Statistica Sinica Preprint No: SS-2026-0065	
Title	Sparse Tucker Decomposition and Graph Regularization for High-dimensional Time Series Forecasting
Manuscript ID	SS-2026-0065
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0065
Complete List of Authors	Sijia Xia, Michael K. Ng and Xiongjun Zhang
Corresponding Authors	Xiongjun Zhang
E-mails	xjzhang@ccnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Statistica Sinica

Sparse Tucker Decomposition and Graph Regularization for High-Dimensional Time Series Forecasting

Sijia Xia[1], Michael K. Ng[2], and Xiongjun Zhang[1]

[1]*Central China Normal University and* [2]*Hong Kong Baptist University*

Abstract: Existing methods of vector autoregressive model for multivariate time series analysis make use of low-rank matrix approximation or Tucker decomposition to reduce the dimension of the over-parameterization issue. In this paper, we propose a sparse Tucker decomposition method with graph regularization for high-dimensional vector autoregressive time series. By stacking the time-series transition matrices into a third-order tensor, the sparse Tucker decomposition is employed to characterize important interactions within the transition third-order tensor and reduce the number of parameters. Moreover, the graph regularization is employed to measure the local consistency of the response, predictor and temporal factor matrices in the vector autoregressive model. The two proposed regularization techniques can be shown to more accurate parameters estimation. A non-asymptotic error bound of the estimator of the proposed method is established, which is lower than those of the existing matrix or tensor based methods when the Tucker core is sufficiently sparse. A proximal alternating linearized minimization algorithm is designed to solve the resulting model and its global convergence is established under very mild conditions. Numerical experiments

on synthetic data and real-world datasets are carried out to verify the superior performance of the proposed method over existing state-of-the-art methods.

Key words and phrases: Graph regularization, high-dimensional vector autoregression, proximal alternating linearized minimization algorithm, sparse Tucker decomposition.

1. Introduction

With the rapid development of information technology, high-dimensional time series data have emerged across various fields in the current data-abundant environment, such as finance (Tsay, 2005), economics (Sims, 1980), and meteorology (Duchon and Hale, 2012). Classical vector autoregressive (VAR) models are widely employed for modeling multivariate time series data due to the flexible ability for capturing cross-variable temporal dynamics, which have been applied in a wide range of areas, such as signal processing (Basu et al., 2019) and traffic state estimation (Chen et al., 2025). Consider an m-dimensional zero-mean VAR model of lag order p (denoted by VAR(p)) in the following form

$$\mathbf{y}_t = \mathbf{W}_1\mathbf{y}_{t-1} + \mathbf{W}_2\mathbf{y}_{t-2} + \cdots + \mathbf{W}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad 1 \le t \le T, \tag{1.1}$$

where $\{\mathbf{y}_t\}$ is the observed data with $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})' \in \mathbb{R}^m$, $\mathbf{W}_i \in \mathbb{R}^{m\times m}$ are the transition matrices, $i = 1, \ldots, p$, T denotes the sample size, $\boldsymbol{\varepsilon}_t$

denotes independent and identically distributed (i.i.d.) error with $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{mt})' \in \mathbb{R}^m, \mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$, and $\text{var}(\boldsymbol{\varepsilon}_t) < \infty$.

In real-world applications, the dimension m in (1.1) is often very large, which implies that the number of coefficient parameters (i.e., m^2p) is large. Hence an unrestricted VAR(p) model is likely to encounter the difficulty of over-parameterization, where the number of parameters (e.g., coefficient matrices in time-varying VAR) inevitably exceeds the number of observations. As a result, the corresponding VAR methods cannot provide reliable estimates nor accurate forecasts without further restrictions (De Mol et al., 2008). Estimation consistency of high-dimensional VAR models is achievable under certain regularity conditions about the transition matrices (Wang and Tsay, 2023). For example, if the coefficient matrices have an unobserved low-dimensional structure, such as sparsity or low-rankness, the structure-inducing regularization methods, including least absolute shrinkage and selection operator (Lasso) (Basu and Michailidis, 2015) and nuclear norm penalty (Negahban and Wainwright, 2011) give consistent estimates under the Gaussian assumption of the time series.

On the other hand, the number of parameters in $\mathbf{W}_i$ increases quadratically with the dimension m, which makes it difficult to apply VAR models to high-dimensional data. To overcome it, a commonly used approach is to

Statistica Sinica: Newly accepted Paper

assume sparsity in parameter matrices, and many sparsity-imposing or inducing methods can then be employed for estimation and variable selection, including ℓ_1 regularization (Basu and Michailidis, 2015; Han et al., 2015; Kock et al., 2026; Wong et al., 2020), weighted ℓ_1 regularization (Wang et al., 2023), weakly sparse constraint on the transition matrix (Han et al., 2015), and nonconcave penalization method based on ℓ_1 norm (Zhu, 2020). However, unlike the linear regression for Lasso, the time series data exhibit temporal and cross-sectional dependence, while the ℓ_1 regularization based methods neglect this property, which will seriously affect the accuracy of estimators of the sparse regularization based methods.

To address the over-parameterization issue of the VAR model and the data dependence for time series data, some reduced-rank regression based methods were proposed, see Negahban and Wainwright (2011); Basu et al. (2019); Raskutti et al. (2019); Miao et al. (2023) and references therein. In these papers, the transition matrices are low-rank, which can reduce the number of parameters. For example, by forming the transition matrices into a larger matrix $\mathbf{W}^{(C)} = (\mathbf{W}_1, \ldots, \mathbf{W}_p)$, Basu et al. (2019) proposed a low-rank plus sparse estimation for high-dimensional VAR models, where only order-one was discussed in this model. Samadi and Herath (2024) proposed a reduced-rank envelope VAR model by combining the envelope model into

the reduced-rank VAR model to extract relevant information from complex data efficiently. Additionally, Reinsel (1983) proposed an autoregressive index model based on the low-rankness assumption on the stacking matrices $\mathbf{W}^{(R)} = (\mathbf{W}_1', \ldots, \mathbf{W}_p')$. However, the reduced-rank models based on matrix methods only consider the low-rankness of $\mathbf{W}^{(C)}$ or $\mathbf{W}^{(R)}$, which can only reduce the dimensionality of these parameter matrices along one direction.

By reformulating the transition matrices into a third-order tensor, Wang et al. (2022) developed a tensor Tucker decomposition method for VAR time series modeling, where the transition tensor is decomposed into a core tensor and three sparse factor matrices, and the low-rankness can be explored along different directions. However, the above tensor based method did not consider the local patterns of the time series data. Bahadori et al. (2014) proposed a unified low-rank tensor learning framework for multivariate spatio-temporal analysis by using low Tucker rank constraint and spatial Laplacian regularization, which can be applied for VAR time series. Moreover, Harris et al. (2021) proposed a time-varying autoregressive model by incorporating CANDECOMP/PARAFAC (CP) decomposition and smoothness priors over time for multivariate time series. However, the CP rank is NP-hard to determine in general (Hillar and Lim, 2013). Besides, there is no theoretical guarantee in the above two work.

Statistica Sinica: Newly accepted Paper

1.1 Our Proposal

In this paper, we propose a sparse Tucker decomposition method with graph regularization for high-dimensional VAR time series. Specifically, by stacking the transition matrices of (1.1) into a third-order tensor, the sparse Tucker decomposition is employed to explore the low-rankness of the transition tensor along different dimensions, which can reduce the parameters greatly for small Tucker rank. Moreover, the sparsity is imposed on the core tensor, which can reduce the parameters further and select significant variables along the response, predictor, and temporal factors for VAR time series. Besides, the graph regularization on the factor matrices is utilized to characterize the local patterns of the transition tensor, which is capable of preserving the intrinsic manifold structure of data in the response, predictor and temporal factors. Compared with the existing tensor decomposition method in Wang et al. (2022), the proposed framework jointly integrates Tucker low-rank modeling, sparse core tensor, and graph regularization within a autoregressive framework. A non-asymptotic error bound of the estimator of the proposed model is established under some conditions, which is smaller than those of the existing sparse and low-rank based methods when the Tucker core is sufficiently sparse. A proximal alternating linearized minimization (PALM) algorithm is developed to solve the

Statistica Sinica: Newly accepted Paper

resulting model and its global convergence is established under mild conditions. Numerical experiments on simulated data and real-world datasets substantiate the superiority of the proposed method compared with other competition methods. In Table 1, we summarize existing VAR models and our contribution.

Table 1: Comparison of our method with existing matrix- and tensor-based VAR methods.

Method	Coefficient	Factor sparsity	Core sparsity	Graph	Theoretical guarantee
Basu and Michailidis (2015)	Sparse	×	×	×	✓
Negahban and Wainwright (2011)	low-rank matrix	×	×	×	✓
Bahadori et al. (2014)	Tucker	×	×	✓	×
Harris et al. (2021)	CP	×	×	×	×
Wang et al. (2022)	Tucker	✓	×	×	✓
Our method	Tucker	×	✓	✓	✓

The remaining parts of this paper are organized as follows. In the next section, some notations and preliminaries about tensors are given. In Section 2, we propose a sparse Tucker decomposition with graph regularization method for high-dimensional VAR time series. Moreover, a non-asymptotic error bound of the proposed model is established. In Section 3, a PALM

Statistica Sinica: Newly accepted Paper

algorithm with global convergence guarantee is designed to solve the resulting model. In Section 4, some numerical experiments are conducted to demonstrate the effectiveness of the proposed method. Section 5 offers concluding remarks. All technical proofs of main results and additional numerical experiments are deferred to the supplementary materials.

