CENTRE FOR ECONOMETRIC ANALYSIS
CEA@BAYES

https://www.bayes.city.ac.uk/faculties-and-research/centres/cea

Bayes Business School (formerly Cass)
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Tensor Principal Component Analysis

Andrii Babii, Eric Ghysels and Junsu Pan

CEA@Bayes Working Paper Series

WP–CEA–01-2023

Tensor Principal Component Analysis

Andrii Babii* Eric Ghysels† Junsu Pan‡

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

January 10, 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 (PCA), 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 tensor PCA estimator of loadings and factors. The algorithm demonstrates good performance in Mote Carlo experiments and is applied to sorted portfolios.

Keywords: Principal component analysis, tensor data, singular value and canonical polyadic decompositions.

*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 US 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 illustrate 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.

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 dependencies in a tensor datasets. In traditional factor models, the dependencies 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

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

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 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 for a d-way tensor dataset also describes a tensor as a sum of a low-rank component and an idiosyncratic shocks tensor. The low-rank component is usually approximated using either the Tucker or the Canonical Polyadic (CP) decompositions; see Tucker (1966) and Hitchcock (1927). The former is more general and covers the CP decomposition as a special case. Similar to how a matrix can be approximated by a sum of the outer products of vectors with the SVD, the CP decomposition approximates a tensor by a sum of the outer products of vectors multiplied by scale components that correspond to singular values in the SVD. The Tucker decomposition replaces the scale components with a "core" tensor and does not lead to the same interpretation as the SVD for traditional factor models. Therefore, in this paper, we focus on the tensor factor models related to the CP decomposition; see Han, Chen, Yang, and Zhang (2020), Wang, Zheng, and Li (2021), Chen, Yang, and Zhang (2022), Han, Chen, and Zhang (2022) for tensor factor models based on the Tucker decomposition, and Richard and Montanari (2014) for a symmetric rank-1 model.

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 new estimation procedure for tensor factor models that is an extension of PCA, which we call tensor PCA, and discuss the associated sampling properties of the estimator. 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 to obtain the components of the tensor decomposition. Therefore, the tensor PCA leads to the closed form expressions for factors and factor loadings. We show that the tensor PCA can identify and consistently estimate the factors

and loadings, describe the associated convergence rates, and the asymptotic distribution for large dimensional tensors with growing dimensions. We find that our tensor PCA algorithm is more accurate than ALS. We also illustrate that the d-way tensor factor model achieves more efficient dimensionality reduction than the naively pooled 2-way factor model.

The paper is organized as follows. Section 2 introduces a class of tensor factor models. The PCA estimators for tensor data are covered in Section 3. This section contains convergence rates and large sample distributions. Small sample simulation evidence is reported in Section 4. An illustrative empirical example appears in Section 5. Section 6 concludes. Lastly, in the Appendix, we collect several illustrative examples of tensors, discuss tensor unfoldings, provide proofs of all main and auxiliary results.

Notation: The Khatri-Rao product of two matrices $A = (a_1, \ldots, a_R)$ and $B = (b_1, \ldots, b_R)$ is defined as $A \odot B = (a_1 \otimes_{\mathrm{K}} b_1, \ldots, a_R \otimes_{\mathrm{K}} b_R)$, where $\otimes_{\mathrm{K}}$ denotes the Kronecker product. In addition, for a collection of matrices $(V_j)_{j=1}^d$, we define $\bigodot_{k \neq j} V_k = V_d \odot \cdots \odot V_{k+1} \odot V_{k-1} \odot \cdots \odot V_1$. 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. For two tensors $A, B \in \mathbb{R}^{N_1 \times \cdots \times N_d}$, the Frobenius inner product is defined as $\langle A, B\rangle_F = \sum_{i_1,\ldots,i_d} A_{i_1,\ldots,i_d} B_{i_1,\ldots,i_d}$. Let $\|A\|_{\mathrm{F}} = \sqrt{\langle A, A\rangle_{\mathrm{F}}}$ be the Frobenius norm of a tensor A induced by the inner product. For a matrix A with columns $(a_1, \ldots, a_n)$, 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[2] and a matrix $\mathbf{U} \in \mathbb{R}^{N\times T}$ of idiosyncratic shocks:

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

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 principal component analysis (PCA).

[2]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.

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):

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

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

Therefore, in this paper, we will focus on the tensor factor model in equation (3), which is a direct generalization of the widely used 2-way factor models in econometrics and statistics; cf. equation (1).

3 Tensor PCA

In this section, we introduce the tensor PCA 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 tensor PCA estimator of loadings/factors.

3.1 Unfolding

Since the scale of loadings/factors is not identified,[3] we 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

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

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 of a d-way tensor.

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, $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_R)$, 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.

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

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.

$$V_j^\top V_j \qquad \text{is a diagonal matrix,} \qquad 1 \leq j \leq d.$$

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.[4]

Proposition 3.1. *Under Assumption 3.1*

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

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 tensor PCA estimation algorithm:

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

2) Estimate M_j as $\widehat{M}_j = (\hat{m}_{j,1}, \ldots, \hat{m}_{j,R})$ 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.

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

3) Recover the scale components as

$$\hat{\sigma}_r = \left\langle \mathbf{Y}, \bigotimes_{j=1}^{d} \hat{m}_{j,r} \right\rangle_{\mathrm{F}}, \qquad \forall r \geq 1.$$

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 $\mathbf{Y}$: $\hat{\sigma}_r = \hat{m}_{1,r}^{\top} \mathbf{Y} \hat{m}_{2,r}$. The estimator in step 3) can also be characterized as the least-squares solution; see Proposition A.4.1 in the Appendix.

