JADE for Tensor-Valued Observations

Joni Virta, Bing Li, Klaus Nordhausen and Hannu Oja

Abstract

Independent component analysis is a standard tool in modern data analysis and numerous different techniques for applying it exist. The standard methods however quickly lose their effectiveness when the data are made up of structures of higher order than vectors, namely matrices or tensors (for example, images or videos), being unable to handle the high amounts of noise. Recently, an extension of the classic fourth order blind identification (FOBI) specially suited for tensor-valued observations was proposed and showed to outperform its vector version for tensor data. In this paper we extend another popular independent component analysis method, the joint approximate diagonalization of eigen-matrices (JADE), for tensor observations. In addition to the theoretical background we also provide the asymptotic properties of the proposed estimator and use both simulations and real data to show its usefulness and superiority over its competitors.

Keywords: Independent component analysis, multilinear algebra, kurtosis, limiting normality, minimum distance index.

J. Virta, K. Nordhausen and H. Oja are with the Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland (e-mail: joni.virta@utu.fi).

B. Li is with the Department of Statistics, Pennsylvania State University, 326 Thomas Building, University Park, Pennsylvania 16802, USA.

1 Introduction

In the following “tensor” is used to refer to an array in $\mathbb{R}^{p_1 \times \cdots \times p_r}$ and before the actual ideas are described we first review some key properties of tensors and matrices needed later.

A tensor of rth order $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ can be seen as a higher order analogy of vectors and matrices. Whereas a matrix can be viewed either as a collection of rows or that of columns, a tensor of rth order has in total r *modes*. The *m-mode vectors* of a tensor are given by letting the mth index vary while keeping all other indices fixed, $m = 1, \ldots, r$. A tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ thus contains $\rho_m := \Pi_{s \neq m}^r p_s$ m-mode vectors of length p_m. The opposite construct, fixing a single index i_m and varying the others, then gives what we call the *m-mode faces* of a tensor. The number of m-mode faces then totals p_m and each is a tensor of size $p_1 \times \cdots \times p_{m-1} \times p_{m+1} \times \cdots \times p_r$.

For representing tensor contraction we use the Einstein summation convention in which a twice-appearing index in a product implies summation over the range of the index. For example, for a tensor $\mathbf{X} = \{x_{i_1 i_2 i_3}\}$ we have

$$x_{i_1 i_2 j} x_{i_1 i_2 k} := \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} x_{i_1 i_2 j} x_{i_1 i_2 k}.$$

Two special cases of tensor contraction prove especially useful for us. The product $\mathbf{X} \odot_m \mathbf{A}$ of a tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ with a matrix $\mathbf{A} \in \mathbb{R}^{p_m \times p_m}$, $m = 1, \ldots, r$, is defined as the $p_1 \times \cdots \times p_r$-dimensional tensor with the elements

$$(\mathbf{X} \odot_m \mathbf{A})_{i_1 \cdots i_r} = x_{i_1 \cdots i_{m-1} j_m i_{m+1} \cdots i_r} a_{i_m j_m}. \tag{1}$$

That is, the multiplication $\mathbf{X} \odot_m \mathbf{A}$ linearly transforms $\mathbf{X}$ from the direction of the mth mode without changing the size of the tensor. The operation can alternatively be viewed as applying the linear transformation given by $\mathbf{A}$ separately to each m-mode vector of the tensor. The second useful product, $\mathbf{X} \odot_{-m} \mathbf{Y}$, of two tensors of the same size, $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ is defined as the $p_m \times p_m$-dimensional matrix with the elements

$$(\mathbf{X} \odot_{-m} \mathbf{Y})_{jk} = x_{i_1 \cdots i_{m-1} j i_{m+1} \cdots i_r} y_{i_1 \cdots i_{m-1} k i_{m+1} \cdots i_r}. \tag{2}$$

The special case $\mathbf{X} \odot_{-m} \mathbf{X}$ provides higher order counterparts for the products of a vector $\mathbf{x} \in \mathbb{R}^{p_1}$ or a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ with itself, such as $\mathbf{x}\mathbf{x}^T$, $\mathbf{X}\mathbf{X}^T$ or $\mathbf{X}^T\mathbf{X}$, and proves useful in defining the “covariance matrix” of a tensor.

Finally, define the vectorization $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{p_1 \cdots p_r}$ of a tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ as the stacking of the elements $x_{i_1 \cdots i_r}$ in such a way that the leftmost index goes through its cycle the quickest and the rightmost index the slowest. Then it holds for a tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ and matrices $\boldsymbol{A}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \boldsymbol{A}_r \in \mathbb{R}^{p_r \times p_r}$ that

$$\mathrm{vec}(\mathbf{X} \odot_1 \boldsymbol{A}_1 \cdots \odot_r \boldsymbol{A}_r) = (\mathbf{A}_r \otimes \cdots \otimes \mathbf{A}_1)\mathrm{vec}(\mathbf{X}),$$

where $\otimes$ is the Kronecker product.

In this paper we assume that the tensor-valued i.i.d. random elements $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, $i = 1, \ldots, n$, are observed from the recently suggested (Virta et al., 2016) *tensor independent component (IC) model*:

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{Z} \odot_1 \boldsymbol{\Omega}_1 \cdots \odot_r \boldsymbol{\Omega}_r, \tag{3}$$

where $\boldsymbol{\Omega}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \boldsymbol{\Omega}_r \in \mathbb{R}^{p_r \times p_r}$ are full rank *mixing matrices*, $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ is the location center, and $\mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ is an unobserved random tensor. The model (3) is further equipped with the following assumptions.

Assumption 1. *The components of $\boldsymbol{Z}$ are mutually independent.*

Assumption 2. *The components of $\boldsymbol{Z}$ are standardized in the sense that $E[\mathrm{vec}(\boldsymbol{Z})] = \boldsymbol{0}$ and $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Z})] = \mathbf{I}$.*

Assumption 3. *For each $m = 1, \ldots, r$, at most one m-mode face of $\boldsymbol{Z}$ consists entirely of Gaussian components.*

Assumption 2 implies that $E[\mathbf{X}] = \boldsymbol{\mu}$ and that

$$\mathrm{Cov}[\mathrm{vec}(\mathbf{X})] = (\boldsymbol{\Omega}_r \boldsymbol{\Omega}_r^T) \otimes \cdots \otimes (\boldsymbol{\Omega}_1 \boldsymbol{\Omega}_1^T)$$

has the so-called Kronecker structure. Assumption 3 is a tensor analogy for the usual vector independent component model assumption on maximally one Gaussian component and without it some column blocks of some of the matrices $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_r$ could be identifiable only up to a rotation. After the above assumptions we can still freely change the signs and orders of the columns of all $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_r$, or multiply any $\boldsymbol{\Omega}_s$ by a constant and divide any $\boldsymbol{\Omega}_t$ by the same constant, but this indeterminacy is acceptable in practice. The model along with its assumptions now provides a natural extension for the standard independent component model which is obtained as a special case when $r = 1$.

Alternatively, the model can be seen as an extension of the general location-scatter model for tensor-valued data, which is equivalent to (3) with only Assumption 2 and is often, for $r = 1, 2$, combined with the assumption on Gaussianity or sphericity of $\mathrm{vec}(\mathbf{Z})$. Under the location-scatter model the covariance matrix of $\mathrm{vec}(\mathbf{X})$ again has the above Kronecker structure. In addition to requiring less parameters to estimate than a full $p_1 \cdots p_r \times p_1 \cdots p_r$ covariance matrix, the assumption on Kronecker structure is a natural choice in many applications, see e.g. Werner et al. (2008). One particular example is multivariate repeated measures data where the observations are matrices with each row coinciding to one of a set of p_1 variables and the columns correspond to the p_2 time points on which the variables are measured. In that case the matrix $\mathbf{\Omega}_1$ in (3) specifies the covariance structure between the variables and $\mathbf{\Omega}_2$ the covariance structure between the time points. For the estimation of covariance parameters under the assumption on Kronecker structure in the matrix case, $r = 2$, see Srivastava et al. (2008); Wiesel (2012); Sun et al. (2015). For the general tensor Gaussian distribution and the estimation of its parameters see Hoff et al. (2011).

The extension of dimension reduction methods from vector to matrix or tensor observations is in signal processing usually done via tensor decompositions such as the CP-decomposition and the Tucker decomposition. A review of them with a plethora of references for applications is given in Kolda and Bader (2009), see also Lu et al. (2011) for more applications. For examples of particular dimension reduction methods incorporating matrix or tensor predictors, see e.g. Vasilescu and Terzopoulos (2005); Zhang et al. (2008); Virta et al. (2016) for independent component analysis, Li et al. (2010); Pfeiffer et al. (2012); Xue and Yin (2014); Ding and Cook (2015) for sufficient dimension reduction and Ding and Cook (2014); Greenewald and Hero (2015) for principal components analysis-based techniques. More references are also given in Li et al. (2010); Virta et al. (2016).

In tensor independent component analysis the objective is to estimate, based on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, some unmixing matrices $\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_r$ such that $\mathbf{X} \odot_1 \mathbf{\Phi}_1 \cdots \odot_r \mathbf{\Phi}_r$ has mutually independent components. A naïve method for accomplishing this would be to vectorize the observations and resort to some standard method of independent component analysis, but in doing so the resulting estimate lacks the desired Kronecker structure. In addition, vectorizing and using standard tools meant for vector-valued data requires the stronger, component-wise version of Assumption 3, inflates the number of parameters and can make the dimension of the data too large for standard methods to handle. To circumvent this, Vasilescu and Terzopoulos (2005); Zhang et al. (2008); Virta et al. (2016) proposed estimating an unmixing matrix separately for each of the modes and Virta et al. (2016) presented an extension of the classic fourth order blind identification (FOBI) (Cardoso, 1989) for tensor observations called TFOBI.

