Independent component analysis for tensor-valued data☆

Joni Virta[a,*], Bing Li[b], Klaus Nordhausen[a,c], Hannu Oja[a]

[a]*Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland*
[b]*Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, Pennsylvania 16802, USA*
[c]*CSTAT - Computational Statistics, Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstraße 7, A-1040 Vienna, Austria*

Abstract

In preprocessing tensor-valued data, e.g., images and videos, a common procedure is to vectorize the observations and subject the resulting vectors to one of the many methods used for independent component analysis (ICA). However, the tensor structure of the original data is lost in the vectorization and, as a more suitable alternative, we propose the matrix- and tensor fourth order blind identification (MFOBI and TFOBI). In these tensorial extensions of the classic fourth order blind identification (FOBI) we assume a Kronecker structure for the mixing and perform FOBI simultaneously on each direction of the observed tensors. We discuss the theory and assumptions behind MFOBI and TFOBI and provide two different algorithms and related estimates of the unmixing matrices along with their asymptotic properties. Finally, simulations are used to compare the method's performance with that of classical FOBI for vectorized data and we end with a real data clustering example.

Keywords: FOBI, Kronecker structure, Matrix-valued data, Multilinear algebra
2000 MSC: 62H12, 62G20, 62H10

1. Introduction

1.1. Review of matrix-valued data with the Kronecker structure

In this paper we develop the theory and algorithms for *independent component analysis* (ICA) for tensor-valued data. As the main ideas are best illustrated in the special case where the observations are matrix-valued, we begin by considering the following location-scatter model incorporating Kronecker structure for matrix-valued random elements:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}_L \mathbf{Z} \boldsymbol{\Omega}_R^\top, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{p\times q}$ is the observed matrix, $\boldsymbol{\mu} \in \mathbb{R}^{p\times q}$ is a location center, $\boldsymbol{\Omega}_L \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Omega}_R \in \mathbb{R}^{q\times q}$ are *mixing matrices* that specify linear row and column dependencies, respectively, and $\mathbf{Z} \in \mathbb{R}^{p\times q}$ is a matrix of standardized uncorrelated random variables, $\mathrm{E}\{\mathrm{vec}(\mathbf{Z})\} = \mathbf{0}_{pq}$ and $\mathrm{cov}\{\mathrm{vec}(\mathbf{Z})\} = \mathbf{I}_{pq}$.

It follows that $\mathrm{E}\{\mathrm{vec}(\mathbf{X})\} = \mathrm{vec}(\boldsymbol{\mu})$ and the covariance matrix of the vectorized observation has the Kronecker covariance structure,

$$\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\} = \boldsymbol{\Sigma}_R \otimes \boldsymbol{\Sigma}_L,$$

with $\boldsymbol{\Sigma}_R = \boldsymbol{\Omega}_R \boldsymbol{\Omega}_R^\top$ and $\boldsymbol{\Sigma}_L = \boldsymbol{\Omega}_L \boldsymbol{\Omega}_L^\top$. Note that the structured $\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\}$ has $p(p+1)/2 + q(q+1)/2 - 1$ parameters while the number of parameters in the general unstructured case is as large as $pq(pq+1)/2$.

*Corresponding author
URL: joni.virta@utu.fi (Joni Virta), bing@stat.psu.edu (Bing Li), klaus.nordhausen@tuwien.ac.at (Klaus Nordhausen), hannu.oja@utu.fi (Hannu Oja)

Many examples of matrix-valued data with Kronecker structure exist. For example, in the case of clustered multivariate data the iid observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ represent the n clusters with q individuals in each cluster and p variables measured on each individual, whereas in repeated measures analysis one considers n individuals $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with p measured variables and q repetitions on each individual. If the columns of $\mathbf{X}$ are exchangeable random vectors, as is the case with clustered data, then $\mathbf{\Sigma}_R$ has the intraclass correlation structure, $\mathbf{\Sigma}_R \propto (1-\rho)\mathbf{I}_q + \rho \mathbf{1}_q \mathbf{1}_q^\top$. In applications of matrix or tensor-valued data such as channel modeling for multiple-input multiple-output (MIMO) communication, analysis of spatio-temporal EEG (electroencephalography) data, fMRI (functional Magnetic Resonance Imaging) data, or general image or video clip data, for example, the problem itself often suggests Kronecker structure [51].

Consider next applying distributional assumptions for $\mathbf{Z}$ in the model (1). The (parametric) *multivariate Normal model* or the wider (semiparametric) *elliptical model* are obtained if one assumes that $\mathrm{vec}(\mathbf{Z}) \sim \mathcal{N}_{pq}(\mathbf{0}_{pq}, \mathbf{I}_{pq})$ or that the distribution of $\mathrm{vec}(\mathbf{Z})$ is spherically symmetric, respectively. In these models $\mathbf{\Omega}_L$ and $\mathbf{\Omega}_R$ are well-defined only up to postmultiplication by orthogonal matrices and the number of free mixing parameters is therefore $p(p+1)/2 + q(q+1)/2 - 1$. See, e.g., [8] for an overview of matrix-valued distributions. In this paper we assume that the pq components of $\mathrm{vec}(\mathbf{Z})$ are mutually independent. This semiparametric model, called the *independent component model*, provides an alternative extension of the multivariate Normal model. In this case $\mathbf{\Omega}_L$ and $\mathbf{\Omega}_R$ are well-defined up to permutations and signs of their columns making the number of free mixing parameters $p^2 + q^2 - 1$. In independent component analysis for matrix-valued data, the objective is then to use the realizations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the model (1) to estimate *unmixing matrices* $\mathbf{\Gamma}_L \in \mathbb{R}^{p \times p}$ and $\mathbf{\Gamma}_R \in \mathbb{R}^{q \times q}$ such that $\mathbf{\Gamma}_L \mathbf{X} \mathbf{\Gamma}_R^\top$ has mutually independent components.

In the multivariate Normal case, Srivastava et al. [41] introduced a likelihood ratio test for the null hypothesis of Kronecker covariance structure and used the so-called flip-flop algorithm to find maximum likelihood estimates of $\mathbf{\Sigma}_L$ and $\mathbf{\Sigma}_R$ under the null hypothesis. For another approach to this estimation problem, see [38, 52]. Srivastava et al. [41] also tested the hypothesis that $\mathbf{\Sigma}_R$ is an identity matrix, a diagonal matrix or of intraclass correlation structure, see their paper for further references. Sun et al. [42] considered robust estimation of a structured covariance matrix, including Kronecker covariance structure, under heavy-tailed elliptical distributions and [7] modeled the covariance matrix of spatio-temporal data as a sum of low-rank Kronecker products and a sparse matrix.

1.2. Review of methods for general tensor-valued data

Like matrices, tensor-valued observations have become a prevalent form of modern data and some fields of application include psychometrics, chemometrics and computer vision; see [17, 22] along with the references therein for more examples. For modeling tensor data, e.g., the tensor Normal distribution has been proposed; see [23, 31]. Also a general location-scatter model and an independent component model for tensor-valued data are easily defined; see Section 5. In both cases, for a tensor-valued random element $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, the covariance matrix of the vectorized observation again exhibits a Kronecker structure, viz. $\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\} = \mathbf{\Sigma}_r \otimes \cdots \otimes \mathbf{\Sigma}_1$.

Tensor-based methods have a long history in, e.g., signal processing in the form of different tensor decompositions. The two most prevalent ones are CP-decomposition and the Tucker decomposition which provide tensor analogies for singular value decomposition and principal component analysis, respectively. Both are thoroughly discussed in [17], where a review of numerous other tensor decompositions is also given. See also Beckmann and Smith [1], who introduce tensor PICA, an independent component analysis method for fMRI data based on the CP-decomposition and Kim et al. [15], who present various robust and sparse tensor decompositions for coping with outliers and sparsity.

In recent years, methods for tensor-valued observations have also been increasingly discussed in the statistical literature. For example, Li et al. [19] expanded the sliced inverse regression methodology developed in [20] to create dimension folding, a supervised dimension reduction method for matrix and tensor-valued predictors. Sufficient dimension reduction for longitudinal predictors was considered in [34], logistic regression and generalized linear models for tensor-valued predictors were developed in [9, 59], and regularized linear regression and generalized linear models for tensor-valued predictors were addressed in [56, 58]. Ding and Cook [4] discussed matrix versions of principal component analysis and principal fitted components (PFC). Xue and Yin [53] introduced central mean dimension folding subspace and proposed several methods to estimate it. Ding and Cook [6] further developed tensor-valued sliced inverse regression. An alternative perspective for sufficient dimension reduction for tensors was considered in [5, 57]; see also [10, 40, 54].

High dimensionality is common to modern, naturally tensor-valued data sets and in many cases the number of variables further exceeds the number of observations, preventing the use of vector-valued methods. In such cases

tensorial methods of dimension reduction, such as those listed above, provide an especially attractive course of action, allowing the reduction of the data while taking into account its special tensor structure; see [47, 50]. In this paper we tackle this problem from the viewpoint of independent component analysis.

1.3. Independent component analysis for tensor-valued data

Extending independent component analysis to tensors has also attracted some attention but, to our knowledge, no model-based treatise has been given. The ICA problem for tensor data is discussed in [44, 55], where it is proposed to unmix each of the modes separately by m-flattening the data tensor and subjecting the matrix of m-mode vectors to standard ICA methods. However, this approach discards all the information on the structural dependence present in the tensors. Our proposed method, TFOBI, a tensor analogy for a popular independent component analysis method called fourth order blind identification (FOBI) [2], also considers each mode separately, but instead takes advantage of this structural information in estimating the independent components.

In the classic independent component analysis for vector-valued data, it is assumed that the observations $\mathbf{x} \in \mathbb{R}^p$ obey the model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}, \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location center, $\boldsymbol{\Omega} \in \mathbb{R}^{p\times p}$ is the so-called mixing matrix and $\mathbf{z} \in \mathbb{R}^p$ is a vector of standardized, mutually independent components. The goal is, given the iid observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, to find an estimate of an unmixing matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{\Gamma}\mathbf{x}$ has mutually independent components. Numerous methods for solving the vector-valued independent component problem can be found in the literature, the most popular ones including FOBI, JADE (joint approximate diagonalization of eigen-matrices) and FastICA; see, e.g., [11, 27].

FOBI is based on the fact that in the independent component model (2), both

$$\mathrm{E}(\mathbf{z}\mathbf{z}^\top) = \mathbf{I}_p \quad \text{and} \quad \mathrm{E}(\mathbf{z}\mathbf{z}^\top\mathbf{z}\mathbf{z}^\top) = \mathrm{E}(\|\mathbf{z}\|^2\mathbf{z}\mathbf{z}^\top)$$

are diagonal matrices. In a similar way our extension of FOBI for matrix-valued observations, called *matrix fourth order blind identification* (MFOBI), makes use of the fact that the matrices $\mathrm{E}(\mathbf{Z}\mathbf{Z}^\top) = q\mathbf{I}_p$, $\mathrm{E}(\mathbf{Z}^\top\mathbf{Z}) = p\mathbf{I}_q$, $\mathrm{E}(\mathbf{Z}\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top)$, $\mathrm{E}(\mathbf{Z}^\top\mathbf{Z}\mathbf{Z}^\top\mathbf{Z})$, $\mathrm{E}(\|\mathbf{Z}\|_F^2\mathbf{Z}\mathbf{Z}^\top)$ and $\mathrm{E}(\|\mathbf{Z}\|_F^2\mathbf{Z}^\top\mathbf{Z})$ are all diagonal. Here $\|\cdot\|_F$ is the Frobenius norm. Similar constructs for tensor-valued data are discussed in Section 5.

This paper is structured as follows. We start with some notation and important concepts in Section 2. In Section 3 we review the classic independent component model for vector-valued observations and then extend the model for matrix-valued data. The identifiability constraints and assumptions regarding both models are also discussed. Next, in Section 4, we first review the basic steps — standardization and rotation — of finding the classical FOBI solution and then by analogy find the MFOBI solution by double standardization and double rotation. Furthermore, we provide two different ways for estimating the double rotation and then show that the MFOBI estimate is Fisher consistent. In Section 5 we further extend the method to tensor-valued data and obtain the general TFOBI method. In Section 6 we provide the asymptotic behavior for the extended FOBI procedures in the case of identity mixing. Orthogonal equivariance of TFOBI implies that the asymptotic variances derived for both versions allow comparisons with FOBI also for any orthogonal mixing matrices. In Section 7 we use simulations to compare TFOBI with vectorizing and using FOBI in both the general case of estimating the correct unmixing matrix and blind classification. Also a real data example is included. Finally, in Section 8 we close with some conclusions and prospective ideas.

Our route of exposition from MFOBI to TFOBI is not the most parsimonious one, as MFOBI is logically a special case of TFOBI. We choose this path not only because the core ideas are best explained in the matrix setting; they would be hard to discern among the complicated tensor manipulations, but also because the asymptotic behavior of TFOBI reverts to that of MFOBI for tensors of all orders.

2. Notation

2.1. Some moments and cross-moments

Next, we define some particular moments and expressions based on the moments of the elements of the iid random vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ from the distribution of $\mathbf{z} \in \mathbb{R}^p$ and iid random matrices $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ from the distribution of $\mathbf{Z} \in \mathbb{R}^{p\times q}$.

The components of $\mathbf{z}$ and $\mathbf{Z}$ are mutually independent and standardized to have zero mean and unit variance. Beginning with the marginal moments of the vectors we write, for each $k \in \{1, \ldots, p\}$,

$$\gamma_k = \mathrm{E}(z_k^3), \quad \beta_k = \mathrm{E}(z_k^4), \quad \text{and} \quad \omega_k = \mathrm{var}(z_k^3).$$

For the matrix version we require the same moments and thus define, for all $k \in \{1, \ldots, p\}$ and $\ell \in \{1, \ldots, q\}$,

$$\gamma_{k\ell} = \mathrm{E}(z_{k\ell}^3), \quad \beta_{k\ell} = \mathrm{E}(z_{k\ell}^4), \quad \omega_{k\ell} = \mathrm{var}(z_{k\ell}^3).$$

Interestingly, MFOBI involves the row and column means of the previously defined moments and we use the notation $\bar{\alpha}_{k\bullet}$ to denote taking the average over the values of the bulleted index, e.g., $\bar{\omega}_{k\bullet} = \sum_\ell \omega_{k\ell}/q$. Additionally, we are going to need the covariance of two rows of kurtoses and define $\delta_{kk'} = \sum_\ell \beta_{k\ell}\beta_{k'\ell}/q - \bar{\beta}_{k\bullet}\bar{\beta}_{k'\bullet}$.

For the asymptotic behavior of FOBI we require the following cross-moment estimates for distinct $k, k', m \in \{1, \ldots, p\}$:

$$\hat{s}_{kk'} = \frac{1}{n}\sum_{i=1}^n z_{i,k}z_{i,k'}, \quad \hat{q}_{kk'} = \frac{1}{n}\sum_{i=1}^n (z_{i,k}^3 - \gamma_k)z_{i,k'} \quad \hat{q}_{mkk'} = \frac{1}{n}\sum_{i=1}^n z_{i,m}^2 z_{i,k}z_{i,k'}.$$

For their matrix counterparts, we need both the "left" and "right" versions, e.g.,

$$\bar{s}_{kk'}^L = \frac{1}{q}\sum_{\ell=1}^q \left(\frac{1}{n}\sum_{i=1}^n z_{i,k\ell}z_{i,k'\ell}\right), \quad \bar{s}_{\ell\ell'}^R = \frac{1}{p}\sum_{k=1}^p \left(\frac{1}{n}\sum_{i=1}^n z_{i,k\ell}z_{i,k\ell'}\right),$$

where a bar ($\bar{a}$ instead of $\hat{a}$) is used to emphasize the taking of the mean and to avoid confusion with $\hat{s}_{kk'}$. Notice also how $\bar{s}_{kk'}^L$ and $\bar{s}_{\ell\ell'}^R$ are again the row and column averages of the corresponding vector quantities. We also see that the right-hand side version of the quantity is obtained from the left-hand side version by simply reversing the roles of rows and columns (or transposing the matrices $\mathbf{Z}_i$). Due to this connection, we next state only the left-hand side versions of the remaining needed quantities, also omitting the superscript "L":

$$\bar{q}_{kk'} = \frac{1}{q}\sum_{\ell=1}^q \left\{\frac{1}{n}\sum_{i=1}^n (z_{i,k\ell}^3 - \gamma_{k\ell})z_{i,k'\ell}\right\}, \quad \bar{q}_{mkk'} = \frac{1}{q}\sum_{\ell=1}^q \left\{\frac{1}{n}\sum_{i=1}^n z_{i,m\ell}^2 z_{i,k\ell}z_{i,k'\ell}\right\},$$

and the following which lack a vector counterpart:

$$\bar{r}_{kk'} = \frac{1}{q}\sum_{\ell=1}^q \sum_{\ell'=1,\,\ell'\neq\ell}^q \left(\frac{1}{n}\sum_{i=1}^n z_{i,k\ell}^2 z_{i,k\ell'}z_{i,k'\ell'}\right), \quad \bar{r}_{mkk'}^0 = \frac{1}{q}\sum_{\ell=1}^q \sum_{\ell'=1,\,\ell'\neq\ell}^q \left(\frac{1}{n}\sum_{i=1}^n z_{i,k\ell}z_{i,m\ell}z_{i,m\ell'}z_{i,k'\ell'}\right) \quad \text{and}$$

$$\bar{r}_{mkk'}^1 = \frac{1}{q}\sum_{\ell=1}^q \sum_{\ell'=1,\,\ell'\neq\ell}^q \left(\frac{1}{n}\sum_{i=1}^n z_{i,m\ell}^2 z_{i,k\ell'}z_{i,k'\ell'}\right).$$

Assuming that the eighth moments of $\mathbf{Z}$ exist, the joint limiting distribution of the above quantities can be shown to be multivariate Normal. Additional properties of the quantities are discussed in the proof of Theorem 5 in Section 6. Furthermore, similar quantities could also be defined for random tensors, but they are not needed in the exposition as it is later shown that the asymptotical behavior of TFOBI reduces to that of MFOBI.