Our tensor PCA 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 $d \geq 3$; see Kolda and Bader (2009). Second, the best rank-R approximation to a tensor requires solving a non-convex optimization problem whereas the tensor PCA leads to the closed-form expressions. Third, the asymptotic properties of ALS are not known whereas we obtain the consistency and the large sample distributions for tensor PCA 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 values of R. In contrast, the tensor PCA 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* $\mathbb{E}(u_{i_1,\dots,i_d}) = 0$, $\mathrm{Var}(u_{i_1,\dots,i_d}) = \sigma^2$, *and* $\mathbb{E}|u_{i_1,\dots,i_d}|^4 < \infty$; *(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

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

be the eigengap around the r^{th} eigenvalue of $M_1 D^2 M_1^\top$. Then, the following result holds:[5]

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

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

According to Theorem 3.1, the larger eigengap around the r^{th} eigenvalue makes it easier to estimate the r^{th} loading/factor. Therefore, δ_r can be understood as a measure of loading/factor strength.

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*

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

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 Asymptotic Distribution

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:

[5] See Appendix for proofs of all results.

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.

The following result holds:[6]

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*

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

Our result can be compared to the asymptotic distribution of the 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 $\mathbb{E}[Y_i Y_i^\top] = \Sigma$ 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).

[6] Note that for $d = 3$, conditions of Theorem 3.2 are satisfied when the tensor dimensions grow proportionally, i.e. $N_1 \sim N_2 \sim N_3$.

3.6 Model Selection

The natural extension of the Mallow's Cp would probably be

$$\min_R \left\| \mathbf{Y} - \sum_{r=1}^{R} \hat{\sigma}_r^R \bigotimes_{j=1}^{d} \hat{m}_{j,r}^R \right\|_{\mathrm{F}}^2 + 2R\hat{\sigma}^2 \frac{\sum_{j=1}^d N_j}{\prod_{j=1}^d N_j},$$

where $(\hat{\sigma}_r^R, \hat{m}_{j,r}^R)$ are the tensor PCA estimates for the model with R factors and $\hat{\sigma}^2$ is a consistent estimator of σ^2. This roughly corresponds to the $AIC_3(R)$ objective in Bai and Ng (2002b).

Issues:

1. They claim on p.202 that the penalty function does not satisfy conditions of their Theorem 2. But that theorem does not establish necessary and sufficient conditions, so the claim that it is wrong is false.

2. Generalizing Bai and Ng (2002a), the right objective could be conjectured

$$\min_R \left\| \mathbf{Y} - \sum_{r=1}^{R} \hat{\sigma}_r^R \bigotimes_{j=1}^{d} \hat{m}_{j,r}^R \right\|_{\mathrm{F}}^2 + \hat{\sigma}^2 R \frac{\sum_{j=1}^d N_j}{\prod_{j=1}^d N_j} \log\left(\frac{\sum_{j=1}^d N_j}{\prod_{j=1}^d N_j} \right).$$

   But we also do not know if our estimators are MLE for the joint optimization problem.

3. We could also directly apply Bai and Ng (2002a) to one of the d unfoldings. But this choice is ad-hoc and different unfoldings may estimate a different number of factors. For example, if we use the jth unfolding, we get

$$\min_R \left\| \mathbf{Y}_{(j)} - \hat{M}_j \hat{D} \left(\bigodot_{k \neq j} \hat{M}_k \right)^\top \right\|_{\mathrm{F}}^2 + \hat{\sigma}^2 R \frac{N_j + \sum_{k \neq j} N_k}{\prod_{j=1}^d N_j} \log\left(\frac{N_j + \sum_{k \neq j} N_k}{\prod_{j=1}^d N_j} \right).$$

4 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.

4.1 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:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \lambda_{i,r}\mu_{j,r}f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0, \tag{8}$$

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

The model in equation (8) 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 the PCA to the 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 start with comparing the 2- and 3-way approaches using the notion of *Model Complexity*, defined as the number of parameters expressed as the percent of the data size. 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.

Table 1 reports on the comparison of the model complexity of the 3- and 2-way factor models under different sizes of the three data dimensions. The numbes in the table are simple algebraic calcutions which show the model complexities, with 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 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

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

This table reports on the comparison of the model complexity of the 3- and 2-way factor models under different sizes of the three data dimensions, where model complexity is defined as the number of parameters expressed as the percent of the data size.

	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%

reduction for the latter is 98.8% , while that of the 2-way 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 one can identify the loadings specific to each dimension.

4.2 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 (8), 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:

$$\begin{aligned} y_{i,j,t} &= \sum_{r=1}^{R} \sigma_r \lambda_{i,r} \mu_{j,r} f_{t,r} + u_{i,j,t}, \qquad u_{i,j,t} \sim_{\text{i.i.d.}} N(0, s_u^2), \\ \dot{f}_{t,r} &= \rho \dot{f}_{t-1,r} + \varepsilon_{t,r}, \qquad \varepsilon_{t,r} \sim_{\text{i.i.d.}} N(0, s_\varepsilon^2), \end{aligned} \tag{9}$$

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 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 $s_u = 1$.