1.2 Preliminaries

Throughout this paper, we denote scalars by lowercase letters (e.g., a), vectors by bold lowercase letters (e.g., $\mathbf{a}$), matrices by bold uppercase letters (e.g., $\mathbf{A}$), and tensors by Euler script letters (e.g., $\mathcal{A}$). The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$. For a third-order tensor $\mathcal{A}$, its (i, j, l)-th entry is denoted by $\mathcal{A}_{ijl}$. Moreover, $(\cdot)'$ and $(\cdot)^{-1}$ denote the transpose and inverse operators, respectively, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. We use $\mathcal{A}^{\langle i \rangle}$ to denote the i-th frontal slice of $\mathcal{A}$. The ℓ_2 norm of a vector $\mathbf{a}$ is denoted by $\|\mathbf{a}\|_2$, and for a matrix $\mathbf{A}$, $\sigma_j(\mathbf{A})$ denotes its j-th largest singular value. The ℓ_0 norm, denoted by $\|\cdot\|_0$, is defined as the number of nonzero entries. The nuclear norm, spectral norm, and Frobenius norm of matrix $\mathbf{A}$ are defined by $\|\mathbf{A}\|_* = \sum_{j=1}^{\min\{n_1,n_2\}} \sigma_j(\mathbf{A}), \|\mathbf{A}\| = \sigma_1(\mathbf{A}), \|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})}$, respectively. For two tensors $\mathcal{A}$ and $\mathcal{B}$, their inner product is defined by $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{i=1}^{n_3} \mathrm{tr}((\mathcal{A}^{\langle i \rangle})' \mathcal{B}^{\langle i \rangle})$, and $\mathrm{vec}(\mathcal{A})$ stacks the entries of $\mathcal{A}$ into a vec-

tor in lexicographical order. The tensor ℓ_∞ norm is $\|\mathcal{A}\|_\infty = \max_{i,j,l} |\mathcal{A}_{ijl}|$, the tensor ℓ_1 norm is $\|\mathcal{A}\|_1 = \sum_{i,j,l} |\mathcal{A}_{ijl}|$, and the tensor Frobenius norm is $\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$. Let $\delta_D(\cdot)$ denote the indicator function of a set D, defined by $\delta_D(x) = 0$ if $x \in D$ and $\delta_D(x) = +\infty$ otherwise.

The k-mode product of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ with a matrix $\mathbf{A} \in \mathbb{R}^{J \times n_k}$ is denoted by $\mathcal{X} \times_k \mathbf{A}$ and is of size $n_1 \times \cdots \times n_{k-1} \times J \times n_{k+1} \times \cdots \times n_d$ (Kolda and Bader, 2009). The point-wise manner is defined as $(\mathcal{X} \times_k \mathbf{A})_{i_1 \cdots i_{k-1} j i_{k+1} \cdots i_d} = \sum_{i_k=1}^{n_k} \mathcal{X}_{i_1 \cdots i_d} \mathbf{A}_{j i_k}$. The mode-$k$ matricization of a tensor $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ is denoted by $\mathcal{X}_{(k)}$ and arranges the mode-k fibers to be the columns of the resulting matrix. Moreover, $\mathcal{Y} = \mathcal{X} \times_k \mathbf{A}$ is equivalent to $\mathcal{Y}_{(k)} = \mathbf{A} \mathcal{X}_{(k)}$.

2. Sparse Tucker Decomposition with Graph Regularization

Let each frontal slice of $\mathcal{W} \in \mathbb{R}^{m \times m \times p}$ be $\mathcal{W}^{\langle i \rangle} = \mathbf{W}_i, i = 1, \ldots, p$. Then problem (1.1) can be rewritten as

$$\mathbf{y}_t = (\mathbf{W}_1, \ldots, \mathbf{W}_p) \begin{pmatrix} \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p} \end{pmatrix} + \boldsymbol{\varepsilon}_t = \mathcal{W}_{(1)} \mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{2.2}$$

where $\mathbf{x}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})' \in \mathbb{R}^{mp \times 1}$. Therefore, the Tucker decomposition of $\mathcal{W}$ can be employed to explore the low-rankness of different modes in the

VAR model, i.e.,

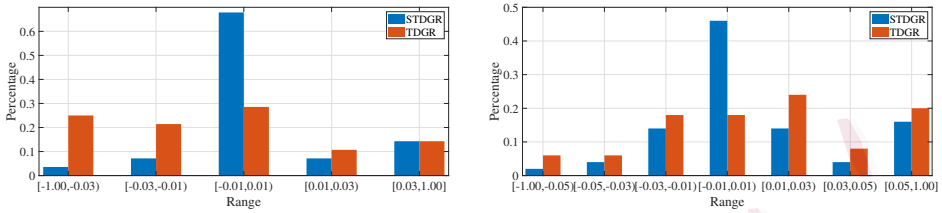
$$\mathcal{W} = \mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3 := [\![\mathcal{G}; \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!], \tag{2.3}$$

where $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor, $\mathbf{A}_1 \in \mathbb{R}^{m \times r_1}$, $\mathbf{A}_2 \in \mathbb{R}^{m \times r_2}$ and $\mathbf{A}_3 \in \mathbb{R}^{p \times r_3}$ are the response, predictor and temporal factor matrices, respectively. Consequently, problem (2.2) can be expressed as

$$\mathbf{y}_t = \mathcal{W}_{(1)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t = (\mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3)_{(1)}\mathbf{x}_t + \boldsymbol{\varepsilon}_t.$$

Moreover, in order to obtain a unique Tucker decomposition of $\mathcal{W}$, we consider the column orthogonal constraints on the factor matrices (Kolda and Bader, 2009), i.e., $\mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_{r_i}, i = 1, 2, 3$. Note that $\mathbf{A}_1'\mathbf{y}_t = \mathcal{G}_{(1)}(\mathbf{A}_3 \otimes \mathbf{A}_2)'\mathbf{x}_t + \mathbf{A}_1'\boldsymbol{\varepsilon}_t$. Let $\mathbf{A}_1'\mathbf{y}_t := \mathbf{f}_t = (f_{1t}, \ldots, f_{r_1 t})' \in \mathbb{R}^{r_1}$, which denotes r_1 response factors across m variables of $\mathbf{y}_t$. Additionally, $f_{it} = \sum_{j=1}^{m}(\mathbf{A}_1)_{ji}y_{jt}$ is the i-th response factor. If $(\mathbf{A}_1)_{ji} = 0$, then f_{it} is irrelevant to y_{jt}. In this case, $\mathbf{A}_1$ can be interpreted as the loadings of the response factors. Similarly, $\mathbf{A}_2$ and $\mathbf{A}_3$ can be interpreted as the loadings of the predictor and temporal factors, respectively, see Wang et al. (2022) for more discussions.

In the Tucker decomposition, the elements of the core tensor $\mathcal{G}$ reflect the interactions and connections between the components (columns) in different mode matrices, which only keeps the most significant connections between the components in different modes (response, predictor and tem-



(a) CCD (b) PeMS04

Figure 1: Distributions of the elements of $\mathcal{G}$ obtained by non-sparse model (TDGR) and sparse model (STDGR) for the CCD and PeMS04 datasets, respectively.

poral). Therefore, we impose the sparsity constraint on the core tensor. In Figure 1, we demonstrate the effectiveness of the sparsity constraint on the core tensor by numerical examples on the CCD and PeMS04 datasets (see the details in Section S5 of the supplementary materials and Section 4.4). In the figure, we show the distributions of the elements of $\mathcal{G}$ obtained by sparse model (STDGR) and non-sparse model (TDGR). It is clear that many elements of $\mathcal{G}$ are close to zero (in $[-0.01, 0.01)$), which illustrates the sparsity pattern of the estimated core tensor. The sparsity of the core tensor reduces redundant interactions among response, predictor, and temporal components and provides a more parsimonious representation of the tensor autoregressive structure.

Notice that the coefficient matrices $\mathbf{A}_i$ are the low-dimensional representation of the sampled data in the response, predictor or temporal mode,

which can reflect the similarity of sampled data in the low-dimensional subspace. Hence, we utilize the graph prior to each factor matrix of Tucker decomposition of $\mathbf{A}_i$ to characterize the similarity of the sampled data along the response, predictor and temporal modes for VAR.