2.2. Notations for matrices and sets of matrices

An inverse square root $\mathbf{S}^{-1/2}$ of a symmetric, positive definite matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ is any matrix $\mathbf{G} \in \mathbb{R}^{p\times p}$ satisfying $\mathbf{GSG}^\top = \mathbf{I}_p$. Given the eigendecomposition of the matrix $\mathbf{S} = \mathbf{UDU}^\top$, all possible inverse square root matrices of $\mathbf{S}$ are of the form $\mathbf{VD}^{-1/2}\mathbf{U}^\top$, where $\mathbf{V} \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. If $\mathbf{S}$ has distinct eigenvalues, then a unique, symmetric choice for $\mathbf{S}^{-1/2}$ is $\mathbf{UD}^{-1/2}\mathbf{U}^\top$, see, e.g., [14].

For each $k \in \{1, \ldots, p\}$, the p-vector $\mathbf{e}_k$ is a vector with kth element 1 and other elements 0, and $\mathbf{E}^{k\ell} = \mathbf{e}_k\mathbf{e}_\ell^\top$ is a $p \times p$ matrix with (k, ℓ)-element 1 and other elements 0, for $k, \ell \in \{1, \ldots, p\}$. Note that $\mathbf{I}_p = \sum_{k=1}^p \mathbf{E}^{kk}$ and all diagonal matrices with diagonal elements $c_1, \ldots, c_p$ can be written as $\sum_{k=1}^p c_k\mathbf{E}^{kk}$.

Table 1 lists some particular sets of (affine transformation) matrices used in the following sections. A permutation matrix is obtained if we permute the rows and/or columns of an identity matrix. A heterogeneous sign-change matrix is a diagonal matrix with diagonal entries ± 1. A heterogeneous scaling matrix is a diagonal matrix with positive diagonal entries.

Table 1: Some useful sets of square matrices

Set	Description
$\mathcal{A}^r$	The set of all $r \times r$ non-singular matrices.
$\mathcal{U}^r$	The set of all $r \times r$ orthogonal matrices.
$\mathcal{P}^r$	The set of all $r \times r$ permutation matrices.
$\mathcal{J}^r$	The set of all $r \times r$ heterogeneous sign-change matrices.
$\mathcal{D}^r$	The set of all $r \times r$ heterogeneous scaling matrices.
C^r	The set of all matrices $\mathbf{PJD}$, where $\mathbf{P} \in \mathcal{P}^r$, $\mathbf{J} \in \mathcal{J}^r$ and $\mathbf{D} \in \mathcal{D}^r$.

3. Independent component models

In this section we derive the basic model behind MFOBI by expanding the classic *independent component model* from vector-valued to matrix-valued observations.

3.1. Vector-valued independent component model

Definition 1. *The vector-valued independent component model assumes that the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ are iid realizations of a random vector $\boldsymbol{x}$ satisfying $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\boldsymbol{z}$, where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Omega} \in \mathcal{A}^p$ and the random vector $\boldsymbol{z} \in \mathbb{R}^p$ satisfies Assumptions V1 and V2 below.*

Assumption V1. *The components z_k of $\boldsymbol{z}$ are mutually independent and standardized in the sense that* $\mathrm{E}(z_k) = 0$ *and* $\mathrm{var}(z_k) = 1$.

Assumption V2. *At most one of the components z_k of $\boldsymbol{z}$ is normally distributed.*

Without Assumption V1 the model itself in Definition 1 is not well-defined in the sense that replacing $\boldsymbol{\Omega}$ and $\mathbf{z}$ with $\boldsymbol{\Omega}^* = \boldsymbol{\Omega}\mathbf{C}$ and $\mathbf{z}^* = \mathbf{C}^{-1}\mathbf{z}$, for some $\mathbf{C} \in C^p$, yields exactly the same model for $\mathbf{x}$. The standardization part of Assumption V1 can thus be regarded as an identification constraint that removes some of the ambiguity present in the formulation of the model by fixing the locations and scales of the components of $\mathbf{z}$. As for Assumption V2, it is required by the rotational invariance of the multivariate Gaussian distribution. Namely, assume, e.g., that the first two components of $\mathbf{z}$ are Gaussian. Then the corresponding subvector is distributionally invariant under rotations and the first two columns of $\boldsymbol{\Omega}$ could be identified only up to a 2×2 rotation. Thus only a single normally distributed component is allowed. Given Assumptions V1 and V2, we are then left with ambiguity regarding the signs and the order of the independent components which is satisfactory in most applications.

3.2. Matrix-valued independent component model

The matrix-valued independent component model is now obtained simply by adding right-hand side mixing to the vector-valued independent component model.

Definition 2. *The matrix-valued independent component model assumes that the variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{p \times q}$ are iid realizations of a random matrix $\boldsymbol{X}$ satisfying $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Omega}_L \boldsymbol{Z} \boldsymbol{\Omega}_R^\top$, where $\boldsymbol{\mu} \in \mathbb{R}^{p \times q}$, $\boldsymbol{\Omega}_L \in \mathcal{A}^p$, $\boldsymbol{\Omega}_R \in \mathcal{A}^q$ and the random matrix $\boldsymbol{Z}_i \in \mathbb{R}^{p \times q}$ satisfies Assumptions M1 and M2 below.*

Assumption M1. *The components $z_{k\ell}$ of $\mathbf{Z}$ are mutually independent and standardized in the sense that* $\mathrm{E}(z_{k\ell}) = 0$ *and* $\mathrm{var}(z_{k\ell}) = 1$.

Assumption M2. *At most one row of $\mathbf{Z}$ consists entirely of Gaussian components and at most one column of $\mathbf{Z}$ consists entirely of Gaussian components.*

The assumptions now guarantee that $\boldsymbol{\Omega}_L$ and $\boldsymbol{\Omega}_R$ are well-defined up to postmultiplication by any matrices $\mathbf{PJ}$, $\mathbf{P} \in \mathcal{P}^p, \mathbf{J} \in \mathcal{J}^p$ or $\mathbf{P} \in \mathcal{P}^q, \mathbf{J} \in \mathcal{J}^q$, respectively. Thus the first assumption serves again to remove the ambiguity concerning the location of $\mathbf{Z}$ and the scales of the columns of $\boldsymbol{\Omega}_L$ and $\boldsymbol{\Omega}_R$, leaving us with the acceptable uncertainty of the signs and order. Again without assumption M2, if, e.g., the first two rows of $\mathbf{Z}$ were Gaussian then the first two columns of $\boldsymbol{\Omega}_L$ could be identified only up to a 2×2 rotation. Note that we could still estimate those columns

of $\mathbf{\Omega}_L$ that correspond to non-Gaussian rows of $\mathbf{Z}$ but the successful use of such a method in practice would require some way of estimating or testing for the number of non-Gaussian rows in $\mathbf{Z}$. Such a problem is considered for vector-valued ICA in [29, 30] and extending the method to matrix and tensor observations constitutes an interesting future challenge. Given these assumptions, there is still ambiguity in the proportional sizes of the mixing matrices as the transformations $\mathbf{\Omega}_L \to c\mathbf{\Omega}_L$ and $\mathbf{\Omega}_R \to c^{-1}\mathbf{\Omega}_R$, $c \neq 0$, do not change the distribution of $\mathbf{X}$. The number of free mixing parameters is therefore $p^2 + q^2 - 1$.

4. From FOBI to MFOBI

Taking the same approach as with the independent component models in the previous section, we first review the steps of the classic FOBI procedure, i.e., standardization and rotation, for vector-valued data and then suggest a similar procedure for matrix-valued data, called MFOBI, using similar but separate steps from both sides of the matrices.

4.1. Fourth order blind identification (FOBI)

Without loss of generality, we assume in the following that the random vector $\mathbf{x} \in \mathbb{R}^p$ has zero mean, i.e., $\boldsymbol{\mu} = \mathbf{0}_p$ in the model of Definition 1. Note that the following exposition is not the standard way to approach FOBI. However, presenting it this way makes the formulation of MFOBI more intuitive.

We piece together the FOBI-solution by considering the singular value decomposition of the mixing matrix $\mathbf{\Omega} = \mathbf{UDV}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathcal{U}^p$ and $\mathbf{D} \in \mathcal{D}^p$ (the diagonal elements of $\mathbf{D}$ can be chosen to be positive as the matrix $\mathbf{\Omega}$ was assumed to have full rank). The model then has the form $\mathbf{x} = \mathbf{UDV}^\top\mathbf{z}$. In this form it is easy to break down the steps in which we gradually "lose" the independence of the components of $\mathbf{z}$ and move towards the observed $\mathbf{x}$.

Step 0. The vector of independent components $\mathbf{z}$ has independent components and unit component variances: $\mathrm{cov}(\mathbf{z}) = \mathbf{I}_p$.

Step 1. The vector of standardized components $\mathbf{x}^{st} = \mathbf{V}^\top\mathbf{z}$ has uncorrelated components and unit component variances: $\mathrm{cov}(\mathbf{x}^{st}) = \mathbf{I}_p$.

Step 2. The vector of uncorrelated components $\mathbf{x}^{un} = \mathbf{D}\mathbf{x}^{st}$ has uncorrelated components: $\mathrm{cov}(\mathbf{x}^{un}) = \mathbf{D}^2$.

Step 3. The observed vector $\mathbf{x} = \mathbf{U}\mathbf{x}^{un}$ has (generally) correlated components: $\mathrm{cov}(\mathbf{x}) = \mathbf{UD}^2\mathbf{U}^\top$.

That is, in Step 1 we lose independence, in the second step the unit variances, and finally in the third step the uncorrelatedness. For the solution we then hope to carry out these steps in the reversed order.

4.1.1. Standardization

The first step in FOBI consists of standardizing $\mathbf{x}$ with an inverse square root of its covariance matrix $\mathrm{cov}(\mathbf{x}) = \mathbf{S}$. As $\mathbf{S} = \mathbf{UD}^2\mathbf{U}^\top$ one can choose any matrix of the form $\mathbf{S}^{-1/2} = \mathbf{MD}^{-1}\mathbf{U}^\top$, where $\mathbf{M} \in \mathcal{U}^p$. This yields the transformation

$$\mathbf{x} \mapsto \mathbf{S}^{-1/2}\mathbf{x} = \mathbf{M}\mathbf{x}^{st} = \mathbf{W}\mathbf{z}, \tag{3}$$

where $\mathbf{W} = \mathbf{MV}^\top \in \mathcal{U}^p$. Thus the standardization part moves us directly from $\mathbf{x}$ to a standardized random vector and leaves us a rotation away from the independent components.

4.1.2. Rotation

To estimate the orthogonal matrix $\mathbf{W}^\top$ that rotates the standardized observation in (3) to the vector of independent components, we use the so-called FOBI-matrix functional, $\mathbf{B}(\mathbf{x}) = \mathrm{E}(\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top)$. Plugging the standardized vector in, we have

$$\mathbf{B} = \mathbf{B}(\mathbf{W}\mathbf{z}) = \mathbf{W}\mathbf{B}(\mathbf{z})\mathbf{W}^\top$$

where

$$\mathbf{B}(\mathbf{z}) = \mathrm{E}(\mathbf{z}\mathbf{z}^\top\mathbf{z}\mathbf{z}^\top) = \sum_{k=1}^{p}(\beta_k + p - 1)\mathbf{E}^{kk}$$

is a diagonal matrix. Therefore, the orthogonal matrix $\mathbf{W}$ can be found from the eigendecomposition of the matrix $\mathbf{B}$. However, for the eigenbasis of $\mathbf{B}$ to be identifiable, we must make the following assumption that can be seen as a stronger version of Assumption V2.

Assumption V3. *The kurtosis values $\beta_1, \ldots, \beta_p$ of the components of $\mathbf{z}$ are distinct.*

The recovering of the independent components by FOBI is then captured by the following formula:

$$\mathbf{x} \mapsto \mathbf{W}^\top \mathbf{S}^{-1/2} \mathbf{x}.$$

This process consisting of standardization and rotation will next be translated for matrix-valued observations in an intuitively appealing manner.

4.2. Matrix fourth order blind identification (MFOBI)

Without loss of generality, assume that the random matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$ has zero mean, i.e., $\boldsymbol{\mu} = \mathbf{0}_{p \times q}$ in the model of Definition 2. Resorting again to the singular value decompositions of the full-rank mixing matrices $\boldsymbol{\Omega}_L$ and $\boldsymbol{\Omega}_R$, the model in Definition 2 takes the form

$$\mathbf{X} = \boldsymbol{\Omega}_L \mathbf{Z} \boldsymbol{\Omega}_R^\top = \mathbf{U}_L \mathbf{D}_L \mathbf{V}_L^\top \mathbf{Z} \mathbf{V}_R \mathbf{D}_R \mathbf{U}_R^\top.$$

Again the diagonal elements of $\mathbf{D}_L$ and $\mathbf{D}_R$ can be chosen to be positive.

We then apply a similar analysis for the double mixing process of $\mathbf{Z}$ as was done with FOBI previously.

Step 0. The random matrix $\mathbf{Z}$ has independent components and unit component variances: $\mathrm{cov}\{\mathrm{vec}(\mathbf{Z})\} = \mathbf{I}_{pq}$.

Step 1. The matrix of standardized components $\mathbf{X}^{st} = \mathbf{V}_L^\top \mathbf{Z} \mathbf{V}_R$ has uncorrelated components and unit component variances: $\mathrm{cov}\{\mathrm{vec}(\mathbf{X}^{st})\} = \mathbf{I}_{pq}$.

Step 2. The matrix of uncorrelated components $\mathbf{X}^{un} = \mathbf{D}_L \mathbf{X}^{st} \mathbf{D}_R$ has uncorrelated components: $\mathrm{cov}\{\mathrm{vec}(\mathbf{X}^{un})\} = (\mathbf{D}_R^2 \otimes \mathbf{D}_L^2)$.

Step 3. The observed matrix $\mathbf{X} = \mathbf{U}_L \mathbf{X}^{un} \mathbf{U}_R^\top$ has (generally) correlated components: $\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\} = (\mathbf{U}_R \mathbf{D}_R^2 \mathbf{U}_R^\top) \otimes (\mathbf{U}_L \mathbf{D}_L^2 \mathbf{U}_L^\top)$.

We see that the observed matrix $\mathbf{X}$ is built from the matrix of independent components $\mathbf{Z}$ in three steps exactly corresponding to the likewise process on random vectors outlined in the section before. Again our objective is to reverse this process.

4.2.1. Double standardization

We begin by finding a matrix counterpart for the standardization that provides the first step in FOBI. The presence of a double-sided mixing makes it clear that the standardization has to be performed on $\mathbf{X}$ from both left and right. Define the left and right covariance matrices of a zero-mean random matrix $\mathbf{X} \in \mathbb{R}^{p \times q}$ as

$$\mathrm{cov}_L(\mathbf{X}) = \frac{1}{q} \mathrm{E}(\mathbf{X}\mathbf{X}^\top) \quad \text{and} \quad \mathrm{cov}_R(\mathbf{X}) = \frac{1}{p} \mathrm{E}(\mathbf{X}^\top \mathbf{X}),$$

The use of cov_L and cov_R for matrix observations has been considered already in [41]. Consider then the left covariance matrix of $\mathbf{X}$ in the matrix independent component model of Definition 2,

$$\mathbf{S}_L = \mathrm{cov}_L(\mathbf{X}) = \frac{1}{q} \mathrm{E}(\mathbf{X}\mathbf{X}^\top) = \frac{1}{q} \mathbf{U}_L \mathrm{E}\{\mathbf{X}^{un} (\mathbf{X}^{un})^\top\} \mathbf{U}_L^\top, \tag{4}$$

where straightforward calculations show that $\mathrm{E}\{\mathbf{X}^{un} (\mathbf{X}^{un})^\top\} = \mathrm{tr}(\mathbf{D}_R^2) \mathbf{D}_L^2$. Thus (4) provides the eigendecomposition of $\mathbf{S}_L$ and all of its inverse square roots are precisely of the form $\sqrt{q}\, \|\mathbf{D}_R\|_F^{-1} \mathbf{M}_L \mathbf{D}_L^{-1} \mathbf{U}_L^\top$, where $\mathbf{M}_L \in \mathcal{U}^p$ and $\|\cdot\|_F$ denotes the Frobenius norm. The exact same procedure for the right covariance matrix of $\mathbf{X}$ yields

$$\mathbf{S}_R = \mathrm{cov}_R(\mathbf{X}) = \frac{1}{p} \mathrm{E}(\mathbf{X}^\top \mathbf{X}) = \frac{1}{p} \mathbf{U}_R \mathrm{E}\{(\mathbf{X}^{un})^\top \mathbf{X}^{un}\} \mathbf{U}_R^\top,$$

where $\mathrm{E}\{(\mathbf{X}^{un})^\top \mathbf{X}^{un}\} = \mathrm{tr}(\mathbf{D}_L^2)\mathbf{D}_R^2$ and the inverse square roots of $\mathbf{S}_R$ are precisely of the form $\sqrt{p}\,\|\mathbf{D}_L\|_F^{-1}\mathbf{M}_R\mathbf{D}_R^{-1}\mathbf{U}_R^\top$, where $\mathbf{M}_R \in \mathcal{U}^q$.