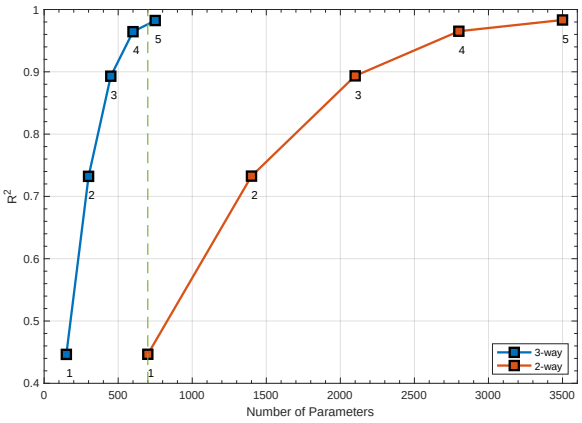
(3) We consider the cases where we generate the true number of factors $R \in \{5, 10\}$, and the dimension sizes of the tensor $(T, N, J) \in \{(100, 30, 20), (50, 50, 50), (50, 100, 100)\}$.

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.

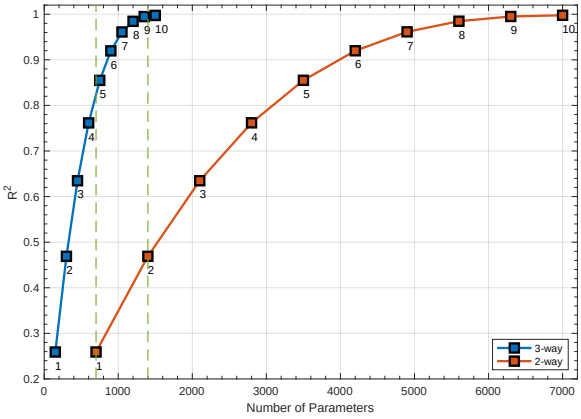
Figure 1 plots the average R^2 against the number of parameters for the 2- and 3-way factor approaches under a 5 and 10 factor model and three different size parameterizations. All panels show that the complexity of the 2-way factor model grows significantly faster than that of the 3-way factor model. 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.

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

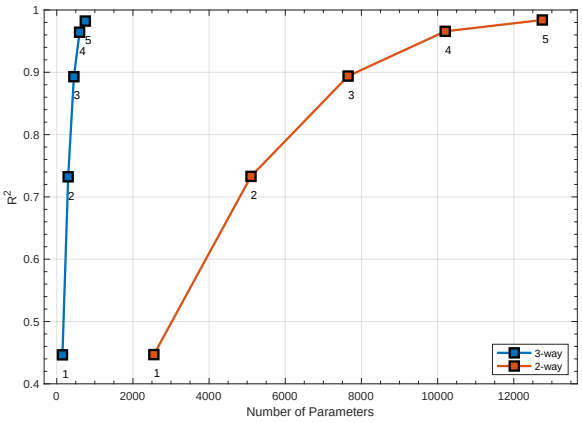
The DGP appears in equation (9). We consider the cases where $R \in \{5, 10\}$ and $(T, N, J) \in \{(100, 30, 20), (50, 50, 50), (50, 100, 100)\}$, and we estimate the number of factors from 1 to R without knowing the true number of factors. We repeat the simulation for 5000 times, and report the average R^2.



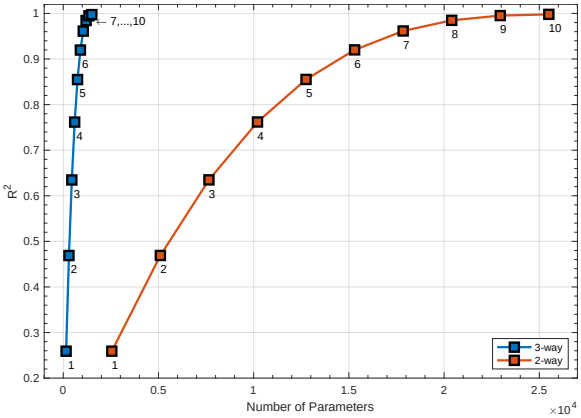
(a) **Sample size $100 \times 30 \times 20$ - 5 Factors**



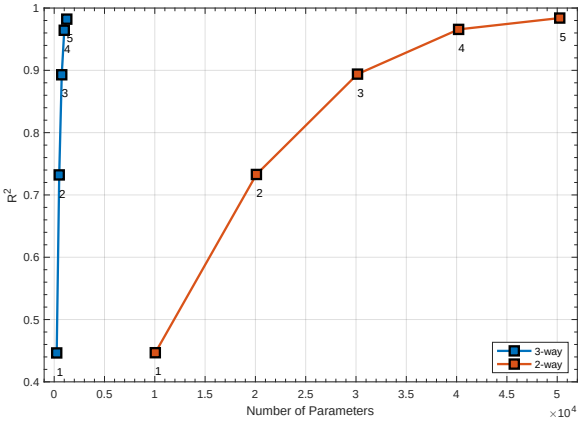
(b) **Sample size $100 \times 30 \times 20$ - 10 Factors**



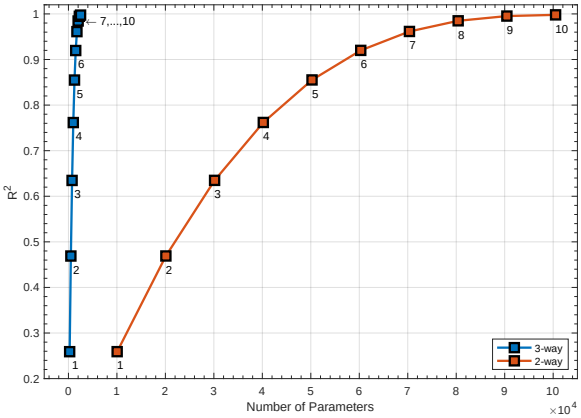
(c) **Sample size $50 \times 50 \times 50$ - 5 Factors**



