JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 539, 1338–1356: Theory and Methods
https://doi.org/10.1080/01621459.2020.1855183

High-Dimensional Vector Autoregressive Time Series Modeling via Tensor Decomposition

Di Wang[a], Yao Zheng[b], Heng Lian[c], and Guodong Li[a]

[a]Department of Statistics and Actuarial Science, University of Hong Kong, Pok Fu Lam, Hong Kong; [b]Department of Statistics, University of Connecticut, Storrs, CT; [c]Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong

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 article proposes to rearrange the transition matrices of the model into a tensor form such that the parameter space can be restricted along three directions simultaneously via tensor decomposition. In contrast, the reduced-rank regression method can restrict the parameter space in only one direction. Besides achieving substantial dimension reduction, the proposed model is interpretable from the factor modeling perspective. Moreover, to handle high-dimensional time series, this article considers imposing sparsity on factor matrices to improve the model 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 nonasymptotic properties of the sparsity-inducing estimator and propose an ADMM algorithm for regularized estimation. Simulation experiments and a real data example demonstrate the advantages of the proposed approach over various existing methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted November 2020

KEYWORDS
Factor model; High-dimensional time series; Reduced-rank regression; 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, for example, the banded structure (Guo, Wang, and Yao 2016), network structure (Zhu et al. 2017), and linear restrictions (Zheng and Cheng 2020). Consider the VAR model of the form (Lütkepohl 2005; Tsay 2010):

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

where $\{\boldsymbol{y}_t\}$ is the observed time series with $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})' \in \mathbb{R}^N$, $\{\boldsymbol{\epsilon}_t\}$ are independent and identically distributed (iid) innovations with $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$, $\mathbb{E}(\boldsymbol{\epsilon}_t) = 0$ and $\mathrm{var}(\boldsymbol{\epsilon}_t) < \infty$, $\boldsymbol{A}_j$'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 dimensions N and P are moderately large (De Mol, Giannone, and Reichlin 2008; Carriero, Kapetanios, and Marcellino 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, Eisenstat, and Koop 2016). However, the VARMA model may have a serious identification problem (Chan, Eisenstat, and Koop 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 to provide a better fit for the data (Ravenna 2007). For example, to guarantee the approximation accuracy, we need to assume that $P \to \infty$ and $PT^{-1/3} \to 0$ as $T \to \infty$ for univariate and multivariate cases (Said and Dickey 1984; Li, Leng, and Tsai 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 $\boldsymbol{A}_j$'s are sparse and apply sparsity-inducing regularized estimation, for example, the ℓ_1-regularization (Lasso or Dantzig selector) for VAR models (Basu and Michailidis 2015; Han, Lu, and Liu 2015; Kock and Callot 2015; Davis, Zang, and Zheng 2016; Wu and Wu 2016). However, unlike the traditional linear regression, time series data have nonnegligible temporal and cross-sectional dependencies, which will seriously affect the accuracy of the regularized estimation. Moreover, as 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

CONTACT Heng Lian henglian@cityu.edu.hk Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong; Guodong Li gdli@hku.hk Department of Statistics and Actuarial Science, University of Hong Kong, Pok Fu Lam, Hong Kong.
Lian and Li are contributed to the article equally.

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

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, Dong, and Chan 2013; Basu, Li, and Michailidis 2019; Raskutti, Yuan, and Chen 2019). The VAR model in (1) can be rewritten as

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

where $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-P}')'$, and $\boldsymbol{A}^{(C)} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_P)$ is assumed to have a low rank (Velu, Reinsel, and Wichern 1986; Velu and Reinsel 2013). Based on the reduced-rank VAR model in (2), Carriero, Kapetanios, and Marcellino (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 $\boldsymbol{A}_j$'s with $P > 1$. Specifically, the rank of $\boldsymbol{A}^{(C)}$ is the dimension of the column space of $\boldsymbol{A}_j$'s. Denote $\boldsymbol{A}^{(R)} = (\boldsymbol{A}_1', \boldsymbol{A}_2', \ldots, \boldsymbol{A}_P')$ and $\boldsymbol{A}^{(L)} = (\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_2), \ldots, \mathrm{vec}(\boldsymbol{A}_P))'$, where $\mathrm{vec}(\boldsymbol{A}_j)$ is the vectorization of $\boldsymbol{A}_j$. The ranks of $\boldsymbol{A}^{(R)}$ and $\boldsymbol{A}^{(L)}$ are then the dimensions of the row space and vectorized matrix space of $\boldsymbol{A}_j$'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 $\boldsymbol{A}^{(R)}$. Moreover, the transition matrices $\boldsymbol{A}_j$'s may have a low-rank structure along different lags, that is, $\boldsymbol{A}^{(L)}$ may be 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 $\boldsymbol{A}^{(C)}$, $\boldsymbol{A}^{(R)}$, and $\boldsymbol{A}^{(L)}$ 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 $\boldsymbol{A}_j$'s into a tensor. Interestingly, the corresponding mode-1, -2, and -3 matricizations of the tensor happen to be $\boldsymbol{A}^{(C)}$, $\boldsymbol{A}^{(R)}$, and $\boldsymbol{A}^{(L)}$, 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.

In the literature, low-rank structures of high-dimensional time series are commonly explored through factor models (Stock and Watson 2005; Bai and Ng 2008; Stock and Watson 2011; Bai and Wang 2016). Similarly, as a means of low-rank discovery for VAR processes, the proposed model is naturally interpretable from the factor modeling perspective. As we will discuss in Section 2.2, by imposing the low-rankness along three directions, the proposed model can extract different dynamic factors across response variables, predictor variables, and predictor time lags. Indeed, the proposed model can be written as a static factor model (SFM, Bai and Wang 2016) yet endowed with additional low-rank structures for more substantial dimension reduction. However, in contrast to factor models which are mainly used for interpretation, it is worth noting that the proposed model can be used for forecasting. On the other hand, the dynamic factor model (DFM) in the literature can be constructed by combining the SFM with a certain dynamic structure for the latent factors (Stock and Watson 2011). Compared to the DFM with VAR latent factors (Amengual and Watson 2007), the proposed model may be more flexible in the sense that it can extract different sets of dynamic factors from the response $\boldsymbol{y}_t$ and the lagged predictors $\boldsymbol{y}_{t-j}$'s, whereas the DFM restricts them to be identical. In addition, the proposed model can capture the possible low-rank structure across the P time lags.

Another important contribution of this article is to introduce a sparse decomposition for the transition tensor to further increase the estimation efficiency for much higher-dimensional time series data. 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, She, and Wegkamp (2012) considered row-wise sparsity in singular value decomposition, where zero rows imply irrelevance of the corresponding predictors to the responses; Lian, Feng, and Zhao (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, Chan, and Stenseth (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 estimation 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, Li, and Zhu 2013; Li and Zhang 2017; Sun and Li 2017; Li et al. 2018; Raskutti, Yuan, and Chen 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 methods have the following attractive features:

(a) The proposed model substantially reduces the dimension along three directions of the transition tensor, allowing each direction to have a different low-rank structure. This results in interpretable physical structures and interesting connections with factor models in the literature, and allows us to handle much higher dimensional data than the reduced-rank VAR model in (2).
(b) Through the sparsity assumption on the three factor matrices, the proposed high-dimensional method further improves the model interpretability and estimation efficiency by selecting important variables for each response, predictor or temporal factor. The corresponding estimation can be accomplished by an ADMM algorithm which effectively untangles the ℓ_1-regularization and orthogonality constraints.

The rest of the article is organized as follows. Section 2 introduces the proposed model and discusses its connections with factor models. Section 3 presents asymptotic properties

of the least squares estimator in low dimensions and an alternating least squares algorithm. For the high-dimensional case, the sparse higher-order reduced-rank estimation is proposed in Section 4, taking into account both the orthogonality and sparsity. Its nonasymptotic properties are established, and an ADMM algorithm is developed. A consistent rank selection method is proposed in Section 5. Simulation experiments and real data analysis are presented in Sections 6 and 7, respectively. A short discussion is given in Section 7. All technical proofs are given in a separate online supplementary file, and all codes and datasets in this article can be found at *https://github.com/diwangstat/VAR-Tensor/*.

2. Multilinear Low-Rank Vector Autoregression

2.1. Tensor Decomposition

Tensors, also known as multidimensional arrays, are natural higher-order extensions of matrices. A multidimensional array $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is called a Kth-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 article will focus on third-order tensors.

Throughout the article, we denote vectors by small boldface letters $\boldsymbol{y}, \boldsymbol{x}, \ldots$, matrices by capital letters $\boldsymbol{Y}, \boldsymbol{X}, \ldots$, and tensors by Euler script capital letters $\mathcal{Y}, \mathcal{X}, \ldots$. For a vector $\boldsymbol{x}$, denote by $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_2$ its ℓ_1 and ℓ_2 norms, respectively. For a matrix $\boldsymbol{X}$, denote by $\|\boldsymbol{X}\|_{\mathrm{F}}$, $\|\boldsymbol{X}\|_1$, $\|\boldsymbol{X}\|_0$, $\|\boldsymbol{X}\|_{\mathrm{op}}$, $\|\boldsymbol{X}\|_*$, $\mathrm{vec}(\boldsymbol{X})$, $\boldsymbol{X}'$ and $\sigma_j(\boldsymbol{X})$ its Frobenius norm, vectorized ℓ_1 norm (i.e., $\|\boldsymbol{X}\|_1 = \|\mathrm{vec}(\boldsymbol{X})\|_1$), ℓ_0 "norm," spectral norm, nuclear norm, vectorization, transpose and the jth largest singular value, respectively. For two symmetric matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, we write $\boldsymbol{X} \preceq \boldsymbol{Y}$ if $\boldsymbol{Y} - \boldsymbol{X}$ is positive semidefinite. Furthermore, for a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, let $\|\mathcal{X}\|_{\mathrm{F}} = \left(\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\sum_{k=1}^{p_3}\mathcal{X}_{ijk}^2\right)^{1/2}$ 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 $1 \le i \le p_1, 1 \le j \le p_2$, and $1 \le k \le p_3$, and $\mathcal{X}_{(1)}$ contains all mode-1 fibers $\{(\mathcal{X}_{[:,i_2,i_3]}) \in \mathbb{R}^{p_1} : 1 \le i_2 \le p_2, 1 \le i_3 \le p_3\}$. 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 $\boldsymbol{Y} \in \mathbb{R}^{q_1 \times p_1}$ is defined as

$$\mathcal{X} \times_1 \boldsymbol{Y} = \left(\sum_{i=1}^{p_1} \mathcal{X}_{ijk}\boldsymbol{Y}_{si}\right)_{1 \le s \le q_1, 1 \le j \le p_2, 1 \le k \le p_3}. \tag{3}$$

Multiplications $\times_2$ and $\times_3$ can be defined similarly.

Unlike matrices, there is no universal definition of the rank for tensors. In this article, 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

$$\begin{aligned} r_1 &= \mathrm{rank}_1(\mathcal{X}) := \mathrm{rank}(\mathcal{X}_{(1)}) \\ &= \dim(\mathrm{span}\{\mathcal{X}_{[:,i_2,i_3]} \in \mathbb{R}^{p_1} : 1 \le i_2 \le p_2, 1 \le i_3 \le p_3\}), \end{aligned} \tag{4}$$

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 $\mathrm{rank}_j(\mathcal{X}) = r_j$ for $1 \le j \le 3$, then there exists a Tucker decomposition (Tucker 1966; De Lathauwer, De Moor, and Vandewalle 2000),

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

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

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 $1 \le j \le 3$. It can be verified that $\mathcal{A}_{(1)} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_P)$, $\mathcal{A}_{(2)} = (\boldsymbol{A}_1', \boldsymbol{A}_2', \ldots, \boldsymbol{A}_P')$ and $\mathcal{A}_{(3)} = (\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_2), \ldots, \mathrm{vec}(\boldsymbol{A}_P))'$, which correspond to the column, row and vectorized matrix spaces of $\boldsymbol{A}_j$'s, respectively.

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

$$\boldsymbol{y}_t = (\mathcal{G} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3)_{(1)}\boldsymbol{x}_t + \boldsymbol{\epsilon}_t, \tag{5}$$

where $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-P}')'$. For simplicity, we call model (5) the multilinear low-rank VAR model.

In addition, since $(\mathcal{G} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3)_{(1)} = \boldsymbol{U}_1 \mathcal{G}_{(1)}(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'$, where $\otimes$ is the Kronecker product, model (5) also has the following equivalent forms

$$\boldsymbol{y}_t = \boldsymbol{U}_1 \mathcal{G}_{(1)}(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'\boldsymbol{x}_t + \boldsymbol{\epsilon}_t = \boldsymbol{U}_1 \mathcal{G}_{(1)} \mathrm{vec}(\boldsymbol{U}_2' \boldsymbol{X}_t \boldsymbol{U}_3) + \boldsymbol{\epsilon}_t, \tag{6}$$

where $\boldsymbol{X}_t = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-P})$.

Assumption 1. All roots of the matrix polynomial $\mathcal{A}(z) = \boldsymbol{I}_N - \boldsymbol{A}_1 z - \cdots - \boldsymbol{A}_P z^P$, $z \in \mathbb{C}$, 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 $\rho(\boldsymbol{A}_1) < 1$, where $\rho(\boldsymbol{A}_1)$ denotes the spectral radius of $\boldsymbol{A}_1$.

Remark 1. To gain insight into the effect of the stationarity condition on the entries of $\boldsymbol{A}_1$, we may consider the following result regarding random matrices. Suppose that the entries of $\boldsymbol{A}_1$ are iid with mean zero and variance σ^2, that is, they are equally important. Then, by Bai (1997), $N^{-1/2}\rho(\boldsymbol{A}_1) \to \sigma$ in probability as $N \to \infty$. In other words, when $\rho(\boldsymbol{A}_1) < 1$, a larger N will shrink the entries of $\boldsymbol{A}_1$ toward zero.

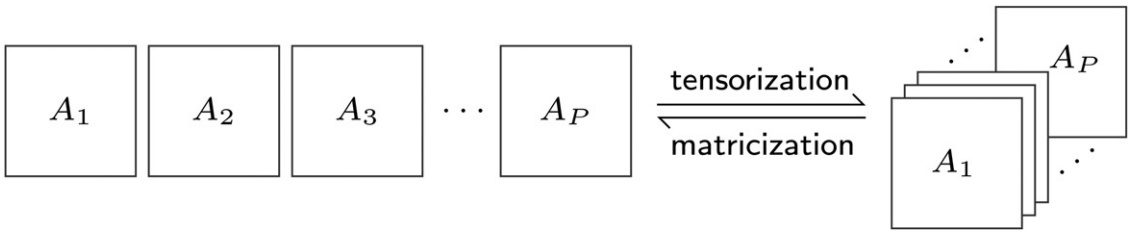

Figure 1. Rearranging P transition matrices of a VAR model into a tensor.