Using then the inverse square roots of $\mathbf{S}_L$ and $\mathbf{S}_R$ to doubly standardize the data, we obtain the transformation

$$\mathbf{X} \mapsto \mathbf{S}_L^{-1/2}\mathbf{X}(\mathbf{S}_R^{-1/2})^\top = \sqrt{pq}\,\|\mathbf{D}_L\|_F^{-1}\|\mathbf{D}_R\|_F^{-1}\mathbf{M}_L\mathbf{V}_L^\top\mathbf{Z}\mathbf{V}_R\mathbf{M}_R^\top.$$

Denoting $\mathbf{W}_L = \mathbf{M}_L\mathbf{V}_L^\top \in \mathcal{U}^p$ and $\mathbf{W}_R = \mathbf{M}_R\mathbf{V}_R^\top \in \mathcal{U}^q$, we have the following theorem.

Theorem 1. *Denote by $\mathbf{S}_L^{-1/2}$ and $\mathbf{S}_R^{-1/2}$ any inverse square roots of the matrices $\mathrm{cov}_L(\boldsymbol{X})$ and $\mathrm{cov}_R(\boldsymbol{X})$, respectively. Then, under the matrix independent component model of Definition 2, $\mathbf{S}_L^{-1/2}\boldsymbol{X}(\mathbf{S}_R^{-1/2})^\top \propto \boldsymbol{W}_L\boldsymbol{Z}\boldsymbol{W}_R^\top$, where $\boldsymbol{W}_L \in \mathcal{U}^p$ and $\boldsymbol{W}_R \in \mathcal{U}^q$.*

Theorem 1 thus says that the double standardization by $\mathbf{S}_L^{-1/2}$ and $\mathbf{S}_R^{-1/2}$ is a natural counterpart of the standardization of a random vector $\mathbf{z}$ by $\mathbf{S}^{-1/2}$, again leaving us only a (double) rotation away from independent components.

4.2.2. Double rotation

We next approach the rotation part with the same mindset. First, notice that we have two logical matrix counterparts for the FOBI functional $\mathbf{B}(\mathbf{x})$, namely

$$\mathbf{B}^0(\mathbf{X}) = \mathrm{E}(\mathbf{X}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top) \quad \text{and} \quad \mathbf{B}^1(\mathbf{X}) = \mathrm{E}(\|\mathbf{X}\|_F^2\mathbf{X}\mathbf{X}^\top),$$

both reducing to the ordinary FOBI-matrix functional $\mathbf{B}(\mathbf{x})$ if $\mathbf{X}$ has only one column. For finding the rotations we then use either the pair

$$\mathbf{B}_L^0 = \frac{1}{q}\,\mathbf{B}^0\{\mathbf{S}_L^{-1/2}\mathbf{X}(\mathbf{S}_R^{-1/2})^\top\} \quad \text{and} \quad \mathbf{B}_R^0 = \frac{1}{p}\,\mathbf{B}^0\{\mathbf{S}_R^{-1/2}\mathbf{X}^\top(\mathbf{S}_L^{-1/2})^\top\},$$

or the pair

$$\mathbf{B}_L^1 = \frac{1}{q}\,\mathbf{B}^1\{\mathbf{S}_L^{-1/2}\mathbf{X}(\mathbf{S}_R^{-1/2})^\top\} \quad \text{and} \quad \mathbf{B}_R^1 = \frac{1}{p}\,\mathbf{B}^1\{\mathbf{S}_R^{-1/2}\mathbf{X}^\top(\mathbf{S}_L^{-1/2})^\top\}.$$

Write next $\tau = \sqrt{pq}\,\|\mathbf{D}_L\|_F^{-1}\|\mathbf{D}_R\|_F^{-1}$, $a_0 = (p-1)+(q-1)$ and $a_1 = pq-1$. Plugging in the standardized matrix $\mathbf{S}_L^{-1/2}\mathbf{X}(\mathbf{S}_R^{-1/2})^\top = \tau\mathbf{W}_L\mathbf{Z}\mathbf{W}_R^\top$ we obtain, for $N \in \{0,1\}$,

$$\mathbf{B}_L^N = \tau^4\mathbf{W}_L\left(a_N\mathbf{I}_p + \sum_{k=1}^{p}\bar{\beta}_{k\bullet}\mathbf{E}^{kk}\right)\mathbf{W}_L^\top \quad \text{and} \quad \mathbf{B}_R^N = \tau^4\mathbf{W}_R\left(a_N\mathbf{I}_q + \sum_{\ell=1}^{q}\bar{\beta}_{\bullet\ell}\mathbf{E}^{\ell\ell}\right)\mathbf{W}_R^\top, \tag{5}$$

which are precisely the eigendecompositions of the matrices $\mathbf{B}_L^N$ and $\mathbf{B}_R^N$, giving us a way of finding the missing double rotation by $\mathbf{W}_L^\top$ and $\mathbf{W}_R^\top$. To identify the needed eigenbases, the matrix counterpart for Assumption V3 is then as follows.

Assumption M3. *Both the row averages $\bar{\beta}_{1\bullet}, \ldots, \bar{\beta}_{p\bullet}$ and the column averages $\bar{\beta}_{\bullet 1}, \ldots, \bar{\beta}_{\bullet q}$ of the kurtosis values of $z_{k\ell}$ are distinct.*

Interestingly, the number of constraints on the distinctness of the kurtoses of the components does not grow linearly with the number of components but is rather proportional to its square root (assuming the number of rows and the number of columns grow linearly). In a sense MFOBI thus allows for more freedom for the individual marginal distributions. Note that for $q = 1$ Assumption M3 reduces to Assumption V3.

4.2.3. The method in total

The similarity between FOBI and MFOBI is now particularly easy to see if we first write the formula for FOBI as

$$\mathbf{z} = \mathbf{W}^\top\mathbf{S}^{-1/2}\mathbf{x},$$

where $\mathbf{W}$ has the eigenvectors of $\mathbf{B}$ as its columns and the equality sign means equality up to sign-change and permutation. Compare then the above to the same expression for MFOBI:

$$\mathbf{Z} \propto (\mathbf{W}_L^\top \mathbf{S}_L^{-1/2})\mathbf{X}(\mathbf{W}_R^\top \mathbf{S}_R^{-1/2})^\top,$$

where $\mathbf{W}_L$ and $\mathbf{W}_R$ have respectively the eigenvectors of $\mathbf{B}_L$ and $\mathbf{B}_R$ as their columns and the proportionality is up to permutation and sign-change from both left and right. Seen this way, MFOBI can simply be regarded as FOBI applied from both sides simultaneously. Recovering the matrix $\mathbf{Z}$ only up to proportionality is not a problem as we can always estimate the constant of proportionality using the assumption that $\mathrm{cov}\{\mathrm{vec}(\mathbf{Z})\} = \mathbf{I}_{pq}$.

5. Extension to tensor-valued data

In this section we further extend FOBI to tensor-valued data, producing a method we refer to as TFOBI. To handle summations over multiple indices we use Einstein's summation convention [24]; i.e., whenever an index appears twice, summation over that index is implied. For example, for a 4-dimensional tensor $\boldsymbol{A} = \{a_{ijk\ell}\}$, the symbol $a_{abjk}a_{cdjk}$ stands for

$$\textstyle\sum_j \sum_k a_{abjk}a_{cdjk}.$$

That is, $\{a_{abjk}a_{cdjk}\}$ is a 4-dimensional tensor, the (a, b, c, d)th entry of which is given above.

5.1. Tensor independent component model

Let $\boldsymbol{X}$ be a random element in $\mathbb{R}^{p_1\times\cdots\times p_r}$, i.e., a random tensor of order r. Following [18], for a given tensor $\boldsymbol{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$, we call any p_m-vector obtained by letting i_m vary over $\{1, \ldots, p_m\}$ while fixing all the other indices an m-mode vector. For each $m \in \{1, \ldots, r\}$, the term mth mode (or "m-mode") refers to the mth direction of a tensor of order r. In some sense the opposite operation, fixing a value of one of the indices, $i_m \in \{1, \ldots, p_m\}$, while varying the others produces what we call the m-mode faces of a tensor. For any given $m \in \{1, \ldots, r\}$, a tensor $\boldsymbol{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$ thus has in total p_m m-mode faces of size $p_1 \times \cdots \times p_{m-1} \times p_{m+1} \times \cdots \times p_r$. Notice that for each $i_m \in \{1, \ldots, p_m\}$, the set of i_mth elements of all m-mode vectors of a tensor $\mathbf{A}$ contains the same elements as the i_mth m-mode face of $\mathbf{A}$.

To work with tensors we next introduce a product operation between a tensor and a matrix that provides a higher order generalization of a linear transformation of a vector by matrix. Following again [18], for $\boldsymbol{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$ and $\boldsymbol{B} \in \mathbb{R}^{q_m\times p_m}$, let $\boldsymbol{A} \odot_m \boldsymbol{B}$ be the $p_1 \times \cdots \times p_{m-1} \times q_m \times p_{m+1} \times \cdots \times p_r$ dimensional tensor whose $(i_1, \ldots, j_m, \ldots, i_r)$th entry is

$$(\boldsymbol{A} \odot_m \boldsymbol{B})_{i_1\ldots j_m\ldots i_r} = a_{i_1\ldots i_m\ldots i_r} b_{j_m i_m}. \tag{6}$$

Let $\boldsymbol{B}_1 \in \mathbb{R}^{q_1\times p_1}, \ldots, \boldsymbol{B}_r \in \mathbb{R}^{q_r\times p_r}$. We use the notation $\boldsymbol{A} \odot_1 \boldsymbol{B}_1 \odot \cdots \odot_r \boldsymbol{B}_r$ to abbreviate the tensor

$$(\ldots(\boldsymbol{A} \odot_1 \boldsymbol{B}_1) \odot_2 \boldsymbol{B}_2 \ldots) \odot_r \boldsymbol{B}_r) = \{a_{j_1\ldots j_r} b_{i_1 j_1} \ldots b_{i_r j_r}\}.$$

It is easy to see that for a vector $\mathbf{a} \in \mathbb{R}^{p_1}$ we have $(\mathbf{a}\odot_1 \mathbf{B}_1) = \mathbf{B}_1\mathbf{a}$ and for a matrix $\mathbf{A} \in \mathbb{R}^{p_1\times p_2}$ similarly $(\mathbf{A}\odot_1 \mathbf{B}_1) = \mathbf{B}_1\mathbf{A}$ and $(\mathbf{A}\odot_2 \mathbf{B}_2) = \mathbf{A}\mathbf{B}_2^\top$, assuming $\mathbf{B}_1$ and $\mathbf{B}_2$ are of appropriate size. Thus $\odot_m$ can be seen as a linear transformation from the direction of the mth mode. Using m-mode vectors the multiplication has also a second interpretation; $(\mathbf{A} \odot_m \mathbf{B}_m)$ applies the linear transformation given by $\mathbf{B}_m$ individually to each m-mode vector of $\mathbf{A}$.

The previous multiplication operation is also commutative in the sense that for distinct values of m, the order we apply the multiplications $\odot_m$ has no effect on the outcome. If we want to multiply multiple times from the direction of the same mode commutativity fails and we instead have the following lemma.

Lemma 1. *For any $\boldsymbol{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$, $\boldsymbol{B}_1 \in \mathbb{R}^{p_1\times p_1}, \ldots, \boldsymbol{B}_r \in \mathbb{R}^{p_r\times p_r}$, $\boldsymbol{C}_1 \in \mathbb{R}^{p_1\times p_1}, \ldots, \boldsymbol{C}_r \in \mathbb{R}^{p_r\times p_r}$, we have*

$$\boldsymbol{A} \odot_1 (\boldsymbol{B}_1\boldsymbol{C}_1) \odot \cdots \odot_r (\boldsymbol{B}_r\boldsymbol{C}_r) = \boldsymbol{A} \odot_1 \boldsymbol{C}_1 \odot \cdots \odot_r \boldsymbol{C}_r \odot_1 \boldsymbol{B}_1 \odot \cdots \odot_r \boldsymbol{B}_r. \tag{7}$$

Proof. By definition, the right-hand side is the tensor in $\mathbb{R}^{p_1\times\cdots\times p_r}$ whose $(i_1 \ldots i_r)$th entry is

$$a_{k_1\ldots k_r} b_{j_1 k_1} \ldots b_{j_r k_r} c_{i_1 j_1} \ldots c_{i_r j_r} = a_{k_1\ldots k_r}(c_{i_1 j_1} b_{j_1 k_1}) \ldots (c_{i_r j_r} b_{j_r k_r}).$$

The right-hand side is precisely the $(i_1 \ldots i_r)$th entry of the tensor on the left-hand side of (7). □

We now have sufficient tools to define the independent component model for tensors.

Definition 3. *The tensor-valued independent component model assumes that the tensors* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *are iid realizations of a random tensor* $\boldsymbol{X}$ *satisfying*

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{Z} \odot_1 \boldsymbol{\Omega}_1 \odot \cdots \odot_r \boldsymbol{\Omega}_r. \tag{8}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, $\boldsymbol{\Omega}_1 \in \mathcal{A}^{p_1}, \ldots, \boldsymbol{\Omega}_r \in \mathcal{A}^{p_r}$, *and* $\boldsymbol{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *satisfies Assumptions T1 and T2 below.*

Assumption T1. *The components* $z_{k_1 \ldots k_r}$ *of* $\boldsymbol{Z}$ *are mutually independent and standardized in the sense that* $\mathrm{E}(z_{k_1 \ldots k_r}) = 0$ *and* $\mathrm{var}(z_{k_1 \ldots k_r}) = 1$.

Assumption T2. *For each* $m \in \{1, \ldots, r\}$*, at most one m-mode face of* $\boldsymbol{Z}$ *consists entirely of Gaussian components.*

The above assumptions serve the same purpose as the corresponding assumptions of the vector and matrix independent component models in Section 3. The need for Assumption T2 can be seen by considering the product operation $\odot_m \boldsymbol{\Omega}_m$ as a linear transformation of the m-mode vectors by $\boldsymbol{\Omega}_m$ and observing that if two or more m-mode faces had only Gaussian components, the corresponding columns of $\boldsymbol{\Omega}_m$ would be rotationally invariant.

5.2. The m-mode moment matrices of a random tensor

The matrix unmixing procedure described in Section 4 involves left and right standardization and then left and right rotation. We need to generalize these to m-mode standardization and m-mode rotation. We first define the m-mode product between two tensors: for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, the m-mode product, $\boldsymbol{A} \odot_{-m} \boldsymbol{B}$, is the $p_m \times p_m$ matrix, the (s,t)th entry of which is

$$(\boldsymbol{A} \odot_{-m} \boldsymbol{B})_{st} = a_{i_1 \ldots i_{m-1}\, s\, i_{m+1} \ldots i_r} b_{i_1 \ldots i_{m-1}\, t\, i_{m+1} \ldots i_r}.$$

In some sense the operation $\odot_{-m}$ is opposite to the operation $\odot_m$ in (6); whereas $\odot_m$ involves the sum over the mth index of $\boldsymbol{A}$, $\odot_{-m}$ involves the sum over all indices except the mth index of $\boldsymbol{A}$ and $\boldsymbol{B}$.

We now extend moment matrices such as

$$\mathrm{cov}\,(\boldsymbol{X}), \quad \mathrm{cov}(\boldsymbol{X}^\top), \quad \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top) \quad \text{and} \quad \mathrm{E}(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X})$$

to the tensor case. Again, for convenience and without loss of generality, assume that $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{0}$. As in the matrix case, there are two generalizations of the FOBI functional.

Definition 4. *The m-mode covariance and two types of m-mode FOBI functionals of a random tensor* $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ *are the following* $p_m \times p_m$ *matrices*

$$\mathrm{cov}_{(m)}(\boldsymbol{X}) = \left(\textstyle\prod_{s \neq m} p_s\right)^{-1} \mathrm{E}\,(\boldsymbol{X} \odot_{-m} \boldsymbol{X}), \quad \boldsymbol{B}^0_{(m)}(\boldsymbol{X}) = \left(\textstyle\prod_{s \neq m} p_s\right)^{-1} \mathrm{E}\{(\boldsymbol{X} \odot_{-m} \boldsymbol{X})^2\}$$

and

$$\boldsymbol{B}^1_{(m)}(\boldsymbol{X}) = \left(\textstyle\prod_{s \neq m} p_s\right)^{-1} \mathrm{E}\{\|\boldsymbol{X}\|_F^2 (\boldsymbol{X} \odot_{-m} \boldsymbol{X})\},$$

where $\|\cdot\|_F^2$ *is the squared Frobenius norm of a tensor (the sum of squared elements).*

Define further

$$\rho_m = \textstyle\prod_{s \neq m} p_s. \tag{9}$$

This proportionality constant reflects the fact that $\odot_{-m}$ involves the sum of ρ_m terms.