In this paper, we propose a sparse Tucker decomposition model combined with graph regularization for the factor matrices (called STDGR for short) as follows:

$$\begin{aligned}\min_{\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3} \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t-\left(\mathcal{G}\times_1\mathbf{A}_1\times_2\mathbf{A}_2\times_3\mathbf{A}_3\right)_{(1)}\mathbf{x}_t\right\|_2^2+\beta\|\mathcal{G}\|_1 \\ +\sum_{i=1}^{3}\alpha_i\operatorname{tr}\left(\mathbf{A}_i'\mathbf{L}_i\mathbf{A}_i\right) \\ \text{s.t. } \|\mathcal{G}\|_\infty\le c,\mathbf{A}_i'\mathbf{A}_i=\mathbf{I}_{r_i},i=1,2,3,\end{aligned} \tag{2.4}$$

where $\beta,\alpha_i,c>0$ are given parameters and $\mathbf{L}_i$ is the graph Laplacian matrix, $i=1,2,3$. Here we note that for any matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, its graph Laplacian regularization is defined as (Cai et al., 2011)

$$\sum_{i,j=1}^{m}\frac{1}{2}z_{ij}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|_2^2=\operatorname{tr}\left(\mathbf{A}'\mathbf{L}\mathbf{A}\right), \tag{2.5}$$

where $\mathbf{a}_i$ denotes the i-th row of $\mathbf{A}$,

$$\mathbf{L}=\mathbf{D}-\mathbf{Z} \tag{2.6}$$

represents the Laplacian matrix with each element of $\mathbf{Z}$ being z_{ij}, and $\mathbf{D}\in\mathbb{R}^{m\times m}$ is a diagonal matrix with the i-th diagonal element being

$\sum_{j=1}^{m} z_{ij}, i = 1, \ldots, m$ (Chung, 1997). In practice, $\mathbf{L}_i$ are given or constructed in advance. We will discuss the detailed construction of $\mathbf{L}_i$ in Section 4.2.

To prevent the spiky coefficient tensor in the VAR process, we assume that the core tensor $\mathcal{G}$ is upper bounded, that is, $\|\mathcal{G}\|_\infty \leq c$. The graph Laplacian regularization on the factor matrices in model (2.4) is used to explore the local consistency of the low-dimensional subspace of the coefficient tensor. The factor matrices $\mathbf{A}_1, \mathbf{A}_2$, and $\mathbf{A}_3$ represent the low-dimensional subspace loadings of $\mathcal{W}$ along the response, predictor, and temporal modes, respectively (Wang et al., 2022). Conversely, it can reflect the similarity of the sampled data for VAR. For traffic data as an example, adjacent sensors show similar patterns and present temporal correlation properties and can exhibit similar low-dimensional loadings (Cai et al., 2011), which can be effectively captured by imposing graph Laplacian regularization on the response loading $\mathbf{A}_1$ and the predictor loading $\mathbf{A}_2$. The Laplacian regularization on the temporal loading matrix $\mathbf{A}_3$ can capture temporal similarity across p time lags, which facilitates adjacent lags to have similar subspace loadings. Moreover, the sparse Tucker decomposition in model (2.4) is used to reduce the dimension and can address the over-parameterization issue in VAR. In fact, the parameters in VAR with sparse Tucker decomposition are

Statistica Sinica: Newly accepted Paper

only $mr_1+mr_2+pr_3+s$, where s denotes the number of nonzero entries of the core tensor. This is less than that of model (1.1) (i.e., m^2p) or the reduced-rank model (Negahban and Wainwright, 2011) (i.e., $(mp+m-r_1)r_1$).

Remark 1. Compared with the model in Wang et al. (2022), the model in (2.4) used the sparse core tensor in Tucker decomposition, which can reduce the parameters of the transition tensor further. Moreover, the graph regularization is employed in model (2.4) to characterize the local consistency of the factor matrices, which is capable of reflecting the similarity of the sampled data in response, predictor and temporal modes. Besides, the model (2.4) jointly considers tensor low-rank structure, sparse latent interactions, and graph information within a unified tensor VAR model.

2.1 Non-asymptotic Error Bound

In this subsection, we establish the non-asymptotic error bound of the estimator of model (2.4). Denote the optimal solution of (2.4) by $(\widehat{\mathcal{G}}, \widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2, \widehat{\mathbf{A}}_3)$. Let $\widehat{\mathcal{W}} = \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}}_1 \times_2 \widehat{\mathbf{A}}_2 \times_3 \widehat{\mathbf{A}}_3$. To derive the non-asymptotic error bound of $\widehat{\mathcal{W}}$, we first make the following assumptions.

Assumption 1. The matrix polynomial $\mathbf{W}(z) = \mathbf{I}_m - \sum_{i=1}^{p} \mathbf{W}_i z^i, z \in \mathbb{C}$ satisfies $\det(\mathbf{W}(z)) \neq 0$ for all $|z| = 1$, where $\mathbb{C}$ denotes the complex field and $\mathbf{W}_i$ is defined in (1.1).

Statistica Sinica: Newly accepted Paper

Assumption 2. The error vector $\{\boldsymbol{\varepsilon}_t\}$ in (1.1) are i.i.d. Gaussian random vectors with mean zero and positive definite covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$.

Assumption 3. The core tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ has at most s nonzero entries. And define $\tau = 1 - s/(r_1 r_2 r_3) \in [0, 1]$, which represents the sparse level of $\mathcal{G}$, i.e., the proportion of zero entries.

Assumption 4. For $i = 1, 2, 3$, the nonzero singular values of $\mathcal{W}_{(i)}$ in (2.2) satisfy $\sigma_{j-1}^2(\mathcal{W}_{(i)}) - \sigma_j^2(\mathcal{W}_{(i)}) \geq \rho \sigma_{j-1}^2(\mathcal{W}_{(i)})$, $j = 2, \ldots, r_i$, where $\rho > 0$ is a constant.

Assumption 5. Denote $\Delta_{\mathbf{A}_i} := \widehat{\mathbf{A}}_i - \mathbf{A}_i$, for $i = 1, 2, 3$. Suppose that $\Delta_{\mathbf{A}_i}$ satisfies $\sum_{l=1}^{r_i} \sum_{j=1}^{n_i} (\Delta_{\mathbf{A}_i})_{jl}^2 \gg \sum_{l=1}^{r_i} \frac{1}{n_i} (\sum_{j=1}^{n_i} (\Delta_{\mathbf{A}_i})_{jl})^2$ with $n_1 = n_2 = m$ and $n_3 = p$.

Assumption 1 guarantees the stability of the VAR process (Basu and Michailidis, 2015). Assumption 2 is used to employ the concentration inequalities for VAR models (Basu and Michailidis, 2015). And Assumption 3 reflects the sparse level of core tensor $\mathcal{G}$. Assumption 4 further ensures that the singular values of each unfolding matrices of $\mathcal{W}$ are sufficiently separated, which can eliminate non-identifiability and enable us to determine the upper bound for perturbation errors from Lemma 5 of the supplementary materials. Finally, Assumption 5 is essential for the analysis of the

Statistica Sinica: Newly accepted Paper

graph regularization terms.

Let $\mathbf{W}'(z)$ denote the conjugate transpose of $\mathbf{W}(z)$. Following Basu and Michailidis (2015) and Wang et al. (2022), we define $\mu_{\min}(\mathbf{W}) := \min_{|z|=1} \lambda_{\min}(\mathbf{W}'(z)\mathbf{W}(z))$, $\mu_{\max}(\mathbf{W}) := \max_{|z|=1} \lambda_{\max}(\mathbf{W}'(z)\mathbf{W}(z))$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a matrix, respectively. Let $a_1, a_2 > 1, a, \hat{a}, a_3, c_0, c, b > 0$ be absolute constants independent of m, p, and T. Let $s_{\text{aux}} = s_1 m^2 p + s_2 m^2 p$, where $s_1 \geq \|\mathbf{Ag}\|_0/(m^2 p)$, $s_2 \geq \|\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_0/(m^2 p)$. Here $\mathbf{A} = \mathbf{A}_1 \otimes \mathbf{A}_3 \otimes \mathbf{A}_2, \mathbf{g} = \text{vec}(\mathcal{G}'_{(1)})$, and $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_1 \otimes \widehat{\mathbf{A}}_3 \otimes \widehat{\mathbf{A}}_2, \widehat{\mathbf{g}} = \text{vec}(\widehat{\mathcal{G}}'_{(1)})$. Let $\widehat{\mathbf{A}}\widehat{\mathbf{g}} - \mathbf{Ag} = \widehat{\Delta} = (\widehat{\Delta}'_1, \ldots, \widehat{\Delta}'_m)' \in \mathbb{R}^{m^2 p}$, where $\widehat{\Delta}_i \in \mathbb{R}^{mp}$. $\bar{p} = \max\{q_1 mp, q_2 mp, \ldots, q_m mp\}$, where $q_i \geq \frac{\|\widehat{\Delta}_i\|_0}{mp} \in [0, 1], i = 1, \ldots, m$. $\kappa = \sqrt{(\log(m^2 p))/T}$, $l_i = \|\mathbf{L}_i\|_F$, $\varpi = \lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})/\mu_{\max}(\mathbf{W})$, $\eta_i = (3\sigma_1(\mathcal{W}_{(i)}) + c\sqrt{r_1 r_2 r_3})\sqrt{\sum_{j=1}^{r_i} \frac{1}{\sigma_j^4(\mathcal{W}_{(i)})}}$, $\varsigma = \frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})/\mu_{\max}(\mathbf{W})}{\lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})/\mu_{\min}(\mathbf{W})}$, $\vartheta = \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}})(1 + \mu_{\max}(\mathbf{W})/\mu_{\min}(\mathbf{W}))$. $\lambda_2(\mathbf{L}_i) > 0$ denotes the second smallest eigenvalue of the Laplacian matrix $\mathbf{L}_i$, $i = 1, 2, 3$. Therefore, $\bar{p}$, s_{aux}, κ, $l_i, \varpi, \eta_i, \varsigma, \vartheta$ are problem-dependent quantities.

