High-Dimensional Low-Rank Tensor Autoregressive Time Series Modeling

Di Wang[†], Yao Zheng[‡] and Guodong Li[†]

[†] *University of Hong Kong and* [‡] *University of Connecticut*

January 19, 2021

Abstract

Modern technological advances have enabled an unprecedented amount of structured data with complex temporal dependence, urging the need for new methods to efficiently model and forecast high-dimensional tensor-valued time series. This paper provides the first practical tool to accomplish this task via autoregression (AR). By considering a low-rank Tucker decomposition for the transition tensor, the proposed tensor autoregression can flexibly capture the underlying low-dimensional tensor dynamics, providing both substantial dimension reduction and meaningful dynamic factor interpretation. For this model, we introduce both low-dimensional rank-constrained estimator and high-dimensional regularized estimators, and derive their asymptotic and non-asymptotic properties. In particular, by leveraging the special balanced structure of the AR transition tensor, a novel convex regularization approach, based on the sum of nuclear norms of square matricizations, is proposed to efficiently encourage low-rankness of the coefficient tensor. A truncation method is further introduced to consistently select the Tucker ranks. Simulation experiments and real data analysis demonstrate the advantages of the proposed approach over various competing ones.

Keywords: dimension reduction; high-dimensional time series; nuclear norm; tensor decomposition; tensor-valued data

1 Introduction

Modern technological development has made available enormous quantities of data, many of which are structured and collected over time. Tensor-valued time series data, namely observations on a set of variables structured in a tensor form collected over time, have become increasingly common in a wide variety of areas. Examples include multiple macroeconomic indices time series for multiple countries (?), dynamic inter-regional transport flow data (?), and sequential image and video processing (?), among many others.

To model temporal dependencies of tensor-valued time series data, naturally one might resort to vectorization of the tensor-valued observations so that conventional vector-valued time series models can be directly applied. For instance, let $\mathcal{Y}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ be the tensor-valued observation at time t, for $t = 1, \ldots, T$, where T is the sample size. Then an obvious first step is to conduct the following vector autoregression (VAR) for its vectorization:

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

where $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ is the unknown transition matrix, with $p = \prod_{i=1}^{d} p_i$, and $\mathcal{E}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the tensor-valued random error at time t. Although model (??) is potent in that it takes into account the linear association between every variable in $\mathcal{Y}_t$ and that in $\mathcal{Y}_{t-1}$, it clearly suffers from two major drawbacks:

- The vectorization will destroy the important multidimensional structural information inherently embedded in the tensor-valued observations, resulting in a lack of interpretability;
- The number of unknown parameters, $p^2 = (\prod_{i=1}^{d} p_i)^2$, can be formidable even for small d and p_i; e.g., even when $d = 3$ and $p_1 = p_2 = p_3 = 3$, the number of unknown parameters will be as large as $(3 \times 3 \times 3)^2 = 729$, while the sample size T often has a similar magnitude in practice; see, e.g., the real data examples in Section ??.

In this work, motivated by the idea of Tucker decomposition (?), we propose the Low-Rank Tensor Autoregressive (LRTAR) model through folding the $p \times p$ transition matrix

$\boldsymbol{A}$ in (**??**) into the $2d$-th-order transition tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times\cdots\times p_d\times p_1\times\cdots\times p_d}$ which is assumed to have low multilinear ranks $(r_1,\ldots,r_{2d})$. That is, r_i can be much smaller than p_i, where $p_{d+i} = p_i$ for $i = 1,\ldots,d$. As a result, the aforementioned drawbacks of the vectorization approach can be overcome by the proposed model:

- As we will show in Section **??**, the proposed model can flexibly retain the distinct structural information across all modes of observed tensor process, encapsulating an interpretable low-dimensional tensor dynamic relationship between $\boldsymbol{\mathcal{Y}}_t$ and $\boldsymbol{\mathcal{Y}}_{t-1}$.

- As the transition tensor $\boldsymbol{\mathcal{A}}$ is assumed to have low multilinear ranks, the parameter space is simultaneously constrained in $2d$ directions, reducing its dimension from $(\prod_{i=1}^{d} p_i)^2$ drastically to $\prod_{i=1}^{2d} r_i + \sum_{i=1}^{d} r_i(p_i - r_i) + \sum_{i=1}^{d} r_{d+i}(p_i - r_{d+i})$.

Recently there has been a rapidly emerging interest in high-dimensional matrix- and tensor-valued time series analysis, particularly through factor models, e.g., the matrix factor model proposed in **?**, the constrained matrix factor model in **?**, and the tensor factor model in **?**. Factor models are powerful tools for dimension reduction with great interpretability. However, unlike autoregression, factor models do not seek to explicitly model the temporal dependency and thus by themselves cannot be directly used for forecasting. On the other hand, despite the extensive literature on high-dimensional VAR models in recent years (e.g. **?????**), counterparts able to meet the particular needs of matrix- and tensor-valued time series modeling tasks have been rarely explored. The most relevant existing work in this direction so far might be the matrix autoregressive (MAR) model in **?**, where the focus is on low-dimensional modeling; see also **?**. As we will discuss in Section **??**, the proposed model includes the MAR model as a special case, but enjoys greater flexibility and a more drastic reduction of the dimensionality; see also Figure **??** for an illustration.

The estimation of the proposed model is studied under both low- and high-dimensional scaling. When the sample size T is sufficiently large and the transition tensor $\boldsymbol{\mathcal{A}}$ is exactly low-rank, we consider the rank-constrained estimation method and prove the asymptotic normality for the proposed low-Tucker-rank (LTR) estimator. Under the high-dimensional setup, we consider a more general and natural setting where $\mathcal{A}$ can be well approximated by a low-Tucker-rank tensor, and develop regularized estimation methods based on nuclear-norm-type penalties. We first study the Sum of Nuclear (SN) norm regularizer, defined as the sum of nuclear norms of all one-mode matricizations of $\mathcal{A}$, and derive the non-asymptotic estimation error bound for the corresponding estimator. The SN norm regularizer has been widely used in the literature for various low-rank tensor problems (**????**). Its major strength lies in the fact that the summation of nuclear norms allows enforcing the low-rankness simultaneously across all modes of the tensor. In contrast, if only a single nuclear norm is used, the low-rankness of only one mode will be accounted for, obviously leading to a much less efficient estimator.

However, although penalizing multiple one-mode matricizations of $\mathcal{A}$ simultaneously is far more efficient than penalizing only one of them, the SN regularized estimator suffers from serious efficiency loss due to the *fat-and-short* shape of the one-mode matricizations. Note that the low-rankness in fact can also be encouraged by penalizing nuclear norms of multi-mode matricizations. For instance, the conventional Matrix Nuclear (MN) norm regularized estimator (**?**) simply penalizes the nuclear norm of the transition matrix $\boldsymbol{A}$ in representation (**??**), which under the proposed LRTAR model actually is a square-shaped multi-mode matricization, namely square matricization, of the transition tensor $\mathcal{A}$. As we will show in Theorem **??** and simulations, even though the MN regularizer incorporates only one single square matricization, it still clearly beats the SN regularizer, since the former avoids the efficiency bottleneck caused by the imbalance of the one-mode matricization.

Indeed, due to the autoregressive form of the proposed model, the transition tensor $\mathcal{A}$ has a special balanced structure in the sense that its first d modes have exactly the same dimensions as its other d modes. As a result, actually many different square matricizations of $\mathcal{A}$ can be formed by appropriately pairing up its modes; see Section **??** for details. This naturally motivates us to propose a new regularizer that combines the strengths of both SN and MN norms. Specifically, for the proposed tensor autoregression, we introduce a novel Sum of Square-matrix Nuclear (SSN) norm regularizer, defined as the sum of nuclear norms of all the $p \times p$ square matricizations of the transition tensor $\boldsymbol{\mathcal{A}}$. The SSN regularizer is expected to be superior to the MN, since it simultaneously encourages the low-rankness across all possible square matricizations of $\boldsymbol{\mathcal{A}}$ rather than only one of them. Moreover, thanks to the use of square matricizations, the SSN regularized estimator is provably more efficient than the SN regularized one; see Theorem **??** and the last simulation experiment in Section **??**. Note that the adoption of a more balanced (square) matricization to improve estimation performance was proposed in **?** for low-rank tensor completion problems, where only a single square matricization was penalized, similarly to the MN regularizer.

This work is also related to the literature of matrix-variate regression and tensor regression for independent data. The matrix-variate regression model in **?** has the same basic bilinear form as that of the MAR model mentioned earlier, while an envelope method was introduced to further reduce the dimension. **?** proposed a multi-response tensor regression model, where they mainly studied the third-order coefficient tensor and the SN regularization which is known to be statistically sub-optimal for higher-order tensor estimation. In contrast, we study the model for general higher-order tensor-valued time series. Moreover, our SSN regularized estimator has a much faster statistical convergence rate than the SN estimator. Recently, **?** and **?** studied non-convex projected gradient descent methods for tensor regression. While their non-convex approaches require exact low-rankness with known Tucker ranks, our methods can handle both exact and approximate low-rankness and select the unknown ranks consistently in the exactly low-rank case. In addition, existing literature on tensor regression has only considered independent data and Gaussian time series data, whereas we allow sub-Gaussianity of the time series. This is a non-trivial relaxation, since unlike the Gaussian case, sub-Gaussian time series cannot be linearly transformed into independent samples.

We summarize the most important contributions of this paper as follows:

(i) This paper provides the first practical tool to model and forecast general structured, high-dimensional data with complex temporal dependence via tensor autoregression. By flexibly and efficiently capturing the underlying low-dimensional tensor dynamics, the proposed model delivers significant dimension reduction, meaningful structural interpretation and favorable forecast performance.

(ii) By exploiting the special balanced structure of the transition tensor $\boldsymbol{\mathcal{A}}$, a novel SSN regularization approach is introduced to simultaneously and efficiently encourage low-rankness across all square matricizations of $\boldsymbol{\mathcal{A}}$, outperforming both the SN and MN methods under both exact and approximate low-rankness. For exactly low-rank $\boldsymbol{\mathcal{A}}$, a truncated estimator is further introduced for consistent rank selection.

(iii) On the technical front, by establishing a novel martingale-based concentration bound, this paper relaxes the conventional Gaussian assumption in the literature on high-dimensional time series to sub-Gaussianity. This technique is generally applicable to the non-asymptotic estimation theory for high-dimensional time series models with a VAR representation and hence is of independent interest.

The rest of the paper is organized as follows. Section **??** introduces basic notation and tensor algebra. Section **??** presents the proposed LRTAR model. Section **??** studies the low-dimensional least squares estimator and its asymptotic properties. The high-dimensional regularized estimation is covered in Section **??**, where we develop the non-asymptotic theory for three regularized estimators and rank selection consistency for the truncated estimator. Sections **??** and **??** present simulation studies and real data analysis, respectively. Section **??** concludes with a brief discussion. All technical proofs are relegated to the Appendix.

2 Preliminaries: Notation and Tensor Algebra

Tensors, also known as multidimensional arrays, are natural higher-order extensions of matrices. The order of a tensor is known as the dimension, way or mode, so a multidimensional array $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is called a d-th-order tensor. We refer readers to **?** for a detailed review of basic tensor algebra.

Throughout this paper, we denote vectors by boldface small letters, e.g. $\boldsymbol{x}$, $\boldsymbol{y}$, matrices by boldface capital letters, e.g. $\boldsymbol{X}$, $\boldsymbol{Y}$, and tensors by boldface Euler capital letters, e.g. $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$. For any two real-valued sequences x_k and y_k, we write $x_k \gtrsim y_k$ if there exists a constant $c > 0$ such that $x_k \geq c y_k$ for all k, and write $x_k \gg y_k$ if $\lim_{k\to\infty} y_k/x_k = 0$. In addition, write $x_k \asymp y_k$ if $x_k \gtrsim y_k$ and $y_k \gtrsim x_k$. We use C to denote a generic positive constant, which is independent of the dimensions and the sample size.

For a generic matrix $\boldsymbol{X}$, we let $\boldsymbol{X}^\top$, $\|\boldsymbol{X}\|_{\mathrm{F}}$, $\|\boldsymbol{X}\|_{\mathrm{op}}$, $\|\boldsymbol{X}\|_*$, $\mathrm{vec}(\boldsymbol{X})$, and $\sigma_j(\boldsymbol{X})$ denote its transpose, Frobenius norm, operator norm, nuclear norm, vectorization, and j-th largest singular value, respectively. For any matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, recall that the nuclear norm and its dual norm, the operator norm, are defined as

$$\|\boldsymbol{X}\|_* = \sum_{j=1}^{\min(m,n)} \sigma_j(\boldsymbol{X}) \quad \text{and} \quad \|\boldsymbol{X}\|_{\mathrm{op}} = \sigma_1(\boldsymbol{X}).$$

For any square matrix $\boldsymbol{X}$, we let $\lambda_{\min}(\boldsymbol{X})$ and $\lambda_{\max}(\boldsymbol{X})$ denote its minimum and maximum eigenvalues. For any real symmetric matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, we write $\boldsymbol{X} \leq \boldsymbol{Y}$ if $\boldsymbol{Y} - \boldsymbol{X}$ is a positive semidefinite matrix.

Matricization, also known as unfolding, is the process of reordering the elements of a third- or higher-order tensor into a matrix. The most commonly used matricization is the one-mode matricization defined as follows. For any d-th-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$, its mode-$s$ matricization $\boldsymbol{\mathcal{X}}_{(s)} \in \mathbb{R}^{p_s \times p_{-s}}$, with $p_{-s} = \prod_{i=1,i\neq s}^d p_i$, is the matrix obtained by setting the s-th tensor mode as its rows and collapsing all the others into its columns, for $s = 1, \ldots, d$. Specifically, the $(i_1, \ldots, i_d)$-th element of $\boldsymbol{\mathcal{X}}$ is mapped to the (i_s, j)-th element of $\boldsymbol{\mathcal{X}}_{(s)}$, where

$$j = 1 + \sum_{\substack{k=1\\k\neq s}}^{d} (i_k - 1) J_k \quad \text{with} \quad J_k = \prod_{\substack{\ell=1\\ \ell\neq s}}^{k-1} p_\ell.$$

The above one-mode matricization can be extended to the multi-mode matricization by combining multiple modes to rows and combining the rest to columns of a matrix. For any index subset $S \subset \{1, \ldots, d\}$, the multi-mode matricization $\boldsymbol{\mathcal{X}}_{[S]}$ is the $\prod_{i\in S} p_i$-by-$\prod_{i\notin S} p_i$ matrix whose (i, j)-th element is mapped from the $(i_1, \ldots, i_d)$-th element of $\boldsymbol{\mathcal{X}}$, where

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

Note that the modes in the multi-mode matricization are collapsed following their original order $1, \ldots, d$. Moreover, it holds $\mathcal{X}_{[S]} = \mathcal{X}_{[S^\complement]}^\top$, where $S^\complement = \{1, \ldots, d\} \setminus S$ is the complement of S. In addition, the one-mode matricization $\mathcal{X}_{(s)}$ defined above is simply $\mathcal{X}_{[\{s\}]}$.

We next review the concepts of tensor-matrix multiplication, tensor generalized inner product and norm. For any d-th-order tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ and matrix $\boldsymbol{Y} \in \mathbb{R}^{q_k \times p_k}$ with $1 \leq k \leq d$, the mode-k multiplication $\mathcal{X} \times_k \boldsymbol{Y}$ produces a d-th-order tensor in $\mathbb{R}^{p_1 \times \cdots \times p_{k-1} \times q_k \times p_{k+1} \times \cdots \times p_d}$ defined by

$$(\mathcal{X} \times_k \boldsymbol{Y})_{i_1 \cdots i_{k-1} j i_{k+1} \cdots i_d} = \sum_{i_k=1}^{p_k} \mathcal{X}_{i_1 \cdots i_d} \boldsymbol{Y}_{j i_k}.$$

For any two tensors $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_n}$ and $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times \cdots p_m}$ with $n \geq m$, their generalized inner product $\langle \mathcal{X}, \mathcal{Y} \rangle$ is the $(n-m)$-th-order tensor in $\mathbb{R}^{p_{m+1} \times \cdots \times p_n}$ defined by

$$\langle \mathcal{X}, \mathcal{Y} \rangle_{i_{m+1} \ldots i_n} = \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_m=1}^{p_m} \mathcal{X}_{i_1 i_2 \ldots i_m i_{m+1} \ldots i_n} \mathcal{Y}_{i_1 i_2 \ldots i_m},$$

where $1 \leq i_{m+1} \leq p_{m+1}, \ldots, 1 \leq i_n \leq p_n$. In particular, when $n = m$, it reduces to the conventional real-valued inner product. In addition, the Frobenius norm of any tensor $\mathcal{X}$ is defined as $\|\mathcal{X}\|_{\mathrm{F}} = \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$.

Some basic properties of the tensor generalized inner product are as follows. Let $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_n}$, $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times \cdots p_m}$ and $\mathcal{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_{k-1} \times q_k \times p_{k+1} \cdots \times p_m}$ be tensors with $n \geq m \geq k \geq 1$. If $\boldsymbol{Y} \in \mathbb{R}^{q_k \times p_k}$, then

$$\langle \mathcal{X} \times_k \boldsymbol{Y}, \mathcal{Z} \rangle = \langle \mathcal{X}, \mathcal{Z} \times_k \boldsymbol{Y}^\top \rangle.$$

If $\boldsymbol{Z} \in \mathbb{R}^{q_{m+j} \times p_{m+j}}$ with $1 \leq j \leq n - m$, then

$$\langle \mathcal{X}, \mathcal{Y} \rangle \times_j \boldsymbol{Z} = \langle \mathcal{X} \times_{m+j} \boldsymbol{Z}, \mathcal{Y} \rangle.$$

Moreover,

$$\mathrm{vec}(\langle \mathcal{X}, \mathcal{Y} \rangle) = \mathcal{X}_{[S]} \mathrm{vec}(\mathcal{Y}),$$

where $S = \{m+1, \ldots, n\}$, and when $m = n$, $\mathcal{X}_{[\emptyset]} = \mathrm{vec}(\mathcal{X})^\top$.

Finally, we summarize some concepts and useful results of the Tucker decomposition (**??**). For any tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, its multilinear ranks $(r_1, \ldots, r_d)$ are defined as the matrix ranks of its one-mode matricizations, namely $r_i = \text{rank}(\boldsymbol{\mathcal{X}}_{(i)})$, for $i = 1, \ldots, d$. Note that r_i's are analogous to the row and column ranks of a matrix, but are not necessarily equal for third- and higher-order tensors. Suppose that $\boldsymbol{\mathcal{X}}$ has multilinear ranks $(r_1, \ldots, r_d)$. Then $\boldsymbol{\mathcal{X}}$ has the following Tucker decomposition:

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}} \times_1 \boldsymbol{Y}_1 \times_2 \boldsymbol{Y}_2 \cdots \times_d \boldsymbol{Y}_d = \boldsymbol{\mathcal{Y}} \times_{i=1}^d \boldsymbol{Y}_i,$$

where $\boldsymbol{Y}_i \in \mathbb{R}^{p_i \times r_i}$ for $i = 1, \ldots, d$ are the factor matrices and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ is the core tensor. Hence, the multilinear ranks are also called Tucker ranks.

If $\boldsymbol{\mathcal{X}}$ has the Tucker decomposition in (**??**), then we have the following results for its one- and multi-mode matricizations:

$$\boldsymbol{\mathcal{X}}_{(s)} = \boldsymbol{Y}_s \boldsymbol{\mathcal{Y}}_{(s)} (\boldsymbol{Y}_d \otimes \cdots \otimes \boldsymbol{Y}_{s+1} \otimes \boldsymbol{Y}_{s-1} \cdots \otimes \boldsymbol{Y}_1)^\top = \boldsymbol{Y}_s \boldsymbol{\mathcal{Y}}_{(s)} (\otimes_{i \neq s} \boldsymbol{Y}_i)^\top, \quad s = 1, \ldots, d,$$

and

$$\boldsymbol{\mathcal{X}}_{[S]} = (\otimes_{i \in S} \boldsymbol{Y}_i) \boldsymbol{\mathcal{Y}}_{[S]} (\otimes_{i \notin S} \boldsymbol{Y}_i)^\top, \quad S \subset \{1, \ldots, d\},$$

where $\otimes_{i \neq s}, \otimes_{i \in S}$ and $\otimes_{i \notin S}$ are matrix Kronecker products operating in the reverse order within the corresponding index sets.

Note that for any nonsingular matrices $\boldsymbol{O}_i \in \mathbb{R}^{r_i \times r_1}$ for $i = 1, \ldots, d$, it holds

$$\boldsymbol{\mathcal{Y}} \times_{i=1}^d \boldsymbol{Y}_i = (\boldsymbol{\mathcal{Y}} \times_{i=1}^d \boldsymbol{O}_i) \times_{i=1}^d (\boldsymbol{Y}_i \boldsymbol{O}_i^{-1}).$$

This indicates that the Tucker decomposition in (**??**) is not unique unless appropriate identifiability constraints are imposed. In this paper, to fix ideas, we will focus on a special Tucker decomposition called the higher-order singular value decomposition (HOSVD). In the HOSVD, the factor matrix $\boldsymbol{Y}_i$ in (**??**) is defined as the tall orthonormal matrix consisting of the top r_i left singular vectors of $\boldsymbol{\mathcal{X}}_{(i)}$, for $i = 1, \ldots, d$. Consequently, the core tensor $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} \times_{i=1}^d \boldsymbol{Y}_i^\top$ has the all-orthogonal property that $\boldsymbol{\mathcal{Y}}_{(i)} \boldsymbol{\mathcal{Y}}_{(i)}^\top$ is a diagonal matrix for $i = 1, \ldots, d$.

3 Low-Rank Tensor Autoregression

For the tensor-valued time series $\{\boldsymbol{\mathcal{Y}}_t\}_{t=1}^T$, we propose the following Low-Rank Tensor Autoregressive (LRTAR) model:

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

where $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_d \times p_1 \times \cdots \times p_d}$ is the $2d$-th-order transition tensor which is assumed to have multilinear ranks $(r_1, \dots, r_{2d})$, with $r_i = \text{rank}(\boldsymbol{\mathcal{A}}_{(i)})$, $\langle \cdot, \cdot \rangle$ is the generalized tensor inner product defined in (**??**), and $\boldsymbol{\mathcal{E}}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the mean-zero random error at time t with possible dependencies among its contemporaneous elements. Denote $S_1 = \{1, 2, \dots, d\}$ and $S_2 = \{d+1, d+2, \dots, 2d\}$. Note that by (**??**), model (**??**) can be written into the VAR form in (**??**) with transition matrix $\boldsymbol{A} = \boldsymbol{\mathcal{A}}_{[S_2]}$.

Then, we have the higher-order singular value decomposition (HOSVD) of $\boldsymbol{\mathcal{A}}$,

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_{i=1}^{2d} \boldsymbol{U}_i,$$

where $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times \cdots \times r_{2d}}$ is the core tensor, and each $\boldsymbol{U}_i \in \mathbb{R}^{p_i \times r_i}$ is the orthonormal factor matrix defined as the top r_i left singular vectors of $\boldsymbol{\mathcal{A}}_{(i)}$, for $1 \leq i \leq 2d$. Thus, by (**??**), the VAR representation of model (**??**) can be written as

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

where $\boldsymbol{y}_t = \text{vec}(\boldsymbol{\mathcal{Y}}_t)$ and $\boldsymbol{e}_t = \text{vec}(\boldsymbol{\mathcal{E}}_t)$.

In contrast to the conventional VAR model in (**??**) which has p^2 unknown parameters, where $p = \prod_{i=1}^d p_i$, the dimension of the parameter space for model (**??**) is reduced substantially to

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

which is computed by subtracting the number of constraints due to the orthonormality of $\boldsymbol{U}_i$'s and the all-orthogonal property of $\boldsymbol{\mathcal{G}}$ from the total number of parameters.

