Tensor Principal Component Analysis[*]

Andrii Babii[†] *UNC Chapel Hill*

Eric Ghysels[‡] *UNC Chapel Hill*

Junsu Pan[§] Job Market Candidate *UNC Chapel Hill*

October 20, 2023

Abstract

In this paper, we develop new methods for analyzing high-dimensional tensor datasets. A tensor factor model describes a high-dimensional dataset as a sum of a low-rank component and an idiosyncratic noise, generalizing traditional factor models for panel data. We propose an estimation algorithm, called tensor principal component analysis (TPCA), which generalizes the traditional PCA applicable to panel data. The algorithm involves unfolding the tensor into a sequence of matrices along different dimensions and applying PCA to the unfolded matrices. We provide theoretical results on the consistency and asymptotic distribution for the TPCA estimator of loadings and factors. We also introduce a novel test for the number of factors in a tensor factor model. The TPCA and the test feature good performance in Monte Carlo experiments and are applied to sorted portfolios.

Keywords: Principal component analysis, SVD and CP decomposition, tensor data, tensor factor model, testing number of factors.

[*] 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 and 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

Factor analysis is a more than century-old dimension reduction technique, originally introduced in the psychology literature by Spearman (1904). Factor models are commonly used in economics and finance to analyze a large set of correlated variables and to explore the latent factors driving the dependencies in the data. Traditional factor models apply to two-dimensional panel data consisting of cross-sectional observations evolving over time. The econometric analysis covers situations where the cross-sectional and/or time series sample sizes grow asymptotically; see Stock and Watson (2002) and Bai (2003) among others.

Many economic data, however, feature more than two dimensions. For example, a typical panel data set of macroeconomic series covering a collection of real activity and price series involves regional aggregation of state- or county-level observations. Hence, there is a geographical dimension in addition to the cross-sectional and time series dimensions. Likewise, asset pricing models pertaining to the cross-section of equities typically involve characteristic-based portfolio sorts. The sorting into deciles is common, but only the lowest and highest deciles are used and combined in a high minus low return spread. Again, a third dimension which in this case is the return on each of the decile portfolio sorts is muted. If we go beyond national borders we have a fourth international dimension in addition to time, sorting characteristics, and deciles. Adding the international dimension to the macro data set also yields a four-dimensional data structure. Each of these example illustrates the fact that we often aggregate high-dimensional data, and by doing so suppress more granular information, to obtain matrix representations of the observations; see Matyas (2017) for more examples.

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).[1] 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.

The scope of our paper is to apply insights, methods, and theory of the familiar PCA to tensor type data. More specifically, we introduce PCA-type estimators for

[1] Henceforth, we will use the term 'traditional' for the 2-way factor models applied to panel data.

d-way factor models. The estimators have closed-form expressions and do not require solving the non-convex optimization problems. We call our procedure tensor PCA - or TPCA - and (a) we show that it can identify and consistently estimate the factors and loadings, (b) we describe the associated convergence rates and asymptotic distribution, (c) we show that tensor dimensions can improve the estimation accuracy for factors/loadings, and (d) develop a formal test for a number of factors in the tensor factor model.

A d-way tensor is a d-dimensional array generalizing vectors and matrices introduced in Ricci and Levi-Civita (1900). 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, the correlations are captured with vector components in the time dimension (factors) and the vector components in the cross-sectional dimension (factor loadings). Likewise, one can decompose a tensor into vector components with respect to each dimension. Hence, the vector components in the 'time' dimension correspond to factors that vary over time, and the vector components in the other dimensions correspond to loadings determining the heterogeneous exposure of each dimension to the factors.

Traditional factor models can also be viewed as decomposing a matrix representing a panel dataset as a sum of a low-rank loading and factor matrix 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. The low-rank component is usually modeled using either the Tucker or the Canonical Polyadic (CP) decompositions; see Tucker (1966) and Hitchcock (1927). While the former is more general and covers the CP decomposition as a special case, it is also overparametrized and leads to difficult identifiability issues. Therefore, in this paper, we focus on the tensor factor models related to the CP decomposition.[2] For tensor factor models based on the Tucker decomposition, see Han, Chen, Yang, and Zhang (2020), Wang, Zheng, Lian, and Li (2022), Chen, Yang, and Zhang (2022), Han, Chen, and Zhang (2022) and for a symmetric rank-1 tensor factor

[2] Strictly speaking, the tensor factor model introduced in this paper features orthogonal factors/loadings which is not the case for the CP decomposition.

model Richard and Montanari (2014).

There exist several methods to compute the CP decomposition of a tensor. Carroll and Chang (1970) and Harshman (1970) proposed an iterative algorithm, known as *alternating least squares* (ALS), which is the most widely used in practice procedure; see also Kolda and Bader (2009). The ALS algorithm computes the least-squares approximation to the observed tensor which is a non-convex optimization problem that can be unstable in practice. It is also worth mentioning that the best rank-R approximation to a tensor may not even exist and that most of the optimization problems related to tensors, including the rank determination, are NP-hard; see De Silva and Lim (2008) and Hillar and Lim (2013). As a result, the statistical properties of the ALS algorithm are, to the best of our knowledge, largely unknown.

To deal with the aforementioned computational challenges, we propose a tensor factor model with *orthogonal* factors/loadings and a new estimation procedure for tensor factor models that is an extension of PCA. The algorithm consists of steps that first unfold the d-way tensor into a d matrices in different directions, and then apply PCA to the unfolded matrices, or matricizations, to obtain the components of the tensor decomposition. Therefore, the TPCA leads to the closed form expressions for factors and factor loadings and can identify and consistently estimate the factors and loadings. We describe the associated asymptotic properties for large dimensional tensors with growing dimensions. Our results show that the tensor dimensions can improve the estimation accuracy for factors/loadings.

Lastly, we develop a novel test for the number of factors in the tensor factor which to the best of our knowledge is the first formal statistical tests in the tensor factor literature. Monte Carlo simulations support our asymptotic results in finite samples. We find that our TPCA algorithm is more accurate than the ALS algorithm and that the d-way tensor factor model reduces dimensions more efficiently than the naively pooled 2-way factor model.

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. 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 collect several illustrative examples of tensors, discuss tensor unfoldings, provide the proofs for all the main and auxiliary results.

Notation: The Khatri-Rao product of two matrices $A = (a_1, \dots, a_R)$ and $B = (b_1, \dots, b_R)$ is defined as $A \odot B = (a_1 \otimes_{\mathrm{K}} b_1, \dots, a_R \otimes_{\mathrm{K}} b_R)$, where $\otimes_{\mathrm{K}}$ denotes the Kronecker product. In addition, for a collection of matrices $(V_j)_{j=1}^d$, we define $\bigodot_{k \neq j} V_k = V_d \odot \dots \odot V_{k+1} \odot V_{k-1} \odot \dots \odot V_1$. 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 $a_n \leq C b_n$ for all $n \in \mathbb{N}$. The operator norm of a matrix A is defined as $\|A\|_{\mathrm{op}} = \sup_{\|x\|=1} \|Ax\|$, 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 $\langle A, B \rangle_F = \sum_{i_1,\dots,i_d} A_{i_1,\dots,i_d} B_{i_1,\dots,i_d}$. Let $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A \rangle_{\mathrm{F}}}$ be the Frobenius norm of a tensor A induced by the inner product. For a matrix A with columns $(a_1, \dots, a_n)$, the $\ell_{2,1}$ matrix norm defined as $\|A\|_{2,1} = \sum_{j=1}^n \|a_j\|$. 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 two-dimensional panel data described by a matrix. For a dataset $\mathbf{Y} \in \mathbb{R}^{N \times T}$, the factor model with R factors can be expressed as a sum of a low-rank matrix and a matrix $\mathbf{U} \in \mathbb{R}^{N \times T}$ of idiosyncratic shocks.[3] Hence:

$$\mathbf{Y} = \sum_{r=1}^{R} \lambda_r \otimes f_r + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0, \tag{1}$$

where $f_r \in \mathbb{R}^T$ are the common factors, $\lambda_r \in \mathbb{R}^N$ are the factor loadings, $\lambda_r \otimes f_r = \lambda_r f_r^\top$ is the outer product. Estimating the factors and their loadings can be done via PCA.

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

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

see Appendix Figure A.1 for graphical illustrations. The generalizations of matrix rows and columns to tensors are called *fibers* and *slices.* 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

[3] Recall that a matrix has rank-R if and only if it can be expressed as a sum of R outer products of two vectors.

is a mode-2 fiber; see Appendix Figure A.2 for a graphical illustration. Fixing all but two indices of a tensor, we obtain a matrix, called slice; see Appendix Figure A.3 for a graphical illustration.

The notion of a rank-1 matrix also has a natural generalization to tensors. A d-way tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ has rank 1 if it can be expressed as an outer product of d vectors

$$\mathbf{Y} = v_1 \otimes v_2 \otimes \ldots \cdots \otimes v_d \equiv \bigotimes_{j=1}^{d} v_j,$$

where $v_j \in \mathbb{R}^{N_j}$ and $\otimes$ is the vector outer product, i.e. each element of $\mathbf{Y}$ is the product of corresponding vector elements: $\mathbf{Y}_{i_1,\ldots,i_d} = v_{1,i_1} v_{2,i_2} \ldots v_{d,i_d}$. Every tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ can be expressed as a sum of rank-1 tensors

$$\mathbf{Y} = \sum_{r=1}^{R} \bigotimes_{j=1}^{d} v_{j,r}. \tag{2}$$

The smallest number R such that the decomposition in equation (2) holds is called the *rank* of the tensor and the corresponding decomposition is called the Canonical Polyadic (CP) Decomposition; Kolda and Bader (2009) for a comprehensive review.

Similarly to the 2-way factor model, the d-way factor model for a tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ can be defined as a sum of a low-rank tensor and an idiosyncratic noise tensor $\mathbf{U} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$

$$\mathbf{Y} = \sum_{r=1}^{R} \bigotimes_{j=1}^{d} v_{j,r} + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0. \tag{3}$$

The vectors $v_{j,r} \in \mathbb{R}^{N_j}, 1 \leq j \leq d$ are factors and factor loadings depending on the context of the application. For instance in economics, the vectors in the “time” dimension are often treated as factors that evolve through time, and the vectors in other dimensions can be treated as factor loadings that determine the heterogeneous cross-sectional exposure to the time dimension.

Example 2.1 (Three-way factor model)**.** *Suppose that $d = 3$, N_1 is the number of portfolio characteristics, N_2 is the number of mutual funds, and N_3 is the number of* *time periods. Then we obtain a 3-way factor model*

$$\mathbf{Y} = \sum_{r=1}^{R} v_{1,r} \otimes v_{2,r} \otimes v_{3,r} + \mathbf{U},$$

where $v_{3,r} \in \mathbb{R}^{N_3}$ is a vector of time-series factors, $v_{2,r} \in \mathbb{R}^{N_2}$ are the loadings of mutual funds, and $v_{1,r} \in \mathbb{R}^{N_1}$ measure the heterogeneous exposures to different characteristics; see Lettau (2022).

There exists another decomposition of a tensor into a sum of rank-1 tensors, called Tucker decomposition; see Tucker (1958). The Tucker decomposition decomposes a d-way tensor as

$$\mathbf{Y} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_d=1}^{R_d} g_{r_1 r_2 \cdots r_d} \bigotimes_{j=1}^{d} v_{j,r_j},$$

where the tensor $\mathbf{G} = \{g_{r_1,\dots,r_d} : 1 \leq r_j \leq R_j, 1 \leq j \leq d\}$ is called a core tensor. The CP decomposition is a special case of the Tucker decomposition with a diagonal core tensor, $g_{r_1 \dots r_d} = \mathbb{1}_{r_1 = \dots = r_d}$. However, the Tucker decomposition features a substantially larger number of parameters in the core tensor and is in general not unique, which results in non-trivial identification issues.

Example 2.2. *Suppose that $d = 2$ with the first dimension corresponding to the cross-section and the second to the time-series. Then the Tucker factor model with $R_1 = 2$ and $R_2 = 1$ is observationally equivalent to a model with $R_1 = R_2 = 1$*

$$\begin{aligned} y_{it} &= \lambda_{i1} f_t + \lambda_{i2} f_t + u_{it} \\ &= \tilde{\lambda}_i f_t + u_{it}, \end{aligned}$$

where $\tilde{\lambda}_i = \lambda_{i1} + \lambda_{i2}$.

The above example also illustrates that having a different number of factors/loadings in different dimensions does not often lead to intuitive economic interpretation. Moreover, there does not appear to be a broadly accepted solution to the identification problem arising from Tucker decompositions. Some impose sparsity, see e.g. Wang et al. (2022) while often the non-uniqueness of the solution is ignored. In this paper, we will focus on the tensor factor model in equation (3), which (a) is a direct generalization of the widely used 2-way factor models in econometrics and statistics; cf. equation (1), (b) has a well understood identification problem for which a widely accepted normalization is used, and (c) involves a convex optimization with a closed-form solution.

3 Tensor PCA

In this section, we introduce the TPCA algorithm to estimate a tensor factor model. We begin by explaining the process of unfolding in the first subsection. The next one discusses the identification issues. We present the estimation algorithm in the third subsection. The final subsection covers the asymptotic properties—rates of consistency and large sample distributions—for the TPCA estimator of loadings/factors.

3.1 Unfolding

Similar to traditional factor models, the scale of loadings/factors in tensor factor models is not identified.[4] We therefore will focus on the normalized model

$$\mathbf{Y} = \sum_{r=1}^{R} \sigma_r \bigotimes_{j=1}^{d} m_{j,r} + \mathbf{U}, \qquad \mathbb{E}\mathbf{U} = 0, \tag{4}$$

where $\sigma_r = \prod_{j=1}^{d} \|v_{j,r}\|$ is a scale component and $m_{j,r} = v_{j,r}/\|v_{j,r}\|$ are the normalized loadings/factors. The objective is to identify and estimate the normalized matrices of loadings/factors

$$M_j = (m_{j,1}, \ldots, m_{j,R}), \qquad 1 \leq j \leq d \tag{5}$$

and the scale components $(\sigma_r)_{r=1}^{R}$.