Now we give the error bound about the estimator of (2.4) in the following theorem.

Theorem 1. *Suppose that Assumptions 1-5 hold. If the sample size of the VAR model satisfies* $T \geq \max\{a_1 \log(m^2 p), \frac{8\varsigma^2}{a_3}\bar{p}\min(\log(mp), \log(21emp/\bar{p}))\}$, *and the regularization parameters satisfy* $\alpha_i \geq \frac{6c^2}{\lambda_2(\mathbf{L}_i)}\sqrt{s\log(m^2 p)/T}$, $\beta \geq$

Statistica Sinica: Newly accepted Paper

$2\pi\vartheta a_2\sqrt{\log(m^2p)/T}$, *then with probability at least* $1-6\exp(-a\log(m^2p))-2\exp(-\bar{p}\min(\log mp, \log(21emp/\bar{p})))$, *the estimator of (2.4) satisfies*

$$\|\widehat{\mathcal{W}}-\mathcal{W}\|_F \leq \frac{4\beta(\sqrt{s_{\text{aux}}}+b\sqrt{s})+\hat{a}\kappa+c_0\sum_{i=1}^{3}\alpha_i l_i\eta_i}{\varpi+2\kappa\sqrt{s/(r_1r_2r_3)^2}}, \tag{2.7}$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathcal{W}}_{(1)}\mathbf{x}_t-\mathcal{W}_{(1)}\mathbf{x}_t\|_2^2 \leq \frac{\lambda_{\max}\left(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}\right)\left(4\beta(\sqrt{s_{\text{aux}}}+b\sqrt{s})+\hat{a}\kappa+c_0\sum_{i=1}^{3}\alpha_i l_i\eta_i\right)^2}{\mu_{\min}(\mathbf{W})\left(\varpi+2\kappa\sqrt{s/(r_1r_2r_3)^2}\right)^2}. \tag{2.8}$$

The numerator of the Frobenius error bound in (2.7) can be decomposed into three components: $4\beta(\sqrt{s_{\text{aux}}}+b\sqrt{s})+\hat{a}\kappa+c_0\sum_{i=1}^{3}\alpha_i l_i\eta_i$. For the first component, $4\beta(\sqrt{s_{\text{aux}}}+b\sqrt{s})$ is associated with the sparsity structure of the model. Since $\beta \geq \sqrt{\log(m^2p)/T}$, this component has order $O((\sqrt{s_{\text{aux}}}+\sqrt{s})\sqrt{\log(m^2p)/T})$. Thus, the estimation error decreases at the $T^{-1/2}$ rate up to logarithmic factors and increases with the effective sparsity level. By the definition of κ, the second component has order $O(\sqrt{\log(m^2p)/T})$, which represents the stochastic fluctuation arising from the finite sample. Importantly, the dependence on the ambient dimensions m and p in this term is only logarithmic. The third component $c_0\sum_{i=1}^{3}\alpha_i l_i\eta_i$ is induced by the graph regularization. Here, $l_i=\|\mathbf{L}_i\|_F$ measures the magnitude of the graph Laplacian, whereas η_i captures the geometry of the Tucker factors through the ranks (r_1,r_2,r_3) and the singular values of the

Statistica Sinica: Newly accepted Paper

tensor unfoldings of the underlying tensor. Hence, this term reflects the interaction between the imposed graph structure and the underlying low Tucker rank structure. Moreover, the algebraic connectivity $\lambda_2(\mathbf{L}_i)$ determines the lower bound of the graph regularization parameters α_i. A larger $\lambda_2(\mathbf{L}_i)$ indicates a better-connected graph structure and leads to a smaller required regularization parameter, which reduces the graph-related contribution $c_0 \sum_{i=1}^{3} \alpha_i l_i \eta_i$ in the estimation error bound. Therefore, informative and well-connected graphs can improve estimation accuracy by providing stronger structural constraints while avoiding excessive regularization bias.

The Tucker ranks do not appear only as an isolated dimension factor. They enter the bound through both η_i and the denominator: $\eta_i = (3\sigma_1(\mathcal{W}_{(i)}) + c\sqrt{r_1 r_2 r_3})\Big(\sum_{j=1}^{r_i} \frac{1}{\sigma_j^4(\mathcal{W}_{(i)})}\Big)^{1/2}$ and $\varpi + 2\kappa\sqrt{\frac{s}{(r_1 r_2 r_3)^2}}$. Therefore, the theoretical guarantee explicitly reflects the complexity of the Tucker representation. In general, smaller Tucker ranks lead to a simpler low-rank structure and a more favorable error bound, provided that the corresponding singular values are sufficiently well separated from zero.

In Theorem 1, we establish a non-asymptotic error bound for model (2.4). When the Tucker rank (r_1, r_2, r_3) is fixed, the inequality in (2.7) shows that $\widehat{\mathcal{W}}$ is a consistent estimator as $T \to \infty$. In fact, the estimation and prediction error bounds in (2.7) and (2.8) are given by $O_p((\sqrt{s} +$

Statistica Sinica: Newly accepted Paper

$\sqrt{s_{\text{aux}}})\sqrt{\log(m^2p)/T})$ and $O_p((\sqrt{s}+\sqrt{s_{\text{aux}}})^2\log(m^2p)/T)$, respectively.

Remark 2. If the additional sparsity condition $s_{\text{aux}} \leq s$ holds, then the error bound in (2.7) reduces to

$$O_p\left(\sqrt{s\log(m^2p)/T}\right). \tag{2.9}$$

This condition should be interpreted as a favorable sparsity regime rather than a general requirement. Since $s_{\text{aux}} = (s_1+s_2)m^2p$, with $s_1 \geq \|\mathbf{Ag}\|_0/(m^2p)$, $s_2 \geq \|\widehat{\mathbf{A}}\widehat{\mathbf{g}}\|_0/(m^2p)$, the magnitude of s_{aux} depends on how the effective sparsity changes with the ambient dimension. When the factor matrices and core tensor are sufficiently sparse, s_{aux} can remain comparable to or smaller than s, and the above simplified rate applies. In contrast, when the factor matrices and core tensor become dense as the dimension increases, this condition may fail, and the full error bound in Theorem 1 should be used. Under this favorable sparsity regime, the resulting rate provides a convenient basis for comparison with existing estimation error bounds for high-dimensional VAR models.

- Basu and Michailidis (2015) studied the estimation of stationary Gaussian VAR(p) model with the sparse transition matrix. The corresponding Lasso estimator is given by $\widehat{\mathcal{W}}_{\text{Lasso}} = \arg\min_{\mathcal{W}} \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathcal{W}_{(1)}\mathbf{x}_t\|_2^2 + \lambda_{\text{Lasso}}\|\mathcal{W}_{(1)}\|_1$. It follows from (Basu and Michailidis, 2015,

Statistica Sinica: Newly accepted Paper

Proposition 4.1) that the estimator $\widehat{\mathcal{W}}_{\text{Lasso}}$ satisfies $\|\widehat{\mathcal{W}}_{\text{Lasso}} - \mathcal{W}\|_F = O_p(\sqrt{k_0 \log(m^2 p)/T})$, where $k_0 = \|\mathcal{W}\|_0$. In this case, it can be seen that the error bound in Theorem 1 is smaller than that of the Lasso estimator in Basu and Michailidis (2015) when $s \leq k_0$. In general, the number of nonzero entries of the core tensor is much smaller than that of the full tensor in practice.

- Negahban and Wainwright (2011) proposed to use the nuclear norm (NN) to characterize the low-rankness of the transition matrix in the VAR model, but their analysis was restricted to the VAR(1) case. We extend the VAR(1) model to the VAR(p) case, yielding

$$\widehat{\mathcal{W}}_{\text{NN}} = \arg\min \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{y}_t - \mathcal{W}_{(1)} \mathbf{x}_t\|_2^2 + \lambda_{\text{NN}} \|\mathcal{W}_{(1)}\|_*. \tag{2.10}$$

Furthermore, it follows from Remark 10 of Wang et al. (2022) that $\|\widehat{\mathcal{W}}_{\text{NN}} - \mathcal{W}\|_F = O_p(\sqrt{r_1 mp/T})$, where r_1 is the rank of mode-1 unfolding matrix of $\mathcal{W}$. Therefore, if $s/r_1 \leq mp/\log(m^2 p)$, the estimation error bound of the nuclear norm method is larger than the error bound in (2.9). This condition can be satisfied easily in real-world scenarios since s is smaller than $r_1 r_2 r_3$, r_1, r_2 are much smaller than m, and r_3 is much smaller than p.

