This project consists of two papers:

Paper 1: Estimation and Inference for CP Tensor Factor Model

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 tensor dimension diverges. Our focus is on a factor model that admits CP-type tensor decomposition, which allows for non-orthogonal loading vectors. We propose an iterative simultaneous projection estimation method based on the contemporary covariance matrix and establish an inferential theory. Applying this method to sorted portfolio data reveals three key factors: a market factor, a long-short factor, and a volatility factor.

Paper 2: Diffusion Index Forecast with Tensor Data

This paper studies diffusion index forecasts with both tensor and non-tensor predictors, where the tensor structure is maintained using a CP tensor factor model. When the number of non-tensor predictors is small, we establish the asymptotic properties of the least squares estimator in this tensor factor-augmented regression framework, accommodating factors of varying strengths. For settings where the number of non-tensor predictors is large or diverges with the sample size, we adopt a penalized regression approach with the LASSO penalty and demonstrate the consistency of the resulting estimator. Simulation studies validate our theoretical results, and an empirical application to US trade flows illustrates the advantages of our method over existing approaches.

# Estimation and Inference for CP Tensor Factor Models*

Bin Chen[†1], Yuefeng Han [‡2], and Qiyang Yu[§3]

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

March 20, 2025

**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. Simulation studies confirm the theoretical results and an empirical application to sorted portfolios reveals three important factors: a market factor, a long-short factor, and a volatility factor.

*We thank Yoosoon Chang, Rong Chen, Joon Park, Mahrad Sharifvaghefi and seminar participants at Indiana University and University of Pittsburgh, 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 emerged as 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)). Classical factor models handle a large panel data model, where both cross-sectional and time series dimensions are allowed to increase. Factor models admit a low-rank structure and have a common-idiosyncratic decomposition. These features enable a formal way of discerning significant variations within the panel of economic data.

In contemporary economics, researchers increasingly encounter vast, multi-dimensional datasets, or tensor. For instance, consider monthly import-export volume time series spanning various product categories (e.g., chemical, food, machinery, etc.) among countries, constituting a dynamic sequence represented as a three-dimensional tensor, with the diagonal elements for each product category unavailable. Similarly, in portfolio selection, data often involve firms, stock prices, and a multitude of firm characteristics across time, forming a two-dimensional tensor. Additionally, in macroeconomic studies on growth and productivity, researchers analyze multiple macro variables at the country-industry level, facilitating 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 works have explored various estimation approaches and extensions. For example, working with Tucker decomposition, Chen et al. (2022) consider two estimators based on the autocovariance matrices. Han et al. (2022a) extend them with the iterative procedure using the matrix unfolding mechanism. Chen and Fan (2023) propose an estimation method called $\alpha$-PCA that preserves the matrix structure and aggregates mean and contemporary covariance through a hyper-parameter $\alpha$. Chen and Lam (2024) introduce 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) introduce a semiparametric tensor factor model leveraging mode-wise covariates. In the context of CP decomposition, Han et al. (2023) propose an iterative simultaneous orthogonalization algorithm with warm-start initialization, while Babii et al. (2023) employ tensor principal component analysis (TPCA), assuming orthogonal factor loadings. Chang et al. (2023) develop 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 its 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 a generalized eigenvalue ratio-based estimator for the number of the latent factors. Section 4 establishes consistency and the 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 $U \in \mathbb{R}^{m_k \times r_k}$ is an order $K$-tensor of size $d_1 \times \cdots \times d_{k-1} \times m_k \times r_{k+1} \times \cdots \times r_K$, 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 $\text{mat}_k(\mathcal{A}) \in \mathbb{R}^{d_k \times (d/d_k)}$, where $d = \prod_{j=1}^{K} d_j$ and $d/d_k = \prod_{j=1, j\neq k}^{K} d_j$. 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 $\text{vec}(\mathcal{A}) \in \mathbb{R}^m$. Note that $\text{mat}_k(\text{vec}(\mathcal{A})) = \text{mat}_k(\mathcal{A})$.

## 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 lies 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 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 sorted portfolios, detailed in Section 6. The observed $\mathcal{Y}_t$ is represented as a matrix, where $d_1$ is the number of characteristic-sorted portfolios, $d_2 = 10$, and $K = 2$. Each entry $\mathcal{Y}_{t,jl}$ corresponds to the excess return of the $l^{\text{th}}$-decile of the $j^{\text{th}}$ characteristic at time $t = 1, \ldots, T$. Figure 1 shows a time series plot of $\mathcal{Y}_t$ for ten characteristics from January 1990 to December 2022. Model (1) identifies $r$ latent factors, which can be interpreted as systematic risk factors. The element $a_{i1,j}$ of the loading vector

$a_{i1}$, where $i = 1, \cdots, r$ and $j = 1, \cdots, d_1$, captures the heterogeneous exposure of the $j^{\text{th}}$ characteristic to the $i^{\text{th}}$ risk factor. Similarly, the entry $a_{i2,l}$ of the loading vector $a_{i2}$, where $i = 1, \cdots, r$ and $l = 1, \cdots, 10$, determines the exposure of the $l^{\text{th}}$ decile to the $i^{\text{th}}$ risk factor. We allow the number of risk factors $r$ to increase with the dimensions $d_1$, $d_2$ and the sample size $T$.

Figure 1: Time series plots of characteristic decile portfolio returns




Note: sample period from January 1990 to December 2022.

In the literature, an alternative tensor factor model based on Tucker decomposition has been 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 $\Sigma$ 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 integrate randomized projection into our composite PCA approach (Procedure 2). Random projection or random slicing (Anandkumar et al., 2014b; Sun et al., 2017) is a well-recognized initialization method in tensor CP decomposition, allowing for repeated eigenvalues.

**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 $\mathcal{Y}_t$:

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

where

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

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

**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 $\nu$.

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

  Randomly draw a $d_1 \times d_1$ Gaussian matrix $\theta$ 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 $\widehat{a}_{ik}^{\mathrm{rcpca}}, 1 \leqslant i \leqslant s, 1 \leqslant k \leqslant K$

$B_k = A_k(A_k^\top A_k)^{-1} = (b_{1k}, ..., b_{rk}) \in \mathbb{R}^{d_k \times r}$, $A_k = (a_{1k}, \ldots, a_{rk}) \in \mathbb{R}^{d_k \times r}$, 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 $k$th 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 $\widetilde{\Sigma}$. 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

**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

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

factors $r$. Thus, the eigenvalue ratio-based estimator based on $\widetilde{\Sigma}$ 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 based on 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

**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 $\widehat{r}^{\mathrm{ip}}$.

consider a CP tensor factor model with orthogonal loading vectors. Unlike Tucker factor models, the identification 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 $\xi_i$ 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 $\alpha$-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 $t$. While incorporating weak temporal correlation among the noise, as suggested by Bai and Ng (2002), is plausible, it substantially complicates our theoretical analysis, thus deferring 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 $\lambda_i$ 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 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* $\delta$ *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 satisfies. 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 $\lambda_1/\lambda_r$ for part (ii), is detailed in the appendix. In practical scenarios, 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 situations where sample spiked eigenvalues are nearly identical. In our simulation study, we will demonstrate that the original composite PCA still yields viable initializations, when $\lambda_1 = \lambda_r$, although its performance is not as good as that of randomized composite PCA utilizing 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 $\psi_0$ (for example, the right hand side of (18)), and also let

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

**Theorem 4.2.** *Suppose Assumptions 4.1, 4.2, 4.3 hold. Assume that* $\delta_{\max}=\max_{k\leqslant K}\delta_k<1$ *with* $\delta_k$ *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*

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

*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 $\psi_0$ 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 quicker than that found in the vector factor model.*

We now demonstrate the feasibility of obtaining a more precise bound by a detailed

examination of the leading order term. This process allows us to ascertain the asymptotic behavior of the estimator $a_{ik}$. More concretely, 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 $\Theta$ 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 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*

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

*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*

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

In Theorem 4.3, our focus is limited to 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, in scenarios where $\sin \angle(u, a_{ik}) = 0$, the convergence rate of the estimated linear form is quicker, and its long-term distribution is identifiable as a blend of $\chi_1^2$ 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 $1 - T^{-C} - d^{-C}$,*

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

*where $\psi^{ideal}$ is defined in (22).*

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

$$T\left(1-(\widehat{a}_{ik}^{\mathrm{iso}\top}a_{ik})^2\right) \xrightarrow{d} \sum_{j=1}^{d_k}\varpi_j\chi_1^2, \tag{29}$$

*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* $\Sigma_{ik} = \Theta_{ii}^{-2}\Theta_{ii}\mathbb{E}[(\mathcal{E}_t \times_{\ell\neq k}^{K} b_{i\ell})(\mathcal{E}_t \times_{\ell\neq k}^{K} b_{i\ell})^\top]$.

*(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, where a debiasing process is often necessary to achieve asymptotic normality in linear combinations of the principal components, as discussed in works by Koltchinskii and Lounici (2016, 2017); Koltchinskii et al. (2020), proves insightful. 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 $\widehat{r}^{\mathrm{uer}}$ and $\widehat{r}^{\mathrm{ip}}$ 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. 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 TPCA by Babii et al. (2023). The auto-covariance considered by Han et al. (2023) is defined by the following lagged-cross product operator:

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

with $h = 1$. 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 $\delta$ to change the angles between loading vectors. It is evident that an increase in $\delta$ 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

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

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.

The following three configurations are adapted from Babii et al. (2023) and Han et al. (2022a) 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 $r = 3$.

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

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 $\delta$ 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 $\sigma_{e,ij} = 0.5^{|i-j|}$;
- $\text{vec}(Z_t) = \Phi \text{vec}(Z_{t-1}) + U_t$ 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 $\rho$.

We vary $\rho$ 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 $\alpha$ varies in the set $\{2.5, 3, 3.5, 4\}$. Note that when $\alpha = 2$, the factor structure is considered strong. A larger $\alpha$ 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 algorithms, 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.

In the experiment of Configuration II, we assess the impact of non-orthogonal factor loadings on estimations using ISO algorithms and TPCA algorithm. Figure 3 shows the

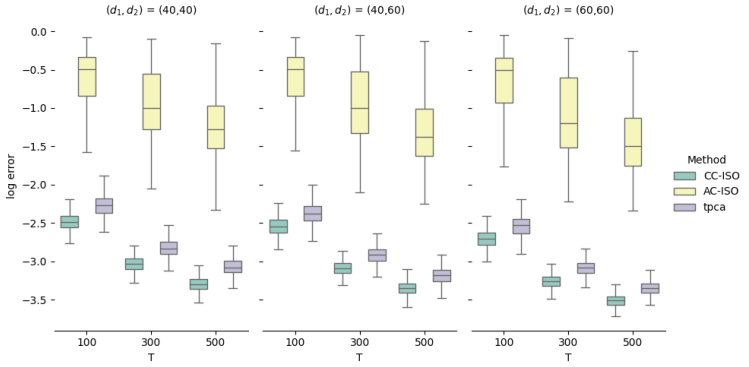



Figure 2: Boxplots of the estimation error over 500 replications under configuration I

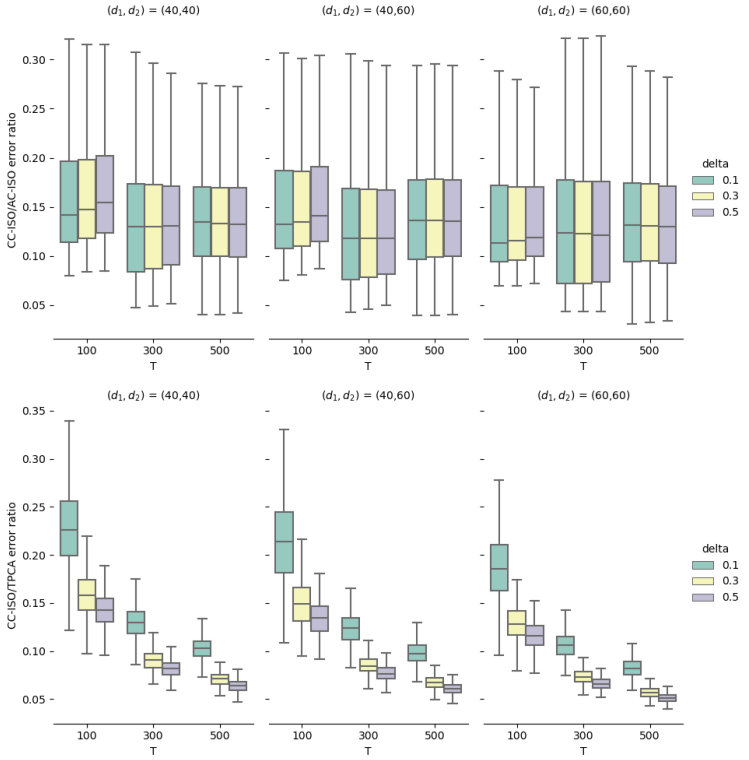



Figure 3: Boxplots of the estimation error over 500 replications under configuration II. Note: The first panel shows the ratio of the estimation error of CC-ISO on AC-ISO. The second panel shows the ratio of the estimation error of CC-ISO on TPCA.

ratio of the estimation errors of CC-ISO to AC-ISO (first panel) and to TPCA (second panel) across different values of $\delta$ and dimensions. It is evident that 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 attributed to TPCA's inability to 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.

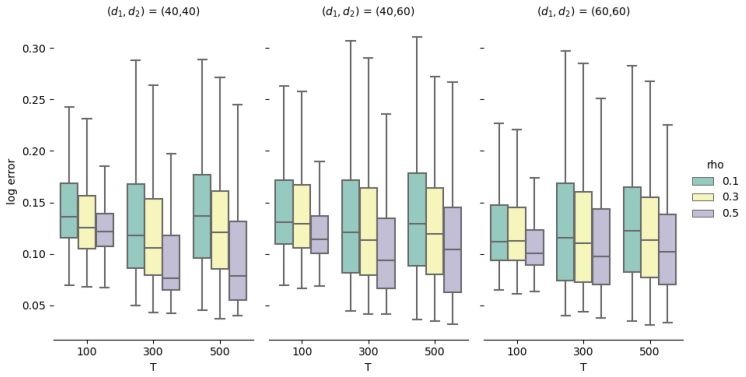



Figure 4: Boxplots of the estimation error over 500 replications under configuration III

Figure 4 presents 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 note that the 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 common components by the serial correlations in the error terms. However, CC-ISO demonstrates its robustness against such serial correlations in the error term.

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 with increasing $T$. Additionally, the rate of decrease in estimation errors is contingent upon the value of $\alpha$, with a higher $\alpha$ leading to a faster rate of 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.

In addition to the previously outlined configurations, 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

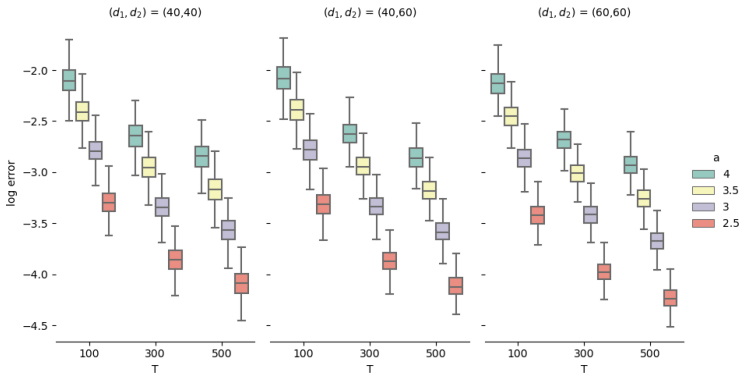



Figure 5: Boxplots of the estimation error over 500 replications under configuration IV

(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 $L = 2r^2$.

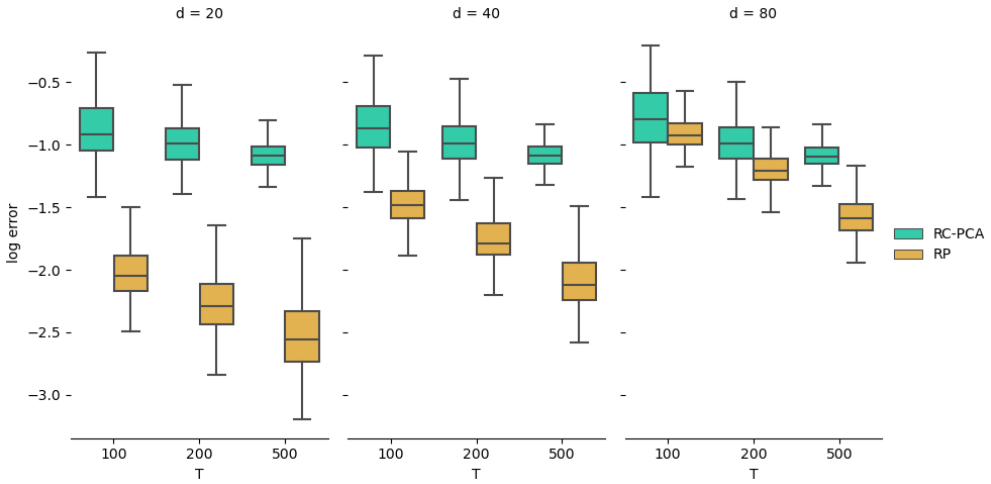



Figure 6: Boxplots of the estimation error over 500 replications under configuration V

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. We observe that the RP algorithm outperforms the RC-PCA algorithm in terms of the accuracy of initial estimations. This domination becomes more pronounced when $d$ is smaller and $T$ is larger. This phenomenon arises because the sample covariance of $\text{Vec}(\mathcal{E}_t)$ approaches the identity matrix as $d$ decreases and $T$ increases. Consequently, $\widetilde{\Sigma}$ is more likely to have eigenvalues that are closer together.

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 of the DGP 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 $\alpha$ indicates a more severe mis-specification in the model.

Figure 7 shows the results under configuration VI. Given $\alpha$, 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 (26) $a_{ik}$ 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 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.

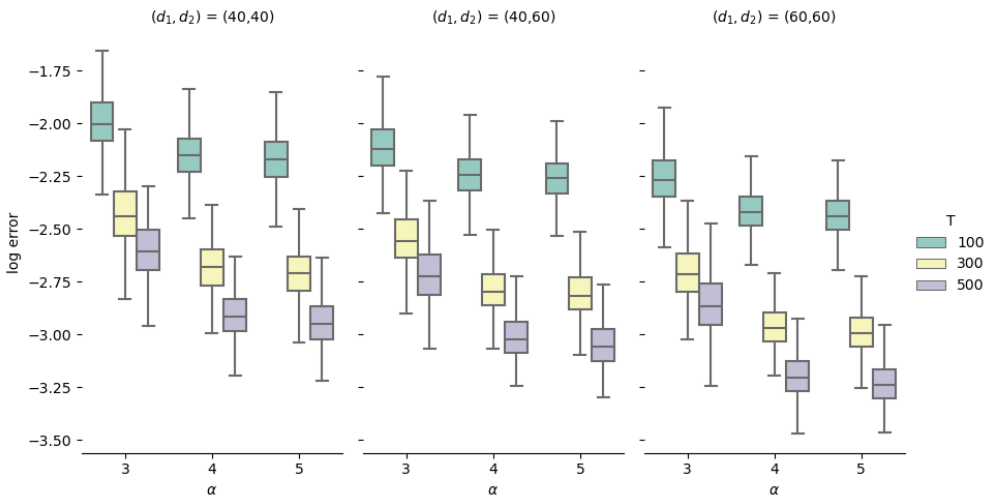



Figure 7: Boxplots of the estimation error over 500 replications under configuration VI

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 Equation (31) with $w_i = (r - i + 1)\bar{d}$ and $\phi \in \{0.1, 0.5\}$.

The results are presented in Table 1. The figures in the table denote the ratio of correct rank estimations to the total number of replications. We can observe that both methods perform very well with accuracy levels close to 1.

# 6 Empirical Application: Characteristic decile portfolios

Latent factor models have been extensively employed in the asset pricing literature to identify risk factors and measure risk exposures. In this section, we conduct an empirical analysis using the dataset compiled by Chen and Zimmermann (2022), which consists of over 200 characteristic-sorted portfolios drawn from previous studies on stock market anomalies. This data has also been analyzed by Babii et al. (2023). We utilize the August 2023 release of the database, which contains monthly portfolio excess returns sorted into 10 deciles based on firm-level characteristics, spanning from January 1990 to December 2022. To ensure a balanced panel of portfolios throughout the sample period, we restrict our analysis to 127 characteristics. Consequently, the dimension of the tensor-valued time series $\mathcal{Y}_t$ is $127 \times 10$

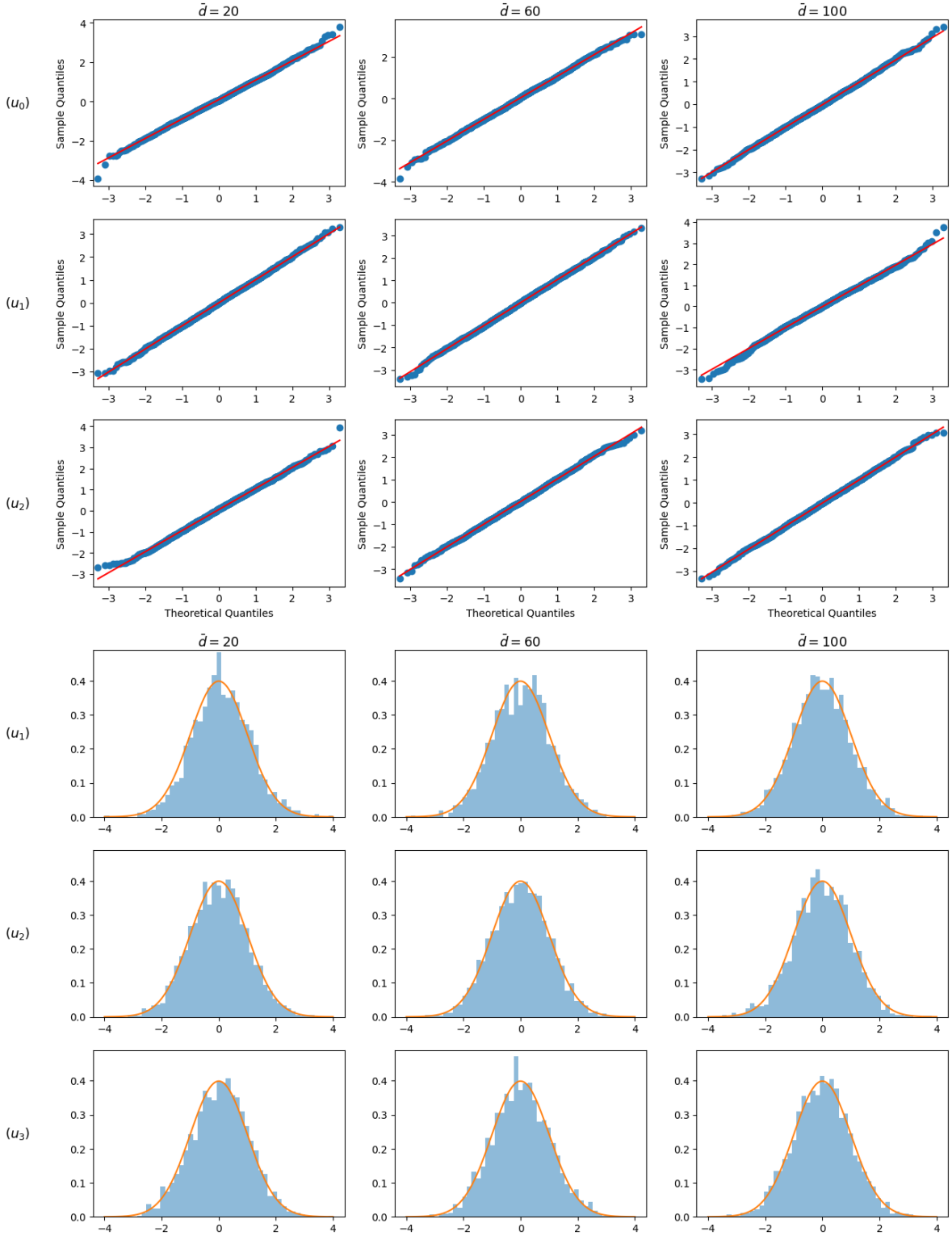



Figure 8: QQ plots and histograms of $\sqrt{T}u^\top \left(\widehat{a}_{ik}^{\text{iso}} - \text{sign}(\widehat{a}_{ik}^{\text{iso}\top} a_{ik}) \cdot a_{ik}\right) / \sigma_{u,ik}$ under Configuration VII. Note: 1. The row displays the results for $u_i$. The column shows the results for different dimension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

Table 1: Rank estimation

| | $\rho$ | 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.

with a sample size of $T = 396$. The risk-free rate, obtained from the Kenneth French data library, is used to compute the excess return of each portfolio.

