Estimation and Inference for CP Tensor Factor Models*

Bin Chen[†1], Yuefeng Han [‡2], and Qiyang Yu[§1]

[1]University of Rochester
[2]University of Notre Dame

June, 2024

Abstract

High-dimensional tensor-valued data have recently gained attention from researchers in economics and finance. We consider the estimation and inference of high-dimensional tensor factor models, where each dimension of the tensor diverges. Our focus is on a factor model that admits CP-type tensor decomposition, which allows for non-orthogonal loading vectors. Based on the contemporary covariance matrix, we propose an iterative simultaneous projection estimation method. Our estimator is robust to weak dependence among factors and weak correlation across different dimensions in the idiosyncratic shocks. We establish an inferential theory, demonstrating both consistency and asymptotic normality under relaxed assumptions. Within a unified framework, we consider two eigenvalue ratio-based estimators for the number of factors in a tensor factor model and justify their consistency. Through a simulation study and two empirical applications featuring sorted portfolios and international trade flows, we illustrate the advantages of our proposed estimator over existing methodologies in the literature.

JEL Classifications: *C13, C32, C55*

Keywords: Asymptotic normality, Canonical Polyadic Decompositions, Factor models, High-dimensional, Tensor data.

*We thank Yoosoon Chang, Rong Chen, Joon Park, Mahrad Sharifvaghefi and seminar participants at Indiana University, University of Pittsburgh, University of Rochester, the 2024 Econometric Society Summer Meeting, and the workshop on Analysis of Complex Data: Tensors, Networks and Dynamic Systems for their useful comments and discussions. Any remaining errors are solely ours.

[†]binchen@rochester.edu
[‡]yuefeng.han@nd.edu
[§]qyu13@ur.rochester.edu

1 Introduction

Factor models have become one of the most popular tools for summarizing and extracting information from high-dimensional data in economics and finance (Fan et al. (2021), Bai and Wang (2016), Stock and Watson (2016)). Traditional factor models are designed to manage large panel data, where both cross-sectional and time series dimensions increase. These models admit a low-rank structure and have a common-idiosyncratic decomposition, allowing for the identification of significant variations within the panel of economic data.

In modern economics, researchers increasingly encounter vast, multi-dimensional datasets, or tensor. For example, monthly import-export volume time series spanning various product categories among countries can be represented as a three-dimensional tensor, with unavailable diagonal elements for each product category. Similarly, in portfolio selection, data often involve stock prices and various firm characteristics over time across different firms, forming a two-dimensional tensor. Additionally, macroeconomic studies on growth and productivity analyze multiple macro variables at the country-industry level, enabling cross-country comparative analyses, which are challenging with traditional panel data.

Statistical methods and economic applications for the high-dimensional tensor factor analysis are still in their early stages of development. As in the classical panel setting, tensor factor models typically assume low-rank structures, with Canonical Polyadic (CP) and Tucker structures being the most common choices (see, e.g., Kolda and Bader (2009)). Recent studies have explored various estimation approaches and extensions. For example, working with Tucker decomposition, Chen et al. (2022) considered two estimators based on the autocovariance matrices, while Han et al. (2022a) extended these methods using an iterative procedure with the matrix unfolding mechanism. Chen and Fan (2023) proposed an estimation method called α-PCA that preserves the matrix structure and aggregates mean and contemporary covariance through a hyper-parameter α. Chen and Lam (2024) introduced a pre-averaging technique for the Tucker tensor factor model that significantly enhances the model's inherent signal strength under certain conditions. Chen et al. (2024) introduced a semiparametric tensor factor model leveraging mode-wise covariates. In the context of CP decomposition, Han et al. (2023) proposed an iterative simultaneous orthogonalization algorithm with warm-start initialization, while Babii et al. (2023) employed tensor principal component analysis (TPCA), assuming orthogonal factor loadings. Chang et al. (2023) developed estimation procedure based on a generalized eigenanalysis constructed from the serial dependence structure of the underlying process.

In this paper, we focus on a tensor factor model with a CP low-rank structure due to its parsimonious features. We propose an iterative projection estimation based on contemporary covariance rather than autocovariance matrices. As highlighted by Chen and Fan (2023), autocovariance-based methods rely on the assumption of non-zero autocovariances among individual factors, limiting their effectiveness in scenarios with serially independent factors or weak autocorrelations in tensor data.

We develop inferential theory, establishing consistency, convergence rates, and limiting distributions under relaxed assumptions. Additionally, we extend the eigenvalue ratio-based estimator (Ahn and Horenstein (2013)) for latent dimensions to tensor factor models and show estimation consistency.

The remaining sections of this paper are organized as follows. Section 2 introduces the high-dimensional tensor factor model with a CP low rank structure allowing for nonorthogonal loading vectors. In Section 3, we present an iterative projection estimation procedure and two generalized eigenvalue ratio-based estimators for the number of the latent factors. Section 4 establishes the consistency and limiting distributions of the estimated loading vectors. We assess the finite sample performance through simulation in Section 5 and provide two empirical applications in Section 6. Finally, Section 7 concludes the paper with all mathematical proofs included in the Appendix.

1.1 Notations and preliminaries

In this subsection, we introduce essential notations and basic tensor operations. For an in-depth review, readers may refer to Kolda and Bader (2009).

Let $\|x\|_q = (x_1^q + ... + x_p^q)^{1/q}$, $q \geqslant 1$, for any vector $x = (x_1, ..., x_p)^\top$. We employ the following matrix norms: matrix spectral norm $\|M\|_2 = \max_{\|x\|_2=1, \|y\|_2=1} \|x^\top M y\|_2 = \sigma_1(M)$, where $\sigma_1(M)$ is the largest singular value of M. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there exists a constant C such that $|a_n| \leqslant C|b_n|$ holds for all sufficiently large n, and $a_n \lesssim b_n$ if there exists a constant C such that $a_n \leqslant C b_n$.

Consider two tensors $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}, \mathcal{B} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_N}$. The tensor product $\otimes$ is defined as $\mathcal{A} \otimes \mathcal{B} \in \mathbb{R}^{d_1 \times \cdots \times d_K \times r_1 \times \cdots \times r_N}$, where

$$(\mathcal{A} \otimes \mathcal{B})_{i_1,...,i_K,j_1,...,j_N} = (\mathcal{A})_{i_1,...,i_K} (\mathcal{B})_{j_1,...,j_N}.$$

The k-mode product of $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$ with a matrix $U \in \mathbb{R}^{m_k \times d_k}$ is an order K tensor of dimension $d_1 \times \cdots \times d_{k-1} \times m_k \times d_{k+1} \times \cdots \times d_K$, denoted as $\mathcal{A} \times_k U$, where

$$(\mathcal{A} \times_k U)_{i_1,...,i_{k-1},j,i_{k+1},...,i_K} = \sum_{i_k=1}^{d_k} \mathcal{A}_{i_1,i_2,...,i_K} U_{j,i_k}.$$

The mode-k matricization of a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is denoted as $\mathrm{mat}_k(\mathcal{A}) \in \mathbb{R}^{d_k \times d_{-k}}$, where $d = \prod_{j=1}^K d_j$ and $d_{-k} = d/d_k = \prod_{j=1, j\neq k}^K d_j$. It is obtained by setting the k-th tensor mode as its rows and collapsing all the others into its columns. And the vectorization of the matrix/tensor $\mathcal{A}$ is denoted as $\mathrm{vec}(\mathcal{A}) \in \mathbb{R}^d$. Note that $\mathrm{mat}_k(\mathrm{vec}(\mathcal{A})) = \mathrm{mat}_k(\mathcal{A})$.

2 Model

We consider a tensor-valued time series $\mathcal{Y}_t \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_K}$, where $1 \leqslant t \leqslant T$. Our focus is on a tensor factor model with a CP low-rank structure:

$$\mathcal{Y}_t = \sum_{i=1}^{r} f_{it} \left(a_{i1} \otimes a_{i2} \cdots \otimes a_{iK} \right) + \mathcal{E}_t, \quad t \leqslant T, \tag{1}$$

where $\otimes$ denotes the tensor product, f_{it} is a one-dimensional latent factor, a_{ik} denotes the d_k-dimensional loading vector, which needs not to be orthogonal. Unlike Han et al. (2023), we permit arbitrary correlation structures among individual factors. Without loss of generality, we assume $\|a_{ik}\|_2 = 1$, for all $1 \leqslant i \leqslant r$ and $1 \leqslant k \leqslant K$. The noise tensor $\mathcal{E}_t$ is assumed to be uncorrelated with the latent factors but may exhibit weak correlations across different dimensions. The rank r may either be fixed or divergent.

When $K = 1$, $\mathcal{Y}_t$ reduces to a vector, and model (1) becomes the classical factor model, extensively studied in the literature (Bai and Ng (2002) and Stock and Watson (2002)). For $K > 1$, an alternative approach is to vectorize data:

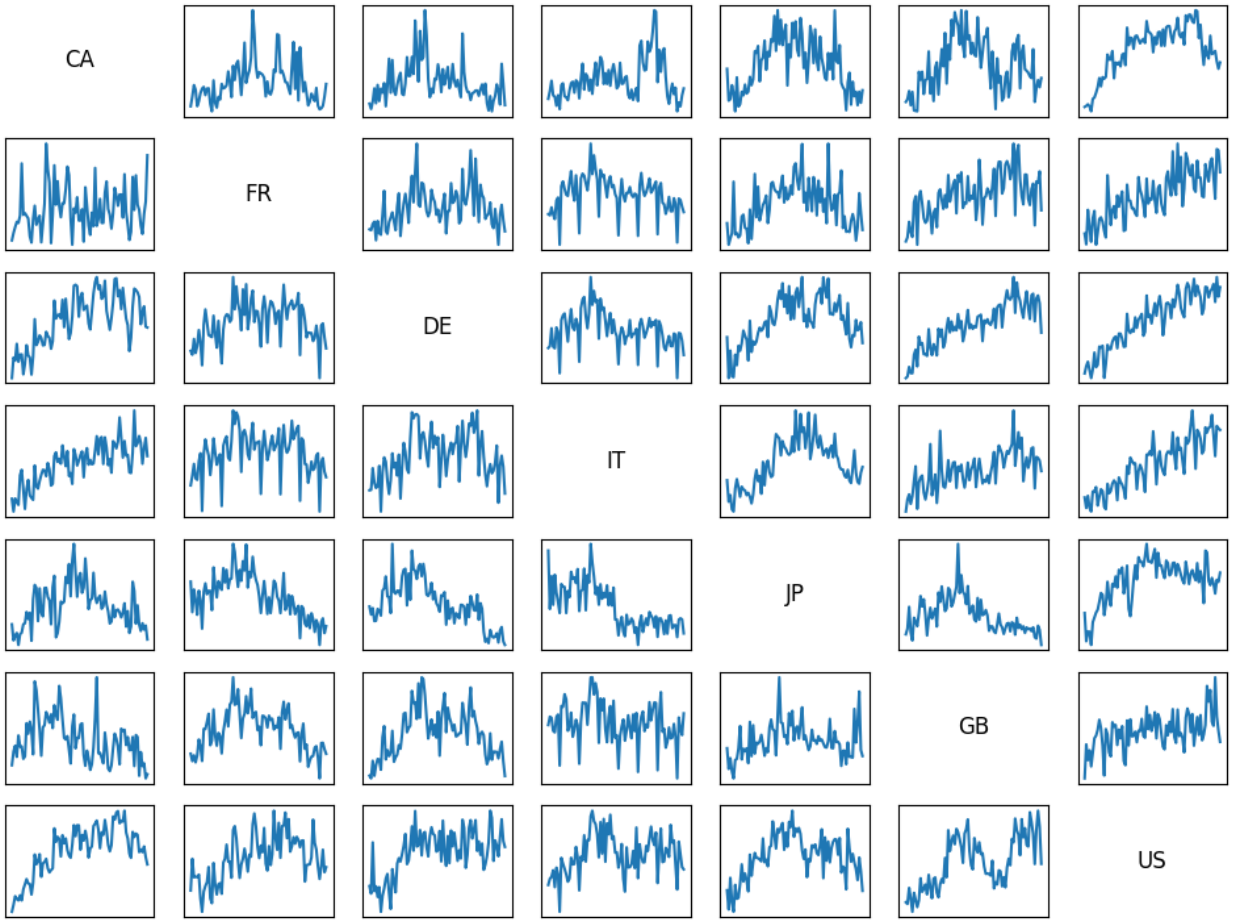
$$\text{vec}(\mathcal{Y}_t) = \Xi F_t + \text{vec}(\mathcal{E}_t), \tag{2}$$

where $\text{vec}(\mathcal{Y}_t) \in \mathbb{R}^d$ with $d = d_1 d_2 \cdots d_K$ and $F_t = (f_{1t}, f_{2t}, \cdots, f_{rt})^\top \in \mathbb{R}^r$. However, this method ignores the tensor structure of the data and hence substantially increases the number of parameters in the loading matrices from $(d_1 + d_2 + \cdots + d_K)r$ in the tensor case to $(d_1 d_2 \cdots d_K)r$ in the stacked vector version. Our proposed approach, modeling $\text{vec}(\mathcal{Y}_t)$ as $AF_t + \text{vec}(\mathcal{E}_t)$, $A = (a_1, \cdots, a_r)$ and $a_i = \text{vec}(a_{i1} \otimes a_{i2} \otimes \cdots \otimes a_{iK})$, within our specific framework, yields improved convergence rates due to its unique structure.

Consider an illustrative example of international trade flows, detailed in Section 6. The observed $\mathcal{Y}_t$ forms a square matrix, where $d_1 = d_2 = n$ and $K = 2$. Each entry $\mathcal{Y}_{t,ij}$ of $\mathcal{Y}_t$, with $i, j = 1, 2, \cdots, n$, represents the volume of trade flow from country i to country j at time t. Thus, the ith row represents data where country i is the exporter, while the jth column represents data where country j is the importer. Figure 1 shows a time series plot of $\mathcal{Y}_t$ for G7 countries excluding EU spanning from January 2008 to December 2014. Model (1) identifies r latent factors, analogous to r trading hubs. Each country exports to these hubs with certain distributions (determined by the loading matrix A_1) and imports from them likewise (determined by the loading matrix A_2). The element $a_{1,il}$ of A_1, where $i = 1, \cdots, r$ and $l = 1, \cdots, n$, represents the export contribution of country l to trading hub i. Similarly, the entry $a_{2,jm}$ of A_2, where $j = 1, \cdots, r$ and $m = 1, \cdots, n$, can be interpreted as the import contribution of country m to trading hub j. We allow the number of trading hubs r to increase with the increase of n and T.

In the literature, an alternative tensor factor model based on Tucker decomposition has been

Figure 1: Time series plots of the value of goods traded among G7 countriesU



Notes: (1) sample period: January 2008- December 2014. (2) The plots only show the patterns of the time series while the magnitudes are not comparable between plots because the ranges of the y-axis are different.

explored (see, e.g., Han et al. (2022a), Wang and Lu (2017), Lettau (2023)):

$$\mathcal{Y}_t = \mathcal{F}_t \times_1 A_1 \times \cdots \times A_K + \mathcal{E}_t, \tag{3}$$

where the core tensor $\mathcal{F}_t \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the latent factor process in a tensor form, and A_i's are $d_i \times r_i$ loading matrices. As discussed in Babii et al. (2023) and Han et al. (2023), unlike the CP decomposition, the Tucker decomposition is generally non-unique, leading to significant identification issues. Consequently, estimation results from model (3) may exhibit ambiguity, undermining meaningful discussions of individual factors (Stock and Watson (2002)). In contrast, the CP tensor factor model (1) yields a unique set of one-dimensional latent factors, which serve as natural inputs for diffusion index forecasts and factor-augmented regressions (Bai and Ng (2006)). We regard the CP tensor factor as a more parsimonious yet flexible and effective alternative. Further comparison of the performance of these two tensor factor models will be presented in Section 6.

3 Estimation

We consider a two-step estimation procedure to derive the loading vectors and latent factors. This approach begins with initialization through randomized composite PCA, followed by an iterative refinement step utilizing an iterative simultaneous orthogonalization procedure.

We start by defining the contemporary covariance as the expected value of the outer product of $\mathcal{Y}_t$:

$$\begin{aligned} \Sigma &= \mathbb{E}\left[\mathcal{Y}_t \otimes \mathcal{Y}_t\right] \\ &= \sum_{i,j=1}^{r} \Theta_{ij} \otimes_{l=1}^{K} a_{il} \otimes_{l=1}^{K} a_{jl} + \mathbb{E}\left[\mathcal{E}_t \otimes \mathcal{E}_t\right], \end{aligned} \tag{4}$$

where $\Theta_{ij} = \mathbb{E}\left[f_{it} f_{jt}\right]$. Its sample analogue, denoted as $\widehat{\Sigma}$, is computed as the average outer product over T observations:

$$\widehat{\Sigma} = \sum_{t=1}^{T} \frac{\mathcal{Y}_t \otimes \mathcal{Y}_t}{T}. \tag{5}$$

We aim to estimate the loading vectors by minimizing the empirical quadratic loss, formulated as:

$$\min_{\substack{a_{i1}, a_{i2}, \ldots, a_{iK}, 1 \leqslant i \leqslant r, \\ \|a_{i1}\|_2 = \ldots = \|a_{iK}\|_2 = 1}} \left\| \widehat{\Sigma} - \sum_{i,j=1}^{r} \Theta_{ij} \otimes_{l=1}^{K} a_{il} \otimes_{l=1}^{K} a_{jl} \right\|_{\mathrm{F}}^2, \tag{6}$$

where $\|\mathcal{A}\|_F$ denotes the Frobenius norm of a tensor $\mathcal{A}$. However, this optimization problem is non-convex and prone to multiple local optima. To counter this problem, we employ a two-step approach. The first step focuses on obtaining a suitable initialization close to the global optimum.

The contemporary covariance Σ in (4) can be unfolded to a $d \times d$ matrix

$$\Sigma_0 = A\Theta A^\top, \tag{7}$$

where $\Theta = \mathbb{E}F_t F_t^\top$, $F_t = (f_{1t}, \cdots, f_{rt})^\top$. This unfolding enables classical PCA estimation if the columns of the loading matrix A are orthogonal. Our framework accommodates general non-orthogonal a_i's and hence the PCA procedure introduces a bias component, which motivates the second stage refinement. The accuracy of the PCA estimator hinges on the maximum correlation among the loading vectors. When the additional orthogonality condition is imposed as in Babii et al. (2023), the maximum correlation reduces to 0 and hence bias disappears. The first step, termed intializaiton via randomized composite PCA, is detailed in Algorithm 1.

To further relax the eigengap assumption imposed in Babii et al. (2023) and Han et al. (2023), we incorporate randomized projection into our composite PCA approach (Procedure 2). Random projection, also known as random slicing (Anandkumar et al., 2014b; Sun et al., 2017) is a well-recognized initialization method in noiseless tensor CP decomposition, which accommodates repeated eigenvalues. We extend this approach to the tensor CP factor model.

Algorithm 1: Initialization via Randomized Composite PCA

Input : The observations $\mathcal{Y}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t = 1, ..., T$, the number of factors r, small constant $0 < c_0 < 1$.

Evaluate $\widehat{\Sigma}$ in (5), and unfold it to $d \times d$ matrix $\widetilde{\Sigma}$.

Obtain $\widehat{\lambda}_i, \widehat{u}_i, 1 \leqslant i \leqslant r$, the top r eigenvalues and eigenvectors of $\widetilde{\Sigma}$. Set $\widehat{\lambda}_0 = \infty$ and $\widehat{\lambda}_{r+1} = 0$.

if $\min\{|\widehat{\lambda}_i - \widehat{\lambda}_{i-1}|, |\widehat{\lambda}_i - \widehat{\lambda}_{i+1}|\} > c_0\widehat{\lambda}_r$ **then**

  Compute $\widehat{a}_{ik}^{\text{rcpca}}$ as the top left singular vector of $\text{mat}_k(\widehat{u}_i) \in \mathbb{R}^{d_k \times (d/d_k)}$, for all $1 \leqslant k \leqslant K$.

else

  Form disjoint index sets $I_1, ..., I_N$ from all contiguous indices $1 \leqslant i \leqslant r$ that do not satisfy the above criteria of the eigengap.

  For each I_j, form $d \times d$ matrix $\widetilde{\Sigma}_j = \sum_{\ell \in I_j} \widehat{\lambda}_\ell \widehat{u}_\ell \widehat{u}_\ell^\top$, and formulate it into a tensor $\widehat{\Sigma}_j \in \mathbb{R}^{d_1 \times \cdots \times d_K \times d_1 \times \cdots \times d_K}$. Then run Procedure 2 on $\widehat{\Sigma}_j$ to obtain $\widehat{a}_{ik}^{\text{rcpca}}$ for all $i \in I_j, 1 \leqslant k \leqslant K$.

Output: Warm initialization $\widehat{a}_{ik}^{\text{rcpca}}, 1 \leqslant i \leqslant r, 1 \leqslant k \leqslant K$

Following initialization, we refine the estimation using an iterative simultaneous orthogonalization procedure (Algorithm 3). This step aims to enhance estimation accuracy and extract latent factors. The procedure is motivated by the vector factor structure of the denoised $\mathcal{Y}_t$:

$$\mathcal{Z}_{t,ik} = f_{it} a_{ik} + \mathcal{V}_{t,ik}, \tag{8}$$

Procedure 2: Randomized Projection

Input : Noisy tensor $\Xi \in \mathbb{R}^{d_1 \times \cdots \times d_K \times d_1 \times \cdots \times d_K}$, rank s, number of random projections L, tuning parameter ν.

for $\ell = 1$ *to* L **do**

- Randomly draw a $d_1 \times d_1$ Gaussian matrix θ whose entries are i.i.d. $N(0,1)$.
- Compute $\Xi \times_1 \times_{K+1} \theta$ and compute its leading singular value and left singular vector $\eta_\ell, \widetilde{u}_\ell$.
- Compute $\widetilde{a}_{\ell k}$ as the top left singular vector of $\mathrm{mat}_k(\widetilde{u}_\ell) \in \mathbb{R}^{d_k \times (d/(d_k d_1))}$, for all $2 \leqslant k \leqslant K$.
- Compute $\widetilde{a}_{\ell 1}$ as the top left singular vector of $\Xi \times_{k=2}^{K} \widetilde{a}_{\ell k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell,k-K}$.
- Add the tuple $(\widetilde{a}_{\ell k}, 1 \leqslant k \leqslant K)$ to $\mathcal{S}_L$.

for $i = 1$ *to* s **do**

- Among the remaining tuples in $\mathcal{S}_L$, choose one tuple $(\widetilde{a}_{\ell k}, 1 \leqslant k \leqslant K)$ that correspond to the largest $\|\Xi \times_{k=1}^{K} \widetilde{a}_{\ell k} \times_{k=K+1}^{2K} \widetilde{a}_{\ell,k-K}\|_2$. Set it to be $\widehat{a}_{ik}^{\mathrm{rcpca}} = \widetilde{a}_{\ell k}$.
- Remove all the tuples with $\max_{1 \leqslant k \leqslant K} |\widetilde{a}_{\ell' k}^\top \widehat{a}_{ik}^{\mathrm{rcpca}}| > \nu$.

Output: Warm initialization $\widehat{a}_{ik}^{\mathrm{rcpca}}, 1 \leqslant i \leqslant s, 1 \leqslant k \leqslant K$

where

$$\mathcal{Z}_{t,ik} = \mathcal{Y}_t \times_1 b_{i1}^\top \times_2 \cdots \times_{k-1} b_{i,k-1}^\top \times_{k+1} b_{i,k+1}^\top \times_{k+2} \cdots \times_K b_{iK}^\top, \tag{9}$$

$$\mathcal{V}_{t,ik} = \mathcal{E}_t \times_1 b_{i1}^\top \times_2 \cdots \times_{k-1} b_{i,k-1}^\top \times_{k+1} b_{i,k+1}^\top \times_{k+2} \cdots \times_K b_{iK}^\top, \tag{10}$$

$B_k = A_k(A_k^\top A_k)^{-1} = (b_{1k}, ..., b_{rk}) \in \mathbb{R}^{d_k \times r}$, $A_k = (a_{1k}, \ldots, a_{rk}) \in \mathbb{R}^{d_k \times r}$, and we have used the fact that b_{ik} is orthogonal to all $a_{jk}, j \neq i$ by construction. Note that the orthogonalization projection, which takes place in all except the kth mode simultaneously in each computational iteration, transforms the tensor $\mathcal{Y}_t$ to a $d_k \times 1$ vector, reducing dimensions and noise substantially. This transformation enables easy and accurate estimation of the classical vector factor model in equation (8).

In practice, we don't observe b_{ik} and iterations can be applied to update the estimations. Given the previous estimates $\widehat{a}_{ik}^{(m-1)}$, where m is the iteration number, $\mathcal{Y}_t$ can be denoised via

$$\mathcal{Z}_{t,ik}^{(m)} = \mathcal{Y}_t \times_1 \widehat{b}_{i1}^{(m)\top} \times_2 \cdots \times_{k-1} \widehat{b}_{i,k-1}^{(m)\top} \times_{k+1} \widehat{b}_{i,k+1}^{(m-1)\top} \times_{k+2} \cdots \times_K \widehat{b}_{iK}^{(m-1)\top},$$

for $t = 1, ..., T$, and consequently, updated loading vectors $\widehat{a}_{ik}^{(m)}$ are obtained through eigenanalysis based on the contemporary covariance $\widehat{\Sigma}(\mathcal{Z}_{1:T,ik}^{(m)}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{Z}_{t,ik}^{(m)} \mathcal{Z}_{t,ik}^{(m)\top}$. The iteration continues until convergence or the maximum number of iterations is reached.