- Recently, Wang et al. (2022) proposed an ℓ_1-penalized sparse higher-

order reduced-rank estimator for VAR, where the coefficient tensor $\mathcal{W}$ has the formulation of high-order singular value decomposition (HOSVD) and the factor matrices are sparse. When each column of $\mathbf{A}_i$ has at most s_i nonzero entries, $i = 1, 2, 3$, the resulting estimator, denoted by $\widehat{\mathcal{W}}_{\text{SHORR}}$, satisfies $\|\widehat{\mathcal{W}}_{\text{SHORR}} - \mathcal{W}\|_F = O_p\big(\sqrt{s_1 s_2 s_3 \log(m^2 p)/T}\big)$ (Wang et al., 2022, Theorem 2). Therefore, if $s \leq s_1 s_2 s_3$, the error bound of our estimator in (2.9) is smaller than that in Wang et al. (2022). Here the condition $s \leq s_1 s_2 s_3$ can be easily satisfied in real-world applications since s is much smaller than $r_1 r_2 r_3$, $1 \leq s_i \leq m, i = 1, 2, 1 \leq s_3 \leq p$, r_i is much smaller than $m, i = 1, 2$, r_3 is much smaller than p, and s_i is not too small in general, $i = 1, 2, 3$.

3. Proximal Alternating Linearized Minimization Algorithm

In this section, we use the well-known variable-splitting and penalty techniques in optimization to solve model (2.4). Specifically, three auxiliary variables $\mathbf{U}_i$ are introduced to transfer $\mathbf{A}_i$ out of the graph Laplacian regularization term, and the difference between $\mathbf{U}_i$ and $\mathbf{A}_i$ is penalized,

$i = 1, 2, 3$, yielding the following optimization model:

$$\begin{aligned}\min_{\substack{\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\\ \mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3}} & \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G}\times_1\mathbf{A}_1\times_2\mathbf{A}_2\times_3\mathbf{A}_3)_{(1)}\,\mathbf{x}_t\right\|_2^2 + \beta\|\mathcal{G}\|_1 \\ & + \sum_{i=1}^{3}\alpha_i\,\mathrm{tr}\left(\mathbf{U}_i'\mathbf{L}_i\mathbf{U}_i\right) + \sum_{i=1}^{3}\frac{\gamma_i}{2}\left\|\mathbf{U}_i - \mathbf{A}_i\right\|_F^2 \\ \text{s.t. } & \|\mathcal{G}\|_\infty \le c, \mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_{r_i}, i = 1, 2, 3,\end{aligned} \tag{3.11}$$

where $\gamma_i > 0$ are given constants. By using the definition of the indicator function, model (3.11) can be reformulated into the following unconstrained optimization problem:

$$\begin{aligned}\min_{\substack{\mathcal{G},\mathbf{A}_1,\mathbf{A}_2,\mathbf{A}_3,\\ \mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3}} & \frac{1}{2T}\sum_{t=1}^{T}\left\|\mathbf{y}_t - (\mathcal{G}\times_1\mathbf{A}_1\times_2\mathbf{A}_2\times_3\mathbf{A}_3)_{(1)}\,\mathbf{x}_t\right\|_2^2 + \beta\|\mathcal{G}\|_1 \\ & + \sum_{i=1}^{3}\alpha_i\,\mathrm{tr}\left(\mathbf{U}_i'\mathbf{L}_i\mathbf{U}_i\right) + \sum_{i=1}^{3}\frac{\gamma_i}{2}\left\|\mathbf{U}_i - \mathbf{A}_i\right\|_F^2 + \sum_{i=1}^{3}\delta_{\mathfrak{B}_i}(\mathbf{A}_i) + \delta_{\mathcal{D}}(\mathcal{G}),\end{aligned} \tag{3.12}$$

where $\mathfrak{B}_i := \{\mathbf{A}_i : \mathbf{A}_i'\mathbf{A}_i = \mathbf{I}_{r_i}\}, i = 1, 2, 3, \mathcal{D} := \{\mathcal{G} : \|\mathcal{G}\|_\infty \le c\}$.

Define $\Psi(\mathcal{G}, \mathbf{A}_i, \mathbf{U}_i) = \frac{1}{2T}\sum_{t=1}^{T}\|\mathbf{y}_t - (\mathcal{G}\times_1\mathbf{A}_1\times_2\mathbf{A}_2\times_3\mathbf{A}_3)_{(1)}\,\mathbf{x}_t\|_2^2 + \sum_{i=1}^{3}\frac{\gamma_i}{2}\|\mathbf{U}_i - \mathbf{A}_i\|_F^2$. Note that Ψ is a smooth function with respect to each variable. We now apply a PALM algorithm to solve (3.12). The main idea of PALM is to linearize each variable of the coupled function Ψ and add a proximal quadratic term (Bolte et al., 2014). Then we update each variable alternatingly. In this case, each subproblem can be solved in a closed-form.

The iterative framework of the PALM algorithm is given as follows:

$$
\begin{cases}
\mathcal{G}^{k+1} \in \arg\min\limits_{\mathcal{G}} \left\langle \nabla_{\mathcal{G}} \Psi(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathcal{G} - \mathcal{G}^k \right\rangle + \beta \|\mathcal{G}\|_1 \\
\qquad\qquad + \delta_{\mathcal{D}}(\mathcal{G}) + \dfrac{\rho_1}{2} \|\mathcal{G} - \mathcal{G}^k\|_F^2, \\
\mathbf{A}_1^{k+1} \in \arg\min\limits_{\mathbf{A}_1} \left\langle \nabla_{\mathbf{A}_1} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{A}_1 - \mathbf{A}_1^k \right\rangle \\
\qquad\qquad + \delta_{\mathfrak{B}_1}(\mathbf{A}_1) + \dfrac{\rho_2}{2} \|\mathbf{A}_1 - \mathbf{A}_1^k\|_F^2, \\
\mathbf{A}_2^{k+1} \in \arg\min\limits_{\mathbf{A}_2} \left\langle \nabla_{\mathbf{A}_2} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{A}_2 - \mathbf{A}_2^k \right\rangle \\
\qquad\qquad + \delta_{\mathfrak{B}_2}(\mathbf{A}_2) + \dfrac{\rho_3}{2} \|\mathbf{A}_2 - \mathbf{A}_2^k\|_F^2, \\
\mathbf{A}_3^{k+1} \in \arg\min\limits_{\mathbf{A}_3} \left\langle \nabla_{\mathbf{A}_3} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{A}_3 - \mathbf{A}_3^k \right\rangle \\
\qquad\qquad + \delta_{\mathfrak{B}_3}(\mathbf{A}_3) + \dfrac{\rho_4}{2} \|\mathbf{A}_3 - \mathbf{A}_3^k\|_F^2, \\
\mathbf{U}_1^{k+1} \in \arg\min\limits_{\mathbf{U}_1} \left\langle \nabla_{\mathbf{U}_1} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{U}_1 - \mathbf{U}_1^k \right\rangle \\
\qquad\qquad + \alpha_1 \operatorname{tr}(\mathbf{U}_1' \mathbf{L}_1 \mathbf{U}_1) + \dfrac{\rho_5}{2} \|\mathbf{U}_1 - \mathbf{U}_1^k\|_F^2, \\
\mathbf{U}_2^{k+1} \in \arg\min\limits_{\mathbf{U}_2} \left\langle \nabla_{\mathbf{U}_2} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^k, \mathbf{U}_3^k), \mathbf{U}_2 - \mathbf{U}_2^k \right\rangle \\
\qquad\qquad + \alpha_2 \operatorname{tr}(\mathbf{U}_2' \mathbf{L}_2 \mathbf{U}_2) + \dfrac{\rho_6}{2} \|\mathbf{U}_2 - \mathbf{U}_2^k\|_F^2, \\
\mathbf{U}_3^{k+1} \in \arg\min\limits_{\mathbf{U}_3} \left\langle \nabla_{\mathbf{U}_3} \Psi(\mathcal{G}^{k+1}, \mathbf{A}_1^{k+1}, \mathbf{A}_2^{k+1}, \mathbf{A}_3^{k+1}, \mathbf{U}_1^{k+1}, \mathbf{U}_2^{k+1}, \mathbf{U}_3^k), \mathbf{U}_3 - \mathbf{U}_3^k \right\rangle \\
\qquad\qquad + \alpha_3 \operatorname{tr}(\mathbf{U}_3' \mathbf{L}_3 \mathbf{U}_3) + \dfrac{\rho_7}{2} \|\mathbf{U}_3 - \mathbf{U}_3^k\|_F^2,
\end{cases}
\tag{3.13}
$$

where $\rho_i > 0$ are given constants, $i = 1, \ldots, 7$. The gradients of Ψ with respect to $\mathcal{G}$, $\mathbf{A}_i$ and $\mathbf{U}_i$, as well as the detailed solutions for each subproblem

in (3.13) are provided in Section S3 of the supplementary materials.

