Tensor Methods for High-Dimensional Time Series Modeling

Yao Zheng

(based on joint works with Di Wang, Heng Lian, and Guodong Li)

May 10, 2022

Department of Statistics

University of Connecticut

Outline

- Introduction
- Topic 1: High-Dimensional Vector Autoregression via Tensor Decomposition
  - Motivation
  - Proposed Model
  - Estimation Methods
  - Empirical Example
- Topic 2: Low-Rank Tensor Autoregression
  - Motivation
  - Proposed Model
  - Estimation Methods
  - Empirical Example
- Conclusion

Introduction

High Dimensional Time Series

Big data is everywhere, and many big datasets are temporally dependent.

Needs for high-dimensional time series models:

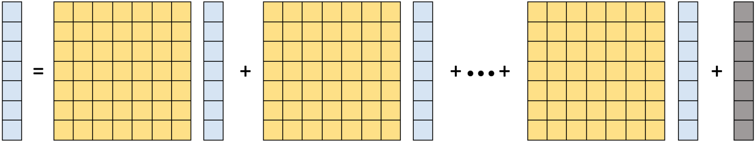
- Economics: forecast with many predictors and understand causal relationships
- Finance: build large scale systemic risk models
- Functional Genomics: reconstruct gene regulatory networks based on limited experimental data
- Neuroscience: build detailed connectivity maps on temporal data exhibiting multiple structural changes

Vector AutoRegression

- VAR is a fundamental model for multivariate time series analysis.

- VAR with N variables, lag order P, and time length T:

$$\boldsymbol{y}_t = \sum_{j=1}^{P} \boldsymbol{A}_j \boldsymbol{y}_{t-j} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})', \; t = 1, \ldots, T.$$



- This is called a VAR(P) model.

- Curse of dimensionality ($N^2 P \gg T$) even when the dimension N is moderately large.

Topic 1: High-Dimensional Vector Autoregression via Tensor Decomposition

Motivation: Large Lag Order

- Compared with VAR, the vector autoregressive moving average (VARMA) model usually performs better in practice since it provides a more flexible autocorrelation structure.

- However, VARMA has a serious identification problem when $N \geq 2$. (Estimation is unstable since the objective function involves a high-order polynomial.)

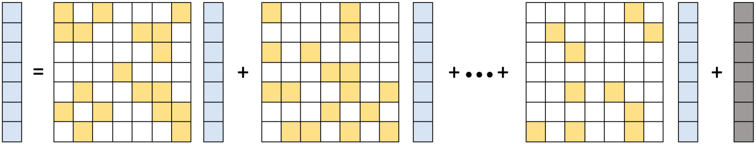
- It is common to employ a VAR(P) model to approximate the VARMA process, and the lag order P may be very large to provide a better fit.

- As $T \to \infty$, we need $P \to \infty$ and $PT^{-1/3} \to 0$.

- Curse of dimensionality

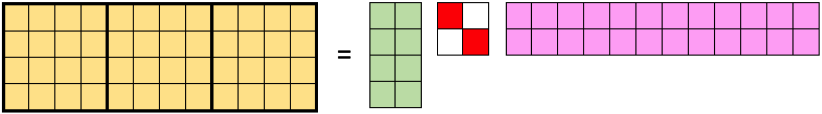
Low-Dimensional Structures in VAR

- Sparse VAR model (Lasso, Dantzig selector, SCAD, etc.)



Basu and Michailidis (2015); Han et al. (2015); Wu and Wu (2016)

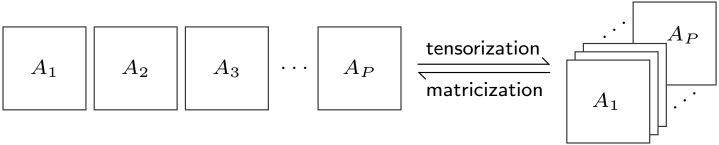
- Reduced-rank VAR model (SVD, nuclear norm)



Velu et al. (1986); Negahban and Wainwright (2011); Chen et al. (2013)
Constraint on column space of $[\boldsymbol{A}_1, \boldsymbol{A}_2, \dots, \boldsymbol{A}_P]$.

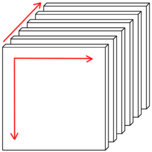
VAR in Tensor Form

We propose to rearrange the transition matrices into a tensor



and consider dimensionality reduction in three different directions:

- column-wise $[\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_P]$ (reduced-rank model)
- row-wise $[\boldsymbol{A}_1', \boldsymbol{A}_2', \ldots, \boldsymbol{A}_P']$ (autoregressive index model)
- temporal $[\mathrm{vec}(\boldsymbol{A}_1), \mathrm{vec}(\boldsymbol{A}_2), \ldots, \mathrm{vec}(\boldsymbol{A}_P)]$ (new temporal structure)



Vector, Matrix and Tensor

Tensors are higher-order extensions of matrices.

