Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Linear Algebra and its Applications 391 (2004) 31–55

LINEAR ALGEBRA AND ITS APPLICATIONS

www.elsevier.com/locate/laa

Dimensionality reduction in higher-order signal processing and rank-($R_1, R_2, \ldots, R_N$) reduction in multilinear algebra

Lieven De Lathauwer [a,*], Joos Vandewalle [b,1]

[a] *ETIS, UMR 8051, 6 avenue du Ponceau, BP 44, F 95014 Cergy-Pontoise Cedex, France*
[b] *SCD-SISTA, E.E. Dept. (ESAT), K.U.Leuven, Kasteelpark Arenberg 10,*
B-3001 Leuven (Heverlee), Belgium

Received 19 May 2003; accepted 22 January 2004

Submitted by S. van Huffel

Abstract

In this paper we review a multilinear generalization of the singular value decomposition and the best rank-($R_1, R_2, \ldots, R_N$) approximation of higher-order tensors. We show that they are important tools for dimensionality reduction in higher-order signal processing. We discuss applications in independent component analysis, simultaneous matrix diagonalization and subspace variants of algorithms based on higher-order statistics.

Keywords: Multilinear algebra; Higher-order tensors; Higher-order statistics; Independent component analysis; Singular value decomposition; Source separation; Rank reduction

1. Introduction

Multilinear algebra is the algebra of higher-order tensors, which are the higher-order equivalents of vectors (first-order) and matrices (second-order), i.e., quantities of which the elements are addressed by more than two indices. Multilinear algebra

[*] Corresponding author. Tel.: +33-1-3073-6610; fax: +33-1-3073-6282.
E-mail addresses: delathau@ensea.fr (L. De Lathauwer), vdwalle@esat.kuleuven.ac.be (J. Vandewalle).
[1] Tel.: +32-16-321709; fax: +32-16-321970.

0024-3795/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2004.01.016

is gaining more and more interest, largely due to its applications in higher-order statistics (HOS) [34]. Moreover, multilinear algebra is the framework for multi-way data analysis (going from Chemometrics [5] to DS-CDMA techniques in telecommunications [37]). There are also important links with non-linear (polynomial, Volterra) modelling [11]. More generally, multilinear algebra is an important emerging discipline in non-Gaussian, non-linear and non-stationary signal processing.

The first part of this paper consists of a review of two important multilinear algebraic concepts: a multilinear generalization of the singular value decomposition (SVD) [18] (Section 3) and the best approximation, in least squares sense, of a given tensor by a tensor of lower column rank, row rank, etc. [19] (Section 4). We will discuss the way they are related. We will pay special attention to their use in the estimation of the column space, row space, etc., of a tensor contaminated by noise. Some necessary background material is first provided in Section 2.

The goal of the second part of this paper is to show that these concepts are important tools for dimensionality reduction in higher-order signal processing. Working with the original, high-dimensional data may be too time-consuming or even computationally infeasible. Moreover, it is known that low-dimensional estimators often have a smaller variance than high-dimensional estimators, which may lead to more accurate results. Important application domains are biomedical engineering, data analysis, image processing, etc.

We will pay special attention to the problem of independent component analysis (ICA). The necessary material on HOS is provided in Section 5; the problem itself is discussed in Section 6. A dimensionality reduction in ICA is usually based on an eigenvalue decomposition (EVD) of the covariance matrix of the data or, directly, on an SVD of the data matrix. The fact that this approach is explicitly or implicitly based on second-order statistics involves certain drawbacks, as will be clarified in Section 6. Section 7 explains how one can proceed when the ICA solution is obtained from HOS only. Section 8 addresses the case where both second- and higher-order statistics are used. Many algorithms for signal processing, including ICA variants, are nowadays based on a simultaneous diagonalization of a set of matrices. Section 9 explains how a dimensionality reduction can be realized in this context. In [2] subspace processing was discussed for general HOS-based signal processing. In Section 10 the results are cast in the multilinear algebraic framework of Sections 3 and 4, which leads to some complementary insights and techniques.

The content of Section 7 has already appeared as the conference paper [15].

Before starting the actual exposition, we would like to comment on our notation. To facilitate the distinction between scalars, vectors, matrices and higher-order tensors, the type of a given quantity will be reflected by its representation: scalars are denoted by lower-case letters ($a, b, \ldots; \alpha, \beta, \ldots$), vectors are written as capitals ($A, B, \ldots$) (italic shaped), matrices correspond to bold-face capitals ($\mathbf{A}, \mathbf{B}, \ldots$) and tensors are written as calligraphic letters ($\mathscr{A}, \mathscr{B}, \ldots$). This notation is consistently used for lower-order parts of a given structure. For example, the entry with row index i and column index j in a matrix $\mathbf{A}$, i.e., $(\mathbf{A})_{ij}$, is symbolized by a_{ij} (also

$(A)_i = a_i$ and $(\mathscr{A})_{i_1 i_2 \cdots i_N} = a_{i_1 i_2 \cdots i_N}$); furthermore, the ith column vector of a matrix $\mathbf{A}$ is denoted as A_i, i.e., $\mathbf{A} = [A_1 \;\; A_2 \;\; \cdots]$. To enhance the overall readability, we have made one exception to this rule: as we frequently use the characters i, r and n in the meaning of indices (counters), I, R and N will be reserved to denote the index upper bounds, unless stated otherwise.

2. Basic definitions

In this section we introduce some elementary notations and definitions in multilinear algebra, which will be needed in the further developments.

2.1. Multiplication of a higher-order tensor by a matrix

We have the following definition.

Definition 1 (*n-mode product* [41]). The n-mode product of a tensor $\mathscr{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{C}^{J_n \times I_n}$, denoted by $\mathscr{A} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \cdots \times I_N)$-tensor of which the entries are given by

$$(\mathscr{A} \times_n \mathbf{U})_{i_1 i_2 \cdots i_{n-1} j_n i_{n+1} \ldots i_N} \overset{\text{def}}{=} \sum_{i_n} a_{i_1 i_2 \cdots i_{n-1} i_n i_{n+1} \cdots i_N} u_{j_n i_n}.$$

In this notation, the matrix product $\mathbf{G} = \mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^{\mathrm{T}}$, involving matrices $\mathbf{F} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{U} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{V} \in \mathbb{R}^{J_2 \times I_2}$ and $\mathbf{G} \in \mathbb{R}^{J_1 \times J_2}$, is written as $\mathbf{G} = \mathbf{F} \times_1 \mathbf{U} \times_2 \mathbf{V}$. This is meaningful: the relationship between $\mathbf{U}$ and $\mathbf{F}$ and the relationship between $\mathbf{V}$ (not $\mathbf{V}^{\mathrm{T}}$) and $\mathbf{F}$ are in fact completely similar: in the same way as $\mathbf{U}$ makes linear combinations of the rows of $\mathbf{F}$, $\mathbf{V}$ makes linear combinations of the columns of $\mathbf{F}$; in the same way as the columns of $\mathbf{F}$ are multiplied by $\mathbf{U}$, the rows of $\mathbf{F}$ are multiplied by $\mathbf{V}$; in the same way as the columns of $\mathbf{U}$ are associated with the column space of $\mathbf{G}$, the columns of $\mathbf{V}$ are associated with the row space of $\mathbf{G}$. This typical relationship is denoted by means of the $\times_n$-symbol.

We have the following properties:

$$(\mathscr{A} \times_n \mathbf{U}) \times_m \mathbf{V} = (\mathscr{A} \times_m \mathbf{V}) \times_n \mathbf{U} = \mathscr{A} \times_n \mathbf{U} \times_m \mathbf{V},$$
$$(\mathscr{A} \times_n \mathbf{V}) \times_n \mathbf{U} = \mathscr{A} \times_n (\mathbf{U} \cdot \mathbf{V}).$$

Note that in general $\mathscr{A} \times_n \mathbf{U} \times_2 \mathbf{V} \neq \mathscr{A} \times_n (\mathbf{U} \cdot \mathbf{V}^{\mathrm{T}})$.

