Autoregressive Models for Tensor-Valued Time Series

Zebang Li
Joint with Prof.Han Xiao

Rutgers, The State University of New Jersey

zl326@stat.rutgers.edu; hxiao@stat.rutgers.edu

JSM 2020, Aug 03

Introduction

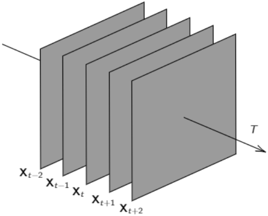


Figure 1: Matrix-valued time series.

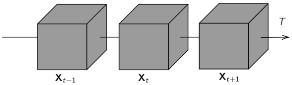
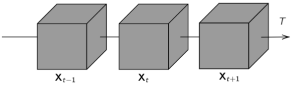


Figure 2: Tensor-valued time series.

High dimensional time series observed in tensor form are becoming more and more commonly seen in various fields.

Introduction



For example, Average Value Weighted Returns of **Fama French portfolios** are allocated to

- Two **Size groups** (Small and Big) using NYSE median market cap breakpoint.
- Stocks in each Size group are allocated independently to four **B/M groups** (Book-to-Market, low B/M to High B/M)
- and four **OP groups** (Operating Profitability, Low OP to High OP) using NYSE quartile breakpoints specific to the Size group

which formed a $4 \times 4 \times 2$ tensor time series, from July 1963 to June 2020.

Autoregressive Models for Tensor-Valued Time Series

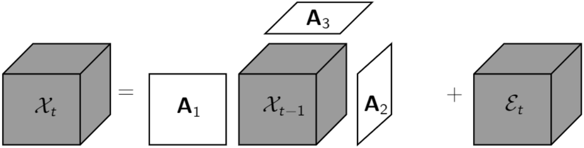
At each time t, a matrix $\mathbf{X}_t \in \mathbb{R}^{d_1 \times d_2}$ is observed. Recall Matrix Autoregression Model, proposed by Chen, et al., 2020.

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}' + \mathbf{E}_t, \ \ \mathbf{X}_t \in \mathbb{R}^{d_1 \times d_2}$$

Now at each time t, a mode-K tensor $\mathcal{X}_t \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ is observed. We proposed Tensor Autoregression Model (of order 1), in the form

$$\mathcal{X}_t = \mathcal{X}_{t-1} \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K + \mathcal{E}_t \tag{1}$$

where $\mathbf{A}_k \in \mathbb{R}^{d_k \times d_k}$, $1 \le k \le K$ are autoregressive coefficient matrices.



$\times_k$ is k-mode product and $\mathcal{E}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a tensor white noise.

$$\mathcal{X}_t = \mathcal{X}_{t-1} \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K + \mathcal{E}_t$$

This model is consistent with vector and matrix AR model.

- when mode $K = 1$, it is the VAR(1) model.

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{E}_t, \ \mathbf{X}_t \in \mathbb{R}^{d_1}$$

- when mode $K = 2$, it is the MAR(1) model.

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1}\mathbf{B}' + \mathbf{E}_t, \ \mathbf{X}_t \in \mathbb{R}^{d_1 \times d_2}$$

Probabilistic Properties

When the condition of proposition 1 is fulfilled, the model has the following causal representation after vectorization:

$$\text{vec}(\mathcal{X}_t) = \sum_{k=0}^{\infty} \left(\mathbf{A}_K^k \otimes \cdots \otimes \mathbf{A}_1^k \right) \text{vec}(\mathcal{E}_{t-k}). \tag{2}$$

Proposition 1

If $\prod_{i=1}^{K} \rho(\mathbf{A}_i) < 1$, then the tensor autoregressive model is stationary and causal, where ρ denotes spectral radius.

Multi-term Models

The model can be extended to have R terms. That is

$$\mathcal{X}_t = \sum_{i=1}^{R} \mathcal{X}_{t-1} \times_1 \mathbf{A}_1^{(i)} \times_2 \cdots \times_K \mathbf{A}_K^{(i)} + \mathcal{E}_t$$

This is still an order-1 autoregressive model, but with more parallel terms. Such a structure provides more flexibility to capture the different interactions among fibbers of the tensor.

Estimation

- Projection method.
- Iterated least squares.
- MLE under a structured covariance tensor.