Since PCA is based on matrix representations of data, we will reshape the tensor into matrices. The process is called *unfolding* and can be understood as a generalization of matrix vectorization. A d-way tensor can be unfolded in d different directions. The mode-j unfolding of a tensor $\mathbf{Y} \in \mathbb{R}^{N_1 \times \ldots \times N_d}$, denoted $\mathbf{Y}_{(j)} \in \mathbb{R}^{N_j \times (N_1 \ldots N_{j-1} N_{j+1} \ldots N_d)}$, reshapes the mode-$j$ *fibers* of $\mathbf{Y}$ into the columns of $\mathbf{Y}_{(j)}$; see Appendix Section A.2 for an illustrative examples and equation (A.1) for a generic formula for mode-j unfoldings, i.e. matricizations, of a d-way tensor.[5]

[4] For example, $(v_{1,r}, v_{2,r})$ is observationally equivalent to $(av_{1,r}, v_{2,r}/a)$ for every $a \neq 0$.

[5] Sometimes the term "flattening" of a tensor is used. We prefer unfolding and in particular 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.

For the 3-way tensor factor model in equation (4), using the mode-1 unfolding, see equation (A.1), we obtain a 2-way factor model:

$$\mathbf{Y}_{(1)} = M_1 D (M_3 \odot M_2)^\top + \mathbf{U}_{(1)}, \tag{6}$$

where $\mathbf{Y}_{(1)}$ and $\mathbf{U}_{(1)}$ are $N_1 \times N_2 N_3$ matrices, $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_R)$, and $M_3 \odot M_2$ is the Khatri-Rao product of M_3 and M_2. This unfolding allows us to estimate M_1 and $M_3 \odot M_2$ using PCA. More specifically, the PCA applied to the unfolded tensor in equation (6) estimates the product $M_3 \odot M_2$, but does not estimate M_2 and M_3 separately. However, we can also matricize the 3-way tensor using the mode-j unfolding for $j = 2$ and 3:

$$\mathbf{Y}_{(2)} = M_2 D (M_3 \odot M_1)^\top + \mathbf{U}_{(2)} \qquad \text{and} \qquad \mathbf{Y}_{(3)} = M_3 D (M_2 \odot M_1)^\top + \mathbf{U}_{(3)},$$

where $\mathbf{Y}_{(2)}, \mathbf{U}_{(2)} \in \mathbb{R}^{N_2 \times N_1 N_3}$ and $\mathbf{Y}_{(3)}, \mathbf{U}_{(3)} \in \mathbb{R}^{N_3 \times N_1 N_2}$, which allow us to estimate M_2 and M_3 respectively.

For the general d-way factor model with $d \geq 3$, the mode-j unfoldings of equation (4) are

$$\mathbf{Y}_{(j)} = M_j D \left(\bigodot_{k \neq j} M_k \right)^\top + \mathbf{U}_{(j)}, \qquad 1 \leq j \leq d, \tag{7}$$

where $\bigodot_{k \neq j} M_k = M_d \odot \cdots \odot M_{j+1} \odot M_{j-1} \odot \cdots \odot M_1$.

3.2 Identification

For each $j = 1, \ldots, d$, let $V_j = (v_{j,1}, \ldots, v_{j,R})$ be the $N_j \times R$ matrices of loadings/factors. Following the convention in the factor literature, we assume that the loadings/factors are orthogonal:

Assumption 3.1. *For $1 \leq j \leq d$: $V_j^\top V_j$ is a diagonal matrix.*

Under Assumption 3.1, the matrices of normalized loadings/factors in equation (5) are unitary:

$$M_j^\top M_j = I_R, \qquad 1 \leq j \leq d.$$

The following result shows that the matrices $\left(\bigodot_{k \neq j} M_k \right)^\top$ in equation (7) are also unitary.[6]

Proposition 3.1. *Under Assumption 3.1*

$$\left(\bigodot_{k\neq j} M_k\right)^\top \left(\bigodot_{k\neq j} M_k\right) = I_R, \qquad 1 \le j \le d.$$

Therefore, under Assumption 3.1, the tensor factor model is identified, cf. Bai and Ng (2013).

3.3 Estimation Algorithm

The discussion in the previous subsection leads to the following TPCA estimation algorithm:

1) Unfold the tensor $\mathbf{Y}$ into matrices $\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \dots, \mathbf{Y}_{(d)}$ along each of its dimensions.

2) Estimate M_j as $\widehat{M}_j = (\hat{m}_{j,1}, \dots, \hat{m}_{j,R})$ via PCA, i.e. take $\hat{m}_{j,r}$ is the unit norm eigenvector of $\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top$ corresponding to the r^{th} largest eigenvalue.

3) Recover the scale components $(\hat{\sigma}_{r,j}^2)_{r=1}^R$ as the first R largest eigenvalues of $\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top$.

When $d = 2$, we have a traditional 2-way factor model, and $\mathbf{Y}$ is a matrix. In this case, step 1) is vacuous; step 2) is standard PCA estimation of the 2-way factor model; and in step 3) we get the singular values of $\mathbf{Y}$: $\hat{\sigma}_r = \hat{m}_{1,r}^\top \mathbf{Y} \hat{m}_{2,r}$.

Our TPCA algorithm features several advantages relative to the widely used ALS algorithm that computes the best rank-R approximation to $\mathbf{Y} \in \mathbb{R}^{N_1\times\cdots\times N_d}$. First, the best rank-$R$ approximation to a tensor may not exist when $d \ge 3$; see Kolda and Bader (2009). Second, the best rank-R approximation to a tensor requires solving a non-convex optimization problem whereas the TPCA leads to convex optimization with closed-form solutions. Third, the asymptotic properties of ALS are not known whereas we obtain the consistency and the large sample distributions for TPCA in the following subsections. Lastly, the best rank-R approximation computed with ALS is sensitive to the choice of R and the first extracted factor will be different for different

[6] All proofs appear in Appendix Section A.4.

values of R. In contrast, the TPCA computes all factors at once and the first factor will always be the same regardless of the number of specified factors.

3.4 Rates of Consistency

Write $\mathbf{U}_{(j)} = (\mathbf{U}_1^{(j)}, \mathbf{U}_2^{(j)}, \dots)$, where $\mathbf{U}_i^{(j)} \in \mathbb{R}^{N_j}$ is the i^{th} column of the unfolded tensor $\mathbf{U}_{(j)}$. The following assumption imposes mild restrictions on the data generating process.

Assumption 3.2. *The idiosyncratic errors $\mathbf{U} = \{u_{i_1,\dots,i_d} : 1 \leq i_j \leq N_j, 1 \leq j \leq d\}$ are i.i.d. with $\mathbb{E}(u_{i_1,\dots,i_d}) = 0$, $\operatorname{Var}(u_{i_1,\dots,i_d}) = \sigma^2$, and $\mathbb{E}|u_{i_1,\dots,i_d}|^{4+\epsilon} < \infty$ for some $\epsilon > 0$; (ii) $\mathbb{E}|\langle \mathbf{U}_i^{(j)}, m_{1,k}\rangle \langle \mathbf{U}_i^{(j)}, m_{1,r}\rangle|^2 = O(1)$ for every $k \neq r$; (iii) $\|m_{j,r}\|_\infty = o(1)$ for every j, r.*

The i.i.d. assumption can be relaxed to heterogeneous and dependent arrays at the costs of heavier notations and proofs. For the condition (ii), note that if $\mathbf{U}_i^{(j)}$ is a Gaussian vector, then $\mathbb{E}|\langle \mathbf{U}_i^{(j)}, m_{1,k}\rangle \langle \mathbf{U}_i^{(j)}, m_{1,r}\rangle|^2 = \sigma^2$. Lastly, condition (iii) is not restrictive given that the loadings/factors are normalized so that $\|m_{j,r}\| = 1$ for all values of N_j, though it rules out the case when $(m_{j,r})_{r=1}^R$ is a canonical basis of $\mathbb{R}^{N_j}$.

Since PCA has a sign indeterminacy, it is worth noting that we can always assume that the signs of the sample eigenvectors $(\hat{m}_{j,r})_{r=1}^R$ are properly selected. Let

$$\delta_r = \min_{k \neq r} |\sigma_k^2 - \sigma_r^2|$$

be a measure of how the contribution of the r^{th} factor is distinguishable from the rest.[7] Then, the following result holds:

Theorem 3.1. *Suppose that Assumptions 3.1 and 3.2 (i) are satisfied. Then*

$$\|\hat{m}_{j,r} - m_{j,r}\| = O_P\left(\frac{\sqrt{N_j}\operatorname{trace}(D) + N_j \vee \prod_{k\neq j} N_k}{\delta_r}\right), \qquad \forall j, r \geq 1.$$

The proof appears in the Appendix. According to Theorem 3.1, the more the contribution of r^{th} is distinguishable from the rest, as measured by δ_r, the easier it is to estimate it.

[7] We set $\sigma_k^2 = 0$ for all $k \geq R + 1$.

Next, we generalize the pervasive factors assumption to the tensor factor model. This assumption is not the weakest possible, see Onatski (2012, 2022). Nevertheless, it is commonly used in the literature and can be justified for random factors/loadings by the law of large numbers; see Fan and Wang (2015).

Assumption 3.3. *There exist constants $d_1 > d_2 > \cdots > d_R > 0$ such that*

$$\lim_{N_1,\ldots,N_d\to\infty} \frac{\sigma_r^2}{\prod_{j=1}^d N_j} = d_r, \qquad 1 \leq \forall r \leq R.$$

Theorem 3.1 leads to the following results.

Corollary 3.1. *Suppose that Assumptions 3.1, 3.2 (i), and 3.3 are satisfied. Then*

$$\|\widehat{M}_j - M_j\|_{2,1} = O_P\left(\sqrt{\frac{1}{\prod_{k\neq j} N_k}} \vee \frac{1}{N_j}\right).$$

According to this result, the additional tensor dimensions allow us to estimate the factors and loadings more accurately. For instance, for the 3-way factor model when all tensor dimensions grow proportionally to some N, we obtain a very fast rate of order $O(1/N)$.

3.5 Large Sample Approximations

We will focus on the asymptotic distribution of a linear functional of the loading/factor vector $\hat{m}_{j,r} \in \mathbb{R}^{N_j}$. Recall that every linear functional $\Phi : \mathbb{R}^{N_j} \to \mathbb{R}$ can be represented as $\Phi(x) = \langle x, \nu \rangle$ for some $\nu \in \mathbb{R}^{N_j}$. The following two examples are of interest:

1. i^{th} element of loading/factor vector $m_{j,r}$: ν is an "all zeros" vectors, except for i^{th} element equal to 1.

2. The average loading/factor vector: $\nu = (1, 1, \ldots, 1)/N_j$.

We assume that the population counterpart $\langle m_{j,r}, \nu \rangle$ is asymptotically well-defined.

Assumption 3.4. *Suppose that ν is such that for every $k \neq r$,*

$$\omega_{j,k}(\nu) \equiv \lim_{N_j\to\infty} \sqrt{N_j} \langle m_{j,k}, \nu \rangle$$

exists and is strictly positive.

Then the following result holds:

Theorem 3.2. *Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied and that $N_j/\prod_{k\neq j} N_k = o(1)$, and $\prod_{k\neq j} N_k/N_j^3 = o(1)$. Then*

$$\prod_{k\neq j}\sqrt{N_k}\langle \hat{m}_{j,r} - m_{j,r}, \nu\rangle \xrightarrow{d} N\left(0, \sigma^2\sum_{k\neq r}\omega_{j,k}^2(\nu)\frac{d_r + d_k}{(d_r - d_k)^2}\right), \qquad N_1, \dots, N_d \to \infty.$$

Note that for $d = 3$, conditions of Theorem 3.2 are satisfied when the tensor dimensions grow proportionally, i.e. $N_1 \sim N_2 \sim N_3$. Our result can be compared to the asymptotic distribution of PCA when there is no underlying factor structure as follows. Let $Y_i \in \mathbb{R}^{N_1}$ be a random vector with $\mathbb{E}Y_i = 0$ and $\mathbb{E}[Y_iY_i^\top] = \Sigma$ with the eigendecomposition $\Sigma = \sum_{r=1}^{N_1} d_r m_{1,r}\otimes m_{1,r}$. Then it follows from Dauxois, Pousse, and Romain (1982) that the eigenvectors of the sample covariance matrix satisfy

$$\sqrt{N_2}\langle \hat{m}_{1,r} - m_{1,r}, \nu\rangle \xrightarrow{d} N\left(0, \sum_{j\neq r}\sum_{k\neq r}\omega_{1,k}^2(\nu)\frac{\mathbb{E}\left[\langle Y_i, m_r\rangle^2\langle Y_i, m_j\rangle\langle Y_i, m_k\rangle\right]}{(d_r - d_j)(d_r - d_k)}\right), \qquad N_2 \to \infty.$$

The expression of the asymptotic variance simplifies when $Y_i \sim N(0, \Sigma)$, in which case $\langle Y_i, m_k\rangle_{k\geq 1}$ are independent and

$$\sqrt{N_2}\langle \hat{m}_{1,r} - m_{1,r}, \nu\rangle \xrightarrow{d} N\left(0, \sum_{k\neq r}\omega_{1,k}^2(\nu)\frac{d_k d_r}{(d_r - d_k)^2}\right), \qquad N_2 \to \infty;$$

see also Anderson (1963).

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

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

$$\left(\frac{\hat{\sigma}_{r,j}^2 - \sigma_r^2}{\sigma_r}\right)_{1\leq r\leq R} \xrightarrow{d} N(0, 4\sigma^2 I_R),$$

provided that $N_j/\prod_{k\neq j} N_k = o(1)$ and $\prod_{k\neq j} N_k/N_j^3 = o(1)$ as $N_1, \dots, N_d \to \infty$.

It immediately follows from Theorem 3.3 that $\hat{\sigma}_{r,j}^2/\prod_{j=1}^d N_j$ is a consistent estimator of d_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. The test builds on the eigenvalue ratio statistics of Onatski (2009) for different unfoldings and the p-value combinations; see Vovk and Wang (2020). Specifically, we consider the following hypotheses

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

Let $\hat{\sigma}^2_{1,j} \ge \hat{\sigma}^2_{2,j} \ge \cdots \ge \hat{\sigma}^2_{N_j,j}$ be the eigenvalues of $\mathbf{Y}_{(j)}\mathbf{Y}_{(j)}^\top$. Under Assumption 3.3, the first R eigenvalues diverge from the rest to infinity. Consider the following statistics

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

and put

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

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

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

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

Theorem 4.1. *Suppose that Assumptions 3.1, 3.2 (i), and 3.3 are satisfied and that* $u_{i_1,\dots,i_d} \sim N(0,\sigma^2)$. *Suppose also that* $N_j/\tau + \prod_{k\ne j} N_k/(N_j\tau) = o(1)$. *Then under* H_0, $S_j \xrightarrow{d} Z$, *while under* H_1, *we have* $S_j \uparrow \infty$ *for every* $j \le d$.