(d) **Sample size $50 \times 50 \times 50$ - 10 Factors**



(e) **Sample size $50 \times 100 \times 100$ - 5 Factors**



(f) **Sample size $50 \times 100 \times 100$ - 10 Factors**

4.3 Comparison to ALS

In this subsection, we compare our tensor PCA 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 tensor PCA 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.[7]

The data are generated the same as equation (9), 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 estimates by taking the losses as follows:

$$\begin{aligned} \mathbb{L}_\lambda &= \|\hat{\lambda}_r \times \text{sign}(\hat{\lambda}_r^\top \lambda_r) - \lambda_r\|, \\ \mathbb{L}_\mu &= \|\hat{\mu}_r \times \text{sign}(\hat{\mu}_r^\top \mu_r) - \mu_r\|, \\ \mathbb{L}_f &= \|\hat{f}_r \times \text{sign}(\hat{f}_r^\top f_r) - f_r\|, \end{aligned} \tag{10}$$

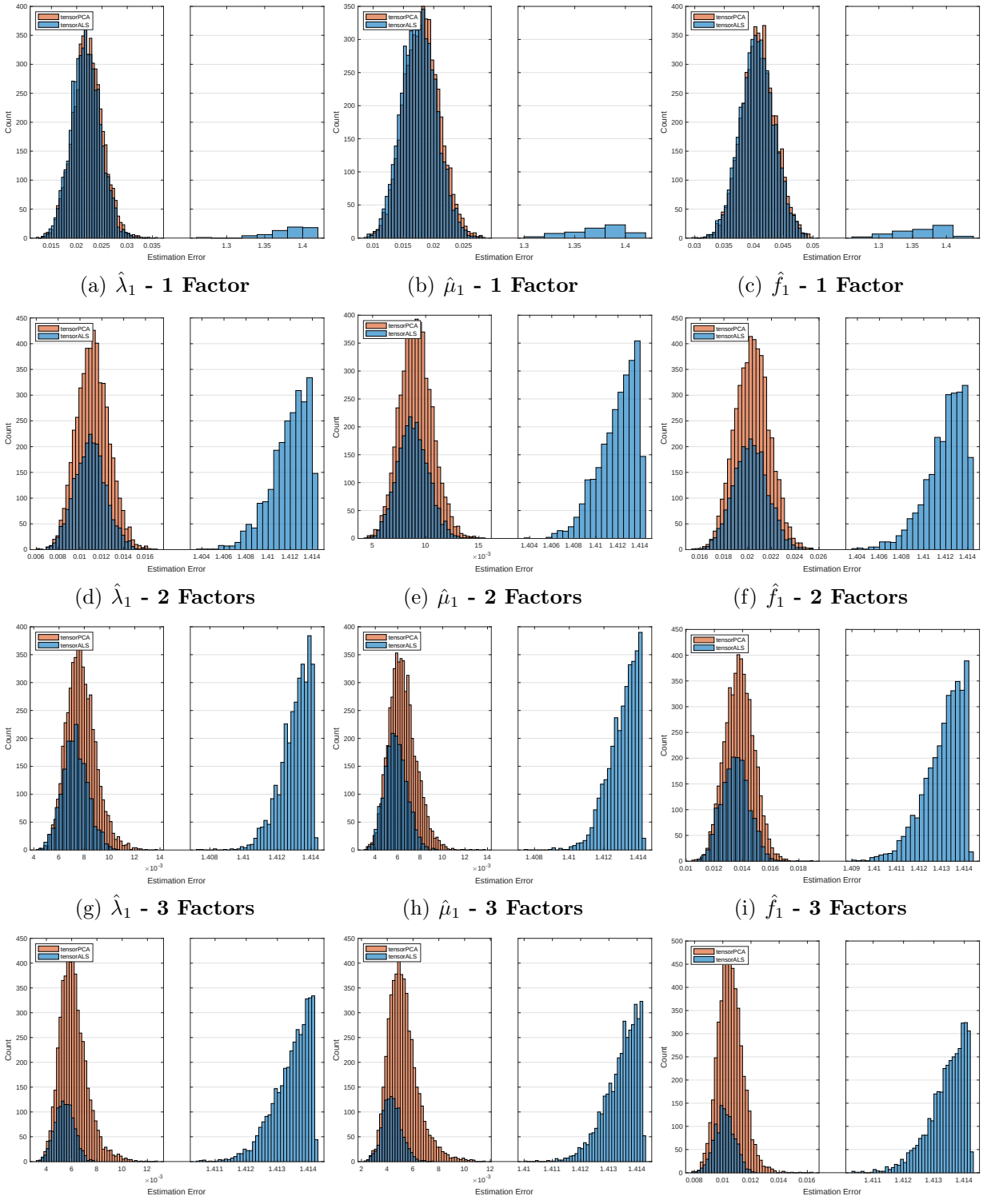
where $\text{sign}(a) = \mathbb{1}_{a>0} - \mathbb{1}_{a<0}$.

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 tensor PCA have comparable accuracy. Nonetheless, the distribution of ALS errors (blue histogram) has a heavy right tail due to outliers. 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 1 factor, we see that the performance of ALS quickly deteriorates 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.

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

Figure 2: Tensor Factor Model Fit via tensor PCA versus ALS

The DGP appears in equation (9). We plot the histograms of the ℓ_2 losses of the estimated first factor/loading of tensor PCA vs ALS in 5000 MC simulations. The horizontal axis is discontinuous because estimates 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**

4.4 Finite Sample Properties

We conclude with an assessment of how changing sample sizes affect 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 is $O_P(1/\sqrt{NT})$, and for the factor vector f_r is $O_P(1/\sqrt{NJ})$.