- 1st-order tensors are vectors ($\boldsymbol{a}$)
- 2nd-order tensors are matrices ($\boldsymbol{A}$)
- higher-order tensors ($\boldsymbol{\mathcal{A}}$)

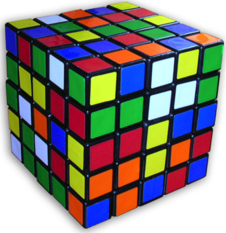
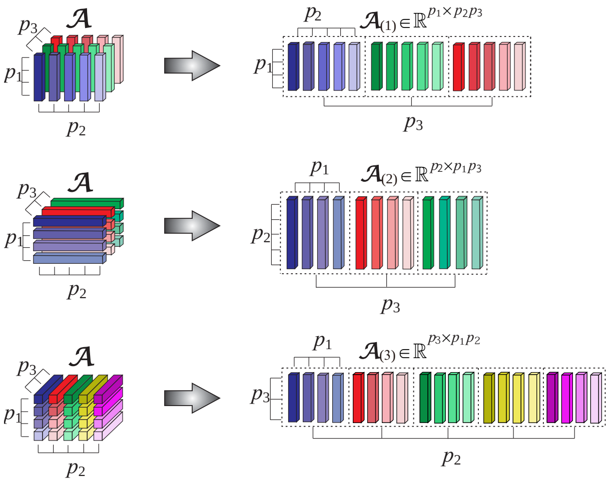


Figure 1: $5 \times 5 \times 5$ tensor. This is a third-order tensor.

Matricization and Tucker Ranks of a Tensor

Consider $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$. For $i = 1, 2, 3$, its mode-i matricization $\mathcal{A}_{(i)}$ is



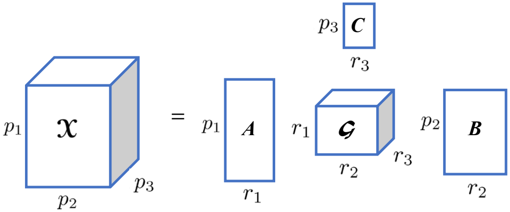
Let $r_i = \text{rank}(\mathcal{A}_{(i)})$ be the matrix rank of $\mathcal{A}_{(i)}$. (r_1, r_2, r_3) are analogous to column and row ranks of a matrix, but they are not always equal.

Tucker Decomposition

- For a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, the Tucker decomposition is

$$\mathcal{X} = \mathcal{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} \equiv [\![\mathcal{G}; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!]$$

where $\mathcal{G}$ is a $r_1 \times r_2 \times r_3$ core tensor, $\boldsymbol{A}$ is a $p_1 \times r_1$ matrix, $\boldsymbol{B}$ is a $p_2 \times r_2$ matrix, and $\boldsymbol{C}$ is a $p_3 \times r_3$ matrix.



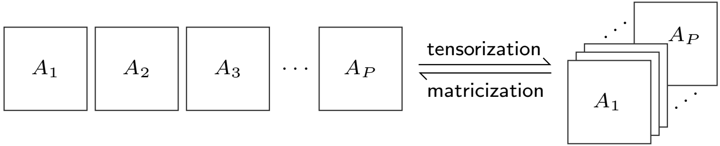
- Higher-order SVD: $\mathcal{G}$ is all-orthogonal; $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are orthonormal

- (r_1, r_2, r_3): Tucker ranks or multilinear ranks.

Proposed Model: Multilinear Low-Rank VAR

- For a VAR(P) model, we stack $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_P$ into an $N \times N \times P$ tensor $\boldsymbol{\mathcal{A}}$, where

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3.$$



- If $(r_1, r_2, r_3) \ll (N, N, P)$,

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

- The reduced-rank model is a special case of multilinear low-rank model with ranks (r, N, P).

Connection with the factor model

- Factor model:

$$\boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \tag{1}$$

  where $\boldsymbol{f}_t$ is a set of latent factor with dimension $r \ll N$, $\boldsymbol{\Lambda}$ is an N-by-r factor loading matrix, and $\boldsymbol{\xi}_t$ is the noise series.

- Estimated factor: $\widehat{\boldsymbol{f}}_t = \widehat{\boldsymbol{\Lambda}}' \boldsymbol{y}_t$.

- Supervised factor interpretation: since $\boldsymbol{U}_i$ is orthonormal

$$\boldsymbol{U}_1' \boldsymbol{y}_t = \mathcal{G}_{(1)} \text{vec}(\boldsymbol{U}_2' \boldsymbol{X}_t \boldsymbol{U}_3) + \boldsymbol{U}_1' \boldsymbol{\epsilon}_t, \tag{2}$$

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

- $\boldsymbol{U}_1' \boldsymbol{y}_t$: r_1 response factors
- $\boldsymbol{U}_2' \boldsymbol{X}_t \boldsymbol{U}_3$: $r_2 \times r_3$ bilinear predictor factors.
- r_1: response rank, r_2: predictor rank and r_3: temporal rank.