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 $(\xi_1, \xi_2, \dots)$, 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 $1 \leq j \leq d$, where $F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{Z_i \leq z}$.

3. Combine the p-values e.g. using the min, median, or mean:

$$\begin{aligned} p_{\text{min}} &= d \min(p_1, \ldots, p_d), \\ p_{\text{median}} &= 2 \times \text{median}(p_1, \ldots, p_d), \\ p_{\text{mean}} &= \frac{2}{d}\sum_{j=1}^{d} p_j. \end{aligned} \tag{8}$$

The scaling by 2 is needed to obtain valid p-values; see Vovk and Wang (2020). Small p-values indicate that there is some important factor left out.

Remark 4.1. *Note that the reason why we introduce a novel test, while there have been many proposed in the 2-way standard factor model PCA literature, is the fact that we can combine p-values across the matricizations. We rely on p-value combination schemes because the different matricizations create complex dependence structures across the test statistics and the joint asymptotic distribution of* $(S_1, S_2, \ldots, S_d)$ *is unknown.*

Remark 4.2. *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 as well as to compare it to ALS. Before studying the finite sample properties of our estimator, we elaborate first on the issue of model fit and complexity and illustrate the effectiveness of dimension reduction with tensor data factor models compared to standard 2-way factor models.

5.1 Model Complexity

In anticipation of the empirical application we will use a slightly modified notation. We consider the following data generating process (DGP) for $\mathbf{Y} = \{y_{i,j,t}\} \in \mathbb{R}^{N\times J\times T}$, a 3-dimensional array of data:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \lambda_{i,r} \mu_{j,r} f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0, \tag{9}$$

where $u_{i,j,t}$ are idiosyncratic errors, $f_r = (f_{1,r}, \dots, f_{T,r})^\top$ are the factors, $\lambda_r = (\lambda_{1,r}, \dots, \lambda_{N,r})^\top$ and $\mu_r = (\mu_{1,r}, \dots, \mu_{J,r})^\top$ are the factor loadings.

The model in equation (9) can also be estimated using 2-way factor approach, which can be done by simply pooling all the data in the (i, j) dimensions into a single dimension. Formally, this can be achieved by unfolding the tensor $\mathbf{Y}$ into a matrix $\mathbf{Y}_{(3)} = \{y_{i,j,t}\} \in \mathbb{R}^{NJ\times T}$, and applying PCA to the resulting covariance matrix $\mathbf{Y}_{(3)}\mathbf{Y}_{(3)}^\top$. This would estimate the pooled loadings $\beta_{i,j,r} = \lambda_{i,r}\mu_{j,r}$ and factors $f_{t,r}$ in

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,j,r} f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0.$$

With the 3-way tensor factor model the number of parameters is $R \times (N + J + T)$ while the number of parameters in the 2-way factor model is $R \times (NJ + T)$ which is significantly larger. In addition, in the 2-way factor model one cannot separately identifying the loadings $\lambda_r \in \mathbb{R}^N$ and $\mu_r \in \mathbb{R}^J$ specific to each dimension. We use a notion of *Model Complexity*, defined as the number of parameters expressed as the percent of the data size, to compare the 2- and 3-way approaches. The lower the model complexity is, the better the dimensionality of the original data is being reduced. The model complexity of the 3-way factor model is $R \times (N + J + T)/(NJT)$, and the model complexity of the 2-way factor model is $R \times (NJ + T)/(NJT)$. Dimensionality Reduction is then defined as 1 - Model Complexity.

The entries in Table 1 are the model complexities of 3- and 2-way factor models for different sizes of the three data dimensions. The numbers in the table show the model complexities as the number of factors ranging from 1 to 10. In Panel A, we can see that when N and J are small, the 2-way factor model can be 4.67 times more complex than the 3-way factor model. For a 10 factor model, the 3-way approach

Table 1: Model Complexity of 3- versus 2-way Factor Model

Entries pertain to model complexity, defined as the number of parameters expressed as the percent of the data size, for 3- and 2-way factor models for different sizes of the three data dimensions.

	Number of factors									
Model	1	2	3	4	5	6	7	8	9	10
Panel A: $T = 100, N = 30, J = 20$										
3-way	0.25%	0.5%	0.75%	1%	1.25%	1.5%	1.75%	2%	2.25%	2.5%
2-way	1.17%	2.33%	3.5%	4.67%	5.83%	7%	8.17%	9.33%	10.5%	11.67%
Panel B: $T = 50, N = 50, J = 50$										
3-way	0.12%	0.24%	0.36%	0.48%	0.60%	0.72%	0.84%	0.96%	1.08%	1.2%
2-way	2.04%	4.08%	6.12%	8.16%	10.2%	12.24%	14.28%	16.32%	18.36%	20.4%
Panel C: $T = 50, N = 100, J = 100$										
3-way	0.05%	0.1%	0.15%	0.2%	0.25%	0.3%	0.35%	0.4%	0.45%	0.5%
2-way	2.01%	4.02%	6.03%	8.04%	10.05%	12.06%	14.07%	16.08%	18.09%	20.1%

achieves a dimension reduction of 97.5%, whereas the 2-way approach only achieves a reduction of 88.33%. In Panel B, where N and J are of the same size as T, the 2-way model is 17 times more complex than the corresponding 3-way model. The dimension reduction for the latter is 98.8% , whereas that of the 2-way model is only 79.6%. Finally, in Panel C, where the dimensions of N and J are twice as large as T, the 2-way factor model is 40.2 times more complex, which means that the number of parameters to be estimated in the 2-way model are 40.2 times the number for the 3-way model, with a dimension reduction of the latter equal to 99.5%, while it is only 79.9% for the former.

The calculations reported in Table 1 show that when 3-dimensional tensor data is available it is advisable to forgo using the traditional 2-way factor model in favor of 3-way factor models. The latter has the additional benefit that one can identify the loadings specific to each dimension.

5.2 Model Fit

While the 3-way factor model is less complex, it may not be good at data fitting. To assess model fit we compare 2- and 3-way factor models in terms R^2 defined as $1 - \mathrm{RSS}/\mathrm{TSS}$, where $\mathrm{RSS} = \sum_{i,j,t} \hat{u}_{i,j,t}^2$ and $\mathrm{TSS} = \sum_{i,j,t} y_{i,j,t}^2$ The higher the R^2 is, the better the variation of the data is being explained by the model. To make the assessments of model fit, we conduct a simulation study. We need to expand the DGP in equation (9), namely we need to specify a model for the factors to simulate their sample paths. We consider a 3-way factor model with R number of factors/loadings, namely factors $f_r \in \mathbb{R}^T$, and loadings $\lambda_r \in \mathbb{R}^N$, $\mu_r \in \mathbb{R}^J$, where $r= 1, \ldots, R$. The data are generated as follows:

$$\begin{aligned} y_{i,j,t} &= \sum_{r=1}^{R} \sigma_r \lambda_{i,r} \mu_{j,r} f_{t,r} + u_{i,j,t}, \qquad u_{i,j,t} \sim_{\text{i.i.d.}} N(0, s_u^2), \\ \dot{f}_{t,r} &= \rho \dot{f}_{t-1,r} + \varepsilon_{t,r}, \qquad \varepsilon_{t,r} \sim_{\text{i.i.d.}} N(0, s_\varepsilon^2), \end{aligned} \tag{10}$$

where factors are normalized to be unit-norm by taking $f_r = \dot{f}_r / \|\dot{f}_r\|$. The loadings λ, μ are randomly generated orthonormal vectors by using the following procedure: (a) generate $N \times N$ (or $J \times J$) matrix A with entries uniformly distributed on $[0, 1]$, (b) compute a symmetric matrix $B = A^\top A$, and finally (c) define λ_r (or μ_r) as the r^{th} (orthonormal) eigenvector of B. The parameters are set as follows:

(1) The AR(1) process of $\dot{f}_r$ takes $\rho = 0.5$ and $s_\varepsilon = 0.1$.

(2) We set the signal strength $\sigma_r = d_r \times \sqrt{NJT}$ with decreasing $d_r = R - r + 1$ to ensure the model is correctly identified, and the strength of the noise $s_u = 1$.

(3) We consider the cases where we generate the true number of factors $R \in \{5, 10\}$, and the dimension sizes of the tensor $(T, N, J) \in \{(100, 30, 20), (50, 50, 50), (50, 100, 100)\}$.

We estimate the number of factors from 1 to R without knowing the true number of factors. In each repetition of the 5000 Monte Carlo simulations, we only allow $u_{i,j,t}$ to be changing, so the factors and loadings are generated only once and kept the same for all repetitions. And we repeat the simulation for 5000 times, and report the average R^2.

Figure 1 plots the average R^2 against the number of parameters for the 2- and 3-way factor approaches under a 5 and 10 factor model and three different size parameterizations. All panels show that the complexity of the 2-way factor model grows significantly faster than that of the 3-way factor model. More importantly, with the same number of parameters, 3-way factor model is capable of explain much more variation of the data. In panel (a) and (b), the green dashed lines locate the intersections of the two approaches estimating the same number of parameters. In panel (a), the green dashed line shows when 3-way factor model explains almost 97% of the variation, the 2-way is only 45%. Similarly, in panel (b), the two dashed lines show when 3-way explain about 82% and 99%, the 2-way is still at about 26% and 47%. There is no dashed line plotted in panels (c) - (f) because the simplest one factor 2-way model is more complex than a 10 factor 3-way model. This means the 2-way model is even more over-parameterized when dimensions N, J are larger than T.

5.3 Comparison to ALS

In this subsection, we compare our TPCA algorithm with the benchmark Alternating Least Squares (ALS) method. The objective is to show that, without the knowledge of the true number of factors R, ALS will not correctly identify the factors/loadings. For example, the first R_0 factors computed with ALS will be different depending on the number of factors assumed. In contrast, our TPCA relies on the eigendecomposition which computes all factors at once and the first R_0 factor will always be numerically the same regardless of the total number of specified factors.[8]

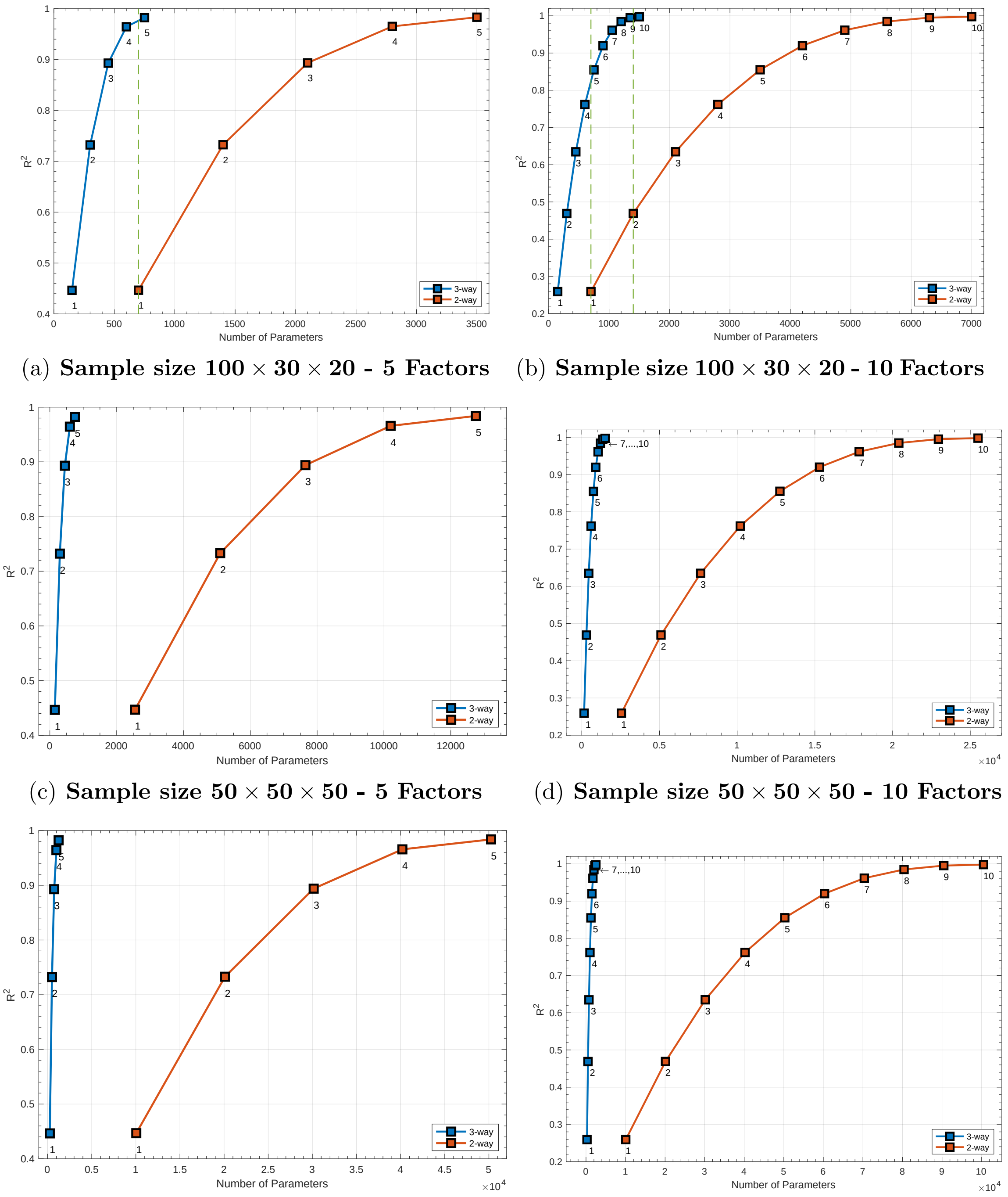
The data are generated the same as equation (10), with the parameters set as follows: (1) the AR(1) process of $\dot{f}_r$ takes $\rho = 0.5$ and $s_\varepsilon = 0.1$, (2) we set the signal strength $\sigma_r = d_r \times \sqrt{NJT}$ with decreasing $d_r = R - r + 1$, and the noise strength $s_u = 1$, (3) the size of each dimension of the tensor $T = 100$, $N = 30$, $J = 20$ and finally (4) we consider cases where we generate the true number of factors $R \in \{1, 2, 3, 4, 5\}$, and we always estimate a 1 factor model without the knowledge of the true R.

