Heterogeneous Autoregressive Modeling with Flexible Cascade Structures

Huiling Yuan[1], Kexin Lu[2], Guodong Li[2], Alan T.K. Wan[3], and Yong Zhou[1]

[1] *East China Normal University,* [2] *University of Hong Kong, and* [3] *City University of Hong Kong*

Supplementary Material

This online Supplementary Material includes seven sections. Section S1 provides the tensor notations and Tucker decomposition. Section S2 gives the technical proofs of Theorems 1-2. Section S3 provides four useful lemmas. Lemma 1 establishes covering number and discretization of low-multilinear-rank tensors. Lemma 2 derives restricted strong convexity and smoothness. Lemma 3 derives the deviation bound, and they will be used in the proof of Theorem 1 and 2. Lemma 4 derives the contractive projection property, which is used in the proof of Theorem 2. Section S4 presents the technical proofs of Corollary 1 in Section 3. Section S5 gives two useful lemmas. Lemma 5 derives the restricted strong convexity of the estimated low-Tucker-rank linear form, Lemma 6 derives the deviation bound of the estimated low-Tucker-rank linear form, and they will be used in the proof of Corollary 1. Section S6 presents simulation results for the MLR-TT-HAR model, and Section S7 gives one Table for the selected ranks of the MLR-FT-HAR, MLR-TT-HAR and VHARI models in Real data analysis. Throughout this subsection, we will use C to represent generic positive numbers, whose value may vary from line to line.

S1 Tensor notations and Tucker decomposition

The section gives a brief introduction to tensor notations and Tucker decomposition, and a detailed review on tensor notations and operations can be referred to in (Kolda and Bader, 2009).

Tensors, also known as multidimensional arrays, are higher-order extensions of matrices, and a multidimensional array $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{d_1 \times \ldots \times d_K}$ is called a K-th-order tensor, where the order of a tensor is known as the dimension, way or mode. This paper concentrates on fourth-order tensors.

For a fourth-order tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{p_1 \times p_2 \times p_3 \times p_4}$, its element is denoted by $\boldsymbol{\mathcal{A}}_{ijkl}$ for $1 \leqslant i \leqslant p_1, 1 \leqslant j \leqslant p_2$, $1 \leqslant k \leqslant p_3$ and $1 \leqslant l \leqslant p_4$, and the Frobenius norm is defined as $\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}} = \left(\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \sum_{k=1}^{p_3} \sum_{l=1}^{p_4} X_{ijkl}^2\right)^{1/2}$.

Matricization, also known as unfolding or flattening, involves rearranging the elements of a higher-order tensor into a matrix format. This process treats the first mode of the tensor as the rows of the matrix, while collapsing all other modes into the columns. Specifically, the element at position (i_1, i_2, i_3, i_4) in the tensor $\boldsymbol{\mathcal{A}}$ corresponds to the (i_1, j)-th element in the matricized form $\boldsymbol{\mathcal{A}}_{(1)} \in \mathbb{R}^{p_1 \times p_2 p_3 p_4}$, where

$$j = 1 + \sum_{k=2}^{4} (i_k - 1) J_k \quad \text{with} \quad J_k = \prod_{\ell=2}^{k-1} p_\ell.$$

Similarly, mode-2, mode-3, and mode-4 matricizations can be defined. Matricizing tensors allows for establishing connections between matrix concepts and properties with those of tensors. The mode-1 multiplication, denoted as $\times_1$, between a tensor $\boldsymbol{\mathcal{A}}$

and a matrix $\mathbf{B} \in \mathbb{R}^{q_1 \times p_1}$, is defined as follows

$$\boldsymbol{\mathcal{A}} \times_1 \mathbf{B} = \left(\sum_{i=1}^{p_1} \boldsymbol{\mathcal{A}}_{ijkl} \mathbf{B}_{si} \right)_{1 \leqslant s \leqslant q_1, 1 \leqslant j \leqslant p_2, 1 \leqslant k \leqslant p_3, 1 \leqslant l \leqslant p_4}.$$

The mode-k multiplication, denoted as $\times_k$, where $k = 2, 3, 4$, can be defined similarly.

The multilinear ranks of a tensor $\boldsymbol{\mathcal{A}}$ is defined as (r_1, r_2, r_3, r_4), where

$$r_1 = \operatorname{rank}\left(\boldsymbol{\mathcal{A}}_{(1)}\right), \quad r_2 = \operatorname{rank}\left(\boldsymbol{\mathcal{A}}_{(2)}\right), \quad r_3 = \operatorname{rank}\left(\boldsymbol{\mathcal{A}}_{(3)}\right), \quad \text{and} \quad r_4 = \operatorname{rank}\left(\boldsymbol{\mathcal{A}}_{(4)}\right).$$

Accordingly, there exists a Tucker decomposition (Tucker, 1966; De Lathauwer et al., 2000):

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3 \times_4 \mathbf{U}_4 = \boldsymbol{\mathcal{G}} \times_{i=1}^{4} \mathbf{U}_i, \tag{S1.1}$$

where $\mathbf{U}_i \in \mathbb{R}^{p_i \times r_i}$ for $i = 1, 2, 3, 4$ represent the factor matrices, and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times r_3 \times r_4}$ denotes the core tensor. Alternatively, this decomposition can be represented as $\boldsymbol{\mathcal{A}} = [\![\boldsymbol{\mathcal{G}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]$.

Note that the Tucker decomposition is not unique, since

$$[\![\boldsymbol{\mathcal{G}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!] = [\![\boldsymbol{\mathcal{G}} \times_1 \mathbf{O}_1 \times_2 \mathbf{O}_2 \times_3 \mathbf{O}_3 \times_4 \mathbf{O}_4; \mathbf{U}_1\mathbf{O}_1^{-1}, \mathbf{U}_2\mathbf{O}_2^{-1}, \mathbf{U}_3\mathbf{O}_3^{-1}, \mathbf{U}_4\mathbf{O}_4^{-1}]\!].$$

for any invertible matrices $\mathbf{O}_i \in \mathbb{R}^{r_i \times r_i}$ with $1 \leqslant i \leqslant 4$. As a special Tucker decomposition, the higher-order singular value decomposition (HOSVD) is defined by choosing $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ and $\mathbf{U}_4$ as the tall matrices that consist of the top (r_1, r_2, r_3, r_4) left singular vectors of $\boldsymbol{\mathcal{A}}_{(1)}$, $\boldsymbol{\mathcal{A}}_{(2)}$, $\boldsymbol{\mathcal{A}}_{(3)}$ and $\boldsymbol{\mathcal{A}}_{(4)}$ respectively, and then $\boldsymbol{\mathcal{G}} = \boldsymbol{\mathcal{A}} \times_1 \mathbf{U}_1^\top \times_2 \mathbf{U}_2^\top \times_3 \mathbf{U}_3^\top \times_4 \mathbf{U}_4^\top$. As a result, factor matrices $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$ and $\mathbf{U}_4$ are all orthonormal, and $\boldsymbol{\mathcal{G}}$ possesses

the all-orthogonal property, i.e., for each $1 \leqslant j \leqslant 4$, the rows of $\mathcal{G}_{(j)}$ are pairwise orthogonal. Note that the HOSVD is still not unique unless we impose more constraints (Wang et al., 2022).

The four ranks, r_1, r_2, r_3 and r_4, are not equal in general. In particular, when $r_4 = p_4$, the multilinear ranks of $\boldsymbol{\mathcal{A}}$ are denoted by (r_1, r_2, r_3) instead, omitting the rank of mode-4 matricization. The multilinear ranks are also known as Tucker ranks, as they are closely related to the Tucker decomposition. There are many other tensor decomposition methods, such as CP decomposition (Kolda and Bader, 2009), and the ranks of a tensor can be defined in many different ways.

S2 Proofs of two Theorems

Proof of Theorem 1. For simplicity, denote the multilinear low-rank estimator $\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}}$ by $\widehat{\boldsymbol{\mathcal{A}}}$, and let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\mathcal{A}}} - \boldsymbol{\mathcal{A}}$, where $\boldsymbol{\mathcal{A}}$ is the true parameter tensor. The loss function has the form of

$$L(\boldsymbol{\mathcal{A}}) = \frac{1}{T}\sum_{n=1}^{T} ||\mathbf{y}_n - \boldsymbol{\mathcal{A}}_{(1)}\mathbf{x}_n||_2^2,$$

where $\mathbf{x}_n = (\mathbf{y}_{n-1}^\top, \ldots, \mathbf{y}_{n-SQ}^\top)^\top \in \mathbb{R}^{NSQ}$. Due to the optimality of the $\widehat{\boldsymbol{\mathcal{A}}}$, it holds that

$$\frac{1}{T}\sum_{n=1}^{T} \|\mathbf{y}_n - \widehat{\boldsymbol{\mathcal{A}}}_{(1)}\mathbf{x}_n\|_2^2 \leqslant \frac{1}{T}\sum_{n=1}^{T} \|\mathbf{y}_n - \boldsymbol{\mathcal{A}}_{(1)}\mathbf{x}_n\|_2^2,$$

which implies that

$$\frac{1}{T}\sum_{n=1}^{T} \|\boldsymbol{\Delta}_{(1)}\mathbf{x}_n\|_2^2 \leqslant \frac{2}{T}\sum_{n=1}^{T}\langle \boldsymbol{\varepsilon}_n, \boldsymbol{\Delta}_{(1)}\mathbf{x}_n\rangle \leqslant 2\langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta}\rangle, \tag{S2.1}$$

where $\mathbf{X}_n = (\mathbf{y}_{n-1}, \ldots, \mathbf{y}_{n-SQ}) \in \mathbb{R}^{N \times SQ}$, $\sum_{n=1}^{T} \langle \boldsymbol{\varepsilon}_n, \boldsymbol{\Delta}_{(1)} \mathbf{x}_n \rangle = \langle \sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \rangle$, and $\circ$ denotes the outer product.

Denote the set of tensors

$$\mathcal{S}(r_1, r_2, r_3, r_4) = \{\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N \times N \times S \times Q} : \|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}} = 1, \mathrm{rank}_i(\boldsymbol{\mathcal{A}}_i) \leqslant r_i, 1 \leqslant i \leqslant 4\}.$$

Note that the Tucker ranks of both $\widehat{\boldsymbol{\mathcal{A}}}$ and $\boldsymbol{\mathcal{A}}$ are (r_1, r_2, r_3, r_4), and hence the Tucker ranks of $\boldsymbol{\Delta}$ are at most $(2r_1, 2r_2, 2r_3, 2r_4)$. As a result, from (S2.1),

$$\frac{1}{T} \sum_{n=1}^{T} \|\boldsymbol{\Delta}_{(1)} \mathbf{x}_n\|_2^2 \leqslant 2\|\boldsymbol{\Delta}\|_{\mathrm{F}} \sup_{\boldsymbol{\Delta} \in \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)} \langle \frac{1}{T} \sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \rangle,$$

and we hence can derive the estimation error bound by applying Lemma 2 and Lemma 3. The prediction error bound can also be established from the above inequality, estimation error bound and Lemma 3. □

Proof of Theorem 2. For a fixed $1 \leqslant k \leqslant K$, define a linear space,

$$\mathcal{A} = \{\alpha_1 \widehat{\boldsymbol{\mathcal{A}}}_k + \alpha_2 \boldsymbol{\mathcal{A}}, \ \ \alpha_1, \alpha_2 \in \mathbb{R}\},$$

and denote by $(\boldsymbol{\mathcal{B}})_{\mathcal{A}}$ the projection of $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{N \times N \times S \times Q}$ onto the space $\mathcal{A}$, where the dependence of $\mathcal{A}$ on k is suppressed for simplicity. Since $\boldsymbol{\mathcal{A}} \in \boldsymbol{\Theta}(r_1, r_2, r_3, r_4)$ and $\widehat{\boldsymbol{\mathcal{A}}}_k \in \boldsymbol{\Theta}(r_1^{'}, r_2^{'}, r_3^{'}, r_4^{'})$, it holds that $\mathcal{A} \subset \boldsymbol{\Theta}(r_1 + r_1^{'}, r_2 + r_2^{'}, r_3 + r_3^{'}, r_4 + r_4^{'})$.