As noted by Babii et al. (2023), **?** also study a similar dataset, which includes 10 single-sorted decile portfolios for each of 37 characteristics. However, rather than utilizing the natural tensor structure, they flatten the data into a total of 370 portfolios and apply risk-premium PCA based on a vector factor model. In contrast, we exploit the inherent tensor structure by estimating a CP tenor factor model via the proposed CC-ISO method. The CP tensor factor model effectively identifies risk factors and factor exposures without rotation ambiguity. This key property ensures that the estimated factors and loadings are uniquely identified up to sign, independent of rotations or the choice of a varimax algorithm, therefore enhancing the reliability of their economic interpretation.

To implement this, we estimate the CP tensor factor model introduced in Equation 1. As discussed in Section 2, the latent factors can be interpreted as systematic risk factors, while the factor loadings represent heterogeneous risk exposures.

To determine the number of factors, we examine the eigenvalues of the sample covariance matrix. The first factor is substantially stronger than the others, with its corresponding eigenvalue accounting for 86% of the total sum of eigenvalues, which might suggest the presence of a dominant market factor. To ensure that our model captures effects beyond this dominant factor, we set the minimum number of factors to two. Guided by the eigenvalue

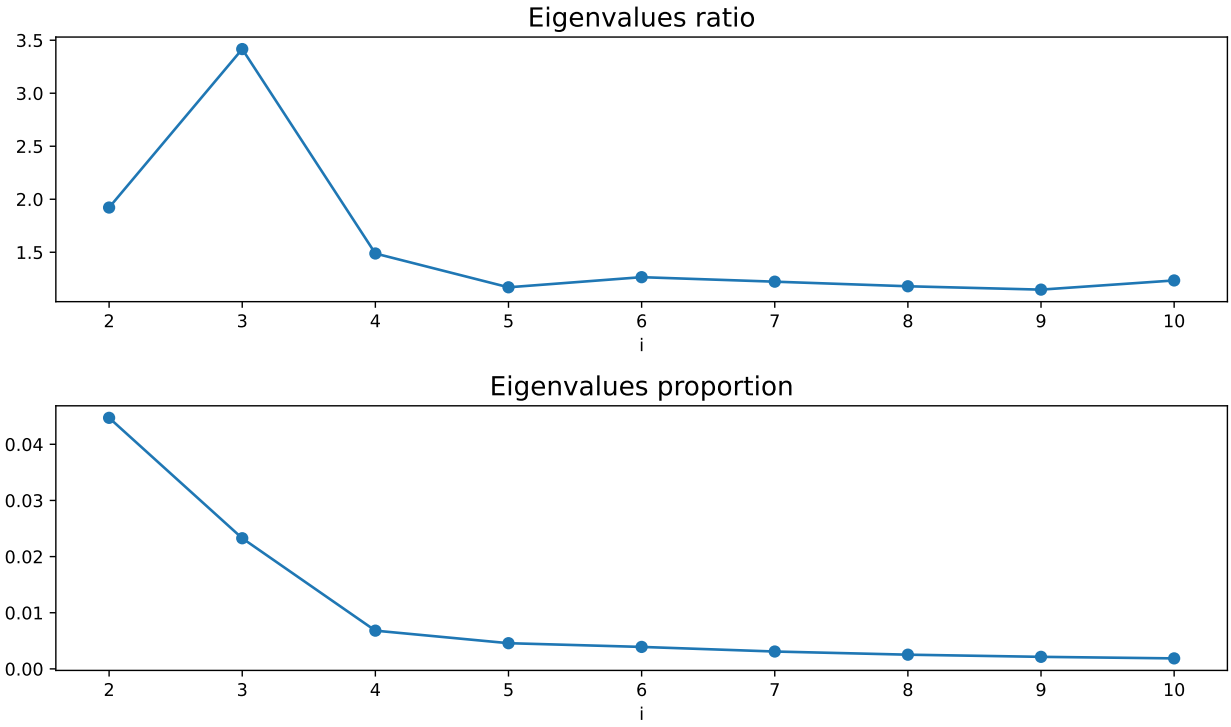



Figure 9: The plot of aggregate eigenvalue ratio (upper panel) and the aggregate scree plot (lower panel) of the excess return of characteristic decile portfolios, starting from the second largest eigenvalue. In the scree plot, the y-axis indicates the proportion of the $i^{th}$ largest eigenvalue to the sum of eigenvalues. In the eigenvalue ratio plot, the y-axis indicates the ratio of the $i^{th}$ eigenvalue to the $(i+1)^{th}$ eigenvalue. The plots suggest 3 factors for our analysis.

ratio plot and the scree plot (Figure 9), we select three factors for our analysis. For the Tucker factor models, the mode-wise eigenvalue ratio and scree plots (Figure 17 in Appendix D) suggest the rank pair $(3, 3)$.

We estimate the CP factor model in Equation (1) using various algorithms: CC-ISO, tensor PCA, AC-ISO with $h = 1$, generalized eigen-analysis based estimation (GE) proposed by Chang et al. (2023) with $K = 1$, and the CC-based and AC-based iterative projection methods based on Tucker decomposition (Tucker-CC-IP & Tucker-AC-IP) proposed by Han et al. (2022a) with rank $(3, 3)$ and $h = 1$ for AC-based projection method. We compare the in-sample performance by $R^2$ defined as:

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

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

To evaluate out-of-sample performance, we conduct one-step-ahead forecasts over the final two years of the sample, following the scheme proposed by Chang et al. (2023). Specifically, for each $s \in \{1, 2, \ldots, 24\}$, we estimate the model using the sample $\{\mathcal{Y}_t\}_{t=1}^{371+s}$ and fit the estimated latent factors with a VAR(1) model. The one-period-ahead excess return $\mathcal{Y}_{372+s}$ is then predicted using the estimated loadings and the predicted one-period-ahead factors. The out-of-sample MSE is defined as:

$$\text{Out-of-sample MSE} = \frac{1}{24}\sum_{s=1}^{24}\left\|\mathcal{Y}_{372+s} - \widehat{\mathcal{Y}}_{372+s}\right\|_F^2.$$

Table 2 displays the $R^2$ and out-of-sample MSE ratio to Tucker-CC-IP across six algorithms along with p-values from the forecast comparison tests of Diebold and Mariano (1995). The tests are one-sided with alternative hypothesis in row "DM I" being that competing algorithms are more accurate than Tucker-AC-IP, while for row "DM II", the alternative is that CC-ISO is more accurate than the competing algorithm. Among the CP factor algorithms, the CC-ISO achieves the best in-sample fit, with the highest $R^2$. Within the ISO methods, CC-ISO outperforms AC-ISO, yielding higher $R^2$ values. The GE algorithm displays the lowest $R^2$ value of 0.768.

Given that Tucker decomposition generally involves a much larger number of factors than CP decomposition (9 vs 3 in our example), it is not surprising that Tucker-CC-IP achieves a bit higher $R^2$ than the CP-based methods. Nonetheless, the improvement is modest, with Tucker-CC-IP increasing $R^2$ by only 3 to 4 percentage points compared to CP models. This observation suggests that the characteristic decile portfolio data might exhibit a CP-like factor structure, where a low-rank CP model effectively captures the underlying risk factors.

Regarding out-of-sample forecast performance, we present the out-of-sample MSE ratios of various methods relative to Tucker-CC-IP, which has the highest MSE. The smaller the ratio, the better the out-of-sample prediction accuracy. Table 2 reports these ratios along with p-values from the Diebold and Mariano (1995) (DM) test. The DM test is one-sided, with DM I testing the null hypothesis that the competing method performs no better than Tucker-CC-IP and DM II comparing CC-ISO against the method in the corresponding column.

Our results show that CC-ISO demonstrates statistically significant superior predictive performance compared to Tucker-CC-IP, Tucker-AC-IP, and GE, while performing similarly to AC-ISO and TPCA. Interestingly, although TPCA achieves the lowest out-of-sample MSE, its advantage over Tucker-CC-IP is not statistically significant. Overall, the Tucker-based algorithms (Tucker-CC-IP and Tucker-AC-IP) yield relatively large MSEs, likely reflecting

overfitting due to the increased number of estimated factors. This result highlights the advantage of CP-based methods in balancing model complexity and predictive accuracy.

| Method | CC-ISO | AC-ISO | TPCA | GE | Tucker-CC-IP | Tucker-AC-IP |
|---|---|---|---|---|---|---|
| $R^2$ | 0.853 | 0.844 | 0.810 | 0.768 | 0.897 | 0.878 |
| OOS MSE ratio | 0.928 | 0.935 | 0.924 | 0.995 | 1.000 | 0.999 |
| DM I | 0.009 | 0.035 | 0.154 | 0.444 | - | 0.492 |
| DM II | - | 0.337 | 0.523 | 0.008 | 0.009 | 0.031 |

Table 2: Comparison of different algorithms for characteristic decile portfolio excess returns. Notes: $R^2$: the in-sample $R^2$ across all algorithms. OOS MSE ratio: the out-of-sample mean square error ratios relative to Tucker-CC-IP, which has the largest MSE. DM I: one-sided p-value from Diebold and Mariano (1995) test, with the alternative that the algorithm in each column is more accurate than Tucker-CC-IP. DM II: one-sided p-value from Diebold and Mariano (1995) test based on the pairwise comparison between the CC-ISO and the algorithm on the column.

Next, we further examine the estimated factors and loadings in our CP factor model. The estimated loadings reveal several notable patterns. Figure 10 presents the estimated loadings along with their 95% confidence intervals across characteristics. The loadings on the first factor are uniformly positive, resembling "long-only" portfolio weights. The loadings on the second and third factors display more variation, with predominantly positive elements. These loadings are closely linked to the volatilities of the decile portfolios. In particular, Figure 11 compares the volatilities of decile 10 portfolios with the corresponding characteristic loadings on the third factor. The figure highlights a strong correlation (0.875) between the two: characteristics with higher decile 10 portfolio volatility tend to have larger third-factor loadings, while those with lower volatility have smaller or even negative loadings.

The factor loadings along the decile mode follow a distinct level-slope-curvature pattern, as shown in Figure 12. The first factor follows a "long-only" structure, with strictly positive and nearly uniform weights across deciles. The second factor shows a slope pattern, with decile-10 and decile-1 loadings having opposite signs but similar magnitudes. This structure resembles a "long-short factor", where high-return deciles are offset by low-return deciles. The third loading shows a convex pattern, forming the "curvature" component.

The consistent "long-only" pattern of the first loading across both modes, combined with the dominant eigenvalue of the first factor, suggests a strong association with the market factor. Figure 13 presents the time series of the first factor along with the excess market return from the Kenneth French data library. The two series exhibit a high correlation (above 0.9), reinforcing the interpretation that this factor primarily captures systematic market risk. In summary, the three estimated factors from the CP factor model can be interpreted as the market factor, the long-short factor, and the volatility factor, respectively.

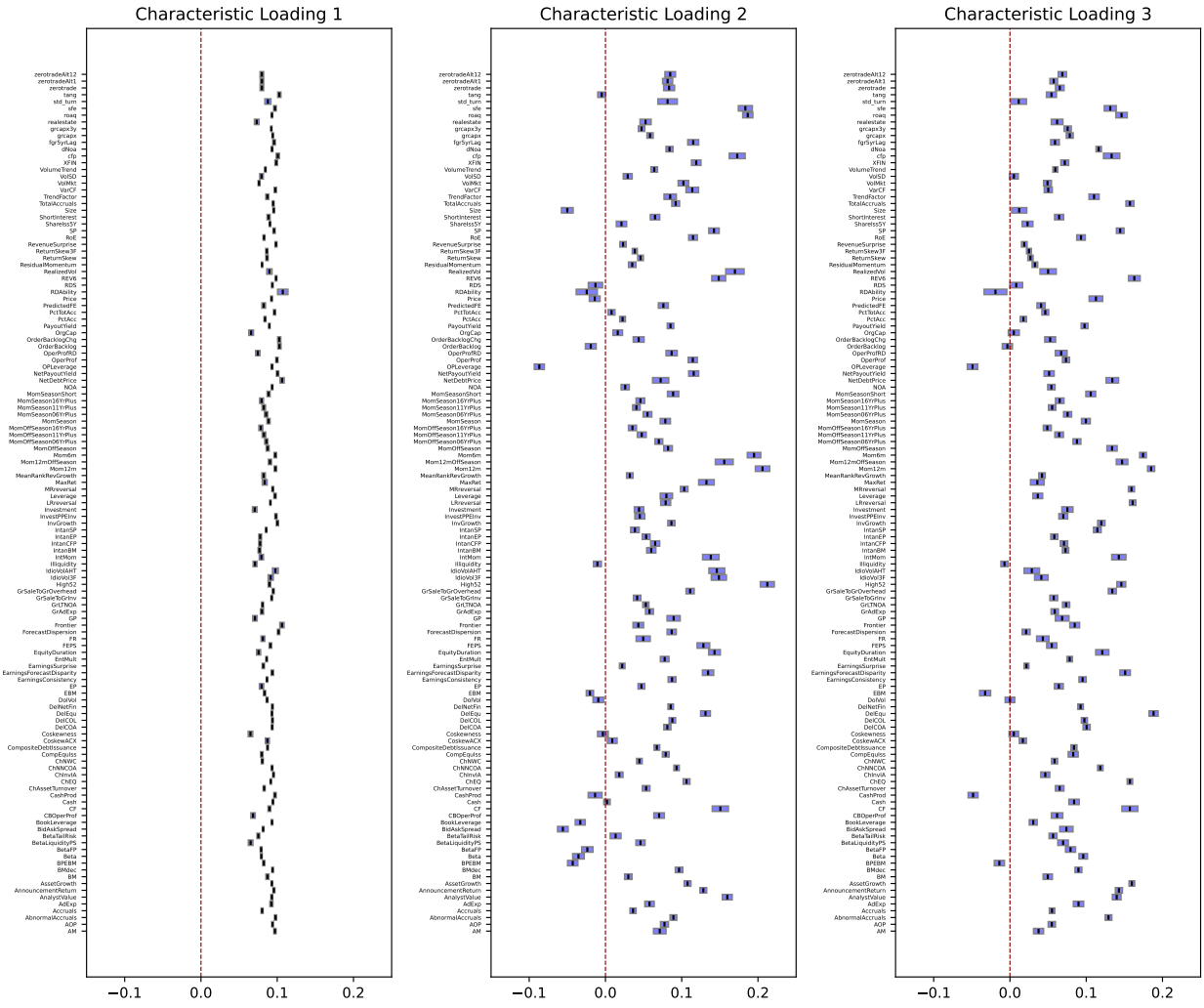



Figure 10: Estimated loadings on characteristics with 95% confidence interval.

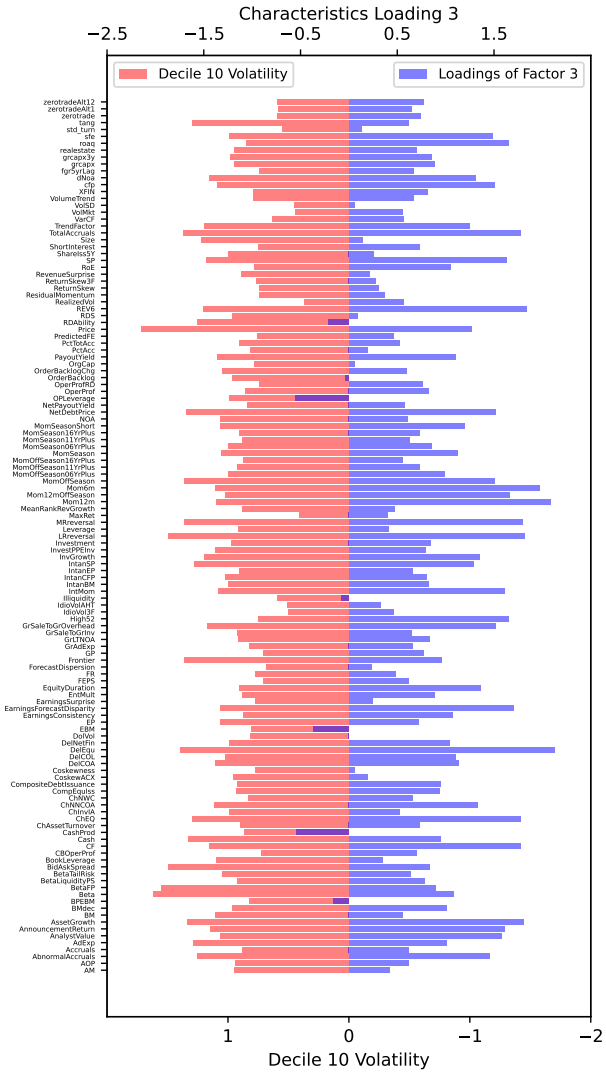



Figure 11: Characteristics loading 3 vs. Decile 10 portfolio volatility.

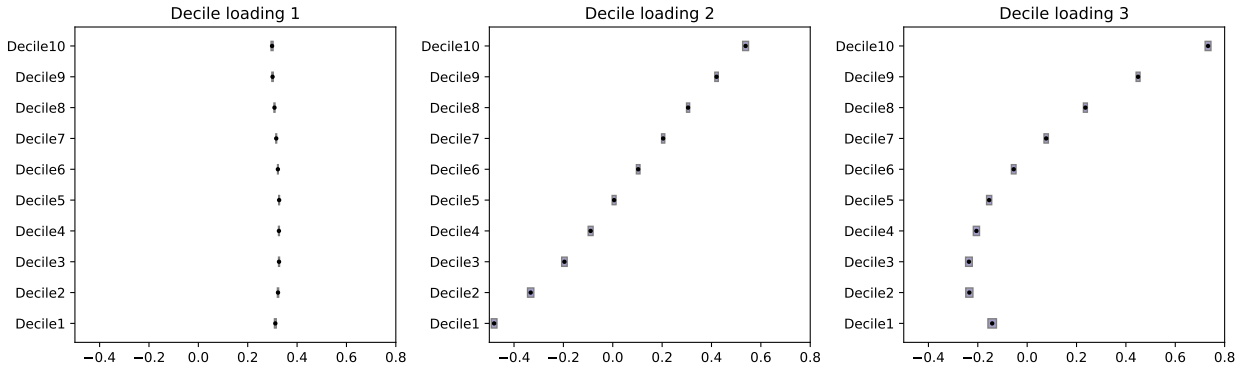



Figure 12: Estimated loadings on decile with 95% confidence interval.

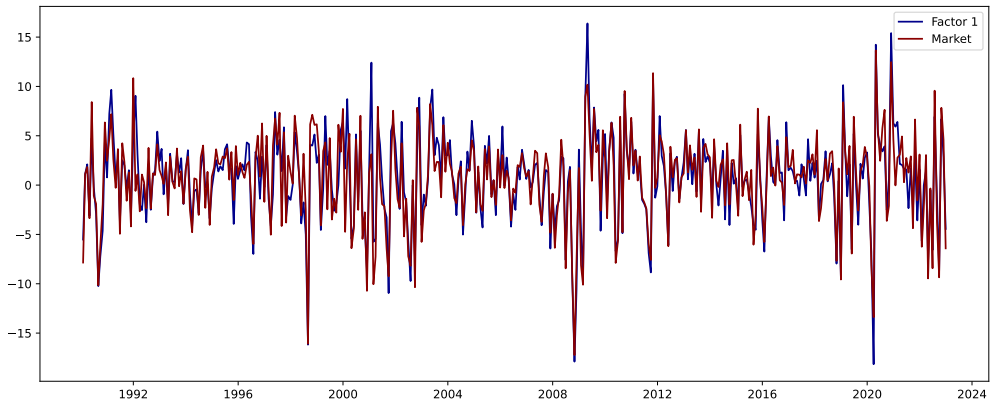



Figure 13: Factor 1 vs. Excess market return. Factor 1 is scaled so that its volatility is equal to the volatility of the excess market return. The correlation between Factor 1 and the market factor is 0.904.

# 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*.

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 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.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

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

# 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 $1 \leqslant j \leqslant r$,

$$\|\widehat{u}_j\widehat{u}_j^\top - u_ju_j^\top\|_2 \leqslant 2\|A\Theta A^\top - U\widetilde{\Lambda}U^\top + \Psi^*\|_2/\lambda_{j,\pm} \leqslant \big(2\lambda_1\delta + 2\|\Psi^*\|_2\big)/\lambda_{j,\pm}, \tag{32}$$

where $\lambda_{j,\pm} = \min\{\lambda_{j-1}-\lambda_j, \lambda_j-\lambda_{j+1}\}$. Combining (32) and the inequality $\|AA^\top - UU^\top\|_2 \leqslant \delta$, we have

$$\|\widehat{u}_j\widehat{u}_j^\top - a_ja_j^\top\|_2 \leqslant \delta + \big(2\lambda_1\delta + 2\|\Psi^*\|_2\big)/\lambda_{j,\pm}. \tag{33}$$

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 $a_{jk}^{\mathrm{rcpca}}$ is the top left singular vector of $\widehat{U}_{jk}$. By Lemma B.2,

$$\|a_{jk}^{\mathrm{rcpca}}a_{jk}^{\mathrm{rcpca}\top} - a_{jk}a_{jk}^\top\|_2^2 \wedge (1/2) \leqslant \|\widehat{u}_j\widehat{u}_j^\top - a_ja_j^\top\|_2^2. \tag{34}$$

Substituting (33) 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

$$\left\|\widehat{a}_{jk}^{\mathrm{rcpca}}\widehat{a}_{jk}^{\mathrm{rcpca}\top}-a_{jk}a_{jk}^{\top}\right\|_2\leqslant C\phi_0,\qquad 1\leqslant j\leqslant i-1, 1\leqslant k\leqslant K, \tag{35}$$

in an event $\Omega$ with high probability, where

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

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 $\Omega$ 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 $\Psi$ is defined by unfolding $\Psi^*$ 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 (52) 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

$$\left\|\widehat{\Sigma}\times_{k=1}^{2K}\widetilde{a}_{\ell_*k}\right\|_2\geqslant(1-2\phi_0^2)\Theta_{ii}. \tag{36}$$

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

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

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

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

where $C'$ is a sufficiently large constant. It contradicts (36) 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 $\Omega$ with probability at least $1-T^{-c}-d^{-c}$,

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

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*

$$\begin{aligned}\|\Psi^*\|_2 \leqslant C\lambda_1\left(\sqrt{\frac{r+\log T}{T}}+\frac{(r+\log T)^{1/\gamma}}{T}\right) + C\left(\sqrt{\frac{d\log(d)}{T}}+\frac{d\log(d)}{T}+1\right)\\ + C\lambda_1^{1/2}\left(\sqrt{\frac{d\log(d)}{T}}+\frac{d\log(d)}{T}\right).\end{aligned} \tag{37}$$

*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

$e_t = \mathrm{vec}(\mathcal{E}_t)$. Write

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

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),

$$\begin{aligned}
\mathbb{P}\left(T\left|u_j^\top \Theta^{-1/2}(\widehat{\Theta} - \Theta)\Theta^{-1/2}u_j\right| \geqslant x\right) &\leqslant T\exp\left(-\frac{x^\gamma}{c_1}\right) + \exp\left(-\frac{x^2}{c_2 T}\right) \\
&\quad + \exp\left(-\frac{x^2}{c_3 T}\exp\left(\frac{x^{\gamma(1-\gamma)}}{c_4(\log x)^\gamma}\right)\right). \qquad (38)
\end{aligned}$$

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 $\Omega_1$ 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 $\Omega_1$,

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

and,

$$\begin{aligned}\|\Upsilon_0\|_2 &\leqslant \left\|A\Theta A^\top\right\|_2 + \left\|A\left(\widehat{\Theta} - \Theta\right)A^\top\right\|_2 \\ &\leqslant \|A\|_2^2\lambda_1 + \left\|A\left(\widehat{\Theta} - \Theta\right)A^\top\right\|_2 \\ &\leqslant (1+\delta)\lambda_1 + C\lambda_1\left(\sqrt{\frac{r + \log T}{T}} + \frac{(r + \log T)^{1/\gamma}}{T}\right) \\ &:= \Delta_\Upsilon.\end{aligned} \tag{39}$$

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

4.1, $N \leqslant C^2 d$ and

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

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

$$\mathbb{P}\left(\left\|\sum_{t=1}^T\left[e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}} - \mathbb{E}e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}\right]\right\|_2 \geqslant x\right) \leqslant 2d\exp\left(-\frac{x^2/2}{\sigma_0^2 + Nx/3}\right). \tag{40}$$

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

$$\left\|\frac{1}{T}\sum_{t=1}^T e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}} - \mathbb{E}e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2\leqslant C\sqrt{d}\}}\right\|_2 \leqslant C_1\sqrt{\frac{d\log(d)}{T}} + C_1\cdot\frac{d\log(d)}{T} \tag{41}$$