Projection Method: One-Term Models

Our first approach is to view it as the structured VAR(1) model.

$$\mathrm{vec}(\mathcal{X}_t) = (\mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1)\mathrm{vec}(\mathcal{X}_t) + \mathrm{vec}(\mathcal{E}_t)$$

Let $\mathbf{\Phi} = \mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1$. First obtain the maximum likelihood estimate or the least square estimate $\hat{\mathbf{\Phi}}$ of $\mathbf{\Phi}$ without the structure constraint then we find the estimators by projecting $\hat{\mathbf{\Phi}}$ onto the space of Kronecker products under the Frobenius norm:

$$(\hat{\mathbf{A}}_1, \cdots, \hat{\mathbf{A}}_K) = \underset{A_1, \cdots, A_K}{\mathrm{argmin}} \|\hat{\mathbf{\Phi}} - \mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1\|_F^2$$

Projection Method: One-Term Models

Lemma 1

There exist a rearrangement operator $\mathcal{T} : \mathbb{R}^{d_1 \cdots d_K \times d_1 \cdots d_K} \to \mathbb{R}^{d_1^2 \times \cdots \times d_K^2}$ *such that*

$$\|\hat{\mathbf{\Phi}} - \mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1\|_F^2 = \|\mathcal{T}(\Phi) - \mathbf{a}_K \circ \cdots \circ \mathbf{a}_1\|_F^2$$

where $\mathbf{a}_1 = vec(\mathbf{A}_1), \cdots, \mathbf{a}_K = vec(\mathbf{A}_K)$ *and* $\circ$ *is outer product.*

- When $K = 2$ this minimization problem is called *Nearest Kronecker Product* (NKP) problem in matrix computation (Van Loan, 2000).
- When $K > 2$, consider it as the best rank-1 approximation problem for tensor.

One-Term Models LSE estimators:

$$(\tilde{\mathbf{A}}_1, \cdots, \tilde{\mathbf{A}}_K) = \underset{\mathbf{A}_1, \cdots, \mathbf{A}_K}{\operatorname{argmin}} \sum_t \|\mathcal{X}_t - \mathcal{X}_{t-1} \times_1 \mathbf{A}_1 \times_2 \cdots \times_K \mathbf{A}_K\|_F^2$$

To solve it numerically, we can update the K matrices $\tilde{\mathbf{A}}_1, \cdots, \tilde{\mathbf{A}}_K$ iteratively. By gradient conditions, the iteration of updating $\mathbf{A}_i$ given $\mathbf{A}_2, \cdots, \mathbf{A}_K$ is

$$\mathbf{A}_i \leftarrow \Big(\sum_t \mathcal{X}_{t,(i)} \mathbf{W}_{t,(i)}\Big)\Big(\sum_t \mathbf{W}'_{t,(i)} \mathbf{W}_{t,(i)}\Big)^{-1}$$

where $\mathbf{W}_{t,(i)} := (\mathcal{X}_{t-1} \times_1 \cdots \times_{i-1} \mathbf{A}_{i-1} \times_{i+1} \mathbf{A}_{i+1} \cdots \times_K \mathbf{A}_K)'_{(i)}$, and $\mathcal{X}_{t,(i)}$ is the i-th unfolding of tensor $\mathcal{X}_t$, $1 \leq i \leq K$.

MLE under a structured covariance tensor

We also consider a structured covariance matrix

$$\mathcal{E}_t = \mathcal{Z}_t \times_1 \mathbf{\Sigma}_1^{1/2} \cdots \times_K \mathbf{\Sigma}_K^{1/2}$$

where tensor $\mathcal{Z}_t$ has iid standard normal entries,

$$\text{cov}(\text{vec}(\mathcal{E}_t)) = \mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_1$$

The log likelihood under normality can be written as, for any $1 \le k \le K$,

$$-\sum_i^K (T-1) \prod_{l \neq i} d_l \log |\mathbf{\Sigma}_i| - \sum_t \text{tr}[\mathbf{\Sigma}_k^{-1} \mathcal{R}_{t,(1)} \mathbf{S}_k^{-1} \mathcal{R}'_{t,(1)}]$$

where

$$\mathbf{S}_i = \mathbf{\Sigma}_K \otimes \cdots \otimes \mathbf{\Sigma}_{i+1} \otimes \mathbf{\Sigma}_{i-1} \otimes \cdots \otimes \mathbf{\Sigma}_1$$

$$\mathcal{R}_t = \mathcal{X}_t - \sum_{j=1}^R \mathcal{X}_{t-1} \times_1 \mathbf{A}_1^{(j)} \times_2 \cdots \times_K \mathbf{A}_K^{(j)}$$