By the VAR representation in (**??**), we immediately have the necessary and sufficient condition for the existence of a unique strictly stationary solution to model (**??**) as follows.

Assumption 1. *The spectral radius of $\boldsymbol{\mathcal{A}}_{[S_2]}$ is strictly less than one.*

The proposed model implies an interesting low-dimensional tensor dynamical structure. To be specific, by (**??**), (**??**) and the orthonormality of $\boldsymbol{U}_i$, it can be shown that (**??**) together with (**??**) implies

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

Note that in (**??**), $\boldsymbol{\mathcal{Y}}_t$ and $\boldsymbol{\mathcal{E}}_t$ are both projected onto a low-dimensional space via the $\boldsymbol{U}_{d+i}$'s, while $\boldsymbol{\mathcal{Y}}_{t-1}$ is projected onto another low-dimensional space via the $\boldsymbol{U}_i$'s, with $1 \leq i \leq d$. Hence, (**??**) is essentially a low-dimensional tensor regression defined on the projections of $\boldsymbol{\mathcal{Y}}_t$ and $\boldsymbol{\mathcal{Y}}_{t-1}$. Element-wisely, the low-dimensional tensor $\boldsymbol{\mathcal{Y}}_t \times_{i=1}^{d} \boldsymbol{U}_{d+i}^\top$ can be interpreted as $\prod_{i=1}^{d} r_{d+i}$ multilinear response factors, $\boldsymbol{\mathcal{Y}}_{t-1} \times_{i=1}^{d} \boldsymbol{U}_{i}^\top$ as $\prod_{i=1}^{d} r_i$ multilinear predictor factors, and $\boldsymbol{\mathcal{E}}_t \times_{i=1}^{d} \boldsymbol{U}_{d+i}^\top$ as multilinear error factors. For this reason, we call $\boldsymbol{U}_{d+1}, \dots, \boldsymbol{U}_{2d}$ the response factor matrices, and $\boldsymbol{U}_1, \dots, \boldsymbol{U}_d$ the predictor factor matrices.

We discuss some special cases of the proposed model below.

Example 1. *For simplicity, we first consider the case with $d = 2$, so $\boldsymbol{\mathcal{Y}}_t \equiv \boldsymbol{Y}_t, \boldsymbol{\mathcal{E}}_t \equiv \boldsymbol{E}_t \in \mathbb{R}^{p_1\times p_2}$ are matrices. Then the VAR representation in* (**??**) *becomes*

$$\text{vec}(\boldsymbol{Y}_t) = (\boldsymbol{U}_4 \otimes \boldsymbol{U}_3)\boldsymbol{\mathcal{G}}_{[\{3,4\}]}(\boldsymbol{U}_2^\top \otimes \boldsymbol{U}_1^\top)\text{vec}(\boldsymbol{Y}_{t-1}) + \text{vec}(\boldsymbol{E}_t),$$

and the low-dimensional representation in (**??**) *becomes*

$$\boldsymbol{U}_3^\top \boldsymbol{Y}_t \boldsymbol{U}_4 = \langle \boldsymbol{\mathcal{G}}, \boldsymbol{U}_1^\top \boldsymbol{Y}_{t-1} \boldsymbol{U}_2 \rangle + \boldsymbol{U}_3^\top \boldsymbol{E}_t \boldsymbol{U}_4,$$

where $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1\times\cdots\times r_4}$. It is interesting to compare this model with the matrix autoregressive (MAR) model in **?** *and* **?**, *which is defined by*

$$\boldsymbol{Y}_t = \boldsymbol{B}_1 \boldsymbol{Y}_{t-1} \boldsymbol{B}_2^\top + \boldsymbol{E}_t,$$

where $\boldsymbol{B}_1 \in \mathbb{R}^{p_1\times p_1}$ and $\boldsymbol{B}_2 \in \mathbb{R}^{p_2\times p_2}$, whose vector form is

$$\text{vec}(\boldsymbol{Y}_t) = (\boldsymbol{B}_2 \otimes \boldsymbol{B}_1)\text{vec}(\boldsymbol{Y}_{t-1}) + \text{vec}(\boldsymbol{E}_t).$$

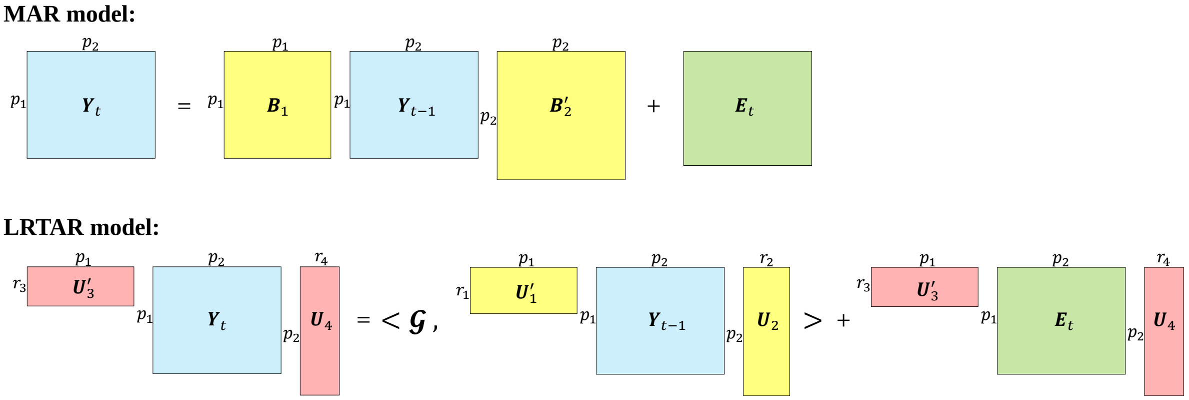

Figure 1: Illustration of the MAR model and the proposed LRTAR model in the case of $d = 2$.

It can be easily seen that if $r_1 = r_3 = p_1$, $r_2 = r_4 = p_2$, $\boldsymbol{U}_3 = \boldsymbol{I}_{p_1}$, $\boldsymbol{U}_4 = \boldsymbol{I}_{p_2}$, *and* $\mathcal{G}_{[\{3,4\}]} = (\boldsymbol{B}_2 \otimes \boldsymbol{B}_1)(\boldsymbol{U}_2 \otimes \boldsymbol{U}_1)$, *then* (**??**) *becomes exactly* (**??**). *Thus, the MAR model in* (**??**) *can be viewed as a special case of the proposed model without reducing dimensions* p_i*'s to* r_i*'s and without transforming* $\boldsymbol{Y}_t$*; see Figure* **??** *for an illustration. The above comparison also applies to the general case with* $d \geq 3$. *The tensor version of the MAR model is considered in* **?** *and is defined as*

$$\mathcal{Y}_t = \mathcal{Y}_{t-1} \times_{i=1}^{d} \boldsymbol{B}_i + \mathcal{E}_t,$$

where $\boldsymbol{B}_i \in \mathbb{R}^{p_i \times p_i}$ *for* $i = 1, \ldots, d$. *We call* (**??**) *the multilinear tensor autoregressive (MTAR) model. Note that its vector form is*

$$\text{vec}(\mathcal{Y}_t) = (\boldsymbol{B}_d \otimes \cdots \otimes \boldsymbol{B}_1)\text{vec}(\mathcal{Y}_{t-1}) + \text{vec}(\mathcal{E}_t).$$

Similarly, (**??**) *is a special case of* (**??**) *with* $r_i = r_{d+i} = p_i$, $\boldsymbol{U}_{d+i} = \boldsymbol{I}_{p_i}$, *for* $i = 1, \ldots, d$, *and* $\mathcal{G}_{[S_2]} = (\otimes_{i \in S_1} \boldsymbol{B}_i)(\otimes_{i \in S_1} \boldsymbol{U}_i)$. *Obviously, the number of unknown parameters in the MTAR model,* $\sum_{i=1}^{d} p_i^2$, *is much larger than that of the proposed model as shown in* (**??**). *Also note that* **?** *focuses on the low-dimensional estimation and its asymptotic theory, while* **?** *considers a Bayesian estimation method.*

Example 2. *In the special case where* $\boldsymbol{U}_{d+i} = \boldsymbol{U}_i$ *and* $r_{d+i} = r_i$ *for* $i = 1, \ldots, d$, *the proposed model may be understood from the perspective of dynamic factor modeling (**??**) for* *tensor-valued time series. Specifically, consider the following model:*

$$\mathcal{Y}_t = \mathcal{F}_t \times_{i=1}^{d} \boldsymbol{U}_i, \quad \mathcal{F}_t = \langle \mathcal{G}, \mathcal{F}_{t-1} \rangle + \mathcal{H}_t,$$

where $\mathcal{Y}_t \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is the observed tensor-valued time series, $\mathcal{F}_t \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ represents $\prod_{i=1}^{d} r_i$ factors, and $\boldsymbol{U}_i \in \mathbb{R}^{p_i \times r_i}$ are orthonormal matrices for $i = 1, \ldots, d$. Here $\mathcal{F}_t$ follows the tensor autoregression (TAR) with transition tensor $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d \times r_1 \times \cdots \times r_d}$ and random error $\mathcal{H}_t$. Note that (??) can be rewritten as

$$\mathcal{Y}_t = \left\langle \mathcal{G} \times_{i=1}^{d} \boldsymbol{U}_i \times_{i=d+1}^{2d} \boldsymbol{U}_i, \mathcal{Y}_{t-1} \right\rangle + \mathcal{H}_t \times_{i=1}^{d} \boldsymbol{U}_i.$$

Thus, model (??) is a special case of the proposed model with $\boldsymbol{U}_{d+i} = \boldsymbol{U}_i$ and $r_{d+i} = r_i$ for $i = 1, \ldots, d$, and $\mathcal{E}_t = \mathcal{H}_t \times_{i=1}^{d} \boldsymbol{U}_i$. ? introduces the tensor factor model in the form of $\mathcal{Y}_t = \mathcal{F}_t \times_{i=1}^{d} \boldsymbol{U}_i + \mathcal{E}_t$ without an explicit modeling of the latent factors $\mathcal{F}_t$. Hence, model (??) may be regarded as a special tensor factor model with autoregressive dynamic factors, but without any random error in the model equation of $\mathcal{Y}_t$.

4 Low-Dimensional Least Squares Estimation

We consider the parameter estimation for the low-dimensional case where the sample size T is sufficiently large such that the dimension of the parameter space can be assumed fixed. Throughout this section, we assume that the data are generated from the proposed model in Section 3, where the true transition tensor $\boldsymbol{\mathcal{A}}$ is exactly low-Tucker rank with multilinear ranks $(r_1, \ldots, r_{2d})$. This assumption will be relaxed under the high-dimensional setup in the next section.

Suppose that the true ranks $(r_1, \ldots, r_{2d})$ of the exactly low-rank tensor $\boldsymbol{\mathcal{A}}$ are known; we relegate the rank selection to Section ??. Then the parameters can be estimated by

$$\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}} = \widehat{\mathcal{G}} \times_{i=1}^{2d} \widehat{\boldsymbol{U}}_i = \underset{\mathcal{G}, \boldsymbol{U}_i}{\arg\min} \sum_{t=1}^{T} \left\| \mathcal{Y}_t - \langle \mathcal{G} \times_{i=1}^{2d} \boldsymbol{U}_i, \mathcal{Y}_{t-1} \rangle \right\|_{\text{F}}^2.$$

We call $\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}}$ the low-Tucker-rank (LTR) estimator. Note that the minimization in (??) is unconstrained, so the Tucker decomposition of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}}$ is not unique. Indeed, there are more than one solution of $\widehat{\mathcal{G}}$ and $\widehat{\boldsymbol{U}}_i$'s corresponding to the same $\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}}$. Due to the lack of identifiability of the Tucker decompositions, standard asymptotics of the maximum likelihood estimation cannot apply directly. Nevertheless, we can still derive the asymptotic distribution of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}}$ using the asymptotic theory for overparameterized models in **?**.

Recall that $S_1 = \{1, 2, \ldots, d\}$ $S_2 = \{d+1, d+2, \ldots, 2d\}$, $\boldsymbol{y}_t = \text{vec}(\boldsymbol{\mathcal{Y}}_t)$, and $\boldsymbol{e}_t = \text{vec}(\boldsymbol{\mathcal{E}}_t)$. Let $\boldsymbol{\theta} = (\text{vec}(\mathcal{G}_{[S_2]})^\top, \text{vec}(\boldsymbol{U}_1)^\top, \cdots, \text{vec}(\boldsymbol{U}_{2d})^\top)^\top$ be the parameter vector, and let $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{\theta}) = \text{vec}(\boldsymbol{\mathcal{A}}_{[S_2]}) = \text{vec}((\otimes_{i\in S_2}\boldsymbol{U}_i)\mathcal{G}_{[S_2]}(\otimes_{i\in S_1}\boldsymbol{U}_i)^\top)$ be the vectorization of the transition matrix. Denote $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \text{var}(\boldsymbol{y}_t)$, $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \text{var}(\boldsymbol{e}_t)$, and $\boldsymbol{J} = \boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}}$. In addition, for $i = 1, \ldots, 2d$, let $\boldsymbol{P}_{[S_2]}^{(i)}$ be the $p^2 \times p^2$ permutation matrix such that $\text{vec}(\boldsymbol{\mathcal{A}}_{[S_2]}) = \boldsymbol{P}_{[S_2]}^{(i)}\text{vec}(\boldsymbol{\mathcal{A}}_{(i)})$. Then, it can be shown that the Jacobian matrix $\boldsymbol{H} := \partial\boldsymbol{h}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is given by

$$\begin{aligned}\boldsymbol{H} = \big(&(\otimes_{i\in S_1}\boldsymbol{U}_i) \otimes (\otimes_{i\in S_2}\boldsymbol{U}_i), \boldsymbol{P}_{[S_2]}^{(1)}\left\{\left[(\otimes_{i=1,i\neq 1}^{2d}\boldsymbol{U}_i)\mathcal{G}_{(1)}^\top\right] \otimes \boldsymbol{I}_{p_1}\right\}, \\ &\boldsymbol{P}_{[S_2]}^{(2)}\left\{\left[(\otimes_{i=1,i\neq 2}^{2d}\boldsymbol{U}_i)\mathcal{G}_{(2)}^\top\right] \otimes \boldsymbol{I}_{p_2}\right\}, \ldots, \boldsymbol{P}_{[S_2]}^{(2d)}\left\{\left[(\otimes_{i=1,i\neq 2d}^{2d}\boldsymbol{U}_i)\mathcal{G}_{(2d)}^\top\right] \otimes \boldsymbol{I}_{p_{2d}}\right\}\big),\end{aligned}$$

where $p_{d+i} = p_i$ for $i = 1, \ldots, d$.

Theorem 1. *Suppose that the time series $\{\boldsymbol{\mathcal{Y}}_t\}$ is generated by model* (**??**) *with $\mathbb{E}\|\boldsymbol{e}_t\|_2^4 < \infty$, and Assumption* **??** *holds. Then,*

$$\sqrt{T}\text{vec}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}} - \boldsymbol{\mathcal{A}})_{[S_2]}) \to N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{LTR}}),$$

in distribution as $T \to \infty$, where $\boldsymbol{\Sigma}_{\text{LTR}} = \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{J}\boldsymbol{H})^\dagger\boldsymbol{H}^\top$, and † is the Moore-Penrose inverse.

Since the asymptotic theory for overparameterized models in **?** allows for unidentifiability of the components $\mathcal{G}$ and $\boldsymbol{U}_i$'s in the decomposition of $\boldsymbol{\mathcal{A}}$, Theorem **??** does not require that the Tucker decomposition of $\boldsymbol{\mathcal{A}}$ is unique; see the proof of Theorem **??** in Appendix A for more details.

Next we compare the result in Theorem **??** to those of two other consistent estimators for the proposed model in the low-dimensional setup. Note that the rank of $\boldsymbol{\mathcal{A}}_{[S_2]}$ in (**??**) is at most $s_1 := \min(\prod_{i=1}^d r_i, \prod_{i=d+1}^{2d} r_i)$. Thus, $\boldsymbol{\mathcal{A}}_{[S_2]}$ can be estimated by both the reduced-rank regression (RRR) and ordinary least squares (OLS) methods,

$$\widehat{\boldsymbol{\mathcal{A}}}_{\text{RRR}} = \underset{\text{rank}(\boldsymbol{\mathcal{B}}_{[S_2]})\leq s_1}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{\mathcal{Y}}_t - \langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2$$

and

$$\widehat{\boldsymbol{\mathcal{A}}}_{\text{OLS}} = \underset{\boldsymbol{\mathcal{B}}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \|\boldsymbol{\mathcal{Y}}_t - \langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2.$$

Naturally, under model (**??**), $\widehat{\boldsymbol{\mathcal{A}}}_{\text{LTR}}$ is asymptotically more efficient than $\widehat{\boldsymbol{\mathcal{A}}}_{\text{RRR}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{OLS}}$:

Corollary 1. *If the conditions of Theorem* **??** *hold, then* $\sqrt{T}\text{vec}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{OLS}}-\boldsymbol{\mathcal{A}})_{[S_2]}) \to N(0, \boldsymbol{\Sigma}_{\text{OLS}})$ *and* $\sqrt{T}\text{vec}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{RRR}}-\boldsymbol{\mathcal{A}})_{[S_2]}) \to N(0, \boldsymbol{\Sigma}_{\text{RRR}})$ *in distribution as* $T \to \infty$*. Moreover, it holds that* $\boldsymbol{\Sigma}_{\text{LTR}} \leq \boldsymbol{\Sigma}_{\text{RRR}} \leq \boldsymbol{\Sigma}_{\text{OLS}}$.

To solve the minimization problem in (**??**), we propose an alternating least squares (ALS) method. Specifically, by the vector representation of the proposed model in (**??**), the loss function in (**??**) can be rewritten as

$$L(\boldsymbol{\mathcal{G}}, \boldsymbol{U}_1, \ldots, \boldsymbol{U}_{2d}) = \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - (\otimes_{i\in S_2}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{[S_2]}(\otimes_{i\in S_1}\boldsymbol{U}_i)^\top \boldsymbol{y}_{t-1}\right\|_2^2 = \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \boldsymbol{m}(\boldsymbol{\theta})\right\|_2^2,$$

where $\boldsymbol{m}(\boldsymbol{\theta}) = (\otimes_{i\in S_2}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{[S_2]}(\otimes_{i\in S_1}\boldsymbol{U}_i)^\top \boldsymbol{y}_{t-1}$ is the conditional mean of $\boldsymbol{y}_t = \text{vec}(\boldsymbol{\mathcal{Y}}_t)$ and is a function of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_{2d}^\top)^\top$, with $\boldsymbol{\theta}_0 = \text{vec}(\boldsymbol{\mathcal{G}}_{[S_2]})$ and $\boldsymbol{\theta}_i = \text{vec}(\boldsymbol{U}_i)$ for $i = 1, \ldots, 2d$. Note that $\boldsymbol{m}(\boldsymbol{\theta})$ is linear in each $\boldsymbol{\theta}_i$ while keeping the other components $\boldsymbol{\theta}_j$ with $0 \leq j \neq i \leq 2d$ fixed. This is indeed because

$$\begin{aligned}
\boldsymbol{m}(\boldsymbol{\theta}) &= (\otimes_{i\in S_2}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{[S_2]}\boldsymbol{P}_{k,1}\left\{\left[(\otimes_{i\in S_1, i\neq k}\boldsymbol{U}_i)^\top(\boldsymbol{\mathcal{Y}}_{t-1})_{(k)}^\top\right]\otimes \boldsymbol{I}_{r_k}\right\}\boldsymbol{P}_{k,3}\text{vec}(\boldsymbol{U}_k)\\
&= \boldsymbol{P}_{k,2}\left\{\left[(\otimes_{i\in S_2, i\neq d+k}\boldsymbol{U}_i)(\boldsymbol{\mathcal{M}}_{t-1})_{(k)}^\top\right]\otimes \boldsymbol{I}_{p_k}\right\}\text{vec}(\boldsymbol{U}_{d+k})\\
&= \left\{\left[\boldsymbol{y}_{t-1}^\top(\otimes_{i\in S_1}\boldsymbol{U}_i)\right]\otimes(\otimes_{i\in S_2}\boldsymbol{U}_i)\right\}\text{vec}(\boldsymbol{\mathcal{G}}_{[S_2]}),
\end{aligned}$$

for $k = 1, \ldots, d$, where $\boldsymbol{P}_{k,1} \in \mathbb{R}^{\prod_{i=1}^d r_i \times \prod_{i=1}^d r_i}$, $\boldsymbol{P}_{k,2} \in \mathbb{R}^{p\times p}$ and $\boldsymbol{P}_{k,3} \in \mathbb{R}^{r_k p_k \times r_k p_k}$ are permutation matrices defined such that $\boldsymbol{P}_{k,1}\text{vec}(\boldsymbol{\mathcal{T}}_{(k)}) = \text{vec}(\boldsymbol{\mathcal{T}})$ for any $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{r_1\times\cdots\times r_d}$, $\boldsymbol{P}_{k,2}\text{vec}(\boldsymbol{\mathcal{T}}_{(k)}) = \text{vec}(\boldsymbol{\mathcal{T}})$ for any $\boldsymbol{\mathcal{T}} \in \mathbb{R}^{p_1\times\cdots\times p_d}$, and $\boldsymbol{P}_{k,3}\text{vec}(\boldsymbol{U}_k) = \text{vec}(\boldsymbol{U}_k^\top)$, and $\boldsymbol{\mathcal{M}}_{t-1} \in \mathbb{R}^{r_{d+1}\times\cdots\times r_{2d}}$ is defined such that $\text{vec}(\boldsymbol{\mathcal{M}}_t) = \boldsymbol{\mathcal{G}}_{[S_2]}(\otimes_{i\in S_1}\boldsymbol{U}_i)^\top \boldsymbol{y}_{t-1}$. Therefore, we can update

Algorithm 1 ALS algorithm for LTR estimator

```
Initialize: $\boldsymbol{\mathcal{A}}^{(0)} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{RRR}}$ or $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{OLS}}$
HOSVD: $\boldsymbol{\mathcal{A}}^{(0)} \approx \boldsymbol{\mathcal{G}}^{(0)} \times_{i=1}^{2d} \boldsymbol{U}_i^{(0)}$,  with multilinear ranks $(r_1, \dots, r_{2d})$.
repeat $s = 0, 1, 2, \dots$
    for $k = 1, \dots, d$
        $\boldsymbol{U}_k^{(s+1)} \leftarrow \underset{\boldsymbol{U}_k}{\arg\min} \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - (\otimes_{i \in S_2} \boldsymbol{U}_i^{(s)}) \boldsymbol{\mathcal{G}}_{[S_2]} \boldsymbol{P}_{k,1} \left\{ \left[ (\otimes_{i \in S_1, i \neq k} \boldsymbol{U}_i^{(s)})^\top (\boldsymbol{\mathcal{Y}}_{t-1})_{(k)}^\top \right] \otimes \boldsymbol{I}_{r_k} \right\} \mathrm{vec}(\boldsymbol{U}_k) \right\|_2^2$
    end for
    for $k = 1, \dots, d$
        $\boldsymbol{U}_{d+k}^{(s+1)} \leftarrow \underset{\boldsymbol{U}_{d+k}}{\arg\min} \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \boldsymbol{P}_{k,2} \{ \left[ (\otimes_{i \in S_2, i \neq d+k} \boldsymbol{U}_i^{(s)}) (\boldsymbol{\mathcal{M}}_{t-1}^{(s+1)})_{(k)}^\top \right] \otimes \boldsymbol{I}_{p_k} \} \mathrm{vec}(\boldsymbol{U}_{d+k}) \right\|_2^2$
    end for
    $\boldsymbol{\mathcal{G}}^{(s+1)} \leftarrow \underset{\boldsymbol{\mathcal{G}}}{\arg\min} \sum_{t=1}^{T} \left\| \boldsymbol{y}_t - \left\{ \left[ \boldsymbol{y}_{t-1}^\top (\otimes_{i \in S_1} \boldsymbol{U}_i^{(s+1)}) \right] \otimes (\otimes_{i \in S_2} \boldsymbol{U}_i^{(s+1)}) \right\} \mathrm{vec}(\boldsymbol{\mathcal{G}}_{[S_2]}) \right\|_2^2$
    $\boldsymbol{\mathcal{A}}^{(s+1)} \leftarrow \boldsymbol{\mathcal{G}}^{(s+1)} \times_{i=1}^{2d} \boldsymbol{U}_i^{(s+1)}$
until convergence
```

each component parameter vector $\boldsymbol{\theta}_i$, and hence the corresponding $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{U}_i$'s, iteratively by the least squares method. The resulting ALS algorithm is shown in Algorithm **??**.