Note that $\widetilde{\boldsymbol{\mathcal{A}}}_k = \widehat{\boldsymbol{\mathcal{A}}}_{k-1} - \eta \nabla L(\widehat{\boldsymbol{\mathcal{A}}}_{k-1})$, and $r_i^{'} \geqslant \left(\sqrt[4]{1 + \frac{\kappa_L}{24\kappa_U}} - 1\right)^{-2} r_i$ with $1 \leqslant i \leqslant 4$. From Lemma 4, we have

$$\|\widehat{\boldsymbol{\mathcal{A}}}_k - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}} \leqslant [\prod_{i=1}^{4} (\sqrt{\frac{r_i}{r_i^{'}}} + 1) - 1] \|\boldsymbol{\mathcal{A}} - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}} \leqslant \frac{\kappa_L}{24\kappa_U} \|\boldsymbol{\mathcal{A}} - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}},$$

which, together with the fact that $1 < 1 + \frac{\kappa_L}{24\kappa_U} < 2$, implies that

$$\begin{aligned}\|\widehat{\boldsymbol{\mathcal{A}}}_k - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}} \leqslant & \|\widehat{\boldsymbol{\mathcal{A}}}_k - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}} + \|\boldsymbol{\mathcal{A}} - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}} \leqslant (1 + \frac{\kappa_L}{24\kappa_U})\|\boldsymbol{\mathcal{A}} - (\widetilde{\boldsymbol{\mathcal{A}}}_k)_{\mathcal{A}}\|_{\mathrm{F}} \\ \leqslant & (1 + \frac{\kappa_L}{24\kappa_U})\|(\boldsymbol{\mathcal{A}} - \widehat{\boldsymbol{\mathcal{A}}}_{k-1} - \eta[\nabla L(\boldsymbol{\mathcal{A}}) - \nabla L(\widehat{\boldsymbol{\mathcal{A}}}_{k-1})])_{\mathcal{A}}\|_{\mathrm{F}} + 2\eta\|(\nabla L(\boldsymbol{\mathcal{A}}))_{\mathcal{A}}\|_{\mathrm{F}} \\ := & A_1 + A_2.\end{aligned} \tag{S2.2}$$

We first handle the term of A_1. Let $\mathbf{H} = T^{-1}\sum_{n=1}^{T}(\mathbf{x}_n\mathbf{x}_n^{\top} \otimes \mathbf{I}_N)$, and it holds that $T^{-1}\sum_{n=1}^{T}\|\boldsymbol{\Delta}_{(1)}\mathbf{x}_n\|_2^2 = T^{-1}\sum_{n=1}^{T}\|(\mathbf{x}_n^{\top} \otimes \mathbf{I}_N)\mathrm{vec}(\boldsymbol{\Delta})\|_2^2 = \mathrm{vec}(\Delta)^{\top}\mathbf{H}\mathrm{vec}(\Delta)$. Then, from Lemma 2 and for all $\boldsymbol{\Delta} \in \boldsymbol{\Theta}(r_1 + r_1^{'}, r_2 + r_2^{'}, r_3 + r_3^{'}, r_4 + r_4^{'})$,

$$\frac{1}{8}\kappa_L\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leqslant \mathrm{vec}(\Delta)^{\top}\mathbf{H}\mathrm{vec}(\Delta) \leqslant \frac{8}{3}\kappa_{\mathrm{U}}\|\Delta\|_{\mathrm{F}}^2$$

with a probability at least

$$1 - \exp(-Cd_{\mathcal{M}}^{'}) - 2\exp(Cd_{\mathcal{M}}^{'} - CT(\kappa_L/\kappa_U)^2\min\{\kappa^{-2}, \kappa^{-4}\}). \tag{S2.3}$$

Note that $\eta = 2/3\kappa_U$, and $\mathbf{H}$ is the Hessian matrix of the loss function $L(\boldsymbol{\mathcal{B}})$ with respect to $\mathrm{vec}(\boldsymbol{\mathcal{B}})$. It holds that $\mathrm{vec}(\nabla L(\boldsymbol{\mathcal{A}}) - \nabla L(\widehat{\boldsymbol{\mathcal{A}}}_{k-1})) = \mathbf{H}\mathrm{vec}(\boldsymbol{\mathcal{A}} - \widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{k-1}})$, and

$$\begin{aligned}A_1 &= (1 + \frac{\kappa_L}{24\kappa_U})\|((\mathbf{I} - \eta\mathbf{H})\mathrm{vec}(\boldsymbol{\mathcal{A}} - \widehat{\boldsymbol{\mathcal{A}}}_{k-1}))_{\mathcal{A}}\|_2 \\ &\leqslant (1 + \frac{\kappa_L}{24\kappa_U})(1 - \frac{\kappa_L}{12\kappa_U})\|\widehat{\boldsymbol{\mathcal{A}}}_{k-1} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}} \\ &\leqslant (1 - \frac{\kappa_L}{24\kappa_U})\|\widehat{\boldsymbol{\mathcal{A}}}_{k-1} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}},\end{aligned} \tag{S2.4}$$

with the probability at (S2.3), where $(\mathrm{vec}(\boldsymbol{\mathcal{B}}))_{\mathcal{A}} = (\boldsymbol{\mathcal{B}})_{\mathcal{A}}$, and the first inequality is by Lemma 4 of Chen et al. (2019).

We next handle the term of A_2, and by Lemma 5 in Chen et al. (2019),

$$A_2 \leqslant \frac{32}{\kappa_L}\|(\nabla L(\boldsymbol{\mathcal{A}}))_{\mathcal{A}}\|_{\mathrm{F}} = \frac{32}{\kappa_L}\sup_{\boldsymbol{\mathcal{S}}\in\mathcal{A},\|\boldsymbol{\mathcal{S}}\|_{\mathrm{F}}=1}\langle\nabla L(\boldsymbol{\mathcal{A}}),\boldsymbol{\mathcal{S}}\rangle \leqslant \frac{32}{\kappa_L}\xi,$$

where $\mathcal{S}(r_1+r_1^{'},r_2+r_2^{'},r_3+r_3^{'},r_4+r_4^{'}) = \boldsymbol{\Theta}(r_1+r_1^{'},r_2+r_2^{'},r_3+r_3^{'},r_4+r_4^{'})\bigcap\{\|\boldsymbol{\mathcal{S}}\|_{\mathrm{F}}=1\}$ and

$$\begin{aligned}\xi &= \sup_{\boldsymbol{\mathcal{S}}\in\mathcal{S}(r_1+r_1^{'},r_2+r_2^{'},r_3+r_3^{'},r_4+r_4^{'})}\langle\nabla L(\boldsymbol{\mathcal{A}}),\boldsymbol{\mathcal{S}}\rangle \\ &= \sup_{\boldsymbol{\mathcal{S}}\in\mathcal{S}(r_1+r_1^{'},r_2+r_2^{'},r_3+r_3^{'},r_4+r_4^{'})}\left\langle\frac{2}{T}\sum_{n=SQ+1}^{T}\boldsymbol{\varepsilon}_n\circ\widetilde{\mathbf{X}}_n,\boldsymbol{\mathcal{S}}\right\rangle \\ &\leqslant C\left(\kappa^2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\kappa_U}\sqrt{\frac{d_{\mathcal{M}}^{'}}{T}}\right)\end{aligned}$$

with a probability at least $1-\exp(-Cd_{\mathcal{M}}^{'})-2\exp(Cd_{\mathcal{M}}^{'}-CT(\kappa_L/\kappa_U)^2\min\{\kappa^{-2},\kappa^{-4}\})$. This, together with (S2.2) and (S2.4), accomplishes the proof.

□

S3 Four useful lemmas for Theorems 1 and 2

We provide four useful lemmas for Theorems 1, 2.

Lemma 1. *(Covering number and discretization of low-multilinear-rank tensors). Suppose that $\bar{\mathcal{S}}(r_1,r_2,r_3,r_4)$ is an ϵ-net of the set $S(r_1,r_2,r_3,r_4) := \{\boldsymbol{\mathcal{A}}\in\mathbb{R}^{N\times N\times S\times Q} : \|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}=1,\mathrm{rank}_i(\boldsymbol{\mathcal{A}}_{(i)})\leqslant r_i, 1\leqslant i\leqslant 4\}$.*

(i) The cardinality of $\bar{\mathcal{S}}(r_1,r_2,r_3,r_4)$ satisfies

$$\left|\bar{\mathcal{S}}(r_1,r_2,r_3,r_4)\right| \leqslant (15/\epsilon)^{(r_1r_2r_3+Nr_1+Nr_2+Sr_3+Qr_4)}.$$

(ii) For any tensor $\boldsymbol{\mathcal{N}} \in \mathbb{R}^{N\times N\times S\times Q}$ *and matrix* $\mathbf{Z} \in \mathbb{R}^{NSQ\times T}$*, it holds that,*

$$\sup_{\boldsymbol{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \langle \boldsymbol{\mathcal{N}}, \boldsymbol{\Delta}\rangle \leqslant (1-4\epsilon)^{-1} \max_{\bar{\boldsymbol{\Delta}}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \left\langle \boldsymbol{\mathcal{N}}, \bar{\boldsymbol{\Delta}}\right\rangle, \quad \textit{and}$$

$$\sup_{\boldsymbol{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \left\|\boldsymbol{\Delta}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} \leqslant (1-4\epsilon)^{-1} \max_{\bar{\boldsymbol{\Delta}}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \left\|\bar{\boldsymbol{\Delta}}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}}.$$

Proof of Lemma 1. (i) The proof hinges on the covering number for the low-rank matrix studied by Candès and Plan (2011).

Let the HOSVD $\boldsymbol{\mathcal{A}} = [\![\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]$, where $\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}} = 1$ and each $\mathbf{U}_i$ is an orthonormal matrix. We construct an ϵ-net for $\boldsymbol{\mathcal{A}}$ by covering the set of $\mathcal{G}$ and all $\mathbf{U}_i$'s. We take $\bar{G}$ to be an $\epsilon/5$-net for $\mathcal{G}$ with $|\bar{G}| \leqslant (15/\epsilon)^{r_1r_2r_3r_4}$. Next, let $O_{n,r} = \left\{\mathbf{U} \in \mathbb{R}^{n\times r} : \mathbf{U}^\top\mathbf{U} = \mathbf{I}_r\right\}$. To cover $O_{n,r}$, it is beneficial to use the $\|\cdot\|_{1,2}$ norm, defined as

$$\|\mathbf{X}\|_{1,2} = \max_i \|\mathbf{X}_i\|_2$$

where $\mathbf{X}_i$ denotes the i th column of $\mathbf{X}$. Let $Q_{n,r} = \{\mathbf{X} \in \mathbb{R}^{n\times r} : \|\mathbf{X}\|_{1,2} \leqslant 1\}$. One can easily check that $O_{n,r} \subset Q_{n,r}$, and thus an $\epsilon/5$-net $\bar{O}_{n,r}$ for $O_{n,r}$ obeying $\left|\bar{O}_{n,r}\right| \leqslant (15/\epsilon)^{nr}$.

Denote $\bar{\mathcal{S}} = \left\{[\![\bar{\mathcal{G}}; \bar{\mathbf{U}}_1, \bar{\mathbf{U}}_2, \bar{\mathbf{U}}_3, \bar{\mathbf{U}}_4]\!] : \bar{\mathcal{G}} \in \bar{G}, \bar{\mathbf{U}}_i \in \bar{O}_{n_i,r_i}, i = 1,2,3,4\right\}$ and we have $|\bar{\mathcal{S}}| \leqslant |\bar{\mathcal{G}}| \times \left|\bar{O}_{N\times r_1}\right| \times \left|\bar{O}_{N\times r_2}\right| \times \left|\bar{O}_{S\times r_3}\right| \times \left|\bar{O}_{Q\times r_4}\right| = (15/\epsilon)^{r_1r_2r_3r_4+Nr_1+Nr_2+Sr_3+Qr_4}$.

