Tensor PCA for Factor Models*

Andrii Babii† Eric Ghysels‡ Junsu Pan§

UNC Chapel Hill *UNC Chapel Hill* *UNC Chapel Hill*

February 27, 2025

Abstract

Modern empirical analysis often relies on high-dimensional panel datasets with non-negligible cross-sectional and time-series correlations. Factor models are natural for capturing such dependencies. A tensor factor model describes the d-dimensional panel as a sum of a reduced rank component and an idiosyncratic noise, generalizing traditional factor models for two-dimensional panels. We consider a tensor factor model corresponding to the notion of a reduced multilinear rank of a tensor. We show that for a strong factor model, a simple tensor principal component analysis algorithm is optimal for estimating factors and loadings. When the factors are weak, the convergence rate of simple TPCA can be improved with alternating least-squares iterations. We also provide inferential results for factors and loadings and propose the first test to select the number of factors. The new tools are applied to the problem of imputing missing values in a multidimensional panel of firm characteristics.

Keywords: Multidimensional panel data, tensors, cross-sectional dependence, dynamic networks, spatial data, factor models, principal component analysis, asset pricing, imputation, firm characteristics.

*We would like to thank Alex Belloni, Peter Bossaerts, George Kapetanios, Oliver Linton, Alexei Onatski, Mirco Rubin, Giovanni Urga, as well as seminar participants at the Bayes Business School, University of Cambridge, and Triangle Econometrics Conference for helpful comments.

†Department of Economics, University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: babii.andrii@gmail.com.

‡Department of Economics, University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305 and Department of Finance, Kenan-Flagler Business School, Email: eghysels@gmail.com.

§Department of Economics, University of North Carolina at Chapel Hill - Gardner Hall, CB 3305 Chapel Hill, NC 27599-3305. Email: junsupan@live.unc.edu.

1 Introduction

Modern empirical analysis often relies on high-dimensional panel datasets with non-negligible dependencies across one or several dimensions. Various economic reasons exist for such dependencies, including common shocks in macroeconomics and finance, interactions in a network, or spatial distribution of economic activity. Since their original introduction in the psychology literature by Spearman (1904), factor models have become a ubiquitous tool in economics for modeling dependencies in two-dimensional panel datasets.

Among the many applications of factor analysis in macroeconomics and finance, we may cite asset pricing, see Ross (1976) and Chamberlain and Rothschild (1982); business cycle analysis Sargent, Sims, et al. (1977); industrial production analysis, see Foerster, Sarte, and Watson (2011) and Andreou, Gagliardini, Ghysels, and Rubin (2019); forecasting with big data; see Stock and Watson (2002b). Factor models are also widely used for causal inference in panel data; see Pesaran (2006), Bai (2009), Abadie, Diamond, and Hainmueller (2010), Gobillon and Magnac (2016), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021), or more generally to model unobserved heterogeneity in microeconometrics; see Cunha, Heckman, and Schennach (2010) and Bonhomme and Robin (2010). Lastly, structural economic models often naturally lead to a factor structure as seen in Liu and Tsyvinski (2024) for the input-output production networks or Lewbel (1991) for the demand systems.

Traditional factor models are designed for two-dimensional panel datasets consisting of cross-sectional units tracked over time, where both dimensions can potentially be large; see Stock and Watson (2002a) and Bai (2003).[1] However, many economic data feature more than two dimensions. For example, the input-output production or trade network usually evolves over time, which is not captured by the traditional factor model.[2] A panel of macroeconomic indicators often involves regional aggregation of state- or county-level observations, introducing a geographical dimension in addition to the traditional cross-section and time series. Asset pricing models for the cross-section of equities typically involve characteristic-based portfolio sorts, introducing a third dimension. The sorting into deciles is common, but only the lowest and highest deciles are used and combined in a high minus low return spread. Going beyond national borders introduces an international dimension to macro and financial datasets, resulting in a four-dimensional data structure. Each of these examples illustrates that to obtain matrix representations of the data, we often aggregate the multidimensional panel datasets, suppressing more granular information.

[1] We will use the term 'traditional' for 2-dimensional factor models applicable to 2-dimensional panel datasets.

[2] See, e.g. Graham (2020) for a recent review of static networks.

Principal component analysis (PCA) is a commonly used method for identifying and estimating traditional factor models for two-dimensional panel data sets; see Pearson (1901). PCA extracts latent factors and their loadings using either the singular value decomposition (SVD) of the original panel dataset collected in a matrix or, equivalently, the eigendecomposition of the associated sample covariance matrices; see Jolliffe (2002) for a review of PCA and factor analysis.

A d-way tensor is a d-dimensional array generalizing vectors and matrices introduced in Ricci and Levi-Civita (1900).[3] In this paper, we consider an extension of traditional factor models to multidimensional datasets, called tensor factor models. Similarly to their 2-way counterpart, the d-way tensor factor model can be used to identify the latent factors driving correlations in tensor datasets. In traditional factor models, correlations are captured with a matrix of factors in the time dimension and a matrix of factor loadings in the cross-sectional dimension. Likewise, one can decompose a tensor into a collection of matrices with respect to each dimension. Hence, the matrix component in the 'time' dimension correspond to factors that vary over time, and the matrix component in the other dimensions correspond to loadings determining the heterogeneous exposure of each dimension to the latent factors.

Traditional factor models can also be viewed as decomposing a matrix representing a panel dataset as a sum of a low-rank matrix product of loading and factor matrices and a matrix of idiosyncratic shocks. The low-rank component is then approximated using the truncated SVD decomposition of the observed data. The d-way factor model also describes a d-way tensor dataset as a sum of a low-rank component and an idiosyncratic shocks tensor. However, there several different ways to describe the rank of a tensor. The two most widely known definitions of tensor ranks are called the Canonical Polyadic (CP) rank and the *multilinear rank*; see Carroll and Chang (1970), Harshman (1970), and Tucker (1966).

The multilinear rank of a d-dimensional tensor is described by the d-tuple $(R_1, R_2 \ldots, R_d)$ corresponding to ranks of each of its mode-j matricizations and is more general than than the CP rank.[4] It leads to the so-called Tucker model that has been previously considered in statistics literature; see Han, Chen, Yang, and Zhang (2025), Wang, Zheng, Lian, and Li (2022), Chen, Yang, and Zhang (2022), Han, Chen, and Zhang (2022b), and Zhang and Xia (2018) among others. Our paper introduces the factor dynamics differently in contrast to these papers. More importantly, we show that under the strong factor model assumption, commonly used in economics and finance, the simple PCA-type estimators for tensors have

[3]Tensor tools have already been used by economists, e.g., for the identification of finite mixture models; see Bonhomme, Jochmans, and Robin (2016).

[4]The concepts of multilinear rank and CP rank for tensors are implicitly used in Hitchcock (1927). In fact, Hitchcock (1927) considered even more general concept of a tensor rank based on arbitrary matricizations, not just along one of its d modes.

optimal convergence rates and derive the corresponding asymptotic distributions.

The simple tensor PCA (TPCA) estimators are easy to compute and consists of reshaping (or matricizing) a d-way tensor into d matrices along each of its d dimensions and applying the standard PCA these matrices. For this simple TPCA estimators, we (a) demonstrate that they can identify and consistently estimate the factors and loadings; (b) describe the associated convergence rates and asymptotic distribution; (c) show that tensor dimensions can improve the estimation accuracy for factors/loadings and the estimator is rate-optimal for strong factor models; and (d) develop a formal test for the number of factors in the tensor factor model. For the moderately weak factor model, the TPCA procedure is not optimal and we also consider additional results for the optimal iterative alternating least-squares (ALS) algorithm, where the TPCA estimators are used as a starting value. Our theoretical results rely on the powerful perturbation theory results for singular subspaces and singular values which are not commonly used in econometrics, cf. Bai and Ng (2023) and references therein.

Monte Carlo simulations support our asymptotic results in finite samples. We find that the d-way tensor factor model reduces dimensions more efficiently than the naively pooled 2-way factor model. We also verify the convergence rates and the distribution theory.

In the empirical application we study the issue of missing firm characteristics in widely used data sources which are the cornerstone of research in finance. Bryzgalova, Lerner, Lettau, and Pelger (2024) provide a comprehensive analysis of missing data in firm characteristics, and propose a statistical model for imputing missing values and investigate the impact of missingness on asset returns. Firm characteristic data is a 3-dimension tensor, with (a) firm, (b) characteristic and (c) time as the three dimensions. The statistical methods proposed by Bryzgalova et al. (2024) do not exploit the tensor data structure, while in this empirical application we do and we show that using our proposed tensor data models improve on their time series of cross-sectional models.

The paper is organized as follows. Section 2 introduces a new class of tensor factor models. The convergence rates and large sample distributions of the TPCA estimator are covered in Section 3 which also covers the improved ALS estimator. Section 4 presents a novel testing procedure for the number of factors. Small sample simulation evidence is reported in Section 5. An illustrative empirical example appears in Section 6. Section 7 concludes. Lastly, in the Appendix, we provide proofs for all the main and auxiliary results.

Notation: For two matrices $A \in \mathbb{R}^{N_1 \times R_1}$ and $B \in \mathbb{R}^{N_2 \times R_2}$, we use $A \otimes B \in \mathbb{R}^{N_1 N_2 \times R_1 R_2}$ to denote their Kronecker product. In addition, for a collection of matrices $(\Lambda_j)_{j=1}^d$, we define $\bigotimes_{k \neq j} \Lambda_k = \Lambda_d \otimes \cdots \otimes \Lambda_{j+1} \otimes \Lambda_{j-1} \otimes \cdots \otimes \Lambda_1$. For a positive integer p, we use

I_p to denote the $p \times p$ identity matrix. For a matrix A, we use $P_A = A(A^\top A)^\dagger A^\top$ to denote the the projection matrix on the column space of A, where † denotes the generalized inverse. For two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \lesssim b_n$ if and only if there exists $C < \infty$ such that $a_n \leqslant Cb_n$ for all $n \in \mathbb{N}$. The operator norm of a matrix A is defined as $\|A\|_{\text{op}} = \sup_{\|x\|=1} \|Ax\|$, where $\|.\|$ is the Euclidean norm. More generally, we use $\|.\|_p$ to denote the ℓ_p norm. For two tensors $A, B \in \mathbb{R}^{N_1\times\cdots\times N_d}$, the Frobenius inner product is defined as $\langle A, B\rangle_F = \sum_{i_1,\ldots,i_d} A_{i_1,\ldots,i_d} B_{i_1,\ldots,i_d}$. Let $\|A\|_{\text{F}} = \sqrt{\langle A, A\rangle_{\text{F}}}$ be the Frobenius norm of a tensor A induced by the inner product. For a matrix A with columns $(a_1, \ldots, a_n)$, the $\ell_{2,1}$ matrix norm defined as $\|A\|_{2,1} = \sum_{j=1}^n \|a_j\|$. We also use $\mathbb{O}_N$ to denote the set of $N \times N$ orthogonal matrices, i.e. matrices A such that $A^\top A = I_N$, where I_N is $N \times N$ identity matrix. Lastly, for $a, b \in \mathbb{R}$, put $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$.

2 Tensor Factor Models

Traditional factor models apply to 2-dimensional panel data represented by a matrix $\mathbf{Y} \in \mathbb{R}^{N\times T}$. The factor model with R factors can be expressed as a sum of a low-rank matrix and a matrix of idiosyncratic shocks:

$$\mathbf{Y} = \Lambda F^\top + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0, \tag{1}$$

where $F \in \mathbb{R}^{T\times R}$ is a matrix of R factors, $\Lambda \in \mathbb{R}^{N\times R}$ is a matrix of corresponding factor loadings, and $\mathbf{U} \in \mathbb{R}^{N\times T}$ are idiosyncratic random shocks.[5] The estimation of factors and their loadings can be done via PCA which can be computed with the singular-value decomposition (SVD) of $\mathbf{Y}$. More precisely, the fators/loadings can be estimated as the leading R right/left singular vectors of $\mathbf{Y}$. Equivalently, the loadings can be estimated as the leading R eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$ while and the factors as the leading R eigenvectors of $\mathbf{Y}^\top\mathbf{Y}$.

A d-dimensional panel dataset can be represented by a d-way tensor, denoted $\mathbf{Y} \in \mathbb{R}^{N_1\times\ldots\times N_d}$. We can describe $\mathbf{Y}$ by enumerating all its elements along the d ways (or modes):

$$\mathbf{Y} = \{y_{i_1,i_2,\ldots,i_d},\ 1 \leqslant i_j \leqslant N_j,\ 1 \leqslant j \leqslant d\}\,;$$

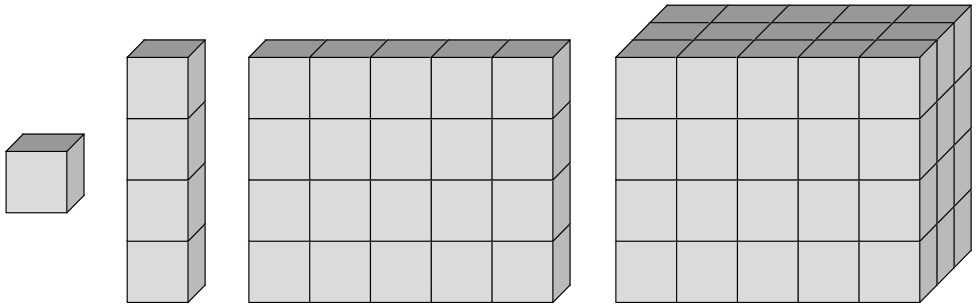
see Figure 1 for a graphical illustration. Similarly to a 2-way factor model, we can define a d-way factor model for a d-way tensor $\mathbf{Y} \in \mathbb{R}^{N_1\times\cdots\times N_d}$ as follows

$$\mathbf{Y} = \text{low-rank tensor } \mathbf{X} + \mathbf{U}.$$

[5] The factors and loadings can be random, but all results are stated conditional on their realization.

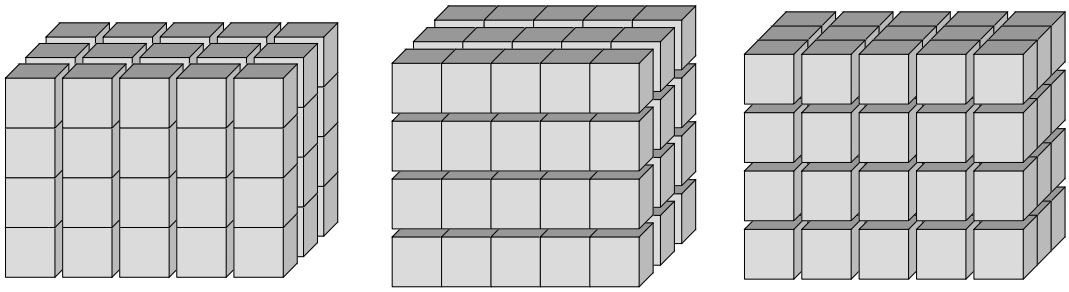
Unlike for matrices, the rank of a tensor can be defined in several different ways. To describe the notion of rank used in this paper, we need to introduce the notion of tensor matricization.

Figure 1: A scalar, and tensors of order 1 (vector), 2 (matrix), and 3



A tensor matricization operation can be described in terms of tensor *fibers*—a generalization of matrix rows and columns. A fiber is defined by fixing all but one of its dimensions, e.g., a matrix column is a mode-1 fiber and a matrix row is a mode-2 fiber. The mode-j fibers of a higher-order tensor are defined similarly; see Figure 2 for an illustration in the case of a 3-way tensor.

Figure 2: Mode-$1, 2$ and 3 *fibers* of a $4 \times 5 \times 3$ tensor



Recall that a matrix $\mathbf{Y} \in \mathbb{R}^{N_1 \times N_2}$ can be vectorized by stacking all its columns to obtain an $N_1 N_2 \times 1$ vector. There is also a second way to vectorize a matrix by stacking all its rows into an $1 \times N_1 N_2$ vector. Similarly, a d-way tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ can be matricized in d different ways across each of its d modes.[6] The mode-j matricizations of a tensor are

[6]Sometimes the term "flattening" of a tensor is used. We prefer the term "matricization" as it makes clear we are creating matrices, i.e., tensors could be flattened to lower dimensions that are not necessarily matrices.

obtained by stacking its mode-j fibers as columns of a $N_j \times \prod_{l \neq j} N_l$ matrix, denoted $\mathbf{Y}_{(j)}$. Formally, the elements of a mode-j matricization of a tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ is obtained by the following mapping:

$$y_{i_1,i_2,\ldots,i_d} \mapsto \dot{y}_{i_j,k} \quad \text{with} \quad k = 1 + \sum_{\substack{n=1 \\ n \neq j}}^{d} \left((i_n - 1) \prod_{\substack{m=1 \\ m \neq j}}^{n-1} N_m \right), \tag{2}$$

where $\dot{y}_{i_j,k}$ is (i_j, k) element of $\mathbf{Y}_{(j)} \in \mathbb{R}^{N_j \times \prod_{l \neq j} N_l}$; see Appendix A.1 for a numerical example of how a 3-way tensor is matricized along each of its three ways.

As we have already mentioned there are several different definitions of tensor rank, each leading to a different d-way factor model. The two most widely used definitions of tensor rank go back at least to Hitchcock (1927) and lead to the so-called Canonical Polyadic (CP) and Tucker decompositions.[7] The Tucker decomposition is more general than the CP decomposition and is based on a notion of a multilinear rank which corresponds to the d-tuple of ranks of each of its mode-j matricizations:

Definition 2.1. *A tensor $\mathbf{X} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ has a multilinear rank $(R_1, \ldots, R_d)$ if*

$$\operatorname{rank}(\mathbf{X}_{(j)}) = R_j, \qquad 1 \leqslant j \leqslant d.$$

Note that the matricizations of a tensor can have in general different ranks which is allowed in the Tucker decomposition and is not allowed in the CP decomposition. We can decompose a tensor $\mathbf{X} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ with multilinear rank $(R_1, \ldots, R_d)$ as

$$\mathbf{X}_{i_1,\ldots,i_d} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_d=1}^{R_d} g_{r_1,r_2,\cdots,r_d} \lambda^{(1)}_{i_1,r_1} \lambda^{(2)}_{i_2,r_2} \ldots \lambda^{(d)}_{i_d,r_d}$$

or more concisely as

$$\begin{aligned} \mathbf{X} &= \mathbf{G} \times_1 \Lambda_1 \times_2 \Lambda_2 \cdots \times_d \Lambda_d \\ &= \mathbf{G} \bigtimes_{j=1}^{d} \Lambda_j, \end{aligned}$$

where $\Lambda_j \in \mathbb{R}^{N_j \times R_j}$ is a matrix of factors/loadings with entries $\lambda^{(j)}_{i_j,r_j}$; $\mathbf{G} \in \mathbb{R}^{R_1 \times \cdots \times R_d}$ is the so-called *core tensor* with elements $g_{r_1,\ldots,r_d}$; and $\times_j$ is the mode-j product. The mode-j product is defined as a multiplication along the mode-j, e.g., the mode-1 product is $\times_1$:

[7] The CP decomposition is also known as CANDECOMP or PARAFAC; see Carroll and Chang (1970), Harshman (1970). The name Tucker decomposition comes from Tucker (1966).

$\mathbb{R}^{R_1 \times R_2 \cdots \times R_d} \times \mathbb{R}^{N_1 \times R_1} \to \mathbb{R}^{N_1 \times R_2 \cdots \times R_d}$ is

$$\mathbf{G} \times_1 \Lambda_1 = \left(\sum_{r=1}^{R_1} g_{r,r_2,\ldots,r_d} \lambda^{(1)}_{i_1,r} \right)_{1 \leqslant i_1 \leqslant N_1, 1 \leqslant r_j \leqslant R_j, 2 \leqslant j \leqslant d}$$

Following the idea of defining a tensor factor model as a sum of a low-rank tensor and a tensor of idiosyncratic shocks, we define the Tucker tensor factor model as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0 \\ &= \mathbf{G} \bigtimes_{j=1}^{d} \Lambda_j + \mathbf{U}. \end{aligned}$$

It is known that the Tucker decomposition is in general not unique, however, it is possible to identify the core tensor and all factors/loadings in the Tucker tensor model under additional orthogonality restrictions; see Section 3.1. The orthogonality restrictions are also commonly used to identify 2-way factor models.