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,

$$\mathbb{P}(|M| \geqslant C) \leqslant \exp\left(-T^{c_2}\right). \tag{42}$$

It follows that

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

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

$$\left\| \frac{1}{T} \sum_{t=1}^{T} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 \leqslant C_2 \cdot \frac{d}{T}. \tag{43}$$

Similarly,

$$\mathbb{P}\left( \left\| \mathbb{E} e_t e_t^\top \mathbf{1}_{\{\|e_t\|_2 \geqslant C\sqrt{d}\}} \right\|_2 > 0 \right) = \mathbb{P}\left( \|e_t\|_2 \geqslant C\sqrt{d} \right) \leqslant 4 \exp\left( -C' d^{\frac{\vartheta}{2\vartheta+4}} \right) \leqslant T^{-c_3}. \tag{44}$$

Combing (41), (43), (44), in an event $\Omega_2$ 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 (39), in the event $\Omega_1$, $\sigma_2^2 \lesssim Td\lambda_1$. 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 (43),

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

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 $\Omega_3$ 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*

$$\left\|\widetilde{a}_{j_i,k}\widetilde{a}_{j_i,k}^\top - a_{ik}a_{ik}^\top\right\|_2 \leqslant \psi_i, \qquad 2 \leqslant k \leqslant K, \tag{45}$$

$$\left\|\widetilde{a}_{j_i,1}\widetilde{a}_{j_i,1}^\top - a_{i1}a_{i1}^\top\right\|_2 \leqslant \psi_i + (\delta/\delta_1)w + \psi_i^{K-1}w, \tag{46}$$

*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 $\Xi(\theta) = \mathrm{mat}_{[K-1]}\widehat{\Sigma} \times_1 \times_{K+1}\theta$. 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 $\widetilde{a}_i = \mathrm{vec}(a_{i2} \otimes \cdots a_{iK})$. 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 $1 - d^{-c}$

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

where $\theta_{(kl)}$ is the $(k,l)$th element of $\theta$, $(\widetilde{\Delta}_4)_{..kl}$ represents the $(k,l)$th (3,4) slice of $\widetilde{\Delta}_4$, and $\mathrm{mat}_{(1),(234)}(\cdot)$ 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 $1 - T^{-c} - d^{-c}$,

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

As $d_1 \lesssim \lambda_1$, it follows that in an event $\Omega_0$ with probability at least $1 - T^{-c} - d^{-c}$,

$$\left\|\widetilde{\Psi} \times_3 \times_4 \theta\right\|_2 \leqslant C\|\Psi^*\|_2\sqrt{\log(d)}. \tag{47}$$

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

$$\Xi(\theta) = \Theta_{ii}(a_{i1}^\top \theta a_{i1})\widetilde{a}_i\widetilde{a}_i^\top + \sum_{(j_1,j_2)\neq(i,i)} \Theta_{j_1j_2}(a_{j_11}^\top \theta a_{j_21})\widetilde{a}_{j_1}\widetilde{a}_{j_2}^\top + \widetilde{\Psi} \times_3 \times_4 \theta. \tag{48}$$

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

$$\mathbb{P}\left(\max_{1\leqslant\ell\leqslant L}(a_{i1} \odot a_{i1})^\top \mathrm{vec}(\theta_\ell) \leqslant \sqrt{2\log(L)} - \frac{\log\log(L)}{4\sqrt{\log(L)}} - \sqrt{2\log(8)}\right) \leqslant \frac{1}{4}, \tag{49}$$

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

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

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

Moreover, for the reminder bias term $(a_{j_1 1}\odot a_{j_2 1})^\top (a_{i1}\odot a_{i1})(a_{i1}\odot a_{i1})^\top\theta_\ell$, we have,

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

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

$$\begin{aligned}
&(a_{i1}\odot a_{i1})^\top \mathrm{vec}(\theta_{\ell_*})\Theta_{ii} - \left\| \sum_{(j_1,j_2)\neq(i,i)} \Theta_{j_1 j_2} \left((a_{j_1 1}\odot a_{j_2 1})^\top \mathrm{vec}(\theta_{\ell_*})\right) \widetilde{a}_{j_1}\widetilde{a}_{j_2}^\top \right\|_2 \\
&\geqslant (1-2\delta_1^2 w)\left(\sqrt{2\log(L)} - \frac{\log\log(L)}{4\sqrt{\log(L)}} - \sqrt{2\log(8)}\right)\Theta_{ii} - \left(\sqrt{4\log(r)} + \sqrt{2\log(8)}\right) w\Theta_{ii} \\
&\geqslant C_0\sqrt{\log(d)}\Theta_{ii}, \qquad (50)
\end{aligned}$$

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

Since $\theta_\ell$ 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 $\theta_\ell$, that the above statement (50) for the $i$-th factor holds in an event $\Omega_i$ with probability at least $1 - d^{-C_1}$. By Wedin's perturbation theory, we have in the event $\Omega_0 \cap \Omega_i$,

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

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,

$$\left\|\widetilde{a}_{\ell_*,k}\widehat{a}_{\ell_*,k}^\top - a_{ik}a_{ik}^\top\right\|_2 \leqslant \frac{CR^{(0)}}{\Theta_{ii}} + \delta_1^2 w, \qquad 2 \leqslant k \leqslant K. \tag{51}$$

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

$$\begin{aligned}\widehat{\Sigma} \times_{k=2}^K \widetilde{a}_{\ell_*,k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell_*,k} = &\prod_{k=2}^K \left(\widetilde{a}_{\ell_*,k}^\top a_{ik}\right)^2 \Theta_{ii} a_{i1} a_{i1}^\top + \Psi \times_{k=2}^K \widetilde{a}_{\ell_*,k} \times_{k=K+2}^{2K} \widetilde{a}_{\ell_*,k} \\ &+ \sum_{(j_1,j_2)\neq(i,i)} \prod_{k=2}^K \left(\widetilde{a}_{\ell_*,k}^\top a_{j_1k}\right)\left(\widetilde{a}_{\ell_*,k}^\top a_{j_2k}\right) \Theta_{j_1j_2} a_{j_11} a_{j_21}^\top.\end{aligned}$$

By Lemma A.1 and (51), 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

$$\begin{aligned}\max_{j_1\neq i} \left|a_{j_1k}^\top \widetilde{a}_{\ell_*,k}\right| &= \max_{j_1\neq i} \left|\widetilde{a}_{\ell_*,k}^\top a_{ik} a_{ik}^\top a_{j_1k} + \widetilde{a}_{\ell_*,k}^\top (I - a_{ik}a_{ik}^\top) a_{j_1k}\right| \\ &\leqslant \max_{j_1\neq i} \left|\widetilde{a}_{\ell_*,k}^\top a_{ik}\right| \left|a_{ik}^\top a_{j_1k}\right| + \max_{j_1\neq i} \left\|\widetilde{a}_{\ell_*,k}^\top (I - a_{ik}a_{ik}^\top)\right\|_2 \left\|(I - a_{ik}a_{ik}^\top) a_{j_1k}\right\|_2 \\ &\leqslant \sqrt{1-\psi_i^2}\delta_k + \psi_i\sqrt{1-\delta_k^2} \leqslant \delta_k + \psi_i,\end{aligned} \tag{52}$$

we have

$$\begin{aligned}\left\|\sum_{(j_1,j_2)\neq(i,i)} \prod_{k=2}^K \left(\widetilde{a}_{\ell_*,k}^\top a_{j_1k}\right)\left(\widetilde{a}_{\ell_*,k}^\top a_{j_2k}\right) \Theta_{j_1j_2} a_{j_11} a_{j_21}^\top\right\|_2 &\leqslant (1+\delta_1)\prod_{k=2}^K (\delta_k + \psi_i)\lambda_1 \\ &\leqslant C_K(\delta/\delta_1 + \psi_i^{K-1}) w\Theta_{ii}.\end{aligned}$$

By Wedin's perturbation theory,

$$\left\|\widetilde{a}_{\ell_*,1}\widehat{a}_{\ell_*,1}^\top - a_{i1}a_{i1}^\top\right\|_2 \leqslant \frac{CR^{(0)}}{\Theta_{ii}} + (\delta/\delta_1)w + \psi_i^{K-1}w. \tag{53}$$

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 $1 - T^{-c} - d^{-c}$, (51) and (53) 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

$$\begin{aligned}
\widehat{\Sigma} &= \frac{1}{T}\sum_{t=1}^T \mathcal{Y}_t \otimes \mathcal{Y}_t \\
&= \sum_{i=1}^r \overline{\lambda}_i \otimes_{k=1}^{2K} a_{ik} + \sum_{i\neq j}^r \frac{1}{T}\sum_{t=1}^T f_{it}f_{jt} \otimes_{k=1}^K a_{ik} \otimes_{k=K+1}^{2K} a_{jk} + \frac{1}{T}\sum_{t=1}^T \mathcal{E}_t \otimes \mathcal{E}_t \\
&\quad + \frac{1}{T}\sum_{t=1}^T\sum_{i=1}^r f_{it} \otimes_{k=1}^K a_{ik} \otimes \mathcal{E}_t + \frac{1}{T}\sum_{t=1}^T\sum_{i=1}^r f_{it}\mathcal{E}_t \otimes_{k=K+1}^{2K} a_{ik} \\
&:= \sum_{i=1}^r \overline{\lambda}_i \otimes_{k=1}^{2K} a_{ik} + \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4,
\end{aligned} \tag{54}$$

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 $\Omega_0$ 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

$$\psi_{m,i,k} := \|\widehat{a}_{ik}^{(m)}\widehat{a}_{ik}^{(m)\top} - a_{ik}a_{ik}^\top\|_2, \qquad \psi_{m,k} := \max_i \psi_{m,i,k}, \qquad \psi_m = \max_k \psi_{m,k}. \tag{55}$$

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

$$\widehat{\Sigma} \times_{\ell \in [2K]\setminus\{k,K+k\}} g_{i\ell}^{\top} = \overline{\lambda}_i a_{ik} a_{ik}^{\top} + \Psi \times_{\ell \in [2K]\setminus\{k,K+k\}} g_{i\ell}^{\top}. \tag{56}$$

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

$$\widetilde{\lambda}_{j,i} = \overline{\lambda}_j \prod_{\ell \in [2K]\setminus\{k,K+k\}} a_{j\ell}^{\top} \widehat{g}_{i\ell}^{(m)}. \tag{57}$$

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 $(j,\ell)$. Similar to the proofs of Theorem 3 in Han and Zhang (2023), we can show

$$\max_{j \leqslant r} \|\widehat{a}_{j\ell}^{(m)} - a_{j\ell}\|_2 \leqslant \psi_{m,\ell} / \sqrt{1-1/(4r)}, \quad \|\widehat{b}_{j\ell}^{(m)}\|_2 \leqslant \|\widehat{B}_{\ell}^{(m)}\|_2 \leqslant \left(\sqrt{1-\delta_\ell} - \frac{r^{1/2}\psi_0}{\sqrt{1-1/(4r)}}\right)^{-1}, \tag{58}$$

$$\left\|\widehat{g}_{j\ell}^{(m)} - b_{j\ell}/\|b_{j\ell}\|_2\right\|_2 \leqslant (\psi_{m,\ell}/\alpha)\sqrt{2r/(1-1/(4r))}. \tag{59}$$

Moreover, (58) provides

$$\max_{i \neq j} \left|a_{i\ell}^{\top} \widehat{g}_{j\ell}^{(m)}\right| \leqslant \psi_{m,\ell} / \sqrt{1-1/(4r)}, \quad \left|a_{j\ell}^{\top} \widehat{g}_{j\ell}^{(m)}\right| \geqslant \alpha, \tag{60}$$

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

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

Employing similar arguments in the proof of Lemma A.1, in an event $\Omega_1$ with probability at least $1 - T^{-c_1}$, we have

$$\left\| \widehat{\Theta} - \Theta \right\|_2 \leqslant C_1 \lambda_1 \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right), \tag{61}$$

In the event $\Omega_1$, 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 $\Omega_1$, 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 $\Omega_1$,

$$\begin{aligned} \|\widehat{a}_{ik}^{(m+1)}\widehat{a}_{ik}^{(m+1)\top} - a_{ik}a_{ik}^{\top}\|_2 &\leqslant \frac{2\left\|\sum_{j\neq i}^{r} \widetilde{\lambda}_{j,i} a_{jk} a_{jk}^{\top}\right\|_2 + 2\|\Psi \times_{\ell \in [2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2}{\widetilde{\lambda}_{i,i}} \\ &\leqslant \frac{4\|A_k\|_2^2 \max_{j\neq i}|\lambda_{j,i}| + 4\|\Psi \times_{\ell \in [2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2}{\alpha^{2k-2}\Theta_{ii}}. \end{aligned} \tag{62}$$

To bound the numerator of (62), we write

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

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 (60), the numerator on the right hand side of (62) can be bounded by

$$\begin{aligned}
&\|\Psi \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2 \\
\leqslant &\|\Delta_1 \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2 + \|\Delta_{2,k}(\widehat{g}_{i\ell}^{(m)}, \widehat{g}_{i\ell}^{(m)}, \ell \neq k)\|_2 + 2\|\Delta_{3,k}(\widehat{g}_{i\ell}^{(m)}, \ell \neq k)\|_2 \\
&\quad + 2\|A_k\|_2\|\Delta_{4,k}(\widehat{g}_{i\ell}^{(m)}, \ell \neq k)\|_2 \max_{j\neq i}\prod_{\ell\neq k}^{K}\left|a_{j\ell}^\top \widehat{g}_{i\ell}^{(m)}\right| \\
\leqslant &\sum_{q=1,2,3}\|\Delta_{1,q}\|_2 + \|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell \neq k)\|_2 + (2K-2)\phi_{m,k}\|\Delta_{2,k}\|_{\mathrm{S}} \\
&\quad + 2\|\Delta_{3,k}(g_{i\ell}, \ell \neq k)\|_2 + (4K-4)\phi_{m,k}\|\Delta_{3,k}\|_{\mathrm{S}} \\
&\quad + 2\|A_k\|_2\prod_{\ell\neq k}^{K}\left(\psi_{m,\ell}/\sqrt{1-1/(4r)}\right)\|\Delta_{4,k}(g_{i\ell}, \ell \neq k)\|_2 \\
&\quad + 2\|A_k\|_2(2K-2)\phi_{m,k}\prod_{\ell\neq k}^{K}\left(\psi_{m,\ell}/\sqrt{1-1/(4r)}\right)\|\Delta_{4,k}\|_{\mathrm{S}},
\end{aligned} \tag{63}$$

where

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

$$\|\Delta_{2,k}\|_{\mathrm{S}} = \max_{\substack{\|\widetilde{g}_{i\ell}\|_2=\|\breve{g}_{i\ell}\|_2=1,\\ \widetilde{g}_{i\ell},\breve{g}_{i\ell}\in\mathbb{R}^{d_\ell}}} \|\Delta_{6,k}(\widetilde{g}_{i\ell}, \breve{g}_{i\ell}, \ell \neq k)\|_2.$$

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 (61), in the event $\Omega_1$,

$$\|\Delta_{1,1}\|_2 \lesssim \sqrt{\lambda_1 \Theta_{ii}} \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right) \prod_{\ell \neq k}^{K} \psi_{m,\ell}. \tag{64}$$

Similarly, in the event $\Omega_1$,

$$\|\Delta_{1,2}\|_2 \lesssim \sqrt{\lambda_1 \Theta_{ii}} \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right) \prod_{\ell \neq k}^{K} \psi_{m,\ell}, \tag{65}$$

$$\|\Delta_{1,3}\|_2 \lesssim \lambda_1 \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right) \prod_{\ell \neq k}^{K} \psi_{m,\ell}^2. \tag{66}$$

Let $\Upsilon_{0,i,k} = T^{-1} \sum_{t=1}^{T} f_{it}^2 a_{ik} a_{ik}^\top$ and $\Upsilon_{0,-i,k} = T^{-1} \sum_{t=1}^{T} \sum_{j_1, j_2 \neq i}^{r} f_{j_1 t} f_{j_2 t} a_{j_1 k} a_{j_2 k}^\top$. Then, in the event $\Omega_1$,

$$\|\Upsilon_{0,i,k}\|_2 \leqslant \Theta_{ii} + C_1 \Theta_{ii} \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right) := \Delta_{\Upsilon_i} \asymp \Theta_{ii}, \tag{67}$$

$$\|\Upsilon_{0,-i,k}\|_2 \leqslant \lambda_1 + C_1 \lambda_1 \left( \sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T} \right) := \Delta_{\Upsilon_{-i}} \asymp \lambda_1. \tag{68}$$

Recall $e_t = \mathrm{vec}(\mathcal{E}_t)$. Similar to the proof of Lemma A.1, we can show, in an event $\Omega_2$ with probability at least $1 - T^{-c_2} - d^{-c_2}$,

$$\begin{aligned}
\|\Delta_{2,k}\|_{\mathrm{S}} &\leqslant \left\| \frac{1}{T} \sum_{t=1}^{T} e_t e_t^\top \right\|_2 \lesssim \sqrt{\frac{d \log(d)}{T}} + \frac{d \log(d)}{T} + 1, \\
\|\Delta_{3,k}\|_{\mathrm{S}} &\leqslant \left\| \frac{1}{T} \sum_{t=1}^{T} f_{it} e_t^\top \right\|_2 \lesssim \sqrt{\frac{d \log(d)}{T}} \sqrt{\Theta_{ii}} + \frac{d \log(d)}{T} \sqrt{\Theta_{ii}}, \\
\|\Delta_{4,k}\|_{\mathrm{S}} &\leqslant \left\| \frac{1}{T} \sum_{t=1}^{T} (f_{jt},\ j \neq i)^\top e_t^\top \right\|_2 \lesssim \sqrt{\frac{d \log(d)}{T}} \sqrt{\lambda_1} + \frac{d \log(d)}{T} \sqrt{\lambda_1}.
\end{aligned} \tag{69}$$

We claim that in certain events $\Omega_3$, with probability at least $1 - T^{-c_3} - d^{-c_3}$, for any $1 \leqslant \ell \leqslant K$,

the following bounds hold,

$$\begin{aligned}
\|\Delta_{2,k}(g_{i\ell}, g_{i\ell}, \ell \neq k)\|_2 &\leqslant \frac{C_1 d_k \log(d)}{T} + C_1\sqrt{\frac{d_k \log(d)}{T}} + C_1, \\
\|\Delta_{3,k}(g_{i\ell}, \ell \neq k)\|_2 &\leqslant C_1\sqrt{\frac{d_k \log(d)}{T}}\sqrt{\Theta_{ii}} + \frac{C_1 d_k \log(d)}{T}\sqrt{\Theta_{ii}}, \\
\|\Delta_{4,k}(g_{i\ell}, \ell \neq k)\|_2 &\leqslant C_1\sqrt{\frac{d_k \log(d)}{T}}\sqrt{\lambda_1} + \frac{C_1 d_k \log(d)}{T}\sqrt{\lambda_1}.
\end{aligned} \tag{70}$$

Define

$$R_{k,i} = \sqrt{\frac{d_k \log(d)}{T}} + \frac{d_k \log(d)}{T} + 1 + \sqrt{\frac{\Theta_{ii} d_k \log(d)}{T}} + \frac{\sqrt{\Theta_{ii}} d_k \log(d)}{T}, \ R^* = \max_i \max_k R_{k,i}/\Theta_{ii} \tag{71}$$

As $g_{i\ell}$ is true and deterministic, it follows from (63), (64), (65), (66), (69), (70), in the event $\cap_{q=0}^3 \Omega_q$, for some numeric constant $C_2 > 0$

$$\begin{aligned}
&\|\Psi \times_{\ell \in [2K]\setminus\{k, K+k\}} \widehat{g}_{i\ell}^{(m)\top}\|_2 \\
\leqslant& C_2 R_{k,i} + C_{1,K} R^{(0)} \phi_{m,k} + C_{1,K}\sqrt{\lambda_1 \Theta_{ii}} \prod_{\ell \neq k} \psi_{m,\ell} \left(\sqrt{\frac{r + \log T}{T}} + \frac{(r + \log T)^{1/\gamma}}{T}\right) \\
&+ C_{1,K}\lambda_1 \prod_{\ell \neq k} \psi_{m,\ell}^2 \left(\sqrt{\frac{r + \log T}{T}} + \frac{(r + \log T)^{1/\gamma}}{T}\right) + C_{1,K}\sqrt{\frac{d_k \log(d)}{T}}\sqrt{\lambda_1} \prod_{\ell \neq k} \psi_{m,\ell},
\end{aligned} \tag{72}$$

where $R^{(0)} = \phi^{(0)}$ is defined in (19) in Theorem 4.1. Substituting (72) into (62), by the

definition of $\phi_{m,k}$, we have, in the event $\cap_{q=0}^{3}\Omega_q$,

$$
\begin{aligned}
&\|\widehat{a}_{ik}^{(m+1)}\widehat{a}_{ik}^{(m+1)\top} - a_{ik}a_{ik}^{\top}\|_2 \\
&\leqslant \frac{4(1+\delta_{\max})\lambda_1\prod_{\ell\neq k}\psi_{m,\ell}^2}{\alpha^{2K-2}\Theta_{ii}[\sqrt{(1-\delta_{\max})(1-1/(4r))}\alpha]^{2K-2}} + \frac{4C_2R_{k,i}}{\alpha^{2K-2}\Theta_{ii}} + \frac{4C_{1,K}R^{(0)}\phi_{m,k}}{\alpha^{2K-2}\Theta_{ii}} \\
&\quad + \frac{4C_{1,K}\sqrt{\lambda_1}}{\alpha^{2K-2}\Theta_{ii}}\prod_{\ell\neq k}\psi_{m,\ell}\sqrt{\frac{d_k\log(d)}{T}} \\
&\quad + \frac{4C_{1,K}}{\alpha^{2K-2}}\sqrt{\lambda_1/\Theta_{ii}}\prod_{\ell\neq k}\psi_{m,\ell}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\quad + \frac{4C_{1,K}}{\alpha^{2K-2}}(\lambda_1/\Theta_{ii})\prod_{\ell\neq k}\psi_{m,\ell}^2\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\leqslant C_{\alpha,K}R_{k,i}/\Theta_{ii} + C_{\alpha,K}(\sqrt{r}\psi_0)\psi_{m,k} + C_{\alpha,K}\sqrt{\lambda_1/\lambda_r}\prod_{\ell\neq k}\psi_{m,\ell}R_{k,i}/\Theta_{ii} \\
&\quad + C_{\alpha,K}(\lambda_1/\lambda_r)\prod_{\ell\neq k}\psi_{m,\ell}^2 + C_{\alpha,K}\sqrt{\lambda_1/\lambda_r}\prod_{\ell\neq k}\psi_{m,\ell}\left(\sqrt{\frac{r+\log T}{T}} + \frac{(r+\log T)^{1/\gamma}}{T}\right) \\
&\leqslant C_{\alpha,K}R^* + \rho\psi_m, \qquad (73)
\end{aligned}
$$

where the last inequality comes from condition (23) with $\rho < 1$. As $R^* \lesssim 1$, we have $R^* \lesssim \psi^{\text{ideal}}$. It follows that, after $O(\log(\psi_0/\psi^{\text{ideal}}))$ iterations,

$$\psi_{m,i,k} \lesssim \psi^{\text{ideal}}. \qquad (74)$$

In the end, we divide the rest of the proof into 3 steps to prove (70).