It is shown that the gradient of $\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ is Lipschitz continuous with respect to each variable in Lemmas 8-11 in the supplementary material. Let L_1 be the Lipschitz constant of $\nabla_{\mathcal{G}}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ with respect to $\mathcal{G}$, L_2 be the Lipschitz constant of $\nabla_{\mathbf{A}_1}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ with respect to $\mathbf{A}_1$, L_3 be the Lipschitz constant of $\nabla_{\mathbf{A}_2}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ with respect to $\mathbf{A}_2$, and L_4 be the Lipschitz constant of $\nabla_{\mathbf{A}_3}\Psi(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ with $\mathbf{A}_3$. Next, we establish the global convergence of the PALM algorithm.

Theorem 2. *Let the sequence* $\{(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)\}$ *be generated by PALM Algorithm. Suppose that* $\rho_1 > L_1, \rho_2 > L_2, \rho_3 > L_3, \rho_4 > L_4, \rho_5 > \gamma_1, \rho_6 > \gamma_2, \rho_7 > \gamma_3$. *Then the sequence* $\{(\mathcal{G}^k, \mathbf{A}_1^k, \mathbf{A}_2^k, \mathbf{A}_3^k, \mathbf{U}_1^k, \mathbf{U}_2^k, \mathbf{U}_3^k)\}$ *converges to a critical point of (3.12).*

In Lemmas 8-11 of the supplementary material, we show that $L_1 = \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2, L_j = r_1 r_2 r_3 c^2 L_1 + \gamma_{j-1}, j = 2, 3, 4$. Moreover, the assumptions in Theorem 2 are very weak. We only need $\rho_1 > L_1, \rho_2 > L_2, \rho_3 > L_3, \rho_4 > L_4, \rho_5 > \gamma_1, \rho_6 > \gamma_2, \rho_7 > \gamma_3$, where these conditions are used to guarantee the sufficient decrease of the objective function in (3.12). In the experiments, we will give ρ_i in detail for the implementation of PALM.

Remark 3. Theorem 1 establishes a statistical error bound for a global

Statistica Sinica: Newly accepted Paper

minimizer of the problem in (2.4), whereas Theorem 2 only guarantees that PALM converges to a critical point. Since the problem (2.4) is nonconvex, we do not prove that the PALM converges to a global minimizer at present. Therefore, the statistical error bound does not apply to every estimator generated by PALM.

4. Numerical Experiments

In this section, some numerical experiments are conducted to demonstrate the effectiveness of the proposed STDGR model for high-dimensional time series forecasting. In addition, more experiment results and discussions about parameters analysis and different constructions of graphs are provided in Sections S5 and S6 of the supplementary materials. We compare the STDGR with the following five models: the Lasso type method for time series models (Lasso, Basu and Michailidis (2015)), nuclear norm minimization of the transition matrix (NNM, Negahban and Wainwright (2011)), low-rank tensor learning via Tucker rank constraint (LRTLT, Bahadori et al. (2014)), time-varying autoregression with low-rank tensors (TVART, Harris et al. (2021)), sparse higher-order reduced-rank VAR (SHORR, Wang et al. (2022)). To ensure a fair comparison, all competing approaches are carefully tuned based on the same criterion and under the same experimental

Statistica Sinica: Newly accepted Paper

settings. For each method, the tuning parameters are selected from their recommended ranges according to the prediction performance on the validation data. Moreover, the multiple-origin one-step-ahead forecast is used in the experiments. All experiments are performed in MATLAB R2020b with a 12-core, 2.6 GHz Intel Core CPU and 16 GB of RAM.

The mean squared error (MSE) is used to evaluate the forecasting accuracy for high-dimensional time series data, which is defined as

$$\text{MSE} := \frac{1}{mT_0}\sum_{t=1}^{T_0}\|\tilde{\mathbf{y}}_t - \hat{\mathbf{y}}_t\|_2^2,$$

where T_0 is the number of testing samples, $\tilde{\mathbf{y}}_t \in \mathbb{R}^m$ denotes the ground truth value and $\hat{\mathbf{y}}_t \in \mathbb{R}^m$ denotes the forecasting value.

For any $i \in \{1,2,3\}$, denote $\Lambda_1 = \frac{\|\mathcal{G}^{k+1}-\mathcal{G}^k\|_F}{\|\mathcal{G}^k\|_F}, \Lambda_{i+1} = \frac{\|\mathbf{A}_i^{k+1}-\mathbf{A}_i^k\|_F}{\|\mathbf{A}_i^k\|_F}$, $\Lambda_{i+4} = \frac{\|\mathbf{U}_i^{k+1}-\mathbf{U}_i^k\|_F}{\|\mathbf{U}_i^k\|_F}$. The PALM algorithm will be terminated if $\max_{1\le j\le 7}\{\Lambda_j\} \le 3\times 10^{-3}$ or the number of iterations reaches 200.

4.1 The Selection of Tucker Rank

In this subsection, we give the choice of Tucker rank in PALM. The NNM in (2.10) is used to get an initial estimator of our model, where the mode-1 unfolding of the transition tensor is used in the NNM model. The estimator in (2.10) is a consistent initial estimator (Negahban and Wainwright, 2011; Wang et al., 2022). We apply the ridge-type ratio estimator (Wang et al.,

Statistica Sinica: Newly accepted Paper

2022) to estimate the Tucker rank in PALM, which is defined as

$$\widehat{r}_i = \arg \min_{1 \le j \le n_i - 1} \frac{\sigma_{j+1}((\widehat{\mathcal{W}}_{\mathrm{NN}})_{(i)}) + \bar{c}}{\sigma_j((\widehat{\mathcal{W}}_{\mathrm{NN}})_{(i)}) + \bar{c}}, \tag{4.14}$$

where $i = 1, 2, 3$, $n_1 = n_2 = m, n_3 = p$, and $\bar{c} = o_p(\sqrt{r_1 mp/T})$. It follows from Theorem 3 in Wang et al. (2022) that the rank selection in (4.14) is consistent, i.e., $\mathbb{P}(\widehat{r}_1 = r_1, \widehat{r}_2 = r_2, \widehat{r}_3 = r_3) \to 1$ as $T \to \infty$. In the experiments, we set $\bar{c} = \sqrt{mp \log(T)/(50T)}$.

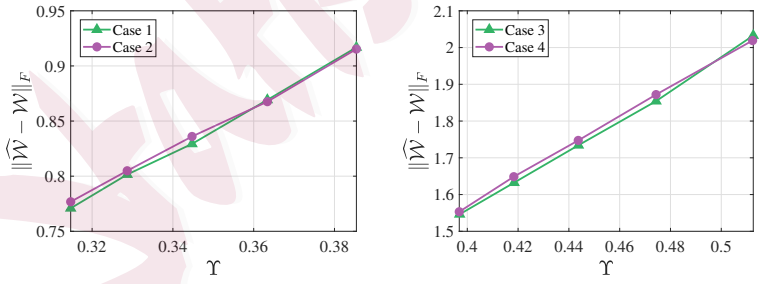
4.2 The Construction of Laplacian Matrices

The Laplacian matrices $\mathbf{L}_i$ in model (2.4), $i = 1, 2, 3$, are constructed as follows: First, we use the NNM model in (2.10) to get an initial estimator $\widehat{\mathcal{W}}_{\mathrm{NN}}$ and then take the HOSVD (Kolda and Bader, 2009) to obtain the factor matrices $\widehat{\mathbf{A}}_i$ with rank $\widehat{r}_i$, $i = 1, 2, 3$, where $\widehat{r}_i$ is computed via (4.14). Let $(\widehat{\mathbf{a}}_i)_l$ denote the l-th row of $\widehat{\mathbf{A}}_i$. Then the weight $(z_i)_{lt}$ in (2.5) between $(\widehat{\mathbf{a}}_i)_l$ and $(\widehat{\mathbf{a}}_i)_t$ is given by $(z_i)_{lt} = e^{-d_{lt}^2/(2\epsilon^2)}$ and $d_{lt}^2 = \|(\widehat{\mathbf{a}}_i)_l - (\widehat{\mathbf{a}}_i)_t\|_2^2$, where we set $\epsilon = 0.2$ in all experiments for simplicity. Then the Laplacian matrix $\mathbf{L}_i$ is given by $\mathbf{L}_i = \mathbf{D}_i - \mathbf{Z}_i$, where the (l, t)-th element of $\mathbf{Z}_i$ is $(z_i)_{lt}$ and $\mathbf{D}_i$ is a diagonal matrix with the l-th diagonal element being the sum of elements of the l-th row of $\mathbf{Z}_i$. Note that the initial estimator of NNM is constructed by only the training data and the resulting Gaussian-kernel graphs are constructed by the initial estimator.

Statistica Sinica: Newly accepted Paper

4.3 Synthetic Data

In this subsection, we first verify the non-asymptotic error bound of the estimator of STDGR via synthetic data. The reported estimation error is averaged over 10 independent replications in the experiments. We set $\gamma_1 = \gamma_2 = \gamma_3$ simply and choose them from the set $\{0.001, 0.1, 10, 100\}$ to get the best performance. By Theorem 2, we simply set $\rho_i = \bar{a}_1 L_i$ for $i = 1, 2, 3, 4$ and $\rho_{j+4} = \bar{a}_2 \gamma_j$ for $j = 1, 2, 3$ to guarantee the convergence of PALM algorithm, where $\bar{a}_1$ is set as 1.1 and $\bar{a}_2$ is selected from the set $\{1.1, 3, 5, 10\}$. The parameter c defined in (2.4) is chosen from $\{1, 2\}$ and β is set to be 0.001. For simplicity, we set $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$ in all experiments, where $\bar{\alpha}$ is selected from the set $\{0.001, 0.1, 1\}$ to get the best performance.