2.2. Matrix representation of a higher-order tensor

Some of the results can be conveniently expressed in matrix terms. To this end, we must stack the elements of a higher-order tensor in a matrix. There are several

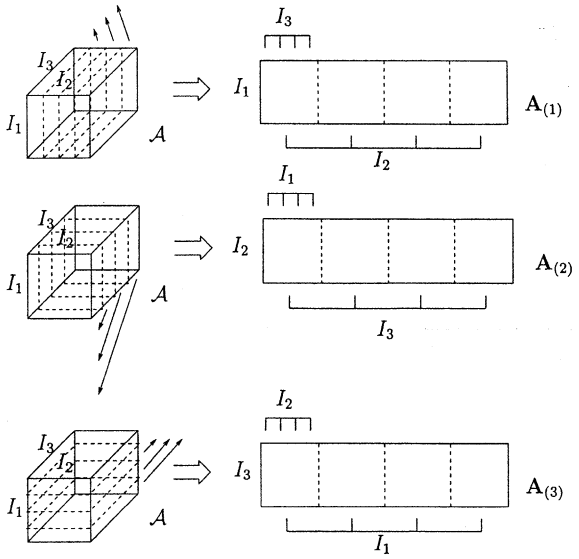

Fig. 1. Unfolding of the $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ to the $(I_1 \times I_2 I_3)$-matrix $\mathbf{A}_{(1)}$, the $(I_2 \times I_3 I_1)$-matrix $\mathbf{A}_{(2)}$ and the $(I_3 \times I_1 I_2)$-matrix $\mathbf{A}_{(3)}$ $(I_1 = I_2 = I_3 = 4)$.

ways to do so. One particular type of "matrix unfolding" will prove to be particularly useful, namely, the matrix representation of a given tensor in which all its column (row, ...) vectors are simply stacked one after another [41]. To avoid confusion, we will retain one particular ordering of the column (row, ...) vectors; for order three, these unfolding procedures are visualized in Fig. 1. Notice that the definitions of the matrix unfoldings involve the tensor dimensions I_1, I_2, I_3 in a cyclic way and that, when dealing with an unfolding of dimensionality $I_c \times I_a I_b$, we formally assume that the index i_a varies more slowly than i_b. In general, we define:

Definition 2. Assume an Nth-order tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$. The matrix unfolding $\mathbf{A}_{(n)} \in \mathbb{C}^{I_n \times (I_{n+1} I_{n+2} \cdots I_N I_1 I_2 \cdots I_{n-1})}$ contains the element $a_{i_1 i_2 \cdots i_N}$ at the position with row number i_n and column number equal to $(i_{n+1} - 1) I_{n+2} I_{n+3} \cdots I_N I_1 I_2 \cdots I_{n-1} + (i_{n+2} - 1) I_{n+3} I_{n+4} \cdots I_N I_1 I_2 \cdots I_{n-1} + \cdots + (i_N - 1) I_1 I_2 \cdots I_{n-1} + (i_1 - 1) I_2 I_3 \cdots I_{n-1} + (i_2 - 1) I_3 I_4 \cdots I_{n-1} + \cdots + i_{n-1}$.

Example 1. Define a tensor $\mathcal{A} \in \mathbb{R}^{3 \times 2 \times 3}$ by $a_{111} = a_{112} = a_{211} = -a_{212} = 1$, $a_{213} = a_{311} = a_{313} = a_{121} = a_{122} = a_{221} = -a_{222} = 2$, $a_{223} = a_{321} = a_{323} = 4$, $a_{113} = a_{312} = a_{123} = a_{322} = 0$. The matrix unfolding $\mathbf{A}_{(1)}$ is given by:

$$\mathbf{A}_{(1)} = \left(\begin{array}{ccc|ccc} 1 & 1 & 0 & 2 & 2 & 0 \\ 1 & -1 & 2 & 2 & -2 & 4 \\ 2 & 0 & 2 & 4 & 0 & 4 \end{array}\right).$$

2.3. Scalar product, orthogonality and Frobenius-norm

The scalar product $\langle \mathscr{A}, \mathscr{B} \rangle$ of two tensors $\mathscr{A}, \mathscr{B} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined in a straightforward way as $\langle \mathscr{A}, \mathscr{B} \rangle \overset{\text{def}}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} a_{i_1 i_2 \cdots i_N} b^*_{i_1 i_2 \cdots i_N}$. The Frobenius-norm of a tensor $\mathscr{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is then defined as $\|\mathscr{A}\| \overset{\text{def}}{=} \sqrt{\langle \mathscr{A}, \mathscr{A} \rangle}$. Two tensors are called orthogonal when their scalar product is 0.

2.4. Rank properties of a higher-order tensor

There are major differences between matrices and higher-order tensors when rank properties are concerned. A rank-1 tensor is a tensor that consists of the outer product of a number of vectors [5,19,27]. For an Nth-order tensor $\mathscr{A}$ and N vectors $U^{(1)}, U^{(2)}, \ldots, U^{(N)}$, this means that $a_{i_1 i_2 \cdots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \cdots u^{(N)}_{i_N}$ for all values of the indices, which will be concisely written as $\mathscr{A} = U^{(1)} \circ U^{(2)} \circ \cdots \circ U^{(N)}$. An n-mode vector of an $(I_1 \times I_2 \times \cdots \times I_N)$-tensor $\mathscr{A}$ is an I_n-dimensional vector obtained from $\mathscr{A}$ by varying the index i_n and keeping the other indices fixed. The n-rank of a higher-order tensor is the obvious generalization of the column (row) rank of matrices: it equals the dimension of the vector space spanned by the n-mode vectors. An important difference with the rank of matrices, is that the different n-ranks of a higher-order tensor are not necessarily the same. The n-rank will be denoted as $\text{rank}_n(\mathscr{A})$. A tensor of which the n-ranks are equal to R_n $(1 \leqslant n \leqslant N)$ is called a rank-$(R_1, R_2, \ldots, R_N)$ tensor. Even when all the n-ranks are the same, they can still be different from *the* rank of the tensor, denoted as $\text{rank}(\mathscr{A})$; $\mathscr{A}$ having rank R generally means that it can be decomposed in a sum of R, but not less than R, rank-1 terms (see e.g. [31]). From the definition of n-rank and rank follows that $R_n \leqslant R$ for all n.

Example 2. Consider the $(2 \times 2 \times 2)$-tensor $\mathscr{A}$ defined by

$$\begin{cases} a_{111} = a_{112} = 1, \\ a_{221} = a_{222} = 2, \\ a_{211} = a_{121} = a_{212} = a_{122} = 0. \end{cases}$$

The 1-mode vectors are the columns of the matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 2 \end{pmatrix}.$$

Because of the symmetry, the set of 2-mode vectors is the same as the set of 1-mode vectors. The 3-mode vectors are the columns of the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 \end{pmatrix}.$$

Hence, we have that $R_1 = R_2 = 2$ but $R_3 = 1$.

The rank $R = 2$, because $\mathscr{A}$ can be decomposed as

$$\mathscr{A} = \begin{pmatrix}1\\0\end{pmatrix} \circ \begin{pmatrix}1\\0\end{pmatrix} \circ \begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} \circ \begin{pmatrix}0\\1\end{pmatrix} \circ \begin{pmatrix}2\\2\end{pmatrix}.$$

The n-rank of a given tensor can be analyzed by means of matrix techniques:

Property 1. *The n-mode vectors of $\mathscr{A}$ are the column vectors of the matrix unfolding $\mathbf{A}_{(n)}$ and*

$$\mathrm{rank}_n(\mathscr{A}) = \mathrm{rank}(\mathbf{A}_{(n)}).$$

3. The higher-order singular value decomposition

In [18,40,41] the following tensor decomposition was discussed.