In the vector independent component model, $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}$, the standardized vector $\mathbf{x}_{st} := \mathrm{Cov}[\mathbf{x}]^{-1/2}(\mathbf{x} - E[\mathbf{x}])$ equals $\mathbf{U}\mathbf{z}$ for some orthogonal matrix $\mathbf{U}$. In FOBI the rotation $\mathbf{U}$ is then found using the eigendecomposition of the matrix of fourth moments $\mathbf{B} := E[\mathbf{x}_{st}\mathbf{x}_{st}^T\mathbf{x}_{st}\mathbf{x}_{st}^T]$. This same approach is taken in TFOBI by performing both steps of the procedure, the standardization and the rotation, on all r modes of $\mathbf{X}$. Assuming centered $\mathbf{X}$, in Virta et al. (2016) the m-mode covariance matrices,

$$\boldsymbol{\Sigma}_m(\mathbf{X}) := \rho_m^{-1} E\left[\mathbf{X} \odot_{-m} \mathbf{X}\right], \quad m = 1, \ldots, r, \tag{4}$$

are first used to standardize the observations as $\mathbf{X}_{st} := \mathbf{X} \odot_1 \boldsymbol{\Sigma}_1^{-1/2} \cdots \odot_r \boldsymbol{\Sigma}_r^{-1/2}$. The tensor $\mathbf{Z}$ is then found by rotating $\mathbf{X}_{st}$ from all r modes and the rotation matrices can be found from the eigendecompositions of the m-mode matrices of fourth moments:

$$\mathbf{B}_m := \rho_m^{-1} E\left[(\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st})(\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st})\right].$$

Another widely used independent component analysis method for vector-valued data, called the joint approximate diagonalization of eigen-matrices (JADE) (Cardoso and Souloumiac, 1993), also uses fourth moments to estimate the required final rotation but utilizes them in the form of cumulant matrices (assuming $E\left[\mathbf{x}\right] = \mathbf{0}$),

$$\begin{aligned} \mathbf{C}^{ij}(\mathbf{x}) := {} & E\left[x_i x_j \cdot \mathbf{x}\mathbf{x}^T\right] - E[x_i x_j] E\left[\mathbf{x}\mathbf{x}^T\right] \\ & - E\left[x_i \cdot \mathbf{x}\right] E\left[x_j \cdot \mathbf{x}^T\right] - E\left[x_j \cdot \mathbf{x}\right] E\left[x_i \cdot \mathbf{x}^T\right]. \end{aligned} \tag{5}$$

The final rotation from $\mathbf{x}_{st}$ to $\mathbf{z}$ is in JADE obtained by jointly diagonalizing the matrices

$$\mathbf{C}^{ij}(\mathbf{x}_{st}) = E\left[x_{st,i} x_{st,j} \cdot \mathbf{x}_{st}\mathbf{x}_{st}^T\right] - \delta_{ij}\mathbf{I} - \mathbf{E}^{ij} - \mathbf{E}^{ji}, \tag{6}$$

where $\mathbf{E}^{ij}$ is a matrix with a single one as element (i, j) and zeroes elsewhere and δ_{ij} is the Kronecker delta. Compared to FOBI which only uses $p(p + 1)/2$ sums of fourth joint moments of $\mathbf{x}_{st}$, JADE thus has a clear advantage in using all possible fourth joint cumulants of $\mathbf{x}_{st}$ in the estimation of the rotation matrix.

Because of the well-known fact that JADE outperforms FOBI in most cases (see e.g. Miettinen et al. (2015)) it is natural to expect that the extension of JADE to tensor-valued data would similarly be superior to TFOBI. This is indeed the case, and in the following sections we formulate the tensor joint diagonalization of eigen-matrices (TJADE) which is obtained from JADE by applying very much the same extensions as required when moving from FOBI to TFOBI. We first briefly discuss the standard vector-valued independent component model and review the theory and assumptions behind the original JADE in Section 2. The corresponding aspects of TJADE are presented in Section 3 and the asymptotical properties of both methods in Section 4. Simulations comparing TJADE to TFOBI and both the original JADE and original FOBI are presented in Section 5 along with a real data example and we close in Section 6 with some discussion. The proofs can be found in Appendix A.

2 Original JADE

The original JADE assumes that the vector-valued observations are generated by the vector independent component model

$$\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_i, \quad i = 1, \ldots, n, \tag{7}$$

where the mixing matrix $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ has full rank, $\boldsymbol{\mu} \in \mathbb{R}^p$ and the i.i.d. random vectors $\mathbf{z}_i \in \mathbb{R}^p$ have mutually independent components standardized to have zero means and unit variances. To ensure the existence of the JADE solution we must further assume that at most one of the components of $\mathbf{z}$ has zero excess kurtosis (Cardoso and Souloumiac, 1993).

Assuming next that the data are centered, that is, $E[\mathbf{x}] = \mathbf{0}$, we standardize the vectors as $\mathbf{x}_{st} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$. The standardized vectors can be shown to satisfy $\mathbf{x}_{st} = \mathbf{U}\mathbf{z}$ for some orthogonal matrix $\mathbf{U}$, see Cardoso and Souloumiac (1993). To estimate $\mathbf{U}$, JADE uses the cumulant matrices $\mathbf{C}^{ij}(\mathbf{x}_{st})$, $i, j = 1, \ldots, p$, in (6). Under the independent component model the cumulant matrices can be shown to satisfy, for all $i, j = 1, \ldots, p$,

$$\mathbf{C}^{ij}(\mathbf{x}_{st}) = \mathbf{U} \left(\sum_{k=1}^{p} u_{ik} u_{jk} \kappa_k \mathbf{E}^{kk} \right) \mathbf{U}^T, \tag{8}$$

where $\kappa_k := E(z_k^4) - 3$, the excess kurtosis of the kth component, and u_{ab} are the components of $\mathbf{U}$. The expression in (8) is the eigendecomposition of $\mathbf{C}^{ij}(\mathbf{x}_{st})$ and while the rank of any single matrix $\mathbf{C}^{ij}(\mathbf{x}_{st})$ may not be large enough to estimate $\mathbf{U}$, JADE instead simultaneously (approximately) diagonalizes them all, that is, finds $\mathbf{U}^T$ as

$$\mathbf{U}^T = \underset{\mathbf{U}:\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}{\operatorname{argmax}} \sum_{i=1}^{p} \sum_{j=1}^{p} \| \mathrm{diag}(\mathbf{U}\mathbf{C}^{ij}(\mathbf{x}_{st})\mathbf{U}^T) \|^2. \tag{9}$$

This has the advantage of considering all fourth joint cumulants in the estimation. Optimization problems of type (9) are called joint diagonalization problems, see e.g. Cardoso and Souloumiac (1993); Bunse-Gerstner et al. (1993); Cardoso and Souloumiac (1996); Belouchrani et al. (1997).

In Miettinen et al. (2015) a thorough analysis of the statistical properties of JADE is given and the authors show the JADE estimator is an *independent component functional*, that is, the resulting components are invariant up to sign-change and permutation under affine transformations, even outside the independent component model. See also Moreau (2001) who discusses the higher-order extensions of cumulant-based joint diagonalization methods for blind source separation.

3 Tensor JADE

In formulating TJADE we assume that the data are generated by the tensor IC model (3) and satisfy Assumptions 1, 2 and 3. Assuming $E[\mathbf{X}] = \mathbf{0}$, we next go separately through the tensor analogies of the standardization and rotation steps of the original JADE.

3.1 Standardization step

We take the same approach for standardization of $\mathbf{X}$ as in Virta et al. (2016), that is, use the m-mode covariance matrices, $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_r$, to standardize $\mathbf{X}$ simultaneously from all r

modes. This gives us the standardized tensor

$$\mathbf{X}_{st} := \mathbf{X} \odot_1 \boldsymbol{\Sigma}_1^{-1/2} \cdots \odot_r \boldsymbol{\Sigma}_r^{-1/2}.$$

where, for the asymptotics, we assume that the standardization functionals $\boldsymbol{\Sigma}_m^{-1/2}$, $m = 1, \ldots, r$, are chosen to be symmetric, see e.g. Ilmonen et al. (2012). Estimates $\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_r$ of the m-mode covariance matrices are obtained by applying (4) to the empirical distribution of $\mathbf{X}$. The next step towards $\mathbf{Z}$ is guided by Theorem 5.3.1 in Virta et al. (2016) which states that

$$\mathbf{X}_{st} = \tau \cdot \mathbf{Z} \odot_1 \mathbf{U}_1 \cdots \odot_r \mathbf{U}_r, \tag{10}$$

for some orthogonal matrices $\mathbf{U}_1 \in \mathbb{R}^{p_1 \times p_1}, \ldots, \mathbf{U}_r \in \mathbb{R}^{p_r \times p_r}$ and for $\tau = (\prod_{i=1}^m p_m^{1/2})^{r-1} \|\boldsymbol{\Omega}_r \otimes \cdots \otimes \boldsymbol{\Omega}_1\|_F^{1-r}$, where $\| \cdot \|_F$ is the Frobenius norm.