Note that the Tucker decomposition in (5) is not unique since $[\![\mathcal{G}; \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3]\!] = [\![\mathcal{G} \times_1 \boldsymbol{O}_1 \times_2 \boldsymbol{O}_2 \times_3 \boldsymbol{O}_3; \boldsymbol{U}_1\boldsymbol{O}_1^{-1}, \boldsymbol{U}_2\boldsymbol{O}_2^{-1}, \boldsymbol{U}_3\boldsymbol{O}_3^{-1}]\!]$ for any nonsingular matrices $\boldsymbol{O}_1 \in \mathbb{R}^{r_1\times r_1}$, $\boldsymbol{O}_2 \in \mathbb{R}^{r_2\times r_2}$ and $\boldsymbol{O}_3 \in \mathbb{R}^{r_3\times r_3}$. Hence, we consider a special Tucker decomposition: the higher-order singular value decomposition (HOSVD); see De Lathauwer, De Moor, and Vandewalle (2000). Specifically, we let $\boldsymbol{U}_j$ be a tall matrix consisting of the top r_j left singular vectors of $\mathcal{A}_{(j)}$ for each $1 \le j \le 3$, where (r_1, r_2, r_3) are the multilinear ranks of the tensor $\mathcal{A}$. Let the core tensor $\mathcal{G} = \mathcal{A} \times_1 \boldsymbol{U}_1' \times_2 \boldsymbol{U}_2' \times_3 \boldsymbol{U}_3'$. Then $\mathcal{G}$ has the following *all-orthogonal* property: for each $1 \le j \le 3$, the rows of $\mathcal{G}_{(j)}$ are pairwise orthogonal.

Remark 2. Due to the HOSVD, the proposed multilinear low-rank VAR model in (5) has only $r_1r_2r_3 + (N - r_1)r_1 + (N - r_2)r_2 + (P - r_3)r_3$ parameters, that is, the dimension increases linearly in N and P; see Zhang (2019). By contrast, model (1) has N^2P parameters, while the reduced-rank VAR model in (2) has $(NP + N - r_1)r_1$ parameters, where $r_1 = \text{rank}(\mathcal{A}_{(1)})$.

Since $\boldsymbol{U}_1$ is orthonormal, it follows from (6) that

$$\boldsymbol{U}_1'\boldsymbol{y}_t = \mathcal{G}_{(1)}(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'\boldsymbol{x}_t + \boldsymbol{U}_1'\boldsymbol{\epsilon}_t = \mathcal{G}_{(1)}\text{vec}(\boldsymbol{U}_2'\boldsymbol{X}_t\boldsymbol{U}_3) + \boldsymbol{U}_1'\boldsymbol{\epsilon}_t. \tag{7}$$

The above representation reveals an interesting dynamic factor based interpretation for the proposed model. Specifically, $\boldsymbol{U}_1'\boldsymbol{y}_t := \boldsymbol{f}_t^{\text{Response}} = (f_{1,t}^{\text{Response}}, \ldots, f_{r_1,t}^{\text{Response}})' \in \mathbb{R}^{r_1}$ represents r_1 *response factors* across the N variables of $\boldsymbol{y}_t$, where $f_{j,t}^{\text{Response}} = \boldsymbol{u}_{1,j}'\boldsymbol{y}_t = \sum_{i=1}^N (\boldsymbol{U}_1)_{ij}y_{it}$ is the jth response factor, for $1 \le j \le r_1$. Thus, if the (i, j)th entry of $\boldsymbol{U}_1$ is zero, that is, $(\boldsymbol{U}_1)_{ij} = 0$, then y_{it} is irrelevant to $\boldsymbol{f}_{j,t}^{\text{Response}}$. In other words, $\boldsymbol{U}_1$ can be interpreted as the loadings of the response factors.

On the right side of (7), the predictor has the bilinear form $\boldsymbol{U}_2'\boldsymbol{X}_t\boldsymbol{U}_3$. On the one hand, $\boldsymbol{U}_2'\boldsymbol{X}_t := \boldsymbol{F}_t^{\text{Predictor}} = (\boldsymbol{f}_{1,t}^{\text{Predictor}}, \ldots, \boldsymbol{f}_{r_2,t}^{\text{Predictor}})' \in \mathbb{R}^{r_2\times P}$ represents r_2 *predictor factors* across the N variables (rows) of the predictor matrix $\boldsymbol{X}_t$, where $\boldsymbol{f}_{j,t}^{\text{Predictor}} = \sum_{i=1}^N (\boldsymbol{U}_2)_{ij}\boldsymbol{x}_{it}$ is the jth predictor factor, for $j = 1, \ldots, r_2$, with $\boldsymbol{x}_{it} = (y_{i,t-1}, \ldots, y_{i,t-P})'$ for $1 \le i \le N$. Hence, if $(\boldsymbol{U}_2)_{ij} = 0$, then $\boldsymbol{x}_{it}$ is irrelevant to $\boldsymbol{f}_{j,t}^{\text{Predictor}}$. On the other hand, $\boldsymbol{U}_3'\boldsymbol{X}_t' := \boldsymbol{F}_t^{\text{Lag}} = (\boldsymbol{f}_{1,t}^{\text{Lag}}, \ldots, \boldsymbol{f}_{r_3,t}^{\text{Lag}})' \in \mathbb{R}^{r_3\times N}$ represents r_3 *temporal factors* across the P time lags (columns) of the predictor matrix $\boldsymbol{X}_t$, where $\boldsymbol{f}_{j,t}^{\text{Lag}} = \sum_{i=1}^P (\boldsymbol{U}_3)_{ij}\boldsymbol{y}_{t-i}$ is the jth temporal factor, for $j = 1, \ldots, r_3$. As a result, $(\boldsymbol{U}_3)_{ij} = 0$ implies that the ith lagged predictor $\boldsymbol{y}_{t-i}$ is irrelevant to $\boldsymbol{f}_{j,t}^{\text{Lag}}$. Therefore, $\boldsymbol{U}_2$ and $\boldsymbol{U}_3$ can be interpreted as the loadings of the predictor and temporal factors, respectively.

For simplicity, we call r_1, r_2 and r_3 the response, predictor and temporal ranks, respectively. Similar formulations can be found in matrix variate regressions (e.g., Zhao and Leng 2014; Ding and Cook 2018). The response, predictor and temporal factors interpretations of (7) reveal that the proposed model is related to factor modeling, one of the most widely used techniques for high-dimensional time series. We will explore the similarities and differences between them in the next subsection.

2.3. Connections With Factor Modeling for Time Series

In the literature, low-rank structures of high-dimensional time series are commonly explored through factor models (Stock and Watson 2005; Bai and Ng 2008; Stock and Watson 2011; Bai and Wang 2016). The multilinear low-rank assumption of $\mathcal{A}$ in the proposed model fulfills a similar purpose as it extracts dynamic factors along three dimensions, as shown in our discussion about (7). Meanwhile, the proposed model can be used directly for forecasting, which is another attractive feature compared to factor models. In the following, we take a closer look at the factor structures of the proposed model and both SFM and DFM in the literature, and discuss some interesting connections between them.

The SFM is commonly written as

$$\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t, \tag{8}$$

where $\boldsymbol{y}_t \in \mathbb{R}^N$ is the observed time series, $\boldsymbol{f}_t \in \mathbb{R}^r$ are r latent factors with $r \ll N$, $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times r}$ is the factor loading matrix, and $\boldsymbol{e}_t \in \mathbb{R}^N$ is the random error. The usual normalization restrictions require that $\boldsymbol{F}'\boldsymbol{F}/T = \boldsymbol{I}_r$ and that $\boldsymbol{\Lambda}'\boldsymbol{\Lambda} \in \mathbb{R}^{r\times r}$ is a full-rank diagonal matrix, where $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)'$; see Bai and Wang (2016).

We can show that the proposed model in (6) has an SFM representation. Specifically, as shown in Section E of the supplementary materials, there exist $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times r_1}$ and $\boldsymbol{f}_t \in \mathbb{R}^{r_1}$ such that

$$\boldsymbol{y}_t = \boldsymbol{U}_1\mathcal{G}_{(1)}(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'\boldsymbol{x}_t + \boldsymbol{\epsilon}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t, \tag{9}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\Lambda}$ and the resulting $\boldsymbol{F}$ satisfy the aforementioned normalization restrictions, and $\boldsymbol{f}_t$ is the normalized version of $\mathcal{G}_{(1)}(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'\boldsymbol{x}_t \in \mathbb{R}^{r_1}$. Let span$(\cdot)$ denote the column space of a matrix, and it can be verified that span$(\boldsymbol{\Lambda})$ = span$(\boldsymbol{U}_1)$.

Remark 3. Consider $\{\boldsymbol{y}_t\}$ generated by the proposed model. A useful by-product of representation (9) is that the low-dimensional subspace span$(\boldsymbol{U}_1)$ can actually be estimated by span$(\widehat{\boldsymbol{\Lambda}})$, where $\widehat{\boldsymbol{\Lambda}}$ is the estimator of $\boldsymbol{\Lambda}$ obtained by fitting

an SFM with $r = r_1$. Moreover, let $\mathbf{\Lambda}_1 = \mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Lambda})^{-1/2}$ be the orthonormalization of $\mathbf{\Lambda}$. Then $\text{span}(\mathbf{\Lambda}_1) = \text{span}(\mathbf{\Lambda}) = \text{span}(\boldsymbol{U}_1)$, and their orthogonal projectors are identical, namely $\mathbf{\Lambda}_1\mathbf{\Lambda}_1' = \boldsymbol{U}_1\boldsymbol{U}_1'$. Thus, the estimation error of $\text{span}(\boldsymbol{U}_1)$ can be measured by the commonly used subspace distance $\|\widehat{\mathbf{\Lambda}}_1\widehat{\mathbf{\Lambda}}_1' - \boldsymbol{U}_1\boldsymbol{U}_1'\|_{\text{F}}^2$, where $\widehat{\mathbf{\Lambda}}_1 = \widehat{\mathbf{\Lambda}}(\widehat{\mathbf{\Lambda}}'\widehat{\mathbf{\Lambda}})^{-1/2}$; see Vu and Lei (2013).

On the other hand, the DFM can be defined by combining model (8) with a certain dynamic structure, for example, the VAR, for the latent factor process $\boldsymbol{f}_t$ (Amengual and Watson 2007). To fix ideas, suppose that $\boldsymbol{f}_t$ evolves as the VAR(1),

$$\boldsymbol{f}_t = \boldsymbol{B}\boldsymbol{f}_{t-1} + \boldsymbol{\xi}_t, \tag{10}$$

where $\boldsymbol{B} \in \mathbb{R}^{r\times r}$ is the transition matrix, and $\boldsymbol{\xi}_t \in \mathbb{R}^r$ is the random error. Let $\boldsymbol{w}_t = \mathbf{\Lambda}\boldsymbol{f}_t$ and $\boldsymbol{u}_t = \mathbf{\Lambda}\boldsymbol{\xi}_t$. Then, the conjunction of (8) and (10) can also be written as

$$\boldsymbol{y}_t = \boldsymbol{w}_t + \boldsymbol{e}_t, \quad \boldsymbol{w}_t = \boldsymbol{V}\boldsymbol{C}\boldsymbol{V}'\boldsymbol{w}_{t-1} + \boldsymbol{u}_t, \tag{11}$$

where $\boldsymbol{D} = \mathbf{\Lambda}'\mathbf{\Lambda}$ is diagonal, $\boldsymbol{V} = \mathbf{\Lambda}\boldsymbol{D}^{-1/2}$ is orthonormal, and $\boldsymbol{C} = \boldsymbol{D}^{1/2}\boldsymbol{B}\boldsymbol{D}^{-1/2} \in \mathbb{R}^{r\times r}$. Interestingly, (11) resembles the VAR with measurement error, where $\boldsymbol{y}_t$ is the observed outcome of the true VAR(1) process $\boldsymbol{w}_t$ subject to measurement error $\boldsymbol{e}_t$. Note that the naive estimation ignoring the measurement error of the autoregressive process will result in asymptotic biases (see, e.g., Staudenmayer and Buonaccorsi 2005).

However, if $\boldsymbol{e}_t = 0$, we may gain more insights by comparing the DFM in (11) to the proposed model of lag order one. Note that when $P = 1$ the latter reduces to the reduced-rank VAR,

$$\boldsymbol{y}_t = \boldsymbol{U}_1\mathcal{G}_{(1)}\boldsymbol{U}_2'\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t \quad \text{or} \quad \boldsymbol{U}_1'\boldsymbol{y}_t = \mathcal{G}_{(1)}\boldsymbol{U}_2'\boldsymbol{y}_{t-1} + \boldsymbol{U}_1'\boldsymbol{\epsilon}_t,$$

with $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ being orthonormal and $r_1 = r_2$, while the DFM model in (11) with $\boldsymbol{e}_t = 0$ has the form of

$$\boldsymbol{V}'\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{V}'\boldsymbol{y}_{t-1} + \boldsymbol{V}'\boldsymbol{u}_t.$$

Hence, we may argue that the proposed model is more flexible than the DFM in (9), as the former can accommodate different Two y_t and two y_{t-j}'s in the highlighted sentence need to be changed to bold format to be identical. It is also worth noting that when $P > 1$, another advantage of the proposed model is that it can capture the possible low-rank structure across time lags of the predictors; see (7) in the previous subsection. Lastly, we note that the proposed model may be extended along the line of the factor augmented VAR models (FAVAR) (Bernanke, Boivin, and Eliasz 2005) by incorporating known low-dimensional factors.

Remark 4. In contrast to the proposed model, the classical factor model in the general form of (8) is not specific to VAR models, since it allows for general latent factors. However, the general factor model in (8) cannot be directly used for forecasting unless an additional dynamic structure is imposed on the latent factor process, for example, (10). As discussed above, if the multilinear low-rank assumption holds, the proposed model can be more favorable than the DFM.

3. Low-Dimensional Time Series Modeling

3.1. Multilinear Low-Rank Least Squares Estimation

For the multilinear low-rank VAR model in (5) with ranks (r_1, r_2, r_3), the multilinear low-rank (MLR) least squares estimator can be defined as

$$\widehat{\mathcal{A}}_{\text{MLR}} \equiv [\![\widehat{\mathcal{G}}; \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3]\!] = \arg\min L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3), \tag{12}$$

where

$$L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3) = \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{y}_t - (\mathcal{G}\times_1\boldsymbol{U}_1\times_2\boldsymbol{U}_2\times_3\boldsymbol{U}_3)_{(1)}\boldsymbol{x}_t\|_2^2. \tag{13}$$

We will derive asymptotic properties of $\widehat{\mathcal{A}}_{\text{MLR}}$ when both N and P are fixed and the true multilinear ranks (r_1, r_2, r_3) are known. Note that the minimization in (12) is unconstrained, so the Tucker decomposition $[\![\widehat{\mathcal{G}}; \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3]\!]$ of $\widehat{\mathcal{A}}_{\text{MLR}}$ is not unique.

