\mathbf{A}}_{\text{RRR}}$, respectively.

Theorem 2. *If the conditions of Theorem 1 hold, then $\sqrt{T}\{\text{vec}(\hat{\mathbf{A}}_{\text{OLS}}) - \mathbf{h}(\phi)\} \rightarrow N(0, \Sigma_{\text{OLS}})$ and $\sqrt{T}\{\text{vec}(\hat{\mathbf{A}}_{\text{RRR}}) - \mathbf{h}(\phi)\} \rightarrow N(0, \Sigma_{\text{RRR}})$ in distribution as $T \rightarrow \infty$. Moreover, it holds that $\Sigma_{\text{MLR}} \leq \Sigma_{\text{RRR}} \leq \Sigma_{\text{OLS}}$.*

3.2 Alternating least squares algorithm

Let $\mathcal{F}_t = \sigma(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, \dots)$ be the σ -field generated by $\{\boldsymbol{\epsilon}_s, s \leq t\}$ and recall that $\mathbf{X}_t = (\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-P})$. The objective function in (6) is a nonlinear function of \mathbf{G} , \mathbf{U}_1 , \mathbf{U}_2 and \mathbf{U}_3 . However, from model (3), we have

$$\begin{aligned} \mathbb{E}(\mathbf{y}_t | \mathcal{F}_{t-1}) &= (\mathbf{x}'_t(\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{G}'_{(1)} \otimes \mathbf{I}_N)\text{vec}(\mathbf{U}_1) \\ &= \mathbf{U}_1\mathbf{G}_{(1)}((\mathbf{U}'_3\mathbf{X}'_t) \otimes \mathbf{I}_{r_2})\text{vec}(\mathbf{U}'_2) \\ &= \mathbf{U}_1\mathbf{G}_{(1)}(\mathbf{I}_{r_3} \otimes (\mathbf{U}'_2\mathbf{X}_t))\text{vec}(\mathbf{U}_3) \\ &= (((\mathbf{U}_3 \otimes \mathbf{U}_2)'\mathbf{x}_t)' \otimes \mathbf{U}_1)\text{vec}(\mathbf{G}_{(1)}), \end{aligned}$$

which implies that the objective function in (6) is linear with respect to any of \mathcal{G} , \mathbf{U}_1 , \mathbf{U}_2 and \mathbf{U}_3 when the other three are fixed. Hence, we can employ the alternating least squares algorithm to find $\hat{\mathcal{A}}_{\text{MLR}}$; see Algorithm 1. The convergence of the algorithm is shown by Proposition 3.

Algorithm 1 Alternating least squares algorithm for $\hat{\mathcal{A}}_{\text{MLR}}$

Initialize: $\mathcal{A}^{(0)} = \hat{\mathcal{A}}_{\text{OLS}}$ or $\hat{\mathcal{A}}_{\text{RRR}}$

HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \mathbf{U}_1^{(0)} \times_2 \mathbf{U}_2^{(0)} \times_3 \mathbf{U}_3^{(0)}$ with multilinear ranks (r_1, r_2, r_3) .

repeat $k = 0, 1, 2, \dots$

$$\mathbf{U}_1^{(k+1)} \leftarrow \arg \min_{\mathbf{U}_1} \sum_{t=1}^T \|\mathbf{y}_t - (\mathbf{x}_t'(\mathbf{U}_3^{(k)} \otimes \mathbf{U}_2^{(k)})(\mathcal{G}_{(1)}^{(k)})' \otimes \mathbf{I}_K) \text{vec}(\mathbf{U}_1)\|_2^2$$

$$\mathbf{U}_2^{(k+1)} \leftarrow \arg \min_{\mathbf{U}_2} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} ((\mathbf{X}_t \mathbf{U}_3^{(k)})' \otimes \mathbf{I}_{r_2}) \text{vec}(\mathbf{U}_2')\|_2^2$$

$$\mathbf{U}_3^{(k+1)} \leftarrow \arg \min_{\mathbf{U}_3} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} (\mathbf{I}_{r_3} \otimes (\mathbf{U}_2^{(k+1)})' \mathbf{X}_t) \text{vec}(\mathbf{U}_3)\|_2^2$$

$$\mathcal{G}^{(k+1)} \leftarrow \arg \min_{\mathcal{G}} \sum_{t=1}^T \|\mathbf{y}_t - (((\mathbf{U}_3^{(k+1)} \otimes \mathbf{U}_2^{(k+1)})' \mathbf{x}_t)' \otimes \mathbf{U}_1^{(k+1)}) \text{vec}(\mathcal{G}_{(1)})\|_2^2$$

$$\mathcal{A}^{(k+1)} \leftarrow \mathcal{G}^{(k+1)} \times_1 \mathbf{U}_1^{(k+1)} \times_2 \mathbf{U}_2^{(k+1)} \times_3 \mathbf{U}_3^{(k+1)}$$

until convergence

Finalize: $\hat{\mathbf{U}}_i \leftarrow$ top r_i left singular vectors of $\hat{\mathcal{A}}_{(i)}$, for $1 \leq i \leq 3$

$$\hat{\mathcal{G}} \leftarrow \llbracket \hat{\mathcal{A}}; \hat{\mathbf{U}}_1', \hat{\mathbf{U}}_2', \hat{\mathbf{U}}_3' \rrbracket$$

Proposition 3. *Suppose that the stationary points of the objective function in (6) are isolated, up to an arbitrary nonsingular linear transformation. Then $\phi^{(k)}$ converges to a stationary point as $k \rightarrow \infty$, where $\phi^{(k)} = (\text{vec}(\mathcal{G}^{(k)})', \text{vec}(\mathbf{U}_1^{(k)})', \text{vec}(\mathbf{U}_2^{(k)})', \text{vec}(\mathbf{U}_3^{(k)})')'$. Moreover, let $\phi^{(\infty)} = (\text{vec}(\mathcal{G}^{(\infty)})', \text{vec}(\mathbf{U}_1^{(\infty)})', \text{vec}(\mathbf{U}_2^{(\infty)})', \text{vec}(\mathbf{U}_3^{(\infty)})')'$ be a local minimum of the objective function. Then $\{\phi^{(k)}\}$ will be attracted to $\phi^{(\infty)}$ if the initial value $\phi^{(0)}$ is sufficiently close to $\phi^{(\infty)}$.*

Remark 3. *In Algorithm 1, we do not need the all-orthogonal constraint on \mathcal{G} or orthonormal constraints on the \mathbf{U}_i s. Instead, we obtain $\hat{\mathcal{G}}$ by projecting the solution $\hat{\mathcal{A}}$ onto the tensor space $\mathcal{S} = \{\mathcal{B} \in \mathbb{R}^{N \times N \times P} : \mathcal{B} \text{ has multilinear ranks } (r_1, r_2, r_3)\}$ via the HOSVD; that is, we let $\hat{\mathbf{U}}_i$ be the matrix of the top r_i left singular vectors of $\hat{\mathcal{A}}_{(i)}$ for $1 \leq i \leq 3$, and then set $\hat{\mathcal{G}} = \llbracket \hat{\mathcal{A}}; \hat{\mathbf{U}}_1', \hat{\mathbf{U}}_2', \hat{\mathbf{U}}_3' \rrbracket$.*

3.3 Rank selection based on the BIC

Algorithm 1 requires predetermined multilinear ranks (r_1, r_2, r_3) . To select these ranks, we propose to minimize the following Bayesian information criterion (BIC):

$$\text{BIC} = T \log \left(\frac{1}{T} \sum_{t=1}^T \|\mathbf{y}_t - \hat{\mathcal{A}}_{(1)} \mathbf{x}_t\|_2^2 \right) + (\text{df} + 1) \log T,$$

where the number of degrees of freedom is defined as the number of free parameters in the Tucker decomposition of the transition tensor \mathcal{A} :

$$\text{df} = r_1 r_2 r_3 + r_1(N - r_1) + r_2(N - r_2) + r_3(P - r_3).$$

Remark 4. To reduce the computational cost, before minimizing the BIC, we may first select the plausible multilinear ranks by the following procedure: (1) for a given initial value $\tilde{\mathcal{A}}$, say $\hat{\mathcal{A}}_{\text{OLS}}$, of the transition tensor \mathcal{A} , compute the singular values for the matricizations $\mathcal{A}_{(i)}$ for each mode $1 \leq i \leq 3$; (2) select the plausible multilinear ranks by maximizing the ratio of two adjacent singular values (Ahn and Horenstein, 2013). Then, we can search the optimal ranks in the neighborhood of the plausible ranks.

4 High-dimensional time series modeling

4.1 Sparse higher-order reduced-rank estimation

The proposed multilinear low-rank VAR model allows us to reduce a large number of predictors and responses to a few factors, while each factor is a combination of all variables. However, when the number of variables N is very large, it may be more desirable to incorporate variable selection into those factors to improve interpretability. In this section, we assume that there exists an HOSVD for the transition tensor \mathcal{A} such that the factor matrices \mathbf{U}_1 , \mathbf{U}_2 and \mathbf{U}_3 are sparse orthogonal matrices, and hence each factor consists of only a subset of the N variables.

Specifically, we introduce the following ℓ_1 -penalized Sparse Higher-Order Reduced-

Rank (SHORR) estimator:

$$\begin{aligned} \hat{\mathcal{A}}_{\text{SHORR}} &\equiv \llbracket \hat{\mathbf{G}}; \hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2, \hat{\mathbf{U}}_3 \rrbracket \\ &= \arg \min_{\mathbf{G}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3} \left\{ \frac{1}{T} \sum_{t=1}^T \|\mathbf{y}_t - (\mathbf{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3)_{(1)} \mathbf{x}_t\|_2^2 + \lambda \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1 \right\} \quad (7) \\ &\text{subject to } \mathbf{G} \in \text{AO}(r_1, r_2, r_3) \text{ and } \mathbf{U}_i' \mathbf{U}_i = \mathbf{I}_{r_i}, \quad i = 1, 2, 3, \end{aligned}$$

where $\text{AO}(r_1, r_2, r_3) = \{\mathbf{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \mathbf{G}_{(i)} \text{ is row-orthogonal, } i = 1, 2, 3\}$. The all-orthogonal constraint on \mathbf{G} and the orthonormal constraints on \mathbf{U}_i are imposed for the sake of identifiability.

Remark 5. *The proposed SHORR estimation method is different from the row-sparse reduced-rank regression that has been studied extensively in the literature (Chen and Huang, 2012; Bunea et al., 2012). We avoid imposing the row-sparsity because (1) it would restrict the flexibility and interpretability of the VAR model, and (2) with a row-sparse response factor matrix \mathbf{U}_1 , those unselected time series cannot be predicted at all. Thus, we consider the general sparsity structure for the \mathbf{U}_i s rather than the row-sparsity.*

Remark 6. *Alternatively, one might consider penalizing each \mathbf{U}_i individually, e.g. with the penalty term $\sum_{i=1}^3 \lambda_i \|\mathbf{U}_i\|_1$. However, the three tuning parameters would cause much higher computational costs. Instead, the SHORR estimator induces sparsity for \mathbf{U}_1 , \mathbf{U}_2 and \mathbf{U}_3 jointly since $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1 = \|\mathbf{U}_3\|_1 \|\mathbf{U}_2\|_1 \|\mathbf{U}_1\|_1$. Implementation of this joint penalty is convenient through the alternating algorithm to be introduced in Section 4.3. Moreover, when P is relatively small, we might wish to impose sparsity on \mathbf{U}_1 and \mathbf{U}_2 only, and then $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1$ can be replaced by $\|\mathbf{U}_2 \otimes \mathbf{U}_1\|_1$.*

4.2 Theoretical properties of the SHORR estimator

To derive the non-asymptotic estimation and prediction error bounds of the SHORR estimator, we make the following assumptions.

Assumption 3. *(Gaussian error) The errors $\{\epsilon_t\}$ are i.i.d. Gaussian random vectors with mean zero and positive definite covariance matrix Σ_ϵ .*

Assumption 4. *(Sparsity) The factor matrices satisfy that $\|\mathbf{U}_i\|_0 \leq s_i$, for $1 \leq i \leq 3$.*

Assumption 5. (*Restricted parameter space*) The parameter space for \mathbf{G} and \mathbf{U}_i with $1 \leq i \leq 3$ is $\Omega = \{\mathbf{G} \in \text{AO}(r_1, r_2, r_3) : \|\mathbf{G}\|_\infty \leq \bar{g} < \infty\} \times \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3$, where $\mathcal{U}_i = \{\mathbf{U} \in \mathbb{R}^{p_i \times r_i} : \mathbf{U}'\mathbf{U} = \mathbf{I}_{r_i}, \text{ and } \mathbf{U}_{ij}^2 \geq \nu > 0 \text{ or } \mathbf{U}_{ij} = 0\}$ with $p_1 = p_2 = N$ and $p_3 = P$, and ν is a uniform lower threshold for elements of \mathbf{U}_i s.

Assumption 6. (*Relative spectral gap*) The nonzero singular values of $\mathbf{A}_{(i)}$ satisfy that $\sigma_{j-1}^2(\mathbf{A}_{(i)}) - \sigma_j^2(\mathbf{A}_{(i)}) \geq \delta \sigma_{j-1}^2(\mathbf{A}_{(i)})$ for $2 \leq j \leq r_i$ and $1 \leq i \leq 3$, where δ is a positive constant.

Assumption 3 enables us to apply the concentration inequalities for VAR models in Basu and Michailidis (2015). The Gaussian condition may be relaxed to sub-Gaussianity by techniques in Zheng and Raskutti (2018). Assumption 4 states the sparsity of each factor matrix. Assumption 5 imposes an upper bound on the core tensor \mathbf{G} , which is not a stringent assumption since large values in \mathbf{G} could cause nonstationarity of the VAR process. The lower threshold ν for the \mathbf{U}_i s is essential to restrict the estimation error to a subspace such that the restricted eigenvalue condition (Bickel et al., 2009) can be established. Note that ν may shrink to zero as the dimension increases, so this condition is not too stringent. Assumption 6 guarantees that the singular values of each $\mathbf{A}_{(i)}$ are well separated. This rules out non-identifiability and allows us to derive the upper bound for the perturbation errors in Lemma 1 of Appendix A.3.