3.2 Rotation step

We extend the cumulant matrices by noting that the operation $\odot_{-m}$ provides an m-mode analogy for the product of a vector and its transpose. By writing the random quantity $x_i x_j \cdot \mathbf{x}\mathbf{x}^T$ in (5) either as $\mathbf{e}_i^T \mathbf{x}\mathbf{x}^T \mathbf{e}_j \cdot \mathbf{x}\mathbf{x}^T$ or as $\mathbf{x}\mathbf{x}^T \mathbf{e}_i \mathbf{e}_j^T \mathbf{x}\mathbf{x}^T$, where $\mathbf{e}_i$ is the ith standard basis vector, two straightforward tensor m-mode analogies for the matrices of fourth cumulants $\mathbf{C}^{ij}$, $i, j = 1, \ldots, p_m$, in (5) are then given by

$$\begin{aligned} \mathbf{C}_{1,m}^{ij}(\mathbf{X}) = {} & \rho_m^{-1} E\left[\mathbf{e}_i^T (\mathbf{X} \odot_{-m} \mathbf{X}) \mathbf{e}_j \cdot (\mathbf{X} \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[\mathbf{e}_i^T (\mathbf{X}^* \odot_{-m} \mathbf{X}^*) \mathbf{e}_j \cdot (\mathbf{X} \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[\mathbf{e}_i^T (\mathbf{X}^* \odot_{-m} \mathbf{X}) \mathbf{e}_j \cdot (\mathbf{X}^* \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[\mathbf{e}_i^T (\mathbf{X}^* \odot_{-m} \mathbf{X}) \mathbf{e}_j \cdot (\mathbf{X} \odot_{-m} \mathbf{X}^*)\right], \end{aligned} \tag{11}$$

and

$$\begin{aligned} \mathbf{C}_{2,m}^{ij}(\mathbf{X}) = {} & \rho_m^{-1} E\left[(\mathbf{X} \odot_{-m} \mathbf{X}) \mathbf{E}^{ij} (\mathbf{X} \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[(\mathbf{X}^* \odot_{-m} \mathbf{X}^*) \mathbf{E}^{ij} (\mathbf{X} \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[(\mathbf{X}^* \odot_{-m} \mathbf{X}) \mathbf{E}^{ij} (\mathbf{X}^* \odot_{-m} \mathbf{X})\right] \\ & - \rho_m^{-1} E\left[(\mathbf{X}^* \odot_{-m} \mathbf{X}) \mathbf{E}^{ij} (\mathbf{X} \odot_{-m} \mathbf{X}^*)\right], \end{aligned} \tag{12}$$

with $m = 1, \ldots, r$, where $\mathbf{X}^*$ is an independent copy of $\mathbf{X}$, that is, a random variable that is independent of $\mathbf{X}$ and has the same distribution as $\mathbf{X}$. Note that the expressions (11) and (12) are not cumulant matrices in the true sense of the word but rather consist of sums of certain joint cumulants. Theoretically, a third way to generalize the idea is obtained by considering $\mathbf{x}\mathbf{x}^T\mathbf{e}_j\mathbf{e}_i^T\mathbf{x}\mathbf{x}^T$. However, that would be redundant as the resulting set of matrices for $i, j = 1, \ldots, p_m$ is the same as with (12) and the individual matrices can be obtained by just reversing i and j in (12). Naturally, for vector observations, $r = 1$, both (11) and (12) are equivalent.

Define next for the model (3) its kurtosis tensor $\boldsymbol{\kappa} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ as $(\boldsymbol{\kappa})_{i_1 \cdots i_r} := E[z^4_{i_1 \cdots i_r}] - 3$ and its m-mode average kurtosis vector as $\bar{\boldsymbol{\kappa}}^{(m)} := (\bar{\kappa}^{(m)}_1, \ldots, \bar{\kappa}^{(m)}_{p_m})$, where $\bar{\kappa}^{(m)}_k$ is the average of the excess kurtoses of the random variables in the kth m-mode face of the tensor $\mathbf{Z}$, $k = 1, \ldots, p_m$. The following theorem then shows that (11) and (12) actually serve in TJADE the same purpose as their vector counterparts do in JADE.

Theorem 1. *If τ, $\boldsymbol{U}_1$, ..., $\boldsymbol{U}_r$ are as defined in (10), then, for $c = 1, 2$ and $m = 1, \ldots, r$, the matrices of fourth cumulants $\boldsymbol{C}^{ij}_{c,m}, i, j = 1, \ldots, p$ satisfy*

$$\boldsymbol{C}^{ij}_{c,m}(\boldsymbol{X}_{st}) = \tau^4 \cdot \boldsymbol{U}_m \left(\sum_{k=1}^{p_m} u^{(m)}_{ik} u^{(m)}_{jk} \bar{\kappa}^{(m)}_k \boldsymbol{E}^{kk} \right) \boldsymbol{U}^T_m.$$

According to Theorem 1, $\mathbf{U}^T_m$ simultaneously diagonalizes all matrices $\mathbf{C}^{ij}_{c,m}(\mathbf{X}_{st})$, $i, j = 1, \ldots, p_m$, regardless of c, giving two straightforward ways of estimating the m-mode rotation $\mathbf{U}_m$ by using (9) with $\mathbf{C}^{ij}(\mathbf{x}_{st})$ replaced by $\mathbf{C}^{ij}_{c,m}(\mathbf{X}_{st})$ for the chosen value of c. However, in estimating an individual matrix $\mathbf{C}^{ij}_{c,m}(\mathbf{X}_{st})$ in (11) or (12) we have to estimate four matrices in total, the last two of which are costly to estimate because of the independent copies $\mathbf{X}^*$. Using the method of the proof of Theorem 1 one can show that, analogously to the vector-valued case,

$$\mathbf{C}^{ij}_{1,m}(\mathbf{X}_{st}) = \mathbf{B}^{ij}_{1,m} - \boldsymbol{\Xi}_m \left(\delta_{ij}\rho_m \mathbf{I} + \mathbf{E}^{ij} + \mathbf{E}^{ji} \right) \boldsymbol{\Xi}^T_m,$$

where $\mathbf{B}^{ij}_{1,m} := \rho_m^{-1} E\left[\mathbf{e}_i^T (\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st}) \mathbf{e}^j \cdot (\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st}) \right]$ and $\boldsymbol{\Xi}_m := \rho_m^{-1} E\left[\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st} \right] = \tau^2 \mathbf{I}$, which provides a natural estimator for τ^2. Similarly

$$\mathbf{C}^{ij}_{2,m}(\mathbf{X}_{st}) = \mathbf{B}^{ij}_{2,m} - \boldsymbol{\Xi}_m \left(\delta_{ij} \mathbf{I} + \rho_m \mathbf{E}^{ij} + \mathbf{E}^{ji} \right) \boldsymbol{\Xi}^T_m,$$

where $\mathbf{B}_{2,m}^{ij} := \rho_m^{-1} E \left[(\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st}) \mathbf{E}^{ij} (\mathbf{X}_{st} \odot_{-m} \mathbf{X}_{st}) \right]$ and $\boldsymbol{\Xi}_m$ is as above.

Natural estimates for the previous matrices are provided by

$$\hat{\mathbf{C}}_{1,m}^{ij} := \hat{\mathbf{B}}_{1,m}^{ij} - \hat{\boldsymbol{\Xi}}_m \left(\delta_{ij} \rho_m \mathbf{I} + \mathbf{E}^{ij} + \mathbf{E}^{ji} \right) \hat{\boldsymbol{\Xi}}_m^T \quad \text{and} \tag{13}$$

$$\hat{\mathbf{C}}_{2,m}^{ij} := \hat{\mathbf{B}}_{2,m}^{ij} - \hat{\boldsymbol{\Xi}}_m \left(\delta_{ij} \mathbf{I} + \rho_m \mathbf{E}^{ij} + \mathbf{E}^{ji} \right) \hat{\boldsymbol{\Xi}}_m^T, \tag{14}$$

where $i, j = 1, \ldots, p_m$, and the estimates $\hat{\mathbf{B}}_{1,m}^{ij}$, $\hat{\mathbf{B}}_{2,m}^{ij}$ and $\hat{\boldsymbol{\Xi}}_m$ are obtained by applying the definitions of $\mathbf{B}_{1,m}^{ij}$, $\mathbf{B}_{2,m}^{ij}$ and $\boldsymbol{\Xi}_m$ to the empirical distribution of $\mathbf{X}$, including an empirical standardization by $\hat{\boldsymbol{\Sigma}}_1^{-1/2}, \ldots, \hat{\boldsymbol{\Sigma}}_r^{-1/2}$. Choosing then either of the sets, $c = 1, 2$, the rotation matrix $\mathbf{U}_m^T$, $m = 1, \ldots, r$, is found by simultaneous (approximate) diagonalization as

$$\mathbf{U}_m^T = \underset{\mathbf{U}:\ \mathbf{U}^T\mathbf{U} = \mathbf{I}}{\operatorname{argmax}} \sum_{i=1}^{p_m} \sum_{j=1}^{p_m} \| \operatorname{diag}(\mathbf{U} \mathbf{C}_{c,m}^{ij}(\mathbf{X}_{st}) \mathbf{U}^T) \|^2. \tag{15}$$

The corresponding estimates $\hat{\mathbf{U}}_m^T$, $m = 1, \ldots, r$, are obtained by replacing in (15) the matrices $\mathbf{C}_{c,m}^{ij}(\mathbf{X}_{st})$ with their estimates $\hat{\mathbf{C}}_{c,m}^{ij}$.

Combining the standardization and the rotation, the final TJADE algorithm for a sample, $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_r}$, $i = 1, \ldots, n$, consists of the following steps.

1) Center $\mathbf{X}_i$ and estimate $\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_r$.