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

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

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

$$\begin{aligned} \boldsymbol{H} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\phi}} = \Big(& (\boldsymbol{U}_3\otimes\boldsymbol{U}_2\otimes\boldsymbol{U}_1), \\ & [(\boldsymbol{U}_3\otimes\boldsymbol{U}_2)\mathcal{G}_{(1)}'] \otimes \boldsymbol{I}_N, \boldsymbol{T}_{21}\{[(\boldsymbol{U}_1\otimes\boldsymbol{U}_3)\mathcal{G}_{(2)}'] \otimes \boldsymbol{I}_N\}, \\ & \boldsymbol{T}_{31}\{[(\boldsymbol{U}_1\otimes\boldsymbol{U}_2)\mathcal{G}_{(3)}'] \otimes \boldsymbol{I}_P\}\Big), \end{aligned} \tag{14}$$

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

Theorem 1. Suppose that the time series $\{\boldsymbol{y}_t\}$ is generated by model (5) with $\mathbb{E}\|\boldsymbol{\epsilon}_t\|_2^4 < \infty$, both N and P are fixed, and (r_1, r_2, r_3) are known. Then, under Assumption 1,

$$\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{MLR}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\} \to N(\boldsymbol{0}, \mathbf{\Sigma}_{\text{MLR}}) \tag{15}$$

in distribution as $T \to \infty$, where $\mathbf{\Sigma}_{\text{MLR}} = \boldsymbol{H}(\boldsymbol{H}'\boldsymbol{J}\boldsymbol{H})^{\dagger}\boldsymbol{H}'$, 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). It does not require that $\mathcal{G}$ and $\boldsymbol{U}_j$'s are identifiable, nor does it require imposing identification constraints on the estimation in (12).

However, if we are further interested in estimating the true components $\mathcal{G}$ and $\boldsymbol{U}_j$'s in the HOSVD of $\mathcal{A}$, the identifiability of these components, that is, the uniqueness of the HOSVD, will be required. This is guaranteed by the following assumption.

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

In Assumption 2, Condition (i) avoids indeterminacy of the factor loading vectors and holds generally in practice. Condition (ii) rules out sign switches in $\boldsymbol{U}_j$ and is commonly used in low-rank matrix models (Li et al. 2016).

Accordingly, based on the unconstrained estimator $\widehat{\mathcal{A}}_{\text{MLR}}$, we can define each $\widehat{\boldsymbol{U}}_j$ uniquely as the top r_j left singular vectors of $(\widehat{\mathcal{A}}_{\text{MLR}})_{(j)}$ such that the first element in each column of $\widehat{\boldsymbol{U}}_j$ is positive, and set $\widehat{\mathcal{G}} = [\![\widehat{\mathcal{A}}_{\text{MLR}}; \widehat{\boldsymbol{U}}_1', \widehat{\boldsymbol{U}}_2', \widehat{\boldsymbol{U}}_3']\!]$. As a result, the estimators $\widehat{\mathcal{G}}$ and $\widehat{\boldsymbol{U}}_j$'s are consistent and asymptotically normal.

Corollary 1. Suppose that the conditions of Theorem 1 and Assumption 2 hold. Then $\sqrt{T}\{\text{vec}(\widehat{\mathcal{G}}) - \text{vec}(\mathcal{G})\}$, $\sqrt{T}\{\text{vec}(\widehat{\boldsymbol{U}}_1) - \text{vec}(\boldsymbol{U}_1)\}$, $\sqrt{T}\{\text{vec}(\widehat{\boldsymbol{U}}_2) - \text{vec}(\boldsymbol{U}_2)\}$, and $\sqrt{T}\{\text{vec}(\widehat{\boldsymbol{U}}_3) - \text{vec}(\boldsymbol{U}_3)\}$ converge to normal distributions with mean zero as $T \to \infty$.

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

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

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

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

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

Corollary 2. Under the conditions of Theorem 1, $\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{OLS}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\} \to N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{OLS}})$ and $\sqrt{T}\{\text{vec}((\widehat{\mathcal{A}}_{\text{RRR}})_{(1)}) - \text{vec}(\mathcal{A}_{(1)})\} \to N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{RRR}})$ in distribution as $T \to \infty$. Moreover, it holds that $\boldsymbol{\Sigma}_{\text{MLR}} \le \boldsymbol{\Sigma}_{\text{RRR}} \le \boldsymbol{\Sigma}_{\text{OLS}}$.

3.2. Alternating Least Squares Algorithm

Let $\mathcal{F}_t = \sigma(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, \ldots)$ be the σ-field generated by $\{\boldsymbol{\epsilon}_s, s \le t\}$ and recall that $\boldsymbol{X}_t = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-P})$. The objective function in (12) is a nonlinear function of $\mathcal{G}$, $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$. However, from model (5), we have

$$\begin{aligned}
\mathbb{E}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) &= \left((\boldsymbol{x}_t'(\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)\mathcal{G}_{(1)}') \otimes \boldsymbol{I}_N\right)\text{vec}(\boldsymbol{U}_1) \\
&= \boldsymbol{U}_1\mathcal{G}_{(1)}((\boldsymbol{U}_3'\boldsymbol{X}_t') \otimes \boldsymbol{I}_{r_2})\text{vec}(\boldsymbol{U}_2') \\
&= \boldsymbol{U}_1\mathcal{G}_{(1)}(\boldsymbol{I}_{r_3} \otimes (\boldsymbol{U}_2'\boldsymbol{X}_t))\text{vec}(\boldsymbol{U}_3) \\
&= (((\boldsymbol{U}_3 \otimes \boldsymbol{U}_2)'\boldsymbol{x}_t)' \otimes \boldsymbol{U}_1)\text{vec}(\mathcal{G}_{(1)}),
\end{aligned} \tag{16}$$

which implies that the objective function in (12) is linear with respect to any of $\mathcal{G}$, $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$ when the other three are fixed.

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

Initialize: $\mathcal{A}^{(0)}$

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

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

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

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

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

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

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

until convergence

Finalize: $\widehat{\boldsymbol{U}}_i \leftarrow$ top r_i left singular vectors of $\widehat{\mathcal{A}}_{(i)}$ with positive first elements, $1 \le i \le 3$

$\widehat{\mathcal{G}} \leftarrow [\![\widehat{\mathcal{A}}; \widehat{\boldsymbol{U}}_1', \widehat{\boldsymbol{U}}_2', \widehat{\boldsymbol{U}}_3']\!]$

Given the multilinear ranks (r_1, r_2, r_3), we can employ Algorithm 1 to find $\widehat{\mathcal{A}}_{\text{MLR}}$. Note that this is an alternating least squares algorithm where each step has a closed-form solution. In practice, the multilinear ranks need to be selected consistently, and we relegate the details to Section 5. The following proposition gives the convergence property of Algorithm 1.

Proposition 1. Suppose that the stationary points of the objective function in (12) are isolated, up to an arbitrary nonsingular linear transformation. Then $\boldsymbol{\phi}^{(k)}$ converges to a stationary point as $k \to \infty$, where $\boldsymbol{\phi}^{(k)} = (\text{vec}(\mathcal{G}^{(k)})', \text{vec}(\boldsymbol{U}_1^{(k)})', \text{vec}(\boldsymbol{U}_2^{(k)})', \text{vec}(\boldsymbol{U}_3^{(k)})')'$. Moreover, let $\boldsymbol{\phi}^{(\infty)} = (\text{vec}(\mathcal{G}^{(\infty)})', \text{vec}(\boldsymbol{U}_1^{(\infty)})', \text{vec}(\boldsymbol{U}_2^{(\infty)})', \text{vec}(\boldsymbol{U}_3^{(\infty)})')'$ be a strict local minimum of the objective function. Then $\{\boldsymbol{\phi}^{(k)}\}$ will be attracted to $\boldsymbol{\phi}^{(\infty)}$ if the initial value $\boldsymbol{\phi}^{(0)}$ is sufficiently close to $\boldsymbol{\phi}^{(\infty)}$.

Remark 5. If the sample size is sufficiently large, by Corollary 2, $\widehat{\mathcal{A}}_{\text{OLS}}$ can be used as the initial value $\mathcal{A}^{(0)}$ of Algorithm 1. For smaller sample sizes, $\widehat{\mathcal{A}}_{\text{RRR}}$ or the nuclear norm estimator to be discussed in Section 5 can be employed instead. Moreover, Algorithm 1 does not guarantee convergence to the global solution defined in (12). As a result, in practice, we recommend a random initialization method with $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{pre}} + T^{-1/2}\mathcal{T}$, where $\widehat{\mathcal{A}}_{\text{pre}}$ is a preliminary estimate, say, $\widehat{\mathcal{A}}_{\text{OLS}}$ or $\widehat{\mathcal{A}}_{\text{RRR}}$, and $\mathcal{T} \in \mathbb{R}^{N \times N \times P}$ is a random perturbation whose entries are drawn independently from $N(0, 1)$. Many randomized initial values can be tried, and the solution which yields the smallest value for the objective function will be adopted.

Remark 6. Algorithm 1 corresponds to the unconstrained estimation in (12). Thus, we do not need the orthogonality constraints of $\mathcal{G}$ and $\boldsymbol{U}_i$'s. The unidentifiability of the Tucker decomposition does not affect the convergence of the algorithm, since

Proposition 1 does not require that the convergent sequence $\boldsymbol{\phi}^{(k)}$ is unique. Moreover, note that the final estimates $\widehat{\mathcal{G}}$ and $\widehat{\boldsymbol{U}}_i$'s in Algorithm 1 are obtained from the unconstrained estimate of $\mathcal{A}$, which is consistent with the definitions of $\widehat{\mathcal{G}}$ and $\widehat{\boldsymbol{U}}_i$'s in Corollary 1. Similar alternating algorithms without imposing identification constraints can be found in the literature of tensor decomposition (see, e.g., Zhou, Li, and Zhu 2013; Li et al. 2018).

4. High-Dimensional Time Series Modeling

4.1. Sparse Higher-Order Reduced-Rank VAR

As discussed in Section 2.2, the proposed model can effectively capture the dynamic information along three dimensions by response, predictor, and temporal factors, with $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$ representing the corresponding factor loadings. However, when the dimensions N and/or P are very large, the fitted loading matrices often contain many small values, indicating relatively insignificant contribution of certain variables or lags to the factors. For example, if the (i,j)th entry of $\boldsymbol{U}_1$ is very small, then y_{it} may be irrelevant to the jth response factor, with $1 \le i \le N$ and $1 \le j \le r_1$; see also the discussion below (7) for similar interpretations regarding $\boldsymbol{U}_2$ and $\boldsymbol{U}_3$.

To improve the interpretability, we may shrink the small values in the factor loading matrices to zero by imposing sparsity assumptions on $\boldsymbol{U}_i$'s. This allows us to substantially reduce the number of unknown parameters while performing data-driven variable selection for each factor, and hence the estimation efficiency is also improved; see Chen, Chan, and Stenseth (2012) and Uematsu et al. (2019).

Specifically, we introduce the following ℓ_1-penalized sparse higher-order reduced-rank (SHORR) estimator:

$$\widehat{\mathcal{A}}_{\text{SHORR}} \equiv [\![\widehat{\mathcal{G}}; \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3]\!] = \underset{\mathcal{G},\boldsymbol{U}_1,\boldsymbol{U}_2,\boldsymbol{U}_3}{\arg\min} \big\{ L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3) + \lambda \|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1 \big\} \tag{17}$$

subject to

$$\mathcal{G} \in \text{AO}(r_1, r_2, r_3) \quad \text{and} \quad \boldsymbol{U}_i'\boldsymbol{U}_i = \boldsymbol{I}_{r_i}, \quad i = 1, 2, 3, \tag{18}$$

where $L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3)$ is defined as in (13), and $\text{AO}(r_1, r_2, r_3) = \{\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \mathcal{G}_{(i)}$ is row-orthogonal, $i = 1, 2, 3\}$. Unlike the unconstrained estimation in (12), the orthogonality constraints in (18) are necessary; otherwise, the sparsity patterns of $\boldsymbol{U}_i$ cannot be identified. As in Section 3, we will derive the statistical properties of the proposed estimator under the true multilinear ranks (r_1, r_2, r_3), while a consistent rank selection procedure will be discussed in Section 5.

Remark 7. The proposed SHORR estimation method is different from the row-sparse reduced-rank regression that has been studied extensively in the literature (Bunea, She, and Wegkamp 2012; Chen and Huang 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 $\boldsymbol{U}_1$, those unselected time series cannot be predicted at all. Thus, we consider the general sparsity structure for $\boldsymbol{U}_i$'s rather than the row-sparsity.

Remark 8. Alternatively, one might consider penalizing each $\boldsymbol{U}_i$ individually, for example, with the penalty term $\sum_{i=1}^{3} \lambda_i \|\boldsymbol{U}_i\|_1$. Unfortunately, the three tuning parameters will bring about much higher computational costs and significant theoretical difficulties. To circumvent this problem, the SHORR estimator induces sparsity for $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$ jointly since $\|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1 = \|\boldsymbol{U}_3\|_1 \|\boldsymbol{U}_2\|_1 \|\boldsymbol{U}_1\|_1$. Implementation of this joint penalty is convenient through the alternating algorithm to be introduced in Section 4.3. Similar ideas of joint penalization can be found in the literature, for example, the joint Lasso penalty in Zhao and Leng (2014) and the joint penalty for left and right singular vectors for sparse SVD in Chen, Chan, and Stenseth (2012). Moreover, when P is relatively small, we might wish to impose sparsity on $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ only, and then $\|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1$ can be replaced by $\|\boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1$.

4.2. Theoretical Properties of the SHORR Estimator

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

Assumption 3 (Gaussian error). The errors $\{\boldsymbol{\epsilon}_t\}$ are iid Gaussian random vectors with mean zero and positive definite covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

Assumption 4 (Sparsity). Each column of the factor matrices $\boldsymbol{U}_i$ has at most s_i nonzero entries, for $i = 1, 2, 3$.

Assumption 5 (Restricted parameter space). The parameter space for $\mathcal{G}$ and $\boldsymbol{U}_i$ with $1 \le i \le 3$ is $\Omega = \{\mathcal{G} \in \text{AO}(r_1, r_2, r_3) : \sigma_1(\mathcal{G}_{(j)}) \le \bar{g} < \infty$, for $1 \le j \le 3\} \times \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{U}_3$, where $\mathcal{U}_i = \{\boldsymbol{U} \in \mathbb{R}^{p_i \times r_i} : \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_{r_i}$, and $\boldsymbol{U}_{ij}^2 \ge \nu > 0$ or $\boldsymbol{U}_{ij} = 0\}$ with $p_1 = p_2 = N$ and $p_3 = P$, and ν is a uniform lower threshold for elements of $\boldsymbol{U}_i$'s.