As mentioned, the minimization in (**??**) is unconstrained. Accordingly, to obtain $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{LTR}}$, no orthogonality constraint of $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{U}_i$'s is needed in Algorithm **??**. Instead, we compute the final unique estimates of $\widehat{\boldsymbol{\mathcal{G}}}$ and $\widehat{\boldsymbol{U}}_i$'s by the HOSVD operation based on the unconstrained solution $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{LTR}}$ obtained from Algorithm **??**. Specifically, we compute each $\widehat{\boldsymbol{U}}_i$ uniquely as the top r_i left singular vectors of $(\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{LTR}})_{(i)}$ such that the first element in each column of $\widehat{\boldsymbol{U}}_i$ is positive, and set $\widehat{\boldsymbol{\mathcal{G}}} = [\![\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{LTR}}; \widehat{\boldsymbol{U}}_1^\top, \cdots, \widehat{\boldsymbol{U}}_{2d}^\top]\!]$. Similar alternating algorithms without imposing identification constraints can be found in the literature of tensor decomposition; see, e.g. **?** and **?**.

5 High-Dimensional Regularized Estimation

5.1 Regularization via One-Mode Matricization

The methods in the previous section rely on the assumptions that the dimension is fixed and that the true transition tensor is exactly low-rank with known Tucker ranks. We next relax both assumptions. Specifically, under the high-dimensional setup, we consider regularized estimation of model (??) via different nuclear-norm-type penalties and develop the corresponding non-asymptotic theory under only an approximately low-Tucker-rank assumption on the underlying true transition tensor.

In model (??), the exactly low-rank transition tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times\cdots\times p_{2d}}$ is subject to the constraints $r_i = \text{rank}(\boldsymbol{\mathcal{A}}_{(i)})$, for $i = 1, \ldots, 2d$. A commonly used convex relaxation of such multilinear rank constraints is the regularization via the sum of nuclear (SN) norms of all the one-mode matricizations,

$$\|\boldsymbol{\mathcal{A}}\|_{\text{SN}} = \sum_{i=1}^{2d} \|\boldsymbol{\mathcal{A}}_{(i)}\|_*.$$

The SN norm has been widely used in the literature (????) to simultaneously encourage the low-rankness for all modes of a tensor. This leads us to the SN norm regularized estimator

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

where λ_{SN} is the tuning parameter. Note that if instead of $\|\boldsymbol{\mathcal{A}}\|_{\text{SN}}$, only one single nuclear norm, say $\|\boldsymbol{\mathcal{A}}_{(1)}\|_*$, is penalized, then the resulting estimator will only enforce the low-rankness for the first mode of $\boldsymbol{\mathcal{A}}$, while failing to do so for all the other $2d-1$ modes, and hence will be less efficient than the above SN estimator.

To derive the estimation error bound for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}$, we make the following assumption on the random error $\boldsymbol{e}_t = \text{vec}(\boldsymbol{\mathcal{E}}_t)$.

Assumption 2. *Let $\boldsymbol{e}_t = \boldsymbol{\Sigma}_{\boldsymbol{e}}^{1/2}\boldsymbol{\xi}_t$, where $\{\boldsymbol{\xi}_t\}$ is a sequence of i.i.d. random vectors, with $\mathbb{E}(\boldsymbol{\xi}_t) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{\xi}_t) = \boldsymbol{I}_p$, and $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \text{var}(\boldsymbol{e}_t)$ is a positive definite matrix. In addition, the entries $(\boldsymbol{\xi}_{it})_{1\leq i\leq p}$ of $\boldsymbol{\xi}_t$ are mutually independent and κ^2-sub-Gaussian, i.e., $\mathbb{E}(e^{\mu\xi_{it}}) \leq e^{\kappa^2\mu^2/2}$, for any $\mu \in \mathbb{R}$ and $i = 1, \ldots, p$.*

The sub-Gaussianity condition in Assumption **??** is milder than the commonly used normality assumption in the literature on high-dimensional stationary vector autoregressive models (**??**). This relaxation is made possible through establishing a novel martingale-based concentration bound in the proof of the deviation bound; see Lemma **??** in Appendix **??**. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ captures the contemporaneous dependency in $\boldsymbol{\mathcal{E}}_t$, and the constant κ controls the tail heaviness of the marginal distributions.

For any $z \in \mathbb{C}$, let $\mathcal{A}(z) = \boldsymbol{I}_p - \boldsymbol{\mathcal{A}}_{[S_2]} z$ be a matrix polynomial, where $\mathbb{C}$ is the set of complex numbers. Let $\mu_{\min}(\mathcal{A}) = \min_{|z|=1} \lambda_{\min}(\mathcal{A}^*(z)\mathcal{A}(z))$ and $\mu_{\max}(\mathcal{A}) = \max_{|z|=1} \lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))$, where $\mathcal{A}^*(z)$ is the conjugate transpose of $\mathcal{A}(z)$; see **?** for more discussions on the connection between the spectral density of the VAR process and the two quantities. In addition, let

$$\alpha_{\text{RSC}} = \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})}{\mu_{\max}(\mathcal{A})}, \quad M_1 = \frac{\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})}{\mu_{\min}^{1/2}(\mathcal{A})}, \quad \text{and} \quad M_2 = \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\mu_{\max}(\mathcal{A})}{\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\mu_{\min}(\mathcal{A})}.$$

The exactly low-rankness of the true transition tensor $\boldsymbol{\mathcal{A}}$ assumed in Section **??** could be too stringent in real-world applications. In what follows, we relax it to the following approximately low-rank assumption: We assume that all one-mode matricizations of the underlying true transition tensor $\boldsymbol{\mathcal{A}}$ belong to the set of approximately low-rank matrices, namely $\boldsymbol{\mathcal{A}}_{(i)} \in \mathbb{B}_q(r_q^{(i)}; p_i, p_{-i}p)$ for some $q \in [0, 1)$, where

$$\mathbb{B}_q(r; d_1, d_2) := \left\{ \boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2} : \sum_{i=1}^{\min(d_1,d_2)} \sigma_i(\boldsymbol{M})^q \le r \right\},$$

and $p_{-i} = p/p_i = \prod_{j=1, j\neq i}^{d} p_j$. For the convenience of notation, here we require that $0^0 = 0$. Note that when $q = 0$, $\mathbb{B}_0(r; d_1, d_2)$ is the set of d_1-by-d_2 rank-r matrices. For $q > 0$, the restriction on $\sum_{i=1}^{\min(d_1,d_2)} \sigma_i(\boldsymbol{M})^q \le r$ requires that the singular values decay fast, and it is more general and natural than the exactly low-rank assumption.

Theorem 2. *Under Assumptions* **??** *and* **??**, *if* $T \gtrsim \max_{1\le i\le d} p_{-i}p + \max(\kappa^2, \kappa^4) M_2^{-2} p$, $\lambda_{\text{SN}} \gtrsim \kappa^2 M_1 d^{-2} \sum_{i=1}^{d} \sqrt{p_{-i}p/T}$, *and* $\boldsymbol{\mathcal{A}}_{(i)} \in \mathbb{B}_q(r_q^{(i)}; p_i, p_{-i}p)$ *for some* $q \in [0, 1)$ *and all* $i = 1, \ldots, 2d$, *then with probability at least* $1 - 2\sum_{i=1}^{d} \exp(-Cp_{-i}p) - \exp[-C\min(\kappa^{-2}, \kappa^{-4}) M_2^2 p]$,

$$\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} \lesssim \sqrt{r_q} \left(\frac{2d \cdot \lambda_{\text{SN}}}{\alpha_{\text{RSC}}} \right)^{1-q/2}$$

where $r_q = (2d)^{-1}\sum_{i=1}^{2d} r_q^{(i)}$.

By Theorem **??**, when $\lambda_{\text{SN}} \asymp \kappa^2 M_1 d^{-2}\sum_{i=1}^{d}\sqrt{p_{-i}p/T}$, the estimation error bound scales as $\sqrt{r_q}(\max_{1\leq i\leq d} p_{-i}p/T)^{1/2-q/4}$; note that the factor d in the error bounds is canceled by the d^{-2} in the rate of λ_{SN}. When $q=0$, namely $\boldsymbol{\mathcal{A}}$ is exactly low-rank with Tucker ranks $(r_0^{(1)},\dots,r_0^{(2d)})$, the error bound reduces to $\sqrt{r_0\max_{1\leq i\leq d} p_{-i}p/T}$ and it is comparable to that in **?** for *i.i.d.* tensor regression.

However, recent research (e.g., **??**) has shown that the SN norm regularization approach is suboptimal, mainly because $\boldsymbol{\mathcal{A}}_{(i)}$ is an unbalanced *fat-and-short* matricization of a higher-order tensor. Technically, in the proof of Theorem **??**, an essential intermediate step is to establish the deviation bound, where we need to upper bound the operator norm of a sub-Gaussian random matrix with the same dimensions as $\boldsymbol{\mathcal{A}}_{(i)}$; see Lemma **??** in Appendix **??** for details. Undesirably, the order of this operator norm will be dominated by the larger of the row and column dimensions of the matrix $\boldsymbol{\mathcal{A}}_{(i)}\in\mathbb{R}^{p_i\times p_{-i}p}$, and hence by the column dimension $p_{-i}p$, which eventually appears in the error bound. This indicates that the imbalance of the matricization leads to the efficiency bottleneck of the SN estimator.

On the other hand, similarly to (**??**), since the the reduced-rank VAR model can be regarded as an overparameterization of the proposed LRTAR model, alternatively one may focus on the approximately low-rank structure of the transition matrix $\boldsymbol{\mathcal{A}}_{[S_2]}$ in the VAR representation in (**??**), and adopt the matrix nuclear (MN) estimator (**?**) to estimate $\boldsymbol{\mathcal{A}}$,

$$\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}} = \underset{\boldsymbol{\mathcal{A}}}{\arg\min}\left\{\frac{1}{T}\sum_{t=1}^{T}\|\boldsymbol{\mathcal{Y}}_t - \langle\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2 + \lambda_{\text{MN}}\|\boldsymbol{\mathcal{A}}_{[S_1]}\|_*\right\}.$$

Note that the multi-mode matricization $\boldsymbol{\mathcal{A}}_{[S_2]} = \boldsymbol{\mathcal{A}}_{[S_1]}^\top$ is a $p\times p$ square matrix. Thus, the loss of efficiency due to the unbalanced matricization can be avoided, which is confirmed by the following theorem.

Theorem 3. *Under Assumptions* **??** *and* **??**, *if* $T \gtrsim [1+\max(\kappa^2,\kappa^4)M_2^{-2}]p$, $\lambda_{\text{MN}} \gtrsim \kappa^2 M_1\sqrt{p/T}$, *and* $\boldsymbol{\mathcal{A}}_{[S_1]}\in\mathbb{B}_q(s_q^{(1)};p,p)$ *for some* $q=[0,1)$, *then with probability at least* $1-\exp(-Cp)-$

$\exp[-C\min(\kappa^{-2},\kappa^{-4})M_2^2 p]$,

$$\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}-\boldsymbol{\mathcal{A}}\|_{\text{F}}\lesssim\sqrt{s_q^{(1)}}\left(\frac{\lambda_{\text{MN}}}{\alpha_{\text{RSC}}}\right)^{1-q/2}.$$

Theorem **??** shows that, with $\lambda_{\text{MN}}\asymp\kappa^2 M_1\sqrt{p/T}$, the estimation error bound for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ scales as $\sqrt{s_q^{(1)}}(p/T)^{1/2-q/4}$. This result is comparable to that in **?** for reduced-rank VAR models, yet we relax both the constraint $\|\boldsymbol{A}\|_{\text{op}}<1$ on the transition matrix $\boldsymbol{A}$ and the normality assumption on the random error in their paper. This estimation error bound is clearly smaller than that in Theorem **??**, as $(\max_{1\leq i\leq d}p_{-i}p/T)^{1/2-q/4}$ in general can be much larger than $(p/T)^{1/2-q/4}$ when $s_q^{(1)}\asymp r_q$. Therefore, adopting square matricization can indeed improve the estimation performance.

The idea of using square matricization to improve efficiency was adopted by **?** in low-rank tensor completion problems. Their proposed method, called the square deal, is to first unfold a general higher-order tensor into a matrix with similar numbers of rows and columns, and then use the MN norm as the regularizer. However, for our estimation problem, despite the advantage of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ over $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}$, Theorem **??** reveals another drawback of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$. That is, the error bounds for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ depend on the ℓ_q radius $s_q^{(1)}$ of the singular values of $\boldsymbol{\mathcal{A}}_{[S_1]}$, suggesting that $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ may perform badly when $s_q^{(1)}$ is relatively large. In other words, unless we have prior knowledge that the ℓ_q "norm" of singular values of particular matricization $\boldsymbol{\mathcal{A}}_{[S_1]}$ is truly small, $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ may not be desirable in practice.

On the other hand, although the SN regularizer in (**??**) suffers from inefficiency due to the imbalance of one-mode matricizations, it has the attractive feature of simultaneously enforcing the low-rankness across all modes of the tensor $\boldsymbol{\mathcal{A}}$, and thus is more efficient than its counterpart which considers only one single one-mode matricization, say, $\|\boldsymbol{\mathcal{A}}_{(1)}\|_*$. Similarly, if we can enforce the approximate low-rankness across all possible square matricizations of $\boldsymbol{\mathcal{A}}$, the estimation performance may be further improved upon $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$. This motivates us to propose a new regularization approach in the next subsection.

5.2 Regularization via Square Matricization

In this subsection, we propose a novel convex regularizer which improves upon both the SN and MN regularizers in (??) and (??), respectively, by simultaneously encouraging the low-rankness across all possible square matricizations of the transition tensor $\boldsymbol{\mathcal{A}}$.

For any $2d$-th-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1\times\cdots\times p_d\times p_1\times\cdots\times p_d}$, its multi-mode matricization $\boldsymbol{\mathcal{A}}_{[I]}$ will be a $p \times p$ square matrix, with $p = \prod_{i=1}^{d} p_i$, if the index set is chosen as

$$I = \{\ell_1, \ldots, \ell_d\},$$

where each index ℓ_i is set to either i or $d+i$, for $i = 1, \ldots, d$. For instance, $\boldsymbol{\mathcal{A}}_{[S_1]}$ is the square matricization formed by setting $\ell_i = i$ for all $i = 1, \ldots, d$. Moreover, if $\boldsymbol{\mathcal{A}}$ has multilinear ranks $(r_1, \ldots, r_{2d})$, then the rank of the matricization $\boldsymbol{\mathcal{A}}_{[I]}$ is at most $\min(\prod_{i=1,i\in I}^{2d} r_i, \prod_{i=1,i\notin I}^{2d} r_i)$. Therefore, if we penalize the sum of nuclear norms of all such squares matricizations, which we call the sum of square-matrix nuclear (SSN) norms for simplicity, then the resulting estimator would enjoy the efficiency gain from both the use of square matricizations and simultaneous incorporation of many rank constraints.

Obviously, there are 2^d possible choices of the index set I that corresponds to a square matricization $\boldsymbol{\mathcal{A}}_{[I]}$. However, since $\boldsymbol{\mathcal{A}}_{[I]} = \boldsymbol{\mathcal{A}}_{[I^\complement]}^\top$, when defining the SSN norm, we only need to include one of I and its complement $I^\complement$. A simple way to do so is to choose only sets containing the index one. That is, fix $\ell_1 = 1$ and choose $\ell_i = i$ or $d+i$ for $i = 2, \ldots, d$. This results in totally 2^{d-1} chosen index sets, denoted by $I_1, I_2, \ldots, I_{2^{d-1}}$. Note that $I_1 = S_1 = \{1, \ldots, d\}$. For example, when $d = 3$, we have four chosen index sets, $I_1 = \{1,2,3\}, I_2 = \{1,5,3\}, I_3 = \{1,2,6\}$ and $I_4 = \{1,5,6\}$.

Based on the above choice of the 2^{d-1} index sets, we introduce the following SSN norm,

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

The corresponding estimator is defined as

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

	SN	MN	SSN
Sample size	$T \gtrsim (\max_{1\leq i\leq d} p_{-i} + M_2^{-2})p$	$T \gtrsim (1 + M_2^{-2})p$	$T \gtrsim (1 + M_2^{-2})p$
Estimation error	$\sqrt{r_q}(\max_{1\leq i\leq d} p_{-i}p/T)^{1/2-q/4}$	$\sqrt{s_q^{(1)}}(p/T)^{1/2-q/4}$	$\sqrt{s_q}(p/T)^{1/2-q/4}$

Table 1: Summary of the sample size conditions and error upper bounds in Theorems **??**–**??**, where $p_{-i} = \prod_{j=1,j\neq i}^{d} p_j$, $r_q = (2d)^{-1}\sum_{i=1}^{2d} r_q^{(i)}$, and $s_q = 2^{1-d}\sum_{k=1}^{2^{d-1}} s_q^{(k)}$.

where λ_{SSN} is the tuning parameter.

If the rank of one-mode matricizations $\text{rank}(\boldsymbol{\mathcal{A}}_{(i)}) = r_i$, each square matricization $\boldsymbol{\mathcal{A}}_{[I_k]}$ is also low-rank with $\text{rank}(\boldsymbol{\mathcal{A}}_{[I_k]}) \leq \min(\prod_{i=1,i\in I_k}^{2d} r_i, \prod_{i=1,i\notin I_k}^{2d} r_i)$. Similarly, if all $\boldsymbol{\mathcal{A}}_{(i)}$s are approximately low-rank, the square matricizations are approximately low-rank as well. In contrast to the SN norm in (**??**) which directly matches the multilinear ranks $\text{rank}(\boldsymbol{\mathcal{A}}_{(i)})$ for $i = 1, \ldots, d$, the SSN norm encourages the multilinear low-rank structure of $\boldsymbol{\mathcal{A}}$ by simultaneously enforcing the low-rankness of all the square matricizations $\boldsymbol{\mathcal{A}}_{[I_k]}$. The following theorem gives the theoretical results for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$.

Theorem 4. *Under Assumptions* **??** *and* **??**, *if* $T \gtrsim [1 + \max(\kappa^2, \kappa^4)M_2^{-2}]p$, $\lambda_{\text{SSN}} \gtrsim \kappa^2 M_1 2^{1-d}\sqrt{p/T}$, *and* $\boldsymbol{\mathcal{A}}_{[I_k]} \in \mathbb{B}(s_q^{(k)}; p, p)$ *for some* $q \in [0, 1)$ *and all* $k = 1, \ldots, 2^{d-1}$, *with probability at least* $1 - \exp[-C(p-d)] - \exp[-C\min(\kappa^{-2}, \kappa^{-4})M_2^2 p]$,

$$\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} \lesssim \sqrt{s_q}\left(\frac{2^{d-1}\lambda_{\text{SSN}}}{\alpha_{\text{RSC}}}\right)^{1-q/2}$$

where $s_q = 2^{1-d}\sum_{k=1}^{2^{d-1}} s_q^{(k)}$.

By Theorem **??**, when $\lambda_{\text{SSN}} \asymp \kappa^2 M_1 2^{1-d}\sqrt{p/T}$, the estimation error bound scales as $\sqrt{s_q}(p/T)^{1/2-q/4}$ and reduces to $\sqrt{s_0 p/T}$ in the exactly low-rank setting for $q = 0$. For a clearer comparison among the three estimators $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}, \widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$, we summarize the main results of Theorems **??**–**??** in Table **??**. First, both $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ have much smaller error bounds and less stringent sample size requirements than $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}$, due to the diverging dimension p_{-i} in the results of the latter. This reaffirms the advantage of the square matricizations.

Secondly, comparing $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ to $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$, since the factor s_q in the error bounds of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ is the average of all $s_q^{(k)}$ for $k = 1, \ldots, 2^{d-1}$, $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ can protect us from the bad scenarios where the ℓ_q "norm" of the singular values of $\boldsymbol{\mathcal{A}}_{[S_1]}$ is relatively large. If all the $s_q^{(k)}$'s are of the same order, then the error upper bounds for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ in Table **??** will be similar. However, our simulation results in Section **??** show that $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ clearly outperforms $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ under various settings, even when $s_q^{(1)} = \cdots = s_q^{(2d)}$. Indeed, the error bounds for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ in Theorem **??** is likely to be loose, which is believed to be caused by taking the upper bounds on the dual norm of the SSN norm in the proof of Lemma **??**; see Appendix **??** for details. By contrast, the error bounds for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ are minimax-optimal (**?**). Therefore, although our theoretical results are not sharp enough to distinguish clearly between the error rates of $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$, we conjecture that the actual rate of the former is generally smaller than that of the latter. Methodologically, this is also easy to understand because, unlike $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$, $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ simultaneously enforces the low-rankness across all square matricizations of $\boldsymbol{\mathcal{A}}$ rather than just on $\boldsymbol{\mathcal{A}}_{[S_1]}$.

Remark 1. *While our SSN regularization is proposed in the time series context, the idea of imposing joint penalties on all (close to) square matricizations of the coefficient tensor can potentially be extended to general higher-order tensor estimation problems. Moreover, it can be refined to accommodate particular structures of the data. For example, if some of the d modes of the tensor-value time series $\boldsymbol{\mathcal{Y}}_t$, namely $p_1, \ldots, p_d$, are equal, then even a greater number of possible square matricizations of $\boldsymbol{\mathcal{A}}$ can be formed, resulting in improved estimation efficiency.*

5.3 Truncated Regularized Estimation

While the proposed regularized estimation methods do not require exact low-rankness of the true transition tensor $\boldsymbol{\mathcal{A}}$, sometimes imposing exact low-rankness may be more desirable if one wants to interpret the underlying low-dimensional tensor dynamics. As discussed in Section **??**, the Tucker ranks can be interpreted as the numbers of dynamic factors in each mode. In this section, we propose a truncation method to consistently estimate the true multilinear ranks $(r_1, \ldots, r_{2d})$ under the exact low-rank assumption.

Let γ be a threshold parameter to be chosen properly. Given the estimator $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$, for each $i = 1, \dots, 2d$, we calculate the singular value decomposition (SVD) of the mode-i matricization $(\widehat{\mathcal{A}}_{\text{SSN}})_{(i)}$ with the singular values arranged in descending order. Next we truncate the SVD by retaining only singular values greater than γ, and take their corresponding left singular vectors to define the matrix $\widetilde{\boldsymbol{U}}_i$. Then, the truncated core tensor is defined as

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

based on which we propose the truncated sum of square-matrix nuclear (TSSN) estimator

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

To derive the theoretical results on rank selection, we make the following assumption on the exact Tucker ranks and the magnitude of the singular values.

Assumption 3. *For all $i = 1, \dots, 2d$, $\sigma_r(\mathcal{A}_{(i)}) = 0$ for all $r > r_i$, and there exists a constant $C > 1$ such that $\min_{1 \le i \le 2d} \sigma_{r_i}\left(\mathcal{A}_{(i)}\right) \ge C\gamma$. As $T \to \infty$, the threshold parameter satisfies $\gamma \gg (\kappa^2 M_1/\alpha_{\text{RSC}})\sqrt{s_0 p/T}$, where $s_0 = 2^{1-d}\sum_{k=1}^{2^{d-1}} \text{rank}(\mathcal{A}_{[I_k]})$.*

Assumption **??** requires that $\mathcal{A}$ has exact Tucker ranks $(r_1, \dots, r_{2d})$ which do not diverge too fast. The smallest positive singular value for each $\mathcal{A}_{(i)}$ is assumed to be bounded away from the threshold γ when the sample size is sufficiently large. Since Assumption **??** involves unknown quantities, it cannot be used directly for determining γ in practice. Instead, we recommend using the data-driven threshold parameter γ described at the end of the next subsection.