Theorem 1 (Nth-order singular value decomposition). *Every complex $(I_1 \times I_2 \times \cdots \times I_N)$-tensor $\mathscr{A}$ can be written as the product*

$$\mathscr{A} = \mathscr{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}, \tag{1}$$

in which:

- $\mathbf{U}^{(n)} = \left[U_1^{(n)}\ U_2^{(n)}\ \cdots\ U_{I_n}^{(n)}\right]$ *is a unitary $(I_n \times I_n)$-matrix,*
- $\mathscr{S}$ *is a complex $(I_1 \times I_2 \times \cdots \times I_N)$-tensor of which the subtensors $\mathscr{S}_{i_n=\alpha}$, obtained by fixing the* nth *index to α, have the properties of*:
  - *all-orthogonality: two subtensors $\mathscr{S}_{i_n=\alpha}$ and $\mathscr{S}_{i_n=\beta}$ are orthogonal for all possible values of n, α and β subject to $\alpha \neq \beta$*:

    $$\langle \mathscr{S}_{i_n=\alpha}, \mathscr{S}_{i_n=\beta} \rangle = 0 \quad \textit{when } \alpha \neq \beta, \tag{2}$$

  - *ordering*:

    $$\|\mathscr{S}_{i_n=1}\| \geqslant \|\mathscr{S}_{i_n=2}\| \geqslant \cdots \geqslant \|\mathscr{S}_{i_n=I_n}\| \geqslant 0 \tag{3}$$

    for all possible values of n.

The Frobenius-norms $\|\mathscr{S}_{i_n=i}\|$, symbolized by $\sigma_i^{(n)}$, are n-mode singular values of $\mathscr{A}$ and the vector $U_i^{(n)}$ is an ith *n-mode singular vector. The decomposition is visualized for third-order tensors in Fig.* 2.

Applied to a tensor $\mathscr{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, Theorem 1 says that it is always possible to find orthogonal transformations of the column, row and 3-mode space such that $\mathscr{S} = \mathscr{A} \times_1 \mathbf{U}^{(1)^{\mathrm{T}}} \times_2 \mathbf{U}^{(2)^{\mathrm{T}}} \times_3 \mathbf{U}^{(3)^{\mathrm{T}}}$ is all-orthogonal and ordered (the new basis vectors are the columns of $\mathbf{U}^{(1)}$, $\mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$). All-orthogonality means that the different

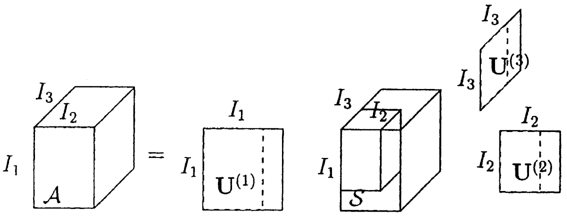

Fig. 2. Visualization of the HOSVD for a third-order tensor.

“horizontal matrices” of $\mathscr{S}$ (the first index i_1 is kept fixed, whilst the two other indices, i_2 and i_3, are free) are mutually orthogonal with respect to the scalar product of matrices (i.e. the sum of the products of the corresponding entries vanishes); at the same time, the different “frontal” matrices (i_2 fixed) and the different “vertical” matrices (i_3 fixed) should be mutually orthogonal as well. The ordering constraint imposes that the Frobenius-norm of the horizontal (frontal resp. vertical) matrices does not increase as the index i_1 (i_2 resp. i_3) is increased.

This decomposition is clearly a generalization of the matrix SVD. Note, w.r.t. the condition of all-orthogonality on $\mathscr{S}$, that in the matrix case the matrix of singular values is all-orthogonal as well: due to its diagonal structure, the scalar product of two different rows or columns also vanishes. On the other hand, one cannot define a higher-order SVD by imposing diagonality on the core tensor $\mathscr{S}$: in general, it is impossible to reduce higher-order tensors to a diagonal form by means of unitary transformations, because the number of degrees of freedom in such a hypothetical decomposition is smaller than the number of entries of the tensor that has to be decomposed.

For convenience, we will refer to the decomposition in Theorem 1 as the “higher-order SVD” (HOSVD). This is substantiated by the many striking analogies with the matrix SVD, established in [18]. Nevertheless, one should be aware that focusing on different properties of the matrix SVD can lead to the definition of different (formally less striking) multilinear SVD-generalizations. One possibility is to look for unitary transformations that make the core tensor as diagonal as possible (in a least squares sense) [10]. Another alternative is the decomposition of the tensor in a minimal number of rank-1 terms; this decomposition is often called the canonical decomposition (CANDECOMP) or parallel factors decomposition (PARAFAC) [5,12]. Imposing orthogonality constraints on these rank-1 terms may lead to yet other generalizations [28].

The matrix of n-mode singular vectors $\mathbf{U}^{(n)}$ can directly be found as the matrix of left singular vectors of the matrix unfolding $\mathbf{A}_{(n)}$. The n-mode singular values correspond to the singular values of this matrix unfolding. The core tensor $\mathscr{S}$ can then be computed by bringing the matrices of singular vectors to the left side of Eq. (1):

$$\mathscr{S} = \mathscr{A} \times_1 \mathbf{U}^{(1)^{\mathrm{H}}} \times_2 \mathbf{U}^{(2)^{\mathrm{H}}} \cdots \times_N \mathbf{U}^{(N)^{\mathrm{H}}}. \tag{4}$$

The fact that the number of non-vanishing singular values of a given matrix equals its (column/row) rank, carries over to the n-mode singular values and the n-rank values of a given tensor [18]. The fact that we have N different sets of n-mode singular values is conform the fact that we also have N different n-ranks. Like for matrices, this link even holds in a numerical sense: the number of significant n-mode singular values of a given tensor equals its numerical n-rank.

4. Best rank-($R_1, R_2, \ldots, R_N$) approximation

In this section we consider a multilinear generalization of the best rank-R approximation of a given matrix. Formally, the problem we want to solve, can be formulated as follows:

Given an Nth-*order tensor* $\mathscr{A} \in \mathbb{C}^{I_1 \times I_2 \times \cdots I_N}$, *find a tensor* $\hat{\mathscr{A}} \in \mathbb{C}^{I_1 \times I_2 \times \cdots I_N}$, *with* $\mathrm{rank}_1(\hat{\mathscr{A}}) = R_1, \mathrm{rank}_2(\hat{\mathscr{A}}) = R_2, \ldots, \mathrm{rank}_N(\hat{\mathscr{A}}) = R_N$, *that minimizes the least-squares cost function*

$$f(\hat{\mathscr{A}}) = \|\mathscr{A} - \hat{\mathscr{A}}\|^2. \tag{5}$$

The matrix counterpart is also known as the total least squares problem [46].

The n-rank conditions imply that $\hat{\mathscr{A}}$ can be decomposed as

$$\hat{\mathscr{A}} = \mathscr{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)}, \tag{6}$$

in which $\mathbf{U}^{(1)} \in \mathbb{C}^{I_1 \times R_1}, \mathbf{U}^{(2)} \in \mathbb{C}^{I_2 \times R_2}, \ldots, \mathbf{U}^{(N)} \in \mathbb{C}^{I_N \times R_N}$ each have orthonormal columns and $\mathscr{B} \in \mathbb{C}^{R_1 \times R_2 \times \cdots \times R_N}$.

Similarly to the second-order case, where the best approximation of a given matrix $\mathbf{A} \in \mathbb{C}^{I_1 \times I_2}$ by a matrix $\hat{\mathbf{A}} = \mathbf{U}^{(1)} \cdot \mathbf{B} \cdot \mathbf{U}^{(2)^{\mathrm{H}}}$, with $\mathbf{U}^{(1)} \in \mathbb{C}^{I_1 \times R}$ and $\mathbf{U}^{(1)} \in \mathbb{C}^{I_2 \times R}$ column-wise orthonormal, is equivalent to the maximization of $\left\|\mathbf{U}^{(1)^{\mathrm{H}}} \cdot \mathbf{A} \cdot \mathbf{U}^{(2)}\right\|$, we have that the minimization of f is equivalent to the maximization of

$$g\left(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}\right) = \left\|\mathscr{A} \times_1 \mathbf{U}^{(1)^{\mathrm{H}}} \times_2 \mathbf{U}^{(2)^{\mathrm{H}}} \cdots \times_N \mathbf{U}^{(N)^{\mathrm{H}}}\right\|^2. \tag{7}$$

The optimal core tensor follows from

$$\mathscr{B} = \mathscr{A} \times_1 \mathbf{U}^{(1)^{\mathrm{H}}} \times_2 \mathbf{U}^{(2)^{\mathrm{H}}} \cdots \times_N \mathbf{U}^{(N)^{\mathrm{H}}}. \tag{8}$$

It is natural to ask whether the best rank-($R_1, R_2, \ldots, R_N$) approximation of a higher-order tensor can be obtained by truncation of the HOSVD. The situation turns out to be quite different for tensors here [30]. By discarding the smallest n-mode singular values, one obtains a tensor $\hat{\mathscr{A}}$ that is in general not the best possible approximation under the given n-mode rank constraints (see e.g. Example 3). Nevertheless, the ordering assumption (3) implies that the "energy" of $\mathscr{A}$ is mainly concentrated in the part corresponding to low values of $i_1, i_2, \ldots, i_N$. Consequently, if $\sigma^{(n)}_{R_n} \gg \sigma^{(n)}_{R_n+1}$, $\hat{\mathscr{A}}$ is still to be considered as a good approximation of $\mathscr{A}$. The error is bounded as follows [18].