Assumption 6 (Relative spectral gap). The nonzero singular values of $\mathcal{A}_{(i)}$ satisfy that $\sigma_{j-1}^2(\mathcal{A}_{(i)}) - \sigma_j^2(\mathcal{A}_{(i)}) \ge \delta \sigma_{j-1}^2(\mathcal{A}_{(i)})$ for $2 \le j \le r_i$ and $1 \le i \le 3$, 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 (2019). Assumption 4 states the sparsity of each factor matrix. Assumption 5 imposes an upper bound on the core tensor $\mathcal{G}$, which is not a stringent assumption since large singular values in $\mathcal{G}$ could cause nonstationarity of the VAR process. The lower threshold ν for the $\boldsymbol{U}_i$'s is essential to restrict the estimation error to a subspace such that the restricted eigenvalue condition (Bickel, Ritov, and Tsybakov 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 $\mathcal{A}_{(i)}$ are well separated. This rules out unidentifiability and allows us to derive the upper bound for the perturbation errors in Lemma 1 in Section D of the supplementary materials.

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

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

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

$$\mu_{\min}(\mathcal{A}) = \min_{\theta\in[-\pi,\pi]} \lambda_{\min}\left(\left(\boldsymbol{I}_N - \sum_{p=1}^{P} \boldsymbol{A}_p' e^{ip\theta}\right) \times \left(\boldsymbol{I}_N - \sum_{p=1}^{P} \boldsymbol{A}_p' e^{-ip\theta}\right)\right) \tag{19}$$

and

$$\mu_{\max}(\mathcal{A}) = \max_{\theta\in[-\pi,\pi]} \lambda_{\max}\left(\left(\boldsymbol{I}_N - \sum_{p=1}^{P} \boldsymbol{A}_p' e^{ip\theta}\right) \times \left(\boldsymbol{I}_N - \sum_{p=1}^{P} \boldsymbol{A}_p' e^{-ip\theta}\right)\right). \tag{20}$$

Theorem 2. Suppose that Assumptions 1 and 3–6 hold, and (r_1, r_2, r_3) are known. If $\lambda \gtrsim \mathcal{M}\sqrt{\log(N^2P)/T}$ and $T \gtrsim \log(N^2P) + \mathcal{M}^2 d \min[\log(NP), \log(cNP/d)]$, then

$$\|\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}\|_{\text{F}} \le C_1 \tau \sqrt{S}\lambda/\alpha, \tag{21}$$

and

$$T^{-1}\sum_{t=1}^{T} \|(\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A})_{(1)}\boldsymbol{x}_t\|_2^2 \le C_2 \tau^2 S\lambda^2/\alpha, \tag{22}$$

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, $\mathcal{M} = \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})(1 + \mu_{\max}(\mathcal{A})/\mu_{\min}(\mathcal{A}))$, $d = \nu^{-2} r_1 r_2 r_3$, $\tau = \delta^{-1} r_1 r_2 r_3 \sum_{i=1}^{3} \eta_i/\sqrt{r_i}$ with $\eta_i = (\sum_{j=1}^{r_i} \sigma_1^2(\mathcal{A}_{(i)})/\sigma_j^2(\mathcal{A}_{(i)}))^{1/2}$, $S = s_1 s_2 s_3$ and $\alpha = \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})/\mu_{\max}(\mathcal{A})$.

Theorem 2 gives the nonasymptotic error upper bounds under high-dimensional scaling. When the multilinear ranks (r_1, r_2, r_3) and lower threshold ν are fixed, (21) shows that $\widehat{\mathcal{A}}_{\text{SHORR}}$ is a consistent estimator if $T \gtrsim S\log(N^2P)$. In this setting, the estimation and prediction error bounds in (21) and (22) become $O_p(\sqrt{S\log(N^2P)/T})$ and $O_p(S\log(N^2P)/T)$, respectively.

Remark 9. Basu and Michailidis (2015) considered estimation of stationary Gaussian VAR(P) models with sparse transition matrices such that $\|\mathcal{A}\|_0 = k$. For the Lasso estimator $\widehat{\mathcal{A}}_{\text{LASSO}} = \arg\min T^{-1}\sum_{t=1}^{T} \|\boldsymbol{y}_t - \mathcal{A}_{(1)}\boldsymbol{x}_t\|_2^2 + \lambda\|\mathcal{A}_{(1)}\|_1$, it was shown that $\|\widehat{\mathcal{A}}_{\text{LASSO}} - \mathcal{A}\|_{\text{F}} \lesssim \sqrt{k\log(N^2P)/T}$ and $T^{-1}\sum_{t=1}^{T} \|(\widehat{\mathcal{A}}_{\text{LASSO}} - \mathcal{A})_{(1)}\boldsymbol{x}_t\|_2^2 \lesssim k\log(N^2P)/T$ with high probability, which are consistent with the regular error bounds for the Lasso as N^2P corresponds to the number of parameters (e.g., Wang et al. 2015). In contrast, we assume that $\mathcal{A}$ admits an HOSVD with sparse factor matrices $\boldsymbol{U}_i$, but $\mathcal{A}$ itself is not necessarily sparse. When each $\boldsymbol{U}_i$ is row-sparse with s_i nonzero rows, for $i = 1, 2, 3$, it can be checked that $\mathcal{A}$ is a sparse tensor with sparsity level S. In this case, the SHORR estimator has the same error bounds as the Lasso estimator. However, in the general case, even when $\mathcal{A}$ has a sparse HOSVD, $\mathcal{A}$ may not be highly sparse, that is, k is larger than S, so $\widehat{\mathcal{A}}_{\text{SHORR}}$ may be more efficient than $\widehat{\mathcal{A}}_{\text{LASSO}}$.

4.3. ADMM Algorithm

There are two major challenges in developing an efficient algorithm for the SHORR estimator. First, the core tensor $\mathcal{G}$ is subject to the all-orthogonal constraint in (18) which cannot be handled in a straightforward way. Second, the ℓ_1-regularization in (17) and the orthogonality constraints in (18) are imposed jointly on $\boldsymbol{U}_i$'s. The former is nonsmooth while the latter is nonconvex. To deal with these challenges, we adopt the alternating direction method of multipliers (ADMM) algorithm (Boyd et al. 2011) to update $\boldsymbol{U}_i$'s and $\mathcal{G}$ alternatingly; see Algorithm 2.

First, to tackle the all-orthogonal constraint of $\mathcal{G}$, our idea is to separate it into three orthogonality constraints on the matricizations $\mathcal{G}_{(i)}$ for $1 \le i \le 3$. This is to say that $\mathcal{G}_{(i)}$ can be decomposed as $\mathcal{G}_{(i)} = \boldsymbol{D}_i \boldsymbol{V}_i'$, where $\boldsymbol{D}_i \in \mathbb{R}^{r_i \times r_i}$ is a diagonal matrix, and $\boldsymbol{V}_1 \in \mathbb{R}^{r_2 r_3 \times r_1}$, $\boldsymbol{V}_2 \in \mathbb{R}^{r_1 r_3 \times r_2}$, and $\boldsymbol{V}_3 \in \mathbb{R}^{r_1 r_2 \times r_3}$ are orthonormal matrices with $\boldsymbol{V}_i'\boldsymbol{V}_i = \boldsymbol{I}_{r_i}$. Then, the augmented Lagrangian corresponding to the objective function in (17) can be written as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\varrho}}(\mathcal{G}, \{\boldsymbol{U}_i\}, \{\boldsymbol{D}_i\}, \{\boldsymbol{V}_i\}; \{\mathcal{C}_i\}) &= L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3) + \lambda\|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1 \\ &\quad + 2\sum_{i=1}^{3} \varrho_i \langle (\mathcal{C}_i)_{(i)}, \mathcal{G}_{(i)} - \boldsymbol{D}_i \boldsymbol{V}_i' \rangle + \sum_{i=1}^{3} \varrho_i \|\mathcal{G}_{(i)} - \boldsymbol{D}_i \boldsymbol{V}_i'\|_{\text{F}}^2,\end{aligned}$$

where $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ are the tensor-valued dual variables, and $\boldsymbol{\varrho} = (\varrho_1, \varrho_2, \varrho_3)'$ is the set of regularization parameters. This leads us to Algorithm 2. Note that all-orthogonal constraint of $\mathcal{G}$ has been transferred to the matrices $\boldsymbol{V}_i$'s in line 10, so no constraint is needed for updating $\mathcal{G}$ in line 7 of Algorithm 2.

Second, we consider the update of $\boldsymbol{U}_i$'s. Since $L(\mathcal{G}, \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3)$ in (13) is a least squares loss function with respect to each $\boldsymbol{U}_i$, the $\boldsymbol{U}_i$-update steps in lines 4–6 of Algorithm 2 are ℓ_1-regularized least squares problems subject to an orthogonality constraint, which can be written in the general form:

$$\min_{B} \left\{n^{-1}\|\boldsymbol{y} - \boldsymbol{X}\text{vec}(\boldsymbol{B})\|_2^2 + \lambda\|\boldsymbol{B}\|_1\right\}, \text{ s.t. } \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}. \tag{23}$$

Since the ℓ_1-regularization and the orthogonality constraint for $\boldsymbol{B}$ are difficult to handle jointly, we adopt an ADMM subroutine to separate them into two steps. Specifically, we introduce the dummy variable $\boldsymbol{W}$ as a surrogate for $\boldsymbol{B}$ and write problem (23) into the equivalent form as follows:

$$\min_{\boldsymbol{B},\boldsymbol{W}} \{n^{-1}\|\boldsymbol{y} - \boldsymbol{X}\text{vec}(\boldsymbol{B})\|_2^2 + \lambda\|\boldsymbol{W}\|_1\}, \text{ s.t. } \boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I} \text{ and } \boldsymbol{B} = \boldsymbol{W}. \tag{24}$$

Then the corresponding augmented Lagrangian formulation is

$$\min_{\boldsymbol{B},\boldsymbol{W}} \{n^{-1}\|\boldsymbol{y} - \boldsymbol{X}\text{vec}(\boldsymbol{B})\|_2^2 + \lambda\|\boldsymbol{W}\|_1 + 2\kappa\langle \boldsymbol{M}, \boldsymbol{B} - \boldsymbol{W}\rangle + \kappa\|\boldsymbol{B} - \boldsymbol{W}\|_{\text{F}}^2\}, \tag{25}$$

Algorithm 2 ADMM algorithm for SHORR estimator

1: Initialize: $\mathcal{A}^{(0)}$
2: HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \boldsymbol{U}_1^{(0)} \times_2 \boldsymbol{U}_2^{(0)} \times_3 \boldsymbol{U}_3^{(0)}$ with multilinear ranks (r_1, r_2, r_3).
3: **repeat**
4: $\quad \boldsymbol{U}_1^{(k+1)} \leftarrow \underset{\boldsymbol{U}_1'\boldsymbol{U}_1=\boldsymbol{I}_{r_1}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1, \boldsymbol{U}_2^{(k)}, \boldsymbol{U}_3^{(k)}) + \lambda \|\boldsymbol{U}_1\|_1 \|\boldsymbol{U}_2^{(k)}\|_1 \|\boldsymbol{U}_3^{(k)}\|_1 \right\}$
5: $\quad \boldsymbol{U}_2^{(k+1)} \leftarrow \underset{\boldsymbol{U}_2'\boldsymbol{U}_2=\boldsymbol{I}_{r_2}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2, \boldsymbol{U}_3^{(k)}) + \lambda \|\boldsymbol{U}_1^{(k+1)}\|_1 \|\boldsymbol{U}_2\|_1 \|\boldsymbol{U}_3^{(k)}\|_1 \right\}$
6: $\quad \boldsymbol{U}_3^{(k+1)} \leftarrow \underset{\boldsymbol{U}_3'\boldsymbol{U}_3=\boldsymbol{I}_{r_3}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2^{(k+1)}, \boldsymbol{U}_3) + \lambda \|\boldsymbol{U}_1^{(k+1)}\|_1 \|\boldsymbol{U}_2^{(k+1)}\|_1 \|\boldsymbol{U}_3\|_1 \right\}$
7: $\quad \mathcal{G}^{(k+1)} \leftarrow \arg\min \left\{ L(\mathcal{G}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2^{(k+1)}, \boldsymbol{U}_3^{(k+1)}) + \sum_{i=1}^3 \varrho_i \|\mathcal{G}_{(i)} - \boldsymbol{D}_i^{(k)} \boldsymbol{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2 \right\}$
8: $\quad$ **for** $i \in \{1, 2, 3\}$ **do**
9: $\quad\quad \boldsymbol{D}_i^{(k+1)} \leftarrow \underset{\boldsymbol{D}_i=\mathrm{diag}(\boldsymbol{d}_i)}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i \boldsymbol{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2$
10: $\quad\quad \boldsymbol{V}_i^{(k+1)} \leftarrow \underset{\boldsymbol{V}_i'\boldsymbol{V}_i=\boldsymbol{I}_{r_i}}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i^{(k+1)} \boldsymbol{V}_i' + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2$
11: $\quad\quad (\mathcal{C}_i^{(k+1)})_{(i)} \leftarrow (\mathcal{C}_i^{(k)})_{(i)} + \mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i^{(k+1)} \boldsymbol{V}_i^{(k+1)\prime}$
12: $\quad$ **end for**
13: $\quad \mathcal{A}^{(k+1)} \leftarrow \mathcal{G}^{(k+1)} \times_1 \boldsymbol{U}_1^{(k+1)} \times_2 \boldsymbol{U}_2^{(k+1)} \times_3 \boldsymbol{U}_3^{(k+1)}$
14: **until convergence**

Algorithm 3 ADMM subroutine for sparse and orthogonal regression

1: Initialize: $\boldsymbol{B}^{(0)} = \boldsymbol{W}^{(0)}, \boldsymbol{M}^{(0)} = \boldsymbol{0}$
2: **repeat**
3: $\quad \boldsymbol{B}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{B}'\boldsymbol{B}=\boldsymbol{I}} \left\{ n^{-1} \|\boldsymbol{y} - \boldsymbol{X}\mathrm{vec}(\boldsymbol{B})\|_2^2 + \kappa \|\boldsymbol{B} - \boldsymbol{W}^{(k)} + \boldsymbol{M}^{(k)}\|_{\mathrm{F}}^2 \right\}$
4: $\quad \boldsymbol{W}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{W}} \left\{ \kappa \|\boldsymbol{B}^{(k+1)} - \boldsymbol{W} + \boldsymbol{M}^{(k)}\|_{\mathrm{F}}^2 + \lambda \|\boldsymbol{W}\|_1 \right\}$
5: $\quad \boldsymbol{M}^{(k+1)} \leftarrow \boldsymbol{M}^{(k)} + \boldsymbol{B}^{(k+1)} - \boldsymbol{W}^{(k+1)}$
6: **until convergence**

where $\boldsymbol{M}$ is the dual variable, and κ is a regularization parameter. The ADMM subroutine for (25) is presented in Algorithm 3. This yields solutions to the $\boldsymbol{U}_i$-update subproblems in Algorithm 2.