Asymptotics: Some Notations

- Let $\mathbf{a}_i := \text{vec}(\mathbf{A}_i)$.
- $\gamma_i := (0', \mathbf{a}_i', 0')'$ be a vector in $\mathbb{R}^{d_1^2+\cdots+d_K^2}$ for $1 \leq i \leq K$.
- $\mathbf{\Sigma}$ is the covariance matrix of $\text{vec}(\mathcal{E}_t)$.
- Permutation matrix $\mathbf{Q}_i$ are such that: for tensor $\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_n}$,

$$\text{vec}(\mathcal{X}_{(i)}) = \mathbf{Q}_i \text{vec}(\mathcal{X})$$

where $\mathcal{X}_{(i)}$ is i-th unfolding of tensor $\mathcal{X}$.

Asymptotics: Least Squares Estimators

Theorem 1 (CLT for One-Term Model Least Squares Estimators)

Define $\mathbf{H} := \mathbb{E}(\mathbf{W}_t\mathbf{W}_t') + \sum_{i=1}^{K-1} \gamma_i\gamma_i'$, $\Xi_2 =: \mathbf{H}^{-1}\mathbb{E}(\mathbf{W}_t\Sigma\mathbf{W}_t')\mathbf{H}^{-1}$ *where*

$$\mathbf{W}_t = \begin{pmatrix} (((\mathcal{X}_t)_{(1)}\mathbf{\Phi}_1') \otimes \mathbf{I}_{d_1})\mathbf{Q}_1 \\ \cdots \\ (((\mathcal{X}_t)_{(K)}\mathbf{\Phi}_K') \otimes \mathbf{I}_{d_K})\mathbf{Q}_K \end{pmatrix}$$

$\mathbf{\Phi}_i = \mathbf{A}_K \otimes \cdots \otimes \mathbf{A}_1$. *Assume that* $\mathcal{E}_t$, $1 \le t \le T$, *are iid with mean zero and finite second moments. Also assume the causality condition, and* $\mathbf{A}_k$, $1 \le k \le K$, $\mathbf{\Sigma}$ *are nonsigular. Then it holds that*

$$\sqrt{T}\begin{pmatrix} vec(\tilde{\mathbf{A}}_1 - \mathbf{A}_1) \\ \cdots \\ vec(\tilde{\mathbf{A}}_K - \mathbf{A}_K) \end{pmatrix} \to \mathcal{N}(0, \Xi_2)$$

Asymptotics: Least Squares Estimators

Theorem 2 (CLT for Multi-Term Model Least Squares Estimators)

Define $\mathbf{H} := \mathbb{E}(\mathbf{W}_t\mathbf{W}_t') + \sum_{j=1}^{r}\sum_{i=1}^{K-1}\gamma_{ij}\gamma_{ij}'$, $\Xi_2 =: \mathbf{H}^{-1}\mathbb{E}(\mathbf{W}_t\mathbf{\Sigma}\mathbf{W}_t')\mathbf{H}^{-1}$,

$$\mathbf{W}_t = \begin{pmatrix} \mathbf{W}_t^{(1)} \\ \cdots \\ \mathbf{W}_t^{(R)} \end{pmatrix}, \quad \mathbf{W}_t^{(i)} = \begin{pmatrix} (((\mathcal{X}_t)_{(1)}\mathbf{\Phi}_1^{(i)'}) \otimes \mathbf{I}_{d_1})\mathbf{Q}_1 \\ \cdots \\ (((\mathcal{X}_t)_{(K)}\mathbf{\Phi}_K^{(i)'}) \otimes \mathbf{I}_{d_K})\mathbf{Q}_K \end{pmatrix}, \quad 1 \le i \le R$$

where $\mathbf{\Phi}_i^{(j)} = \mathbf{A}_K^{(j)} \otimes \cdots \otimes \mathbf{A}_1^{(j)}$, $1 \le i \le K$, $1 \le j \le R$. *Then it holds that*

$$\sqrt{T}\begin{pmatrix} vec(\hat{\mathbf{A}}_1^{(1)} - \mathbf{A}_1) \\ \cdots \\ vec(\hat{\mathbf{A}}_K^{(R)} - \mathbf{A}_K) \end{pmatrix} \to \mathcal{N}(0, \Xi_2)$$