Assumption 1 guarantees that the eigenvalues of the Hermitian matrix $\mathcal{A}^*(z)\mathcal{A}(z)$ over the unit circle $\{z \in \mathbb{C} : |z| = 1\}$ are all positive, where $\mathcal{A}^*(z)$ denotes the conjugate transpose of $\mathcal{A}(z)$. Following Basu and Michailidis (2015), let

$$\mu_{\min}(\mathcal{A}) = \min_{|z|=1} \lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z)) \quad \text{and} \quad \mu_{\max}(\mathcal{A}) = \max_{|z|=1} \lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)),$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a matrix, respectively. It holds that

$$\mu_{\min}(\mathcal{A}) = \min_{\theta \in [-\pi, \pi]} \lambda_{\min} \left(\left(\mathbf{I}_N - \sum_{p=1}^P \mathbf{A}'_p e^{ip\theta} \right) \left(\mathbf{I}_N - \sum_{p=1}^P \mathbf{A}'_p e^{-ip\theta} \right) \right)$$

and

$$\mu_{\max}(\mathcal{A}) = \max_{\theta \in [-\pi, \pi]} \lambda_{\max} \left(\left(\mathbf{I}_N - \sum_{p=1}^P \mathbf{A}'_p e^{ip\theta} \right) \left(\mathbf{I}_N - \sum_{p=1}^P \mathbf{A}'_p e^{-ip\theta} \right) \right).$$

Theorem 3. Suppose that Assumptions 1 and 3-6 hold. If $\lambda \gtrsim \mathcal{M}\sqrt{\log(N^2P)/T}$ and $T \gtrsim \log(N^2P) + \mathcal{M}^2d \min[\log(NP), \log(cNP/d)]$, then

$$\|\hat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}\|_F \leq C_1 \tau \sqrt{S} \lambda / \alpha, \quad (8)$$

and

$$T^{-1} \sum_{t=1}^T \|(\hat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A})_{(1)} \mathbf{x}_t\|_2^2 \leq C_2 \tau^2 S \lambda^2 / \alpha, \quad (9)$$

with probability at least $1 - C \exp[-c \log(N^2P)] - C \exp\{-cd \min[\log(NP), \log(cNP/d)]\}$, where $c, C, C_1, C_2 > 0$ are absolute constants, $\mathcal{M} = \lambda_{\max}(\mathbf{\Sigma}_{\epsilon}) (1 + \mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A}))$, $d = \nu^{-2} r_1 r_2 r_3$, $\tau = \delta^{-1}(\eta_1 \sqrt{r_2 r_3} + \eta_2 \sqrt{r_1 r_3} + \eta_3 \sqrt{r_1 r_2})$ with $\eta_i = (\sum_{j=1}^{r_i} \sigma_1^2(\mathcal{A}_{(i)}) / \sigma_j^4(\mathcal{A}_{(i)}))^{1/2}$, $S = (\sqrt{s_1 s_2 s_3} + r_1 r_2 r_3)^2$ and $\alpha = \lambda_{\min}(\mathbf{\Sigma}_{\epsilon}) / \mu_{\max}(\mathcal{A})$.

From Theorem 3, when the multilinear ranks (r_1, r_2, r_3) , degrees of sparsity (s_1, s_2, s_3) and lower threshold ν are fixed, the proposed SHORR estimator is consistent if $T \gtrsim \log(N^2P)$, which is the same as the sample size requirement for the Lasso estimator for VAR models (Basu and Michailidis, 2015). In this case, the estimation and prediction error bounds in (8) and (9) become $O(\sqrt{S \log(N^2P)/T})$ and $O(S \log(N^2P)/T)$, respectively. Note that they are both smaller than the corresponding error bounds for the Lasso estimator when $S \lesssim \|\mathcal{A}\|_0$, i.e. when the true transition tensor \mathcal{A} is dense but has a sparse low-rank Tucker decomposition.

4.3 ADMM-based algorithm

The optimization problem in (7) is nonconvex due to the nonconvex ℓ_1 penalty term $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1 = \|\mathbf{U}_3\|_1 \|\mathbf{U}_2\|_1 \|\mathbf{U}_1\|_1$. However, similarly to Algorithm 1, we can update each \mathbf{U}_i by an alternating regularized algorithm; see Algorithm 2.

Note that the subproblems of updating \mathbf{U}_i for $1 \leq i \leq 3$ in Algorithm 2 have the general form:

$$\min_{\mathbf{B}} \{n^{-1} \|\mathbf{y} - \mathbf{X} \text{vec}(\mathbf{B})\|_2^2 + \lambda \|\mathbf{B}\|_1\}, \quad \text{s.t. } \mathbf{B}' \mathbf{B} = \mathbf{I}. \quad (10)$$

It involves both the ℓ_1 regularization and the orthogonality constraint, which are difficult to handle jointly since the orthogonality constraint for factor matrices is nonconvex.

To solve this subproblem, we propose an alternating direction method of multipliers

Algorithm 2 Alternating regularized algorithm for SHORR estimator

Initialize: $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{RRR}}$

HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \mathbf{U}_1^{(0)} \times_2 \mathbf{U}_2^{(0)} \times_3 \mathbf{U}_3^{(0)}$ with multilinear ranks (r_1, r_2, r_3) .

repeat $k = 0, 1, 2, \dots$

$$\begin{aligned}
\mathbf{U}_1^{(k+1)} &\leftarrow \arg \min_{\mathbf{U}_1' \mathbf{U}_1 = \mathbf{I}_{r_1}} \left\{ T^{-1} \sum_{t=1}^T \|\mathbf{y}_t - (\mathbf{x}_t' (\mathbf{U}_3^{(k)} \otimes \mathbf{U}_2^{(k)}) (\mathcal{G}_{(1)}^{(k)})' \otimes \mathbf{I}_N) \text{vec}(\mathbf{U}_1)\|_2^2 \right. \\
&\quad \left. + \lambda \|\mathbf{U}_1\|_1 \|\mathbf{U}_2^{(k)}\|_1 \|\mathbf{U}_3^{(k)}\|_1 \right\} \\
\mathbf{U}_2^{(k+1)} &\leftarrow \arg \min_{\mathbf{U}_2' \mathbf{U}_2 = \mathbf{I}_{r_2}} \left\{ T^{-1} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} ((\mathbf{X}_t \mathbf{U}_3^{(k)})' \otimes \mathbf{I}_{r_2}) \text{vec}(\mathbf{U}_2')\|_2^2 \right. \\
&\quad \left. + \lambda \|\mathbf{U}_1^{(k+1)}\|_1 \|\mathbf{U}_2\|_1 \|\mathbf{U}_3^{(k)}\|_1 \right\} \\
\mathbf{U}_3^{(k+1)} &\leftarrow \arg \min_{\mathbf{U}_3' \mathbf{U}_3 = \mathbf{I}_{r_3}} \left\{ T^{-1} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} (\mathbf{I}_{r_3} \otimes (\mathbf{U}_2^{(k+1)})' \mathbf{X}_t) \text{vec}(\mathbf{U}_3)\|_2^2 \right. \\
&\quad \left. + \lambda \|\mathbf{U}_1^{(k+1)}\|_1 \|\mathbf{U}_2^{(k+1)}\|_1 \|\mathbf{U}_3\|_1 \right\} \\
\mathcal{G}^{(k+1)} &\leftarrow \arg \min_{\mathcal{G} \in \text{AO}(r_1, r_2, r_3)} \sum_{t=1}^T \|\mathbf{y}_t - (((\mathbf{U}_3^{(k+1)} \otimes \mathbf{U}_2^{(k+1)})' \mathbf{x}_t)' \otimes \mathbf{U}_1^{(k+1)}) \text{vec}(\mathcal{G}_{(1)})\|_2^2 \\
\mathcal{A}^{(k+1)} &\leftarrow \mathcal{G}^{(k+1)} \times_1 \mathbf{U}_1^{(k+1)} \times_2 \mathbf{U}_2^{(k+1)} \times_3 \mathbf{U}_3^{(k+1)}
\end{aligned}$$

until convergence

(ADMM) algorithm (Boyd et al., 2011) to separate the ℓ_1 regularization and the orthogonality constraint into two steps. Specifically, we define the dummy variable \mathbf{W} as a surrogate for \mathbf{B} and write problem (10) into the equivalent form as follows:

$$\min_{\mathbf{B}, \mathbf{W}} \{n^{-1} \|\mathbf{y} - \mathbf{X} \text{vec}(\mathbf{B})\|_2^2 + \lambda \|\mathbf{W}\|_1\}, \quad \text{s.t. } \mathbf{B}' \mathbf{B} = \mathbf{I}, \text{ and } \mathbf{B} = \mathbf{W}.$$

Then the corresponding augmented Lagrangian formulation is

$$\min_{\mathbf{B}, \mathbf{W}} \{n^{-1} \|\mathbf{y} - \mathbf{X} \text{vec}(\mathbf{B})\|_2^2 + \lambda \|\mathbf{W}\|_1 + 2\rho \langle \mathbf{M}, \mathbf{B} - \mathbf{W} \rangle + \rho \|\mathbf{B} - \mathbf{W}\|_F^2\}, \quad (11)$$

where \mathbf{M} is the dual variable and ρ is a regularization parameter. Algorithm 3 presents the proposed ADMM subroutine for solving problem (11).

In Algorithm 3, the ℓ_1 regularization and the orthogonality constraint are separated into two simple updates. Note that both the \mathbf{B} -update step in Algorithm 3 and the \mathcal{G} -update step in Algorithm 2 are least squares problems with orthogonal constraints, which can be solved by the **splitting orthogonality constraint method** (Lai and Osher, 2014). To update \mathbf{W} , we apply the explicit soft-thresholding for the ℓ_1 -regularized optimization.

Algorithm 2 requires predetermination of the multilinear ranks (r_1, r_2, r_3) and tuning parameter λ . Here we again employ the BIC for their selection, as standard cross-

Algorithm 3 ADMM subroutine for sparse and orthogonal regression

Initialize: $\mathbf{B}^{(0)} = \mathbf{W}^{(0)}$, $\mathbf{M}^{(0)} = \mathbf{0}$

repeat $k = 0, 1, 2, \dots$

$$\mathbf{B}^{(k+1)} \leftarrow \arg \min_{\mathbf{B}'\mathbf{B}=\mathbf{I}} \left\{ n^{-1} \|\mathbf{y} - \mathbf{X} \text{vec}(\mathbf{B})\|_2^2 + \rho \|\mathbf{B} - \mathbf{W}^{(k)} + \mathbf{M}^{(k)}\|_F^2 \right\}$$

$$\mathbf{W}^{(k+1)} \leftarrow \arg \min \left\{ \rho \|\mathbf{B}^{(k+1)} - \mathbf{W} + \mathbf{M}^{(k)}\|_F^2 + \lambda \|\mathbf{W}\|_1 \right\}$$

$$\mathbf{M}^{(k+1)} \leftarrow \mathbf{M}^{(k)} + \mathbf{B}^{(k+1)} - \mathbf{W}^{(k+1)}$$

until convergence

validation techniques cannot be applied for time series data. Although the degrees of freedom in a sparse and orthogonal matrix are unclear, we may use the total number of nonzero elements in \mathcal{G} , \mathbf{U}_1 , \mathbf{U}_2 and \mathbf{U}_3 as the pseudo degrees of freedom. Moreover, the method in Remark 4 can be used to narrow down the range of plausible ranks, where the reduced-rank regression via adaptive nuclear norm penalization (Chen et al., 2013) can be used to obtain an initial estimator of \mathcal{A} .

5 Numerical studies

5.1 Simulation experiments

We conduct three simulation experiments to evaluate (1) the performance of the MLR estimator $\hat{\mathcal{A}}_{\text{MLR}}$ in Section 3, (2) that of the SHORR estimator $\hat{\mathcal{A}}_{\text{SHORR}}$ in Section 4, and (3) the robustness of these estimators against VARMA misspecifications. Throughout the experiments, we let the innovations $\{\epsilon_t\}$ be *i.i.d.* Gaussian random vectors with mean zero and covariance matrix Σ_ρ , where Σ_ρ has diagonal entries equal to one and off-diagonal entries equal to $\rho = 0, 0.3$ or 0.5 , corresponding to zero, small and moderate error associations, respectively.

In the first two experiments, we compare the estimation performance of $\hat{\mathcal{A}}_{\text{MLR}}$ and $\hat{\mathcal{A}}_{\text{SHORR}}$ with existing commonly used estimators for large VAR models. The data are generated by the multilinear low-rank VAR model in (3) with $r_1 = r_2 = 3$. The core tensor \mathcal{G} is generated by scaling an $r_1 \times r_2 \times r_3$ tensor of independent standard normal random numbers such that $\sigma_1(\mathcal{G}_{(1)}) = 0.9$. The orthonormal matrices \mathbf{U}_i s are generated

randomly by methods detailed in Appendix A.4. For a fair comparison, both \mathcal{G} and \mathbf{U}_i s are generated repeatedly for different replications.

The first experiment focuses on the comparison of the proposed estimator $\hat{\mathcal{A}}_{\text{MLR}}$ for the low-dimensional case with the OLS estimator $\hat{\mathcal{A}}_{\text{OLS}}$ and the RRR estimator $\hat{\mathcal{A}}_{\text{RRR}}$; see Section 3.1 for details. We set $r_3 = 1, 2, \dots, 5$ and $(N, P, T) = (10, 5, 1000)$. The \mathbf{U}_i s are non-sparse orthonormal matrices. For both $\hat{\mathcal{A}}_{\text{MLR}}$ and $\hat{\mathcal{A}}_{\text{RRR}}$, we assume that the true ranks r_i s are known. Table 1 presents the average estimation error in Frobenius norm based on 400 replications. Clearly the proposed estimator $\hat{\mathcal{A}}_{\text{MLR}}$ has much smaller estimation errors than the other two estimators, while $\hat{\mathcal{A}}_{\text{RRR}}$ outperforms $\hat{\mathcal{A}}_{\text{OLS}}$. This confirms the theoretical results in Theorem 2 and can be intuitively explained by the capability of $\hat{\mathcal{A}}_{\text{MLR}}$ to exploit low-rank structures in all three directions.