2) Standardize: $\mathbf{X}_i \leftarrow \mathbf{X}_i \odot_1 \hat{\boldsymbol{\Sigma}}_1^{-1/2} \cdots \odot_r \hat{\boldsymbol{\Sigma}}_r^{-1/2}$.

3) Choose c and estimate the r rotations $\hat{\mathbf{U}}_1^T, \ldots, \hat{\mathbf{U}}_r^T$ by diagonalizing for each $m = 1, \ldots, r$ simultaneously the sets $\hat{\mathbf{C}}_{c,m}^{ij}$, $i, j = 1, \ldots, p_m$.

4) Rotate: $\mathbf{X}_i \leftarrow \mathbf{X}_i \odot_1 \hat{\mathbf{U}}_1^T \cdots \odot_r \hat{\mathbf{U}}_r^T$.

Using Lemma 5.1.1 from Virta et al. (2016) the final result can be written as the product $\mathbf{X}_i \odot_1 \hat{\boldsymbol{\Phi}}_1 \cdots \odot_r \hat{\boldsymbol{\Phi}}_r$, where $\hat{\boldsymbol{\Phi}}_m := \hat{\mathbf{U}}_m^T \hat{\boldsymbol{\Sigma}}_m^{-1/2}$, $m = 1, \ldots, r$, is the *m-mode TJADE estimate.*

Remark 1. *Technically, there is no reason why we could not use different c for estimating different rotations $\boldsymbol{U}_m$. However, the asymptotic properties of the different approaches are in the next section shown to be equivalent and thus the choice of c is for large enough samples irrelevant.*

For a vector valued $\mathbf{x} \in \mathbb{R}^p$ and a full-rank matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, $(\mathbf{A}\mathbf{x})_{st} = \mathbf{U}\mathbf{x}_{st}$ for some orthogonal $\mathbf{U}$ (Ilmonen et al., 2012). Unfortunately, the analogous relation in the tensor setting,

$$(\mathbf{X} \odot_1 \mathbf{A}_1 \cdots \odot_r \mathbf{A}_r)_{st} = \mathbf{X}_{st} \odot_1 \mathbf{U}_1 \cdots \odot_r \mathbf{U}_r \tag{16}$$

for some orthogonal $\mathbf{U}_1, \ldots, \mathbf{U}_r$, holds only for orthogonal $\mathbf{A}_1, \ldots, \mathbf{A}_r$. This lack of *$m$-affine equivariance* of $\mathbf{\Sigma}_m(\mathbf{X})$, $m = 1, \ldots, r$, is discussed in Virta et al. (2016) along with a conjecture that in the general tensor case, $r > 1$, no standardization functional leading into the property (16) exists. In practice this means that outside the model (3) a change (other than rotation or reflection) in the coordinate system leads into different estimated components. However, the TJADE estimator is still Fisher consistent by Theorem 1.

4 Asymptotic properties

The asymptotical properties of JADE were considered in Bonhomme and Robin (2009), Miettinen et al. (2015), Virta et al. (2015) and are in Miettinen et al. (2015), Virta et al. (2015) based on the fact that the JADE functional is affine equivariant, allowing them to consider only the case of no mixing, $\mathbf{\Omega} = \mathbf{I}$. In the following we consider the analogous case of $\mathbf{\Omega}_1 = \mathbf{I}, \ldots, \mathbf{\Omega}_r = \mathbf{I}$ for TJADE. However, because of the lack of full affine equivariance, the results generalize only to orthogonal mixing from all r modes.

For a tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_r}$ define its *m-flattening* $\mathbf{X}_{(m)} \in \mathbb{R}^{p_m \times \rho_m}$ as the horizontal stacking of all m-mode vectors of the tensor into a matrix in a predefined order, see De Lathauwer et al. (2000) for a rigorous definition. If the stacking order is assumed to be cyclical in the dimensions in the sense of De Lathauwer et al. (2000) we have for $\mathbf{X}^* := \mathbf{X} \odot_1 \mathbf{A}_1 \cdots \odot_r \mathbf{A}_r$ the identity

$$\mathbf{X}^*_{(m)} = \mathbf{A}_m \mathbf{X}_{(m)} \left(\mathbf{A}_{m+1} \otimes \cdots \otimes \mathbf{A}_r \otimes \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_{m-1} \right)^T. \tag{17}$$

The reason why m-flattening is particularly useful for us is that it allows us to write the m-mode product of a tensor with itself as an ordinary matrix product, namely $\mathbf{X} \odot_{-m} \mathbf{X} = \mathbf{X}_{(m)} \mathbf{X}_{(m)}^T$, regardless of the stacking order. This, combined with the fact that the matrices $\mathbf{C}^{ij}_{c,m}(\mathbf{X})$ depend on $\mathbf{X}$ only via the previous product, implies that it is sufficient to derive the asymptotics for the case $r = 2$ only. The results for tensors of order $r > 2$ are then obtained by applying the case $r = 2$ for each of the m-flattened matrices $\mathbf{X}_{(1)}, \dots, \mathbf{X}_{(r)}$. Similarly, even for the case $r = 2$ we only need to consider the 1-mode TJADE estimate $\hat{\mathbf{\Phi}}_1$ (matrix multiplication from left) as the results for $\hat{\mathbf{\Phi}}_2$ follow by simply transposing $\mathbf{X}$. Interestingly, we also have no need to specify the used set of cumulant matrices c, as the two choices, $c = 1$ and $c = 2$, are shown to lead into asymptotically equivalent estimators.

We next provide the asymptotic expressions for the elements of the TJADE estimate $\hat{\mathbf{\Phi}}_1 =: \hat{\mathbf{\Phi}}$ in the case of a matrix-valued sample $\mathbf{X}_i \in \mathbb{R}^{p_1 \times p_2}$, $i = 1, \dots, n$. The asymptotic properties of $\hat{\mathbf{\Phi}}$ can be shown to depend on row means of various moments of $\mathbf{Z}$, particularly on the elements of $\bar{\boldsymbol{\kappa}}^{(1)}$ but also on

$$\bar{\boldsymbol{\beta}}^{(1)} := \frac{1}{p_2} \sum_{l=1}^{p_2} \left(\mathrm{E}[z_{1l}^4], \dots, \mathrm{E}[z_{p_1 l}^4] \right)^T \quad \text{and} \quad \bar{\boldsymbol{\omega}}^{(1)} := \frac{1}{p_2} \sum_{l=1}^{p_2} \left(\mathrm{Var}[z_{1l}^3], \dots, \mathrm{Var}[z_{p_1 l}^3] \right)^T.$$

Define further the covariance of two rows of kurtoses as

$$\rho_{kk'} = \frac{1}{p_2} \sum_{l=1}^{p_2} (\beta_{kl} \beta_{k'l}) - \bar{\beta}_k^{(1)} \bar{\beta}_{k'}^{(1)},$$

where $\beta_{kl} := \mathrm{E}[z_{kl}^4]$. For the asymptotic expression of $\hat{\mathbf{\Phi}}$ in Theorem 2 we need the terms

$$\hat{s}_{kk'} := \frac{1}{p_2} \sum_{l=1}^{p_2} \left(\frac{1}{n} \sum_{i=1}^{n} z_{i,kl} z_{i,k'l} \right), \qquad \hat{q}_{kk'} := \frac{1}{p_2} \sum_{l=1}^{p_2} \left(\frac{1}{n} \sum_{i=1}^{n} \left(z_{i,kl}^3 - E[z_{kl}^3] \right) z_{i,k'l} \right),$$

$$\hat{r}_{kk'} := \frac{1}{p_2} \sum_{l=1}^{p_2} \sum_{\substack{l'=1 \\ l' \neq l}}^{p_2} \left(\frac{1}{n} \sum_{i=1}^{n} z_{i,kl}^2 z_{i,kl'} z_{i,k'l'} \right),$$

the joint limiting normality of which is easy to show, assuming the eighth moments of $\mathbf{Z}$ exist.

Theorem 2. *Let $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ be a random sample from a distribution with finite eighth moments and satisfying Assumptions 1, 2 and 4 (see below). Then there exists a sequence of TJADE estimates such that $\hat{\mathbf{\Phi}} \to_P \mathbf{I}$ and*

$$\sqrt{n}(\hat{\phi}_{kk} - 1) = -\frac{1}{2}\sqrt{n}(\hat{s}_{kk} - 1) + o_P(1),$$

$$\sqrt{n}\hat{\phi}_{kk'} = \frac{\sqrt{n}\hat{\psi}_{kk'} - \sqrt{n}\hat{\psi}_{k'k} - d_{kk'}\sqrt{n}\hat{s}_{kk'}}{(\bar{\kappa}_k^{(1)})^2 + (\bar{\kappa}_{k'}^{(1)})^2} + o_P(1),$$

where $k \neq k'$, $\hat{\psi}_{kk'} := \bar{\kappa}_k^{(1)}(\hat{r}_{kk'} + \hat{q}_{kk'})$ and $d_{kk'} := (p_2 + 2)(\bar{\kappa}_k^{(1)} - \bar{\kappa}_{k'}^{(1)}) + (\bar{\kappa}_k^{(1)})^2$.

Using the expressions of Theorem 2 the asymptotic variances of the elements of $\hat{\boldsymbol{\Phi}}$ can now be computed.