Asymptotics: MLE under Kronecker Structured Covariance

Theorem 3 (CLT for One-Term Model MLE Estimators)

Under the same condition of Theorem 1, and the additional assumption (12). Define $\bar{\mathbf{H}} := \mathbb{E}(\mathbf{W}_t \Sigma^{-1} \mathbf{W}_t') + \sum_{i=1}^{K-1} \gamma_i \gamma_i'$,
$\Xi_3 =: \bar{\mathbf{H}}^{-1} \mathbb{E}(\mathbf{W}_t \Sigma^{-1} \mathbf{W}_t') \bar{\mathbf{H}}^{-1}$. *Then it holds that*

$$\sqrt{T} \begin{pmatrix} vec(\bar{\mathbf{A}}_1 - \mathbf{A}_1) \\ \cdots \\ vec(\bar{\mathbf{A}}_K - \mathbf{A}_K) \end{pmatrix} \to \mathcal{N}(0, \Xi_3)$$

There are similar results for Muti-term Model MLE estimators.

Determining The Number of Terms

Multi-term model with R terms:

$$\mathcal{X}_t = \sum_{i=1}^{R} \mathcal{X}_{t-1} \times_1 \mathbf{A}_1^{(i)} \times_2 \mathbf{A}_2^{(i)} \times_3 \cdots \times_K \mathbf{A}_K^{(i)} + \mathcal{E}_t \tag{3}$$

We discussed the information criteria about determining the number of terms.

$$\mathrm{CP}(k) = \frac{1}{NT} \sum_t \|\mathrm{vec}(\mathcal{X}_t) - \mathbf{\Phi}\mathrm{vec}(\mathcal{X}_{t-1})\|_F^2 + k \cdot g(N, T) \tag{4}$$

where $\mathbf{\Phi} = \sum_{i=1}^R \mathbf{A}_K^{(i)} \otimes \mathbf{A}_{K-1}^{(i)} \otimes \cdots \otimes \mathbf{A}_1^{(i)}$, $N = d_1 d_2 \cdots d_K$, and $g(N, T)$ controls the weight of penalty.

Determining The Number of Terms

Assumption 1

Assume that $N/T \to 0$, *as* $N, T \to \infty$.

Assumption 2

Under Assumption 1. Assume that $\liminf_{T,N\to\infty} \lambda_{min}(\frac{\mathbf{X}'\mathbf{X}}{T}) \xrightarrow{a.s.} c > 0$, *where the columns of the matrix* $\mathbf{X} \in \mathbb{R}^{T\times N}$ *are those* $vec[(\mathcal{X}_t)]$

Assumption 3

Consider the model (3). Assume that for $1 \le r \le R$, *there esists some consitant* $\eta > 0$ *such that* $\|\mathbf{A}_K^{(r)} \otimes \cdots \otimes \mathbf{A}_1^{(r)}\|_F^2 > \eta N$.

Assumption 3 is mild and can be often satisfied by generic $\mathbf{A}_1, \cdots, \mathbf{A}_K$.

Determining The Number of Terms

Theorem 4

Suppose Assumption 1 to 3 holds. Set R be the true number of terms and $\hat{k} = \operatorname{argmin} CP(k)$. If

- $g(N, T) \to 0$;
- $T \cdot g(N, T) \to \infty$, *as* $N, T \to \infty$.

Then $\lim_{N,T\to\infty} P(\hat{k} = R) = 1$

Corollary 1

Under the Assumptions of Theorem 4, the class of criteria defined by

$$IC(k) := \log\left(\frac{1}{NT}\|vec(\mathcal{X}_t) - \mathbf{\Phi} vec(\mathcal{X}_{t-1})\|_F^2\right) + k \cdot g(N, T) \tag{5}$$

will also consistently estimate R.

Experiment I: Comparison of Estimators

- Experiment I: Comparison of estimators PROJ, LSE, MLE and VAR.
  - Setting I: $\text{cov}(\text{vec}(\mathcal{E}_t)) = \mathbf{\Sigma} = \mathbf{I}$.
  - Setting II: $\text{cov}(\text{vec}(\mathcal{E}_t)) = \mathbf{\Sigma}$ is arbitrary.
  - Setting III: $\text{cov}(\text{vec}(\mathcal{E}_t))$ takes the kronecker product form.