The rank selection consistency of the truncation method and the asymptotic estimation error rate of $\widehat{\mathcal{A}}_{\text{TSSN}}$ are given by the following theorem.

Theorem 5. *Under the conditions of Theorem **??** and Assumption **??**, if the tuning parameter $\lambda_{\text{SSN}} \asymp \kappa^2 M_1 2^{1-d}\sqrt{p/T}$, then*

$$\mathbb{P}\left\{\text{rank}\left((\widehat{\mathcal{A}}_{\text{TSSN}})_{(i)}\right) = \text{rank}(\mathcal{A}_{(i)}),\ \textit{for } i = 1, \dots, 2d\right\} \to 1,$$

as $T \to \infty$, and for any fixed d,

$$\|\widehat{\mathcal{A}}_{\text{TSSN}} - \mathcal{A}\|_{\text{F}} = O_p\left(\sqrt{s_0 p/T}\right),$$

*where s_0 is defined as in Assumption **??**.*

Algorithm 2 ADMM algorithm for (T)SSN estimator

```
Initialize: $\boldsymbol{\mathcal{C}}_k^{(0)}$, $\boldsymbol{\mathcal{W}}_k^{(0)} = \boldsymbol{\mathcal{A}}^{(0)} = \widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$, for $k = 1, \dots, 2^{d-1}$, threshold parameter $\gamma$
for $j \in \{0, 1, \dots, J-1\}$ do
    $\boldsymbol{\mathcal{A}}^{(j+1)} \leftarrow \arg\min \left\{ \mathcal{L}_T(\boldsymbol{\mathcal{A}}) + \sum_{k=1}^{2^{d-1}} \rho \|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{W}}_k^{(j)} + \boldsymbol{\mathcal{C}}_k^{(j)}\|_{\text{F}}^2 \right\}$
    for $k \in \{1, 2, \dots, 2^{d-1}\}$ do
        $\boldsymbol{\mathcal{W}}_k^{(j+1)} \leftarrow \arg\min \left\{ \rho \|\boldsymbol{\mathcal{A}}^{(j+1)} - \boldsymbol{\mathcal{W}}_k + \boldsymbol{\mathcal{C}}_k^{(j)}\|_{\text{F}}^2 + \lambda_{\text{SSN}} \|(\boldsymbol{\mathcal{W}}_k)_{[I_k]}\|_* \right\}$
        $\boldsymbol{\mathcal{C}}_k^{(j+1)} \leftarrow \boldsymbol{\mathcal{C}}_k^{(j)} + \boldsymbol{\mathcal{A}}^{(j+1)} - \boldsymbol{\mathcal{W}}_k^{(j+1)}$
    end for
end for
$\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} \leftarrow \boldsymbol{\mathcal{A}}^{(J)}$
for $i \in \{1, 2, \dots, 2d\}$ do
    $\widetilde{\boldsymbol{U}}_i \leftarrow \text{Truncated\_SVD}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}})_{(i)}, \gamma)$
end for
$\widetilde{\boldsymbol{\mathcal{G}}} \leftarrow \widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} \times_{i=1}^{2d} \widetilde{\boldsymbol{U}}_i^\top$
$\widehat{\boldsymbol{\mathcal{A}}}_{\text{TSSN}} \leftarrow \widetilde{\boldsymbol{\mathcal{G}}} \times_{i=1}^{2d} \widetilde{\boldsymbol{U}}_i$
```

5.4 ADMM Algorithm

This subsection presents the algorithm for the proposed (T)SSN regularized estimator. The algorithm for $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}$ can be developed analogously, while $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$ can be obtained easily as in **?**.

The objective function for the estimator $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ in (**??**) can be rewritten as

$$\mathcal{L}_T(\boldsymbol{\mathcal{A}}) + \lambda_{\text{SSN}} \|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} = \mathcal{L}_T(\boldsymbol{\mathcal{A}}) + \lambda_{\text{SSN}} \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{[I_k]}\|_*,$$

where $\mathcal{L}_T(\boldsymbol{\mathcal{A}}) = T^{-1} \sum_{t=1}^{T} \|\boldsymbol{\mathcal{Y}}_t - \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{Y}}_{t-1} \rangle\|_{\text{F}}^2$ is the quadratic loss function. In (**??**), the regularizer $\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}}$ involves 2^{d-1} nuclear norms $\|\boldsymbol{\mathcal{A}}_{[I_k]}\|_*$, which are challenging to handle at the same time. A similar difficulty also occurs in low-rank tensor completion, for which **?** applied the alternating direction method of multipliers (ADMM) algorithm (**?**) to efficiently separate the different nuclear norms. Borrowing the idea of **?**, we develop an ADMM algorithm for the miminization of (**??**).

To separate the 2^{d-1} nuclear norms in $\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}}$, for each $\boldsymbol{\mathcal{A}}_{[I_k]}$, we introduce a different dummy variable $\boldsymbol{\mathcal{W}}_k$ as a surrogate for $\boldsymbol{\mathcal{A}}$, where $k = 1, \dots, 2^{d-1}$. Then the augmented Lagrangian is

$$\mathcal{L}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{C}}) = \mathcal{L}_T(\boldsymbol{\mathcal{A}}) + \sum_{k=1}^{2^{d-1}} \Big[\lambda_{\text{SSN}}\|(\boldsymbol{\mathcal{W}}_k)_{[I_k]}\|_* + 2\rho\langle\boldsymbol{\mathcal{C}}_k, \boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{W}}_k\rangle + \rho\|\boldsymbol{\mathcal{A}} - \boldsymbol{\mathcal{W}}_k\|_{\text{F}}^2\Big],$$

where $\boldsymbol{\mathcal{C}}_k$ are the Lagrangian multipliers, for $k = 1, \dots, 2^{d-1}$, and ρ is the regularization parameter. Then we can iteratively update $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{W}}_k$ and $\boldsymbol{\mathcal{C}}_k$ by the ADMM, as shown in Algorithm **??**.

In Algorithm **??**, the $\boldsymbol{\mathcal{A}}$-update step is an ℓ_2-regularized least squares problem. Similarly to **?**, the $\boldsymbol{\mathcal{W}}_k$-update step can be solved by applying the explicit soft-thresholding operator to the singular values of $(\boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{C}}_k)_{[I_k]}$. Both subproblems have close-form solutions. Thus, the miminization of (**??**) can be solved efficiently.

For the tuning parameter selection, since the cross-validation method is unsuitable for time series or intrinsically ordered data, we apply the Bayesian information criterion (BIC) to select the optimal λ_{SSN} from a sequence of tuning parameter values, where the number of degrees of freedom is defined as $2^{-(d-1)}\sum_{k=1}^{2^{d-1}} s_k(2p - s_k)$. For the threshold parameter γ of the truncated estimator, we recommend $\gamma = 2^{d-1}\lambda_{\text{SSN}}/4$ to practitioners, where λ_{SSN} is the optimal tuning parameter selected by the BIC.

6 Simulation Studies

We conduct three simulation experiments to examine the finite-sample performance of the proposed low- and high-dimensional estimation methods for the LRTAR model in previous sections. Throughout, we generate the data from model (**??**) with $\text{vec}(\boldsymbol{\mathcal{E}}_t) \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{I}_p)$. The entries of the core tensor $\boldsymbol{\mathcal{G}}$ are generated independently from $N(0,1)$ and rescaled such that $\|\boldsymbol{\mathcal{G}}\|_{\text{F}} = 5$. The factor matrices $\boldsymbol{U}_i$'s are generated by extracting the leading singular vectors from Gaussian random matrices while ensuring the stationarity condition in Assumption **??**.

The first experiment focuses on the proposed low-dimensional estimation method considered in Section **??**. We consider four cases of data generating processes. In both cases (1a) and (1b), we consider $d = 2$ and multilinear ranks $(r_1, r_2, r_3, r_4) = (1, 1, 1, 1)$, $(2, 2, 2, 2)$, or $(2, 3, 2, 3)$. In both cases (2a) and (2b), we consider $d = 3$ and multilinear ranks $(r_1, r_2, r_3, r_4, r_5, r_6) = (1, 1, 1, 1, 1, 1)$, $(2, 2, 2, 1, 1, 1)$, or $(2, 2, 2, 2, 2, 2)$. Both pairs of cases differ in the setting of the dimensions p_i's: (1a) $p_1 = p_2 = 5$; (1b) $p_1 = p_2 = 10$; (2a) $p_1 = p_2 = p_3 = 5$; and (2b) $p_1 = p_2 = p_3 = 7$. We repeat each data generating process 300 times, and conduct the estimation using true multilinear ranks. Figure **??** displays the average estimation error against $T \in [1000, 2000]$. The LTR estimator clearly outperforms both the RRR and OLS estimators throughout all cases, confirming the theoretical efficiency comparison in Corollary **??**.

The second experiment aims to verify the estimation error bound for the proposed SSN estimator in high dimensions. By Theorem **??**, when the true transition tensor $\boldsymbol{\mathcal{A}}$ is exactly low-rank, we have $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}^2 = O_p(s_0 p/T)$, where $s_0 = 2^{1-d} \sum_{k=1}^{2^{d-1}} s_0^{(k)}$ with $s_0^{(k)} = \text{rank}(\boldsymbol{\mathcal{A}}_{[I_k]})$. Note that this rate is dependent on the overall dimension, $p = \prod_{i=1}^{d} p_i$, but independent of the individual dimensions p_i. To examine the relationship between the estimation error and the overal dimension p, sample size T, multilinear ranks r_i and individual dimensions p_i, we generate the data under the eight settings listed in Table **??**, and plot $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}^2$ against the varying parameter in each case in Figure **??**, including $p, 1/T, s_0$ and different settings of p_i's under a fixed overall dimension p. The first three columns of Figure **??** show that $\|\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}^2$ roughly scales linearly in p, T and s_0, while the last column suggests that the estimation error is invariant under different settings of individual p_i as long as p and the other parameters are fixed. This lends support to the theoretical error bound for the SSN estimator.

In the third experiment, we compare the performance of all the four high-dimensional estimators discussed in Section **??**, namely the SN, MN, SSN and TSSN estimators. We consider four cases of data generating processes. In cases (3a) and (3b), we consider $d = 2$ and multilinear ranks $(r_1, r_2, r_3, r_4) = (1, 1, 1, 1)$, $(2, 2, 1, 1)$, or $(2, 2, 2, 2)$. In cases (4a) and (4b), we consider $d = 3$ and multilinear ranks $(r_1, r_2, r_3, r_4, r_5, r_6) = (1, 1, 1, 1, 1, 1)$, $(2, 2, 2, 1, 1, 1)$, or $(2, 2, 2, 2, 2, 2)$. Similarly to the first experiment, both pairs of cases differ in the setting for p_i's: (3a) $p_1 = p_2 = 5$; (3b) $p_1 = p_2 = 10$; (4a) $p_1 = p_2 = p_3 = 5$; and (4b) $p_1 = p_2 = p_3 = 7$. For each setting, we repeat 300 times, and conduct the estimation using SN, MN, SSN and TSSN. The tuning parameter and the truncation threshold parameter are selected by the method described in Section **??**. In Figure **??**, the average estimation error is plotted against $T \in [400, 1000]$ for cases (3a) and (3b), and $T \in [600, 1200]$ for cases (4a) and (4b). First, it can be seen that the SN estimator is much inferior to the other three estimators, which is due to its use of the unbalanced one-mode matricizations. Secondly, the SSN and TSSN estimators outperform or are at least as good as the MN estimator in all cases, and their advantage is remarkably clear even when $r_1 = \cdots = r_{2d}$. In addition, the TSSN estimator generally performs better than the SSN, probably because the former yields a more parsimonious model which further improves the estimation efficiency.

7 Real Data Analysis

7.1 Matrix-Valued Time Series

We first consider the modeling of a matrix-valued time series. The data is the monthly market-adjusted return series of Fama–French 10×10 portfolios from January 1979 to December 2019, obtained from **?**. The portfolios are constructed as the intersections of 10 portfolios formed by the book-to-market (B/M) ratio and 10 portfolios formed by the size (market value of equity). Hence, $d = 2$, $p_1 = p_2 = 10$ and $T = 492$ months. We remove the market effect for each portfolio by subtracting its average return from the original return series.

Let $\boldsymbol{Y}_t \in \mathbb{R}^{10 \times 10}$ be the observed matrix-valued time series, with its rows and columns corresponding to different B/M ratios (sorted from lowest to highest) and sizes (sorted from smallest to largest), respectively, and denote $\boldsymbol{y}_t = \text{vec}(\boldsymbol{Y}_t)$. For comparison, we consider the following five candidate models:

- Vector autoregression (VAR): $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{e}_t$, where $\boldsymbol{A} \in \mathbb{R}^{100 \times 100}$. The model is estimated by the least squares method.

- 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. The model is estimated by the method in **?**, and for prediction, the estimated factors $\widehat{\boldsymbol{f}}_t$ are then fitted by a VAR(1) model.

- Matrix autoregression (MAR): $\boldsymbol{Y}_t = \boldsymbol{B}_1 \boldsymbol{Y}_{t-1} \boldsymbol{B}_2^\top + \boldsymbol{E}_t$, where $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{10\times 10}$ are coefficient matrices; see Example **??**. The model is estimated by the iterated least squares method in **?**.

- Matrix factor model (MFM): $\boldsymbol{Y}_t = \boldsymbol{R} \boldsymbol{F}_t \boldsymbol{C}^\top + \boldsymbol{E}_t$, where $\boldsymbol{F}_t$ is the low-dimensional matrix-valued latent factor, and $\boldsymbol{R}$ and $\boldsymbol{C}$ are the loading matrices. The model is estimated by the method in **?**, and for prediction, the estimated factors $\widehat{\boldsymbol{F}}_t$ are then fitted by a VAR(1) model.

- The proposed LRTAR model: $\boldsymbol{Y}_t = \langle \boldsymbol{\mathcal{A}}, \boldsymbol{Y}_{t-1} \rangle + \boldsymbol{E}_t$, with $\boldsymbol{\mathcal{A}} = \mathcal{G} \times_{i=1}^4 \boldsymbol{U}_i$. The model is estimated using the SSN and TSSN regularized estimation methods in Section **??**.

We first focus on the results for proposed LRTAR model. By the proposed truncation method, we obtain the estimated multilinear ranks $(\widehat{r}_1, \widehat{r}_2, \widehat{r}_3, \widehat{r}_4) = (8, 8, 2, 2)$. As shown in Example **??**, the model can be written equivalently as $\boldsymbol{U}_3^\top \boldsymbol{Y}_t \boldsymbol{U}_4 = \langle \mathcal{G}, \boldsymbol{U}_1^\top \boldsymbol{Y}_{t-1} \boldsymbol{U}_2 \rangle + \boldsymbol{U}_3^\top \boldsymbol{E}_t \boldsymbol{U}_4$. Thus, $\boldsymbol{U}_3$ and $\boldsymbol{U}_1$ can be viewed as the factor loadings across different B/M ratios, while $\boldsymbol{U}_4$ and $\boldsymbol{U}_2$ represent those across different sizes. By the factor interpretation below (**??**), this result indicates that the information in $\boldsymbol{Y}_t$ can be effectively summarized into the 2×2 response factors $\boldsymbol{U}_3^\top \boldsymbol{Y}_t \boldsymbol{U}_4$, while the low-rank structure associated with the predictor $\boldsymbol{Y}_{t-1}$ is not strong. Moreover, the estimated multilinear ranks suggest that the MAR model might be inefficient for the data, since the MAR imposes that $r_1 = r_2 = r_3 = r_4 = 10$ and $\boldsymbol{U}_3 = \boldsymbol{U}_4 = \boldsymbol{I}_{10}$; see Example **??**. In addition, by an argument similar to the discussion in Example **??**, the MFM can be regarded as a special case of the proposed model with $(\boldsymbol{U}_1, \boldsymbol{U}_2) = (\boldsymbol{U}_3, \boldsymbol{U}_4)$ and $(r_1, r_2) = (r_3, r_4)$. Thus, the fact that $\widehat{r}_3$ and $\widehat{r}_4$ are much smaller than $\widehat{r}_1$ and $\widehat{r}_2$ suggests that the MFM may be too restrictive for the data.

The TSSN estimates of the factor matrices are shown in Figure **??**. The patterns of the estimated response factor matrices $\widetilde{\boldsymbol{U}}_3, \widetilde{\boldsymbol{U}}_4 \in \mathbb{R}^{10\times 2}$ are particularly interesting. First, for both $\widehat{\boldsymbol{U}}_3$ and $\widehat{\boldsymbol{U}}_4$, the uniform pattern of the first column indicate that portfolios across different B/M ratios (or sizes) contribute approximately equally to the first B/M (or size) response factor. Thus, this factor represents the component of the market performance which is invariant to the size and B/M ratio of the portfolios. The significance of this factor may be partially because we fit the model to market-adjusted returns, where the average return is subtracted for all portfolios. Meanwhile, for both response factor matrices, the second column has a smoothly increasing pattern, suggesting that part of the return variation in the market has a monotonic relationship with the size and B/M ratio. Moreover, the above interpretations are consistent with the famous Fama-French three-factor model, where the return of a portfolio is expected to be affected by the market premium, outperformance of small versus big companies, and outperformance of high B/M versus small B/M companies.

The performance of the five models are compared through both average in-sample and out-of-sample forecasting errors. The average in-sample forecasting error is calculated based on the fitted models for the entire data, while the average out-of-sample forecasting error is calculated based on the rolling forecast procedure as follows. From January 2016 ($t = 445$) to December 2019 ($t = 492$), we fit the models using all the available data until time $t-1$ and obtain the one-step-ahead forecast $\widehat{\boldsymbol{Y}}_t$. Then, we obtain the average of the rolling forecasting errors for this period.

The average forecasting errors in ℓ_2 and ℓ_0 norms are presented in Table **??**. Firstly, the VAR model has the smallest in-sample forecasting error among all models, which is as expected because the VAR model is highly overparametrized. The bad in-sample performance of the MFM agrees with our discussion about its restrictiveness for the data due to the mismatch of the multilinear ranks. It is worth noting that the proposed LRTAR model has competitive in-sample forecasting performance among all models.

The out-of-sample forecasting results provides a fuller picture of the efficiency of different methods. It can be seen that the VAR and VFM models perform worst among all, as they both completely ignore the matrix structure of the data. Notably the proposed LRTAR model, fitted by either the SSN or the TSSN methods, have the smallest out-of-sample forecasting errors. This suggests that the proposed LRTAR model can indeed efficiently capture the dynamic structural information in the data.

7.2 Three-Way Tensor-Valued Time Series

? extended the classical Fama-French three-factor model to a five-factor model which further incorporates the effect of operating profitability (OP) and investment (Inv). This motivates tensor-valued stock returns data formed according to the size (small and big), B/M ratio (four groups sorted from lowest to highest), OP (four groups sorted from lowest to highest), and Inv (four groups sorted from lowest to highest). We consider two datasets retrieved from **?**. The first dataset consists of monthly market-adjusted returns series of $4 \times 4 \times 2$ portfolios from July 1963 to December 2019, formed by the OP, B/M ratio and size. The second dataset consists of those of $4 \times 4 \times 2$ portfolios formed by the Inv, B/M ratio and size. Hence, both are tensor-valued time series with $d = 3$, $p_1 = p_2 = 4$, $p_3 = 2$ and $T = 678$ months.

Similar to the analysis in the previous subsection, five candidate models are considered:

- Vector autoregression (VAR): $\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{e}_t$, where $\boldsymbol{A} \in \mathbb{R}^{32\times32}$.
- 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): $\boldsymbol{\mathcal{Y}}_t = \boldsymbol{\mathcal{Y}}_{t-1} \times_{i=1}^{3} \boldsymbol{B}_i + \boldsymbol{\mathcal{E}}_t$, where $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{4\times4}$ and $\boldsymbol{B}_3 \in \mathbb{R}^{2\times2}$ are coefficient matrices; see Example **??**.
- Tensor factor model (TFM): $\boldsymbol{\mathcal{Y}}_t = \boldsymbol{\mathcal{F}}_t \times_{i=1}^{3} \boldsymbol{U}_i + \boldsymbol{\mathcal{E}}_t$, where $\boldsymbol{\mathcal{F}}_t$ is the low-dimensional tensor-valued latent factor, and $\boldsymbol{U}_i$'s are the loading matrices.
- The proposed LRTAR model: $\boldsymbol{\mathcal{Y}}_t = \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{Y}}_{t-1} \rangle + \boldsymbol{\mathcal{E}}_t$, with $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_{i=1}^{6} \boldsymbol{U}_i$.

The TFM is estimated by the method in **?**, and for prediction, the estimated factors $\widehat{\mathfrak{F}}_t$ are fitted by a VAR(1) model. The other four models are fitted by the same methods as in the previous subsection. The multilinear ranks selected by the truncation method are $(\widehat{r}_1, \widehat{r}_2, \widehat{r}_3, \widehat{r}_4, \widehat{r}_5, \widehat{r}_6) = (3, 3, 2, 1, 1, 1)$ and $(2, 2, 2, 1, 1, 1)$ for the OP-BM-Size portfolio return series and the Inv-BM-Size portfolio return series, respectively. Note that similar to the BM-Size 10 × 10 series in the previous subsection, the low-rank structure for the response in these two tensor-valued datasets is more evident than that for the predictor. Figure **??** shows the TSSN estimates of the factor matrices. We find that the estimated response factors have a uniform pattern similar to that in Figure **??** for the matrix-valued data. The fact that only one response factor is extracted in each direction suggests that there might not be substantial effect of OP, Inv, B/M ratio or size on the returns for these datasets.

Finally, using the same methods as in the previous subsection, we calculate the average in-sample and out-of-sample forecasting errors for both datasets fitted with the five models. As shown in Table **??**, the comparison results for the two datasets are quite similar. It can be clearly observed that the VAR model always has the smallest in-sample forecasting error, yet the largest out-of-sample forecasting error. On the contrary, the proposed LRTAR model, fitted by either the SSN or the TSSN methods, has the smallest out-of-sample forecasting error. Moreover, the in-sample forecasting performance of the LRTAR model is competitive even compared to the VAR model. Similarly to the MFM in the previous subsection, the TFM model has poor in-sample performance, possibly due to the discrepancy between (r_1, r_2, r_3) and (r_4, r_5, r_6), as reflected by the estimated multilinear ranks. In sum, the results support the efficiency and flexibility of the proposed model and estimation methods for tensor-valued time series data.

8 Conclusion and Discussion

Efficient modeling and forecasting of general structured (tensor-valued), high-dimensional time series data is an important research topic which however has been rarely explored in the literature so far. This paper makes the first thorough attempt to address this problem by introducing the low-rank tensor autoregressive model. By assuming the exactly or approximately low-Tucker-rank structure of the transition tensor, the model exploits the low-dimensional tensor dynamic structure of the high-dimensional time series data, and summarizes the complex temporal dependencies into interpretable dynamic factors.

Asymptotic and non-asymptotic properties are derived for the proposed low- and high-dimensional estimators, respectively. For the latter, we relax the conventional Gaussian assumption in the high-dimensional time series literature to sub-Gaussianity via a new martingale-based concentration technique. Moreover, based on the special structure of the transition tensor, a novel convex regularizer, the SSN, is proposed, gaining efficiencies from both the square matricization and simultaneous penalization across modes. A truncation method, the TSSN, is further introduced to consistently select the multilinear ranks and enhance model interpretability.