Corollary 1. *Under the assumptions of Theorem 2 the limiting distribution of* $\sqrt{n}\,\mathrm{vec}(\hat{\boldsymbol{\Phi}} - \mathbf{I})$ *is multivariate normal with mean vector* $\mathbf{0}$ *and the following asymptotic variances.*

$$
\begin{aligned}
ASV\,(\hat{\phi}_{kk}) &= \frac{\bar{\beta}_k^{(1)} - 1}{4p_2}, \\
ASV(\hat{\phi}_{kk'}) &= \frac{\zeta_k + \zeta_{k'} + (\bar{\kappa}_{k'}^{(1)})^4 - 2\bar{\kappa}_k^{(1)}\bar{\kappa}_{k'}^{(1)}\rho_{kk'}}{p_2((\bar{\kappa}_k^{(1)})^2 + (\bar{\kappa}_{k'}^{(1)})^2)^2}, \quad k \neq k',
\end{aligned}
$$

where $\zeta_k := (\bar{\kappa}_k^{(1)})^2[\bar{\omega}_k^{(1)} - (\bar{\beta}_k^{(1)})^2] + (\bar{\kappa}_k^{(1)})^2(\bar{\kappa}_k^{(1)} + 2)(p_2 - 1)$.

It is easily seen that the expressions in Corollary 1 revert to the forms of Corollary 4 in Miettinen et al. (2015) when $r = 1$, that is, we observe just a vector $\mathbf{x}$. In this case $\bar{\boldsymbol{\kappa}}^{(1)}$ contains just the element-wise kurtoses of the elements of $\mathbf{z}$. Of the popular ICA methods, FastICA, FOBI and JADE, it is well-known that only for FOBI does the asymptotic behavior of $\hat{\phi}_{kk'}$ depend on components other than z_k and $z_{k'}$. The analogous result holds also for TFOBI and TJADE in the sense that in TFOBI the asymptotic behavior of $\hat{\phi}_{kk'}^{(m)}$ depends on the whole tensor $\mathbf{Z}$ (Virta et al., 2016) and in TJADE only on the kth and k'th m-mode faces of $\mathbf{Z}$.

The denominators in Theorem 2 imply that for the existence of the limiting distributions we need the following assumption.

Assumption 4. *For each* $m = 1, \ldots, r$, *at most one of the components of* $\bar{\boldsymbol{\kappa}}^{(m)}$ *is zero.*

Assumption 4 for TJADE is much less restrictive than the assumption needed for TFOBI, for each $m = 1, .., r$ the components of $\bar{\boldsymbol{\kappa}}^{(m)}$are distinct (Virta et al., 2016), and the one needed for vector JADE, at most one element of $\boldsymbol{\kappa}$ is zero (Miettinen et al., 2015). More specifically, in TJADE, and in tensor independent component analysis in general, several individual elements of $\mathbf{Z}$ are allowed to be Gaussian, as long as Assumption 3 is not violated. Conveniently located, a majority of the elements of $\mathbf{Z}$ can thus be Gaussian.

The analytical comparison of TJADE and TFOBI via the asymptotic variances involves in general case rather complicated expressions and thus we resort to simulations for their comparison in the next section.

5 Simulations and examples

In the following all computations were done in R 3.1.2 (R Core Team, 2014) especially using the R-packages JADE (Miettinen et al., 2015), Rcpp (Eddelbuettel et al., 2011; Eddelbuettel, 2013) and ggplot2 (Wickham, 2009). For the approximate joint diagonalization, an algorithm based on Jacobi angles was used, see e.g Cardoso and Souloumiac (1996). Testing the algorithms in various settings showed that both $c = 1$ and $c = 2$ yield almost identical results with respect to the MDI-values (see below) but the former is computationally more efficient and thus the TJADE solution in the simulations is computed with the choice $c = 1$.

5.1 Efficiency comparisons

We compared the separation performance of TJADE with its nearest competitor, TFOBI, and also with regular FOBI and JADE as applied to vectorized tensor data, called here VFOBI and VJADE. Note that VFOBI and VJADE do not use the prior information on the data structure and are therefore expected to be worse than TFOBI and TJADE. The simulation setting was the same as in Virta et al. (2016): we simulated n independent 3×4 matrix observations with individual elements coming from a diverse array of distributions. The excess kurtoses of the distributions used were -1.2, -0.6, 0, 1, 2, 3, 4, 5, 6, 8, 10 and 15 and the exact distributions used are given in Appendix A.

We generated 2000 repetitions for each sample size, $n = 1000, 2000, 4000, 8000, 16000, 32000$, and for each sample the same data was mixed using three different distributions for the elements of the 1-mode and 2-mode mixing matrices, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$. In the first case the mixing matrices were random orthogonal matrices of sizes 3×3 and 4×4 distributed uniformly with respect to the Haar measure. In the second and third case the elements of both matrices were generated independently from $\mathcal{N}(0, 1)$ and Uniform$(-1, 1)$ distributions, respectively.

The mixed data were then subjected to each of the four methods producing the four unmixing matrix estimates, $\hat{\boldsymbol{\Phi}}_{VF}$, $(\hat{\boldsymbol{\Phi}}_{2,MF} \otimes \hat{\boldsymbol{\Phi}}_{1,MF})$, $\hat{\boldsymbol{\Phi}}_{VJ}$ and $(\hat{\boldsymbol{\Phi}}_{2,MJ} \otimes \hat{\boldsymbol{\Phi}}_{1,MJ})$. To allow comparison we took the Kronecker product of the 2-mode and 1-mode unmixing matrices of TFOBI and TJADE meaning that all the four previous matrices estimate the inverse of the same matrix $(\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1)$, up to scaling, sign-change and permutation of its columns.

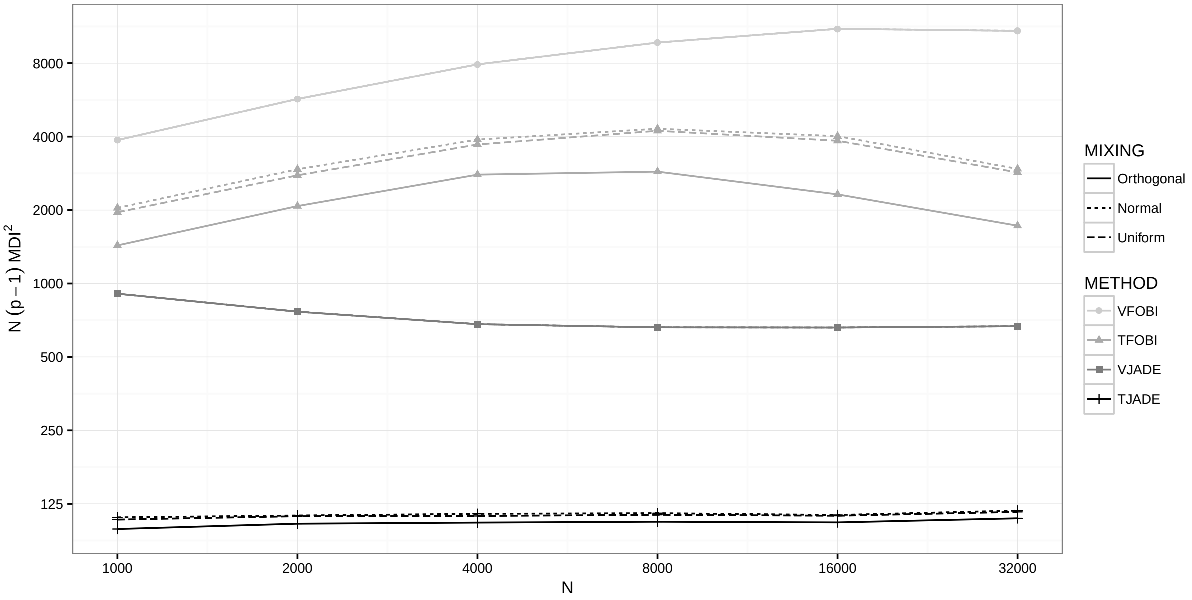

Figure 1: The plot of sample size versus the transformation $n(p-1)\mathrm{MDI}^2$ under combinations of the four methods and three different distributions for the mixing matrices.

The actual comparison was done by first computing the *minimum distance index* (MDI) (Ilmonen et al., 2010) of the estimates $D(\hat{\mathbf{\Phi}}\mathbf{\Omega}) = (p-1)^{-1/2} \inf_{\mathbf{C} \in \mathcal{C}} \| \mathbf{C}\hat{\mathbf{\Phi}}\mathbf{\Omega} - \mathbf{I} \|_F$, where $\hat{\mathbf{\Phi}} \in \mathbb{R}^{p \times p}$ is the estimated unmixing matrix, $\mathbf{\Omega} \in \mathbb{R}^{p \times p}$ is the mixing matrix and $\mathcal{C}$ is the set of all $p \times p$ matrices with a single non-zero element in each row and column. MDI thus measures how far away $\hat{\mathbf{\Phi}}\mathbf{\Omega}$ is from the set $\mathcal{C}$. The index varies from 0 to 1 with 0 indicating a perfect separation. In our simulation we further transformed the MDI-values as $n(p-1)\mathrm{MDI}^2$ which in vector-valued independent component analysis converges in distribution to a random variable with finite mean and variance (Ilmonen et al., 2010).

The mean transformed MDI-values for different sample sizes, methods and mixing matrices are shown in Figure 1. The lines for both VFOBI and VJADE are for all mixings identical since both methods are affine equivariant. For TFOBI and TJADE the separation is best under orthogonal mixing, the results for normal and uniform mixing being a bit worse. But the main implication of the plot is that none of the other methods can really compete with TJADE in matrix independent component analysis. Interestingly, also regular JADE combined with vectorization is better than TFOBI.