Connection with VARMA

- The VARMA(1,1) process

$$\boldsymbol{y}_t = \boldsymbol{\Psi}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t - \boldsymbol{\Theta}\boldsymbol{\epsilon}_{t-1} \tag{3}$$

has the VAR(∞) form with $\boldsymbol{A}_j = -\boldsymbol{\Theta}^{j-1}(\boldsymbol{\Theta} - \boldsymbol{\Psi})$ for $j \geq 1$:

$$\boldsymbol{y}_t = \boldsymbol{\epsilon}_t + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \boldsymbol{A}_3\boldsymbol{y}_{t-3} + \cdots . \tag{4}$$

Accordingly, we can define an $N \times N \times \infty$ tensor $\boldsymbol{\mathcal{A}}_{\text{VARMA}}$.

Proposition 1

Under regularity conditions, if $\text{rank}(\boldsymbol{\Theta}) = r$ and $\text{rank}(\boldsymbol{\Psi}) = s$, then $\boldsymbol{\mathcal{A}}_{\text{VARMA}}$ has multilinear ranks at most $(r+s, r+s, r+1)$.

- VAR(P) approximation is easier to implement but involves more parameters.
- Tucker decomposition reduces the dimensionality and alleviates the overparametrization.

Multilinear Low-Rank Estimator

- Denote $\boldsymbol{x}_t = (\boldsymbol{y}_{t-1}', \ldots, \boldsymbol{y}_{t-P}')'$. Given Tucker ranks (r_1, r_2, r_3), consider the multilinear low-rank (MLR) estimator

$$\begin{aligned}\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}} &\equiv [\![\widehat{\mathcal{G}}; \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3]\!] \\ &= \arg\min \sum_{t=1}^{T} \|\boldsymbol{y}_t - (\mathcal{G} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3)_{(1)} \boldsymbol{x}_t\|_2^2.\end{aligned}$$

- Alternating least squares algorithm:
  - Update $\mathcal{G}$, $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, $\boldsymbol{U}_3$ alternatingly.
  - Each step is an OLS problem.

Theorem 1 (Asymptotic Normality)

Under regularity conditions, if N and P are fixed, then as $T \to \infty$,

$$\sqrt{T}(\text{vec}(\widehat{\boldsymbol{\mathcal{A}}}_{\text{MLR}}) - \text{vec}(\boldsymbol{\mathcal{A}})) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\text{MLR}}).$$

Rank Selection

- We propose a ridge-type ratio estimator to determine (r_1, r_2, r_3).
- Based on an initial estimator $\widehat{\mathcal{A}}$ (e.g., the OLS estimator or the nuclear norm estimator), we estimate (r_1, r_2, r_3) by

$$\widehat{r}_i = \underset{1 \le j \le p_i - 1}{\arg\min} \frac{\sigma_{j+1}(\widehat{\mathcal{A}}_{(i)}) + c}{\sigma_j(\widehat{\mathcal{A}}_{(i)}) + c}, \quad 1 \le i \le 3,$$

where $p_1 = p_2 = N$, $p_3 = P$, and c is a well-chosen parameter.

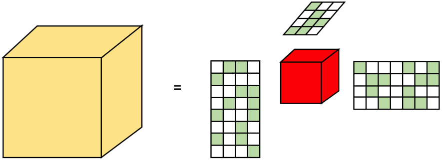
Theorem 2 (Rank Selection Consistency)

Under the conditions of Theorem 1, if c is chosen such that $T^{-1/2} \ll c \ll \sigma_{r_i}(\mathcal{A}_{(i)}) \cdot \min_{1 \le j < r_i} \sigma_{j+1}(\mathcal{A}_{(i)}) / \sigma_j(\mathcal{A}_{(i)})$, for $1 \le i \le 3$,

$$\mathbb{P}(\widehat{r}_1 = r_1, \widehat{r}_2 = r_2, \widehat{r}_3 = r_3) \to 1, \quad \text{as } T \to \infty.$$

SHORR Estimator

- Sparsity in $\boldsymbol{U}_i \Rightarrow$ variable selection in factor loadings



- Sparse Higher-Order Reduced Rank (SHORR) estimator:

$$\widehat{\mathcal{A}}_{\text{SHORR}} \equiv [\![\widehat{\mathcal{G}}; \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2, \widehat{\boldsymbol{U}}_3]\!]$$

$$= \arg\min \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\boldsymbol{y}_t - (\mathcal{G} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3)_{(1)} \boldsymbol{x}_t\|^2 + \lambda \|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1 \right\}$$

subject to $\mathcal{G}$ is all-orthogonal and $\boldsymbol{U}_i$ is orthonormal, where $\|\cdot\|_1 = \|\text{vec}(\cdot)\|_1$ for matrices.

- $\|\boldsymbol{U}_3 \otimes \boldsymbol{U}_2 \otimes \boldsymbol{U}_1\|_1$ induces sparsity for three factor matrices jointly.

SHORR Estimator

- We propose an alternating direction method of multipliers (ADMM) algorithm.