**Step 1.** We prove (70) 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 $1 - d^{-c_1}$,

$$\left\|\frac{1}{T}\sum_{t=1}^T e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}} - \mathbb{E}e_{t,ik}e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \leqslant C\sqrt{d_k}\}}\right\|_2 \leqslant C_1\sqrt{\frac{d_k\log(d)}{T}} + C_1 \cdot \frac{d_k\log(d)}{T} \tag{75}$$

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 $1 - \exp(-T^{c_2}) - T^{-c_2}$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} e_{t,ik} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 \leqslant C_2 \cdot \frac{d_k}{T}. \tag{76}$$

Similarly,

$$\mathbb{P}\left( \left\| \mathbb{E} e_{t,ik} e_{t,ik}^\top \mathbf{1}_{\{\|e_{t,ik}\|_2 \geqslant C\sqrt{d_k}\}} \right\|_2 > 0 \right) = \mathbb{P}\left( \|e_{t,ik}\|_2 \geqslant C\sqrt{d_k} \right) \leqslant 4\exp\left( -C' d_k^{\frac{\vartheta}{2\vartheta+4}} \right) \leqslant T^{-c_3}. \tag{77}$$

Combing (75), (76), (77), in an event with probability at least $1 - T^{-c_4} - d^{-c_1}$,

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

**Step 2.** Now we prove (70) 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

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

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 $\Omega_1$, $\sigma_2^2 \lesssim T d_k \Theta_{ii}$. 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 (76),

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

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 (70) 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 stationary point, $\widehat{a}_{ik}$ is the top eigenvector of the matrix

$$
\begin{aligned}
\widehat{\Sigma}_{ik} =& \widehat{\Sigma} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
=& \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \frac{1}{T}\sum_{t=1}^{T} f_{it} \left( \prod_{\ell\neq k}^{K} a_{i\ell}^{\top} \widehat{g}_{i\ell}^{(m)\top} \right) a_{ik} \otimes \left( \mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} \widehat{g}_{i\ell}^{\top} \right) \\
&+ \frac{1}{T}\sum_{t=1}^{T} f_{it} \left( \prod_{\ell\neq k}^{K} a_{i\ell}^{\top} \widehat{g}_{i\ell}^{(m)\top} \right) \left( \mathcal{E}_t \times_{\ell\in[K]\setminus\{k\}} \widehat{g}_{i\ell}^{\top} \right) \otimes a_{ik} + \frac{1}{T}\sum_{t=1}^{T} \mathcal{E}_t \otimes \mathcal{E}_t \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
&+ \sum_{j\neq i}^{r} \widetilde{\lambda}_{j,i} a_{jk} a_{jk}^{\top} + \sum_{j_1\neq j_2}^{r} \frac{1}{T}\sum_{t=1}^{T} f_{j_1 t} f_{j_2 t} \otimes_{k=1}^{K} a_{j_1 k} \otimes_{k=K+1}^{2K} a_{j_2 k} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
&+ \frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq i}^{r} f_{jt} \otimes_{k=1}^{K} a_{jk} \otimes \mathcal{E}_t \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} + \frac{1}{T}\sum_{t=1}^{T}\sum_{j\neq i}^{r} f_{jt} \mathcal{E}_t \otimes_{k=K+1}^{2K} a_{jk} \times_{\ell\in[2K]\setminus\{k,K+k\}} \widehat{g}_{i\ell}^{\top} \\
&:= \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \Psi_1 + \Psi_2 + \Psi_3 + \Psi_4 + \Psi_5 + \Psi_6 + \Psi_7 \\
&:= \widetilde{\lambda}_{i,i} a_{ik} a_{ik}^{\top} + \Psi, \qquad (78)
\end{aligned}
$$

where $\widetilde{\lambda}_{j,i}$ is defined in (57), $\widehat{g}_{i\ell} = \widehat{b}_{i\ell}/\|\widehat{b}_{i\ell}\|_2$, 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 $\Omega$ with probability at least $1 - T^{-C} - d^{-C}$,

$$
\|\widehat{a}_{ik}\widehat{a}_{ik}^{\top} - a_{ik}a_{ik}^{\top}\|_2 \leqslant C_0 \psi^{\text{ideal}}, \qquad (79)
$$

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

$$
\begin{aligned}
\widehat{a}_{ik}\widehat{a}_{ik}^{\top} - a_{ik}a_{ik}^{\top} =& \frac{1}{\widetilde{\lambda}_{i,i}} P_{a_{ik},\perp} \Psi P_{a_{ik}} + \frac{1}{\widetilde{\lambda}_{i,i}} P_{a_{ik}} \Psi P_{a_{ik},\perp} \\
&+ \frac{1}{\widetilde{\lambda}_{i,i}^2} \left( P_{a_{ik}} \Psi P_{a_{ik},\perp} \Psi P_{a_{ik},\perp} + P_{a_{ik},\perp} \Psi P_{a_{ik}} \Psi P_{a_{ik},\perp} + P_{a_{ik},\perp} \Psi P_{a_{ik},\perp} \Psi P_{a_{ik}} \right) \\
&- \frac{1}{\widetilde{\lambda}_{i,i}^2} \left( P_{a_{ik},\perp} \Psi P_{a_{ik}} \Psi P_{a_{ik}} + P_{a_{ik}} \Psi P_{a_{ik},\perp} \Psi P_{a_{ik}} + P_{a_{ik}} \Psi P_{a_{ik}} \Psi P_{a_{ik},\perp} \right) \\
&+ \mathcal{R}_3(\Psi). \qquad (80)
\end{aligned}
$$

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 $\Omega$.

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

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

By Theorem 4.2 and (79), in the event $\Omega$, $\|\widetilde{\lambda}_{i,i}^{-1} a_{ik,\perp}^\top \Psi a_{ik}\|_2 \leqslant C_0 \psi^{\text{ideal}}$. It follows that, in the event $\Omega$,

$$\left(\widehat{a}_{ik}^\top a_{ik}\right)^2 - 1 \leqslant C_3 (\psi^{\text{ideal}})^2. \tag{81}$$

From the condition (23) and (73), we have $\psi^{\text{ideal}} \lesssim \psi_0^2$, where $\psi_0$ is the error bound for the initialization. By the proofs of Theorem 4.2, i.e. the derivation of (73), we can show, in the event $\Omega$,

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \left(\Psi_4 + \Psi_5 + \Psi_6 + \Psi_7\right) \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2, \tag{82}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_3 \right\|_2 \leqslant C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}} + C_4 \frac{1}{\Theta_{ii}} \left( \frac{d_k \log(d)}{T} + \sqrt{\frac{d_k \log(d)}{T}} + 1 \right), \tag{83}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_1 - \frac{1}{\Theta_{ii} \prod_{\ell \neq k} (a_{i\ell}^\top g_{il}) T} \sum_{t=1}^T f_{it} a_{ik} \otimes \left(\mathcal{E}_t \times_{\ell \in [K]\setminus\{k\}} g_{i\ell}^\top\right) \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2 + C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}}, \tag{84}$$

$$\left\| \frac{1}{\widetilde{\lambda}_{i,i}} \Psi_2 - \frac{1}{\Theta_{ii} \prod_{\ell \neq k} (a_{i\ell}^\top g_{il}) T} \sum_{t=1}^T f_{it} \left(\mathcal{E}_t \times_{\ell \in [K]\setminus\{k\}} g_{i\ell}^\top\right) \otimes a_{ik} \right\|_2 \leqslant C_4 (\psi^{\text{ideal}})^2 + C_4 (\sqrt{r}\psi_0)\psi^{\text{ideal}}. \tag{85}$$

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

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

**Case (ii)**. Let $u \perp a_{ik}$. Define $v = uP_{a_{ik},\perp}$. 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 $\Omega$, $\|\widetilde{\lambda}_{i,i}^{-1}\Psi\|_2 \leqslant C_0\psi^{\text{ideal}}$. By (82), (83), (84), (85), we have,

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

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 (81),

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

In addition, by (87) and (79),

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

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 (70) 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 (81), we have (28). Moreover, by (86) 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* $\Lambda$ *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* $\|M\|_{\mathrm{F}} = 1$ *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,*

$$\left(\|\widehat{a}\widehat{a}^\top - aa^\top\|_2^2\right) \wedge (1/2) \leqslant \|vec(M)vec(M)^\top - vec(ab^\top)vec(ab^\top)^\top\|_2^2. \tag{88}$$

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 $g_{it}$, $\phi$, 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 $\phi$. CC-ISO algorithm, however, outperforms AC-ISO algorithm even with stronger serial correlation in the factor process.

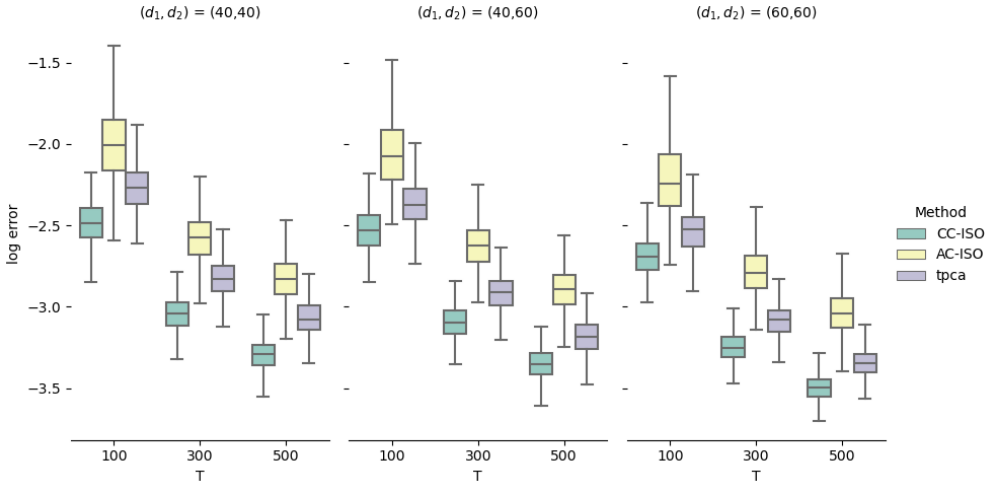



Figure 14: Boxplots of estimation errors over 500 replications under Configuration I with $\phi = 0.5$

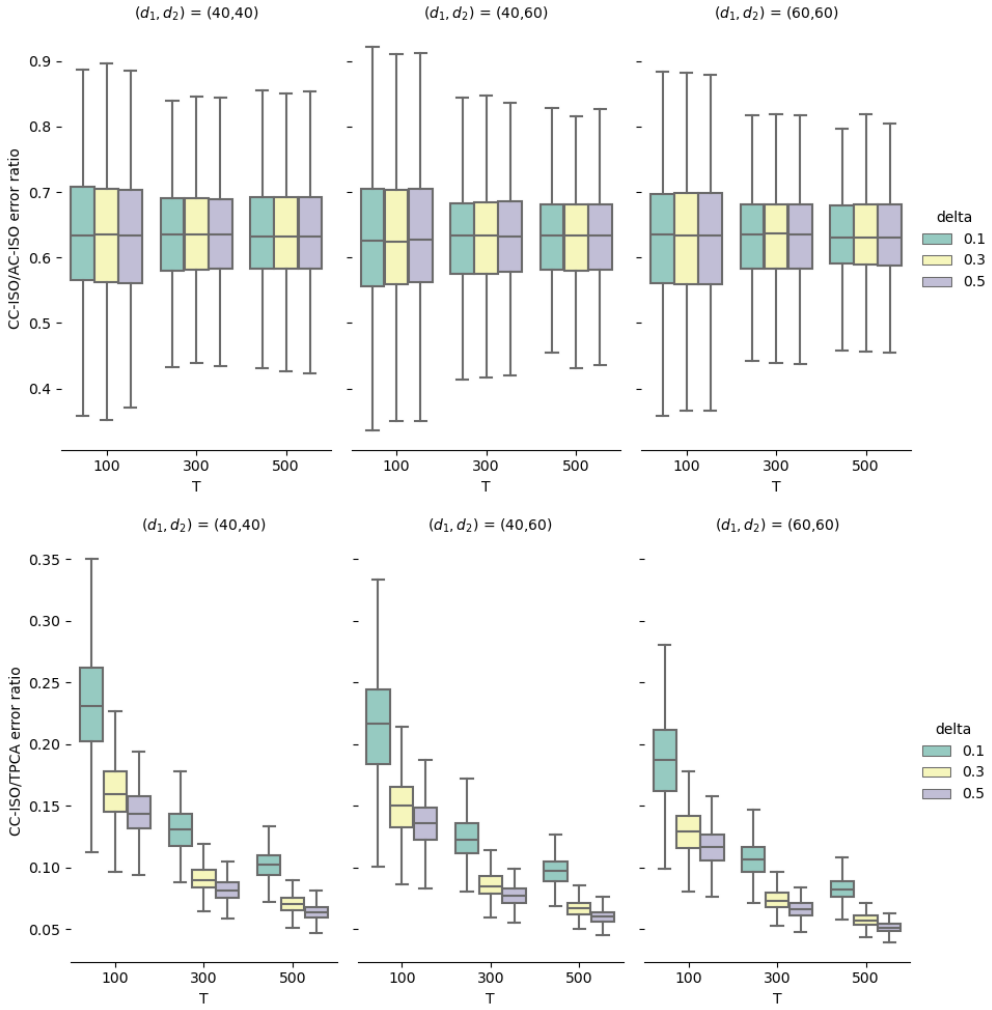



Figure 15: Boxplots of the estimation error over 500 replications under configuration II with $\phi = 0.5$. Note: The first panel shows the ratio of the estimation error of CC-ISO on AC-ISO. The second panel shows the ratio of the estimation error of CC-ISO on TPCA.

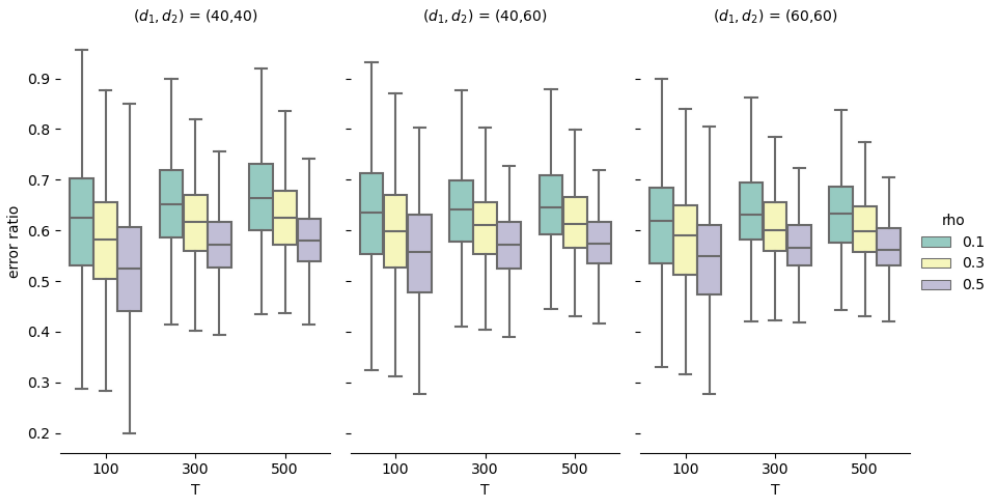



Figure 16: Boxplots of estimation errors over 500 replications under Configuration III with $\phi = 0.5$

# Appendix D More Empirical Results

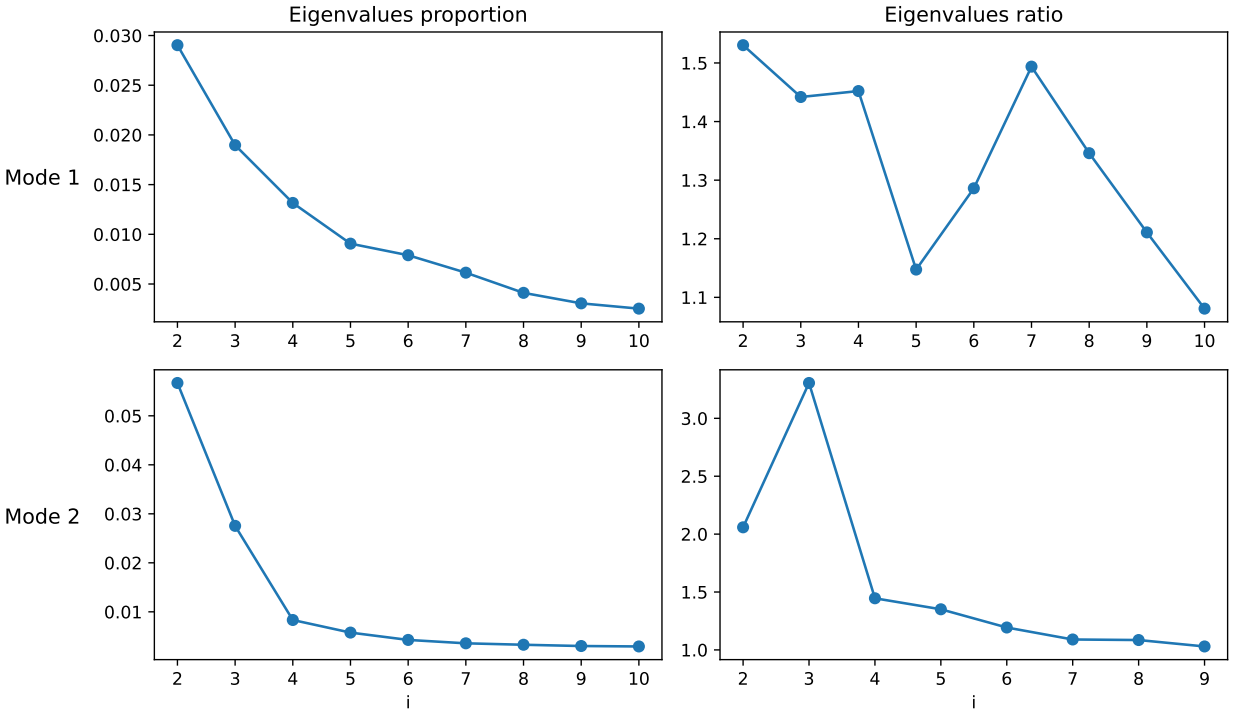



Figure 17: The mode-wide eigenvalue ratio plot (the first column) and the mode-wide scree plot (the second column) of the excess return of characteristic decile portfolios, starting from the second largest eigenvalue. The first row shows the plot of the first mode, i.e., the characteristic mode, and the second row shows the second mode, i.e., the decile mode. In the scree plot, the y-axis indicates the proportion of the $i^{th}$ largest eigenvalue to the sum of eigenvalues. In the eigenvalue ratio plot, the y-axis indicates the ratio of the $i^{th}$ eigenvalue to the $(i+1)^{th}$ eigenvalue. The plots suggest the rank pair $(3, 3)$ for the Tucker factor model.

# Diffusion Index Forecast with Tensor Data*

Bin Chen[1], Yuefeng Han[2], Qiyang Yu[1]

[1]Department of Economics, University of Rochester
[2]Department of Applied and Computational Mathematics and Statistics, University of Notre Dame

**Abstract**

This paper studies diffusion index forecast with both tensor and non-tensor predictors, where the tensor structure is maintained using a Canonical Polyadic tensor factor model. When the number of non-tensor predictors is small, we establish the asymptotic properties of the least squares estimator in this tensor factor-augmented regression framework, accommodating factors of varying strengths. We derive an analytical expression for prediction intervals that accounts for the estimation uncertainty of the latent factors. For settings where the number of non-tensor predictors is large or diverges with the sample size, we adopt a penalized regression approach with the LASSO penalty and demonstrate the consistency of the resulting estimator. Simulation studies validate our theoretical results, and an empirical application to US trade flows illustrates the advantages of our method over existing approaches.

## 1 Introduction

Since the seminal work of Stock and Watson (2002) and Bai and Ng (2006), diffusion index forecast has been widely adopted by government agencies, policy institutes, and academic researchers around the world (see, e.g., Ludvigson and Ng (2007),Ludvigson and Ng (2009),Jurado et al. (2015)). The classical diffusion index model predicts the target variable as a linear

*We thank Rong Chen, Tae-Hwy Lee and seminar participants at Duke University, UC Riverside and the ASSA 2025 Meeting for their useful comments and discussions. Any remaining errors are solely ours.

combination of factors extracted from a large panel of time series data, as well as other important predictors. Its strength lies in the ability to significantly reduce the dimensionality of the predictor space by summarizing it into a small number of factors, enabling effective use of large datasets while keeping the size of the forecasting model small.

However, as the availability and complexity of economic data have expanded, new challenges have emerged for forecasting models. In particular, multidimensional data - panel data with more than two dimensions, have attracted increasing attention in economics due to their ability to capture richer and more intricate relationships. For example, consider predicting U.S. import/export volumes with China using monthly time series data. While the traditional gravity model focuses on bilateral trade flows, it may overlook the influence of trade patterns between the U.S., China, and other countries due to substitution effects. Such data can be structured as a three-dimensional tensor, where the observed time series $\mathcal{X}_t$ is of dimension $N \times N$ for each period $t$, with $N$ denoting the number of countries in the dataset. This type of multidimensional structure poses challenges to diffusion index forecasting, which is based on vector factor models.

A natural approach to tackling this challenge is to flatten or vectorize the tensor time series (see, e.g., Ludvigson and Ng 2007) to fit within the framework of vector factor models. However, this process changes the original data structure, potentially diminishing the interpretability of how information from different dimensions interacts. Furthermore, vectorizing tensors often leads to a significant increase in the number of parameters to estimate, which can result in high computational costs.

In this paper, we consider diffusion index forecast with tensor and non-tensor predictors, where the tensor structure is preserved with a Canonical Polyadic (CP) tensor factor model[1]. Common factors are extracted from tensor data using the contemporary covariance-based iterative simultaneous orthogonalization (CC-ISO) procedure proposed in Chen et al. (2024a). When the number of potential non-tensor predictors is small, we estimate the diffusion index model with ordinary least square (OLS) and establish the consistency and asymptotic normality of the estimator. Unlike Bai and Ng (2006), we allow factors to exhibit different strengths. The convergence rate of the conditional mean prediction for the target variable depends on both the strength of the weakest factor and the sample size. To conduct valid

[1]CP and Tucker structures are the two most commonly assumed low-rank structures for tensor factor models (see, e.g. Kolda and Bader (2009)). We adopt the CP low-rank structure due to its parsimonious features.

inferences, we propose a thresholding-based covariance matrix estimator that is robust to cross-sectional correlation in the idiosyncratic component and demonstrate its consistency.

When the number of potential non-tensor predictors is large - potentially comparable to or exceeding the sample size—we propose a two-step penalized regression approach, applying least absolute shrinkage and selection operator (LASSO) to select important non-tensor predictors. The combination of factor models and sparse regression has been explored in the literature. In a panel data context, Fan et al. (2024) consider the factor augumented sparse linear regression model which includes the vector latent factor model and sparse regression as special cases. Chen et al. (2024b) extend this framework to matrix-variate data and propose two new algorithms for estimation. However, both papers focus on a single type of predictor, either panel or matrix data. In economic forecasting, researchers often have access to mixed types of data. Consider the trade example again. While trade flows among various countries provide valuable information for predicting U.S. import/export volumes, other economic variables such as GDP, unemployment rates, tariffs, and interest rates also play a critical role. These different types of data may reflect distinct sources of predictability. The tensor data on trade flows captures global factors, while macroeconomic variables act as proxies for local predictability. Our model offers a novel framework for integrating these diverse data sources to improve forecast accuracy.

The rest of the paper is organized as follows. In Section 2, we introduce the diffusion index forecast based on the CP tensor facor model and develop the estimator when the number of non-tensor predictor is small. Section 3 derives the inferential theories for the diffusion index model and propose a robust covariance matrix estimator. Section 4 introduces the two-stage penalized estimation for the high-dimensional case and discusses the consistency of the estimator. In Section 5, a simulation study is conducted to assess the reliability of the low- and high-dimensional estimators in finite samples. In Section 6, an empirical example on US export/import forecasting highlights the merits of our approach in comparison with some popular methods in the literature. All mathematical proofs are contained in the appendix.

## 1.1 Notation 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 $\|x\|_q = (x_1^q + ... + x_p^q)^{1/q}$, $q \geq 1$, for any vector $x = (x_1, ..., x_p)^\top$. In particular,

$\|x\|_\infty = \max_{1\le j\le p}|x_j|$. We employ the following matrix norms: matrix spectral norm $\|M\|_2 = \max_{\|x\|_2=1,\|y\|_2=1}\|x^\top My\|_2 = \sigma_1(M)$, where $\sigma_1(M)$ is the largest singular value of $M$; max entry norm: $\|M\|_{\max} = \max_{1\le i\le p,1\le j\le q}|M_{ij}|$ for $M\in\mathbb{R}^{p\times q}$, where $M_{ij}$ denotes the $(i,j)$ entry 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 $|a_n|\le C|b_n|$ holds for all sufficiently large $n$, and $a_n \lesssim b_n$ if there exists a constant $C$ such that $a_n \le Cb_n$.

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 outer 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

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

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

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

The Khatri-Rao (or column-wise Kronecker) product of two matrices $A=(a_1,a_2,\cdots,a_r)$ and $B=(b_1,b_2,\cdots,b_r)$ is defined as $A*B=(a_1\odot b_1,\cdots,a_r\odot b_r)$, where $\odot$ denotes the Kronecker product. Denote $d=d_1\times d_2\times\cdots\times d_K$, $d_{\min}=\min_{k\le K}d_k$ and $d_{\max}=\max_{k\le K}d_k$.