We will next show that for any $\boldsymbol{\mathcal{A}} \in \mathcal{S}(r_1, r_2, r_3, r_4)$, there exists a $\bar{\boldsymbol{\mathcal{A}}} \in \bar{\mathcal{S}}$ such that $\|\boldsymbol{\mathcal{A}} - \bar{\boldsymbol{\mathcal{A}}}\|_{\mathrm{F}} \leqslant \epsilon$.

For any fixed $\boldsymbol{\mathcal{A}} \in \mathcal{S}(r_1, r_2, r_3, r_4)$, decompose it by HOSVD as $\boldsymbol{\mathcal{A}} = [\![\mathcal{G}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]$. Then, there exist $\bar{\boldsymbol{\mathcal{A}}} = [\![\bar{\mathcal{G}}; \bar{\mathbf{U}}_1, \bar{\mathbf{U}}_2, \bar{\mathbf{U}}_3, \bar{\mathbf{U}}_4]\!]$ with $\bar{\mathcal{G}} \in \bar{G}, \bar{\mathbf{U}}_i \in \bar{O}_{n_i,r_i}$ satisfying that

$\left\|\mathbf{U}_i - \bar{\mathbf{U}}_i\right\|_{1,2} \leqslant \epsilon/5$ and $\|\mathcal{G} - \bar{\mathcal{G}}\|_{\mathrm{F}} \leqslant \epsilon/5$. This gives

$$\begin{aligned}\|\mathcal{A} - \bar{\mathcal{A}}\|_{\mathrm{F}} \leqslant & \left\|[\![\mathcal{G} - \bar{\mathcal{G}}; \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]\right\|_{\mathrm{F}} + \left\|[\![\bar{\mathcal{G}}; \mathbf{U}_1 - \bar{\mathbf{U}}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]\right\|_{\mathrm{F}} \\ & + \left\|[\![\bar{\mathcal{G}}; \bar{\mathbf{U}}_1, \mathbf{U}_2 - \bar{\mathbf{U}}_2, \mathbf{U}_3, \mathbf{U}_4]\!]\right\|_{\mathrm{F}} + \left\|[\![\bar{\mathcal{G}}; \bar{\mathbf{U}}_1, \bar{\mathbf{U}}_2, \mathbf{U}_3 - \bar{\mathbf{U}}_3, \mathbf{U}_4]\!]\right\|_{\mathrm{F}} \\ & + \left\|[\![\bar{\mathcal{G}}; \bar{\mathbf{U}}_1, \bar{\mathbf{U}}_2, \bar{\mathbf{U}}_3, \mathbf{U}_4 - \bar{\mathbf{U}}_4]\!]\right\|_{\mathrm{F}}.\end{aligned}$$

Since each $\mathbf{U}_i$ is an orthonormal matrix, the first term is $\|\mathcal{G} - \bar{\mathcal{G}}\|_F \leqslant \epsilon/5$. For the second term, by the all-orthogonal property of $\bar{\mathcal{G}}$ and the orthonormal property of $\mathbf{U}_2$, $\mathbf{U}_3$ and $\mathbf{U}_4$,

$$\left\|[\![\bar{\mathcal{G}}; \mathbf{U}_1 - \bar{\mathbf{U}}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4]\!]\right\|_{\mathrm{F}} = \left\|\bar{\mathcal{G}} \times_1 \left(\mathbf{U}_1 - \bar{\mathbf{U}}_1\right)\right\|_{\mathrm{F}} \leqslant \|\bar{\mathcal{G}}\|_{\mathrm{F}} \left\|\mathbf{U}_1 - \bar{\mathbf{U}}_1\right\|_{2,1} \leqslant \epsilon/5.$$

Similarly, we can obtain the upper bound for the third, fourth and the last term, and thus show that $\|\mathcal{A} - \bar{\mathcal{A}}\|_{\mathrm{F}} \leqslant \epsilon$.

(ii) Consider an ϵ-net $\bar{\mathcal{S}}(2r_1, 2r_2, 2r_3, 2r_4)$ for $\mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$. Then for any tensor $\boldsymbol{\Delta} \in \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$, there exists a $\bar{\boldsymbol{\Delta}} \in \bar{\mathcal{S}}(2r_1, 2r_2, 2r_3, 2r_4)$ such that $\|\boldsymbol{\Delta} - \bar{\boldsymbol{\Delta}}\|_{\mathrm{F}} \leqslant \epsilon$. Since the rank of $\overline{\boldsymbol{\mathcal{W}}} = \boldsymbol{\Delta} - \bar{\boldsymbol{\Delta}}$ are at most $(4r_1, 4r_2, 4r_3, 4r_4)$, we can split the HOSVD of $\overline{\boldsymbol{\mathcal{W}}}$ into 16 parts such that $\overline{\boldsymbol{\mathcal{W}}} = \sum_{i=1}^{16} \overline{\boldsymbol{\mathcal{W}}}_i$, where $\mathrm{rank}_j(\overline{\boldsymbol{\mathcal{W}}}_i) \leqslant 2r_j$ for $1 \leqslant i \leqslant 16$ and $1 \leqslant j \leqslant 4$, and $\left\langle \overline{\boldsymbol{\mathcal{W}}}_j, \overline{\boldsymbol{\mathcal{W}}}_k \right\rangle = 0$ for any $j \neq k$. Then for any $\boldsymbol{\mathcal{N}} \in \mathbb{R}^{N \times N \times S \times Q}$, we have

$$\langle \boldsymbol{\mathcal{N}}, \boldsymbol{\Delta} \rangle = \left\langle \boldsymbol{\mathcal{N}}, \bar{\boldsymbol{\Delta}} \right\rangle + \sum_{i=1}^{16} \langle \boldsymbol{\mathcal{N}}, \overline{\boldsymbol{\mathcal{W}}}_i \rangle = \left\langle \boldsymbol{\mathcal{N}}, \bar{\boldsymbol{\Delta}} \right\rangle + \sum_{i=1}^{16} \left\langle \boldsymbol{\mathcal{N}}, \overline{\boldsymbol{\mathcal{W}}}_i / \|\overline{\boldsymbol{\mathcal{W}}}_i\|_{\mathrm{F}} \right\rangle \left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}}, \qquad \text{(S3.1)}$$

where $\overline{\boldsymbol{\mathcal{W}}}_i / \left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}} \in S(2r_1, 2r_2, 2r_3, 2r_4)$, and $\left\langle \boldsymbol{\mathcal{N}}, \overline{\boldsymbol{\mathcal{W}}}_i / \|\overline{\boldsymbol{\mathcal{W}}}_i\|_{\mathrm{F}} \right\rangle \leqslant \sup_{\boldsymbol{\Delta} \in S(2r_1, 2r_2, 2r_3, 2r_4)} \langle \boldsymbol{\mathcal{N}}, \boldsymbol{\Delta} \rangle$.

Note that $\left\|\overline{\boldsymbol{\mathcal{W}}}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{16} \left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}}^2$, and it holds that $\sum_{i=1}^{16} \left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}} \leqslant 4 \left\|\overline{\boldsymbol{\mathcal{W}}}\right\|_{\mathrm{F}} \leqslant 4\epsilon$,

which, together with (S3.1), implies that

$$\gamma := \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \langle \mathcal{N}, \mathbf{\Delta}\rangle \leqslant \max_{\overline{\mathbf{\Delta}}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \left\langle \mathcal{N}, \bar{\mathbf{\Delta}}\right\rangle + 4\gamma\epsilon, \quad \text{or}$$

$$\gamma = \sup_{\mathbf{\Delta}\in S(2r_1,2r_2,2r_3,2r_4)} \langle \mathcal{N}, \mathbf{\Delta}\rangle \leqslant (1-4\epsilon)^{-1} \max_{\overline{\mathbf{\Delta}}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \left\langle \mathcal{N}, \bar{\mathbf{\Delta}}\right\rangle.$$

For matrix $\mathbf{Z}\in\mathbb{R}^{NSQ\times T}$, it holds that

$$\begin{aligned}
\left\|\mathbf{\Delta}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} &\leqslant \left\|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} + \sum_{i=1}^{16}\left\|(\overline{\boldsymbol{\mathcal{W}}}_i)_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} \\
&= \left\|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} + \sum_{i=1}^{16}\left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}}\left\|(\overline{\boldsymbol{\mathcal{W}}}_i)_{(1)}/\|\overline{\boldsymbol{\mathcal{W}}}_i\|_{\mathrm{F}}\mathbf{Z}\right\|_{\mathrm{F}} \\
&\leqslant \left\|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} + \sum_{i=1}^{16}\left\|\overline{\boldsymbol{\mathcal{W}}}_i\right\|_{\mathrm{F}}\sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)}\left\|\mathbf{\Delta}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} \\
&\leqslant \left\|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}} + 4\epsilon\sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)}\left\|\mathbf{\Delta}_{(1)}\mathbf{Z}\right\|_{\mathrm{F}},
\end{aligned}$$

and we accomplish the proof by taking supremum on both sides. □

Lemma 2. *(Restricted strong convexity and smoothness). Suppose that Assumptions 1 and 2 hold, if* $T \gtrsim \max(\kappa^2,\kappa^4)(\kappa_U/\kappa_L)^2 d_{\mathcal{M}}$*, then*

$$\frac{1}{8}\kappa_L\left\|\mathbf{\Delta}\right\|_{\mathrm{F}}^2 \leqslant \frac{1}{T}\sum_{n=1}^{T}\left\|\mathbf{\Delta}_{(1)}\mathbf{x}_n\right\|_2^2 \leqslant \frac{8}{3}\kappa_U\left\|\mathbf{\Delta}\right\|_{\mathrm{F}}^2,$$

for all $\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)$ *with probability at least*

$$1-2\exp(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2\min\{\kappa^{-2},\kappa^{-4}\}),$$

where κ*,* κ_L*,* κ_U *and* $d_{\mathcal{M}}$ *are defined in Theorem 1.*

Proof of Lemma 2. Denote $R_T(\boldsymbol{\Delta}) = \sum_{n=1}^{T} \|\boldsymbol{\Delta}_{(1)}\mathbf{x}_n\|_2^2$, and it holds that

$$\begin{aligned} R_T(\boldsymbol{\Delta}) &= \sum_{n=1}^{T} \left\|\Delta_{(1)}\mathbf{x}_{\mathrm{n}}\right\|_2^2 \\ &= \sum_{n=1}^{T} \mathbf{x}_n^\top \boldsymbol{\Delta}_{(1)}^\top \boldsymbol{\Delta}_{(1)} \mathbf{x}_n. \end{aligned} \tag{S3.2}$$

By the spectral measure of ARMA processes in Basu and Michailidis (2015), we have $\lambda_{\min}\{\mathbb{E}(\mathbf{x}_n\mathbf{x}_n^\top)\} \geqslant \lambda_{\min}(\boldsymbol{\Sigma}_\varepsilon)/\mu_{\max}(\boldsymbol{\mathcal{A}}) = \kappa_L$ and $\lambda_{\max}\{\mathbb{E}(\mathbf{x}_n\mathbf{x}_n^\top)\} \leqslant \lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)/\mu_{\min}(\boldsymbol{\mathcal{A}}) = \kappa_U$, and it then holds that

$$T\kappa_L \leqslant \mathbb{E}(R_T(\boldsymbol{\Delta})) = \mathbb{E}\left(\sum_{n=1}^{T} \mathrm{vec}(\boldsymbol{\Delta})^\top (\mathbf{I}_N \otimes \mathbf{x}_n\mathbf{x}_n^\top)\mathrm{vec}(\boldsymbol{\Delta})\right) \leqslant T\kappa_U, \tag{S3.3}$$

as $\|\boldsymbol{\Delta}\|_{\mathrm{F}} = 1$. Furthermore, $R_T(\boldsymbol{\Delta}) \geqslant \mathbb{E}(R_T(\boldsymbol{\Delta})) - \sup_{\boldsymbol{\Delta}\in\mathcal{S}}\{|R_T(\boldsymbol{\Delta}) - \mathbb{E}(R_T(\boldsymbol{\Delta}))|\}$, and $R_T(\boldsymbol{\Delta}) \leqslant \mathbb{E}(R_T(\boldsymbol{\Delta})) + \sup_{\boldsymbol{\Delta}\in\mathcal{S}}\{|R_T(\boldsymbol{\Delta}) - \mathbb{E}(R_T(\boldsymbol{\Delta}))|\}$, where $\mathcal{S} = \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$. We next first bound $|R_T(\boldsymbol{\Delta}) - \mathbb{E}(R_T(\boldsymbol{\Delta}))|$ for each fixed $\boldsymbol{\Delta} \in \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$.