- Nonasymptotic error bounds:

Theorem 3 (Simplified by assuming (r_1, r_2, r_3) are fixed)

Under regularity conditions, if $\lambda \gtrsim \sqrt{\log(N^2P)/T}$ and $T \gtrsim \log(N^2P)$, then with high probability,

$$||\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A}||_{\text{F}} \lesssim \sqrt{s_1 s_2 s_3}\lambda,$$

$$\frac{1}{T}\sum_{t=1}^{T} \|(\widehat{\mathcal{A}}_{\text{SHORR}} - \mathcal{A})_{(1)}\boldsymbol{x}_t\|_2^2 \lesssim \tau^2 s_1 s_2 s_3 \lambda^2,$$

where s_i is the maximum number of nonzero entries in each column of $\boldsymbol{U}_i$, for $1 \le i \le 3$.

- Estimation convergence rate is $\sqrt{s_1 s_2 s_3 \log(N^2P)/T}$.

Comparison of Estimation Efficiency

Estimator	Structure	Estimation error rate
SHORR	low-rank & sparsity	$\sqrt{s_1 s_2 s_3 \log(N^2 P)/T}$
Lasso	sparsity	$\sqrt{\|\mathcal{A}\|_0 \log(N^2 P)/T}$
Nuclear	low-rank	$\sqrt{rNP/T}$

Introducing sparsity into the low-rank decomposition can improve the estimation efficiency.

Macroeconomic Forecasting

- A list of 40 major U.S. quarterly macroeconomic variables from Q1-1959 to Q4-2007, seasonally adjusted and transformed to be stationary. Eight categories:

(1) GDP and its decomposition	(2) NAPM indices
(3) industrial production	(4) housing
(5) money, credit and interest rate	(6) employment
(7) prices and wages	(8) others

- Apply VAR(4) model. Select $(r_1, r_2, r_3) = (4, 3, 2)$.

- Perform rolling forecast from Q4-2000 to Q4-2006. Forecast error:

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

- SHORR and MLR have impressive forecasting accuracy compared to competing methods.

Response factors $\boldsymbol{U}_1$

- Almost all variables are selected.
- Each factor covers multiple categories of macroeconomic indices.
- No group structure can be observed.

Response Factors				Short Name	Predictor Factors			Category
-0.195	0.195	0.008		GDP251		-0.310	0.155	GDP Decomposition
	0.465	0.023	0.001	GDP252			-0.012	GDP Decomposition
	0.351			GDP253			0.014	GDP Decomposition
-0.297			-0.034	GDP256	-0.014	0.332	-0.118	GDP Decomposition
-0.063	-0.172			GDP263				GDP Decomposition
-0.298	-0.114		-0.076	GDP264				GDP Decomposition
		0.028		GDP265				GDP Decomposition
-0.094	0.368	0.020		GDP270				GDP Decomposition
	-0.059		0.413	PMCP	0.065			NAPM Indices
-0.029	-0.207		0.425	PMDEL	0.172			NAPM Indices
-0.249			0.183	PMI	-0.021	-0.870	0.006	NAPM Indices
-0.320		-0.026		PMNO	0.761	-0.020		NAPM Indices
	-0.109		0.473	PMNV				NAPM Indices
-0.363				PMP	-0.621			NAPM Indices
-0.302				IPS10				Industrial Production
-0.004	0.071	0.785	0.178	UTL11		-0.007	0.849	Industrial Production
0.049	0.347	-0.424	0.459	HSFR		-0.196	-0.494	Housing
-0.086				BUSLOANS				Money, Credit, Interest Rate
-0.077			-0.035	CCINRV				Money, Credit, Interest Rate
0.097	0.107			FM1				Money, Credit, Interest Rate
0.087				FM2				Money, Credit, Interest Rate
				FMRNBA				Money, Credit, Interest Rate
				FMRRA				Money, Credit, Interest Rate
	0.131		-0.162	FSPIN				Money, Credit, Interest Rate
-0.123			0.082	FYFF				Money, Credit, Interest Rate
-0.016				FYGT10				Money, Credit, Interest Rate
-0.126	-0.077	-0.383	-0.202	SEYGT10	-0.069			Money, Credit, Interest Rate
-0.269		0.132	0.046	CES002				Employment
-0.278				LBMNU				Employment
0.002	0.346		-0.016	LBOUT				Employment
-0.228	0.120			LHEL				Employment
0.331				LHUR				Employment
	0.267	0.145		CES275R				Prices and Wages
-0.068				CPIAUCSL				Prices and Wages
			0.025	GDP273				Prices and Wages
	0.005		0.018	GDP276				Prices and Wages
	-0.037	-0.105	0.218	PSCCOMR				Prices and Wages
-0.076	-0.146			PWFSA				Prices and Wages
		0.082	-0.082	EXRUS				Others
	0.008		-0.106	HHSNTN				Others

Response Factors **Short Name** **Predictor Factors**

Category: GDP Decomposition; NAPM Indices; Industrial Production; Housing; Money, Credit, Interest Rate; Employment; Prices and Wages; Others