We discuss several directions for future research. First, the proposed estimators cannot adapt to the contemporaneous correlation among elements of the random error $\boldsymbol{\mathcal{E}}_t$, which may lead to efficiency loss. This issue can be addressed by considering the generalized least squares loss function, $\mathcal{L}_T(\boldsymbol{\mathcal{A}};\boldsymbol{\Sigma}_{\boldsymbol{e}}) = T^{-1}\sum_{t=1}^{T}\text{vec}(\boldsymbol{\mathcal{Y}}_t - \langle\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{Y}}_{t-1}\rangle)^\top\boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1}\text{vec}(\boldsymbol{\mathcal{Y}}_t - \langle\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{Y}}_{t-1}\rangle)$, where $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \text{var}(\text{vec}(\boldsymbol{\mathcal{E}}_t))$. Then $\boldsymbol{\mathcal{A}}$ may be estimated jointly with $\boldsymbol{\Sigma}_{\boldsymbol{e}}$ or by a two-step approach based on a consistent estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ (**??**).

Secondly, the proposed methods can be generalized to the LRTAR model of finite lag order L, defined as $\boldsymbol{\mathcal{Y}}_t = \langle\boldsymbol{\mathcal{A}}_1,\boldsymbol{\mathcal{Y}}_{t-1}\rangle+\cdots+\langle\boldsymbol{\mathcal{A}}_L,\boldsymbol{\mathcal{Y}}_{t-L}\rangle+\boldsymbol{\mathcal{E}}_t$, where $\boldsymbol{\mathcal{A}}_1,\dots,\boldsymbol{\mathcal{A}}_L$ are $2d$-th-order multilinear low-rank coefficient tensors. Then, one may consider the SSN regularized estimation by minimizing $T^{-1}\sum_{t=1}^{T}\|\boldsymbol{\mathcal{Y}}_t-\sum_{j=1}^{L}\langle\boldsymbol{\mathcal{A}}_j,\boldsymbol{\mathcal{Y}}_{t-j}\rangle\|_{\text{F}}^2+\sum_{j=1}^{L}\lambda_j\|\boldsymbol{\mathcal{A}}_j\|_{\text{SSN}}$. Alternatively, the $\boldsymbol{\mathcal{A}}_j$'s can be combined into a $(2d+1)$-th-order tensor $\boldsymbol{\mathcal{A}}\in\mathbb{R}^{p_1\times\cdots p_d\times p_1\times\cdots\times p_d\times L}$ whose mode-$(2d+1)$ matricization is $\boldsymbol{\mathcal{A}}_{(2d+1)} = (\boldsymbol{\mathcal{A}}_1,\dots,\boldsymbol{\mathcal{A}}_L)$; see **?** for a similar idea. Even though $\boldsymbol{\mathcal{A}}$ may not have exactly square matricizations for $L>1$, the proposed SSN can still be adapted by employing approximately square matricizations. For instance, consider the $pL\times p$ multi-mode matricizations $\boldsymbol{\mathcal{A}}_{[I_k\cup\{2d+1\}]}$, where the index sets I_k for $k=1,\dots,2^{d-1}$ are defined as in this paper.

Then, a generalized SSN norm can be constructed as $\|\boldsymbol{\mathcal{A}}\|_{\text{GSSN}} = \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{[I_k \cup \{2d+1\}]}\|_*$, which will reduce to $\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}}$ when $L = 1$.

Thirdly, the LRTAR model can be readily extended to incorporate exogenous tensor-valued predictors, giving rise to the LRTAR-X model, $\boldsymbol{\mathcal{Y}}_t = \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{Y}}_{t-1} \rangle + \langle \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{X}}_t \rangle + \boldsymbol{\mathcal{E}}_t$, where $\boldsymbol{\mathcal{X}}_t$ is an m-th-order tensor of exogenous variables, and $\boldsymbol{\mathcal{B}}$ is a $(d+m)$-th-order coefficient tensor. When the dimension of $\boldsymbol{\mathcal{X}}_t$ is high, a low-dimensional structure, such as sparsity, group sparsity or low-rankness, can be imposed on $\boldsymbol{\mathcal{B}}$ to improve the estimation efficiency.

Moreover, an open question for matrix- or tensor-valued time series models is how to incorporate additional structures or constraints; for instance, transport or trade flow data have unspecified diagonal entries (**?**), and realized covariance matrix data are subject to positive definite constraints. Beyond the time series context, it is also worth investigating the generalization of the SSN regularization method in higher-order tensor estimation and completion applications, such as neuroimaging analysis (**?**), recommender sytem (**?**) and natural language processing (**?**).

Appendix A: Proofs for Low-Dimensional Estimation

Below we give the proofs of Theorem **??** and Corollary **??** in Section **??**, which generally follow from Proposition 4.1 in **?** for overparameterized models.

Proof of Theorem **??**. The proposed model in (**??**) can be written in the matrix form

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1^\top \\ \boldsymbol{y}_2^\top \\ \vdots \\ \boldsymbol{y}_T^\top \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} \boldsymbol{y}_0^\top \\ \boldsymbol{y}_1^\top \\ \vdots \\ \boldsymbol{y}_{T-1}^\top \end{bmatrix}}_{\boldsymbol{X}} \boldsymbol{\mathcal{A}}_{[S_2]}^\top + \underbrace{\begin{bmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{e}_2^\top \\ \vdots \\ \boldsymbol{e}_T^\top \end{bmatrix}}_{\boldsymbol{E}}.$$

Let $\widehat{\boldsymbol{h}}_{\text{OLS}} = \text{vec}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{OLS}})_{[S_2]})$. Under Assumption **??**, by the classical asymptotic theory for stationary VAR models, as $T \to \infty$, we have

$$\sqrt{T}(\widehat{\boldsymbol{h}}_{\text{OLS}} - \boldsymbol{h}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{OLS}}),$$

where $\boldsymbol{\Sigma}_{\text{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{e}} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}$, with $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \text{var}(\boldsymbol{e}_t)$ and $\boldsymbol{\Sigma}_{\boldsymbol{y}} = \text{var}(\boldsymbol{y}_t)$.

Following **?**, consider the discrepancy function

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

Note that $F(\widehat{\boldsymbol{h}}_{\text{OLS}}, \boldsymbol{h})$ is nonnegative and twice continuously differentiable, and equals zero if and only if $\widehat{\boldsymbol{h}}_{\text{OLS}} = \boldsymbol{h}$.

The Jacobian matrix $\boldsymbol{H} = \partial \boldsymbol{h}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ in (**??**) can be verified by noting that

$$\boldsymbol{h} = \text{vec}(\boldsymbol{\mathcal{A}}_{[S_2]}) = \text{vec}((\otimes_{i\in S_2}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{[S_2]}(\otimes_{i\in S_1}\boldsymbol{U}_i)^\top) = ((\otimes_{i\in S_1}\boldsymbol{U}_i) \otimes (\otimes_{i\in S_2}\boldsymbol{U}_i))\,\text{vec}(\boldsymbol{\mathcal{G}}_{[S_2]})$$

and that for any $1 \le i \le 2d$,

$$\begin{aligned}\boldsymbol{h} = \text{vec}(\boldsymbol{\mathcal{A}}_{[S_2]}) = \boldsymbol{P}_{[S_2]}^{(i)}\text{vec}(\boldsymbol{\mathcal{A}}_{(i)}) &= \boldsymbol{P}_{[S_2]}^{(i)}\text{vec}\left(\boldsymbol{U}_i\boldsymbol{\mathcal{G}}_{(i)}(\otimes_{j=1,j\neq i}^{2d}\boldsymbol{U}_j^\top)\right) \\ &= \boldsymbol{P}_{[S_2]}^{(i)}\left\{\left[(\otimes_{j=1,j\neq i}^{2d}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{(i)}^\top\right] \otimes \boldsymbol{I}_{p_i}\right\}\text{vec}(\boldsymbol{U}_i).\end{aligned}$$

Then, by Proposition 4.1 in **?**, we obtain that the minimizer of $F(\widehat{\boldsymbol{h}}_{\text{OLS}}, \cdot)$, namely the LTR estimator, has the asymptotic normality,

$$\sqrt{T}(\widehat{\boldsymbol{h}}_{\text{LTR}} - \boldsymbol{h}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{LTR}})$$

and $\boldsymbol{\Sigma}_{\text{LTR}} = \boldsymbol{P}\boldsymbol{\Sigma}_{\text{OLS}}\boldsymbol{P}^\top$, where $\boldsymbol{P} = \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{J}\boldsymbol{H})^\dagger\boldsymbol{H}^\top\boldsymbol{J}$ is the projection matrix, $\boldsymbol{J} = \boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1} \otimes \boldsymbol{\Sigma}_{\boldsymbol{y}}$ is the Fisher information matrix of $\boldsymbol{h}$, and † denotes the Moore-Penrose inverse. Since $\boldsymbol{\Sigma}_{\text{OLS}} = \boldsymbol{J}^{-1}$, we can obtain that $\boldsymbol{\Sigma}_{\text{LTR}} = \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{J}\boldsymbol{H})^\dagger\boldsymbol{H}^\top$. □

Proof of Corollary **??**. As discussed in the proof of Theorem **??**, $\boldsymbol{\Sigma}_{\text{OLS}} = \boldsymbol{J}^{-1}$, and observe that

$$\boldsymbol{\Sigma}_{\text{LTR}} = \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{J}\boldsymbol{H})^\dagger\boldsymbol{H}^\top = \boldsymbol{J}^{-1/2}\boldsymbol{Q}_{\boldsymbol{J}^{1/2}\boldsymbol{H}}\boldsymbol{J}^{-1/2},$$

where $\boldsymbol{Q}_{\boldsymbol{J}^{1/2}\boldsymbol{H}}$ is the projection matrix onto the orthogonal compliment of $\text{span}(\boldsymbol{J}^{1/2}\boldsymbol{H})$.

On the other hand, under the proposed model, the transition matrix can be decomposed as $\boldsymbol{\mathcal{A}}_{[S_2]} = \boldsymbol{V}_1\boldsymbol{V}_2^\top$, where $\boldsymbol{V}_1 = \otimes_{i\in S_1}\boldsymbol{U}_i$, $\boldsymbol{V}_2 = (\otimes_{i\in S_2}\boldsymbol{U}_i)\boldsymbol{\mathcal{G}}_{[S_2]}$, and $\text{rank}(\boldsymbol{\mathcal{A}}_{[S_2]}) \leq s_1 = \min(\prod_{i=1}^d r_i, \prod_{i=d+1}^{2d} r_i)$. Thus, we can write $\boldsymbol{h} = \text{vec}(\boldsymbol{\mathcal{A}}_{[S_2]}) = \boldsymbol{h}(\boldsymbol{\phi})$, where $\boldsymbol{\phi} = (\text{vec}(\boldsymbol{V}_1)^\top, \text{vec}(\boldsymbol{V}_2)^\top)^\top$ is the parameter vector for the RRR estimator. Then, similarly to the proof of Theorem **??**, we denote the Jacobian matrix by $\boldsymbol{R} = \partial\boldsymbol{h}/\partial\boldsymbol{\phi}$. By similar arguments, we have $\sqrt{T}(\widehat{\boldsymbol{h}}_{\text{RRR}} - \boldsymbol{h}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_{\text{RRR}})$, where $\widehat{\boldsymbol{h}}_{\text{RRR}} = \text{vec}((\widehat{\boldsymbol{\mathcal{A}}}_{\text{RRR}})_{[S_2]})$, and

$$\boldsymbol{\Sigma}_{\text{RRR}} = \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{J}\boldsymbol{R})^\dagger\boldsymbol{R}^\top = \boldsymbol{J}^{-1/2}\boldsymbol{Q}_{\boldsymbol{J}^{1/2}\boldsymbol{R}}\boldsymbol{J}^{-1/2},$$

where $\boldsymbol{Q}_{\boldsymbol{J}^{1/2}\boldsymbol{R}}$ is the projection matrix onto the orthogonal compliment of $\text{span}(\boldsymbol{J}^{1/2}\boldsymbol{R})$. Hence, it is clear that $\boldsymbol{\Sigma}_{\text{RRR}} \leq \boldsymbol{J}^{-1} = \boldsymbol{\Sigma}_{\text{OLS}}$.

Moreover, since the Tucker decomposition can be viewed as a further decomposition of the low-rank decomposition $\boldsymbol{V}\boldsymbol{U}^\top$ for $\boldsymbol{\mathcal{A}}_{[S_2]}$, we have $\boldsymbol{H} = \partial\boldsymbol{h}/\partial\boldsymbol{\theta} = \boldsymbol{R}\cdot\partial\boldsymbol{\phi}/\partial\boldsymbol{\theta}$. By (**??**) and (**??**), we have $\boldsymbol{\Sigma}_{\text{LTR}} \leq \boldsymbol{\Sigma}_{\text{RRR}}$, since $\text{span}(\boldsymbol{J}^{1/2}\boldsymbol{H}) \subset \text{span}(\boldsymbol{J}^{1/2}\boldsymbol{R})$. □

Appendix B: Proofs for High-Dimensional Estimation

In this appendix, we provide the proofs of Theorems **??**–**??** in Section **??**. We start with a preliminary analysis in Appendix **??** which lays out the common technical framework for proving the estimation and prediction error bounds of the SN, MN and SSN regularized estimators, and four lemmas, Lemmas **??**–**??**, are introduced herein. Then in Appendix

?? we give the proofs of Theorems ??–??. The proofs of Lemmas ??–?? are provided in Appendix ??, and three auxiliary lemmas are collected in Appendix ??

B.1 Preliminary Analysis

The technical framework for proving the error bounds in Theorem ??–?? consists of two main steps, a deterministic analysis and a stochastic analysis, given in Sections ?? and ??, respectively. The goal of the first one is to derive the error bounds given the deterministic realization of the time series, assuming that the parameters satisfy certain regularity conditions. The goal of the second one is to verify that under stochasticity these regularity conditions are satisfied with high probability.

B.1.1 Deterministic Analysis

Throughout the appendix, we adopt the following notations. We use C to denote a generic positive constant, which is independent of the dimensions and the sample size. For any matrix $\boldsymbol{M}$ and a compatible subspace $\mathcal{S}$, we denote by $\boldsymbol{M}_{\mathcal{S}}$ the projection of $\boldsymbol{M}$ onto $\mathcal{S}$. In addition, let $\text{col}(\boldsymbol{M})$ be the column space of $\boldsymbol{M}$, and let $\mathcal{S}^{\perp}$ be the complement of the subspace $\mathcal{S}$. For a generic tensor $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{p_1\times\cdots\times p_{2d}}$, the dual norms of its SSN norm and SN norm, denoted by $\|\boldsymbol{\mathcal{W}}\|_{\text{SSN}^*}$ and $\|\boldsymbol{\mathcal{W}}\|_{\text{SN}^*}$, respectively, are defined as

$$\|\boldsymbol{\mathcal{W}}\|_{\text{SSN}^*} = \sup_{\boldsymbol{\mathcal{T}}\in\mathbb{R}^{p_1\times\cdots\times p_{2d}}, \|\boldsymbol{\mathcal{T}}\|_{\text{SSN}}\leq 1} \langle \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{T}}\rangle, \quad \text{and} \quad \|\boldsymbol{\mathcal{W}}\|_{\text{SN}^*} = \sup_{\boldsymbol{\mathcal{T}}\in\mathbb{R}^{p_1\times\cdots\times p_{2d}}, \|\boldsymbol{\mathcal{T}}\|_{\text{SN}}\leq 1} \langle \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{T}}\rangle.$$

Moreover, for any two tensors $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1\times\cdots\times p_m}$ and $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{p_{m+1}\times\cdots\times p_{m_n}}$, their tensor outer product is defined as $(\boldsymbol{\mathcal{X}} \circ \boldsymbol{\mathcal{Y}}) \in \mathbb{R}^{p_1\times\cdots\times p_m\times p_{m+1}\times\cdots\times p_{m+n}}$ where

$$(\boldsymbol{\mathcal{X}} \circ \boldsymbol{\mathcal{Y}})_{i_1\ldots i_m i_{m+1}\ldots i_{m+n}} = \boldsymbol{\mathcal{X}}_{i_1\ldots i_m}\boldsymbol{\mathcal{Y}}_{i_{m+1}\ldots i_{m+n}},$$

for any $1 \leq i_1 \leq p_1, \ldots, 1 \leq i_{m+n} \leq p_{m+n}$.

For the theory of regularized M-estimators, restricted error sets and restricted strong convexity are essential definitions. To define the former, we need to first introduce the following restricted model subspaces.

For $i = 1, \dots, 2d$, denote by $\widetilde{\mathcal{U}}_i$ and $\widetilde{\mathcal{V}}_i$ the spaces spanned by the first r_i left and right singular vectors in the SVD of $\boldsymbol{\mathcal{A}}_{(i)}$, respectively. Define the collections of subspaces

$$\mathcal{N} = (\mathcal{N}_1, \dots, \mathcal{N}_{2d}) \quad \text{and} \quad \overline{\mathcal{N}}^{\perp} = (\overline{\mathcal{N}}_1^{\perp}, \dots, \overline{\mathcal{N}}_{2d}^{\perp}),$$

where

$$\mathcal{N}_i = \{\boldsymbol{M} \in \mathbb{R}^{p_i \times p_{-i}p} | \text{col}(\boldsymbol{M}) \subset \widetilde{\mathcal{U}}_i, \text{col}(\boldsymbol{M}^\top) \subset \widetilde{\mathcal{V}}_i\},$$
$$\overline{\mathcal{N}}_i^{\perp} = \{\boldsymbol{M} \in \mathbb{R}^{p_i \times p_{-i}p} | \text{col}(\boldsymbol{M}) \perp \widetilde{\mathcal{U}}_i, \text{col}(\boldsymbol{M}^\top) \perp \widetilde{\mathcal{V}}_i\},$$

for $i = 1, \dots, 2d$. Note that $\mathcal{N}_i \subset \overline{\mathcal{N}}_i$.

Furthermore, for $k = 1, \dots, 2^{d-1}$, denote by $\mathcal{U}_k$ and $\mathcal{V}_k$ the spaces spanned by the first s_k^* left and right singular vectors in the SVD of the square matricization $\boldsymbol{\mathcal{A}}_{[I_k]}$, respectively, where $s_k^* = \text{rank}(\boldsymbol{\mathcal{A}})_{[I_k]}$. Similarly, define the collections of subspaces

$$\mathcal{M} := (\mathcal{M}_1, \dots, \mathcal{M}_{2^{d-1}}) \quad \text{and} \quad \overline{\mathcal{M}}^{\perp} = (\overline{\mathcal{M}}_1^{\perp}, \dots, \overline{\mathcal{M}}_{2^{d-1}}^{\perp}),$$

where

$$\mathcal{M}_k = \{\boldsymbol{M} \in \mathbb{R}^{p \times p} | \text{col}(\boldsymbol{M}) \subset \mathcal{U}_k, \ \text{col}(\boldsymbol{M}^\top) \subset \mathcal{V}_k\},$$
$$\overline{\mathcal{M}}_k^{\perp} = \{\boldsymbol{M} \in \mathbb{R}^{p \times p} | \text{col}(\boldsymbol{M}) \perp \mathcal{U}_k, \ \text{col}(\boldsymbol{M}^\top) \perp \mathcal{V}_k\},$$

for $k = 1, \dots, 2^{d-1}$. In particular, as described in Section **??**, $I_1 = S_1 = \{1, \dots, d\}$. Thus, $\mathcal{M}_1$ and $\overline{\mathcal{M}}_1^{\perp}$ are the subspaces associated with the square matricization $\boldsymbol{\mathcal{A}}_{[S_1]}$.

Then, for simplicity, for any $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}}$, we denote

$$\boldsymbol{\mathcal{W}}_{\mathcal{N}}^{(i)} = (\boldsymbol{\mathcal{W}}_{(i)})_{\mathcal{N}_i}, \quad \boldsymbol{\mathcal{W}}_{\mathcal{N}^\perp}^{(i)} = (\boldsymbol{\mathcal{W}}_{(i)})_{\mathcal{N}_i^\perp}, \quad \boldsymbol{\mathcal{W}}_{\overline{\mathcal{N}}}^{(i)} = (\boldsymbol{\mathcal{W}}_{(i)})_{\overline{\mathcal{N}}_i}, \quad \boldsymbol{\mathcal{W}}_{\overline{\mathcal{N}}^\perp}^{(i)} = (\boldsymbol{\mathcal{W}}_{(i)})_{\overline{\mathcal{N}}_i^\perp}$$
$$\boldsymbol{\mathcal{W}}_{\mathcal{M}}^{(k)} = (\boldsymbol{\mathcal{W}}_{[I_k]})_{\mathcal{M}_k}, \quad \boldsymbol{\mathcal{W}}_{\mathcal{M}^\perp}^{(k)} = (\boldsymbol{\mathcal{W}}_{[I_k]})_{\mathcal{M}_k^\perp}, \quad \boldsymbol{\mathcal{W}}_{\overline{\mathcal{M}}}^{(k)} = (\boldsymbol{\mathcal{W}}_{[I_k]})_{\overline{\mathcal{M}}_k}, \quad \boldsymbol{\mathcal{W}}_{\overline{\mathcal{M}}^\perp}^{(k)} = (\boldsymbol{\mathcal{W}}_{[I_k]})_{\overline{\mathcal{M}}_k^\perp},$$

where $i = 1, \dots, 2d$ and $k = 1, \dots, 2^{d-1}$. Based on the subspaces defined in (**??**) and (**??**), we can define the restricted error sets corresponding to the three regularized estimators as follows.

Definition 1. *The restricted error set corresponding to $\overline{\mathcal{M}}$ is defined as*

$$\mathbb{C}_{\text{SSN}}(\overline{\mathcal{M}}) := \left\{ \boldsymbol{\Delta} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}} : \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* \leq 3 \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* + 4 \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)}\|_* \right\}.$$

The restricted error set corresponding to $\overline{\mathcal{N}}$ is defined as

$$\mathbb{C}_{\mathrm{SN}}(\overline{\mathcal{N}}) := \left\{ \boldsymbol{\Delta} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}} : \sum_{i=1}^{2d} \|\boldsymbol{\Delta}^{(i)}_{\overline{\mathcal{N}}^{\perp}}\|_* \leq 3 \sum_{i=1}^{2d} \|\boldsymbol{\Delta}^{(i)}_{\overline{\mathcal{N}}}\|_* + 4 \sum_{i=1}^{2d} \|\boldsymbol{\mathcal{A}}^{(i)}_{\mathcal{N}^{\perp}}\|_* \right\}.$$

The restricted error set corresponding to $\overline{\mathcal{M}}_1$ is defined as

$$\mathbb{C}_{\mathrm{MN}}(\overline{\mathcal{M}}_1) := \left\{ \boldsymbol{\Delta} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}} : \|\boldsymbol{\Delta}^{(1)}_{\overline{\mathcal{M}}^{\perp}}\|_* \leq 3 \|\boldsymbol{\Delta}^{(1)}_{\overline{\mathcal{M}}}\|_* + 4 \|\boldsymbol{\mathcal{A}}^{(1)}_{\mathcal{M}^{\perp}}\|_* \right\}.$$

The first lemma shows that if the tuning parameter is well chosen for each regularized estimator, the estimation error belongs to the corresponding restricted error set.

Lemma B.1. *For the SSN estimator, if the regularization parameter $\lambda_{\mathrm{SSN}} \geq 4\|T^{-1}\sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t\|_{\mathrm{SSN}^*}$, the error $\boldsymbol{\Delta}_{\mathrm{SSN}} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{SSN}} - \boldsymbol{\mathcal{A}}$ belongs to the set $\mathbb{C}_{\mathrm{SSN}}(\overline{\mathcal{M}})$.*

For the SN estimator, if the regularization parameter $\lambda_{\mathrm{SN}} \geq 4\|T^{-1}\sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t\|_{\mathrm{SN}^*}$, the error $\boldsymbol{\Delta}_{\mathrm{SN}} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{SN}} - \boldsymbol{\mathcal{A}}$ belongs to the set $\mathbb{C}_{\mathrm{SN}}(\overline{\mathcal{N}})$.