Note that the $\boldsymbol{B}$-update step in Algorithm 3 and the $\boldsymbol{V}_i$-update step in line 10 of Algorithm 2 are least squares problems with an orthogonality constraint. Hence, they can be solved efficiently by the splitting orthogonality constraint (SOC) method (Lai and Osher 2014). The $\boldsymbol{W}$-update step in Algorithm 3 is an ℓ_1-regularized minimization, which can be solved by the explicit soft-thresholding. The $\mathcal{G}$- and $\boldsymbol{D}_i$-update steps in lines 7 and 9 of Algorithm 2 are simple least squares problems.

For general nonconvex problems, it is well known that ADMM algorithms need not converge, and even if they do, they need not converge to an optimal solution. A comprehensive algorithmic convergence analysis for Algorithm 2 is challenging due to both the nested ADMM subroutine, Algorithm 3, and its interplay with the outer loop of Algorithm 2.

Wang, Yin, and Zeng (2019) gave a rigorous convergence analysis of multi-block ADMMs for nonconvex nonsmooth optimization with linear equality constraints. Their theory would be applicable to Algorithm 3 if the $\boldsymbol{B}$-update step in line 3 were exact. The extension to the inexact $\boldsymbol{B}$-update step would require a sophisticated analysis of the optimization error of the SOC method. We do not delve into the development of the convergence theory further in this article. Nonetheless, similarly to the analysis in Uematsu et al. (2019), under some high-level assumptions on $\mathcal{L}_{\boldsymbol{\varrho}}(\cdot)$, we can still obtain the following convergence result for Algorithm 2.

Proposition 2. Let $\Delta\mathcal{L}_{\boldsymbol{\varrho}}(\cdot)$ be the decrease in the augmented Lagrangian $\mathcal{L}_{\boldsymbol{\varrho}}(\cdot)$ by a block update. If $\sum_{k=1}^{\infty}\{[\Delta\mathcal{L}_{\boldsymbol{\varrho}}(\mathcal{G}^{(k)})]^{1/2} + \sum_{i=1}^3 [\Delta\mathcal{L}_{\boldsymbol{\varrho}}(\boldsymbol{U}_i^{(k)})]^{1/2} + \sum_{i=1}^3 [\Delta\mathcal{L}_{\boldsymbol{\varrho}}(\boldsymbol{D}_i^{(k)})]^{1/2} + \sum_{i=1}^3 [\Delta\mathcal{L}_{\boldsymbol{\varrho}}(\boldsymbol{V}_i^{(k)})]^{1/2}\} < \infty$, then the sequence generated by Algorithm 2 converges to a local solution of problem (17).

Remark 10. The initial value $\mathcal{A}^{(0)}$ for Algorithm 2 can be set to the nuclear norm (NN) estimator $\widehat{\mathcal{A}}_{\mathrm{NN}}$ for low-rank VAR models (Negahban and Wainwright 2011), and it holds $\|\widehat{\mathcal{A}}_{\mathrm{NN}} - \mathcal{A}\|_{\mathrm{F}} = O_p(\sqrt{r_1 NP/T})$; see also Section 5. Consequently, if one searches the SHORR estimator within a neighborhood of $\widehat{\mathcal{A}}_{\mathrm{NN}}$ of radius $O(\sqrt{r_1 NP/T})$, then all iterates $\mathcal{A}^{(k)}$ will satisfy $\|\mathcal{A}^{(k)} - \mathcal{A}\|_{\mathrm{F}} \le \|\mathcal{A}^{(k)} - \widehat{\mathcal{A}}_{\mathrm{NN}}\|_{\mathrm{F}} + \|\widehat{\mathcal{A}}_{\mathrm{NN}} - \mathcal{A}\|_{\mathrm{F}} = O_p(\sqrt{r_1 NP/T})$. Additionally, Theorem 2 implies $\|\mathcal{A}^{(k)} - \widehat{\mathcal{A}}_{\mathrm{SHORR}}\|_{\mathrm{F}} \le \|\mathcal{A}^{(k)} - \mathcal{A}\|_{\mathrm{F}} + \|\widehat{\mathcal{A}}_{\mathrm{SHORR}} - \mathcal{A}\|_{\mathrm{F}} = O_p(\sqrt{r_1 NP/T})$, where $\widehat{\mathcal{A}}_{\mathrm{SHORR}}$ is the global solution. A similar convex relaxation based initialization approach is used by Uematsu et al. (2019) for a nonconvex optimization problem with jointly imposed sparsity and orthogonality constraints. Moreover, since Algorithm 2

and Proposition 2 do not guarantee the convergence to a global solution, similarly to the random initialization method in Remark 5, in practice we can try many randomized initial values $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{NN}} + (NP/T)^{1/2}\mathcal{T}$, where the entries of the perturbation $\mathcal{T} \in \mathbb{R}^{N\times N\times P}$ are drawn independently from $N(0, (N^2P)^{-1})$ such that $\|\mathcal{T}\|_{\text{F}} = O_p(1)$, and then select the final solution as the one with the smallest value for the objective function.

Remark 11. The above algorithms are presented under known multilinear ranks and a fixed tuning parameter λ. In practice, to save computational costs, we recommend a two-step procedure: first select the ranks by the method to be introduced in Section 5, and then fixing these rank, select the tuning parameter λ by a fine grid search with information criterion such as the BIC or its high-dimensional extensions. Although the degrees of freedom in a sparse and orthogonal matrix are unclear, the total number of nonzero elements in $\mathcal{G}$, $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$ could be used as proxies.

5. Rank Selection

The theoretical results we derived for MLR and SHORR estimators hinge on correct multilinear ranks. This section introduces a procedure for consistent rank selection.

Suppose that $\widehat{\mathcal{A}}$ is a consistent initial estimator of $\mathcal{A}$. We propose the following ridge-type ratio estimator (Xia, Xu, and Zhu 2015) to estimate the multilinear ranks,

$$\widehat{r}_i = \arg\min_{1\le j\le p_i-1} \frac{\sigma_{j+1}(\widehat{\mathcal{A}}_{(i)}) + c}{\sigma_j(\widehat{\mathcal{A}}_{(i)}) + c},$$

for $1 \le i \le 3$, where $p_1 = p_2 = N$, $p_3 = P$, and c is a parameter that needs to be well chosen; see the assumption below. Here we allow N, P and the multilinear ranks to diverge with T. For $i = 1, 2, 3$, denote

$$\varsigma_i = \frac{1}{\sigma_{r_i}(\mathcal{A}_{(i)})} \cdot \max_{1\le j<r_i} \frac{\sigma_j(\mathcal{A}_{(i)})}{\sigma_{j+1}(\mathcal{A}_{(i)})}.$$

Assumption 7. The parameter $c > 0$ is chosen such that (i) $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\text{F}}/c = o_p(1)$ and (ii) $c \max_{1\le i\le 3} \varsigma_i = o(1)$.

Remark 12. In Assumption 7, Condition (i) states that the estimation error is dominated by c, while Condition (ii) requires that c grows much slower than ς_i's. Roughly speaking, Condition (ii) may be violated if the smallest nonzero singular value of $\mathcal{A}_{(i)}$ is too small, or if there is a big drop from $\sigma_j(\mathcal{A}_{(i)})$ to $\sigma_{j+1}(\mathcal{A}_{(i)})$, for some $1 \le j < r_i$ and $1 \le i \le 3$. In either case, it will be more difficult for the ridge-type ratio to select the rank correctly. Note that if all the nonzero singular values are bounded above and away from zero, then Condition (ii) simply becomes $c = o(1)$.

Similar to the minimal signal assumption for variable selection consistency of sparsity-inducing estimators, Assumption 7 is essential to the rank selection consistency:

Theorem 3. Under Assumption 7 and the conditions of Theorem 2, $\mathbb{P}(\widehat{r}_1 = r_1, \widehat{r}_2 = r_2, \widehat{r}_3 = r_3) \to 1$ as $T \to \infty$.

For the initial estimator, in this article, we use the nuclear norm (NN) estimator for low-rank VAR models defined as

$$\widehat{\mathcal{A}}_{\text{NN}} = \arg\min \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{y}_t - \mathcal{A}_{(1)}\boldsymbol{x}_t\|_2^2 + \lambda\|\mathcal{A}_{(1)}\|_*.$$

Note that the estimation error rate derived in Negahban and Wainwright (2011) for VAR(1) models can be readily extended to VAR(P) cases, which yields $\|\widehat{\mathcal{A}}_{\text{NN}} - \mathcal{A}\|_{\text{F}} = O_p(\sqrt{r_1 NP/T})$; see also Remark 10. Then, the rank selection consistency in Theorem 3 would hold for a relatively large range of c. In practice, we recommend using $c = \sqrt{NP\log(T)/10T}$, which is shown to perform satisfactorily in the first simulation experiment of Section 6.

6. Simulation Experiments

6.1. Rank Selection Consistency

As the rank selection method proposed in Section 5 will be used throughout all the following simulations and real data analysis in the next section, we first conduct an experiment to evaluate its consistency.

The data are generated from the proposed model in (5) with dimensions $(N, P) = (10, 5)$, multilinear ranks $(r_1, r_2, r_3) = (3, 3, 3)$, and $\boldsymbol{\epsilon}_t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_N)$. To examine how the singular values of $\mathcal{A}_{(i)}$'s impact the rank selection performance, we let $\mathcal{G}$ be a diagonal cube with superdiagonal elements $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}) = (2, 2, 2)$ (case a), $(4, 3, 2)$ (case b), $(1, 1, 1)$ (case c), or $(2, 1, 0.5)$ (case d). As a result, the three nonzero singular values of every $\mathcal{A}_{(i)}$ are exactly $\mathcal{G}_{111}, \mathcal{G}_{222}$, and $\mathcal{G}_{333}$. We generate the orthonormal factor matrices $\boldsymbol{U}_i$'s as the first r_i left singular vectors of Gaussian random matrices while ensuring that the stationarity condition in Assumption 1 holds. The parameter c for the proposed ridge-type ratio estimator is set to $\sqrt{NP\log(T)/10T}$. Figure 2 presents the proportion of correct rank selection, that is, the event $\{(\widehat{r}_1, \widehat{r}_2, \widehat{r}_3) = (r_1, r_2, r_3)\}$, across different sample sizes $T \in [50, 400]$ based on 1000 replications for each setting. First, it can be seen that the proportion increases as T increases and reaches almost one when $T = 400$ for all cases. Second, as noted in Remark 12, the rank selection may be more difficult if the smallest nonzero singular value $\sigma_{r_i}(\mathcal{A}_{(i)})$ is too small, or if there is a big gap between any two consecutive nonzero singular values. Thus, the better performance of cases a and b may be due to their larger $\sigma_{r_i}(\mathcal{A}_{(i)})$ compared to the other two cases. Moreover, it can be seen that cases a and c outperform cases b and d, respectively, which may be explained by the equality of the singular values $\mathcal{G}_{111}, \mathcal{G}_{222}$, and $\mathcal{G}_{333}$ in the former cases.

6.2. Performance of MLR and SHORR Estimators

We conduct two experiments to verify the theoretical properties of the proposed MLR and SHORR estimators.

We first verify the asymptotic results in Section 3 for the proposed MLR estimator $\widehat{\mathcal{A}}_{\text{MLR}}$ in comparison with the other two low-dimensional estimators, $\widehat{\mathcal{A}}_{\text{OLS}}$ and $\widehat{\mathcal{A}}_{\text{RRR}}$. The data are

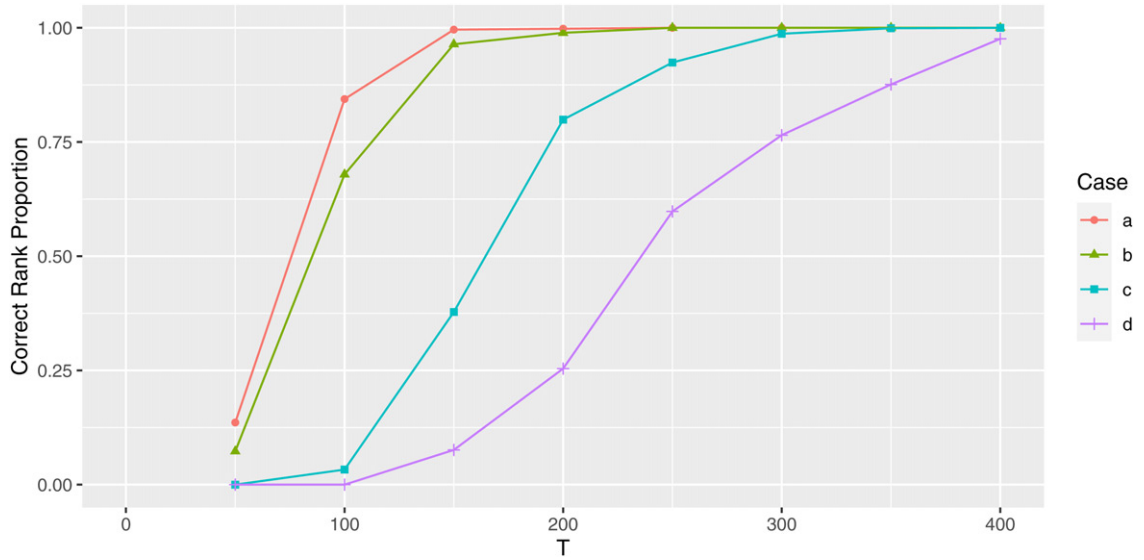

Figure 2. Proportion of correct rank selection when the three nonzero singular values of each $\mathcal{A}_{(i)}$ are (2, 2, 2) (case a), (4, 3, 2) (case b), (1, 1, 1) (case c), or (2, 1, 0.5) (case d).

generated from model (5) with $(N, P) = (10, 5)$, $\boldsymbol{\epsilon}_t \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{I}_N)$, $r_1 = r_2 = 3$, and $r_3 = 2, 3$, or 4. We generate $\mathcal{G}$ by scaling a randomly generated tensor with independent standard normal entries such that $\min_{1\le i\le 3} \sigma_{r_i}(\mathcal{G}_{(i)}) = 1$, and generate $\boldsymbol{U}_i$'s by the same method as in the previous experiment. There are 1000 replications for each setting. Throughout this and all following experiments, the multilinear ranks are selected by the method in Section 5. For each estimator, that is, $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}_{\text{OLS}}, \widehat{\mathcal{A}}_{\text{RRR}}$, or $\widehat{\mathcal{A}}_{\text{MLR}}$, we calculate the average bias across all elements of $\widehat{\mathcal{A}}$ and all replications. The square of this average bias is plotted against $T \in [2000, 4000]$ in the upper panels of Figure 3. We also calculate the empirical and asymptotic variances for each element of $\widehat{\mathcal{A}}$ according to Theorem 1 and Corollary 2. The averages of these empirical and asymptotic variances over all elements of $\widehat{\mathcal{A}}$ and all replications, denoted by EVar and AVar, respectively, are plotted against T in the lower panels of Figure 3. It can be seen that $\widehat{\mathcal{A}}_{\text{MLR}}$ has much smaller squared bias, EVar and AVar than $\widehat{\mathcal{A}}_{\text{OLS}}$ and $\widehat{\mathcal{A}}_{\text{RRR}}$. In addition, the EVar generally matches the corresponding AVar well, with their difference getting smaller as T increases, although the EVar tends to overestimate the variances for all cases due to the large (N, P) relative to the sample size. In sum, the asymptotic theory of the proposed MLR estimator in Section 3 is confirmed by this experiment.