Remark 2.1. *In economics and finance, we often have 3-dimensional panels, where one of the dimensions corresponds to time. Setting $F := \Lambda_3$ and $T := N_3$, the model in equation (2) becomes*

$$\mathbf{Y} = \mathbf{G} \times_1 \Lambda_1 \times_2 \Lambda_2 \times_3 F + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0.$$

If f_{t,r_3} is the (t, r_3) element of the factor matrix $F \in \mathbb{R}^{T \times R_d}$, then the entries of $\mathbf{Y}$ can be written equivalently as

$$\begin{aligned} y_{i_1,i_2,t} &= \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} g_{r_1,r_2,r_3} \lambda^{(1)}_{i_1,r_1} \lambda^{(2)}_{i_2,r_2} f_{t,r_3} + u_{i_1,i_2,t} \\ &= \sum_{r_1=1}^{R_1} \lambda^{(1)}_{i_1,r_1} f^{(1)}_{i_2,t,r_1} + u_{i_1,i_2,t} = \sum_{r_2=1}^{R_2} \lambda^{(2)}_{i_2,r_2} f^{(2)}_{i_1,t,r_2} + u_{i_1,i_2,t} = \sum_{r_3=1}^{R_3} \lambda^{(3)}_{i_1,i_2,r_3} f_{t,r_3} + u_{i_1,i_2,t}, \end{aligned}$$

where $f^{(1)}_{i_2,t,r_1}, f^{(2)}_{i_1,t,r_2}, \lambda^{(3)}_{i_1,i_2,r_3}$ are suitably defined. The last expression suggests that the Tucker tensor factor model has R_3 underlying time series factors with heterogeneous exposures $\lambda^{(3)}_{i_1,i_2,r_3}$ for modes-1 and 2. However, we could also rewrite it as a factor model with R_1 and R_2 factors for modes-1 and 2 respectively. The model is consistent with Lettau (2022, 2023) who considers the exact factor model with $\mathbf{U} = 0$. In contrast, in statistics it is common to consider a model, where the dynamics is driven by the core tensor $\mathbf{G}_t \in \mathbb{R}^{R_1 \times R_2}$ and there are $R_1 \times R_2$ time series factors; see Han et al. (2025), Chen et al. (2022), and Barigozzi, He, Li, and Trapani (2023) among others.

Remark 2.2. *When $R_1 = R_2 = \cdots = R_d$ and the core tensor $\mathbf{G}$ is diagonal with elements $g_{r_1,\ldots,r_d} = \mathbb{1}_{r_1=\cdots=r_d}$, we obtain the CP tensor factor model:*

$$\mathbf{Y} = \sum_{r=1}^{R} \lambda_r^{(1)} \circ \cdots \circ \lambda_r^{(d)} + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0,$$

where $\lambda_r^{(j)} \in \mathbb{R}^{N_j}$ are some vectors and $\circ$ denotes the tensor outer product. The CP factor model corresponds to the notion of CP rank. Formally, we say that a d-way tensor is a rank-1 tensor if it can be expressed as an outer product of d vectors. Every tensor $\mathbf{X} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ can be expressed as a finite sum of rank-1 tensors and the smallest number R of such rank-1 tensors is called the CP rank of $\mathbf{X}$. The CP model requires that all mode-j matricizations $\mathbf{Y}_{(j)}, j \leqslant d$ have the same rank R which may be restrictive in some applications.[8]

In the remaining part of this section, we consider several examples of tensor data in economics and finance, showing that such type of data appear in many applications. Using 3-dimensional examples with the notation $\mathbf{Y} = y_{i,j,k}$ or $y_{i,j,t}$, they include:

- Input-Output Models: i is industry sector, j type of input (e.g., labor, materials, capital), and k region or country
- Macroeconomic Data Across Countries i country, j macroeconomic variables (e.g., GDP, inflation, unemployment rate) and t time period (e.g., quarterly, annually)
- High-Frequency Trading Data: i asset (e.g., stocks, commodities), j attributes (e.g., bid price, ask price, volume), and t timestamps (e.g., milliseconds)
- Energy Markets and Economics: i country, j energy types (e.g., coal, wind, solar, oil), and t time periods
- Supply Chain Analysis: i products, j locations (e.g., factories, warehouses, retail outlets), and t time periods
- Real Estate and Urban Economics: i region or city, j property types (e.g., residential, commercial), and t time periods.

As well as 4-dimensional examples such as:

- Banking and Credit Risk Models: i customer, j loan types, k credit scores, and t time periods

[8] A previous version of our paper, Babii, Ghysels, and Pan (2022), considered the CP factor model with orthogonal loadings and factors which is now a special case of a more general framework.

- Consumer Behavior Analysis: i consumer demographics (e.g., age group, income level), j product categories (e.g., electronics, groceries), k regions, and t time periods.

3 Tensor PCA

In this section, we consider two tensor PCA (TPCA) algorithms to estimate the Tucker tensor factor model in equation (2). We begin by discussing the sufficient identifying conditions and present a simple TPCA algorithm. We argue that the simple TPCA is optimal when factors are strong in the second subsection. The third subsection describes an improved iterative alternating least-squares algorithm and discusses the improvements for the weak factor model. The next subsection provides the large sample distributions for loadings/factors estimated with simple TPCA.

3.1 Identification and Simple Tensor PCA

It is known that the Tucker decomposition of a tensor is not in general unique. In this section, we argue that the loadings/factors in the Tucker tensor factor model in equation (2) can be identified under the following assumption:

Assumption 3.1. *For every* $1 \leqslant j \leqslant d$, *(i)* $\Lambda_j^\top \Lambda_j = I_{R_j}$; *and (ii)* $\mathbf{G}_{(j)}\mathbf{G}_{(j)}^\top = \operatorname{diag}(\sigma_{j,1}^2, \ldots, \sigma_{j,R_j}^2)$ *for some* $\sigma_{j,1} > \cdots > \sigma_{j,R_j} > 0$.

By Appendix Lemma A.4.2, the mode-j matricization of equation (2) is

$$\mathbf{Y}_{(j)} = \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l \neq j} \Lambda_l \right)^\top + \mathbf{U}_{(j)}, \qquad 1 \leqslant j \leqslant d, \tag{3}$$

where $\bigotimes_{l \neq j} \Lambda_l = \Lambda_d \otimes \cdots \otimes \Lambda_{j+1} \otimes \Lambda_{j-1} \otimes \cdots \otimes \Lambda_1$. The latter satisfies the following property:

Proposition 3.1. *Under Assumption 3.1 (i)*

$$\left(\bigotimes_{k \neq j} \Lambda_k \right)^\top \left(\bigotimes_{k \neq j} \Lambda_k \right) = I_{\prod_{k \neq j} R_k}, \qquad 1 \leqslant j \leqslant d.$$

Then the tensor factor model is identified from the PCA applied to $\mathbf{Y}_{(j)}$. Indeed, if $\mathbf{U} = 0$, then by Proposition 3.1

$$\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top = \Lambda_j D_j \Lambda_j^\top,$$

where under Assumption 3.1, (ii) $D_j = \mathbf{G}_{(j)}\mathbf{G}_{(j)}^\top$ is a diagonal matrix with distinct elements and $\Lambda_j^\top \Lambda_j = I_{R_j}$. Then Λ_j is identified as eigenvectors of $\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top$ or equivalently as the left

singular vectors of $\mathbf{Y}_{(j)}$. Moreover, we have

$$\mathbf{Y} \bigtimes_{j=1}^{d} \Lambda_j = \mathbf{G} \tag{4}$$

since $\Lambda_j^\top \Lambda_j = I_{R_j}$ for all $1 \leqslant j \leqslant d$. If $\mathbf{U} \neq 0$, then assuming that the factor part dominates the noise asymptotically, cf. Assumption 3.3 below, identification is achieved in large samples.

This leads us to the TPCA estimation Algorithm 1 often credited to Tucker (1966) with some aspects already in Hitchcock (1927).[9]

Algorithm 1 TPCA

procedure TPCA($\mathbf{Y}, R_1, \ldots, R_d$)
  for $j = 1, \ldots, d$ **do**
    $\hat{\Lambda}_j \leftarrow N_j \times R_j$ matrix of R_j leading left singular vectors of $\mathbf{Y}_{(j)}$
  end for
  $\hat{\mathbf{G}} \leftarrow \mathbf{Y} \times_{j=1}^{d} \hat{\Lambda}_j^\top$ ▷ Estimate the core tensor
  return $\hat{\mathbf{G}}, \hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_d$
end procedure

3.2 Rates of Consistency

The following assumption imposes mild restrictions on the data generating process.

Assumption 3.2. *The idiosyncratic errors* $\mathbf{U} = \{u_{i_1,\ldots,i_d} : 1 \leqslant i_j \leqslant N_j, 1 \leqslant j \leqslant d\}$ *are i.i.d. such that* $\mathbb{E}(u_{i_1,\ldots,i_d}) = 0$, $\mathrm{Var}(u_{i_1,\ldots,i_d}) = \sigma^2$, *and for some* $c > 0$, $\mathbb{E}[e^{\lambda u_{i_1,\ldots,i_d}}] \leqslant e^{c\lambda}, \forall \lambda \in \mathbb{R}$.

Assumption 3.2 does not impose any restrictions on the dependence structure in factors and loadings. The factors may be generated by a non-stationary process and the loadings may fail to be i.i.d. It is plausible that after controlling for cross-sectional and time-series dependence through the factor structure, the idiosyncratic errors are i.i.d. Nonetheless, Assumption 3.2 can also be relaxed to heterogeneous and dependent arrays at costs of heavier notation and proofs and modified TPCA algorithm, see Zhang, Cai, and Wu (2022), which is left for future work.

Next, we make an assumption about the factor strength. Let σ_{j,R_j} be the smallest non-zero singular value of the matricized tensor $\mathbf{X}_{(j)}$ and let $\delta = \min_{1 \leqslant j \leqslant d} \sigma_{j,R_j}$ be a measure of

[9] It is also known as higher-order singular values decomposition in the numerical analysis; see De Lathauwer, De Moor, and Vandewalle (2000b).

factor strength in the tensor factor model. We assume next that factors are not extremely weak in the sense that the smallest singular values of matricizations are well-separated from zero:

Assumption 3.3. *The factor strength is such that* $\delta^2 \gtrsim \prod_{j=1}^d \sqrt{N_j} + \max_{1\leqslant j\leqslant d} N_j$.

The strong factors model in the tensor setting can be described as $\delta^2 \sim \prod_{j=1}^d N_j$.[10] In this case, Assumption 3.3 is automatically satisfied. Let O_j be the optimal rotation matrix solving $\min_{O\in\mathbb{O}_{R_j}} \|\hat{\Lambda}_j O - \Lambda_j\|_{\mathrm{F}}^2$.[11] The following result holds:

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied. Then with probability at least* $1 - Ce^{-cN_j}$, *we have*

$$\|\hat{\Lambda}_j O_j - \Lambda_j\|_{\mathrm{F}}^2 \lesssim \frac{R_j N_j}{\delta^2} + \frac{R_j \prod_{l=1}^d N_l}{\delta^4}, \qquad 1 \leqslant \forall j \leqslant d.$$

The proof appears in the Appendix. Theorem 3.1 is a non-asymptotic result valid for any values of tensor dimensions $(N_1, \ldots, N_d)$, provided that Assumption 3.3 is satisfied. It also applies to the CP-factor model, when matricizations have the same ranks and factors/loadings are assumed to be orthogonal. The orthogonality restriction for the CP factor model can also be relaxed; see Chang, He, Yang, and Yao (2023) and Chen, Han, and Yu (2024) for recent contributions.

Remark 3.1. *Consider the 2-way case,* $\mathbf{Y} = \Lambda F^\top + \mathbf{U}$ *with* $\mathbf{Y} \in \mathbb{R}^{N\times T}$, $\Lambda \in \mathbb{R}^{N\times R}$, *and* $F \in \mathbb{R}^{T\times R}$, *where* R *is the number of factors. In the strong factor model, the smallest singular value of* $\Lambda F^\top$ *is* $\delta \sim \sqrt{NT}$. *Theorem 3.1 implies that*

$$\|\hat{\Lambda} O_1 - \Lambda\|_{\mathrm{F}}^2 = O_P\left(\frac{R}{T}\right) \qquad \textit{and} \qquad \|\hat{F} O_2 - F\|_{\mathrm{F}}^2 = O_P\left(\frac{R}{N}\right),$$

To the best of our knowledge, these rates are faster than the best currently known results in the factor literature, cf. Bai and Ng (2023). In particular, the factor loadings are consistently estimated when N *is fixed while factors are consistently estimated when* T *is fixed. We also allow for the number of factors to diverge slowly with the sample size; see also Freeman and Weidner (2023) and Beyhum and Gautier (2022b) for results with a diverging number of factors in the 2-way case.*

[10] In the 2-way case, this corresponds to assuming that the squared singular values of $\mathbf{Y}$ scale at the NT-rate; see also Onatski (2012, 2022) for the weak factors in the 2-way case.

[11] It is known that the optimal rotation matrix for the Frobenius norm has a closed-form expression $O_j = \mathrm{sign}(\hat{\Lambda}_j^\top \Lambda_j)$, where the sign of a matrix A with SVD $A = \Lambda D V^\top$ is defined as $\mathrm{sign}(A) = \Lambda V^\top$.

Remark 3.2. *For the strong factor model, when* $\delta^2 \sim \prod_{j=1}^d N_j$*, Theorem 3.1 shows that*

$$\|\hat{\Lambda}_j O_j - \Lambda_j\|_{\mathrm{F}}^2 = O_P\left(\frac{R_j}{\prod_{l\neq j} N_l}\right), \qquad 1 \leqslant \forall j \leqslant d.$$

More generally, the $O_P(R_j N_j/\delta^2)$ *rate is obtained provided that the factor/loadings strength is such that* $\delta^2 \gtrsim \prod_{l\neq j} N_l$*. This rate is known to be minimax-optimal when* $d = 3$ *see Zhang and Xia (2018), Theorem 3. Therefore, the naive TPCA Algorithm 1 is optimal under the strong factors asymptotics.*

Remark 3.3. *If the last tensor dimension corresponds to time, i.e.,* $N_d = T$ *and* $\Lambda_d = F$*, then for the strong factor model, Theorem 3.1 shows that factors are estimated at the rate*

$$\|\hat{F} O_d - F\|_{\mathrm{F}}^2 = O_P\left(\frac{R_j}{N_1 N_2 \dots N_{d-1}}\right).$$

3.3 Alternating Least-Squares

For weak factors/loadings, the estimation accuracy of factors/loadings in Theorem 3.1 does not always improve when more data is available in all tensor dimensions because the second term with $\prod_{l=1}^d N_l$ can dominate, creating a bottleneck. This comes from the fact that the matricized tensor $\mathbf{Y}_{(j)}$ has a very large number of columns, namely $\prod_{l\neq j} N_l$. Under Assumption 3.1, using the mode-l multiplication of $\mathbf{Y}$ by loadings $\Lambda_l, l \neq j$, we obtain

$$\mathbf{Y} \bigtimes_{l\neq j} \Lambda_l^\top = \mathbf{G} \times_j \Lambda_j^\top + \mathbf{U} \bigtimes_{l\neq j} \Lambda_l^\top; \tag{5}$$

cf. equation (4). The mode-j matricization of equation (5),

$$\left(\mathbf{Y} \bigtimes_{l\neq j} \Lambda_l^\top\right)_{(j)} = \Lambda_j \mathbf{G}_{(j)} + \left(\mathbf{U} \bigtimes_{l\neq j} \Lambda_l^\top\right)_{(j)},$$

has a much smaller number of columns, namely $\prod_{l\neq j} R_l$ instead of $\prod_{l\neq j} N_l$. Therefore, applying PCA to the mode-j matricization of the left-hand side in the equation (5), may allow us to estimate Λ_j more precisely, provided that the noise is negligible; see also equation (A.13). Iterating this procedure and updating the factors/loadings leads to the alternating least-squares Algorithm 2 for Tucker decomposition; see Kroonenberg and De Leeuw (1980).[12] The following result establishes the convergence rate of estimators computed using

[12]It is also known as the higher-order orthogonal iterations; see De Lathauwer, De Moor, and Vandewalle (2000a) for more details.

this algorithm:

Theorem 3.2. *Suppose that assumptions of Theorem 3.1 are satisfied and for all $j \leqslant d$, we have $\sigma_{1,j} \lesssim \delta$, $N_j \sim N \uparrow \infty$, and $R_j = O(1)$. Then if the factor strength is such that $N^d/\delta^4 = o(1)$ and the number of iterations is $\bar{k} \geqslant \log(N^{(d-1)/2}/\delta)$, we obtain with probability at least $1 - Ce^{-cN}$ that:*

$$\left\| \hat{\Lambda}_j^{(\bar{k})} O_j - \Lambda_j \right\|_{\mathrm{F}}^2 \lesssim \frac{N}{\delta^2}, \qquad 1 \leqslant \forall j \leqslant d.$$

The proof of this result appears in the Appendix.

Algorithm 2 ALS

procedure ALS($\mathbf{Y}, R_1, \ldots, R_d$)
  Initialize $\hat{\Lambda}_j^{(0)}$ for $1 \leqslant j \leqslant d$ using Algorithm 1.
  for $k = 1, 2, \ldots, \bar{k}$ **do**
    for $j = 1, 2, \ldots, d$ **do**
      $\hat{\mathbf{Y}}^{(k)} \leftarrow \mathbf{Y} \times_{l \neq j} \hat{\Lambda}_l^{(k-1)\top}$
      $\hat{\Lambda}_j^{(k)} \leftarrow R_j$ leading left singular vectors of $\hat{\mathbf{Y}}_{(j)}^{(k)}$ ▷ Update loadings/factors
    end for
  end for
  $\hat{\mathbf{G}} \leftarrow \mathbf{Y} \times_{j=1}^{d} \hat{\Lambda}_j^{(\bar{k})\top}$ ▷ Estimate the core tensor
  return $\hat{\mathbf{G}}, \hat{\Lambda}_1^{(\bar{k})}, \hat{\Lambda}_2^{(\bar{k})}, \ldots, \hat{\Lambda}_d^{(\bar{k})}$
end procedure

Theorem 3.2 makes a simplifying assumption that the tensor dimensions increase proportionally to each other which in the 2-way factor case requires that $N/T \to c > 0$. Then for the strong factor model, we have $\delta^2 \sim N^d$ and the condition $N^d/\delta^4 = o(1)$ simply requires that $N \to \infty$ which is trivially satisfied. Consequently, we obtain the rate of order $O_P(1/N^{d-1})$ which is the same as the optimal rate of the naive TPCA Algorithm 1. However, when the factors are weak, the rate of the ALS Algorithm 2 can be faster than that of the naive TPCA. Therefore, we can expect that the ALS with naive TPCA used as a starting point may estimate factors/loadings more accurately, provided that a sufficiently large number of iterations is made.

Remark 3.4. *It is known that the rate achieved by the ALS Algorithm 2 in Theorem 3.2 is minimax-optimal; see Zhang and Xia (2018), Theorem 3. Interestingly, if the factors become so weak, that the $N^d/\delta^4 = o(1)$ condition fails, the rate-optimal MLE estimator, is a solution to the NP-hard optimization problem and there is a gap between statistical and computational limits; see also Richard and Montanari (2014).*

3.4 Inference

In this section, we consider inference on loadings and factors in the general Tucker model, covering the orthogonal CP model as a special case. The loadings/factors are estimated with the TPCA Algorithm 1. Let $\hat{\Lambda}_{j,i:}$ and $\Lambda_{j,i:}$ be $R_j \times 1$ vectors corresponding to the transposed i^{th} rows of $\hat{\Lambda}_j O_j$ and Λ_j respectively with $O_j = \hat{\Lambda}_j^\top \Lambda_j$. Let $\|.\|_{2,\infty}$ be the entry-wise $\ell_{2,\infty}$ matrix norm.

We need the following set of assumptions for inference:

Assumption 3.4. *(i) the factor strength is* $\delta \sim \prod_{j=1}^d N_j$*; (ii) the coherence is* $\|\Lambda_j \Lambda_j^\top\|_{2,\infty}^2 = O(R_j/N_j)$ *and* $\|V_j\|_{2,\infty} = o(1)$ *for all* $j \leqslant d$*.*

The coherence condition roughly requires that the values of factors/loadings are spread out instead of being concentrated in a finite number of entries; see Candes and Recht (2012), Definition 1.8.

The following result holds for factors/loadings, depending on the dimension $j = 1, \ldots, d$:

Theorem 3.3. *Suppose that Assumptions 3.1, 3.2, and 3.4 are satisfied. Then*

$$D_j^{1/2}(\hat{\Lambda}_{j,i:} - \Lambda_{j,i:} - B_{j,i}) \xrightarrow{d} N(0, \sigma^2 I_{R_j}),$$

provided that $\prod_{l \neq j} N_l / N_j^3 = o(1)$ *and the expression of* $B_{j,i}$ *can be found in the proof.*

Note that for $d = 3$, conditions of Theorem 3.3 are satisfied when the tensor dimensions grow proportionally, i.e., $N_j \sim N \uparrow \infty$ for $j = 1, 2, 3$. Note also that the results of Bai (2003), Theorem 2, for 2-way factor models are not applicable to tensors because the condition $\sqrt{T}/N = o(1)$ does not hold for $d = 3$ when the tensor dimensions grow proportionally since $T \sim N^2$ in this case. Lastly, eliminating the bias and/or relaxing the i.i.d. homoskedastic errors may require modifying the TPCA estimator, see Zhang et al. (2022), and is left for future research.

We now turn to the estimation of scale components. Let $(\hat{\sigma}_{j,r}^2)_{1 \leqslant r \leqslant R}$ be the eigenvalues of $\mathbf{Y}_{(j)} \mathbf{Y}_{(j)}^\top$. The following result holds:

Theorem 3.4. *Under Assumptions of Theorem 3.3, we have*

$$\left(\frac{\hat{\sigma}_{j,r}^2 - \sigma_{j,r}^2}{\sigma_{j,r}} \right)_{1 \leqslant r \leqslant R_j} \xrightarrow{d} N(0, 4\sigma^2 I_{R_j}).$$

It immediately follows from Theorem 3.4 that $\hat{\sigma}_{j,r}^2 / \prod_{j=1}^d N_j$ is a consistent estimator of the asymptotic factor strength constant $d_{j,r}$.