For the MN estimator, if the regularization parameter $\lambda_{\mathrm{MN}} \geq 4\|T^{-1}\sum_{t=1}^{T} \mathrm{vec}(\boldsymbol{\mathcal{Y}}_{t-1})\mathrm{vec}(\boldsymbol{\mathcal{E}}_t)^{\top}\|_*$, the error $\boldsymbol{\Delta}_{\mathrm{MN}} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MN}} - \boldsymbol{\mathcal{A}}$ belongs to the set $\mathbb{C}_{\mathrm{MN}}(\overline{\mathcal{M}}_{(1)})$.

Following **?** and **?**, a restricted strong convexity (RSC) condition for the square loss function can be defined as follows.

Definition 2. *The loss function satisfies the RSC condition with curvature $\alpha_{\mathrm{RSC}} > 0$ and restricted error set $\mathbb{C}$, if*

$$\frac{1}{T}\sum_{t=1}^{T} \|\langle \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}_{t-1} \rangle\|_F^2 \geq \alpha_{\mathrm{RSC}} \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{\Delta} \in \mathbb{C}.$$

Based on the restricted error sets and RSC conditions, the estimation errors have the following deterministic upper bounds.

Lemma B.2. *Suppose that $\lambda_{\mathrm{SSN}} \geq 4\|T^{-1}\sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t\|_{\mathrm{SSN}^*}$, the RSC condition holds with the parameter α_{RSC} and restricted error set $\mathbb{C}_{\mathrm{SSN}}(\overline{\mathcal{M}})$, and $\boldsymbol{\mathcal{A}}_{[I_k]} \in \mathbb{B}_q(s_q^{(k)}; p, p)$ for some $q \in [0, 1)$ and all $k = 1, \ldots, 2^{d-1}$,*

$$\|\boldsymbol{\Delta}_{\mathrm{SSN}}\|_{\mathrm{F}} \lesssim \sqrt{s_q} \left(\frac{2^{d-1}\lambda_{\mathrm{SSN}}}{\alpha_{\mathrm{RSC}}} \right)^{1-q/2},$$

where $s_q = 2^{1-d} \sum_{k=1}^{2^{d-1}} s_q^{(k)}$.

Suppose that $\lambda_{\text{SN}} \geq 4\|T^{-1} \sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t\|_{\text{SN}^*}$, *the RSC condition holds with the parameter* α_{RSC} *and restricted error set* $\mathbb{C}_{\text{SN}}(\overline{\mathcal{N}})$, *and* $\boldsymbol{\mathcal{A}}_{(i)} \in \mathbb{B}_q(r_q^{(i)}; p_i, p_{-i}p)$ *for some* $q \in [0, 1)$ *and all* $i = 1, \ldots, 2d$,

$$\|\boldsymbol{\Delta}_{\text{SN}}\|_{\text{F}} \lesssim \sqrt{r_q} \left(\frac{2d \cdot \lambda_{\text{SN}}}{\alpha_{\text{RSC}}} \right)^{1-q/2},$$

where $r_q = (2d)^{-1} \sum_{i=1}^{2d} r_q^{(i)}$.

Suppose that $\lambda_{\text{MN}} \geq 4\|T^{-1} \sum_{t=1}^{T} \text{vec}(\boldsymbol{\mathcal{Y}}_{t-1})\text{vec}(\boldsymbol{\mathcal{E}}_t)\|_*$, *the RSC condition holds with the parameter* α_{RSC} *and restricted error set* $\mathbb{C}_{\text{MN}}(\overline{\mathcal{M}}_1)$, *and* $\boldsymbol{\mathcal{A}}_{[S_1]} \in \mathbb{B}_q(s_q^{(1)}; p, p)$ *for some* $q \in [0, 1)$,

$$\|\boldsymbol{\Delta}_{\text{MN}}\|_{\text{F}} \lesssim \sqrt{s_q^{(1)}} \left(\frac{\lambda_{\text{MN}}}{\alpha_{\text{RSC}}} \right)^{1-q/2}.$$

Note that Lemma **??** is deterministic and the radius s_q, r_q, and $s_q^{(1)}$ can also diverge to infinity.

B.1.2 Stochastic Analysis

We continue with the stochastic analysis to show that the deviation bound and the RSC condition hold simultaneously with high probability.

Lemma B.3 (Deviation bound)**.** *Suppose that Assumptions* **??** *and* **??** *hold. If* $T \gtrsim p$ *and* $\lambda_{\text{SSN}} \gtrsim \kappa^2 M_1 2^{1-d} \sqrt{p/T}$, *with probability at least* $1 - \exp[-C(p - d)]$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t \right\|_{\text{SSN}^*} \leq \frac{\lambda_{\text{SSN}}}{4}$$

where $M_1 = \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})/\mu_{\min}^{1/2}(\mathcal{A})$.

If $T \gtrsim \max_{1 \leq i \leq d} p_{-i}p$ *and* $\lambda_{\text{SN}} \gtrsim \kappa^2 M_1 d^{-2} \sum_{i=1}^{d} \sqrt{p_{-i}p/T}$, *with probability at least* $1 - 2\sum_{i=1}^{d} \exp(-Cp_{-i}p)$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t \right\|_{\text{SN}^*} \leq \frac{\lambda_{\text{SN}}}{4}.$$

Moreover, if $T \gtrsim p$ *and* $\lambda_{\text{MN}} \gtrsim \kappa^2 M_1 \sqrt{p/T}$, *with probability at least* $1 - \exp(-Cp)$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \text{vec}(\boldsymbol{\mathcal{Y}}_{t-1})\text{vec}(\boldsymbol{\mathcal{E}}_t)^\top \right\|_* \leq \frac{\lambda_{\text{MN}}}{4}.$$

Next, we prove the restricted strong convexity for regularized estimators. According to Lemma **??**, we need the sample size $T \gtrsim p$ for all three estimators. In this case, we can establish the strong convexity condition that is stronger than the RSC condition.

Lemma B.4 (Strong convexity)**.** *Under Assumptions* **??** *and* **??***, for* $T \gtrsim \max(\kappa^2, \kappa^4)M_2^{-2}p$*, with probability at least* $1 - \exp[-C\min(\kappa^{-2}, \kappa^{-4})M_2^2 p]$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\langle\boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_F^2 \geq \alpha_{\text{RSC}}\|\boldsymbol{\Delta}\|_{\text{F}}^2,$$

where $M_2 = [\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\mu_{\max}(\mathcal{A})]/[\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\mu_{\min}(\mathcal{A})]$ *and* $\alpha_{\text{RSC}} = \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})/(2\mu_{\max}(\mathcal{A}))$*.*

B.2 Proofs of Theorems ??–??

Proof of Theorems **??** *and* **??***.* Theorems **??** and **??** can be proved based on Lemmas **??**–**??** following the same line of the proof of Theorem **??** given below. Therefore, we omit the details here. □

Proof of Theorem **??***.* The proof of Theorem **??** has been split into Lemmas **??**–**??**. By Lemma **??**, for deterministic realization with sample size T of a tensor autoregressive process, if we choose $\lambda_{\text{SSN}} \geq 4\|T^{-1}\sum_{t=1}^{T}\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\|_{\text{SSN}^*}$ and RSC condition holds for the square loss with the parameter α_{RSC}, the following error upper bound can be established

$$\|\boldsymbol{\Delta}\|_{\text{F}} \lesssim \sqrt{s_q}\left(\frac{2^{d-1}\lambda_{\text{SSN}}}{\alpha_{\text{RSC}}}\right)^{1-q/2}.$$

Denote the events $E_1(\beta) = \{\beta \geq 4\|T^{-1}\sum_{t=1}^{T}\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\|_{\text{SSN}^*}\}$ and $E_2(\alpha) = \{\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^\top/T) \geq \alpha\}$. If we take $\lambda_{\text{SSN}} \gtrsim \kappa^2 M_1 2^{1-d}\sqrt{p/T}$, it suffices to show that $E_1(C\kappa^2 M_1 2^{1-d}\sqrt{p/T})$ and $E_2(\alpha_{\text{RSC}}/2)$ occur simultaneously with high probability.

By Lemma **??**, when $T \gtrsim p$,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\right\|_{\text{SSN}^*} \lesssim \kappa^2 M_1 2^{1-d}\sqrt{\frac{p}{T}}$$

with probability at least $1 - \exp[-C(p-d)]$.

By Lemma **??**, when $T \gtrsim \max(\kappa^2, \kappa^4) M_2^{-2} p$, for any $\boldsymbol{\Delta} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}}$,

$$\frac{1}{T} \sum_{t=1}^{T} \|\langle \boldsymbol{\Delta}, \boldsymbol{\mathcal{Y}}_{t-1} \rangle\|_{\mathrm{F}}^2 \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})}{2\mu_{\max}(\mathcal{A})} \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2$$

with probability at least $1 - \exp[-C \min(\kappa^{-2}, \kappa^{-4}) M_2^2 p]$.

Hence, when $T \gtrsim [1 + \max(\kappa^2, \kappa^4) M_2^{-2}] p$ and $\lambda \gtrsim \kappa^2 M_1 2^{1-d} \sqrt{p/T}$, with probability at least $1 - \exp[-C(p-d)] - \exp[-C \min(\kappa^{-2}, \kappa^{-4}) M_2^2 p]$, the condition $\lambda \geq 4 \|T^{-1} \sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t\|_{\mathrm{SSN}^*}$ and the RSC condition with the parameter $\alpha_{\mathrm{RSC}} = \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})/\mu_{\max}(\mathcal{A})$ hold. □

Proof of Theorem **??**. Theorem **??** gives the Frobenius estimation error bound. For simplicity, we write $\widehat{\boldsymbol{\mathcal{A}}} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{SSN}}$ and $\widetilde{\boldsymbol{\mathcal{A}}} = \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{TSSN}}$ in this proof. By definition, for any tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}}$,

$$\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2 = \|\boldsymbol{\mathcal{A}}_{(i)}\|_{\mathrm{F}}^2 = \sum_{j=1}^{p_i} \sigma_j^2(\boldsymbol{\mathcal{A}}_{(i)}), \quad i = 1, 2, \ldots, 2d.$$

In other words, the Frobenius norm of the error tensor is equivalent to the ℓ_2 norm of singular values of the one-mode matricization. By Mirsky's singular value inequality (**?**),

$$\sum_{j=1}^{p_i} [\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}) - \sigma_j(\boldsymbol{\mathcal{A}}_{(i)})]^2 \leq \sum_{j=1}^{p_i} \sigma_j^2(\widehat{\boldsymbol{\mathcal{A}}}_{(i)} - \boldsymbol{\mathcal{A}}_{(i)}) = \|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2, \quad i = 1, 2, \ldots, 2d.$$

Obviously, the ℓ_∞ error bound is smaller than the ℓ_2 error bound, so it follows the same upper bound. By Theorem **??**, when $\lambda_{\mathrm{SSN}} \asymp \kappa^2 M_1 2^{1-d} \sqrt{p/T}$, with probability approaching one,

$$\begin{aligned} \max_{1 \leq i \leq 2d} \max_{1 \leq j \leq p_i} |\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}) - \sigma_j(\boldsymbol{\mathcal{A}}_{(i)})| \leq & \max_{1 \leq i \leq 2d} \left\{ \sum_{j=1}^{p_i} [\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}) - \sigma_j(\boldsymbol{\mathcal{A}}_{(i)})]^2 \right\}^{1/2} \\ \leq & \|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}} \lesssim \frac{\kappa^2 M_1}{\alpha_{\mathrm{RSC}}} \sqrt{\frac{s_0 p}{T}}. \end{aligned}$$

Therefore, by Assumption **??**, as $T \to \infty$,

$$\gamma \gg \max_{1 \leq i \leq 2d} \max_{1 \leq j \leq p_i} |\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}) - \sigma_j(\boldsymbol{\mathcal{A}}_{(i)})|.$$

Then, for any $j > r_i$, since $\sigma_j(\boldsymbol{\mathcal{A}}_{(i)}) = 0$, we have $\gamma \gg \sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)})$. Thus, for all $i = 1, \ldots, 2d$, $\sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(i)})$ will be truncated for all $j > r_i$. Meanwhile, by Assumption **??** and (**??**), we have

$\sigma_{r_i}(\widehat{\boldsymbol{\mathcal{A}}}_{(i)}) > \gamma$ for T sufficiently large, for all $i = 1, \dots, 2d$. Therefore, the rank selection consistency of the truncated estimator $\widetilde{\boldsymbol{\mathcal{A}}}$ can be established.

Denote the event $E = \{\text{rank}(\widetilde{\boldsymbol{\mathcal{A}}}_{(i)}) = r_i, \text{ for } i = 1, \dots, 2d\}$. For a generic tensor $\boldsymbol{\mathfrak{T}} \in \mathbb{R}^{p_1 \times \cdots \times p_{2d}}$, denote the sub-tensor $\boldsymbol{\mathfrak{T}}_{i_k = j}$, a $p_1 \times \cdots \times p_{k-1} \times 1 \times p_{k+1} \times \cdots \times p_{2d}$ tensor such that

$$(\boldsymbol{\mathfrak{T}}_{i_k=j})_{i_1 \dots i_{k-1} 1 i_{k+1} \dots i_{2d}} = \boldsymbol{\mathfrak{T}}_{i_1 \dots i_{k-1} j i_{k+1} \dots i_{2d}},$$

and sub-tensor $\boldsymbol{\mathfrak{T}}_{i_k > j}$, a $p_1 \times \cdots \times p_{k-1} \times (p_k - j) \times p_{k+1} \times \cdots \times p_{2d}$ tensor such that

$$(\boldsymbol{\mathfrak{T}}_{i_k>j})_{i_1 \dots i_{k-1} \ell i_{k+1} \dots i_{2d}} = \boldsymbol{\mathfrak{T}}_{i_1 \dots i_{k-1} (\ell+j) i_{k+1} \dots i_{2d}}.$$

Let the HOSVD of $\widehat{\boldsymbol{\mathcal{A}}}$ be $\widehat{\boldsymbol{\mathfrak{G}}} \times_{i=1}^{2d} \widehat{\boldsymbol{U}}_i$. By definition, $\widehat{\boldsymbol{\mathfrak{G}}}$ is a $p_1 \times \cdots \times p_{2d}$ all-orthogonal and sorted tensor such that

$$\|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k=1}\|_{\text{F}} \geq \|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k=2}\|_{\text{F}} \geq \cdots \geq \|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k=p_k}\|_{\text{F}},$$

for $k = 1, \dots, 2d$. On E, the truncation procedure is equivalent to truncating all the sub-tensors $\widehat{\boldsymbol{\mathfrak{G}}}_{i_k > r_k}$ to zeros. Thus, $\|\widehat{\boldsymbol{\mathcal{A}}} - \widetilde{\boldsymbol{\mathcal{A}}}\|_{\text{F}} = \|\widehat{\boldsymbol{\mathfrak{G}}} - \widetilde{\boldsymbol{\mathfrak{G}}}\|_{\text{F}}^2 \leq \sum_{k=1}^{2d} \|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k > r_k}\|_{\text{F}}^2$.

By the definition of HOSVD, $\|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k=j}\|_{\text{F}} = \sigma_j(\widehat{\boldsymbol{\mathfrak{G}}}_{(k)}) = \sigma_j(\widehat{\boldsymbol{\mathcal{A}}}_{(k)})$, and then

$$\begin{aligned}\|\widehat{\boldsymbol{\mathfrak{G}}}_{i_k > r_k}\|_{\text{F}}^2 = \sum_{i=r_{k+1}}^{p_k} \sigma_i^2(\widehat{\boldsymbol{\mathcal{A}}}_{(k)}) &= \sum_{i=r_{k+1}}^{p_k} [\sigma_i(\widehat{\boldsymbol{\mathcal{A}}}_{(k)}) - \sigma_i(\boldsymbol{\mathcal{A}}_{(k)})]^2 \\ &\leq \sum_{i=1}^{p_k} [\sigma_i(\widehat{\boldsymbol{\mathcal{A}}}_{(k)}) - \sigma_i(\boldsymbol{\mathcal{A}}_{(k)})]^2 \leq \|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}^2,\end{aligned}$$

where the last inequality follows from (**??**).

Finally, on the event E, $\|\widetilde{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} \leq \|\widetilde{\boldsymbol{\mathcal{A}}} - \widehat{\boldsymbol{\mathcal{A}}}\|_{\text{F}} + \|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} \leq (1 + \sqrt{2d})\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\text{F}}$, where d is fixed. Note that Theorem **??** implies the asymptotic rate $\|\widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}\|_{\text{F}} = O_p(\sqrt{s_0 p / T})$ and the first part of this proof shows that $\mathbb{P}(E) \to 1$, as $T \to \infty$. The proof is complete. □

B.3 Proofs of Lemmas ??–??

Proof of Lemma **??**. In this part, we focus on $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SSN}}$ and simplify it to $\widehat{\boldsymbol{\mathcal{A}}}$. The tuning parameter λ_{SSN} is simplified to λ. The proof can be readily extended to $\widehat{\boldsymbol{\mathcal{A}}}_{\text{SN}}$ and $\widehat{\boldsymbol{\mathcal{A}}}_{\text{MN}}$.

Note that the quadratic loss function can be rewritten as $\mathcal{L}_T(\boldsymbol{\mathcal{A}}) = T^{-1}\sum_{t=1}^T \|\boldsymbol{\mathcal{Y}}_t - \langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2 = T^{-1}\sum_{t=1}^T \|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2$, where $\boldsymbol{y}_t = \text{vec}(\boldsymbol{\mathcal{Y}}_t)$. By the optimality of the SSN estimator,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\boldsymbol{y}_t - \widehat{\boldsymbol{\mathcal{A}}}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 + \lambda\|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}} \leq \frac{1}{T}\sum_{t=1}^T \|\boldsymbol{y}_t - \boldsymbol{\mathcal{A}}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 + \lambda\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} \\
\Rightarrow\ &\frac{1}{T}\sum_{t=1}^T \|\boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 \leq \frac{2}{T}\sum_{t=1}^T \langle \boldsymbol{e}_t, \boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\rangle + \lambda(\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} - \|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}}) \\
\Rightarrow\ &\frac{1}{T}\sum_{t=1}^T \|\boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 \leq 2\left\langle T^{-1}\sum_{t=1}^T \boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t, \boldsymbol{\Delta}\right\rangle + \lambda(\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} - \|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}}) \\
\Rightarrow\ &\frac{1}{T}\sum_{t=1}^T \|\boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 \leq 2\|\boldsymbol{\Delta}\|_{\text{SSN}}\left\|T^{-1}\sum_{t=1}^T \boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\right\|_{\text{SSN}^*} + \lambda(\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} - \|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}}),
\end{aligned}$$

where $\|\cdot\|_{\text{SSN}^*}$ refers to the dual norm of the SSN norm.

By triangle inequality and decomposability, we have

$$\begin{aligned}
&\|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}} - \|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} = \|\boldsymbol{\mathcal{A}} + \boldsymbol{\Delta}\|_{\text{SSN}} - \|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} = \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{[I_k]} + \boldsymbol{\Delta}_{[I_k]}\|_* - \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{[I_k]}\|_* \\
=&\sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{\mathcal{M}}^{(k)} + \boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)} + \boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)} + \boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* - \sum_{k=1}^{2^{d-1}} \|\boldsymbol{\mathcal{A}}_{[I_k]}\|_* \\
\geq&\sum_{k=1}^{2^{d-1}} \left[\|\boldsymbol{\mathcal{A}}_{\mathcal{M}}^{(k)} + \boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* - \|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)} + \boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* - \|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)}\|_* - \|\boldsymbol{\mathcal{A}}_{\mathcal{M}}^{(k)}\|_*\right] \\
\geq&\sum_{k=1}^{2^{d-1}} \left[\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* - 2\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)}\|_* - \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_*\right].
\end{aligned}$$

If $\lambda \geq 4\|T^{-1}\sum_{t=1}^T \boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\|_{\text{SSN}^*}$, we have

$$\begin{aligned}
0 \leq \frac{1}{T}\sum_{t=1}^T \|\boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2 &\leq \frac{\lambda}{2}\|\boldsymbol{\Delta}\|_{\text{SSN}} - \lambda(\|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}} - \|\boldsymbol{\mathcal{A}}\|_{\text{SSN}}) \\
&\leq \frac{\lambda}{2}\sum_{k=1}^{2^{d-1}} \left[\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* + \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* - 2\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_* + 4\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)}\|_* + 2\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_*\right] \\
&= \frac{\lambda}{2}\sum_{k=1}^{2^{d-1}} \left[3\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* + 4\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^\perp}^{(k)}\|_* - \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp}^{(k)}\|_*\right].
\end{aligned}$$

Hence, the error $\boldsymbol{\Delta}$ lies in the restricted error set $\mathbb{C}_{\text{SSN}}(\overline{\mathcal{M}})$. □

Proof of Lemma **??**. Similar to Lemma **??**, we focus on the SSN estimator, and the results for SN and MN estimators can be extended in a similar way.

Note that $T^{-1}\sum_{t=1}^{T}\|\langle\boldsymbol{\Delta},\boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2 = T^{-1}\sum_{t=1}^{T}\|\boldsymbol{\Delta}_{[S_2]}\boldsymbol{y}_{t-1}\|_2^2$. Following the proof of Lemma **??**, $\boldsymbol{\Delta}\in\mathbb{C}_{\text{SSN}}(\overline{\mathcal{M}})$ and

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\langle\boldsymbol{\Delta},\boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2 &\leq \frac{\lambda}{2}\|\boldsymbol{\Delta}\|_{\text{SSN}} + \lambda(\|\boldsymbol{\mathcal{A}}\|_{\text{SSN}} - \|\widehat{\boldsymbol{\mathcal{A}}}\|_{\text{SSN}}) \leq \frac{3\lambda}{2}\|\boldsymbol{\Delta}\|_{\text{SSN}}\\
&= \frac{3\lambda}{2}\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\Delta}_{[I_k]}\|_* \leq \frac{3\lambda}{2}\sum_{k=1}^{2^{d-1}}\left(\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* + \|\boldsymbol{\Delta}_{\overline{\mathcal{M}}^{\perp}}^{(k)}\|_*\right)\\
&\leq 6\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_* + 6\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*\\
&\leq 6\lambda\sum_{k=1}^{2^{d-1}}\sqrt{2s_k}\|\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}\|_{\text{F}} + 6\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*\\
&\lesssim \lambda\sum_{k=1}^{2^{d-1}}\sqrt{2s_k}\|\boldsymbol{\Delta}\|_{\text{F}} + 6\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*\end{aligned}$$

where the last inequality stems from the fact that $\boldsymbol{\Delta}_{\overline{\mathcal{M}}}^{(k)}$ has a matrix rank at most $2s_k$, similar to Lemma 1 in **?**.

As the RSC condition holds with the parameter α_{RSC} and restricted error set $\mathbb{C}_{\text{SSN}}(\overline{\mathcal{M}})$,

$$\alpha_{\text{RSC}}\|\boldsymbol{\Delta}\|_{\text{F}}^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\langle\boldsymbol{\Delta},\boldsymbol{\mathcal{Y}}_{t-1}\rangle\|_{\text{F}}^2 \lesssim \lambda\sum_{k=1}^{2^{d-1}}\sqrt{s_k}\|\boldsymbol{\Delta}\|_{\text{F}} + \lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*.$$

Thus, by Cauchy inequality,

$$\|\boldsymbol{\Delta}\|_{\text{F}}^2 \lesssim \frac{\lambda^2(\sum_{k=1}^{2^{d-1}}\sqrt{s_k})^2}{\alpha_{\text{RSC}}^2} + \frac{\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*}{\alpha_{\text{RSC}}} \lesssim \frac{\lambda^2 2^{d-1}\sum_{k=1}^{2^{d-1}}s_k}{\alpha_{\text{RSC}}^2} + \frac{\lambda\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_*}{\alpha_{\text{RSC}}}.$$

Consider any threhold $\tau_k \geq 0$ and define the thresholded subspace $\mathcal{M}^{(k)}$ corresponding to the column and row spaces spanned by the first $r^{(k)}$ singular vectors of $\boldsymbol{\mathcal{A}}_{[I_k]}$ where $\sigma_1(\boldsymbol{\mathcal{A}}_{[I_k]}) \geq \cdots \geq \sigma_{r^{(k)}}(\boldsymbol{\mathcal{A}}_{[I_k]}) > \tau_k \geq \sigma_{r^{(k)}+1}(\boldsymbol{\mathcal{A}}_{[I_k]})$. By the definition of $\mathbb{B}_q(s_q^{(k)};p,p)$, we have $s_q^{(k)} \geq r^{(k)}\cdot\tau_k^q$ and thus $r^{(k)} \leq s_q^{(k)}\cdot\tau_k^{-q}$.