We evaluate the estimates using the ℓ_2 criterion from the theory. As the signs of $\hat{\lambda}_r, \hat{\mu}_r$, and $\hat{f}_r$ are undetermined, we calculate the error and select the signs of the

[8] The ALS algorithm toolbox we use is composed by Bader and Kolda (2022), and is accessible via https://www.tensortoolbox.org/

Figure 1: Plots of R^2 against Number of Parameters of 3- vs. pooled 2-way Factor Model

The DGP appears in equation (10). We consider the cases where $R \in \{5, 10\}$ and $(T, N, J) \in \{(100, 30, 20), (50, 50, 50), (50, 100, 100)\}$, and we estimate the number of factors from 1 to R without knowing the true number of factors. We repeat the simulation for 5000 times, and report the average R^2.



(a) **Sample size 100 × 30 × 20 - 5 Factors**

(b) **Sample size 100 × 30 × 20 - 10 Factors**

(c) **Sample size 50 × 50 × 50 - 5 Factors**

(d) **Sample size 50 × 50 × 50 - 10 Factors**

(e) **Sample size 50 × 100 × 100 - 5 Factors**

(f) **Sample size 50 × 100 × 100 - 10 Factors** estimates by taking the losses as follows:

$$\begin{aligned}\mathbb{L}_\lambda &= \|\hat{\lambda}_r \times \operatorname{sign}(\hat{\lambda}_r^\top \lambda_r) - \lambda_r\|,\\ \mathbb{L}_\mu &= \|\hat{\mu}_r \times \operatorname{sign}(\hat{\mu}_r^\top \mu_r) - \mu_r\|,\\ \mathbb{L}_f &= \|\hat{f}_r \times \operatorname{sign}(\hat{f}_r^\top f_r) - f_r\|,\end{aligned} \tag{11}$$

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

Figure 2 plots the histograms of the ℓ_2 losses of the two algorithms using 5000 simulation repetitions, where in each panel we have discontinuous horizontal axis because the estimation errors are large when the estimates deviate too far from the true parameters. Panels (a) - (c) show that, when the number of factors is correctly, i.e. $R = 1$, then the estimates from ALS and TPCA have comparable accuracy. Nonetheless, the distribution of ALS errors (blue histogram) has a few outliers - which is why we break up the horizontal axis and display the results across two plots. The outliers occur because the ALS algorithm can be trapped in local optima. As we move on to the tensor factor model with more than one factor, we see that the performance of ALS quickly deteriorates to become heavily right-tailed and it is worst when $R = 4$ in our experiments. This is because the ALS algorithm requires the knowledge of the true number of factors to correctly identify the parameters, and it does not allow to compute the factors/loadings sequentially. Among other things, it is an important advantage of our TPCA estimator that it does not require the number of selected factors to coincide with the true number.

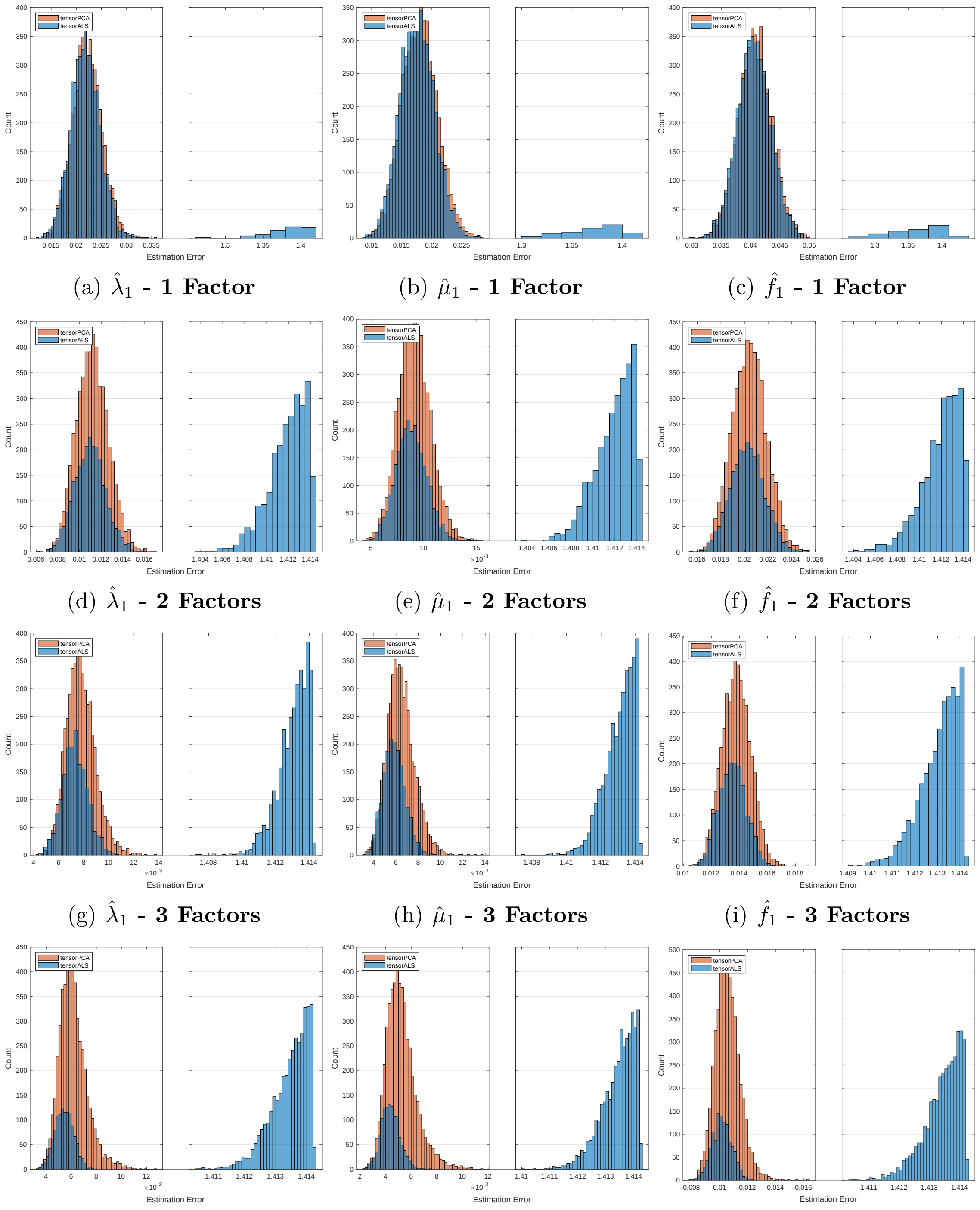
5.4 Finite Sample Properties

In this subsection, we 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 Corollary 3.1. The corollary states that the convergence rate is roughly $O_P(1/\prod_{k\neq j}\sqrt{N_k})$ for the loading/factor vector in the j^{th} dimension, which means the rate for the loading vector λ_r is $O_P(1/\sqrt{JT})$, for the loading vector μ_r it is $O_P(1/\sqrt{NT})$, and $O_P(1/\sqrt{NJ})$ for the factor vector f_r.

We use again the data generating process appearing in equation (10), with the parameters set as follows: (1) the AR(1) process $\dot{f}_r$ takes $\rho = 0.5$ and $s_\varepsilon = 0.1$, (2) we generate and estimate only a one-factor model for the purpose of demonstrating

Figure 2: Estimation accuracy: TPCA vs. ALS

The DGP appears in equation (10). We plot the histograms of the ℓ_2 losses of the estimated first factor/loading of TPCA (orange) vs. ALS (blue) in 5000 MC simulations. The horizontal axis is split across two plots because estimates (for ALS) sometimes deviate very far from the true parameters.



(a) $\hat{\lambda}_1$ **- 1 Factor**

(b) $\hat{\mu}_1$ **- 1 Factor**

(c) $\hat{f}_1$ **- 1 Factor**

(d) $\hat{\lambda}_1$ **- 2 Factors**

(e) $\hat{\mu}_1$ **- 2 Factors**

(f) $\hat{f}_1$ **- 2 Factors**

(g) $\hat{\lambda}_1$ **- 3 Factors**

(h) $\hat{\mu}_1$ **- 3 Factors**

(i) $\hat{f}_1$ **- 3 Factors**

(j) $\hat{\lambda}_1$ **- 4 Factors**

(k) $\hat{\mu}_1$ **- 4 Factors**

(l) $\hat{f}_1$ **- 4 Factors**

finite sample properties, i.e. $R = 1$, (3) we set the signal strength $\sigma_1 = \sqrt{NJT}$, and the noise strength $s_u = 1$, (4) the sample size of the baseline model is $T = 100$, $N = 30$, $J = 20$, 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, $(T, N, J) = (200{,}60{,}40)$, (b) doubling the sample sizes of two dimensions, $(T, N, J) = (100{,}60{,}40)$, (c) doubling the sample size of only one dimension, $(T, N, J) = (100{,}60{,}20)$. Finally, we evaluate the estimates using the MSE as in equation (11).

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 Corollary 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 0.016/0.022 while the improvement for $\hat{\mu}_1$ is 0.013/0.018. 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 0.02/0.041 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 $\hat{\mu}_1$ is 0.013/0.018 while the improvement for $\hat{f}_1$ is 0.029/0.041. Both are aligned with the $1/\sqrt{2}$ improvement factor. Overall these results show that the predictions of the asymptotic theory are valid in finite samples.

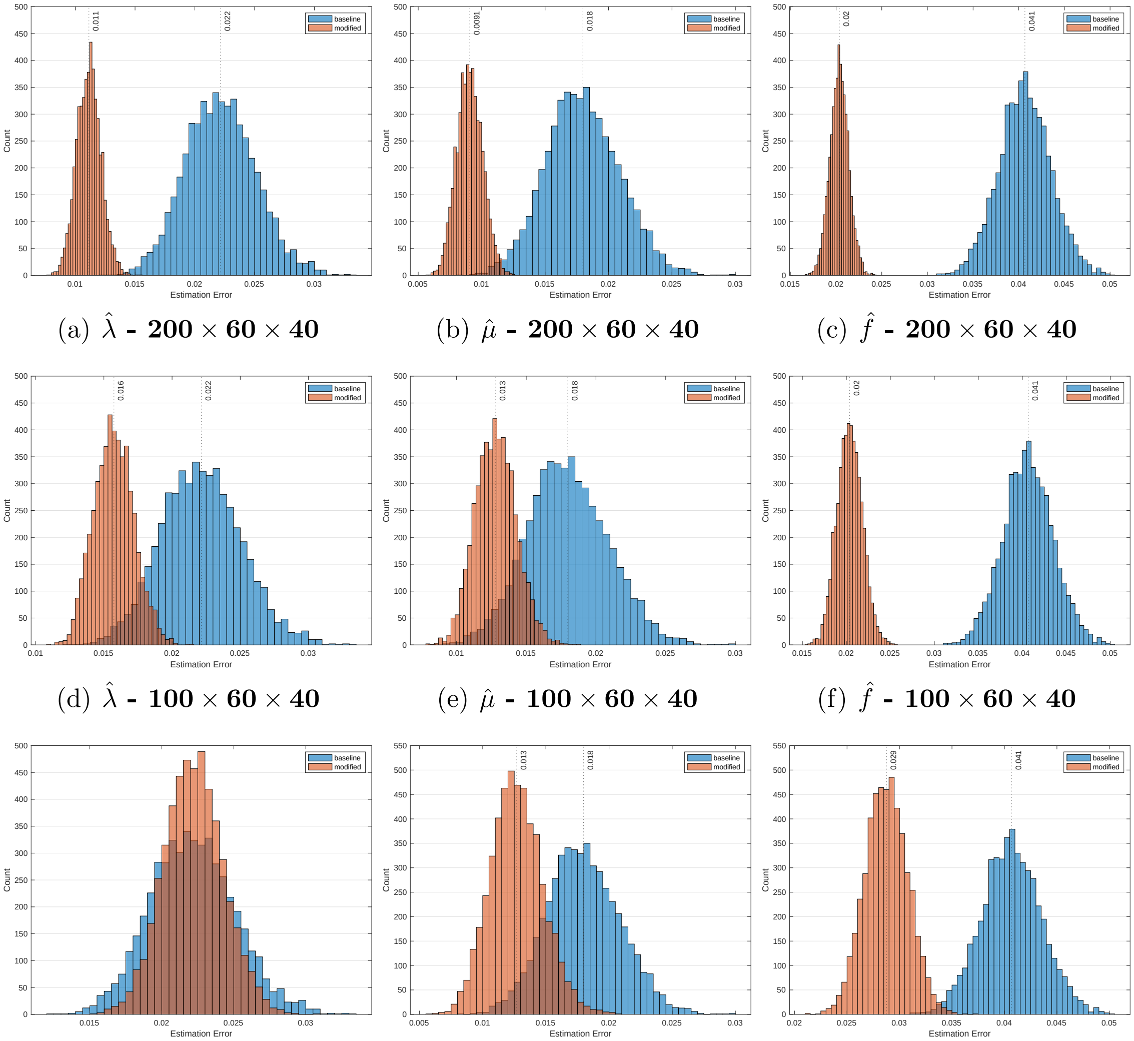
5.5 Testing the Number of Factors

We conclude with power properties of our maximum eigenvalue ratio test for the number of factors discussed in Section 4. The DGP is generated using equation (10) as in previous simulations. We generate a 2-factor model, and test the null hypothesis that there is only 1 factor against the alternative that there are more than 1 but less than K factors. Since the test is performed on each matricization of the tensor, we also examine the effectiveness of p-value combinations as studied by Vovk and Wang (2020).

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

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

The DGP appears in equation (10). 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 (T, N, J) = (100,30,20), and modified DGP has sizes (T, N, J) 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 plots the mean of the ℓ_2 errors.