The above estimation procedure assumes that the rank r is known. However, we need to estimate r in practice. We consider two estimation procedures based on the eigenvalue ratio method proposed by Ahn and Horenstein (2013).

For the first procedure, we unfold the sample contemporary covariance $\widehat{\Sigma}$ in (5) to a $d \times d$ matrix

Algorithm 3: Iterative Simultaneous Orthogonalization (ISO)

Input : The observations $\mathcal{Y}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t = 1, ..., T$, the number of factors r, the warm-start initial estimates $\widehat{a}_{ik}^{(0)}$, $1 \leqslant i \leqslant r$ and $1 \leqslant k \leqslant K$, the tolerance parameter $\epsilon > 0$, and the maximum number of iterations M.

Compute $\widehat{B}_k^{(0)} = \widehat{A}_k^{(0)}(\widehat{A}_k^{(0)\top}\widehat{A}_k^{(0)})^{-1} = (\widehat{b}_{1k}^{(0)}, ..., \widehat{b}_{rk}^{(0)})$ with $\widehat{A}_k^{(0)} = (\widehat{a}_{1k}^{(0)}, \ldots, \widehat{a}_{rk}^{(0)}) \in \mathbb{R}^{d_k \times r}$ for $k = 1, \ldots, K$. Set $m = 0$.

repeat

  Let $m = m + 1$.

  for $k = 1$ *to* K **do**

    for $i = 1$ *to* r **do**

      Given previous estimates $\widehat{a}_{ik}^{(m-1)}$, calculate

$$\mathcal{Z}_{t,ik}^{(m)} = \mathcal{Y}_t \times_1 \widehat{b}_{i1}^{(m)\top} \times_2 \cdots \times_{k-1} \widehat{b}_{i,k-1}^{(m)\top} \times_{k+1} \widehat{b}_{i,k+1}^{(m-1)\top} \times_{k+2} \cdots \times_K \widehat{b}_{iK}^{(m-1)\top},$$

      for $t = 1, ..., T$. Let $\widehat{\Sigma}\left(\mathcal{Z}_{1:T,ik}^{(m)}\right) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{Z}_{t,ik}^{(m)} \mathcal{Z}_{t,ik}^{(m)\top}$.

      Compute $\widehat{a}_{ik}^{(m)}$ as the top eigenvector of $\widehat{\Sigma}(\mathcal{Z}_{1:T,ik}^{(m)})$.

    Compute $\widehat{B}_k^{(m)} = \widehat{A}_k^{(m)}(\widehat{A}_k^{(m)\top}\widehat{A}_k^{(m)})^{-1} = (\widehat{b}_{1k}^{(m)}, ..., \widehat{b}_{rk}^{(m)})$ with $\widehat{A}_k^{(m)} = (\widehat{a}_{1k}^{(m)}, \ldots, \widehat{a}_{rk}^{(m)})$.

until $m = M$ *or* $\max_{1 \leqslant i \leqslant r} \max_{1 \leqslant k \leqslant K} \|\widehat{a}_{ik}^{(m)}\widehat{a}_{ik}^{(m)\top} - \widehat{a}_{ik}^{(m-1)}\widehat{a}_{ik}^{(m-1)\top}\|_2 \leqslant \epsilon$;

Output: Estimates

$$\begin{aligned}
\widehat{a}_{ik}^{\text{iso}} &= \widehat{a}_{ik}^{(m)}, \quad i = 1, ..., r, \quad k = 1, ..., K, \\
\widehat{f}_{it} &= \mathcal{Y}_t \times_{k=1}^{K} \widehat{b}_{ik}^{(m)\top}, \quad i = 1, ..., r, \quad t = 1, ..., T, \\
\widehat{\mathcal{Y}}_t &= \sum_{i=1}^{r} \widehat{f}_{it} \otimes_{k=1}^{K} \widehat{a}_{ik}^{(m)}, \quad t = 1, ..., T.
\end{aligned}$$

$\widetilde{\Sigma}$. Let $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_r \geqslant 0$ be the ordered eigenvalues of $\widetilde{\Sigma}$. The CP tensor factor model (1) can also be adapted to a vector factor model (2) with the same number of factors r. Thus, the eigenvalue ratio-based estimator derived from the unfolded covariance matrix $\widetilde{\Sigma}$ can be defined as

$$\widehat{r}^{\text{uer}} = \underset{1 \leqslant i \leqslant r_{\max}}{\arg\max} \frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}}, \tag{11}$$

and $r_{\max}$ is a selected upper bound.

Alternatively, we can define the mode-k covariance with the inner product:

$$\widehat{\Sigma}_k = \sum_{t=1}^{T} \frac{\text{mat}_k(\mathcal{Y}_t)\,\text{mat}_k^{\top}(\mathcal{Y}_t)}{T} \in \mathbb{R}^{d_k \times d_k}.$$

Let $\widehat{\lambda}_{1k} \geqslant \widehat{\lambda}_{2k} \geqslant \cdots \geqslant \widehat{\lambda}_{rk} \geqslant 0$ be the ordered eigenvalues of $\widehat{\Sigma}_k$. The eigenvalue ratio-based estimator using the inner product can be defined as

$$\widehat{r}^{\text{ip}} = \max(\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_K), \tag{12}$$

where $\widehat{r}_k = \arg\max_{1 \leqslant i \leqslant r_{\max}} \frac{\widehat{\lambda}_{ik}}{\widehat{\lambda}_{i+1,k}}$. We have adopted the setup in the CP tensor factor model (1) where the number of spiked eigenvalues r_k remains constant across different mode-k covariance. Further details of these two procedures can be found in Algorithms 4 and 5.

Algorithm 4: Unfolded Eigenvalue Ratio Method

Input : The observations $\mathcal{Y}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t = 1, ..., T$, the upper bound of the number of factors $r_{\max}$.

1 Evaluate $\widehat{\Sigma}$ in (5), and unfold it to $d \times d$ matrix $\widetilde{\Sigma}$, i.e. $\widetilde{\Sigma} = \text{mat}_{[K]}(\widehat{\Sigma})$.

2 Obtain $\widehat{\lambda}_i, 1 \leqslant i \leqslant r_{\max} + 1$, the top $r_{\max} + 1$ eigenvalues of $\widetilde{\Sigma}$.

3 Obtain $\widehat{r}^{\text{uer}}$ by

$$\widehat{r}^{\text{uer}} = \underset{1 \leqslant i \leqslant r_{\max}}{\arg\max} \frac{\widehat{\lambda}_i}{\widehat{\lambda}_{i+1}}.$$

Output: Estimate of the number of factors $\widehat{r}^{\text{uer}}$.

It is noteworthy that Han et al. (2023) explore a similar tensor CP factor model as (1), albeit within a distinct setting where latent factors are assumed uncorrelated and noise follows a white noise process. Methodologically, their approach rely on the autocovariance between $\mathcal{Y}_{t-h}$ and $\mathcal{Y}_t$, where $h \geqslant 1$, whereas our method employs contemporaneous covariance. The autocovariance-based method may not be ideal for datasets with low temporal dependence, such as asset return data, which often exhibit minimal serial correlation possibly due to market efficiency.

Another closely related approach is tensor PCA proposed in Babii et al. (2023). They consider a CP tensor factor model with orthogonal loading vectors. Unlike Tucker factor models, the identifica-

Algorithm 5: Eigenvalue Ratio Method through Inner Product

Input : The observations $\mathcal{Y}_t \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $t = 1, ..., T$, the upper bound of the number of factors $r_{\max}$.

for $k = 1$ *to* K **do**

Evaluate

$$\widehat{\Sigma}_k = \sum_{t=1}^{T} \frac{\mathrm{mat}_k(\mathcal{Y}_t)\,\mathrm{mat}_k^\top(\mathcal{Y}_t)}{T} \in \mathbb{R}^{d_k \times d_k}.$$

Obtain $\widehat{\lambda}_{ik}, 1 \leqslant i \leqslant r_{\max}+1$, the top $r_{\max}+1$ eigenvalues of $\widehat{\Sigma}_k$.

Obtain $\widehat{r}_k$ by

$$\widehat{r}_k = \underset{1 \leqslant i \leqslant r_{\max}}{\arg\max} \frac{\widehat{\lambda}_{ik}}{\widehat{\lambda}_{i+1,k}}.$$

Calculate $\widehat{r}^{\mathrm{ip}}$ by

$$\widehat{r}^{\mathrm{ip}} = \max(\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_K).$$

Output: Estimate of the number of factors $\widehat{r}^{\mathrm{ip}}$.

tion of CP factor models does not necessarily require orthogonality. Applying tensor PCA to models with non-orthogonal loadings introduces a bias component of higher order than our first-stage randomized composite PCA. Even when the loadings are orthogonal, our contemporary variance-based iterative estimation exhibits a faster convergence rate than tensor PCA due to dimension and noise reduction. A comparison of our estimator with the autocovariance-based estimator and tensor PCA through simulation will be presented in Section 5.

4 Theory

In this section, we delve into the statistical attributes of the algorithms introduced previously. Our theoretical framework offers guarantees for consistency and outlines the statistical error rates for estimating the factor loading vectors a_{ik}, where $1 \leqslant i \leqslant r, 1 \leqslant k \leqslant K$, given certain regularity conditions. Considering that the loading vector a_{ik} can only be identified with a change in sign, we employ

$$\|\widehat{a}_{ik}\widehat{a}_{ik}^\top - a_{ik}a_{ik}^\top\|_2 = \sqrt{1-(\widehat{a}_{ik}^\top a_{ik})^2} = \sup_{\boldsymbol{z} \perp a_{ik}} |\boldsymbol{z}^\top \widehat{a}_{ik}|$$

to quantify the discrepancy between $\widehat{a}_{ik}$ and a_{ik}.

To present theoretical properties of the proposed procedures, we impose the following assumptions.

Assumption 4.1. *Let $\xi_t = (\xi_{1t}, \xi_{2t}, \ldots, \xi_{pt})$ be independent p-dimensional random vector with each entry ξ_i independent and satisfying $\mathbb{E}(\xi_{it}) = 0$, $\mathbb{E}(\xi_{it}^2) = 1$ and for $0 < \vartheta \leqslant 2$*

$$\max_i \mathbb{P}\left(|\xi_{it}| \geqslant x\right) \leqslant c_1 \exp\left(-c_2 x^{\vartheta}\right). \tag{13}$$

Let $vec(\mathcal{E}_t) = H\xi_t$, where H is a deterministic matrix and $p \geqslant d$. The eigenvalues of the covariance matrix of $vec(\mathcal{E}_t)$ satisfies $C_0^{-1} \leqslant \lambda_d(\Sigma_e) \leqslant \cdots \leqslant \lambda_1(\Sigma_e) \leqslant C_0$ where $\Sigma_e = \mathbb{E}vec(\mathcal{E}_t)vec(\mathcal{E}_t)^\top$ and C_0 is a constant.

Assumption 4.2. *Recall $F_t = (f_{1t}, ..., f_{rt})^\top$, $\Theta = \mathbb{E}(F_t F_t^\top)$ and $\lambda_i = \lambda_i(\Theta)$ for $1 \leqslant i \leqslant r$. Assume $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$. For any $v \in \mathbb{R}^r$ with $\|v\|_2 = 1$,*

$$\max_t \mathbb{P}\left(\left|v^\top \Theta^{-1/2} F_t\right| \geqslant x\right) \leqslant c_1 \exp\left(-c_2 x^{\gamma_1}\right), \tag{14}$$

where c_1, c_2 are some positive constants and $0 < \gamma_1 \leqslant 2$.

Assumption 4.3. *Assume the factor process $f_{it}, 1 \leqslant i \leqslant r$, is stationary and strong α-mixing in t. The mixing coefficient satisfies*

$$\alpha(m) \leqslant \exp\left(-c_0 m^{\gamma_2}\right) \tag{15}$$

for some constant $c_0 > 0$ and $\gamma_2 \geqslant 0$, where

$$\alpha(m) = \sup_t \left\{\left|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| : A \in \sigma(f_{is}, 1 \leqslant i \leqslant r, s \leqslant t), B \in \sigma(f_{is}, 1 \leqslant i \leqslant r, s \geqslant t + m)\right\}.$$

Assumption 4.1 aligns closely with the noise conditions presented in seminal works such as Bai and Ng (2002), Bai (2003), Lam et al. (2011), Lam and Yao (2012), and others within the factor model literature. For simplicity, we assume that the noise tensor remains independent across time t, allowing for weak cross-sectional dependence. While incorporating weak temporal correlation among the noise, as suggested by Bai and Ng (2002), is plausible, it substantially complicates our theoretical analysis. Therefore, we defer this exploration to future research. Nonetheless, our simulation studies demonstrate the robust performance of the proposed methods even under conditions of weak temporal dependence.

Assumption 4.2 ensures the unique identification of all factor loading vectors a_{ik} up to sign changes. Unlike the eigen decomposition of a matrix, if some λ_i are equal, the estimation of the loading vectors a_{ik} isn't subject to rotational ambiguity but only to the signed permutation of loading vectors. Furthermore, Assumption 4.2 specifies that the tail probability of f_{it} must exhibit exponential decay. Specifically, when $\gamma_1 = 2$, it implies that f_{it} follows a sub-Gaussian distribution.

Assumption 4.3 is a widely recognized standard condition that accommodates a broad range of time series models, including causal ARMA processes with continuously distributed innovations, as further detailed in works such as Tong (1990); Bradley (2005); Tsay (2005); Fan and Yao (2003);

Rosenblatt (2012); Tsay and Chen (2018), among others.

While Assumptions 4.1 and 4.2 currently assume exponential tails for both noise and factor processes, these conditions can be extended to accommodate polynomial-type tails (under bounded moment conditions) when the number of factors r is fixed, albeit at the cost of a more complex theoretical analysis.

Recall A_k defined in equation (10) with a_{ik} as its columns, and $A_k^\top A_k = (\sigma_{ij,k})_{r\times r}$. As $\sigma_{ii,k} = \|a_{ik}\|_2^2 = 1$, the correlation among columns of A_k can be measured by

$$\delta_k = \|A_k^\top A_k - I_r\|_2. \tag{16}$$

Similarly we use

$$\delta = \|A^\top A - I_r\|_2 \tag{17}$$

to measure the correlation of the matrix $A = (a_1, \ldots, a_r) \in \mathbb{R}^{d\times r}$ with $a_i = \mathrm{vec}(\otimes_{k=1}^K a_{ik})$ and $d = \prod_{k=1}^K d_k$. Let $\delta_{\max} = \max\{\delta_1, \cdots \delta_K\}$.

Theorem 4.1 below presents the performance bounds, which depends on the coherence (the degree of non-orthogonality) of the factor loading vectors.

Theorem 4.1. *Suppose Assumptions 4.1, 4.2, 4.3 hold. Let $1/\gamma = 2/\gamma_1 + 1/\gamma_2$, and $\delta < 1$ with δ defined in (17). Assume $T \leqslant C\exp\left(d^{\vartheta/(2\vartheta+4)}\right)$ and $T \leqslant C\exp\left((d/r)^{\gamma_1/2}\right)$.*
(i). The eigengaps satisfy $\min\{\lambda_i - \lambda_{i+1}, \lambda_i - \lambda_{i-1}\} \leqslant c\lambda_r$ for all $1 \leqslant i \leqslant r$, with $\lambda_0 = \infty, \lambda_{r+1} = 0$, and c is sufficiently small constant. With probability at least $1 - T^{-C_1} - d^{-C_1}$, the following error bound holds for the estimation of the loading vectors a_{ik} using Algorithm 1,

$$\|\widehat{a}_{ik}^{\mathrm{rcpca}}\widehat{a}_{ik}^{\mathrm{rcpca}\top} - a_{ik}a_{ik}^\top\|_2 \leqslant \left(1 + \frac{2\lambda_1}{\lambda_r}\right)\delta + \frac{C_2\phi^{(0)}}{\lambda_r}, \tag{18}$$

for all $1 \leqslant i \leqslant r$, $1 \leqslant k \leqslant K$, where C_1, C_2 are some positive constants, and

$$\phi^{(0)} = \lambda_1\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) + \sqrt{\frac{\lambda_1 d\log d}{T}} + \frac{\sqrt{\lambda_1}(d\log d)}{T} + 1. \tag{19}$$

(ii). The eigengaps condition in (i) is not satisfied. Assume $\lambda_1 \asymp \lambda_r$ and the number of random projections $L \geqslant Cd^2 \vee Cdr^{2(\lambda_1/\lambda_r)^2}$. With probability at least $1 - T^{-C_1} - d^{-C_1}$, the following error bound holds for the estimation of the loading vectors a_{ik} using Algorithm 1,

$$\|\widehat{a}_{ik}^{\mathrm{rcpca}}\widehat{a}_{ik}^{\mathrm{rcpca}\top} - a_{ik}a_{ik}^\top\|_2 \leqslant C_3\sqrt{\delta_{\max}} + C_3\sqrt{\frac{\phi^{(0)}}{\lambda_r}}. \tag{20}$$

The first term of the upper limit in (18) and (20) arises due to the loading vectors a_{ik} not being orthogonal, which may be seen as bias. Meanwhile, the subsequent term in (18) is derived from

a concentration bound concerning random noise, thereby being describable as a form of stochastic error. The condition $T \leqslant C\exp\left(d^{\vartheta/(2\vartheta+4)}\right)$ and $T \leqslant C\exp\left((d/r)^{\gamma_1/2}\right)$ is imposed primarily for convenience. Eliminating this condition would result in a more complex convergence rate.

When the eigengap condition is not met, we employ randomized projection to determine the statistical convergence rate as shown in (20), which is slower than the rate in (18). A broader result than (20), permitting a more general eigen ratio λ_1/λ_r for part (ii), is detailed in the appendix. In practice, since the sample covariance tensor includes both the average of signal-by-noise cross-products and the average of noise-by-noise cross-products, it is uncommon to encounter nearly identical sample spiked eigenvalues. Our simulation study demonstrates that while the original composite PCA provides viable initializations when $\lambda_1 = \lambda_r$, its performance is not as good as that of randomized composite PCA using Procedure 2.

Remark 4.1. *With minor modifications to the proof of Theorem 4.1(i), we are able to show*

$$\|\widehat{a}_{ik}^{\mathrm{rcpca}}\widehat{a}_{ik}^{\mathrm{rcpca}\top} - a_{ik}a_{ik}^{\top}\|_2 = O_{\mathbb{P}}\left((\lambda_1/\lambda_r)\delta + \frac{\lambda_1}{\lambda_r}\left(\sqrt{\frac{r}{T}} + \frac{r^{1/\gamma}}{T}\right) + \frac{\sqrt{\lambda_1 d}}{\lambda_r\sqrt{T}} + \frac{1}{\lambda_r}\right). \tag{21}$$

In the typical strong factor models where $\lambda_1 \asymp \lambda_r \asymp d$ and r fixed, the rate becomes $O_{\mathbb{P}}(\delta + \sqrt{1/T} + 1/d)$, aligning with the convergence rate for the vector factor model when $\delta = 0$.

Let the statistical error bound of the initialization used in Algorithm 3 be ψ_0 (for example, the right hand side of (18)), and also let

$$\psi^{\mathrm{ideal}} = \max_{1\leqslant k\leqslant K}\left(\frac{1}{\lambda_r}\sqrt{\frac{d_k\log d}{T}} + \sqrt{\frac{d_k\log d}{\lambda_r T}} + \frac{1}{\lambda_r}\right). \tag{22}$$

Theorem 4.2. *Suppose Assumptions 4.1, 4.2, 4.3 hold. Assume that $\delta_{\max} = \max_{k\leqslant K}\delta_k < 1$ with δ_k defined in (16), and $r = O(T)$. Let $1/\gamma = 2/\gamma_1 + 1/\gamma_2$, $d = d_1\cdots d_K$, and $d_{\min} = \min_{k\leqslant K} d_k$. Assume $T \leqslant C\exp\left(d_{\min}^{\vartheta/(2\vartheta+4)}\right)$ and $T \leqslant C\exp\left((d_{\min}/r)^{\gamma_1/2}\right)$. Suppose that for a proper numeric constant $C_{1,K}$ depending on K only, we have*

$$C_{1,k}\sqrt{r}\psi_0 + C_{1,K}\left(\frac{\lambda_1}{\lambda_r}\right)\psi_0^{2K-3} + C_{1,K}\sqrt{\frac{\lambda_1}{\lambda_r}}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right)\psi_0^{K-2} \leqslant \rho < 1. \tag{23}$$

Then, after at most $M = O(\log(\psi_0/\psi^{ideal}))$ iterations of Algorithm 3, with probability at least $1 - T^{-C} - d^{-C}$, the final estimator satisfies

$$\|\widehat{a}_{ik}^{\mathrm{iso}}\widehat{a}_{ik}^{\mathrm{iso}\top} - a_{ik}a_{ik}^{\top}\|_2 \leqslant C_{0,K}\psi^{ideal}, \tag{24}$$

for all $1 \leqslant i \leqslant r$, $1 \leqslant k \leqslant K$, where $C_{0,K}$ is a constant depending on K only and C is a positive numeric constant.

Again, we assume $T \leqslant C\exp\left(d_{\min}^{\vartheta/(2\vartheta+4)}\right)$ and $T \leqslant C\exp\left((d_{\min}/r)^{\gamma_1/2}\right)$ primarily for simplifying the convergence rate. It is important to note that the error bound ψ_0 for initialization is intended for each individual loading vector a_{ik}. When applying Algorithm 3, which requires the inverse of $\widehat{A}_k^\top \widehat{A}_k$, the condition $\sqrt{r}\psi_0 \lesssim 1$ ensures a reliable initial estimate of the loading matrix $\widehat{A}_k$. Although the other components in (23) may seem complex, they are designed to ensure the error contraction effect in each iteration. This ensures that as iterations progress, the error bound will approach the desired statistical upper bound.

Remark 4.2. *With slightly modifications to the proof of Theorem 4.2, we can show*

$$\|\widehat{a}_{ik}^{\text{iso}}\widehat{a}_{ik}^{\text{iso}\top} - a_{ik}a_{ik}^\top\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{d_{\max}}{\lambda_r T}} + \frac{1}{\lambda_r}\right). \tag{25}$$

In the typical strong factor models where $\lambda_1 \asymp \lambda_r \asymp d$, the rate simplifies to $O_{\mathbb{P}}(\sqrt{d_{\max}/(dT)} + 1/d)$. This rate is significantly faster than that found in the vector factor model.

We now demonstrate the feasibility of obtaining a more precise bound by closely examining the leading order term. This process allows us to ascertain the asymptotic behavior of the estimator a_{ik}. Specifically, we will establish that

$$\widehat{a}_{ik}^{\text{iso}} - \text{sign}(a_{ik}^\top \widehat{a}_{ik}^{\text{iso}})a_{ik} = P_{a_{ik},\perp}\left[\frac{1}{\Theta_{ii}T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t \times_{\ell\in[K]\setminus k} b_{i\ell}^\top\right)\right] + O_{\mathbb{P}}\left(\frac{1}{\Theta_{ii}}\left(\frac{d_k}{T} + \sqrt{\frac{d_k}{T}} + 1\right)\right),$$

where $P_{a_{ik},\perp} = I_{d_k} - a_{ik}a_{ik}^\top$ and $\Theta = (\Theta_{ij})_{r\times r}$, where Θ is defined in Assumption 4.2. This enables the determination of asymptotic distributions for linear forms of a_{ik}.

The following theorem shows the asymptotic distribution of a linear form of the factor loading vector $u^\top a_{ik}$ for some fixed vector u. Note that in the strong factor model, we have $\Theta_{ii} \asymp d$ for all $1 \leqslant i \leqslant r$.

Theorem 4.3. *Suppose the conditions in Theorem 4.2 are satisfied. Let $\lambda_1 \asymp \lambda_r$. Assume that $\liminf_{d_k\to\infty} \|P_{a_{ik},\perp}u\|_2 > 0$, for each $1 \leqslant i \leqslant r, 1 \leqslant k \leqslant K$, we have:*

(i) If $T/(d_k\Theta_{ii}) \to 0$, then

$$\sqrt{T}u^\top\left(\widehat{a}_{ik}^{\text{iso}} - \text{sign}(\widehat{a}_{ik}^{\text{iso}\top}a_{ik})\cdot a_{ik}\right) \xrightarrow{d} N(0,\sigma_{u,ik}^2), \tag{26}$$

where $\sigma_{u,ik}^2 = h_{ik}^\top \Sigma_e h_{ik}/\Theta_{ii}$, $h_{ik} = b_{iK}\odot\cdots\odot b_{i,k+1}\odot P_{a_{ik},\perp}u\odot b_{i,k-1}\odot\cdots\odot b_{i1} \in \mathbb{R}^d$ and $\odot$ represents Kronecker product.

(ii) If $d_k\Theta_{ii} = O(T)$, then

$$\Theta_{ii}u^\top\left(\widehat{a}_{ik}^{\text{iso}} - \text{sign}(\widehat{a}_{ik}^{\text{iso}\top}a_{ik})\cdot a_{ik}\right) = O_{\mathbb{P}}(1). \tag{27}$$

In Theorem 4.3, we focus on vectors u with the property that $\|P_{a_{ik},\perp}u\|_2 > 0$ when d_k is large, effectively presupposing that $\sin\angle(u, a_{ik}) > 0$. Conversely, when $\sin\angle(u, a_{ik}) = 0$, the convergence rate of the estimated linear form is faster, and its asymptotic distribution is a blend of χ^2_1 distributions.

Theorem 4.4. *Suppose the conditions in Theorem 4.2 are satisfied. Let $\lambda_1 \asymp \lambda_r$. For each $1 \leqslant i \leqslant r, 1 \leqslant k \leqslant K$, we have:*

(i) With probability at least $1 - T^{-C} - d^{-C}$,

$$1 - (\widehat{a}_{ik}^{\mathrm{iso}\top} a_{ik})^2 \leqslant C_{0,K}(\psi^{ideal})^2, \tag{28}$$

where ψ^{ideal} is defined in (22).