# 2 Model and Estimation

Assume that a decision maker is interested in predicting some univariate series $y_{t+h}$, conditional on $I_t$, the information available at time t, which consists a tensor-variate predictor $\mathcal{X}_t\in\mathbb{R}^{d_1\times d_2\times\cdots\times d_K}$ and a set of other observable variables $w_t\in\mathbb{R}^p$, such as lags of $y_t$. We consider a diffusion index forecast model as

$$y_{t+h} = \beta_0^\top w_t + \beta_1^\top f_t + \epsilon_{t+h}, \quad t=1,...,T. \tag{1}$$

where $h\ge 0$ is the lead time between information available and the target variable. The vector $f_t=(f_{1t},\ldots,f_{rt})'$ consists of $r$ latent factors extracted from the observed tensor data

$\mathcal{X}_t$. Specifically, we model $\mathcal{X}_t$ as a tensor factor model with a CP low-rank structure:

$$\mathcal{X}_t = \sum_{i=1}^{r} s_i f_{it} (a_{i1} \otimes a_{i2} \otimes \cdots \otimes a_{iK}) + \mathcal{E}_t, \tag{2}$$

where $r$ denotes the fixed number of factors and $a_{ik}$ denotes the $d_k$-dimensional loading vector, which needs not to be orthogonal. Without loss of generality, we assume $\mathbb{E}f_{it}^2 = 1$, $\|a_{ik}\|_2 = 1$, for all $1 \leq i \leq r$ and $1 \leq k \leq K$. Then all the signal strengths are captured in $s_i$. The noise tensor $\mathcal{E}_t$ is assumed to be uncorrelated with the latent factors but may exhibit weak correlations across different dimensions[2]. Unlike classical vector factor models which suffer from rotation ambiguity, the CP tensor factors are uniquely identified up to sign changes (Kruskal (1977),Kruskal (1989),Sidiropoulos and Bro (2000)).

To construct forecasts for $y_{T+h}$, the CP factor model (2) needs to be estimated first. We adopt the CC-ISO method proposed by Chen et al. (2024a) in our context. Specifically, we estimate $f_t$ via the following algorithm.

Step 1. Obtain the initial value $\widehat{A}_k^{(0)} = (\widehat{a}_{1k}^{(0)}, \ldots, \widehat{a}_{rk}^{(0)}) \in \mathbb{R}^{d_k \times r}$ via randomized composite PCA (Chen et al. (2024a)) or tensor PCA (Babii et al. (2023)) and compute $\widehat{B}_k^{(0)} = \widehat{A}_k^{(0)} (\widehat{A}_k^{(0)\top} \widehat{A}_k^{(0)})^{-1} = (\widehat{b}_{1k}^{(0)}, ..., \widehat{b}_{rk}^{(0)})$, where $1 \leq k \leq K$.

Step 2. Given the previous estimates $\widehat{a}_{ik}^{(m-1)}$, where $m$ is the iteration number, calculate

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

for $t = 1, \cdots, T$. Then the updated loading vectors $\widehat{a}_{ik}^{(m)}$ are obtained as the top eigenvector of the contemporary covariance $\widehat{\Sigma}(\mathcal{Z}_{1:T,ik}^{(m)}) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{Z}_{t,ik}^{(m)} \mathcal{Z}_{t,ik}^{(m)\top}$, where $1 \leq i \leq r$ and $1 \leq k \leq K$.

[2]Note that if $y_t$ in (1) is a vector of $d$ series and $f_t$ is a vector of $r$ univariate factors obtained from $\mathcal{X}$ via (2), a tensor CP factor-augmented vector autoregressions (TFAVAR) of order $q$ can be constructed as

$$y_{t+1} = \sum_{k=0}^{q} \alpha_{11,k} y_{t-k} + \sum_{k=0}^{q} \alpha_{12,k} f_{t-k} + \epsilon_{1t+1},$$
$$f_{t+1} = \sum_{k=0}^{q} \alpha_{21,k} y_{t-k} + \sum_{k=0}^{q} \alpha_{22,k} f_{t-k} + \epsilon_{2t+1},$$

where $\alpha_{11,k}$, $\alpha_{12,k}$, $\alpha_{21,k}$ and $\alpha_{22,k}$ are model parameters. The inference can be conducted following Bai and Ng (2006). To stay focused, we only consider diffusion index forecast and leave TFAVAR for future research.

Step 3. Update $\widehat{B}_k^{(m)} = \widehat{A}_k^{(m)}(\widehat{A}_k^{(m)\top}\widehat{A}_k^{(m)})^{-1} = (\widehat{b}_{1k}^{(m)}, ..., \widehat{b}_{rk}^{(m)})$ with $\widehat{A}_k^{(m)} = (\widehat{a}_{1k}^{(m)}, \ldots, \widehat{a}_{rk}^{(m)})$.

Step 4. Repeat Steps 2 and 3 until convergence or the maximum number of iterations is reached.

Step 5. Obtain the estimated signal as $\widehat{s}_i = \sqrt{\frac{\sum_{t=1}^{T}\left(\mathcal{X}_t \times_{k=1}^{K} \widehat{b}_{ik}^{\top}\right)^2}{T}}$ and the estimated factors as $\widehat{f}_{it} = \widehat{s}_i^{-1}\left(\mathcal{X}_t \times_{k=1}^{K} \widehat{b}_{ik}^{\top}\right)$, for $i = 1, \cdots, r$ and $t = 1, \cdots, T$.

The estimated factors, $\widehat{f}_t$, along with $w_t$, are then used to estimate the coefficients in Equation (1). When the dimension of the non-tensor predictors $w_t$ is small, we estimate (1) with OLS and the forecast for $y_{T+h}$ is obtained as

$$\widehat{y}_{T+h} = \widehat{\beta}_0^{\top} w_T + \widehat{\beta}_1^{\top} \widehat{f}_T,$$

where $\widehat{\beta}_0^{\top}$ and $\widehat{\beta}_1^{\top}$ are OLS estimates.

The above forecast procedure assumes that the rank of $\mathcal{X}_t$ is known. However, we need to estimate it in practice. We adopt the contemporary covariance-based unfolded eigenvalue ratio estimator considered in Chen et al. (2024a). Other estimator such as inner-product based eigenvalue ratio estimator or autocovariance-based eigen ratio estimator works as well. More details can be found in Han et al. (2023) and Chen et al. (2024a).

# 3 Asymptotic properties

In this section, we consider the asymptotic properties of our estimation when the number of non-tensor predictors is relatively small ($p \ll T$) and thus no regularization is required. Chen et al. (2024a) propose the CC-ISO method and focus on the estimation and inference of loadings while the asymptotic properties of latent factors are unknown. Hence we first fill in the gap by presenting the consistency and asymptotic normality of the estimated latent factors in Section 3.1. Then we derive the inferential theories for the diffusion index model in Section 3.2. Section 3.3 introduces a robust covariance matrix estimator of the factor process for conducting inference on the conditional mean forecasts.

## 3.1 Estimation of Factors

We start with some assumptions that are necessary for our theoretical development.

**Assumption 3.1.** *Denote* $e_t = vec(\mathcal{E}_t) \in \mathbb{R}^d$ *where* $d = \prod_{k=1}^K d_k$ *and* $f_t = (f_{1t}, ..., f_{rt})^\top$,

*(i) For any* $v \in \mathbb{R}^r$ *with* $\|v\|_2 = 1$ *and any* $u \in \mathbb{R}^d$ *with* $\|u\|_2 = 1$,

$$\max_t \mathbb{P}\left(|u^\top e_t| \geq x\right) \leq c_1 \exp(-c_2 x^{\nu_1}), \tag{3}$$

$$\max_t \mathbb{P}\left(\left|v^\top f_t\right| \geq x\right) \leq c_1 \exp\left(-c_2 x^{\nu_2}\right), \tag{4}$$

*for some constants* $c_1, c_2, \nu_1, \nu_2 > 0$.

*(ii) Assume* $(f_t, e_t)$ *is stationary and* $\alpha$*-mixing in* $t$. *The mixing coefficient satisfies*

$$\alpha(m) \leq \exp\left(-c_0 m^\gamma\right) \tag{5}$$

*for some constant* $c_0 > 0$ *and* $\gamma_1 \geq 0$, *where*

$$\alpha(m) = \sup_t \Big\{ \Big| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \Big| : A \in \sigma\left((f_s, e_s), s \leq t\right), B \in \sigma\left((f_s, e_s), 1 \leq i \leq r, s \geq t + m\right) \Big\}.$$

*(iii) Denote* $\Sigma_e = \mathbb{E}(e_t e_t^\top)$. *There exists a constant* $C_0 > 0$ *such that* $C_0^{-1} \leq \lambda_d(\Sigma_e) \leq \cdots \leq \lambda_1(\Sigma_e) \leq C_0$.

*(iv) The factor process* $f_t$ *is independent with the errors* $e_t$.

**Assumption 3.2.** *Denote* $d_{\max} = \max_{1 \leq k \leq K} d_k$, $\frac{1}{\eta_1} = \frac{2}{\nu_1} + \frac{1}{\gamma}$, $\frac{1}{\eta_2} = \frac{\nu_1 + \nu_2}{\nu_1 \nu_2} + \frac{1}{\gamma}$ *and* $\eta_3 = \frac{2}{\nu_2} + \frac{1}{\gamma}$. *Assume* $\frac{1}{\eta_1} > 1$ *and* $\frac{1}{\eta_3} > 1$. *The signal components satisfy* $s_i^2 \asymp d^{\alpha_i}$ *for some* $0 < \alpha_r \leq \alpha_{r-1} \leq \cdots \leq \alpha_1 \leq 1$ *such that:*

$$\sqrt{\frac{d_{\max}}{d^{\alpha_r} T}} + \frac{d_{\max}^{1/\eta_1}}{d^{\alpha_r} T} + \frac{d_{\max}^{1/\eta_2}}{d^{\alpha_r/2} T} + \frac{1}{\sqrt{T}} = o(1).$$

Assumption 3.1(i) assume that the tails of the error and factor processes exhibit exponential decay, which includes sub-Gaussian distribution as an important example. This assumption could be extended to account for polynomial-type tails with bounded moment conditions. Unlike Lam and Yao (2012), Han et al. (2023) and Chen et al. (2024a), Assumption 3.1(ii) and (iii) allow both weak cross-sectional and serial correlations in the error term. Assumption 3.1(ii) assumes the strong-mixing property for the factor process, a standard assumption in

the factor model literature that is flexible enough to accommodate a wide range of time series models.

Assumption 3.1 (iv) imposes independence between factors and errors, which simplifies the analysis of both the ISO algorithm and the forecast model. A more general assumption allowing for limited dependence, as suggested by Bai (2003), could also be considered, though it would introduce significantly greater theoretical complexity.

Unlike Bai (2003) and Fan et al. (2024), Assumption 3.2 allows for varying factor strengths by incorporating a mix of strong and weak factors, with certain conditions on the weakest signal strength, the dimensions of tensor data, and the sample size. Specifically, it ensures that, as $d, T \to \infty$, $\max_{i\le r, k\le K} \|\widehat{a}_{ik}\widehat{a}_{ik}^\top - a_{ik}a_{ik}^\top\|_2 \to 0$, thereby guaranteeing the consistency of the factor estimation.

Let $\psi_0$ be the estimation error of the warm-start initial estimates. Denote $H = \mathrm{diag}(s_1\widehat{s_1}^{-1}, \cdots, s_r\widehat{s_r}^{-1})$. We first present the performance bounds of $\widehat{f_t}$ below.

**Theorem 3.1.** *Suppose Assumptions 3.1- 3.2 hold. Assume that* $\max_{k\le K}\|A_k^\top A_k - I_r\|_2 < 1$ *and* $T \le C\exp(d_{\max})$ *for some constant* $C$*. Suppose that the initial estimation error bounds satisfy the condition:*

$$C_{1,K}\left(\frac{s_1^2}{s_r^2}\right)\psi_0^{2K-3} + C_{1,K}\frac{s_1}{s_r}\left(\sqrt{\frac{\log T}{T}} + \frac{(\log T)^{1/\eta_3}}{T}\right)\psi_0^{K-2} \le \rho < 1, \tag{6}$$

*where* $C_{1,K}$ *is some constant depending on* $K$ *only. Then the estimated tensor factors satisfy*

$$\begin{aligned}
&(i)\ \|\widehat{f_t} - Hf_t\|_2 = O_p(\psi + \frac{1}{s_r}) = O_p\left(\sqrt{\frac{d_{\max}}{d^{\alpha_r}T}} + \frac{d_{\max}^{1/\eta_1}}{d^{\alpha_r}T} + \frac{d_{\max}^{1/\eta_2}}{d^{\alpha_r/2}T} + \frac{1}{d^{\alpha_r/2}}\right)\\
&(ii)\ \|\widehat{f_t} - f_t\|_2 = O_p\left(\psi + \frac{1}{s_r} + \sqrt{\frac{1}{T}}\right) = O_p\left(\sqrt{\frac{d_{\max}}{d^{\alpha_r}T}} + \frac{d_{\max}^{1/\eta_1}}{d^{\alpha_r}T} + \frac{d_{\max}^{1/\eta_2}}{d^{\alpha_r/2}T} + \frac{1}{d^{\alpha_r/2}} + \sqrt{\frac{1}{T}}\right).
\end{aligned} \tag{7}$$

Theorem 3.1 shows that $\widehat{f_t}$ is a consistent estimator of the latent factor $f_t$. However, the convergence rate of $\widehat{f_t}$ to $f_t$ may be slower compared to its convergence to $Hf_t$. This discrepancy arises due to the non-negligible estimation error associated with the factor signal $s_i$. Nevertheless, this does not affect the prediction of $y_{t+h}$, as the impact is absorbed by the

coefficient $\beta_1$. We will provide further discussion on this point in Section 6. When all factors are strong, i.e., $\alpha_i = 1$ for all $1 \leq i \leq r$, Theorem 3.1 implies the following:

$$\|\widehat{f}_t - Hf_t\|_2 = O_p\left(\sqrt{\frac{d_{\max}}{dT}} + \frac{d_{\max}^{1/\eta_1}}{dT} + \frac{d_{\max}^{1/\eta_2}}{d^{1/2}T} + \sqrt{\frac{1}{d}}\right)$$
$$\|\widehat{f}_t - f_t\|_2 = O_p\left(\sqrt{\frac{d_{\max}}{dT}} + \frac{d_{\max}^{1/\eta_1}}{dT} + \frac{d_{\max}^{1/\eta_2}}{d^{1/2}T} + \sqrt{\frac{1}{d}} + \sqrt{\frac{1}{T}}\right).$$

If we further assume that the error term is serially uncorrelated and follows a sub-Gaussian distribution, then the rates simplify to:

$$\|\widehat{f}_t - Hf_t\|_2 = O_p\left(\sqrt{\frac{d_{\max}}{dT}} + \sqrt{\frac{1}{d}}\right)$$
$$\|\widehat{f}_t - f_t\|_2 = O_p\left(\sqrt{\frac{1}{T}} + \sqrt{\frac{1}{d}}\right).$$

.

**Remark 3.1.** *Under the assumption that the error $\mathcal{E}_t$ is serially uncorrelated and both $d$ and $T$ go to infinity, the consistency results require $\sqrt{d_{\max}/(d^{\alpha_r}T)} \to 0$. Setting $d_{\max} = d^{\vartheta_d}$ and $T = d^{\vartheta_T}$, this condition simplifies to $\alpha_r + \vartheta_T > \vartheta_d$. If PCA is applied to the vectorized $\mathcal{X}_t$, with some modifications to the proofs in Bai and Ng (2023), Huang et al. (2022) and Gao and Tsay (2024), it can be shown that consistency requires $\alpha_r + \vartheta_T > 1$. Since $\vartheta_d < 1$, the CC-ISO algorithm imposes a weaker sample size requirement than PCA. Specifically, CC-ISO remains consistent in the range where $\vartheta_d < \alpha_r + \vartheta_T < 1$, whereas PCA does not. Appendix B provides numerical examples and simulations to illustrate this point.*

Denote $B = (b_1, b_2, \ldots, b_r) \in \mathbb{R}^{d\times r}$ where $b_i = b_{iK} \odot b_{iK-1} \odot \cdots \odot b_{i1}$ with $b_{ik}$ defined as $B_k = A_k(A_k^\top A_k)^{-1} = (b_{1k}, ..., b_{rk}) \in \mathbb{R}^{d_k\times r}$, $A_k = (a_{1k}, \ldots, a_{rk}) \in \mathbb{R}^{d_k\times r}$. And denote $\widehat{S} = \mathrm{diag}(\widehat{s}_1, \cdots, \widehat{s}_r)$.

**Assumption 3.3.** *Assume $\sum_{j=1}^{d} B_{j\cdot}e_{jt} \xrightarrow{d} N(0, \Sigma_{Be})$, where $B_{j\cdot}$ is the $j^{th}$ row of $B$ and $\Sigma_{Be} = \lim_{d\to\infty}\sum_{j=1}^{d}\sum_{l=1}^{d}\mathbb{E}\left[B_{j\cdot}B_{l\cdot}^\top e_{jt}e_{lt}\right]$.*

**Theorem 3.2.** *Under Assumptions 3.1- 3.3 and further assume $s_1\psi = o(1)$, as $d, T \to \infty$, we have*

$$\widehat{S}\left(\widehat{f}_t - Hf_t\right) \xrightarrow{d} N(0, \Sigma_{Be}). \tag{8}$$

Theorem 3.2 establishes the asymptotic normality of the estimated factors, confirming that the normal approximation is valid in this context. This result is consistent with the findings of Bai (2003) for vector factor models. Additionally, Theorem 3.2 derives the asymptotic variance of $\widehat{f}_T$, which provides a theoretical basis for inference in the diffusion index model discussed below.

## 3.2 Inference for Diffusion Index Model

We first consider the properties of the OLS estimates when the CC-ISO estimates of the latent factors are used as regressors, and then discuss how to construct a confidence interval for the conditional mean of (1).

To take advantage of the faster convergence rate of $\widehat{f}_t$ to $f_t$, we rewrite the diffusion index model (1) as

$$y_{t+h} = \beta_0^\top w_t + \widetilde{\beta}_1^\top H f_t + \epsilon_{t+h}, \qquad t = 1, \ldots, T. \tag{9}$$

where $\widetilde{\beta}_1 = H^{-1}\beta_1$. The conditional mean of $y_{T+h}$ given the information available at time $T$ is

$$y_{T+h|T} = \beta_0^\top w_T + \beta_1^\top f_T = \beta_0^\top w_T + \widetilde{\beta}_1^\top H f_T, \tag{10}$$

which is an infeasible predictor since it involves the unknown parameters $\beta_0$, $\widetilde{\beta}_1$ and latent factors $f_T$.

To obtain a feasible forecast, the factor process $f_t$ is first estimated using the CC-ISO algorithm discussed in Section 2. Then the coefficients $\widetilde{\beta} = (\beta_0^\top, \widetilde{\beta}_1^\top)^\top$ are estimated via OLS:

$$\widehat{\beta} = \left( \frac{1}{T} \sum_{t=1}^{T-h} \widehat{z}_t \widehat{z}_t^\top \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T-h} \widehat{z}_t y_{t+h} \right), \tag{11}$$

where $\widehat{z}_t = (w_t^\top, \widehat{f}_t^\top)^\top$ and the feasible prediction of $y_{T+h|T}$ is then given by

$$\widehat{y}_{T+h|T} = \widehat{\beta}^\top \widehat{z}_T = \widehat{\beta}_0 w_T + \widehat{\beta}_1^\top \widehat{f}_T. \tag{12}$$

Denote $z_t = (w_t^\top, f_t^\top)^\top$. To study the asymptotic normality of the OLS estimator $\widehat{\beta}$, we

impose the following assumptions.

**Assumption 3.4.** *(i) $z_t$ and $\epsilon_{t+h}$ are independent with $\mathcal{E}_s$ for all $t$ and $s$.*

*(ii) For any $u_w \in \mathbb{R}^p$ with $\|u_w\|_2 = 1$, $w_t$ satisfies:*

$$\max_t \mathbb{P}\left(\left|u_w^\top w_t\right| \geq x\right) \leq c_1 \exp\left(-c_2 x^{\nu_3}\right),$$

*and $\epsilon_{t+h}$ satisfies*

$$\max_t \mathbb{P}\left(|\epsilon_{t+h}| \geq x\right) \leq c_1 \exp\left(-c_2 x^{\nu_4}\right),$$

*for some constant $c_1, c_2, \nu_3, \nu_4 > 0$.*

*(iii) $(z_t, e_t, \epsilon_t)$ is stationary and $\alpha$-mixing in $t$. The mixing coefficients satisfy*

$$\alpha(m) \leq \exp\left(-c_0 m^\gamma\right)$$

*for some constant $c_0 > 0$, where $\gamma$ is defined in Assumption 3.1.*

*(iv) $\mathbb{E}\left[\epsilon_{t+h} | y_t, z_t, y_{t-1}, z_{t-1}, \ldots\right] = 0$ for all $t$.*

These assumptions are standard in both factor and regression analysis. Assumption 3.4 (ii) is weaker than the common assumption that regressors and errors are sub-Gaussian with $\eta_1 = \xi_1 = 2$ (see, for example, Fan et al. (2024), Huang et al. (2022), Gao and Tsay (2024)). Given this weaker condition, Assumption 3.4 (v) imposes additional conditions on the dimensionality and strength of the signals to ensure the consistency and asymptotic normality of $\widehat{f}_t$, $\widehat{\beta}$, and $\widehat{y}_{T+h|T}$. While Assumption 3.4 (ii) could be further relaxed to require only bounded fourth moments for errors and regressors, as in Bai and Ng (2006), doing so would necessitate more complex restrictions on dimension and signal strengths. To simplify the analysis and maintain interpretability, we adopt the current assumption framework.

**Theorem 3.3.** *Under Assumptions 3.1 to 3.4, and $\min\{\frac{2}{\nu_3}, \frac{2}{\nu_4}\} + \frac{1}{\gamma} > 1$, we have*

$$\sqrt{T}(\widehat{\beta} - \widetilde{\beta}) \xrightarrow{d} N(0, \Sigma_{zz}^{-1} \Sigma_{zz,\epsilon} \Sigma_{zz}^{-1})$$

Theorem 3.3 shows the asymptotic normality of $\widehat{\beta}$, centered by $\widetilde{\beta}$, the scaled true coefficient. This result does not hold for the unscaled true coefficient $\beta = (\beta_0^\top, \beta_1^\top)^\top$ because the estimation error of $\widehat{f}_t$ with respect to $f_t$ is of order $\sqrt{T}$. Nonetheless, it does not affect the inference

for the prediction $\widehat{y}_{T+h|T}$ as shown below. A consistent estimator of the asymptotic variance of $\widehat{\beta}$ can be obtained by the sample covariance matrix of the residuals:

$$\widehat{\text{Avar}\,\widehat{\beta}} = \left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t^\top\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t^\top\widehat{\epsilon}_{t+h}^2\right)\left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t^\top\right)^{-1}. \tag{13}$$

Under conditional homoskedasticity such that $\mathbb{E}\left[\epsilon_{t+h}^2|z_t\right] = \sigma_\epsilon^2$, Equation (13) can be simplified to

$$\widehat{\text{Avar}\,\widehat{\beta}} = \widehat{\sigma}_\epsilon^2\left(\frac{1}{T}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t^\top\right)^{-1}, \tag{14}$$

where $\widehat{\sigma}_\epsilon^2 = \frac{1}{T}\sum_{t=1}^{T-h}\widehat{\epsilon}_{t+h}^2$.

**Theorem 3.4.** *Under the assumptions of Theorem 3.3, we have*

$$\widehat{y}_{T+h|T} - y_{T+h|T} \xrightarrow{d} N(0, \sigma^2_{y_{T+h|T}}),$$

*where* $\sigma^2_{y_{T+h|T}} = \frac{1}{T}z_T^\top \text{Avar}(\widehat{\beta})z_T + \beta_1^\top S^{-1}\,\text{Avar}(\widehat{f}_T)S^{-1}\beta_1$ *with* $\text{Avar}(\widehat{f}_T)$ *defined in* (8).

The two terms in the asymptotic variance of $\widehat{y}_{T+h|T}$ decay at different rates, so the convergence rate of $\widehat{y}_{T+h|T}$ is $\min\{\sqrt{d^{\alpha_r}}, \sqrt{T}\}$. Theorem 3.3 assumes $T/d^{\alpha_r} \to 0$, which implies that the first term in $\sigma^2_{y_{T+h|T}}$ dominates. However, simulation results suggest that incorporating the second term in the asymptotic variance of $\widehat{y}_{T+h|T}$ improves inference for the forecast.