(a) $\hat{\lambda}$ **- 200 × 60 × 40**

(b) $\hat{\mu}$ **- 200 × 60 × 40**

(c) $\hat{f}$ **- 200 × 60 × 40**

(d) $\hat{\lambda}$ **- 100 × 60 × 40**

(e) $\hat{\mu}$ **- 100 × 60 × 40**

(f) $\hat{f}$ **- 100 × 60 × 40**

(g) $\hat{\lambda}$ **- 100 × 60 × 20**

(h) $\hat{\mu}$ **- 100 × 60 × 20**

(i) $\hat{f}$ **- 100 × 60 × 20**

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, (2) we study cases where the idiosyncratic errors are generated with Gaussian distribution or student's t distribution (the degree of freedom is 5), and in both cases the variance of the errors is normalized to be $\sigma_u = 1$, (3) we study p-value combinations using both 3-way and 5-way tensor, and respectively the sizes of the tensors are $60 \times 80 \times 100$ and $10 \times 20 \times 30 \times 40 \times 50$, with the first dimension being the smallest.

The p-value combination strategies we consider are maximum, minimum, median, and mean. The p-value combinations have to be scaled properly in order to be considered as valid p-values, please see Vovk and Wang (2020) for discussion of different p-value combinations. Specifically, for a d-way we combine the p-values using the $p_{\min}$, p_{median}, and p_{mean} rules - see equation (8). We also a fourth combination rule: $p_{\max} := \max(p_1, \ldots, p_d)$.

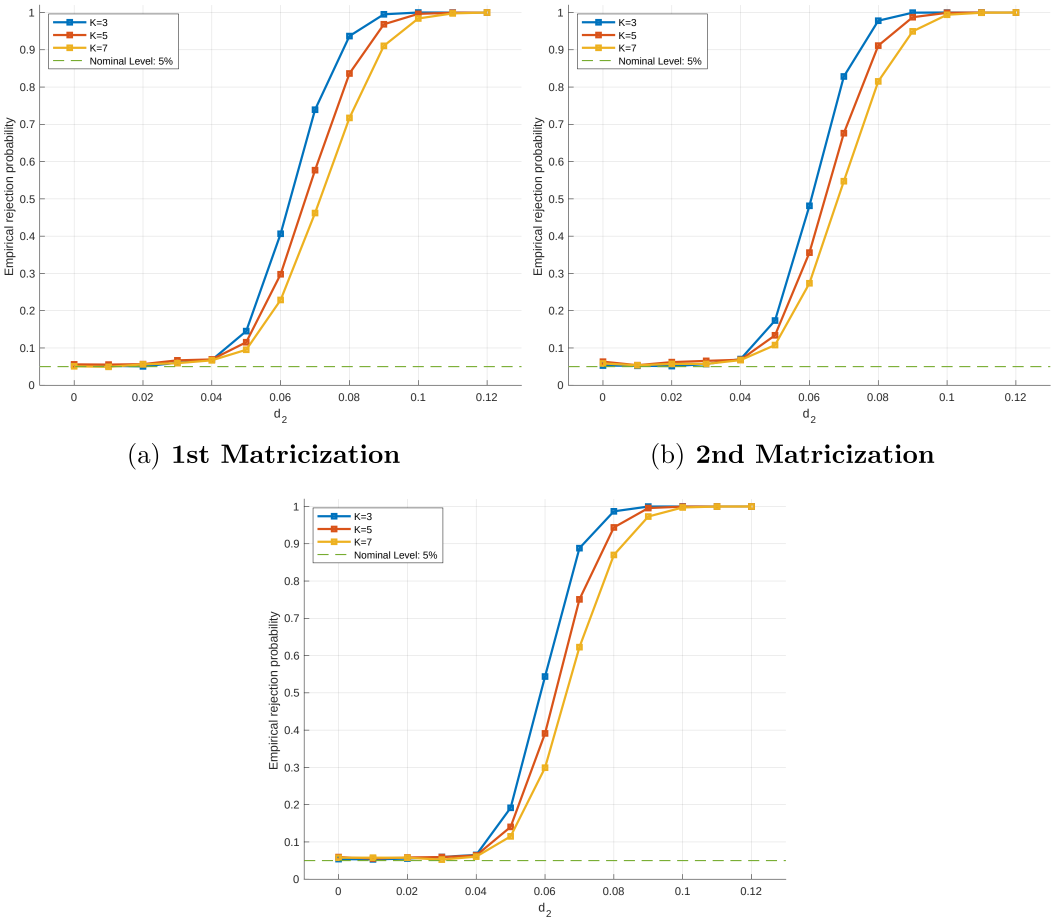
We perform the test for the number of factors as discussed in Section 4, where we approximate the asymptotic distribution of the statistics by randomly generating Gaussian matrices 5,000 times, and we also replicate the simulations for 5,000 times to calculate the empirical rejection probability for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

Figure 4 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%, and the empirical power of test reaches 1 as 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 among different matricizations and different p-value combination strategies under 3 cases. The three rows of Figure 5 correspond to three different cases: (1) 3$^{\text{rd}}$ order tensor with Gaussian errors, (2) 3$^{\text{rd}}$ order tensor with student's t distributed errors, (3) 5$^{\text{th}}$ order tensor with Gaussian errors. Among different matricizations, the dimension with the largest size ('mat3' for the 3$^{\text{rd}}$ order tensor, and 'mat5' for the 5$^{\text{th}}$ order tensor) tend to climb fastest to one, meaning the matricization with the largest dimension gives the highest empirical

Figure 4: 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 factor 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.



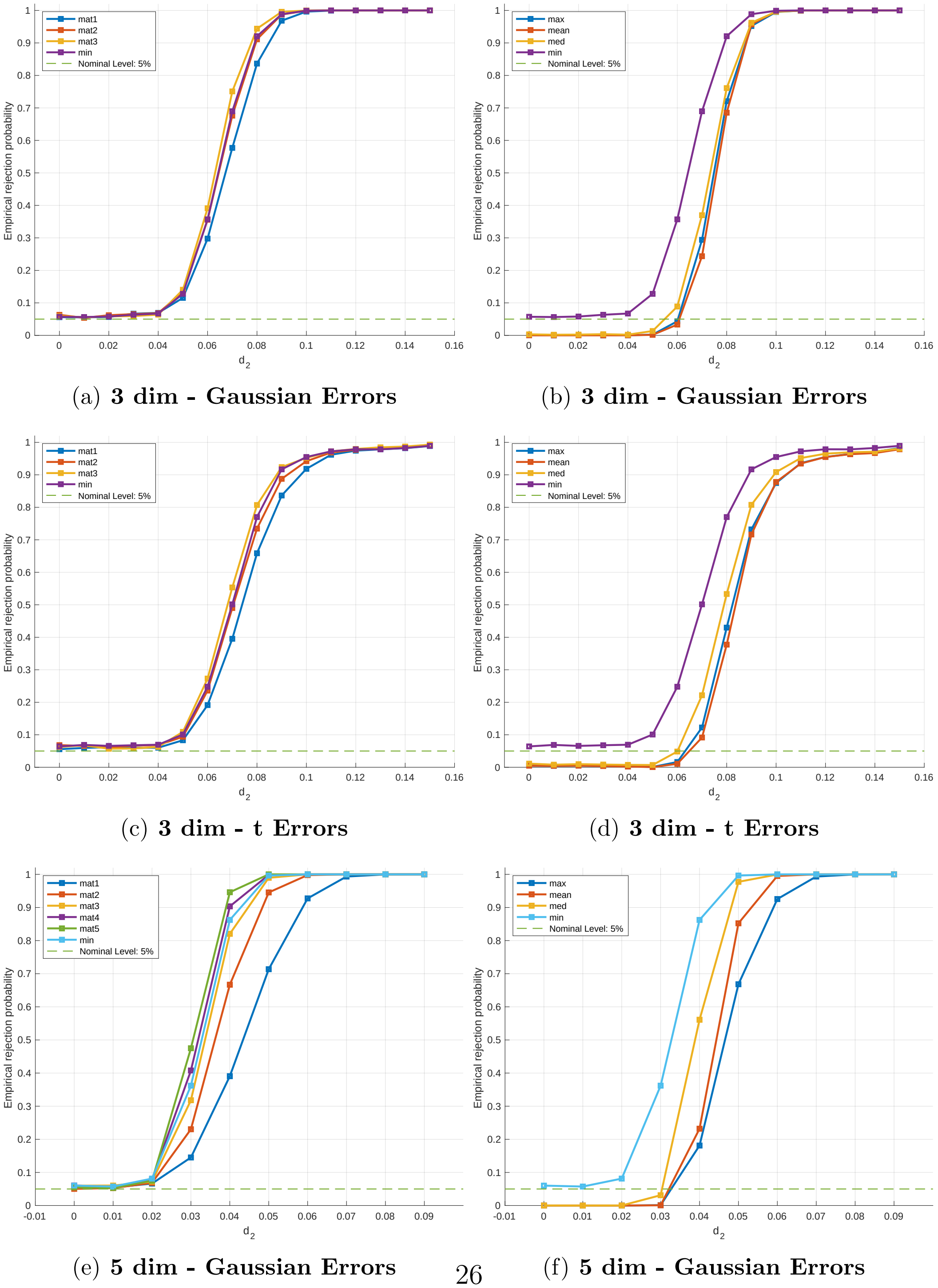
(a) **1st Matricization**

(b) **2nd Matricization**

(c) **3rd Matricization**

Figure 5: 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 and 5-dim tensor. We report the empirical power of testing on individual matricizations (left) as well as different p-value combination schemes (right).



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

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

(c) **3 dim - t Errors**

(d) **3 dim - t Errors**

(e) **5 dim - Gaussian Errors**

(f) **5 dim - Gaussian Errors**

power. When compared to various p-value combinations, the “minimum” performs similar to individual matricizations, but not as good as the largest dimension. The other combinations perform overly conservative, and thus distort the empirical size downwards.

We also note from Figure 5 plots (e)-(f) that the power of the test improves when the number of dimensions of a tensor increases. Finally, while our simulation results rely on the critical values that assume Gaussian errors, we learn from from Figure 5 plots (c)-(d) that the test performs well for non-Gaussian distributions albeit with some loss of power.

6 Empirical Illustration: Sorted Portfolios

One can think of many applications of tensor factor models discussed in the previous sections. In this section we only provide an illustrative financial application which focuses on the common practice of characteristic-based sorted portfolios to estimate systematic risks in asset pricing models.[9] We start from a specific recent example, namely Lettau and Pelger (2020) use 37 anomaly sorted portfolios to estimate systematic risk factors via PCA (and a variation called RP-PCA), where the data is a 3-dimensional tensor of anomalies, deciles, and time. They treat all decile portfolios of the anomalies as individual assets, which adds up to 370 assets. This in fact is the same procedure as the (pooled) 2-way factor model approach mentioned in Section 5: unfold a 3-way tensor into a matrix and apply PCA to estimate the factors/loadings. As previously discussed, there are two main issues with this approach: (1) the model is over-parameterized, the (2) loadings specific to anomalies and deciles are not identified. Therefore, we propose a better solution to such a 3-dimensional array using our tensor factor model. To that end, consider a 3-way factor model:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \lambda_{i,r} \mu_{j,r} f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0, \tag{12}$$

where $y_{i,j,t}$ is the excess return of the j^{th}, $j = 1, \ldots, J$ quantile of the i^{th}, $i = 1, \ldots, N$ characteristic at time $t = 1, \ldots, T$, $u_{i,j,t}$ is the idiosyncratic shock, $f_r \in \mathbb{R}^T$ are

[9] A more substantial empirical analysis of tensor factor asset pricing models appears in Babii, Ghysels, and Pan (2023).

the systematic risk factors driving the excess returns, loadings $\lambda_r \in \mathbb{R}^N$ determines the heterogeneous exposure of each characteristic to the r^{th} risk factor, loadings μ_r determines the exposure of each quantile to the r^{th} risk factor, and σ_r absorbs all the scales of factors and loadings. The number of parameters to be estimated in this model is $R \times (N + J + T)$. The traditional 2-way factor model approach pools all the characteristics deciles:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \beta_{i,j,r} f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0,$$

where $\beta_{i,j,r} = \lambda_{i,r}\mu_{j,r}$. The number of estimated parameters in this model is $R \times (NJ + T)$. Therefore, the goal in this section is to estimate the loadings that are specific to characteristics and deciles using the 3-way factor model, and provide interpretations of the loadings which is unique to a tensor factor model.

To conduct the empirical analysis, we rely on Chen and Zimmermann (2021) who collected over 200 characteristic-sorted portfolios from previous studies of stock market anomalies.[10] We consider the monthly portfolio returns, which are sorted into 10 deciles based on firm level characteristics, from Jan. 1990 to Dec. 2019. We only consider a balanced set of portfolios, and therefore the number of characteristics throughout the entire sample period is 133. Hence, the 3-dimensional tensor we consider is of size $N \times J \times T$, with $N = 133$, $J = 10$, $T = 360$, and the total number of observations is $NJT = 478{,}800$. We also use the risk-free rate from the March 2022 release of Kenneth French data library to compute excess returns. We estimate the 3-way factor model with two factors in equation (12) using both TPCA and ALS, and compare the estimates from the two different algorithms.[11]

Since the market portfolio is a dominant factor we report results for the residuals of the CAPM, i.e. returns projected on the market portfolio. We find that there are two additional factor, when we rely on the largest dimension (time T) testing strategy. Namely, the test results indicate that we reject the null $R \leq 1$ and while accepting $R \leq 2$ and higher.[12]

[10] The data we use is the March 2022 release of the database "Open Source Cross-Sectional Asset Pricing" created by Chen and Zimmermann (2021).

[11] Although ALS does not impose the orthogonality restriction, we obtain orthogonalized loadings and factors by applying the Gram-Schmidt transformation.