Consider the term of $R_T(\boldsymbol{\Delta}) - \mathbb{E}(R_T(\boldsymbol{\Delta}))$. Note that $\mathbf{x}_n = (\mathbf{y}_{n-1}^\top, \ldots, \mathbf{y}_{n-SQ}^\top)^\top \in \mathbb{R}^{NSQ}$, and $\boldsymbol{\mathcal{A}}_{(1)} = (\mathbf{A}_1, \ldots, \mathbf{A}_{SQ})$ with each $\mathbf{A}_i$ being an N-by-N matrix. From (2.1), we have $\mathbf{y}_n = \boldsymbol{\mathcal{A}}_{(1)}\mathbf{x}_n + \boldsymbol{\varepsilon}_n$. It can be further rewritten into an VAR(1) form, $\mathbf{x}_n = \mathbf{B}\mathbf{x}_{n-1} + \boldsymbol{e}_n$, and hence the VMA representation of $\mathbf{x}_n = \sum_{j=0}^{\infty} \mathbf{B}^j \boldsymbol{e}_{n-j}$, or $\mathbf{z} = \mathbf{P}\boldsymbol{e}$, where

$$\mathbf{B} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{SQ-1} & \mathbf{A}_{SQ} \\ \mathbf{I}_N & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_N & \mathbf{0} \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} \mathbf{I}_{NSQ} & \mathbf{B} & \mathbf{B}^2 & \cdots & \mathbf{B}^{T-1} & \cdots \\ \mathbf{0} & \mathbf{I}_{NSQ} & \mathbf{B} & \cdots & \mathbf{B}^{T-2} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{NSQ} & \cdots & \mathbf{B}^{T-3} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{NSQ} & \cdots \end{pmatrix},$$

$\boldsymbol{e}_n = (\boldsymbol{\varepsilon}_n^\top, \ldots, \mathbf{0})^\top \in \mathbb{R}^{NSQ}$, $\boldsymbol{e} = (\boldsymbol{e}_T^\top, \ldots, \boldsymbol{e}_{SQ}^\top, \ldots)^\top$, and $\mathbf{z} = (\mathbf{x}_T^\top, \ldots, \mathbf{x}_1^\top)^\top \in \mathbb{R}^{NSQT}$. Moreover, by Assumption 2, the error term has the form of $\boldsymbol{e} = \bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{\xi}}$, where $\bar{\boldsymbol{\xi}}_n = (\boldsymbol{\xi}_n^\top, \ldots, \mathbf{0})^\top \in \mathbb{R}^{NSQ}$, $\bar{\boldsymbol{\xi}} = (\bar{\boldsymbol{\xi}}_T^\top, \ldots, \bar{\boldsymbol{\xi}}_{SQ}^\top, \ldots)^\top$,

$$\bar{\boldsymbol{\Sigma}}_\varepsilon = \begin{pmatrix} \boldsymbol{\Sigma}_\varepsilon & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\varepsilon & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_\varepsilon \end{pmatrix} \in \mathbb{R}^{NSQ\times NSQ} \quad \text{and} \quad \bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \bar{\boldsymbol{\Sigma}}_\varepsilon^{1/2} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \bar{\boldsymbol{\Sigma}}_\varepsilon^{1/2} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \bar{\boldsymbol{\Sigma}}_\varepsilon^{1/2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Denote $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}} = \bar{\boldsymbol{\Sigma}}\mathbf{P}^\top(\mathbf{I}_T \otimes \boldsymbol{\Delta}_{(1)}^\top\boldsymbol{\Delta}_{(1)})\mathbf{P}\bar{\boldsymbol{\Sigma}}$, and then

$$\mathrm{R}_T = \sum_{n=1}^{T} \mathbf{x}_n^\top\boldsymbol{\Delta}_{(1)}^\top\boldsymbol{\Delta}_{(1)}\mathbf{x}_n = \bar{\boldsymbol{\xi}}^\top\bar{\boldsymbol{\Sigma}}\mathbf{P}^\top(\mathbf{I}_T \otimes \boldsymbol{\Delta}_{(1)}^\top\boldsymbol{\Delta}_{(1)})\mathbf{P}\bar{\boldsymbol{\Sigma}}\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}^\top\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\bar{\boldsymbol{\xi}}.$$

Note that $\lambda_{\max}(\mathbf{P}\mathbf{P}^\top) = 1/\mu_{\min}(\boldsymbol{\mathcal{A}})$, $\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{op}} \leqslant \kappa_U$ and

$$\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{F}} \leqslant \|\bar{\boldsymbol{\Sigma}}\|_{\mathrm{op}}^2\|\mathbf{P}\|_{\mathrm{op}}\|\mathbf{P}^\top\|_{\mathrm{op}}\|\mathbf{I}_T \otimes \boldsymbol{\Delta}_{(1)}^\top\boldsymbol{\Delta}_{(1)}\|_{\mathrm{F}} \leqslant \sqrt{T}\kappa_U.$$

For any $t > 0$, by Hanson-Wright inequality, we can bound $\mathrm{R}_T(\Delta) - \mathbb{E}(\mathrm{R}_T(\Delta))$ below,

$$\begin{aligned} \mathbb{P}\left[|\mathrm{R}_T(\Delta) - \mathbb{E}(\mathrm{R}_T(\Delta))| \geqslant t\right] &\leqslant 2\exp\left(-C\min\left(\frac{t}{\kappa^2\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{op}}}, \frac{t^2}{\kappa^4\|\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2}\right)\right) \\ &\leqslant 2\exp\left(-C\min\left(\frac{t}{\kappa^2\kappa_U}, \frac{t^2}{\kappa^4 T\kappa_U^2}\right)\right). \end{aligned} \tag{S3.4}$$

Let $t_1 = T\kappa_L/2$ and, from (S3.3), it holds that

$$\begin{aligned} &\mathbb{P}\left[0.5\kappa_L \leqslant T^{-1}R_T(\boldsymbol{\Delta}) \leqslant 1.5\kappa_U\right] \\ &\qquad \geqslant 1 - 2\exp\left(-C\min\left(\frac{T\kappa_L}{\kappa^2\kappa_U}, \frac{T\kappa_L^2}{\kappa^4\kappa_U^2}\right)\right), \end{aligned} \tag{S3.5}$$

Let $\bar{\mathcal{S}}$ to be an ϵ-covering net of $S(2r_1, 2r_2, 2r_3, 2r_4)$. To construct the union bound, we rewrite $R_T(\boldsymbol{\Delta})$ as $R_T(\boldsymbol{\Delta}) = \|\boldsymbol{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}}^2$, where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_T) \in \mathbb{R}^{NSQ\times T}$. Define the

event

$$\mathcal{E}(\epsilon) = \left\{ \forall \mathbf{\Delta} \in \bar{\mathcal{S}}(2r_1, 2r_2, 2r_3, 2r_4) : \sqrt{0.5\kappa_L} \leqslant \frac{1}{\sqrt{T}} \|\mathbf{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}} \leqslant \sqrt{1.5\kappa_U} \right\}.$$

Then, by the pointwise bound in (S3.5) and the covering number in Lemma 1(i),

$$\mathbb{P}\left[\mathcal{E}^c(\epsilon)\right] \leqslant 2\exp\left(C d_{\mathcal{M}} - C\min\left(\frac{T\kappa_L}{\kappa^2\kappa_U}, \frac{T\kappa_L^2}{\kappa^4\kappa_U^2} \right) \right).$$

Note that, by Lemma 1 (ii),

$$\begin{aligned} \mathcal{E}(\epsilon) &\subset \left\{ \max_{\mathbf{\Delta}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \frac{1}{\sqrt{T}} \|\mathbf{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}} \leqslant \sqrt{1.5\kappa_U} \right\} \\ &\subset \left\{ \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \frac{1}{\sqrt{T}} \|\mathbf{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}} \leqslant \frac{\sqrt{1.5\kappa_U}}{1-4\epsilon} \right\}. \end{aligned}$$

Moreover, similarly to Lemma 1(ii), we can show that

$$\begin{aligned} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}} &\geqslant \|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\|_{\mathrm{F}} - \sum_{i=1}^{16} \|(\overline{\mathbf{\mathcal{W}}}_i)_{(1)}\mathbf{Z}\|_{\mathrm{F}} \\ &\geqslant \|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\|_{\mathrm{F}} - \sum_{i=1}^{16} \|\overline{\mathbf{\mathcal{W}}}_i\|_{\mathrm{F}} \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}} \\ &\geqslant \|\bar{\mathbf{\Delta}}_{(1)}\mathbf{Z}\|_{\mathrm{F}} - 4\epsilon \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}}, \end{aligned}$$

where the last inequality is due to $\sum_{i=1}^{16} \|\overline{\mathbf{\mathcal{W}}}_i\|_{\mathrm{F}} \leqslant 4\epsilon$. Taking infimum on both sides, if $0 \leqslant \epsilon \leqslant \frac{1}{8}$, we have

$$\begin{aligned} \inf_{\mathbf{\Delta}\in S(2r_1,2r_2,2r_3,2r_4)} \frac{1}{\sqrt{T}} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}} \geqslant & \min_{\mathbf{\Delta}\in\bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \frac{1}{\sqrt{T}} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}} \\ & - 4\epsilon \frac{1}{\sqrt{T}} \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \|\mathbf{\Delta}_{(1)}\mathbf{Z}\|_{\mathrm{F}} \\ \geqslant & \sqrt{0.5\kappa_L} - 4\epsilon \frac{\sqrt{1.5\kappa_U}}{1-4\epsilon} \geqslant \sqrt{0.5\kappa_L} - 4\epsilon\sqrt{6\kappa_U}. \end{aligned}$$

When ϵ is chosen to be $\frac{1}{8}\sqrt{\frac{\kappa_L}{12\kappa_U}}$, As a result, with the above choice of ϵ,

$$\mathcal{E}(\epsilon) \subset \left\{ \frac{\kappa_L}{8} \leqslant \inf_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \frac{1}{T}\|\mathbf{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}}^2 \leqslant \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \frac{1}{T}\|\mathbf{\Delta}_{(1)}\mathbf{X}\|_{\mathrm{F}}^2 \leqslant \frac{8\kappa_U}{3} \right\}.$$

Given the conditions that $T \gtrsim (\kappa_U/\kappa_L)^2 \max(\kappa^2, \kappa^4) d_{\mathcal{M}}$, we have that for all $\mathbf{\Delta} \in \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$,

$$\mathbb{P}\left[\frac{\kappa_L}{8} \leqslant \frac{1}{T}\sum_{n=1}^{T} \|\mathbf{\Delta}_{(1)}\mathbf{x}_n\|_2^2 \leqslant \frac{8\kappa_U}{3} \right] \geqslant 1 - 2\exp\left(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2 \min(\kappa^{-2}, \kappa^{-4})\right).$$

This accomplishes the proof. □

Lemma 3. *(Deviation bound) Suppose that Assumptions 1 and 2 hold. If sample size* $T \gtrsim \max(\kappa^2, \kappa^4)(\kappa_U/\kappa_L)^2 d_{\mathcal{M}}$, *then*

$$\sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,,2r_4)} \langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \mathbf{\Delta} \rangle \leqslant C\kappa^2 \sqrt{\lambda_{\max}(\mathbf{\Sigma}_\varepsilon)\kappa_U} \sqrt{\frac{d_{\mathcal{M}}}{T}}$$

with probability at least

$$1 - \exp(-Cd_{\mathcal{M}}) - 2\exp(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2 \min\{\kappa^{-2}, \kappa^{-4}\})$$