Property 2. *Let the HOSVD of* $\mathscr{A}$ *be given as in Theorem* 1 *and let the n-mode rank of* $\mathscr{A}$ *be equal to* $\tilde{R}_n$ $(1 \leqslant n \leqslant N)$*. Define a tensor* $\hat{\mathscr{A}}$ *by discarding the smallest n-mode singular values* $\sigma^{(n)}_{R_n+1}, \sigma^{(n)}_{R_n+2}, \ldots, \sigma^{(n)}_{\tilde{R}_n}$*, for given values of* R_n $(1 \leqslant n \leqslant N)$*, i.e., set the corresponding parts of* $\mathscr{S}$ *equal to zero. Then we have*

$$\|\mathscr{A} - \hat{\mathscr{A}}\|^2 \leqslant \sum_{i_1=R_1+1}^{\tilde{R}_1} \sigma^{(1)^2}_{i_1} + \sum_{i_2=R_2+1}^{\tilde{R}_2} \sigma^{(2)^2}_{i_2} + \cdots + \sum_{i_N=R_N+1}^{\tilde{R}_N} \sigma^{(N)^2}_{i_N}. \tag{9}$$

(For completeness, we mention that the truncation of $\mathscr{S}$ may destroy its all-orthogonality. In this sense, the components of the truncated HOSVD of $\mathscr{A}$ may be different from the components of the HOSVD of $\hat{\mathscr{A}}$, as opposed to the matrix case.)

In [19,29,30,38] the following approach was followed for the computation of the best rank-$(R_1, R_2, \ldots, R_N)$ approximation. Imagine that the matrices $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(n-1)}, \mathbf{U}^{(n+1)}, \ldots, \mathbf{U}^{(N)}$ are fixed and that the only unknown is the column-wise orthonormal matrix $\mathbf{U}^{(n)}$. We have:

$$g = \left\|\tilde{\mathscr{A}}^{(n)} \times_n \mathbf{U}^{(n)^{\mathrm{H}}}\right\|^2, \tag{10}$$

in which

$$\tilde{\mathscr{A}}^{(n)} \overset{\text{def}}{=} \mathscr{A} \times_1 \mathbf{U}^{(1)^{\mathrm{H}}} \cdots \times_{n-1} \mathbf{U}^{(n-1)^{\mathrm{H}}} \times_{n+1} \mathbf{U}^{(n+1)^{\mathrm{H}}} \cdots \times_N \mathbf{U}^{(N)^{\mathrm{H}}}. \tag{11}$$

In a matrix format we have:

$$g = \left\|\mathbf{U}^{(n)^{\mathrm{H}}} \cdot \tilde{\mathbf{A}}^{(n)}_{(n)}\right\|^2, \tag{12}$$

with

$$\tilde{\mathbf{A}}^{(n)}_{(n)} = \mathbf{A}_{(n)} \cdot \left(\mathbf{U}^{(n+1)} \otimes \cdots \otimes \mathbf{U}^{(N)} \otimes \mathbf{U}^{(1)} \otimes \cdots \otimes \mathbf{U}^{(n-1)}\right),$$

in which $\otimes$ represents the Kronecker product. Hence the columns of $\mathbf{U}^{(n)}$ can be found as an orthonormal basis for the dominant subspace of the n-mode space of $\tilde{\mathscr{A}}^{(n)}$. Repeating this procedure for different mode numbers leads to an alternating least squares (ALS) algorithm for the (local) maximization of $f(\hat{\mathscr{A}})$: in each step the estimate of one of the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ is optimized, while the other matrix estimates are kept constant. This technique is a higher-order extension of the orthogonal iteration for matrices [24].

It makes sense to initialize the higher-order orthogonal iteration with the truncated HOSVD. The HOSVD-estimate usually belongs to the attraction region of the best rank-$(R_1, R_2, \ldots, R_N)$ approximation, although there is no absolute guarantee of convergence to the global optimum [19].

Example 3. Fig. 3 visualizes the ALS algorithm for the best rank-1 approximation of a $(2 \times 2 \times 2)$-tensor. For different choices of θ_0, determining initial vectors $U^{(2)}_0 = U^{(3)}_0 = (\cos\theta_0 \;\; \sin\theta_0)^{\mathrm{T}}$, we plotted, after each iteration step k, the angle $\theta^{(3)}_k$

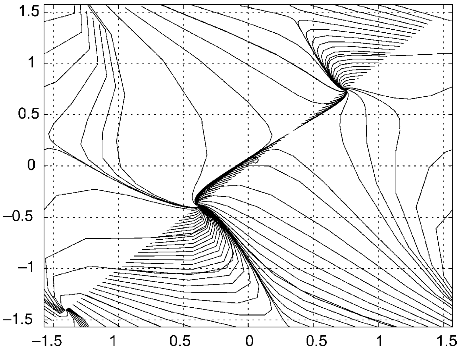

Fig. 3. Visualization of the ALS algorithm for the computation of the best rank-1 approximation of a supersymmetric tensor in $\mathbb{R}^{2\times 2\times 2}$. Abscis: the angle $\theta_k^{(2)}$ in $U_k^{(2)} = \left(\cos\theta_k^{(2)} \quad \sin\theta_k^{(2)}\right)^{\mathrm{T}}$ (in radians). Ordinate: the angle $\theta_k^{(3)}$ in $U_k^{(3)} = \left(\cos\theta_k^{(3)} \quad \sin\theta_k^{(3)}\right)^{\mathrm{T}}$ (in radians). Both angles are normalized to the interval $(-\pi/2, +\pi/2]$. The small circle shows the initial guess obtained by HOSVD. The global optimum is $(-0.3860, -0.3860)$.

in $U_k^{(3)} = (\cos\theta_k^{(3)} \quad \sin\theta_k^{(3)})^{\mathrm{T}}$ versus the angle $\theta_k^{(2)}$ in $U_k^{(2)} = (\cos\theta_k^{(2)} \quad \sin\theta_k^{(2)})^{\mathrm{T}}$. The angles were normalized to the interval $(-\pi/2, +\pi/2]$. The tensor $\mathscr{A}$ that we consider, is real supersymmetric (invariant for arbitrary index permutations) and defined by:

$$\begin{cases} a_{111} = 1.5578, & a_{222} = 1.1226, \\ a_{112} = -2.4443, & a_{221} = -1.0982. \end{cases}$$

We observe that the algorithm leads to unsymmetric intermediate results, not located on the main diagonal of the figure. We also remark that there are two stable ((−0.3860, −0.3860), (0.7413, 0.7413)) and two unstable ((−1.4052, −1.4052), (0.3347, 0.3347)) symmetric stationary points. The global optimum is (−0.3860, −0.3860).