5.2 Assumption comparisons

In the second simulation we compared the four methods of the previous simulation via their assumptions. For this we used three simulation settings of $3 \times 3 \times 2$ tensors with independent elements having either Gaussian (N), Laplace (L), exponential (E), or continuous uniform (U) distributions standardized to have zero means and unit variances. The distributions of the tensors are shown in the following by the two $3 \times 3 \times 1$ faces of each setting:

$$\left(\begin{array}{ccc|ccc} \mathtt{N} & \mathtt{L} & \mathtt{E} & \mathtt{U} & \mathtt{U} & \mathtt{U} \\ \mathtt{L} & \mathtt{L} & \mathtt{E} & \mathtt{U} & \mathtt{L} & \mathtt{L} \\ \mathtt{E} & \mathtt{E} & \mathtt{E} & \mathtt{U} & \mathtt{L} & \mathtt{E} \end{array}\right) \qquad \left(\begin{array}{ccc|ccc} \mathtt{N} & \mathtt{L} & \mathtt{L} & \mathtt{U} & \mathtt{U} & \mathtt{U} \\ \mathtt{L} & \mathtt{L} & \mathtt{L} & \mathtt{U} & \mathtt{L} & \mathtt{L} \\ \mathtt{L} & \mathtt{L} & \mathtt{L} & \mathtt{U} & \mathtt{L} & \mathtt{L} \end{array}\right) \qquad \left(\begin{array}{ccc|ccc} \mathtt{E} & \mathtt{E} & \mathtt{N} & \mathtt{N} & \mathtt{N} & \mathtt{N} \\ \mathtt{E} & \mathtt{E} & \mathtt{N} & \mathtt{N} & \mathtt{N} & \mathtt{N} \\ \mathtt{N} & \mathtt{N} & \mathtt{N} & \mathtt{N} & \mathtt{N} & \mathtt{N} \end{array}\right)$$

It is easy to see that none of the above settings satisfies the assumptions of VFOBI as all of them have at least two identical components. Only setting 1 satifies the assumption of TFOBI on distinct kurtosis means in all modes and settings 1 and 2 satisfy the assumption on maximally one component having zero excess kurtosis required by VJADE. All three settings satisfy Assumption 4 on maximally one zero kurtosis mean in each mode required by TJADE.

We simulated 2000 repetitions of all three settings for different sample sizes using identity mixing and the resulting transformed MDI-values of the four methods are depicted in Figure 2. The above reasoning about the violation of assumptions is clearly visible in the plots. The mean transformed MDI-values of the different methods break one-by-one when the setting changes from 1 to 2 to 3 leaving TJADE as the only method able to handle all three settings. Interestingly, VJADE failed to converge 4601 times out of the 36000 total repetitions across all settings, the majority of failures occurring in the third setting.

The plot for setting 1 further indicates that there exist cases where TFOBI beats VJADE, proving that, though very efficient, the JADE methodology itself is not the only factor in the superior performance of TJADE; the tensor structure also plays a role.

5.3 Real data example

Extreme kurtosis can be shown to be associated with multimodal distributions and thus independent component analysis is commonly used as preprocessing step in classification

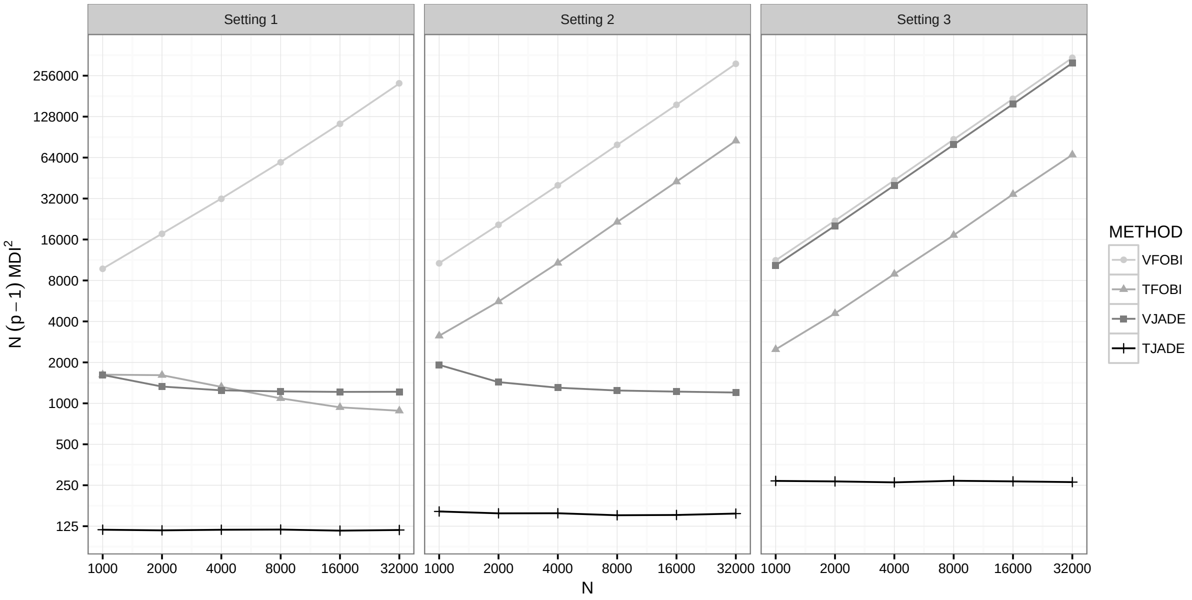

Figure 2: The means of transformed MDI-values for different combinations of setting, sample-size and method. Moving from left to right, all other methods but TJADE break down one-by-one.

to obtain directions of interest. In this spirit we consider the *semeion*[1] data set, available in the UCI Machine Learning Repository (Lichman, 2013) as a classification problem. The data consist of 1593 binary 16 × 16 pixel images of hand-written digits. For this example we chose only the images representing the digits 0, 1 and 7, having respective group sizes of 161, 162 and 158. The objective is to find a few components separating the three digits.

Subjecting the data to TJADE gives the results depicted in Figure 3. The left-hand side plot shows the scatter plot of the two resulting components with the lowest kurtoses using the individual digit images as plot markers. Clearly the two found directions are sufficient to separate all three groups of digits. The same conclusion can be drawn from the corresponding density estimators and rug plots on the right-hand side of Figure 3. As a next step, some low-dimensional classification algorithm could be applied to the extracted components to create a classification rule.

[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.

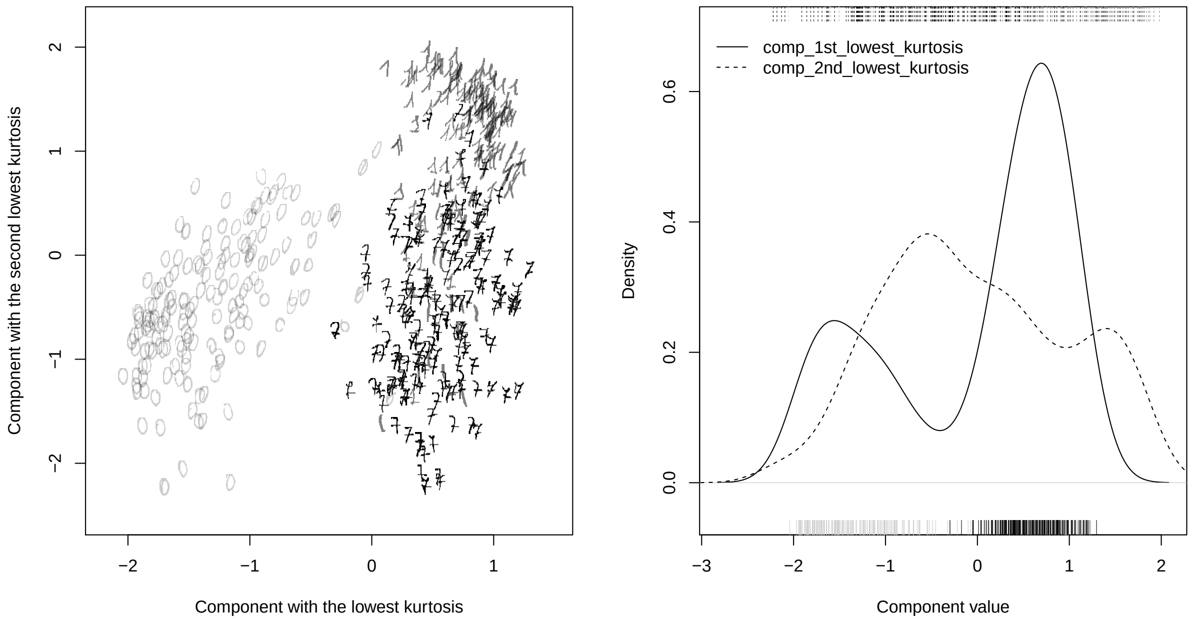

Figure 3: The results of applying TJADE on the semeion data. The plot on the left-hand side shows the scatter plot of the two components having the lowest kurtoses found by TJADE with the individual images as markers. The three digits clearly form three groups in the plane. The density plots along with the rugs on the right-hand side imply the same. The lower rug corresponds to the component with the lowest kurtosis (min_kurtosis_1) and the coloring of the groups in the rugs is the same as in the scatter plot.

6 Discussion

In this paper we proposed TJADE, an extension of the classic JADE suited for tensor-valued observations. In the course of the paper we first reviewed the theory and the algorithm behind JADE and then formulated TJADE analogously giving two different, although asymptotically equivalent, ways of estimating the needed rotations. The asymptotic behaviors of the elements of the TJADE-estimates under orthogonal mixing were next provided allowing theoretical comparison to other methods. Finally, simulation studies comparing TJADE to TFOBI, and the naïve approaches combining vectorization with either FOBI or JADE showed that TJADE is superior to all the previous competitors in tensor independent component analysis.