, where κ, κ_L, κ_U *and* $d_{\mathcal{M}}$ *are defined in Theorem 1.*

Proof of Lemma 3. We let $\mathbf{X}_n = (\mathbf{y}_{n-1}, \ldots, \mathbf{y}_{n-SQ}) \in \mathbb{R}^{N\times SQ}$, then

$$\sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \mathbf{\Delta} \rangle = \sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n, \mathbf{\Delta}_{(1)}\mathbf{x}_n \rangle. \tag{S3.6}$$

Since it is easily verified that $\langle \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \mathbf{\Delta} \rangle = \langle \boldsymbol{\varepsilon}_n, \mathbf{\Delta}_{(1)}\mathbf{x}_n \rangle$. Denote $S_t(\mathbf{\Delta}) = \sum_{n=1}^{t} \langle \boldsymbol{\varepsilon}_n, \mathbf{\Delta}_{(1)}\mathbf{x}_n \rangle$ and $R_t(\mathbf{\Delta}) = \sum_{n=1}^{t} \|\mathbf{\Delta}_{(1)}\mathbf{x}_n\|_2^2$ for $1 \leqslant n \leqslant T$. By the Chernoff bound of errors, for any

$\alpha > 0, \beta > 0$ and $c > 0$, there exists $\eta > 0$,

$$\begin{aligned}
&\mathbb{P}\Big[\{S_T(\boldsymbol{\Delta}) \geqslant \alpha\} \bigcap \{R_T(\boldsymbol{\Delta}) \leqslant \beta\}\Big] \\
=&\inf_{\eta>0} \mathbb{P}\left[\{\exp\left(\eta S_T(\boldsymbol{\Delta})\right) \geqslant \exp\left(\eta\alpha\right)\} \bigcap \{R_T(\boldsymbol{\Delta}) \leqslant \beta\}\right] \\
=&\inf_{\eta>0} \mathbb{P}\left[\exp\left(\eta S_T\left(\boldsymbol{\Delta}\right)\right) \mathbb{I}\left(R_T(\boldsymbol{\Delta}) \leqslant \beta\right) \geqslant \exp(\eta\alpha)\right] \\
\leqslant&\inf_{\eta>0} \exp\left(-\eta\alpha\right) \mathbb{E}\left[\exp(\eta S_T(\boldsymbol{\Delta}))\mathbb{I}(R_T(\boldsymbol{\Delta}) \leqslant \beta)\right] \\
=&\inf_{\eta>0} \exp(-\eta\alpha + c\eta^2\beta)\mathbb{E}\left[\exp(\eta S_T(\boldsymbol{\Delta}) - c\eta^2\beta)\mathbb{I}(R_T(\boldsymbol{\Delta}) \leqslant \beta)\right] \\
\leqslant&\inf_{\eta>0} \exp(-\eta\alpha + c\eta^2\beta)\mathbb{E}\left[\exp(\eta S_T(\boldsymbol{\Delta}) - c\eta^2 R_T(\boldsymbol{\Delta}))\right].
\end{aligned} \tag{S3.7}$$

By the tower property for conditional expectations, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\eta S_T(\boldsymbol{\Delta}) - c\eta^2 R_T(\boldsymbol{\Delta})\right)\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\exp\left(\eta S_T(\boldsymbol{\Delta}) - c\eta^2 R_T(\boldsymbol{\Delta})\right)|\mathcal{F}_{T-1}\right]\right] \\
=&\mathbb{E}\left[\exp\left(\eta S_{T-1}(\boldsymbol{\Delta}) - c\eta^2 R_{T-1}(\boldsymbol{\Delta})\right)\mathbb{E}\left[\exp\left(\eta\langle\boldsymbol{\varepsilon}_T, \boldsymbol{\Delta}_{(1)}\mathbf{x}_T\rangle - c\eta^2\|\boldsymbol{\Delta}_{(1)}\mathbf{x}_T\|_2^2\right)|\mathcal{F}_{T-1}\right]\right].
\end{aligned}$$

With the sub-Gaussianity condition in Assumption 2, then $\langle\boldsymbol{\varepsilon}_T, \boldsymbol{\Delta}_{(1)}\mathbf{x}_T\rangle = \langle\boldsymbol{\xi}_T, \boldsymbol{\Sigma}_\varepsilon^{1/2}\boldsymbol{\Delta}_{(1)}\mathbf{x}_T\rangle$, and $\mathbb{E}\left[\exp(\eta\langle\boldsymbol{\varepsilon}_T, \boldsymbol{\Delta}_{(1)}\mathbf{x}_T\rangle)\right] \leqslant \exp\left(\eta^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\|\boldsymbol{\Delta}_{(1)}\mathbf{x}_T\|_2^2/2\right)$. Since $\mathbf{x}_n$ is $\mathcal{F}_{n-1}$-measurable, $\boldsymbol{\varepsilon}_n$ is $\mathcal{F}_n$-measurable and $\boldsymbol{\varepsilon}_n|\mathcal{F}_{n-1}$ is mean-zero, let $c = \kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)/2$, and the following inequalities can be easily deduced,

$$\begin{aligned}
&\mathbb{E}[\exp(\eta S_T(\boldsymbol{\Delta}) - \eta^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)R_T(\boldsymbol{\Delta})/2)] \\
\leqslant&\mathbb{E}[\exp(\eta S_{T-1}(\boldsymbol{\Delta}) - \eta^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)R_{T-1}(\boldsymbol{\Delta})/2)] \\
\leqslant&\cdots \leqslant \mathbb{E}[\exp(\eta S_{SQ+1}(\boldsymbol{\Delta}) - \eta^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)R_{SQ+1}(\boldsymbol{\Delta})/2)] \leqslant 1.
\end{aligned}$$

As a result, for any $\alpha > 0$ and $\beta > 0$, we can have the following inequality of (S3.7),

$$\begin{aligned}
&\mathbb{P}[\{S_T(\boldsymbol{\Delta}) \geqslant \alpha\} \bigcap \{R_T(\boldsymbol{\Delta}) \leqslant \beta\}] \\
\leqslant &\inf_{\eta>0} \exp(-\eta\alpha + \eta^2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\beta/2) \\
= &\exp\left(-\frac{\alpha^2}{2\kappa^2\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\beta}\right).
\end{aligned} \tag{S3.8}$$

Moreover, according to Lemma 2, the following bounds for $R_T(\boldsymbol{\Delta})$ hold that

$$\frac{T}{8}\kappa_L \leqslant R_T(\boldsymbol{\Delta}) \leqslant \frac{8T}{3}\kappa_U \tag{S3.9}$$

with probability at least $1 - 2\exp(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2 \min\{\kappa^{-2}, \kappa^{-4}\})$.

By Lemma 1 (ii), for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{\Delta}\in S(2r_1,2r_2,2r_3,2r_4)} \left\langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \right\rangle \geqslant x\right] \\
\leqslant &\mathbb{P}\left[\max_{\boldsymbol{\Delta}\in \bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \left\langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \right\rangle \geqslant (1-4\epsilon)x\right] \\
\leqslant &|\bar{\mathcal{S}}(2r_1, 2r_2, 2r_3, 2r_4)| \cdot \mathbb{P}\left[\left\langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \right\rangle \geqslant (1-4\epsilon)x\right],
\end{aligned} \tag{S3.10}$$

which, together with (S3.8) and (S3.9), implies that

$$\begin{aligned}
&\mathbb{P}\left[\left\langle \frac{1}{T}\sum_{n=1}^{T} \boldsymbol{\varepsilon}_n \circ \mathbf{X}_n, \boldsymbol{\Delta} \right\rangle \geqslant (1-4\epsilon)x\right] \\
\leqslant &\mathbb{P}[\{S_T(\boldsymbol{\Delta}) \geqslant T(1-4\epsilon)x\} \bigcap \{R_T(\boldsymbol{\Delta}) \leqslant CT\kappa^2\kappa_U\}] + \mathbb{P}[R_T(\boldsymbol{\Delta}) \geqslant CT\kappa^2\kappa_U] \\
\leqslant &\exp\left[-\frac{(1-4\epsilon)^2 T x^2}{2C\kappa^4\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)\kappa_U}\right] + 2\exp(Cd_{\mathcal{M}} - CT(\kappa_L/\kappa_U)^2 \min\{\kappa^{-2}, \kappa^{-4}\}),
\end{aligned}$$

for any $x > 0$. Note that, from Lemma 1, $|\bar{\mathcal{S}}(r_1, r_2, r_3, r_4)| \leqslant (15/\epsilon)^{d_{\mathcal{M}}}$. By letting

$\epsilon = 1/10$, and $x = C\sqrt{d_{\mathcal{M}}\kappa^4\lambda_{\max}(\mathbf{\Sigma}_\varepsilon)\kappa_U/T}$, we then have

$$\begin{aligned}&\mathbb{P}\left[\sup_{\mathbf{\Delta}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)}\left\langle\frac{1}{T}\sum_{n=1}^{T}\boldsymbol{\varepsilon}_n\circ\mathbf{X}_n,\mathbf{\Delta}\right\rangle\geqslant C\kappa^2\sqrt{\lambda_{\max}(\mathbf{\Sigma}_\varepsilon)\kappa_U}\sqrt{\frac{d_{\mathcal{M}}}{T}}\right]\\&\leqslant\exp(-Cd_{\mathcal{M}})+2\exp(Cd_{\mathcal{M}}-CT(\kappa_L/\kappa_U)^2\min\{\kappa^{-2},\kappa^{-4}\}).\end{aligned}\tag{S3.11}$$

We hence complete the proof. □

Lemma 4. *(Contractive projection property) Suppose that* $\mathcal{X}\in\mathbf{\Theta}(r_1^{(0)},r_2^{(0)},r_3^{(0)},r_4^{(0)})$ *and* $r_i^{(1)}<r_i^{(2)}<r_i^{(0)}$ *with* $1\leqslant i\leqslant 4$*, then for any* $\mathcal{Y}\in\mathbf{\Theta}(r_1^{(1)},r_2^{(1)},r_3^{(1)},r_4^{(1)})$*,*

$$\|P_{\mathbf{\Theta}(r_1^{(2)},r_2^{(2)},r_3^{(2)},r_4^{(2)})}(\mathcal{X})-\mathcal{X}\|_{\mathrm{F}}\leqslant[\Pi_{i=1}^4(\beta_i+1)-1]\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}},\tag{S3.12}$$

where $\beta_i=\sqrt{(r_i^{(0)}-r_i^{(2)})/(r_i^{(0)}-r_i^{(1)})}$.

Proof of Lemma 4. The proof could be divided into two parts. First, we show a matrix low-rank projection result, i.e. for two matrices $\mathbf{W},\mathbf{A}\in\mathbb{R}^{N_1\times N_2},\mathrm{rank}(\mathbf{A})=\mathrm{r}^{(1)}<\mathrm{r}^{(2)}<\mathrm{r}^{(0)}=\mathrm{rank}(\mathbf{W})$, we have $\|P_{r^{(2)}}(\mathbf{W})-\mathbf{W}\|_{\mathrm{F}}^2\leqslant\left(r^{(0)}-r^{(2)}\right)/\left(r^{(0)}-r^{(1)}\right)\|\mathbf{A}-\mathbf{W}\|_{\mathrm{F}}^2$ where $P_{r^{(2)}}$ denotes projection to matrix subspace with $\mathrm{rank(W)}\leqslant\mathrm{r}^{(2)}$. Second, we extend the result to tensors with the approximate projection operator $P_{\mathbf{\Theta}(r_1^{(2)},r_2^{(2)},r_3^{(2)},r_4^{(2)})}$.