(a) (b)

Figure 2: Estimation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ versus $\Upsilon = (\sqrt{s} + \sqrt{s_{\text{aux}}}) \sqrt{\frac{\log(m^2 p)}{T}}$ for different synthetic data. (a) $(m, p) = (50, 8)$. (b) $(m, p) = (80, 12)$.

Statistica Sinica: Newly accepted Paper

Four cases are discussed for the synthetic data. Specifically, for the core tensor, we choose a diagonal cube with superdiagonal elements being nonzero, and $\mathbf{A}_i$ is generated as the r_i left singular vectors of Gaussian random matrices. Other parameters about the dimensions and Tucker rank are set as follows: Case 1. $(m, p) = (50, 8)$, Tucker rank $(r_1, r_2, r_3) = (3, 3, 3)$, $\mathcal{G}_{111} = \mathcal{G}_{222} = \mathcal{G}_{333} = 2$. Case 2. $(m, p) = (50, 8)$, Tucker rank $(r_1, r_2, r_3) = (6, 6, 6)$, $\mathcal{G}_{111} = \mathcal{G}_{222} = \mathcal{G}_{333} = \mathcal{G}_{444} = \mathcal{G}_{555} = \mathcal{G}_{666} = 2$. Case 3. $(m, p) = (80, 12)$, Tucker rank $(r_1, r_2, r_3) = (4, 4, 4)$, $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}) = (2, 2, 2, 1)$. Case 4. $(m, p) = (80, 12)$, Tucker rank $(r_1, r_2, r_3) = (7, 7, 7)$, and $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}, \mathcal{G}_{555}, \mathcal{G}_{666}, \mathcal{G}_{777}) = (2, 2, 2, 1, 1, 1, 0.5)$. Moreover, $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_m)$. It follows from Theorem 1 that the error bound $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ is on the order of $O_p((\sqrt{s} + \sqrt{s_{\text{aux}}})\sqrt{\log(m^2 p)/T})$ for fixed Tucker rank, where we set $s_{\text{aux}} = s$. Let $\Upsilon = (\sqrt{s} + \sqrt{s_{\text{aux}}})\sqrt{\log(m^2 p)/T}$ and we choose different sample sizes T such as $\Upsilon = 0.314, 0.328, 0.344, 0.363, 0.385$ for Case 1 and Case 2, and $\Upsilon = 0.397, 0.418, 0.443, 0.474, 0.512$ for Case 3 and Case 4. In Figure 2, we plot $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ versus Υ for the four cases. It can be seen that the estimation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ generally increases linearly in Υ, and the two lines in each figure almost coincide, which demonstrates the theoretical finding in Theorem 1.

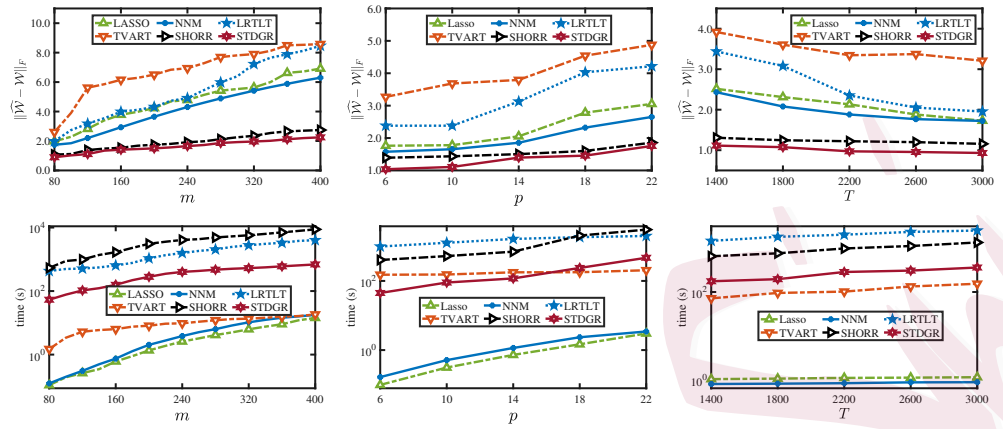
Additional synthetic data experiments are conducted to compare the es-

Statistica Sinica: Newly accepted Paper

timation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ and computational time (in seconds) of STDGR with those of Lasso, NNM, LRTLT, TVART, and SHORR under different dimensions m, lag orders p, and sample sizes T. For the noise term, we set $\boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, 0.25\mathbf{I}_m)$. Additionally, the factor matrices $\mathbf{A}_i$ are constructed as follows (Belkin and Niyogi, 2001): we first generate the weight matrices $\mathbf{Z}_i$ randomly and construct the corresponding Laplacian matrices as $\mathbf{L}_i = \mathbf{D}_i - \mathbf{Z}_i$, where $\mathbf{D}_i$ is the corresponding degree matrix given by (2.6). Then the columns of $\mathbf{A}_i$ are chosen as the r_i eigenvectors of $\mathbf{L}_i$ associated with its r_i smallest eigenvalues. In spectral clustering, the entries of $\mathbf{A}_i$ measure the similarities between data samples and clusters, as they indicate how likely data samples belong to specific clusters Wang et al. (2014). Therefore, this construction of $\mathbf{A}_i$ can reflect the similarity of the rows of $\mathbf{A}_i$. The Tucker rank is set as $(r_1, r_2, r_3) = (6, 6, 6)$ and the core tensor is set as a diagonal cube with nonzero superdiagonal elements $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}, \mathcal{G}_{444}, \mathcal{G}_{555}, \mathcal{G}_{666}) = (0.5, 1, 0.5, 1, 0.5, 1)$.

Three cases are considered. Case 5. $(p, T) = (14, 2200)$, where m is varied from 80 to 400 with step size 80. Case 6. $(m, T) = (120, 2200)$, where p is varied from 6 to 22 with step size 4. Case 7. $(m, p) = (120, 14)$, where T is varied from 1400 to 3000 with step size 400. The corresponding estimation error $\|\widehat{\mathcal{W}} - \mathcal{W}\|_F$ and computational time of different methods

Statistica Sinica: Newly accepted Paper



(a) (b) (c)

Figure 3: Estimation errors and computational time of different methods under varying dimensions, lag orders, and sample sizes. First row: Estimation error versus m, p, T. Second row: Computational time versus m, p, T. (a) Different m with $p = 14$ and $T = 2200$. (b) Different p with $m = 120$ and $T = 2200$. (c) Different T with $m = 120$ and $p = 14$.

are presented in Figure 3, we take the logarithm of the computational times to facilitate better visual comparisons across different methods.

It can be observed from Figure 3 that the estimation errors of all methods gradually increase as m or p increases. In contrast, the estimation errors gradually decrease as T increases. Moreover, STDGR achieves the lowest estimation error across the testing cases. And SHORR performs better than Lasso, NNM, LRTLT, and TVART in terms of estimation errors. Moreover, we can see from the second row of Figure 3 that the computational time re-

Statistica Sinica: Newly accepted Paper

quired by Lasso, NNM, STDGR, SHORR, and LRTLT increases as m or p increases, whereas the computational time of TVART changes only slightly. The computational time of LRTLT, TVART, SHORR, and STDGR increases as T increases, whereas the computational time of Lasso and NNM remains relatively stable. Although the computational time required by STDGR is more than Lasso, NNM, and TVART, the estimation error of STDGR is much smaller than these competition methods.

4.4 Traffic Flow Dataset

In this subsection, we test two traffic flow datasets to demonstrate the effectiveness of STDGR for VAR. The first one is the Hangzhou metro passenger flow dataset, which records incoming passenger flow of 80 metro stations over 25 days from January 1 to January 25, 2019, with a 10-minute resolution in Hangzhou, China. The dataset is publicly available at `https://tianchi.aliyun.com/competition/entrance/231708/information`. We discard the interval from 00:00 to 06:00 since metro service is unavailable at that time, retaining the remaining 108 time intervals per day. We select 2500 time intervals in this dataset. The second one is the PeMS04 dataset, which is a public traffic flow dataset collected by the California Transportation Agencies Performance Measurement System (PeMS) in the Bay Area

Statistica Sinica: Newly accepted Paper

of the United States. The PeMS04 dataset contains flow volumes of 307 sensors from January 1 to February 28, 2018, which are recorded every five minutes. As a result, there are 16992 time intervals and we select 7500 time intervals in the following experiments. The dataset is publicly available at `https://ieee-dataport.org/documents/pems03-and-pems04`. For the two traffic flow datasets, we just scale each dataset to $[0, 1]$.