(ii) If $T/(d_k\Theta_{ii}) \to 0$, then

$$T\left(1 - (\widehat{a}_{ik}^{\mathrm{iso}\top} a_{ik})^2\right) \xrightarrow{d} \sum_{j=1}^{d_k} \varpi_j \chi^2_1, \tag{29}$$

where $\varpi_j, 1 \leqslant j \leqslant d_k$ are the eigenvalues of $\Sigma_{ik}^{1/2} P_{a_{ik},\perp} \Sigma_{ik}^{1/2}$, with $\Sigma_{ik} = \Theta_{ii}^{-1}\mathbb{E}[(\mathcal{E}_t \times_{\ell\neq k}^{K} b_{i\ell})(\mathcal{E}_t \times_{\ell\neq k}^{K} b_{i\ell})^\top]$.

(iii) If $d_k\Theta_{ii} = O(T)$, then

$$\Theta_{ii}^2\left(1 - (\widehat{a}_{ik}^{\mathrm{iso}\top} a_{ik})^2\right) = O_{\mathbb{P}}(1). \tag{30}$$

Drawing parallels with traditional PCA is insightful; in PCA, a debiasing process is often necessary to achieve asymptotic normality in linear combinations of the principal components, as discussed in Koltchinskii and Lounici (2016, 2017); Koltchinskii et al. (2020). For the CP tensor factor model, however, merely meeting the signal strength requirement $T/(d_k\Theta_{ii}) \to 0$ is enough to render the bias inconsequential. This observation aligns with findings by Bai (2003) regarding vector factor models.

The estimators are constructed with a specified rank r, although in the theoretical analysis, r is allowed to increase. Practically, $\widehat{r}$ can be estimated using the generalized eigenvalue ratio-based estimators detailed in Algorithms 4 or 5. The asymptotic validity of $\widehat{r}^{\mathrm{uer}}$ and $\widehat{r}^{\mathrm{ip}}$ are established in Theorem 4.5 below.

Theorem 4.5. *Suppose Assumptions 4.1, 4.2, 4.3 hold and $r_{\max}$ is a predetermined constant no smaller than r. Assume $r = O(T)$ and $\lambda_r^{-1/2} d^{1/2} T^{-1/2} + \lambda_r^{-1} = o(1)$. Then*

$$\mathbb{P}(\widehat{r}^{\mathrm{uer}} = r) \to 1,$$
$$\mathbb{P}(\widehat{r}^{\mathrm{ip}} = r) \to 1,$$

as $d_k \to \infty$ and $T \to \infty$.

Theorem 4.5 derives the consistency of rank estimators $\widehat{r}^{\text{uer}}$ and $\widehat{r}^{\text{ip}}$ based on the eigenvalue ratios. This can be viewed as a generalization of Theorem 1 of Ahn and Horenstein (2013) from vector factor models to CP tensor factor models.

5 Simulation

In this section, we conduct empirical comparisons among different methods for estimating loading vectors across various simulation scenarios, and verify the limiting distribution of the estimated loading vectors. We assess the performance of the contemporary covariance-based iterative simultaneous orthogonalization procedure (CC-ISO) proposed in this paper, auto-covariance-based iterative simultaneous orthogonalization procedure by Han et al. (2023) (AC-ISO), and tensor principle component analysis (TPCA) by Babii et al. (2023). The auto-covariance considered by Han et al. (2023) is defined by the following lagged-cross product operator:

$$\mathbf{\Sigma}_h = \mathbb{E}\left[\frac{\mathcal{Y}_{t-h} \otimes \mathcal{Y}_t}{T-h}\right] \in \mathbb{R}^{d_1 \times \cdots d_K \times d_1 \times \cdots \times d_K}.$$

In this section, we fix $h = 1$. The estimation error measures the angle between the estimated loading vector and the true loading vector, computed as:

$$\max_{i,k} \|\widehat{a}_{ik}\widehat{a}_{ik}^\top - a_{ik}a_{ik}^\top\|_2.$$

Throughout our analysis, the observations $\mathcal{Y}_t$'s are simulated according to model (1) with $K = 2$. The true loading vectors are generated as follows: The elements of matrices $\widetilde{A}_k = (\widetilde{a}_{1k}, \ldots, \widetilde{a}_{rk}) \in \mathbb{R}^{d_k \times r}, 1 \leqslant k \leqslant K$, are drawn from i.i.d. $N(0,1)$ distributions and then orthonormalized via QR decomposition. If $\delta = 0$, set $A_k = \widetilde{A}_k$; otherwise, set $a_{1k} = \widetilde{a}_{1k}$ and $a_{ik} = (\widetilde{a}_{1k} + \theta\widetilde{a}_{ik}) / \|\widetilde{a}_{1k} + \theta\widetilde{a}_{ik}\|_2$ for all $i \geqslant 2$ and $1 \leqslant k \leqslant K$, with $\vartheta = \delta/(r-1)$ and $\theta = \left(\vartheta^{-2/K} - 1\right)^{1/2}$. In our simulation study, we vary the correlations between loading vectors through δ. It is evident that an increase in δ leads to a higher degree of linear dependence among the loading vectors.

The factor processes f_{it} exhibit weak temporal dependence and are generated as an independent AR(1) process multiplied by a scalar depending on d_1, d_2 and r:

$$f_{it} = w_i g_{it}, \tag{31}$$

where

- $g_{i,t+1} = \phi g_{it} + \epsilon_{it}$ with $\phi = 0.1$, $\text{Var}(\epsilon) = 1 - \phi^2 = 0.99$[1];
- $w_i = \frac{1}{5} \times (r - i + 1)\sqrt{d_1 \times d_2}$.

[1]The results with $\phi = 0.5$ are reported in the appendix.

In this case, the tensor factor model in (1) represents a typical strong factor model, where $\lambda_i = w_i^2 = (r-i+1)^2 d_1 d_2/25 = O(d)$ when r is fixed. In Appendix C, we provide the result for $\phi = 0.5$.

The following three configurations are adapted from Babii et al. (2023) and Han et al. (2023) with modifications made for comparative analysis of the empirical performances among TPCA, AC-ISO and CC-ISO. In these configurations, $(d_1, d_2) \in \{(40,40),(40,60),(60,60)\}$, $T \in \{100, 300, 500\}$ and $r = 3$.

I. (Orthogonal loading matrix) Set $\delta = 0$ so that the columns of loading matrix A_k are orthonormal. Each entry of error term $\mathcal{E}_t$ is generated independently from $N(0,1)$.

II. (Non-orthogonal loading matrix) Vary δ in the set $\{0.1, 0.3, 0.5\}$ so that the columns of loading matrix A_k are not orthogonal. Each entry of error term $\mathcal{E}_t$ is generated independently from $N(0,1)$.

III. (Serial correlation in $\mathcal{E}_t$) Set $\delta = 0.2$. The errors $\mathcal{E}_t$ are generated according to $\mathcal{E}_t = \Psi_1^{1/2} Z_t \Psi_2^{1/2}$, where

   - $\Psi_1 = \Psi_2 = \{\sigma_{e,ij}\}$ with $\sigma_{e,ij} = 0.5^{|i-j|}$;
   - $\mathrm{vec}(Z_t) = \Phi \mathrm{vec}(Z_{t-1}) + U_t$ where $U_t \sim i.i.d.\ N(0, I_{d_1 d_2})$ and $\Phi \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ is a diagonal matrix with all diagonal elements equal to ρ.

   We vary ρ in the set $\{0.1, 0.3, 0.5\}$ to investigate the robustness of our algorithm under weak cross-sectional correlation and serial correlation in the error term.

The following configuration aims to assess the robustness of our proposed algorithm under weak factor structures.

IV. (Weak factors) Set $r = 3$, and $\delta = 0.2$. The error terms are generated according to $\mathcal{E}_t = \Psi_1^{1/2} Z_t \Psi_2^{1/2}$, where

   - $\Psi_1 = \Psi_2 = \{\sigma_{e,ij}\}$ with $\sigma_{e,ij} = 0.5^{|i-j|}$;
   - $Z_{ijt} \sim i.i.d.\ N(0,1)$.

   The scaling multiplier in factor process $w_i = (r-i+1) \times (d_1 d_2)^{1/\alpha}$, where α varies in the set $\{2.5, 3, 3.5, 4\}$. Note that when $\alpha = 2$, the factor structure is considered strong. A larger α indicates a weaker factor structure.

For each configuration, we conduct the experiment 500 times and present the box plots of the results. Figure 2 shows the estimation errors for CC-ISO, AC-ISO and TPCA under configuration I. Notably, CC-ISO consistently outperforms the other two algorithms across various dimensions. The estimation by AC-ISO deviates significantly from the true value due to the weak signal in the auto-covariance matrix resulting from the weak temporal dependence in the factor process.

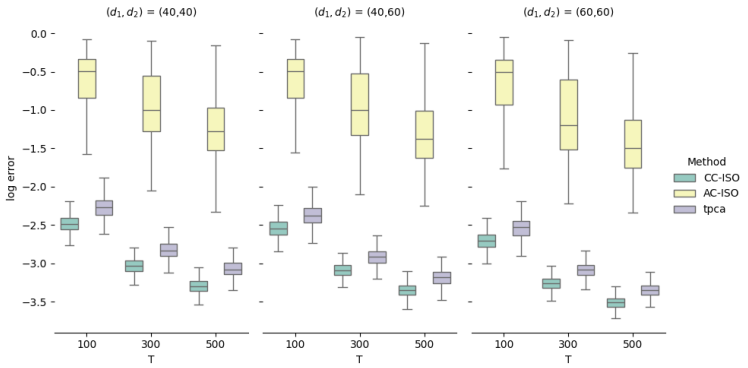


Figure 2: Boxplots of the estimation error over 500 replications under configuration I

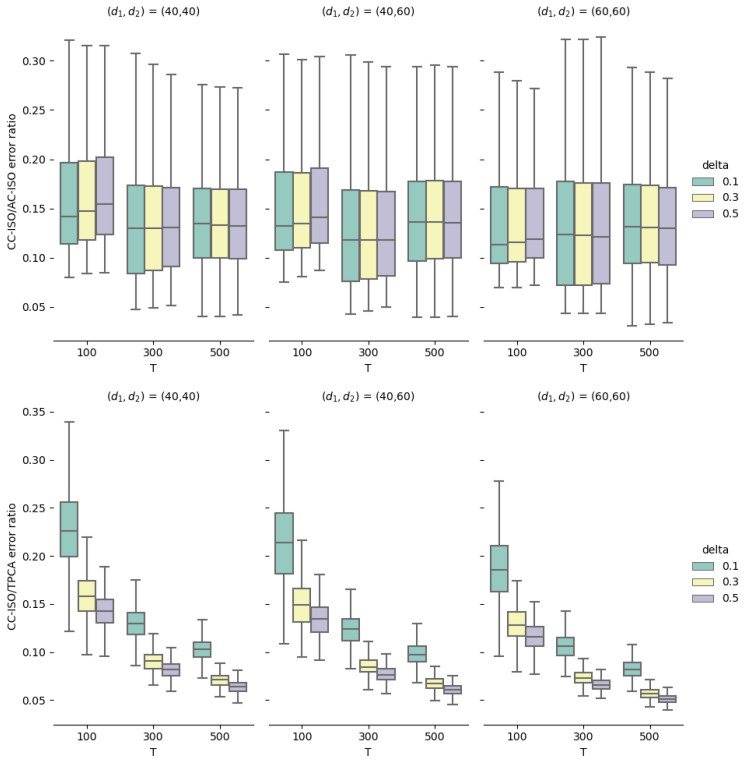


Figure 3: Boxplots of the estimation error over 500 replications under configuration II. Note: The first panel shows the ratio of the estimation error of CC-ISO on AC-ISO. The second panel shows the ratio of the estimation error of CC-ISO on TPCA.

In Configuration II, we assess the impact of non-orthogonal factor loadings on estimations using ISO algorithms and TPCA algorithm. Figure 3 shows the ratio of the estimation errors of CC-ISO to AC-ISO (first panel) and to TPCA (second panel) across different values of δ and dimensions. The error ratio of CC-ISO to AC-ISO remains around 0.15, indicating the superior accuracy of CC-ISO. The ratio remains relatively stable because the signal part in AC-ISO, albeit small, also increases with dimensions, resulting in limited improvements on the estimation. However, in the second panel, the error ratio of CC-ISO to TPCA converges as dimensions increase. This is because TPCA cannot identify non-orthogonal factor loading vectors, leading to stable estimation errors across varying dimensions. In contrast, CC-ISO successfully identifies non-orthogonal loading vectors, resulting in estimation errors converging to 0.

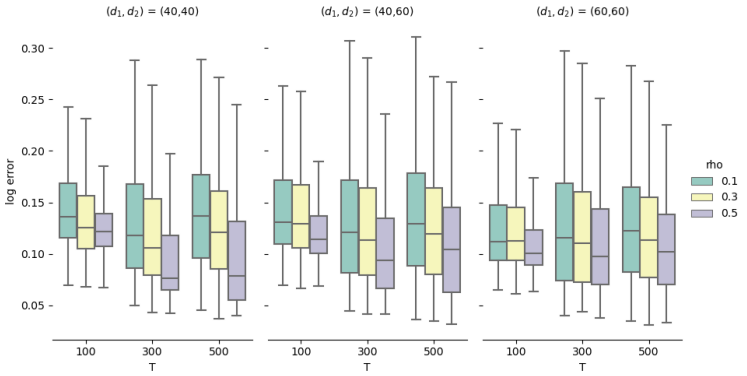


Figure 4: Boxplots of the estimation error over 500 replications under configuration III

Figure 4 shows the ratio of estimation errors of CC-ISO to AC-ISO under configuration III, designed to evaluate the robustness of proposed CC-ISO algorithm against serial correlation in the error term. We observe that CC-ISO's performance improves monotonically as T increases. In contrast, AC-ISO's performance deteriorates as the serial correlations in the error term strengthens. This decline is due to the contamination of signals in the auto-covariance by the serial correlations in the error terms. However, CC-ISO demonstrates robustness against such serial correlations.

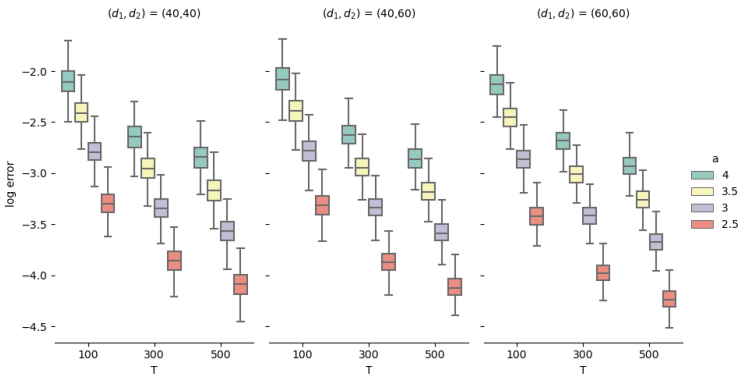


Figure 5: Boxplots of the estimation error over 500 replications under configuration IV

In Figure 5, we show the box plots of the logarithm of the estimation errors of CC-ISO algorithm under a weak factor configuration. It is evident that the estimation errors decrease as T increases. Additionally, the rate of decrease in estimation errors depends on the value of α: a higher α leads to a faster decrease. These results validate the robustness of the CC-ISO algorithm against a certain degree of weak factor structure, in line with the conclusions drawn in Theorem 4.2.

We also examine the performance of Randomized Projection (RP) from Procedure 2 and compare it with Composite PCA (C-PCA), which corresponds to Algorithm 1 but without the steps for detecting close eigenvalues (Step 3, 5, 6, and 7 in Algorithm 1).

V. (C-PCA vs. RP-PCA) $r = 5$. $d_1 = d_2 = \bar{d}$ with $\bar{d} \in \{20, 40, 80\}$ and $T \in \{100, 200, 500\}$. The columns of factor loadings A_k are orthonormal and are generated as described in Configuration I. Furthermore, the factors f_{it} are also orthonormal, generated using QR decomposition after deriving from AR(1) processes. In this setting, the singular values of the common components $\sum_{i=1}^{r} w_i g_{it} a_{i1} \otimes a_{i2}$ are solely determined by w_i. We set $w_i = w = 10$ to ensure identical eigenvalues of common components. Error terms are generated from i.i.d. $N(0, 1)$. Though the top r eigenvalues of the $\widetilde{\Sigma}$ are not identical due to noise, their differences are relatively small, allowing randomized projection algorithms to ensure the accuracy of initial estimations. For the remaining parameters, we set $\nu = 0.8$, $c_0 = 0.1$ and $L = 2r^2$.

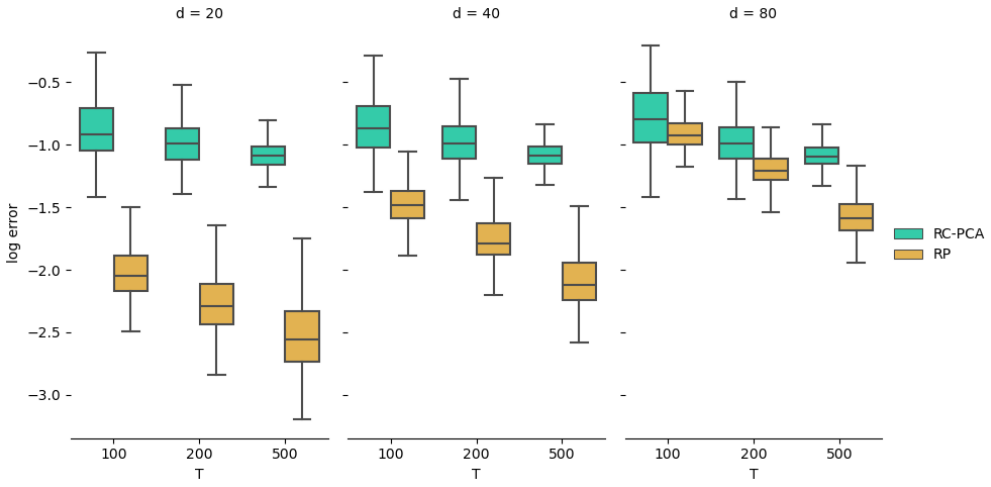


Figure 6: Boxplots of the estimation error over 500 replications under configuration V

Given the close empirical performances of CC-ISO under both initialization methods under configuration V, our focus shifts to the estimation errors of the initial estimations, as illustrated in Figure 6. RP algorithm outperforms the RC-PCA algorithm in terms of the accuracy of initial estimations, particularly pronounced when d is smaller and T is larger. This occurs because the sample covariance of $\mathrm{Vec}(\mathcal{E}_t)$ approaches the identity matrix as d decreases and T increases. Consequently, $\widetilde{\Sigma}$ is more likely to have eigenvalues that are closer together.

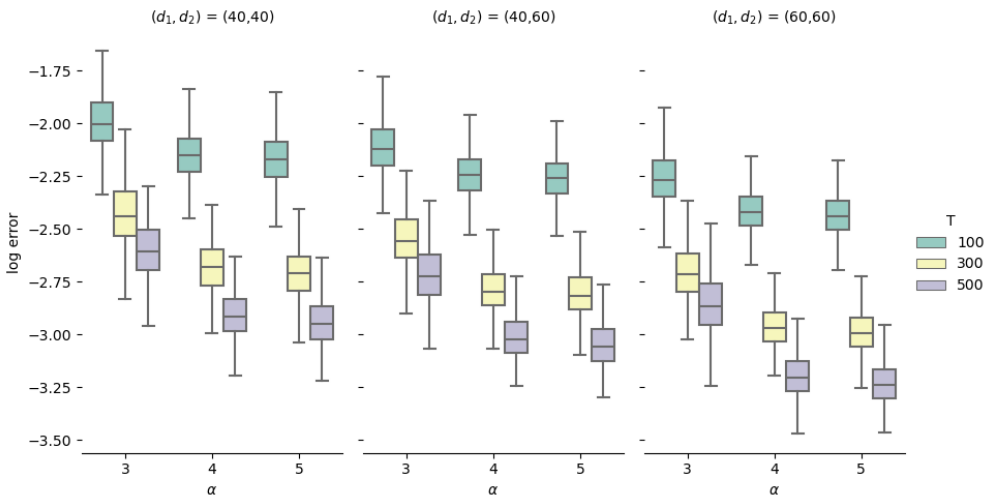


Figure 7: Boxplots of the estimation error over 500 replications under configuration VI

The subsequent simulation verifies the robustness of the CC-ISO algorithm against weak mis-specification of the model. The data is generated following the tucker factor model with $K = 2$:

$$\mathcal{Y}_t = \boldsymbol{A}_1 \mathcal{F}_t \boldsymbol{A}_2^\top + \mathcal{E}_t,$$

where $\mathcal{F}_t \in \mathbb{R}^{r \times r}$ is the factor process in the Tucker factor model. In the CP factor model, $\mathcal{F}_t$ is diagonal with the i^{th} diagonal element equal to f_{it}. In the mis-specification setting, we allow the off-diagonal entries to deviate from 0. Denote the $(i,j)^{th}$ entry of $\mathcal{F}_t$ by f_{ijt}. Let $f_{ijt} = w_{ijt} g_{ijt}$, where g_{ijt} is generated as specified in (31). The configuration is as follows:

VI. (Mis-specification) $r = 3, (d_1, d_2) \in \{(40,40), (40,60), (60,60)\}$ and $T \in \{100, 300, 500\}$. The loading vectors and error terms are generated as in Configuration IV, allowing for correlation between loading vectors and weak cross-sectional correlation in the error term. $w_{ijt} = \sqrt{d_1 d_2}/5$ if $i = j$ and $w_{ijt} = (d_1 d_2)^{1/\alpha}/5$ with $\alpha \in \{3, 4, 5\}$. A smaller α indicates a more severe mis-specification in the model.

Figure 7 shows the results under configuration VI. Given α, the estimation error decreases in T or in d, which illustrates the robustness of CC-ISO against weak mis-specification.

Next simulation is conducted to verify the results in Theorem 4.3(i). The configuration is as follows:

VII. (CLT) $r = 3$. $d_1 = d_2 = \bar{d} \in \{20, 60, 100\}$. For each $\bar{d}$, we set $T = 200$ and $w_i = (r - i + 1)\sqrt{d_1 d_2}$. For factor loading vectors, we let $\delta = 0.2$ to allow for non-orthogonal loading vectors. The error $\mathcal{E}_{i,j,t}$ are generated as in Configuration IV to allow for weak cross-sectional correlations. We simulate the distribution of a_{ik} in (26) with $i = 1$, $k = 1$ under three choices of u: $u_1 = 1/\sqrt{\bar{d}}$, $u_2 = [1, 0, 0, \ldots, 0]^\top$ and $u_3 = [0, 1, 0, 0, \ldots, 0]^\top$.

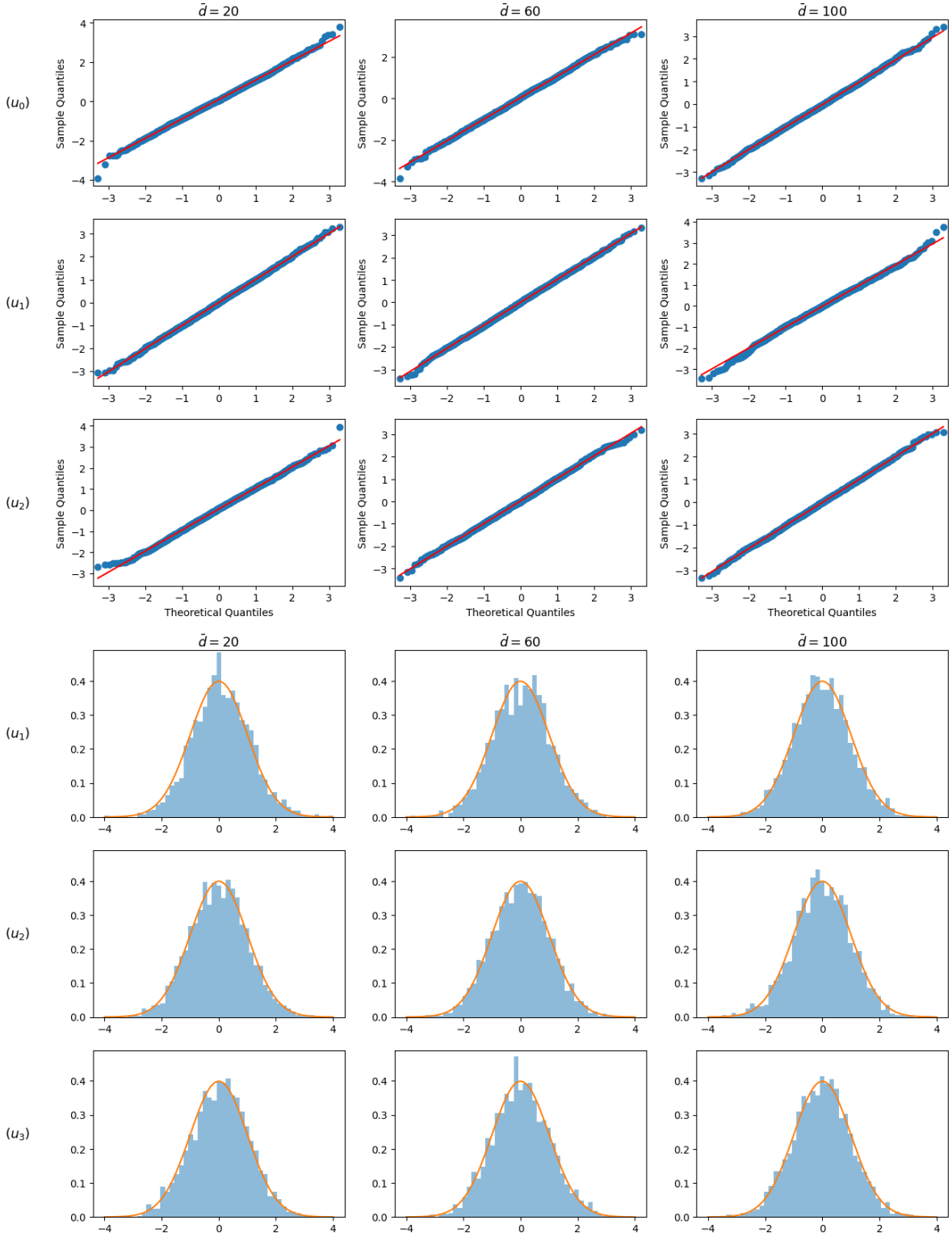


Figure 8: QQ plots and histograms of $\sqrt{T}u^\top \left(\widehat{a}_{ik}^{\text{iso}} - \text{sign}(\widehat{a}_{ik}^{\text{iso}\top} a_{ik}) \cdot a_{ik}\right)/\sigma_{u,ik}$ under Configuration VII. Note: 1. The row displays the results for u_i. The column shows the results for different dimension $\bar{d}$. 2. The red curve plots the distribution of standard normal distribution.