The first part mainly follows Lemma 1 and 2 in Jain et al. (2014). Consider a SVD $\mathbf{W}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with singular values $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_{r^{(0)}}$, and it then holds that

$$\|P_{r^{(2)}}(\mathbf{W})-\mathbf{W}\|_{\mathrm{F}}^2=\sum_{i=r^{(2)}+1}^{r^{(0)}}\sigma_i^2=\|p_{r^{(2)}}(\mathrm{diag}(\mathbf{\Sigma}))-\mathrm{diag}(\mathbf{\Sigma})\|_2^2,$$

where $\mathrm{diag}(\Sigma)$ returns a column vector of the elements on the diagonal of Σ and $p_{r^{(2)}}(\mathrm{diag}(\Sigma))$ takes the $r^{(2)}$ largest elements of the vector $\mathrm{diag}(\Sigma)$. Then we consider

the expression

$$\begin{aligned}&\frac{\|p_{r^{(1)}}(\mathrm{diag}(\Sigma))-\mathrm{diag}(\Sigma)\|_2^2}{r^{(0)}-r^{(1)}}-\frac{\|p_{r^{(2)}}(\mathrm{diag}(\Sigma))-\mathrm{diag}(\Sigma)\|_2^2}{r^{(0)}-r^{(2)}}\\=&\frac{1}{r^{(0)}-r^{(1)}}\sum_{i=r^{(1)}+1}^{r^{(2)}}\sigma_i^2+\left(\frac{1}{r^{(0)}-r^{(1)}}-\frac{1}{r^{(0)}-r^{(2)}}\right)\sum_{i=r^{(2)}+1}^{r^{(0)}}\sigma_i^2\geqslant 0.\end{aligned}$$

Hence, $\|p_{r^{(2)}}(\mathrm{diag}(\Sigma))-\mathrm{diag}(\Sigma)\|_2^2\leqslant\left(r^{(0)}-r^{(2)}\right)/\left(r^{(0)}-r^{(1)}\right)\|p_{r^{(1)}}(\mathrm{diag}(\Sigma))-\mathrm{diag}(\Sigma)\|_2^2=\left(r^{(0)}-r^{(2)}\right)/\left(r^{(0)}-r^{(1)}\right)\|P_{r^{(1)}}(\mathbf{W})-\mathbf{W}\|_{\mathrm{F}}^2\leqslant\left(r^{(0)}-r^{(2)}\right)/\left(r^{(0)}-r^{(1)}\right)\|\mathbf{A}-\mathbf{W}\|_{\mathrm{F}}^2$. The last inequality is due to Eckart Young Theorem and finishes the proof for the first part.

Second we consider the approximate projection of tensor $P_{\Theta\left(r_1^{(2)},r_2^{(2)},r_3^{(2)},r_4^{(2)}\right)}(\mathcal{X})$. Recall that

$$P_{\mathbf{\Theta}(r_1^{(2)},r_2^{(2)},r_3^{(2)},r_4^{(2)})}(\mathcal{X})=\left(\mathcal{M}_4^{-1}\circ P_{r_4^{(2)}}\circ\mathcal{M}_4\right)\circ\cdots\circ\left(\mathcal{M}_1^{-1}\circ P_{r_1^{(2)}}\circ\mathcal{M}_1\right)\circ\mathcal{X}.$$

We then introduce following notation for projection operator sequentially

$$\begin{aligned}\mathcal{X}_{[1]}&=\left(\mathcal{M}_1^{-1}\circ P_{r_1^{(2)}}\circ\mathcal{M}_1\right)\circ\mathcal{X},\\\mathcal{X}_{[i]}&=\left(\mathcal{M}_i^{-1}\circ P_{r_i^{(2)}}\circ\mathcal{M}_i\right)\circ\mathcal{X}_{[i-1]},\end{aligned}$$

for $i=2,3,4$. $\mathcal{M}_i$ represents mode- i sequential matricization. So it is obvious that $P_{\Theta\left(r_1^{(2)},r_2^{(2)},r_3^{(2)},r_4^{(2)}\right)}(\mathcal{X})=\mathcal{X}_{[4]}$. By triangle inequality,

$$\left\|\mathcal{X}_{[4]}-\mathcal{X}\right\|_{\mathrm{F}}\leqslant\left\|\mathcal{X}_{[1]}-\mathcal{X}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[2]}-\mathcal{X}_{[1]}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[3]}-\mathcal{X}_{[2]}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[4]}-\mathcal{X}_{[3]}\right\|_{\mathrm{F}}.\quad\text{(S3.13)}$$

Let $\beta_i=\sqrt{\left(r_i^{(0)}-r_i^{(2)}\right)/\left(r_i^{(0)}-r_i^{(1)}\right)}$ for $i=1,2,3,4$. Now we use the result in the first part to analyze every term on the right side of the above inequality (S3.13).

For any $\mathcal{Y}$ such that $\operatorname{rank}\left(\mathcal{Y}_{(1)}\right) \leqslant \mathrm{r}_1^{(1)}$,

$$\begin{aligned}\left\|\mathcal{X}_{[1]}-\mathcal{X}\right\|_{\mathrm{F}} =& \left\|\left(\mathcal{M}_1^{-1} \circ P_{r_1^{(2)}} \circ \mathcal{M}_1\right) \circ \mathcal{X}-\mathcal{X}\right\|_{\mathrm{F}} \\ =& \left\|P_{r_1^{(2)}}\left(\mathcal{M}_1(\mathcal{X})\right)-\mathcal{M}_1(\mathcal{X})\right\|_{\mathrm{F}} \\ \leqslant & \beta_1\left\|\mathcal{Y}_{(1)}-\mathcal{M}_1(\mathcal{X})\right\|_{\mathrm{F}} \\ =& \beta_1\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}.\end{aligned}$$

Similarly, we have

$$\left\|\mathcal{X}_{[2]}-\mathcal{X}_{[1]}\right\|_{\mathrm{F}} \leqslant \beta_2\left\|\mathcal{Y}-\mathcal{X}_{[1]}\right\|_{\mathrm{F}} \leqslant \beta_2\left(\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}+\left\|\mathcal{X}_{[1]}-\mathcal{X}\right\|_{\mathrm{F}}\right) \leqslant \beta_2\left(1+\beta_1\right)\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}.$$

Furthermore,

$$\begin{aligned}\left\|\mathcal{X}_{[3]}-\mathcal{X}_{[2]}\right\|_{\mathrm{F}} & \leqslant \beta_3\left\|\mathcal{Y}-\mathcal{X}_{[2]}\right\|_{\mathrm{F}} \\ & \leqslant \beta_3\left(\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}+\left\|\mathcal{X}_{[1]}-\mathcal{X}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[2]}-\mathcal{X}_{[1]}\right\|_{\mathrm{F}}\right) \\ & \leqslant \beta_3\left(1+\beta_2\right)\left(1+\beta_1\right)\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}},\end{aligned}$$

and

$$\begin{aligned}\left\|\mathcal{X}_{[4]}-\mathcal{X}_{[3]}\right\|_{\mathrm{F}} \leqslant & \beta_4\left\|\mathcal{Y}-\mathcal{X}_{[3]}\right\|_{\mathrm{F}} \\ \leqslant & \beta_4\left(\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}+\left\|\mathcal{X}_{[1]}-\mathcal{X}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[2]}-\mathcal{X}_{[1]}\right\|_{\mathrm{F}}+\left\|\mathcal{X}_{[3]}-\mathcal{X}_{[2]}\right\|_{\mathrm{F}}\right) \\ \leqslant & \beta_4 \prod_{i=1}^3\left(1+\beta_i\right)\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}.\end{aligned}$$

Sum up these terms and we have

$$\left\|P_{\boldsymbol{\Theta}\left(r_1^{(2)}, r_2^{(2)}, r_3^{(2)}, r_4^{(2)}\right)}(\mathcal{X})-\mathcal{X}\right\|_{\mathrm{F}} \leqslant\left[\prod_{i=1}^4\left(\beta_i+1\right)-1\right]\|\mathcal{Y}-\mathcal{X}\|_{\mathrm{F}}.$$

We hence complete the proof. □

S4 Proofs of Corollary 1

Proof. Since the low-rank estimator $\widehat{\boldsymbol{\mathcal{A}}}$ is typically biased, we then compute a debiased estimator using residuals from the initial model fit

$$\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} = \widehat{\boldsymbol{\mathcal{A}}}_{(1)} + \frac{1}{T}\sum_{n=1}^{T}(\mathbf{y}_n - \widehat{\boldsymbol{\mathcal{A}}}_{(1)}\mathbf{x}_n)\mathbf{x}_n^{\top}\mathbf{K},$$

where $\mathbf{K}$ is the precision matrix of $\mathbf{x}_n$. Let $\widehat{\boldsymbol{\Delta}}_{(1)} = \boldsymbol{\mathcal{A}}_{(1)} - \widehat{\boldsymbol{\mathcal{A}}}_{(1)}$ be the estimation error, The debiased estimator can be decomposed as

$$\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} = \boldsymbol{\mathcal{A}}_{(1)} + \frac{1}{T}\sum_{n=1}^{T}\boldsymbol{\varepsilon}_n\mathbf{x}_n^{\top}\mathbf{K} + \frac{1}{T}\sum_{n=1}^{T}(\widehat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n\mathbf{x}_n^{\top}\mathbf{K} - \widehat{\boldsymbol{\Delta}}_{(1)}).$$

For a prespecified loading tensor $\boldsymbol{\mathcal{B}}$, let $d_{\mathcal{B}}$ denotes the size of the low-rank space that the prespecified loading tensor $\boldsymbol{\mathcal{B}}$ resides, and $c = T^{-1}\mathrm{tr}(\boldsymbol{\mathcal{B}}_{(1)}\boldsymbol{K}\boldsymbol{\mathcal{B}}_{(1)}^{\top}\boldsymbol{\Sigma}_{\varepsilon})$, we have

$$\begin{aligned}
&\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\langle\widehat{\boldsymbol{\mathcal{A}}}^{u} - \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}\rangle \leqslant u\right) - \mathbb{P}\left(g \leqslant u\right)\right| \\
&= \sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\langle\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}, \boldsymbol{\mathcal{B}}_{(1)}\rangle \leqslant u\right) - \mathbb{P}\left(g \leqslant u\right)\right| \\
&\leqslant \sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\langle\widehat{\boldsymbol{\mathcal{A}}}^{u}_{(1)} - \boldsymbol{\mathcal{A}}_{(1)}, \boldsymbol{\mathcal{B}}_{(1)}\rangle \leqslant u\right) - \mathbb{P}\left(\frac{1}{T}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^{\top}\mathbf{K}, \boldsymbol{\mathcal{B}}_{(1)}\rangle \leqslant u - \delta\right)\right| \\
&\quad + \sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\frac{1}{T}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^{\top}\mathbf{K}, \boldsymbol{\mathcal{B}}_{(1)}\rangle \leqslant u - \delta\right) - \mathbb{P}\left(g \leqslant u - \delta\right)\right| \\
&\quad + \sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(g \leqslant u - \delta\right) - \mathbb{P}\left(g \leqslant u\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \sup_{u\in\mathbb{R}}\left\{\mathbb{P}\left(\frac{1}{T}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle\leqslant u\right)-\mathbb{P}\left(\frac{1}{T}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle\leqslant u-\delta\right)\right\}\\
&\quad+\mathbb{P}\left\{\left|\langle\frac{1}{T}\sum_{n=1}^{T}(\hat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n\mathbf{x}_n^\top\mathbf{K}-\hat{\boldsymbol{\Delta}}_{(1)}),\boldsymbol{\mathcal{B}}_{(1)}\rangle\right|\geqslant\delta\right\}\\
&\quad+\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\frac{1}{T}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle\leqslant u-\delta\right)-\mathbb{P}\left(g\leqslant u-\delta\right)\right|\\
&\quad+\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(g\leqslant u-\delta\right)-\mathbb{P}\left(g\leqslant u\right)\right|\\
&\leqslant 2\exp\left(Cd_{\mathcal{M}}d_B-C\min\left(\frac{T\delta}{\kappa^2\kappa_U^k},\frac{T\delta^2}{\kappa^4{\kappa_U^k}^2}\right)\right)+3\frac{\sigma}{T^{1/9}}\|\mathbf{K}\|_2\|\boldsymbol{\mathcal{B}}_{(1)}\|_{\mathrm{F}}+2\frac{\delta}{\sqrt{2\pi c}}e^{-u^2/2c}.
\end{aligned}
$$

The last inequality is due to Lemmas 5 and 6. Let $\delta=T^{-1/4}$, and $T^{1/2}\gtrsim{\kappa_U^k}^2\max(\kappa^2,\kappa^4)d_{\mathcal{M}}d_{\mathcal{B}}$, the above bound is dominated by the second term with $T^{-1/9}$. □

S5 Two useful lemmas for Corollary 1

We provide two useful lemmas for Corollary 1.