Given consistent estimators of $\text{Avar}(\widehat{\beta})$ and $\text{Avar}(\widehat{f}_T)$, the confidence interval for $y_{T+h|T}$ with confidence level $\alpha$ can be constructed as

$$\left(\widehat{y}_{T|T-h} - q_{1-2/\alpha}\widehat{\sigma}_{y_{T+h|T}},\quad \widehat{y}_{T|T-h} + q_{1-2/\alpha}\widehat{\sigma}_{y_{T+h|T}}\right),$$

where $q_{1-2/\alpha}$ is the $1-2/\alpha$ quantile of the standard normal distribution, and

$$\widehat{\sigma}^2_{y_{T+h|T}} = \frac{1}{T}\widehat{z}_T^\top\widehat{\text{Avar}}(\widehat{\beta})\widehat{z}_T + \widehat{\beta}_1^\top\widehat{S}^{-1}\widehat{\text{Avar}}(\widehat{f}_T)\widehat{S}^{-1}\widehat{\beta}_1.$$

With $\widehat{\text{Avar}}(\widehat{\beta})$ given in Equation (13), a consistent estimator of $\text{Avar}(\widehat{f}_T)$ is still needed. Assuming the components of $\mathcal{E}_t$ are cross-sectionally independent, such that $\Sigma_e = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$,

$\text{Avar}(\widehat{f}_T)$ can be consistently estimated by

$$\widehat{\Gamma}_1 = \frac{1}{d}\sum_{j=1}^{d} \widehat{\widetilde{B}}_{j\cdot}\widehat{\widetilde{B}}_{j\cdot}^{\top} \frac{1}{T}\sum_{t=1}^{T-h} \widehat{e}_{jt}^2, \tag{15}$$

where $\widehat{\widetilde{B}}$ is the estimated $\widetilde{B}$ defined on Page 9, with the CC-ISO estimator $\widehat{a}_{jk}$ replacing the unknown $a_{jk}$ with $\widehat{e}_t = \text{vec}(\widehat{\mathcal{E}}_t)$, $\widehat{\mathcal{E}}_t = \mathcal{X}_t - \sum_{i=1}^{r} \widehat{s}_i \widehat{f}_{it}(\widehat{a}_{i1} \otimes \widehat{a}_{i2} \otimes \cdots \otimes \widehat{a}_{iK})$. If cross-sectional dependence is allowed, a robust variance estimator will be introduced in the next section.

## 3.3 Covariance matrix estimation of factor process by thresholding

In the context of vector factor models, Bai and Ng (2006) proposes the CS-HAC estimator of $\text{Avar}(\widehat{f}_T)$ robust to cross-sectional correlation as

$$\widehat{\text{Avar}}(\widehat{f}_T) = \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \widehat{\lambda}_j \widehat{\lambda}_l^{\top} \frac{1}{T}\sum_{t=1}^{T-h} \widehat{e}_{jt}\widehat{e}_{lt},$$

where $n \to \infty$ grows slower than $\min\{d, T\}$, and $\widehat{\lambda}_j$ denotes the estimated factor loading. This estimator could be extended to the CP factor model by replacing $\lambda_j$ with $\widetilde{B}_{j\cdot}$. The consistency of this estimator hinges on the implicit but critical assumption that

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbb{E}\left[\lambda_j \lambda_l^{\top} e_{jt} e_{lt}\right] \to \Sigma_{\lambda,e,t}, \tag{16}$$

for any $n \to \infty$ growing slower than $\min\{d, T\}$. This assumption, however, is too strong for the CP factor model as the cross-sectional autocovariance $\gamma_q = \mathbb{E}\left[e_{it}e_{jt}\right]$, where $q = |i - j|$, is not generally monotonically decreasing with $q$[3].

Recall from Theorem 3.2 that the asymptotic variance of $\widehat{f}_t$ is given by

$$\text{Avar}(\widehat{f}_T) = \Sigma_{Be} = B^{\top}\Sigma_e B,$$

where $B$ can be consistently estimated using the CC-ISO estimator $\widehat{B} = (\widehat{b}_1, \ldots, \widehat{b}_r)$, as

[3]In Appendix A, we provide an illustrative example and simulation results with a commonly assumed dependence structure. Figure 7 shows that $\gamma_q$ exhibits cyclical patterns, which lead to the inconsistency of the CS-HAC estimator. The simulation result is in line with the observation

shown in Chen et al. (2024a). The primary challenge lies in estimating the high-dimensional covariance matrix $\Sigma_e$ in the presence of cross-sectional dependence. To address this, we propose a thresholding estimator $\widehat{\Sigma}_e^{\mathcal{T}}$:

$$\widehat{\Sigma}_e^{\mathcal{T}} = \mathcal{T}_\lambda\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{e}_t\widehat{e}_t^\top\right),$$

where $\left(\widehat{\Sigma}_e^{\mathcal{T}}\right)_{(j,l)} = \mathcal{T}_\lambda\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{e}_{jt}\widehat{e}_{lt}\right)$, $\mathcal{T}_\lambda(\cdot)$ is a thresholding operator and $\widehat{e}_t$ is the vectorized estimated error using the CC-ISO algothrithm. Following Rothman et al. (2009), the thresholding operator $\mathcal{T}_\lambda(\cdot)$ is defined to satisfy the following conditions:

(i) $|\mathcal{T}_\lambda(z)| \leq |z|$;

(ii) $\mathcal{T}_\lambda(z) = 0$ for $|z| \leq \lambda$;

(iii) $|\mathcal{T}_\lambda(z) - z| \leq \lambda$ for all $z$.

Examples of generalized thresholding include the LASSO penalty rule:

$$\mathcal{T}_\lambda(z) = sign(z)\left(|z| - \lambda\right)_+$$

and the SCAD thresholding rule proposed by Fan and Li (2001):

$$\mathcal{T}_\lambda(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+ & \text{if } |z| > 2\lambda \\ \left[(a-1)z - \text{sgn}(z)a\lambda\right]/(a-2) & \text{if } 2\lambda < |z| \leq a\lambda \\ z & \text{if } |z| \leq 2\lambda. \end{cases}$$

The bound for the estimation error of $\widehat{\Sigma}_e^{\mathcal{T}}$ is established uniformly over a class of covariance matrices, as introduced by Bickel and Levina (2008) and Rothman et al. (2009):

$$\mathcal{U}(q, c_0(d), M) = \left\{\Sigma : \sigma_{ii} < M,\ \max_i \sum_{j=1}^{d} |\sigma_{ij}|^q \leq c_0(d)\right\}, \tag{17}$$

for $0 \leq q < 1$. When $q = 0$, this class represents exact sparse covariance matrices, where the number of non-zero entries per column is bounded by $c_0(d)$. For $q > 0$, this class defines approximately sparse covariance matrices, where most of the entries in each column are small. Additional assumptions are imposed to derive the bound for the estimation error of $\widehat{\Sigma}_e^{\mathcal{T}}$.

**Assumption 3.5.** *(i) For all $i$ and $k$, $\max_{1\leq j\leq d_k} {d_k}^{\alpha_i/2} a_{ik,j} \leq C$ for some constant $C > 0$.*

*(ii)* $\log(d)^{2/\mu-1} = o(T)$ *where* $\mu = \min\{\eta_1, \eta_2\}$.

*(iii)* $\frac{d_{\max}}{d^{\alpha_r}} + \frac{d_{\max}^{2/\eta_1}}{d^{2\alpha_r}T} + \frac{d_{\max}^{2/\eta_2}}{d^{\alpha_r}T} = O(\log(d))$.

Assumption 3.5 (i) bounds the max entry of the factor loadings. Similar conditions are used in the strong factor model literature such as Bai (2003) and Fan et al. (2013). In strong factor models, Assumption 3.5 (i) ensures that the factor loadings for each mode are "dense", i.e. the number of zero entries of each column of $\widetilde{A}$ does not increase with $d$. In weaker factor model, however, this number is allowed to increase in $d$ with the rate depending the factor strength $s_i$. Assumption 3.5 (ii) is imposed to ensure that the bound of $|e_{it}e_{jt} - \mathbb{E}[e_{it}e_{jt}]|$ is the same as in Bickel and Levina (2008) and Rothman et al. (2009) to accommodate stationary and ergodic errors. This assumption is also imposed in Fan et al. (2011) and Fan et al. (2013).

The following theorem provides the rate of convergence for $\widehat{\Sigma}_e^{\mathcal{T}}$ over the class $\mathcal{U}(q, c_0(d), M)$:

**Theorem 3.5.** *Suppose Assumptions 3.1-3.2 and 3.5 hold. Assume the true covariance matrix $\Sigma_e$ lies in the set $\mathcal{U}(q, c_0(d), M)$ defined in Equation (17) with parameter $q$, $c_0(d)$ and $M$, and the threshold $\lambda = C'\left(\sqrt{\frac{\log(d)}{T}} + \frac{1}{d^{\alpha_r/2}}\right)$, where $C' > 0$ is a sufficiently large constant. Then we have*

$$\|\widehat{\Sigma}_e^{\mathcal{T}} - \Sigma_e\|_2 = O_p\left(c_0(d)\left(\sqrt{\frac{\log(d)}{T}} + \frac{1}{d^{\alpha_r/2}}\right)^{1-q}\right).$$

**Remark 3.2.** *If Assumption 3.5 (i) is replaced with a "dense" factor loading assumption, that is, there exists a constant $C > 0$ such that $\max_j a_{ik,j} \leq \frac{C}{\sqrt{d_k}}$ for all $i$ and $k$, where $a_{ik,j}$ denotes the $j^{th}$ entry of $a_{ik}$, Theorem 3.5 could be strengthened by replacing $d^{\alpha_r}$ with $d$ in both the threshold and rate. In particular, let $\lambda = C'\left(\sqrt{\frac{\log(d)}{T}} + \sqrt{\frac{1}{d}}\right)$ and we can obtain*

$$\|\widehat{\Sigma}_e^{\mathcal{T}} - \Sigma_e\|_2 = O_p\left(c_0(d)\left(\sqrt{\frac{\log(d)}{T}} + \sqrt{\frac{1}{d}}\right)^{1-q}\right).$$

Fan et al. (2013) show that the thresholding estimator $\widehat{\Sigma}_e^{\mathcal{T}}$ with the adaptive threshold developed by Cai and Liu (2011) achieves the same rate as in Theorem 3.5 within the strong vector factor model framework. While these results could, in principle, be extended to the CP fac-

tor model, the adaptive threshold method presents significant computational challenges when applied to tensor data. Specifically, it requires estimating $\text{var}(e_{jt}e_{lt})$ for all $j$ and $l$, which substantially increases the computational cost due to the high dimensionality of the tensor data. In addition, the adaptive thresholding approach allows $\Sigma_e$ to have diverging diagonal entries, whereas in the CP factor model, the spectral norm of $\Sigma_e$ is typically assumed to be bounded (Chen et al., 2024a; Han et al., 2023). This boundedness assumption aligns with both the theoretical framework and practical considerations of the CP factor model, making the results in Theorem 3.5 sufficient for inference in the diffusion index model.

Define $\widehat{\Gamma}_2 = \widehat{B}^\top \widehat{\Sigma}_e^{\mathcal{T}} \widehat{B}$. Theorem 3.5 implies the consistency of $\widehat{\Gamma}_2$, as summarized below.

**Corollary 3.1.** *Under the assumptions of Theorem 3.5, suppose* $c_0(d)\left(\sqrt{\frac{\log(d)}{T}} + \sqrt{\frac{1}{d^{\alpha_r}}}\right)^{1-q} = o(1)$*, then* $\|\widehat{\Gamma}_2 - \Sigma_{Be}\|_2 = o_p(1)$.

Corollary 3.1 suggests a valid confidence interval for $y_{T+h|T}$ robust to potential cross-sectional correlation in errors.

# 4 Penalized Regression

While diffusion index forecasting with OLS is effective when the number of predictors is relatively small, some real-world applications might involve a large number of potential predictors, sometimes exceeding the sample size. This high-dimensional setting arises in macroeconomic forecasting, financial modeling and trade analysis, where policymakers and researchers need to integrate information from multiple sources. In such contexts, OLS estimation might become unreliable. Moreover, some predictors may be irrelevant, introducing noise rather than improving forecast accuracy. Therefore, it is important to employ variable selection techniques that identify the most relevant predictors while preserving the predictive power of the model. In this section, we extend diffusion index forecasting to accommodate high-dimensional predictors by incorporating regularization, ensuring robust estimation and improved out-of-sample performance.

Let $w_t \in \mathbb{R}^p$ denote the set of high-dimensional predictors, alongside the tensor time series

$\mathcal{X}_t$. We consider the diffusion index forecast model:

$$y_{t+h} = \beta_0^\top w_t + \beta_1^\top f_t + \epsilon_{t+h}, \tag{18}$$

$$w_t = \Lambda f_t + V_t, \tag{19}$$

$$\mathcal{X}_t = \sum_{i=1}^{r} s_i f_{it}(a_{i1} \otimes a_{i2} \otimes \cdots \otimes a_{iK}) + \mathcal{E}_t, \tag{20}$$

where $p$ is allowed to diverge with the sample size $T$.

Substituting Equation 19 into Equation 18, we obtain:

$$y_{t+h} = \beta_0 V_t + \beta_1^{*\top} f_t + \epsilon_{t+h},$$

where $\beta_1^* = \Lambda^\top \beta_0 + \beta_1$. After estimating the factors $f_t$ and $V_t$ from Equation (20) and (19), we obtain the estimators of the unknown parameters $\beta_0$ and $\beta_1^*$ via the following penalized regression:

$$\left(\widehat{\beta}_0, \widehat{\beta}_1^*\right) = \text{argmin}_{\beta_0,\beta_1} \frac{1}{2T} \sum_{t=1}^{T-h} \left(y_{t+h} - \beta_0^\top \widehat{V}_t - \beta_1^{*\top} \widehat{f}_t\right)^2 + \lambda \|\beta_0\|_1, \tag{21}$$

where $\lambda > 0$ is a tuning parameter. Since $\widehat{V}_t$ is orthogonal to $\widehat{f}_t$, the solution to the penalized regression can be obtained via the following steps:

Step 1. Obtain $\widehat{f}_t$ using the CC-ISO algorithm described in Section 2.

Step 2. Estimate $\Lambda$ and $V_t$ via OLS:

$$\widehat{\Lambda} = \sum_{t=1}^{T} w_t \widehat{f}_t^\top \left(\sum_{t=1}^{T} \widehat{f}_t \widehat{f}_t^\top\right)^{-1},$$

$$\widehat{V}_t = w_t - \widehat{\Lambda} \widehat{f}_t.$$

Step 3. Obtain the projection residuals $\widetilde{y}_{t+h}$ by regressing $y_{t+h}$ on $\widehat{f}_t$:

$$\widehat{\beta}_1^* = \left(\sum_{t=1}^{T-h} \widehat{f}_t \widehat{f}_t^\top\right)^{-1} \left(\sum_{t=1}^{T-h} \widehat{f}_t y_{t+h}\right),$$
$$\widetilde{y}_{t+h} = y_{t+h} - \widehat{\beta}_1^{*\top} \widehat{f}_t.$$

Step 4. Estimate $\beta_0$ by regressing $\widetilde{y}_{t+h}$ on $\widehat{V}_t$ using LASSO:

$$\widehat{\beta}_0 = \text{argmin}_{\beta_0} \frac{1}{2T} \|\widetilde{Y} - \widehat{V}\beta_0\|_2^2 + \lambda \|\beta_0\|_1,$$

where $\widehat{V} = (\widehat{V}_1, \ldots, \widehat{V}_{T-h})^\top \in \mathbb{R}^{(T-h)\times p}$ and $\widetilde{Y} = (\widetilde{y}_{1+h}, \ldots, \widetilde{y}_T) \in \mathbb{R}^{T-h}$.

Step 5. Forecast the conditional mean $y_{T+h|T}$ by

$$\widehat{y}_{T+h|T} := \widehat{\beta}_0^\top \widehat{V}_T + \widehat{\beta}_1^{*\top} \widehat{f}_T.$$

The algorithm is based on residual-on-residual regression. By the Frisch-Waugh-Lovell Theorem, it also provides the least squares solution to the original model without explicitly specifying the relationship in Equation (19). While the formulation simplifies theoretical analysis, simulation results in Section 5 confirm its robustness even when the relationship between $w_t$ and $f_t$ is nonlinear.

For $\varsigma \geq 0$, define the sparsity index set $\mathcal{S}_\varsigma := \{j : |\beta_{0,j}| > \varsigma\}$. Let $p_0 := |\mathcal{S}_0|$ denote the cardinality of the support set of $\beta_0$. The following additional assumptions are imposed.

**Assumption 4.1.** *(i) For any $u \in \mathbb{R}^p$ with $\|u\|_2 = 1$, $V_t$ satisfies:*

$$\max_t \mathbb{P}\left(\left|u^\top V_t\right| \geq x\right) \leq c_1 \exp\left(-c_2 x^{\nu_5}\right),$$

*and $\epsilon_{t+h}$ satisfies*

$$\max_t \mathbb{P}\left(|\epsilon_{t+h}| \geq x\right) \leq c_1 \exp\left(-c_2 x^{\nu_6}\right),$$

*for some constant $c_1, c_2, \nu_5, \nu_6 > 0$.*

*(ii)* $(f_t, e_t, V_t, \epsilon_t)$ *is stationary and* $\alpha$*-mixing in* $t$*. The mixing coefficients satisfy*

$$\alpha(m) \leq \exp\left(-c_0 m^{\gamma}\right)$$

*for some constant* $c_0 > 0$*, where* $\gamma$ *is defined in Assumption 3.1.*

*(iii) For a general index set* $\mathcal{S}$*, define the compatibility constant*

$$\phi_{\Sigma_V}(\mathcal{S}) = \min_{\beta \in \mathcal{C}(\mathcal{S},3)} \frac{|\mathcal{S}| \beta^{\top} \Sigma_V \beta}{\|\beta_S\|_1^2},$$

*where* $\Sigma_V = \mathbb{E}\left[V_t V_t^{\top}\right]$*,* $\mathcal{C}(\mathcal{S}, 3) = \{\beta \in \mathbb{R}^p : \|\beta_{\mathcal{S}^C}\|_1 \leq \|\beta_{\mathcal{S}}\|_1\}$ *and* $\beta_{\mathcal{S}} = (\beta_j)_{j \in \mathcal{S}}$*. Assume that* $\phi^2_{\Sigma_V}(\mathcal{S}_\lambda) \geq 1/C$ *for some constant* $C > 0$*.*

*(iv)* $\mathbb{E}\left[V_t f_t\right] = \mathbb{E}\left[V_t \epsilon_{t+h}\right] = \mathbb{E}\left[f_t \epsilon_{t+h}\right] = \mathbb{E}\left[V_t e_t\right] = 0$*.*

*(v) Let* $\Lambda_j$ *denotes the* $j^{th}$ *row of* $\Lambda$*.* $\max_{j=1,\ldots,p} \|\Lambda_j\|_2 \leq C$ *for some constant* $C$*.*

*(vi)* $\beta_0$ *satisfies* $\|\beta_0\|_1 = O(p_0)$ *and* $p_0\left(\psi^2 + d^{-\alpha_r} + \log(p)/T\right) = o(1)$*.*

*(vii) Let* $\eta_6 = 2/\nu_5 + 1/\gamma > 1$ *and* $\eta_7 = 2/\nu_6 + 1/\gamma > 1$*, and assume* $\log(p)^{2/\max\{\eta_3, \eta_4\} - 1} = o(T)$*.*

These assumptions are standard in the analysis of high-dimensional regressions. Assumption 4.1(i) is weaker than the common assumption that regressors and errors are sub-Gaussian, as seen in the high-dimensional regression literature (e.g., Loh and Wainwright (2012) and Fan et al. (2024)). Assumption 4.1(iii) imposes a compatibility condition, which is less restrictive than directly assuming the positive definiteness of the population covariance matrix. Since $V_t$ is not directly observable in the data, it is more natural to impose the compatibility condition on the population covariance matrix rather than its sample counterpart, as is often done in the high-dimensional regression literature. This approach is also adopted in Adamek et al. (2023).

**Theorem 4.1.** *Under Theorem 3.1, 3.1, 3.2, 4.1 and conditions on Theorem 3.1 and* $p = O\left(\exp\left(\left(\psi + d^{\alpha_r/2}\right)^{\nu_5}\right) + \exp\left(d_{max}\right)\right)$*, if the tuning parameter* $\lambda = C\left(\psi^2 + \frac{1}{s_r^2} + \sqrt{\log(p)/T}\right)$

*for some constant $C$ that is large enough, we have*

$$\|\widehat{\beta}_0 - \beta_0\|_1 = O_p\left(p_0\left(\sqrt{\frac{\log(p)}{T}} + \frac{1}{d^{\alpha_r}} + \psi^2\right)\right),$$
$$\|\widehat{\beta}_1 - \beta_1\|_2 = O_p\left(p_0\left(\psi + \frac{1}{d^{\alpha_r/2}} + \sqrt{\frac{\log(p)}{T}}\right)\right),$$
$$|\widehat{y}_{T+h|T} - y_{T+h|T}| = O_p\left(p_0\left((\log(p))^{1/\nu_5}\sqrt{\frac{\log(p)}{T}} + \psi + \frac{1}{d^{\alpha_r/2}}\right)\right),$$

*where* $\psi = \max_{i\le i\le r, 1\le k\le K} = \left\|\widehat{a}_{ik}\widehat{a}_{ik}^\top - a_{ik}a_{ik}^\top\right\|_2 = O_p\left(\sqrt{\frac{d_{max}}{d^{\alpha_r}T}} + \frac{d_{max}^{1/\eta_2}}{d^{\alpha_r/2}T} + \frac{d_{max}^{1/\eta_1}}{d^{\alpha_r}T} + \frac{1}{d^{\alpha_r}}\right)$.

Theorem 4.1 shows that diffusion index forecasting remains consistent even in the presence of a large number of potential predictors. The rate condition on $p$ is imposed to simplify the consistency result. Furthermore, by assuming $d^{\alpha_r}\psi^2 = o(1)$, the result can be improved by eliminating the $\psi$ term in the rates. While selection consistency of the penalized regression could be established with much more involved theoretical derivations and additional assumptions, our primary focus is on prediction. Therefore, we leave this extension for future research to maintain clarity and focus.

**Remark 4.1.** *If we further let the restricted eigenvalue condition in Assumption 4.1(iii) hold with* $\phi^*_{\Sigma_V}(\mathcal{S}) := \min_{\beta\in\mathcal{C}(\mathcal{S},3)} \frac{|\mathcal{S}|\beta^\top\Sigma_V\beta}{\|\beta\|_2^2}$*, we can bound the estimation error of* $\beta_0$ *with 2-norm as in Fan et al. (2024):*

$$\|\widehat{\beta}_0 - \beta_0\|_2 = O_p\left(\sqrt{p_0}\left(\sqrt{\frac{\log(p)}{T}} + \frac{1}{d^{\alpha_r}} + \psi^2\right)\right).$$

**Remark 4.2.** *Suppose there exist low-dimensional predictors* $g_t$ *that are strong predictors for* $y_{t+h}$ *and should be selected for sure. The proposed model can be extended to incorporate* $g_t$ *by including* $g_t$ *in the regression equations* (18) *and* (19)*. The theoretical results in Theorem 4.1 remain valid in this extended setting, provided that* $g_t$ *satisfies additional tail conditions, mixing properties, and moment conditions, corresponding to Assumption 4.1(i), (ii) and (iv).*

# 5 Simulation

In this section we examine the finite-sample properties of the proposed estimators through a simulation study. We consider the following data generating process (DGP) for $\mathcal{X}_t$ with $r = 3$ and $K = 2$:

$$\begin{aligned}
\mathcal{X}_t &= \sum_{i=1}^{r} f_{it} s_i \otimes_{k=1}^{K} a_{ik} + \mathcal{E}_t, \\
f_{it} &= \rho_i f_{it-1} + \sqrt{1-\rho_i^2} u_{it}, \qquad (\rho_1, \rho_2, \rho_3) = (0.6, 0.5, 0.4) \\
\mathcal{E}_t &= \Sigma_{\mathcal{E},1}^{1/2} Z_t \Sigma_{\mathcal{E},2}^{1/2},
\end{aligned}$$

where $u_{it}$ is generated independently from $\mathcal{N}(0,1)$ and $Z_t \sim \mathcal{MN}(0, I_d, I_d)$. Throughout the section, we let $d_1 = d_2$ and let $\Sigma_{\varepsilon,k} = Toeplitz(0.5, d_k)$ such that the $(j,l)^{th}$ entry of $\Sigma_{\mathcal{E},k}$ is equal to $0.5^{|j-l|}$. Factor loadings $A_k = (a_{1k}, \ldots, a_{rk})$ are generated as follows: let $\widetilde{A}_k^{(\mathcal{N})} \in \mathbb{R}^{d_k \times r}$ whose elements are generated independently from $\mathcal{N}(0,1)$. We first generate $\widetilde{A}_k$ by orthonormalizing $\widetilde{A}_k^{(\mathcal{N})}$ through QR decomposition, i.e. $\widetilde{A}_k = (\widetilde{a}_{1k}, \ldots, \widetilde{a}_{rk}) = \text{QR}(\widetilde{A}_k^{(\mathcal{N})})$. Then $A_k = (a_{1k}, \ldots, a_{rk})$ is generated by $a_{ik} = \Sigma_{\mathcal{E},k}^{1/2} \widetilde{a}_{ik} / \sqrt{\widetilde{a}_{ik}^\top \Sigma_{\mathcal{E},k} \widetilde{a}_{ik}}$. We set the factor strength $s_i = (r - i + 1)\sqrt{d^\alpha}$ with $\alpha \in \{0.6, 0.3\}$.