Figure 8 shows the QQ plots and histograms of $\sqrt{T}u^\top\left(\widehat{a}_{ik}^{\text{iso}} - \text{sign}(\widehat{a}_{ik}^{\text{iso}\top}a_{ik})\cdot a_{ik}\right)/\sigma_{u,ik}$ derived from Theorem 4.3 under Configuration VII. It is observed that the normalized empirical distribution closely approximates the standard normal distribution.

Finally, we evaluate the performance of two proposed rank estimation algorithms: the unfolded eigenvalue ratio method and the eigenvalue ratio method via inner product, over the following DGP configuration:

VIII (Rank Estimation) $r = 3$. $d_1 = d_2 = \bar{d} \in \{20, 40, 60, 80\}$ and $T \in \{100, 300, 500\}$. Set $\delta = 0.2$ to allow for correlation among factor loading vectors. Error terms are generated as in Configuration IV to accomodate weak cross-sectional correlation. The factors are generated according to (31) with $w_i = (r - i + 1)\bar{d}$ and $\phi \in \{0.1, 0.5\}$.

The results are presented in Table 1. The numbers in the table denote the relative frequency of correct rank estimation over 500 replications. Both methods perform very well with accuracy levels close to 1.

Table 1: Rank estimation

	ρ	0.1		0.5	
(d_1, d_2)	T	$\widehat{r}^{uer}$	$\widehat{r}^{ip}$	$\widehat{r}^{uer}$	$\widehat{r}^{ip}$
(20,20)	100	1	0.98	1	0.95
	300	1	1	1	1
	500	1	1	1	1
(40,40)	100	1	1	1	1
	300	1	1	1	1
	500	1	1	1	1
(60,60)	100	1	1	1	1
	300	1	1	1	1
	500	1	1	1	1
(80,80)	100	1	1	1	1
	300	1	1	1	1
	500	1	1	1	1

Note: Relative frequency of correct rank estimation over 500 replications.

6 Empirical Application

6.1 Characteristic decile portfolios

In this section, in line with Babii et al. (2023), we conduct empirical analysis on the dataset collected by Chen and Zimmermann (2022), consisting of over 200 characteristic-sorted portfolios from previous studies of stock market anomalies. We utilize the August 2023 release of the database,

focusing on monthly portfolio returns sorted into 10 deciles based on firm-level characteristics spanning from January 1990 to December 2022. As we only consider a balanced panel of portfolios, the number of characteristics throughout the entire sample period is 127. Therefore, the dimension of the tensor-valued time series $\mathcal{Y}_t$ we considered is 127×10 with sample size $T = 396$. Additionally, we obtain the risk-free rate from the Kenneth French data library to compute the excess return of each portfolio.

We rewrite model (1) as:

$$\mathcal{Y}_{t,jl} = \sum_{i=1}^{r} f_{it} a_{i1,j} a_{i2,l} + \mathcal{E}_{t,jl}, \tag{32}$$

where $\mathcal{Y}_{t,jl}$ is the excess return of the l^{th}-decile of the j^{th} characteristic at time $t = 1, \ldots, T$; f_{it} are the systematic risk factors; factor loading $a_{i1,j}$ determines the heterogeneous exposure of the j^{th} characteristics to the i^{th} risk factor; loading $a_{i2,l}$ determines the exposure of the l^{th} decile to the i^{th} risk factor. In this model, all loading vectors are normalized to 1, with f_{it} absorbing all scales of loadings. We use the generalized ratio-based method as well as the screen plot to select 3 as the number of factors.

We estimate the factor model in equation (32) employing various algorithms: TPCA, CC-ISO, AC-ISO with $h = 1$, generalized eigen-analysis based estimation (GE) proposed by Chang et al. (2023) with $K = 1$, and the AC iterative projection based on tucker decomposition (Tucker-AC-IP) by Han et al. (2022a) with $h = 1$ and rank $(3, 3)$. We compare the estimates as well as the R-squared obtained from these different algorithms.

Table 2 reports the summary statistics of the estimated loadings $\widehat{a}_{i1}$, which determine the exposure of characteristics to the three latent factors. The statistics for TPCA closely align with the empirical results documented in Babii et al. (2023). Across all algorithms, the loadings on the first factor are consistently positive with relatively small standard deviations. However, while $\widehat{a}_{2,1}$ and $\widehat{a}_{3,1}$ demonstrate approximate symmetry around zero in TPCA, with approximately half of the loadings being positive, they exhibit significant skewness in the ISO algorithms, with their means deviating from zero. In the case of the GE algorithm, $\widehat{a}_{2,1}$ displays high skewness whereas $\widehat{a}_{3,1}$ shows an approximate symmetry. Specifically, around half of the loadings are positive, and the maximum and minimum values of $\widehat{a}_{3,1,i}$ are approximately symmetric about zero. Similarly, Tucker-AC-IP algorithm also exhibits approximate symmetric patterns in the second factor loading vector.

Table 3 displays the R^2 of the estimation of Model (32) across five algorithms. In this section, R^2 is defined as:

$$R^2 = 1 - \frac{\|\mathcal{Y} - \widehat{\mathcal{Y}}\|_F^2}{\|\widetilde{\mathcal{Y}}\|_F^2},$$

where $\widetilde{\mathcal{Y}}$ is the demeaned tensor of $\mathcal{Y}$ with the t^{th} entry defined as $\widetilde{\mathcal{Y}}_t = \mathcal{Y}_t - \frac{1}{T}\sum_{t=1}^{T} \mathcal{Y}_t$.

Among the four algorithms for the CP factor model, it is observed that the CC-ISO algorithms achieve the best fit to the model, exhibiting the highest R-squared values. Comparatively, within the ISO algorithms, CC-ISO outperforms AC-ISO, yielding higher R-squared values. However, the

Table 2: Summary statistics of estimated loadings $\widehat{a}_{i1}$ specific to characteristics

	Max	Mean	Min	Std	>0
			CC-ISO		
$\widehat{a}_{1,1}$	0.065	0.088	0.107	0.009	1
$\widehat{a}_{2,1}$	0.087	-0.068	-0.212	0.058	0.134
$\widehat{a}_{3,1}$	0.049	-0.073	-0.188	0.050	0.063
			AC-ISO		
$\widehat{a}_{1,1}$	0.038	0.087	0.126	0.019	1
$\widehat{a}_{2,1}$	0.063	-0.070	-0.208	0.055	0.087
$\widehat{a}_{3,1}$	0.148	-0.044	-0.230	0.078	0.276
			TPCA		
$\widehat{a}_{1,1}$	0.108	0.088	0.064	0.009	1
$\widehat{a}_{2,1}$	0.275	0.003	-0.230	0.089	0.559
$\widehat{a}_{3,1}$	0.186	0.006	-0.348	0.089	0.496
			GE		
$\widehat{a}_{1,1}$	0.181	0.08	0.001	0.033	1
$\widehat{a}_{2,1}$	0.245	0.059	-0.087	0.067	0.803
$\widehat{a}_{3,1}$	0.301	0.006	-0.291	0.089	0.520
			Tucker-AC-IP		
$\widehat{a}_{1,1}$	0.14	0.09	0.066	0.01	1
$\widehat{a}_{2,1}$	0.286	0	-0.295	0.089	0.409
$\widehat{a}_{3,1}$	0.263	-0.002	-0.187	0.089	0.457

Note: > 0 denotes the ratio of positive entries in each loading vector.

Table 3: R-squared across different algorithms

Method	CC-ISO	AC-ISO	TPCA	GE	Tucker-AC-IP
R^2	0.853	0.844	0.759	0.776	0.878

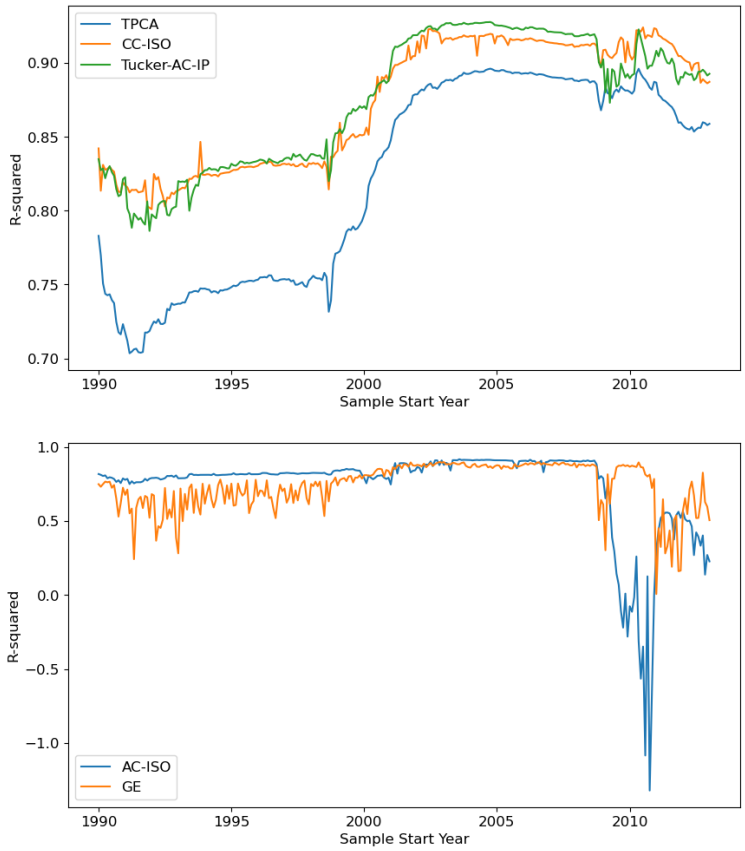


Figure 9: R-squared of the rolling-sample study on characteristic decile portfolio returns across five algorithms. Note: The first panel shows the results of CC-ISO, TPCA and Tucker-AC-IP. The second panel shows the results of AC-ISO and GE.

TPCA algorithm displays the lowest R-squared value at 0.759.

Given that Tucker decomposition entails a larger number of factors compared to CP decomposition, it is not surprising that Tucker-AC-IP yields higher R^2 than all algorithms based on CP decomposition. Nonetheless, the degree of improvement is modest, with R^2 increasing by only 2 to 4 percent when adopting the Tucker factor model. This observation suggests that the characteristic decile portfolio data might possess a CP-like factor structure.

We also conduct a rolling-sample study on portfolio excess returns, where each rolling sample spans 120 months, resulting in a total of $T - 120 = 276$ rolling samples with the first rolling sample from January 1990 to December 2000. Within each rolling sample, we estimate the CP factor model in equation (32) using four algorithms and compute the sample R^2. For comparison, we also calculate the sample R^2 under tucker tensor factor model using Tucker-AC-IP algorithm .

Figure 9 illustrates the results of the rolling R^2. The first panel shows the results of the algorithms for the CP factor model based on contemporary covariance matrix, alongside the Tucker-AC-IP algorithm for the Tucker factor model. Meanwhile, the second panel presents the results of algorithms for CP factor model based on auto-covariance matrix.

The three curves in the first panel exhibits a similar pattern: fluctuations with an overall increasing trend between 1990 and 2000, followed by a sharp rise between 2000 and 2003. Subsequently, they revert to fluctuation after 2003. However, the rolling R^2 of CC-ISO consistently outperforms that of TPCA across all rolling samples. Interestingly, while the Tucker factor model demonstrates the best fit in the full sample, it does not outperform CC-ISO in the rolling sample study, particularly in the rolling samples prior to 2000 and after 2008.

In the second panel, both AC-ISO and GE algorithms show significant fluctuations in R^2 starting after 2010, which coincides with the COVID pandemic period. Tucker-AC-IP, which also relies on auto covariance matrix estimation, exhibit milder fluctuations during this period. Before 2008, AC-ISO's R^2 follows a smoother pattern compared to the GE algorithm.

6.2 Aggregate international trade flow

Understanding the pattern and evolution of international trade flow is essential for a broad range of economic activities including policy-making, economic forecast, and firm-level optimization. In this study, we apply the CP tensor factor model to the international trade flow data, where loading vectors and latent factors are estimated simultaneously.

We use monthly aggregate import and export volumes of commodity goods from January 1991 to December 2015, including a total of $172,800$ observations. The data comes from the International Monetary Fund (IMF) *Direction of Trade Statistics* (DOTS) and involves trade among 24 countries and regions. These include Australia (AU), Canada (CA), China Mainland (CN), Denmark (DK), Finland (FI), France (FR), Germany (DE), Hong Kong (HK), Indonesia (ID), Ireland (IE), Italy (IT), Japan (JP), Korea (KR), Malaysia (MY), Mexico (MX), Netherlands (NL), New Zealand (NZ), Singapore (SG), Spain (ES), Sweden (SE), Taiwan (TW), Thailand (TH), United Kingdom (GB), and the United States (US).

As discussed in Section 2, the latent factors can be interpreted as trading hubs while the factor loadings represent import/export contributions to these hubs. Employing the generalized ratio-based method and the screen plot, we determine 6 as the optimal number of trading hubs.

Table 4 presents the R^2 across 4 different algorithms for the CP factor model and 2 algorithms for the Tucker factor model: the auto-covariance based algorithm for the Tucker factor model (Tucker-AC-IP) with $h = 1$ by Han et al. (2022a) and the contemporary-covariance based counterpart (Tucker-CC-IP). In the Tucker factor model, the number of dimensions of the latent hubs is selected as $(4, 4)$, following Chen and Chen (2022). Among the algorithms for the CP factor model, CC-ISO achieves the highest R^2. AC-ISO follows closely behind, with a slightly lower R^2 compared to CC-ISO. However, the TPCA algorithm fails to adequately fit the model to the data, resulting in a negative R^2 value of -2.56.

To gain deeper insights into dynamic patterns, we conduct a five-year rolling study on the trade flow data. Each rolling sample spans five years, starting from 1991 through 1995 for the first sample, and progressing consecutively. Within each rolling sample, we estimate the factor loadings using

Table 4: R-squared of fitting aggregate international trade flow data across different algorithms

Method	CC-PCA	AC-PCA	TPCA	GE	Tucker-AC-IP	Tucker-CC-IP
R^2	0.7951	0.7921	-2.56	0.5782	0.521	0.521

the CC-ISO algorithm proposed in this paper. We assume that the factor loadings remain constant in each rolling sample and fix the number of factors at $r = 6$ across all samples. Each sample is indexed by the mid-year of the five-year span.

Unlike the matrix factor model considered in Chen and Chen (2022), the factor loading vectors in CP factor model are uniquely identified up to the sign change. Therefore, rotation analysis, such as varimax rotation, is not applicable. Instead of applying varimax and interpreting latent hubs by the dominant country/region, we analyze latent hubs based directly on their estimates from the model. In our proposed algorithm, latent hubs are ranked by the corresponding eigenvalues of the unfolded covariance matrix and we fix, with the first latent hub contributing the most to the export/import volumes and variances. However, country contributions to each latent hub can vary across different time periods.

Figure 10 illustrates the relationships between countries and latent hubs for three selected years. The sizes of latent hub nodes are proportional to the strengths of the corresponding factors. The relationships between countries and latent hubs, shown as dotted lines, are plotted using the loading matrix on the export/import side after a truncation transformation. This is achieved by normalizing the loading matrix so that the sum of all entries equals one. Therefore, each entry represents the relative contribution of a country to a latent hub. The figure includes countries with the top four contributions to any latent hubs. Some countries significantly contribute to more than one latent hub, resulting in a stable number of countries on the export/import side, approximately 10. Countries are ranked by their total export/import volume to/from in-sample countries in the selected sub-sample.

As shown in the plot, in 1995, the US had the highest export and import volume, dominating Hub 1 on the export side and Hubs 1 to 4 on the import side. Hub 3 was significantly dominated by China on the export side, with other participants including Japan, Germany, and Taiwan on the export side, and the US, Hong Kong, and Korea on the import side, indicating deep trade connections among these countries. Hub 6 was mainly utilized by European countries on both the export and import sides, reflecting the impact of the foundation of the European Union in 1993.

In 2003, the US remained the country with the highest export and import volumes. However, China surpassed Japan and Germany, becoming the country with the second highest export and import volumes. On the export side, China actively participated in international trade within Hubs 1 to 4, along with the US, Germany, Japan, and European countries. The hub mainly used by European countries experienced growth from 1995 and became the fifth largest hub in 2003. Hub 6 can be interpreted as the APEC hub, as it was primarily used by the US and Asian countries/regions,

including Japan, China, Korea, and Taiwan.

By 2012, China had surpassed the US and became the country with the highest trade volume. On the export side, Hubs 1 and 2 were dominated by China. Participants of Hubs 1 and 2 on the import side were mainly from North America and Europe, indicating the growing importance of China in international trade in the 2010s. Additionally, the composition of the hubs became less geographically concentrated: while Hub 6 remained primarily used by European countries on the export side, Canada and Mexico became significant participants in this hub on the import side. Moreover, there was no hub predominantly dominated by Asian countries as in 2003. Japan, Korea, and Singapore were important members of Hub 3, which was shared with the US, Germany, the Netherlands, and France. This suggests that international trade became more global and less regional by 2012.

7 Conclusion

Modeling high-dimensional tensor time series has gathered increasing attention recently, owing to the availability of multidimensional datasets beyond the classical panel data structure. This paper considers matrix and tensor factor models with a CP low-rank structure, offering a generalization of classical vector factor models. We develop iterative simultaneous orthogonalization estimation procedures based on contemporary covariance, preserving the tensor data structure. Theoretical properties such as the rate of convergence and limiting distributions are investigated, assuming each tensor dimension is comparable to or greater than the number of observations, and the tensor rank might be fixed or divergent.

In contrast to auto-covariance-based estimation methods, we explore information from contemporary data and are also able to consistently estimate loadings and factors for uncorrelated tensor observations where auto-covariance methods might fail. Additionally, we propose two generalized eigenvalue-ratio estimators for rank selection and justify their consistency. A comprehensive simulation study underscores the merits of our proposed method compared to existing methods. Furthermore, empirical applications regarding sorted portfolios and international trade flows showcase the practical relevance of our approach.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014a). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

Anandkumar, A., Ge, R., and Janzamin, M. (2014b). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*.

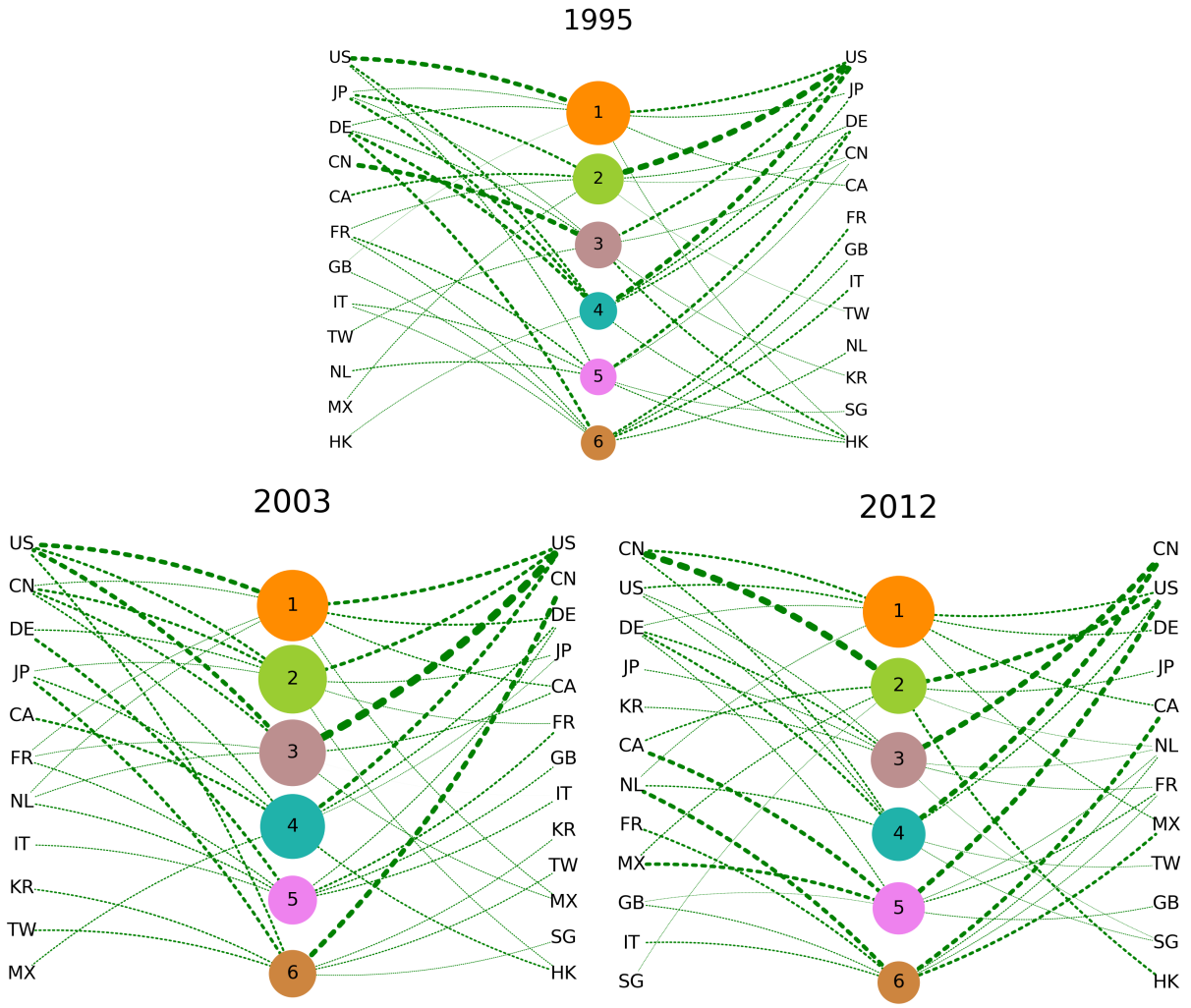


Figure 10: Trading volume network plots of 3 selected sub-samples. Notes: 1. Relations of the export (import) side between countries and latent hubs are represented on the left (right) half of the figure. 2. On the left(right)-hand side are the union of countries with top 4 contributions to each hub on the export (import) side. 3. Countries are ranked by the total volume of export (left) and import (right) to/from countries in the sample. 4. Sizes of latent hub nodes represent the strengths of the latent hub. 5. Thickness of dotted lines between countries and latent hubs represent the level of contribution.

Babii, A., Ghysels, E., and Pan, J. (2023). Tensor principal component analysis. *Working paper*.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

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

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

Chang, J., He, J., Yang, L., and Yao, Q. (2023). Modelling matrix time series via a tensor CP-decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):127–148.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264.

Chen, E. Y. and Chen, R. (2022). Modeling dynamic transport network with matrix factor models: an application to international trade flow. *Journal of Data Science*, 21(3):490–507.

Chen, E. Y. and Fan, J. (2023). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055.

Chen, E. Y., Xia, D., Cai, C., and Fan, J. (2024). Semi-parametric tensor factor analysis by iteratively projected singular value decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae001.

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116.

Chen, W. and Lam, C. (2024). Rank and factor loadings estimation in time series tensor factor model by pre-averaging. *The Annals of Statistics*, 52(1):364–391.

Fan, J., Li, K., and Liao, Y. (2021). Recent developments in factor models and applications in econometric learning. *Annual Review of Financial Economics*, 13:401–430.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. Springer-Verlag, New York.

Han, Y., Chen, R., Yang, D., and Zhang, C.-H. (2022a). Tensor factor model estimation by iterative projection. *working paper*.

Han, Y., Chen, R., and Zhang, C.-H. (2022b). Rank determination in tensor factor model. *Electronic Journal of Statistics*, 16(1):1726–1803.

Han, Y., Yang, D., Zhang, C.-H., and Chen, R. (2023). CP factor model for dynamic tensors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, in press.

Han, Y. and Zhang, C.-H. (2023). Tensor principal component analysis in high dimensional cp models. *IEEE Transactions on Information Theory*, 69(2):1147–1167.

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

Koltchinskii, V., Löffler, M., and Nickl, R. (2020). Efficient estimation of linear functionals of principal components. *The Annals of Statistics*, 48(1):464 – 490.

Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976–2013.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Lettau, M. (2023). High-dimensional factor models and the factor zoo. *working paper*.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474.

Rosenblatt, M. (2012). *Markov Processes, Structure and Asymptotic Behavior: Structure and Asymptotic Behavior*, volume 184. Springer Science & Business Media.

Shu, H. and Nan, B. (2019). Estimation of large covariance and precision matrices from temporally dependent observations. *The Annals of Statistics*, 47(3):1321–1350.