Some further research ideas concerning ICA and tensors include: As the number of matrices to jointly diagonalize in estimating the m-mode rotation in TJADE grows proportional to the square of the corresponding dimension p_m, an extension like k-JADE (Miettinen et al., 2013) is worth considering for TJADE. Also, as a competing alternative a tensor version of the FastICA algorithm (Hyvärinen et al., 2001) will be investigated, opening many possibilities via allowing choosing both the non-linearity function g and the norm used in the maximization problem, see Miettinen et al. (2015).

Acknowledgments

The work of Virta, Nordhausen and Oja was supported by the Academy of Finland Grant 268703. The work of Li was supported by the National Science Foundation Grant DMS-1407537.

A Simulation details and theorem proofs

The distributions used in the first simulation of Section 5 are, starting from the upper left corner of the matrix and moving down and right, Uniform$(-\sqrt{3}, \sqrt{3})$, Triangular$(-\sqrt{6}, \sqrt{6}, 0)$, $\mathcal{N}(0,1)$, t_{10}, Gamma$(3, \sqrt{3})$, Laplace$(0, 1/\sqrt{2})$, χ^2_3, Gamma$(1.2, \sqrt{1.2})$, Exp(1), $\chi^2_{1.5}$, $\chi^2_{1.2}$ and InverseGaussian$(1,1)$. The distributions were further standardized to have zero means and unit variances.

The proof of Theorem 1. Consider first the case $c = 1$ and the four terms in (11) separately fixing the choice of m. Denoting the first term of (11) by $\mathbf{B}^{ij}_{1,m}(\mathbf{X})$, then according to Lemma 5.4.1 in Virta et al. (2016) we have

$$\mathbf{B}^{ij}_{1,m}(\mathbf{X}_{st}) = \frac{\tau^4}{\rho_m} \mathbf{U}_m E \left[(\mathbf{u}^{(m)}_i)^T \mathbf{Z}_{(m)} \mathbf{Z}^T_{(m)} \mathbf{u}^{(m)}_j \cdot \mathbf{Z}_{(m)} \mathbf{Z}^T_{(m)} \right] \mathbf{U}^T_m,$$

where $\mathbf{Z}_{(m)}$ is the flattened matrix defined in Section 4 and $(\mathbf{u}^{(m)}_i)^T$ is the ith row of $\mathbf{U}_m$. Using the standard properties of expected value and independent random variables the (k, k') element of the inner expectation can be shown to be for $k \neq k'$ equal to $u^{(m)}_{ik} u^{(m)}_{jk'} + u^{(m)}_{jk} u^{(m)}_{ik'}$ and for $k = k'$ equal to $\delta_{ij} \rho_m + u^{(m)}_{ik} u^{(m)}_{jk} (\bar{\kappa}^{(m)}_k + 2)$. Using these to construct a matrix form for the expectation we have

$$\mathbf{B}^{ij}_{1,m}(\mathbf{X}_{st}) = \tau^4 \mathbf{U}_m \left(\sum_{k=1}^{p} u^{(m)}_{ik} u^{(m)}_{jk} \bar{\kappa}^{(m)}_k \mathbf{E}^{kk} \right) \mathbf{U}_m^T + \tau^4 \delta_{ij} \rho_m \mathbf{I} + \tau^4 \mathbf{E}^{ij} + \tau^4 \mathbf{E}^{ji}.$$

The second, third and fourth terms in (11) then serve to remove the extra constant terms above. That they indeed cancel one-by-one the final terms can easily be shown by examining them in the above manner using the independence of $\mathbf{X}$ and $\mathbf{X}^*$. This concludes the proof for $c = 1$ and the corresponding result for $c = 2$ can be proven in precisely the same manner. □

The proof of Theorem 2. The consistency of the TJADE estimator is proven similarly as the consistency of the TFOBI estimator in the proof of Theorem 5.2.1 in Virta et al. (2016).

In the following we assume that $r = 2$ and we are interested in the asymptotical behavior of the 1-mode unmixing matrix. As discussed in Section 4, for the general case of arbitrary r and m-mode unmixing matrix, it suffices to m-flatten the tensor and replace in the following $\hat{\mathbf{\Sigma}}_1^{-1/2}$ with $\hat{\mathbf{\Sigma}}_m^{-1/2}$, $\hat{\mathbf{\Sigma}}_2^{-1/2}$ with $\hat{\mathbf{\Sigma}}_{m+1}^{-1/2} \otimes \cdots \otimes \hat{\mathbf{\Sigma}}_r^{-1/2} \otimes \hat{\mathbf{\Sigma}}_1^{-1/2} \otimes \cdots \otimes \hat{\mathbf{\Sigma}}_{m-1}^{-1/2}$, p_2 with ρ_m and use the corresponding row mean quantities.

For the asymptotic expressions of the diagonal elements of $\sqrt{n}(\hat{\mathbf{\Phi}} - \mathbf{I})$ it suffices to use the same arguments as in the proof of Theorem 5.2.1 in Virta et al. (2016) and for the off-diagonal elements we aim to use Lemma 2 from Miettinen et al. (2015).

But first, define the *symmetric* standardization functionals $\hat{\mathbf{L}} = (\hat{l}_{kk'}) := \hat{\mathbf{\Sigma}}_1^{-1/2}$ and $\hat{\mathbf{R}} = (\hat{r}_{ll'}) := \hat{\mathbf{\Sigma}}_2^{-1/2}$ giving the standardized identity-mixed observations as $\mathbf{X}_{st,i} = \hat{\mathbf{L}} \tilde{\mathbf{Z}}_i \hat{\mathbf{R}}^T$, where $\tilde{\mathbf{Z}}_i = \mathbf{Z}_i - \bar{\mathbf{Z}}$. We then have

$$\sqrt{n}(\hat{l}_{kk'} - \delta_{kk'}) = -(1/2)\sqrt{n}(\hat{s}_{kk'} - \delta_{kk'}) + o_P(1),$$

see Virta et al. (2016), and as simple moment-based estimators we have both $\sqrt{n}(\hat{\mathbf{L}} - \mathbf{I}) = O_P(1)$ and $\sqrt{n}(\hat{\mathbf{R}} - \mathbf{I}) = O_P(1)$, regardless of whether we really have $r = 2$ or use flattened tensors of higher order.

Assume then first that $c = 1$. The matrices $\hat{\mathbf{C}}^{kk'}_{1,1}$, $k, k' = 1, \ldots, p$, in (13) to be simultaneously diagonalized satisfy $\hat{\mathbf{C}}^{kk'} := \hat{\mathbf{C}}^{kk'}_{1,1} \rightarrow_P \mathbf{C}^{kk'}_{1,1}(\mathbf{Z}_i) = \delta_{kk'} \bar{\kappa}^{(1)}_k \mathbf{E}^{kk}$. In the view of Lemma 2 in Miettinen et al. (2015) this means that the only matrices $\mathbf{C}^{rs}_{1,1}(\mathbf{Z}_i)$, $r, s = 1, \ldots, p$, having non-zero kth or k'th diagonal elements are $\mathbf{C}^{kk}_{1,1}(\mathbf{Z}_i)$ and $\mathbf{C}^{k'k'}_{1,1}(\mathbf{Z}_i)$, respectively, yielding the following form for the (k,k'), $k \neq k'$, element of $\hat{\mathbf{U}} := \hat{\mathbf{U}}_1^T$ estimated by (15).

$$\sqrt{n}\hat{u}_{kk'} = \frac{\bar{\kappa}_k^{(1)}\sqrt{n}\hat{\mathbf{C}}_{kk'}^{kk} - \bar{\kappa}_{k'}^{(1)}\sqrt{n}\hat{\mathbf{C}}_{kk'}^{k'k'}}{(\bar{\kappa}_k^{(1)})^2 + (\bar{\kappa}_{k'}^{(1)})^2} + o_P(1),$$

where $\hat{\mathbf{C}}_{rs}^{kk}$ is the (r,s) element of $\hat{\mathbf{C}}^{kk}$. The above expression then together with the (k,k'), $k \neq k'$, element of the left standardization matrix $\hat{\mathbf{L}}$ gives an asymptotic expression for the off-diagonal elements of the estimated left TJADE matrix, see Virta et al. (2016):

$$\sqrt{n}\hat{\phi}_{kk'} = \sqrt{n}\hat{u}_{kk'} + \sqrt{n}\hat{l}_{kk'} + o_P(1), \tag{18}$$

reducing the problem of finding the asymptotics of TJADE into the task of finding the asymptotic behaviors of $\sqrt{n}\hat{\mathbf{C}}_{kk'}^{kk}$ and $\sqrt{n}\hat{\mathbf{C}}_{kk'}^{k'k'}$. Dropping the subscripts for clarity, note that $\hat{\mathbf{C}}^{aa} = \hat{\mathbf{B}}^{aa} - \hat{\boldsymbol{\Xi}}(p_2\mathbf{I} + 2\mathbf{E}^{aa})\hat{\boldsymbol{\Xi}}^T$ and starting from $\hat{\mathbf{B}}^{aa}$ write it out as

$$\hat{\mathbf{B}}^{aa} = \frac{1}{p_2 n}\sum_{i=1}^{n}(\hat{\mathbf{L}}_a^T\tilde{\mathbf{Z}}_i\hat{\mathbf{R}}^*\tilde{\mathbf{Z}}_i^T\hat{\mathbf{L}}_a)\cdot\hat{\mathbf{L}}\tilde{\mathbf{Z}}_i\hat{\mathbf{R}}^*\tilde{\mathbf{Z}}_i^T\hat{\mathbf{L}}^T,$$