Legend: 0.8, 0.6, 0.4, 0.2, 0, -0.2, -0.4, -0.6, -0.8

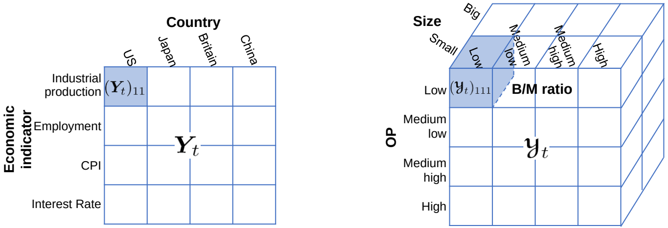
Predictor factors $\boldsymbol{U}_2$

- Only 12 variables are selected, all but one from the first four categories.
- Activeness of production and investment serves as the driving force of the whole economy.

Topic 2: Low-Rank Tensor Autoregression

Tensor-Value Time Series Data

- Tensor-valued time series can be found in many fields: economics, portfolio analysis, neuroscience, bioinformatics, computer vision, ...
- Denoted by $\{\mathcal{Y}_t, t = 1, \ldots, T\}$, where $\mathcal{Y}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d}$. When $d = 1$, vector-valued time series $\{\boldsymbol{y}_t\}$. When $d = 2$, matrix-valued time series $\{\boldsymbol{Y}_t\}$.



(a) *Matrix-valued* $\boldsymbol{Y}_t \in \mathbb{R}^{4\times 4}$ (b) *Tensor-valued* $\mathcal{Y}_t \in \mathbb{R}^{4\times 4\times 2}$

Figure 2: *Observation at time t for (a) a 4×4 matrix-valued macroeconomic indicators time series $\{\boldsymbol{Y}_t\}$ and (b) a $4 \times 4 \times 2$ tensor-valued stock portfolio returns time series $\{\mathcal{Y}_t\}$. OP: operating profitability; B/M: book-to-market.*

How to model tensor-valued time series?

- Consider $\mathcal{Y}_t \in \mathbb{R}^{4\times4\times2}$ in Figure 1(b). A simple approach is the VAR:

$$\text{vec}(\mathcal{Y}_t) = \boldsymbol{A}\text{vec}(\mathcal{Y}_{t-1}) + \text{vec}(\mathcal{E}_t), \tag{5}$$

  where $\boldsymbol{A} \in \mathbb{R}^{32\times32}$ is the unknown transition matrix. It can incorporate linear associations between every variable in $\mathcal{Y}_t$ and that in $\mathcal{Y}_{t-1}$.

- Even with only one lag, # of parameters $= 32^2 = 1024$. (curse of dimensionality)

- The vectorization will destroy the intrinsic multidimensional structural information of the observed tensors $\mathcal{Y}_t$. (lack of interpretability)

Multi-Mode Matricization

- For a fixed index set $S \subset \{1, 2, \ldots, d\}$, the multi-mode matricization of $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the matrix

$$\mathcal{X}_{[S]} \in \mathbb{R}^{\prod_{i \in S} p_i \times \prod_{i \notin S} p_i},$$

  with $\prod_{i \in S} p_i$ rows and $\prod_{i \notin S} p_i$ columns[a].

- One-mode matriciation: The mode-i matricization of $\mathcal{X}$, $\mathcal{X}_{(i)}$, is simply $\mathcal{X}_{[\{i\}]}$.

[a]Specifically, its (i, j)-th entry is

$$\left(\mathcal{X}_{[S]}\right)_{i,j} = \mathcal{X}_{i_1, \ldots, i_d},$$

where $i = 1 + \sum_{k \in S}(i_k - 1)I_k$ and $j = 1 + \sum_{k \notin S}(i_k - 1)J_k$, with $I_k = \prod_{\ell \in S, \ell < k} p_\ell$, and $J_k = \prod_{\ell \notin S, \ell < k} p_\ell$.

Proposed Model: Low-Rank Tensor Autoregression (LRTAR)

We propose

$$\mathcal{Y}_t = \langle \mathcal{A}, \mathcal{Y}_{t-1} \rangle + \mathcal{E}_t,$$

where

$$\mathcal{Y}_t, \mathcal{E}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d},$$

and

$$\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_d \times p_1 \times \cdots \times p_d}$$

is a $2d$-th-order transition tensor with Tucker ranks $(r_1, \ldots, r_{2d})$, i.e.,

$$r_i = \text{rank}(\mathcal{A}_{(i)}), \quad i = 1, \ldots, 2d.$$

Tucker Decomposition and Connection with VAR

- $\mathcal{A}$ has the Tucker decomposition $\mathcal{A} = \mathcal{G} \times_{i=1}^{2d} \boldsymbol{U}_i$, with core tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_{2d}}$ and factor matrices $\boldsymbol{U}_i \in \mathbb{R}^{p_i \times r_i}$, $1 \le i \le 2d$.