Stock, J. and Watson, M. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. *Handbook of macroeconomics*, 2:415–512.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Sun, W. W., Lu, J., Liu, H., and Cheng, G. (2017). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(79):899–916.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

Tsay, R. S. (2005). *Analysis of Financial Time Series*, volume 543. John Wiley & Sons.

Tsay, R. S. and Chen, R. (2018). *Nonlinear Time Series Analysis*, volume 891. John Wiley & Sons.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, P.-A. and Lu, C.-J. (2017). Tensor decomposition via simultaneous power iteration. In *International Conference on Machine Learning*, pages 3665–3673. PMLR.

Wedin, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Xia, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851.

Supplementary Material of "Estimation and Inference for CP Tensor Factor Models"

Appendix A Proofs of Main Theorem

Proof of Theorem 4.1. **Part (i)**. Recall $\widehat{\Sigma} = T^{-1}\sum_{t=1}^{T}\mathcal{Y}_t\otimes\mathcal{Y}_t$, $\Theta = \mathbb{E}F_tF_t^\top$, $a_i = \mathrm{vec}(a_{i1}\otimes a_{i2}\otimes\cdots\otimes a_{iK})$, $d = d_1d_2...d_K$. Let $e_t = \mathrm{vec}(\mathcal{E}_t)$. Write

$$\widetilde{\Sigma} := \mathrm{mat}_{[K]}\big(\widehat{\Sigma}\big) = A\Theta A^\top + \Psi^*.$$

Let $\Theta \in \mathbb{R}^{r\times r}$ have the eigenvalue decomposition $\Theta = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^\top$, where the diagonal matrix $\widetilde{\Lambda}$ denotes its eigenvalues. Let $U = (u_1,\ldots,u_r)$ be the orthonormal matrix corresponding to $A\widetilde{V}$ as in Lemma B.1. We have $\|AA^\top - UU^\top\|_2 \leqslant \delta$ and $\|A\Theta A^\top - U\widetilde{\Lambda}U^\top\|_2 \leqslant \lambda_1\delta$ by the applications of the error bound in Lemma B.1 with $\Lambda = \widetilde{\Lambda}$ the first time.

Let the top r eigenvectors of $\widetilde{\Sigma}$ be $\widehat{U} = (\widehat{u}_1,...,\widehat{u}_r) \in \mathbb{R}^{d\times r}$. By Wedin's perturbation theorem (Wedin, 1972) for any $1 \leqslant j \leqslant r$,

$$\|\widehat{u}_j\widehat{u}_j^\top - u_ju_j^\top\|_2 \leqslant 2\|A\Theta A^\top - U\widetilde{\Lambda}U^\top + \Psi^*\|_2/\lambda_{j,\pm} \leqslant \big(2\lambda_1\delta + 2\|\Psi^*\|_2\big)/\lambda_{j,\pm}, \tag{33}$$

where $\lambda_{j,\pm} = \min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}$. Combining (33) and the inequality $\|AA^\top - UU^\top\|_2 \leqslant \delta$, we have

$$\|\widehat{u}_j\widehat{u}_j^\top - a_ja_j^\top\|_2 \leqslant \delta + \big(2\lambda_1\delta + 2\|\Psi^*\|_2\big)/\lambda_{j,\pm}. \tag{34}$$

We formulate each $\widehat{u}_j \in \mathbb{R}^d$ to be a K-way tensor $\widehat{U}_j \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Let $\widehat{U}_{jk} = \mathrm{mat}_k(\widehat{U}_j)$, which is viewed as an estimate of $a_{jk}\mathrm{vec}(\otimes_{l\neq k}^K a_{jl})^\top \in \mathbb{R}^{d_k\times(d/d_k)}$. Then a_{jk}^{rcpca} is the top left singular vector of $\widehat{U}_{jk}$. By Lemma B.2,

$$\|a_{jk}^{\mathrm{rcpca}}a_{jk}^{\mathrm{rcpca}\top} - a_{jk}a_{jk}^\top\|_2^2 \wedge (1/2) \leqslant \|\widehat{u}_j\widehat{u}_j^\top - a_ja_j^\top\|_2^2. \tag{35}$$

Substituting (34) and Lemma A.1 into the above equation, we have the desired results.

Part (ii). For simplicity, consider the most extreme case where $\min\{\lambda_i-\lambda_{i+1}, \lambda_i-\lambda_{i-1}\} \leqslant c\lambda_r$ for all i, with $\lambda_0 = \infty, \lambda_{r+1} = 0$, and c is sufficiently small constant. In such cases, we need to employ Procedure 2 to the entire sample covariance tensor $\widehat{\Sigma}$. Let the eigenvalue ratio $w := \lambda_1/\lambda_r = O(r)$. Without loss of generality, assume $\Theta_{11} \geqslant \Theta_{22} \geqslant \cdots \geqslant \Theta_{rr}$. In general, the statement in the theorem holds for number of initialization $L \geqslant Cd^2 \vee Cdr^{2w^2}$, where $a\vee b = \max\{a,b\}$. We prove the statements through induction on factor index i starting from $i = 1$ proceeding to $i = r$. By the

induction hypothesis, we already have estimators such that

$$\left\|\widehat{a}_{jk}^{\mathrm{rcpca}}\widehat{a}_{jk}^{\mathrm{rcpca}\top}-a_{jk}a_{jk}^{\top}\right\|_2\leqslant C\phi_0,\qquad 1\leqslant j\leqslant i-1, 1\leqslant k\leqslant K, \tag{36}$$

in an event Ω with high probability, where

$$\phi_0^2=C\left(\frac{R^{(0)}}{\lambda_r}+\left(\frac{R^{(0)}}{\lambda_r}\right)^{K-1}\left(\frac{\lambda_1}{\lambda_r}\right)+(\delta_1^2+\delta/\delta_1)\left(\frac{\lambda_1}{\lambda_r}\right)+\delta_1\right),$$

and $R^{(0)}=\phi^{(0)}$ is defined in (19) in Theorem 4.1.

Applying Lemma A.2, we obtain that at the i-th step (i-th factor), we have

$$\left\|\widetilde{a}_{\ell k}\widetilde{a}_{\ell k}^{\top}-a_{ik}a_{ik}^{\top}\right\|_2\leqslant\phi_0^2,\qquad 1\leqslant k\leqslant K,$$

in the event Ω with probability at least $1-T^{-c}-d^{-c}$ for at least one $\ell\in[L]$. It follows that this estimator $\widetilde{a}_{\ell k}$ satisfies

$$\begin{aligned}\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell k}\right\|_2&\geqslant\left\|\sum_{j_1,j_2=1}^{r}\Theta_{j_1,j_2}\prod_{k=1}^{K}(a_{j_1k}^{\top}\widetilde{a}_{\ell k})(a_{j_2k}^{\top}\widetilde{a}_{\ell k})\right\|_2-\left\|\Psi\times_{k=1}^{2K}\widetilde{a}_{\ell k}\right\|_2\\&\geqslant\left\|\Theta_{ii}\prod_{k=1}^{K}(a_{ik}^{\top}\widetilde{a}_{\ell k})^2\right\|_2-\left\|\sum_{(j_1,j_2)\neq(i,i)}^{r}\Theta_{j_1,j_2}\prod_{k=1}^{K}(a_{j_1k}^{\top}\widetilde{a}_{\ell k})(a_{j_2k}^{\top}\widetilde{a}_{\ell k})\right\|_2-\left\|\Psi\times_{k=1}^{2K}\widetilde{a}_{\ell k}\right\|_2\end{aligned}$$

where Ψ is defined by unfolding Ψ^* into a $d_1\times d_2\times\cdots\times d_K\times d_1\times d_2\times\cdots\times d_K$ tensor. Let $\psi_i=CR^{(0)}/\Theta_{ii}+\delta_1^2w$. By (53) and the last part of the proof of Lemma A.2, as $\|\Psi^*\|_2/\lambda_r\leqslant\phi_0^2$ and $(1+\delta_1)\prod_{k=2}^{K}(\delta_k+\psi_i)w\leqslant\phi_0^2$, it follows that

$$\begin{aligned}\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell k}\right\|_2&\geqslant(1-\phi_0^4)^K\Theta_{ii}-(1+\delta_1)\prod_{k=2}^{K}(\delta_k+\psi_i)\lambda_1-\phi_0^2\Theta_{ii}\\&\geqslant(1-3\phi_0^2)\Theta_{ii}.\end{aligned}$$

Now consider the best initialization $\ell_*\in[L]$ by using $\ell_*=\arg\max_s|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{sk}|$. By the calculation above, it is immediate that

$$\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell_*k}\right\|_2\geqslant(1-3\phi_0^2)\Theta_{ii}. \tag{37}$$

Let $\widetilde{a}_{\ell_*}=\mathrm{vec}(\widetilde{a}_{\ell_*1}\otimes\cdots\otimes\widetilde{a}_{\ell_*K})$. If $\|a_ia_i^{\top}-\widetilde{a}_{\ell_*}\widetilde{a}_{\ell_*}^{\top}\|_2\geqslant C\phi_0$ for a sufficiently large constant C, we

have that

$$\begin{aligned}\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell_* k}\right\|_2 &\leqslant \left\|\sum_{j_1,j_2=1}^{r}\Theta_{j_1,j_2}(a_{j_1}^\top\widetilde{a}_{\ell_*})(a_{j_2}^\top\widetilde{a}_{\ell_*})\right\|_2 + \left\|\Psi\times_{k=1}^{2K}\widetilde{a}_{\ell k}\right\|_2\\ &\leqslant \left\|\sum_{j_1,j_2=i}^{r}\Theta_{j_1,j_2}(a_{j_1}^\top\widetilde{a}_{\ell_*})(a_{j_2}^\top\widetilde{a}_{\ell_*})\right\|_2 + \phi_0^2\Theta_{ii} + r\nu^{2K}\lambda_1\\ &\leqslant (1+\delta_1)(1-C^2\phi_0^2)\Theta_{ii} + \phi_0^2\Theta_{ii} + r\nu^{2K}\lambda_1.\end{aligned}$$

If ν satisfies $r\nu^{2K}(\lambda_1/\lambda_r)\leqslant c\phi_0^2$ for a small positive constant c, as $\delta_1\leqslant\phi_0^2$, we have

$$\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell_* k}\right\|_2 \leqslant (1-C'\phi_0^2)\Theta_{ii},$$

where C' is a sufficiently large constant. It contradicts (37) above. This implies that for $\ell=\ell_*$, we have

$$\|a_i a_i^\top - \widetilde{a}_{\ell_*}\widetilde{a}_{\ell_*}^\top\|_2 \leqslant C\phi_0.$$

By Lemma B.2, with $\widehat{a}_{ik}^{\mathrm{rcpca}} = \widetilde{a}_{\ell_* k}$, in the event Ω with probability at least $1-T^{-c}-d^{-c}$,

$$\left\|\widehat{a}_{ik}^{\mathrm{rcpca}}\widehat{a}_{ik}^{\mathrm{rcpca}\top} - a_{ik}a_{ik}^\top\right\|_2 \leqslant C\phi_0, \quad 1\leqslant k\leqslant K.$$

This finishes the proof of part (ii) by an induction argument along with the requirements $r\nu^{2K}(\lambda_1/\lambda_r)\leqslant c\phi_0^2$. □

Lemma A.1. *Suppose Assumptions 4.1, 4.2, 4.3 hold and $\delta<1$. Let $\widetilde{\Sigma}=A\Theta A^\top+\Psi^*$ and $1/\gamma=2/\gamma_1+1/\gamma_2$. In an event with probability at least $1-T^{-c}-d^{-c}$, we have*

$$\begin{aligned}\|\Psi^*\|_2 \leqslant C\lambda_1\left(\sqrt{\frac{r+\log T}{T}}+\frac{(r+\log T)^{1/\gamma}}{T}\right) + C\left(\sqrt{\frac{d\log(d)}{T}}+\frac{d\log(d)}{T}+1\right)\\ + C\lambda_1^{1/2}\left(\sqrt{\frac{d\log(d)}{T}}+\frac{d\log(d)}{T}\right).\end{aligned} \tag{38}$$

Proof. Let $\Upsilon_0 = T^{-1}\sum_{t=1}^{T}\sum_{i,j=1}^{r} f_{it}f_{jt}a_i a_j^\top$, $\overline{\mathbb{E}}(\cdot)=\mathbb{E}(\cdot|f_{it},1\leqslant i\leqslant r,1\leqslant t\leqslant T)$. Define $e_t=$

$\text{vec}(\mathcal{E}_t)$. Write

$$\begin{aligned}\widetilde{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T}\text{vec}(\mathcal{Y}_t)\text{vec}(\mathcal{Y}_t)^\top \\ &= A\Theta A^\top + \sum_{i,j=1}^{r}\frac{1}{T}\sum_{t=1}^{T}\left(f_{i,t}f_{j,t} - \mathbb{E}f_{i,t}f_{j,t}\right)a_i a_j^\top + \frac{1}{T}\sum_{t=1}^{T}e_t e_t^\top \\ &\quad + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r}f_{i,t}a_i e_t^\top + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r}f_{i,t}e_t a_i^\top \\ &:= A\Theta A^\top + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4.\end{aligned}$$

That is, $\Psi^* = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$.

We first bound $\|\Delta_1\|_2$. Note that $\Delta_1 = A\left(\widehat{\Theta} - \Theta\right)A^\top$. For any unit vector u in $\mathbb{R}^r$, there exist $u_j \in \mathbb{R}^r$ with $\|u_j\|_2 \leqslant 1$, $j = 1, ..., N_{r,\epsilon}$ such that $\max_{\|u\|_2\leqslant 1}\min_{1\leqslant j\leqslant N_{r,\epsilon}}\|u - u_j\|_2 \leqslant \epsilon$. The standard volume comparison argument implies that the covering number $N_{r,\epsilon} = \lfloor(1 + 2/\epsilon)^r\rfloor$. Then, there exist $u_j \in \mathbb{R}^r$, $1 \leqslant j \leqslant N_{r,1/3} := 7^r$, such that $\|u_j\|_2 = 1$ and

$$\|\widehat{\Theta} - \Theta\|_2 - \max_{1\leqslant j\leqslant N_{r,1/3}}\left|u_j^\top(\widehat{\Theta} - \Theta)u_j\right| \leqslant (2/3)\|\widehat{\Theta} - \Theta\|_2.$$

It follows that

$$\|\widehat{\Theta} - \Theta\|_2 \leqslant 3\max_{1\leqslant j\leqslant N_{r,1/3}}\left|u_j^\top(\widehat{\Theta} - \Theta)u_j\right|.$$

As $1/\gamma = 2/\gamma_1 + 1/\gamma_2$, by Theorem 1 in Merlevède et al. (2011),

$$\begin{aligned}\mathbb{P}\left(T\left|u_j^\top\Theta^{-1/2}(\widehat{\Theta} - \Theta)\Theta^{-1/2}u_j\right| \geqslant x\right) &\leqslant T\exp\left(-\frac{x^\gamma}{c_1}\right) + \exp\left(-\frac{x^2}{c_2 T}\right) \\ &\quad + \exp\left(-\frac{x^2}{c_3 T}\exp\left(\frac{x^{\gamma(1-\gamma)}}{c_4(\log x)^\gamma}\right)\right). \qquad (39)\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\left(T\left\|\Theta^{-1/2}(\widehat{\Theta} - \Theta)\Theta^{-1/2}\right\|_2/3 \geqslant x\right) &\leqslant N_{r,1/3}^2 T\exp\left(-\frac{x^\gamma}{c_1}\right) + N_{r,1/3}^2\exp\left(-\frac{x^2}{c_2 T}\right) \\ &\quad + N_{r,1/3}^2\exp\left(-\frac{x^2}{c_3 T}\exp\left(\frac{x^{\gamma(1-\gamma)}}{c_4(\log x)^\gamma}\right)\right).\end{aligned}$$

Choosing $x \asymp \sqrt{T(r + \log T)} + (r + \log T)^{1/\gamma}$, in an event Ω_1 with probability at least $1 - T^{-c_1}/2$,

$$\left\|\Theta^{-1/2}(\widehat{\Theta} - \Theta)\Theta^{-1/2}\right\|_2 \leqslant C\sqrt{\frac{r + \log(T)}{T}} + \frac{C(r + \log T)^{1/\gamma}}{T}.$$

It follows that, in the event Ω_1,

$$\begin{aligned}\|\Delta_1\|_2 &\leqslant \|A\|_2^2\lambda_1 \cdot \left\|\Theta^{-1/2}(\widehat{\Theta}-\Theta)\Theta^{-1/2}\right\|_2 \\ &\leqslant C\lambda_1\left(\sqrt{\frac{r+\log T}{T}}+\frac{(r+\log T)^{1/\gamma}}{T}\right),\end{aligned}$$

and,

$$\begin{aligned}\|\Upsilon_0\|_2 &\leqslant \left\|A\Theta A^\top\right\|_2 + \left\|A\left(\widehat{\Theta}-\Theta\right)A^\top\right\|_2 \\ &\leqslant \|A\|_2^2\lambda_1 + \left\|A\left(\widehat{\Theta}-\Theta\right)A^\top\right\|_2 \\ &\leqslant (1+\delta)\lambda_1 + C\lambda_1\left(\sqrt{\frac{r+\log T}{T}}+\frac{(r+\log T)^{1/\gamma}}{T}\right) \\ &:= \Delta_\Upsilon.\end{aligned} \tag{40}$$

Note that $\Delta_\Upsilon \lesssim \lambda_1$.

Next, consider $\|\Delta_2\|_2$. By Assumption 4.1 and Lemma A.1 in Shu and Nan (2019), we have

$$\begin{aligned}&\mathbb{P}\left(\|e_t\|_2^2 - \mathbb{E}\|e_t\|_2^2 \geqslant x\right) = \mathbb{P}\left(\xi_t^\top H^\top H\xi_t - \mathbb{E}\xi_t^\top H^\top H\xi_t \geqslant x\right) \\ &\leqslant 4\exp\left(-C'\left(\frac{x}{\|H^\top H\|_{\mathrm{F}}}\right)^{\frac{1}{1+2/\vartheta}}\right) \leqslant 4\exp\left(-C'\left(\frac{x}{\sqrt{d}}\right)^{\frac{\vartheta}{\vartheta+2}}\right).\end{aligned}$$

Note that $\mathbb{E}\|e_t\|_2^2 = \mathbb{E}\xi_t^\top H^\top H\xi_t = \mathbb{E}\mathrm{tr}(H^\top H\xi_t\xi_t^\top) = \mathrm{tr}(H^\top H) = \mathrm{tr}(HH^\top) \asymp d$, and $\|H^\top H\|_{\mathrm{F}}^2 = \|HH^\top\|_{\mathrm{F}}^2 = \|\Sigma_e\|_{\mathrm{F}}^2 \asymp d$. Choosing $x \asymp d$, we have

$$\mathbb{P}\left(\|e_t\|_2 \geqslant C\sqrt{d}\right) \leqslant 4\exp\left(-C'd^{\frac{\vartheta}{2\vartheta+4}}\right).$$

Let $N := \|e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}\|_2$ and $\sigma_0^2 := \|\sum_{t=1}^T \mathbb{E}(e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}})^2\|_2$. Then, by Assumption 4.1, $N \leqslant C^2 d$ and

$$\begin{aligned}\sigma_0^2 &\leqslant T\|\mathbb{E}(e_te_t^\top e_te_t^\top)\|_2 = T\|\mathbb{E}(H\xi_t\xi_t^\top H^\top H\xi_t\xi_t^\top H^\top)\|_2 \\ &\leqslant T\|\mathbb{E}(H\xi_t\xi_t^\top H^\top H\xi_t\xi_t^\top H^\top)\|_{\mathrm{F}} = T\|\mathbb{E}(\xi_t^\top H^\top H\xi_t\xi_t^\top H^\top H\xi_t)\|_{\mathrm{F}} \\ &= T\mathbb{E}\sum_{j,l}(H^\top H)_{jl}\xi_{jt}\xi_{lt}\sum_{j'l'}(H^\top H)_{j'l'}\xi_{j't}\xi_{l't} \\ &\leqslant C_0'T\left(\sum_{j,l}(H^\top H)_{jl}^2 + \sum_{j,l}(H^\top H)_{jj}(H^\top H)_{ll}\right) \\ &= C_0'T\left(\|H^\top H\|_{\mathrm{F}}^2 + [\mathrm{tr}(H^\top H)]^2\right) \\ &\leqslant C_0Td.\end{aligned}$$

By matrix Bernstein inequality (see, e.g., Theorem 5.4.1 of Vershynin (2018)),

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T}\left[e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}-\mathbb{E}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}\right]\right\|_2\geqslant x\right)\leqslant 2d\exp\left(-\frac{x^2/2}{\sigma_0^2+Nx/3}\right). \tag{41}$$

Choosing $x\asymp\sqrt{Td\log(d)}+d\log(d)$, with probability at least $1-d^{-c_1}$,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}-\mathbb{E}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}\right\|_2\leqslant C_1\sqrt{\frac{d\log(d)}{T}}+C_1\cdot\frac{d\log(d)}{T} \tag{42}$$

Define $M:=\{1\leqslant t\leqslant T:\|e_t\|_2\geqslant C\sqrt{d}\}$. Since $\mathbf{1}_{\{\|e_t\|_2\geqslant C\sqrt{d}\}}$ are independent Bernoulli random variable and $\log(T)\leqslant d^{\vartheta/(2\vartheta+4)}$, we have

$$\mathbb{E}|M|=T\mathbb{P}\left(\|e_t\|_2\geqslant C\sqrt{d}\right)\leqslant 4T\exp\left(-C'd^{\frac{\vartheta}{2\vartheta+4}}\right)\leqslant T^{-c_2}.$$

By Chernoff bound for Bernoulli random variables,

$$\mathbb{P}(|M|\geqslant C)\leqslant\exp\left(-T^{c_2}\right). \tag{43}$$

It follows that

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{t=1}^{T}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\geqslant C\sqrt{d}\}}\right\|_2\geqslant x\right)&\leqslant\mathbb{P}\left(|M|\max_t\|e_t\|_2^2\geqslant x\right)\\&\leqslant\mathbb{P}(|M|\geqslant C)+\mathbb{P}\left(|M|<C,|M|\max_t\|e_t\|_2^2\geqslant x\right)\\&\leqslant\exp\left(-T^{c_2}\right)+\mathbb{P}\left(\max_t\|e_t\|_2^2\geqslant x/C\right)\end{aligned}$$

Choosing $x\asymp d$, we have, with probability at least $1-\exp(-T^{c_2})-T^{-c_2}$,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\geqslant C\sqrt{d}\}}\right\|_2\leqslant C_2\cdot\frac{d}{T}. \tag{44}$$

Similarly,

$$\mathbb{P}\left(\left\|\mathbb{E}e_te_t^\top \mathbf{1}_{\{\|e_t\|_2\geqslant C\sqrt{d}\}}\right\|_2>0\right)=\mathbb{P}\left(\|e_t\|_2\geqslant C\sqrt{d}\right)\leqslant 4\exp\left(-C'd^{\frac{\vartheta}{2\vartheta+4}}\right)\leqslant T^{-c_3}. \tag{45}$$

Combing (42), (44), (45), in an event Ω_2 with probability at least $1 - T^{-c_4} - d^{-c_1}$,

$$\begin{aligned}
\|\Delta_2\|_2 &\leqslant \left\| \frac{1}{T}\sum_{t=1}^{T} e_t e_t^\top - \Sigma_e \right\|_2 + \|\Sigma_e\|_2 \\
&\leqslant \left\| \frac{1}{T}\sum_{t=1}^{T} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} - \mathbb{E} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2 + \left\| \frac{1}{T}\sum_{t=1}^{T} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 \\
&\quad + \left\| \mathbb{E} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 + \|\Sigma_e\|_2 \\
&\leqslant C_2 \sqrt{\frac{d\log(d)}{T}} + C_2 \cdot \frac{d\log(d)}{T} + C_2
\end{aligned}$$

Next, consider $\|\Delta_3\|_2$. Note that $\|\Delta_4\|_2$ is the same as $\|\Delta_3\|_2$. Let $\overline{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot|F_1, ..., F_T)$ and $\overline{\mathbb{E}}(\cdot) = \mathbb{E}(\cdot|F_1, ..., F_T)$ be the conditional probability and conditional expectation given the factor process, respectively. Similar to the derivation for $\|\Delta_2\|_2$, let

$$\begin{aligned}
N_1 &:= \left\| \sum_{i=1}^{r} f_{it} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2, \\
\sigma_1^2 &:= \max\left\{ \left\| \sum_{t=1}^{T} \overline{\mathbb{E}} \sum_{i,j=1}^{r} f_{it} f_{jt} a_i e_t^\top e_t a_j^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2, \left\| \sum_{t=1}^{T} \overline{\mathbb{E}} \sum_{i,j=1}^{r} f_{it} f_{jt} e_t a_i^\top a_j e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2 \right\}.
\end{aligned}$$

It is easy to show

$$\begin{aligned}
N_1 &\leqslant C\sqrt{d} \left\| \sum_{i=1}^{r} f_{it} a_i \right\|_2, \\
\sigma_1^2 &\leqslant C_3 T d \max\left\{ \left\| \frac{1}{T}\sum_{t=1}^{T}\sum_{i,j=1}^{r} f_{it} f_{jt} a_i a_j^\top \right\|_2, \left\| \frac{1}{Td}\sum_{t=1}^{T}\sum_{i,j=1}^{r} f_{it} f_{jt} a_i^\top a_j \right\|_2 \right\} := \sigma_2^2.
\end{aligned}$$

By matrix Bernstein inequality,

$$\overline{\mathbb{P}}\left(\left\| \sum_{t=1}^{T} \left[\sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} - \overline{\mathbb{E}} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right] \right\|_2 \geqslant x \right) \leqslant 2d \exp\left(-\frac{x^2/2}{\sigma_1^2 + N_1 x/3} \right).$$

Choosing $x \asymp \sqrt{d}\log(d)\|\sum_{i=1}^{r} f_{it} a_i\|_2 + \sqrt{\log(d)}\sigma_2$, with probability at least $1 - d^{-c_4}$,

$$\left\| \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} - \overline{\mathbb{E}} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2 \leqslant C_4 \frac{\sqrt{\log(d)}}{T}\sigma_2 + C_4 \frac{\sqrt{d}\log(d)\|\sum_{i=1}^{r} f_{it} a_i\|_2}{T}.$$

As $\sqrt{r}\log(T)^{1/\gamma_1} \lesssim \sqrt{d}$, by Assumption 4.2, with probability at least $1 - T^{-c_5}$,

$$\left\| \sum_{i=1}^{r} f_{it} a_i \right\|_2 = \|F_t^\top A^\top\|_2 \leqslant (1+\delta)\|F_t\|_2 \lesssim \sqrt{r}(\log(T))^{1/\gamma_1}\sqrt{\lambda_1} \lesssim \sqrt{d\lambda_1}.$$

By (40), in the event Ω_1, $\sigma_2^2 \lesssim Td\lambda_1$. Then, with probability at least $1 - T^{-c_1}/2 - T^{-c_5} - d^{-c_4}$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} - \mathbb{E} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2 \leqslant C_5 \sqrt{\frac{d\log(d)}{T}} \sqrt{\lambda_1} + C_5 \cdot \frac{d\log(d)\sqrt{\lambda_1}}{T}.$$

Similar to (44),

$$\mathbb{P}\left(\left\| \sum_{t=1}^{T} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 \geqslant x \right) \leqslant \mathbb{P}(|M| > C) + \mathbb{P}\left(|M| < C, |M| \max_t \left\| \sum_{i=1}^{r} f_{it} a_i \right\|_2 \cdot \|e_t\|_2 \geqslant x \right).$$

Choosing $x \asymp d\sqrt{\lambda_1}$, we have with probability at least $1 - \exp(-T^{c_2}) - T^{-c_2} - T^{-c_5}$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 \leqslant C_6 \cdot \frac{d\sqrt{\lambda_1}}{T}$$

Thus, in an event Ω_3 with probability $1 - T^{-c_1}/2 - T^{-c_6} - d^{-c_4}$,

$$\begin{aligned} \|\Delta_3\|_2 \leqslant & \left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} - \mathbb{E} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \leqslant C\sqrt{d}\}} \right\|_2 \\ & + \left\| \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 + \left\| \mathbb{E} \sum_{i=1}^{r} f_{i,t} a_i e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 \\ \leqslant & C_7 \sqrt{\frac{d\log(d)}{T}} \sqrt{\lambda_1} + C_7 \cdot \frac{d\log(d)\sqrt{\lambda_1}}{T}. \end{aligned}$$

Therefore, in the event $\Omega_1 \cap \Omega_2 \cap \Omega_3$ with probability at least $1 - T^{-c} - d^{-c}$, we have the desired bound for $\|\Psi^*\|_2$.