4 Testing the Number of Factors

In this section, we develop a novel test for the number of factors in the tensor factor model. Specifically, we consider the following hypotheses:

$$\mathrm{H}_0: \leqslant k \text{ factors} \qquad \text{vs.} \qquad \mathrm{H}_1: \text{ the number of factors is } > k, \text{ but } \leqslant K.$$

The quantity K is selected by the user and reflects the upper bound on the total possible number of factors.

Let $\hat{\sigma}_{j,1}^2 \geqslant \hat{\sigma}_{j,2}^2 \geqslant \cdots \geqslant \hat{\sigma}_{j,N_j}^2 \geqslant 0$ be the eigenvalues of $\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top$. Under Assumption 3.3, the first R_j eigenvalues diverge from zero, at least in population. Consider the following statistics

$$S_j = \max_{k<r\leqslant K} \frac{\hat{\sigma}_{j,r}^2 - \hat{\sigma}_{j,r+1}^2}{\hat{\sigma}_{j,r+1}^2 - \hat{\sigma}_{j,r+2}^2}, \qquad 1 \leqslant j \leqslant d,$$

and put

$$Z = \max_{0<r\leqslant K-k} \frac{\xi_r - \xi_{r+1}}{\xi_{r+1} - \xi_{r+2}},$$

where $(\xi_1, \ldots, \xi_{K-k+2})$ follow the joint type-1 Tracy-Widom distribution; see El Karoui (2003) and Soshnikov (2002).

Then, consider the sequence:

$$\tau = \left(\sqrt{N_j \vee \prod_{k\neq j} N_k - 1} + \sqrt{N_j \wedge \prod_{k\neq j} N_k}\right)\left(\left(N_j \vee \prod_{k\neq j} N_k - 1\right)^{-1/2} + \left(N_j \wedge \prod_{k\neq j} N_k\right)^{-1/2}\right)^{1/3}.$$

The following result holds provided that $N_j \lesssim \prod_{k\neq j} N_k$:

Theorem 4.1. *Suppose that Assumptions 3.1 and 3.2 are satisfied, $\sigma_{j,r}^2 \sim \prod_{j=1}^d N_j$, and $u_{i_1,\ldots,i_d} \sim N(0,\sigma^2)$. Suppose also that $N_j/\tau + \prod_{k\neq j} N_k/(N_j\tau) = o(1)$. Then under H_0, $S_j \xrightarrow{d} Z$, while under H_1, we have $S_j \uparrow \infty$ for every $j \leqslant d$.*

Note that the rate condition for τ is satisfied in the 3-dimensional case when $N_1 \sim N_2 \sim N_3$. Theorem 4.1 leads to the following testing procedure:

1. Let $(Z_i)_{i=1}^m$ be m independent random variables drawn from the same distribution as Z. To approximate the distribution of $(\xi_1, \xi_2, \ldots)$, we use the eigenvalues of a symmetric $N_j \times N_j$ Gaussian matrix $\Xi = (\zeta_{i,j})$ with $\zeta_{i,j} \sim_{i.i.d.} N(0, \tau_{i,j})$ with $\tau_{i,j} = 1$ if $i < j$ and $\tau_{i,j} = 2$ for $i = j$.

2. Compute the p-value $p_j = 1 - F_m(S_j)$ for each $1 \leqslant j \leqslant d$, where $F_m(x) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}_{Z_i \leqslant z}$.

The p-values would correspond to the null hypothesis that there are at most k factors for the mode-j matricization. The joint tests for the number of factors across all matricizations can be obtained using the Bonferroni's correction. The α-level test would reject H_0 whenever

$$p_j \leqslant \alpha/d.$$

In practice, one could run the test for several different values of k to determine the number of factors and control the size with Bonferroni correction. Note also that the test statistic S_j is decreasing with k, which implies that p-values are increasing with k.

Remark 4.1. *In contrast to Onatski (2009), the dimensions of matrices obtained from tensor matricization do not grow proportionally and the Tracy-Widom asymptotics is recovered thanks to El Karoui (2003). Note also that in our case we have the type-1 Tracy-Widom distribution.*

5 Monte Carlo Experiments

The objective of this section is to assess the finite sample properties of our estimation procedure.

5.1 Simulation Design

We consider the Tucker model for the 3-way tensor for $N \times J \times T$ tensor

$$\mathbf{Y} = \mathbf{G} \times_1 \Lambda \times_2 M \times_3 F + \mathbf{U},$$

where the core tensor is $\mathbf{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$, the loadings matrices are $\Lambda \in \mathbb{R}^{N \times R_1}$ and $M \in \mathbb{R}^{J \times R_2}$, and the factor matrix is $F \in \mathbb{R}^{T \times R_3}$.

We set $R_1 = 1$ and $R_2 = R_3 = 2$ and generate the $1 \times 2 \times 2$ core tensor $\mathbf{G}$ with 1×2 slices

$$G_1 = [\sigma_1, 0] \qquad \text{and} \qquad G_2 = [0, \sigma_2],$$

where $\sigma_1 > \sigma_2 > 0$. The corresponding matricizations are

$$\mathbf{G}_{(1)} = [\sigma_1, 0, 0, \sigma_2], \qquad \mathbf{G}_{(2)} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}, \qquad \mathbf{G}_{(3)} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}.$$

Clearly, we have $\mathbf{G}_{(j)}\mathbf{G}_{(j)}^\top$ for $j = 1, 2, 3$ begin diagonal, hence, the core tensor satisfies Assumption 3.1. The smallest singular values of $\mathbf{G}_{(j)}$ are $\sqrt{\sigma_1^2 + \sigma_2^2}$, σ_2, and σ_2 respectively

for $j = 1, 2, 3$. The factor strength is $\delta = \sigma_2$. We first set $\sigma_1 = d_1\sqrt{NJT}$ and $\sigma_2 = d_2\sqrt{NJT}$ which corresponds to the strong tensor factor model.

Therefore, we simulate the $\mathbf{Y}$ tensor with $(i, j, t)^{\text{th}}$ observation

$$y_{i,j,t} = \sigma_1\lambda_i\mu_{j,1}f_{t,1} + \sigma_2\lambda_i\mu_{j,2}f_{t,2} + u_{i,j,t}, \qquad u_{i,j,t} \sim_{\text{i.i.d.}} N(0, s_u^2).$$

Note that this rank-$(1, 2, 2)$ Tucker tensor factor model features fewer parameters than the rank-2 CP model, where the first loading vector would be different in the two terms.

The rest of the DGP is as follows. We generate $T \times 2$ matrix of factors from two independent AR(1) processes

$$\dot{f}_{t,r} = \rho\dot{f}_{t-1,r} + \varepsilon_{t,r}, \qquad \varepsilon_{t,r} \sim_{\text{i.i.d.}} N(0, s_\varepsilon^2), \qquad r = 1, 2.$$

For the factor series $\dot{f}_r \in \mathbb{R}^T$, we compute $f_r = \dot{f}_r/\|\dot{f}_r\|$ to ensure that the factor matrix satisfies Assumption 3.1. All loading vectors are generated by taking norm-1 eigenvectors of a symmetric positive semi-definite matrix $A^\top A$, where each element is $A_{i,j} \sim_{\text{i.i.d.}} U(0, 1)$.

5.2 Small Sample Properties

In this subsection, we first assess how changing sample sizes affects the estimation accuracy of factors and loadings, and show that the estimation improvements are perfectly aligned with the convergence rates given in Theorem 3.1. We also show that the finite sample distribution of the estimation error is aligned with the asymptotic distribution shown in Theorem 3.3.

For convergence rate, we simulate the 3-way tensor as described in Section 5.1 where the parameters are set to be: (1) the AR(1) process $\dot{f}_r$ takes $\rho = 0.5$ and $s_\varepsilon = 0.1$; (2) the signal strength $d_1 = 2, d_2 = 1$ and the noise strength $s_u = 1$; (4) the sample size of the baseline model is (N, J, T)=(30,30,30), and we compare the estimation error of the baseline model with that of the modified model. We consider three different cases of modification: (a) doubling sample sizes of all dimensions, $(N, J, T) = (60,60,60)$, (b) doubling the sample sizes of two dimensions, $(N, J, T) = (60,60,30)$, (c) doubling the sample size of only one dimension, $(N, J, T) = (60,30,30)$. We evaluate the estimates using the ℓ_2 norm. Although the rank for the M, F is 2, the convergence rates are the same for the first and the second rank and it is enough to examine the first rank. As the signs of $\hat{\lambda}_r, \hat{\mu}_r$, and $\hat{f}_r$ are undetermined, we

calculate the errors as follows:

$$\begin{aligned}\mathbb{L}_\lambda &= \|\hat{\lambda}_r \times \text{sign}(\hat{\lambda}_r^\top \lambda_r) - \lambda_r\|,\\ \mathbb{L}_\mu &= \|\hat{\mu}_r \times \text{sign}(\hat{\mu}_r^\top \mu_r) - \mu_r\|,\\ \mathbb{L}_f &= \|\hat{f}_r \times \text{sign}(\hat{f}_r^\top f_r) - f_r\|,\end{aligned} \tag{6}$$

where $\text{sign}(a) = \mathbb{1}_{a>0} - \mathbb{1}_{a<0}$.

Figure 3 plots the histograms of the ℓ_2 losses of the baseline versus modified DGPs. In panel (a) - (c), as we double the sizes of all dimensions, the estimation of the factor $\hat{f}_1$ and loadings $\hat{\lambda}_1$, $\hat{\mu}_1$ all improve. The average error is reduced roughly by a half for the factor and two loadings vectors. This is aligned with the convergence rate in Theorem 3.1, since doubling all 3 dimensions of a tensor reduces the ℓ_2 error by 1/2. In panels (d) - (f), when we only double N and J, the improvement for the average error of $\hat{\lambda}_1$ is 0.01/0.015, while the improvement for $\hat{\mu}_1$ is 0.012/0.017. Both are roughly aligned with the reduction in the ℓ_2 error by $1/\sqrt{2}$. On the other hand, the improvement for $\hat{f}_1$ is 0.0083/0.017, which is aligned with the reduction of the ℓ_2 error by 1/2. In panels (g) - (i), when we only double N, there is no improvement for $\hat{\lambda}_1$ because J and T are unchanged in the $O_P(1/\sqrt{JT})$ rate; the improvement for both $\hat{\mu}_1$ and $\hat{f}_1$ are 0.012/0.017, which is aligned with the $1/\sqrt{2}$ improvement factor.

For the asymptotic distribution, due to symmetry, we focus on inference for the loading vector $\mu_{j,r} \in \mathbb{R}$. We simulate the 3-way tensor using the same parameters as previous simulation except that we fix the sample size at $(N, J, T) = (30, 30, 30)$. Theorem 3.3 shows that

$$d_r s_u^{-1} \sqrt{NJT}(\hat{\mu}_{j,r} - \mu_{j,r} - b_{j,r}) \xrightarrow{d} N(0, 1)$$

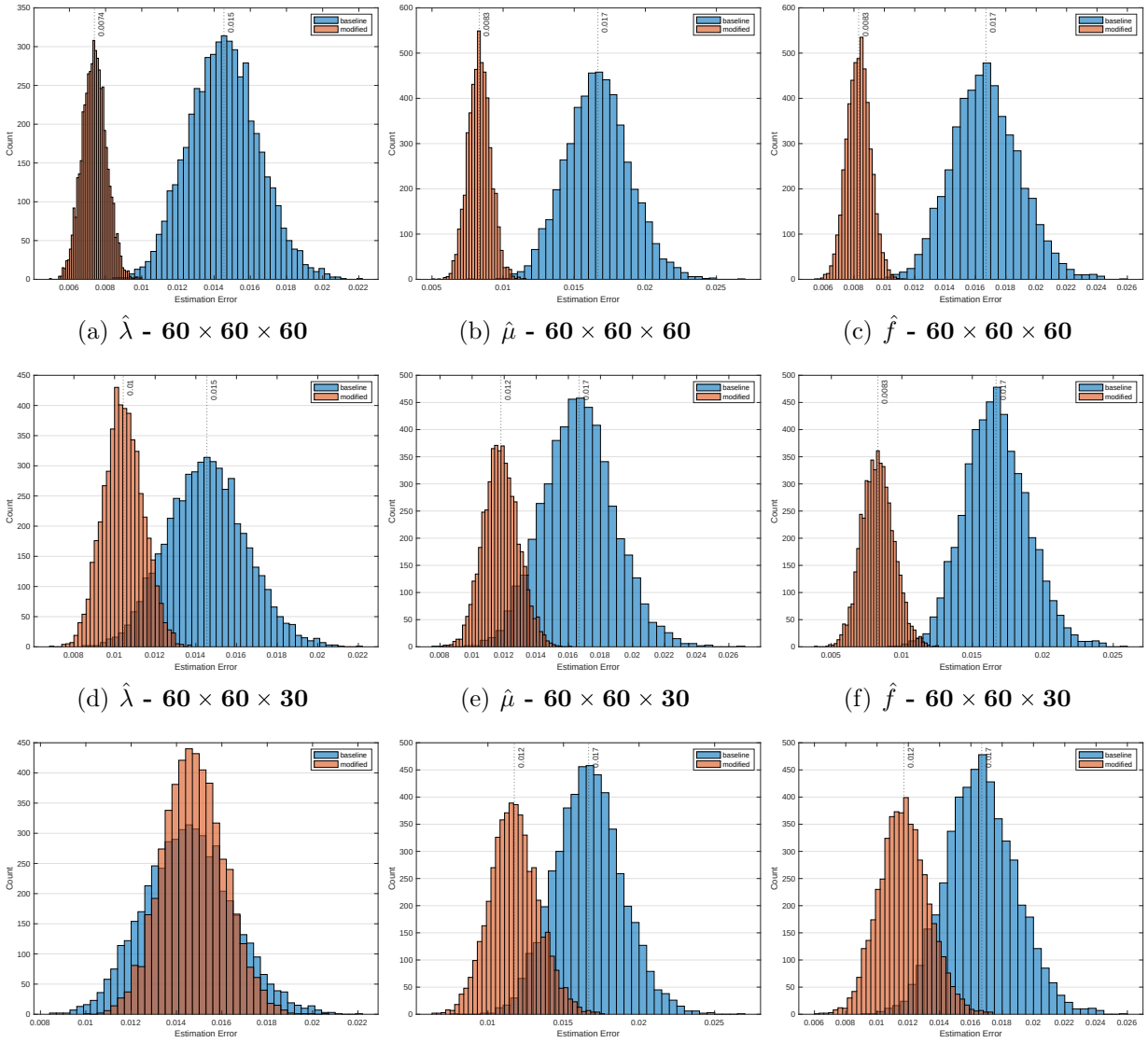
which is what we aim to verify with simulations. Figure 4 reports the histograms of the scaled estimation error $d_r s_u^{-1} \sqrt{NJT}(\hat{\mu}_{j,r} - \mu_{j,r})$ in 5000 MC simulations. For convenience, we only report the results of the first two elements of the loadings $\mu_{j,r}$ with $r = 1, 2$ and $j = 1, 2$. It can be seen that the finite sample distribution closely matches the asymptotic distribution, the QQ plot of the errors against the asymptotic distribution again confirms the finding. Interestingly we find that the bias appears to be negligible. Overall, these results show that the predictions of the asymptotic theory are valid in finite samples.

5.3 TPCA vs. ALS

In this subsection, we show that ALS (Algorithm 2) has faster convergence rate than TPCA (Algorithm 1) under weak factor scenarios. According to Theorem 3.1 and Theorem 3.2,

Figure 3: Estimation Accuracy of Tensor PCA: Changing Sizes of Dimensions

The DGP appears in Section 5.1. We plot the histograms of ℓ_2 losses of the estimated factor/loading of baseline vs modified DGP in 5000 MC simulations. The baseline DGP has sizes $(N, J, T) = (30,30,30)$, and modified DGP has sizes (N, J, T) shown in the subtitles. The blue histogram corresponds to the baseline DGP while the orange histogram to the modified DGP with the increased sample size. The dotted line marks the mean of the ℓ_2 errors.



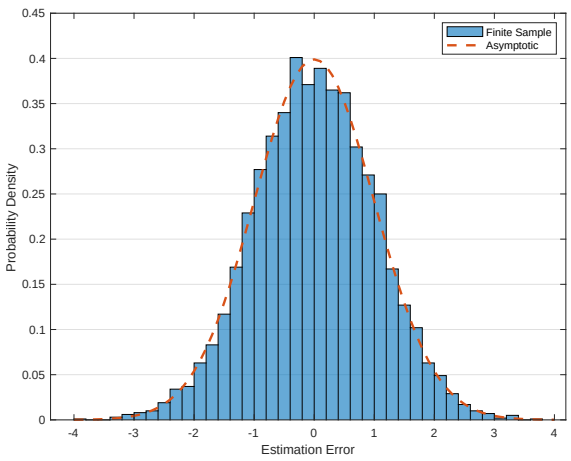
(a) $\hat{\lambda}$ - $\mathbf{60 \times 60 \times 60}$ (b) $\hat{\mu}$ - $\mathbf{60 \times 60 \times 60}$ (c) $\hat{f}$ - $\mathbf{60 \times 60 \times 60}$

(d) $\hat{\lambda}$ - $\mathbf{60 \times 60 \times 30}$ (e) $\hat{\mu}$ - $\mathbf{60 \times 60 \times 30}$ (f) $\hat{f}$ - $\mathbf{60 \times 60 \times 30}$

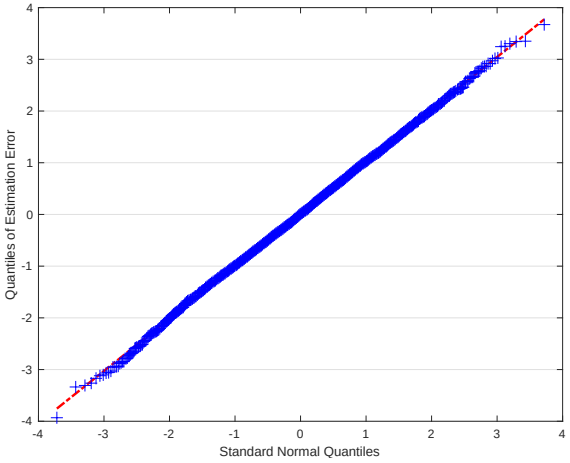
(g) $\hat{\lambda}$ - $\mathbf{60 \times 30 \times 30}$ (h) $\hat{\mu}$ - $\mathbf{60 \times 30 \times 30}$ (i) $\hat{f}$ - $\mathbf{60 \times 30 \times 30}$

Figure 4: Asymptotic Distribution Reconciliation

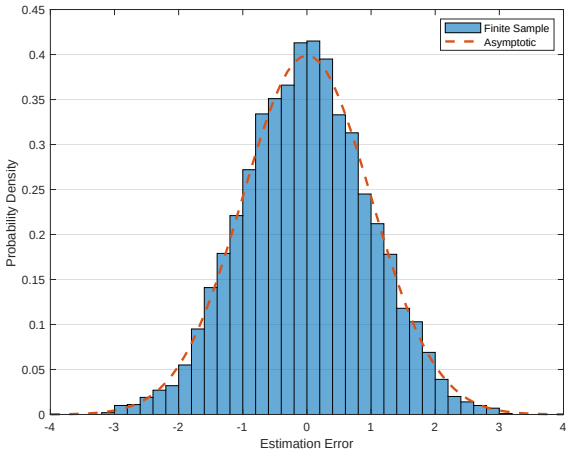
The DGP appears in Section 5.1. We plot the histograms of the scaled estimation error $d_r s_u^{-1}\sqrt{NJT}(\hat{\mu}_{j,r} - \mu_{j,r})$ in 5000 MC simulations and compare it with the asymptotic distribution - standard normal. We also show the QQ plot of the errors against the asymptotic. For simplicity, we only report the results of the first two elements of the loadings $\mu_{j,r}, r = 1, 2$ and $j = 1, 2$ as the theoretical results hold for all elements.