Example 4. In this example tensors $\mathscr{A} \in \mathbb{R}^{10\times 10\times 10\times 10}$ are generated in the following way:

$$\mathscr{A} = \tilde{\mathscr{A}}/\|\tilde{\mathscr{A}}\| + \sigma_E \mathscr{E}/\|\mathscr{E}\|, \tag{13}$$

in which $\tilde{\mathscr{A}}$ is a rank-(2, 2, 2, 2) tensor:

$$\tilde{\mathscr{A}} = \mathscr{B} \times_1 \mathbf{M}^{(1)} \times_2 \mathbf{M}^{(2)} \times_3 \mathbf{M}^{(3)} \times_4 \mathbf{M}^{(4)}, \tag{14}$$

with $\mathscr{B} \in \mathbb{R}^{2\times 2\times 2\times 2}$ and $\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \mathbf{M}^{(3)}, \mathbf{M}^{(4)} \in \mathbb{R}^{10\times 2}$. All entries of $\mathscr{B}$, $\{\mathbf{M}^{(n)}\}$ and $\mathscr{E}$ are drawn from a zero-mean unit-variance Gaussian distribution. The second term in Eq. (13) is to be considered as an error (noise) term and σ_E^{-1} is the signal-to-noise ratio (SNR). We compute the HOSVD of $\mathscr{A}$ and truncate it after the second n-mode singular values, yielding an estimate $\hat{\mathscr{A}}_H \in \mathbb{R}^{10\times 10\times 10\times 10}$ and matrices $\{\mathbf{U}_H^{(n)}\} \in \mathbb{R}^{10\times 2}$. We also compute the best rank-(2, 2, 2, 2) approximation

Table 1
Norm of the HOSVD-truncate and the best rank-(2, 2, 2, 2) approximation in Example 4

SNR (dB)	$\|\hat{\mathscr{A}}_H\|$	$\|\hat{\mathscr{A}}_B\|$
20	0.9999900	0.9999901
10	1.000234	1.000247
0	1.00224	1.00357
−10	0.9476	1.0404

The variance over 100 Monte Carlo simulations has been taken into account in the number of digits that are displayed.

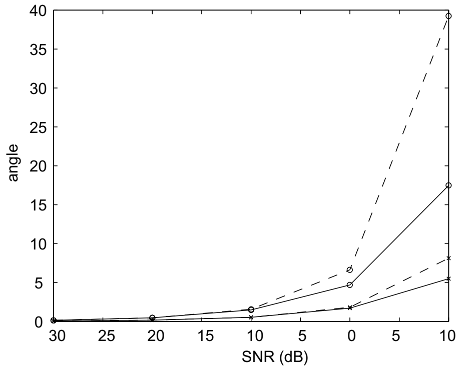

Fig. 4. Principal angles (in degrees) between the column spaces of $\mathbf{M}^{(1)}$ on one hand and $\mathbf{U}_H^{(1)}$ (dashed), $\mathbf{U}_B^{(1)}$ (solid) on the other hand, in Example 4. First principal angle: ×-marks; second principal angle: ○-marks.

of $\mathscr{A}$, yielding an estimate $\hat{\mathscr{A}}_B \in \mathbb{R}^{10\times10\times10\times10}$ and matrices $\{\mathbf{U}_B^{(n)}\} \in \mathbb{R}^{10\times2}$. (Although $\tilde{\mathscr{A}}$ is a rank-4 tensor, we do not calculate the best rank-4 approximation of $\mathscr{A}$, because this approximation is in general not of the form (14); the reason is that the n-mode factors of the rank-1 terms are not necessarily restricted to a subspace of dimension 2.) In Table 1 we show $\|\hat{\mathscr{A}}_H\|$ and $\|\hat{\mathscr{A}}_B\|$ for different SNR values, averaged over 100 Monte Carlo simulations. We see that the best rank-(2, 2, 2, 2) approximation is indeed able to explain more of the “energy” of $\mathscr{A}$ than the HOSVD-truncate, as put forward in Eq. (7).

In Fig. 4 we plot the average principal angles [13,44,45] between the column space of $\mathbf{M}^{(1)}$ on one hand and $\mathbf{U}_H^{(1)}$, $\mathbf{U}_B^{(1)}$ on the other hand. We observe that the best rank-(2, 2, 2, 2) approximation is more robust. The reason is that, in the computation of $\mathbf{U}_H^{(1)}$ by means of the SVD of $\mathbf{A}_{(1)}$, the structure in the row space of $\mathbf{A}_{(1)}$ is not taken into account. In the absence of noise we have

$$\mathbf{A}_{(1)} = \mathbf{M}^{(1)} \cdot \mathbf{B}_{(1)} \cdot \left(\mathbf{M}^{(2)} \otimes \mathbf{M}^{(3)} \otimes \mathbf{M}^{(4)}\right)^{\mathrm{T}}. \tag{15}$$

However, in the best rank-(2, 2, 2, 2) approximation we explicitly look for a tensor that has the same structure: Eq. (15) is a matrix formulation of the structure imposed by Eq. (6) for the dimensionalities in this example.

5. Higher-order statistics

From here on, we assume that the data are real-valued. The generalization to complex data is straightforward but more cumbersome from a notational point of view.

The basic HOS are higher-order moments and cumulants. For random vectors these quantities are higher-order tensors. We have the following definitions.

Definition 3 (*Moment*). The Nth-order moment of a real stochastic vector X is defined by

$$\mathscr{M}_X^{(N)} \overset{\text{def}}{=} E\{X \circ X \circ \cdots \circ X\}, \tag{16}$$

in which the expectation over the Nth-order outer product is performed in a component-wise manner.

The first-order moment is the mean of the stochastic vector. The second-order moment is the correlation matrix.

Definition 4 (*Cumulant*). For a real zero-mean stochastic vector X the cumulants up to order 4 are given by:

$$(C_X)_i = \mathrm{Cum}(x_i) \overset{\text{def}}{=} E\{x_i\}, \tag{17}$$

$$(\mathbf{C}_X)_{i_1 i_2} = \mathrm{Cum}(x_{i_1}, x_{i_2}) \overset{\text{def}}{=} E\{x_{i_1} x_{i_2}\}, \tag{18}$$

$$\left(\mathscr{C}_X^{(3)}\right)_{i_1 i_2 i_3} = \mathrm{Cum}(x_{i_1}, x_{i_2}, x_{i_3}) \overset{\text{def}}{=} E\{x_{i_1} x_{i_2} x_{i_3}\}, \tag{19}$$

$$\begin{aligned}\left(\mathscr{C}_X^{(4)}\right)_{i_1 i_2 i_3 i_4} = \mathrm{Cum}(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}) \overset{\text{def}}{=} {} & E\{x_{i_1} x_{i_2} x_{i_3} x_{i_4}\} \\ & - E\{x_{i_1} x_{i_2}\} E\{x_{i_3} x_{i_4}\} \\ & - E\{x_{i_1} x_{i_3}\} E\{x_{i_2} x_{i_4}\} \\ & - E\{x_{i_1} x_{i_4}\} E\{x_{i_2} x_{i_3}\}.\end{aligned} \tag{20}$$

For every component x_i of X that has a non-zero mean, x_i has to be replaced in these formulas, except Eq. (17), by $x_i - E\{x_i\}$.

The first-order cumulant is also equal to the mean of the stochastic vector. The second-order cumulant is the covariance matrix. The definition for arbitrary cumulant orders is complicated; it can be found in [20,33,34].

Some crucial properties are [33,34]:

1. *Multilinearity*: If a real stochastic vector X is transformed into a stochastic vector $\tilde{X}$ by a matrix multiplication $\tilde{X} = \mathbf{A} \cdot X$, with $\mathbf{A} \in \mathbb{R}^{J \times I}$, then we have:

$$\mathscr{M}_N^{\tilde{X}} = \mathscr{M}_N^{X} \times_1 \mathbf{A} \times_2 \mathbf{A} \cdots \times_N \mathbf{A}, \tag{21}$$

$$\mathscr{C}_N^{\tilde{X}} = \mathscr{C}_N^{X} \times_1 \mathbf{A} \times_2 \mathbf{A} \cdots \times_N \mathbf{A}. \tag{22}$$

2. *Even distribution*: If a real random variable x has an even probability density function $p_x(x)$, i.e., $p_x(x)$ is symmetric about the origin, then the odd moments and cumulants of x vanish.
3. *Partitioning of independent variables*: If in a set of I stochastic variables $x_1, x_2, \ldots, x_I$ a subset is independent of the other variables, then we have:

$$\mathrm{Cum}(x_1, x_2, \ldots, x_I) = 0. \tag{23}$$

This property does not hold in general for moments. A consequence of the property is that a higher-order cumulant of a stochastic vector having mutually independent components, is a diagonal tensor, i.e., only the entries of which all the indices are equal can be different from zero. This very strong algebraic condition is the basis of all algebraic ICA techniques.

4. *Sum of independent variables*: If the random variables $x_1, x_2, \ldots, x_I$ are mutually independent from the variables $y_1, y_2, \ldots, y_I$, then we have:

$$\begin{aligned}&\mathrm{Cum}(x_1 + y_1, x_2 + y_2, \ldots, x_k + y_k)\\ &\quad = \mathrm{Cum}(x_1, x_2, \ldots, x_k) + \mathrm{Cum}(y_1, y_2, \ldots, y_k).\end{aligned} \tag{24}$$

This property does not hold for moments either; as a matter of fact, it explains the term “cumulant”.