□

Lemma A.2. *Let $\lambda_1/\lambda_r = w$. Assume $d_1 \lesssim \lambda_1$ and $\delta_1^2 w \leqslant c$ for a sufficiently small positive constant c. Apply random projection in Procedure 2 to the whole sample covariance tensor $\widehat{\Sigma}$ with $L \geqslant Cd^2 \vee Cdr^{2w^2}$. Denote the estimated CP basis vectors as $\widetilde{a}_{\ell k}$, for $1 \leqslant \ell \leqslant L, 1 \leqslant k \leqslant K$. Then in an event with probability at least $1 - T^{-c} - d^{-c}$, we have for any CP factor loading vectors tuple $(a_{ik}, 1 \leqslant k \leqslant K)$, there exist $j_i \in [L]$ such that*

$$\left\| \widetilde{a}_{j_i,k} \widetilde{a}_{j_i,k}^\top - a_{ik} a_{ik}^\top \right\|_2 \leqslant \psi_i, \qquad 2 \leqslant k \leqslant K, \tag{46}$$

$$\left\| \widetilde{a}_{j_i,1} \widetilde{a}_{j_i,1}^\top - a_{i1} a_{i1}^\top \right\|_2 \leqslant \psi_i + (\delta/\delta_1) w + \psi_i^{K-1} w, \tag{47}$$

where $\psi_i = CR^{(0)}/\Theta_{ii} + \delta_1^2 w$, $R^{(0)} = \phi^{(0)}$ is defined in (19) in Theorem 4.1, and $1 \leqslant i \leqslant r$.

Proof. Without loss of generality, assume $\Theta_{11} \geqslant \Theta_{22} \geqslant \cdots \geqslant \Theta_{rr}$. Then $\Theta_{rr} \geqslant \lambda_r, \Theta_{11} \leqslant \lambda_1$. Let

$\widehat{\Theta}_{ij} = T^{-1}\sum_{t=1}^{T} f_{it}f_{jt}$ and $\Sigma_{\mathcal{E}} = \mathbb{E}\mathcal{E}_t \otimes \mathcal{E}_t$. Write

$$\begin{aligned}
\widehat{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T} \mathcal{Y}_t \otimes \mathcal{Y}_t \\
&= \sum_{i,j=1}^{r} \Theta_{ij} \otimes_{k=1}^{K} a_{ik} \otimes_{k=K+1}^{2K} a_{jk} + \sum_{i,j=1}^{r} (\widehat{\Theta}_{ij} - \Theta_{ij}) \otimes_{k=1}^{K} a_{ik} \otimes_{k=K+1}^{2K} a_{jk} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r} f_{it} \otimes_{k=1}^{K} a_{ik} \otimes \mathcal{E}_t + \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r} f_{it}\mathcal{E}_t \otimes_{k=K+1}^{2K} a_{ik} + \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{E}_t \otimes \mathcal{E}_t - \Sigma_{\mathcal{E}}\right) + \Sigma_{\mathcal{E}} \\
&:= \sum_{i,j=1}^{r} \Theta_{ij} \otimes_{k=1}^{K} a_{ik} \otimes_{k=K+1}^{2K} a_{jk} + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5,
\end{aligned}$$

with $a_{i,K+k} = a_{ik}$ for all $1 \leqslant k \leqslant K$. Let $\Psi = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5$. Let $\Xi(\theta) = \text{mat}_{[K-1]}\widehat{\Sigma} \times_1 \times_{K+1}\theta$. Unfold $\Psi \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_K\times d_1\times d_2\times\cdots\times d_K}$ to be an order 4 tensor of dimension $(d/d_1) \times (d/d_1) \times d_1 \times d_1$ and denote it as $\widetilde{\Psi}$, and also define $\widetilde{\Delta}_k, k = 1, ..., 5$ in a similar way. Then

$$\Xi(\theta) = \sum_{i,j=1}^{r} \Theta_{ij}(a_{i1}^{\top}\theta a_{j1})\widetilde{a}_i\widetilde{a}_j^{\top} + \widetilde{\Psi} \times_3 \times_4 \theta,$$

where $\widetilde{a}_i = \text{vec}(a_{i2} \otimes \cdots a_{iK})$. Let $\widetilde{A} = (\widetilde{a}_1, ..., \widetilde{a}_r) \in \mathbb{R}^{(d/d_1)\times r}$.

First, consider the upper bound of $\|\widetilde{\Psi} \times_3 \times_4 \theta\|_2$. By concentration inequality for matrix Gaussian sequence (see, for example Theorem 4.1.1 in Tropp et al. (2015)) and employing similar arguments in the proof of Lemma A.1, we have, with probability at least $1 - d^{-c}$

$$\begin{aligned}
\left\|\widetilde{\Delta}_4 \times_3 \times_4 \theta\right\|_2 &= \left\|\sum_{k,l} \theta_{(kl)}(\widetilde{\Delta}_4)_{\cdot\cdot kl}\right\|_2 \\
&\leqslant C \max\left\{\left\|\text{mat}_{(1),(234)}(\widetilde{\Delta}_4)\right\|_2, \left\|\text{mat}_{(2),(134)}(\widetilde{\Delta}_4)\right\|_2\right\} \cdot \sqrt{\log(d)} \\
&\leqslant C\sqrt{d_1}\left(\sqrt{\frac{d\log(d)}{T}} + \frac{d\log(d)}{T}\right) \cdot \sqrt{\log(d)},
\end{aligned}$$

where $\theta_{(kl)}$ is the (k,l)th element of θ, $(\widetilde{\Delta}_4)_{\cdot\cdot kl}$ represents the (k,l)th (3,4) slice of $\widetilde{\Delta}_4$, and $\text{mat}_{(1),(234)}(\cdot)$ denotes the reshaping of fourth-order tensor into a matrix by collapsing its first indices as rows, and the second, third, fourth indices as columns. In the last step, we apply the arguments in the proof of Lemma A.1. Similarly, we have, with probability at least $1 - d^{-c}$,

$$\|\Sigma_{\mathcal{E}} \times_3 \times_4 \theta\|_2 \leqslant C\sqrt{\log(d)}.$$

And, with probability at least $1 - T^{-c} - d^{-c}$,

$$\left\|(\widetilde{\Delta}_1 + \widetilde{\Delta}_2 + \widetilde{\Delta}_3 + \widetilde{\Delta}_4) \times_3 \times_4 \theta\right\|_2 \leqslant C\lambda_1 \left(\sqrt{\frac{r + \log T}{T}} + \frac{(r + \log T)^{1/\gamma}}{T}\right) \cdot \sqrt{\log(d)} + C\lambda_1^{1/2} \left(\sqrt{\frac{d\log(d)}{T}} + \frac{d\log(d)}{T}\right) \cdot \sqrt{\log(d)}.$$

As $d_1 \lesssim \lambda_1$, it follows that in an event Ω_0 with probability at least $1 - T^{-c} - d^{-c}$,

$$\left\|\widetilde{\Psi} \times_3 \times_4 \theta\right\|_2 \leqslant C\|\Psi^*\|_2\sqrt{\log(d)}. \tag{48}$$

Consider the i-th factor and rewrite $\Xi(\theta)$ as follows

$$\Xi(\theta) = \Theta_{ii}(a_{i1}^\top \theta a_{i1})\widetilde{a}_i\widetilde{a}_i^\top + \sum_{(j_1,j_2)\neq(i,i)} \Theta_{j_1j_2}(a_{j_11}^\top \theta a_{j_21})\widetilde{a}_{j_1}\widetilde{a}_{j_2}^\top + \widetilde{\Psi} \times_3 \times_4 \theta. \tag{49}$$

Suppose now we repeatedly sample $\theta_\ell \sim \theta$, for $\ell = 1, ..., L$. By the anti-concentration inequality for Gaussian random variables (see Lemma B.1 in Anandkumar et al. (2014a)), we have

$$\mathbb{P}\left(\max_{1\leqslant\ell\leqslant L}(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell) \leqslant \sqrt{2\log(L)} - \frac{\log\log(L)}{4\sqrt{\log(L)}} - \sqrt{2\log(8)}\right) \leqslant \frac{1}{4}, \tag{50}$$

where $\odot$ denotes Kronecker product. Let

$$\ell_* = \arg\max_{1\leqslant\ell\leqslant L}(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell).$$

Note that $(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell)$ and $(I_{d_1^2} - (a_{i1} \odot a_{i1})(a_{i1} \odot a_{i1})^\top)\mathrm{vec}(\theta_\ell)$ are independent. Since the definition of ℓ_* depends only on $(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell)$, this implies that the distribution of $(I_{d_1^2} - (a_{i1} \odot a_{i1})(a_{i1} \odot a_{i1})^\top)\mathrm{vec}(\theta_\ell)$ does not depend on ℓ_*.

By Gaussian concentration inequality of 1-Lipschitz function, we have

$$\mathbb{P}\left(\max_{j_1,j_2\leqslant r}\left(a_{j_11} \odot a_{j_21}\right)^\top \left(I_{d_1^2} - (a_{i1} \odot a_{i1})(a_{i1} \odot a_{i1})^\top\right)\mathrm{vec}(\theta_\ell) \geqslant \sqrt{4\log(r)} + \sqrt{2\log(8)}\right) \leqslant \frac{1}{4}.$$

Moreover, for the reminder bias term $(a_{j_11} \odot a_{j_21})^\top (a_{i1} \odot a_{i1})(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell)$, we have,

$$\begin{aligned}
&\left\|\sum_{(j_1,j_2)\neq(i,i)} \Theta_{j_1,j_2}(a_{j_11} \odot a_{j_21})^\top (a_{i1} \odot a_{i1})(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell) \cdot \widetilde{a}_{j_1}\widetilde{a}_{j_2}^\top\right\|_2 \\
&\leqslant (a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell) \cdot \left\|\widetilde{A}\left(\Theta \circ \left(A_1^\top a_{i1}a_{i1}^\top A_1 - e_{ii}\right)\right)\widetilde{A}^\top\right\|_2 \\
&\leqslant (a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell)\|\widetilde{A}\|_2^2\|\Theta\|_2 \left\|A_1^\top a_{i1}a_{i1}^\top A_1 - e_{ii}\right\|_2 \\
&\leqslant (1 + \delta/\delta_1)\delta_1^2\lambda_1(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell),
\end{aligned}$$

where $\circ$ denotes Hadamard product and e_{ii} is a $d_1 \times d_1$ matrix with the (i,i)-th element be 1 and all the others be 0.

Thus, we obtain the top eigengap

$$\begin{aligned}
&(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_{\ell *}) \Theta_{ii} - \left\| \sum_{(j_1,j_2) \neq (i,i)} \Theta_{j_1 j_2} \left((a_{j_1 1} \odot a_{j_2 1})^\top \mathrm{vec}(\theta_{\ell *}) \right) \widetilde{a}_{j_1} \widetilde{a}_{j_2}^\top \right\|_2 \\
&\geqslant (1 - 2\delta_1^2 w) \left(\sqrt{2\log(L)} - \frac{\log\log(L)}{4\sqrt{\log(L)}} - \sqrt{2\log(8)} \right) \Theta_{ii} - \left(\sqrt{4\log(r)} + \sqrt{2\log(8)} \right) w \Theta_{ii} \\
&\geqslant C_0 \sqrt{\log(d)} \Theta_{ii}, \qquad (51)
\end{aligned}$$

with probability at least $\frac{1}{2}$, by letting $L \geqslant Cd \vee Cr^{2w^2}$.

Since θ_ℓ are independent samples, we instead take $L_i = L_{i1} + \cdots + L_{iM}$ for $M = \lceil C_1 \log(d)/\log(2) \rceil$ and $L_{i1}, ..., L_{iM} \geqslant Cd \vee Cr^{2w^2}$. We define

$$\ell_*^{(m)} = \arg \max_{1 \leqslant \ell \leqslant L_{im}} (a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell), \quad \ell_* = \arg \max_{1 \leqslant \ell \leqslant L_i} (a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell).$$

We then have, by independence of θ_ℓ, that the above statement (51) for the i-th factor holds in an event Ω_i with probability at least $1 - d^{-C_1}$. By Wedin's perturbation theory, we have in the event $\Omega_0 \cap \Omega_i$,

$$\left\| \widehat{a}_{\ell_*} \widehat{a}_{\ell_*}^\top - \widetilde{a}_i \widetilde{a}_i^\top \right\|_2 \leqslant \frac{C R^{(0)}}{\Theta_{ii}} + \delta_1^2 w,$$

where $\widehat{a}_{\ell_*}$ is the top left singular vector of $\Xi(\theta_{\ell_*})$, and $R^{(0)} = \phi^{(0)}$ is defined in (19) in Theorem 4.1. By Lemma B.2,

$$\left\| \widetilde{a}_{\ell_*,k} \widetilde{a}_{\ell_*,k}^\top - a_{ik} a_{ik}^\top \right\|_2 \leqslant \frac{C R^{(0)}}{\Theta_{ii}} + \delta_1^2 w, \qquad 2 \leqslant k \leqslant K. \qquad (52)$$

Now consider to obtain $\widetilde{a}_{\ell_*,1}$. Write $\psi_i = CR^{(0)}/\Theta_{ii} + \delta_1^2 w$. Note that

$$\begin{aligned}
\widehat{\Sigma} \times_{k=2}^K \widetilde{a}_{\ell_*,k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell_*,k} = &\prod_{k=2}^K \left(\widetilde{a}_{\ell_*,k}^\top a_{ik} \right)^2 \Theta_{ii} a_{i1} a_{i1}^\top + \Psi \times_{k=2}^K \widetilde{a}_{\ell_*,k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell_*,k} \\
&+ \sum_{(j_1,j_2) \neq (i,i)} \prod_{k=2}^K \left(\widetilde{a}_{\ell_*,k}^\top a_{j_1 k} \right) \left(\widetilde{a}_{\ell_*,k}^\top a_{j_2 k} \right) \Theta_{j_1 j_2} a_{j_1 1} a_{j_2 1}^\top.
\end{aligned}$$

By Lemma A.1 and (52), in the event $\Omega_0 \cap \Omega_1$,

$$\begin{aligned}
&\left\|\Psi \times_{k=2}^{K} \widetilde{a}_{\ell_*,k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell_*,k}\right\|_2 \leqslant \|\Psi^*\|_2, \\
&\prod_{k=2}^{K} \left(\widetilde{a}_{\ell_*,k}^\top a_{ik}\right)^2 \geqslant (1-\psi_i^2)^{K-1}.
\end{aligned}$$

Since

$$\begin{aligned}
\max_{j_1 \neq i} \left|a_{j_1 k}^\top \widetilde{a}_{\ell_*,k}\right| &= \max_{j_1 \neq i} \left|\widetilde{a}_{\ell_*,k}^\top a_{ik} a_{ik}^\top a_{j_1 k} + \widetilde{a}_{\ell_*,k}^\top (I - a_{ik}a_{ik}^\top) a_{j_1 k}\right| \\
&\leqslant \max_{j_1 \neq i} \left|\widetilde{a}_{\ell_*,k}^\top a_{ik}\right| \left|a_{ik}^\top a_{j_1 k}\right| + \max_{j_1 \neq i} \left\|\widetilde{a}_{\ell_*,k}^\top (I - a_{ik}a_{ik}^\top)\right\|_2 \left\|(I - a_{ik}a_{ik}^\top) a_{j_1 k}\right\|_2 \\
&\leqslant \sqrt{1-\psi_i^2}\delta_k + \psi_i \sqrt{1-\delta_k^2} \leqslant \delta_k + \psi_i, \qquad (53)
\end{aligned}$$

we have

$$\begin{aligned}
\left\| \sum_{(j_1,j_2) \neq (i,i)} \prod_{k=2}^{K} \left(\widetilde{a}_{\ell_*,k}^\top a_{j_1 k}\right) \left(\widetilde{a}_{\ell_*,k}^\top a_{j_2 k}\right) \Theta_{j_1 j_2} a_{j_1 1} a_{j_2 1}^\top \right\|_2 &\leqslant (1+\delta_1) \prod_{k=2}^{K} (\delta_k + \psi_i) \lambda_1 \\
&\leqslant C_K (\delta/\delta_1 + \psi_i^{K-1}) w \Theta_{ii}.
\end{aligned}$$

By Wedin's perturbation theory,

$$\left\|\widetilde{a}_{\ell_*,1} \widetilde{a}_{\ell_*,1}^\top - a_{i1} a_{i1}^\top\right\|_2 \leqslant \frac{C R^{(0)}}{\Theta_{ii}} + (\delta/\delta_1) w + \psi_i^{K-1} w. \qquad (54)$$

Repeat the same argument again for all $1 \leqslant i \leqslant r$ factors, and let $L = \sum_i L_i \geqslant Cd^2 \vee Cdr^{2w^2} \geqslant Cdr\log(d) \vee Cr^{2w^2+1}\log(d)$. We have, in the event $\Omega_0 \cap \Omega_1 \cap \cdots \cap \Omega_r$ with probability at least $1 - T^{-c} - d^{-c}$, (52) and (54) hold for all i.

□

Proof of Theorem 4.2. Recall $\widehat{A}_k^{(m)} = (\widehat{a}_{1k}^{(m)}, \ldots, \widehat{a}_{rk}^{(m)}) \in \mathbb{R}^{d_k \times r}$, $\widehat{\Sigma}_k^{(m)} = \widehat{A}_k^{(m)\top} \widehat{A}_k^{(m)}$, and $\widehat{B}_k^{(m)} = \widehat{A}_k^{(m)} (\widehat{\Sigma}_k^{(m)})^{-1} = (\widehat{b}_{1k}^{(m)}, ..., \widehat{b}_{rk}^{(m)}) \in \mathbb{R}^{d_k \times r}$. Let $\overline{\mathbb{E}}(\cdot) = \mathbb{E}(\cdot | F_t, 1 \leqslant t \leqslant T)$ and $\overline{\mathbb{P}}(\cdot) = \mathbb{P}(\cdot | F_t, 1 \leqslant t \leqslant T)$. Also let

$$\overline{\lambda}_i = \frac{1}{T} \sum_{t=1}^{T} f_{it} f_{it}.$$

Without loss of generality, assume $\Theta_{11} \geqslant \Theta_{22} \geqslant \cdots \geqslant \Theta_{rr}$. Then $\mathbb{E}\overline{\lambda}_r = \Theta_{rr} \geqslant \lambda_r, \mathbb{E}\overline{\lambda}_1 = \Theta_{11} \leqslant \lambda_1$.

Write

$$\begin{aligned}\widehat{\Sigma} &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{Y}_t\otimes\mathcal{Y}_t\\ &= \sum_{i=1}^{r}\overline{\lambda}_i\otimes_{k=1}^{2K}a_{ik}+\sum_{i\neq j}^{r}\frac{1}{T}\sum_{t=1}^{T}f_{it}f_{jt}\otimes_{k=1}^{K}a_{ik}\otimes_{k=K+1}^{2K}a_{jk}+\frac{1}{T}\sum_{t=1}^{T}\mathcal{E}_t\otimes\mathcal{E}_t\\ &\quad+\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r}f_{it}\otimes_{k=1}^{K}a_{ik}\otimes\mathcal{E}_t+\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{r}f_{it}\mathcal{E}_t\otimes_{k=K+1}^{2K}a_{ik}\\ &:= \sum_{i=1}^{r}\overline{\lambda}_i\otimes_{k=1}^{2K}a_{ik}+\Delta_1+\Delta_2+\Delta_3+\Delta_4,\end{aligned}\tag{55}$$

with $a_{i,K+k}=a_{ik}$ for all $1\leqslant k\leqslant K$. Let $\Psi=\Delta_1+\Delta_2+\Delta_3+\Delta_4$.