[12] This evidence also corresponds to findings reported recently by Andreou, Gagliardini, Ghysels, and Rubin (2022) who show that the true factors are those common between conditional asset pricing

Table 2 reports summary statistics of estimated loadings $\hat{\lambda}$ that determine the exposure of all 133 characteristics to the three factors. For TPCA, the last column shows that the values of $\hat{\lambda}_{i,1}$'s are strictly positive for all characteristics $i = 1, \ldots, 133$, while the $\hat{\lambda}_{i,2}$'s and $\hat{\lambda}_{i,3}$'s are about half positive and half negative. Moreover, $\hat{\lambda}_{i,1}$ has a maximum of 0.1133 and a minimum of 0.0658 with a very small standard deviation of 0.0087. In comparison, the $\hat{\lambda}_{i,2}$ are symmetric around zero, with a maximum of 0.2715 and a minimum of -0.2681 and relatively larger standard deviation of 0.0870. While the $\hat{\lambda}_{i,3}$ estimates aren't symmetric around zero, it is worth noting that on average both $\hat{\lambda}_{i,2}$ and $\hat{\lambda}_{i,3}$ are roughly zero.

As for ALS, the values of $\hat{\lambda}_{i,1}$'s are no longer strictly positive for all characteristics, with about 20% of them being negatively exposed to the first factor (i.e. the market), and also about 96% are positively exposed to the second factor, which is considerably different from TPCA. This is obviously not an issue of sign indeterminacy, which would imply a different sign uniformly across all characteristics. The differences are most evident in $\hat{\lambda}_1$ where ALS has a much larger standard deviation of 0.0759 than 0.0087 for TPCA.

Table 3 reports the estimates of the loadings $\hat{\mu}$ that determine the exposure of all 10 deciles. For TPCA, the values of $\hat{\mu}_{j,1}$'s are around 0.3 for $j = 1, \ldots, 10$, and they are largest on the two extremes and smaller in the middle. In contrast, $\hat{\mu}_{j,2}$ is largest in the first decile and is monotonically decreasing from the 1^{st} to the 10^{th} decile, with the absolute values of $\hat{\mu}_{1,2}$ and $\hat{\mu}_{10,2}$ being very close and around 0.5. In contrast, $\hat{\mu}_{j,3}$ is again U-shaped and in fact the last column, showing 10 minus 1, reveals that the two extremes cancel out for $\hat{\mu}_{1,1}$ versus $\hat{\mu}_{10,1}$ and $\hat{\mu}_{1,3}$ versus $\hat{\mu}_{10,3}$. Hence, high-minus-low portfolio sorts will bring out the exposure to the second factor, while canceling the first and third.

The asset pricing literature commonly uses high-minus-low portfolios as a proxy for risk premium, which is basically using the returns of the last decile portfolio minus the returns of the first. In the context of ten deciles, the high-minus-low portfolios are

$$y_{it}^{10-1} = \sigma_1 \lambda_{i,1} (\mu_{10,1} - \mu_{1,1}) f_{t,1} + \sigma_2 \lambda_{i,2} (\mu_{10,2} - \mu_{1,2}) f_{t,2} + u_{it}^{10-1},$$

models for individual stocks and models estimated from sorted portfolios. They find three common factors. These are not the Fama-French 3 factors and they are not even spanned by the Fama-French 5 factors. More importantly, they feature superior out-of-sample pricing performance compared to standard asset pricing models.

where $y_{it}^{10-1} = y_{i,10,t} - y_{i,1,t}$ and $u_{it}^{10-1} = u_{i,10,t} - u_{i,1,t}$. For TPCA, since $\hat{\mu}_{1,1}$ is very close to $\hat{\mu}_{10,1}$, the first term on the right-hand side cancels out when taking the difference between the two extreme deciles. Therefore the risk premium associated with characteristics are driven solely by the second risk factor. Note that we did not impose symmetry. Instead TPCA yields such estimates which support the common practice.

As for ALS, the pattern looks similar to that of TPCA, where $\hat{\mu}_1$ has smaller decile loadings in the middle and has the largest decile loadings on two extremes, and $\hat{\mu}_2$ is monotonically increasing. The difference is that $\hat{\mu}_1$ and the absolute value of $\hat{\mu}_3$ are no longer symmetric compared to the TPCA estimates. Hence, ALS does not coincide with the common practice of computing high-minus-low portfolio sorts. This means ALS does not support the common practice of high-low sorting.

The tensor factor model estimated by TPCA confirms that taking the difference between two extreme deciles is actually a proper proxy for the risk premium of the second factor beyond the market. The reason for some $\hat{\lambda}_{i,2}$ being positive and some being negative is because some characteristics have a positive while others have a negative association with risk beyond the market. The tensor factor model also confirms that using the difference of two extreme deciles is better than the difference of middle deciles, e.g. 9-2 or 8-3, because this gives the highest risk premium associated with characteristics. The right panel of Table 4 lists the top 10 and bottom 10 characteristics exposed to the second systematic risk factor in absolute value terms, estimated by TPCA. Among the 133 characteristics, "Bid-ask spread", "Idiosyncratic risk (AHT)", and "CAPM beta" have the highest exposure to the second systematic risk, whereas "Earnings Surprise", "Market leverage", and "Real estate holdings" have the lowest exposure. Higher exposure to the second systematic risk means higher risk premium associated with this characteristic.[13] Among the 133 characteristics, "Firm Age - Momentum", "Idiosyncratic risk (AHT)", and "Price" have the highest exposure to the market, whereas "Volume Variance", "Frazzini-Pedersen Beta", and "Price delay r square" have the lowest exposure.

[13] Table A.1 in the Appendix provides the descriptions of the acronyms and references.

Table 2: Summary Statistics of Estimated Loadings $\hat{\lambda}_r$ specific to Characteristics

The table reports the summary statistics of $\hat{\lambda}$ for the 3-factor model appearing in equation (12) estimated with TPCA versus ALS. The columns each report the maximum, average, minimum, standard deviation, and the percent of values greater than zero.

	Max	Mean	Min	Std.	> 0
			Tensor PCA		
$\hat{\lambda}_1$	0.1133	0.0863	0.0658	0.0087	100%
$\hat{\lambda}_2$	0.2715	0.0027	-0.2681	0.0870	58.65%
$\hat{\lambda}_3$	0.4605	-0.0055	-0.1549	0.0869	43.61%
			ALS		
$\hat{\lambda}_1$	0.2740	0.0424	-0.2060	0.0759	81.20%
$\hat{\lambda}_2$	0.2447	0.0748	-0.0254	0.0441	96.24%
$\hat{\lambda}_3$	0.1626	0.0105	-0.3130	0.0864	60.15%

7 Conclusion

Modern datasets are often multidimensional beyond the 2-dimensional panel data structure 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 widely used 2-way factor models. We show that the d-way factor models can be estimated with a variation of the PCA estimator which we call TPCA. Unlike the ALS algorithm, which is commonly used for this purpose, our TPCA does not involve solving a non-convex optimization problem and has a closed-form expression.

We provide convergence rates and large sample approximations to distribution for our estimator, demonstrating its advantages over ALS. We also propose the first formal statistical test for the number of factors in a tensor factor model. We find

Table 3: Estimated Loadings $\hat{\mu}_r$ specific to Deciles

The table reports $\hat{\mu}$ of a 3-factor model in equation (12) estimated by TPCA and ALS. The values in the $j^{\text{th}}, j = 1, \ldots, 10$ column represent the estimated exposure of the j^{th} decile to the first two factors.

Decile	1	2	3	4	5	6	7	8	9	10	10 - 1
						Tensor PCA					
$\hat{\mu}_1$	0.3779	0.3423	0.3170	0.3021	0.2924	0.2846	0.2866	0.2961	0.3097	0.3406	-0.0373
$\hat{\mu}_2$	0.5259	0.3719	0.2289	0.1225	0.0216	-0.0689	-0.1660	-0.2769	-0.3823	-0.5119	-1.0378
$\hat{\mu}_3$	0.5475	0.0477	-0.1901	-0.2932	-0.3225	-0.3133	-0.2486	-0.1217	0.1118	0.5335	-0.0140
						ALS					
$\hat{\mu}_1$	0.4417	0.3925	0.3481	0.3186	0.2927	0.2700	0.2567	0.2517	0.2541	0.2752	-0.1665
$\hat{\mu}_2$	-0.3585	-0.3308	-0.2346	-0.1433	-0.0369	0.0752	0.2003	0.3261	0.4483	0.5762	0.9347
$\hat{\mu}_3$	-0.5753	-0.1228	0.1510	0.2659	0.3422	0.3537	0.3000	0.1412	-0.0761	-0.4500	0.1253

that our tensor factor model provides an efficient dimensional reduction relatively to the naively pooled traditional factor models. At the same time, the model is parsimonious with easily identifiable factors and loadings in contrast to the tensor factor model based on the Tucker decomposition. These findings are supported by the extensive simulation results. Lastly, we also consider an empirical application to sorted portfolios.

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

Table 4: Top and Bottom 10 Characteristics of Loadings $\hat{\lambda}_1$ and $|\hat{\lambda}_2|$

This table lists the top 10 and bottom 10 characteristics in terms of exposure to the first two factors estimated by TPCA. The left panel is sorted by $\hat{\lambda}_1$, and the right panel is sorted by $|\hat{\lambda}_2|$. The top 10 chacteristics are sorted in descending order, whereas bottom 10 are sorted in ascending order.

	$\hat{\lambda}_1$		$\lvert\hat{\lambda}_2\rvert$	
	Top 10	Bottom 10	Top 10	Bottom 10
1	FirmAgeMom	VolSD	BidAskSpread	EarningsSurprise
2	IdioVolAHT	BetaFP	IdioVolAHT	Leverage
3	Price	PriceDelayRsq	Beta	realestate
4	OrderBacklogChg	MomOffSeason16YrPlus	IdioVol3F	InvGrowth
5	IdioVol3F	DolVol	IdioRisk	PriceDelaySlope
6	RDAbility	MomSeason16YrPlus	Price	ShareIss5Y
7	IdioRisk	PriceDelaySlope	MaxRet	grcapx
8	OrderBacklog	MeanRankRevGrowth	High52	VolumeTrend
9	High52	FR	BetaFP	VolSD
10	Mom12m	EP	FEPS	ChEQ

References

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

T. W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

E. Andreou, P. Gagliardini, E. Ghysels, and M. Rubin. Three common factors. CEPR Discussion Paper No. DP17225, 2022.

A. Babii, E. Ghysels, and J. Pan. Tensor factor asset pricing models. Discussion Paper, UNC, 2023.

B. Bader and T. Kolda. Tensor Toolbox for MATLAB, Version 3.4. https://www.tensortoolbox.org/, 2022.

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

J. Bai and S. Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

J. Beyhum and E. Gautier. Factor and factor loading augmented estimators for panel regression. *arXiv preprint arXiv:2010.01837*, 2020.

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

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.

A. Y. Chen and T. Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review, Forthcoming*, 2021.

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.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, 12(1):136–154, 1982.

V. De Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30 (3):1084–1127, 2008.

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.

J. Fan and W. Wang. Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733*, 2015.

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

Y. Han, R. Chen, D. Yang, and C.-H. Zhang. Tensor factor model estimation by iterative projection. *arXiv preprint arXiv:2006.02611*, 2020.

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

R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

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.

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

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

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

M. Lettau and M. Pelger. Factors that fit the time series and cross-section of stock returns. *Review of Financial Studies*, 33:2274–2325, 2020.

L. Matyas. The econometrics of multi-dimensional panels. *Advanced studies in theoretical and applied econometrics. Berlin: Springer*, 2017.

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.

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.

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.

C. Squires, D. Shen, A. Agarwal, D. Shah, and C. Uhler. Causal imputation via synthetic interventions. In *Conference on Causal Learning and Reasoning*, pages 688–711. PMLR, 2022.

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

L. R. Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23:111–136, 1958.

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.

V. Vovk and R. Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

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.

APPENDIX

A.1 Graphical Illustration of Tensors

Figure A.1: A scalar, 1st order, 2nd order, and 3rd order tensors

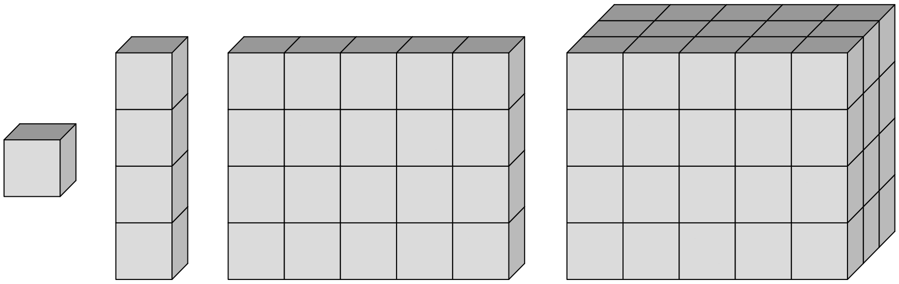
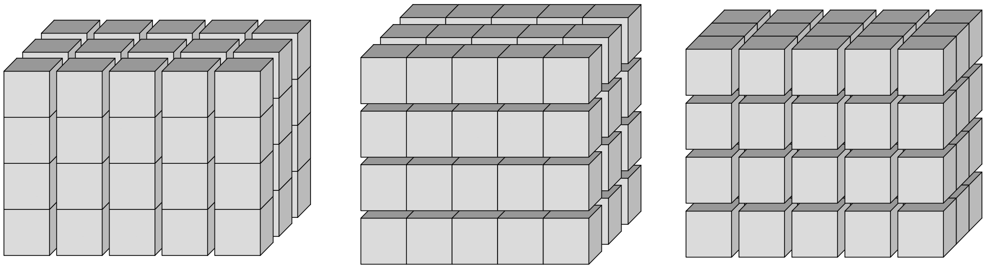


Figure A.2: Mode-$1, 2$ and 3 *fibers* of a $4 \times 5 \times 3$ tensor

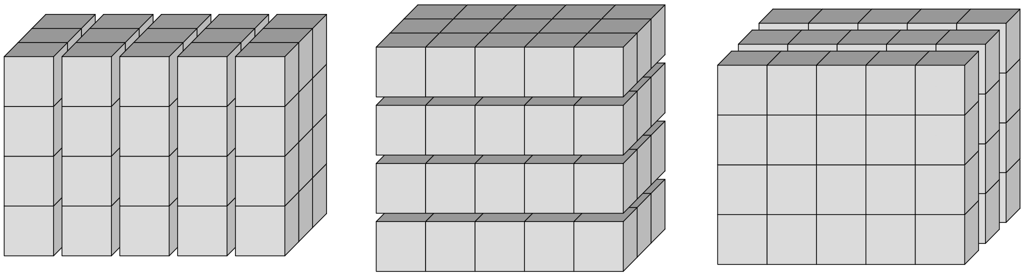


A.2 Unfolding of Tensors: Illustrative Examples

Example 1:

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

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



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

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

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

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

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

Finally, for the generic case let y be an element of tensor $\mathbf{Y}$ and $\bar{y}$ be an element of the unfolded tensor $\mathbf{Y}_{(j)}$, then we can define mode-j unfolding as the following mapping:

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

A.3 Supplementary Tables

Table A.1: Description of Characteristics Acronyms

This table provides simple description for the acronyms listed in table 4. For more detailed description of the characteristics, see Chen and Zimmermann (2021).