where $\hat{\mathbf{L}}_a^T$ is the ath row of $\hat{\mathbf{L}}$ and $\hat{\mathbf{R}}^* := \hat{\mathbf{R}}^T\hat{\mathbf{R}}$. An arbitrary off-diagonal element of $\sqrt{n}(\hat{\mathbf{B}}^{aa} - \mathbf{B}^{aa}(\mathbf{Z}_i))$ then has after the matrix multiplication the form

$$\sqrt{n}\hat{\mathbf{B}}_{kk'}^{aa} = \frac{1}{p_2 n}\sum_{defgstuv}\sqrt{n}\hat{r}_{ef}^*\hat{r}_{tu}^*\hat{l}_{ad}\hat{l}_{ag}\hat{l}_{ks}\hat{l}_{k'v}\hat{H}_{de,gf,st,vu}, \tag{19}$$

where $\hat{H}_{de,gf,st,vu} = (1/n)\sum_{i=1}^{n}\tilde{z}_{i,de}\tilde{z}_{i,gf}\tilde{z}_{i,st}\tilde{z}_{i,vu} \rightarrow_P E(z_{i,de}z_{i,gf}z_{i,st}z_{i,vu})$. Next we expand the multiplicands $\hat{r}_{..}^*$ and $\hat{l}_{..}$ in (19) one-by-one such as $\hat{l}_{ab} = (\hat{l}_{ab} - \delta_{ab}) + \delta_{ab}$, the first term of which is $O_P(1)$ when combined with $\sqrt{n}$ allowing the use of Slutsky's theorem to the whole multiple sum and the second term of which produces an expression like (19) only with one summation index less.

Starting from left this process then produces the terms $o_P(1)$; $o_P(1)$; $\delta_{ak}\sqrt{n}\hat{l}_{kk'} + \delta_{ak'}\sqrt{n}\hat{l}_{k'k} + o_P(1)$; $\delta_{ak}\sqrt{n}\hat{l}_{kk'} + \delta_{ak'}\sqrt{n}\hat{l}_{k'k} + o_P(1)$; $\delta_{ak'}(\bar{\kappa}_{k'}^{(1)} + p_2 + 2)\sqrt{n}\hat{l}_{kk'} + (1-\delta_{ak'})p_2\sqrt{n}\hat{l}_{kk'} + o_P(1)$ and $\delta_{ak}(\bar{\kappa}_k^{(1)} + p_2 + 2)\sqrt{n}\hat{l}_{k'k} + (1-\delta_{ak})p_2\sqrt{n}\hat{l}_{k'k} + o_P(1)$ finally leaving us with the expression

$$\frac{1}{p_2}\sum_{et}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{z}_{i,ae}^2\tilde{z}_{i,kt}\tilde{z}_{i,k't} + o_P(1). \tag{20}$$

Substituting now either $a = k$ or $a = k'$, expanding $\tilde{z}_{i,ab} = z_{i,ab} - \bar{z}_{ab}$ and using the quantities defined in Section 4 the expression in (20) gets the forms $\sqrt{n}\hat{r}_{kk'} + \sqrt{n}\hat{q}_{kk'} + o_P(1)$ and $\sqrt{n}\hat{r}_{k'k} + \sqrt{n}\hat{q}_{k'k} + o_P(1)$, respectively.

Using the above, e.g. $\sqrt{n}\hat{\mathbf{B}}^{kk}_{kk'}$ gets the form

$$(p_2 + 2)\sqrt{n}\hat{l}_{kk'} + (\bar{\kappa}^{(1)}_k + p_2 + 2)\sqrt{n}\hat{l}_{k'k} + \sqrt{n}\hat{r}_{kk'} + \sqrt{n}\hat{q}_{kk'} + o_P(1).$$

For the asymptotic behavior of the remaining term $\hat{\mathbf{\Xi}}(p_2\mathbf{I} + 2\mathbf{E}^{aa})\hat{\mathbf{\Xi}}^T$ one can first use techniques similar to the above to show for $\hat{\mathbf{\Xi}} = (\hat{\xi}_{kk'})$ that $\sqrt{n}(\hat{\xi}_{kk'} - \delta_{kk'}) = o_P(1)$ for $k \neq k'$. Consequently an arbitrary off-diagonal element of $\sqrt{n}(\hat{\mathbf{\Xi}}(p_2\mathbf{I} + 2\mathbf{E}^{aa})\hat{\mathbf{\Xi}}^T - p_2\mathbf{I} - 2\mathbf{E}^{aa})$ is also $o_P(1)$ implying that the term actually contributes nothing to the asymptotic variances of the estimator. Thus $\sqrt{n}\hat{\mathbf{C}}^{aa}_{kk'} = \sqrt{n}\hat{\mathbf{B}}^{aa}_{kk'} + o_P(1)$ and the result of Theorem 2 is obtained by plugging everything in into (18) and using the fact that the standardization functionals are symmetric. The asymptotic variances of Corollary 1 are then straightforward to obtain, e.g. using the table of covariances in the proof of Theorem 5.2.1 in Virta et al. (2016).

Although the starting expressions for $c = 1$ and $c = 2$ are different the final expressions for both $\sqrt{n}\hat{\mathbf{C}}^{kk}_{kk'}$ and $\sqrt{n}\hat{\mathbf{C}}^{k'k'}_{kk'}$ actually match exactly. The corresponding proof for $c = 2$ is obtained in exactly likewise manner, expanding the terms suitably and using Slustky's theorem and is thus omitted here. □

References

Belouchrani, A., K. Abed-Meraim, J.-F. Cardoso, and E. Moulines (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing 45*(2), 434–444.

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*(1), 12 – 25.

Bunse-Gerstner, A., R. Byers, and V. Mehrmann (1993). Numerical methods for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications 14*(4), 927–949.

Cardoso, J.-F. (1989). Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89.*, pp. 2109–2112. IEEE.

Cardoso, J.-F. and A. Souloumiac (1993). Blind beamforming for non-gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, Volume 140, pp. 362–370. IET.

Cardoso, J.-F. and A. Souloumiac (1996). Jacobi angles for simultaneous diagonalization. *SIAM journal on matrix analysis and applications 17*(1), 161–164.

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

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

Ding, S. and R. D. Cook (2015). Tensor sliced inverse regression. *Journal of Multivariate Analysis 133*, 216–231.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp.* Springer.

Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*(8), 1–18.

Greenewald, K. and A. Hero (2015). Robust kronecker product PCA for spatio-temporal covariance estimation. *IEEE Transactions on Signal Processing 63*(23), 6368–6378.

Hoff, P. D. et al. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis 6*(2), 179–196.

Hyvärinen, A., J. Karhunen, and E. Oja (2001). *Independent Component Analysis.* New York, USA: John Wiley & Sons.

Ilmonen, P., K. Nordhausen, H. Oja, and E. Ollila (2010). A new performance index for ICA: properties, computation and asymptotic analysis. In *Latent Variable Analysis and Signal Separation*, pp. 229–236. Springer.

Ilmonen, P., H. Oja, and R. Serfling (2012). On invariant coordinate system (ICS) functionals. *International Statistical Review 80*(1), 93–110.

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

Li, B., M. K. Kim, and N. Altman (2010). On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, 1094–1121.

Lichman, M. (2013). UCI machine learning repository.

Lu, H., K. N. Plataniotis, and A. N. Venetsanopoulos (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognition 44*(7), 1540–1551.

Miettinen, J., K. Nordhausen, H. Oja, and S. Taskinen (2013, May). Fast equivariant JADE. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, pp. 6153–6157.

Miettinen, J., K. Nordhausen, H. Oja, S. Taskinen, and J. Virta (2015). The squared symmetric FastICA estimator. *arXiv preprint arXiv:1512.05534*.

Miettinen, J., K. Nordhausen, and S. Taskinen (2015). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasymp. *Conditional accepted for publication in the Journal of Statistical Software*.

Miettinen, J., S. Taskinen, K. Nordhausen, and H. Oja (2015). Fourth moments and independent component analysis. *Statistical Science 30*(3), 372–390.

Moreau, E. (2001). A generalization of joint-diagonalization criteria for source separation. *IEEE Transactions on Signal Processing 49*(3), 530–541.

Pfeiffer, R. M., L. Forzani, and E. Bura (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine 31*(22), 2414–2427.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Srivastava, M. S., T. von Rosen, and D. Von Rosen (2008). Models with a kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics 17*(4), 357–370.

Sun, Y., P. Babu, and D. P. Palomar (2015). Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing 64*(14), 3576–3590.

Vasilescu, M. A. O. and D. Terzopoulos (2005). Multilinear independent components analysis. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 1, pp. 547–553. IEEE.

Virta, J., B. Li, K. Nordhausen, and H. Oja (2016). Independent component analysis for tensor-valued data. *arXiv preprint arXiv:1602.00879*.

Virta, J., K. Nordhausen, and H. Oja (2015). Joint use of third and fourth cumulants in independent component analysis. *arXiv preprint arXiv:1505.02613*.

Werner, K., M. Jansson, and P. Stoica (2008). On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing 56*(2), 478–491.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.* Springer New York.

Wiesel, A. (2012). Geodesic convexity and covariance estimation. *IEEE Transactions on Signal Processing 60*(12), 6182–6189.

Xue, Y. and X. Yin (2014). Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics 23*(4), 1028–1043.

Zhang, L., Q. Gao, and L. Zhang (2008). Directional independent component analysis with tensor representation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–7. IEEE.