The second experiment is conducted to compare the performance of $\hat{\mathcal{A}}_{\text{SHORR}}$ with that of the following five existing estimators for the high-dimensional case: (i) the LASSO estimator in Tibshirani (1996); (ii) the row-sparse reduced-rank regression (RSRRR) estimator in Chen and Huang (2012); (iii) the regression with a sparse SVD (RSSVD) estimator in Chen et al. (2012); (iv) the reduced-rank regression via adaptive nuclear norm penalization (RRANN) estimator in Chen et al. (2013); and (v) the sparse and orthogonal factor regression (SOFAR) in Uematsu et al. (2019). Note that all these existing methods are based on model (2) and consequently are able to exploit the low-rankness in only one direction.

For this experiment, we fix $r_3 = 3$ and consider two cases with relatively large (N, P) and small T : $(N, P, T) = (10, 5, 400)$ and $(20, 8, 400)$. The \mathbf{U}_i s are sparse orthonormal matrices. The ranks and tuning parameters are selected based on the BIC proposed in Section 4.3. Table 2 presents the average estimation error in Frobenius norm based on 400 replications. It can be seen that in both cases $\hat{\mathcal{A}}_{\text{SHORR}}$ significantly outperforms the other estimators, as the latter takes into account only sparsity and/or low-rankness in one direction. Moreover, in both Tables 1 and 2, the estimation error generally increases as ρ becomes larger, which is due to the lack of ability of all methods to take into account the possible correlation structure among the N elements of the innovation ϵ_t .

The third experiment aims to examine the robustness of the proposed estimators to

VARMA misspecifications. The data are generated by the VARMA(1, 1) process in (5) with $\text{rank}(\Phi) = 3$ and $\text{rank}(\Theta) = 1$. The matrices Θ and Φ are generated randomly as follows: $\Theta = 0.95\mathbf{u}\mathbf{v}'$, where \mathbf{u} and \mathbf{v} are generated by scaling standard normal random vectors such that $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$; and $\Phi = \mathbf{U}\text{diag}(0.8, 0.6, 0.4)\mathbf{V}'$, where \mathbf{U} and \mathbf{V} are $N \times 3$ orthonormal matrices generated by a method in Appendix A.4. As in the first two experiments, both Φ and Ψ are generated repeatedly for different replications.

For this experiment, we consider the following two cases: $(N, T) = (10, 1000)$ and $(20, 500)$. The estimator $\hat{\mathcal{A}}_{\text{MLR}}$ is examined in the former (low-dimensional) case and $\hat{\mathcal{A}}_{\text{SHORR}}$ in the latter (high-dimensional) case. Comparisons with existing estimators in low and high dimensions are made as in the first and second experiments, respectively. Note that by Proposition 2, the transition tensor of the VAR(∞) representation of the VARMA(1, 1) process has multilinear low ranks (4, 4, 2). Thus we fit VAR(4) models using different estimation methods. The estimation error is calculated as the difference between the estimated VAR coefficients and the true coefficients in the truncated VAR(4) representation. The average estimation error in Frobenius norm based on 500 replications is displayed in Table 3 for the aforementioned two cases.

Interestingly, the clear advantage of $\hat{\mathcal{A}}_{\text{MLR}}$ and $\hat{\mathcal{A}}_{\text{SHORR}}$ over existing estimators is preserved even when the model is misspecified. According to Section 2.3, the transition tensor of the truncated VAR(4) representation of the generated VARMA process has a multilinear low-rank structure. This directly explains why $\hat{\mathcal{A}}_{\text{MLR}}$ outperforms the competing estimators in the low-dimensional case. For the high-dimensional case, the transition tensor has multilinear low ranks. However, neither the tensor nor its Tucker decomposition is sparse. As a result, the RRANN estimator outperforms all the other estimators except for $\hat{\mathcal{A}}_{\text{SHORR}}$ since the former imposes low-rankness but not sparsity, whereas the other estimators falsely impose sparsity. The estimation performance of $\hat{\mathcal{A}}_{\text{SHORR}}$ remains the best among all estimators, although it also wrongly assumes the sparsity. This indicates that the benefit of exploiting the multilinear low-rank structure substantially outweighs the cost of falsely imposing sparsity on the model.

5.2 Real data analysis

This subsection applies the proposed estimation methods to jointly model 40 quarterly macroeconomic sequences of the United States from 1959 to 2007, with 194 observed values for each variable (Koop, 2013). All series are seasonally adjusted except for financial variables, transformed to stationarity, and standardized to zero mean and unit variance. These variables capture many aspects of the economy, and can be classified into eight categories: (i) GDP and its decomposition, (ii) National Association of Purchasing Managers (NAPM) indices, (iii) industrial production, (iv) housing, (v) money, credit and interest rate, (vi) employment, (vii) prices and wages, and (viii) others. The VAR model has been widely applied to fit these series in empirical econometric studies for structural analysis and forecasting; see Stock and Watson (2009) and Koop (2013). Table 5 gives more details about these macroeconomic variables.

We first apply the SHORR estimation to the entire data set, with the lag order fixed at $P = 4$ for the fitted VAR model as suggested by Koop (2013). Since the number of variables $N = 40$ is much larger than the lag order $P = 4$, we do not perform variable selection for the factor matrix related to lags; that is, we replace $\|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_1$ with $\|\mathbf{U}_2 \otimes \mathbf{U}_1\|_1$ in the penalty term. The ranks and the tuning parameter λ are selected based on the BIC in Section 4.3, which results in $(r_1, r_2, r_3) = (4, 3, 2)$.

The ℓ_1 penalty yields sparse estimated factor matrices $\hat{\mathbf{U}}_1$ and $\hat{\mathbf{U}}_2$, and the estimated coefficients are presented in Figure 2. The factor loading provides insights into the dynamic relationship among the 40 macroeconomic variables. The four response factors, denoted by R_i for $1 \leq i \leq 4$, contain nearly all of the variables and encapsulate different aspects of the economy: R_1 is mostly related to investments, imports, industrial production and employments; R_2 includes personal consumption, housing starts, and labor productivity; R_3 includes manufacturing, housing starts, and treasury bill yield rates; and R_4 includes NAPM indices, housing starts, and price index. Each response factor covers multiple categories of macroeconomic indices, and no clear group structure can be observed. However, it is noteworthy that only twelve variables are selected by the three predictor factors, and the sparse formulations of the predictor factors mainly consist of variables from the first four categories, including real GDP, private investment, NAPM

indices, manufacturing and housing starts. The above result leads to an interesting interpretation: the activeness of production and investment serves as the driving force of the whole economy and usually precedes changes in other economic aspects such as the price indices, financial indices, and labor markets.

We next evaluate the forecasting performance of $\hat{\mathcal{A}}_{\text{MLR}}$ and $\hat{\mathcal{A}}_{\text{SHORR}}$ in comparison with the competing estimators considered in Section 5.1. The following rolling forecasting procedure is adopted: first, use the historical data with the end point rolling from Q4-2000 to Q3-2007 to fit the models; and then, conduct one-step-ahead forecasts based on the fitted models. The selected ranks and tuning parameters for $\hat{\mathcal{A}}_{\text{SHORR}}$ are preserved from the analysis of the entire data set, i.e. $(r_1, r_2, r_3) = (4, 3, 2)$, and the selected ranks for MLR and RRR estimation are also fixed accordingly.

The ℓ_2 and ℓ_∞ norms of the forecast errors for various methods are displayed in Table 4. It can be seen that the proposed MLR and SHORR estimators have much smaller forecast errors than competing ones. This can be explained by the capability of the proposed estimators to substantially reduce the dimensionality and approximate the potential moving average effects. The SHORR estimator performs best among all estimators as it enforces sparsity of the factor matrices and hence prevents overfitting most effectively.

6 Conclusion and discussion

For a large $\text{VAR}(P)$ model, its reduced-rank structure can be defined in three different ways. The novelty of the proposed approach lies in its ability to jointly enforce the three different reduced-rank structures. This is made possible by rearranging the transition matrices of the VAR model into a tensor such that the Tucker decomposition can be conducted. As a result, the parameter space is restricted effectively in three directions, and the capability of the classical VAR model for modeling large-scale time series is substantially expanded.

Moreover, for the high-dimensional setup, this paper further proposes a sparsity-inducing estimator to improve the estimation efficiency and model interpretability. An

ADMM-based algorithm is introduced to separately implement the ℓ_1 regularization and orthogonality constraints. It is worth noting that this paper has a different focus than most work on tensor regression: here we employ the tensor technique as a novel solution to problems in classical time series modeling.

This paper may be extended in three possible directions. Firstly, the proposed estimators do not take into account the possible correlation structure among components of ϵ_t , which will reduce the estimation efficiency; see the simulation results in Section 5.1. Let $\hat{\Sigma}_\epsilon$ be an estimator of Σ_ϵ . As in Davis et al. (2016), we may alternatively consider the generalized least squares loss $\sum_{t=1}^T (\mathbf{y}_t - \mathcal{A}_{(1)} \mathbf{x}_t) \hat{\Sigma}_\epsilon^{-1} (\mathbf{y}_t - \mathcal{A}_{(1)} \mathbf{x}_t)$ rather than $\sum_{t=1}^T \|\mathbf{y}_t - \mathcal{A}_{(1)} \mathbf{x}_t\|_2^2$. However, the difficulty would be to find a good estimator $\hat{\Sigma}_\epsilon$. Secondly, the tensor technique potentially can be applied to many variants of the VAR model, e.g., those with a nonlinear dynamic structure such as the threshold VAR model (Tsay, 1998) and the varying coefficient VAR model (Lütkepohl, 2005). For instance, consider the time-varying coefficient VAR model with lag one, $\mathbf{y}_t = \mathbf{A}_t \mathbf{y}_{t-1} + \epsilon_t$. Similarly, the coefficient matrices can be rearranged into a tensor \mathcal{A} with $\mathcal{A}_{(1)} = (\mathbf{A}_1, \dots, \mathbf{A}_T)$. If \mathcal{A} has multilinear low ranks (r_1, r_2, r_3) , then the number of parameters will be $r_1 r_2 r_3 + (N - r_1) r_1 + (N - r_2) r_2 + (T - r_3) r_3 \lesssim NT$. Moreover, a fourth-order tensor can be used to handle the case of lag order $P > 1$. **Lastly, the proposed model can be generalized to a tensor autoregressive model for matrix-valued or tensor-valued time series; see Wang et al. (2019) for a related work.**

Appendix: technical details

This appendix gives the technical proofs of theorems, propositions and a corollary, five lemmas used in proving Theorem 3, and methods for generating orthonormal matrices.

A.1 Proofs of Theorems 1-3

Proof of Theorem 1. The proof generally follows from Proposition 4.1 in Shapiro (1986) for overparameterized models. The VAR(P) model can be written as the linear regression

problem

$$\underbrace{\begin{bmatrix} \mathbf{y}'_1 \\ \mathbf{y}'_2 \\ \vdots \\ \mathbf{y}'_T \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \mathbf{y}'_0 & \mathbf{y}'_{-1} & \cdots & \mathbf{y}'_{-P+1} \\ \mathbf{y}'_1 & \mathbf{y}'_0 & \cdots & \mathbf{y}'_{-P+2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}'_{T-1} & \mathbf{y}'_{T-2} & \cdots & \mathbf{y}'_{T-P} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \mathbf{A}'_1 \\ \mathbf{A}'_2 \\ \vdots \\ \mathbf{A}'_P \end{bmatrix}}_{\mathbf{A}'_{(1)}} + \underbrace{\begin{bmatrix} \boldsymbol{\epsilon}'_{P+1} \\ \boldsymbol{\epsilon}'_{P+2} \\ \vdots \\ \boldsymbol{\epsilon}'_T \end{bmatrix}}_{\mathbf{E}} \quad (12)$$

Let $\boldsymbol{\phi}$ denote the component parameters in the Tucker decomposition forms, let $\mathbf{h}(\boldsymbol{\phi})$ denote the true parameter $\text{vec}(\mathbf{A}_{(1)}) = \text{vec}(\mathbf{U}_1 \mathbf{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)')$, and let $\hat{\mathbf{h}}_{\text{OLS}}$ denote the vectorized OLS estimates $\text{vec}(\hat{\mathbf{A}}_{\text{OLS}})$ without constraint. With Assumption 1, according to classical asymptotic theory for stationary VAR model (Tsay, 2013), as $T \rightarrow \infty$,

$$\begin{aligned} (i). \quad & \mathbf{X}'\mathbf{X}/T \xrightarrow{p} \boldsymbol{\Gamma}^* \equiv \begin{bmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{P-1} \\ \boldsymbol{\Gamma}'_1 & \boldsymbol{\Gamma}_0 & \cdots & \boldsymbol{\Gamma}_{P-2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}'_{P-1} & \boldsymbol{\Gamma}'_{P-2} & \cdots & \boldsymbol{\Gamma}_0 \end{bmatrix}; \\ (ii). \quad & T^{-1/2} \text{vec}(\mathbf{X}'\mathbf{E}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon \otimes \boldsymbol{\Gamma}^*); \\ (iii). \quad & \hat{\mathbf{h}}_{\text{OLS}} \xrightarrow{p} \mathbf{h}; \\ (iv). \quad & \sqrt{T}(\hat{\mathbf{h}}_{\text{OLS}} - \mathbf{h}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon \otimes (\boldsymbol{\Gamma}^*)^{-1}). \end{aligned}$$

Consider the discrepancy function for any $\mathbf{h}(\boldsymbol{\phi})$,

$$F(\hat{\mathbf{h}}_{\text{OLS}}, \mathbf{h}) = \|\text{vec}(\mathbf{Y}) - (\mathbf{I}_N \otimes \mathbf{X})\mathbf{h}\|_2^2 - \|\text{vec}(\mathbf{Y}) - (\mathbf{I}_N \otimes \mathbf{X})\hat{\mathbf{h}}_{\text{OLS}}\|_2^2.$$

Obviously, $F(\hat{\mathbf{h}}_{\text{OLS}}, \mathbf{h})$ is a nonnegative and twice continuously differentiable function, and equals to zero if and only if $\hat{\mathbf{h}}_{\text{OLS}} = \mathbf{h}$.