The goal of the next experiment is to verify the nonasymptotic error bound of the proposed SHORR estimator. We consider two settings of the multilinear ranks, $(r_1, r_2, r_3) = (2, 2, 2)$ and $(3, 3, 3)$, and the following four cases of (N, P, s_1, s_2, s_3) for model (5). For case a, we set $(N, P) = (10, 5)$ and $(s_1, s_2, s_3) = (3, 3, 2)$. Then, cases b–d are defined by changing one of the settings in case a while keeping all others fixed. Specifically, we set $(s_1, s_2, s_3) = (2, 2, 2)$ in case b, $N = 20$ in case c, and $P = 10$ in case d. The core tensor $\mathcal{G}$ is generated in the same way as in the previous experiment, and the sparse orthonormal factor matrices $\boldsymbol{U}_i$'s are generated randomly by the method given in Section F of the supplementary materials. The regularization parameter λ is selected by the BIC. By Theorem 2, fixing the multilinear ranks, it holds $\|\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}\|_{\text{F}}^2 = O_p(S \log(N^2 P)/T)$, where $S = s_1 s_2 s_3$. Thus, we denote $\gamma = S \log(N^2 P)/T$ and set the sample size T such that $\gamma = 0.05, 0.1, 0.15, 0.2$, and 0.25. The mean squared error $\|\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}\|_{\text{F}}^2$, averaged over 500 replications, is plotted against γ in Figure 4. It is shown that the mean squared error generally increases linearly in γ, and the four lines in each plot almost coincide. These findings support the error bound in Theorem 2.

6.3. Comparison With Existing Estimation Methods

In the following experiment, we compare the performance of the proposed MLR and SHORR estimators with those of four existing ones for low-rank and/or sparse VAR models, including (i) Lasso (Tibshirani 1996; Basu and Michailidis 2015); (ii) nuclear norm (NN, Negahban and Wainwright 2011); (iii) regression with a sparse SVD (RSSVD, Chen, Chan, and Stenseth 2012); and (iv) sparse and orthogonal factor regression (SOFAR, Uematsu et al. 2019).

The data are generated from model (5) with $(N, P) = (10, 5)$ (case a) or $(15, 8)$ (case b). For both cases, we let $(r_1, r_2, r_3) = (3, 3, 3)$, $(s_1, s_2, s_3) = (3, 3, 2)$, and $\boldsymbol{\epsilon}_t \overset{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{I}_N)$. For case a, $\mathcal{G}$ and $\boldsymbol{U}_i$'s are generated by the same methods as in the previous subsection, and in case b, zeros rows are added below the $\boldsymbol{U}_i$'s in case a. In both cases, entry-wisely $\|\mathcal{A}\|_0 = 500$. Hence, it is not sparse in case a, but is sparse in case b due to the zero rows of $\boldsymbol{U}_i$'s. Figure 5 plots the estimation error $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\text{F}}$ averaged over 500 replications against $T \in [500, 900]$ and $T \in [800, 1200]$ for the smaller and larger (N, P) cases, respectively. The error bars representing $\pm$ one standard deviation are also displayed for the proposed estimators, and suppressed for the others for clearer presentation.

Under both smaller and larger (N, P), Figure 5 shows that both $\widehat{\mathcal{A}}_{\text{MLR}}$ and $\widehat{\mathcal{A}}_{\text{SHORR}}$ significantly outperform the other estimators which either consider the low-rankness along only

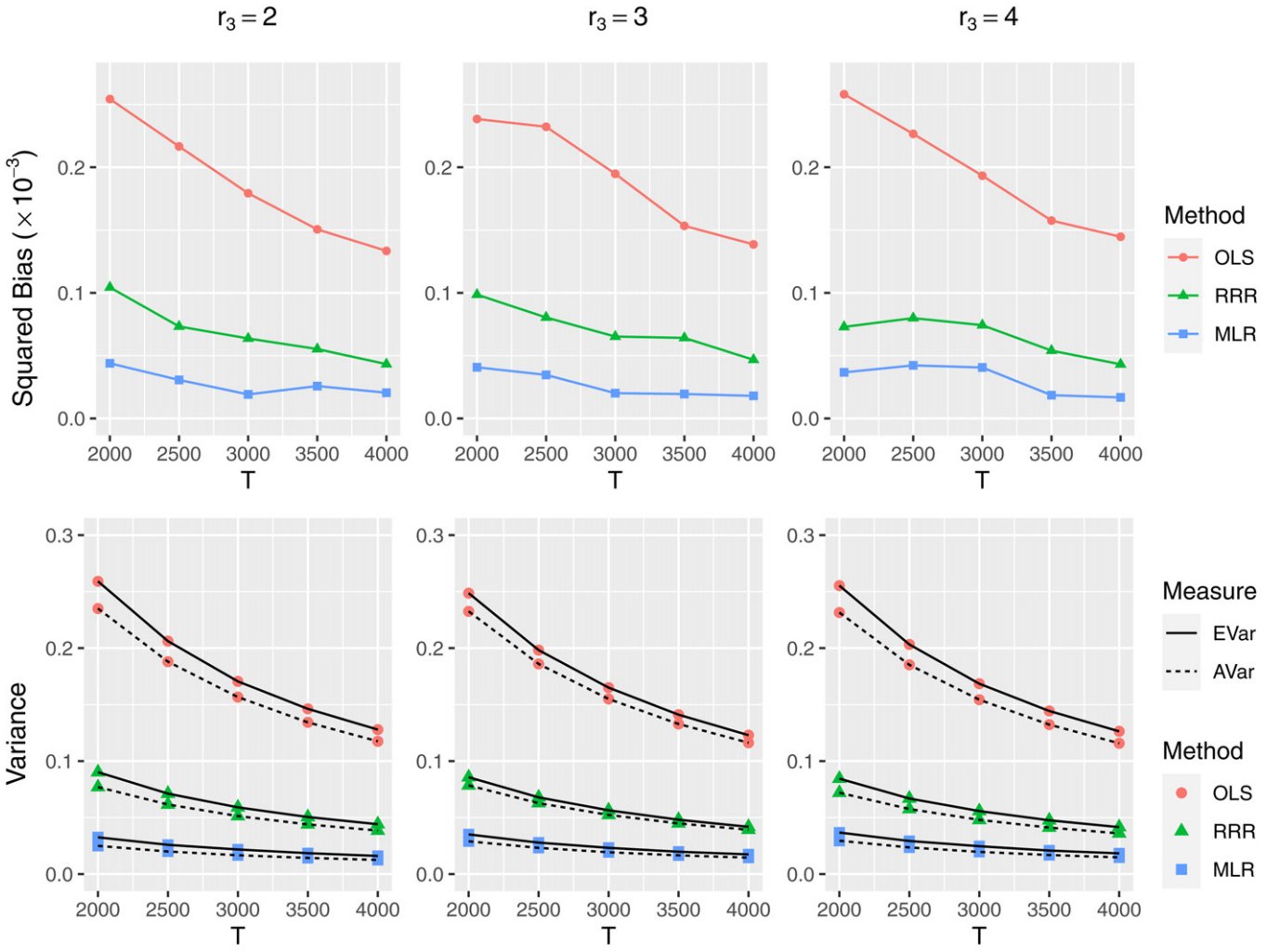

Figure 3. Squared bias, empirical variance (EVar) and asymptotic variance (AVar) for $\widehat{\mathcal{A}}_{\text{OLS}}$ $\widehat{\mathcal{A}}_{\text{RRR}}$, and $\widehat{\mathcal{A}}_{\text{MLR}}$ under various multilinear ranks.

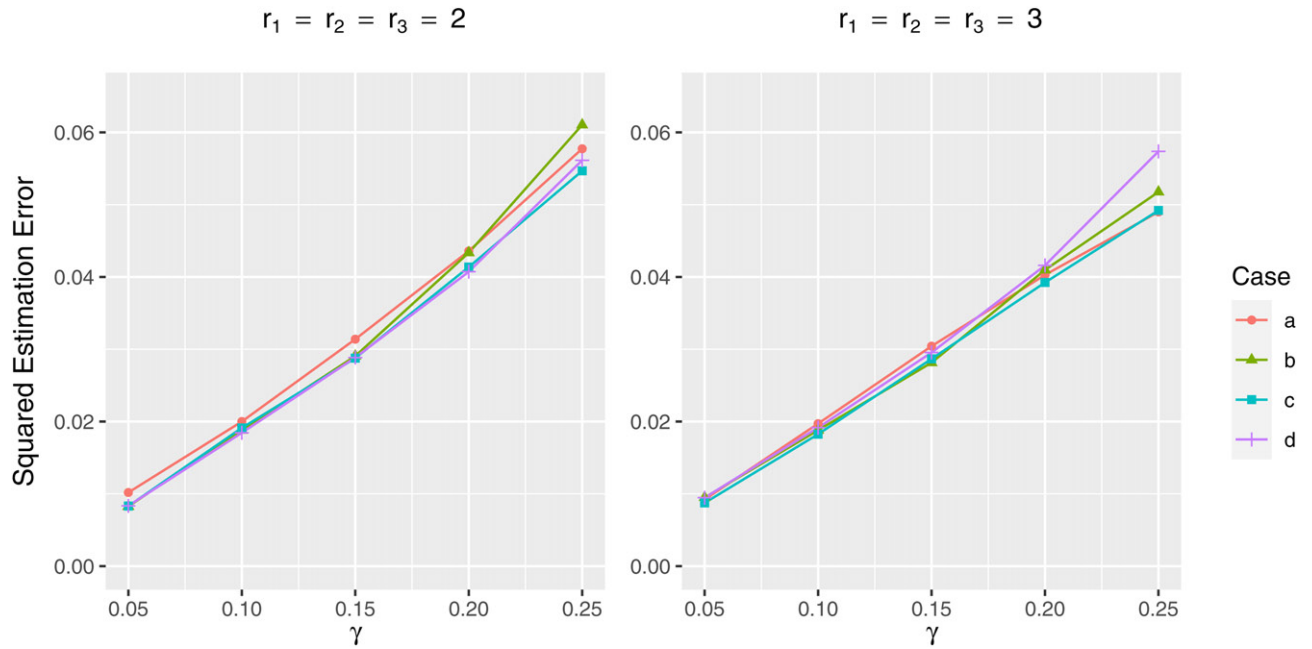

Figure 4. Plots of the squared estimation error $\|\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}\|_{\text{F}}^2$ against $\gamma = S\log(N^2P)/T$ for four cases of (N, P, s_1, s_2, s_3) under two settings of multilinear ranks.

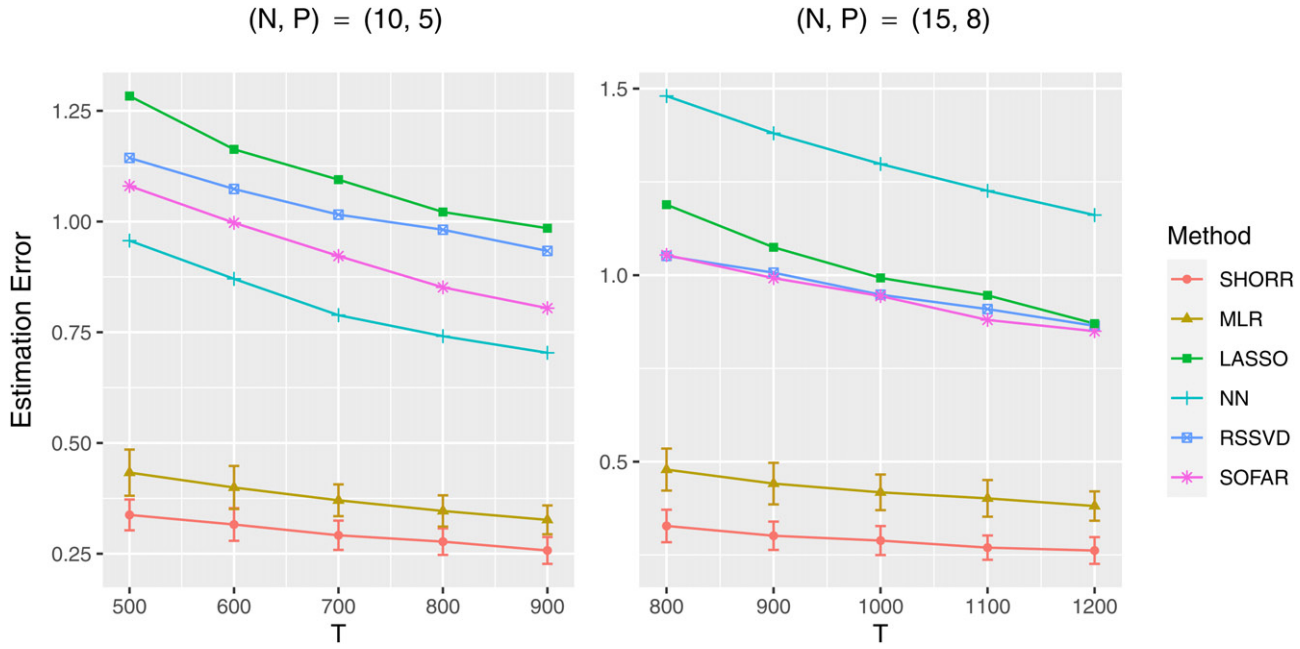

Figure 5. Plots of the estimation error $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\mathrm{F}}$ against T for six estimation methods under two settings of (N, P).

one direction or ignore it completely. Moreover, $\widehat{\mathcal{A}}_{\mathrm{SHORR}}$ consistently outperforms $\widehat{\mathcal{A}}_{\mathrm{MLR}}$ as the former exploits the sparsity of $\boldsymbol{U}_i$'s in addition to the low-rankness along three dimensions. It is also interesting to note the different performances of $\widehat{\mathcal{A}}_{\mathrm{LASSO}}$ and $\widehat{\mathcal{A}}_{\mathrm{NN}}$ in Figure 5. Since $\widehat{\mathcal{A}}_{\mathrm{LASSO}}$ only exploits entry-wise sparsity of $\mathcal{A}$, it has the worst performance when $\mathcal{A}$ is not sparse, as shown in the left panel. In contrast, $\widehat{\mathcal{A}}_{\mathrm{NN}}$ only takes into account the low-rankness, so it performs best among the four existing estimators when $\mathcal{A}$ is not sparse, and yet becomes the worst when $\mathcal{A}$ is sparse as is the case for the right panel. This suggests that higher efficiency can be achieved by incorporating both the low-rankness and sparsity, which is the key advantage of the proposed $\widehat{\mathcal{A}}_{\mathrm{SHORR}}$.