The data are again generated the same as equation (9), 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 generate and estimate only a 1 factor model for the purpose of demonstrating 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 (10).

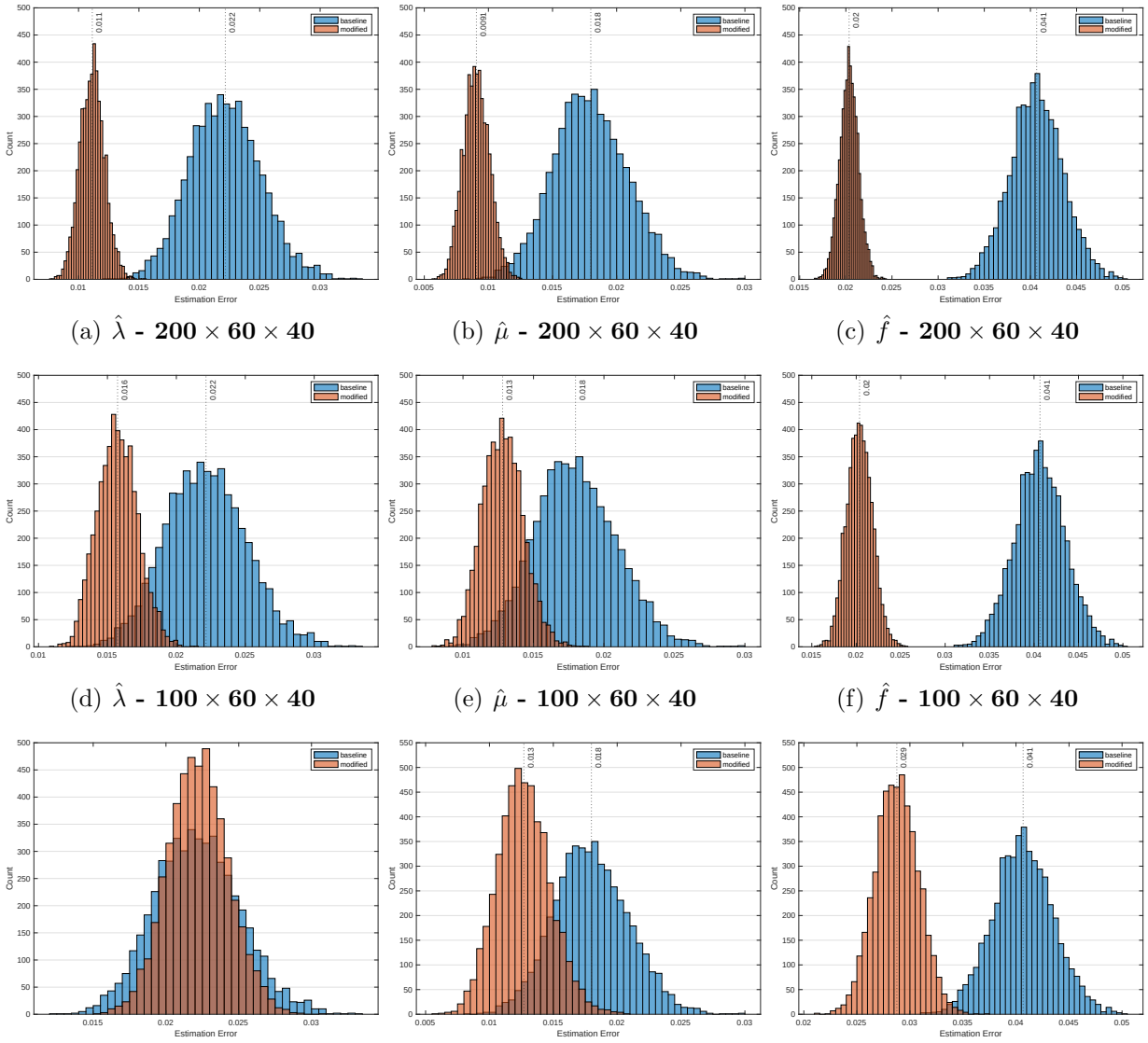
Figure 3 plots the histograms of the ℓ_2 losses of the baseline versus modified DGPs. In panel (a) - (c), as we double the sizes of all dimensions, the estimation of the factor $\hat{f}_1$ and loadings $\hat{\lambda}_1$, $\hat{\mu}_1$ all improve. The average error is reduced roughly by a half for the 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 Empirical Illustration: Sorted Portfolios

One can think of many applications of the type of tensor factor models discussed in the previous sections. In this section we only provide an illustrative financial application which focuses

Figure 3: Tensor Factor Model Fit of Changing Sizes of Dimensions

The DGP appears in equation (9). 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) $\hat{\lambda}$ - $\mathbf{200 \times 60 \times 40}$ (b) $\hat{\mu}$ - $\mathbf{200 \times 60 \times 40}$ (c) $\hat{f}$ - $\mathbf{200 \times 60 \times 40}$

(d) $\hat{\lambda}$ - $\mathbf{100 \times 60 \times 40}$ (e) $\hat{\mu}$ - $\mathbf{100 \times 60 \times 40}$ (f) $\hat{f}$ - $\mathbf{100 \times 60 \times 40}$

(g) $\hat{\lambda}$ - $\mathbf{100 \times 60 \times 20}$ (h) $\hat{\mu}$ - $\mathbf{100 \times 60 \times 20}$ (i) $\hat{f}$ - $\mathbf{100 \times 60 \times 20}$

on the common practice of characteristic-based sorted portfolios to estimate systematic risks in asset pricing models.[8] For instance, 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 4: 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:

$$y_{i,j,t} = \sum_{r=1}^{R} \sigma_r \lambda_{i,r} \mu_{j,r} f_{t,r} + u_{i,j,t}, \qquad \mathbb{E}(u_{i,j,t}) = 0, \tag{11}$$

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 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 deciles of the characteristics together:

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

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.[9] We consider the monthly portfolio returns, which are sorted into 10 deciles based on firm level characteristics, from Jan. 1990 to Dec. 2020. We only consider a balanced set of

[8] A more substantial empirical analysis of tensor factor asset pricing models appears in Babii, Ghysels, and Pan (2022).