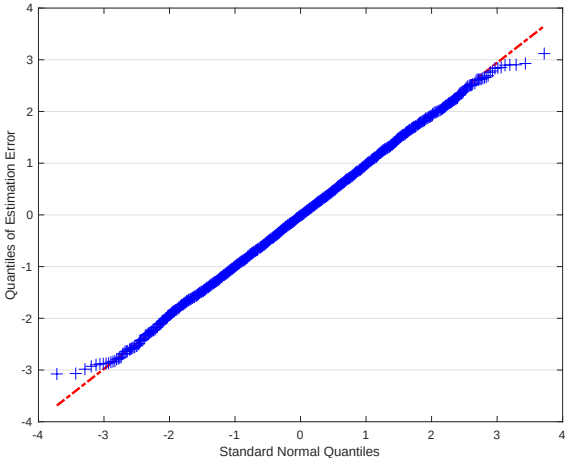
(a) **Estimation Error -** $\mu_{1,1}$



(b) **QQ Plot -** $\mu_{1,1}$



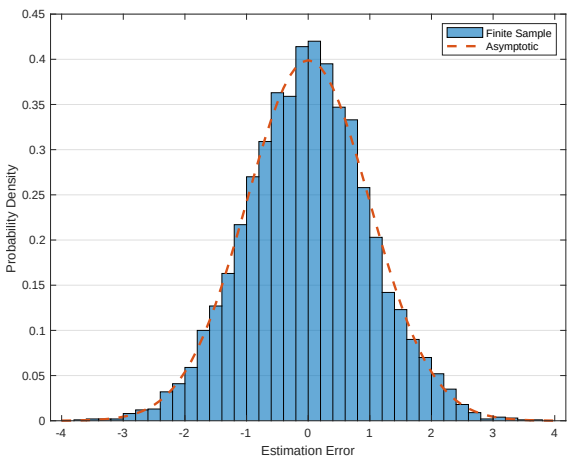
(c) **Estimation Error -** $\mu_{2,1}$



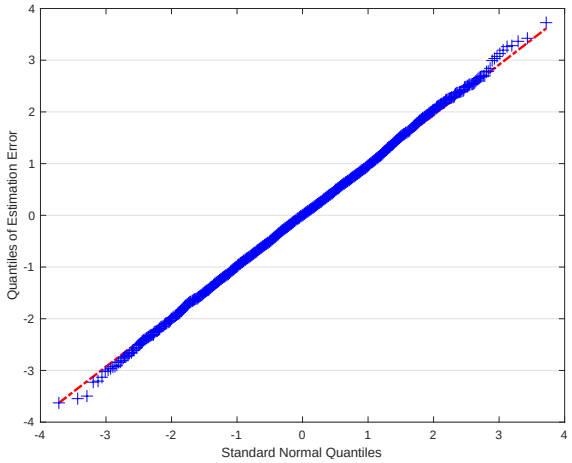
(d) **QQ Plot -** $\mu_{2,1}$

Figure 5: Asymptotic Distribution Reconciliation (continued)

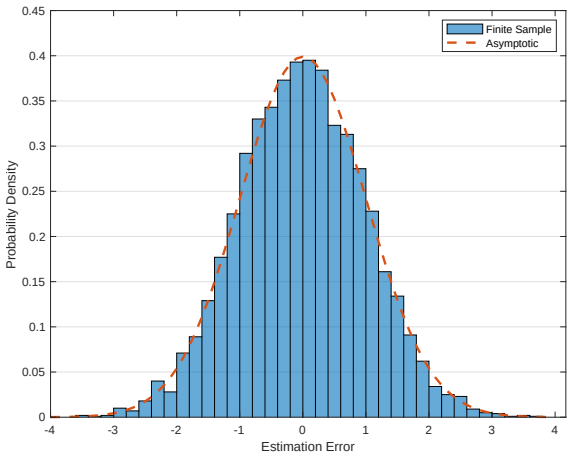
The DGP appears in Section 5.1. We plot the histograms of the scaled estimation error $d_r s_u^{-1}\sqrt{NJT}(\hat{\mu}_{j,r} - \mu_{j,r})$ in 5000 MC simulations and compare it with the asymptotic distribution - standard normal. We also show the QQ plot of the errors against the asymptotic. For simplicity, we only report the results of the first two elements of the loadings $\mu_{j,r}, r = 1, 2$ and $j = 1, 2$ as the theoretical results hold for all elements.



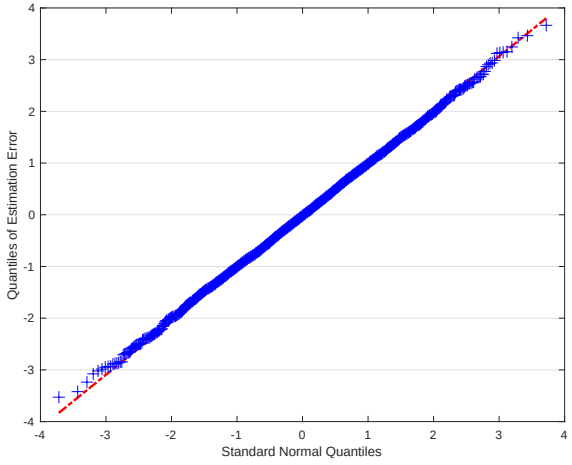
(a) **Estimation Error -** $\mu_{1,2}$



(b) **QQ Plot -** $\mu_{1,2}$



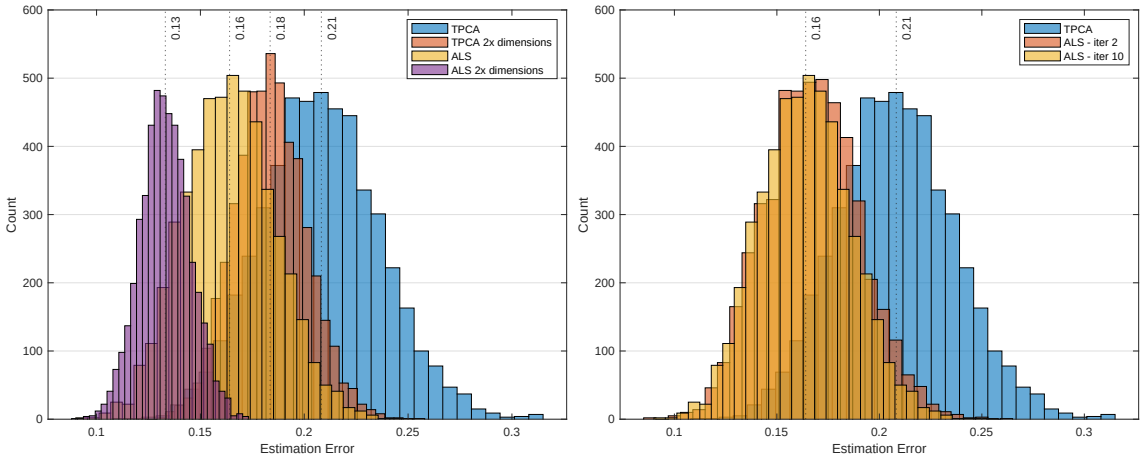
(c) **Estimation Error -** $\mu_{2,2}$



(d) **QQ Plot -** $\mu_{2,2}$

Figure 6: TPCA vs. ALS when doubling dimensions

The weak factor DGP appears in Section 5.3. We plot the histograms of ℓ_2 losses of the estimated factor/loading of TPCA vs ALS in 5000 MC simulations. The baseline DGP has sizes $(N, J, T) = (30,30,30)$, and it is compared to the DGP has double the sizes $(N, J, T) = (60,60,60)$. The dotted line marks the mean of the ℓ_2 errors. The plot (a) shows that ALS has larger improvement than TPCA under weak factors, and the plot (b) shows how ALS iterations improves the convergence rate.



(a) **TPCA vs. ALS** (b) **ALS iterations**

under weak factors, the TPCA rate is $O_P\left(\frac{1}{N^{0.6}} + \frac{1}{N^{0.2}}\right)$ while the ALS rate is $O_P\left(\frac{1}{N^{0.6}}\right)$ for a sufficiently large N, provided that the number of iterations is $\bar{k}$ is large enough. We follow the same DGP as Section 5.2 except we set the signal strength $\sigma_1 = \sqrt{(NJT)^{1.6/3}}$ for weak factor. The sizes of all dimensions are set to be the same so that the finite sample properties are the same for all dimensions, and it is enough to examine only one dimension.

Figure 6 (a) plots the comparison between TPCA and ALS when all dimensions are doubled from $(30, 30, 30)$ to $(60, 60, 60)$. The dotted line marks the mean of the ℓ_2 errors. It can be noted that ALS shows more improvement in error reduction, from the mean of 0.16 to 0.13, while TPCA error is reduced from 0.21 to 0.18. The error reduction rate of ALS - 0.13/0.16 - is roughly in line with the theoretical value of $1/(2^{0.6})$, and the error reduction of TPCA - 0.18/0.21 - is also roughly aligned with the theoretical value of $1/(2^{0.2})$, which is slightly slower under weak factors.

Figure 6 (b) shows the improvement of ALS over TPCA under weak factors. It can be noted that ALS with 2 iterations already improves TPCA, and ALS with 10 iterations further improves the estimation, though by slightly less.

5.4 Testing the Number of Factors

We conclude with power properties of our maximum eigenvalue ratio test for the number of factors discussed in Section 4. The DGP is generated as described in Section 5.1. We generate rank $(1, 2, 2)$ model, and test the null hypothesis that the rank is 1 against the alternative that there are more than 1 but less than K factors for each matricization.

The parameters are designed as follows: (1) the scale component is $\sigma_r = d_r \times \sqrt{NJT}$ with $d_1 = 2$ and we gradually increase d_2 to study the power properties of the test, so when $d_2 = 0$, the empirical rejection probability corresponds to empirical size of the test, and none zero d_2 corresponds to empirical power for rank 2 dimensions, (2) we study cases where the idiosyncratic errors are generated with Gaussian distribution or Student's t distribution (the degrees of freedom are 5), and in both cases the variance of the errors is normalized to be $s_u = 1$, (3) we study finite sample properties of the test using 3-way tensor, and the sizes of the tensors is $(N, J, T) = 30 \times 40 \times 50$, with the first dimension being the smallest.

We perform the test for the number of factors as discussed in Section 4, where we approximate the asymptotic distribution of the statistics by randomly generating Gaussian matrices 5,000 times, and we also replicate the simulations for 5,000 times to calculate the empirical rejection probability for each scenario.

Figure 7 reports the empirical rejections probabilities of the test for the null hypothesis of 1 factor against alternative of more than 1 factor but less than K factors, with $K = 3, 5, 7$, on a 3$^{\text{rd}}$ order tensor. The test is performed on each matricization individually and plotted separately. The power curves show that the empirical size of test is very close to the nominal level of 5%,[13] and the empirical power of the test reaches 1 as the strength of the second factor d_2 increases. The comparison among different K values implies that, when the true number of factors is within range of the alternative hypothesis, the tighter the range of the alternative is, the more likely we reject the null when the alternative is true. This finding is true no matter which matricization we perform the test on.

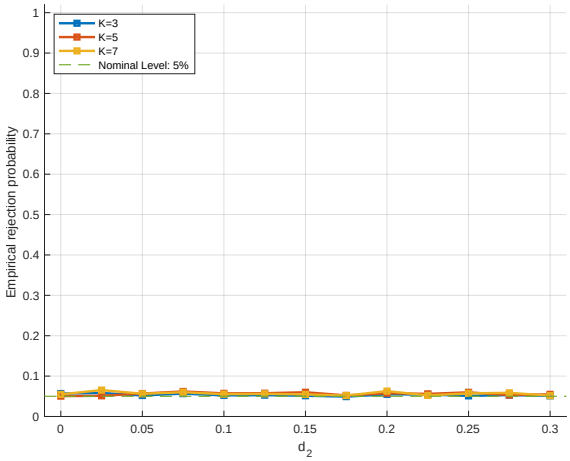
We also make comparisons of the test for different types of errors. The two plots of Figure 8 correspond to tensor with: (a) Gaussian errors, (b) Student's t distributed errors. Among different matricizations, the dimension with the larger size ('mat3') tend to climb faster to one than smaller size ('mat2'), meaning the matricization with the larger dimension gives the higher empirical power.

Finally, while our simulation results rely on the critical values that assume Gaussian errors, we learn from Figure 8 (b) that the test performs well for non-Gaussian distributions albeit with some loss of power.

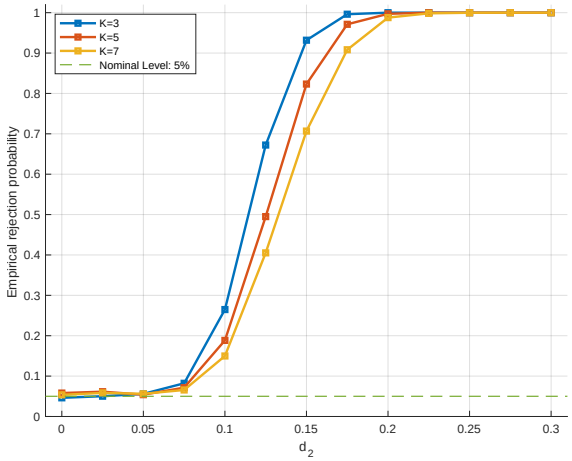
[13] Since the rank is 1 for the first dimension, the power stays at the nominal level.

Figure 7: Testing the Number of Factors: Power Curves

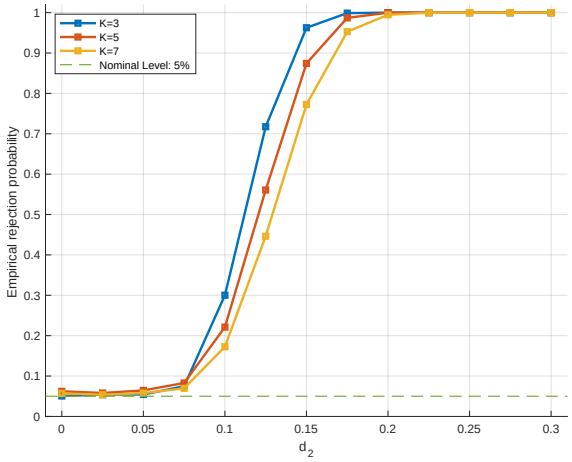
The power curves are plotted for testing the null hypothesis of 1 factor against alternative of more than 1 but less than K factors, with $K = 3, 5, 7$. Plots (a) through (c) report the tests for each matricization separately for a 3rd order tensor with Gaussian errors.



(a) **1st Matricization**



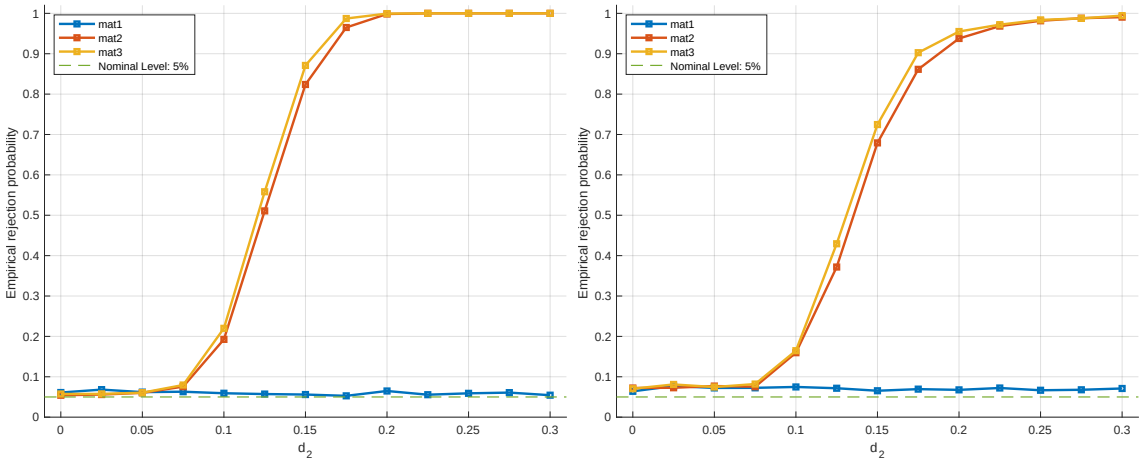
(b) **2nd Matricization**



(c) **3rd Matricization**

Figure 8: Testing the Number of Factors: Power Curves for Different DGPs

The power curves are plotted for testing the null hypothesis of 1 factor against the alternative of 2-5 factors, for cases of 3-dim tensor normal/t-distribution errors. We report the empirical power of testing on individual matricizations.



(a) **3 dim - Gaussian Errors** (b) **3 dim - t Errors**

6 Empirical Illustration: Dealing with Missing Firm Characteristics

Bryzgalova et al. (2024) drew attention to the phenomenon of missing data in firm characteristics, which are the cornerstone of academic research in finance. They provide a comprehensive analysis of missing data in firm characteristics, and show that patterns of missing characteristics vary substantially across characteristics. As a remedy they propose a statistical factor model for imputing missing values and investigate the impact of missingness on asset returns.[14]

Firm characteristics data are of the tensor type. Indeed, for each firm, one has across time a set of observed or missing characteristics. The statistical methods proposed by Bryzgalova et al. (2024) do not exploit the tensor data structure, while in this empirical application we do and we show that using our proposed tensor data models improve on their time series of cross-sectional models.

We conduct the empirical study using the dataset created by Freyberger, Neuhierl, and Weber (2020). The data cover 35 firm characteristics, which are described in Table A.2. The dataset is monthly and ranges from January 1966 to December 2020, and there are a total

[14]Factor models can also be used to impute the missing counterfactual potential outcomes; see Bai and Ng (2021).

of 13588 firms in the entire sample. This number is smaller than 22630 in Bryzgalova et al. (2024), because we only have access to the pre-cleaned dataset that has no missingness in the cross-section of characteristics, i.e., all characteristics exist for all firms at any time period. However, this difference does not affect the empirical results because the RMSE are calculated based on simulated (masked) missing characteristics both in this paper and in Bryzgalova et al. (2024). The results calculated based on cross-sectional factor model are also very close to those in Bryzgalova et al. (2024). This tensor dataset is highly unbalanced where only 15.78% of the data are not missing, and the average length of non-missing characteristics of all firms is 104.16 months.

Prior to estimating the tensor factor and the cross-sectional factor models, the raw characteristics are converted into rank quantiles within range $[-0.5, 0.5]$. This is done because: 1) different characteristics have different scales, and PCA has best performance when applied to data of similar scales, 2) centered rank-normalized characteristics are stationary in the cross-section and over time, and 3) it handles outliers. In practice, the predicted missing rank quantiles can be converted back to the original scale of the raw data.

Instead of fitting cross-sectional two-way factor models to each slice of the tensor (i.e., each matrix at period t), we fit three-way factor models on non-overlapping 5-year horizons. There is a considerable amount of parameter proliferation when estimating time series of cross-sections, thus we would expect a better performance from the parsimonious tensor-based approach. We compare our model to the traditional two-way approach in terms of imputation accuracy, and several other alternatives are also considered. See Section A.2 for detailed description of the model and other approaches being compared.

6.1 Evaluation Metrics

Similar to Bryzgalova et al. (2024), we consider the standard evaluation metrics for model prediction errors - RMSE (root-mean-squared errors). The aggregate RMSE is averaged over all firms, characteristics, and time periods, which is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{N_t}\sum_{i=1}^{N_t}(C_{i,t,\ell} - \hat{C}_{i,t,\ell})^2}.$$

We report both in-sample (IS) and out-of-sample (OOS) RMSE. For measuring out-of-sample RMSE, we randomly mask observed characteristics and calculate RMSE comparing predicted characteristics and the true masked ones. We take the same appraoch as Bryzgalova et al. (2024) to randomly mask 10% of the characteristics using two schemes:

1. Missing Completely at Random (MCAR). 10% of the characteristics are masked com-

pletely at random.

2. Block Missing. We again mask 10% of the characteristics. In order to capture the time-series pattern of the missingness, we randomly mask characteristics in blocks of one year. Based on the observational result of missing characteristics in Bryzgalova et al. (2024), 40% of the blocks are at the beginning.

We also report the R^2 that measures the explained variation of the model relative to the total variation, which is a transformation of the RMSE.

6.2 Number of Factors for Characteristics Tensors

Estimating both the tensor and cross-sectional factor models involves determining the number of factors. As noted in Bryzgalova et al. (2024), the optimal number of factors for the cross-sectional models is between 6 and 8, and they select 6 factors as their final model. In comparison, we use the test introduced in Section 4 to determine the number of factors for each dimension. The test is conducted with the following hypothesis:

$$\mathrm{H}_0: \leqslant k \text{ factors} \qquad \text{vs.} \qquad \mathrm{H}_1: \text{ the number of factors is } > k, \text{ but } \leqslant K.$$

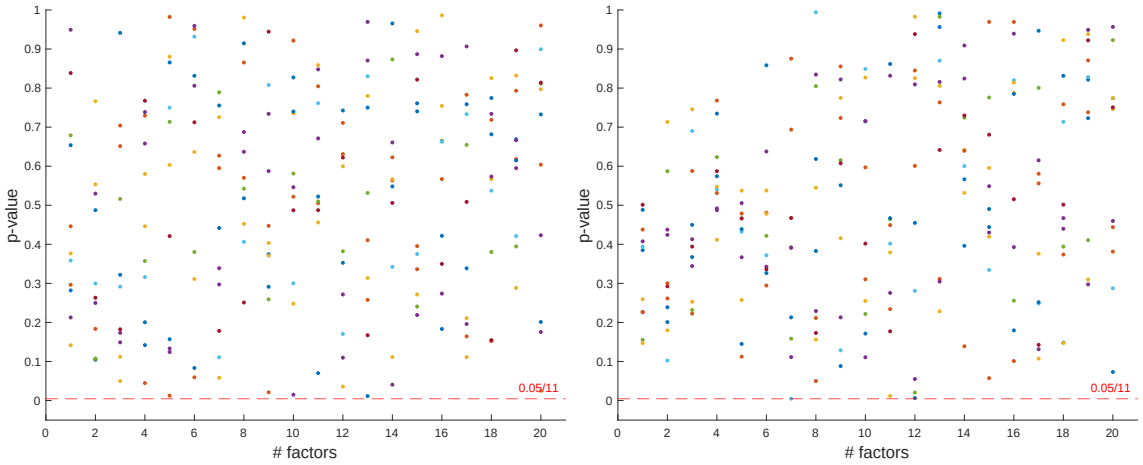
where $K = k + 1$ and we increase k from 1 to 20 (20 is the upper bound of the number of factors we deem reasonable). Since the test is conducted multiple times on the 5-year sub-samples, we use Bonferroni correction to reduce the probability of type I errors.