Then, the approximation error can be bounded by

$$\|\boldsymbol{\mathcal{A}}_{\mathcal{M}^{\perp}}^{(k)}\|_* = \sum_{r=r^{(k)}+1}^{p}\sigma_r(\boldsymbol{\mathcal{A}}_{[I_k]}) = \sum_{r=r^{(k)}+1}^{p}\sigma_r^q(\boldsymbol{\mathcal{A}}_{[I_k]})\cdot\sigma_r^{1-q}(\boldsymbol{\mathcal{A}}_{[I_k]}) \leq s_q^{(k)}\cdot\tau_k^{1-q}.$$

The estimation error can be bounded by

$$\|\boldsymbol{\Delta}\|_{\rm F}^2 \lesssim \frac{\lambda^2 2^{d-1}\sum_{k=1}^{2^{d-1}} s_q^{(k)}\cdot\tau_k^{-q}}{\alpha_{\rm RSC}^2} + \frac{\lambda\sum_{k=1}^{2^{d-1}} s_q^{(k)}\cdot\tau_k^{1-q}}{\alpha_{\rm RSC}}.$$

Setting each $\tau_k \asymp \alpha_{\rm RSC}^{-1}(q/(1-q))2^{d-1}\lambda$, the upper bound can be minimized to

$$\|\boldsymbol{\Delta}\|_{\rm F}^2 \lesssim 2^{1-d}\sum_{k=1}^{2^{d-1}} s_q^{(k)}\left(\frac{\lambda\cdot 2^{d-1}}{\alpha_{\rm RSC}}\right)^{2-q}.$$

The proof is complete. □

Proof of Lemma **??**. First, we derive an upper bound of the dual norm of the SSN norm. By definition, for any tensor $\boldsymbol{\mathcal{A}}$ and collection of index sets $\mathbb{I}=\{I_1,\dots,I_{2^{d-1}}\}$, the SSN norm is

$$\|\boldsymbol{\mathcal{A}}\|_{\rm SSN} = \sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{[I_k]}\|_*,$$

and its dual norm is $\|\boldsymbol{\mathcal{A}}\|_{\rm SSN^*} := \sup\langle\boldsymbol{\mathcal{W}},\boldsymbol{\mathcal{A}}\rangle$ such that $\|\boldsymbol{\mathcal{W}}\|_{\rm SSN}\le 1$. By a method similar to that in **?**, it can be shown that

$$\|\boldsymbol{\mathcal{A}}\|_{\rm SSN^*} = \inf_{\sum_{k=1}^{2^{d-1}}\boldsymbol{\mathcal{X}}_k=\boldsymbol{\mathcal{A}}}\ \max_{k=1,\dots,2^{d-1}}\|(\boldsymbol{\mathcal{X}}_k)_{[I_k]}\|_{\rm op}.$$

Then, we can take $\boldsymbol{\mathcal{X}}_k = (\sum_{k=1}^{2^{d-1}}1/c_k)^{-1}(\boldsymbol{\mathcal{A}}/c_k)$, where $c_k=\|\boldsymbol{\mathcal{A}}_{[I_k]}\|_{\rm op}$, and apply Jensen's inequality so that we have

$$\|\boldsymbol{\mathcal{A}}\|_{\rm SSN^*}\le 2^{-2(d-1)}\sum_{k=1}^{2^{d-1}}\|\boldsymbol{\mathcal{A}}_{[I_k]}\|_{\rm op}.$$

Hence, we have

$$\left\|\frac{1}{T}\sum_{t=1}^T\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\right\|_{\rm SSN^*}\le\frac{1}{2^{2(d-1)}}\sum_{k=1}^{2^{d-1}}\left\|\frac{1}{T}\sum_{t=1}^T(\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t)_{[I_k]}\right\|_{\rm op}.$$

In other words, the dual norm of the SSN norm can be upper bounded by the sum of the scaled matrix operator norms of different matricizations of the tensor $T^{-1}\sum_{t=1}^T\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t$.

All of the square matricizations based on I_k lead to a square p-by-p matrix. Therefore, by the deviation bound in Lemma **??**, we can take a union bound such that

$$\left\|\frac{1}{T}\sum_{t=1}^T\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t\right\|_{\rm SSN^*}\le\frac{C\kappa^2M_1}{2^{d-1}}\sqrt{\frac{p}{T}}$$

with probability at least $1 - \exp[-C(p-d)]$.

Next, for the SN estimator, we can obtain a similar upper bound of the dual norm of the SN norm. The SN norm is defined as

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

and its dual norm has the equivalent form

$$\|\boldsymbol{\mathcal{A}}\|_{\text{SN}^*} = \inf_{\sum_{i=1}^{2d} \boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{A}}} \max_{i=1,\dots,2d} \|(\boldsymbol{\mathcal{Y}}_i)_{(i)}\|_{\text{op}}.$$

Then, we can obtain an upper bound,

$$\|\boldsymbol{\mathcal{A}}\|_{\text{SN}^*} \leq \frac{1}{(2d)^2} \sum_{i=1}^{2d} \|\boldsymbol{\mathcal{A}}_{(i)}\|_{\text{op}}.$$

Then, for each one-mode matricization, we have the deviation bound. Then, we can take a union bound such that

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t \right\|_{\text{SN}^*} \leq \frac{C\kappa^2 M_1}{(2d)^2} \sum_{i=1}^{2d} \sqrt{\frac{p_{-i} p}{T}},$$

with probability at least $1 - 2d \exp[-Cp]$.

Finally, the MN estimator uses a special case of square matricization, and the upper bound for the MN estimator can be obtained by Lemma **??**. □

Proof of Lemma **??**. For any $\boldsymbol{M} \in \mathbb{R}^{m \times p}$, denote $R_T(\boldsymbol{M}) = \sum_{t=0}^{T-1} \|\boldsymbol{M}\boldsymbol{y}_t\|_2^2$. Note that $R_T(\boldsymbol{\Delta}_{[S_2]}) \geq \mathbb{E}R_T(\boldsymbol{\Delta}_{[S_2]}) - \sup_{\boldsymbol{\Delta}} |R_T(\boldsymbol{\Delta}_{[S_2]}) - \mathbb{E}R_T(\boldsymbol{\Delta}_{[S_2]})|$. Following the proof of Lemma **??**, $\mathbb{E}R_T(\boldsymbol{\Delta}_{[S_2]}) = \|(\boldsymbol{I}_T \otimes \boldsymbol{\Delta}_{[S_2]})\boldsymbol{P}\boldsymbol{D}\|_{\text{F}}^2 \geq T\|\boldsymbol{\Delta}\|_{\text{F}}^2 \cdot \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)$.

Similar to Lemma **??**, for any $\boldsymbol{v} \in \mathbb{S}^{p-1}$ and any $t > 0$,

$$\begin{aligned}&\mathbb{P}[|R_T(\boldsymbol{v}^\top) - \mathbb{E}R_T(\boldsymbol{v}^\top)| \geq t] \\ \leq& 2\exp\left(-\min\left(\frac{t^2}{\kappa^4 T \lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}, \frac{t}{\kappa^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)}\right)\right).\end{aligned}$$

Considering an ϵ-covering net of $\mathbb{S}^{p-1}$, by Lemma **??**, we can easily construct the union bound for $T \gtrsim p$,

$$\begin{aligned}&\mathbb{P}\left[\sup_{v \in \mathbb{S}^{p-1}} |R_T(\boldsymbol{v}^\top) - \mathbb{E}R_T(\boldsymbol{v}^\top)| \geq t\right] \\ \leq& C\exp\left(p - \min\left(\frac{t^2}{\kappa^4 T \lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}, \frac{t}{\kappa^2 \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)}\right)\right),\end{aligned}$$

Letting $t = \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)/2$, for $T \gtrsim M_2^{-2}\max(\kappa^4, \kappa^2)p$, we have

$$\mathbb{P}[|R_T(\boldsymbol{v}^\top) - \mathbb{E}R_T(\boldsymbol{v}^\top)| \geq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)/2] \leq 2\exp(-CM_2^2\min(\kappa^{-4}, \kappa^{-2})T),$$

where $M_2 = [\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)]/[\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)]$.

Therefore, with probability at least $1 - 2\exp(-CM_2^2\min(\kappa^{-4}, \kappa^{-2})T)$,

$$R_T(\boldsymbol{\Delta}_{[S_2]}) \geq \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2.$$

Finally, since $\boldsymbol{P}$ is related to the VMA(∞) process, by the spectral measure of ARMA process discussed in **?**, we may replace $\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)$ and $\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)$ with $1/\mu_{\min}(\mathcal{A})$ and $1/\mu_{\max}(\mathcal{A})$, respectively. □

B.4 Three Auxiliary Lemmas

Three auxiliary lemmas used in the proofs of Lemmas **??** and **??** are presented below.

Lemma B.5 (Deviation bound on different matricizations). *For any index set $I \subset \{1, 2, \ldots, 2d\}$, denote $q = \prod_{i=1, i\in I}^{2d} p_i$ and $q' = \prod_{i=1, i\notin I}^{2d} p_i$. If $T \gtrsim (q + q')$, with probability at least $1 - \exp[-C(q + q')]$,*

$$\left\| \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]} \right\|_{\mathrm{op}} < C\kappa^2 M_1\sqrt{(q + q')/T}.$$

where $M_1 = \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})/\mu_{\min}^{1/2}(\mathcal{A})$.

Proof. For any index set $I \subset \{1, 2, \ldots, 2d\}$ and $2d$th-mode tensor $\boldsymbol{\mathcal{T}}$, denote the inverse operation of the multi-mode matricization $\boldsymbol{T} = \boldsymbol{\mathcal{T}}_{[I]}$ by $\boldsymbol{T}^{[I]} = \boldsymbol{\mathcal{T}}$. Denote $\mathcal{W}(r; q, q') = \{\boldsymbol{W} \in \mathbb{R}^{q\times q'} : \mathrm{rank}(\boldsymbol{W}) = r,\ \|\boldsymbol{W}\|_{\mathrm{F}} = 1\}$.

By definition, $\|T^{-1}\sum_{t=1}^{T}(\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}\|_{\mathrm{op}} = \sup_{\boldsymbol{W}\in\mathcal{W}(1;q,q')}\langle T^{-1}\sum_{t=1}^{T}(\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W}\rangle = \sup_{\boldsymbol{W}\in\mathcal{W}(1;q,q')}\langle T^{-1}\sum_{t=1}^{T}\mathrm{vec}(\boldsymbol{\mathcal{E}}_t)\mathrm{vec}(\boldsymbol{\mathcal{Y}}_{t-1})^\top, (\boldsymbol{W}^{[I]})_{[S_1]}^\top\rangle$.

For an arbitrary matrix $\boldsymbol{W} \in \mathbb{R}^{q\times q'}$ such that $\|\boldsymbol{W}\|_{\mathrm{F}} = 1$, denote $\boldsymbol{M} = (\boldsymbol{W}^{[I]})_{[S_1]}^\top$. Then, one can easily check that $\langle(\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W}\rangle = \langle\boldsymbol{e}_t, \boldsymbol{M}\boldsymbol{y}_{t-1}\rangle$.

For a fixed $\boldsymbol{M}$, denote $S_t(\boldsymbol{M}) = \sum_{s=1}^{t}\langle \boldsymbol{e}_s, \boldsymbol{M}\boldsymbol{y}_{s-1}\rangle$ and $R_t(\boldsymbol{M}) = \sum_{s=0}^{t-1}\|\boldsymbol{M}\boldsymbol{y}_s\|_2^2$, for $1 \leq t \leq T$. By the standard Chernoff argument, for any $\alpha > 0$, $\beta > 0$ and $c > 0$,

$$\begin{aligned}
&\mathbb{P}[\{S_T(\boldsymbol{M}) \geq \alpha\} \cap \{R_T(\boldsymbol{M}) \leq \beta\}]\\
&= \inf_{m>0} \mathbb{P}[\{\exp(mS_T(\boldsymbol{M})) \geq \exp(m\alpha)\} \cap \{R_T(\boldsymbol{M}) \leq \beta\}]\\
&= \inf_{m>0} \mathbb{P}[\exp(mS_T(\boldsymbol{M}))\mathbb{I}(R_T(\boldsymbol{M}) \leq \beta) \geq \exp(m\alpha)]\\
&\leq \inf_{m>0} \exp(-m\alpha)\mathbb{E}[\exp(mS_T(\boldsymbol{M}))\mathbb{I}(R_T(\boldsymbol{M}) \leq \beta)]\\
&= \inf_{m>0} \exp(-m\alpha + cm^2\beta)\mathbb{E}[\exp(mS_T(\boldsymbol{M}) - cm^2\beta)\mathbb{I}(R_T(\boldsymbol{M}) \leq \beta)]\\
&\leq \inf_{m>0} \exp(-m\alpha + cm^2\beta)\mathbb{E}[\exp(mS_T(\boldsymbol{M}) - cm^2R_T(\boldsymbol{M}))].
\end{aligned}$$

By the tower rule, we have

$$\begin{aligned}
&\mathbb{E}[\exp(mS_T(\boldsymbol{M}) - cm^2R_T(\boldsymbol{M}))]\\
=&\mathbb{E}[\mathbb{E}[\exp(mS_T(\boldsymbol{M}) - cm^2R_T(\boldsymbol{M}))]|\mathcal{F}_{T-1}]\\
=&\mathbb{E}[\exp(mS_{T-1}(\boldsymbol{M}) - cm^2R_{T-1}(\boldsymbol{M}))\mathbb{E}[\exp(m\langle \boldsymbol{e}_T, \boldsymbol{M}\boldsymbol{y}_{T-1}\rangle - cm^2\|\boldsymbol{M}\boldsymbol{y}_T\|_2^2)|\mathcal{F}_{T-1}]].
\end{aligned}$$

Since $\langle \boldsymbol{e}_T, \boldsymbol{M}\boldsymbol{y}_{T-1}\rangle = \langle \boldsymbol{\xi}_T, \boldsymbol{\Sigma}_{\boldsymbol{e}}^{1/2}\boldsymbol{M}\boldsymbol{y}_{T-1}\rangle$, one can easily check that $\langle \boldsymbol{e}_T, \boldsymbol{M}\boldsymbol{y}_{T-1}\rangle$ is a $\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\|\boldsymbol{M}\boldsymbol{y}_{T-1}\|_2^2$-sub-Gaussian random variable. In other words, $\mathbb{E}[\exp(m\langle \boldsymbol{e}_T, \boldsymbol{M}\boldsymbol{y}_{T-1}\rangle)] \leq \exp(m^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\|\boldsymbol{M}\boldsymbol{y}_{T-1}\|_2^2/2)$. Thus, letting $c = \kappa\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})/2$, we have

$$\begin{aligned}
&\mathbb{E}[\exp(mS_T(\boldsymbol{M}) - m^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})R_T(\boldsymbol{M})/2)]\\
\leq&\mathbb{E}[\exp(mS_{T-1}(\boldsymbol{M}) - m^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})R_{T-1}(\boldsymbol{M})/2)]\\
\leq&\cdots \leq \mathbb{E}[\exp(mS_1(\boldsymbol{M}) - m^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})R_1(\boldsymbol{M})/2)] \leq 1.
\end{aligned}$$

Hence, we have that, for any $\alpha > 0$ and $\beta > 0$,

$$\begin{aligned}
&\mathbb{P}[\{S_T(\boldsymbol{M}) \geq \alpha\} \cap \{R_T(\boldsymbol{M}) \leq \beta\}]\\
&\leq \inf_{m>0} \exp(-m\alpha + m^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\beta/2)\\
&= \exp\left(-\frac{\alpha^2}{2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\beta}\right).
\end{aligned}$$

By Lemma **??**, we have that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}[|R_T(\boldsymbol{M}) - \mathbb{E}R_T(\boldsymbol{M})| \geq t]\\
\leq&2\exp\left(-\min\left(\frac{t^2}{\kappa^4T\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}, \frac{t}{\kappa^2\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}\right)\right).
\end{aligned}$$

In addition, $\mathbb{E}R_T(\boldsymbol{M}) = \operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{M}}) = \|(\boldsymbol{I}_T \otimes \boldsymbol{M})\boldsymbol{P}\boldsymbol{D}\|_{\mathrm{F}}^2 \leq T \cdot \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)$. Letting $t = C\kappa^2 T\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)$, we have

$$\mathbb{P}[R_T(\boldsymbol{M}) \geq C\kappa^2 T\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)] \leq 2\exp(-CT).$$

Next, consider a ϵ-net $\overline{\mathcal{W}}(1;q,q')$ for $\mathcal{W}(1;q,q')$. For any matrix $\boldsymbol{W} \in \mathcal{W}(1;q,q')$, there exist a matrix $\overline{\boldsymbol{W}} \in \overline{\mathcal{W}}(1;q,q')$ such that $\|\boldsymbol{W} - \overline{\boldsymbol{W}}\|_{\mathrm{F}} \leq \epsilon$. Since the rank of $\overline{\boldsymbol{\Delta}} = \boldsymbol{W} - \overline{\boldsymbol{W}}$ is at most 2, we can split the SVD of $\overline{\boldsymbol{\Delta}}$ into 2 parts, such that $\overline{\boldsymbol{\Delta}} = \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_2$, where $\operatorname{rank}(\boldsymbol{\Delta}_1) = \operatorname{rank}(\boldsymbol{\Delta}_2) = 1$ and $\langle \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2 \rangle = 0$. Then, for any matrix $\boldsymbol{N} \in \mathbb{R}^{q\times q'}$, we have

$$\langle \boldsymbol{N}, \boldsymbol{W} \rangle = \langle \boldsymbol{N}, \overline{\boldsymbol{W}} \rangle + \langle \boldsymbol{N}, \overline{\boldsymbol{\Delta}} \rangle = \langle \boldsymbol{N}, \overline{\boldsymbol{W}} \rangle + \sum_{i=1}^{2} \langle \boldsymbol{N}, \boldsymbol{\Delta}_i / \|\boldsymbol{\Delta}_i\|_{\mathrm{F}} \rangle \|\boldsymbol{\Delta}_i\|_{\mathrm{F}},$$

where $\boldsymbol{\Delta}_i/\|\boldsymbol{\Delta}_i\|_{\mathrm{F}} \in \mathcal{W}(1;q,q')$. Since $\|\overline{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Delta}_1\|_{\mathrm{F}}^2 + \|\boldsymbol{\Delta}_2\|_{\mathrm{F}}^2$, by Cauchy inequality, $\|\boldsymbol{\Delta}_1\|_{\mathrm{F}} + \|\boldsymbol{\Delta}_2\|_{\mathrm{F}} \leq \sqrt{2}\|\overline{\boldsymbol{\Delta}}\|_{\mathrm{F}} = \sqrt{2}\epsilon$. Hence, we have

$$\gamma := \sup_{\boldsymbol{W} \in \mathcal{W}(1;q,q')} \langle \boldsymbol{N}, \boldsymbol{W} \rangle \leq \max_{\overline{\boldsymbol{W}} \in \overline{\mathcal{W}}(1;q,q')} \langle \boldsymbol{N}, \overline{\boldsymbol{W}} \rangle + \sqrt{2}\gamma\epsilon.$$

In other words,

$$\sup_{\boldsymbol{W} \in \mathcal{W}(1;q,q')} \langle \boldsymbol{N}, \boldsymbol{W} \rangle \leq (1 - \sqrt{2}\epsilon)^{-1} \max_{\overline{\boldsymbol{W}} \in \overline{\mathcal{W}}(1;q,q')} \langle \boldsymbol{N}, \overline{\boldsymbol{W}} \rangle.$$

Therefore, we have that, for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{W} \in \mathcal{W}(1;q,q')} \left\langle \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W} \right\rangle \geq x\right] \\
\leq &\mathbb{P}\left[\max_{\boldsymbol{W} \in \overline{\mathcal{W}}(1;q,q')} \left\langle \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W} \right\rangle \geq (1-\sqrt{2}\epsilon)x\right] \\
\leq &|\overline{\mathcal{W}}(1;q,q')| \cdot \mathbb{P}\left[\left\langle \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W} \right\rangle \geq (1-\sqrt{2}\epsilon)x\right].
\end{aligned}$$

Note that by (**??**), for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\left\langle \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{\mathcal{Y}}_{t-1} \circ \boldsymbol{\mathcal{E}}_t)_{[I]}, \boldsymbol{W} \right\rangle \geq (1-\sqrt{2}\epsilon)x\right] \\
\leq &\mathbb{P}[\{S_T(\boldsymbol{M}) \geq T(1-\sqrt{2}\epsilon)x\} \cap \{R_T(\boldsymbol{M}) \leq C\kappa^2 T\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)\}] \\
+ &\mathbb{P}[R_T(\boldsymbol{M}) > C\kappa^2 T\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)] \\
\leq &\exp\left[-\frac{CTx^2}{\kappa^4\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)}\right] + 2\exp(-CT).
\end{aligned}$$

By Lemma 3.1 in **?**, for a ϵ-net for $\mathcal{W}(1;q,q')$, the covering number $|\overline{\mathcal{W}}(1;q,q')| \leq (9/\epsilon)^{q+q'}$. Combining (**??**), we have that, when $T \gtrsim q+q'$, for any $x>0$,

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{W}\in\mathcal{W}(1;q,q')}\left\langle\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t)_{[I]},\boldsymbol{W}\right\rangle\geq x\right]\\ \leq&\exp\left[(q+q')\log(9/\epsilon)-\frac{CTx^2}{\kappa^4\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)}\right]+2\exp[(q+q')\log(9/\epsilon)-CT].\end{aligned}$$

Taking $\epsilon=0.1$ and $x=C\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^{1/2}(\boldsymbol{P}\boldsymbol{P}^\top)\cdot\sqrt{(q+q')/T}$, we have

$$\begin{aligned}&\mathbb{P}\left[\sup_{\boldsymbol{W}\in\mathcal{W}(1;q,q')}\left\langle\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mathcal{Y}}_{t-1}\circ\boldsymbol{\mathcal{E}}_t)_{[I]},\boldsymbol{W}\right\rangle\geq C\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^{1/2}(\boldsymbol{P}\boldsymbol{P}^\top)\sqrt{\frac{q+q'}{T}}\right]\\ \leq&\exp[-C(q+q')].\end{aligned}$$

Finally, since $\boldsymbol{P}$ is related to the VMA(∞) process, by the spectral measure of ARMA process discussed in **?**, we may replace $\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)$ with $1/\mu_{\min}(\mathcal{A})$. □

Lemma B.6. *For any $\boldsymbol{M}\in\mathbb{R}^{p\times p}$ such that $\|\boldsymbol{M}\|_{\mathrm{F}}=1$, denote $R_T(\boldsymbol{M})=\sum_{t=0}^{T-1}\|\boldsymbol{M}\boldsymbol{y}_t\|_2^2$. Then, for any $t>0$,*

$$\begin{aligned}&\mathbb{P}[|R_T(\boldsymbol{M})-\mathbb{E}R_T(\boldsymbol{M})|\geq t]\\ \leq&2\exp\left(-\min\left(\frac{t^2}{\kappa^4T\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)},\frac{t}{\kappa^2\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}\right)\right),\end{aligned}$$

where $\boldsymbol{P}$ is defined as

$$\boldsymbol{P}=\begin{bmatrix}\boldsymbol{I}_p & \boldsymbol{A} & \boldsymbol{A}^2 & \boldsymbol{A}^3 & \dots & \boldsymbol{A}^{T-1} & \dots\\ \boldsymbol{O} & \boldsymbol{I}_p & \boldsymbol{A} & \boldsymbol{A}^2 & \dots & \boldsymbol{A}^{T-2} & \dots\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \dots\\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dots & \boldsymbol{I}_p & \dots\end{bmatrix}.$$