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

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 (11) using both tensor PCA and ALS, and compare the estimates from the two different algorithms.[10]

Table 2 reports summary statistics of estimated loadings $\hat{\lambda}$ that determine the exposure of all 133 characteristics to the first two factors. For tensor PCA, the last column shows that the values of $\hat{\lambda}_{i,1}$'s are strictly positive for all characteristics $i = 1, \ldots, 133$, while $\hat{\lambda}_{i,2}$'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. As for ALS, the values of $\hat{\lambda}_{i,1}$'s are no longer strictly positive for all characteristics, with more than 10% of them being negatively exposed with the first factor, and also about 80% are positively exposed to the second factor, which is considerably different from tensor PCA. 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 much larger standard deviation of 0.0566 than 0.0087 for tensor PCA.

Table 3 reports the estimates of the loadings $\hat{\mu}$ that determine the exposure of all 10 deciles to the first two factors. For tensor PCA, 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. As for ALS, the pattern looks very similar to that of tensor PCA, where $\hat{\mu}_1$ is also symmetric around the middle deciles with the largest being the decile on two extremes, and $\hat{\mu}_2$ is monotonically increasing. The difference is that $\hat{\mu}_1$ and the absolute value of $\hat{\mu}_2$ is not as symmetric as that of tensor PCA.

Although there are evident differences in the estimates of tensor PCA and ALS, they share similar interpretations. It is well known that the first systematic risk factor is the market. The values of $\hat{\lambda}_{i,1}$ show that some characteristics tend to have larger exposure to the market than others, and the pattern of $\hat{\mu}_{j,1}$ shows that the tail deciles also tend to have larger exposure to the market than the middle deciles. The left panel of Table 4 lists the top 10 and bottom 10 characteristics in terms of exposure to the market estimated by

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

tensor PCA.[11] 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.

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

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

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 tensor PCA, 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 systematic risk factors beyond the market. However, it is not the case with ALS, as $\hat{\mu}_{1,1}$ is not very close to $\hat{\mu}_{10,1}$. The tensor factor model estimated by tensor PCA confirms that taking the difference between two extreme deciles is actually a proper proxy for the risk premium 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 one characteristic. 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 tensor PCA. 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.

6 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 tensor PCA. Unlike the ALS algorithm, which is commonly used for this purpose, our tensor PCA does not involve solving a non-convex op-

[11] 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 2-factor model appearing in equation (11) estimated with tensor PCA 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%
			ALS		
$\hat{\lambda}_1$	0.2513	0.0659	-0.0536	0.0566	89.47%
$\hat{\lambda}_2$	0.1887	0.0561	-0.1632	0.0664	82.71%

timization problem and has a closed-form expression. Additionally, we provide convergence rates and large sample approximations to distribution for our estimator, demonstrating its advantages over ALS and demonstrating advantages of d-way factor models over the naively pooled traditional factor models for tensors. These findings are supported by the extensive simulation results. Lastly, we also consider an empirical application to sorted portfolios.

Future research directions include addressing the issue of determining the number of factors for d-way factor model, see Han et al. (2022), as well as the application of tensor PCA to multidimensional panel data with interactive fixed effects; see Bai (2009).

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

The table reports $\hat{\mu}$ of a 2-factor model in equation (11) estimated by tensor PCA 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
					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
$\hat{\mu}_2$	0.5259	0.3719	0.2289	0.1225	0.0216	-0.0689	-0.1660	-0.2769	-0.3823	-0.5119
					ALS					
$\hat{\mu}_1$	0.3947	0.3615	0.3316	0.3127	0.2982	0.2838	0.2810	0.2832	0.2879	0.3071
$\hat{\mu}_2$	-0.3469	-0.3682	-0.2638	-0.1727	-0.0822	0.0562	0.1427	0.2505	0.4186	0.6136

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 tensor PCA. 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

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

A. Babii, E. Ghysels, and J. Pan. Tensor factor asset pricing models. Discussion Paper, UNC, 2022.

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

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

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

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

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.

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

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.

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. 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. Liii. 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.

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

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

D. Wang, Y. Zheng, and G. Li. High-dimensional low-rank tensor autoregressive time series modeling. *arXiv preprint arXiv:2101.04276*, 2021.