5.3. The m-mode standardization

Similar to random matrix unmixing our idea of unmixing a random tensor also consists of two steps: standardization and rotation, except now the two operations have to be performed on each of the m modes of the r-tensor. For each $m \in \{1, \dots, r\}$, let

$$\boldsymbol{\Omega}_m = \boldsymbol{U}_m \boldsymbol{D}_m \boldsymbol{V}_m^\top \tag{10}$$

be the singular value decomposition of $\boldsymbol{\Omega}_m$. The next theorem shows that we can recover $\boldsymbol{\Omega}_m \boldsymbol{\Omega}_m^\top$ (up to a proportionality constant) from the m-mode covariance matrix of $\boldsymbol{X}$.

Lemma 2. *Under the tensor IC model in Definition 3 we have*

$$\operatorname{cov}_{(m)}(\boldsymbol{X}) = \rho_m^{-1} (\textstyle\prod_{s \neq m} \|\boldsymbol{D}_s\|_F^2) \boldsymbol{U}_m \boldsymbol{D}_m^2 \boldsymbol{U}_m^\top. \tag{11}$$

Proof. Without loss of generality, assume $m = 1$. By definition, the (a, b)th entry of the matrix $\mathrm{E}(\boldsymbol{X} \odot_{-1} \boldsymbol{X})$ is

$$\mathrm{E}(x_{ai_2 \dots i_r} x_{bi_2 \dots i_r}). \tag{12}$$

By the tensor IC model (8) we have

$$x_{i_1 \dots i_r} = \omega^{(r)}_{i_r j_r} \dots \omega^{(1)}_{i_1 j_1} z_{j_1 \dots j_r}$$

where $\omega^{(m)}_{ij}$ are the entries of $\boldsymbol{\Omega}_m$. Hence (12) can be rewritten as

$$\mathrm{E}(\omega^{(r)}_{i_r j_r} \dots \omega^{(1)}_{a j_1} z_{j_1 \dots j_r} \omega^{(r)}_{i_r k_r} \dots \omega^{(1)}_{b k_1} z_{k_1 \dots k_r}) = \omega^{(r)}_{i_r j_r} \dots \omega^{(1)}_{a j_1} \omega^{(r)}_{i_r k_r} \dots \omega^{(1)}_{b k_1} \delta_{j_1 k_1} \dots \delta_{j_r k_r}.$$

By the properties of the Kronecker delta, we can express the above as

$$\omega^{(r)}_{i_r j_r} \dots \omega^{(1)}_{a j_1} \omega^{(r)}_{i_r j_r} \dots \omega^{(1)}_{b j_1} = (\omega^{(2)}_{i_2 j_2} \omega^{(2)}_{i_2 j_2}) \dots (\omega^{(r)}_{i_r j_r} \omega^{(r)}_{i_r j_r}) (\omega^{(1)}_{a j_1} \omega^{(1)}_{b j_1})$$

which is the (a, b)th entry of the matrix $(\prod_{s \neq 1} \|\boldsymbol{\Omega}_s\|_F^2) \boldsymbol{\Omega}_1 \boldsymbol{\Omega}_1^\top$. Now the assertion of the theorem follows from the singular value decomposition (10). □

Let $\boldsymbol{S}_m = \operatorname{cov}_{(m)}(\boldsymbol{X})$. Relation (11) means that

$$\rho_m^{-1} (\textstyle\prod_{s \neq m} \|\boldsymbol{D}_s\|_F^2) \boldsymbol{U}_m \boldsymbol{D}_m^2 \boldsymbol{U}_m^\top$$

is in fact the eigendecomposition of $\boldsymbol{S}_m$. Thus, all inverse square roots of $\boldsymbol{S}_m$ are of the form

$$(\textstyle\prod_{s \neq m} p_s^{1/2} \|\boldsymbol{D}_s\|_F^{-1}) \boldsymbol{M}_m \boldsymbol{D}_m^{-1} \boldsymbol{U}_m^\top,$$

where $\boldsymbol{M}_m \in \mathcal{U}^{p_m}$. We can use these square roots to recover a rotated version of $\boldsymbol{Z}$, as indicated by the next theorem.

Theorem 2. *Let $\boldsymbol{S}_m$ be as defined in the last paragraph. Then, under the tensor independent component model of Definition 3,*

$$\boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \dots \odot_r \boldsymbol{S}_r^{-1/2} = \tau \boldsymbol{Z} \odot_1 \boldsymbol{W}_1 \odot \dots \odot_r \boldsymbol{W}_r, \tag{13}$$

where, for $m \in \{1, \dots, r\}$,

$$\boldsymbol{W}_m = \boldsymbol{M}_m \boldsymbol{V}_m^\top \in \mathcal{U}^{p_m}, \quad \tau = (\textstyle\prod_{m=1}^r \prod_{s \neq m} p_s^{1/2} \|\boldsymbol{D}_s\|_F^{-1}). \tag{14}$$

Proof. By Lemma 1,

$$\begin{aligned} \boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \dots \odot_r \boldsymbol{S}_r^{-1/2} &= \boldsymbol{Z} \odot_1 \boldsymbol{\Omega}_1 \odot \dots \odot_r \boldsymbol{\Omega}_r \odot_1 \boldsymbol{S}_1^{-1/2} \odot \dots \odot_r \boldsymbol{S}_r^{-1/2} \\ &= \boldsymbol{Z} \odot_1 \boldsymbol{S}_1^{-1/2} \boldsymbol{\Omega}_1 \odot \dots \odot_r \boldsymbol{S}_r^{-1/2} \boldsymbol{\Omega}_r. \end{aligned} \tag{15}$$

However, we note that

$$\boldsymbol{S}_m^{-1/2} \boldsymbol{\Omega}_m = (\textstyle\prod_{s \neq m} p_s^{1/2} \|\boldsymbol{D}_s\|_F^{-1}) \boldsymbol{M}_m \boldsymbol{D}_m^{-1} \boldsymbol{U}_m^\top \boldsymbol{U}_m \boldsymbol{D}_m \boldsymbol{V}_m^\top = (\textstyle\prod_{s \neq m} p_s^{1/2} \|\boldsymbol{D}_s\|_F^{-1}) \boldsymbol{W}_m,$$

where $\boldsymbol{W}_m = \boldsymbol{M}_m \boldsymbol{V}_m^\top \in \mathcal{U}^{p_m}$. Substitute the above into (15) to prove the desired equality. □

The tensor on the right-hand side of (13) is only a rotation away from the independent component tensor $\boldsymbol{Z}$, a step we carry out in the next subsection.

5.4. The m-mode rotation

Let

$$\mathbf{B}_m^0 = \rho_m^{-1}\mathbf{B}_{(m)}^0(\boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \cdots \odot_r \boldsymbol{S}_r^{-1/2}) \quad \text{and} \quad \mathbf{B}_m^1 = \rho_m^{-1}\mathbf{B}_{(m)}^1(\boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \cdots \odot_r \boldsymbol{S}_r^{-1/2}), \tag{16}$$

where $\boldsymbol{B}_{(m)}^0$ and $\boldsymbol{B}_{(m)}^1$ are the FOBI functionals in Definition 4. In order to manipulate them, we need the following lemma.

Lemma 3. *Let* $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p_1\times\cdots\times p_r}$, $\boldsymbol{U}_1 \in \mathcal{U}^{p_1}, \ldots, \boldsymbol{U}_r \in \mathcal{U}^{p_r}$. *Then*

$$(\boldsymbol{A} \odot_1 \boldsymbol{U}_1 \odot \cdots \odot_r \boldsymbol{U}_r) \odot_{-m} (\boldsymbol{B} \odot_1 \boldsymbol{U}_1 \odot \cdots \odot_r \boldsymbol{U}_r) = \boldsymbol{U}_m(\boldsymbol{A} \odot_{-m} \boldsymbol{B})\boldsymbol{U}_m^\top. \tag{17}$$

Proof. Without loss of generality assume that $m = 1$. The (a, b)th entry of the matrix on the left-hand side of (17) is

$$\begin{aligned}(\boldsymbol{A} \odot_1 \boldsymbol{U}_1 \odot \cdots \odot_r \boldsymbol{U}_r)_{a\,i_2\ldots i_r}(\boldsymbol{B} \odot_1 \boldsymbol{U}_1 \odot \cdots \odot_r \boldsymbol{U}_r)_{b\,i_2\ldots i_r} &= a_{j_1\ldots j_r}u_{a\,j_1}^{(1)}u_{i_2\,j_2}^{(2)}\ldots u_{i_r\,j_r}^{(r)}b_{k_1\ldots k_r}u_{b\,k_1}^{(1)}u_{i_2\,k_2}^{(2)}\ldots u_{i_r k_r}^{(r)}\\ &= a_{j_1\ldots j_r}b_{k_1\ldots k_r}(u_{i_2\,j_2}^{(2)}u_{i_2\,k_2}^{(2)})\ldots(u_{i_r\,j_r}^{(r)}u_{i_r k_r}^{(r)})u_{a\,j_1}^{(1)}u_{b\,k_1}^{(1)}\\ &= a_{j_1\ldots j_r}b_{k_1\ldots k_r}\delta_{j_2k_2}\ldots\delta_{j_r k_r}u_{a\,j_1}^{(1)}u_{b\,k_1}^{(1)}.\end{aligned}$$

The above reduces to $a_{j_1 j_2\ldots j_r}b_{k_1 j_2\ldots j_r}u_{a\,j_1}^{(1)}u_{b\,k_1}^{(1)}$, which is the (a, b)th entry of the matrix on the right-hand side of (17). This completes the argument. □

Define the m-flattening, or m-unfolding, of a tensor $\mathbf{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$ to be the matrix $\mathbf{A}_{(m)} \in \mathbb{R}^{p_m\times\rho_m}$ obtained by taking all the m-mode vectors of $\mathbf{A}$ and stacking them horizontally into a matrix. As for the order of stacking we choose to use the cyclical unfolding described in [18]. Then, for $\mathbf{A}^* = \mathbf{A} \odot_1 \mathbf{B}_1 \odot \cdots \odot_r \mathbf{B}_r$, we have

$$\mathbf{A}_{(m)}^* = \mathbf{B}_m\mathbf{A}_{(m)}\,(\mathbf{B}_{m+1} \otimes \cdots \otimes \mathbf{B}_r \otimes \mathbf{B}_1 \otimes \cdots \otimes \mathbf{B}_{m-1})\,. \tag{18}$$

Flattening can also be used to express the m-mode product of a tensor with itself with means of ordinary matrix multiplication. Namely,

$$\boldsymbol{A} \odot_{-m} \boldsymbol{A} = \boldsymbol{A}_{(m)}\boldsymbol{A}_{(m)}^\top. \tag{19}$$

For a tensor $\boldsymbol{A} \in \mathbb{R}^{p_1\times\cdots\times p_r}$ let $\bar{\boldsymbol{A}}_{-m}$ be the p_m-vector whose i_mth element is the mean of the ρ_m elements of the i_mth m-mode face of $\mathbf{A}$ with $i_m \in \{1, \ldots, p_m\}$. Expressed via the previously defined m-flattening $\bar{\boldsymbol{A}}_{-m}$ thus contains the row means of $\mathbf{A}_{(m)}$.

The next theorem shows that the rotations $\boldsymbol{W}_m$ can be recovered from the eigendecompositions of $\boldsymbol{B}_m^0$ and $\boldsymbol{B}_m^1$.

Theorem 3. *Let* ρ_m, $\boldsymbol{W}_m$, τ, $\boldsymbol{B}_m^0$ *and* $\boldsymbol{B}_m^1$ *be as defined in* (9), (14), *and* (16). *Let* $\boldsymbol{\beta} \in \mathbb{R}^{p_1\times\cdots\times p_r}$ *be the tensor with the entries* $\mathrm{E}(z_{i_1\ldots i_r}^4)$. *Then*

$$\boldsymbol{B}_m^0 = \tau^4\boldsymbol{W}_m\{(p_m - 1 + \rho_m - 1)\boldsymbol{I}_{p_m} + \mathrm{diag}(\bar{\boldsymbol{\beta}}_{-m})\}\boldsymbol{W}_m^\top, \quad \boldsymbol{B}_m^1 = \tau^4\boldsymbol{W}_m\{(p_m\rho_m - 1)\boldsymbol{I}_{p_m} + \mathrm{diag}(\bar{\boldsymbol{\beta}}_{-m})\}\boldsymbol{W}_m^\top,$$

where $\mathrm{diag}(\bar{\boldsymbol{\beta}}_{-m})$ *is the diagonal matrix having the elements of* $\bar{\boldsymbol{\beta}}_{-m} \in \mathbb{R}^{p_m}$ *on its diagonal.*

Proof. Again, without loss of generality, assume $m = 1$. By definition,

$$\boldsymbol{B}_1^0 = \rho_1^{-1}\mathrm{E}\left[\left\{(\boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \cdots \odot_r \boldsymbol{S}_r^{-1/2}) \odot_{-1} (\boldsymbol{X} \odot_1 \boldsymbol{S}_1^{-1/2} \odot \cdots \odot_r \boldsymbol{S}_r^{-1/2})\right\}^2\right].$$

By Theorem 2, the right-hand side is

$$\rho_1^{-1}\tau^4\mathrm{E}\left[\{(\boldsymbol{Z} \odot_1 \boldsymbol{W}_1 \odot \cdots \odot_r \boldsymbol{W}_r) \odot_{-1} (\boldsymbol{Z} \odot_1 \boldsymbol{W}_1 \odot \cdots \odot_r \boldsymbol{W}_r)\}^2\right].$$

By Lemma 3, this is $\rho_1^{-1}\tau^4\boldsymbol{W}_1\mathrm{E}\{(\boldsymbol{Z} \odot_{-1} \boldsymbol{Z})^2\}\boldsymbol{W}_1^\top$ and by (19) the expectation can be expressed as $\mathrm{E}\{(\boldsymbol{Z} \odot_{-1} \boldsymbol{Z})^2\} = \mathrm{E}(\boldsymbol{Z}_{(1)}\boldsymbol{Z}_{(1)}^\top\boldsymbol{Z}_{(1)}\boldsymbol{Z}_{(1)}^\top)$. Now applying the matrix identities in (5) completes the proof for $\boldsymbol{B}_m^0$. The proof for $\boldsymbol{B}_m^1$ is carried out similarly by reducing the matter into the matrix case. □

Theorem 3 says that $\boldsymbol{W}_m$ has the eigenvectors of $\boldsymbol{B}_m^0$ and $\boldsymbol{B}_m^1$ as its columns. In other words, we can recover the orthogonal matrices $\boldsymbol{W}_m$ from the eigendecompositions of $\boldsymbol{B}_m^0$ or $\boldsymbol{B}_m^1$ for each $m \in \{1, \ldots, r\}$. Again, to identify the eigenbases we need the following assumption.

Assumption T3. *For each $m \in \{1, \ldots, r\}$, the components of $\bar{\boldsymbol{\beta}}_{-m}$ are distinct.*

The next corollary puts the m-mode standardizations and rotations together to recover the independent components from a random tensor $\boldsymbol{X}$.

Corollary 1. *For each $m \in \{1, \ldots, r\}$, let $\boldsymbol{S}_m^{-1/2}$ be any square root of $\boldsymbol{S}_m$ and let $\boldsymbol{W}_m$ have the eigenvectors of either $\boldsymbol{B}_m^0$ or $\boldsymbol{B}_m^1$ as its columns. Then, under Assumptions T1 and T3, we have*

$$\boldsymbol{X} \odot_1 (\boldsymbol{W}_1^\top \boldsymbol{S}_1^{-1/2}) \odot \cdots \odot_r (\boldsymbol{W}_r^\top \boldsymbol{S}_r^{-1/2}) = \tau \boldsymbol{Z}.$$

Proof. Multiply both sides of the equation (13) from the right by $\odot_1 \boldsymbol{W}_1^\top \odot \cdots \odot_r \boldsymbol{W}_r^\top$ and evoke tensor-matrix product rule in Lemma 1 to prove the result. □

6. Limiting distributions

In this section we pursue the asymptotic distributions of the unmixing estimates given by the extended ICA procedures in the previous sections. We will focus primarily on MFOBI because the corresponding results for TFOBI follow directly from the results for MFOBI, as detailed in Remark 1. However, we first discuss the important concept of equivariance.

6.1. Equivariance and independent component functionals

In the vector-valued case for example [25] state that an unmixing functional $\boldsymbol{\Gamma}$ must satisfy the following two conditions:

(i) For a distribution of $\mathbf{z} \in \mathbb{R}^p$ with standardized and mutually independent components, $\boldsymbol{\Gamma}(\mathbf{z}) = \mathbf{I}_p$.