- Define index sets $S_1 = \{1, 2, \ldots, d\}$ and $S_2 = \{d+1, d+2, \ldots, 2d\}$. Then the LRTAR can be written in the VAR form:

$$\text{vec}(\mathcal{Y}_t) = \overbrace{(\otimes_{i \in S_2} \boldsymbol{U}_i) \mathcal{G}_{[S_2]} (\otimes_{i \in S_1} \boldsymbol{U}_i)^\top}^{\mathcal{A}_{[S_2]}} \text{vec}(\mathcal{Y}_{t-1}) + \text{vec}(\mathcal{E}_t)$$

- The transition matrix is the multi-mode matricization of $\mathcal{A}$, $\mathcal{A}_{[S_2]} \in \mathbb{R}^{\prod_{i=1}^d p_i \times \prod_{i=1}^d p_i}$.

- \# of parameters is reduced from $(\prod_{i=1}^d p_i)^2$ dramatically to

$$\prod_{i=1}^{2d} r_i + \sum_{i=1}^{d} r_i (p_i - r_i) + \sum_{i=1}^{d} r_{d+i} (p_i - r_{d+i}).$$

Dynamic Tensor Factors Interpretation

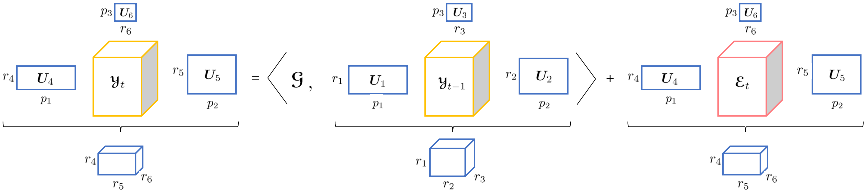


Figure 3: *Low-dimensional dynamic factor structure when $\mathcal{Y}_t$ is a third-order tensor.*

- Consider the HOSVD: all $\boldsymbol{U}_i$ are orthonormal. Then the LRTAR implies a low-dimensional tensor regression:

$$\underbrace{\mathcal{Y}_t \times_{i=d+1}^{2d} \boldsymbol{U}_i^\top}_{r_{d+1}\times r_{d+2}\times\cdots\times r_{2d}} = \langle \mathcal{G}, \underbrace{\mathcal{Y}_{t-1} \times_{i=1}^{d} \boldsymbol{U}_i^\top}_{r_1\times r_2\times\cdots\times r_d} \rangle + \mathcal{E}_t \times_{i=d+1}^{2d} \boldsymbol{U}_i^\top,$$

- $\mathcal{Y}_t \times_{i=d+1}^{2d} \boldsymbol{U}_i^\top$: $r_{d+1} \times r_{d+2} \times \cdots \times r_{2d}$ response factors
- $\mathcal{Y}_{t-1} \times_{i=1}^{d} \boldsymbol{U}_i^\top$: $r_1 \times r_2 \times \cdots \times r_d$ predictor factors

Regularization via Square Matricizations

- $\mathcal{A}$ is a $p_1 \times \cdots \times p_d \times p_1 \times \cdots \times p_d$ tensor. The multi-mode matricization $\mathcal{A}_{[I]}$ will be a $\prod_{i=1}^{d} p_i \times \prod_{i=1}^{d} p_i$ square matrix if the index set is

$$I = \{\ell_1, \ldots, \ell_d\}, \quad \text{where} \quad \ell_i \in \{i, d+i\} \quad \text{for} \quad i = 1, \ldots, d.$$

- There are totally 2^d square matricizations of $\mathcal{A}$, denoted by $\mathcal{A}_{[I_k]}$ with $1 \leq k \leq 2^d$. Note that $\text{rank}(\mathcal{A}_{[I_k]}) \leq \min(\prod_{i=1, i \in I_k}^{2d} r_i, \prod_{i=1, i \notin I_k}^{2d} r_i)$.

- To simultaneously encourage low-rankness across all square matricizations, and hence across all modes, we propose a novel regularizer based on the **Sum of Square-matrix Nuclear (SSN) norm**:

$$\|\mathcal{A}\|_{\text{SSN}} = \sum_{k=1}^{2^d} \left\|\mathcal{A}_{[I_k]}\right\|_*,$$

where $\|\boldsymbol{X}\|_* = \sum_j \sigma_j(\boldsymbol{X})$ is the nuclear norm, with $\sigma_j(\boldsymbol{X})$ being the j-th largest singular value of $\boldsymbol{X}$.