Figure 9 plots the p-values of test where the x-axis is test parameter k, and for each k there are 11 dots representing the p-value of each 5-year sub-sample. The test is conducted on each dimension and for each missing scheme - missing-completely-at-random (MCAR) and missing in blocks of one year (Block). The red dash line is the Bonferroni adjusted significance level. The test result suggests that there is large number of factors in the characteristic dataset (i.e., there is no sharp change in the eigenvalues detected). Therefore, we select 20 factors as the baseline model for optimal results[15], and we also present the results with 6 factors for a comparison with the cross-sectional model.

[15]The test result is aligned with the empirical finding that, when the number of factors is smaller than 20, the more factors the better the imputation performance. We also considered several alternative testing methods leading to the same conclusion.

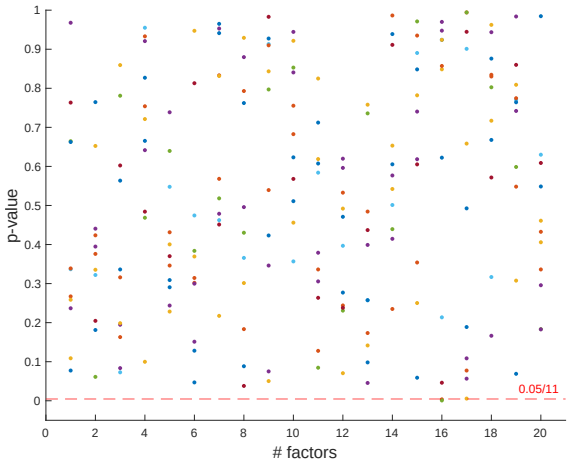
Figure 9: Testing the Number of Factors: p-values

This figure plots the p-values of test where the x-axis is test parameter k, and for each k there are 11 dots representing the p-value of each 5-year sub-sample. The test is conducted on each dimension and for each missing scheme - missing-completely-at-random (MCAR) and missing in blocks of one year (Block). The red dash line is the Bonferroni adjusted significance level.

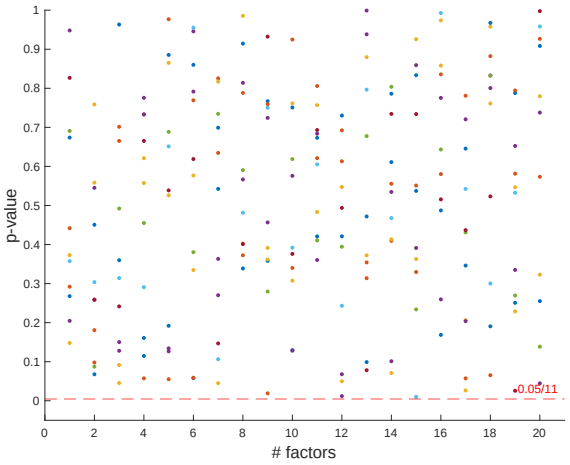


(a) **Dimension N - MCAR**

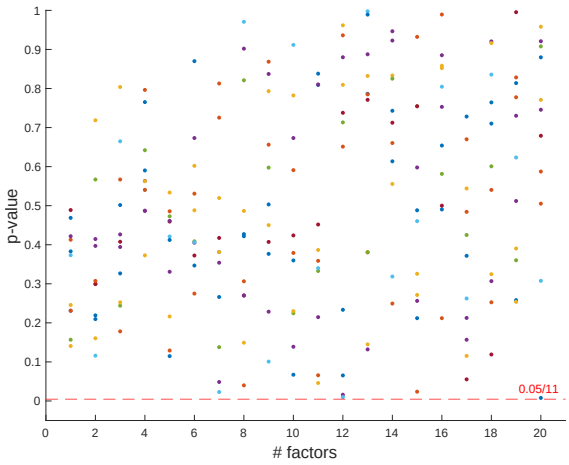
(b) **Dimension T- MCAR**



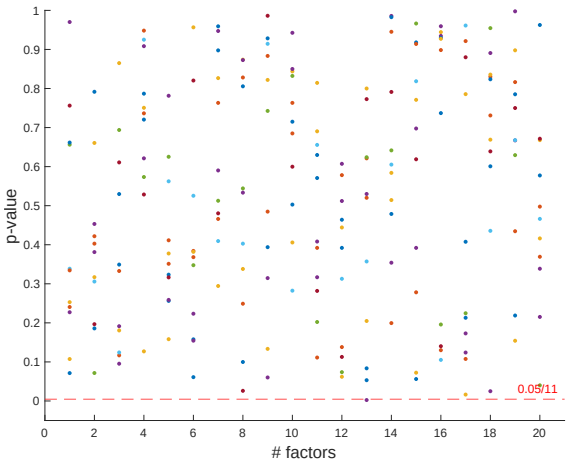
(c) **Dimension L - MCAR**



(d) **Dimension N - Block**



(e) **Dimension T - Block**



(f) **Dimension L - Block**

6.3 Imputation Results

In this subsection, we compare all the methods listed in Table A.1[16]. Some methods require backward and/or forward information, and they become inapplicable when this information is not available. Hence, we use the natural fallback approach, where we replace a method that requires time-series information by a pure lower-level method. For example, B-TPCA is replaced by TPCA and B-XS-ridge is replaced by XS-ridge when the previous characteristic is not observed.

Table 1 summarizes the main imputation results of different methods. We report the in-sample, out-of-sample missing completely at random (OOS MCAR), out-of-sample block missing (OOS Block) for all characteristics using the full sample period. The cross-sectional (XS-ridge) based methods and the Tensor PCA (TPCA) based method are estimated with different number of factors R with $R \in \{6, 20\}$. The first thing to note is that the cross-sectional median method produces imputation errors that are more than twice as large compared to previous value and autoregression methods, both in-sample and OOS missing completely at random. This demonstrates the importance of considering time-series dependency among firm characteristics. The previous value and autoregression methods are relatively reliable when the characteristics are not missing in blocks. However, these two method are no longer applicable when previous values are not available under the case of block missing, which is very common in firm characteristics. Therefore, the results of previous value and autoregression are very close to that of cross-sectional median. Without the combination of backward/forward information, the previous value and autoregression methods are better than XS-ridge in both in-sample and OOS MCAR, but are a lot worse when characteristics are missing in blocks. However, TPCA is consistently better than the standard approaches in both out-of-sample performances.

The results comparing TPCA and XS-ridge methods indicate that XS-ridge method is overfitting the characteristics data. XS-ridge method has lower RMSE and higher R^2 in-sample than TPCA with the same number of factors, while it falls behind TPCA out-of-sample under both missing patterns. The differences are most prominent with 20 factors and under OOS MCAR, where B-XS-ridge is about 40% higher than B-TPCA in terms of RMSE and about 10% lower in terms of R^2. The differences are smaller under block missingness because of lacking in time series information. However, the difference still demonstrates that tensor factor model is superior to cross-sectional factor model by incorporating the time dimension.

[16]BF-TPCA is not reported because it shows slight worse performance than B-TPCA, and B-TPCA is the considered the best model among all.

TPCA remains superior to cross-sectional factor model when combined with backward and forward information. Since TPCA more accurately predicts the missing characteristics than XS-ridge, the superiority is preserved after being combined with the same information. In addition, ALS further improves the imputation accuracy by providing a faster convergence rate.

7 Conclusion

Modern datasets are often multidimensional, extending beyond the 2-dimensional panel data structures used in traditional factor models and PCA. In this paper, we study a class of d-way factor models for high-dimensional tensor data, which are a natural generalization of traditional 2-way factor models. We demonstrate that d-way factor models can be estimated with a variation of PCA, which we call TPCA. This simple algorithm is optimal for the strong factor model. We also consider an improved iterative ALS algorithm which is optimal when the factors are moderately weak.

Additionally, we propose the first formal statistical test for the number of factors in a tensor factor model. Our findings indicate that the tensor factor model offers efficient dimensionality reduction compared to naively pooled traditional factor models. Simultaneously, the model is parsimonious with easily identifiable factors and loadings. These conclusions are supported by extensive simulation results. Lastly, we consider an empirical application to imputing missing firm characteristics.

Interesting applications of TPCA and our results could potentially include more refined panel data models with covariates, e.g., see Freeman (2022) and Beyhum and Gautier (2022a) as well as causal inference and imputations with tensor data; see Agarwal, Shah, and Shen (2020).

Table 1: Imputation Results of All Approaches

This table reports RMSE and R^2 of all methods listed in Table A.1 computed on the full sample. The results are reported for the in-sample, out-of-sample missing-completely-at-random (OOS MCAR), and out-of-sample missing in blocks of one year (OOS Block).

	In-Sample		OOS MCAR		OOS Block	
Method (# factors)	RMSE	R^2	RMSE	R^2	RMSE	R^2
TPCA (6)	0.1841	0.5471	0.1856	0.5402	0.1903	0.5183
ALS (6)	0.1778	0.5777	0.1792	0.5714	0.1848	0.5453
XS-ridge (6)	0.1657	0.6713	0.1980	0.5309	0.2001	0.5232
B-TPCA (6)	0.0904	0.8908	0.1044	0.8545	0.1847	0.5458
B-ALS (6)	0.0900	0.8918	0.1030	0.8585	0.1796	0.5708
B-XS-ridge (6)	0.0862	0.9110	0.1169	0.8363	0.1950	0.5469
BF-XS-ridge (6)	0.0789	0.9254	0.1220	0.8217	0.2001	0.5232
TPCA (20)	0.1279	0.7814	0.1329	0.7644	0.1474	0.7109
ALS (20)	0.1267	0.7855	0.1319	0.7679	0.1471	0.7123
XS-ridge (20)	0.0598	0.9572	0.1613	0.6888	0.1642	0.6790
B-TPCA (20)	0.0835	0.9070	0.0911	0.8893	0.1441	0.7239
B-ALS (20)	0.0830	0.9081	0.0906	0.8904	0.1437	0.7252
B-XS-ridge (20)	0.0317	0.9880	0.1287	0.8018	0.1621	0.6870
BF-XS-ridge (20)	0.0282	0.9905	0.1257	0.8110	0.1642	0.6790
Prev. Value	0.1282	0.8031	0.1522	0.7229	0.2814	0.0570
Median	0.2890	0.0000	0.2890	0.0000	0.2898	0.0000
AR(1)	0.1153	0.8407	0.1426	0.7566	0.2810	0.0596

References

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

K. M. Abadir and J. R. Magnus. Matrix algebra. *Cambridge Books*, 2005.

A. Agarwal, D. Shah, and D. Shen. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

E. Andreou, P. Gagliardini, E. Ghysels, and M. Rubin. Inference in group factor models with an application to mixed-frequency data. *Econometrica*, 87(4):1267–1305, 2019.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

A. Babii, E. Ghysels, and J. Pan. Tensor principal component analysis. *arXiv preprint arXiv:2212.12981*, 2022.

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

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

J. Bai and S. Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021.

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

M. Barigozzi, Y. He, L. Li, and L. Trapani. Statistical inference for large-dimensional tensor factor model by iterative projections. *arXiv preprint arXiv:2206.09800*, 2023.

J. Beyhum and E. Gautier. Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. *Journal of Business & Economic Statistics*, 41 (1):270–281, 2022a.

J. Beyhum and E. Gautier. Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. *Journal of Business & Economic Statistics*, 41 (1):270–281, 2022b.

P. Billingsley. *Probability and Measure*. Wiley, 1995.

S. Bonhomme and J.-M. Robin. Generalized non-parametric deconvolution with an application to earnings dynamics. *The Review of Economic Studies*, 77(2):491–533, 2010.

S. Bonhomme, K. Jochmans, and J.-M. Robin. Estimating multivariate latent-structure models. *The Annals of Statistics*, pages 540–563, 2016.

S. Bryzgalova, S. Lerner, M. Lettau, and M. Pelger. Missing financial data. *Review Financial Studies*, (forthcoming), 2024.

T. T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

E. Candes and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3): 283–319, 1970.

G. Chamberlain and M. Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1982.

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

B. Chen, Y. Han, and Q. Yu. Estimation and inference for cp tensor factor models. *arXiv preprint arXiv:2406.17278*, 2024.

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

F. Cunha, J. J. Heckman, and S. M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000a.

L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000b.

N. El Karoui. On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*, 2003.

A. T. Foerster, P.-D. G. Sarte, and M. W. Watson. Sectoral versus aggregate shocks: A structural factor analysis of industrial production. *Journal of Political Economy*, 119(1): 1–38, 2011.

H. Freeman. Multidimensional interactive fixed-effects. *arXiv preprint arXiv:2209.11691*, 2022.

H. Freeman and M. Weidner. Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics*, 237(1):105498, 2023.

J. Freyberger, A. Neuhierl, and M. Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

L. Gobillon and T. Magnac. Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551, 2016.

B. S. Graham. Network data. In *Handbook of Econometrics*, volume 7, pages 111–218. Elsevier, 2020.

R. Han, Y. Luo, M. Wang, and A. R. Zhang. Exact clustering in tensor block model: Statistical optimality and computational limit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1666–1698, 2022a.

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

Y. Han, R. Chen, D. Yang, and C.-H. Zhang. Tensor factor model estimation by iterative projection. *Annals of Statistics (forthcoming)*, 2025.

R. A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2013.

I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

T. Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 1996.

A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96(454):519–542, 2001.

T. G. Kolda. Multilinear operators for higher-order decompositions. Technical report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA, 2006.

V. Koltchinskii and K. Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157, 2017.

P. M. Kroonenberg and J. De Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45:69–97, 1980.

R. Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.

L. Lei and B. Ross. Estimating counterfactual matrix means with short panel data. *arXiv preprint arXiv:2312.07520*, 2023.

M. Lettau. High dimensional factor models with an application to mutual fund characteristics. *Working Paper*, 2022.

M. Lettau. High-dimensional factor models and the factor zoo. Technical report, National Bureau of Economic Research, 2023.

A. Lewbel. The rank of demand systems: theory and nonparametric estimation. *Econometrica: Journal of the Econometric Society*, pages 711–730, 1991.

E. Liu and A. Tsyvinski. A dynamic model of input-output networks. *Review of Economic Studies*, page rdae012, 2024.

A. Onatski. A formal statistical test for the number of factors in the approximate factor models. *Columbia University. Working Paper*, 2006.

A. Onatski. Testing hypotheses about the number of factors in large factor models. *Econometrica*, 77:1447–1479, 2009.

A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

A. Onatski. Uniform asymptotics for weak and strong factors. *University of Cambridge Working Paper*, 2022.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

M. H. Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012, 2006.

V. V. Petrov. Limit theorems of probability theory; sequences of independent random variables. Technical report, 1995.

M. Ricci and T. Levi-Civita. Méthodes de calcul différentiel absolu et leurs applications. *Mathematische Annalen*, 54(1):125–201, 1900.

E. Richard and A. Montanari. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.

S. A. Ross. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13: 341–360, 1976.

T. J. Sargent, C. A. Sims, et al. Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, 1:145–168, 1977.

A. Soshnikov. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *Journal of Statistical Physics*, 108:1033–1056, 2002.

C. Spearman. General intelligence objectively determined and measured. *American Journal of Psychology*, 15:107–197, 1904.

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

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

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31 (3):279–311, 1966.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

D. Wang, Y. Zheng, H. Lian, and G. Li. High-dimensional vector autoregressive time series modeling via tensor decomposition. *Journal of the American Statistical Association*, 117 (539):1338–1356, 2022.

G. S. Watson. *Statistics on spheres*. Wiley-Interscience, 1983.

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

R. Xiong and M. Pelger. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics* (forthcoming), 2022.

R. Xiong and M. Pelger. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301, 2023.

A. Zhang and D. Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

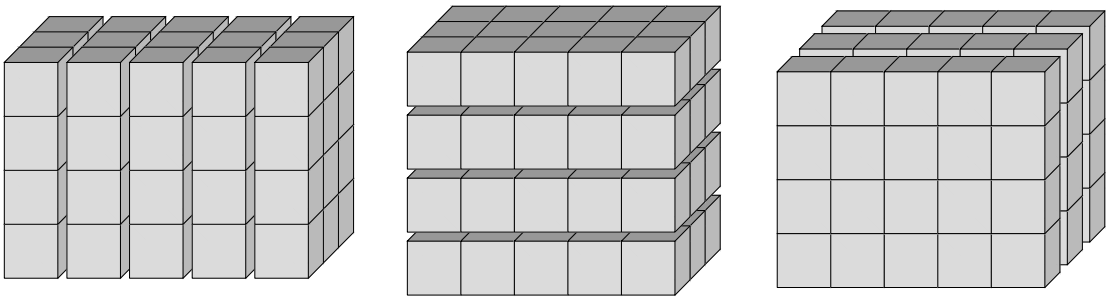
A. R. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80, 2022.

APPENDIX

A.1 Tensor Matricization: An Illustrative Example

Fixing all but two indices of a tensor, we obtain a matrix, called slice; see Figure A.1 for a graphical illustration.

Figure A.1: Lateral, horizontal, and frontal slices of a $4 \times 5 \times 3$ tensor



Example 1:

Let $\mathbf{Y}$ be a $3 \times 4 \times 2$ dimensional tensor of the following two frontal slices:

$$\mathbf{Y}_1 = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix} \qquad \mathbf{Y}_2 = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix}.$$

Then the mode-1, 2 and 3 matricizations of $\mathbf{Y}$ are respectively:

$$\mathbf{Y}_{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix}$$

$$\mathbf{Y}_{(2)} = \begin{bmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{bmatrix}$$

$$\mathbf{Y}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & \cdots & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & \cdots & 21 & 22 & 23 & 24 \end{bmatrix}$$

Example 2:

Let $\mathbf{Y}$ be a $3 \times 3 \times 3$ dimensional tensor of the following three frontal slices:

$$\mathbf{Y}_1 = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \quad \mathbf{Y}_2 = \begin{bmatrix} 10 & 13 & 16 \\ 11 & 14 & 17 \\ 12 & 15 & 18 \end{bmatrix} . \mathbf{Y}_3 = \begin{bmatrix} 19 & 22 & 25 \\ 20 & 23 & 26 \\ 21 & 24 & 27 \end{bmatrix} .$$

Then the mode-1, 2 and 3 matricization of $\mathbf{Y}$ are respectively:

$$\mathbf{Y}_{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 & 25 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 & 26 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 & 27 \end{bmatrix}$$

$$\mathbf{Y}_{(2)} = \begin{bmatrix} 1 & 2 & 3 & 10 & 11 & 12 & 19 & 20 & 21 \\ 4 & 5 & 6 & 13 & 14 & 15 & 22 & 23 & 24 \\ 7 & 8 & 9 & 16 & 17 & 18 & 25 & 26 & 27 \end{bmatrix}$$

$$\mathbf{Y}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 \\ 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 \end{bmatrix}$$

A.2 Imputation Methods for Missing Firm Characteristics

Bryzgalova et al. (2024) model the panel of firm characteristics for each month t as follows:

$$\mathbf{C}^t_{i,\ell} = \mathbf{F}^t_i \mathbf{\Lambda}^{t\top}_\ell + \mathbf{U}^t_{i,\ell} \quad \text{with } i = 1, \ldots, N_t \text{ and } \ell = 1, \ldots, L. \tag{A.1}$$

The above equation makes clear that we de facto have a tensor data set, along three dimensions i, ℓ and t. Hence, firm characteristics are a 3-way or third-order tensor with its three dimensions: (a) firms, (b) characteristics, and (c) time.

A.2.1 Tensor models for characteristics

For firm characteristics $\mathbf{C} \in \mathbb{R}^{N \times T \times L}$ with possibly missing entries, we instead consider the tensor factor model

$$\mathbf{C} = \mathbf{G} \times_1 \mathbf{M} \times_2 \mathbf{F} \times_3 \mathbf{\Lambda} + \mathbf{U} \equiv [\![\mathbf{G}; \mathbf{M}, \mathbf{F}, \mathbf{\Lambda}]\!] + \mathbf{U}, \tag{A.2}$$

where $\mu_r \in \mathbb{R}^N$ are the loadings corresponding to the firms dimension, $\lambda_r \in \mathbb{R}^L$ are the loadings corresponding to the characteristics dimension, and $f_r \in \mathbb{R}^T$ are the factors. We collect the λ_r, μ_r, and f_r in matrices $\mathbf{\Lambda} \in \mathbb{R}^{I \times R}$, $\mathbf{M} \in \mathbb{R}^{L \times R}$, and $\mathbf{F} \in \mathbb{R}^{T \times R}$. The loadings and factors are of unit norm, and $\mathbf{G}$ is the core tensor.

Before discussing our estimator, let us first review the approach of Bryzgalova et al. (2024) to estimate equation (A.1). They follow Xiong and Pelger (2022), and compute the loadings $\mathbf{\Lambda}^t$ as eigenvectors of the estimated characteristic covariance matrix

$$\hat{\mathbf{\Sigma}}^{\text{XS},t}_{\ell,p} = \frac{1}{Q^t_{\ell,p}} \sum_{i \in Q^t_{\ell,p}} \mathbf{C}^t_{\ell,i} \mathbf{C}^t_{p,i}, \tag{A.3}$$

where $Q^t_{\ell,p}$ is the set of all stocks that are observed for the two characteristics ℓ and p at time t. The superscript XS, t refers to the time series of cross-sections approach. For each t, the cross-sectional factors are estimated by a regression on the estimated loadings $\hat{\mathbf{\Lambda}}^t$:

$$\hat{\mathbf{F}}^t_i = \left(\frac{1}{L} \sum_{\ell=1}^{L} W^t_{i,\ell} \hat{\mathbf{\Lambda}}^t_\ell (\hat{\mathbf{\Lambda}}^t_\ell)^\top \right)^{-1} \left(\frac{1}{L} \sum_{\ell=1}^{L} W^t_{i,\ell} \hat{\mathbf{\Lambda}}^t_\ell (\mathbf{C}^t_{\ell,i})^\top \right), \tag{A.4}$$

where $W^t_{i,\ell} = 1$ if characteristic ℓ is observed for stock i at time t and $W^t_{i,\ell} = 0$ otherwise.