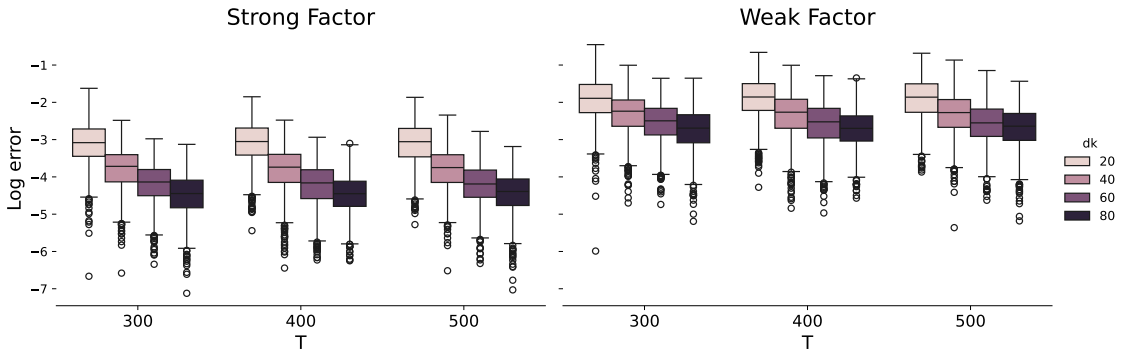
In Section 5.1 and 5.2, we evaluate the consistency and asymptotic distribution of factor estimators. In Section 5.3, we compare the coverage rates of the prediction intervals by CC-ISO and by PCA. In Section 5.4, we illustrate the convergence rates of the LASSO estimators and predictions that are studied in Section 4.

## 5.1 Factor Estimator Consistency

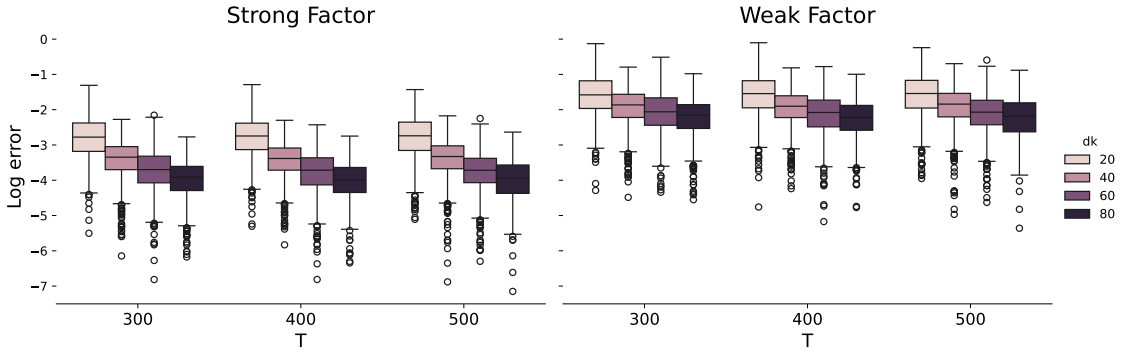
In this section, we evaluate the finite-sample performance of factor estimator $\widehat{f}_t$. Estimation errors are measured as $\|\widehat{f}_t - H f_t\|_2$ at $t = T$ where $H$ is defined in Theorem 3.1[4]. We vary $d_k$ in $\{20, 40, 60, 80\}$ and $T$ in $\{300, 400, 500\}$.

Figure 1 presents boxplots of log estimation errors over 1000 repetitions with strong and weak factors in the settings E1 and E2. In all settings, estimation errors decrease as $d_k$ increases. Additionally, estimation errors are lower in the strong factor setting (S1) than in the weak factor setting (S2). These findings align with Theorem 3.1.

[4]Since our primary interest is in forecasting, we report results for $t = T$. Figures for $t = \frac{T}{2}$ show a similar pattern.




(a) Setting E1




(b) Setting E2

Figure 1: Boxplots of log estimation errors of $\widehat{f}_T$.

## 5.2 Factor Estimator Distribution

Next, we conduct simulations to assess the asymptotic normality of $\widehat{f}_t$, as stated in Theorem 3.1, and to evaluate the proposed covariance matrix estimator in Theorem 3.5. We vary $d_k$ in $\{40, 60, 80\}$ and let $T = 800 + \lceil d^{3/4} \rceil$.

In setting (E1), where the idiosyncratic errors are uncorrelated, the asymptotic variance is estimated using $\widehat{\Gamma}_1$ in Equation(15). In setting (E2), which allows for cross-sectional dependence, the covariance matrix of the factors is then estimated using $\widehat{\Gamma}_2$, as studied in Section 3.3. Specifically, we use the SCAD thresholding function developed by Fan and Li

(2001), defined as

$$\mathcal{T}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+ & \text{if } |z| > a\lambda \\ [(a-1)z - \text{sgn}(z)a\lambda]/(a-2) & \text{if } 2\lambda < |z| \leq a\lambda \\ z & \text{if } |z| \leq 2\lambda, \end{cases}$$

where we set $a = 3.7$ as suggested in Fan and Li (2001).[5] The threshold $\lambda$ is set to $\sqrt{\log(d)/T} + \sqrt{1/d}$.

Figures 2 shows the distribution of $\widehat{\Sigma}_{Be}^{-1/2}\widehat{S}(\widehat{f}_T - Hf_T)$ over 2000 repetitions under various error structures and factor strengths. The results show that the distribution closely approximates the standard normal distribution, which validates Theorem 3.2. Furthermore, the result remains robust to cross-sectional dependence, supporting the effectiveness of the proposed thresholding covariance matrix estimation.

## 5.3 Prediction Interval

In this section, we examine the prediction intervals for $y_{T+1|T}$ constructed based on Theorem 3.4. The target variable $y_{t+1}$ is generated as

$$y_{t+1} = \beta_0 + \beta_1^\top f_t + \epsilon_{t+1},$$

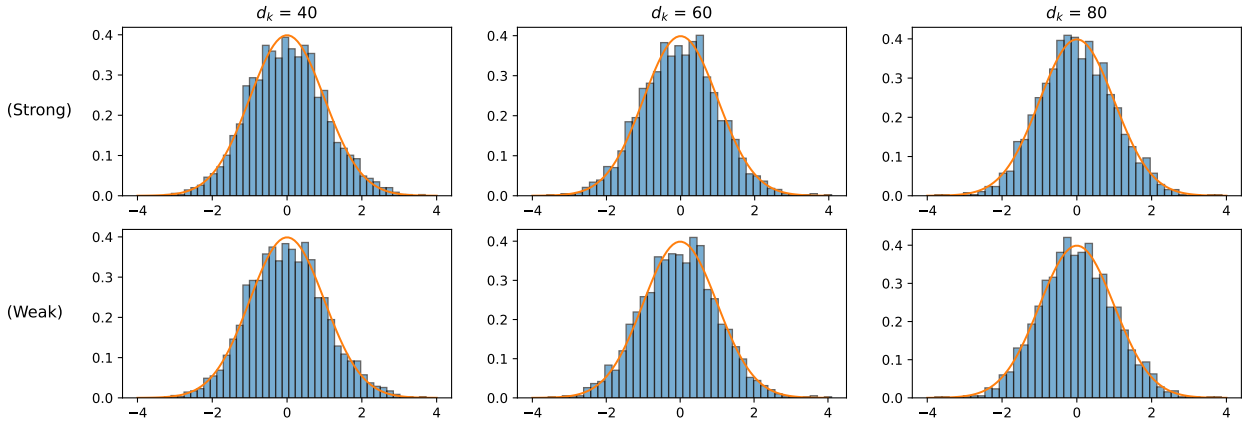
where $\beta_0 = 0.5$ and $\beta_1 = (0.5, 0.5, 0.5)$. The idiocyncratic error $\epsilon_{t+1}$ is drawn independently from $\mathcal{N}(0, \sigma_\epsilon^2)$ under two different settings:

($\epsilon 1$) Homoskedasticity: $\sigma_{\epsilon,t}^2 = 1$;

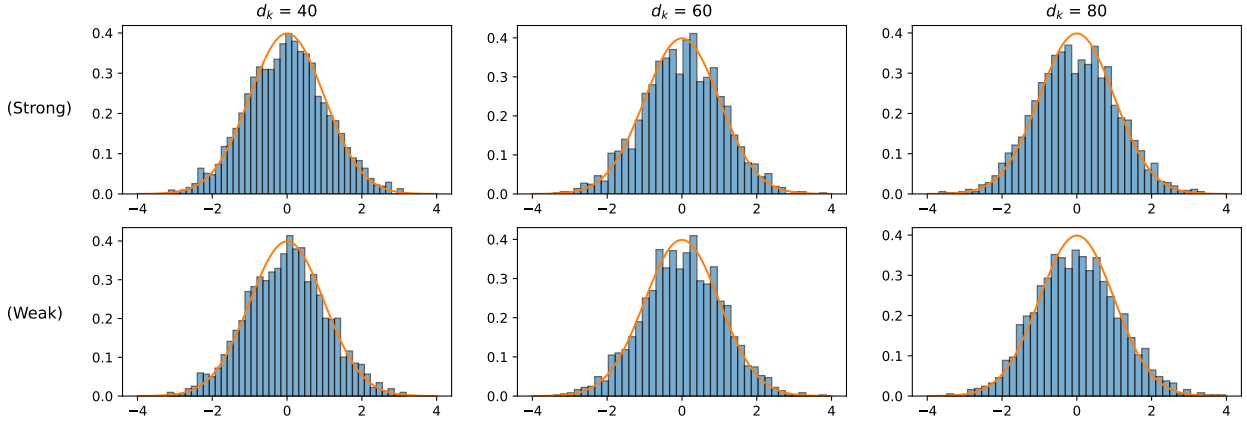
($\epsilon 2$) Heteroskedasticity: $\sigma_{\epsilon,t}^2 = \nu_t$; where $\nu_t \sim U(0.5, 1.5)$.

Set $d_1 = d_2 \in \{20, 40, 60\}$ and $T = 800 + \lceil d^{3/4} \rceil$, with a confidence level of 0.95. We assess the finite-sample performance of the prediction interval for $\widehat{y}_{T+1|T}$ proposed in Equation (**??**) and compare it with the vector PCA method of Bai and Ng (2006). For this comparison, we apply the classical PCA method to the vectorized tensor $x_t := \text{vec}(\mathcal{X}_t) \in \mathbb{R}^d$ and construct

[5]The other three thresholding functions (hard thresholding, soft thresholding and adaptive LASSO) considered in Rothman et al. (2009) are also evaluated, yielding similar simulation results.




(a) Setting E1




(b) Setting E2

Figure 2: Sample distribution of $\widehat{\Sigma}_{Be}^{-1/2}(\widehat{\widetilde{f}}_t - \widetilde{f}_t)$ over 2000 repetitions under different settings. The orange line is the pdf of standard normal.

the confidence interval following Bai and Ng (2006) and Bai and Ng (2023):

$$\left(\widehat{y}_{T+h|T} - q_{1-2/\alpha}\widehat{\sigma}_{y_{T+h|T},pca}, \quad \widehat{y}_{T+h|T} + q_{1-2/\alpha}\widehat{\sigma}_{y_{T+h|T},pca}\right), \tag{22}$$

where $\widehat{\sigma}^2_{y_{T+h|T},pca} = \frac{1}{T}\widehat{z}_T^{(pca)\top}\ \widehat{\text{Avar}}(\widehat{\beta}^{(pca)})\widehat{z}_T^{(pca)\top} + \frac{1}{d}\widehat{\beta}_1^{(pca)\top}\widehat{\text{Avar}}(\widehat{f}_T^{(pca)})\widehat{\beta}_1^{(pca)\top}$. The variance estimator for $\widehat{f}_T^{(pca)}$ is given by

$$\widehat{\text{Avar}}(\widehat{f}_T^{(pca)}) = \widetilde{V}^{-1}\widehat{\Gamma}_t\widetilde{V}^{-1},$$

where $\widetilde{V}$ is a diagonal matrix with diagonal elements equal to the top $r$ eigenvalues of

$\frac{1}{dT}\sum_{t=1}^{T} x_t x_t^\top$, and $\widehat{z}_T^{(pca)}$, $\widehat{\beta}^{(pca)}$, and $\widehat{f}_T^{(pca)}$ are the corresponding PCA estimators. Following Bai and Ng (2006) and Bai and Ng (2023), we adopt the following estimators of $\Sigma_{Be}$: in the case of uncorrelated errors, $\widehat{\Gamma}_t$ is similar to $\widehat{\Gamma}_1$ in (15), except that $\widehat{\widehat{B}}$ is replaced by the factor loadings estimated via PCA; for correlated errors, HAC-type estimator is used:

$$\widehat{\Gamma}_t^{(HAC)} = \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \widehat{A}_{j:}^{(\mathrm{PCA})}\widehat{A}_{l:}^{\top(\mathrm{PCA})}\frac{1}{T}\sum_{t=1}^{T}\widehat{e}_{jt}^{(\mathrm{PCA})}\widehat{e}_{lt}^{(\mathrm{PCA})},$$

where $\widehat{A}_{j:}^{(\mathrm{PCA})}$ and $\widehat{e}_{jt}^{(\mathrm{PCA})}$ are factor loadings and errors estimated via PCA, respectively. The tuning parameter is set as $n = \min\{\sqrt{d}, \sqrt{T}\}$ as suggested by Bai and Ng (2006). For the heteroskedastic setting, $\mathrm{Avar}(\widehat{\beta})$ is estimated using Equation ((13)). In the homoskedastic case, $\mathrm{Avar}(\widehat{\beta})$ follows the same formulation but with $\widehat{\Sigma}_{zz,\epsilon}$ replaced by (14).

Table 1 shows the coverage rate of the three estimated prediction intervals under setting (E1) and (E2) and with both strong and weak factors. The coverage rates of CP are quite close to the nominal level 0.95. For PCA, the prediction interval performs well in setting (E1) with large $d$ and $T$, which is consistent with the theoretical results of Bai and Ng (2006) and Bai and Ng (2023). However, in setting (E2), where $\mathcal{E}_t$ exhibits cross-sectional correlation, PCA is outperformed by CP under both strong and weak factors, with CP's coverage rats remaining closer to 0.95. Under strong factors, however, the coverage rates of PCA approach 0.95 with increasing $d_k$. That is because the second component of $\widehat{\sigma}^2_{y_{T+h|T},pca}$ in Equation (22) is of order $1/d$. As a result, the bias of $\widehat{\mathrm{Avar}}(\widehat{f}_T)$ induced by the HAC estimator becomes negligible as $d$ increases.

## 5.4 Multi-Source Factor-Augmented Sparse Regression

In this section, we evaluate the performance of MS-FASR studied in Section 3.2. We consider the following DGP for $y_{t+h}$ and $w_t \in \mathbb{R}^p$:

$$\begin{aligned} y_{t+1} &= \beta_0^\top w_t + \beta_1^\top z_t + \epsilon_{t+1} \\ w_t &= \Lambda z_t + V_t, \end{aligned}$$

where $z_t = \left(1, f_t^\top\right) \in \mathbb{R}^{r+1}$. We set the dimension of predictors $p$ to be 200, with the first three elements of $\beta_0$ to 0.5 and the remaining elements set to 0. To evaluate the convergence rate in Theorem 4.1, we choose $T$ such that the value of $p_0\left(\sqrt{\log(p)/T}\right)$ takes uniform grids in

| Error Setting | E1 | | | | E2 | | | |
|---|---|---|---|---|---|---|---|---|
| Strength | Strong | | Weak | | Strong | | Weak | |
| $d_k$ | CP | PCA | CP | PCA | CP | PCA | CP | PCA |
| $\sigma^2_{\epsilon,t} = 1$ | | | | | | | | |
| $d_k = 20$ | 0.95 | 0.954 | 0.949 | 0.946 | 0.948 | 0.926 | 0.915 | 0.835 |
| $d_k = 40$ | 0.959 | 0.959 | 0.961 | 0.953 | 0.959 | 0.959 | 0.961 | 0.953 |
| $d_k = 60$ | 0.926 | 0.925 | 0.941 | 0.94 | 0.936 | 0.925 | 0.933 | 0.821 |
| $\sigma^2_{\epsilon,t} \sim U(0.5, 1.5)$ | | | | | | | | |
| $d_k = 20$ | 0.953 | 0.956 | 0.952 | 0.944 | 0.946 | 0.925 | 0.923 | 0.822 |
| $d_k = 40$ | 0.942 | 0.942 | 0.954 | 0.945 | 0.953 | 0.928 | 0.917 | 0.711 |
| $d_k = 60$ | 0.95 | 0.952 | 0.947 | 0.944 | 0.948 | 0.941 | 0.938 | 0.832 |

Table 1: Coverage rate of CP and PCA prediction intervals over 2000 repetitions. Confidence level is 0.95.

$[0.15, 0.5]$. The entry of $\Lambda$ is generated from a uniform distribution $U[-1, 1]$, while the entries of $V_t$ are generated independently from $N(0, 1)$. We set $d_1 = d_2 = 40$. The idiosyncratic errors $\epsilon_{t+h}$ follow setting $(\epsilon 1)$ in Section 5.3 and $\mathcal{E}_t$ is generated according to setting (E1).

The estimation error of $\beta_0$ and the prediction error of $y_{T+h|T}$ are shown in Figure 3. The results further support the theoretical findings in Section 4, which establish that the statistical rates are approximately $p_0\,(\log(p)/T)$.

# 6 Empirical Application

Understanding trade flow patterns and forecasting their dynamics are essential for policy-making, firm optimization, and risk management. Trade data inherently form a dynamic sequence of tensor variates, which can capture network-like structures, underlie common dynamics, and reveal intricate interaction patterns. In this section, we consider a diffusion index forecast based on the CP tensor factor model for international trade data, providing a unified framework to estimate global trade factors and predict future variations in U.S. trade.

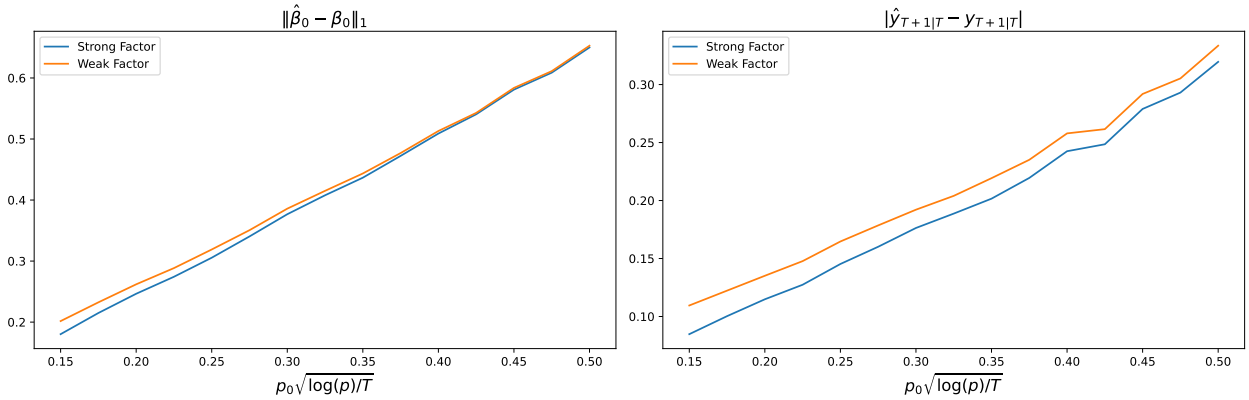



Figure 3: Estimation error of $\beta_0$ and prediction error of $y_{T+h|T}$ over 1000 repetitions under strong and weak factor setting.

## 6.1 Data and sample

We analyze monthly bilateral import and export volumes of commodity goods among 24 countries and regions from January 1999 to December 2018, using data from the International Monetary Fund Direction of Trade Statistics (IMF-DOTS). The countries and regions included in the dataset are: 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).

In our study, we employ the diffusion index forecast model with a CP low rank structure, as defined in (1) and (2). Specifically, we represent the trade data as a $24 \times 24$ two-dimensional tensor, where each element $x_{i,j,t}$ denotes the monthly variation of exports from country $i$ to country $j$ at month $t$. For simplicity, self-exports are set to zero, i.e., $x_{i,i,t} = 0$ for all $i$ and $t$. The target variables for our analysis are the monthly variation of U.S. aggregate export and import to/from in-sample countries, denoted by $y_t^{ex}$ and $y_t^{im}$, respectively.

The number of common factors is determined using the eigen ratio-based method proposed by Ahn and Horenstein (2013) and Chen et al. (2024a), which identifies four common factor explaining 51.1% of the total variance. Let $f_t$ denote the common factor extracted from the growth rate of bilateral trade. We then construct one-month-ahead forecasts for yearly

| | Const | $y_t^{(ex)}$ | $y_t^{(im)}$ | $\widehat{f}_{1t}$ | $\widehat{f}_{2t}$ | $\widehat{f}_{3t}$ | $\widehat{f}_{4t}$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| $y_{t+1}^{(ex)}$ | | | | | | | | |
| (a) | 294.066 | **-0.325** | -0.069 | | | | | 0.163 |
| | (0.97) | (-3.48) | (-1.06) | | | | | |
| (b) | 384.766 | | | **0.726** | **-0.234** | **0.422** | -0.125 | 0.289 |
| | (1.363) | | | (6.73) | (-2.13) | (2.99) | (-0.74) | |
| (c) | 365.062 | **-0.989** | **0.337** | **1.118** | **0.347** | **0.6** | **1.179** | 0.402 |
| | (1.4) | (-5.91) | (3.57) | (7.13) | (2.55) | (4.39) | (3.0) | |
| $y_{t+1}^{(im)}$ | | | | | | | | |
| (a) | 586.208 | -0.056 | **-0.307** | | | | | 0.114 |
| | (1.31) | (-0.41) | (-3.21) | | | | | |
| (b) | 613.43 | | | **0.906** | 0.243 | **0.86** | -0.457 | 0.211 |
| | (1.45) | | | (5.61) | (1.47) | (4.07) | (-1.81) | |
| (c) | 597.329 | **-1.376** | -0.069 | **0.759** | **0.941** | **1.21** | **2.4** | 0.301 |
| | (1.49) | (-5.35) | (-0.47) | (3.15) | (4.5) | (5.76) | (3.97) | |

*Note:* The table reports results from im-sample diffusion index model of monthly variation in US total export and import to countries in the dataset on lagged variables named in the first row. $\widehat{f}_{it}$ is the $i$-th common factor extracted from the bilateral trade flow tensor data. The t-values are reported in parentheses. Coeefficients that are statistically significant at the 5% level are in bold.

growth of U.S. aggregate exports and imports using the following regression:

$$\begin{aligned} y_{t+1}^{(ex)} &= \beta_{00}^{(ex)} + \beta_{01}^{(ex)} y_t^{(ex)} + \beta_{02}^{(ex)} y_t^{(im)} + \beta_1^{(ex)\top} f_t + \epsilon_{t+1}^{(ex)}, \\ y_{t+1}^{(im)} &= \beta_{00}^{(im)} + \beta_{01}^{(im)} y_t^{(ex)} + \beta_{02}^{(im)} y_t^{(im)} + \beta_1^{(im)\top} f_t + \epsilon_{t+1}^{(im)}. \end{aligned} \tag{23}$$

## 6.2 In-sample analysis

Table 6.2 reports the in-sample forecast results based on Equation (23). As a benchmark, regression (a) predicts each target variable using only the first lag of changes in US exports and imports. In contrast, regression (b) demonstrates that incorporating common factors extracted from the tensor data significantly increases predictive power compared to using lagged values alone. Specifically, the common factors explain 29% the variation in monthly export changes and 21% of the variation in import changes. Regression (c) integrates both the lagged target variables and the common factors, leading to a substantial improvement in explanatory power, with R-squared values increasing to 40% for U.S. exports and 30% for imports. Moreover, all four common factors are statistically significant predictors for both

trade flows. These findings highlight the crucial role of common factors derived from the tensor data in enhancing the accuracy of monthly U.S. export and import forecasts.

Since the factors in the CP tensor model are identified only up to sign changes, it is meaningful to explore their economic interpretation. To characterize these factors, we examine their correlations with monthly variations in bilateral trade flows among the selected countries. These correlations are visualized in the heatmap shown in Figure 4, where stronger correlations between a factor and bilateral trade flows are indicated by deeper blue shades.