(ii) For the distribution of any $\mathbf{x} \in \mathbb{R}^p$, it holds that $\boldsymbol{\Gamma}(\mathbf{A}\mathbf{x}) = \boldsymbol{\Gamma}(\mathbf{x})\mathbf{A}^{-1}$, for all $\mathbf{A} \in \mathcal{A}^p$.

In both conditions the equalities are understood up to permutation and sign changes of the rows. The second condition means that the functional is equivariant under affine transformations and $\boldsymbol{\Gamma}(\mathbf{x})\mathbf{x}$ is thus independent of the used coordinate system. Theoretical derivations can then be limited to the case $\boldsymbol{\Omega} = \mathbf{I}_p$.

Consider next the unmixing matrix functionals in the tensor case and for each $m \in \{1, \ldots, r\}$, write $\boldsymbol{\Gamma}_{(m)}(\mathbf{X}) = \boldsymbol{W}_m^\top \boldsymbol{S}_m^{-1/2}$ for the m-mode unmixing matrix functional. The functional $\boldsymbol{\Gamma}_{(m)}$ is said to be *(fully) affine equivariant* if, for all $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ and all $\mathbf{A}_1 \in \mathcal{A}^{p_1}, \ldots, \mathbf{A}_r \in \mathcal{A}^{p_r}$,

$$\boldsymbol{\Gamma}_{(m)}(\mathbf{X} \odot_1 \mathbf{A}_1 \odot \cdots \odot_r \mathbf{A}_r) = \boldsymbol{\Gamma}_{(m)}(\mathbf{X})\mathbf{A}_m^{-1}.$$

This is however true for our unmixing matrix functionals only if $\mathbf{A}_1, \ldots, \mathbf{A}_r$ are all orthogonal. The TFOBI unmixing matrix functionals $\boldsymbol{\Gamma}_{(m)}$ are thus *orthogonally equivariant*. Also the weaker *marginal affine equivariance*

$$\boldsymbol{\Gamma}_{(m)}(\mathbf{X} \odot_m \mathbf{A}_m) = \boldsymbol{\Gamma}_{(m)}(\mathbf{X})\mathbf{A}_m^{-1},$$

for some fixed $m \in \{1, \ldots, r\}$, holds only if all $\mathbf{A}_s$, $s \neq m$ are orthogonal. The reason why both of these conditions fail in the general case is that the m-mode covariance functionals are not fully affine equivariant in the sense that, for all $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, all $\mathbf{A}_1 \in \mathcal{A}^{p_1}, \ldots, \mathbf{A}_r \in \mathcal{A}^{p_r}$ and all $m \in \{1, \ldots, r\}$,

$$\operatorname{cov}_{(m)}(\mathbf{X} \odot_1 \mathbf{A}_1 \odot \cdots \odot_r \mathbf{A}_r) = \mathbf{A}_m \operatorname{cov}_{(m)}(\mathbf{X})\mathbf{A}_m^\top. \tag{20}$$

The condition (20) also holds only if $\mathbf{A}_1, \ldots, \mathbf{A}_r$ are all orthogonal, leading then into the orthogonal equivariance and marginal orthogonal equivariance of $\boldsymbol{\Gamma}_{(m)}$. In fact, (20) in general seems such strict a requirement that we conjecture that no functional satisfying it exists. This would then imply also that no fully affine equivariant tensor unmixing

matrix functionals based on separate standardization and rotation steps exist. Note however, that marginally affine equivariant $\mathbf{\Gamma}_{(m)}$ for a single direction can be obtained if $\text{cov}_{(m)}$ and then $\mathbf{B}^0_{(m)}$ or $\mathbf{B}^1_{(m)}$ are applied separately for each direction.

The lack of full affine equivariance means that the asymptotic results for the unmixing matrix estimates for general $\mathbf{\Omega}_L$ and $\mathbf{\Omega}_R$ no longer follow from the results in the simple case, $\mathbf{\Omega}_L = \mathbf{I}_p$, $\mathbf{\Omega}_R = \mathbf{I}_q$, and thus the comparison of different estimates becomes difficult. In the following, we find the limiting distributions of the FOBI estimate $\hat{\mathbf{\Gamma}}$ and the MFOBI estimates $\hat{\mathbf{\Gamma}}_L$ and $\hat{\mathbf{\Gamma}}_R$ under the assumptions that $\mathbf{\Omega} = \mathbf{I}_p$ (FOBI) and that $\mathbf{\Omega}_L = \mathbf{I}_p$ and $\mathbf{\Omega}_R = \mathbf{I}_q$ (MFOBI). The estimates are obtained by applying the functionals to empirical distributions of sample size n.

6.2. Limiting distribution of the FOBI estimate

The asymptotic behavior of the classic FOBI was first derived in [12] and requires Assumption V3 on the distinct kurtosis values of the components. The following results are however in the form of [27], see their Theorem 8 and Corollary 3.

Theorem 4. *Let $z_1, \ldots, z_n$ be a random sample from a p-variate distribution having finite eighth moments and satisfying assumptions V1 and V3. Assume further that $\mathbf{\Omega} = \boldsymbol{I}_p$ and that the standardization functional $\hat{\boldsymbol{S}}^{-1/2}$ is chosen to be symmetric. Then there exists a sequence of FOBI estimates such that $\hat{\mathbf{\Gamma}} \to_P \boldsymbol{I}_p$ and*

$$\sqrt{n}\,(\hat{\gamma}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1), \qquad \sqrt{n}\,\hat{\gamma}_{kk'} = \frac{\sqrt{n}\hat{Q} - (\beta_k + p + 1)\sqrt{n}\hat{s}_{kk'}}{\beta_k - \beta_{k'}} + o_P(1),$$

where $\hat{Q} = \hat{q}_{kk'} + \hat{q}_{k'k} + \sum_{m\neq k,k'} \hat{q}_{mkk'}$ and $k \neq k'$.

Based on Theorem 4 we can then compute the asymptotic variances of the elements of the estimated unmixing matrix $\hat{\mathbf{\Gamma}}$.

Corollary 2. *Under the assumptions of Theorem 4 the limiting distribution of $\sqrt{n}\,\text{vec}(\hat{\mathbf{\Gamma}} - \boldsymbol{I}_p)$ is multivariate Normal with mean vector $\boldsymbol{0}_{p^2}$ and the following asymptotic variances.*

$$ASV(\hat{\gamma}_{kk}) = \frac{\beta_k - 1}{4}, \qquad ASV(\hat{\gamma}_{kk'}) = \frac{\omega_k + \omega_{k'} - \beta_k^2 - 6\beta_{k'} + 9 + \sum_{m\neq k,k'}(\beta_m - 1)}{(\beta_k - \beta_{k'})^2}, \qquad k \neq k'.$$

As Corollary 2 shows, the asymptotic variance of any off-diagonal element $\hat{\gamma}_{kk'}$ of the unmixing matrix depends also on components other than z_k and $z_{k'}$ (via their kurtoses). Of the commonly used independent component analysis methods, FastICA, FOBI and JADE, FOBI is unique in this sense, partly explaining its inferiority to the other methods.

6.3. Limiting distribution of the MFOBI estimate

We provide the asymptotic properties of only the left-hand side unmixing matrix estimate $\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Gamma}}_L$, the right-hand side version being again easily obtained by reversing the roles of rows and columns. Here $N = 0$ or $N = 1$ depending on the choice of the FOBI functional and the sample left and right covariance matrices are denoted by $\bar{\boldsymbol{S}}_L = (\bar{s}^L_{kk'})$ and $\bar{\boldsymbol{S}}_R = (\bar{s}^R_{kk'})$, respectively.

Theorem 5. *Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be a random sample from a distribution of a matrix-valued $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ having finite eighth moments and satisfying assumptions M1 and M3. Assume further that $\mathbf{\Omega}_L = \boldsymbol{I}_p$ and $\mathbf{\Omega}_R = \boldsymbol{I}_q$, and that the left and right standardization functionals, $\bar{\boldsymbol{S}}_L^{-1/2}$ and $\bar{\boldsymbol{S}}_R^{-1/2}$, are chosen to be symmetric. Then there exists a sequence of left MFOBI estimates such that $\hat{\mathbf{\Gamma}} \to_P \boldsymbol{I}_p$ and*

$$\sqrt{n}\,(\hat{\gamma}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\bar{s}_{kk} - 1) + o_P(1), \qquad \sqrt{n}\hat{\gamma}_{kk'} = \frac{\sqrt{n}\bar{Q} + \sqrt{n}\bar{R}^N - (\bar{\beta}_{k\bullet} + b_N)\sqrt{n}\bar{s}_{kk'}}{(\bar{\beta}_{k\bullet} - \bar{\beta}_{k'\bullet})} + o_P(1), \quad k \neq k',$$

where $\bar{Q} = \bar{q}_{kk'} + \bar{q}_{k'k} + \sum_{m\neq k,k'} \bar{q}_{mkk'}$, $\bar{R}^N = \bar{r}_{kk'} + \bar{r}_{k'k} + \sum_{m\neq k,k'} \bar{r}^N_{mkk'}$, $b_0 = 2q + p - 1$ and $b_1 = qp + 1$.

Corollary 3. *i) Under the assumptions of Theorem 5 the limiting distribution of* $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{\Gamma}}-\boldsymbol{I}_p)$ *is multivariate Normal with mean vector* $\boldsymbol{0}_{p^2}$ *and the following asymptotic variances:*

$$AS\,V\,(\hat{\gamma}_{kk}) = \frac{\bar{\beta}_{k\bullet} - 1}{4q},$$
$$AS\,V(\hat{\gamma}_{kk'}) = \frac{\bar{\omega}_{k\bullet} + \bar{\omega}_{k'\bullet} - \bar{\beta}^2_{k\bullet} + 2\delta_{kk'} + (q-1)\bar{\beta}_{k\bullet} + (q-7)\bar{\beta}_{k'\bullet} + c_N}{q(\bar{\beta}_{k\bullet} - \bar{\beta}_{k'\bullet})^2}, \quad k \neq k',$$

where $c_0 = \sum_{m\neq kk'} \bar{\beta}_{m\bullet} + pq - 2p - 4q + 15$ *and* $c_1 = q\sum_{m\neq kk'} \bar{\beta}_{m\bullet} - pq + 11$.

Proof. The proof for the consistency of the estimator is obtained similarly as in the proof of Theorem 5.1.1 in [48]. Write then

$$\bar{\mathbf{L}} = (\bar{\ell}_{kk'}) = \bar{\mathbf{S}}_L^{-1/2} \to_P \mathbf{I}_p, \quad \bar{\mathbf{L}}^* = \bar{\mathbf{L}}^\top\bar{\mathbf{L}} \to_P \mathbf{I}_p, \quad \bar{\mathbf{R}} = (\bar{r}_{\ell\ell'}) = \bar{\mathbf{S}}_R^{-1/2} \to_P \mathbf{I}_q, \quad \bar{\mathbf{R}}^* = \bar{\mathbf{R}}^\top\bar{\mathbf{R}} \to_P \mathbf{I}_q.$$

Limiting Normal distributions of the components of the sample covariance functionals imply that $\sqrt{n}\,(\bar{\mathbf{S}}_L - \mathbf{I}_p) = O_P(1)$ and $\sqrt{n}\,(\bar{\mathbf{S}}_R - \mathbf{I}_q) = O_P(1)$ and the following two asymptotic expansions are then easy to prove using Slutsky's theorem:

$$\sqrt{n}\,(\bar{\mathbf{L}} - \mathbf{I}_p) = -\frac{1}{2}\sqrt{n}\,(\bar{\mathbf{S}}_L - \mathbf{I}_p) + o_P(1), \quad \sqrt{n}\,(\bar{\mathbf{L}}^\top\bar{\mathbf{L}} - \mathbf{I}_p) = \sqrt{n}\,(\bar{\mathbf{L}} - \mathbf{I}_p) + \sqrt{n}\,(\bar{\mathbf{L}}^\top - \mathbf{I}_p) + o_P(1).$$

see, e.g., the supplementary material to [48].

The estimated left unmixing functional is then $\hat{\mathbf{\Gamma}} = \hat{\mathbf{W}}_L^\top\bar{\mathbf{L}}$, where $\hat{\mathbf{W}}_L^\top$ is obtained from the eigendecomposition of the sample left FOBI functional $\bar{\mathbf{B}}_L^N = (\bar{b}^N_{kk'}) = \hat{\mathbf{W}}_L\hat{\mathbf{\Lambda}}_L^N\hat{\mathbf{W}}_L^\top$, where $\hat{\mathbf{\Lambda}}_L^N \to_P \mathbf{\Lambda}_L^N$. The asymptotic behavior of the diagonal elements $\sqrt{n}\hat{\gamma}_{kk}$ of the estimated left unmixing functional can be derived similarly as in the proof of Theorem 4.1.2 of [48]. For the off-diagonal elements, using Slutsky's theorem and the fact that $\hat{\mathbf{\Lambda}}_L^N$ is diagonal, it is straightforward to show that we have for an arbitrary (k,k')-element of the estimated left unmixing functional

$$\sqrt{n}\hat{\gamma}_{kk'} = \frac{\sqrt{n}\bar{b}^N_{kk'} + (\bar{\beta}_{k\bullet} - \bar{\beta}_{k'\bullet})\sqrt{n}\bar{\ell}_{kk'}}{\bar{\beta}_{k\bullet} - \bar{\beta}_{k'\bullet}} + o_P(1), \quad k \neq k'. \tag{21}$$

The problem lies then in finding the asymptotic behavior of an arbitrary off-diagonal element $\sqrt{n}\hat{b}^N_{kk'}$. Consider first the case $N = 0$ and write $\bar{\mathbf{B}}_L^0$ open according to its definition:

$$\sqrt{n}\,(\bar{\mathbf{B}}_L^0 - \mathbf{\Lambda}_L^0) = \bar{\mathbf{L}}\left(\sqrt{n}\,\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{Z}}_i\bar{\mathbf{R}}^*\tilde{\mathbf{Z}}_i^\top\bar{\mathbf{L}}^*\tilde{\mathbf{Z}}_i\bar{\mathbf{R}}^*\tilde{\mathbf{Z}}_i^\top\right)\bar{\mathbf{L}}^\top - \sqrt{n}\,\mathbf{\Lambda}_L^0,$$

where $\tilde{\mathbf{Z}}_i = \mathbf{Z}_i - \bar{\mathbf{Z}}$. Inspecting a single off-diagonal element yields

$$\sqrt{n}\,\bar{b}^0_{kk'} = \frac{1}{q}\sum_{defgstuv}\left(\sqrt{n}\,\bar{\ell}_{kd}\bar{r}^*_{ef}\bar{\ell}^*_{gs}\bar{r}^*_{tu}\bar{\ell}_{k'v}\frac{1}{n}\sum_{i=1}^n \tilde{z}_{i,de}\tilde{z}_{i,gf}\tilde{z}_{i,st}\tilde{z}_{i,vu}\right).$$

Next, expand each of the covariance terms one-by-one starting with $\sqrt{n}\,\bar{\ell}_{kd} = \sqrt{n}\,(\bar{\ell}_{kd} - \delta_{kd}) + \sqrt{n}\,\delta_{kd}$. After each expansion, the first term has a multiplicand that is $O_P(1)$ and Slutsky's theorem guarantees the convergence of the corresponding product. Note also that

$$\frac{1}{n}\sum_{i=1}^n \tilde{z}_{i,de}\tilde{z}_{i,gf}\tilde{z}_{i,st}\tilde{z}_{i,vu} \to_P \mathrm{E}(z_{de}z_{gf}z_{st}z_{vu}).$$

The number of sums decreases at each step finally resulting into

$$\sqrt{n}\,\bar{b}^0_{kk'} = (2q + p - 1 + \bar{\beta}_{k'\bullet})\sqrt{n}\,\hat{\ell}_{kk'} + (2q + p - 1 + \bar{\beta}_{k\bullet})\sqrt{n}\,\hat{\ell}_{k'k} + \sum_{egt}\sqrt{n}\,\frac{1}{n}\sum_{i=1}^n \tilde{z}_{i,ke}\tilde{z}_{i,ge}\tilde{z}_{i,gt}\tilde{z}_{i,k't} + o_P(1),$$

the last proper term of which partitions into the quantities defined in Section 2 as

$$\sqrt{n}\,\bar{q}_{kk'} + \sqrt{n}\,\bar{q}_{k'k} + \sum_{m\neq k,k'} \sqrt{n}\,\bar{q}_{mkk'} + \sqrt{n}\,\bar{r}_{kk'} + \sqrt{n}\,\bar{r}_{k'k} + \sum_{m\neq k,k'} \sqrt{n}\,\bar{r}^0_{mkk'},$$

after which plugging everything into expression (21) gives the desired result.