We now turn our attention to the estimation of the tensor factor model (A.2). Let $\mathbf{C}^{(j)}$ be the matricization of firm characteristics along the dimension $j \in \{1, 2, 3\}$, and $\mathbf{C}^{(j)}$ is of size $N_j \times \prod_{k \neq j} N_k$. The factor matrices $\mathbf{M}, \mathbf{F}, \mathbf{\Lambda}$ can be estimated by applying PCA to the covariance matrix of the matricized tensor. Due to missingness in the firm characteristics data, we use the approach of Xiong and Pelger (2023) and calculate the covariance matrix based on partially observed data, denoted by $\hat{\Sigma}^{(j)}$, as

$$\hat{\mathbf{\Sigma}}^{(j)}_{\ell,p} = \frac{1}{Q_{\ell,p}} \sum_{i \in Q_{\ell,p}} \mathbf{C}^{(j)}_{\ell,i} \mathbf{C}^{(j)}_{p,i},$$

where $Q_{\ell,p}$ is the set of all columns that are observed for the ℓ^{th} and the p^{th} rows of $\mathbf{C}^{(j)}$.[17] Then the core tensor $\mathbf{G}$ can be estimated by step 5 of the algorithm 1 where the missing values in $\mathbf{Y}$ are filled with $0s$ (cross-sectional median). The ALS algorithm 2 can also be used to estimate the same model A.2 in the same fashion.

To highlight the difference between the time series of cross-sections and our tensor ap-

[17] Xiong and Pelger (2023) provides inferential theory for estimating latent factor models from panel data with missing observations based on PCA. Matricized tensors are essentially panel data.

proach, we can write the former estimator as follows:

$$\hat{\mathbf{C}}_{i,\ell}^{\mathrm{XS},t} = \sum_{r=1}^{R} \hat{\sigma}_{r,t} \hat{\mu}_{r,i,t} \hat{\lambda}_{r,\ell,t}, \tag{A.5}$$

where all parameters $\hat{\sigma}_{r,t}$, $\hat{\mu}_{r,i,t}$, and $\hat{\lambda}_{r,\ell,t}$ have subscript t because the estimation is done in each period in time. In contrast, the missing characteristics are predicted by TPCA as

$$\hat{\mathbf{C}}^{\mathrm{TPCA}} = \hat{\mathbf{G}} \times_1 \hat{\mathbf{M}} \times_2 \hat{\mathbf{F}} \times_3 \hat{\mathbf{\Lambda}}. \tag{A.6}$$

Comparing equation (A.5) with (A.6) we clear see that there is a considerable amount of parameter proliferation when estimating time series of cross-sections. Therefore, we would expect a better performance from the parsimonious tensor-based approach.

A.2.2 Backward and Forward Information

Firm characteristics have strong dependency with their previous values, and filling the missing values with its previous value is one of the most widely used methods for missing imputations. Inspired by Bryzgalova et al. (2024), we combine the tensor factor model with backward information by a cross-sectional regression. We define the covariates in the regression by

$$\mathbf{X}_{i,t,\ell}^{\text{B-TPCA}} = \left(\sum_{r=1}^{R} \hat{\sigma}_r \hat{\mu}_{r,i} \hat{f}_{r,t} \hat{\lambda}_{r,\ell} \quad \mathbf{C}_{i,\ell}^{t-1} \quad \hat{\mathbf{e}}_{i,\ell}^{t-1} \right), \tag{A.7}$$

where $\hat{\mathbf{e}}_{i,\ell}^{t} = \mathbf{C}_{i,t,\ell} - \hat{\mathbf{C}}_{i,t,\ell}^{\mathrm{TPCA}}$ is the residual from TPCA, and we estimate the model

$$\mathbf{C}_{i,t,\ell} = \left(\beta^{\ell,t,\text{B-TPCA}}\right)^{\top} \left(\sum_{r=1}^{R} \hat{\sigma}_r \hat{\mu}_{r,i} \hat{f}_{r,t} \hat{\lambda}_{r,\ell} \quad \mathbf{C}_{i,\ell}^{t-1} \quad \hat{\mathbf{e}}_{i,\ell}^{t-1} \right) \tag{A.8}$$

with a cross-sectional regression on the partially observed data at time t for characteristic ℓ as follows:

$$\hat{\beta}^{\ell,t,\text{B-TPCA}} = \left(\sum_{i=1}^{N_t} \mathbf{W}_i^{t,\ell} \mathbf{X}_{i,t,\ell}^{\text{B-TPCA}} (\mathbf{X}_{i,t,\ell}^{\text{B-TPCA}})^{\top} \right) \left(\sum_{i=1}^{N_t} \mathbf{W}_i^{t,\ell} \mathbf{X}_{i,t,\ell}^{\text{B-TPCA}} \mathbf{C}_{i,t,\ell} \right),$$

where $\mathbf{W}_i^{t,\ell} = 1$ if $\mathbf{X}_{i,t,\ell}^{\text{B-TPCA}}$ and $C_{i,t,\ell}$ are both observed and 0 otherwise. The missing characteristics are predicted as

$$\hat{\mathbf{C}}_{i,t,\ell}^{\text{B-TPCA}} = (\hat{\beta}^{\ell,t,\text{B-TPCA}})^{\top} \mathbf{X}_{i,t,\ell}^{\text{B-TPCA}}.$$

Table A.1: Different Imputation Methods

Method	Estimation
Tucker (TPCA/ALS)	$\hat{\mathbf{C}}^{\text{TPCA}} = \hat{\mathbf{G}} \times_1 \hat{\mathbf{M}} \times_2 \hat{\mathbf{F}} \times_3 \hat{\mathbf{\Lambda}}$
Backward-Forward-TPCA (BF-TPCA)	$\hat{\mathbf{C}}^{\text{BF-TPCA}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{BF-TPCA}})^\top \left(\hat{\mathbf{C}}^{\text{TPCA}}_{i,t,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell} \quad \mathbf{C}^{t+1}_{i,\ell} \quad \hat{\mathbf{e}}^{t+1}_{i,\ell}\right)$
Backward-TPCA (B-TPCA)	$\hat{\mathbf{C}}^{\text{B-TPCA}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{B-TPCA}})^\top \left(\hat{\mathbf{C}}^{\text{TPCA}}_{i,t,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell}\right)$
Cross-sectional (XS)	$\hat{\mathbf{C}}^{\text{XS}}_{i,t,\ell} = \sum_{r=1}^{R} \hat{\sigma}_{r,t}\hat{\mu}_{r,i,t}\hat{\lambda}_{r,\ell,t}$
Backward-Forward-XS (BF-XS)	$\hat{\mathbf{C}}^{\text{BF-XS}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{BF-XS}})^\top \left(\hat{\mathbf{C}}^{\text{XS}}_{i,t,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell} \quad \mathbf{C}^{t+1}_{i,\ell} \quad \hat{\mathbf{e}}^{t+1}_{i,\ell}\right)$
Backward-XS (B-XS)	$\hat{\mathbf{C}}^{\text{B-XS}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{B-XS}})^\top \left(\hat{\mathbf{C}}^{\text{XS}}_{i,t,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell}\right)$
Autoregression (AR)	$\hat{\mathbf{C}}^{\text{AR}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{AR}})^\top \mathbf{C}^{t-1}_{i,\ell}$
Previous Value (PV)	$\hat{\mathbf{C}}^{\text{PV}}_{i,t,\ell} = \mathbf{C}^{t-1}_{i,\ell}$
Cross-sectional median (median)	$\hat{\mathbf{C}}^{\text{median}}_{i,t,\ell} = 0$

The B-TPCA model can be extended to include forward information at time $t+1$, where we define the covariates

$$\mathbf{X}^{\text{BF-TPCA}}_{i,t,\ell} = \left(\sum_{r=1}^{R} \hat{\sigma}_r \hat{\mu}_{r,i} \hat{f}_{r,t} \hat{\lambda}_{r,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell} \quad \mathbf{C}^{t+1}_{i,\ell} \quad \hat{\mathbf{e}}^{t+1}_{i,\ell}\right), \tag{A.9}$$

and estimate the model

$$\mathbf{C}_{i,t,\ell} = \left(\beta^{\ell,t,\text{BF-TPCA}}\right)^\top \left(\sum_{r=1}^{R} \hat{\sigma}_r \hat{\mu}_{r,i} \hat{f}_{r,t} \hat{\lambda}_{r,\ell} \quad \mathbf{C}^{t-1}_{i,\ell} \quad \hat{\mathbf{e}}^{t-1}_{i,\ell} \quad \mathbf{C}^{t+1}_{i,\ell} \quad \hat{\mathbf{e}}^{t+1}_{i,\ell}\right). \tag{A.10}$$

The model can also be estimated with a cross-sectional regression on the partially observed data at time t for characteristic ℓ as follows:

$$\hat{\beta}^{\ell,t,\text{BF-TPCA}} = \left(\sum_{i=1}^{N_t} \mathbf{W}_i^{t,\ell} \mathbf{X}^{\text{BF-TPCA}}_{i,t,\ell} (\mathbf{X}^{\text{BF-TPCA}}_{i,t,\ell})^\top\right) \left(\sum_{i=1}^{N_t} \mathbf{W}_i^{t,\ell} \mathbf{X}^{\text{BF-TPCA}}_{i,t,\ell} \mathbf{C}_{i,t,\ell}\right), \tag{A.11}$$

where $\mathbf{W}_i^{t,\ell} = 1$ if $\mathbf{X}^{\text{BF-TPCA}}_{i,t,l}$ and $C_{i,t,l}$ are both observed and 0 otherwise. The missing characteristics are predicted as

$$\hat{\mathbf{C}}^{\text{BF-TPCA}}_{i,t,\ell} = (\hat{\beta}^{\ell,t,\text{BF-TPCA}})^\top \mathbf{X}^{\text{BF-TPCA}}_{i,t,\ell}.$$

A.2.3 Alternative Imputation Methods

Bryzgalova et al. (2024) consider various imputation methods for firm characteristics including pure cross-sectional factor model (XS), cross-sectional with backward information

(B-XS), cross-sectional with backward and forward information (BF-XS). Essentially, these methods replace the TPCA estimation with XS estimation in the models discussed previously. Due to the fact that cross-sectional models fit each time period separately, these methods can easily result in overfitting. For this reason Bryzgalova et al. (2024) consider a regularized version of the cross-sectional factor model, where they replace the regression in (A.4) with ridge regression:

$$\hat{\mathbf{F}}_i^{t,\gamma} = \left(\frac{1}{L} \sum_{\ell=1}^{L} W_{i,\ell}^t \hat{\mathbf{\Lambda}}_\ell^t (\hat{\mathbf{\Lambda}}_\ell^t)^\top + \gamma I_K \right)^{-1} \left(\frac{1}{L} \sum_{\ell=1}^{L} W_{i,\ell}^t \hat{\mathbf{\Lambda}}_\ell^t (\mathbf{C}_{\ell,i}^t)^\top \right). \tag{A.12}$$

In practice, the most commonly used methods for missing value imputation also include cross-sectional median, previous value, autoregression (AR) of order 1. All the methods are summarized in Table A.1.

A.3 Supplementary Tables

This section details the firm characteristics included in the data set. The data cover 35 firm characteristics, which are described in Table A.2. See Freyberger et al. (2020) for detailed construction of each characteristic. The dataset is monthly and ranges from January 1966 to December 2020, and there are a total of 13,588 firms in the entire sample.

Table A.2: Firm Characteristics

(1)	A2ME	Assets to market cap
(2)	AT	Total assets
(3)	ATO	Net sales over lagged net operating assets
(4)	BEME	Book to market ratio
(5)	Beta	CAPM beta
(6)	C	Ratio of cash and short-term investments to total assets
(7)	CTO	Capital turnover
(8)	D2A	Capital intensity
(9)	DPI2A	Change in property, plants, and equipment
(10)	E2P	Earnings to price
(11)	FC2Y	Fixed costs to sales
(12)	Free CF	Cash flow to book value of equity
(13)	Idio vol	Idiosyncratic volatility
(14)	Investment	Percentage year-on-year growth rate in total assets
(15)	Lev	Leverage
(16)	LME	Total market capitalization
(17)	LTurnover	Last month's volume over shares outstanding
(18)	NOA	Net operating assets
(19)	OA	Operating accruals
(20)	OL	Operating leverage
(21)	PCM	Price-to-cost margin
(22)	PM	Profit margin
(23)	Prof	profitability
(24)	Q	Tobin's Q
(25)	Rel to High	Closeness to 52-week high
(26)	RNA	Return on net operating assets
(27)	ROA	Return-on-assets
(28)	ROE	Return-on-equity
(29)	r_{12-2}	Momentum
(30)	r_{12-7}	Intermediate momentum
(31)	r_{2-1}	Short-term reversal
(32)	r_{36-13}	Long-term reversal
(33)	S2P	Sales-to-price ratio
(34)	SGA2S	SG&A to sales
(35)	Spread	Bid-ask spread

A.4 Proofs

This appendix contains proofs of all results with several supplementary lemmas. The first lemma provides some useful properties of Kronecker products that will be used repeatedly in proofs.

Lemma A.4.1. *If $\Lambda_j, M_j \in \mathbb{R}^{N_j \times R_j}$ for $j \leqslant d$, then*

(a) $\left\| \bigotimes_{j=1}^{d} \Lambda_j \right\|_{\text{op}} = \prod_{j=1}^{d} \|\Lambda_j\|_{\text{op}}$.

(b) $\left(\bigotimes_{j=1}^{d} \Lambda_j \right)^\top \left(\bigotimes_{j=1}^{d} M_j \right) = \bigotimes_{j=1}^{d} (\Lambda_j^\top M_j)$.

Proof. The proof is based on the following three properties of Kronecker product:

1. $(A \otimes B)^\top = A^\top \otimes B^\top$.
2. $(A_1 \otimes A_2)(B_1 \otimes B_2) = (A_1 B_1) \otimes (A_2 B_2)$.
3. $(\alpha A_1 + \beta A_2) \otimes B = \alpha A_1 \otimes B + \beta A_2 \otimes B$, where $\alpha, \beta \in \mathbb{R}$.

These properties can be found in Abadir and Magnus (2005), Exercises 10.3 and 10.7. Then (b) follows from

$$\left(\bigotimes_{j=1}^{d} \Lambda_j \right)^\top \left(\bigotimes_{j=1}^{d} M_j \right) = \left(\bigotimes_{j=1}^{d} \Lambda_j^\top \right) \left(\bigotimes_{j=1}^{d} M_j \right) = \bigotimes_{j=1}^{d} (\Lambda_j^\top M_j).$$

For (a), using the SVD decomposition of $\Lambda_j = \sum_{r_j=1}^{R_j} \sigma_{j,r_j} u_{j,r_j} v_{j,r_j}^\top$

$$\begin{aligned} \bigotimes_{j=1}^{d} \Lambda_j &= \sum_{r_1,\ldots,r_d} \sigma_{1,r_1} \ldots \sigma_{d,r_d} \bigotimes_{j=1}^{d} u_{j,r_j} v_{j,r_j}^\top \\ &= \sum_{r_1,\ldots,r_d} \sigma_{1,r_1} \ldots \sigma_{d,r_d} \left(\bigotimes_{j=1}^{d} u_{j,r_j} \right) \left(\bigotimes_{j=1}^{d} v_{j,r_j} \right)^\top, \end{aligned}$$

where we use properties 1-3. This shows that $\bigotimes_{j=1}^{d} \Lambda_j$ has singular values $\{\prod_{j=1}^{d} \sigma_{j,r_j} : r_j \leqslant R_j, j \leqslant d\}$, which implies (a):

$$\left\| \bigotimes_{j=1}^{d} \Lambda_j \right\|_{\text{op}} = \max_{r_j \leqslant R_j, j \leqslant d} \prod_{j=1}^{d} \sigma_{j,r_j} = \prod_{j=1}^{d} \max_{r_j \leqslant R_j} \sigma_{j,r_j} = \prod_{j=1}^{d} \|\Lambda_j\|_{\text{op}}.$$

□

Proof of Proposition 3.1. By Lemma A.4.1,

$$\left(\bigotimes_{k\neq j}\Lambda_k\right)^\top\left(\bigotimes_{k\neq j}\Lambda_k\right)=\bigotimes_{k\neq j}(\Lambda_j^\top\Lambda_j)=\bigotimes_{k\neq j}I_{R_k}=I_{\prod_{k\neq j}R_k},$$

where the second equality follows under Assumption 3.1, (i). □

Proof of Theorem 3.1. Recall that $\hat{\Lambda}_j$ is $N_j\times R_j$ matrix of R_j leading left singular vectors of

$$\mathbf{Y}_{(j)}=\Lambda_j\mathbf{G}_{(j)}\left(\bigotimes_{k\neq j}\Lambda_k\right)^\top+\mathbf{U}_{(j)}.$$

In what follows, let $c,C>0$ be some numerical constant with values that can be different in different places. By Cai and Zhang (2018), equation (1.17) and the equation below it, we have

$$\Pr\left(\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\text{op}}^2>C\frac{\left(\delta^2+\prod_{l\neq j}N_l\right)\|\Lambda_j^\top\mathbf{Y}_{(j)}P_{\mathbf{Y}_{(j)}^\top\Lambda_j}\|_{\text{op}}^2}{\delta^4}\right)\leqslant Ce^{-\frac{c\delta^4}{\delta^2+\prod_{l\neq j}N_l}},$$

where

$$\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\text{op}}=\max_{1\leqslant r\leqslant R_j}\cos^{-1}(\hat{\sigma}_{r,j})$$

and $\hat{\sigma}_{1,j}\geqslant\cdots\geqslant\hat{\sigma}_{j,R_j}\geqslant 0$ are the singular values of $\Lambda_j^\top\hat{\Lambda}_j$.

By Cai and Zhang (2018), equation (1.16), with $x=C\sqrt{N_j}$, we also have

$$\begin{aligned}\Pr\left(\|\Lambda_j^\top\mathbf{Y}_{(j)}P_{\mathbf{Y}_{(j)}^\top\Lambda_j}\|_{\text{op}}\geqslant C\sqrt{N_j}\right)&\leqslant Ce^{CN_j-c\min\{C^2N_j,C\sqrt{N_j}\sqrt{\delta^2+\prod_{l\neq j}^dN_l}\}}+Ce^{-c(\delta^2+\prod_{l\neq j}N_l)}\\&\lesssim e^{-cN_j}+e^{-c(\delta^2+\prod_{l\neq j}N_l)}\lesssim Ce^{-cN_j},\end{aligned}$$

where we use $\delta^2\gtrsim\prod_{j=1}^d\sqrt{N_j}+\max_{1\leqslant j\leqslant d}N_j$, cf. Assumption 3.3.

Therefore,

$$\|\hat{\Lambda}_jO_j-\Lambda_j\|_{\text{F}}^2\leqslant 2R_j\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\text{op}}^2\lesssim\frac{R_j\left(\delta^2+\prod_{l\neq j}N_l\right)N_j}{\delta^4}$$

with probability at least $1-Ce^{-cN_j}$. □

The following Lemma provides some useful properties of metricized mode-j products.

Lemma A.4.2. *For tensors $\mathbf{Y}\in\mathbb{R}^{N_1\times\cdots\times N_d}$ and $\mathbf{G}\in\mathbb{R}^{R_1\times\cdots\times R_d}$, and matrices $\Lambda_l\in\mathbb{R}^{N_l\times R_l}$ with $1\leqslant l\leqslant d$, we have*

1. $\left(\mathbf{Y} \times_{l\neq j} \Lambda_l^\top\right)_{(j)} = \mathbf{Y}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right)$ *for all* $1 \leqslant j \leqslant d$.

2. $\left(\mathbf{G} \times_{l=1}^{d} \Lambda_l\right)_{(j)} = \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right)^\top$ *for all* $1 \leqslant j \leqslant d$.

Proof. These properties can be verified by looking at the entries of corresponding tensors; see also Proposition 3.7 in Kolda (2006). □

Lemma A.4.3. *Suppose that* $\mathbf{U} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ *is a tensor with i.i.d. mean-zero subgaussian entries. Then*

$$\left\| \mathbf{U}_{(j)} \bigotimes_{l\neq j} \Lambda_l \right\|_{\text{op}} \lesssim \sqrt{N_j} + \sqrt{\prod_{l\neq j} R_l}$$

and

$$\sup_{\|\Lambda_l\|_{\text{op}} \leqslant 1, \forall l \neq j} \left\| \mathbf{U}_{(j)} \bigotimes_{l\neq j} \Lambda_l \right\|_{\text{op}} \lesssim \sqrt{N_j} + \sqrt{\prod_{l\neq j} R_l} + \sum_{l\neq j} \sqrt{N_l R_l}$$

with probability at least $1 - e^{-c \min_{1\leqslant j\leqslant d} N_j}$ *for some* $c > 0$.