The heatmap reveals distinct regional patterns for each factor. Factor 1 is closely associated with exports from Asian countries to the rest of the dataset, with the highest correlation observed in exports from CN. Factor 2 is highly correlated with China's imports from most countries in the dataset and also shows notable correlations with trade flows among key Asian economies, including CN, KR, JP, and SG. Factor 3 is mainly correlated with bilateral trade flows among European countries, while Factor 4 predominantly captures trade flows within North America, specifically among US, CA and MX. In summary, Factors 1 and 2 contain information on trade flows within Asia, particularly involving China. Factor 3 relates to trade dynamics within Europe, and Factor 4 captures variations in trade among North American countries. The in-sample analysis in Section 6.2 demonstrates that these factors are not only economically interpretable but also provide significant predictive power for monthly variations in U.S. exports and imports.

## 6.3 Out-of-sample analysis

In the section, we evaluate the out-of-sample performance of the diffusion index model (23) based on the CP low rank structure and compare it with alternative methods, in particular the vector factor model studied by Bai and Ng (2006). In addition to Model (23), we incorporate 126 macroeconomic variables from FRED-MD (McCracken and Ng, 2016) and up to 12 lags of U.S. aggregate exports and imports. This allows us to access the performance of MS-FASR, introduced in Section 4 and investigate whether including U.S. macroeconomic variables and additional lags of the target variables improves the out-of-sample forecast of U.S. aggregate exports and imports. The MS-FASR model is specified as follows:

$$\begin{aligned} y_{t+1}^{(ex)} &= \beta_{00}^{(ex)} + \beta_{01}^{(ex)} y_t^{(ex)} + \beta_{02}^{(ex)} y_t^{(im)} + \beta_1^{(ex)\top} f_t + \beta_2^{(ex)\top} w_t + \epsilon_{t+1}^{(ex)}, \\ y_{t+1}^{(im)} &= \beta_{00}^{(im)} + \beta_{01}^{(im)} y_t^{(ex)} + \beta_{02}^{(im)} y_t^{(im)} + \beta_1^{(im)\top} f_t + \beta_2^{(im)\top} w_t + \epsilon_{t+1}^{(im)}, \end{aligned} \tag{24}$$

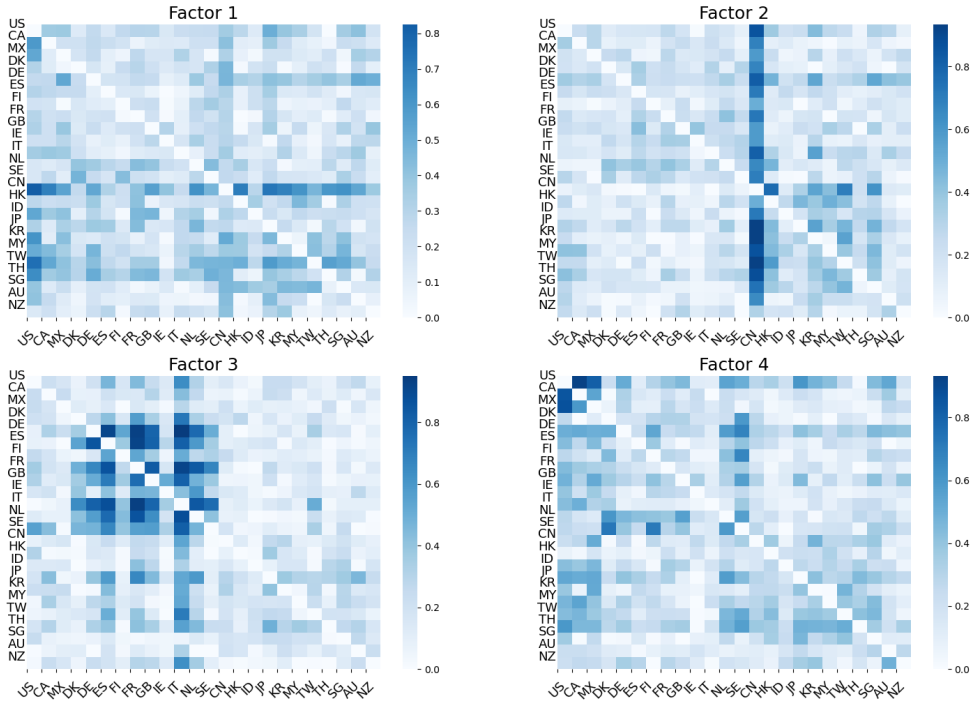



Figure 4: Heatmap of the absolute value of correlation between the common factors and the monthly variation in bilateral tradeflow among selected countries. The saturation represents the correlation strength, with high saturation indicating stronger correlation.

where $w_t \in \mathbb{R}^{148}$ includes 126 macroeconomic variables and lagged U.S. aggregate exports and imports from lag 2 to lag 12.

The out-of-sample analysis follows an expanding-window approach, where model parameters are re-estimated as new data become available. The process begins with an initial five-year sample from December 1999 to December 2004. Factors and parameters are estimated using data from December 1999 to November 2004,[6] and the model is then used to forecast monthly variations in U.S. aggregate exports and imports for December 2004. This procedure is repeated iteratively until the end of the sample, resulting in a total of 169 monthly forecasts from December 2004 to December 2018. We compare the performance of Model (23) and Model (24) against various alternative methods:

[6]The predictor variables span December 1999 to October 2004, while the target variables cover January 2000 to November 2004, forming a five-year training sample.

- Benchmark: Predicts the target variable using only the first lag of U.S. exports and imports along with a constant;
- DI(CP): Model (23) with factors estimated by CC-ISO;
- DI(PCA): Model (23) with factors estimated via PCA on vec $(\mathcal{X}_t)$;
- MS-FASR(CP): Model (24) with factors estimated by CC-ISO;
- MS-FASR(PCA): Model (24) with factors estimated via PCA on vec $(\mathcal{X}_t)$;
- DI(CP) + DI(w): $y_{t+1}^{(\cdot)} = \beta_{00}^{(\cdot)} + \beta_{01}^{(\cdot)} y_t^{(ex)} + \beta_{02}^{(\cdot)} y_t^{(im)} + \beta_1^{(\cdot)\top} f_t + \beta_2^{(\cdot)\top} f_t^{(w)} + \epsilon_{t+1}^{(\cdot)}$, where $f_t^{(w)} \in \mathbb{R}^{r_w}$ consists of factors extracted from $w_t$ via PCA and $f_t$ is estimated using CC-ISO;
- DI(PCA) + DI(w): Same as DI(CP) + DI(w) but with $f_t$ estimated via PCA on vec $(\mathcal{X}_t)$;
- LASSO(w): $y_{t+1}^{(\cdot)} = \beta_{00}^{(\cdot)} + \beta_{01}^{(\cdot)} y_t^{(ex)} + \beta_{02}^{(\cdot)} y_t^{(im)} + \beta_2^{(\cdot)\top} w_t + \epsilon_{t+1}^{(ex)}$, where $\beta_2$ is estimated with an $\ell_1$-norm constraint.

Table 2 presents the Mean square error(MSE) ratios of the one-month-ahead out-of-sample forecast for each model relative to the benchmark. It also presents p-values from the forecast comparison tests of Diebold and Mariano (1995). These tests are one-sided, with the following alternatives:

- DM(Benchmark): Competing methods outperform the benchmark model.
- DM(I): DI(CP) is more accurate than DI(PCA).
- DM(II): MS-FASR(CP) is more accurate than competing methods.

Our findings indicate that, for both exports and imports, MS-FASR(CP) achieves the lowest MSE, significantly outperforming LASSO(w). This reinforces our in-sample results, highlighting that common factors extracted from tensor data are valuable for predicting U.S. export and import variations. Furthermore, MS-FASR(CP) outperforms both DI(CP) and DI(PCA), suggesting that macroeconomic variables provide additional predictive power. Finally, MS-FASR(CP) also outperforms DI(CP) + DI(w) and DI(PCA) + DI(w), which attempt to incorporate factors from multiple sources. This result provides strong empirical support for the combination of the CP tensor model with sparse regression.

Table 2: MSE ratios of out-of-sample forecasts

| | DI(CP) | DI(PCA) | MS-FASR(CP) | MS-FASR(PCA) | DI(CP) + DI(w) | DI(PCA) + DI(w) | LASSO(w) |
|---|---|---|---|---|---|---|---|
| Export | | | | | | | |
| MSE Ratio | 0.7722 | 0.8207 | 0.4271 | 0.4345 | 0.8598 | 0.9205 | 0.7207 |
| DM(Benchmark) | 0.0104 | 0.0371 | $< 0.001$ | $< 0.001$ | 0.1223 | 0.2574 | 0.0059 |
| DM(I) | - | 0.0547 | - | - | - | - | - |
| DM(II) | $< 0.001$ | $< 0.001$ | - | 0.3269 | $< 0.001$ | $< 0.001$ | $< 0.001$ |
| Import | | | | | | | |
| MSE Ratio | 0.8223 | 0.8965 | 0.4937 | 0.4969 | 1.0065 | 1.0526 | 0.7064 |
| DM(Benchmark) | 0.0016 | 0.028 | $< 0.001$ | $< 0.001$ | 0.4796 | 0.286 | 0.0007 |
| DM(I) | - | 0.003 | - | - | - | - | - |
| DM(II) | $< 0.001$ | $< 0.001$ | - | 0.4447 | $< 0.001$ | $< 0.001$ | 0.004 |

*Note:* The table reports the out-of-sample MSE ratios relative to the benchmark model, which only includes $y_t = (y_t^{(ex)}, y_t^{(im)})^\top$ and a constant. The first panel shows the results for U.S. export forecasts, while the second panel shows results for U.S. import forecast. Column names correspond to methods specified in the main text. The MSE Ratio row shows the ratio of each method's MSE to that of the benchmark model. DM(Benchmark) reports p-values from Diebold and Mariano (1995) test with alternative being that competing method is more accurate than the benchmark model. DM(I) reports the p-values from Diebold and Mariano (1995) test with alternative being that DI(CP) outperforms DI(PCA). DM(II) reports the p-values from Diebold and Mariano (1995) test with alternative being that MS-FASR(CP) outperforms the competing method. The number of factors is selected by unfolded eigenvalue ratio method by Chen et al. (2024a). MS-FASR that include $w_t$ are estimated as in Section 4. The tuning parameter $\lambda$ is selectd by 5-fold cross-validation.

# 7 Conclusion

Factor models are effective for extracting meaningful information from a large number of variables. These information, in turn, can be used to predict variables of interest. In this paper, we consider the case when the data is in the form of tensor that can be represented by CP decomposition. We provide the inferential theories for the estimation of factors and the predictive intervals of the diffusion index forecasting model. We show that the least squares estimates obtained from the predictive regressions are $\sqrt{T}$-consistent and asymptotically normal, allowing for the existence of weak factors. The conditional mean is consistent and asymptotically normal with the rate determined by $T$ and the strength of the weakest factor. The construction of prediction interval based on this result requires a consistent estimator of the covariance matrix of the factor. We introduce such an estimator, which is robust to cross-sectional correlations in the idiosyncratic errors, and prove its consistency.

# References


Adamek, R., Smeekes, S., and Wilms, I. (2023). Lasso inference for high-dimensional time series. *Journal of econometrics*, 235(2):1114–1143.

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

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

Bai, J. and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*, 235(2):1893–1916.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Chen, B., Han, Y., and Yu, Q. (2024a). Estimation and inference for cp factor model. *working paper*.

Chen, E. Y., Fan, J., and Zhu, X. (2024b). Factor augmented matrix regression. Working paper, Princeton University.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of statistics*, 39(6):3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 75(4):603–680.

Fan, J., Lou, Z., and Yu, M. (2024). Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association*, 119(546):1076–1088.

Gao, Z. and Tsay, R. S. (2024). Supervised dynamic pca: Linear dynamic forecasting with many predictors. *Journal of the American Statistical Association*, pages 1–15.

Han, Y., Zhang, C.-H., and Chen, R. (2023). Cp factor model for dynamic tensors. *working paper*.

Huang, D., Jiang, F., Li, K., Tong, G., and Zhou, G. (2022). Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3):1177–1216.

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

Kruskal, J. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18:95–138.

Kruskal, J. (1989). Rank, decomposition, and uniqueness for 3-way and n-way arrays. *Multiway Data Analysis*, pages 7–18.

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.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of statistics*, 40(3):1637–1664.

Ludvigson, S. C. and Ng, S. (2007). The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222.

Ludvigson, S. C. and Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067.

McCracken, M. W. and Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Sidiropoulos, N. and Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14:229–239.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

# Appendix A: An illustrative example

To illustrate the invalidity of the assumption in Equation 16, consider the strong matrix factor model with one factor where $\widetilde{A} := \sqrt{d}A$ is a $d$-dimensional vector of ones. In this case, $\Sigma_{Be} = \lim \frac{1}{d} \sum_{j=1}^{d} \sum_{l=1}^{d} \mathbb{E}\left[e_{jt}e_{lt}\right]$. Suppose the idiosyncratic error matrix is generated by:

$$\mathcal{E}_t = \Sigma_{\mathcal{E},1}^{1/2} Z_t \Sigma_{\mathcal{E},2}^{1/2}, \quad Z_t \sim MN(0, I_{d_1}, I_{d_2}).$$

Let $d_1 = d_2$ and $\Sigma_{\mathcal{E},1} = \Sigma_{\mathcal{E},2} = Toeplitz(\tau, d_1)$ such that the $(i,j)^{th}$ entry of $\Sigma_{\mathcal{E},k}$ is equal to $\tau^{|i-j|}$. It can be verified that $\Sigma_e = \Sigma_{\mathcal{E},2} \odot \Sigma_{\mathcal{E},1}$ and for $q = |i-j|$, $\mathbb{E}\left[e_{it}e_{jt}\right] = \gamma_q$, where $\gamma_q = \tau^q$ for $1 \leq q \leq d_1$, $\gamma_q = \tau\gamma_{q-d_1}$ for $d_1 + 1 \leq q \leq 2d_1$, $\gamma_q = \tau^2\gamma_{q-2d_1}$ for $2d_1 + 1 \leq q \leq 3d_1$, and so on. The plot of $\gamma_q$ for $d_1 = 10$ and $\tau = 0.5$ is shown in Figure 7 for illustration. Due to the kronecker product structure of $\Sigma_e$, $\gamma_q$ does not decay monotonically. Let $n = \sqrt{d} = d_1 \to \infty$, then $\lim \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \mathbb{E}\left[e_{jt}e_{lt}\right] = 1 + \lim_{d_1 \to \infty} \sum_{q=1}^{d_1-1} 2\frac{d_1-q}{d_1}\gamma_q$, which is the Newey-West sum of $\gamma_q$ before the second peak in Figure 7. This estimator is not consistent for $\Sigma_{Be}$ as the sum of $\gamma_q$ for $q > d_1$ does not converge to zero. Therefore, CS-HAC estimator proposed by Bai (2003) is not suitable for the CP factor model.

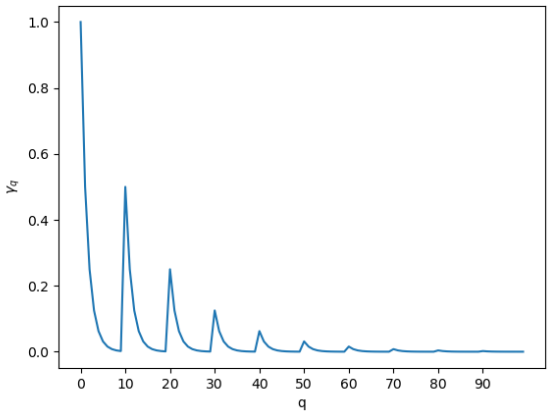



Figure 5: Plot of $\gamma_q$ for $d_k = 10$ and $\tau = 0.5$.

A simulation study is conducted to verify this observation. Consider the following two-way

CP factor model:

$$\begin{aligned}
&\mathcal{X}_t = \sum_{i=1}^{r} s_i f_{it} a_{i1} a_{i2}^\top + \mathcal{E}_t, \\
&f_{it} = \rho_i f_{i,t-1} + \sqrt{1-\rho_i^2} u_{it}, \quad u_{it} \sim N(0,1), \\
&\mathcal{E}_t = \Sigma_{\mathcal{E},1}^{1/2} Z_t \Sigma_{\mathcal{E},2}^{1/2}, \quad Z_t \sim MN(0, I_{d_1}, I_{d_2}) \\
&A_k = [a_{1k}, \ldots, a_{rk}] = \Sigma_{A_k} \widetilde{A}_k,
\end{aligned}$$

The matrix $\widetilde{A}_k$ is generated by QR decomposition of the matrix of $d_k \times r$ where each entry is generated from $N(0,1)$ so that $\widetilde{A}_k$ is orthonormal. I consider the following specifications:

- $d_1 = d_2$, $r = 3$, $T = 1000$;
- $s_i = (r - i + 1)\ \sqrt{d}$;
- $\rho_1, \rho_2, \rho_3 = 0.6, 0.5, 0.4$;
- $\Sigma_{\mathcal{E}_k} = \Sigma_{A_k} = Toeplitz(0.6, d_k)$.

The model is estimated by PCA method on $\textsc{Vec}\left(\mathcal{X}_t\right)$ and I consider three covariance matrices for factor estimator:

- $\widehat{\Gamma}^{HAC} = \widehat{V}^{-1} \left( \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \widehat{A}_{j:} \widehat{A}_{l:}^\top \frac{1}{T} \sum_{t=1}^{T} \widehat{e}_{jt} \widehat{e}_{lt} \right) \widehat{V}^{-1}$;
- $\Gamma^{HAC} = \widehat{V}^{-1} Q \left( \frac{1}{n} \sum_{j=1}^{n} \sum_{l=1}^{n} A_{j:} A_{l:}^\top \frac{1}{T} \sum_{t=1}^{T} e_{jt} e_{lt} \right) Q^\top \widehat{V}^{-1}$;
- $\Gamma = \widehat{V}^{-1} Q \left( A^\top \Sigma_e A / d \right) Q^\top \widehat{V}^{-1}$,

where

- $\Sigma_e = \Sigma_{\mathcal{E}_2} \odot \Sigma_{\mathcal{E}_1}$ is the true covariance matrix of the idiosyncratic error;
- $Q = \widehat{F}^\top F / T$;
- $\widehat{V}$ is the diagonal matrix of the first $r$ eigenvalues of $\frac{1}{dT} \sum_{t=1}^{T} \textsc{Vec}\left(\mathcal{X}_t\right) \textsc{Vec}\left(\mathcal{X}_t\right)^\top$;
- $\widehat{A}_{j:}$ is the $j^{th}$ row of $\widehat{A}$, which is the factor loading estimator by PCA;
- $A_{j:}$ is the $j^{th}$ row of $A$, which is the true factor loading;
- $n = \sqrt{\min\{d, T\}}$.

$\Gamma$ can be considered as the "oracle" estimator of the factor covariance matrix and $\Gamma^{HAC}$ as the "oracle" HAC estimator, where the true factor loadings and errors are used. Figure 6

shows the histogram of the first entry of $\sqrt{d}\ \widehat{\Sigma}_{Be}^{-1/2}\left(\widehat{f}_t - f_t\right)$ for $t = 0$ with choices of $\widehat{\Sigma}_{Be}$ specified above. The sample standard deviation of the histograms are shown at the top right corner of each plot. It can be observed that the HAC estimator as well as the "oracle" version is not consistent for the factor covariance matrix, which is in line with the analysis above.

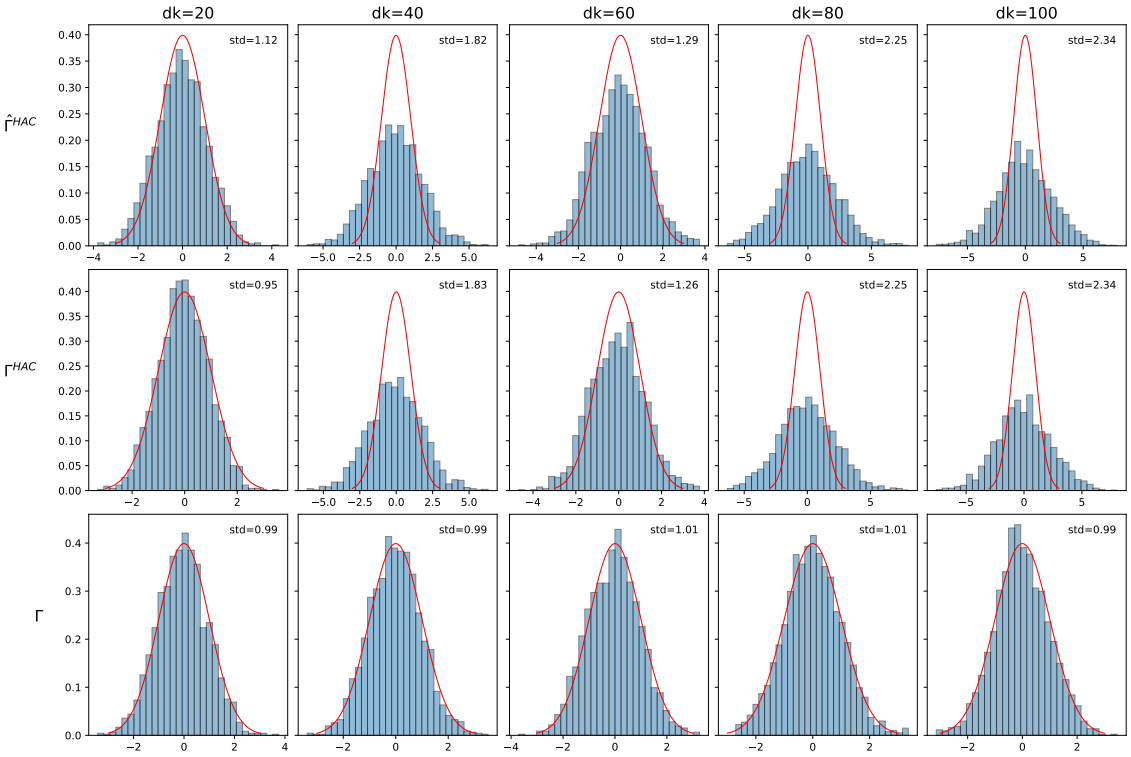



Figure 6: The histogram of the first entry of $\sqrt{d}\ \Sigma_{Be}^{-1/2}\left(\widehat{f}_t - f_t\right)$ for $t = 0$. The first row shows the results for $\widehat{\Sigma}_{Be} = \widehat{\Gamma}^{HAC}$; the second row shows the results for $\widehat{\Sigma}_{Be} = \Gamma^{HAC}$; the third row shows the results for $\widehat{\Sigma}_{Be} = \Gamma$. Columns from left to right show the results for $d = 20, 40, 60, 80, 100$. The sample standard deviation of the histograms are shown at the top right corner of each plot.

# Appendix B: Consistency of the factor estimators

In this appendix, I show that CC-ISO requires a weaker condition for consistency of the factor estimators than PCA. Consider the model for data generating process in Appendix C, the specifications of the model are as follows:

- $d_1 = d_2$, $r = 3$ and $T = 100 + d^{0.3}$;
- Factor loadings are generated as in Section **??**;
- $\rho_1 = 0.6, \rho_2 = 0.5, \rho_3 = 0.4$;
- $\Sigma_{\mathcal{E}_k} = Toeplitz(0.5, d_k)$;
- $s_i = (r - i + 1)\sqrt{d^{\alpha}}$, where $\alpha \in \{0.6, 0.5, 0.4\}$.

Under this data genrating process, we have $\frac{1}{2} < \alpha + 0.3 < 1$. As discussed in Remark 3.1, CC-ISO is consistent but PCA is not in theory. Figure 7 shows the estimation error of factors $\frac{1}{\sqrt{T}}\|\widehat{F} - HF\|_2$ for PCA and CC-ISO. It can be observed that PCA is not consistent for the factor estimation, while CC-ISO is consistent, which is in line with the theoretical results.

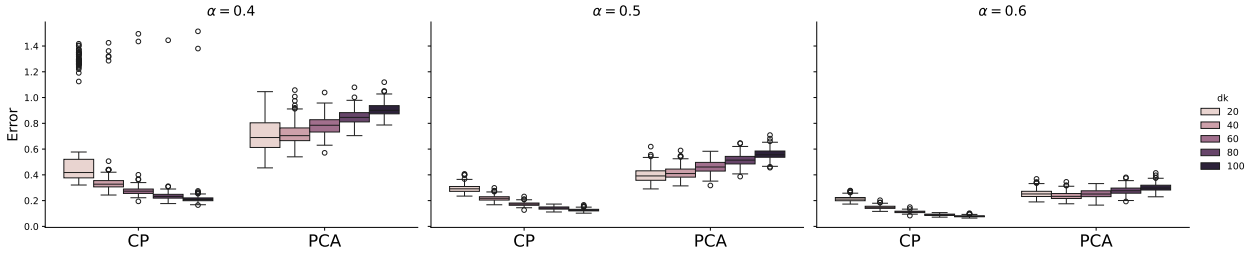



Figure 7: Factor estimation error for PCA and CC-ISO for different choices of $\alpha$ and $d_k$.