For the traffic flow dataset, the order of the VAR(p) model is set as $p = 7$ for all methods, except for TVART, in which we set $p = 1$. The parameters $\gamma_1, \gamma_2, \gamma_3$ are set as the same in each case and chosen from the set $\{0.001, 0.01, 0.1, 1, 100\}$ to obtain the best forecasting performance in the validation data. We fix $c = 1$ and set $\beta = 0.001$. In all experiments, we set $\alpha_1 = \alpha_2 = \alpha_3 = \bar{\alpha}$, where $\bar{\alpha}$ is chosen from the set $\{0.001, 0.1\}$. According to Theorem 2, we set $\rho_i = \bar{a}_1 L_i$ for $i = 1, 2, 3, 4$ and $\rho_{j+4} = \bar{a}_2 \gamma_j$ for $j = 1, 2, 3$ to ensure the convergence of PALM. Here, $\bar{a}_1$ and $\bar{a}_2$ are chosen from the set $\{1.1, 1.3, 10\}$ to get the best forecasting performance in the validation data. In the two datasets, the last 25% samples of training data are used for validation in order to choose suitable parameters.

In Table 2, we report the prediction errors, computation time (in seconds), and number of iterations of different methods for the Hangzhou and PeMS04 datasets, where we use 70% samples for training and the remaining

Statistica Sinica: Newly accepted Paper

Table 2: MSE ($\times 10^{-3}$) of forecasting results, running time (in seconds) and iteration counts of different methods for the traffic flow dataset.

Dataset	Metric	Lasso	NNM	LRTLT	TVART	SHORR	STDGR
Hangzhou	MSE ($\times 10^{-3}$)	1.8872	1.7050	1.4353	4.0145	1.2174	0.6016
	Time (s)	1.92	5.10	4.31	14.77	75.54	24.48
	Iteration	34/200	65/100	11/100	64/500	22/100	47/200
PeMS04	MSE ($\times 10^{-3}$)	3.4640	2.7801	5.2871	5.2348	2.5712	2.0632
	Time (s)	11.51	20.18	743.12	1076.53	4587.20	1087.93
	Iteration	39/200	36/100	41/100	444/500	44/100	35/200

30% samples for testing in each dataset. Here the rows about the iteration are reported by number of iterations/the setting maximum number of iterations in the algorithm. It can be seen that the STDGR achieves the smallest prediction error compared with other methods for the two traffic flow datasets, which exhibits the superior forecasting performance of STDGR. Moreover, the SHORR performs better than Lasso, NNM, LRTLT, and TVART in terms of MSE. The running time required by STDGR is smaller than that required by SHORR for the two large-scale traffic datasets, which demonstrates the effectiveness and efficiency of STDGR.

Statistica Sinica: Newly accepted Paper

5. Concluding Remarks

In this paper, we have proposed an STDGR for high-dimensional VAR time series forecasting. By stacking the transition matrices into a third-order tensor, the Tucker decomposition is employed for the transition tensor to address the over-parameterization issue, which can restrict the parameter space along three directions simultaneously. Besides, the sparsity of the core tensor is enforced to further reduce the parameters. And the graph regularization is employed to characterize the local consistency of the factor matrices, which can characterize the similarity of the sampled data for VAR time series. A PALM algorithm is designed to solve the resulting model with global convergence guarantee. Numerical experiments on simulated and real data show the superior performance of the proposed method over several baseline models.

In the future, we are going to explore the tensor-valued autoregressive time series based on other types of tensor decomposition, which may explore the correlation of data better (cf. Chen et al. (2022); Wang et al. (2024)). Moreover, we will design an efficient algorithm to solve the global minimizer of the proposed model, which will be helpful to support the theory about the non-asymptotic error bound by numerical experiments.

Statistica Sinica: Newly accepted Paper

Supplementary Materials

The supplementary materials contain detailed solutions to the subproblems in PALM, the proofs of Theorems 1 and 2, additional numerical experiments for simulation and real data, and discussions about parameters analysis, choice of lag order, different initialization, choices of Tucker rank, constructions of graphs, ablation analysis, the extensions to uniform noise.

Acknowledgements

The authors would like to thank the Associate Editor and the two anonymous referees for their constructive suggestions and comments, which helped improve the quality of the paper. This research was partially supported by GDSTC: Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme UICR0800008-24, National Key Research and Development Program of China under Grant 2024YFE0202900, RGC GRF 12300125, Joint NSFC and RGC N-HKU769/21, Hubei Provincial Natural Science Foundation of China under Grant 2025AFB966, Fundamental Research Funds for the Central Universities under Grant XJ2026000601, and Academy of Frontier Interdisciplinary Research, Central China Normal University (JC2026PT-002).

Statistica Sinica: Newly accepted Paper

References

Bahadori, M. T., Q. R. Yu, and Y. Liu (2014). Fast multivariate spatio-temporal analysis via low rank tensor learning. *Advances in Neural Information Processing Systems 27*, 3491–3499.

Basu, S., X. Li, and G. Michailidis (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Trans. Signal Process. 67*(5), 1207–1222.

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

Belkin, M. and P. Niyogi (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pp. 585–591.

Bolte, J., S. Sabach, and M. Teboulle (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program. 146*(1-2), 459–494.

Cai, D., X. He, J. Han, and T. S. Huang (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell. 33*(8), 1548–1560.

Chen, R., D. Yang, and C.-H. Zhang (2022). Factor models for high-dimensional tensor time series. *J. Amer. Statist. Assoc. 117*(537), 94–116.

Chen, X., C. Zhang, X.-L. Zhao, N. Saunier, and L. Sun (2025). Forecasting sparse movement speed of urban road networks with nonstationary temporal matrix factorization. *Transp. Sci. 59*(3), 670–687.

Statistica Sinica: Newly accepted Paper

REFERENCES

Chung, F. R. (1997). *Spectral Graph Theory*, Volume 92. Providence: American Mathematical Society.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Econometrics 146*(2), 318–328.

Duchon, C. and R. Hale (2012). *Time Series Analysis in Meteorology and Climatology: An Introduction*. Oxford: Wiley-Blackwell.

Han, F., H. Lu, and H. Liu (2015). A direct estimation of high dimensional stationary vector autoregressions. *J. Mach. Learn. Res. 16*(1), 3115–3150.

Harris, K. D., A. Aravkin, R. Rao, and B. W. Brunton (2021). Time-varying autoregression with low-rank tensors. *SIAM J. Appl. Dyn. Syst. 20*(4), 2335–2358.

Hillar, C. J. and L.-H. Lim (2013). Most tensor problems are NP-hard. *J. ACM 60*(6), 45.

Kock, A. B., R. S. Pedersen, and J. R.-V. Sørensen (2026). Data-driven tuning parameter selection for high-dimensional vector autoregressions. *J. Amer. Statist. Assoc. 121*(553), 289–299.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Rev. 51*(3), 455–500.

Miao, K., P. C. Phillips, and L. Su (2023). High-dimensional VARs with common factors. *J. Econometrics 233*(1), 155–183.

Statistica Sinica: Newly accepted Paper

Negahban, S. and M. J. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist. 39*(2), 1069–1097.

Raskutti, G., M. Yuan, and H. Chen (2019). Convex regularization for high-dimensional multiresponse tensor regression. *Ann. Statist. 47*(3), 1554–1584.

Reinsel, G. (1983). Some results on multivariate autoregressive index models. *Biometrika 70*(1), 145–156.

Samadi, S. Y. and H. W. B. Herath (2024). Reduced-rank envelope vector autoregressive model. *J. Bus. Econom. Statist. 42*(3), 918–932.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica 48*(1), 1–48.

Tsay, R. S. (2005). *Analysis of Financial Time Series*. New York: John Wiley & Sons.

Wang, D. and R. S. Tsay (2023). Rate-optimal robust estimation of high-dimensional vector autoregressive models. *Ann. Statist. 51*(2), 846–877.

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

Wang, D., Y. Zheng, H. Lian, and G. Li (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *J. Amer. Statist. Assoc. 117*(539), 1338–1356.

Wang, H., C. Weng, and J. Yuan (2014). Multi-feature spectral clustering with minimax optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4106–4113.

Statistica Sinica: Newly accepted Paper

REFERENCES

Wang, Z., A. Safikhani, Z. Zhu, and D. S. Matteson (2023). Regularized estimation in high-dimensional vector auto-regressive models using spatio-temporal information. *Statist. Sinica 33*, 1271–1294.

Wong, K. C., Z. Li, and A. Tewari (2020). Lasso guarantees for β-mixing heavy-tailed time series. *Ann. Statist. 48*(2), 1124–1142.

Zhu, X. (2020). Nonconcave penalized estimation in sparse vector autoregression model. *Electron. J. Statist. 14*(1), 1413–1448.

Sijia Xia

School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China

E-mail: sijiax@mails.ccnu.edu.cn ORCID: https://orcid.org/0009-0005-3311-6119

Michael K. Ng

Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

E-mail: michael-ng@hkbu.edu.hk ORCID: https://orcid.org/0000-0001-6833-5227

Xiongjun Zhang

School of Mathematics and Statistics, and Key Laboratory of Nonlinear Analysis & Applications (Ministry of Education), Central China Normal University, Wuhan 430079, China.

E-mail: xjzhang@ccnu.edu.cn ORCID: https://orcid.org/0000-0002-5387-4129