  We repeat the simulation 1000 times and show a box plot of

$$\log \|\hat{\mathbf{A}}_1 \otimes \hat{\mathbf{A}}_2 \otimes \hat{\mathbf{A}}_3 - \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3\|_F^2 \tag{6}$$

- Experiment II: Percentage of Coverages of Confidence Interval.
- Experiment III: Determine the number of terms.

Experiment I: Comparison of Estimators

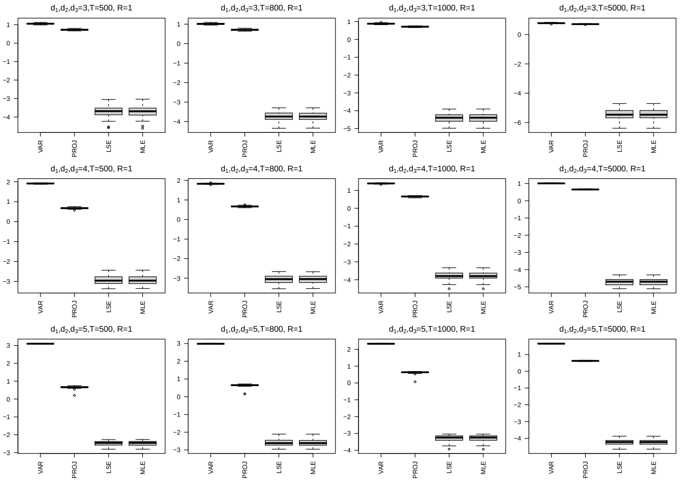


Figure 3: In one-term model, comparison of four estimators, PROJ, LSE, MLEs and VAR, under setting I.

Experiment I: Comparison of Estimators

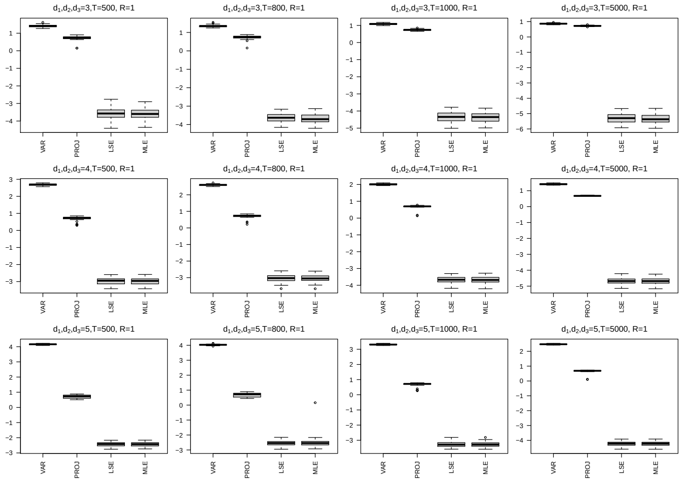


Figure 4: In one-term model, comparison of four estimators, PROJ, LSE, MLEs and VAR, under setting II.

Experiment I: Comparison of Estimators

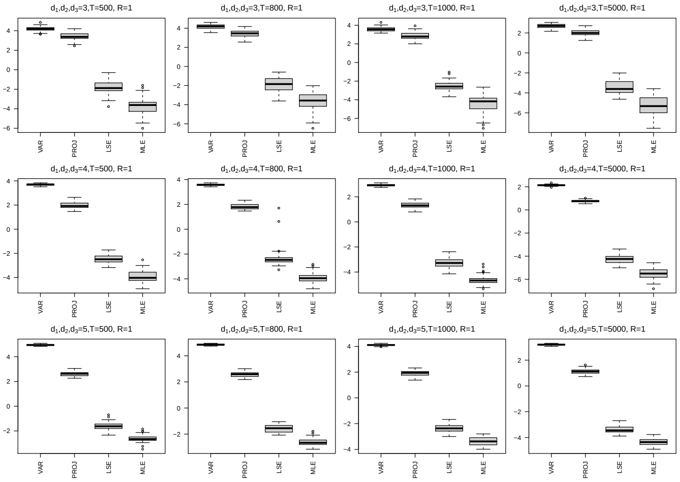


Figure 5: In one-term model, comparison of four estimators, PROJ, LSE, MLEs and VAR, under setting III.