APPENDIX

A.1 Graphical Illustration of Tensors

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

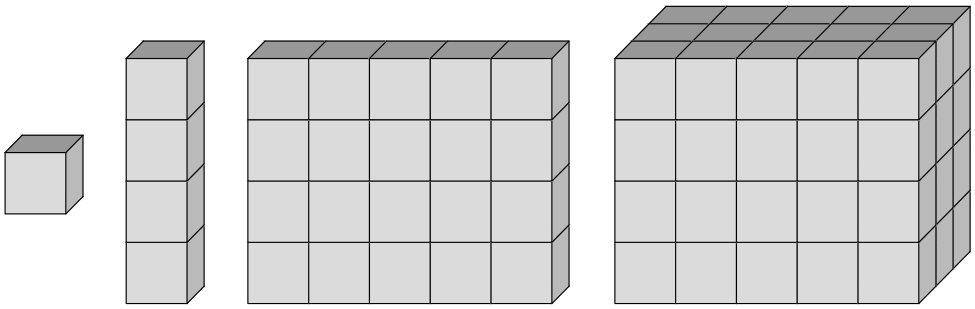
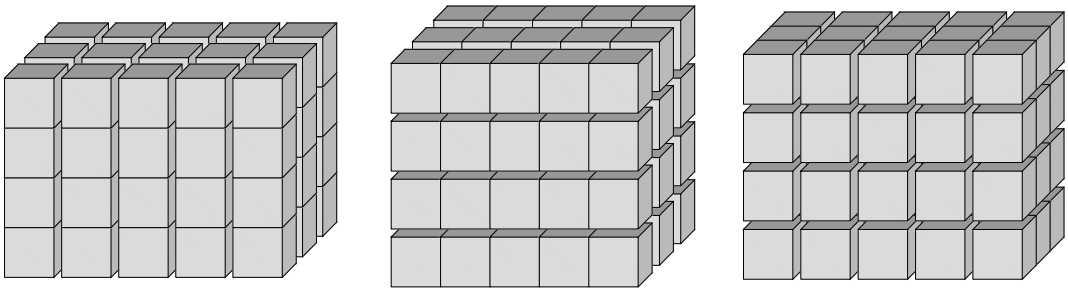


Figure A.2: Mode-1, 2 and 3 *fibers* of a $4 \times 5 \times 3$ tensor

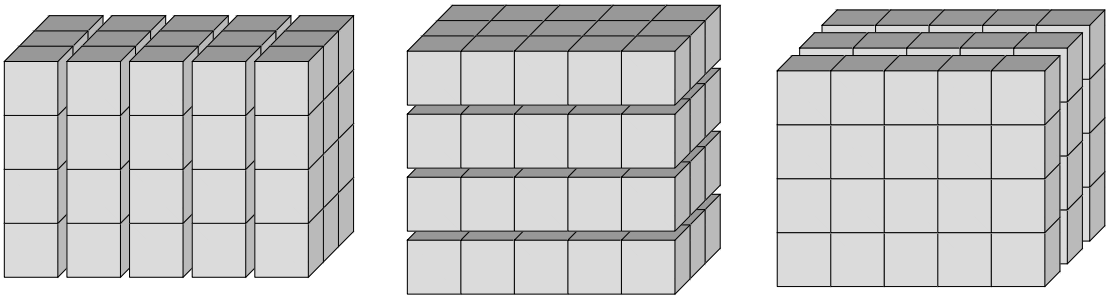


A.2 Unfolding of Tensors: Illustrative Examples

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

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

□

Proposition A.4.1. *Let $\mathbf{Y} \in \mathbb{R}^{N_1\times\cdots\times N_d}$ and let $(m_{j,r})_{j=1}^d$ be a set of orthonormal vectors for each $r \geq 1$. Then the solution to*

$$\hat{\sigma} = \underset{(\sigma_r)_{r=1}^R}{\arg\min} \left\| \mathbf{Y} - \sum_{r=1}^R \sigma_r \bigotimes_{j=1}^d m_{j,r} \right\|_{\mathrm{F}}^2$$

is

$$\hat{\sigma}_r = \left\langle \mathbf{Y}, \bigotimes_{j=1}^d m_{j,r} \right\rangle_{\mathrm{F}}, \qquad r \geq 1.$$

Proof. Put $A_r = \bigotimes_{j=1}^d m_{j,r}$. Using $\|.\|_{\mathrm{F}}^2 = \langle ., . \rangle_{\mathrm{F}}$ and the bilinearity of the inner product

$$\begin{aligned} Q(\sigma_1, \ldots, \sigma_R) &\equiv \left\| \mathbf{Y} - \sum_{r=1}^R \sigma_r A_r \right\|_{\mathrm{F}}^2 \\ &= \left\langle \mathbf{Y} - \sum_{r=1}^R \sigma_r A_r, \mathbf{Y} - \sum_{r=1}^R \sigma_r A_r \right\rangle_{\mathrm{F}} \\ &= \langle \mathbf{Y}, \mathbf{Y} \rangle_{\mathrm{F}} - 2\sum_{r=1}^R \sigma_r \langle \mathbf{Y}, A_r \rangle_{\mathrm{F}} + \sum_{r=1}^R \sum_{s=1}^R \sigma_r \sigma_s \langle A_r, A_s \rangle_{\mathrm{F}}. \end{aligned}$$

Next,

$$\langle A_r, A_s \rangle_{\mathrm{F}} = \left\langle \bigotimes_{j=1}^d m_{j,r}, \bigotimes_{j=1}^d m_{j,r} \right\rangle_{\mathrm{F}} = \prod_{j=1}^d \langle m_{j,r}, m_{j,s} \rangle = \mathbb{1}_{r=s}.$$

Therefore,

$$Q(\sigma_1, \ldots, \sigma_R) = \langle \mathbf{Y}, \mathbf{Y} \rangle_{\mathrm{F}} - 2 \sum_{r=1}^{R} \sigma_r \langle \mathbf{Y}, A_r \rangle_{\mathrm{F}} + \sum_{r=1}^{R} \sigma_r^2.$$