6.4. Comparison With Factor Models

The final experiment aims to compare the proposed model to the SFM and DFM discussed in Section 2.3. Note that the SFM cannot be directly used for forecasting since it does not impose an explicit model on the latent factors. However, for data generated by both the proposed model and the DFM, the SFM can be used to estimate the low-dimensional subspace where the conditional mean $\mathbb{E}(\boldsymbol{y}_t|\mathcal{F}_{t-1})$ lies; see Remark 3.

We consider four data-generating processes for $\{\boldsymbol{y}_t\}$ with dimension $N = 10$. Two of them are generated by DFMs with $\boldsymbol{e}_t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, 0.5\boldsymbol{I}_N)$:

- DFM-1: The DFM with $r = 1$, specified jointly by the SFM $\boldsymbol{y}_t = \boldsymbol{\Lambda} f_t + \boldsymbol{e}_t$ and the autoregressive latent factor $f_t = Bf_{t-1} + \xi_t$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{10\times 1}$ is a randomly generated vector with unit Euclidean norm, $B = 0.5$, and $\xi_t \overset{\text{iid}}{\sim} N(0, 1)$.
- DFM-2: The DFM with $r = 3$, specified jointly by the SFM $\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t$ and the VAR(1) process for the latent factors $\boldsymbol{f}_t = \boldsymbol{B}\boldsymbol{f}_{t-1} + \boldsymbol{\xi}_t$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{10\times 3}$ is a randomly generated orthonormal matrix, $\boldsymbol{B} = \mathrm{diag}(0.6, 0.5, 0.4)$, and $\boldsymbol{\xi}_t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_3)$.

The other two are generated by the proposed model with $P = 3$ and $\boldsymbol{\epsilon}_t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{I}_N)$:

- MLR-1: The proposed multilinear low-rank VAR model in (5) with $(r_1, r_2, r_3) = (2, 2, 2)$. The core tensor $\mathcal{G}$ and factor matrices $\boldsymbol{U}_i$'s are generated in the same way as the first experiment in Section 6.2.
- MLR-2: Same as MLR-1 except for $(r_1, r_2, r_3) = (3, 3, 3)$.

We first compare the performance of the proposed model and the SFM in terms of the estimation accuracy of the conditional mean subspace. The estimation of the SFM is conducted by the principal component method in Bai and Wang (2016). The subspace estimation error can be measured by $\|\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{\Lambda}}_1' - \boldsymbol{\Lambda}\boldsymbol{\Lambda}'\|_{\mathrm{F}}^2$ for DFM-1 and DFM-2, and $\|\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{\Lambda}}_1' - \boldsymbol{U}_1\boldsymbol{U}_1'\|_{\mathrm{F}}^2$ for MLR-1 and MLR-2, where $\widehat{\boldsymbol{\Lambda}}_1 = \widehat{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\Lambda}}'\widehat{\boldsymbol{\Lambda}})^{-1/2}$ is the normalized version of $\widehat{\boldsymbol{\Lambda}}$ for the fitted SFM. The results based on 1000 replications are displayed in Figure 6. It can be seen from the upper panels of the figure that the conditional mean subspaces for DFM-1 and DFM-2 can be consistently estimated by fitting the corresponding SFMs. However, as discussed in Section 2.3, for data generated by the DFM, fitting the proposed model will lead to model misspecification. This may explain why the subspace estimation error for the MLR method is much larger and seems to persist for large T in the upper panels of Figure 6. On the other hand, when the data-generating process is MLR-1 or MLR-2, the lower panels of Figure 6 show that both methods can estimate the subspace consistently, although the MLR method is more efficient. This agrees with our observation that the proposed model admits an SFM representation.

We next compare the performance of the proposed model and the DFM through the prediction error of the conditional

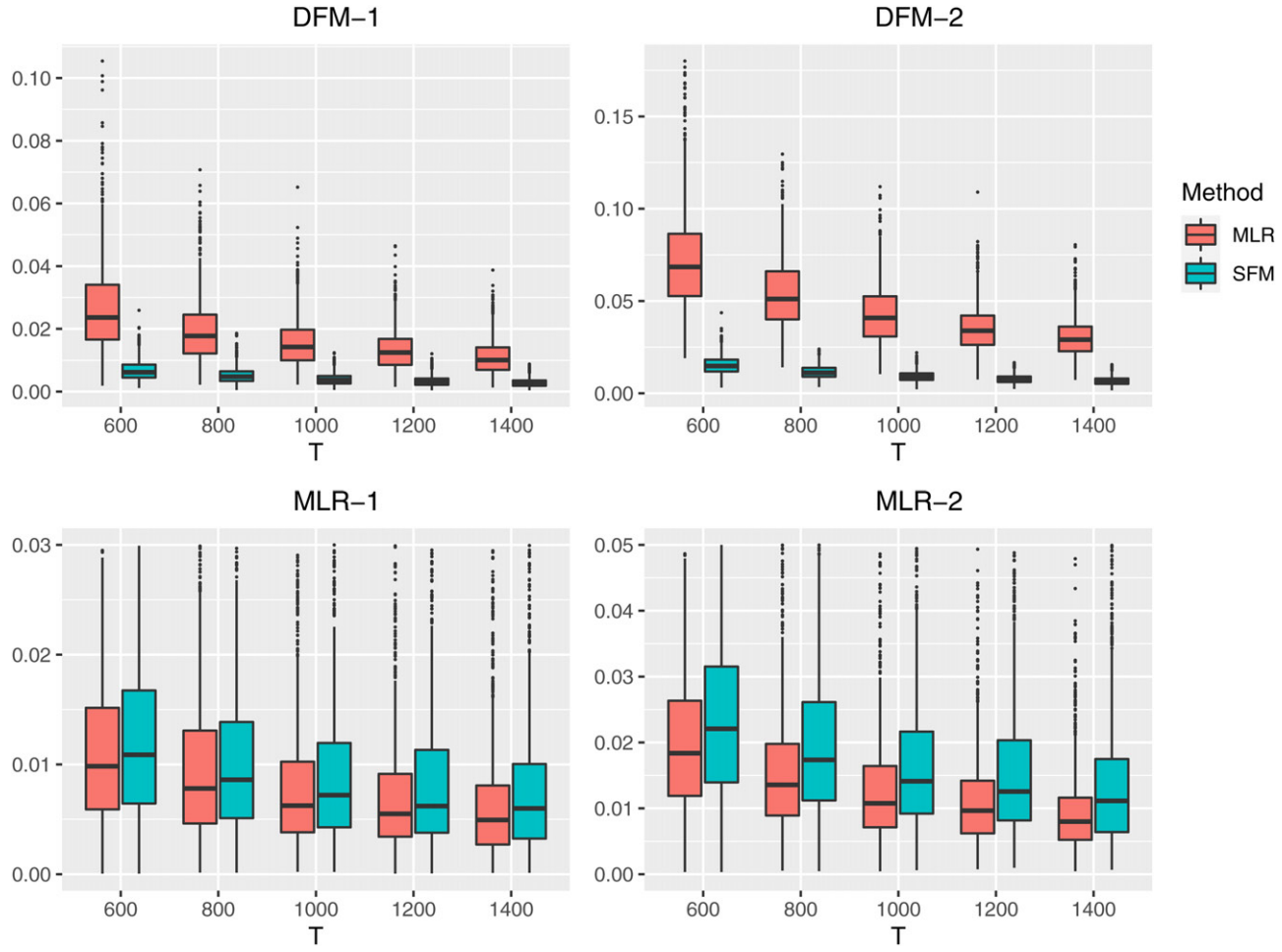

Figure 6. Subspace estimation error for four data-generating processes based on two methods: fitting the proposed model by the MLR method, or fitting the static factor model (SFM).

mean $\mathbb{E}(\boldsymbol{y}_{T+1}|\mathcal{F}_T)$. The DFM is estimated by a two-step approach, where we first obtain the estimated factors $\widehat{\boldsymbol{f}}_t$ by fitting an SFM, and then fit a (vector) autoregressive model to $\{\widehat{\boldsymbol{f}}_t\}$. Figure 7 displays the prediction error $\|\widehat{\mathbb{E}}(\boldsymbol{y}_{T+1}|\mathcal{F}_T) - \mathbb{E}(\boldsymbol{y}_{T+1}|\mathcal{F}_T)\|_2$ based on 1000 replications. Remarkably, as shown in the upper panels, even if the data are generated from the DFM-1 or DFM-2, the proposed model exhibits competitive forecasting performance despite the model misspecfication. On the other hand, as shown in the lower panels, when the data are generated from MLR-1 or MLR-2, the forecasting performance of the fitted DFM is rather poor. As discussed in Section 2.3, the proposed model can accommodate different low-dimensional patterns for the response $\boldsymbol{y}_t$ and predictors $\boldsymbol{y}_{t-j}$'s, whereas the DFM requires the subspaces of $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-j}$'s to be identical. When the DGP is the proposed model with distinct $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, the DFM will forecast $\widehat{\boldsymbol{U}}_1'\boldsymbol{y}_t$ based on $\widehat{\boldsymbol{U}}_1'\boldsymbol{y}_{t-j}$. However, the true conditional expectation of $\boldsymbol{y}_t$ is dependent on $\boldsymbol{U}_2'\boldsymbol{y}_{t-j}$, and the latter could be different from, or even orthogonal to, $\widehat{\boldsymbol{U}}_1'\boldsymbol{y}_{t-j}$. Consequently, when the response's low-dimensional subspace is applied to the predictors, the DFM may have no predictive power at all. This explains why forecasting based on the DFM leads to considerable prediction errors. The robust forecasting performance of the proposed model reflects that its low-dimensional structure can be much more flexible than that of the DFM.

7. Real Data Analysis

This section 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 1 gives more details about these macroeconomic variables.

We first apply the SHORR estimation to the entire dataset, 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 $\|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1$ with $\|\boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1$ in the penalty term. The multilinear ranks are selected by the ridge-type ratio estimator, which results in $(r_1, r_2, r_3) = (4, 3, 2)$, and the tuning parameter λ is selected by BIC.

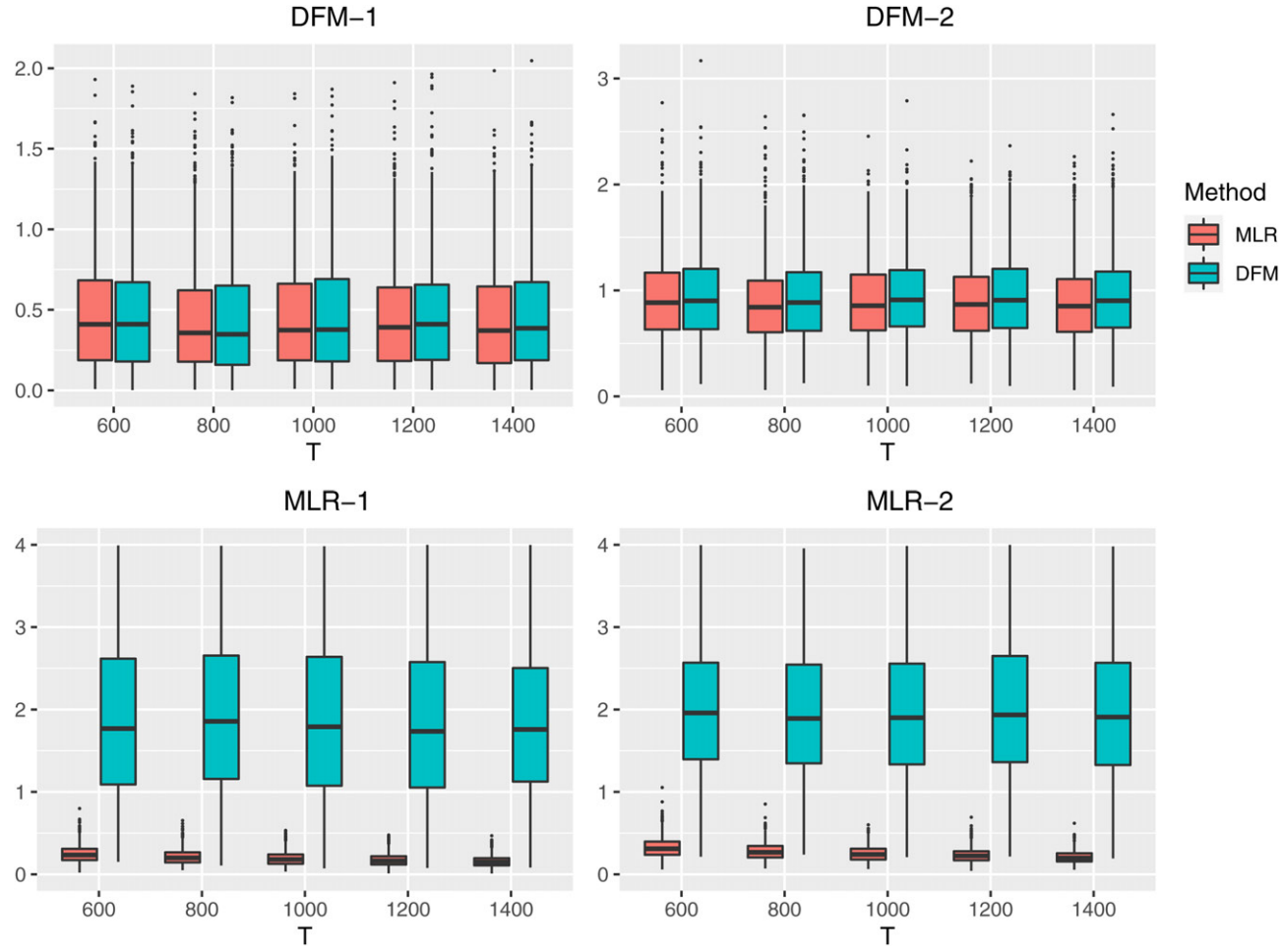

Figure 7. Prediction error for four data-generating processes based on two methods: fitting the proposed model by the MLR method, or fitting the dynamic factor model (DFM).

Table 1. Forty quarterly macroeconomic variables belonging to eight categories.

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

NOTE: 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.