In order to calculate the Jacobian matrix \mathbf{H} , we define the tensor matricization transformation operator $\mathbf{T}_{ij}(N, N, P)$ which is an $N^2P \times N^2P$ matrix and satisfies that $\text{vec}(\mathbf{A}_{(j)}) = \mathbf{T}_{ij}(N, N, P)\text{vec}(\mathbf{A}_{(i)})$ for any tensor $\mathbf{A} \in \mathbb{R}^{N \times N \times P}$. In fact, $\mathbf{T}_{ij}(N, N, P)$ is a full-rank matrix indicating the corresponding position in $\text{vec}(\mathbf{A}_{(i)})$ of \mathbf{A} 's each entry in $\text{vec}(\mathbf{A}_{(j)})$, and can be regarded as the natural extension of the permutation matrix for matrix transpose. Also note that $\mathbf{T}_{ij}(N, N, P)$ only depends on the value of N and P , and since we consider fixed N and P in this part, we simplify it to \mathbf{T}_{ij} .

Therefore,

$$\begin{aligned}\text{vec}(\mathcal{A}_{(1)}) &= \text{vec}(\mathbf{U}_1 \mathbf{G}_{(1)} (\mathbf{U}_3 \otimes \mathbf{U}_2)') = \mathbf{T}_{21} \text{vec}(\mathbf{U}_2 \mathbf{G}_{(2)} (\mathbf{U}_1 \otimes \mathbf{U}_3)') \\ &= \mathbf{T}_{31} \text{vec}(\mathbf{U}_3 \mathbf{G}_{(3)} (\mathbf{U}_1 \otimes \mathbf{U}_2)'),\end{aligned}$$

and the Jacobian matrix of \mathbf{h} is

$$\begin{aligned}\mathbf{H} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\phi}} &= \left((\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1), [(\mathbf{U}_3 \otimes \mathbf{U}_2) \mathbf{G}'_{(1)}] \otimes \mathbf{I}_N, \right. \\ &\quad \left. \mathbf{T}_{21} \{[(\mathbf{U}_1 \otimes \mathbf{U}_3) \mathbf{G}'_{(2)}] \otimes \mathbf{I}_N\}, \mathbf{T}_{31} \{[(\mathbf{U}_1 \otimes \mathbf{U}_2) \mathbf{G}'_{(3)}] \otimes \mathbf{I}_P\} \right).\end{aligned}$$

Then, by Proposition 4.1 in Shapiro (1986), we know that the minimizer of $F(\hat{\mathbf{h}}_{\text{OLS}}, \cdot)$, namely the MLR estimator, has the asymptotic normality,

$$\sqrt{T}(\mathbf{h}(\hat{\boldsymbol{\phi}}_{\text{MLR}}) - \mathbf{h}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{MLR}})$$

and $\boldsymbol{\Sigma}_{\text{MLR}} = \mathbf{P} \boldsymbol{\Gamma} \mathbf{P}'$, where $\mathbf{P} = \mathbf{H}(\mathbf{H}' \mathbf{J} \mathbf{H})^\dagger \mathbf{H}' \mathbf{J}$ is the projection matrix, \mathbf{J} is the Fisher information matrix of \mathbf{h} as T goes to infinity, \mathbf{H} is the Jacobian matrix of \mathbf{h} with respect to the overparameterized model parameters $\boldsymbol{\phi}$, $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_\epsilon \otimes (\boldsymbol{\Gamma}^*)^{-1}$ is the asymptotic covariance matrix for $\hat{\mathbf{h}}_{\text{OLS}}$ and \dagger denotes the Moore-Penrose inverse. Since $\boldsymbol{\Gamma} = \mathbf{J}^{-1}$ in the VAR(P) model, we have $\boldsymbol{\Sigma}_{\text{MLR}} = \mathbf{H}(\mathbf{H}' \mathbf{J} \mathbf{H})^\dagger \mathbf{H}'$. \square

Proof of Theorem 2. The \sqrt{T} -consistency and asymptotic normality of $\hat{\mathcal{A}}_{\text{OLS}}$ has been studied in the proof of Theorem 1, with $\boldsymbol{\Sigma}_{\text{OLS}} = \mathbf{J}^{-1}$.

As discussed previously, $\boldsymbol{\Sigma}_{\text{MLR}} = \mathbf{P} \mathbf{J}^{-1} \mathbf{P}'$ where \mathbf{P} is a projection matrix. Note that $\mathbf{J}^{-1} - \mathbf{H}(\mathbf{H}' \mathbf{J} \mathbf{H})^\dagger \mathbf{H}' = \mathbf{J}^{-1/2} \mathbf{Q}_{\mathbf{J}^{1/2} \mathbf{H}} \mathbf{J}^{-1/2}$, where $\mathbf{Q}_{\mathbf{J}^{1/2} \mathbf{H}}$ is the projection matrix onto the orthogonal complement of $\text{span}(\mathbf{J}^{1/2} \mathbf{H})$. Then, it is clear that $\mathbf{J}^{-1} \geq \mathbf{H}(\mathbf{H}' \mathbf{J} \mathbf{H})^\dagger \mathbf{H}'$.

For the RRR estimator, the components in the SVD, $\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}'$, can be denoted as $\boldsymbol{\theta} = (\text{vec}(\mathbf{U})', \text{diag}(\mathbf{D})', \text{vec}(\mathbf{V})')'$. Therefore, the gradient matrix of the RRR is $\mathbf{R} = \partial \mathbf{h} / \partial \boldsymbol{\theta}$. Since \mathbf{U}_1 in Tucker decomposition is exactly the same as the left singular vectors in the SVD of $\mathcal{A}_{(1)}$, we can view the Tucker decomposition as a further decomposition of the matrix $\mathbf{D} \mathbf{V}'$. Therefore, $\mathbf{H} = \partial \mathbf{h} / \partial \boldsymbol{\phi} = \partial \mathbf{h} / \partial \boldsymbol{\theta} \cdot \partial \boldsymbol{\theta} / \partial \boldsymbol{\phi} = \mathbf{R} \cdot \partial \boldsymbol{\theta} / \partial \boldsymbol{\phi}$. By similar arguments in the proof of Theorem 1, we can obtain that the RRR estimator has the asymptotic covariance $\boldsymbol{\Sigma}_{\text{RRR}} = \mathbf{R}(\mathbf{R}' \mathbf{J} \mathbf{R})^\dagger \mathbf{R}'$ and it is smaller than or equal to $\boldsymbol{\Sigma}_{\text{MLR}}$ since $\text{span}(\mathbf{J}^{1/2} \mathbf{R}) \subset \text{span}(\mathbf{J}^{1/2} \mathbf{H})$. \square

Proof of Theorem 3. The proof of Theorem 3 consists of two parts.

- The first part shows the estimation error bounds given the deterministic realization of the time series process, assuming that the deviation bound condition and restricted eigenvalue condition hold.
- The second step is the stochastic analysis in which we show that these two regulatory conditions are satisfied with high probability converging to 1.

Based on the linear regression form (12), we can rewrite $\mathbf{Y} = \mathbf{X}(\mathbf{U}_3 \otimes \mathbf{U}_2) \mathbf{G}'_{(1)} \mathbf{U}'_1 + \mathbf{E}$ as

$$\underbrace{\text{vec}(\mathbf{Y})}_{\mathbf{y}} = \underbrace{(\mathbf{I}_N \otimes \mathbf{X})}_{\mathbf{Z}} \underbrace{(\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)}_{\mathbf{U}} \underbrace{\text{vec}(\mathbf{G}'_{(1)})}_{\mathbf{g}} + \underbrace{\text{vec}(\mathbf{E})}_{\mathbf{e}}.$$

Denote $\hat{\Delta} = \hat{\mathbf{U}}\hat{\mathbf{g}} - \mathbf{U}\mathbf{g}$, $\hat{\Delta}_u = \hat{\mathbf{U}} - \mathbf{U}$ and $\hat{\Delta}_g = \hat{\mathbf{g}} - \mathbf{g}$. By the optimality of the SHORR estimator,

$$\begin{aligned} T^{-1} \|\mathbf{y} - \mathbf{Z}\hat{\mathbf{U}}\hat{\mathbf{g}}\|_2^2 + \lambda \|\hat{\mathbf{U}}\|_1 &\leq T^{-1} \|\mathbf{y} - \mathbf{Z}\mathbf{U}\mathbf{g}\|_2^2 + \lambda \|\mathbf{U}\|_1 \\ \Rightarrow T^{-1} \|\mathbf{Z}\hat{\Delta}\|_2^2 + \lambda \|\hat{\mathbf{U}}\|_1 &\leq 2\langle T^{-1} \mathbf{Z}'\mathbf{e}, \hat{\Delta} \rangle + \lambda \|\mathbf{U}\|_1. \end{aligned} \quad (13)$$

Note that $\hat{\Delta} = \hat{\mathbf{U}}\hat{\mathbf{g}} - \mathbf{U}\mathbf{g} = (\hat{\mathbf{U}} - \mathbf{U})\hat{\mathbf{g}} + \mathbf{U}(\hat{\mathbf{g}} - \mathbf{g}) = \hat{\Delta}_u\hat{\mathbf{g}} + \mathbf{U}\hat{\Delta}_g$, so we can decompose $\langle T^{-1} \mathbf{Z}'\mathbf{e}, \hat{\Delta} \rangle$ into two parts,

$$\langle T^{-1} \mathbf{Z}'\mathbf{e}, \hat{\Delta} \rangle = \langle T^{-1} \mathbf{Z}'\mathbf{e}, \hat{\Delta}_u\hat{\mathbf{g}} \rangle + \langle T^{-1} \mathbf{Z}'\mathbf{e}, \mathbf{U}\hat{\Delta}_g \rangle.$$

and bound these two parts separately. We denote the event that these two inner products are bounded by $\lambda \|\hat{\Delta}_u\|_1$ and $\lambda \|\hat{\Delta}_g\|_1$ as \mathcal{I}_1 ,

$$\mathcal{I}_1 = \{ \langle T^{-1} \mathbf{Z}'\mathbf{e}, \hat{\Delta}_u\hat{\mathbf{g}} \rangle \leq \lambda \|\hat{\Delta}_u\|_1/4 \} \cap \{ \langle T^{-1} \mathbf{Z}'\mathbf{e}, \mathbf{U}\hat{\Delta}_g \rangle \leq \lambda \|\hat{\Delta}_g\|_1/4 \}.$$

Denote by $\mathbb{S}_{\mathbf{U}}$ the nonzero index set of $\text{vec}(\mathbf{U})$, and by $\mathbb{S}_{\mathbf{U}}^c$ is the complement of $\mathbb{S}_{\mathbf{U}}$. By the sparsity of each \mathbf{U}_i in Assumption 4, $\text{card}(\mathbb{S}_{\mathbf{U}}) = \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_0 \leq s_1 s_2 s_3$. In the following proof, we use the abused notation. For any matrix $\mathbf{M} \in \mathbb{R}^{N^2 P \times r_1 r_2 r_3}$ and any vector norm $\|\cdot\|_*$, we denote $\|(\mathbf{M})_{\mathbb{S}_{\mathbf{U}}}\|_* := \|(\text{vec}(\mathbf{M}))_{\mathbb{S}_{\mathbf{U}}}\|_*$ and $\|(\mathbf{M})_{\mathbb{S}_{\mathbf{U}}^c}\|_* := \|(\text{vec}(\mathbf{M}))_{\mathbb{S}_{\mathbf{U}}^c}\|_*$.

On the event \mathcal{I}_1 , if we multiply 2 to both sides of (13) we can have

$$2T^{-1} \|\mathbf{Z}\hat{\Delta}\|_2^2 + 2\lambda \|\hat{\mathbf{U}}\|_1 \leq \lambda \|\hat{\Delta}_g\|_1 + \lambda \|\hat{\Delta}_u\|_1 + 2\lambda \|\mathbf{U}\|_1$$

On the left-hand side, by triangle inequality,

$$\|\widehat{\mathbf{U}}\|_1 = \|\widehat{\mathbf{U}}_{\mathbb{S}_U}\|_1 + \|\widehat{\mathbf{U}}_{\mathbb{S}_U^c}\|_1 \geq \|\mathbf{U}_{\mathbb{S}_U}\|_1 - \|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U}\|_1 + \|\widehat{\mathbf{U}}_{\mathbb{S}_U^c}\|_1,$$

whereas on the right-hand side, $\|\widehat{\Delta}_{\mathbf{u}}\|_1 = \|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U}\|_1 + \|\widehat{\mathbf{U}}_{\mathbb{S}_U^c}\|_1$. So we have

$$2T^{-1}\|\mathbf{Z}\widehat{\Delta}\|_2^2 + \lambda\|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U^c}\|_1 \leq \lambda\|\widehat{\Delta}_{\mathbf{g}}\|_1 + 3\lambda\|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U}\|_1$$

Next, we assume that there is a lower bound for $2T^{-1}\|\mathbf{Z}\widehat{\Delta}\|_2^2$. Thus, we define the event $\mathcal{I}_2 = \{2T^{-1}\|\mathbf{Z}\widehat{\Delta}\|_2^2 \geq \alpha\|\widehat{\Delta}\|_2^2\}$, where $\alpha = \lambda_{\min}(\Sigma_{\epsilon})/\mu_{\max}(\mathcal{A})$. On the event \mathcal{I}_2 ,

$$\begin{aligned} \alpha\|\widehat{\Delta}\|_2^2 &\leq 2T^{-1}\|\mathbf{Z}\widehat{\Delta}\|_2^2 \leq \lambda\|\widehat{\Delta}_{\mathbf{g}}\|_1 + 3\lambda\|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U}\|_1 \\ &\leq \lambda\sqrt{r_1 r_2 r_3}\|\widehat{\Delta}_{\mathbf{g}}\|_2 + 3\lambda\sqrt{s_1 s_2 s_3}\|(\widehat{\Delta}_{\mathbf{u}})_{\mathbb{S}_U}\|_2 \\ &\leq \lambda\sqrt{r_1 r_2 r_3}\|\widehat{\Delta}_{\mathbf{g}}\|_2 + 3\lambda\sqrt{s_1 s_2 s_3}\|\widehat{\Delta}_{\mathbf{u}}\|_{\text{F}}. \end{aligned}$$