The proof for the case $N = 1$ is almost similar, only the starting expression is somewhat different:

$$\sqrt{n}\,\bar{b}^1_{kk'} = \frac{1}{q} \sum_{defghstuv} \sqrt{n}\,\bar{\ell}_{kd}\bar{r}^*_{ef}\bar{\ell}_{k'g}\bar{\ell}_{hs}\bar{r}^*_{tu}\bar{\ell}_{hv}\frac{1}{n}\sum_{i=1}^{n}\tilde{z}_{i,de}\tilde{z}_{i,gf}\tilde{z}_{i,st}\tilde{z}_{i,vu}.$$

For both choices of N the asymptotic variances of Corollary 3 are then straightforward, albeit a bit tedious, to compute using both Tables 2 and 3 containing covariances between the different terms in addition to the following covariances not fitting into the tables: $nq \times \operatorname{cov}(\bar{q}_{mkk'}, \bar{q}_{m'kk'}) = 1$, $nq \times \operatorname{cov}(\bar{q}_{mkk'}, \bar{r}^0_{m'kk'}) = 0$, $nq \times \operatorname{cov}(\bar{q}_{mkk'}, \bar{r}^1_{m'kk'}) = q^*$, $nq \times \operatorname{cov}(\bar{r}^0_{mkk'}, \bar{r}^0_{m'kk'}) = 0$ and $nq \times \operatorname{cov}(\bar{r}^1_{mkk'}, \bar{r}^1_{m'kk'}) = q^{*2}$, where $m \neq m'$ and $q^* = q - 1$.

Table 2: Covariances of $\sqrt{nq}$ times the row and column quantities, $k \neq k' \neq m$.

	$\bar{q}_{kk'}$	$\bar{q}_{k'k}$	$\bar{q}_{mkk'}$
$\bar{q}_{kk'}$	$\bar{\omega}_{k\bullet}$	$\delta_{kk'} + \bar{\beta}_{k\bullet}\bar{\beta}_{k'\bullet}$	$\bar{\beta}_{k\bullet}$
$\bar{q}_{k'k}$	–	$\bar{\omega}_{k'\bullet}$	$\bar{\beta}_{k'\bullet}$
$\bar{q}_{mkk'}$	–	–	$\bar{\beta}_{m\bullet}$

Table 3: Covariances of $\sqrt{nq}$ times the row and column quantities, $k \neq k' \neq m$ and $q^* = q - 1$.

	$\bar{r}_{kk'}$	$\bar{r}_{k'k}$	$\bar{r}^0_{mkk'}$	$\bar{r}^1_{mkk'}$	$\bar{s}_{kk'}$
$\bar{q}_{kk'}$	$q^*\bar{\beta}_{k\bullet}$	$q^*\bar{\beta}_{k\bullet}$	0	$q^*\bar{\beta}_{k\bullet}$	$\bar{\beta}_{k\bullet}$
$\bar{q}_{k'k}$	$q^*\bar{\beta}_{k'\bullet}$	$q^*\bar{\beta}_{k'\bullet}$	0	$q^*\bar{\beta}_{k'\bullet}$	$\bar{\beta}_{k'\bullet}$
$\bar{q}_{mkk'}$	q^*	q^*	0	q^*	1
$\bar{r}_{kk'}$	$q^*(q-2+\bar{\beta}_{k\bullet})$	q^{*2}	0	q^{*2}	q^*
$\bar{r}_{k'k}$	–	$q^*(q-2+\bar{\beta}_{k'\bullet})$	0	q^{*2}	q^*
$\bar{r}^0_{mkk'}$	–	–	q^*	–	0
$\bar{r}^1_{mkk'}$	–	–	–	$q^*(q-2+\bar{\beta}_{m\bullet})$	q^*
$\bar{s}_{kk'}$	–	–	–	–	1

□

Remark 1. The limiting distributions of the TFOBI estimates, $\hat{\mathbf{\Gamma}}_m = \hat{\mathbf{W}}_m^\top \hat{\mathbf{S}}_m^{-1/2}$ with $m \in \{1, \ldots, r\}$, follow straightforwardly from the results of the matrix case; using the m-flattening of tensors from Section 5 we can express the m-mode tensor product as $\mathbf{Z} \odot_{-m} \mathbf{Z} = \mathbf{Z}_{(m)}\mathbf{Z}_{(m)}^\top$, where the matrices $\mathbf{Z}_{(1)}, \ldots, \mathbf{Z}_{(r)}$ obey the matrix independent component model and have distinct kurtosis row means. Thus the task of finding the mth rotation in TFOBI reduces to that of finding the left rotation in MFOBI. Additionally, (18) shows that the standardization matrices of modes other than m are in the m-flattening of the standardized observations collected to the multiple Kronecker product on the right-hand side both satisfying the assumption $\hat{\mathbf{R}} \to_P \mathbf{I}$ and contributing nothing to the asymptotics of mode m, as shown in the proof of Theorem 5. The limiting distributions for $\hat{\mathbf{\Gamma}}_m$ are thus obtained by applying Theorem 5 into the empirical distributions of $\mathbf{Z}_{(1)}, \ldots, \mathbf{Z}_{(r)}$.

Comparison of the expressions for the two choices of N in Corollary 3 immediately yields the following result.

Corollary 4. *Assume $q > 1$ and denote by $\mathbf{Z}^{kk'}$ the matrix obtained by dropping rows k and k' from $\mathbf{Z}$, $k \neq k'$. Then, for $p > 2$, the choice $N = 1$ is asymptotically superior to the choice $N = 0$ in estimating $\hat{\gamma}_{kk'}$ if and only if the average kurtosis of the elements of $\mathbf{Z}^{kk'}$ is smaller than 2, i.e., when*

$$\frac{1}{p-2} \sum_{m \neq k,k'} \bar{\beta}_{m\bullet} < 2.$$

If $p = 2$ then the methods are asymptotically equivalent regardless of the distribution of $\mathbf{Z}$.

According to Corollary 4, to justify the use of the normed version ($N = 1$) one would have to assume not only one, but several elements of $\mathbf{Z}$ to have kurtosis values below 2. To gain some insight on the strictness of the inequality in Corollary 4, we use the moment inequality of [16] stating that for unimodal distributions with finite fourth moments we have

$$\gamma^2 \leq \beta - \frac{189}{125}.$$

Combining this bound with Corollary 4 then reveals that a necessary condition for the superiority of the normed version is that most elements of $\mathbf{Z}$ must be multimodal or almost symmetric (average squared skewness has to be smaller than 0.488). In the second simulation study of Section 7, we will conduct a comparison of the two versions under different settings but as the condition in Corollary 4 is in general very restrictive and unrealistic the other simulation studies are done using the non-normed versions of MFOBI and TFOBI.

To provide more insight into the second part of Corollary 4 where $p = 2$, recall that the Cayley–Hamilton theorem states that every square matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ is annihilated by its characteristic polynomial [37]. For $p = 2$ this takes the simple form

$$\mathbf{A}^2 - \operatorname{tr}(\mathbf{A})\mathbf{A} + \det(\mathbf{A})\mathbf{I}_2 = \mathbf{0}.$$

Assume now that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a sample of tensors of the same size and that the mth mode of $\mathbf{X}_1$ has length two. Then, $\mathbf{X}_i \odot_{-m} \mathbf{X}_i$ is of size 2×2 for all $i \in \{1, \ldots, n\}$ and we have

$$(\mathbf{X}_i \odot_{-m} \mathbf{X}_i)^2 = \|\mathbf{X}_i\|_F^2 (\mathbf{X}_i \odot_{-m} \mathbf{X}_i) - \det(\mathbf{X}_i \odot_{-m} \mathbf{X}_i)\mathbf{I}_2,$$

where we have utilized the m-flattening, $\operatorname{tr}(\mathbf{X}_i \odot_{-m} \mathbf{X}_i) = \operatorname{tr}(\mathbf{X}_{i(m)}\mathbf{X}_{i(m)}^\top) = \|\mathbf{X}_{i(m)}\|_F^2 = \|\mathbf{X}_i\|_F^2$. Consequently, the sample estimates of $\mathbf{B}_{(m)}^0$ and $\mathbf{B}_{(m)}^1$ in Definition 4 have a difference proportional to the identity matrix, implying that they have the same sets of eigenvectors. Thus for modes of length two the performances of the normed and non-normed version are not only equivalent in the limit, but equivalent for finite samples as well.

6.4. Comparing the limiting efficiencies of the FOBI and TFOBI estimates

As the asymptotic variances in Corollaries 2 and 3 are rather complicated and each of them relates only to a single element of a single matrix, to compare them as a whole a more concise measure of asymptotic accuracy is desired. For this we first review the *minimum distance index* (MDI) [13] computed as

$$\hat{D}_m = D(\hat{\mathbf{\Gamma}}_{(m)}, \mathbf{\Omega}_m) = \frac{1}{\sqrt{p_m - 1}} \inf_{\mathbf{C} \in \mathcal{C}^{p_m}} \|\mathbf{C}\hat{\mathbf{\Gamma}}_{(m)}\mathbf{\Omega}_m - \mathbf{I}_{p_m}\|_F,$$

where $\mathbf{\Omega}_m \in \mathbb{R}^{p_m \times p_m}$ is the true m-mode mixing matrix and $\hat{\mathbf{\Gamma}}_{(m)}$ is the m-mode unmixing matrix estimate. The minimum distance index is a measure of how far away the matrix $\hat{\mathbf{\Gamma}}_{(m)}\mathbf{\Omega}_m$ is from the identity matrix, invariant to order, scales and signs of rows. The index satisfies $0 \leq \hat{D} \leq 1$ with the value 0 indicating that $\hat{\mathbf{\Gamma}}_{(m)} = \mathbf{\Omega}^{-1}$ up to permutation, scaling and sign-change of its rows. The index further obeys the limit result $n(p_m - 1)\hat{D}_m^2 \to_d \mathcal{D}_m$, where $\mathcal{D}_m$ is a distribution with the expected value

$$E_m = \sum_{k=1}^{p_m - 1} \sum_{k'=k+1}^{p_m} \left\{ ASV(\hat{\gamma}_{kk'}^{(m)}) + ASV(\hat{\gamma}_{k'k}^{(m)}) \right\}, \tag{22}$$

where $\hat{\gamma}_{kk'}^{(m)}$ is the (k,k') element of $\hat{\mathbf{\Gamma}}_{(m)}$. Consequently E_m, the sum of asymptotic variances of the off-diagonal elements of $\hat{\mathbf{\Gamma}}_{(m)}$, provides a single-number measure of the asymptotic performance of TFOBI in the mth mode.

However, as FOBI produces only a single number E_1 and TFOBI one for each mode, $E_1, \ldots, E_r$, we still need to somehow combine the latter to allow comparisons between FOBI and TFOBI. Both the FOBI unmixing estimate $\hat{\mathbf{\Gamma}}$ and the Kronecker product $\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)}$ of the TFOBI unmixing estimates estimate the inverse of the same matrix $\mathbf{\Omega} = \mathbf{\Omega}_r \otimes \cdots \otimes \mathbf{\Omega}_1$ and thus the comparison should be done between them. A link connecting the minimum distance indices of the Kronecker product $\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)}$ and its component matrices is given next.

Theorem 6. *Let the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ be generated by the tensor-valued independent component model (8) with identity mixing, $\mathbf{\Omega}_m = \boldsymbol{I}_{p_m}$ for each $m \in \{1, \ldots, r\}$ (in our case also orthogonal mixing suffices, see below). Assume that the unmixing estimates have the limiting Normal distributions $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{\Gamma}}_{(m)} - \boldsymbol{I}_{p_m}) \rightsquigarrow \mathcal{N}(\boldsymbol{0}, \mathbf{\Sigma}_m)$ and denote $p = p_1 \cdots p_r$. Then we have*

$$n(p-1)\hat{D}^2(\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)}, \boldsymbol{I}_p) = \sum_{m=1}^{r} \frac{p}{p_m}\, n(p_m - 1)\hat{D}^2(\hat{\mathbf{\Gamma}}_{(m)}, \boldsymbol{I}_{p_m}) + o_P(1).$$

Proof. By Theorem 1 in [13] the left-hand side of the claim equals

$$\begin{aligned} n\|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)})\|_F^2 + o_P(1) &= n\,\|\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)}\|_F^2 - n\,\|\mathrm{diag}(\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)})\|_F^2 + o_P(1) \\ &= n\prod_{m=1}^{r}\|\hat{\mathbf{\Gamma}}_{(m)}\|_F^2 - n\prod_{m=1}^{r}\|\mathrm{diag}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2 + o_P(1) \\ &= n\prod_{m=1}^{r}\|\hat{\mathbf{\Gamma}}_{(m)}\|_F^2 - n\prod_{m=1}^{r}\left(\|\hat{\mathbf{\Gamma}}_{(m)}\|_F^2 - \|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2\right) + o_P(1). \end{aligned}$$

Focus next on the second product. We have $n\,\|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2 = O_P(1)$, $\|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2 = o_P(1)$ and $\|\hat{\mathbf{\Gamma}}_{(m)}\|_F^2 = p_m + o_P(1)$, meaning that when the product is opened the terms with more than one $\|\mathrm{off}(\cdot)\|_F^2$-term are $o_P(1)$. We are thus left with

$$\sum_{m=1}^{r}\left(n\,\|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2 \prod_{s\neq m}^{r} p_s\right) + o_P(1),$$

and using Theorem 1 in [13] in the other direction, $n\|\mathrm{off}(\hat{\mathbf{\Gamma}}_{(m)})\|_F^2 = n(p_m-1)\hat{D}^2(\hat{\mathbf{\Gamma}}_{(m)}, \mathbf{I}_{p_m}) + o_P(1)$ gives the claim. □

Corollary 5. *Under the assumptions of Theorem 6 the expected value of the limiting distribution of $n(p-1)\hat{D}^2(\hat{\mathbf{\Gamma}}_{(r)} \otimes \cdots \otimes \hat{\mathbf{\Gamma}}_{(1)}, \boldsymbol{I}_p)$ is $\sum_{m=1}^{r}(p/p_m)E_m$, where E_m is as in (22).*

Corollary 5 implies that the comparison between FOBI and TFOBI should be done by comparing the values of E_1^* and $\sum_{m=1}^{r}(p/p_m)E_m$ where E_1^* is the value of (22) for FOBI. These values will later be plotted in the simulations where the orthogonal equivariance of TFOBI guarantees that Corollary 5 holds also when the mixing is orthogonal. Finally, Theorem 6 also provides insight into the general comparison of two arbitrary (transformed) MDI-values, $n(q_1-1)D_1^2$ and $n(q_2-1)D_2^2$. If the respective mixing matrices are of the size $q_1 \times q_1$ and $q_2 \times q_2$ then the quantities $nq_2(q_1-1)D_1^2$ and $nq_1(q_2-1)D_2^2$ are on the same "scale".

7. Simulation studies and a real data example

7.1. On computational issues

Before the simulations we compare the assumptions between MFOBI and first vectorizing and then using FOBI, hereafter referred to just as FOBI. The difference clearly lies in Assumptions V3 and M3, which simply state that MFOBI makes much less assumptions on the kurtosis values. For reasonably large square $p \times p$ matrices, vectorizing and using FOBI roughly squares the amount of constraints needed for MFOBI ($2p$ vs. p^2). However, one has to bear in mind that the nature of the constraints also changes, MFOBI being concerned with the row and column means of kurtoses and FOBI with the individual values.

Table 4: The distributions of the elements of $\mathbf{Z}_i$ in the first simulation. $\mathcal{U}(a,b)$ denotes the continuous uniform distribution from a to b, $Tri(a,b,c)$ the triangular distribution from a to b with the apex located at c, $\mathcal{G}(\alpha,\beta)$ the Gamma distribution with shape α and rate β, $\mathcal{E}(\beta)$ the exponential distribution with rate β and $\mathcal{IG}(\mu,\lambda)$ the inverse Gaussian distribution with mean μ and shape λ.

$\mathcal{U}(-\sqrt{3},\sqrt{3})$	t_{10}	χ^2_3	$\chi^2_{1.5}$
$Tri(-\sqrt{6},\sqrt{6},0)$	$\mathcal{G}(3,\sqrt{3})$	$\mathcal{G}(1.2,\sqrt{1.2})$	$\chi^2_{1.2}$
$\mathcal{N}(0,1)$	$Laplace(0,1/\sqrt{2})$	$\mathcal{E}(1)$	$\mathcal{IG}(1,1)$

Second, the most computationally intensive parts in both FOBI and MFOBI are the eigendecompositions, the computational complexity of finding the eigendecomposition of a $p \times p$ matrix being roughly $O(p^3)$ [32]. Thus assuming again observations of size $p \times p$, MFOBI requires four $O(p^3)$ operations while FOBI needs two $O(p^6)$ operations, a considerable difference with large p. And thirdly, the numbers of estimable parameters are for MFOBI and FOBI $2p^2 - 1$ and p^4, respectively (assuming again that $p = q$).

All the previous issues become even more serious when comparing TFOBI and FOBI: the number of components in FOBI grows exponentially with the order of the tensor while in general TFOBI just has to perform a few more eigendecompositions of much smaller matrices.

All following computations have been made in R [36], especially using the packages `abind` [35], `ICS` [28], `JADE` [26], `MASS` [45] and `tensor` [39]. The implementation of TFOBI and several other tensor methods can be found in the package `tensorBSS` [49].