Lemma 5. *(Restricted strong convexity of the estimated low-Tucker-rank linear form) Suppose Assumptions 1 and 2 hold, $T\gtrsim(\kappa_U^k/\kappa_L^k)^2\max(\kappa^2,\kappa^4)d_{\mathcal{M}}d_{\mathcal{B}}$, then for all $\hat{\boldsymbol{\Delta}}\in\mathcal{S}(2r_1,2r_2,2r_3,2r_4)$, any $t>0$, we have*

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\langle\frac{1}{T}\sum_{n=1}^{T}(\hat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n\mathbf{x}_n^\top\mathbf{K}-\hat{\boldsymbol{\Delta}}_{(1)}),\boldsymbol{\mathcal{B}}_{(1)}\rangle\right|\geqslant C\|\boldsymbol{\mathcal{B}}_{(1)}\|_{\mathrm{F}}\|\hat{\boldsymbol{\Delta}}\|_{\mathrm{F}}t\right\}\\
&\leqslant 2\exp\left(Cd_{\mathcal{M}}d_{\mathcal{B}}-C\min\left(\frac{Tt}{\kappa^2\kappa_U^k},\frac{Tt^2}{\kappa^4{\kappa_U^k}^2}\right)\right).
\end{aligned}
$$

Proof. Without loss of generality, we restrict the $\|\boldsymbol{\mathcal{B}}_{(1)}\|_{\mathrm{F}}$ and $\|\hat{\boldsymbol{\Delta}}\|_{\mathrm{F}}$ to be 1. Let $\mathrm{H}_T=\langle\sum_{n=1}^{T}\hat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle$, then its expectation satisfies $\mathbb{E}(\mathrm{H}_T)=T\langle\hat{\boldsymbol{\Delta}},\boldsymbol{\mathcal{B}}\rangle$.

Following the proof of Lemma 2, $\langle\sum_{n=1}^{T}\hat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle=\sum_{n=1}^{T}\mathbf{x}_n^\top\mathbf{K}\boldsymbol{\mathcal{B}}_{(1)}^\top\hat{\boldsymbol{\Delta}}_{(1)}\mathbf{x}_n=$

$\bar{\boldsymbol{\xi}}^\top \bar{\boldsymbol{\Sigma}} \mathbf{P}^\top (\mathbf{I}_T \otimes \mathbf{K}\boldsymbol{\mathcal{B}}_{(1)} \hat{\boldsymbol{\Delta}}_{(1)}) \mathbf{P} \bar{\boldsymbol{\Sigma}} \bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}^\top \boldsymbol{\Sigma}_\Delta \bar{\boldsymbol{\xi}}$, where $\boldsymbol{\Sigma}_\Delta = \bar{\boldsymbol{\Sigma}} \mathbf{P}^\top (\mathbf{I}_T \otimes \mathbf{K}\boldsymbol{\mathcal{B}}_{(1)} \hat{\boldsymbol{\Delta}}_{(1)}) \mathbf{P} \bar{\boldsymbol{\Sigma}}$, we have $\|\boldsymbol{\Sigma}_\Delta\|_{\text{op}} \leqslant \kappa_U \|\boldsymbol{K}\|_2$ and $\|\boldsymbol{\Sigma}_\Delta\|_{\text{F}} \leqslant \sqrt{T}\kappa_U \|\mathbf{K}\|_2$. Let $\kappa_U^k \asymp \kappa_U \|\mathbf{K}\|_2$, for any $t > 0$, by Hanson-Wright inequality, we can bound $\mathrm{H}_T - \mathbb{E}(\mathrm{H}_T)$ below,

$$\begin{aligned} \mathbb{P}\left[|\mathrm{H}_T - \mathbb{E}(\mathrm{H}_T)| \geqslant t\right] &\leqslant 2\exp\left(-C\min\left(\frac{t}{\kappa^2 \left\|\boldsymbol{\Sigma}_\Delta\right\|_{\text{op}}}, \frac{t^2}{\kappa^4 \left\|\boldsymbol{\Sigma}_\Delta\right\|_{\text{F}}^2}\right)\right) \\ &\leqslant 2\exp\left(-C\min\left(\frac{t}{\kappa^2 \kappa_U^k}, \frac{t^2}{\kappa^4 T {\kappa_U^k}^2}\right)\right). \end{aligned} \tag{S5.1}$$

Then we have the pointwise bound for fixed $\hat{\boldsymbol{\Delta}}$ and $\boldsymbol{B}$,

$$\mathbb{P}\left\{\left|\langle \frac{1}{T}\sum_{n=1}^T (\hat{\boldsymbol{\Delta}}_{(1)} \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K} - \hat{\boldsymbol{\Delta}}_{(1)}), \boldsymbol{\mathcal{B}}_{(1)} \rangle\right| \geqslant t\right\} \leqslant 2\exp\left(-C\min\left(\frac{Tt}{\kappa^2 \kappa_U^k}, \frac{Tt^2}{\kappa^4 {\kappa_U^k}^2}\right)\right). \tag{S5.2}$$

Following the proof of Lemma 2, let $\bar{\mathcal{S}}$ to be an ϵ-covering net of $\mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$. To construct the union bound, we rewrite H_T as $\mathrm{H}_T = \langle \sum_{n=1}^T \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K}, \hat{\boldsymbol{\Delta}}_{(1)}^\top \boldsymbol{\mathcal{B}}_{(1)} \rangle$ and let $\mathbf{X}^k = \sum_{n=1}^T \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K} - \mathbb{E}(\sum_{n=1}^T \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K})$.

By Lemma 1 (ii), then

$$\begin{aligned} &\mathbb{P}\left\{\max_{\boldsymbol{\Delta} \in \bar{\mathcal{S}}(2r_1,2r_2,2r_3,2r_4)} \max_{\boldsymbol{\mathcal{B}} \in \bar{\mathcal{S}}(R_1,R_2,R_3,R_4)} \frac{1}{T}\langle \mathbf{X}^k, \boldsymbol{\Delta}_{(1)}^\top \boldsymbol{\mathcal{B}}_{(1)} \rangle \geqslant t\right\} \\ &\geqslant \mathbb{P}\left\{\sup_{\boldsymbol{\Delta} \in \mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \max_{\boldsymbol{\mathcal{B}} \in \bar{\mathcal{S}}(R_1,R_2,R_3,R_4)} \frac{1}{T}\langle \mathbf{X}^k, \boldsymbol{\Delta}_{(1)}^\top \boldsymbol{\mathcal{B}}_{(1)} \rangle \geqslant t/(1-4\epsilon)\right\} \\ &\geqslant \mathbb{P}\left\{\sup_{\boldsymbol{\Delta} \in \mathcal{S}(2r_1,2r_2,2r_3,2r_4)} \sup_{\boldsymbol{\mathcal{B}} \in \bar{\mathcal{S}}(R_1,R_2,R_3,R_4)} \frac{1}{T}\langle \mathbf{X}^k, \boldsymbol{\Delta}_{(1)}^\top \boldsymbol{\mathcal{B}}_{(1)} \rangle \geqslant t/(1-4\epsilon)^2\right\}. \end{aligned}$$

Choose $\epsilon = 1/8$ and $\boldsymbol{\mathcal{B}} \in \mathcal{S}(R_1, R_2, R_3, R_4)$, we have that for all $\boldsymbol{\Delta} \in \mathcal{S}(2r_1, 2r_2, 2r_3, 2r_4)$,

$$\mathbb{P}\left\{\left|\langle \frac{1}{T}\sum_{n=1}^T (\hat{\boldsymbol{\Delta}}_{(1)} \mathbf{x}_n \mathbf{x}_n^\top \mathbf{K} - \hat{\boldsymbol{\Delta}}_{(1)}), \boldsymbol{\mathcal{B}}_{(1)} \rangle\right| \geqslant t\right\} \leqslant 2\exp\left(Cd_{\mathcal{M}}d_{\mathcal{B}} - C\min\left(\frac{Tt}{\kappa^2 \kappa_U^k}, \frac{Tt^2}{\kappa^4 {\kappa_U^k}^2}\right)\right).$$

This accomplishes the proof.

□

Lemma 6. *(Deviation bound of the estimated low-Tucker-rank linear form) Suppose Assumptions 1 and 2 hold, then*

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle\leqslant u\right)-\mathbb{P}\left(g\leqslant u\right)\right|\lesssim\frac{\sigma}{T^{1/9}}\|\mathbf{K}\|_2\|\boldsymbol{\mathcal{B}}_{(1)}\|_{\mathrm{F}}$$

where $g\sim\mathcal{N}(0,T^{-1}\mathrm{tr}(\boldsymbol{\mathcal{B}}_{(1)}\mathbf{K}\boldsymbol{\mathcal{B}}_{(1)}^\top\boldsymbol{\Sigma}_\varepsilon))$.

Proof. If Assumptions 1 and 2 hold, then the sequence $\{\mathbf{y}_n\}$ is α-mixing, and $\{\mathbf{x}_n\}$ is α-mixing. Since $\{\boldsymbol{\varepsilon}_n\}$ is independent of $\{\mathbf{x}_n\}$, $\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle$ is α-mixing by Theorem 5.2 of Bradley (2005). By Theorem 1 in Chang et al. (2024), and let $\|\boldsymbol{\mathcal{B}}_{(1)}\|_{\mathrm{F}}=1$, we have for all $\boldsymbol{\mathcal{B}}_{(1)}$,

$$\sup_{u\in\mathbb{R}}\left|\mathbb{P}\left(\frac{1}{\sqrt{T}}\sum_{n=1}^{T}\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle\leqslant u\right)-\mathbb{P}\left(g\leqslant u\right)\right|\lesssim\frac{\sigma}{T^{1/9}}\|\boldsymbol{K}\|_2$$

where $g\sim\mathcal{N}(0,\mathrm{tr}(\boldsymbol{\mathcal{B}}_{(1)}\mathbf{K}\boldsymbol{\mathcal{B}}_{(1)}^\top\boldsymbol{\Sigma}_\varepsilon))$, σ is the ϕ_2 norm of $\langle\boldsymbol{\varepsilon}_n\mathbf{x}_n^\top\mathbf{K},\boldsymbol{\mathcal{B}}_{(1)}\rangle$. □

S6 Simulation results for the MLR-TT-HAR model

The third experiment is for evaluating the non-asymptotic estimation error bound for MLR-TT-HAR models. The realized volatilities are generated using the model described by equations (4.1) and (4.2), and the coefficient tensor is generated from (2.5) with $P=22$. The coefficient tensor $\boldsymbol{\mathcal{A}}$ has the form $\boldsymbol{\mathcal{A}}=\boldsymbol{\mathcal{G}}\times\mathbf{U}_1\times\mathbf{U}_2\times\mathbf{U}_3\in\mathbb{R}^{N\times N\times P}$, where the core tensor $\boldsymbol{\mathcal{G}}$, the factor matrices $\mathbf{U}_i$ are generated by the similar method as in the first experiment.

For the MLR-TT-HAR model, $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}}-\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}=O_p(\sqrt{d_{\mathcal{M}}/T})$ with $d_{\mathcal{M}}=r_1r_2r_3+Nr_1+Nr_2+Pr_3$, and hence it is roughly linear with respect to T^{-1} and N, given

fixed values of r_1, r_2, r_3, P. We consider four settings to verify the relationship: (a) (P, r_1, r_2, r_3) is fixed at $(22, 2, 2, 2)$, the dimensionality is limited to $N = 10, 13, 15$, while the sample size T varies among the set of $\{550, 600, 650, 700, 750\}$ such that the values of T^{-1} are approximately and evenly spaced from 13×10^{-4} to 18×10^{-4}; (b) (P, r_1, r_2, r_3) is fixed at $(22, 2, 2, 3)$, with the dimensionality and sample size the same as in (a); (c) (P, r_1, r_2, r_3) is fixed at $(22, 2, 2, 2)$, the sample size is limited to $T = 100, 300, 500$, while the dimensionality N varies among the set to $\{8, 10, 13, 18, 30\}$ such that the values of $\sqrt{N}$ are approximately and evenly spaced from 2.8 to 5.5; (d) (P, r_1, r_2, r_3) is fixed at $(22, 2, 2, 3)$ with the dimensionality and sample size the same as in (c).