By the perturbation bound for HOSVD in Lemma 1, we have

$$\begin{aligned} \|\widehat{\Delta}_{\mathbf{u}}\|_{\text{F}} &= \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \widehat{\mathbf{U}}_1\|_{\text{F}} \\ &\leq \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \widehat{\mathbf{U}}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_{\text{F}} + \|\widehat{\mathbf{U}}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \mathbf{U}_1\|_{\text{F}} \\ &\quad + \|\widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \mathbf{U}_1 - \widehat{\mathbf{U}}_3 \otimes \widehat{\mathbf{U}}_2 \otimes \widehat{\mathbf{U}}_1\|_{\text{F}} \\ &\leq \sqrt{r_1 r_2}\|\mathbf{U}_3 - \widehat{\mathbf{U}}_3\|_{\text{F}} + \sqrt{r_1 r_3}\|\mathbf{U}_2 - \widehat{\mathbf{U}}_2\|_{\text{F}} + \sqrt{r_2 r_3}\|\mathbf{U}_1 - \widehat{\mathbf{U}}_1\|_{\text{F}} \\ &\leq c\tau\|\widehat{\Delta}\|_2, \end{aligned}$$

where $\tau = \delta^{-1}(\eta_1\sqrt{r_2 r_3} + \eta_2\sqrt{r_1 r_3} + \eta_3\sqrt{r_1 r_2})$, and

$$\|\widehat{\Delta}_{\mathbf{g}}\|_2 \leq C\bar{g}\delta^{-1}r_1 r_2 r_3 \sum_{i=1}^3 (\eta_i/\sqrt{r_i})\|\widehat{\Delta}\|_2 = C\bar{g}\sqrt{r_1 r_2 r_3}\tau\|\widehat{\Delta}\|_2.$$

Therefore, we have

$$\alpha\|\widehat{\Delta}\|_2^2 \leq C\tau(\sqrt{s_1 s_2 s_3} + r_1 r_2 r_3)\lambda\|\widehat{\Delta}\|_2.$$

If we denote $S = (\sqrt{s_1 s_2 s_3} + r_1 r_2 r_3)^2$, we can obtain the estimation error bound and in-sample prediction error bound

$$\|\widehat{\Delta}\|_2 \leq C_1\tau\sqrt{S}\lambda/\alpha, \quad \text{and} \quad T^{-1}\|\mathbf{Z}\widehat{\Delta}\|_2^2 \leq C_2\tau^2 S\lambda^2/\alpha,$$

which conclude the deterministic analysis.

In the second part, we show that the events \mathcal{I}_1 and \mathcal{I}_2 occur with high probability. In the high-dimensional regression literature, the conditions in \mathcal{I}_1 and \mathcal{I}_2 are known as deviation bound condition and restricted eigenvalue condition. We defer the proof of both conditions to Lemma 2 and 3, where we show that \mathcal{I}_1 and \mathcal{I}_2 hold simultaneously with probability at least $1 - C \exp[-c \log(N^2 P)] - C \exp\{-cd \min[\log(NP), \log(cNP/d)]\}$, given that the sample size $T \gtrsim \log(N^2 P) + \mathcal{M}^2 d \min[\log(NP), \log(cNP/d)]$. \square

A.2 Proofs of Propositions 1-3 and Corollary 1

Proof of Proposition 1. We denote by \mathcal{A} the transition tensor of the VAR representation. By definition of multilinear ranks, we need to obtain the $\text{rank}_n(\mathcal{A})$, $n = 1, 2, 3$.

Note that $\text{rank}_1(\mathcal{A}) = \text{rank}(\mathcal{A}_{(1)}) = \text{rank}([\Theta, \Theta^2, \Theta^3, \dots])$. Denote by $\mathcal{C}(\mathbf{M})$ the column space of a matrix \mathbf{M} . Since $\mathcal{C}(\Theta^i) \subset \mathcal{C}(\Theta)$, for $i = 1, 2, 3, \dots$, it is obvious that $\mathcal{C}([\Theta, \Theta^2, \Theta^3, \dots]) \subset \mathcal{C}(\Theta)$. Therefore, we have $\text{rank}_1(\mathcal{A}) \leq r$.

Similarly, $\text{rank}_2(\mathcal{A}) = \text{rank}(\mathcal{A}_{(2)}) = \text{rank}([\Theta', (\Theta^2)'], (\Theta^3)', \dots]) \leq r$.

In addition, $\text{rank}_3(\mathcal{A}) = \text{rank}(\mathcal{A}_{(3)}) = \text{rank}([\text{vec}(\Theta), \text{vec}(\Theta^2), \text{vec}(\Theta^3), \dots])$. For the $N \times N$ rank r matrix Θ , denote its Jordan Canonical Form as

$$\Theta = BJB^{-1} = B \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_k \\ & & & & \mathbf{O} \end{bmatrix} B^{-1},$$

where \mathbf{J}_i is a $n_i \times n_i$ matrix, which can take value of

$$\mathbf{J}_i = [\lambda_i], \text{ or } \begin{bmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{bmatrix} \text{ or } \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix} \text{ or } \dots,$$

λ_i is the eigenvalue, and $\sum_{i=1}^k n_i = r$.

Then, $\Theta^m = \mathbf{B}\mathbf{J}^m\mathbf{B}^{-1}$, where

$$\mathbf{J}^m = \begin{bmatrix} \mathbf{J}_1^m & & & \\ & \mathbf{J}_2^m & & \\ & & \ddots & \\ & & & \mathbf{J}_k^m \\ & & & & \mathbf{O} \end{bmatrix}, \quad \mathbf{J}_i^m = \begin{bmatrix} f(\lambda_i) & f^{(1)}(\lambda_i) & \dots & f^{(n_i-1)}(\lambda_i)/(n_i-1)! \\ 0 & f(\lambda_i) & \dots & f^{(n_i-2)}(\lambda_i)/(n_i-2)! \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f(\lambda_i) \end{bmatrix},$$

and $f(x) = x^m$ with its s -th derivative denoted by $f^{(s)}(x)$. Therefore, we denote $\mathbf{B} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N]$ and $\mathbf{B}^{-1} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K]'$, and

$$\begin{aligned} \Theta^m = \sum_{i=1}^k \left[f(\lambda_i) \sum_{s=c_{i-1}+1}^{c_{i-1}+n_i} \mathbf{u}_s \mathbf{v}'_s + f^{(1)}(\lambda_i) \sum_{s=c_{i-1}+1}^{c_{i-1}+n_i-1} \mathbf{u}_s \mathbf{v}'_{s+1} + \dots \right. \\ \left. + \frac{f^{(n_i-1)}(\lambda_i)}{(n_i-1)!} \mathbf{u}_{c_{i-1}+1} \mathbf{v}'_{c_{i-1}+n_i} \right], \end{aligned} \quad (14)$$

where $c_i = \sum_{j=1}^i n_j$.

In other words, for any $m = 1, 2, \dots$,

$$\text{vec}(\Theta^m) \in \text{span} \left(\bigcup_{i=1}^k \left\{ \text{vec} \left(\sum_{s=c_{i-1}+1}^{c_{i-1}+n_i} \mathbf{u}_s \mathbf{v}'_s \right), \dots, \text{vec}(\mathbf{u}_{c_i+1} \mathbf{v}'_{c_{i-1}+n_i}) \right\} \right),$$

which implies that

$$\begin{aligned} & \text{span}(\cup_{m=1}^{\infty} \{\text{vec}(\Theta^m)\}) \\ & \subset \text{span} \left(\bigcup_{i=1}^k \left\{ \text{vec} \left(\sum_{s=c_{i-1}+1}^{c_{i-1}+n_i} \mathbf{u}_s \mathbf{v}'_s \right), \dots, \text{vec}(\mathbf{u}_{c_i+1} \mathbf{v}'_{c_{i-1}+n_i}) \right\} \right), \end{aligned}$$

and thus,

$$\text{rank}_3(\mathcal{A}) = \text{rank}([\text{vec}(\Theta), \text{vec}(\Theta^2), \text{vec}(\Theta^3), \dots]) \leq \sum_{i=1}^k n_i = r.$$

□

Proof of Proposition 2. Similar to Proposition 1, one needs to obtain the n th-mode rank of \mathcal{A} , $n = 1, 2, 3$. One can easily check that $\text{rank}_1(\mathcal{A}) \leq r + s$ and $\text{rank}_2(\mathcal{A}) \leq r + s$.

For the 3rd-mode rank, $\text{rank}_3(\mathcal{A}) = \text{rank}([\text{vec}(\Theta - \Psi), \text{vec}(\Theta(\Theta - \Psi)), \text{vec}(\Theta^2(\Theta - \Psi)), \dots])$. Note that Θ^m has the Jordan Canonical Form representation (14); therefore,

let $\mathbf{M} = \mathbf{\Theta} - \mathbf{\Psi}$ and we have

$$\text{vec}(\mathbf{\Theta}^m \mathbf{M}) \in \text{span} \left(\bigcup_{i=1}^k \left\{ \text{vec} \left(\sum_{s=c_{i-1}+1}^{c_{i-1}+n_i} \mathbf{u}_s \mathbf{v}'_s \mathbf{M} \right), \dots, \text{vec} \left(\mathbf{u}_{c_i+1} \mathbf{v}'_{c_i-1+n_i} \mathbf{M} \right) \right\} \right).$$

Then, it implies that

$$\begin{aligned} & \text{span}(\bigcup_{m=1}^{\infty} \{\text{vec}(\mathbf{\Theta}^m \mathbf{M})\}) \\ & \subset \text{span} \left(\bigcup_{i=1}^k \left\{ \text{vec} \left(\sum_{s=c_{i-1}+1}^{c_{i-1}+n_i} \mathbf{u}_s \mathbf{v}'_s \mathbf{M} \right), \dots, \text{vec} \left(\mathbf{u}_{c_i+1} \mathbf{v}'_{c_i-1+n_i} \mathbf{M} \right) \right\} \right) \end{aligned}$$

and

$$\text{rank}_3(\mathcal{A}) = \text{rank}([\text{vec}(\mathbf{M}), \text{vec}(\mathbf{\Theta} \mathbf{M}), \text{vec}(\mathbf{\Theta}^2 \mathbf{M}), \text{vec}(\mathbf{\Theta}^3 \mathbf{M}), \dots]) \leq 1 + \sum_{i=1}^k n_i = r + 1.$$

□

Proof of Proposition 3. Proof of global convergence hinges on standard arguments for block relaxation algorithm (Lange, 2010). Under the identifiability constraint, the objective function is nonconvex and continuous. However, each updating step is well-defined, differentiable and convex. Since the algorithm decreases the objective function monotonically, the convergence is guaranteed and any convergent point is a stationary point. Denote the objective function as $L(\phi)$, and then global convergence is guaranteed under the following conditions: (i) L is coercive; (ii) the stationary points of L are isolated; (iii) the algorithm mapping is continuous; (iv) ϕ is a fixed point of the algorithm if and only if it is a stationary point of L ; (v) $L(\phi^{(t+1)}) \leq L(\phi^{(t)})$ with equality if and only if $\phi^{(t)}$ is a fixed point of the algorithm.

Condition (i) is guaranteed by the compactness of the set $\{\phi : L(\phi) \leq L(\phi^{(0)})\}$. Condition (ii) is assumed. Condition (iii) follows from the implicit function theorem since the algorithmic map M is a composition of four differentiable and convex maps. A fixed point ϕ satisfies that $\nabla_{\mathbf{g}} L(\phi) = 0$ and $\nabla_{\mathbf{U}_i} L(\phi) = 0$; therefore the fixed point of the mapping $M(\phi)$, i.e., condition (iv) is satisfied. Finally, each step monotonically decreases $L(\phi)$, so they give a strict decrease if and only if they actually change the corresponding components. Hence, condition (v) is satisfied.

Proof of local convergence hinges on the Ostrowski's theorem, which states that the sequence $\boldsymbol{\phi}^{(t+1)} = M(\boldsymbol{\phi}^{(t)})$ is locally attracted to $\boldsymbol{\phi}^{(\infty)}$ if the spectral radius of the differential of the algorithmic map $\rho[dM(\boldsymbol{\phi}^{(\infty)})] < 1$. The condition can be shown to be true based on the local convergence of block relaxation algorithm by Lange (2010), and we omit the detailed proof here. \square

Proof of Corollary 1. Here we prove the asymptotic normality for $\text{vec}(\widehat{\boldsymbol{U}}_1)$, since the proofs for the $\text{vec}(\widehat{\boldsymbol{U}}_2)$ and $\text{vec}(\widehat{\boldsymbol{U}}_3)$ are similar. In this part, we simplify $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}}$ to $\widehat{\boldsymbol{\mathcal{A}}}$. Note that $\widehat{\boldsymbol{U}}_1$ and \boldsymbol{U}_1 are the eigenvectors of $\widehat{\boldsymbol{\mathcal{A}}}_{(1)}\widehat{\boldsymbol{\mathcal{A}}}'_{(1)}$ and $\boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{\mathcal{A}}'_{(1)}$ respectively. By Theorem 1, $\sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma}_h)$. Note that

$$\begin{aligned} & \sqrt{T}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)}\widehat{\boldsymbol{\mathcal{A}}}'_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{\mathcal{A}}'_{(1)}) \\ &= \sqrt{T}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)})\boldsymbol{\mathcal{A}}'_{(1)} + \sqrt{T}\boldsymbol{\mathcal{A}}_{(1)}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)})' + \sqrt{T}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)})(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)})', \end{aligned}$$

so we have

$$\begin{aligned} & \sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)}\widehat{\boldsymbol{\mathcal{A}}}'_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{\mathcal{A}}'_{(1)}) \\ &= (\boldsymbol{\mathcal{A}}_{(1)} \otimes \boldsymbol{I}_N) \sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}) + (\boldsymbol{I}_N \otimes \boldsymbol{\mathcal{A}}_{(1)}) \sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}) + O_p(T^{-1/2}). \end{aligned}$$

Therefore, $\sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)}\widehat{\boldsymbol{\mathcal{A}}}'_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{\mathcal{A}}'_{(1)})$ is asymptotically normally distributed.