5. *Non-Gaussianity*: If Y is a Gaussian variable with the same mean and variance as a given stochastic variable X, then the following relation holds for $N \geqslant 3$:

$$\mathscr{C}_N^{X} = \mathscr{M}_N^{X} - \mathscr{M}_N^{Y}. \tag{25}$$

Higher-order cumulants of a Gaussian variable are 0.

The last three properties make that higher-order cumulants are used more frequently than higher-order moments.

Generally speaking, it becomes harder to estimate HOS from sample data as the order increases, i.e., longer datasets are required to obtain the same accuracy. Hence in practice the use of HOS is usually restricted to third- and fourth-order cumulants. For symmetric distributions fourth-order cumulants are used (cf. property 2).

6. Independent component analysis

Consider the following basic statistical model:

$$Y = \mathbf{M}X + E, \tag{26}$$

in which $Y \in \mathbb{R}^I$ is referred to as the *observation vector*, $X \in \mathbb{R}^R$ is called the *source vector* and $E \in \mathbb{R}^I$ represents additive *noise*. $\mathbf{M} \in \mathbb{R}^{I \times R}$ is called the *mixing matrix*; we assume that its columns are linearly independent. The goal of ICA consists of the estimation of the mixing matrix and/or the corresponding realizations of the source vector X, given only realizations of the observation vector Y. The key assumption is that the components of X are mutually statistically independent, as well as statistically independent from the noise components. This is a very strong hypothesis, but also quite natural in lots of applications. It means that the aim can often be rephrased as splitting the dataset into components "of a different nature", which contributed to the data in a linear way. Application areas include image processing, speech processing, telecommunications, biomedical problems, astrophysics, seismology, chemistry, data analysis, etc. [1,26]. The ICA-problem is also addressed in the literature under the labels *blind source separation* (BSS), *signal copy*, *waveform preserving estimation*, etc. However the precise assumptions on which the solution strategies are based, may sometimes differ from paper to paper.

ICA is subject to two basic indeterminacies. First, it is impossible to determine the norm of the columns of $\mathbf{M}$ in Eq. (26), since a rescaling of these vectors can be compensated by the inverse scaling of the source signal values. Similarly, the ordering of the source signals, having no physical meaning, cannot be identified. For non-Gaussian sources, these indeterminacies are the only way in which the factorization $\mathbf{M}X$ is not unique [10,39]. To guarantee uniqueness, we make in this paper the slightly stronger assumption that the source cumulants of a certain order are different from zero (cf. property 5 in Section 5).

The ICA-assumptions do not allow to distinguish between the signal and the noise term in Eq. (26). Hence the source signals will be estimated as $\hat{X}$, by a simple matrix multiplication:

$$\hat{X} = \mathbf{W}^{\mathrm{T}} Y. \tag{27}$$

As an example, $\mathbf{W}^{\mathrm{T}}$ can take the form of the pseudo-inverse $\hat{\mathbf{M}}^{\dagger}$, in which $\hat{\mathbf{M}}$ is an estimate of the mixing matrix (this choice of $\mathbf{W}^{\mathrm{T}}$ minimizes the interference to signal ratio (ISR)). More generally various beamforming strategies [47] can be applied (e.g. minimizing the interference-plus-noise to signal ratio (INSR)).

Using the properties listed in Section 5 we obtain:

$$\mathbf{C}_Y = \mathbf{C}_X \times_1 \mathbf{M} \times_2 \mathbf{M} + \mathbf{C}_E \tag{28}$$

$$\mathscr{C}_Y^{(N)} = \mathscr{C}_X^{(N)} \times_1 \mathbf{M} \times_2 \mathbf{M} \cdots \times_N \mathbf{M} + \mathscr{C}_E^{(N)} \tag{29}$$

in which $\mathbf{C}_X$ and $\mathscr{C}_X^{(N)}$ are diagonal, and in which $\mathscr{C}_E^{(N)}$ vanishes if the noise is Gaussian.

Different types of ICA-algorithms can be distinguished, depending on how Eqs. (28) and (29) are combined. In most algorithms as much information as possible is extracted from (28), and the remaining degrees of freedom are fixed by resorting to (29). These algorithms are called prewhitening-based.

The prewhitening step amounts to a principal component analysis (PCA) of the observations. Briefly, the goal is to transform the observation vector Y into another stochastic vector, Z, having unit covariance. This involves the multiplication of Y with the pseudo-inverse of a square root of its covariance matrix $\mathbf{C}_Y$. When $R < I$, a projection of Y onto the signal subspace is carried out.

Let us explain this in some more detail. Assuming that the source signals have unit variance (without loss of generality, as we may appropriately rescale the mixing vectors as well), we have:

$$\mathbf{C}_Y = \mathbf{M} \cdot \mathbf{M}^{\mathrm{T}}, \tag{30}$$

in which we have neglected the noise term at this point for clarity. A first observation is that the number of sources can be deduced from the rank of $\mathbf{C}_Y$. Substitution of the SVD of the mixing matrix $\mathbf{M} = \mathbf{USV}^{\mathrm{T}}$ shows that the EVD of the observed covariance allows to estimate the components $\mathbf{U}$ and $\mathbf{S}$ whilst the factor $\mathbf{V}$ remains unknown:

$$\mathbf{C}_Y = \mathbf{U} \cdot \mathbf{S}^2 \cdot \mathbf{U}^{\mathrm{T}} = (\mathbf{US}) \cdot (\mathbf{US})^{\mathrm{T}}. \tag{31}$$

The effect of the additive noise term E can be neutralized by replacing $\mathbf{C}_Y$ by the noise-free covariance $\mathbf{C}_Y - \mathbf{C}_E$, if the noise covariance $\mathbf{C}_E$ is known or can be estimated. If this is not possible, $\mathbf{C}_E$ should be considered as a perturbation of Eq. (30). In the case of spatially white noise, $\mathbf{C}_E$ takes the form of $\sigma_E^2\mathbf{I}$, in which σ_E^2 is the variance of the noise on each data channel. In a more-sensors-than-sources setup, σ_E^2 can be estimated as the mean of the "noise-eigenvalues", i.e., the smallest $I - R$ eigenvalues, of $\mathbf{C}_Y$. The number of sources is estimated as the number of significant eigenvalues of $\mathbf{C}_Y$; for a detailed procedure, we refer to [48].

Assuming that the noise is Gaussian, the unknown matrix $\mathbf{V}$ can now be obtained from

$$\mathscr{C}_Z^{(N)} = \mathscr{C}_X^{(N)} \times_1 \mathbf{V}^{\mathrm{T}} \times_2 \mathbf{V}^{\mathrm{T}} \cdots \times_N \mathbf{V}^{\mathrm{T}}, \tag{32}$$

in which the standardized random vector $Z \in \mathbb{R}^R$ is defined by $Z \overset{\text{def}}{=} \mathbf{S}^{\dagger} \cdot \mathbf{U}^{\mathrm{T}} \cdot Y$. Due to the multilinearity property the cumulants of Z and Y are related as follows:

$$\mathscr{C}_Z^{(N)} = \mathscr{C}_Y^{(N)} \times_1 (\mathbf{US})^{\dagger} \times_2 (\mathbf{US})^{\dagger} \cdots \times_N (\mathbf{US})^{\dagger}. \tag{33}$$

The fact that $\mathscr{C}_X^{(N)}$ is theoretically diagonal can be exploited in several ways. Examples of specific algebraic prewhitening-based ICA-methods are [7,10,21]. A more general reference is [26].

In this paper we focus on the dimensionality reduction when $I \gg R$, i.e., when there are many more observation channels than (significant) sources. Applications can be found in electro-encephalography (EEG), magneto-encephalography (MEG), nuclear magnetic resonance (NMR), hyper-spectral image processing, data analysis, etc.

In contrast to $\mathscr{C}_Y^{(N)}$, $\mathscr{C}_Z^{(N)}$ is a low-dimensional $(R \times R \times \cdots \times R)$ tensor. This is important, because the multilinear algebraic techniques that are used to estimate $\mathbf{V}$ from Eq. (32) are generally much more computationally demanding than classical matrix techniques.