The step size, tolerance and initial values of Algorithm 1 are set as in the first experiment. Both Figure S6.1 and S6.2 displays the average estimation error $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ over 500 replications. Figure S6.1 illustrates that $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ exhibits an approximately linear relationship with respect to $\sqrt{T^{-1}}$, while Figure S6.2 implies that $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ is proportional to $\sqrt{N}$.

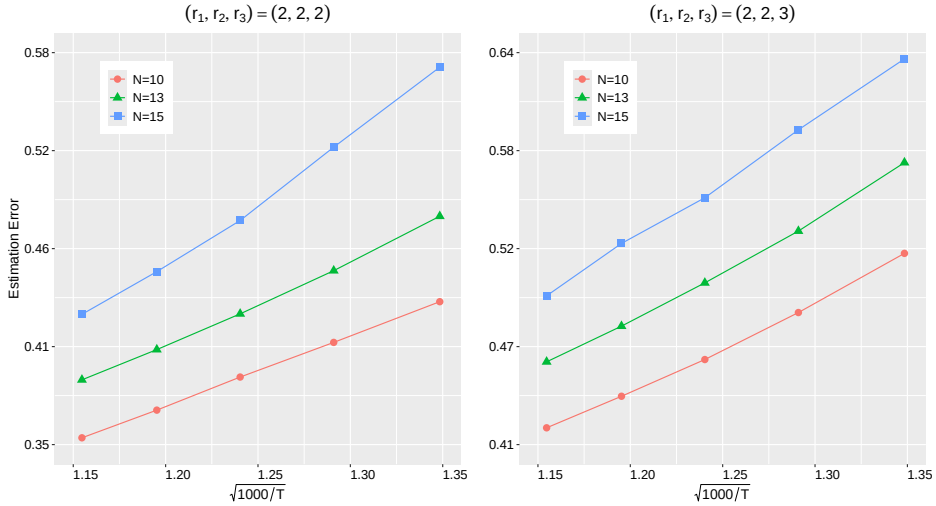

Figure S6.1: Estimation errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ against with $\sqrt{1000/T}$. The ranks are $(r_1, r_2, r_3) = (2, 2, 2)$ in the left panel, and $(r_1, r_2, r_3) = (2, 2, 3)$ in the right panel.

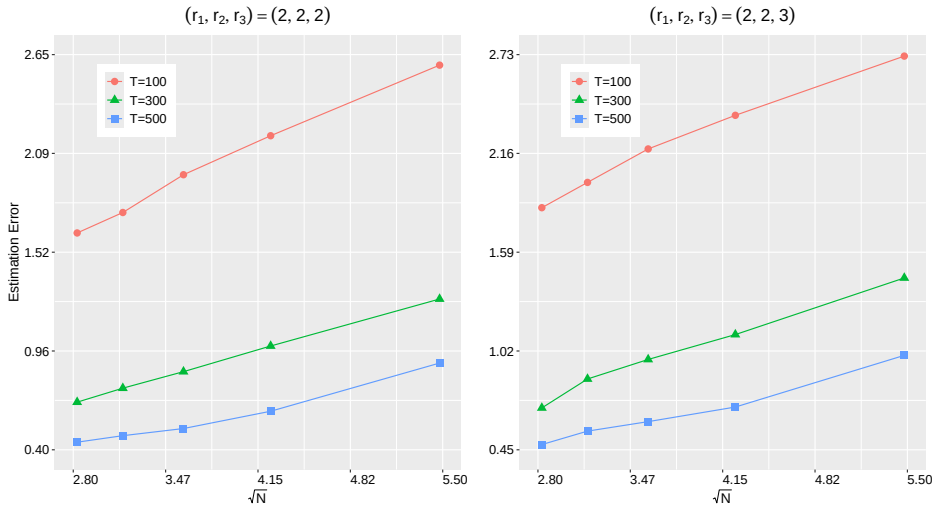

Figure S6.2: Estimation errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ against with $\sqrt{N}$. The ranks are $(r_1, r_2, r_3) = (2, 2, 2)$ in the left panel, and $(r_1, r_2, r_3) = (2, 2, 3)$ in the right panel.

Moreover, we report the estimation accuracy with the dimensionality N varies among the set to $\{8, 10, 13, 18, 30\}$ while holding T fixed at $100, 300, 500$, respectively.

The estimation accuracy is evaluated by the averaged mean squared error (MSE), calculated as $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}}-\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}^2$ over 500 replications. The corresponding runtime (RT) results of first 50 interations in a single replication are also presented. All the results are summarized in Table S6.1. Table S6.1 shows that the MSEs tend to increase as N increases, while they decrease as T increases, and RTs tend to increase as N or T increases.

Table S6.1: MSEs and RTs (seconds) with $N = 8, 10, 13, 18, 30$, while holding $T = 100, 300, 500$.

	$(r_1, r_2, r_3) = (2, 2, 2)$					$(r_1, r_2, r_3) = (2, 2, 3)$				
N	8	10	13	18	30	8	10	13	18	30
					$T = 100$					
MSE	2.71	3.11	3.90	4.82	6.74	3.45	4.00	4.81	5.69	7.43
RT	0.11	0.13	0.17	0.27	0.81	0.10	0.13	0.17	0.27	0.79
					$T = 300$					
MSE	0.46	0.58	0.73	1.00	1.60	0.48	0.75	0.96	1.26	2.10
RT	0.13	0.15	0.20	0.32	0.89	0.13	0.15	0.21	0.33	0.89
					$T = 500$					
MSE	0.20	0.24	0.28	0.40	0.81	0.23	0.32	0.38	0.49	0.99
RT	0.15	0.18	0.25	0.38	1.02	0.15	0.18	0.25	0.39	1.10

The fourth experiment aims to assess the convergence performance of Algorithm 1 for MLR-TT-HAR models. We generate a sample using a similar data generation process to the third experiment. The true Tucker ranks are $(r_1, r_2, r_3) = (2, 2, 2)$. Four different running ranks are considered: $(r_1^{'}, r_2^{'}, r_3^{'}) = (2, 2, 2)$, $(2, 2, 3)$, $(3, 2, 2)$, and $(3, 3, 3)$. Figure S6.3 illustrates the average standardized root mean square errors

$\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ over 500 replications for the first 150 iterations. The plot shows a similar decay pattern across all cases, indicating the convergence of the algorithm. Additionally, specifying more accurate ranks in advance leads to lower estimation errors, with the true ranks yielding the best results.

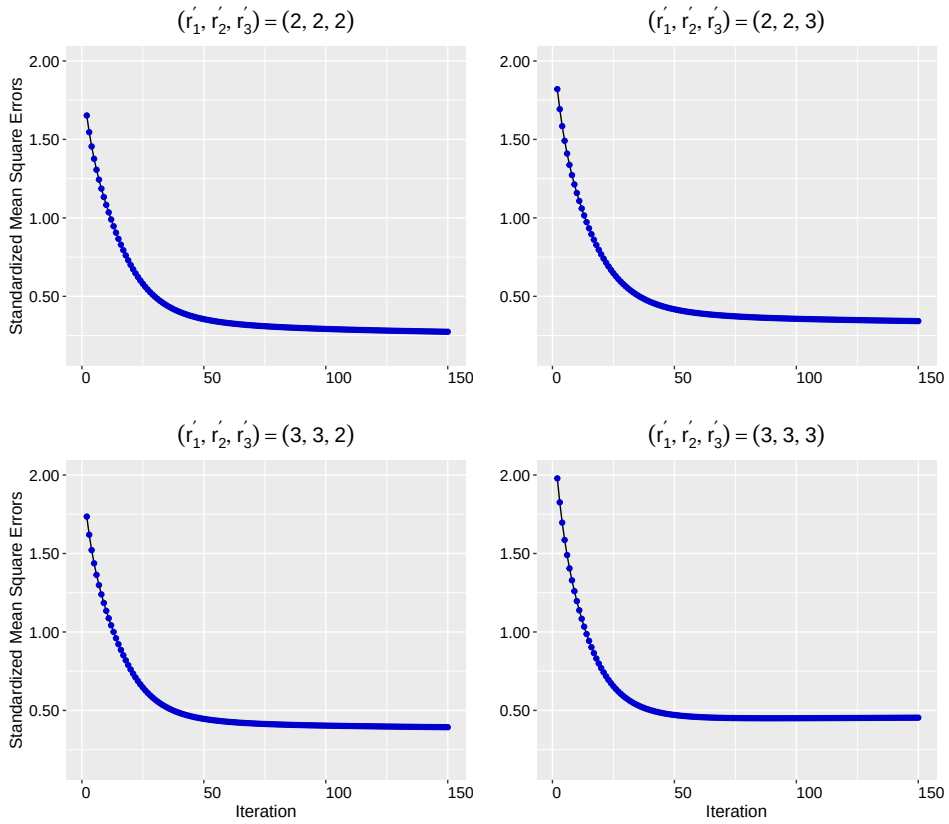

Figure S6.3: Standardized mean squares errors $\|\widehat{\boldsymbol{\mathcal{A}}}_{\mathrm{MLR}} - \boldsymbol{\mathcal{A}}\|_{\mathrm{F}}/\|\boldsymbol{\mathcal{A}}\|_{\mathrm{F}}$ for the first 150 iterations with running ranks $(r_1^{'}, r_2^{'}, r_3^{'}) = (2,2,2)$, $(2,2,3)$, $(3,3,2)$ or $(3,3,3)$.

S7 One Table for the selected ranks of the MLR-FT-HAR, MLR-TT-HAR and VHARI models in Real data analysis

Table S7.1 gives the results of the selected ranks of the MLR-FT-HAR, MLR-TT-HAR and VHARI models.

Table S7.1: Selected ranks of the VHARI, MLR-TT-HAR and MLR-FT-HAR models with 60 stocks and 90 stocks during the short (2011.01 – 2013.12) and long periods (2009.04 – 2013.12).

Models						
N	period	VHARI		MLR-TT-HAR		MLR-FT-HAR
		$P = 22$	$P = 66$	$P = 22$	$P = 66$	$Q = 22$, $S = 3$
60	short	$r = 2$	$r = 2$	$(r_1, r_2, r_3) = (2, 2, 2)$	$(r_1, r_2, r_3) = (2, 2, 2)$	$(r_1, r_2, r_3, r_4) = (3, 3, 1, 2)$
	long	$r = 3$	$r = 3$	$(r_1, r_2, r_3) = (4, 4, 2)$	$(r_1, r_2, r_3) = (3, 3, 2)$	$(r_1, r_2, r_3, r_4) = (3, 3, 1, 4)$
90	short	$r = 2$	$r = 2$	$(r_1, r_2, r_3) = (3, 3, 2)$	$(r_1, r_2, r_3) = (3, 3, 2)$	$(r_1, r_2, r_3, r_4) = (3, 3, 1, 2)$
	long	$r = 3$	$r = 3$	$(r_1, r_2, r_3) = (3, 3, 2)$	$(r_1, r_2, r_3) = (3, 3, 4)$	$(r_1, r_2, r_3, r_4) = (3, 3, 1, 4)$

Bibliography

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

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys 2*, 107–144.

Candès, E. J. and Y. Plan (2011). Tight oracle inequalities for low-rank matrix recov-

ery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory 57*, 2342–2359.

Chang, J., X. Chen, and M. Wu (2024). Central limit theorems for high dimensional dependent data. *Bernoulli 30*, 712–742.

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

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

Jain, P., A. Tewari, and P. Kar (2014). On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems 27*.

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

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

Wang, D., Y. Zheng, H. Lian, and G. Li (2022). High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association 117*, 1338–1356.