By the matrix perturbation expansion (Izenman, 1975; Velu and Reinsel, 2013),

$$\sqrt{T}(\widehat{\boldsymbol{U}}_{1k} - \boldsymbol{U}_{1k}) = \sum_{i \neq k} \frac{1}{d_k^2 - d_i^2} (\boldsymbol{U}'_{1i} \otimes \boldsymbol{U}_{1i} \boldsymbol{U}'_{1i}) \sqrt{T}\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{(1)}\widehat{\boldsymbol{\mathcal{A}}}'_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}\boldsymbol{\mathcal{A}}'_{(1)}) + O_p(T^{-1/2}).$$

Therefore, $\sqrt{T}(\widehat{\boldsymbol{U}}_1 - \boldsymbol{U}_1)$ is also asymptotically normally distributed.

For $\text{vec}(\widehat{\boldsymbol{\mathcal{G}}}_{(1)})$, by the definition of HOSVD,

$$\widehat{\boldsymbol{\mathcal{G}}} = [\widehat{\boldsymbol{\mathcal{A}}}; \widehat{\boldsymbol{U}}'_1, \widehat{\boldsymbol{U}}'_2, \widehat{\boldsymbol{U}}'_3], \quad \text{and} \quad \boldsymbol{\mathcal{G}} = [\boldsymbol{\mathcal{A}}; \boldsymbol{U}'_1, \boldsymbol{U}'_2, \boldsymbol{U}'_3].$$

So we have

$$\begin{aligned}
& \text{vec}(\widehat{\mathbf{G}}_{(1)} - \mathbf{G}_{(1)}) \\
&= (\widehat{\mathbf{U}}'_3 \otimes \widehat{\mathbf{U}}'_2 \otimes \widehat{\mathbf{U}}'_1) \text{vec}(\widehat{\mathbf{A}}_{(1)}) - (\mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1) \text{vec}(\mathbf{A}_{(1)}) \\
&= (\widehat{\mathbf{U}}'_3 \otimes \widehat{\mathbf{U}}'_2 \otimes \widehat{\mathbf{U}}'_1 - \mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1) \text{vec}(\widehat{\mathbf{A}}_{(1)}) \\
&+ (\mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1) \text{vec}(\widehat{\mathbf{A}}_{(1)} - \mathbf{A}_{(1)}) \\
&= [(\widehat{\mathbf{U}}'_3 - \mathbf{U}'_3) \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1] \text{vec}(\mathbf{A}_{(1)}) + [\mathbf{U}'_3 \otimes (\widehat{\mathbf{U}}'_2 - \mathbf{U}'_2) \otimes \mathbf{U}'_1] \text{vec}(\mathbf{A}_{(1)}) \\
&+ [\mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes (\widehat{\mathbf{U}}'_1 - \mathbf{U}'_1)] \text{vec}(\mathbf{A}_{(1)}) + (\mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1) \text{vec}(\widehat{\mathbf{A}}_{(1)} - \mathbf{A}_{(1)}) + o_p(T^{-1/2}) \\
&= [\mathbf{I}_{r_3} \otimes ((\mathbf{U}'_2 \otimes \mathbf{U}'_1) \mathbf{A}'_{(3)})] \text{vec}(\widehat{\mathbf{U}}_3 - \mathbf{U}_3) + [\mathbf{I}_{r_2} \otimes ((\mathbf{U}'_3 \otimes \mathbf{U}'_1) \mathbf{A}'_{(2)})] \text{vec}(\widehat{\mathbf{U}}_2 - \mathbf{U}_2) \\
&+ [\mathbf{I}_{r_1} \otimes ((\mathbf{U}'_3 \otimes \mathbf{U}'_2) \mathbf{A}'_{(1)})] \text{vec}(\widehat{\mathbf{U}}_1 - \mathbf{U}_1) + (\mathbf{U}'_3 \otimes \mathbf{U}'_2 \otimes \mathbf{U}'_1) \text{vec}(\widehat{\mathbf{A}}_{(1)} - \mathbf{A}_{(1)}) + o_p(T^{-1/2}).
\end{aligned}$$

Therefore, $\sqrt{T} \text{vec}(\widehat{\mathbf{G}}_{(1)} - \mathbf{G}_{(1)})$ is also normally distributed with mean zero, as $T \rightarrow \infty$. For simplicity, we omit the covariance of each component, but they can be easily calculated by the above formula. \square

A.3 Five lemmas used for the proof of Theorem 3

Lemma 1. (*HOSVD perturbation bound*) Under Assumption 6, suppose that $\mathbf{A} = [\mathbf{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]$ and $\widetilde{\mathbf{A}} = [\widetilde{\mathbf{G}}; \widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3]$ are two HOSVD for \mathbf{A} and $\widetilde{\mathbf{A}}$, with the same multilinear ranks (r_1, r_2, r_3) . We have

$$\|\widetilde{\mathbf{U}}_i - \mathbf{U}_i\|_F \leq c(\eta_i/\delta) \|\widetilde{\mathbf{A}} - \mathbf{A}\|_F,$$

and

$$\|\widetilde{\mathbf{G}} - \mathbf{G}\|_F \leq \left[\sqrt{r_1 r_2 r_3} + c \bar{g} \delta^{-1} r_1 r_2 r_3 \sum_{i=1}^3 (\eta_i / \sqrt{r_i}) \right] \|\widetilde{\mathbf{A}} - \mathbf{A}\|_F.$$

where $\eta_i = \sum_{j=1}^{r_i} \sigma_1^2(\mathbf{A}_{(i)}) / \sigma_j^2(\mathbf{A}_{(i)})$.

Proof of Lemma 1. Since both $[\mathbf{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3]$ and $[\widetilde{\mathbf{G}}; \widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3]$ are HOSVD for \mathbf{A} and $\widetilde{\mathbf{A}}$. Each factor matrix is exactly the left singular vectors of the corresponding tensor matricization. We can apply the matrix perturbation theory for the factor matrices.

By the extension of the Davis-Kahan theorem for singular decomposition, i.e. Theorem 3 in Yu et al. (2014), under Assumption 6, we have for the j -th singular vector of $\mathbf{A}_{(1)}$,

$$\|\widetilde{\mathbf{U}}_{ij} - \mathbf{U}_{ij}\|_F \leq \frac{c(2\sigma_1(\mathbf{G}_{(i)}) + \|\widetilde{\mathbf{A}}_{(i)} - \mathbf{A}_{(i)}\|_{\text{op}}) \|\widetilde{\mathbf{A}}_{(i)} - \mathbf{A}_{(i)}\|_F}{\min[\sigma_{j-1}^2(\mathbf{G}_{(i)}) - \sigma_j^2(\mathbf{G}_{(i)}), \sigma_j^2(\mathbf{G}_{(i)}) - \sigma_{j+1}^2(\mathbf{G}_{(i)})]} \leq \frac{c\sigma_1(\mathbf{A}_{(i)}) \|\widetilde{\mathbf{A}} - \mathbf{A}\|_F}{\delta\sigma_j^2(\mathbf{A}_{(i)})}.$$

Therefore,

$$\|\tilde{\mathbf{U}}_i - \mathbf{U}_i\|_F^2 = \sum_{j=1}^{r_i} \|\tilde{\mathbf{U}}_{ij} - \mathbf{U}_{ij}\|_F^2 \leq \frac{c\eta_i^2}{\delta^2} \|\tilde{\mathbf{A}} - \mathbf{A}\|_F^2.$$

Note that $\mathbf{G} = \llbracket \mathbf{A}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \rrbracket$, $\tilde{\mathbf{G}} = \llbracket \tilde{\mathbf{A}}; \tilde{\mathbf{U}}_1', \tilde{\mathbf{U}}_2', \tilde{\mathbf{U}}_3' \rrbracket$, and $\|\mathbf{A}\|_F = \|\mathbf{G}\|_F$, so we can derive the upper bound for $\|\tilde{\mathbf{G}} - \mathbf{G}\|_F$

$$\begin{aligned} \|\tilde{\mathbf{G}} - \mathbf{G}\|_F &= \|(\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1)' \text{vec}(\tilde{\mathbf{A}}) - (\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)' \text{vec}(\mathbf{A})\|_F \\ &\leq \|(\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1)' (\text{vec}(\tilde{\mathbf{A}}) - \text{vec}(\mathbf{A}))\|_F + \|(\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1)' \text{vec}(\mathbf{A})\|_F \\ &\leq \|\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1\|_F \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \|\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_F \|\mathbf{G}\|_F \\ &\leq \sqrt{r_1 r_2 r_3} \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \sqrt{r_1 r_2 r_3} \bar{g} \|\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1 - \mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_F. \end{aligned}$$

And the last term can be bounded by

$$\begin{aligned} \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1\|_F &\leq \|\mathbf{U}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \tilde{\mathbf{U}}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1\|_F \\ &+ \|\tilde{\mathbf{U}}_3 \otimes \mathbf{U}_2 \otimes \mathbf{U}_1 - \tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \mathbf{U}_1\|_F + \|\tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \mathbf{U}_1 - \tilde{\mathbf{U}}_3 \otimes \tilde{\mathbf{U}}_2 \otimes \tilde{\mathbf{U}}_1\|_F \\ &\leq \sqrt{r_2 r_3} \|\mathbf{U}_1 - \tilde{\mathbf{U}}_1\|_F + \sqrt{r_1 r_3} \|\mathbf{U}_2 - \tilde{\mathbf{U}}_2\|_F + \sqrt{r_1 r_2} \|\mathbf{U}_3 - \tilde{\mathbf{U}}_3\|_F \\ &\leq \left(\sum_{i=1}^3 \frac{c\eta_i \sqrt{r_1 r_2 r_3}}{\sqrt{r_i} \delta} \right) \|\tilde{\mathbf{A}} - \mathbf{A}\|_F. \end{aligned}$$

Therefore, there exists a constant $C > 1$ such that

$$\|\tilde{\mathbf{G}} - \mathbf{G}\|_F \leq C \bar{g} \delta^{-1} r_1 r_2 r_3 \sum_{i=1}^3 (\eta_i / \sqrt{r_i}) \|\tilde{\mathbf{A}} - \mathbf{A}\|_F.$$

□

Lemma 2. (*Deviation bound*) Under the conditions of Theorem 3, for $T \gtrsim \log(N^2 P)$, the following two inequalities

$$\begin{aligned} \langle T^{-1} \mathbf{Z}' \mathbf{e}, \hat{\Delta}_{\mathbf{u}} \hat{\mathbf{g}} \rangle &\leq C \mathcal{M} \sqrt{\log(N^2 P)/T} \|\hat{\Delta}_{\mathbf{u}}\|_1, \\ \text{and } \langle T^{-1} \mathbf{Z}' \mathbf{e}, \mathbf{U} \hat{\Delta}_{\mathbf{g}} \rangle &\leq C \mathcal{M} \sqrt{\log(N^2 P)/T} \|\hat{\Delta}_{\mathbf{g}}\|_1, \end{aligned}$$

hold with probability at least $1 - C \exp(-c \log(N^2 P))$.

Proof of Lemma 2. For the first inequality,

$$\langle T^{-1} \mathbf{Z}' \mathbf{e}, \hat{\Delta}_{\mathbf{u}} \hat{\mathbf{g}} \rangle = \langle T^{-1} \mathbf{Z}' \mathbf{e} \hat{\mathbf{g}}', \hat{\Delta}_{\mathbf{u}} \rangle \leq \|T^{-1} \mathbf{Z}' \mathbf{e} \hat{\mathbf{g}}'\|_{\infty} \|\hat{\Delta}_{\mathbf{u}}\|_1 \leq \bar{g} \|T^{-1} \mathbf{Z}' \mathbf{e}\|_{\infty} \|\hat{\Delta}_{\mathbf{u}}\|_1,$$

where $\|\mathbf{Z}'\mathbf{e}\|_\infty = \|\mathbf{X}'\mathbf{E}\|_\infty = \max_{1 \leq i \leq NP, 1 \leq j \leq N} |\mathbf{e}'_i \mathbf{X}'\mathbf{E}\mathbf{e}_j|$ where \mathbf{e}_i is a coordinate vector whose i -th entry is 1 and the others are 0.

By Lemma 4, for any vector \mathbf{u} and \mathbf{v} s.t. $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$, and $\eta > 0$,

$$\mathbb{P} \left[|\mathbf{u}'(\mathbf{X}'\mathbf{E}/T)\mathbf{v}| > 2\pi \left(\lambda_{\max}(\Sigma_\epsilon) \left(1 + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \right) \right) \eta \right] \leq 6 \exp[-cT \min(\eta, \eta^2)].$$

Therefore, if we denote $\mathcal{M} \equiv \lambda_{\max}(\Sigma_\epsilon)[1 + \mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A})]$ and take a union bound,

$$\mathbb{P} \left[\max_{1 \leq i \leq NP, 1 \leq j \leq N} |\mathbf{e}'_i \mathbf{X}'\mathbf{E}\mathbf{e}_j/T| > 2\pi\mathcal{M}\eta \right] \leq 6N^2P \exp[-cT \min(\eta, \eta^2)].$$

Take $\eta = \sqrt{\log(N^2P)/T}$ and we obtain

$$\mathbb{P} \left[\max_{1 \leq i \leq NP, 1 \leq j \leq N} |\mathbf{e}'_i \mathbf{X}'\mathbf{E}\mathbf{e}_j/T| > 2\pi\mathcal{M}\sqrt{\log(N^2P)/T} \right] \leq C \exp[-c \log(N^2P)].$$

For the second inequality,

$$\begin{aligned} \langle T^{-1}\mathbf{Z}'\mathbf{e}, \mathbf{U}\hat{\Delta}_g \rangle &\leq \|T^{-1}\mathbf{U}\mathbf{Z}'\mathbf{e}\|_\infty \|\hat{\Delta}_g\|_1 \\ &= \|T^{-1}(\mathbf{U}_3 \otimes \mathbf{U}_2)' \mathbf{X}'\mathbf{E}\mathbf{U}_1\|_\infty \|\hat{\Delta}_g\|_1 \\ &= \max_{1 \leq i \leq r_2 r_3, 1 \leq j \leq r_1} |T^{-1}\mathbf{e}'_i(\mathbf{U}_3 \otimes \mathbf{U}_2)' \mathbf{X}'\mathbf{E}\mathbf{U}_1 \mathbf{e}_j| \cdot \|\hat{\Delta}_g\|_1. \end{aligned}$$