Acronym	Description	Authors
Beta	CAPM beta	Fama and MacBeth
BetaFP	Frazzini-Pedersen Beta	Frazzini and Pedersen
BidAskSpread	Bid-ask spread	Amihud and Mendelsohn
ChEQ	Growth in book equity	Lockwood and Prombutr
DolVol	Past trading volume	Brennan, Chordia, Subra
EP	Earnings-to-Price Ratio	Basu
EarningsSurprise	Earnings Surprise	Foster, Olsen and Shevlin
FEPS	Analyst earnings per share	Cen, Wei, and Zhang
FR	Pension Funding Status	Franzoni and Marin
FirmAgeMom	Firm Age - Momentum	Zhang
High52	52 week high	George and Hwang
IdioRisk	Idiosyncratic risk	Ang et al.
IdioVol3F	Idiosyncratic risk (3 factor)	Ang et al.
IdioVolAHT	Idiosyncratic risk (AHT)	Ali, Hwang, and Trombley
InvGrowth	Inventory Growth	Belo and Lin
Leverage	Market leverage	Bhandari
MaxRet	Maximum return over month	Bali, Cakici, and Whitelaw
MeanRankRevGrowth	Revenue Growth Rank	Lakonishok, Shleifer, Vishny
Mom12m	Momentum (12 month)	Jegadeesh and Titman
MomOffSeason16YrPlus	Off season reversal years 16 to 20	Heston and Sadka
MomSeason16YrPlus	Return seasonality years 16 to 20	Heston and Sadka
OrderBacklog	Order backlog	Rajgopal, Shevlin, Venkatachalam
OrderBacklogChg	Change in order backlog	Baik and Ahn
Price	Price	Blume and Husic
PriceDelayRsq	Price delay r square	Hou and Moskowitz
PriceDelaySlope	Price delay coeff	Hou and Moskowitz
RDAbility	R&D ability	Cohen, Diether and Malloy
ShareIss5Y	Share issuance (5 year)	Daniel and Titman
VolSD	Volume Variance	Chordia, Subra, Anshuman
VolumeTrend	Volume Trend	Haugen and Baker
grcapx	Change in capex (two years)	Anderson and Garcia-Feijoo
realestate	Real estate holdings	Tuzel

A.4 Proofs of Main Results

Proof of Proposition 3.1. Note that

$$\bigodot_{k\neq j} M_k = \left(\bigotimes_{k\neq j}{}_{\mathrm{K}} m_{k,1}, \ldots, \bigotimes_{k\neq j}{}_{\mathrm{K}} m_{k,R} \right).$$

Under Assumption 3.1

$$M_j^\top M_j = I_R, \qquad 1 \leq \forall j \leq d.$$

Therefore, the $(l,m)^{\text{th}}$ element of $\left(\bigodot_{k\neq j} M_k\right)^\top \left(\bigodot_{k\neq j} M_k\right)$ is

$$\left(\bigotimes_{k\neq j}{}_{\mathrm{K}} m_{k,l} \right)^\top \bigotimes_{k\neq j}{}_{\mathrm{K}} m_{k,m} = \prod_{k\neq j} m_{k,l}^\top m_{k,m} = 1.$$

□

Proof of Theorem 3.1. Consider the 2-way factor model

$$\mathbf{Y} = M_1 D M_2^\top + \mathbf{U},$$

where $(\mathbf{Y}, M_1, M_2, \mathbf{U}, N_1, N_2)$ are replaced with $\left(\mathbf{Y}_{(j)}, M_j, \bigotimes_{k\neq j} M_k, \mathbf{U}_{(j)}, N_j, \prod_{k\neq j} N_k\right)$. The result holds by Theorem A.5.1 provided that the required assumptions are verified.

Assumption A.5.1 holds since $M_j^\top M_j = I_R$ and $\left(\bigotimes_{k\neq j} M_k\right)^\top \left(\bigotimes_{k\neq j} M_k\right) = I_R$ by Proposition 3.1 under Assumption 3.1. Lastly, Assumption A.5.2 (i) is verified under Assumption 3.2 (i). □

Proof of Corollary 3.1. Under Assumption 3.3, $\sigma_r^2 = (1+o(1)) d_r \prod_{j=1}^d N_j$. Then $\text{trace}(D) = (1+o(1)) \sum_{r=1}^R \sqrt{d_r} \prod_{j=1}^d \sqrt{N_j}$ and

$$\delta_r = (1+o(1)) \left(\min_{k\neq r} |d_k - d_r| \wedge d_r \right) \prod_{j=1}^d N_j.$$

The result follows from Theorem 3.1. □

Proof of Theorem 3.2. Consider the 2-way factor model

$$\mathbf{Y} = M_1 D M_2^\top + \mathbf{U},$$

where $(\mathbf{Y}, M_1, M_2, \mathbf{U}, N_1, N_2)$ are replaced with $\left(\mathbf{Y}_{(j)}, M_j, \bigotimes_{k \neq j} M_k, \mathbf{U}_{(j)}, N_j, \prod_{k \neq j} N_k\right)$ and $\sigma_r = \|v_{1,r}\|\|v_{2,r}\|$ with $\sigma_r = \prod_{j=1}^d \|v_{j,r}\|$ in the diagonal elements of D. The result holds by Theorem A.5.2 provided that the required assumptions are verified.

Note that Assumptions 3.1 and 3.2 (i)-(ii) imply Assumptions A.5.1 and A.5.2 (i)-(ii) by Proposition 3.1. Recall also that $\bigodot_{k \neq j} M_k$ is a $\prod_{k \neq j} N_k \times R$ matrix with columns $\bigotimes_{\mathrm{K} k \neq j} m_{k,1}$. Therefore, Assumption A.5.2 (iii) is verified under Assumption 3.2 (iii).

Assumption 3.3 implies Assumption A.5.3. Lastly, Assumption A.5.4 is verified under Assumption 3.4. □

Proof of Theorem 3.3. Follows from Theorem A.5.2 given the discussions in the proof of Theorem 3.2. □

Proof of Theorem 4.1. Under Assumption 3.2 (i), if $u_{i_1,\dots,i_d} \sim N(0, \sigma^2)$, then $Q^\top \mathbf{U}_{(j)} \in \mathbb{R}^{N_j - R}$ is a vector of i.i.d. $N(0, \sigma^2)$ for every Q such that $Q^\top Q = I$. Then under Assumptions 3.1, 3.2 (i), and 3.3, by Lemma A.5.3

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

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,

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

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

Recall that

$$\mathbf{Y}_{(j)} = M_j D \left(\bigodot_{k \neq j} M_k \right)^\top + \mathbf{U}_{(j)}.$$

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

$$\left| \frac{\hat{\sigma}_{R,j}}{\sqrt{\prod_{k \neq j}^d N_k}} - \frac{\sigma_{R,j}}{\sqrt{\prod_{k \neq j}^d N_k}} \right| \leq \frac{\|\mathbf{U}_{(j)}\|_{\mathrm{op}}}{\sqrt{\prod_{k \neq j}^d N_k}}.$$

The right-hand side of this equation is $O_P(1)$ by Latała (2005) while $\sigma_{R,j}/\sqrt{\prod_{k \neq j}^d N_k} \uparrow \infty$ under Assumption 3.3. Therefore, $\hat{\sigma}^2_{R,j}/\prod_{k \neq j}^d N_k \uparrow \infty$. This implies that $(\hat{\sigma}^2_{R,j} - \hat{\sigma}^2_{R+1,j})/\tau \to \infty$, and whence $S_j \uparrow \infty$ under H_1. □

A.5 Auxiliary Results

Consider the 2-way factor model

$$\begin{aligned} \mathbf{Y} &= V_1 V_2^\top + \mathbf{U} \\ &= M_1 D M_2^\top + \mathbf{U} \\ &= \sum_{r=1}^R \sigma_r m_{1,r} \otimes m_{2,r} + \mathbf{U}. \end{aligned}$$

The following assumption requires that the factors are orthogonal.

Assumption A.5.1. *Suppose that*

$$M_1^\top M_1 = I_R \qquad \textit{and} \qquad M_2^\top M_2 = I_R.$$

Assumption A.5.1 is without loss of generality in light of identifying assumptions used in the factor literature since the scale of factors is absorbed in $(\sigma_r)_{r=1}^R$.

Let $\mathbf{U}_i = (u_{1,i}, \dots, u_{N_1,i})^\top$ be the i^{th} column of $\mathbf{U}$. The following assumption imposes several mild restrictions on the data generating process.

Assumption A.5.2. *(i)* $\mathbf{U} = \{u_{i,j} : \ 1 \leq i \leq N_1, 1 \leq j \leq N_2\}$ *are i.i.d. with* $\mathbb{E}(u_{i,j}) = 0$, $\mathrm{Var}(u_{i,j}) = \sigma^2$, *and* $\mathbb{E}|u_{i,j}|^4 < \infty$*; (ii)* $\mathbb{E}\left|\langle \mathbf{U}_i, m_{1,k}\rangle \langle \mathbf{U}_i, m_{1,r}\rangle\right|^2 = O(1)$ *for every* $1 \leq k, r \leq R$*; (iii)* $\max_k \|m_{2,k}\|_\infty = o(1)$.

The following result holds:

Theorem A.5.1. *Suppose that Assumptions A.5.1 and A.5.2 (i) are satisfied. Then*

$$\|\hat{m}_{1,r} - m_{1,r}\| = O_P\left(\frac{\sqrt{N_1}\mathrm{trace}(D) + N_1 \vee N_2}{\delta_r}\right), \qquad \forall 1 \leq r \leq R.$$

Proof. Under Assumption A.5.1, by the Davis-Kahan theorem, see Vershynin (2018), Theorem 4.5.5,

$$\|\hat{m}_{1,r} - m_{1,r}\| \leq \frac{2^{3/2}}{\delta_r}\left\|\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top\right\|_{\mathrm{op}}$$

The result follows under Assumption A.5.2 (i) by Lemma A.5.1. □

Lemma A.5.1. *Suppose that* $\mathbf{U} = \{u_{i,j} : 1 \leq i \leq N_1, 1 \leq j \leq N_2\}$ *are i.i.d. with* $\mathbb{E}(u_{i,j}) = 0$, $\mathrm{Var}(u_{i,j}) = \sigma^2$, *and* $\mathbb{E}|u_{i,j}|^4 < \infty$. *Then if* $M_1^\top M_1 = I_R$ *and* $M_2^\top M_2 = I_R$, *we have*

$$\|\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top\|_{\mathrm{op}} = O_P\left(\sqrt{N_1}\mathrm{trace}(D) + N_1 \vee N_2\right).$$

Proof. Since $\mathbf{Y} = M_1 D M_2^\top + \mathbf{U}$, we have

$$\begin{aligned}\left\|\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top\right\|_{\mathrm{op}} &= \left\|M_1 D M_2^\top \mathbf{U}^\top + \mathbf{U} M_2 D M_1^\top + \mathbf{U}\mathbf{U}^\top\right\|_{\mathrm{op}} \\ &\leq \|M_1 D M_2^\top \mathbf{U}^\top\|_{\mathrm{op}} + \|\mathbf{U} M_2 D M_1^\top\|_{\mathrm{op}} + \|\mathbf{U}\mathbf{U}^\top\|_{\mathrm{op}} \\ &= 2\|\mathbf{U} M_2 D M_1^\top\|_{\mathrm{op}} + \|\mathbf{U}\|_{\mathrm{op}}^2,\end{aligned}$$

where the second line uses the triangle inequality and the last $\|A\|_{\mathrm{op}} = \|A^\top\|_{\mathrm{op}}$ and $\|AA^\top\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2$.

Next, by the triangle inequality

$$\begin{aligned}\|\mathbf{U} M_2 D M_1^\top\|_{\mathrm{op}} &= \left\|\sum_{r=1}^R \sigma_r \mathbf{U} m_{2,r} \otimes m_{1,r}\right\|_{\mathrm{op}} \\ &\leq \sum_{r=1}^R \sigma_r \|\mathbf{U} m_{2,r} \otimes m_{1,r}\|_{\mathrm{op}} \\ &= \sum_{r=1}^R \sigma_r \sup_{\|x\| \leq 1} \|\mathbf{U} m_{2,r} \langle m_{1,r}, x\rangle\| \\ &= \sum_{r=1}^R \sigma_r \|\mathbf{U} m_{2,r}\|,\end{aligned}$$

where we use the fact that $\sup_{\|x\|\leq 1} |\langle m_{1,r}, x\rangle| = 1$ given that $M_1^\top M_1 = I_R$. Since $\mathbf{U} \in \mathbb{R}^{N_1\times N_2}$ are i.i.d. with mean zero and variance σ^2 and $M_2^\top M_2 = I_R$, we have

$$\mathbb{E}\|\mathbf{U}m_{2,r}\|^2 = \mathbb{E}\left[m_{2,r}^\top \mathbf{U}^\top \mathbf{U} m_{2,r}\right] = N_1\sigma^2.$$

Therefore, $\|\mathbf{U}M_2DM_1^\top\|_{\mathrm{op}} = O_P\left(\sqrt{N_1}\mathrm{trace}(D)\right)$. Lastly, by Latała (2005)

$$\begin{aligned}\mathbb{E}\|\mathbf{U}\|_{\mathrm{op}} &\lesssim \max_i \sqrt{\sum_j \mathbb{E}u_{i,j}^2} + \max_j \sqrt{\sum_i \mathbb{E}u_{i,j}^2} + \sqrt[4]{\sum_{i,j}\mathbb{E}u_{i,j}^4} \\ &= O\left(\sqrt{N_1}+\sqrt{N_2}\right).\end{aligned}$$

Therefore, $\|\mathbf{U}\|_{\mathrm{op}}^2 = O_P(N_1 \vee N_2)$. The result follows from combining all estimates together. □

Next, we make the following pervasive factor assumption.