Experiment I: Comparison of Estimators

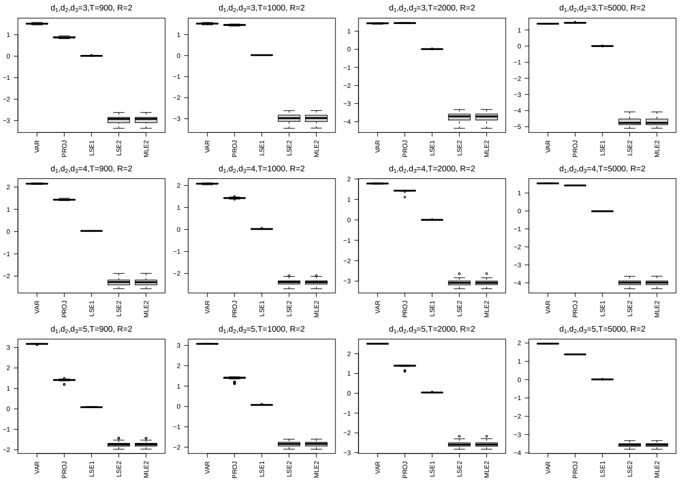


Figure 6: In two-term model, comparison of four estimators, LSE1, LSE2, MLE2, VAR, and PROJ, under setting I.

Experiment I: Comparison of Estimators

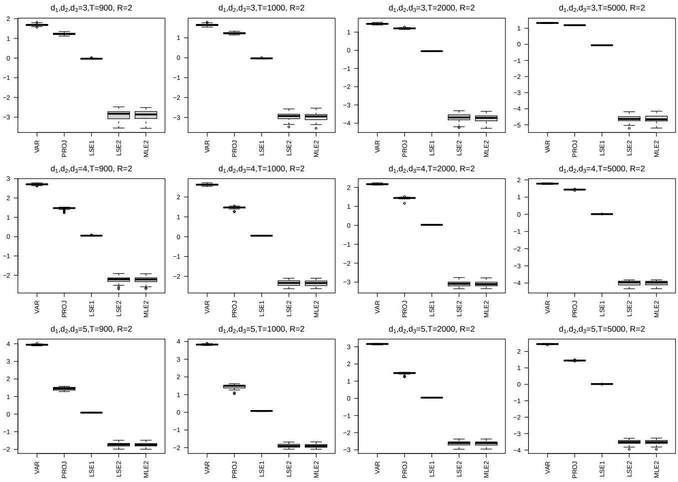


Figure 7: In two-term model, comparison of four estimators, LSE1, LSE2, MLE2, VAR, and PROJ, under setting II.

Experiment I: Comparison of Estimators

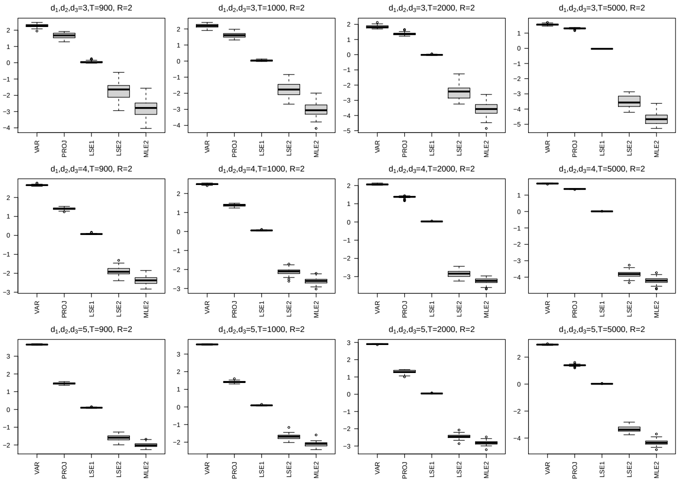


Figure 8: In two-term model, comparison of four estimators, LSE1, LSE2, MLE2, VAR, and PROJ, under setting III.

Experiment II: Percentage of coverages

	Setting	I		II		III	
	Estimator	LSE	MLEs	LSE	MLEs	LSE	MLEs
$R = 1$	T=100	0.945	0.941	0.940	0.771	0.937	0.944
	T=200	0.951	0.950	0.941	0.774	0.946	0.948
	T=1000	0.953	0.952	0.951	0.776	0.955	0.956
$R = 2$	T=500	0.937	0.937	0.934	0.706	0.907	0.933
	T=1000	0.943	0.942	0.941	0.726	0.906	0.935
	T=2000	0.950	0.950	0.943	0.724	0.920	0.936

Table 1: Percentage of coverages of 95% confidence intervals.