For any orthonormal matrix $\mathbf{U}_3 \otimes \mathbf{U}_2$, the spectral density of $\{\mathbf{X}_t(\mathbf{U}_3 \otimes \mathbf{U}_2)\}$ is defined as

$$f_{\mathbf{X}(\mathbf{U}_3 \otimes \mathbf{U}_2)}(\theta) = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} (\mathbf{U}_3 \otimes \mathbf{U}_2)' \Gamma_{\mathbf{X}}(\ell) (\mathbf{U}_3 \otimes \mathbf{U}_2) e^{-i\ell\theta} = (\mathbf{U}_3 \otimes \mathbf{U}_2)' f_{\mathbf{X}}(\theta) (\mathbf{U}_3 \otimes \mathbf{U}_2),$$

so we have $\mathcal{M}(f_{\mathbf{X}(\mathbf{U}_3 \otimes \mathbf{U}_2)}) \leq \mathcal{M}(f_{\mathbf{X}})$. Similarly, $\mathcal{M}(f_{\mathbf{E}\mathbf{U}_1}) \leq \mathcal{M}(f_{\mathbf{E}})$. Therefore, for any $\|\mathbf{u}\|_2 \leq 1$ and $\|\mathbf{v}\|_2 \leq 1$,

$$\mathbb{P}[|T^{-1}\mathbf{u}'((\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{X}'\mathbf{E}\mathbf{U}_1)\mathbf{v}| > 2\pi\mathcal{M}\eta] \leq 6 \exp[-cT \min(\eta, \eta^2)].$$

Taking a union bound, we can have

$$\mathbb{P} \left[\max_{1 \leq i \leq r_2 r_3, 1 \leq j \leq r_1} |T^{-1}\mathbf{e}'_i((\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{X}'\mathbf{E}\mathbf{U}_1)\mathbf{e}_j| > 2\pi\mathcal{M}\eta \right] \leq Cr_1 r_2 r_3 \exp[-cT \min(\eta, \eta^2)].$$

Take $\eta = \sqrt{\log(N^2P)/T}$ and we obtain

$$\begin{aligned} &\mathbb{P} \left[\max_{1 \leq i \leq r_2 r_3, 1 \leq j \leq r_1} |T^{-1}\mathbf{e}'_i((\mathbf{U}_3 \otimes \mathbf{U}_2)\mathbf{X}'\mathbf{E}\mathbf{U}_1)\mathbf{e}_j| > 2\pi\mathcal{M}\sqrt{\log(N^2P)/T} \right] \\ &\leq Cr_1 r_2 r_3 \exp[-c \log(N^2P)] \leq CN^2P \exp[-c \log(N^2P)] \leq C' \exp[-\log(N^2P)]. \end{aligned}$$

□

Lemma 3. (*Restricted eigenvalue*) Under the conditions of Theorem 3, if the sample size $T \gtrsim \mathcal{M}^2 d \min[\log(N^2 P), \log(cN^2 P/d)]$, for $\hat{\Delta} = \hat{U}\hat{g} - U\mathbf{g}$, where $(\hat{\mathcal{G}}, \hat{U}_1, \hat{U}_2, \hat{U}_3)$ and $(\mathcal{G}, U_1, U_2, U_3)$ belong to Ω ,

$$T^{-1} \|(\mathbf{I}_N \otimes \mathbf{X})\hat{\Delta}\|_2^2 \geq \alpha \|\hat{\Delta}\|_2^2 / 2,$$

with probability at least $1 - 2 \exp\{-cd \min[\log(N^2 P), \log(cN^2 P/d)]\}$, where $\alpha = \lambda_{\min}(\Sigma_{\epsilon})/\mu_{\max}(\mathcal{A})$, $d = 2\nu^{-2}r_1 r_2 r_3$.

Proof of Lemma 3. Denote by $\mathcal{K}(s) = \{\mathbf{v} \in \mathbb{R}^{NP} : \|\mathbf{v}\|_0 \leq s, \|\mathbf{v}\|_2 \leq 1\}$ the set of s -sparse vectors.

If we split $\hat{\Delta}$ into N parts, namely $(\hat{\Delta}) = (\hat{\Delta}'_1, \dots, \hat{\Delta}'_N)'$, where $\hat{\Delta}_k \in \mathbb{R}^{NP}$, we have

$$\|(\mathbf{I}_N \otimes \mathbf{X})\hat{\Delta}\|_2^2 = \sum_{i=1}^N \|\mathbf{X}\hat{\Delta}_i\|_2^2.$$

Correspondingly, we split \mathbf{U} and $\hat{\mathbf{U}}$ into N blocks, $\mathbf{U} = (\mathbf{M}'_1, \dots, \mathbf{M}'_N)$ and $\hat{\mathbf{U}} = (\hat{\mathbf{M}}'_1, \dots, \hat{\mathbf{M}}'_N)'$. Since $(\hat{\mathcal{G}}, \hat{U}_1, \hat{U}_2, \hat{U}_3)$ and $(\mathcal{G}, U_1, U_2, U_3)$ belong to Ω , the square of smallest nonzero entries in \mathbf{U}_i and $\hat{\mathbf{U}}_i$ is at least ν . Since each column in \mathbf{U}_i or $\hat{\mathbf{U}}_i$ has unit Euclidean norm, the number of nonzero entries in \mathbf{U}_i or $\hat{\mathbf{U}}_i$ is at most $1/\nu$. By the Kronecker structure in \mathbf{U} and $\hat{\mathbf{U}}$, the number of nonzero entries in each column of \mathbf{M}_i or $\hat{\mathbf{M}}_i$ is at most ν^{-2} . Therefore, $\|\hat{\Delta}_i\|_0 \leq 2\nu^{-2}r_1 r_2 r_3 := d$.

Denote $\hat{\Gamma} = \mathbf{X}'\mathbf{X}/T$ and $\Gamma = \mathbb{E}\hat{\Gamma}$. Since

$$\begin{aligned} T^{-1} \|(\mathbf{I}_N \otimes \mathbf{X})\hat{\Delta}\|_2^2 &= \hat{\Delta}'(\mathbf{I}_N \otimes \hat{\Gamma})\hat{\Delta} \\ &= \hat{\Delta}'(\mathbf{I}_N \otimes \Gamma)\hat{\Delta} + \hat{\Delta}'[\mathbf{I}_N \otimes (\hat{\Gamma} - \Gamma)]\hat{\Delta} \\ &= \hat{\Delta}'(\mathbf{I}_N \otimes \Gamma)\hat{\Delta} + \sum_{i=1}^N \hat{\Delta}'_i(\hat{\Gamma} - \Gamma)\hat{\Delta}_i. \end{aligned}$$

By the property of spectral density, $\lambda_{\min}(\Gamma) \geq \lambda_{\min}(\Sigma_{\epsilon})/\mu_{\max}(\mathcal{A})$, so $T^{-1}\mathbb{E}(\|(\mathbf{I}_N \otimes \mathbf{X})\hat{\Delta}\|_2^2) = T^{-1}\hat{\Delta}'(\mathbf{I}_N \otimes \Gamma)\hat{\Delta} \geq \lambda_{\min}(\Gamma) \geq \lambda_{\min}(\Sigma_{\epsilon})/\mu_{\max}(\mathcal{A})\|\hat{\Delta}\|_2^2 = \alpha\|\hat{\Delta}\|_2^2$.

So it remains to show that $\sup_{\hat{\Delta}_i \in \mathcal{K}(d)} \hat{\Delta}'_i(\hat{\Gamma} - \Gamma)\hat{\Delta}_i$ is close to zero. If we combine Lemma 4 and 5, we can obtain that for any $\eta > 0$,

$$\begin{aligned} &\mathbb{P} \left[\sup_{\mathbf{u} \in \mathcal{K}(d)} \left| \mathbf{u}'(\hat{\Gamma} - \Gamma)\mathbf{u} \right| > 2\pi\mathcal{M}(f_{\mathbf{X}})\eta \right] \\ &\leq 2 \exp\{-cT \min(\eta, \eta^2) + d \min[\log(NP), \log(cNP/d)]\}. \end{aligned}$$

Finally, if we take $\eta = \alpha/(4\pi\mathcal{M})$,

$$\begin{aligned}
& \mathbb{P} \left[\|(\mathbf{I}_N \otimes \mathbf{X})\widehat{\mathbf{\Delta}}\|_2^2/T \geq \alpha \|\widehat{\mathbf{\Delta}}\|_2^2/2 \right] \\
& \geq \mathbb{P} \left[\sup_{\widehat{\mathbf{\Delta}}_i \in \mathcal{K}(d)} \left| \widehat{\mathbf{\Delta}}'_i(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\widehat{\mathbf{\Delta}}_i \right| < \alpha \|\widehat{\mathbf{\Delta}}_i\|_2^2/2 \right] \\
& \geq 1 - 2 \exp\{-cT\mathcal{M}^{-2} + 2d \min[\log(NP), \log(cNP/d)]\} \\
& \geq 1 - 2 \exp\{-cd \min[\log(NP), \log(cNP/d)]\}.
\end{aligned}$$

□

Next, to make the proof self-contained, we state two lemmas to establish concentration inequalities for Gaussian time series from Basu and Michailidis (2015). The first one is Proposition 2.4 in Basu and Michailidis (2015).

Lemma 4. *For a stationary and centered Gaussian time series $\{\mathbf{x}_t\}$ satisfying the bounded spectral density condition, there exists a constant $c > 0$ such that for any vector $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{u}\|_2 \leq 1$, $\|\mathbf{v}\|_2 \leq 1$, and any $\eta \geq 0$,*

$$\mathbb{P}[|\mathbf{v}'(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\mathbf{v}| > 2\pi\mathcal{M}(f_{\mathbf{X}})\eta] \leq 2 \exp[-cT \min(\eta^2, \eta)],$$

where $\widehat{\mathbf{\Gamma}} = T^{-1}\mathbf{X}'\mathbf{X}$ and $\mathbf{X} = [\mathbf{x}_T, \dots, \mathbf{x}_1]'$.

For two p -dimensional, centered, stationary Gaussian processes \mathbf{y}_t and $\mathbf{\epsilon}_t$ such that $\text{Cov}(\mathbf{y}_t, \mathbf{\epsilon}_t) = 0$ for every t . Let $\mathbf{X} = [\mathbf{x}_T, \dots, \mathbf{x}_1]'$ and $\mathbf{E} = [\mathbf{\epsilon}_T, \dots, \mathbf{\epsilon}_1]'$ be the data matrices. Then, there exists a constant $c > 0$ such that for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{u}\|_2 \leq 1$ and $\|\mathbf{v}\|_2 \leq 1$, we have

$$\mathbb{P} \left[|\mathbf{u}'(\mathbf{X}'\mathbf{E}/T)\mathbf{v}| > 2\pi \left(\lambda_{\max}(\mathbf{\Sigma}_{\mathbf{\epsilon}}) \left(1 + \frac{\mu_{\max}(\mathcal{A})}{\mu_{\min}(\mathcal{A})} \right) \right) \right].$$

Finally, we state a union concentration inequality for vectors in a sparse set via discretization from Lemma F.2 in Basu and Michailidis (2015).

Lemma 5. *Consider a symmetric matrix $\mathbf{D}_{p \times p}$. If, for any $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 \leq 1$, and any $\eta \geq 0$,*

$$\mathbb{P}[|\mathbf{v}'\mathbf{D}\mathbf{v}| > C\eta] \leq 2 \exp[-cT \min(\eta, \eta^2)]$$

then, for any integers $s \geq 1$, we have

$$\mathbb{P} \left[\sup_{\mathbf{v} \in \mathcal{K}(s)} |\mathbf{v}'\mathbf{D}\mathbf{v}| > C\eta \right] \leq 2 \exp[-cT \min(\eta^2, \eta) + s \min\{\log(p), \log(cp/s)\}].$$

A.4 Generation of orthonormal matrices

To generate an arbitrary tall orthonormal matrix $\mathbf{O} \in \mathbb{R}^{m \times n}$ with $m > n$, we first generate an $m \times m$ square matrix of independent standard normal random numbers, and then set its top n singular vectors as the columns of \mathbf{O} .

We next generate the sparse orthonormal matrices in the second experiment in Section 5.1. For the case of $(N, P) = (10, 5)$, let

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{A}_{3 \times 2} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{3 \times 1} & \mathbf{B}_{3 \times 2} \\ \mathbf{0}_{4 \times 2} & \mathbf{a}_{4 \times 1} \end{bmatrix} \in \mathbb{R}^{10 \times 3} \quad \text{and} \quad \mathbf{U}_3 = \begin{bmatrix} \mathbf{C}_{3 \times 2} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{2 \times 2} & \mathbf{b}_{2 \times 1} \end{bmatrix} \in \mathbb{R}^{10 \times 3}, \quad (15)$$

where $\mathbf{A}_{3 \times 2}$, $\mathbf{B}_{3 \times 2}$ and $\mathbf{C}_{3 \times 2}$ are 3×2 orthonormal matrices generated by the method mentioned above, $\mathbf{a}_{4 \times 1}$ and $\mathbf{b}_{2 \times 1}$ are column vector of independent standard normal random numbers, which are further scaled such that $\|\mathbf{a}_{4 \times 1}\|_2 = 1$ and $\|\mathbf{b}_{2 \times 1}\|_2 = 1$, and \mathbf{U}_2 is generated by the same method as that for \mathbf{U}_1 . For the case of $(N, P) = (20, 8)$, let

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{a}_{7 \times 1} & \mathbf{0}_{7 \times 2} \\ \mathbf{A}_{3 \times 2} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{3 \times 1} & \mathbf{B}_{3 \times 2} \\ \mathbf{0}_{7 \times 2} & \mathbf{b}_{7 \times 1} \end{bmatrix} \in \mathbb{R}^{20 \times 3}, \quad \mathbf{U}_2 = \begin{bmatrix} \mathbf{U}_{10 \times 3}^* \\ \mathbf{0}_{10 \times 3} \end{bmatrix} \in \mathbb{R}^{20 \times 3}, \quad \mathbf{U}_3 = \begin{bmatrix} \mathbf{C}_{3 \times 2} & \mathbf{0}_{3 \times 1} \\ \mathbf{c}_{2 \times 1} & \mathbf{0}_{2 \times 2} \\ \mathbf{0}_{3 \times 2} & \mathbf{d}_{3 \times 1} \end{bmatrix} \in \mathbb{R}^{8 \times 3},$$

where $\mathbf{U}_{10 \times 3}^*$ is generated by the same method as that for \mathbf{U}_1 in (15), and the other components (orthonormal matrices \mathbf{A} , \mathbf{B} and \mathbf{C} , and column vectors \mathbf{a} , \mathbf{b} , \mathbf{c} and \mathbf{d}) are generated similarly to those in the case of $(N, P) = (10, 5)$. Note that the above constructed matrices are all orthogonal, and we then further standardize them into orthonormal matrices.