7.2. Separation performance comparison

In our first simulation we compared the abilities of MFOBI and FOBI to estimate the unmixing matrix and separate the sources. As our setting we chose samples of independent 3×4 observations $\mathbf{Z}_i$, the 12 components of which, depicted in Table 4, were standardized to have zero mean and unit variance. Starting from the top left corner and moving down and right, the kurtoses of the components are 1.8, 2.4, 3, 4, 5, 6, 7, 8, 9, 11, 13 and 18. The sample sizes considered were $n = 1000, 2000, 4000, 8000, \ldots, 256000$. Furthermore, we considered three types of double mixings, $\mathbf{Z}_i \mapsto \mathbf{X}_i = \mathbf{\Omega}_1 \mathbf{Z}_i \mathbf{\Omega}_2^\top$, namely (i) Normal distribution, (ii) uniform distribution and (iii) orthogonal matrices uniform with respect to the Haar measure. In the first two cases appropriate square matrices were created having random elements from $\mathcal{N}(0,1)$ or $\mathcal{U}(-1,1)$ respectively.

We did a total of 2000 replications per setting and as our performance criteria we used the transformed minimum distance indices discussed in the end of Section 6, $n(p_1p_2-1)D(\hat{\mathbf{\Gamma}}_{(2)} \otimes \hat{\mathbf{\Gamma}}_{(1)}, \mathbf{\Omega}_2 \otimes \mathbf{\Omega}_1)$ and $n(p_1p_2-1)D(\hat{\mathbf{\Gamma}}^*_{(1)}, \mathbf{\Omega}_2 \otimes \mathbf{\Omega}_1)$, where $\hat{\mathbf{\Gamma}}^*_{(1)}$ is the FOBI unmixing estimate. The two values directly measure the accuracies of the methods' separation abilities (lower is better) and under orthogonal mixing (under all mixings for the affine equivariant FOBI), when n grows their means will converge to $\sum_{m=1}^{2}(p_1p_2/p_m)E_m$ and E_1^*, respectively; see (22) and Corollary 5. The mean values of the criteria and their limit values are plotted in Figure 1 and we make the following observations.

Contrary to FOBI, the performance of MFOBI indeed depends on the mixing matrix as is shown by the three distinct lines in Figure 1. The separation is easiest for MFOBI when the mixing is orthogonal (because of its orthogonal equivariance orthogonal mixing is equivalent to no mixing at all) and between normal and uniform mixing the separation is slightly easier under the latter. FOBI, while affine equivariant and independent of the choice of mixing, is clearly inferior to MFOBI both with finite samples (the solid lines) and in the limit (the dashed lines). Both curves under orthogonal mixing approach the corresponding limit values, MFOBI faster than FOBI, giving empirical proof on the correctness of the results of Section 6.

7.3. Comparison between the normed and non-normed versions

Our next simulation study compares the two choices of TFOBI functionals, $N \in \{0,1\}$. By Corollary 4 the value of N makes no difference in modes of length two and, guided by the condition in Corollary 4, we consider two settings, both random samples of independent and identically distributed 3×3 matrices, with the elements

$$\begin{pmatrix} \mathcal{N}(0,1) & \mathcal{B}(-1,1) & \mathcal{B}(-1,1) \\ \mathcal{B}(-1,1) & \mathcal{U}(-\sqrt{3},\sqrt{3}) & \mathcal{B}(-1,1) \\ \mathcal{B}(-1,1) & \mathcal{B}(-1,1) & \mathcal{B}(-1,1) \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathcal{B}(-1,1) & \mathcal{N}(0,1) & \mathcal{N}(0,1) \\ \mathcal{N}(0,1) & \mathcal{U}(-\sqrt{3},\sqrt{3}) & \mathcal{N}(0,1) \\ \mathcal{N}(0,1) & \mathcal{N}(0,1) & \mathcal{N}(0,1) \end{pmatrix},$$

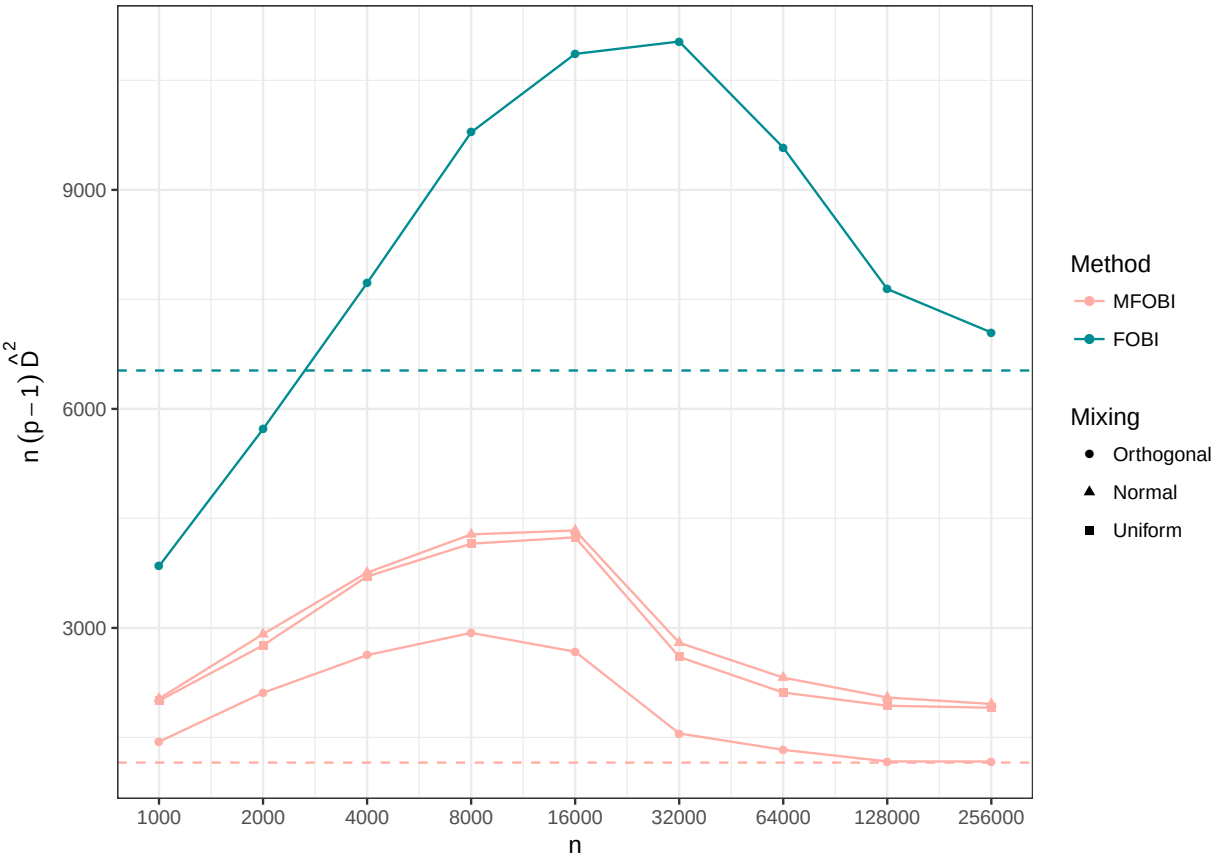

Figure 1: The plot of sample size versus the mean transformed MDI-value with different combinations of method and mixing. The dashed lines give the values of $\sum_{m=1}^{2}(p_1 p_2/p_m)E_m$ and E_1^* towards which the means under orthogonal mixing theoretically converge.

where $\mathcal{N}(0,1)$ is the standardized Normal distribution, $\mathcal{U}(-\sqrt{3},\sqrt{3})$ is the continuous uniform distribution from $-\sqrt{3}$ to $\sqrt{3}$ and $\mathcal{B}(-1,1)$ is the two-point probability distribution taking equally likely each of the values, -1 and 1. The distributions have the respective kurtoses 3, 1.8 and 1 and consequently the condition of Corollary 4 is satisfied for every off-diagonal element in the first setting and is not satisfied for any off-diagonal element in the second setting. Asymptotically the choice $N = 1$ is superior to $N = 0$ in the first setting and vice versa for the second one. To investigate whether this holds also for finite samples, we simulated samples of size $n = 1000, 2000, 4000, 8000, \ldots, 256000$ from the above distributions and applied MFOBI to them in four different forms: using the pairs $(\mathbf{B}_0^L;\mathbf{B}_0^R)$, $(\mathbf{B}_1^L;\mathbf{B}_1^R)$ and the mixed pairs $(\mathbf{B}_0^L;\mathbf{B}_1^R)$ or $(\mathbf{B}_1^L;\mathbf{B}_0^R)$. Intuitively, the performances of the latter two should fall somewhere between those of the former two. To be able to utilize our asymptotic results we did not mix the observations (which is equivalent to using orthogonal mixing).

We again used the minimum distance index as a criterion and the resulting mean transformed MD-indices over 2000 replications are shown in Figure 2. The dashed lines in the plot indicate the limiting expected values computed using the results of Section 6, toward which the solid lines theoretically converge. We have not visually distinguished the limit lines from each other as their order is the same as the order of the empirical lines. The symmetry of the simulated matrices causes the two mixed MFOBI-functionals to have the same limiting values and similar behavior is also visible in the corresponding two empirical lines matching each other closely. Further observations include: The empirical lines approach the limits rather nicely, with some swaying in the setting where the condition is not satisfied. The setting where the condition is satisfied is overall more easily separated (the lines are lower in the plot). Finally, the ordering between the methods is consistent throughout the study and under both settings the two mixed cases are located almost halfway between the non-mixed cases. Despite the success of the choice $N = 1$ here, based on the

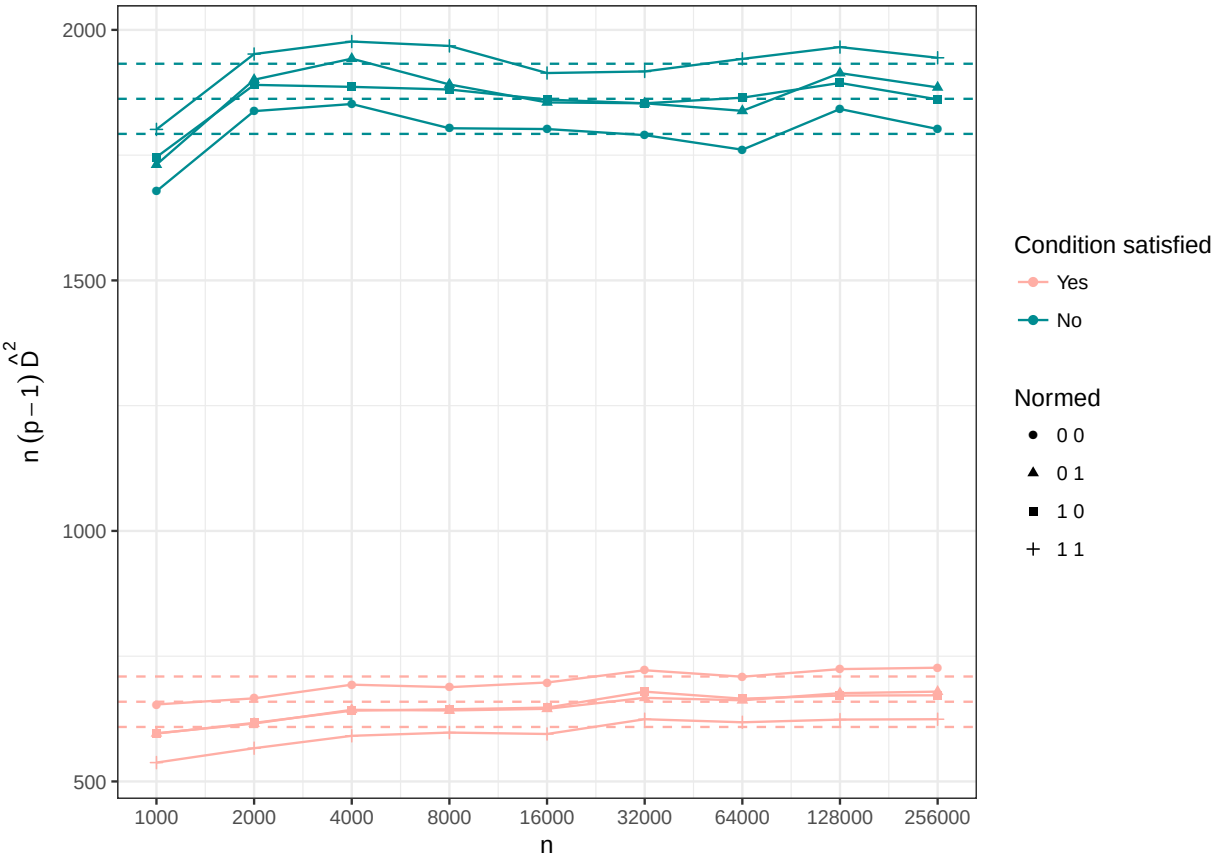

Figure 2: The plot of sample size versus the mean transformed MDI-value with different combinations of setting and $N \in \{0, 1\}$. The value of "Normed" tells which value of N was used for the left and right unmixing matrices, e.g., 1 0 means that the left unmixing matrix used the normed version but the right one did not. The dashed lines give the values of $\sum_{m=1}^{2}(p_1 p_2/p_m)E_m$ towards which the means theoretically converge.

extreme measures that were required to create a setting where the condition of Corollary 4 is satisfied (we needed to resort to the transformed Bernoulli-distribution $\mathcal{B}(-1, 1)$, the probability distribution with the lowest possible kurtosis) we still choose to advocate using primarily the case $N = 0$.

7.4. FOBI and TFOBI in classification

Traditionally, although not consistent with the model assumptions, ICA methods are often used as a preprocessing step for classification as linear combinations of the variables with high or low kurtosis are often the most informative in this sense. Peña et al. [33], for example, used FOBI to reveal cluster structures in the data. Also, Tyler et al. [43] showed that two scatter matrices can be combined to estimate Fisher's linear subspace in the case of mixtures of elliptical distributions with proportional covariance matrices. Following the interpretation of FOBI and TFOBI as a combination of different scatters we compare in this section FOBI and TFOBI for the purposes of classification.

The comparison was done in the following set-up. For each replication we simulated 500 observations of $5 \times 5 \times 5$ tensors $\mathbf{X}_i$ belonging to one of two groups. In Group 1 all elements of the observations $\mathbf{X}_i$ are sampled from independent $\mathcal{N}(0, 1)$-distributions, while in Group 2 the front upper left $2 \times 2 \times 2$ corner has elements sampled from independent $\mathcal{N}(2, 1)$-distributions (and the rest of the elements from $\mathcal{N}(0, 1)$). A proportion π of all observations belonged to Group 2.

We did 2000 replications for each of the values $\pi \in \{0.10, 0.15, \ldots, 0.50\}$ and for each replication we mixed the observations from all three m-modes using the same three types of mixing matrices as in the previous section. Next, we divided the transformed data randomly into training and test sets, with the respective sizes of 400 and 100. Both

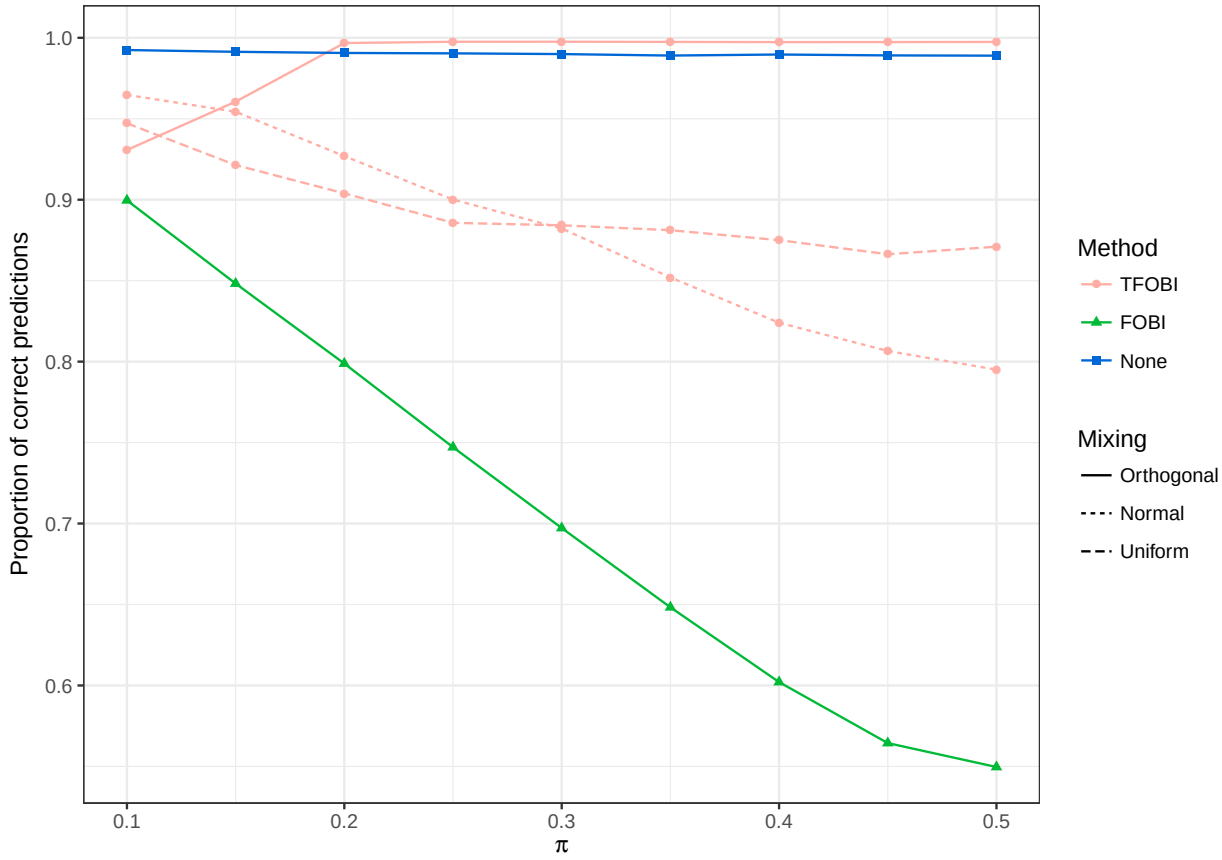

Figure 3: Proportions of correct classifications as a function of proportion π with FOBI and TFOBI as pre-steps and with three types of random mixing matrices.