By Theorem 4.1, in an event Ω_0 with probability at least $1-T^{-C_1}-d^{-C_1}$,

$$\|\widehat{a}_{ik}^{(0)}\widehat{a}_{ik}^{(0)\top}-a_{ik}a_{ik}^{\top}\|_2\leqslant\psi_0.$$

At m-th step, let

$$\psi_{m,i,k}:=\|\widehat{a}_{ik}^{(m)}\widehat{a}_{ik}^{(m)\top}-a_{ik}a_{ik}^{\top}\|_2,\qquad \psi_{m,k}:=\max_i\psi_{m,i,k},\qquad \psi_m=\max_k\psi_{m,k}.\tag{56}$$

Let $g_{i\ell}=b_{i\ell}/\|b_{i\ell}\|_2$ and $\widehat{g}_{i\ell}^{(m)}=\widehat{b}_{i\ell}^{(m)}/\|\widehat{b}_{i\ell}^{(m)}\|_2$. Given $\widehat{a}_{i\ell}^{(m)}$ $(1\leqslant i\leqslant r, 1\leqslant\ell\leqslant K)$, the $(m+1)$th iteration produces estimates $\widehat{a}_{ik}^{(m+1)}$, which is the top left singular vector of $\widehat{\Sigma}\times_{\ell\in[2K]\setminus\{k,K+k\}}\widehat{b}_{i\ell}^{(m)\top}$, or equivalently $\widehat{\Sigma}\times_{\ell\in[2K]\setminus\{k,K+k\}}\widehat{g}_{i\ell}^{(m)\top}$. Note that $\widehat{\Sigma}=\sum_{j=1}^{r}\overline{\lambda}_j\otimes_{\ell=1}^{2K}a_{j\ell}+\Psi$, with $a_{j,\ell+K}=a_{j\ell}$. The "noiseless" version of this update is given by

$$\widehat{\Sigma}\times_{\ell\in[2K]\setminus\{k,K+k\}}g_{i\ell}^{\top}=\overline{\lambda}_i a_{ik}a_{ik}^{\top}+\Psi\times_{\ell\in[2K]\setminus\{k,K+k\}}g_{i\ell}^{\top}.\tag{57}$$

At $(m+1)$-th iteration, for any $1\leqslant i\leqslant r$, we have

$$\widehat{\Sigma}\times_{\ell\in[2K]\setminus\{k,K+k\}}\widehat{g}_{i\ell}^{(m)\top}=\sum_{j=1}^{r}\widetilde{\lambda}_{j,i}a_{jk}a_{jk}^{\top}+\Psi\times_{\ell\in[2K]\setminus\{k,K+k\}}\widehat{g}_{i\ell}^{(m)\top},$$

where

$$\widetilde{\lambda}_{j,i}=\overline{\lambda}_j\prod_{\ell\in[2K]\setminus\{k,K+k\}}a_{j\ell}^{\top}\widehat{g}_{i\ell}^{(m)}.\tag{58}$$

Let

$$\begin{aligned}
\lambda_{j,i} &= \Theta_{jj} \prod_{\ell \in [2K] \setminus \{k, K+k\}} a_{j\ell}^\top \widehat{g}_{i\ell}^{(m)}, \\
\alpha &= \sqrt{1-\delta_{\max}} - (r^{1/2}+1)\psi_0 / \sqrt{1-1/(4r)}, \\
\phi_{m,\ell} &= 1 \wedge \frac{\psi_{m,\ell}\sqrt{2r}}{\alpha\sqrt{1-1/(4r)}}, \\
\phi_m &= \max_\ell \phi_{m,\ell}.
\end{aligned}$$

We may assume without loss of generality $a_{j\ell}^\top \widehat{a}_{j\ell}^{(m)} \geqslant 0$ for all (j,ℓ). Similar to the proofs of Theorem 3 in Han and Zhang (2023), we can show

$$\max_{j \leqslant r} \|\widehat{a}_{j\ell}^{(m)} - a_{j\ell}\|_2 \leqslant \psi_{m,\ell}/\sqrt{1-1/(4r)}, \quad \|\widehat{b}_{j\ell}^{(m)}\|_2 \leqslant \|\widehat{B}_\ell^{(m)}\|_2 \leqslant \left(\sqrt{1-\delta_\ell} - \frac{r^{1/2}\psi_0}{\sqrt{1-1/(4r)}}\right)^{-1}, \tag{59}$$

$$\|\widehat{g}_{j\ell}^{(m)} - b_{j\ell}/\|b_{j\ell}\|_2\|_2 \leqslant (\psi_{m,\ell}/\alpha)\sqrt{2r/(1-1/(4r))}. \tag{60}$$

Moreover, (59) provides

$$\max_{i \neq j} |a_{i\ell}^\top \widehat{g}_{j\ell}^{(m)}| \leqslant \psi_{m,\ell}/\sqrt{1-1/(4r)}, \quad |a_{j\ell}^\top \widehat{g}_{j\ell}^{(m)}| \geqslant \alpha, \tag{61}$$

as $\widehat{a}_{i\ell}^{(m)\top} \widehat{g}_{j\ell}^{(m)} = I\{i=j\}/\|\widehat{b}_{j\ell}^{(m)}\|_2$. Then, for $j \neq i$,

$$\lambda_{j,i}/\lambda_{i,i} \leqslant (\lambda_1/\Theta_{ii}) \prod_{\ell \neq k}^{K} \left(\frac{\psi_{m,\ell}/\sqrt{1-1/(4r)}}{1-\psi_{m,\ell}/\sqrt{1-1/(4r)}}\right)^2.$$

Employing similar arguments in the proof of Lemma A.1, in an event Ω_1 with probability at least $1 - T^{-c_1}$, we have

$$\left\|\widehat{\Theta} - \Theta\right\|_2 \leqslant C_1 \lambda_1 \left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right), \tag{62}$$

In the event Ω_1, we also have

$$\max_{1 \leqslant j_1, j_2 \leqslant r} \left|\frac{1}{T}\sum_{t=1}^{T} f_{j_1,t} f_{j_2,t} - \mathbb{E} f_{j_1,t} f_{j_2,t}\right| \leqslant C_1 \sqrt{\Theta_{j_1,j_1}\Theta_{j_2,j_2}} \left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right).$$

It follows that in the event Ω_1, for any $1 \leqslant j \leqslant r$,

$$|\overline{\lambda}_j - \Theta_{jj}| \leqslant C_1 \Theta_{jj} \left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right).$$

By Wedin's theorem (Wedin, 1972), in the event Ω_1,

$$\|\widehat{a}_{ik}^{(m+1)}\widehat{a}_{ik}^{(m+1)\top} - a_{ik}a_{ik}^\top\|_2 \leqslant \frac{2\left\|\sum_{j\neq i}^r \widetilde{\lambda}_{j,i}a_{jk}a_{jk}^\top\right\|_2 + 2\|\Psi \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2}{\widetilde{\lambda}_{i,i}}$$
$$\leqslant \frac{4\|A_k\|_2^2 \max_{j\neq i}|\lambda_{j,i}| + 4\|\Psi \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2}{\alpha^{2k-2}\Theta_{ii}}. \tag{63}$$

To bound the numerator of (63), we write

$$\Delta_{1,1} = \sum_{j_2\neq i}^r \frac{1}{T}\sum_{t=1}^T f_{it}f_{j_2,t} \otimes_{\ell=1}^K a_{i\ell} \otimes_{\ell=K+1}^{2K} a_{j_2\ell} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top},$$
$$\Delta_{1,2} = \sum_{j_1\neq i}^r \frac{1}{T}\sum_{t=1}^T f_{j_1,t}f_{it} \otimes_{\ell=1}^K a_{j_1\ell} \otimes_{\ell=K+1}^{2K} a_{i\ell} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top},$$
$$\Delta_{1,3} = \sum_{j_1\neq j_2\neq i}^r \frac{1}{T}\sum_{t=1}^T f_{j_1,t}f_{j_2,t} \otimes_{\ell=1}^K a_{j_1\ell} \otimes_{\ell=K+1}^{2K} a_{j_2\ell} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}.$$

For any vectors $\widetilde{g}_{i\ell}, \breve{g}_{i\ell} \in \mathbb{R}^{d_\ell}$, define

$$\Delta_{2,k}(\widetilde{g}_{i\ell}, \breve{g}_{i\ell}, \ell\neq k) = \frac{1}{T}\sum_{t=1}^T \mathcal{E}_t \otimes \mathcal{E}_t \times_{\ell=1,\ell\neq k}^K \widetilde{g}_{i\ell}^\top \times_{\ell=K+1,\ell\neq K+k}^{2K} \breve{g}_{i,\ell-K}^\top \in \mathbb{R}^{d_k\times d_k},$$
$$\Delta_{3,k}(\widetilde{g}_{i\ell}, \ell\neq k) = \frac{1}{T}\sum_{t=1}^T f_{it}\mathcal{E}_t \times_{\ell=1,\ell\neq k}^K \widetilde{g}_{i\ell}^\top \in \mathbb{R}^{d_k},$$
$$\Delta_{4,k}(\widetilde{g}_{i\ell}, \ell\neq k) = \frac{1}{T}\sum_{t=1}^T \mathcal{E}_t \otimes (f_{jt},\ j\neq i)^\top \times_{\ell=1,\ell\neq k}^K \widetilde{g}_{i\ell}^\top \in \mathbb{R}^{d_k\times(r-1)}.$$

As $\Delta_{q,k}(\widetilde{g}_{i\ell}, \breve{g}_{i\ell}, \ell\neq k)$ is linear in $\widetilde{g}_{i\ell}, \breve{g}_{i\ell}$, by (61), the numerator on the right hand side of (63) can

be bounded by

$$
\begin{aligned}
&\|\Psi \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2 \\
\leqslant&\|\Delta_1 \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2 + \|\Delta_{2,k}(\widehat{g}_{i\ell}^{(m)}, \widehat{g}_{i\ell}^{(m)}, \ell\neq k)\|_2 + 2\|\Delta_{3,k}(\widehat{g}_{i\ell}^{(m)}, \ell\neq k)\|_2 \\
&\quad + 2\|A_k\|_2\|\Delta_{4,k}(\widehat{g}_{i\ell}^{(m)}, \ell\neq k)\|_2 \max_{j\neq i}\prod_{\ell\neq k}^{K}\left|a_{j\ell}^\top \widehat{g}_{i\ell}^{(m)}\right| \\
\leqslant&\sum_{q=1,2,3}\|\Delta_{1,q}\|_2 + \|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell\neq k)\|_2 + (2K-2)\phi_{m,k}\|\Delta_{2,k}\|_{\mathrm{S}} \\
&\quad + 2\|\Delta_{3,k}(g_{i\ell}, \ell\neq k)\|_2 + (4K-4)\phi_{m,k}\|\Delta_{3,k}\|_{\mathrm{S}} \\
&\quad + 2\|A_k\|_2\prod_{\ell\neq k}^{K}\left(\psi_{m,\ell}/\sqrt{1-1/(4r)}\right)\|\Delta_{4,k}(g_{i\ell}, \ell\neq k)\|_2 \\
&\quad + 2\|A_k\|_2(2K-2)\phi_{m,k}\prod_{\ell\neq k}^{K}\left(\psi_{m,\ell}/\sqrt{1-1/(4r)}\right)\|\Delta_{4,k}\|_{\mathrm{S}}, \qquad (64)
\end{aligned}
$$

where

$$
\begin{aligned}
\|\Delta_{2,k}\|_{\mathrm{S}} &= \max_{\substack{\|\widetilde{g}_{i\ell}\|_2=\|\breve{g}_{i\ell}\|_2=1,\\ \widetilde{g}_{i\ell},\breve{g}_{i\ell}\in\mathbb{R}^{d_\ell}}} \|\Delta_{2,k}(\widetilde{g}_{i\ell}, \breve{g}_{i\ell}, \ell\neq k)\|_2, \\
\|\Delta_{q,k}\|_{\mathrm{S}} &= \max_{\substack{\|\widetilde{g}_{i\ell}\|_2=1,\\ \widetilde{g}_{i\ell}\in\mathbb{R}^{d_\ell}}} \|\Delta_{q,k}(\widetilde{g}_{i\ell}, \ell\neq k)\|_2, \quad q=3,4.
\end{aligned}
$$

Note that

$$
\Delta_{1,1} = a_{ik}\left(\prod_{\ell\neq k}^{K} a_{i\ell}^\top \widehat{g}_{i\ell}^{(m)}\right)\frac{1}{T}\sum_{t=1}^{T} f_{i,t}\cdot(f_{j_2,t},\ j_2\neq i)\ \mathrm{diag}\left(\prod_{\ell\neq k}^{K} a_{j_2\ell}^\top \widehat{g}_{i\ell}^{(m)}, j_2\neq i\right)(a_{j_2k}, j_2\neq i)^\top.
$$

By (62), in the event Ω_1,

$$
\|\Delta_{1,1}\|_2 \lesssim \sqrt{\lambda_1\Theta_{ii}}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right)\prod_{\ell\neq k}^{K}\psi_{m,\ell}. \qquad (65)
$$

Similarly, in the event Ω_1,

$$
\|\Delta_{1,2}\|_2 \lesssim \sqrt{\lambda_1\Theta_{ii}}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right)\prod_{\ell\neq k}^{K}\psi_{m,\ell}, \qquad (66)
$$

$$
\|\Delta_{1,3}\|_2 \lesssim \lambda_1\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right)\prod_{\ell\neq k}^{K}\psi_{m,\ell}^2. \qquad (67)
$$

Let $\Upsilon_{0,i,k} = T^{-1}\sum_{t=1}^{T} f_{it}^2 a_{ik}a_{ik}^\top$ and $\Upsilon_{0,-i,k} = T^{-1}\sum_{t=1}^{T}\sum_{j_1,j_2\neq i}^{r} f_{j_1t}f_{j_2t}a_{j_1k}a_{j_2k}^\top$. Then, in the event

Ω_1,

$$\|\Upsilon_{0,i,k}\|_2 \leqslant \Theta_{ii} + C_1\Theta_{ii}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) := \Delta_{\Upsilon_i} \asymp \Theta_{ii}, \tag{68}$$

$$\|\Upsilon_{0,-i,k}\|_2 \leqslant \lambda_1 + C_1\lambda_1\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) := \Delta_{\Upsilon_{-i}} \asymp \lambda_1. \tag{69}$$

Recall $e_t = \mathrm{vec}(\mathcal{E}_t)$. Similar to the proof of Lemma A.1, we can show, in an event Ω_2 with probability at least $1 - T^{-c_2} - d^{-c_2}$,

$$\begin{aligned}
\|\Delta_{2,k}\|_{\mathrm{S}} &\leqslant \left\|\frac{1}{T}\sum_{t=1}^{T} e_t e_t^\top\right\|_2 \lesssim \sqrt{\frac{d\log(d)}{T}} + \frac{d\log(d)}{T} + 1,\\
\|\Delta_{3,k}\|_{\mathrm{S}} &\leqslant \left\|\frac{1}{T}\sum_{t=1}^{T} f_{it} e_t^\top\right\|_2 \lesssim \sqrt{\frac{d\log(d)}{T}}\sqrt{\Theta_{ii}} + \frac{d\log(d)}{T}\sqrt{\Theta_{ii}},\\
\|\Delta_{4,k}\|_{\mathrm{S}} &\leqslant \left\|\frac{1}{T}\sum_{t=1}^{T} (f_{jt},\ j\neq i)^\top e_t^\top\right\|_2 \lesssim \sqrt{\frac{d\log(d)}{T}}\sqrt{\lambda_1} + \frac{d\log(d)}{T}\sqrt{\lambda_1}.
\end{aligned} \tag{70}$$

We claim that in certain events Ω_3, with probability at least $1 - T^{-c_3} - d^{-c_3}$, for any $1 \leqslant \ell \leqslant K$, the following bounds hold,

$$\begin{aligned}
\|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell \neq k)\|_2 &\leqslant \frac{C_1 d_k \log(d)}{T} + C_1\sqrt{\frac{d_k\log(d)}{T}} + C_1,\\
\|\Delta_{3,k}(g_{i\ell}, \ell \neq k)\|_2 &\leqslant C_1\sqrt{\frac{d_k\log(d)}{T}}\sqrt{\Theta_{ii}} + \frac{C_1 d_k\log(d)}{T}\sqrt{\Theta_{ii}},\\
\|\Delta_{4,k}(g_{i\ell}, \ell \neq k)\|_2 &\leqslant C_1\sqrt{\frac{d_k\log(d)}{T}}\sqrt{\lambda_1} + \frac{C_1 d_k\log(d)}{T}\sqrt{\lambda_1}.
\end{aligned} \tag{71}$$

Define

$$R_{k,i} = \sqrt{\frac{d_k\log(d)}{T}} + \frac{d_k\log(d)}{T} + 1 + \sqrt{\frac{\Theta_{ii}d_k\log(d)}{T}} + \frac{\sqrt{\Theta_{ii}}d_k\log(d)}{T},\ R^* = \max_i\max_k R_{k,i}/\Theta_{ii}. \tag{72}$$

As $g_{i\ell}$ is true and deterministic, it follows from (64), (65), (66), (67), (70), (71), in the event $\cap_{q=0}^3\Omega_q$, for some numeric constant $C_2 > 0$

$$\begin{aligned}
&\|\Psi \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2\\
\leqslant& C_2 R_{k,i} + C_{1,K}R^{(0)}\phi_{m,k} + C_{1,K}\sqrt{\lambda_1\Theta_{ii}}\prod_{\ell\neq k}\psi_{m,\ell}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right)\\
&+ C_{1,K}\lambda_1\prod_{\ell\neq k}\psi_{m,\ell}^2\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) + C_{1,K}\sqrt{\frac{d_k\log(d)}{T}}\sqrt{\lambda_1}\prod_{\ell\neq k}\psi_{m,\ell},
\end{aligned} \tag{73}$$

where $R^{(0)} = \phi^{(0)}$ is defined in (19) in Theorem 4.1. Substituting (73) into (63), by the definition of $\phi_{m,k}$, we have, in the event $\cap_{q=0}^{3}\Omega_q$,

$$
\begin{aligned}
&\|\widehat{a}_{ik}^{(m+1)}\widehat{a}_{ik}^{(m+1)\top} - a_{ik}a_{ik}^{\top}\|_2 \\
&\leqslant \frac{4(1+\delta_{\max})\lambda_1 \prod_{\ell\neq k}\psi_{m,\ell}^2}{\alpha^{2K-2}\Theta_{ii}[\sqrt{(1-\delta_{\max})(1-1/(4r))}\alpha]^{2K-2}} + \frac{4C_2R_{k,i}}{\alpha^{2K-2}\Theta_{ii}} + \frac{4C_{1,K}R^{(0)}\phi_{m,k}}{\alpha^{2K-2}\Theta_{ii}} \\
&\quad + \frac{4C_{1,K}\sqrt{\lambda_1}}{\alpha^{2K-2}\Theta_{ii}}\prod_{\ell\neq k}\psi_{m,\ell}\sqrt{\frac{d_k\log(d)}{T}} \\
&\quad + \frac{4C_{1,K}}{\alpha^{2K-2}}\sqrt{\lambda_1/\Theta_{ii}}\prod_{\ell\neq k}\psi_{m,\ell}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\quad + \frac{4C_{1,K}}{\alpha^{2K-2}}(\lambda_1/\Theta_{ii})\prod_{\ell\neq k}\psi_{m,\ell}^2\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\leqslant C_{\alpha,K}R_{k,i}/\Theta_{ii} + C_{\alpha,K}(\sqrt{r}\psi_0)\psi_{m,k} + C_{\alpha,K}\sqrt{\lambda_1/\lambda_r}\prod_{\ell\neq k}\psi_{m,\ell}R_{k,i}/\Theta_{ii} \\
&\quad + C_{\alpha,K}(\lambda_1/\lambda_r)\prod_{\ell\neq k}\psi_{m,\ell}^2 + C_{\alpha,K}\sqrt{\lambda_1/\lambda_r}\prod_{\ell\neq k}\psi_{m,\ell}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\leqslant C_{\alpha,K}R^* + \rho\psi_m, \qquad (74)
\end{aligned}
$$

where the last inequality comes from condition (23) with $\rho < 1$. As $R^* \lesssim 1$, we have $R^* \lesssim \psi^{\text{ideal}}$. Note that as $\lambda_r \lesssim d$ and the error bound of ψ_0 in Theorem 4.1, we have $T \gtrsim \sqrt{d} \gtrsim d_{\max}$. It follows that, after $O(\log(\psi_0/\psi^{\text{ideal}}))$ iterations,

$$\psi_{m,i,k} \lesssim \psi^{\text{ideal}}. \qquad (75)$$

In the end, we divide the rest of the proof into 3 steps to prove (71).

Step 1. We prove (71) for the $\|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell \neq k)\|_2$. Let $P_{g_{ik}} = g_{iK}^{\top} \odot \cdots \odot g_{i,k+1}^{\top} \odot I_{d_k} \odot g_{i,k-1}^{\top} \odot \cdots \odot g_{i1}^{\top} \in \mathbb{R}^{d_k \times d}$, where $\odot$ represents Kronecker product. Also let $e_{t,ik} = \mathcal{E}_t \times_{\ell\neq k}^{K} g_{i\ell}$. Then $e_{t,ik} = P_{g_{ik}}H\xi_t \in \mathbb{R}^{d_k}$.

By Assumption 4.1 and Lemma A.1 in Shu and Nan (2019), we have

$$
\begin{aligned}
&\mathbb{P}\left(\|e_{t,ik}\|_2^2 - \mathbb{E}\|e_{t,ik}\|_2^2 \geqslant x\right) = \mathbb{P}\left(\xi_t^{\top}H^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}H\xi_t - \mathbb{E}\xi_t^{\top}H^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}H\xi_t \geqslant x\right) \\
&\leqslant 4\exp\left(-C'\left(\frac{x}{\|H^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}H\|_{\mathrm{F}}}\right)^{\frac{1}{1+2/\vartheta}}\right)
\end{aligned}
$$

Note that $\mathbb{E}\|e_{t,ik}\|_2^2 = \mathbb{E}\xi_t^{\top}H^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}H\xi_t = \mathrm{tr}(HH^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}) \asymp d_k$, and $\|H^{\top}P_{g_{ik}}^{\top}P_{g_{ik}}H\|_{\mathrm{F}}^2 =$

$\|HH^\top P_{g_{ik}}^\top P_{g_{ik}}\|_{\mathrm{F}}^2 \asymp d_k$. Choosing $x \asymp d_k$, we have

$$\mathbb{P}\left(\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\right) \leqslant 4\exp\left(-C' d_k^{\frac{\vartheta}{2\vartheta+4}}\right).$$

Let $N := \|e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}}\|_2$ and $\sigma_0^2 := \|\sum_{t=1}^T \mathbb{E}(e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}})^2\|_2$. Then, by Assumption 4.1, $N \leqslant C^2 d_k$ and $\sigma_0^2 \leqslant C_0 T d_k$. By matrix Bernstein inequality (see, e.g., Theorem 5.4.1 of Vershynin (2018)),

$$\mathbb{P}\left(\left\|\sum_{t=1}^T \left[e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} - \mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}}\right]\right\|_2 \geqslant x\right) \leqslant 2d_k \exp\left(-\frac{x^2/2}{\sigma_0^2 + Nx/3}\right).$$

Choosing $x \asymp \sqrt{Td_k\log(d)} + d_k\log(d)$, with probability at least $1 - d^{-c_1}$,

$$\left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} - \mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}}\right\|_2 \leqslant C_1\sqrt{\frac{d_k\log(d)}{T}} + C_1 \cdot \frac{d_k\log(d)}{T} \quad (76)$$

Define $M := \{1 \leqslant t \leqslant T : \|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}$. Since $\mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}}$ are independent Bernoulli random variable and $\log(T) \leqslant d_k^{\vartheta/(2\vartheta+4)}$, we have

$$\mathbb{E}|M| = T\mathbb{P}\left(\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\right) \leqslant 4T\exp\left(-C' d_k^{\frac{\vartheta}{2\vartheta+4}}\right) \leqslant T^{-c_2}.$$

By Chernoff bound for Bernoulli random variables,

$$\mathbb{P}(|M| \geqslant C) \leqslant \exp\left(-T^{c_2}\right).$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}}\right\|_2 \geqslant x\right) &\leqslant \mathbb{P}\left(|M| \max_t \|e_{t,ik}\|_2^2 \geqslant x\right) \\
&\leqslant \mathbb{P}(|M| \geqslant C) + \mathbb{P}\left(|M| < C, |M| \max_t \|e_{t,ik}\|_2^2 \geqslant x\right) \\
&\leqslant \exp\left(-T^{c_2}\right) + \mathbb{P}\left(\max_t \|e_{t,ik}\|_2^2 \geqslant x/C\right).
\end{aligned}$$

Choosing $x \asymp d_k$, we have, with probability at least $1 - \exp(-T^{c_2}) - T^{-c_2}$,

$$\left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}}\right\|_2 \leqslant C_2 \cdot \frac{d_k}{T}. \quad (77)$$

Similarly,

$$\mathbb{P}\left(\left\|\mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\geqslant C\sqrt{d_k}\}}\right\|_2 > 0\right) = \mathbb{P}\left(\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\right) \leqslant 4\exp\left(-C' d_k^{\frac{\vartheta}{2\vartheta+4}}\right) \leqslant T^{-c_3}. \quad (78)$$

Combing (76), (77), (78), in an event with probability at least $1 - T^{-c_4} - d^{-c_1}$,

$$\begin{aligned}
\|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell \neq k)\|_2 &\leqslant \left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top - \mathbb{E}e_{t,ik}e_{t,ik}^\top\right\|_2 + \|\mathbb{E}e_{t,ik}e_{t,ik}^\top\|_2 \\
&\leqslant \left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}} - \mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right\|_2 + \|P_{g_{ik}}HH^\top P_{g_{ik}}^\top\|_2 \\
&\quad + \left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\geqslant C\sqrt{d_k}\}}\right\|_2 + \left\|\mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\geqslant C\sqrt{d_k}\}}\right\|_2 \\
&\leqslant C_2\sqrt{\frac{d_k\log(d)}{T}} + C_2\cdot\frac{d_k\log(d)}{T} + C_2
\end{aligned}$$

Step 2. Now we prove (71) for $\|\Delta_{3,k}(g_{i\ell}, \ell \neq k)\|_2$. Let

$$\begin{aligned}
N_1 &:= \left\|f_{it}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right\|_2, \\
\sigma_1^2 &:= \max\left\{\left\|\sum_{t=1}^T \overline{\mathbb{E}}f_{it}^2 e_{t,ik}^\top e_{t,ik}\mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right\|_2, \left\|\sum_{t=1}^T \overline{\mathbb{E}}f_{it}^2 e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right\|_2\right\}.
\end{aligned}$$

It is easy to show

$$\begin{aligned}
N_1 &\leqslant C\sqrt{d_k}\,\|f_{it}a_i\|_2, \\
\sigma_1^2 &\leqslant C_3 T d_k \max\left\{\left\|\frac{1}{T}\sum_{t=1}^T f_{it}^2\right\|_2, \left\|\frac{1}{Td_k}\sum_{t=1}^T f_{it}^2\right\|_2\right\} := \sigma_2^2.
\end{aligned}$$

By matrix Bernstein inequality,

$$\overline{\mathbb{P}}\left(\left\|\sum_{t=1}^T\left[f_{i,t}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}} - \overline{\mathbb{E}}f_{i,t}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right]\right\|_2 \geqslant x\right) \leqslant 2d_k\exp\left(-\frac{x^2/2}{\sigma_1^2 + N_1 x/3}\right).$$

Choosing $x \asymp \sqrt{d_k}\log(d)\|f_{it}\|_2 + \sqrt{\log(d)}\sigma_2$, with probability at least $1 - d^{-c_4}$,

$$\left\|\frac{1}{T}\sum_{t=1}^T f_{i,t}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}} - \overline{\mathbb{E}}f_{i,t}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2\leqslant C\sqrt{d_k}\}}\right\|_2 \leqslant C_4\frac{\sqrt{\log(d)}\sigma_2 + \sqrt{d_k}\log(d)\|f_{it}a_i\|_2}{T}.$$

As $\sqrt{r}\log(T)^{1/\gamma_1} \lesssim \sqrt{d_k}$, by Assumption 4.2, with probability at least $1 - T^{-c_5}$,

$$\|f_{it}\|_2 \lesssim \sqrt{r}(\log(T))^{1/\gamma_1}\sqrt{\Theta_{ii}} \lesssim \sqrt{d_k\Theta_{ii}}.$$

Similarly, in the event Ω_1, $\sigma_2^2 \lesssim T d_k \Theta_{ii}$. Then, with probability at least $1 - T^{-c_1}/2 - T^{-c_5} - d^{-c_4}$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} - \mathbb{E} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} \right\|_2 \leqslant C_5 \sqrt{\frac{d_k \log(d)}{T}} \sqrt{\Theta_{ii}} + C_5 \cdot \frac{d_k \log(d) \sqrt{\Theta_{ii}}}{T}.$$

Similar to (77),

$$\mathbb{P}\left(\left\| \sum_{t=1}^{T} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 \geqslant x \right) \leqslant \mathbb{P}(|M| > C) + \mathbb{P}\left(|M| < C, |M| \max_t \|f_{it}\|_2 \cdot \|e_{t,ik}\|_2 \geqslant x \right).$$

Choosing $x \asymp d_k \sqrt{\Theta_{ii}}$, we have with probability at least $1 - \exp(-T^{c_2}) - T^{-c_2} - T^{-c_5}$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 \leqslant C_6 \cdot \frac{d_k \sqrt{\Theta_{ii}}}{T}.$$

Thus, in an event with probability $1 - T^{-c_1}/2 - T^{-c_6} - d^{-c_4}$,

$$\begin{aligned} \|\Delta_3(g_{i\ell}, \ell \neq k)\|_2 \leqslant & \left\| \frac{1}{T} \sum_{t=1}^{T} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} - \mathbb{E} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} \right\|_2 \\ & + \left\| \frac{1}{T} \sum_{t=1}^{T} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 + \left\| \mathbb{E} f_{i,t} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 \\ \leqslant & C_7 \sqrt{\frac{d_k \log(d)}{T}} \sqrt{\Theta_{ii}} + C_7 \cdot \frac{d_k \log(d) \sqrt{\Theta_{ii}}}{T}. \end{aligned}$$

Step 3. Inequality (71) for $\|\Delta_{4,k}(g_{i\ell}, \ell \neq k)\|_2$ follow from the same argument as the above step.