Proof. Denote by $\boldsymbol{y}=(\boldsymbol{y}_{T-1}^\top,\boldsymbol{y}_{T-2}^\top,\dots,\boldsymbol{y}_0^\top)^\top$, $\boldsymbol{e}=(\boldsymbol{e}_{T-1}^\top,\boldsymbol{e}_{T-2}^\top,\dots,\boldsymbol{e}_0^\top,\dots)^\top$, and $\boldsymbol{\xi}=(\boldsymbol{\xi}_{T-1}^\top,\boldsymbol{\xi}_{T-2}^\top,\dots,\boldsymbol{\xi}_0^\top,\dots)^\top$. Based on the moving average representation of VAR(1), we can rewrite $\boldsymbol{y}_t$ to a VMA(∞), $\boldsymbol{y}_t=\boldsymbol{e}_t+\boldsymbol{A}\boldsymbol{e}_{t-1}+\boldsymbol{A}^2\boldsymbol{e}_{t-2}+\boldsymbol{A}^3\boldsymbol{e}_{t-2}+\cdots$. Note that $R_T(\boldsymbol{M})=\boldsymbol{y}^\top(\boldsymbol{I}_T\otimes\boldsymbol{M}^\top\boldsymbol{M})\boldsymbol{y}=\boldsymbol{e}^\top\boldsymbol{P}^\top(\boldsymbol{I}_T\otimes\boldsymbol{M}^\top\boldsymbol{M})\boldsymbol{P}\boldsymbol{e}=\boldsymbol{\xi}^\top\boldsymbol{D}\boldsymbol{P}^\top(\boldsymbol{I}_T\otimes\boldsymbol{M}^\top\boldsymbol{M})\boldsymbol{P}\boldsymbol{D}\boldsymbol{\xi}:=\boldsymbol{\xi}^\top\boldsymbol{\Sigma}_M\boldsymbol{\xi}$, where P is defined in (**??**) and

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{e}}^{1/2} & \boldsymbol{O} & \boldsymbol{O} & \dots \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{\boldsymbol{e}}^{1/2} & \boldsymbol{O} & \dots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\boldsymbol{e}}^{1/2} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

By Hanson-Wright inequality, for any $t > 0$,

$$\mathbb{P}[|R_T(\boldsymbol{M}) - \mathbb{E}R_T(\boldsymbol{M})| \geq t] \leq 2\exp\left(-\min\left(\frac{t^2}{\kappa^4\|\boldsymbol{\Sigma}_{\boldsymbol{M}}\|_{\mathrm{F}}^2}, \frac{t}{\kappa^2\|\boldsymbol{\Sigma}_{\boldsymbol{M}}\|_{\mathrm{op}}}\right)\right).$$

As $\|\boldsymbol{M}\|_{\mathrm{F}} = 1$, by the submultiplicative property of the Frobenius norm and operator norm, we have $\|\boldsymbol{\Sigma}_{\boldsymbol{M}}\|_{\mathrm{F}}^2 \leq T \cdot \lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)$ and $\|\boldsymbol{\Sigma}_{\boldsymbol{M}}\|_{\mathrm{op}} \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)$. These imply that, for any $t > 0$,

$$\begin{aligned} &\mathbb{P}[|R_T(\boldsymbol{M}) - \mathbb{E}R_T(\boldsymbol{M})| \geq t] \\ \leq& 2\exp\left(-\min\left(\frac{t^2}{\kappa^4 T\lambda_{\max}^2(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}^2(\boldsymbol{P}\boldsymbol{P}^\top)}, \frac{t}{\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{e}})\lambda_{\max}(\boldsymbol{P}\boldsymbol{P}^\top)}\right)\right). \end{aligned}$$

The proof of this lemma is accomplished. □

Lemma B.7. *(Covering number of unit sphere) Let $\mathcal{N}$ be an ε-net of the unit sphere $\mathbb{S}^{p-1}$, where $\varepsilon \in (0, 1]$. Then,*

$$|\mathcal{N}| \leq \left(\frac{3}{\varepsilon}\right)^p.$$

Proof. This lemma follows directly from Corollary 4.2.13 of **?**. □

References

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8:53–80.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43:1535–1567.

Bi, X., Qu, A., and Shen, X. (2018). Multilayer tensor factorization with applications to recommender systems. *Annals of Statistics*, 46:3308–3333.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3:1–122.

Candes, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57:2342–2359.

Chen, E. Y. and Chen, R. (2019). Modeling dynamic transport network with matrix factor models: with an application to international trade flow. arXiv:1901.00769 [econ.EM].

Chen, E. Y., Tsay, R. S., and Chen, R. (2020a). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statitical Association*, 115:775–793.

Chen, H., Raskutti, G., and Yuan, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208.

Chen, R., Xiao, H., and Yang, D. (2020b). Autoregressive models for matrix-valued time series. *Journal of Econometrics*. To appear.

Chen, R., Yang, D., and Zhang, C.-H. (2020c). Factor models for high-dimensional tensor time series. arXiv:1905.07530v2 [stat.ME].

Davis, R. A., Zang, P., and Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25:1077–1096.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21:1253–1278.

Ding, S. and Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B*, 80:387–408.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116:1–22.

Frandsen, A. and Ge, R. (2019). Understanding composition of word embeddings via tensor decomposition. In *International Conference on Learning Representations*.

French, K. R. (2020). Data library: U.S. research returns data. Available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

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

Han, F., Lu, H., and Liu, H. (2015a). A direct estimation of high dimensional stationary vector autoregressions. *Journal of Machine Learning Research*, 16:3115–3150.

Han, F., Xu, S., and Liu, H. (2015b). Rate-optimal estimation of a high-dimensional semiparametric time series model. Preprint.

Han, R., Willett, R., and Zhang, A. (2020). An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Annals of Applied Statistics*, 9:1169–1193.

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

Lam, C., Yao, Q., et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics*, 40:694–726.

Li, X., Xu, D., Zhou, H., and Li, L. (2018). Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10:520–545.

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:208–220.

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *Quarterly Journal of Mathematics*, 11:50–59.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *International Conference on Machine Learning*, pages 73–81.

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

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538–557.

Raskutti, G., Yuan, M., and Chen, H. (2019). Convex regularization for high-dimensional multi-response tensor regression. *Annals of Statistics*, 47:1554–1584.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81:142–149.

Stock, J. H. and Watson, M. W. (2011). Dynamic factor models. In Clements, M. P. and Hendry, D. F., editors, *Oxford Handbook of Economic Forecasting*. Oxford University Press.

Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. (2011). Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 972–980.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge.

Walden, A. and Serroukh, A. (2002). Wavelet analysis of matrix–valued time–series. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 458:157–179.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208:231–248.

Wang, D., Zheng, Yao Lian, H., , and Li, G. (2020). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*. To appear.

Zheng, Y. and Cheng, G. (2020). Finite time analysis of vector autoregressive models under linear restrictions. *Biometrika*. To appear.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552.

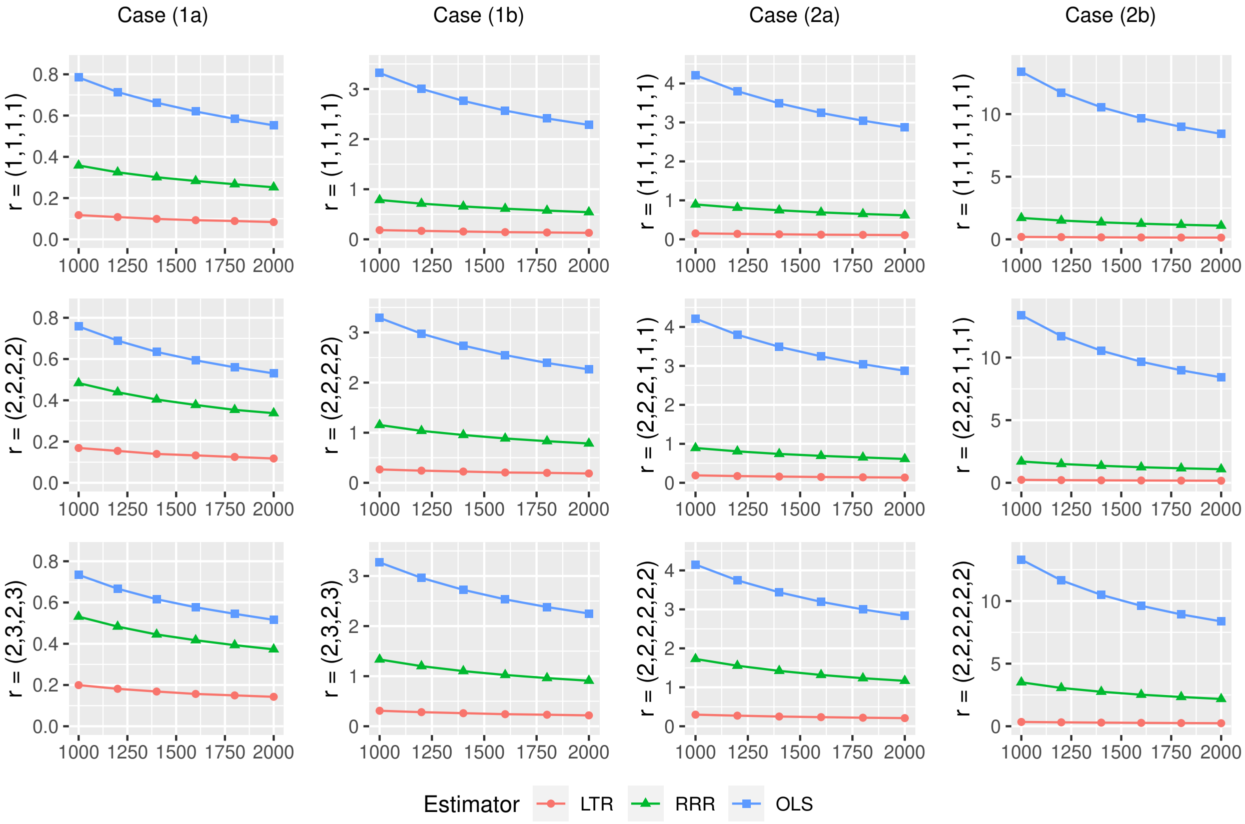

Figure 2: Average estimation error of LTR, OLS and RRR estimators for data generated with different d, p_i's and multilinear ranks.

Case	d	Fixed Parameter	Varying Parameter
(a)	2	$T = 500$, $\boldsymbol{r} = (2, 2, 2, 2)$	$p_1 = p_2 = 5, 7, 9, 10, 11$
(b)	2	$p_1 = p_2 = 8$, $\boldsymbol{r} = (2, 2, 2, 2)$	$T = 200, 400, 600, 800, 1000$
(c)	2	$p_1 = p_2 = 8$, $T = 500$	$\boldsymbol{r} = (1, 1, 1, 1), (1, 2, 1, 2)$, $(2, 2, 2, 2, 2), (2, 3, 2, 3), (3, 3, 3, 3)$
(d)	2	$p = 144$, $T = 1000$, $\boldsymbol{r} = (1, 1, 1, 1)$	$p_1 = 3, 4, 6, 8, 12$
(e)	3	$T = 1000$, $\boldsymbol{r} = (2, 2, 2, 2, 2, 2)$	$\boldsymbol{p} = (4, 4, 4), (4, 4, 5), (4, 5, 5), (5, 5, 5), (5, 5, 6)$
(f)	3	$\boldsymbol{p} = (5, 5, 5)$, $\boldsymbol{r} = (2, 2, 2, 2, 2, 2)$	$T = 600, 800, 1000, 1200, 1400$
(g)	3	$\boldsymbol{p} = (5, 5, 5)$, $T = 1000$	$\boldsymbol{r} = (1, 1, 1, 1, 1, 1), (1, 1, 2, 1, 1, 2)$ $(1, 2, 2, 1, 2, 2), (2, 2, 2, 2, 2, 2), (2, 2, 3, 2, 2, 3)$
(h)	3	$p = 144$, $T = 1000$, $\boldsymbol{r} = (1, 1, 1, 1, 1, 1)$	$\boldsymbol{p} = (2, 2, 36), (3, 3, 16)$, $(4, 4, 9), (3, 4, 12), (4, 6, 6)$

Table 2: Parameter setting for eight cases with different varying parameters, where $\boldsymbol{p} = (p_1, \ldots, p_d)$ and $\boldsymbol{r} = (r_1, \ldots, r_{2d})$.

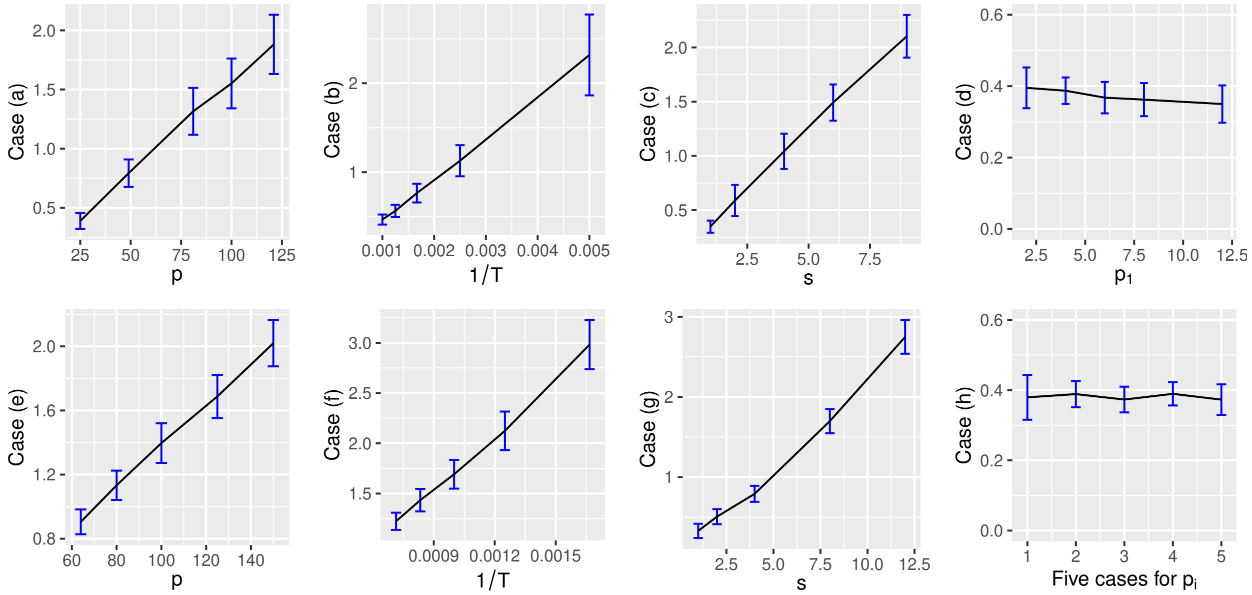

Figure 3: Average squared estimation error for the SSN estimator for eight cases with different varying parameters. The error bars in each panel represent ± one standard deviation.

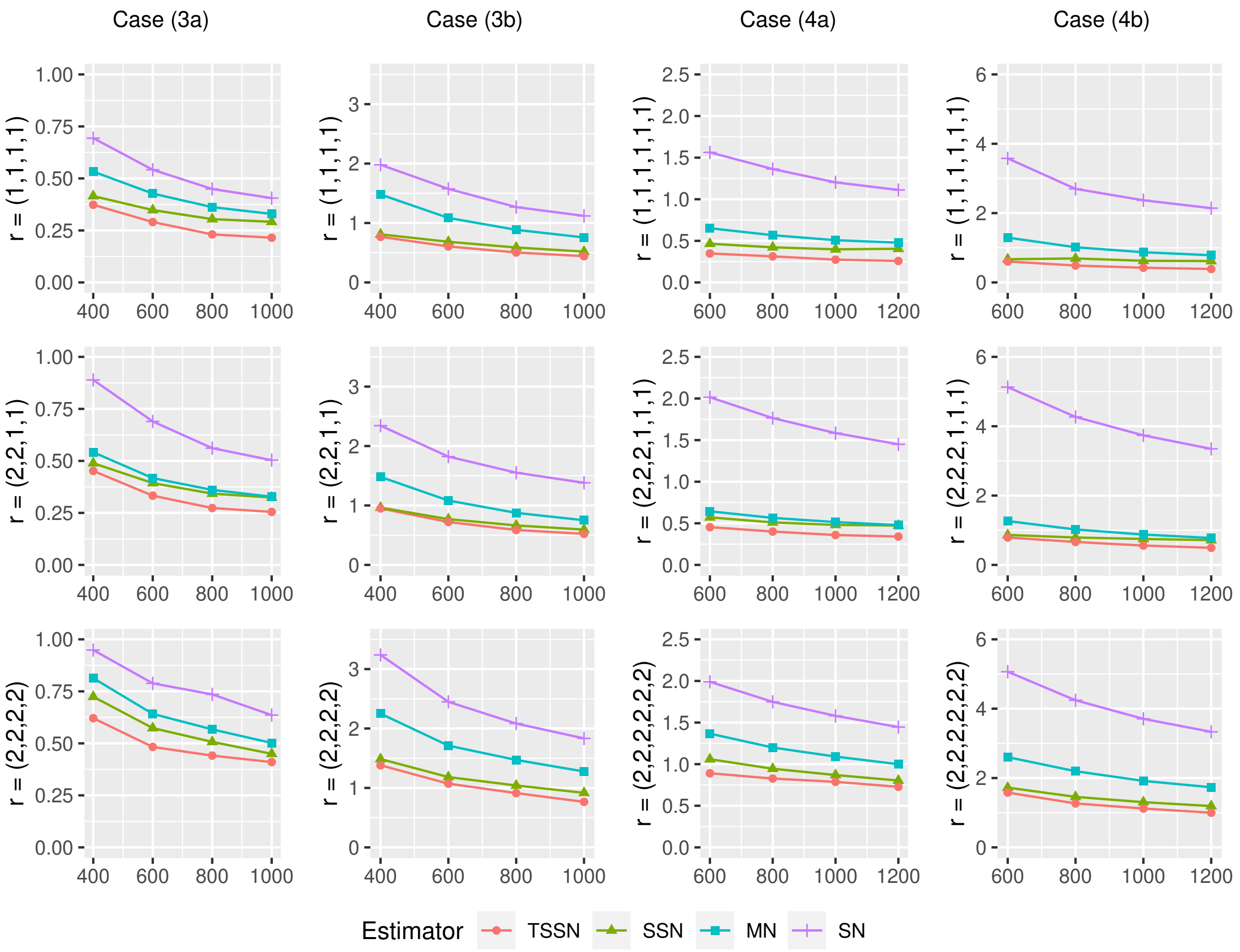

Figure 4: Average estimation error for TSSN, SSN, MN, and SN estimators for data generated with different d, p_i's and multilinear ranks.

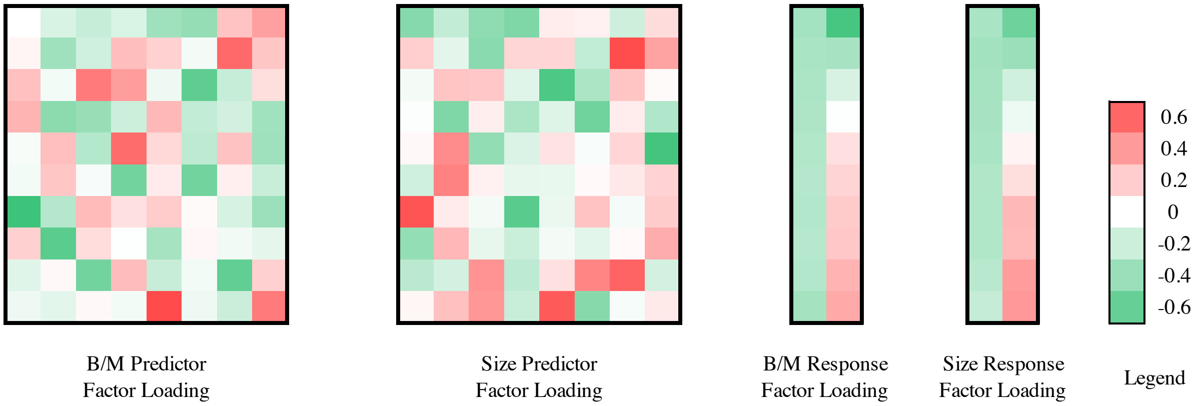

Figure 5: TSSN estimates of predictor and response factor matrices for 10×10 matrix-valued portfolio return series. From left to right: $\widetilde{\boldsymbol{U}}_1, \widetilde{\boldsymbol{U}}_2, \widetilde{\boldsymbol{U}}_3$ and $\widetilde{\boldsymbol{U}}_4$.

	Model	VAR	VFM	MAR	MFM	LRTAR		Best	Worst
						SSN	TSSN		
In-sample	ℓ_2 norm	**31.61**	35.85	34.12	35.86	33.18	33.38	VAR	MFM
	ℓ_0 norm	**8.09**	9.23	9.12	9.24	8.84	8.89	VAR	MFM
Out-of-sample	ℓ_2 norm	39.84	39.11	35.26	38.53	32.60	**31.47**	TSSN	VAR
	ℓ_∞ norm	11.98	12.30	11.47	11.00	9.53	**9.33**	TSSN	VFM

Table 3: Average in-sample forecasting error and out-of-sample rolling forecasting error for 10×10 matrix-valued portfolio return series. The best cases are marked in bold.

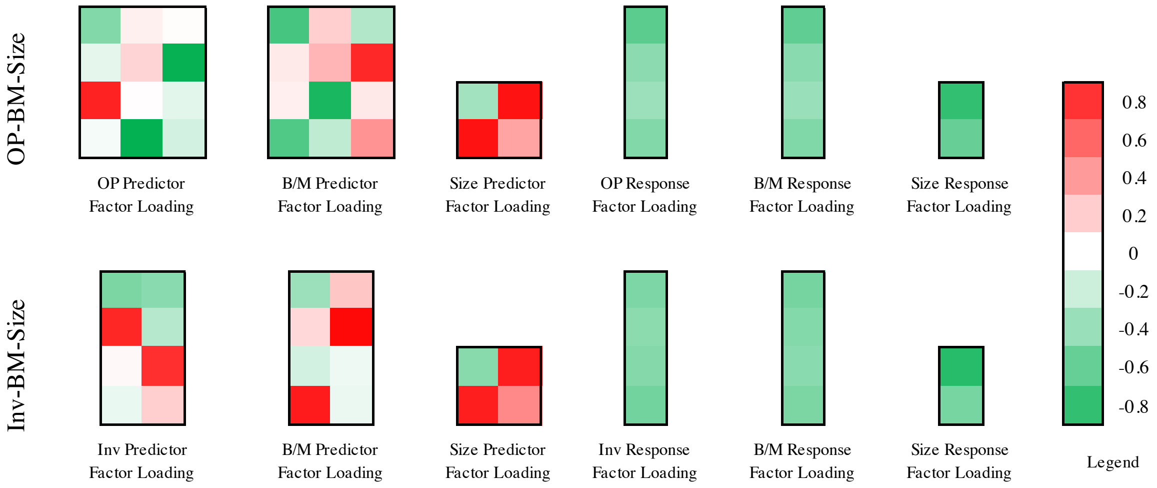

Figure 6: TSSN estimates of predictor and response factor matrices for $4\times4\times2$ tensor-valued portfolio return series. From left to right: $\widetilde{\boldsymbol{U}}_1, \widetilde{\boldsymbol{U}}_2, \widetilde{\boldsymbol{U}}_3$, $\widetilde{\boldsymbol{U}}_4$, $\widetilde{\boldsymbol{U}}_5$ and $\widetilde{\boldsymbol{U}}_6$.

	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 4: 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.