References

- Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81:1203–1227.
- Athanasopoulos, G. and Vahid, F. (2008). VARMA versus VAR for macroeconomic forecasting. *Journal of Business & Economic Statistics*, 26:237–252.

- Bai, Z. D. (1997). Circular law. *The Annals of Probability*, 25:494–529.
- Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43:1535–1567.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3:1–122.
- Bunea, F., She, Y., and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40:2359–2388.
- Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26:735–761.
- Chan, J. C. C., Eisenstat, E., and Koop, G. (2016). Large bayesian VARMA. *Journal of Econometrics*, 192:374–390.
- Chen, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B*, 74:203–221.
- Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100:901–920.
- Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107:1533–1545.
- Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25:1077–1096.

- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21:1253–1278.
- De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146:318–328.
- Dias, G. F. and Kapetanios, G. (2018). Estimation and forecasting in vector autoregressive moving average models for rich datasets. *Journal of Econometrics*, 202:75–91.
- Ding, S. and Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society, Series B*, 80:387–408.
- Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103:889–903.
- Hallin, M. and Lippi, M. (2013). Factor models in high-dimensional time series: A time-domain approach. *Stochastic Processes and their Applications*, 123:2678–2695.
- Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16:3115–3150.
- Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264.
- Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186:325–344.
- Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51:455–500.
- Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics*, 28:177–203.
- Lai, R. and Osher, S. (2014). A splitting method for orthogonality constrained problems. *Journal of Scientific Computing*, 58:431–449.

- Lange, K. (2010). *Numerical analysis for statisticians*. Springer Science & Business Media.
- Li, G., Leng, C., and Tsai, C.-L. (2014). A hybrid bootstrap approach to unit root tests. *Journal of Time Series Analysis*, 35:299–321.
- Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112:1131–1146.
- Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10:520–545.
- Lian, H., Feng, S., and Zhao, K. (2015). Parametric and semiparametric reduced-rank regression with flexible sparsity. *Journal of Multivariate Analysis*, 136:163–174.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.
- Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097.
- Raskutti, G., Yuan, M., and Chen, H. (2019). Convex regularization for high-dimensional multi-response tensor regression. *The Annals of Statistics*, 47:1554–1584.
- Ravenna, F. (2007). Vector autoregressions and reduced form representations of dsge models. *Journal of Monetary Economics*, 54:2048–2064.
- Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika*, 70:145–156.
- Said, E. S. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71:599–607.
- Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81:142–149.

- Stock, J. H. and Watson, M. (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, pages 173–205.
- Sun, W. W. and Li, L. (2017). Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18:4908–4944.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288.
- Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association*, 93:1188–1202.
- Tsay, R. S. (2010). *Analysis of Financial Time Series*. John Wiley & Sons, 3rd edition.
- Tsay, R. S. (2013). *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons.
- Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311.
- Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W. (2019). Sofar: large-scale association network learning. *IEEE Transactions on Information Theory*, 65:4924–4939.
- Velu, R. P. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.
- Velu, R. P., Reinsel, G. C., and Wichern, D. W. (1986). Reduced rank models for multiple time series. *Biometrika*, 73:105–118.
- Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208:231–248.
- Wilms, I., Basu, S., Bien, J., and Matteson, D. S. (2017). Sparse identification and estimation of large-scale vector autoregressive moving averages. *arXiv preprint arXiv:1707.09208*.

- Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10:352–379.
- Yu, Y., Wang, T., and Samworth, R. J. (2014). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323.
- Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*, 69:329–346.
- Zhang, A. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics*, 47:936–964.
- Zhao, J. and Leng, C. (2014). Structure LASSO for regression with matrix covariates. *Statistica Sinica*, 24:799–814.
- Zheng, L. and Raskutti, G. (2018). Testing for high-dimensional network parameters in auto-regressive models. *arXiv preprint arXiv:1812.03659*.
- Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552.
- Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017). Network vector autoregression. *The Annals of Statistics*, 45:1096–1123.

Table 1: Average estimation errors of $\hat{\mathcal{A}}_{\text{MLR}}$ and two competing methods for low-dimensional VAR processes with temporal ranks $1 \leq r_3 \leq 5$.

r_3	1			2			3			4			5		
ρ	MLR	RRR	OLS	MLR	RRR	OLS	MLR	RRR	OLS	MLR	RRR	OLS	MLR	RRR	OLS
0	0.237	0.450	0.727	0.260	0.455	0.725	0.271	0.425	0.725	0.287	0.514	0.725	0.301	0.577	0.725
0.3	0.273	0.633	0.827	0.302	0.626	0.825	0.302	0.509	0.824	0.323	0.645	0.824	0.350	0.700	0.825
0.5	0.348	0.821	0.968	0.406	0.804	0.961	0.365	0.681	0.958	0.395	0.816	0.962	0.444	0.857	0.962

Table 2: Average estimation errors of $\hat{\mathcal{A}}_{\text{SHORR}}$ and five competing methods for high-dimensional VAR processes.

ρ	SHORR	LASSO	RSRRR	RSSVD	RRANN	SOFAR
$(N, P, T) = (10, 5, 400)$						
0	0.4139	0.7103	0.6388	0.8102	0.5702	0.7378
0.3	0.5047	0.7439	0.8588	0.8808	0.7751	0.7720
0.5	0.5951	0.8149	1.1598	0.9615	1.0549	0.9716
$(N, P, T) = (20, 8, 400)$						
0	0.5180	1.1266	1.2025	1.2082	1.0310	1.1907
0.3	0.5771	1.1834	1.2667	1.2433	1.0969	1.2607
0.5	0.6541	1.2485	1.2904	1.2588	1.1423	1.3151

Table 3: Average estimation errors of $\hat{\mathcal{A}}_{\text{MLR}}$, $\hat{\mathcal{A}}_{\text{SHORR}}$ and their competing methods for VARMA processes.

ρ	$(N, T) = (10, 1000)$			$(N, T) = (20, 500)$					
	MLR	RRR	OLS	SHORR	LASSO	RSRRR	RSSVD	RRANN	SOFAR
0	0.348	0.495	0.702	0.5984	1.0375	0.8501	0.9432	0.8029	0.9483
0.3	0.409	0.615	0.799	0.6693	1.2052	0.8947	1.0853	0.8660	1.2108
0.5	0.551	0.818	0.933	0.7850	1.2400	0.9604	1.1389	0.9522	1.2920

Table 4: Forecasting errors of $\hat{\mathcal{A}}_{\text{MLR}}$, $\hat{\mathcal{A}}_{\text{SHORR}}$ and their competing methods.

Criterion	Low dimension			High dimension					
	OLS	RRR	MLR	SHORR	LASSO	RSRRR	RSSVD	RRANN	SOFAR
ℓ_2 norm	20.16	13.31	5.81	5.35	6.72	6.15	6.33	8.16	6.28
ℓ_∞ norm	8.32	4.55	2.56	2.44	3.06	2.93	3.02	3.36	3.02

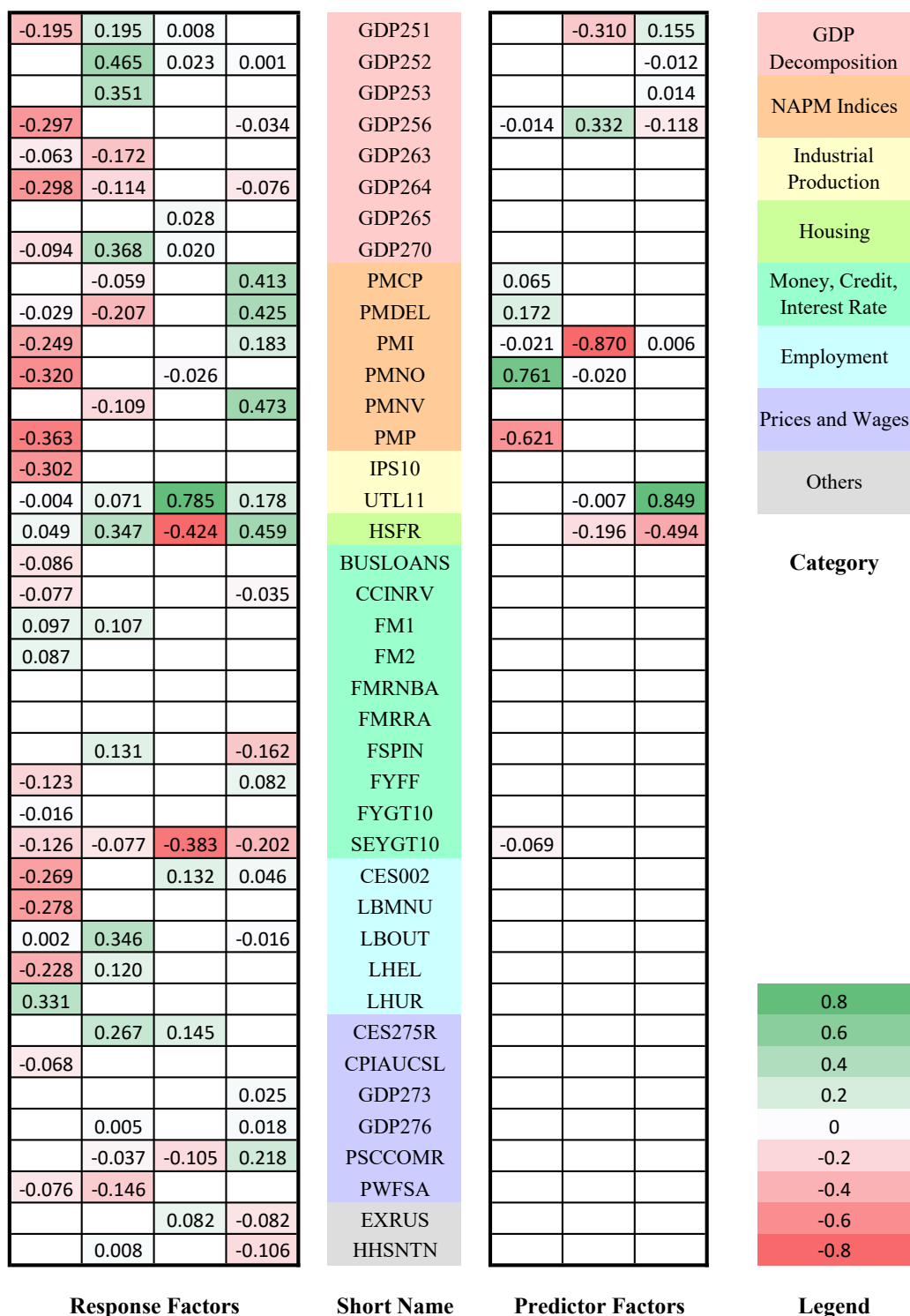


Figure 2: Estimated coefficients in the response and predictor factor loading matrices.

Table 5: Forty quarterly macroeconomic variables belonging to 8 categories. Category code (C) represents: 1 = GDP and its decomposition, 2 = national association of purchasing managers (NAPM) indices, 3 = industrial production, 4 = housing, 5 = money, credit, interest rates, 6 = employment, 7 = prices and wages, 8 = others. Variables are seasonally adjusted except for those in category 5. All variables are transformed to stationarity with the following transformation codes (T): 1 = no transformation, 2 = first difference, 3 = second difference, 4 = log, 5 = first difference of logged variables, 6 = second difference of logged variables.

Short name	C	T	Description	Short name	C	T	Description
GDP251	1	5	Real GDP, quantity index (2000=100)	FM2	5	6	Money stock: M2 (bil\$)
GDP252	1	5	Real Personal Cons. Exp., Quantity Index	FMRNBA	5	3	Depository inst reserves: nonborrowed (mil\$)
GDP253	1	5	Real personal cons expenditures: durable goods	FMRRRA	5	6	Depository inst reserves: total (mil\$)
GDP256	1	5	Real gross private domestic investment	FSPIN	5	5	S&P's common stock price index: industrials
GDP263	1	5	Real exports	FYFF	5	2	Interest rate: federal funds (% per annum)
GDP264	1	5	Real imports	FYGT10	5	2	Interest rate: US treasury const. mat., 10-yr
GDP265	1	5	Real govt cons expenditures & gross investment	SEYGT10	5	1	Spread btwn 10-yr and 3-mth T-bill rates
GDP270	1	5	Real final sales to domestic purchasers	CES002	6	5	Employees, nonfarm: total private
PMCP	2	1	NAPM commodity price index (%)	LBMNU	6	5	Hrs of all persons: nonfarm business sector
PMDEL	2	1	NAPM vendor deliveries index (%)	LBOUT	6	5	Output per hr: all persons, business sec
PMI	2	1	Purchasing managers' index	LHEL	6	2	Index of help-wanted ads in newspapers
PMNO	2	1	NAPM new orders index (%)	LHUR	6	2	Unemp. rate: All workers, 16 and over (%)
PMNV	2	1	NAPM inventories index (%)	CES275R	7	5	Real avg hrly earnings, non-farm prod. workers
PMP	2	1	NAPM production index (%)	CPIAUCSL	7	6	CPI all items
IPS10	3	5	Industrial production index: total	GDP273	7	6	Personal Consumption Exp.: price index
UTL11	3	1	Capacity utilization: manufacturing (SIC)	GDP276	7	6	Housing price index
HSFR	4	4	Housing starts: Total (thousands)	PSCCOMR	7	5	Real spot market price index: all commodities
BUSLOANS	5	6	Comm. and industrial loans at all comm. Banks	PWFSA	7	6	Producer price index: finished goods
CCINRV	5	6	Consumer credit outstanding: nonrevolving	EXRUS	8	5	US effective exchange rate: index number
FM1	5	6	Money stock: M1 (bil\$)	HHSNTN	8	2	Univ of Mich index of consumer expectations