SSN Norm Regularized Estimator

- We propose the SSN norm regularized estimator

$$\widehat{\mathcal{A}}_{\mathsf{SSN}} = \arg\min_{\mathcal{A}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\mathcal{Y}_t - \langle \mathcal{A}, \mathcal{Y}_{t-1} \rangle\|_{\mathsf{F}}^2 + \lambda_{\mathsf{SSN}} \|\mathcal{A}\|_{\mathsf{SSN}} \right\}.$$

Theorem 4

Under regularity conditions, if $\lambda_{\mathsf{SSN}} \gtrsim 2^{-d}\sqrt{p/T}$, and $T \gtrsim p$, then with high probability,

$$\|\widehat{\mathcal{A}}_{\mathsf{SSN}} - \mathcal{A}\|_{\mathsf{F}} \lesssim \sqrt{s}(2^d \lambda_{\mathsf{SSN}}),$$

$$T^{-1} \sum_{t=1}^{T} \|\langle \widehat{\mathcal{A}}_{\mathsf{SSN}} - \mathcal{A}, \mathcal{Y}_{t-1} \rangle\|_{\mathsf{F}}^2 \lesssim Cs(2^d \lambda_{\mathsf{SSN}})^2,$$

where $p = \prod_{i=1}^{d} p_i$ and $\sqrt{s} = 2^{-d} \sum_{k=1}^{2^d} \sqrt{2 s_k}$, with $s_k = \text{rank}(\mathcal{A}_{[I_k]})$.

- Estimation convergence rate is $2^{-d} \sum_{k=1}^{2^d} \sqrt{s_k p/T}$.

Comparison of Estimation Efficiency

- We also considered the Sum of Nuclear (SN) norm (Gandy et al., 2011):

$$\|\mathcal{A}\|_{\mathsf{SN}} = \sum_{i=1}^{2d} \|\mathcal{A}_{(i)}\|_*,$$

$$\widehat{\mathcal{A}}_{\mathsf{SN}} = \arg\min_{\mathcal{A}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \|\mathcal{Y}_t - \langle \mathcal{A}, \mathcal{Y}_{t-1} \rangle\|_{\mathsf{F}}^2 + \lambda_{\mathsf{SN}} \|\mathcal{A}\|_{\mathsf{SN}} \right\},$$

  This is based on the one-mode matricizations.

- The square matricization leads to greater estimation efficiency:

Regularizer	Matricization	Estimation error rate
SN	one-mode	$d^{-1} \sum_{i=1}^{d} \sqrt{r p_{-i} p / T}$
SSN	square	$2^{-d} \sum_{k=1}^{2^d} \sqrt{s_k p / T}$

  - $p = \prod_{i=1}^{d} p_i$, $p_{-i} = \prod_{j=1, j \neq i}^{d} p_j$
  - $\sqrt{r} = (2d)^{-1} \sum_{i=1}^{2d} \sqrt{2 r_i}$, $r_i = \text{rank}(\mathcal{A}^*_{(i)})$, and $s_k = \text{rank}(\mathcal{A}^*_{[I_k]})$ are fixed if $(r_1, \ldots, r_{2d})$ are fixed.

Rank Selection

- The optimization is convex yet the regularizer involves multiple nuclear norms. We propose an ADMM algorithm.

- $\widehat{\mathcal{A}}_{\text{SSN}}$ does not guarantee consistent estimation of the ranks. To this end, we further apply a truncation method:
    - Truncated SVD for each $(\widehat{\mathcal{A}}_{\text{SSN}})_{(i)}$: Retain only singular values exceeding a well-chosen threshold $\gamma > 0$. Obtain the truncated factor matrices, $\widetilde{\boldsymbol{U}}_i$, $1 \leq i \leq 2d$.
    - The truncated core tensor is $\widetilde{\mathcal{G}} = \widehat{\mathcal{A}}_{\text{SSN}} \times_{i=1}^{2d} \widetilde{\boldsymbol{U}}_i^\top$.

  The truncated SSN (TSSN) estimator is

$$\widehat{\mathcal{A}}_{\text{TSSN}} = \widetilde{\mathcal{G}} \times_{i=1}^{2d} \widetilde{\boldsymbol{U}}_i.$$

- The rank selection is consistent if $\sqrt{s^* p/T} \ll \gamma \lesssim \min_{1 \leq i \leq 2d} \sigma_{r_i}\left(\mathcal{A}_{(i)}\right)$.

- $\widehat{\mathcal{A}}_{\text{TSSN}}$ achieves the same asymptotic error rate as $\widehat{\mathcal{A}}_{\text{SSN}}$.

Portfolio Returns Forecasting

- Monthly market-adjusted portfolio return series from July 1963 to Dec. 2019.
  http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
- The portfolios are constructed as the intersections of different levels of
  - size: small and big
  - book-to-market (B/M) ratio: from lowest to highest
  - operating profitability (OP): from lowest to highest
  - Investment (Inv): from lowest to highest
- The first dataset: $4 \times 4 \times 2$ portfolios formed by OP, B/M ratio, and size.
- The second dataset: $4 \times 4 \times 2$ portfolios formed by Inv, B/M ratio, and size.