A drawback of the prewhitening-based approach is that errors introduced in the prewhitening step (due to the presence of (coloured) noise, limited sample size, etc.) cannot be compensated in the higher-order step. Prewhitening errors induce a bound on the overall performance, as explained in [8,17,23].

7. Higher-order-only ICA

In the previous section, we explained that in prewhitening-based ICA the PCA-step has a three-fold goal: (a) reduction of the parameter set of unknowns to the manifold of orthogonal matrices, (b) standardization of the unknown source signals to mutually uncorrelated unit-variance signals, and (c) determination of the number of sources. This scheme has the disadvantage that it is affected by additive (coloured) Gaussian noise.

However, it is possible to identify the mixing matrix by using *only* the higher-order cumulant tensor—Eq. (28) is not strictly necessary to find the solution. For algebraic techniques, we refer to [6,11,14]. Such higher-order-only methods have the interesting feature that they allow to boost signal-to-noise ratios when the noise is Gaussian. Although there is no prewhitening stage, one may still want to reduce the dimensionality of the higher-order cumulant in the more-sensors-than-sources case, as this may significantly reduce the computational load of the actual algorithm; on the other hand, the estimation of the higher-order cumulant itself may form an important part of the global load that cannot be avoided. In this section we will investigate how the dimensionality reduction can be achieved.

Due to the diagonality of $\mathscr{C}_X^{(N)}$, Eq. (29) can up to the noise term be rewritten as

$$\mathscr{C}_Y^{(N)} = \sum_{r=1}^{R} \kappa_r^X M_r \circ M_r \circ \cdots \circ M_r, \tag{34}$$

in which κ_r^X represents the marginal Nth-order cumulant of the rth source. Eq. (34) is a decomposition of $\mathscr{C}_Y^{(N)}$ in a minimal number of rank-1 terms, as the columns of $\mathbf{M}$ are assumed to be linearly independent. It can even be proved that, under the conditions specified in Section 6 (including $R \leqslant I$), the decomposition is unique up to some trivial indeterminacies [22,32]. As a consequence, the aim of higher-order-only ICA can be formulated as the computation of a rank-revealing decomposition of $\mathscr{C}_Y^{(N)}$, taking into account that the sample cumulant equivalent of Eq. (34) may be perturbated by non-Gaussian noise components, finite datalength effects, model misfit, etc.

Every tensor that satisfies Eq. (34) is by definition a rank-R tensor. However, according to the definition of n-rank, such a tensor is also rank-$(R, R, \ldots, R)$. The reason is that every n-mode vector can be written as a linear combination of the R vectors $\{M_r\}$, i.e., the n-mode vector space is R-dimensional. So, to deal with the situation in which $I > R$, we can first project the sample cumulant $\hat{\mathscr{C}}_Y^{(N)}$ on the manifold of rank-$(R, R, \ldots, R)$ tensors, using the techniques explained in Sections 3 and 4. In a subsequent step, we can address the harder problem of the further projection on the submanifold of rank-R tensors and the actual computation of decomposition (34). The latter problem can then be solved in a lower-dimensional space. Note that in the second-order case both projections coincide, as n-rank and rank are necessarily the same. Hence, this kind of preprocessing is only possible because we work in a tensor framework, instead of in a classical vector/matrix framework.

Example 5. Let us consider the following numerical experiment. Data are generated according to the following model:

$$Y = \mathbf{M}_1 X + \mathbf{M}_2 \sigma_E E,$$

in which the entries of $X \in \mathbb{R}^4$ are drawn from a binary distribution with equal probabilities for +1 and -1, and in which $E \in \mathbb{R}^{12}$ is zero-mean unit-variance Gaussian noise. $\mathbf{M}_1 \in \mathbb{R}^{12\times 4}$ and $\mathbf{M}_2 \in \mathbb{R}^{12\times 12}$ are random matrices of which the columns have been normalized to unit length. The data length is 500. A Monte Carlo experiment consisting of 500 runs is carried out for different values of the noise variance σ_E^2.

In each run, the matrix $\mathbf{M}_1$ is estimated from the fourth-order cumulant of Y by first reducing the dimensionality of the problem from 12 to 4, and subsequently matching both sides of Eq. (34) in the least-squares sense by means of the technique described in [5]. This algorithm was initialized with the starting value proposed in [32] and with 9 randomly chosen starting values. The best result was retained. The dimensionality reduction was achieved (1) by means of a simple HOSVD truncation and (2) by calculating the best rank-(4, 4, 4, 4) approximation in the way described in Section 4.

Let the estimate of $\mathbf{M}_1$ be represented by $\hat{\mathbf{M}}_1$ and let the columns of $\hat{\mathbf{M}}_1$ be normalized to unit length and optimally ordered. Then our error measure is defined as the mean of the squared off-diagonal entries of $\hat{\mathbf{M}}_1^{\dagger} \cdot \mathbf{M}_1$; this error measure can be interpreted as an approximate average ISR.

Only the results for which the ISR is smaller than 0.04 are retained; the other results are considered as failures. A failure means that 10 initializations were not enough or that the estimate of the low-dimensional cumulant was simply too bad to get sufficiently accurate results.

The results are listed in Table 2. One can see that calculating the best rank-(4, 4, 4, 4) approximation is more reliable than simple truncation of the HOSVD.

Table 2
Mean m and variance σ^2 of the ISR, and number of successful runs s in Example 5

σ_E^2 (dB)	HOSVD truncation			Best approximation		
	m_H	σ_H^2	s_H	m_B	σ_B^2	s_B
−8	0.0041	7.6e−6	499	0.0039	5.7e−6	499
−7	0.0047	8.9e−6	495	0.0045	8.2e−6	495
−6	0.0066	2.1e−5	481	0.0059	1.6e−5	481
−5	0.0098	5.0e−5	424	0.0083	3.4e−5	440
−4	0.0138	6.5e−5	285	0.0114	5.7e−5	331

8. ICA based on soft whitening

In the prewhitening-based procedure, one computes more than half of the parameters in the SVD of $\mathbf{M}$ from the matrix decomposition (30), which is exactly satisfied. The tensor decomposition (29), which involves many more constraints, as can be verified by counting the degrees of freedom, serves to estimate the factor $\mathbf{V}$; these constraints are only approximately satisfied. Contrarily, in the higher-order-only scheme Eq. (30) is not used at all. It is intuitively clear that, instead, it is preferable to deal with decompositions (28) and (29) in a more balanced way. We call this principle "soft whitening". The idea was first proposed and tested in [49]. In this section we will discuss dimensionality reduction in the context of soft whitening.

The problem can now be formulated as the determination of a matrix $\mathbf{M} \in \mathbb{R}^{I\times R}$ that minimizes the cost function

$$\begin{aligned}\tilde{f}(\mathbf{M}) = {} & w_1^2\|\hat{\mathbf{C}}_Y - \hat{\mathbf{C}}_X \times_1 \mathbf{M} \times_2 \mathbf{M}\|^2 \\ & + w_2^2\left\|\hat{\mathscr{C}}_Y^{(N)} - \hat{\mathscr{C}}_X^{(N)} \times_1 \mathbf{M} \times_2 \mathbf{M} \cdots \times_N \mathbf{M}\right\|^2, \end{aligned} \qquad (35)$$

in which $\hat{\mathbf{C}}_X \in \mathbb{R}^{R\times R}$ is an unknown diagonal matrix and $\hat{\mathscr{C}}_X^{(N)} \in \mathbb{R}^{R\times R\times\cdots\times R}$ an unknown diagonal tensor, and in which $\hat{\mathbf{C}}_Y$ and $\hat{\mathscr{C}}_Y^{(N)}$ are the sample estimates of the covariance matrix and the cumulant tensor of the signal part of Y (a noise compensation, as described for PCA earlier, may have been carried out first). w_1 and w_2 are positive parameters that have to be tuned. A big ratio w_1/w_2 reflects that one has much confidence in the estimate $\hat{\mathbf{C}}_Y$ and little confidence in the estimate $\hat{\mathscr{C}}_Y^{(N)}$ (e.g. short dataset, low noise level); the opposite reflects that one considers $\hat{\mathscr{C}}_Y^{(N)}$ as more reliable than $\hat{\mathbf{C}}_Y$ (e.g. unknown coloured Gaussian noise).