Proof. See Han, Luo, Wang, and Zhang (2022a), Lemma 9. □

Proof of Theorem 3.2. By Lemma A.4.2, the mode-j matricization of $\hat{\mathbf{Y}}^{(k)}$ at step k computed in Algorithm 2 is

$$\begin{aligned}
\hat{\mathbf{Y}}_{(j)}^{(k)} &= \left(\mathbf{Y} \bigtimes_{l\neq j} \hat{\Lambda}_l^{(k-1)\top} \right)_{(j)} \\
&= \mathbf{Y}_{(j)} \bigotimes_{l\neq j} \hat{\Lambda}_l^{(k-1)} \\
&= \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l \right)^\top \bigotimes_{l\neq j} \hat{\Lambda}_l^{(k-1)} + \mathbf{U}_{(j)} \bigotimes_{l\neq j} \hat{\Lambda}_l^{(k-1)} \\
&= \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} O_l \right)^\top + \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} O_l \right)^\top \bigotimes_{l\neq j} \left\{ O_l \Lambda_l^\top \hat{\Lambda}_l^{(k-1)} - I_{R_l} \right\} + \mathbf{U}_{(j)} \bigotimes_{l\neq j} \hat{\Lambda}_l^{(k-1)},
\end{aligned} \tag{A.13}$$

where the last line follows by Lemma A.4.1. Consider the $\sin\Theta$ distance between the singular subspaces

$$\|\sin\Theta(\hat{\Lambda}_j, \Lambda_j)\|_{\text{op}} = \max_{1\leqslant r\leqslant R_j} \cos^{-1}(\hat{\sigma}_{r,j}),$$

where $\hat{\sigma}_{1,j} \geqslant \cdots \geqslant \hat{\sigma}_{j,R_j} \geqslant 0$ are the singular values of $\Lambda_j^\top \hat{\Lambda}_j$.

Since $\hat{\Lambda}_j^{(k)}$ and Λ_j are the R_j leading left-singular vectors of $\hat{\mathbf{Y}}_{(j)}^{(k)}$ and the first term in the last line of equation (A.13) respectively, by Wedin's perturbation theorem, see Wedin

(1972),

$$\begin{aligned}\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\mathrm{op}} &\lesssim \frac{1}{\delta}\left\|\Lambda_j\mathbf{G}_{(j)}\bigotimes_{l\neq j}\left\{\Lambda_l^\top\hat{\Lambda}_l^{(k-1)}-O_l^\top\right\}+\mathbf{U}_{(j)}\bigotimes_{l\neq j}\hat{\Lambda}_l^{(k-1)}\right\|_{\mathrm{op}}\\ &\leqslant \frac{1}{\delta}\left\|\Lambda_j\mathbf{G}_{(j)}\bigotimes_{l\neq j}\left\{\Lambda_l^\top\hat{\Lambda}_l^{(k-1)}-O_l^\top\right\}\right\|_{\mathrm{op}}+\frac{1}{\delta}\left\|\mathbf{U}_{(j)}\bigotimes_{l\neq j}\hat{\Lambda}_l^{(k-1)}\right\|_{\mathrm{op}}.\end{aligned}$$

Under Assumption 3.2, by Lemma A.4.3

$$\begin{aligned}\left\|\mathbf{U}_{(j)}\bigotimes_{l\neq j}\hat{\Lambda}_l^{(k-1)}\right\|_{\mathrm{op}} &\leqslant \sup_{\|\Lambda_l\|_{\mathrm{op}}\leqslant 1,\forall l\neq j}\left\|\mathbf{U}_{(j)}\bigotimes_{l\neq j}\Lambda_l\right\|_{\mathrm{op}}\\ &\lesssim \sum_{j=1}^{d}\sqrt{N_j}+\sqrt{\prod_{l\neq j}R_l}+\sum_{l\neq j}\sqrt{N_lR_l}\\ &\lesssim d\sqrt{N}+R^{(d-1)/2}+(d-1)\sqrt{NR},\end{aligned}$$

where we use $N_j\sim N$ and define $R=\max_{1\leqslant j\leqslant d}R_j$.

Under Assumptions 3.1 and 3.3, by Lemma A.4.1, we also have

$$\begin{aligned}\frac{1}{\delta}\left\|\Lambda_j\mathbf{G}_{(j)}\bigotimes_{l\neq j}\left\{\Lambda_l^\top\hat{\Lambda}_l^{(k-1)}-O_l^\top\right\}\right\|_{\mathrm{op}} &\leqslant \frac{\sigma_{1,j}}{\delta}\left\|\bigotimes_{l\neq j}\left\{\Lambda_l^\top\hat{\Lambda}_l^{(k-1)}-O_l^\top\right\}\right\|_{\mathrm{op}}\\ &\lesssim \prod_{l\neq j}\left\|\Lambda_l^\top(\hat{\Lambda}_l^{(k-1)}O_l-\Lambda_l)\right\|_{\mathrm{op}}\\ &\leqslant \prod_{l\neq j}\left\|\hat{\Lambda}_l^{(k-1)}O_l-\Lambda_l\right\|_{\mathrm{F}}\end{aligned}$$

This shows that

$$\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\mathrm{op}}\lesssim\prod_{l\neq j}\left\|\hat{\Lambda}_l^{(k-1)}O_l-\Lambda_l\right\|_{\mathrm{F}}+\frac{d\sqrt{N}+R^{(d-1)/2}+(d-1)\sqrt{RN}}{\delta}. \tag{A.14}$$

From the proof of Theorem 3.1, we know that the naive TPCA computed with Algorithm 1 satisfies

$$E_0\lesssim\frac{\sqrt{N}}{\delta}+\frac{N^{d/2}}{\delta^2}$$

with probability at least $1-Ce^{-cN}$, provided that $N_j\sim N$ for all $1\leqslant j\leqslant d$. Since $\delta^{-4}N^d=o(1)$, we have $E_0=o(1)$ for a sufficiently large N. Let $E_k:=\max_{1\leqslant j\leqslant d}\|\sin\Theta(\hat{\Lambda}_j,\Lambda_j)\|_{\mathrm{op}}$ be

the largest error across all matricizations at iterations $k = 0, 1, \ldots, \bar{k}$. Then

$$E_k \lesssim E_{k-1}^{d-1} + \frac{d\sqrt{N} + \sqrt{R^{d-1}} + (d-1)\sqrt{RN}}{\delta}$$

with probability at least $1 - Ce^{-cN}$. If $E_{k-1} = o(1)$, then $E_k = o(1)$ since $N^{d/4}/\delta = o(1)$ and $R = O(1)$. This shows that $E_k = o(1), \forall k = 0, 1, \ldots, \bar{k}$ because $E_0 = o(1)$. Therefore, we can bound the ALS error for a sufficiently large N as

$$E_k \leqslant 0.5E_{k-1} + c\frac{\sqrt{N}}{\delta}.$$

Iterating this inequality, we obtain

$$E_k \leqslant 2c\frac{\sqrt{N}}{\delta} + 0.5^k E_0 \lesssim \frac{\sqrt{N}}{\delta} + 0.5^k \left(\frac{\sqrt{N}}{\delta} + \frac{N^{d/2}}{\delta^2} \right) \lesssim \frac{\sqrt{N}}{\delta},$$

where the last inequality follows provided that $k \geqslant \log(N^{(d-1)/2}/\delta)/\log 2$. This shows that

$$\|\hat{\Lambda}_j^{(\bar{k})} O_j - \Lambda_j\|_{\mathrm{F}}^2 \leqslant 2R_j \|\sin\Theta(\hat{\Lambda}_j, \Lambda_j)\|_{\mathrm{op}}^2 \lesssim E_k^2 \lesssim \frac{N}{\delta^2}.$$

□

Next, we prove a central limit theorem for quadratic forms which is an important ingredient in the proof of Theorem 3.3.

Lemma A.4.4. *Suppose that* $\mathbf{U} = \{u_{i,j} : 1 \leqslant i \leqslant N_1, 1 \leqslant j \leqslant N_2\}$ *is a matrix of i.i.d. random variables such that* $\mathbb{E}(u_{i,j}) = 0$, $\mathrm{Var}(u_{i,j}) = \sigma^2$, *and* $\mathbb{E}|u_{i,j}|^{2+\delta} < \infty$ *for some* $\delta > 0$. *Let* $(x_r)_{r=1}^R$ *and* $(y_r)_{r=1}^R$ *be two orthonormal sets in* $\mathbb{R}^{N_1}$ *and* $\mathbb{R}^{N_2}$ *such that* $\max_r \|x_r\|_\infty = o(1)$ *or* $\max_r \|y_r\|_\infty = o(1)$. *Then*

$$(x_r^\top \mathbf{U} y_k)_{1 \leqslant r,k \leqslant R} \xrightarrow{d} Z,$$

where Z *is an* $R \times R$ *matrix with i.i.d.* $N(0, \sigma^2)$ *entries.*

Proof. According to the Cramér-Wold argument, it is enough to show that

$$\sum_{r,k} t_{r,k} x_r^\top \mathbf{U} y_k \xrightarrow{d} \sum_{r,k} t_{r,k} Z_{r,k}$$

for every $(t_{r,k})_{1 \leqslant r,k \leqslant R} \in \mathbb{R}^{R \times R}$. Note that since $\mathbf{U}$ is a matrix of i.i.d. mean zero random

variables,

$$\sum_{r,k} t_{r,k} x_r^\top \mathbf{U} y_k = \sum_{i=1}^{N_1} \underbrace{\sum_{j=1}^{N_2} u_{i,j} \sum_{r,k} t_{r,k} x_{r,i} y_{k,j}}_{=:\xi_{N_2,i}} = \sum_{j=1}^{N_2} \underbrace{\sum_{i=1}^{N_1} u_{i,j} \sum_{r,k} t_{r,k} x_{r,i} y_{k,j}}_{=:\xi_{N_1,j}}$$

is a partial sum of triangular arrays of independent centered random variables. Moreover,

$$\begin{aligned}
\operatorname{Var}\left(\sum_{r,k} t_{r,k} x_r^\top \mathbf{U} y_k\right) &= \sigma^2 \sum_{r,k,r',k'} t_{r,k} t_{r',k'} \langle x_r, x_{r'} \rangle \langle y_k, y_{k'} \rangle \\
&= \sigma^2 \sum_{r,k} t_{r,k}^2 \\
&= \operatorname{Var}\left(\sum_{r,k} t_{r,k} \xi_{r,k}\right),
\end{aligned}$$

where the second line follows under orthonormality. The result follows provided that one of the following Lyapunov's conditions

$$\sum_{i=1}^{N_1} \mathbb{E}\left|\xi_{N_2,i}\right|^{2+\delta} = o(1) \qquad \text{or} \qquad \sum_{j=1}^{N_2} \mathbb{E}\left|\xi_{N_1,j}\right|^{2+\delta} = o(1)$$

holds for some $\delta > 0$; see Billingsley (1995), Theorem 27.3. By Rosenthal's inequality, see Petrov (1995), Theorem 2.9, there exists $c_\delta < \infty$ such that

$$\begin{aligned}
\mathbb{E}\left|\xi_{N_2,i}\right|^{2+\delta} &= \mathbb{E}\left|\sum_{j=1}^{N_2} u_{i,j} \sum_{r,k} t_{r,k} x_{r,i} y_{k,j}\right|^{2+\delta} \\
&\leqslant c_\delta \left\{ \sum_{j=1}^{N_2} \mathbb{E}\left|u_{i,j} \sum_{r,k} t_{r,k} x_{r,i} y_{k,j}\right|^{2+\delta} + \left(\sum_{j=1}^{N_2} \mathbb{E}\left|u_{i,j} \sum_{r,k} t_{r,k} x_{r,i} y_{k,j}\right|^{2} \right)^{1+\delta/2} \right\} \\
&\lesssim \sum_{r,k} t_{r,k}^{2+\delta} \sum_{j=1}^{N_2} x_{r,i}^{2+\delta} y_{k,j}^{2+\delta} + \sum_{r,k} t_{r,k}^{2+\delta} \left(\sum_{j=1}^{N_2} x_{r,i}^2 y_{k,j}^2 \right)^{1+\delta/2},
\end{aligned}$$

where the last line follows by Jensen's inequality and maintained assumptions. Therefore,

$$\begin{aligned}
\sum_{i=1}^{N_1} \mathbb{E}\left|\xi_{N_2,i}\right|^{2+\delta} &\lesssim \max_r \|x_r\|_{2+\delta}^{2+\delta} \max_k \left[\|y_k\|_{2+\delta}^{2+\delta} + \|y_k\|_2^{2+\delta} \right] \\
&\leqslant 2 \max_r \|x_r\|_\infty^\delta = o(1),
\end{aligned}$$

where we use the fact that $\|y_k\|_{2+\delta} \leqslant \|y_k\|_2$ and orthonormality. Similarly, we could verify

the second Lyapunov's condition when $\max_r \|y_r\| = o(1)$. □

Proof of Theorem 3.3. Note that

$$\begin{aligned}\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top - \Lambda_j D_j \Lambda_j^\top &= \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right)^\top \mathbf{U}_{(j)}^\top + \mathbf{U}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right) \mathbf{G}_{(j)}^\top \Lambda_j^\top + \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \\ &=: E_j.\end{aligned}$$

By the triangle inequality

$$\begin{aligned}\|E_j\|_{\text{op}} &\leqslant \left\|\Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right)^\top \mathbf{U}_{(j)}^\top\right\|_{\text{op}} + \left\|\mathbf{U}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right) \mathbf{G}_{(j)}^\top \Lambda_j^\top\right\|_{\text{op}} + \|\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\|_{\text{op}} \\ &= 2\left\|\mathbf{U}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right) \mathbf{G}_{(j)}^\top \Lambda_j^\top\right\|_{\text{op}} + \|\mathbf{U}_{(j)}\|_{\text{op}}^2\end{aligned}$$

Under Assumption 3.2 by Latała (2005)

$$\mathbb{E}\|\mathbf{U}_{(j)}\|_{\text{op}} = O\left(\sqrt{N_j} + \prod_{l\neq j}\sqrt{N_l}\right).$$

Consider the SVD decomposition

$$\left(\bigotimes_{l\neq j} \Lambda_l\right) \mathbf{G}_{(j)}^\top \Lambda_j^\top = V_j D_j^{1/2} \Lambda_j^\top = \sum_{r=1}^{R_j} \sigma_{j,r} v_{j,r} \lambda_{j,r}^\top.$$

Then

$$\begin{aligned}\left\|\mathbf{U}_{(j)} \left(\bigotimes_{l\neq j} \Lambda_l\right) \mathbf{G}_{(j)}^\top \Lambda_j^\top\right\|_{\text{op}} &\leqslant \sum_{r=1}^{R_j} \sigma_{j,r} \|\mathbf{U}_{(j)} v_{j,r} \lambda_{j,r}^\top\|_{\text{op}} \\ &= \sum_{r=1}^{R_j} \sigma_{j,r} \|\mathbf{U}_{(j)} v_{j,r}\| \\ &\lesssim \prod_{j=1}^{d} \sqrt{N_j} \max_{1\leqslant r\leqslant R_j} \|\mathbf{U}_{(j)} v_{j,r}\|,\end{aligned}$$

where we use $\lambda_{j,r}^\top \lambda_{j,r} = 1$ under Assumption 3.1 and $\sigma_{j,R_j}^2 \sim \prod_{j=1}^d N_j$ under Assumption 3.4 (i). Under Assumption 3.2, we also have

$$\mathbb{E}\|\mathbf{U}_{(j)} v_{j,r}\|^2 = v_{j,r}^\top \mathbb{E}[\mathbf{U}_{(j)}^\top \mathbf{U}_{(j)}] v_{j,r} = \sigma^2 N_j,$$

since $v_{j,r}^\top v_{j,r} = 1$. This shows that

$$\|E_j\|_{\text{op}}/\sigma_{j,R_j}^2 = O_P\left(\frac{1}{\prod_{l\neq j}\sqrt{N_l}} + \frac{1}{N_j}\right) = o_P(1), \quad \text{as} \quad (N_1,\ldots,N_d)\to\infty.$$

Therefore, $\|E_j\|_{\text{op}}/\sigma_{j,R_j}^2 < 1/4$ with probability approaching one. Let $\Lambda_{j\perp}$ be the orthonormal basis of the orthogonal complement to the column space of Λ_j. By Theorem A.5.1

$$\hat{\Lambda}_j O - \Lambda_j = \Lambda_{j\perp}\Lambda_{j\perp}^\top E_j \Lambda_j D_j^{-1} + r_j,$$

where the remainder term is

$$\begin{aligned}\|r_j\|_{\text{op}} &\leqslant 8\sigma_{j,1}^2\|E_j\|_{\text{op}}^2/\sigma_{j,R_j}^6 \\ &= O_P\left(\frac{1}{\prod_{j=1}^d N_j} + \frac{1}{N_j^2}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\Lambda}_j O_j - \Lambda_j &= \Lambda_{j\perp}\Lambda_{j\perp}^\top E_j\Lambda_j D_j^{-1} + r_j \\ &= \Lambda_{j\perp}\Lambda_{j\perp}^\top\left[\Lambda_j \mathbf{G}_{(j)}\left(\bigotimes_{l\neq j}\Lambda_l\right)^\top \mathbf{U}_{(j)}^\top + \mathbf{U}_{(j)}\left(\bigotimes_{l\neq j}\Lambda_l\right)\mathbf{G}_{(j)}^\top\Lambda_j^\top + \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\right]\Lambda_j D_j^{-1} + r_j \\ &= \Lambda_{j\perp}\Lambda_{j\perp}^\top \mathbf{U}_{(j)}\left(\bigotimes_{l\neq j}\Lambda_l\right)\mathbf{G}_{(j)}^\top D_j^{-1} + \Lambda_{j\perp}\Lambda_{j\perp}^\top\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\Lambda_j D_j^{-1} + r_j \\ &= \mathbf{U}_{(j)}V_j D_j^{-1/2} + \Lambda_{j\perp}\Lambda_{j\perp}^\top\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\Lambda_j D_j^{-1} - \Lambda_j\Lambda_j^\top\mathbf{U}_{(j)}V_j D_j^{-1/2} + r_j\end{aligned}$$

where we use $\Lambda_j\Lambda_j^\top + \Lambda_{j\perp}\Lambda_{j\perp}^\top = I$ and $\left(\bigotimes_{l\neq j}\Lambda_l\right)\mathbf{G}_{(j)}^\top D_j^{-1} = V_j D_j^{-1/2}$. Under Assumption 3.4 (ii)

$$\|\Lambda_j\Lambda_j^\top\mathbf{U}_{(j)}V_j\|_{2,\infty} \lesssim \sqrt{\frac{R_j}{N_j}}\|\mathbf{U}_{(j)}V_j\|_{2,\infty},$$

where

$$\|\mathbf{U}_{(j)}V_j\|_{2,\infty} = \max_{1\leqslant i\leqslant N_j}\|\mathbf{u}_{(j),i}V_j\| \leqslant R_j\max_{1\leqslant r\leqslant R_j}\max_{1\leqslant i\leqslant N_j}|\mathbf{u}_{(j),i}v_{j,r}|$$

and we partition $\mathbf{U}_{(j)}^\top = [\mathbf{u}_{(j),1},\ldots,\mathbf{u}_{(j),N}]$. Under Assumption 3.2, by Vershynin (2018), Lemma 3.4.2, this is a maximum of sub-Gaussian random variables, so that

$$\|\Lambda_j\Lambda_j^\top\mathbf{U}_{(j)}V_j\|_{2,\infty} = O_P\left(R_j\sqrt{\frac{\log N_j}{N_j}}\right).$$

Putting $B_j = \Lambda_{j\perp}\Lambda_{j\perp}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_j D_j^{-1}$, we obtain

$$(\hat{\Lambda}_j O_j - \Lambda_j - B_j)D_j^{1/2} = \mathbf{U}_{(j)}V_j + s_j$$

with

$$\begin{aligned}\|s_j\|_{2,\infty} &= O_P\left(R_j\sqrt{\frac{\log N_j}{N_j}} + \frac{\sigma_{j,1}}{\prod_{j=1}^d N_j} + \frac{\sigma_{j,1}}{N_j^2}\right) \\ &= O_P\left(R_j\sqrt{\frac{\log N_j}{N_j}} + \frac{1}{\prod_{j=1}^d \sqrt{N_j}} + \frac{\prod_{j=1}^d \sqrt{N_j}}{N_j^2}\right),\end{aligned}$$

where the second line follows under Assumption 3.4. By Theorem A.4.4 the $(i,j)^{\text{th}}$ element of $\mathbf{U}_{(j)}V_j$ is

$$\mathbf{u}_{(j),i}^\top V_j \xrightarrow{d} N(0, \sigma^2 I_{R_j}),$$

given that $\|V_j\|_{2,\infty} = o(1)$. □

Next, we provide a proof of the central limit theorem for scale components.