Models for $4 \times 4 \times 2$ time series

- Vector autoregression (VAR): $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{e}_t$, where $\boldsymbol{A} \in \mathbb{R}^{32\times 32}$.
- Vector factor model (VFM): $\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t$, where $\boldsymbol{f}_t$ is the low-dimensional vector-valued latent factor, and $\boldsymbol{\Lambda}$ is the loading matrix.
- Multilinear tensor autoregression (MTAR): $\mathcal{Y}_t = \mathcal{Y}_{t-1} \times_{i=1}^{3} \boldsymbol{B}_i + \mathcal{E}_t$, where $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{4\times 4}$ and $\boldsymbol{B}_3 \in \mathbb{R}^{2\times 2}$ are coefficient matrices.
- Tensor factor model (TFM): $\mathcal{Y}_t = \mathcal{F}_t \times_{i=1}^{3} \boldsymbol{U}_i + \mathcal{E}_t$, where $\mathcal{F}_t$ is the low-dimensional tensor-valued latent factor, and $\boldsymbol{U}_i$'s are the loading matrices; see Chen et al. (2022). For prediction, the estimated factors $\widehat{\mathcal{F}}_t$ are then fitted by a VAR(1) model.
- Proposed LRTAR: $\mathcal{Y}_t = \langle \mathcal{A}, \mathcal{Y}_{t-1} \rangle + \mathcal{E}_t$, with $\mathcal{A} = \mathcal{G} \times_{i=1}^{6} \boldsymbol{U}_i$.

Results for $4 \times 4 \times 2$ time series

	Model	VAR	VFM	MTAR	TFM	LRTAR		Best	Worst
						SSN	TSSN		
OP-BM-Size $4 \times 4 \times 2$ series									
In-sample	ℓ_2 norm	**19.53**	20.08	19.89	20.09	19.69	19.70	VAR	TFM
	ℓ_0 norm	**7.67**	7.91	7.85	7.92	7.76	7.77	VAR	TFM
Out-of-sample	ℓ_2 norm	22.27	20.17	20.50	20.11	20.32	**19.95**	TSSN	VAR
	ℓ_∞ norm	10.38	10.04	9.86	10.03	**9.29**	9.35	SSN	VAR
Inv-BM-Size $4 \times 4 \times 2$ series									
In-sample	ℓ_2 norm	**16.80**	17.10	17.05	17.11	16.86	16.88	VAR	TFM
	ℓ_0 norm	**6.25**	6.40	6.38	6.41	6.31	6.32	VAR	TFM
Out-of-sample	ℓ_2 norm	18.70	17.70	16.89	17.67	**16.11**	16.29	SSN	VAR
	ℓ_∞ norm	7.42	7.37	6.79	7.33	6.62	**6.43**	TSSN	VAR

Table 1: Average in-sample forecasting error and out-of-sample rolling forecasting error for $4 \times 4 \times 2$ tensor-valued portfolio return series. The best cases are marked in **bold**.

Results for $4 \times 4 \times 2$ time series

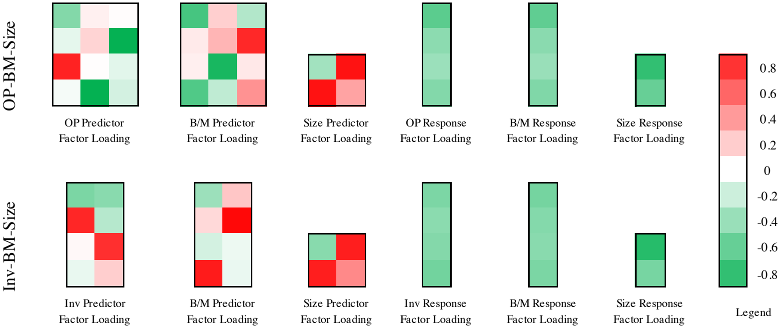


Figure 4: TSSN estimates of predictor and response factor matrices for $4 \times 4 \times 2$ tensor-valued portfolio return series. From left to right: $\widetilde{\boldsymbol{U}}_1, \widetilde{\boldsymbol{U}}_2, \widetilde{\boldsymbol{U}}_3$, $\widetilde{\boldsymbol{U}}_4$, $\widetilde{\boldsymbol{U}}_5$ and $\widetilde{\boldsymbol{U}}_6$.

Conclusion

Conclusion

- In both topics, we leveraged the tensor decomposition for dimensionality reduction of high-dimensional time series models.
- Besides achieving greater estimation efficiency and forecast accuracy, the resulting models admit intepretable dynamic factor structures that enable the extraction of meaningful insights from massive data.
- In topic 1, we developed a new high-dimensional vector autoregressive model - the Multilinear Low-Rank VAR, and further considered imposing sparsity on the factor matrices for automatic variable selection in factor loadings.
- In topic 2, we developed a novel high-dimensional tensor autoregressive model - the Low-Rank TAR, which is one of the first endeavors of statistical modeling for tensor-valued time series data.

Thank you!

References

Topic 1:

Wang, D., Zheng, Y., Lian, H., and Li, G. (2021b). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association.* To appear.

Topic 2:

Wang, D., Zheng, Y., and Li, G. (2021a). High-dimensional low-rank tensor autoregressive time series modelling. *arXiv preprint arXiv:2101.04276.*

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117:94–116.

Gandy, S., Recht, B., and Yamada, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010.