□

Proof of Theorem 4.3. By the definition of the iterative algorithm, after convergence to a sta-

tionary point, $\widehat{a}_{ik}$ is the top eigenvector of the matrix

$$
\begin{aligned}
\widehat{\Sigma}_{ik} =& \widehat{\Sigma} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
=& \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \frac{1}{T}\sum_{t=1}^{T} f_{it} \left(\prod_{\ell\neq k}^{K} a_{i\ell}^{\top} \widehat{g}_{i\ell}^{(m)\top} \right) a_{ik} \otimes \left(\mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} \widehat{g}_{i\ell}^{\top} \right) \\
&+ \frac{1}{T}\sum_{t=1}^{T} f_{it} \left(\prod_{\ell\neq k}^{K} a_{i\ell}^{\top} \widehat{g}_{i\ell}^{(m)\top} \right) \left(\mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} \widehat{g}_{i\ell}^{\top} \right) \otimes a_{ik} + \frac{1}{T}\sum_{t=1}^{T} \mathcal{E}_t \otimes \mathcal{E}_t \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
&+ \sum_{j\neq i}^{r} \widetilde{\lambda}_{j,i} a_{jk} a_{jk}^{\top} + \sum_{j_1\neq j_2}^{r} \frac{1}{T}\sum_{t=1}^{T} f_{j_1 t} f_{j_2 t} \otimes_{k=1}^{K} a_{j_1 k} \otimes_{k=K+1}^{2K} a_{j_2 k} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq i}^{r} f_{jt} \otimes_{k=1}^{K} a_{jk} \otimes \mathcal{E}_t \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} + \frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq i}^{r} f_{jt} \mathcal{E}_t \otimes_{k=K+1}^{2K} a_{jk} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
:=& \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \Psi_1 + \Psi_2 + \Psi_3 + \Psi_4 + \Psi_5 + \Psi_6 + \Psi_7 \\
:=& \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \Psi,
\end{aligned} \tag{79}
$$

where $\widetilde{\lambda}_{j,i}$ is defined in (58), $\widehat{g}_{i\ell} = \widehat{b}_{i\ell}/\|\widehat{b}_{i\ell}\|_2$, and $\Psi = \sum_{j=1}^{7} \Psi_j$.

Let $P_{a_{ik},\perp} = I_{d_k} - a_{ik}a_{ik}^{\top} = a_{ik,\perp}a_{ik,\perp}^{\top}$. By Theorem 4.2, the final estimates of $\widehat{a}_{ik}$ satisfies, in an event Ω with probability at least $1 - T^{-C} - d^{-C}$,

$$
\|\widehat{a}_{ik}\widehat{a}_{ik}^{\top} - a_{ik}a_{ik}^{\top}\|_2 \leqslant C_0 \psi^{\text{ideal}}, \tag{80}
$$

where ψ^{ideal} is defined in (22). Using resolvent based series expansion of projection matrices (e.g., Theorem 1 in Xia (2021)), we have the following expansion,

$$
\begin{aligned}
\widehat{a}_{ik}\widehat{a}_{ik}^{\top} - a_{ik}a_{ik}^{\top} =& \frac{1}{\widetilde{\lambda}_{i,i}} P_{a_{ik},\perp}\Psi P_{a_{ik}} + \frac{1}{\widetilde{\lambda}_{i,i}} P_{a_{ik}}\Psi P_{a_{ik},\perp} \\
&+ \frac{1}{\widetilde{\lambda}_{i,i}^2}\left(P_{a_{ik}}\Psi P_{a_{ik},\perp}\Psi P_{a_{ik},\perp} + P_{a_{ik},\perp}\Psi P_{a_{ik}}\Psi P_{a_{ik},\perp} + P_{a_{ik},\perp}\Psi P_{a_{ik},\perp}\Psi P_{a_{ik}} \right) \\
&- \frac{1}{\widetilde{\lambda}_{i,i}^2}\left(P_{a_{ik},\perp}\Psi P_{a_{ik}}\Psi P_{a_{ik}} + P_{a_{ik}}\Psi P_{a_{ik},\perp}\Psi P_{a_{ik}} + P_{a_{ik}}\Psi P_{a_{ik}}\Psi P_{a_{ik},\perp} \right) \\
&+ \mathcal{R}_3(\Psi).
\end{aligned} \tag{81}
$$

Moreover, $\|\mathcal{R}_3(\Psi)\|_2 \leqslant C_1\|\Psi\|_2^3/\widetilde{\lambda}_{i,i}^3 \leqslant C_2(\psi^{\text{ideal}})^3$ under the event Ω.

Case (i). Let $u = a_{ik}$. Then

$$
\begin{aligned}
u^{\top}\left(\widehat{a}_{ik}\widehat{a}_{ik}^{\top} - a_{ik}a_{ik}^{\top}\right)u = \left(\widehat{a}_{ik}^{\top}a_{ik}\right)^2 - 1 =& -\frac{1}{\widetilde{\lambda}_{i,i}^2} a_{ik}^{\top}\Psi P_{a_{ik},\perp}\Psi a_{ik} + a_{ik}^{\top}\mathcal{R}_3(\Psi)a_{ik} \\
=& -\frac{1}{\widetilde{\lambda}_{i,i}^2}\left(a_{ik,\perp}^{\top}\Psi a_{ik}\right)^{\top}\left(a_{ik,\perp}^{\top}\Psi a_{ik}\right) + a_{ik}^{\top}\mathcal{R}_3(\Psi)a_{ik}.
\end{aligned}
$$

By Theorem 4.2 and (80), in the event Ω, $\|\widetilde{\lambda}_{i,i}^{-1} a_{ik,\perp}^\top \Psi a_{ik}\|_2 \leqslant C_0 \psi^{\text{ideal}}$. It follows that, in the event Ω,

$$\left(\widehat{a}_{ik}^\top a_{ik}\right)^2 - 1 \leqslant C_3 (\psi^{\text{ideal}})^2. \tag{82}$$

From the condition (23) and (74), we have $\psi^{\text{ideal}} \lesssim \psi_0^2$, where ψ_0 is the error bound for the initialization. By the proofs of Theorem 4.2, i.e. the derivation of (74), we can show, in the event Ω,

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \left(\Psi_4 + \Psi_5 + \Psi_6 + \Psi_7\right) \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2, \tag{83}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_3 \right\|_2 \leqslant C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}} + C_4 \frac{1}{\Theta_{ii}} \left(\frac{d_k \log(d)}{T} + \sqrt{\frac{d_k \log(d)}{T}} + 1 \right), \tag{84}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_1 - \frac{1}{\Theta_{ii} \prod_{\ell \neq k} (a_{i\ell}^\top g_{il}) T} \sum_{t=1}^T f_{it} a_{ik} \otimes \left(\mathcal{E}_t \times_{\ell \in [K] \setminus \{k\}} g_{i\ell}^\top \right) \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2 + C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}}, \tag{85}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_2 - \frac{1}{\Theta_{ii} \prod_{\ell \neq k} (a_{i\ell}^\top g_{il}) T} \sum_{t=1}^T f_{it} \left(\mathcal{E}_t \times_{\ell \in [K] \setminus \{k\}} g_{i\ell}^\top \right) \otimes a_{ik} \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2 + C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}}. \tag{86}$$

As $a_{ik,\perp}^\top \Psi_1 a_{ik} = 0$, in the event Ω,

$$\begin{aligned} &\left\| \frac{1}{\widetilde{\lambda}_{i,i}} a_{ik,\perp}^\top \Psi a_{ik} - \frac{1}{\Theta_{ii} \prod_{\ell \neq k} (a_{i\ell}^\top g_{il})} a_{ik,\perp}^\top \left[\frac{1}{T} \sum_{t=1}^T f_{it} \left(\mathcal{E}_t \times_{\ell \in [K] \setminus \{k\}} g_{i\ell}^\top \right) \otimes a_{ik} \right] a_{ik} \right\|_2 \\ \leqslant & C_4 (\psi^{\text{ideal}})^2 + C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}} + C_4 \frac{1}{\Theta_{ii}} \left(\frac{d_k \log(d)}{T} + \sqrt{\frac{d_k \log(d)}{T}} + 1 \right). \end{aligned} \tag{87}$$

Case (ii). Let $u \perp a_{ik}$. Define $v = u P_{a_{ik},\perp}$. Then

$$\begin{aligned} & u^\top \left(\widehat{a}_{ik} \widehat{a}_{ik}^\top - a_{ik} a_{ik}^\top \right) a_{ik} = (u^\top \widehat{a}_{ik})(\widehat{a}_{ik}^\top a_{ik}) \\ = & \frac{1}{\widetilde{\lambda}_{i,i}} u^\top P_{a_{ik},\perp} \Psi a_{ik} + \frac{1}{\widetilde{\lambda}_{i,i}^2} \left(u^\top P_{a_{ik},\perp} \Psi P_{a_{ik},\perp} \Psi a_{ik} - u^\top P_{a_{ik},\perp} \Psi P_{a_{ik}} \Psi a_{ik} \right) + \mathcal{R}_3(\Psi) \\ = & \frac{1}{\widetilde{\lambda}_{i,i}} v^\top \Psi a_{ik} + \frac{1}{\widetilde{\lambda}_{i,i}^2} \left(v^\top \Psi a_{ik,\perp} a_{ik,\perp}^\top \Psi a_{ik} - v^\top \Psi a_{ik} a_{ik}^\top \Psi a_{ik} \right) + \mathcal{R}_3(\Psi). \end{aligned}$$

Note that, in the event Ω, $\|\widetilde{\lambda}_{i,i}^{-1}\Psi\|_2 \leqslant C_0\psi^{\text{ideal}}$. By (83), (84), (85), (86), we have,

$$\begin{aligned}
&\sup_{u\perp a_{ik}}\left|u^\top\left(\widehat{a}_{ik}\widehat{a}_{ik}^\top - a_{ik}a_{ik}^\top\right)a_{ik} - \frac{1}{\Theta_{ii}\prod_{\ell\neq k}(a_{i\ell}^\top g_{il})}u^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t\times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\otimes a_{ik}\right]a_{ik}\right| \\
&\leqslant C_4(\psi^{\text{ideal}})^2 + C_4(\sqrt{r}\psi_0)\psi^{\text{ideal}} + C_4\frac{1}{\Theta_{ii}}\left(\frac{d_k\log(d)}{T} + \sqrt{\frac{d_k\log(d)}{T}} + 1\right). \qquad (88)
\end{aligned}$$

Now, let's move to the proof of Theorem 4.3. Without loss of generality, assume $\widehat{a}_{ik}^\top a_{ik} > 0$. For u such that $\liminf_{d_k\to\infty}\|P_{a_{ik},\perp}u\|_2 > 0$, we have

$$u^\top(\widehat{a}_{ik} - a_{ik}) = u^\top P_{a_{ik},\perp}\widehat{a}_{ik} + (u^\top a_{ik})(a_{ik}^\top\widehat{a}_{ik} - 1).$$

By (82),

$$\left|(u^\top a_{ik})(a_{ik}^\top\widehat{a}_{ik} - 1)\right| \leqslant C\|u^\top a_{ik}\|_2(\psi^{\text{ideal}})^2.$$

In addition, by (88) and (80),

$$\begin{aligned}
&\sup_{u}\left|u^\top P_{a_{ik},\perp}\widehat{a}_{ik} - \frac{1}{\Theta_{ii}\prod_{\ell\neq k}(a_{i\ell}^\top g_{il})}u^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t\times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\otimes a_{ik}\right]a_{ik}\right| \\
\leqslant& C\|P_{a_{ik},\perp}u\|_2\left[(\psi^{\text{ideal}})^2 + (\sqrt{r}\psi_0)\psi^{\text{ideal}} + \frac{1}{\Theta_{ii}}\left(\frac{d_k\log(d)}{T} + \sqrt{\frac{d_k\log(d)}{T}} + 1\right)\right].
\end{aligned}$$

Combing the bound above, we have

$$\begin{aligned}
&\sup_{u}\left|u^\top(\widehat{a}_{ik} - a_{ik}) - \frac{1}{\Theta_{ii}\prod_{\ell\neq k}(a_{i\ell}^\top g_{il})}u^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t\times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\right]\right| \\
\leqslant& C\|P_{a_{ik},\perp}u\|_2\left[(\psi^{\text{ideal}})^2 + (\sqrt{r}\psi_0)\psi^{\text{ideal}} + \frac{1}{\Theta_{ii}}\left(\frac{d_k\log(d)}{T} + \sqrt{\frac{d_k\log(d)}{T}} + 1\right)\right] + C\|u^\top a_{ik}\|_2(\psi^{\text{ideal}})^2.
\end{aligned}$$

By (71) and $\widehat{a}_{ik}^\top a_{ik} > 0$, if $\sqrt{d_k/T} \gg 1/\sqrt{\Theta_{ii}}$, i.e. $\Theta_{ii} \gg T/d_k$, then the leading term in $u^\top(\widehat{a}_{ik} - a_{ik})$ is

$$\frac{1}{\Theta_{ii}\prod_{\ell\neq k}(a_{i\ell}^\top g_{il})}u^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t\times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\right] = \frac{1}{\Theta_{ii}}u^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^{T} f_{it}\left(\mathcal{E}_t\times_{\ell\in[K]\setminus\{k\}} b_{i\ell}^\top\right)\right].$$

Thus, (26) follows from the central limit theorem of the above leading term.

Otherwise, $1/\Theta_{ii}$ is the leading order term of $u^\top(\widehat{a}_{ik} - a_{ik})$. Then we have (27).

□

Proof of Theorem 4.4. First, by (82), we have (28). Moreover, by (87) and case (i) in the proof

of Theorem 4.3, we have

$$\left|\left(\widehat{a}_{ik}^\top a_{ik}\right)^2 - 1 - \frac{1}{\Theta_{ii}^2 \prod_{\ell\neq k}(a_{i\ell}^\top g_{il})^2}\left[\frac{1}{T}\sum_{t=1}^T f_{it}\left(\mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\right]^\top P_{a_{ik},\perp}\left[\frac{1}{T}\sum_{t=1}^T f_{it}\left(\mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} g_{i\ell}^\top\right)\right]\right|$$
$$\leqslant C_4(\psi^{\text{ideal}})^3 + C_4(\sqrt{r}\psi_0)(\psi^{\text{ideal}})^2 + C_4\frac{1}{\Theta_{ii}^2}\left(\frac{d_k\log(d)}{T} + \sqrt{\frac{d_k\log(d)}{T}} + 1\right)^2.$$

Then (29) and (30) can be derived by applying similar arguments in the proof of Theorem 4.3.

□

Proof of Theorem 4.5. The proof of the consistency of $\widehat{r}^{\text{uer}}$ draws on methods similar to those in Ahn and Horenstein (2013) and Han et al. (2022b), given that our CP tensor factor model can be equated to a vector factor model. Moreover, the consistency of $\widehat{r}^{\text{ip}}$ aligns with the proofs in Han et al. (2022b), as our CP tensor factor model can also be regarded as a Tucker factor model with a uniform Tucker rank of $(r, ..., r)$. Specifically, lemmas akin to Lemmas 11 and 12 (or Lemmas 14 and 15) in Han et al. (2022b) can be derived under our assumptions. It leads to $\mathbb{P}(\widehat{r}_k = r, 1 \leqslant k \leqslant K) \to 1$. We omit the detailed proofs as they are laborious, albeit straightforward, adaptations for a specialized case of the Tucker factor model. □

Appendix B Techinical Lemmas

Lemma B.1. *Let $A \in \mathbb{R}^{d_1\times r}$ and $B \in \mathbb{R}^{d_2\times r}$ with $\|A^\top A - I_r\|_2 \vee \|B^\top B - I_r\|_2 \leqslant \delta$ and $d_1 \wedge d_2 \geqslant r$. Let $A = \widetilde{U}_1\widetilde{D}_1\widetilde{U}_2^\top$ be the SVD of A, $U = \widetilde{U}_1\widetilde{U}_2^\top$, $B = \widetilde{V}_1\widetilde{D}_2\widetilde{V}_2^\top$ the SVD of B, and $V = \widetilde{V}_1\widetilde{V}_2^\top$. Then, $\|A\Lambda A^\top - U\Lambda U^\top\|_2 \leqslant \delta\|\Lambda\|_2$ for all nonnegative-definite matrices Λ in $\mathbb{R}^{r\times r}$, and $\|AQB^\top - UQV^\top\|_2 \leqslant \sqrt{2}\delta\|Q\|_2$ for all $r \times r$ matrices Q.*

Lemma B.2. *Let $M \in \mathbb{R}^{d_1\times d_2}$ be a matrix with $\|M\|_{\text{F}} = 1$ and a and b be unit vectors respectively in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$. Let $\widehat{a}$ be the top left singular vector of M. Then,*

$$\left(\|\widehat{a}\widehat{a}^\top - aa^\top\|_2^2\right) \wedge (1/2) \leqslant \|vec(M)vec(M)^\top - vec(ab^\top)vec(ab^\top)^\top\|_2^2. \tag{89}$$

Lemmas B.1 and B.2 are Propositions 5 and 3 in Han and Zhang (2023), respectively.

Appendix C More Simulation Results

In this section we show the simulation results of Configuration I, II and III with AR coefficient on g_{it}, ϕ, equal to 0.5. We can see that AC-ISO algorithm has a better performance since the signal strength in the auto-covariance grow with ϕ. CC-ISO algorithm, however, outperforms AC-ISO algorithm even with stronger serial correlation in the factor process.

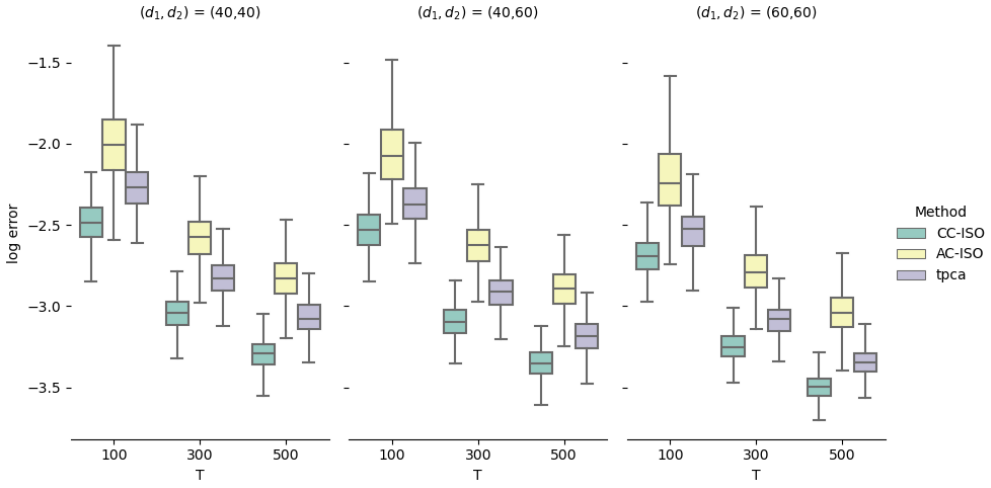


Figure 11: Boxplots of estimation errors over 500 replications under Configuration I with $\phi = 0.5$

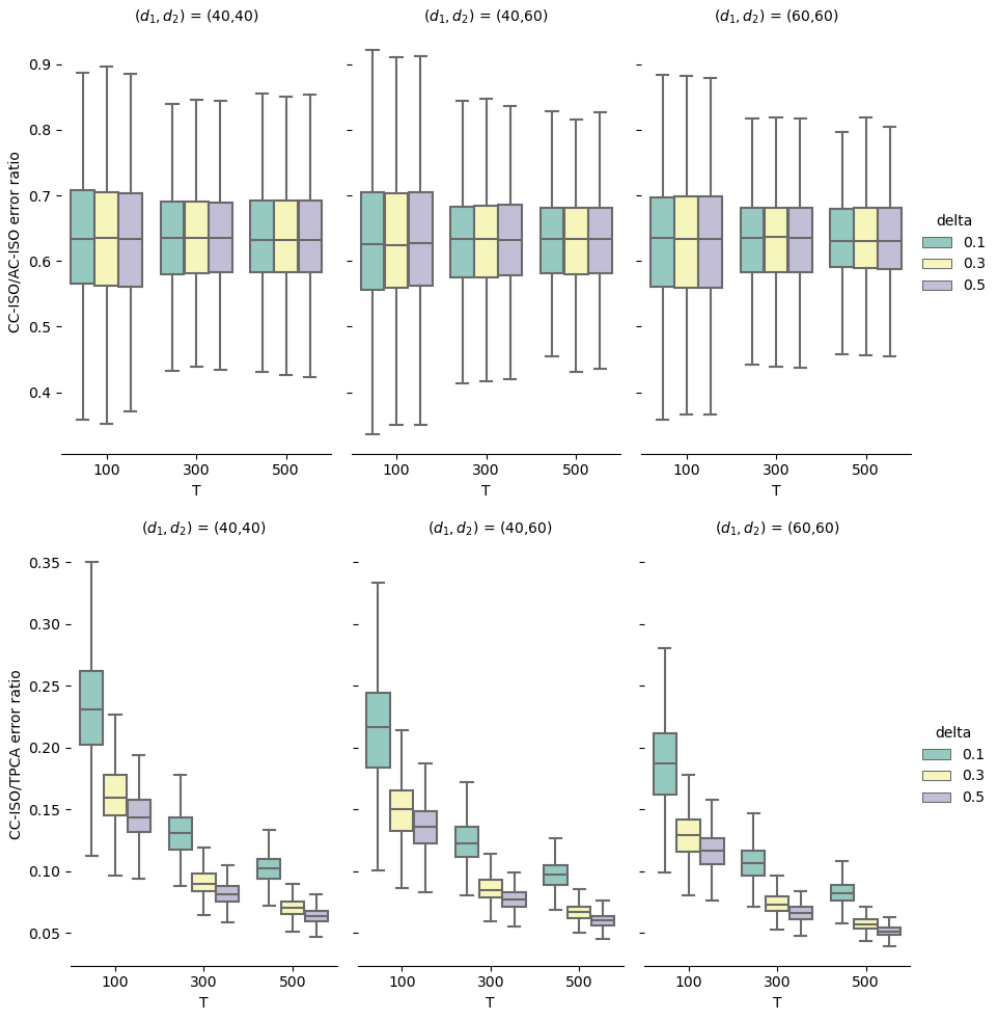


Figure 12: Boxplots of the estimation error over 500 replications under configuration II with $\phi = 0.5$. Note: The first panel shows the ratio of the estimation error of CC-ISO on AC-ISO. The second panel shows the ratio of the estimation error of CC-ISO on TPCA.

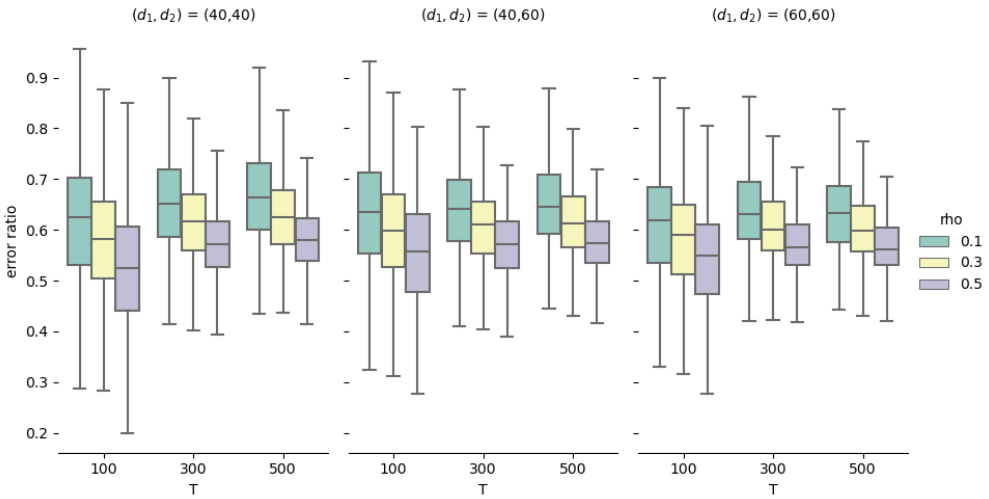


Figure 13: Boxplots of estimation errors over 500 replications under Configuration III with $\phi = 0.5$