Assumption A.5.3. *Suppose that there exist constants $d_1 > d_2 > \cdots > d_R > 0$ such that*

$$\lim_{N_1,N_2\to\infty} \frac{\sigma_r^2}{N_1N_2} = d_r, \qquad 1 \leq \forall r \leq R.$$

We will focus on the central limit theorem for linear functionals of $\hat{m}_{1,r}$. The next assumption states that $\langle m_{1,r}, \nu\rangle$ is asymptotically well-defined.

Assumption A.5.4. *Suppose that for every $k \neq r$,*

$$\omega_k(\nu) \equiv \lim_{N_1\to\infty} \sqrt{N_1}\langle m_{1,k}, \nu\rangle$$

exists and is strictly positive.

The following result holds:

Theorem A.5.2. *Suppose that Assumptions A.5.1, A.5.2, A.5.3, and A.5.4 are satisfied. Then*

$$\sqrt{N_2}\langle \hat{m}_{1,r} - m_{1,r}, \nu\rangle \xrightarrow{d} N\left(0, \sigma^2 \sum_{k\neq r} \omega_k^2(\nu) \frac{d_r + d_k}{(d_r - d_k)^2}\right)$$

provided that $N_1/N_2 = o(1)$ and $N_2/N_1^3 = o(1)$ as $N_1, N_2 \to \infty$.

Proof. By Kneip and Utikal (2001), Lemma A1

$$\hat{m}_{1,r} - m_{1,r} = \sum_{k \neq r} \frac{m_{1,k} \otimes m_{1,k}}{\sigma_r^2 - \sigma_k^2} \left(\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top\right) m_{1,r} + R_r,$$

where by Lemma A.5.1 under Assumption A.5.2 (i)

$$\begin{aligned} \|R_r\| &\leq \frac{6}{\delta_r^2} \left\|\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top\right\|_{\mathrm{op}}^2 \\ &= O_P\left(\frac{N_1 \mathrm{trace}^2(D) + (N_1 \vee N_2)^2}{\delta_r^2}\right). \end{aligned}$$

Under Assumption A.5.3, $\sigma_r^2 \sim N_1 N_2$, so that $\delta_r \sim N_1 N_2$ and $\mathrm{trace}(D) = \sum_{r=1}^R \sigma_r \sim \sqrt{N_1 N_2}$. Therefore,

$$\|R_r\| = O_P\left(\frac{1}{N_2} + \frac{1}{N_1^2}\right) = o_P\left(\frac{1}{\sqrt{N_1 N_2}}\right),$$

which follows since $N_1/N_2 = o(1)$ and $N_2/N_1 = o(N_1^2)$. Since $\mathbf{Y} = M_1 D M_2^\top + \mathbf{U}$, we also have $\mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top = M_1 D M_2^\top \mathbf{U}^\top + \mathbf{U} M_2 D M_1^\top + \mathbf{U}\mathbf{U}^\top$. Therefore,

$$\hat{m}_{1,r} - m_{1,r} = \sum_{k \neq r} \frac{\sigma_k m_{1,r}^\top \mathbf{U} m_{2,k} + \sigma_r m_{1,k}^\top \mathbf{U} m_{2,r} + m_{1,k}^\top \mathbf{U}\mathbf{U}^\top m_{1,r}}{\sigma_r^2 - \sigma_k^2} m_{1,k} + o_P\left(\frac{1}{\sqrt{N_1 N_2}}\right).$$

Under Assumptions A.5.1 and A.5.2 (i)

$$\mathbb{E}[\langle \mathbf{U}_i, m_{1,k}\rangle \langle \mathbf{U}_i, m_{1,r}\rangle] = \sigma^2 \langle m_{1,k}, m_{1,r}\rangle = 0, \qquad \forall k \neq r.$$

Therefore, under Assumption A.5.2 (ii)

$$\begin{aligned} \mathrm{Var}\left(m_{1,k}^\top \mathbf{U}\mathbf{U}^\top m_{1,r}\right) &= N_2 \mathrm{Var}\left(\langle \mathbf{U}_i, m_{1,k}\rangle \langle \mathbf{U}_i, m_{1,r}\rangle\right) \\ &= N_2 \mathbb{E}\left|\langle \mathbf{U}_i, m_{1,k}\rangle \langle \mathbf{U}_i, m_{1,r}\rangle\right|^2 \\ &= O(N_2). \end{aligned}$$

By Chebyshev's inequality, this implies that $m_{1,k}^\top \mathbf{U}\mathbf{U}^\top m_{1,r} = O_P(\sqrt{N_2})$ for all $k \neq r$.

Under Assumption A.5.3, we also know that $\sigma_r^2 = (1+o(1))d_r N_1 N_2$. Therefore,

$$\sqrt{N_1 N_2}(\hat{m}_{1,r} - m_{1,r}) = \sum_{k \neq r} \frac{(1+o(1))}{d_r - d_k} \left\{ \sqrt{d_k} m_{1,r}^\top \mathbf{U} m_{2,k} + \sqrt{d_r} m_{1,k}^\top \mathbf{U} m_{2,r} \right\} m_{1,k} + o_P(1).$$

Under Assumptions A.5.1 and A.5.2 (i) and (iii), by Lemma A.5.2,

$$(m_{1,r}^\top \mathbf{U} m_{2,k})_{1 \le r,k \le R} \xrightarrow{d} Z,$$

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

$$\sqrt{N_2} \langle \hat{m}_{1,r} - m_{1,r}, \nu \rangle \xrightarrow{d} \sum_{k \neq r} \omega_k(\nu) \frac{\sqrt{d_k} Z_{r,k} + \sqrt{d_r} Z_{k,r}}{d_r - d_k}.$$

□

We now turn to the asymptotic distribution of scale components. Let $\hat{\sigma}_r^2$ be the r^{th} eigenvalue of $\mathbf{Y}\mathbf{Y}^\top$. The following result holds:

Theorem A.5.3. *Suppose that Assumptions A.5.1, A.5.2, A.5.3, and A.5.4 are satisfied. Then*

$$\left(\frac{\hat{\sigma}_r^2 - \sigma_r^2}{\sigma_r} \right)_{1 \le r \le R} \xrightarrow{d} N(0, 4\sigma^2 I_R),$$

provided that $N_1/N_2 = o(1)$ and $N_2/N_1^3 = o(1)$ as $N_1, N_2 \to \infty$.

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

$$\begin{aligned} \hat{\sigma}_r^2 - \sigma_r^2 &= \operatorname{trace}(m_{1,r} \otimes m_{1,r}(M_1 D M_2^\top \mathbf{U}^\top + \mathbf{U} M_2 D M_1^\top + \mathbf{U}\mathbf{U}^\top)) + R_r \\ &= m_{1,r}^\top \left(M_1 D M_2^\top \mathbf{U}^\top + \mathbf{U} M_2 D M_1^\top + \mathbf{U}\mathbf{U}^\top \right) m_{1,r} + R_r \\ &= 2\sigma_r m_{1,r}^\top \mathbf{U} m_{2,r} + m_{1,r}^\top \mathbf{U}\mathbf{U}^\top m_{1,r} + R_r, \end{aligned}$$

where the last line follows under Assumption A.5.1. Under Assumption A.5.2 (i) by Lemma A.5.1

$$\begin{aligned} |R_r| &\le \frac{6}{\delta_r} \left\| \mathbf{Y}\mathbf{Y}^\top - M_1 D^2 M_1^\top \right\|_{\text{op}}^2 \\ &= O_P\left(\frac{N_1 \text{trace}^2(D) + (N_1 \vee N_2)^2}{\delta_r} \right). \end{aligned}$$

Under Assumption A.5.3, $\sigma_r^2 \sim N_1 N_2$ and $\delta_r \sim N_1 N_2$. Therefore,

$$R_r = O_P\left(N_1 + \frac{N_1}{N_2} + \frac{N_2}{N_1}\right) = o_P(\sigma_r),$$

provided that $N_1/N_2 = o(1)$ and $N_2/N_1^3 = o(1)$ as $N_1, N_2 \to \infty$. Next, under Assumption A.5.2 (i)-(ii),

$$\mathrm{Var}(m_{1,r}^\top \mathbf{U}\mathbf{U}^\top m_{1,r}) = N_2 \mathrm{Var}(\langle m_{1,r}, \mathbf{U}_i\rangle^2) = N_2\left(\mathbb{E}\langle m_{1,r}, \mathbf{U}_i\rangle^4 - \sigma^4\right) = O(N_2).$$

This shows that $m_{1,r}^\top \mathbf{U}\mathbf{U}^\top m_{1,r} = O_P(\sqrt{N_2}) = o_P(\sigma_r)$ as $N_1 \to \infty$.

Combining all estimates, we obtain

$$\frac{\hat{\sigma}_r^2 - \sigma_r^2}{\sigma_r} = 2 m_{1,r}^\top \mathbf{U} m_{2,r} + o_P(1).$$

Finally, under Assumptions A.5.1 and A.5.2 (i) and (iii), by Lemma A.5.2

$$(m_{1,r}^\top \mathbf{U} m_{2,r})_{1\leq r\leq R} \xrightarrow{d} N(0, \sigma^2 I_R).$$

□

Lemma A.5.2. *Suppose that* $\mathbf{U} = \{u_{i,j} : 1 \leq i \leq N_1, 1 \leq j \leq N_2\}$ *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*

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

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

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

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

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

where the last line follows by Jensen's inequality and Assumption A.5.2 (i). Therefore,

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

where we use the fact that $\|y_k\|_{2+\delta} \leq \|y_k\|_2$ and orthonormality. Similarly, we could verify the second Lyapunov's condition when $\max_r \|y_r\| = o(1)$. □

Lastly, we study the asymptotic expansion for smallest eigenvalues of $\mathbf{Y}\mathbf{Y}^\top$ valid when $N_1, N_2 \to \infty$ without restricting the relative growth of the two dimensions. Let $Q = (m_{1,R+1}, \ldots, m_{1,N_1})$ be $N_1 \times (N_1 - R)$ matrix with columns corresponding to eigenvectors associated with zero eigenvalues of $M_1 D^2 M_1^\top$. The columns of Q are an orthonormal basis for the null space of $M_1 D^2 M_1^\top$.

Lemma A.5.3. *Suppose that Assumptions A.5.1, A.5.2 (i), A.5.3 hold. Then*

$$\hat{\sigma}_{R+r}^2 = \lambda_r(Q^\top \mathbf{U}\mathbf{U}^\top Q) + O_P\left(N_1 + \frac{N_2}{N_1}\right), \qquad 1 \leq \forall r \leq N_1 - R,$$

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

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

$$M = \begin{bmatrix} M_1 & Q \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} D^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then $M\Sigma M^\top$ is the spectral decomposition of $M_1 D^2 M_1^\top$ and we have

$$\begin{aligned} \lambda_{R+r}\left(\frac{\mathbf{Y}\mathbf{Y}^\top}{N_1 N_2}\right) &= \lambda_{R+r}\left(\frac{M^\top \mathbf{Y}\mathbf{Y}^\top M}{N_1 N_2}\right) \\ &= \lambda_{R+r}\left(\frac{1}{N_1 N_2}\begin{bmatrix} M_1^\top \\ Q^\top \end{bmatrix}(M_1 D M_2^\top + \mathbf{U})(M_1 D M_2^\top + \mathbf{U})^\top \begin{bmatrix} M_1 & Q \end{bmatrix}\right) \\ &= \lambda_{R+r}\left(A + \varkappa A^{(1)}\right) \end{aligned}$$

with $\varkappa = 1/N_1$, $A = \begin{bmatrix} \frac{D^2}{N_1 N_2} & 0 \\ 0 & 0 \end{bmatrix}$ and

$$A^{(1)} = \frac{1}{N_2} \begin{bmatrix} M_1^\top \mathbf{U} M_2 D + D M_2^\top \mathbf{U}^\top M_1 + M_1^\top \mathbf{U}\mathbf{U}^\top M_1 & D M_2^\top \mathbf{U}^\top Q + M_1^\top \mathbf{U}\mathbf{U}^\top Q \\ Q^\top \mathbf{U} M_2 D + Q^\top \mathbf{U}\mathbf{U}^\top M_1^\top & Q^\top \mathbf{U}\mathbf{U}^\top Q \end{bmatrix}.$$

By Onatski (2009), Lemma 6

$$\left| \lambda_{R+r}\left(A + \varkappa A^{(1)}\right) - \varkappa \lambda_r \left(\frac{Q^\top \mathbf{U}\mathbf{U}^\top Q}{N_2} \right) \right| \leq \frac{\varkappa^2 \|A^{(1)}\|^2}{0.5\sigma_R^2/(N_1 N_2) - \varkappa \|A^{(1)}\|}$$

provided that $\|A^{(1)}\|/N_1 < 0.5\sigma_R^2/(N_1 N_2)$. To see that this condition holds, note that under Assumption A.5.3, $\lim_{N_1, N_2 \to \infty} \sigma_r^2/(N_1 N_2) = d_r > 0, \forall r \leq R$. Moreover,

$$\|A^{(1)}\|/N_1 \leq \frac{4}{N_1 N_2} \left[\|\mathbf{U} M_2 D\|_{\text{op}} + \left\|\mathbf{U}\mathbf{U}^\top\right\|_{\text{op}} \right].$$

It follows from the proof of Lemma A.5.1 that under Assumption A.5.2 (i), $\|\mathbf{U} M_2 D\|_{\text{op}} = O_P\left(\sum_{r=1}^R \sigma_r \sqrt{N_1} \right)$ and $\|\mathbf{U}\mathbf{U}^\top\|_{\text{op}} = O_P(N_1 + N_2)$. This shows that

$$\|A^{(1)}\|/N_1 = O_P\left(\frac{1}{\sqrt{N_2}} + \frac{1}{N_1} \right) = o_P(1),$$

and whence

$$\lambda_{R+r}\left(\frac{\mathbf{Y}\mathbf{Y}^\top}{N_1 N_2} \right) = \lambda_r \left(\frac{Q^\top \mathbf{U}\mathbf{U}^\top Q}{N_1 N_2} \right) + O_P\left(\frac{1}{N_2} + \frac{1}{N_1^2} \right).$$

□