TFOBI and FOBI were then carried out for the training data and linear discriminant analysis (LDA) was used to create classification rules based on certain selected components. For TFOBI we chose these to be the corner components $z_{1,1,1}$ and $z_{5,5,5}$ and the components having the highest and lowest kurtoses. For FOBI we simply chose the first two and the last two components (ordered according to kurtosis). As a reference, we also created a classification rule with LDA using all the original components. The means of the proportions of correct predictions in the test set for each of the rules are plotted in Figure 3. The reference value is included as the line "NONE".

LDA uses the training set group proportions as a prior and a "baseline" proportion of correct predictions is thus $1 - \pi$, corresponding to classifying all test observations to the dominant group. The plot indicates that FOBI cannot find the direction separating the groups in any systematic way and is actually no better than the baseline. TFOBI, in contrast, is in every case better than FOBI and performs very nicely under all mixings (especially orthogonal). Under orthogonal mixing and for π larger than or equal to 0.20 TFOBI, being able to filter out the noise, is also slightly better than using all the original components. The simulation thus implies that TFOBI provides a reliable way of extracting the separating variables from tensor-valued data.

7.5. Real data example

To see how MFOBI works with real data we use the *semeion*[1] data set available from the UCI Machine Learning Repository [21]. The data consist of 1593 scanned handwritten digits written by 80 persons represented as binary

[1]Semeion Research Center of Sciences of Communication, via Sersale 117, 00128 Rome, Italy; Tattile Via Gaetano Donizetti, 1-3-5,25030 Mairano (Brescia), Italy.

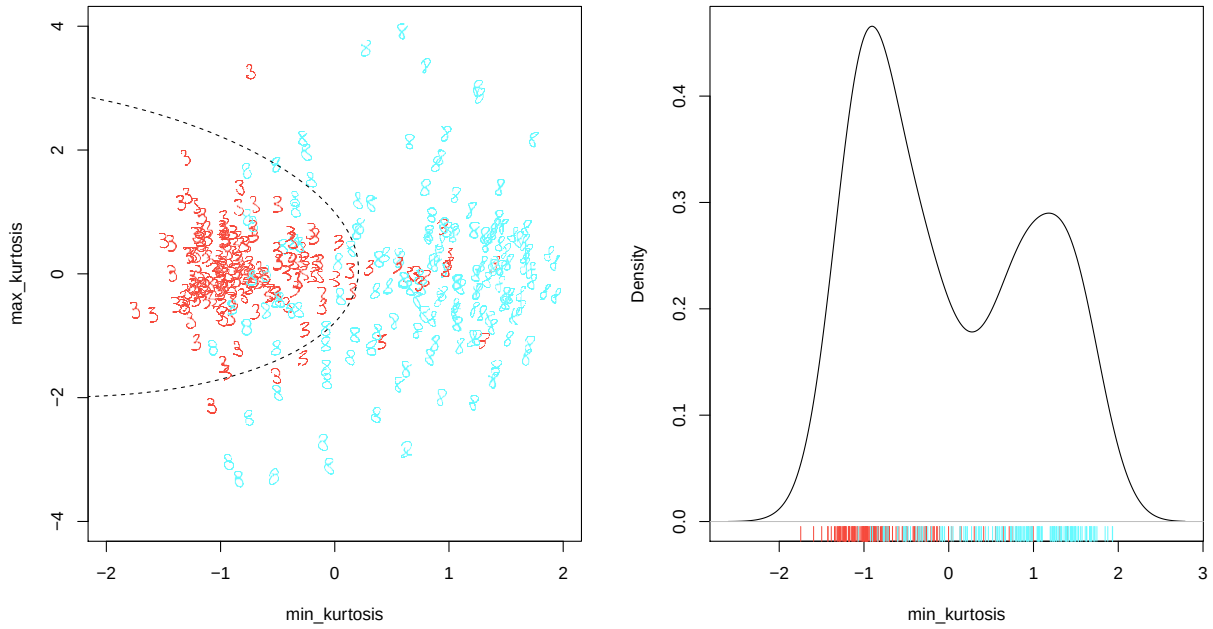

Figure 4: The figure on the left-hand side shows the scatter plot of the two independent components having the lowest and highest kurtoses, dividing the data nicely into two groups. The separation is also visible on the right-hand side in the rug and the bimodal kernel density estimate of the component with the lowest kurtosis.

16×16 matrices. For our analysis we picked only the images of the visually similar digits 3 and 8 hoping to find a direction separating the two digits. The number of observations is then $n = 314$ with almost equal number of threes and eights (159 and 155, respectively).

The results of MFOBI are shown in Figure 4. The scatter plot on the left shows the distributions of the components having the highest and lowest kurtoses ($z_{1,2}$ and $z_{16,16}$, respectively), with the individual images as plotting markers, along with the decision boundary given by quadratic discriminant analysis. Although the two groups of digits overlap a bit the separation is still very clear, as is evidenced also by the kernel density estimate of the minimal kurtosis component on the right-hand side of Figure 4. We also see that the hand-writing is slanting more and more to the right with increasing values of $z_{16,16}$ and that the variable $z_{1,2}$ with highest kurtosis can be used in search for outliers.

For comparison, we also tried applying regular FOBI to the vectorized data with somewhat disappointing results; the covariance matrix of the full data was not invertible and when trying with some subsets of the data, FOBI succeeded only in finding a few outliers.

8. Concluding remarks

In this paper, we presented methods of independent component analysis for matrix- and tensor-valued observations called MFOBI and TFOBI. The total procedure can be seen as a simultaneous application of the classic FOBI on all m-modes of the observed tensors.

Apart from the algorithms and two different ways of estimating the unmixing matrix, we provided the asymptotic variances of the elements of the unmixing matrix estimates in the case of orthogonal mixing. The variance expressions then show that using the non-normed version of TFOBI is in most cases the preferable approach. Regarding the comparison of TFOBI with the often used combination of vectorizing and FOBI, we first stated that the numbers of estimable parameters and assumptions required are of much smaller order in MFOBI and TFOBI. This is because they

are able to exploit the possible tensor structure in the estimation. Next, simulations were used to show TFOBI's superiority to FOBI also in practice, both in estimating the unmixing matrix and as a preprocessing step for discriminant analysis.

With MFOBI and TFOBI being derivatives of FOBI a reasonable conjecture is that, instead of relying on the kurtosis matrices $\mathbf{B}^N$, extending some other standard ICA techniques like projection pursuit or JADE [3] into the tensor case would lead into better estimates. [46] showed that this holds for JADE and some preliminary investigation shows that this is indeed the case for projection pursuit as well and such a take on the problem can then be seen as a tensor version of FastICA [11]. The resulting concept of tensorial projection pursuit will be addressed in future work.

Nevertheless, compared with other perhaps more sophisticated routes of generalization, the FOBI-type extensions enjoy a particularly simple structure for high-dimensional tensors: the higher moment tensors decompose neatly to matrices of reasonably low dimensions. As a result the eigendecompositions only need to be performed on $p_m \times p_m$ matrices individually. This feature makes MFOBI and TFOBI especially attractive when applied on a large scale.

Acknowledgments. First, the authors would like to thank the reviewers for their valuable comments and suggestions. Second, the authors would like to express their sincere gratitude to the Editor, Professor Genest, for the wonderful service he provided us throughout the publication process. The research of Joni Virta, Klaus Nordhausen and Hannu Oja was partially supported by the Academy of Finland Grant 268703. The research of Bing Li was partially supported by the National Science Foundation Grant DMS-1407537.

References

[1] C.F. Beckmann, S.M. Smith, Tensorial extensions of independent component analysis for multisubject FMRI analysis, Neuroimage 25 (2005) 294–311.

[2] J.-F. Cardoso, Source separation using higher order moments, In: International Conference on Acoustics, Speech, and Signal Processing 1989, pp. 2109–2112. IEEE, 1989.

[3] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, In: IEE Proceedings F (Radar and Signal Processing), vol. 140, pp. 362–370, IET, 1993.

[4] S. Ding, R.D. Cook, Dimension folding PCA and PFC for matrix-valued predictors, Statist. Sinica 24 (2014) 463–492

[5] S. Ding, R.D. Cook, Higher-order sliced inverse regressions, Wiley Interdisciplinary Reviews: Comput. Statist. 7 (2015) 249–257.

[6] S. Ding, R.D. Cook, Tensor sliced inverse regression, J. Multivariate Anal. 133 (2015) 216–231.

[7] K. Greenewald, A. Hero, Robust Kronecker product PCA for spatio-temporal covariance estimation, IEEE Trans. Signal Proc. 63 (2015) 6368–6378.

[8] A. Gupta, D. Nagar, Matrix Variate Distributions, Chapman & Hall/CRC, Boca Raton, FL, 2010.

[9] H. Hung, C.-C. Wang, Matrix variate logistic regression model with application to EEG data, Biostatistics 14 (2013) 189–202.

[10] H. Hung, P. Wu, I. Tu, S. Huang, On multilinear principal component analysis, Biometrika 99 (2012) 569–583.

[11] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[12] P. Ilmonen, J. Nevalainen, H. Oja, Characteristics of multivariate distributions and the invariant coordinate system, Statist. Probab. Lett. 80 (2010) 1844–1853.

[13] P. Ilmonen, K. Nordhausen, H. Oja, E. Ollila, A new performance index for ICA: Properties, computation and asymptotic analysis, In: Latent Variable Analysis and Signal Separation, pp. 229–236, Springer, New York, 2010.

[14] P. Ilmonen, H. Oja, R. Serfling, On invariant coordinate system (ICS) functionals, Internat. Statist. Rev. 80 (2012) 93–110.

[15] H.-J. Kim, E. Ollila, V. Koivunen, C. Croux, Robust and sparse estimation of tensor decompositions, In: Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE, pp. 965–968, IEEE, 2013.

[16] C.A. Klaassen, P.J. Mokveld, B. Van Es, Squared skewness minus kurtosis bounded by 186/125 for unimodal distributions, Statist. Probab. Lett. 50 (2000) 131–135.

[17] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (2009) 455–500.

[18] L.D. Lathauwer, B.D. Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (2000) 1253–1278.

[19] B. Li, M.K. Kim, N. Altman, On dimension folding of matrix- or array-valued statistical objects, Ann. Statist. 38 (2010) 1094–1121.

[20] K.-C. Li, Sliced inverse regression for dimension reduction, J. Amer. Statist. Assoc. 86 (1991) 316–327.

[21] M. Lichman, UCI Machine Learning Repository, 2013.

[22] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, A survey of multilinear subspace learning for tensor data, Pattern Recognition 44 (2011) 1540–1551.

[23] A.M. Manceur, P. Dutilleul, Maximum likelihood estimation for the tensor Normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion, J. Comput. Appl. Math. 239 (2013) 37–49.

[24] P. McCullagh, Tensor Methods in Statistics, Chapman & Hall, New York, 1987.

[25] J. Miettinen, K. Nordhausen, H. Oja, S. Taskinen, Deflation-based FastICA with adaptive choices of nonlinearities, IEEE Trans. Signal Proc. 62 (2014) 5716–5724.

[26] J. Miettinen, K. Nordhausen, S. Taskinen, Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp, J. Statist. Software 76, 2017.

[27] J. Miettinen, S. Taskinen, K. Nordhausen, H. Oja, Fourth moments and independent component analysis, Statist. Sci. 30 (2015) 372–390.

[28] K. Nordhausen, H. Oja, D.E. Tyler, Tools for exploring multivariate data: the package ICS, J. Statist. Software 28 (2008) 1–31.
[29] K. Nordhausen, H. Oja, D.E. Tyler, Asymptotic and bootstrap tests for subspace dimension, *arXiv preprint arXiv:1611.04908*, 2016.
[30] K. Nordhausen, H. Oja, D.E. Tyler, J. Virta, Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace, IEEE Signal Proc. Lett. 2017.
[31] M. Ohlson, M.R. Ahmad, D. von Rosen, The multilinear Normal distribution: Introduction and some basic properties, J. Multivariate Anal. 113 (2013) 37–47.
[32] V.Y. Pan, Z. Chen, A. Zheng, The complexity of the algebraic eigenproblem, Mathematical Sciences Research Institute, pp. 1998–71, 1998.
[33] D. Peña, F. Prieto, J. Viladomat, Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure, J. Multivariate Anal. 101 (2010) 1995–2007.
[34] R.M. Pfeiffer, L. Forzani, E. Bura, Sufficient dimension reduction for longitudinally measured predictors, Statist. Medicine 31 (2012) 2414–2427.
[35] T. Plate, R. Heiberger, ABIND: Combine multidimensional arrays, 2015, R Package Version 1.4-3.
[36] R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[37] S. Roman, Advanced Linear Algebra, vol. 3, Springer, 2005.
[38] B. Ros, F. Bijma, J.C. de Munck, M.C. de Gunst, Existence and uniqueness of the maximum likelihood estimator for models with a Kronecker product covariance structure, J. Multivariate Anal. 143 (2016) 345–361.
[39] J. Rougier, TENSOR: Tensor product of arrays, 2012, R Package Version 1.5.
[40] J.R. Schott, Tests for Kronecker envelope models in multilinear principal component analysis, Biometrika 101 (2014) 978–984.
[41] M.S. Srivastava, T. von Rosen, D. von Rosen, Models with a Kronecker product covariance structure: estimation and testing, Math. Methods Statist. 17 (2008) 357–370.
[42] Y. Sun, P. Babu, D. Palomar, Robust estimation of structured covariance matrix for heavy-tailed distributions, In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, pp. 5693–5697, 2015.
[43] D.E. Tyler, F. Critchley, L. Dümbgen, H. Oja, Invariant co-ordinate selection, J. Roy. Statist. Soc. Ser. B 71 (2009) 549–592.
[44] M.A.O. Vasilescu, D. Terzopoulos, Multilinear independent components analysis, In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, vol. 1, pp. 547–553, IEEE, 2005.
[45] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, 4th Ed. Springer, New York, 2002.
[46] J. Virta, B. Li, K. Nordhausen, H. Oja, JADE for tensor-valued observations, Preprint in arXiv:1603.05406, 2016.
[47] J. Virta, K. Nordhausen, Blind source separation of tensor-valued time series, Signal Proc. 141 (2017) 204–216.
[48] J. Virta, K. Nordhausen, H. Oja, Joint use of third and fourth cumulants in independent component analysis, arXiv preprint arXiv:1505.02613, 2015.
[49] J. Virta, K. Nordhausen, H. Oja, B. Li, tensorBSS: Blind Source Separation Methods for Tensor-Valued Observations, 2016. R Package Version 0.3.
[50] J. Virta, S. Taskinen, K. Nordhausen, Applying fully tensorial ICA to fMRI data, In: Signal Processing in Medicine and Biology Symposium (SPMB), 2016 IEEE, pp. 1–6, IEEE, 2016.
[51] K. Werner, M. Jansson, P. Stoica, On estimation of covariance matrices with Kronecker product structure, IEEE Trans. Signal Proc. 56 (2008) 478–491.
[52] A. Wiesel, Geodesic convexity and covariance estimation, IEEE Trans. Signal Proc. 60 (2012) 6182–6189.
[53] Y. Xue, X. Yin, Sufficient dimension folding for regression mean function, J. Comput. Graph. Statist. 23 (2014) 1028–1043.
[54] P. Zeng, W. Zhong. Dimension reduction for tensor classification, Topics Appl. Statist. 55 (2013) 213–227.
[55] L. Zhang, Q. Gao, L. Zhang, Directional independent component analysis with tensor representation, In: IEEE Conference on Computer Vision and Pattern Recognition 2008, pp. 1–7, IEEE, 2008.
[56] J. Zhao, C. Leng, Structured lasso for regression with matrix covariates, Statist. Sinica 24 (2014) 799–814.
[57] W. Zhong, X. Xing, K. Suslick, Tensor sufficient dimension reduction, Wiley Interdisciplinary Reviews: Comput. Statist. 7 (2015) 178–184.
[58] H. Zhou, L. Li, Regularized matrix regression, J. Roy. Statist. Soc. Ser. B 76 (2014) 463–483.
[59] H. Zhou, L. Li, H. Zhu, Tensor regression with applications in neuroimaging, J. Amer. Statist. Assoc. 108 (2013) 540–552.