Proof of Theorem 3.4. Consider the SVD decomposition

$$\Lambda_j \mathbf{G}_{(j)}\left(\bigotimes_{l\neq j}\Lambda_l\right)^\top = \Lambda_j D_j^{1/2}V_j^\top = \sum_{r=1}^{R_j}\sigma_{j,r}\lambda_{j,r}v_{j,r}^\top.$$

By Kneip and Utikal (2001), Lemma A.1 (a),

$$\begin{aligned}\hat{\sigma}_{j,r}^2 - \sigma_{j,r}^2 &= \text{trace}\left\{\lambda_{j,r}\lambda_{j,r}^\top(\Lambda_j D_j^{1/2}V_j^\top\mathbf{U}_{(j)}^\top + \mathbf{U}_{(j)}V_j D_j^{1/2}\Lambda_j^\top + \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top)\right\} + \xi_{j,r} \\ &= \lambda_{j,r}^\top\left(\Lambda_j D_j^{1/2}V_j^\top\mathbf{U}_{(j)}^\top + \mathbf{U}_{(j)}V_j D_j^{1/2}\Lambda_j^\top + \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\right)\lambda_{j,r} + \xi_{j,r} \\ &= 2\sigma_{j,r}\lambda_{j,r}^\top\mathbf{U}_{(j)}v_{j,r} + \lambda_{j,r}^\top\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\lambda_{j,r} + \xi_{j,r},\end{aligned}$$

where

$$\begin{aligned}|\xi_{j,r}| &\leqslant \frac{6}{\min_{s\neq r}|\sigma_{j,s}^2 - \sigma_{j,r}^2|}\left\|\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top - \Lambda_j D_j^{1/2}\Lambda_j^\top\right\|_{\text{op}}^2 \\ &= O_P\left(N_j + \frac{N_j}{\prod_{l\neq j}N_l} + \frac{\prod_{l\neq j}N_l}{N_j}\right) = o_P(\sigma_{j,r}),\end{aligned}$$

which follows from $\sigma_{j,r}^2 \sim \prod_{j=1}^d N_j$ and the proof of Theorem 3.3, provided that $N_j/\prod_{l\neq j}N_l = o(1)$ and $\prod_{l\neq j}N_l/N_j^3 = o(1)$. This shows that

$$\frac{\hat{\sigma}_{j,r}^2 - \sigma_{j,r}^2}{\sigma_{j,r}} = 2\lambda_{j,r}^\top\mathbf{U}_{(j)}v_{j,r} + \sigma_{j,r}^{-1}\lambda_{j,r}^\top\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\lambda_{j,r} + o_P(1).$$

Next, under Assumptions 3.2 and 3.4

$$\mathrm{Var}(\lambda_{j,r}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \lambda_{j,r}) = \prod_{l\neq j} N_l \mathrm{Var}(\langle \lambda_{j,r}, \mathbf{u}_{(j),i}\rangle^2) = \prod_{l\neq j} N_l \left(\mathbb{E}\langle \lambda_{j,r}, \mathbf{u}_{(j),i}\rangle^4 - \sigma^4\right) = O\left(\prod_{l\neq j} N_l\right).$$

This shows that $\sigma_{j,r}^{-1}\lambda_{j,r}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \lambda_{j,r} = o_P(\sigma_{j,r})$, hence,

$$\frac{\hat{\sigma}_{j,r}^2 - \sigma_{j,r}^2}{\sigma_{j,r}} = 2\lambda_{j,r}^\top \mathbf{U}_{(j)} v_{j,r} + o_P(1).$$

Finally, under Assumptions 3.1, 3.2, and 3.4, by Lemma A.4.4

$$(\lambda_{j,r}^\top \mathbf{U}_{(j)} v_{j,r})_{1\leqslant r\leqslant R_j} \xrightarrow{d} N(0, \sigma^2 I_{R_j}).$$

□

Proof of Theorem 4.1. Under Assumption 3.2, if $u_{i_1,\ldots,i_d} \sim N(0,\sigma^2)$, then $\Lambda_{j\perp}^\top \mathbf{u}_i^{(j)} \in \mathbb{R}^{N_j - R_j}$ is a vector of i.i.d. $N(0,\sigma^2)$ for every $\Lambda_{j\perp}$ such that $\Lambda_{j\perp}^\top \Lambda_{j\perp} = I$. Then under Assumptions 3.1, 3.2 (i), and 3.3, by Lemma A.5.1,

$$\hat{\sigma}_{R+r,j}^2 = \lambda_r\left(\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\right) + O_P\left(N_j + \frac{\prod_{k\neq j} N_k}{N_j}\right),$$

where with some abuse of notation $\mathbf{U}_{(j)}$ is $(N_j - R) \times \prod_{k\neq j} N_k$ matrix of i.i.d. $N(0,\sigma^2)$.

By El Karoui (2003),

$$\left(\frac{\lambda_r\left(\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\right) - \lambda_{r+1}\left(\mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top\right)}{\tau}\right)_{r=k+1}^{K+1} \xrightarrow{d} (\xi_r - \xi_{r+1})_{r=1}^{K-k+1}.$$

Therefore,

$$\left(\frac{\hat{\sigma}_{r,j}^2 - \hat{\sigma}_{r+1,j}^2}{\tau}\right)_{r=k+1}^{K+1} \xrightarrow{d} (\xi_r - \xi_{r+1})_{r=1}^{K-k+1},$$

since $N_j/\tau + \prod_{k\neq j} N_k/(N_j\tau) = o(1)$. The result under H_0 follows by the continuous mapping theorem.

For H_1, recall that

$$\mathbf{Y}_{(j)} = \Lambda_j \mathbf{G}_{(j)} \left(\bigotimes_{k\neq j} \Lambda_k\right)^\top + \mathbf{U}_{(j)}.$$

By Weyl's inequality for singular values, see Horn and Johnson (2013), Eq. (7.3.15),

$$\left| \frac{\hat{\sigma}_{j,R_j}}{\sqrt{\prod_{k\neq j} N_k}} - \frac{\sigma_{j,R_j}}{\sqrt{\prod_{k\neq j} N_k}} \right| \leqslant \frac{\|\mathbf{U}_{(j)}\|_{\text{op}}}{\sqrt{\prod_{k\neq j} N_k}}.$$

The right-hand side of this equation is $O_P(1)$ by Latała (2005), while $\sigma_{j,R_j}/\sqrt{\prod_{k\neq j} N_k} \uparrow \infty$ under under the strong factor hypothesis. Therefore, $\hat{\sigma}^2_{j,R_j}/\prod_{k\neq j} N_k \uparrow \infty$. This implies that $(\hat{\sigma}^2_{j,R_j} - \hat{\sigma}^2_{j,R_j+1})/\tau \to \infty$, and whence $S_j \uparrow \infty$ under H_1. □

A.5 Perturbation Theory

In this section we formulate the required perturbation expansions for eigenvectors and eigenvalues following Kato (1996). The perturbation theory has been successfully applied in statistics, see Watson (1983), Kneip and Utikal (2001), or Koltchinskii and Lounici (2017), but is not widely adopted in the factor literature in econometrics; see Onatski (2006) or Lei and Ross (2023) for notable exceptions.

Theorem A.5.1. *Let $\Lambda = [\lambda_1, \ldots, \lambda_R] \in \mathbb{R}^{N\times R}$ be the eigenvectors of a symmetric matrix $\Sigma \in \mathbb{R}^{N\times N}$ with $\operatorname{rank}(\Sigma) = R < N$ corresponding to non-zero eigenvalues $\sigma_1 > \cdots > \sigma_R > 0$. Let $\hat{\Lambda} = [\hat{\lambda}_1, \ldots \hat{\lambda}_R] \in \mathbb{R}^{N\times R}$ be the leading eigenvectors of its noisy measure $\hat{\Sigma} = \Sigma + E$, where the noise satisfies $\|E\|_{\text{op}} < \sigma_R/4$. Then*

$$\hat{\Lambda}O - \Lambda = \Lambda_\perp \Lambda_\perp^\top E \Lambda D^{-1} + r_N,$$

where $O = \hat{\Lambda}^\top \Lambda$, $\Lambda_\perp = [\lambda_{R+1}, \ldots, \lambda_N] \in \mathbb{R}^{N\times(N-R)}$, and $r_N \leqslant 8\sigma_1 \|E\|^2_{\text{op}}/\sigma^3_R$.

Proof. Let $\Sigma = \sum_{j=1}^N \sigma_j \lambda_j \lambda_j^\top$ be the eigendecomposition of Σ. If γ is a contour in $\mathbb{C}$ that does not go through any eigenvalue of Σ, then

$$\begin{aligned}
\frac{1}{2\pi i} \oint_\gamma (\zeta I - \Sigma)^{-1} \mathrm{d}\zeta &= \frac{1}{2\pi i} \oint_\gamma \sum_{j=1}^N \frac{1}{\zeta - \sigma_j} \lambda_j \lambda_j^\top \mathrm{d}\zeta \\
&= \sum_{j:\sigma_j \in \gamma} \lambda_j \lambda_j^\top \frac{1}{2\pi i} \oint_\gamma \frac{1}{\zeta - \sigma_j} \mathrm{d}\zeta + \sum_{j:\sigma_j \notin \gamma} \lambda_j \lambda_j^\top \frac{1}{2\pi i} \oint_\gamma \frac{1}{\zeta - \sigma_j} \mathrm{d}\zeta \\
&= \sum_{j:\sigma_j \in \gamma} \lambda_j \lambda_j^\top,
\end{aligned}$$

where by Cauchy-Goursat theorem the integral is zero if an eigenvalue is outside of γ. The above identity is known as the Riesz formula for spectral projections; see also Kato (1996),

Problem I.5.9. If γ is a circle passing through $\sigma_R/2$ and $\sigma_1 + \sigma_R/2$, then

$$\sum_{r=1}^{R} \lambda_r \lambda_r^\top = \frac{1}{2\pi i} \oint_\gamma (\zeta I - \Sigma)^{-1} \mathrm{d}\zeta.$$

By Weyl's inequality, see Vershynin (2018), Theorem 4.5.3

$$\max_{1 \leqslant j \leqslant N} |\hat{\sigma}_j - \sigma_j| \leqslant \|E\|_{\mathrm{op}} < \frac{\sigma_R}{2}.$$

This shows that the eigenvalues $\hat{\sigma}_1 \geqslant \cdots \geqslant \hat{\sigma}_R$ are inside γ while the eigenvalues $\hat{\sigma}_{R+1} \geqslant \cdots \geqslant \hat{\sigma}_N$ are outside. Therefore, the Riesz formula also gives us

$$\sum_{r=1}^{R} \hat{\lambda}_r \hat{\lambda}_r^\top = \frac{1}{2\pi i} \oint_\gamma (\zeta I - \hat{\Sigma})^{-1} \mathrm{d}\zeta,$$

and whence

$$\begin{aligned} \hat{\Lambda}\hat{\Lambda}^\top - \Lambda\Lambda^\top &= \sum_{r=1}^{R} \hat{\lambda}_r \hat{\lambda}_r - \sum_{r=1}^{R} \lambda_r \lambda_r^\top \\ &= \frac{1}{2\pi i} \oint_\gamma (\zeta I - \hat{\Sigma})^{-1} - (\zeta I - \Sigma)^{-1} \mathrm{d}\zeta. \end{aligned}$$

Since $\hat{\Sigma} = \Sigma + E$, it is easy to see that

$$\begin{aligned} (\zeta I - \hat{\Sigma})^{-1} &= \big\{ (I - E(\Sigma - \zeta I)^{-1})(\zeta I - \Sigma) \big\}^{-1} \\ &= (\zeta I - \Sigma)^{-1} [I - E(\zeta I - E)^{-1}]^{-1} \\ &= (\zeta I - \Sigma)^{-1} \sum_{j=0}^{\infty} E^j (\zeta I - \Sigma)^{-j} \\ &= (\zeta I - \Sigma)^{-1} + \sum_{j=1}^{\infty} (\zeta I - \Sigma)^{-1} E^j (\zeta I - \Sigma)^{-j} \end{aligned}$$

where we use the Neumann series which converges since

$$\|E(\Sigma - \zeta I)^{-1}\|_{\mathrm{op}} \leqslant \|E\|_{\mathrm{op}} \|(\Sigma - \zeta I)^{-1}\|_{\mathrm{op}} \leqslant 2\|E\|_{\mathrm{op}}/\sigma_R < 1, \qquad \forall \zeta \in \gamma.$$

Therefore,

$$\begin{aligned}\hat{\Lambda}\hat{\Lambda}^\top - \Lambda\Lambda^\top &= \frac{1}{2\pi i}\oint_\gamma \sum_{j=1}^{\infty}(\zeta I - \Sigma)^{-1}E^j(\zeta I - \Sigma)^{-j}\mathrm{d}\zeta \\ &= \frac{1}{2\pi i}\oint_\gamma (\zeta I - \Sigma)^{-1}E(\zeta I - \Sigma)^{-1}\mathrm{d}\zeta + \frac{1}{2\pi i}\sum_{j=2}^{\infty}\oint_\gamma (\zeta I - \Sigma)^{-1}E^j(\Sigma - \zeta I)^{-j}\mathrm{d}\zeta \\ &=: L + r_N.\end{aligned}$$

Using the eigendecomposition $\Sigma = \sum_{j=1}^{N}\sigma_j\lambda_j\lambda_j^\top$, we obtain

$$L = \sum_{j=1}^{N}\sum_{k=1}^{N}\frac{1}{2\pi i}\oint_\gamma \frac{\mathrm{d}\zeta}{(\zeta - \sigma_j)(\zeta - \sigma_k)}\lambda_j\lambda_j^\top E\lambda_k\lambda_k^\top.$$

Since the first R eigenvalues of Σ are inside γ and the remaining eigenvalues are zero, by Cauchy's integral formula

$$\frac{1}{2\pi i}\oint_\gamma \frac{\mathrm{d}\zeta}{(\zeta - \sigma_j)(\zeta - \sigma_k)} = \begin{cases}\frac{1}{\sigma_j}, & j \leqslant R \text{ and } k > R, \\ \frac{1}{\sigma_k}, & j > R \text{ and } k \leqslant R, \\ 0, & \text{otherwise.}\end{cases}$$

This shows that

$$\begin{aligned}L &= \sum_{j=1}^{R}\frac{1}{\sigma_j}\lambda_j\lambda_j^\top E\sum_{k=R+1}^{N}\lambda_k\lambda_k^\top + \sum_{j=R+1}^{N}\lambda_j\lambda_j^\top E\sum_{k=1}^{R}\frac{1}{\sigma_k}\lambda_k\lambda_k^\top \\ &= \Lambda D^{-1}\Lambda^\top E\Lambda_\perp\Lambda_\perp^\top + \Lambda_\perp\Lambda_\perp^\top E\Lambda D^{-1}\Lambda^\top.\end{aligned}$$

The result follows since the eigenvectors are orthonormal and

$$\begin{aligned}|r_N| &\leqslant \frac{1}{2\pi}\sum_{j=2}^{\infty}\pi\sigma_1\sup_{\zeta\in\gamma}\|(\zeta I - \Sigma)\|_{\mathrm{op}}^{-(j+1)}\|E\|_{\mathrm{op}}^j \\ &\leqslant \frac{\sigma_1}{2}\|E\|_{\mathrm{op}}^2(2/\sigma_R)^3\sum_{j=0}^{\infty}(2\|E\|_{\mathrm{op}}/\sigma_R)^j \\ &= \frac{2\sigma_1}{\sigma_R^2}\|E\|_{\mathrm{op}}\frac{2\|E\|_{\mathrm{op}}/\sigma_R}{1 - 2\|E\|_{\mathrm{op}}/\sigma_R} \\ &\leqslant \frac{8\sigma_1}{\sigma_R}\frac{\|E\|_{\mathrm{op}}^2}{\sigma_R^2},\end{aligned}$$

provided that $\|E\|_{\text{op}} \leqslant \sigma_R/4$. $\square$

We also need an asymptotic expansion for the smallest singular values. Let $\Lambda_{j\perp} = [\lambda_{j,R_j+1}, \ldots, \lambda_{j,N_j}] \in \mathbb{R}^{N_j \times (N_j - R_j)}$ be eigenvectors corresponding to zero eigenvalues of $\Lambda_j D_j \Lambda_j^\top$. The columns of $\Lambda_{j\perp}$ constitute an orthonormal basis of the null space of $\Lambda_j D_j \Lambda_j^\top$.

Lemma A.5.1. *Suppose that Assumptions 3.1 and 3.2 are satisfied, and* $\sigma^2_{j,R_j} \sim \prod_{j=1}^d N_j$. *Then*

$$\hat{\sigma}^2_{j,R_j+r} = \lambda_r(\Lambda_{j\perp}^\top \mathbf{U}_{(j)} \mathbf{U}_{(j)}^\top \Lambda_{j\perp}) + O_P\left(N_j + \frac{\prod_{l\neq j} N_l}{N_j}\right), \qquad 1 \leqslant \forall r \leqslant N_j - R_j,$$

where $\lambda_r(B)$ *is the* r^{th} *largest eigenvalue of a matrix* B.

Proof. Note that $\lambda_r(\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top) = \lambda_r(M\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top M^\top), r \geqslant 1$ for every orthogonal matrix M. Put

$$M_j = \begin{bmatrix} \Lambda_j & \Lambda_{j\perp} \end{bmatrix} \qquad \text{and} \qquad \Sigma_j = \begin{bmatrix} D_j & 0 \\ 0 & 0 \end{bmatrix}.$$

Then $M_j \Sigma_j M_j^\top$ is the spectral decomposition of $\Lambda_j D_j \Lambda_j^\top$ and we have

$$\begin{aligned}
\lambda_{R_j+r}\left(\frac{\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top}{\prod_{j=1}^d N_j}\right) &= \lambda_{R_j+r}\left(\frac{M^\top \mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top M}{\prod_{j=1}^d N_j}\right) \\
&= \lambda_{R_j+r}\left(\frac{1}{\prod_{j=1}^d N_j}\begin{bmatrix}\Lambda_j^\top \\ \Lambda_{j\perp}^\top\end{bmatrix}(\Lambda_j D_j V_j^\top + \mathbf{U}_{(j)})(\Lambda_j D_j V_j^\top + \mathbf{U}_{(j)})^\top \begin{bmatrix}\Lambda_j & \Lambda_{j\perp}\end{bmatrix}\right) \\
&=: \lambda_{R_j+r}\left(A + \varkappa A^{(1)},\right)
\end{aligned}$$

with $\varkappa = 1/N_j$, $A = \begin{bmatrix} \frac{D_j}{\prod_{j=1}^d N_j} & 0 \\ 0 & 0 \end{bmatrix}$ and

$$A^{(1)} = \frac{1}{\prod_{l\neq j} N_l}\begin{bmatrix} \Lambda_j^\top \mathbf{U}_{(j)} V_j D_j + D_j V_j^\top \mathbf{U}_{(j)}^\top \Lambda_j + \Lambda_j^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_j & D_j V_j^\top \mathbf{U}_{(j)}^\top \Lambda_{j\perp} + \Lambda_j^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_{j\perp} \\ \Lambda_{j\perp}^\top \mathbf{U}_{(j)} V_j D_j + \Lambda_{j\perp}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_j^\top & \Lambda_{j\perp}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_{j\perp} \end{bmatrix}.$$

By Onatski (2009), Lemma 6,

$$\left|\lambda_{R_j+r}\left(A + \varkappa A^{(1)}\right) - \varkappa\lambda_r\left(\frac{\Lambda_{j\perp}^\top \mathbf{U}_{(j)}\mathbf{U}_{(j)}^\top \Lambda_{j\perp}}{\prod_{l\neq j} N_l}\right)\right| \leqslant \frac{\varkappa^2 \|A^{(1)}\|^2}{0.5\sigma^2_{j,R_j}/\prod_{j=1}^d N_j - \varkappa\|A^{(1)}\|},$$

provided that $\|A^{(1)}\|/N_j < 0.5\sigma^2_{j,R_j}/\prod_{j=1}^d N_j$. To see that this condition holds, note that

since $\sigma_{j,r}^2 \sim \prod_{j=1}^d N_j$, we have $\sigma_{j,r}^2 / \prod_{j=1}^d N_j \to d_{j,r} > 0, \forall r \leqslant R_j$. Moreover,

$$\|A^{(1)}\|/N_j \leqslant \frac{4}{\prod_{j=1}^d N_j} \left[\left\| \mathbf{U}_{(j)} V_j D_j \right\|_{\text{op}} + \left\| \mathbf{U}_{(j)} \mathbf{U}_{(j)}^\top \right\|_{\text{op}} \right].$$

It follows from the proof of Theorem 3.3 under Assumption 3.2,

$$\|\mathbf{U}_{(j)} V_j D_j\|_{\text{op}} = O_P\left(\sqrt{N_j} \sum_{r=1}^{R_j} \sigma_{j,r}\right) \qquad \text{and} \qquad \|\mathbf{U}_{(j)} \mathbf{U}_{(j)}^\top\|_{\text{op}} = O_P\left(N_j + \prod_{l \neq j} N_l\right).$$

Since $\sigma_{j,r}^2 \sim \prod_{j=1}^d N_j$, this shows that

$$\|A^{(1)}\|/N_j = O_P\left(\frac{1}{\prod_{l \neq j} \sqrt{N_l}} + \frac{1}{N_j}\right) = o_P(1).$$

Therefore,

$$\lambda_{R+r}\left(\frac{\mathbf{Y}_{(j)} \mathbf{Y}_{(j)}^\top}{\prod_{j=1}^d N_j}\right) = \lambda_r\left(\frac{\Lambda_{j\perp}^\top \mathbf{U}_{(j)} \mathbf{U}_{(j)}^\top \Lambda_{j\perp}}{\prod_{j=1}^d N_j}\right) + O_P\left(\frac{1}{\prod_{l \neq j} N_l} + \frac{1}{N_j^2}\right).$$

□