The ℓ_1 penalty yields sparse estimated factor matrices $\widehat{\boldsymbol{U}}_1$ and $\widehat{\boldsymbol{U}}_2$, and the estimated coefficients are presented in Figure 8. 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

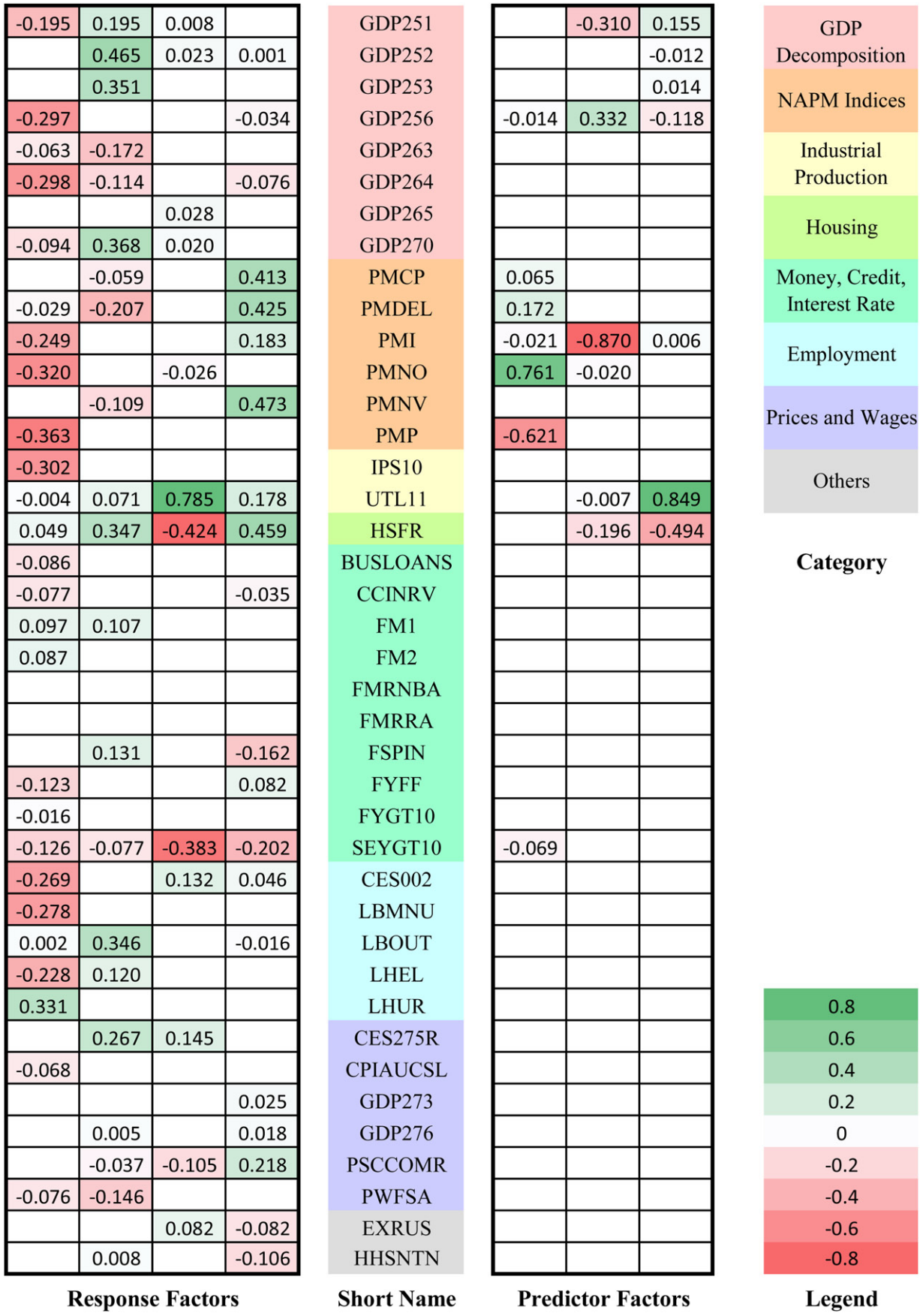

Figure 8. Estimated coefficients in the response and predictor factor loading matrices.

Table 2. Forecasting error for 40 quarterly macroeconomic sequences of the United States from 1959 to 2007.

	Unregularized methods				Regularized methods				
Criterion	OLS	RRR	DFM	MLR	SHORR	LASSO	NN	RSSVD	SOFAR
ℓ_2 norm	20.16	13.31	6.36	**5.81**	**5.35**	6.72	8.16	6.33	6.28
ℓ_∞ norm	8.32	4.55	2.85	**2.56**	**2.44**	3.06	3.36	3.02	3.02

NOTE: The best cases among (un)regularized methods are marked in bold.

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 $\widehat{\mathcal{A}}_{\text{MLR}}$ and $\widehat{\mathcal{A}}_{\text{SHORR}}$ in comparison with the competing estimators considered in Section 6. The following rolling forecasting procedure is adopted: first, use the historical data with the endpoint 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 $\widehat{\mathcal{A}}_{\text{SHORR}}$ are preserved from the analysis of the entire dataset, that is, $(r_1, r_2, r_3) = (4, 3, 2)$, 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 2. It can be seen that the proposed MLR and SHORR estimators have much smaller forecast errors than competing ones, including the DFM with $r = 4$ and the regularized and unregularized estimation methods for the VAR model. This can be explained by the capability of the proposed estimators to substantially reduce the dimensionality along three directions simultaneously. The SHORR estimator performs best among all estimators as it enforces sparsity of the factor matrices and hence prevents overfitting most effectively.

8. Conclusion and Discussion

For a large 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 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 along 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 article further proposes a sparsity-inducing estimator to improve the model interpretability and estimation efficiency. An ADMM algorithm is developed to tackle the computational challenges due to the all-orthogonal constraints on $\mathcal{G}$ as well as the jointly imposed ℓ_1-regularization and orthogonality constraints on $\boldsymbol{U}_i$'s. It is worth noting that this article has a different focus than most work on tensor regression: here we employ the tensor technique as a novel approach to the dimension reduction problem in classical VAR time series modeling.

This article may be extended in three possible directions. First, the proposed estimators do not take into account the possible correlation structure among components of $\boldsymbol{\epsilon}_t$, which will reduce the estimation efficiency. Let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ be an estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. As in Davis, Zang, and Zheng (2016), we may alternatively consider the generalized least squares loss $\sum_{t=1}^{T}(\boldsymbol{y}_t - \mathcal{A}_{(1)}\boldsymbol{x}_t)\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}(\boldsymbol{y}_t - \mathcal{A}_{(1)}\boldsymbol{x}_t)$ rather than $\sum_{t=1}^{T}\|\boldsymbol{y}_t - \mathcal{A}_{(1)}\boldsymbol{x}_t\|_2^2$. However, the difficulty would be to find a good estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Second, the tensor technique potentially can be applied to many variants of the VAR model, for example, 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, $\boldsymbol{y}_t = \boldsymbol{A}_t\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t$. Similarly, the coefficient matrices can be rearranged into a tensor $\mathcal{A}$ with $\mathcal{A}_{(1)} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_T)$. If $\mathcal{A}$ has multilinear low ranks (r_1, r_2, r_3), then the number of parameters will be $r_1r_2r_3 + (N - r_1)r_1 + (N - r_2)r_2 + (T - r_3)r_3 \lesssim NT$. 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, Liu, and Chen (2019) for a related work.

Supplementary Materials

The online supplementary materials contain the proofs of the theoretical results, details about the SFM representation in equation (9), and methods for generating orthonormal matrices in simulation experiments.

Acknowledgments

We are grateful to the joint editor, the associate editor, and three anonymous referees for their valuable comments which led to substantial improvement of this article.

Acknowledgments

This research was partially supported by GRF grants 11300519 and 17305319 from the Hong Kong Research Grant Council and a Key Program grant 72033002 from National Natural Science Foundation of China.

References

Amengual, D., and Watson, M. W. (2007), "Consistent Estimation of the Number of Dynamic Factors in Large N and T Panel," *Journal of Business & Economic Statistics*, 25, 91–96. [1339,1342]

Athanasopoulos, G., and Vahid, F. (2008), "VARMA Versus VAR for Macroeconomic Forecasting," *Journal of Business & Economic Statistics*, 26, 237–252. [1338]

Bai, J., and Ng, S. (2008), "Large Dimensional Factor Analysis," *Foundations and Trends® in Econometrics*, 3, 89–163. [1339,1341]

Bai, J., and Wang, P. (2016), "Econometric Analysis of Large Factor Models," *Annual Review of Economics*, 8, 53–80. [1339,1341,1350]

Bai, Z. D. (1997), "Circular Law," *The Annals of Probability*, 25, 494–529. [1340]

Basu, S., Li, X., and Michailidis, G. (2019), "Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions," *IEEE Transactions on Signal Processing*, 67, 1207–1222. [1339]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-Dimensional Time Series Models," *The Annals of Statistics*, 43, 1535–1567. [1338,1344,1345,1348]

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005), "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120, 387–422. [1342]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1344]

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. [1345]

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. [1339,1344]

Carriero, A., Kapetanios, G., and Marcellino, M. (2011), "Forecasting Large Datasets With Bayesian Reduced Rank Multivariate Models," *Journal of Applied Econometrics*, 26, 735–761. [1338,1339]

Chan, J. C. C., Eisenstat, E., and Koop, G. (2016), "Large Bayesian VARMAs," *Journal of Econometrics*, 192, 374–390. [1338]

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. [1339,1344,1348]

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920. [1339]

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. [1339,1344]

Davis, R. A., Zang, P., and Zheng, T. (2016), "Sparse Vector Autoregressive Modeling," *Journal of Computational and Graphical Statistics*, 25, 1077–1096. [1338,1354]

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. [1340,1341]

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. [1338]

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. [1338]

Ding, S., and Cook, R. D. (2018), "Matrix Variate Regressions and Envelope Models," *Journal of the Royal Statistical Society*, Series B, 80, 387–408. [1341]

Guo, S., Wang, Y., and Yao, Q. (2016), "High-Dimensional and Banded Vector Autoregressions," *Biometrika*, 103, 889–903. [1338]

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. [1338]

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. [1338]

Kock, A. B., and Callot, L. (2015), "Oracle Inequalities for High Dimensional Vector Autoregressions," *Journal of Econometrics*, 186, 325–344. [1338]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [1339,1340]

Koop, G. M. (2013), "Forecasting With Medium and Large Bayesian VARs," *Journal of Applied Econometrics*, 28, 177–203. [1338,1351]

Lai, R., and Osher, S. (2014), "A Splitting Method for Orthogonality Constrained Problems," *Journal of Scientific Computing*, 58, 431–449. [1346]

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. [1338]

Li, G., Yang, D., Nobel, A. B., and Shen, H. (2016), "Supervised Singular Value Decomposition and Its Asymptotic Properties," *Journal of Multivariate Analysis*, 146, 7–17. [1343]

Li, L., and Zhang, X. (2017), "Parsimonious Tensor Response Regression," *Journal of the American Statistical Association*, 112, 1131–1146. [1339]

Li, X., Xu, D., Zhou, H., and Li, L. (2018), "Tucker Tensor Regression and Neuroimaging Analysis," *Statistics in Biosciences*, 10, 520–545. [1339,1344]

Lian, H., Feng, S., and Zhao, K. (2015), "Parametric and Semiparametric Reduced-Rank Regression With Flexible Sparsity," *Journal of Multivariate Analysis*, 136, 163–174. [1339]

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Berlin: Springer. [1338,1354]

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. [1339,1346,1347,1348]

Raskutti, G., Yuan, M., and Chen, H. (2019), "Convex Regularization for High-Dimensional Multi-Response Tensor Regression," *The Annals of Statistics*, 47, 1554–1584. [1339]

Ravenna, F. (2007), "Vector Autoregressions and Reduced Form Representations of DSGE Models," *Journal of Monetary Economics*, 54, 2048–2064. [1338]

Reinsel, G. (1983), "Some Results on Multivariate Autoregressive Index Models," *Biometrika*, 70, 145–156. [1339]

Said, E. S., and Dickey, D. A. (1984), "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order," *Biometrika*, 71, 599–607. [1338]

Shapiro, A. (1986), "Asymptotic Theory of Overparameterized Structural Models," *Journal of the American Statistical Association*, 81, 142–149. [1342]

Staudenmayer, J., and Buonaccorsi, J. P. (2005), "Measurement Error in Linear Autoregressive Models," *Journal of the American Statistical Association*, 100, 841–852. [1342]

Stock, J. H., and Watson, M. W. (2005), "Implications of Dynamic Factor Models for VAR Analysis," National Bureau of Economic Research Working Paper No. 11467. [1339,1341]

——— (2009), "Forecasting in Dynamic Factor Models Subject to Structural Instability," in *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, eds. J. L. Castle and N. Shephard, Oxford: Oxford University Press, pp. 173–205. [1351]

——— (2011), "Dynamic Factor Models," in *The Oxford Handbook of Economic Forecasting*, eds. M. P. Clements and D. F. Hendry, Oxford: Oxford University Press, pp. 35–60. [1339,1341]

Sun, W. W., and Li, L. (2017), "Store: Sparse Tensor Response Regression and Neuroimaging Analysis," *The Journal of Machine Learning Research*, 18, 4908–4944. [1339]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1348]

Tsay, R. S. (1998), "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association*, 93, 1188–1202. [1354]

——— (2010), *Analysis of Financial Time Series* (3rd ed.), Hoboken NJ: Wiley. [1338,1339]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [1340]

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. [1339,1344,1346,1348]

Velu, R. P., and Reinsel, G. C. (2013), *Multivariate Reduced-Rank Regression: Theory and Applications* (Vol. 136), New York: Springer-Verlag. [1339]

Velu, R. P., Reinsel, G. C., and Wichern, D. W. (1986), "Reduced Rank Models for Multiple Time Series," *Biometrika*, 73, 105–118. [1339]

Vu, V. Q., and Lei, J. (2013), "Minimax Sparse Principal Subspace Estimation in High Dimensions," *The Annals of Statistics*, 41, 2905–2947. [1342]

Wang, D., Liu, X., and Chen, R. (2019), "Factor Models for Matrix-Valued High-Dimensional Time Series," *Journal of Econometrics*, 208, 231–248. [1354]

Wang, Y., Yin, W., and Zeng, J. (2019), "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization," *Journal of Scientific Computing*, 78, 29–63. [1346]

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015), "High Dimensional EM Algorithm: Statistical Optimization and Asymptotic Normality," in *Advances in Neural Information Processing Systems*, pp. 2521–2529. [1345]

Wilms, I., Basu, S., Bien, J., and Matteson, D. S. (2017), "Sparse Identification and Estimation of Large-Scale Vector Autoregressive Moving Averages," arXiv no. 1707.09208. [1338]

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. [1338]

Xia, Q., Xu, W., and Zhu, L. (2015), "Consistently Determining the Number of Factors in Multivariate Volatility Modelling," *Statistica Sinica*, 25, 1025–1044. [1347]

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. [1339]

Zhang, A. (2019), "Cross: Efficient Low-Rank Tensor Completion," *The Annals of Statistics*, 47, 936–964. [1341]

Zhao, J., and Leng, C. (2014), "Structure Lasso for Regression With Matrix Covariates," *Statistica Sinica*, 24, 799–814. [1341,1344]

Zheng, L., and Raskutti, G. (2019), "Testing for High-Dimensional Network Parameters in Auto-Regressive Models," *Electronic Journal of Statistics*, 13, 4977–5043. [1344]

Zheng, Y., and Cheng, G. (2020), "Finite Time Analysis of Vector Autoregressive Models Under Linear Restrictions," *Biometrika* (to appear). [1338]

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. [1339,1344]

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017), "Network Vector Autoregression," *The Annals of Statistics*, 45, 1096–1123. [1338]