Experiment III: Determine the number of terms

Consider the following criteria:

$$
\begin{aligned}
\mathrm{IC}_1(k) &= \log\left(\frac{1}{NT}\|\mathbf{Y}^{'} - \Phi\mathbf{X}^{'}\|_F^2\right) + 2k\frac{\log(T)}{T} \\
\mathrm{IC}_2(k) &= \log\left(\frac{1}{NT}\|\mathbf{Y}^{'} - \Phi\mathbf{X}^{'}\|_F^2\right) + 2k\frac{\log(N)}{T} \\
\mathrm{IC}_3(k) &= \log\left(\frac{1}{NT}\|\mathbf{Y}^{'} - \Phi\mathbf{X}^{'}\|_F^2\right) + 2k\frac{\log(d_1^2 + d_2^2 + d_3^2)}{T}
\end{aligned}
$$

The experiments shows that above three criteria can choose the true number of terms 100%, out of 100 repetitions, under different tensor size $3 \times 3 \times 3$, $5 \times 5 \times 5$ and different true number of terms $R = 1, 2, 3$.

Data from Fama-French Factor Model

We are using the 32 Portfolios Formed on Size, Book-to-Market, and Operating Profitability, from Fama-French Data Library, which formed a $4 \times 4 \times 2$ tensor time series, from July 1963 to October 2019.

- Determine the number of terms by IC_1, IC_2, IC_3. $(\hat{R} = 1)$
- Estimate the coefficient matrices.
- Obtain out-sample rolling forecast performances.

Fama-French Factor Model: Coefficient Matrices

	LoOP	50%OP	75%OP	HiOP	LoOP	50%OP	75%OP	HiOP
LoOP	-0.100 (0.112)	-0.080 (0.195)	0.463 (0.126)	-0.483 (0.083)	0	0	+	-
50%OP	0.015 (0.089)	-0.033 (0.164)	0.202 (0.137)	-0.368 (0.065)	0	0	0	-
75%OP	0.069 (0.097)	-0.047 (0.162)	0.151 (0.150)	-0.382 (0.069)	0	0	0	-
HiOP	0.079 (0.111)	0.047 (0.179)	0.116 (0.189)	-0.490 (0.091)	0	0	0	-

Table 2: Estimated coefficient matrix $\mathbf{A}_1$ using LSE method. Standard errors are shown in the parentheses. The right panel indicates the positively significant, negatively significant and insignificant parameters at 5% level using symbols $(+, -, 0)$, respectively.

Fama-French Factor Model: Coefficient Matrices

	LoBM	50%BM	75%BM	HiBM	LoBM	50%BM	75%BM	HiBM
LoBM	0.428 (0.186)	0.173 (0.222)	-0.362 (0.148)	0.226 (0.111)	+	0	-	+
50%BM	0.331 (0.162)	0.148 (0.189)	-0.171 (0.125)	0.171 (0.089)	+	0	0	+
75%BM	0.265 (0.159)	0.145 (0.183)	-0.138 (0.125)	0.192 (0.092)	0	0	0	+
HiBM	0.293 (0.196)	0.315 (0.208)	-0.231 (0.144)	0.251 (0.114)	0	0	0	+

Table 3: Estimated coefficient matrix $\mathbf{A}_2$ using LSE method. Standard errors are shown in the parentheses. The right panel indicates the positively significant, negatively significant and insignificant parameters at 5% level using symbols $(+, -, 0)$, respectively.

Fama-French Factor Model: Coefficient Matrices

	Small	Big	Small	Big
Small	-1.211 (0.579)	-0.508 (0.335)	-	0
Big	-0.470 (0.307)	-0.380 (0.254)	0	0

Table 4: Estimated coefficient matrix $\mathbf{A}_3$ using LSE method. Standard errors are shown in the parentheses. The right panel indicates the positively significant, negatively significant and insignificant parameters at 5% level using symbols $(+, -, 0)$, respectively.

Fama-French Factor Model: Rolling Forecast

LSE1	MLE1	LSE2	MLE2	LSE3	MLE3	iAR(1)	iAR(2)	VAR(1)
1195.36	1202.44	1208.87	1177.74	1198.54	1181.89	1197.11	1204.31	1247.88

Table 5: Rolling forecast mean square errors of LSE, MLE with $R = 1, 2, 3$ as well as univariate AR(1) and AR(2), vector AR(1) models for comparison. Starting from 2013 May ($t = 600$) to 2019 Oct ($t = 677$).

Thanks for listening!