The first-order conditions are

$$\left. \frac{\partial Q(\sigma_1, \ldots, \sigma_R)}{\partial \sigma_r} \right|_{\sigma_r = \hat{\sigma}_r} = -2 \langle \mathbf{Y}, A_r \rangle_{\mathrm{F}} + 2 \hat{\sigma}_r = 0, \qquad \forall r \geq 1.$$

Therefore, $\hat{\sigma}_r = \langle \mathbf{Y}, A_r \rangle_{\mathrm{F}}$ with $r \geq 1$. □

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 $\mathrm{tr}(D) = (1 + o(1)) \sum_{r=1}^{R} \sqrt{d_r} \prod_{j=1}^{d} \sqrt{N_j}$ 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. □

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*

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

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 $\mathbf{U}_i = (u_{1,i}, \ldots, u_{N_1,i})^\top$ 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)* $\mathbf{U} = \{u_{i,j} : 1 \le i \le N_1, 1 \le j \le N_2\}$ *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* $k \neq r$; *(iii)* $\|m_{1,k}\|_\infty = O(1)$ *and* $\|m_{2,k}\|_\infty = o(1)$ *for every* k.

The following result holds:

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

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

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

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

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

Lemma A.5.1. *Suppose that* $\mathbf{U} = \{u_{i,j} : 1 \le i \le N_1, 1 \le j \le N_2\}$ *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*

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

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

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

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

Next, by the triangle inequality

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

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, $\|\mathbf{U} M_2 D M_1^\top\|_{\text{op}} = O_P\left(\sqrt{N_1}\text{tr}(D)\right)$. Lastly, by Latała (2005)

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

Therefore, $\|\mathbf{U}\|_{\text{op}}^2 = O_P(N_1 \vee N_2)$. 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 $d_1 > d_2 > \cdots > d_R$ such that*

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

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)

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

Under Assumption A.5.3, $\sigma_r^2 \sim N_1 N_2$, so that $\delta_r \sim N_1 N_2$ and $\text{tr}(D) = \sum_{r=1}^R \sigma_r \sim \sqrt{N_1 N_2}$. 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 Assumption A.5.2 (i) and (iii), by Lemma A.5.2,

$$m_{1,k}^\top \mathbf{U} m_{2,l} \xrightarrow{d} N(0, \sigma^2), \qquad \forall k \neq l.$$

Moreover, under Assumptions A.5.1 and A.5.2 (i)

$$\begin{aligned}\mathrm{Cov}\left(m_{1,k}^\top \mathbf{U} m_{2,l}, m_{1,r}^\top \mathbf{U} m_{2,s}\right) &= \sigma^2 \langle m_{1,k}, m_{1,l}\rangle\langle m_{2,r}, m_{2,s}\rangle \\ &= 0, \qquad \forall k \neq l \text{ or } r \neq s.\end{aligned}$$

Therefore, $(m_{1,k}^\top \mathbf{U} m_{2,l})_{1\leq k\neq l\leq R}$ are asymptotically independent and

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

where $(\xi_{k,l})_{1\leq k\neq l\leq R}$ are i.i.d. $N(0,\sigma^2)$. □

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* $x \in \mathbb{S}^{N_1-1}$ *and* $y \in \mathbb{S}^{N_2-1}$ *be such that* $\|x\|_\infty = o(1)$ *and* $\|y\|_\infty = O(1)$. *Then*

$$x^\top \mathbf{U} y \xrightarrow{d} N(0, \sigma^2).$$

Proof. Note that

$$x^\top \mathbf{U} y = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} x_i y_j u_{i,j} \triangleq \sum_{i=1}^{N_1} \xi_{N_2,i}$$

is a triangular array of centered independent random variables with variance

$$s_{N_1}^2 \triangleq \sum_{i=1}^{N_1} \mathrm{Var}(\xi_{N_2,i}) = \sigma^2 \|x\|^2 \|y\|^2 = \sigma^2.$$

To show that it is asymptotically Gaussian, we will verify that Lyapunov's condition,

$$\lim_{N_1 \to \infty} \sum_{i=1}^{N_1} \frac{1}{s_{N_1}^{2+\delta}} \mathbb{E} \left| \xi_{N_2,i} \right|^{2+\delta} = 0,$$

holds for some $\delta > 0$; see Billingsley (1995), Theorem 27.3. To that end, it is enough to show that $\max_{1 \le i \le N_1} \mathbb{E}|\xi_{N_2,i}|^{2+\delta} = O(1)$ for some $\delta > 0$. By Rosenthal's inequality, see Petrov (1995), Theorem 2.9, there exists a constant $c(\delta)$ such that

$$\begin{aligned} \mathbb{E} \left| \xi_{N_2,i} \right|^{2+\delta} &= \mathbb{E} \left| \sum_{j=1}^{N_2} x_i y_j u_{i,j} \right|^{2+\delta} \\ &\le x_i^{2+\delta} c(\delta) \left\{ \sum_{j=1}^{N_2} \mathbb{E}|y_j u_{i,j}|^{2+\delta} + \left(\sum_{j=1}^{N_2} \mathbb{E}|y_j u_{i,j}|^2 \right)^{1+\delta/2} \right\}. \end{aligned}$$

Therefore,

$$\sum_{i=1}^{N_1} \mathbb{E} \left| \xi_{N_2,i} \right|^{2+\delta} \lesssim \|x\|^2 \|x\|_\infty^\delta \left\{ \|y\|^2 \|y\|_\infty^\delta + \|y\|^{2+\delta} \right\} = o(1).$$

□