Both the column space of $w_1\hat{\mathbf{C}}_Y$ and the 1-mode vector space of $w_2\hat{\mathscr{C}}_Y^{(N)}$ are theoretically equal to the column space of $\mathbf{M}$. Hence, in comparison with Section 7, it is natural to replace the computation of the subspace of the dominant left singular vectors of the matrix unfolding $\big(\hat{\mathbf{C}}_Y^{(N)}\big)_{(1)}$ (in the truncation of the HOSVD of $\hat{\mathscr{C}}_Y^{(N)}$) by the computation of the subspace of the dominant left singular vectors of the matrix $\big[w_1\hat{\mathbf{C}}_Y \;\; w_2\big(\hat{\mathbf{C}}_Y^{(N)}\big)_{(1)}\big]$.

The ALS approach in Section 4 can easily be modified as follows. In iteration step k we compute now a column-wise orthonormal matrix $\mathbf{X}^{(k)}$ of which the column space is equal to the space generated by the dominant left singular vectors of the matrix containing all the columns of $w_1\hat{\mathbf{C}}_Y\mathbf{X}^{(k-1)}$ and all the 1-mode vectors of $w_2\hat{\mathscr{C}}_Y^{(N)} \times_2 \mathbf{X}^{(k-1)^{\mathrm{T}}} \times_3 \mathbf{X}^{(k-2)^{\mathrm{T}}} \cdots \times_N \mathbf{X}^{(k-N+1)^{\mathrm{T}}}$.

9. Dimensionality reduction for simultaneous matrix diagonalization

Simultaneous diagonalization of a set of matrices has become an important signal processing tool in the last decade. For instance, many variants of ICA and BSS are based on diagonalization by means of a simultaneous congruence transformation [3,4,7,35]. Given a set of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_K \in \mathbb{R}^{I\times I}$, the aim is to find a non-singular matrix $\mathbf{M} \in \mathbb{R}^{I\times R}$ such that, in theory,

$$\begin{aligned} \mathbf{A}_1 &= \mathbf{M}\cdot\mathbf{D}_1\cdot\mathbf{M}^{\mathrm{T}} \\ &\vdots \\ \mathbf{A}_K &= \mathbf{M}\cdot\mathbf{D}_K\cdot\mathbf{M}^{\mathrm{T}}, \end{aligned} \tag{36}$$

with $\mathbf{D}_1, \ldots, \mathbf{D}_K \in \mathbb{R}^{R\times R}$ diagonal. In the presence of noise, the difference between the left- and right-hand side of Eqs. (36) has to be minimized. Examples are the JADE algorithm for the separation of non-Gaussian sources [7], the SOBI algorithm for the separation of sources that are mutually uncorrelated but individually exhibit some correlation in time [3], the algorithm proposed in [35] for the separation of non-stationary sources subject to a constant mixing, etc. In JADE, one of the given matrices is the observed covariance matrix; the corresponding diagonal matrix is the covariance of the sources and the other given matrices are matrix “slices” of the higher-order cumulant of the observations. In SOBI, the matrices $\{\mathbf{A}_k\}$ and the matrices $\{\mathbf{D}_k\}$ are the covariance matrices for different time lags, of the observations and the sources, respectively. In [35] they correspond to covariance matrices measured at different time instances.

The original algorithms to solve (36) are prewhitening-based. In the prewhitening step, one picks a positive (semi-)definite matrix from $\{\mathbf{A}_k\}$, say $\mathbf{A}_1$, and computes its EVD. In the same way as explained in Section 6, this allows to reduce the other equations to a simultaneous unitary diagonalization in a possibly lower-dimensional space. The latter problem can be solved by means of the techniques developed in [3,7,9].

On the other hand, one can also follow the soft whitening approach and solve the different equations in (36) simultaneously, instead of sequentially. In this section we will explain how a dimensionality reduction can be realized here, when $R < I$. For the matrix diagonalization in the lower-dimensional space, one can resort to the techniques developed in [16,25,36,42,43,49].

To see the link with the preceding discussion, let us stack the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_K$ in Eq. (36) in a tensor $\mathscr{A} \in \mathbb{C}^{I \times I \times K}$. Define a matrix $\mathbf{D} \in \mathbb{C}^{K \times R}$ as follows:

$$\mathbf{D} = \begin{pmatrix} \mathrm{diag}(\mathbf{D}_1) \\ \vdots \\ \mathrm{diag}(\mathbf{D}_K) \end{pmatrix}, \tag{37}$$

in which diag$(\cdot)$ is the operator that extracts the diagonal from its argument and puts it in a row. Then we have that

$$\mathscr{A} = \sum_{r=1}^{R} M_r \circ M_r \circ D_r. \tag{38}$$

Let the rank of $\mathbf{D}$ be equal to R_3. Eq. (38) is a decomposition of $\mathscr{A}$ in a minimal sum of rank-1 terms (if no columns of $\mathbf{D}$ are collinear [31]); this problem and the simultaneous diagonalization problem are equivalent. Hence, we can proceed in the same way as in Section 7. The dimensionality reduction can be realized by a rank-(R, R, R_3) reduction of $\mathscr{A}$; the remaining problem is the decomposition of an $(R \times R \times R_3)$-tensor in rank-1 terms.

10. Subspace variants of HOS-based algorithms

In [2] subspace processing is presented as a means to reduce the computational burden of estimating and/or processing HOS and as a way to decrease their variance. Instead of working with the original random variables, one manipulates their projections on some subspace. The application domain is wider than just ICA; the paper contains an example of an application in system identification. By way of illustration, we briefly describe this application. The goal is the estimation of the impulse response of a linear system of which the input and output are contaminated by additive Gaussian noise. The presence of Gaussian noise introduces a bias into solutions based on second-order statistics; instead, the solution is obtained from HOS of the input and cross-HOS of input and output. The vector X below contains a frame of subsequent input values. For many signals it is reasonable to assume that X is restricted to a certain subspace. In some cases one has enough prior knowledge on the input signal to determine this subspace in advance. For more details we refer to [2].

We will first briefly explain the general idea in tensor terms and then present some modifications and further results. Our explanation is in terms of cumulants; other HOS can be treated in the same way.

Let X be an I-dimensional random vector and Z its projection in an R-dimensional subspace, spanned by the columns of a column-wise orthonormal ($I \times R$) matrix $\mathbf{U}$:

$$Z = \mathbf{U}^{\mathrm{T}} \cdot X. \tag{39}$$

Due to the multilinearity property, the cumulants of X and Z are related by

$$\mathscr{C}_Z^{(N)} = \mathscr{C}_X^{(N)} \times_1 \mathbf{U}^{\mathrm{T}} \times_2 \mathbf{U}^{\mathrm{T}} \cdots \times_N \mathbf{U}^{\mathrm{T}}. \tag{40}$$

Processing $\mathscr{C}_Z^{(N)}$ instead of $\mathscr{C}_X^{(N)}$ may lead to a tremendous reduction in computational complexity. The ratio of their number of entries is $\left(\frac{R}{I}\right)^N$. If $\frac{R}{I} = 0.1$ and $N = 4$, then the subspace approach offers a factor of 10,000 fewer statistics.

Let us now project Z back into the full I-dimensional space:

$$Y \overset{\text{def}}{=} \mathbf{U} \cdot Z = \mathbf{U} \cdot \mathbf{U}^{\mathrm{T}} \cdot X \overset{\text{def}}{=} \mathbf{P} \cdot X, \tag{41}$$

in which $\mathbf{P}$ is a projection matrix. Due to the multilinearity property, we have

$$\mathscr{C}_Y^{(N)} = \mathscr{C}_Z^{(N)} \times_1 \mathbf{U} \times_2 \mathbf{U} \cdots \times_N \mathbf{U}, \tag{42}$$

$$= \mathscr{C}_X^{(N)} \times_1 \mathbf{P} \times_2 \mathbf{P} \cdots \times_N \mathbf{P}. \tag{43}$$

By definition, $\mathscr{C}_Y^{(N)}$ is at most rank-($R, R, \ldots, R$).

Let $\hat{\mathscr{C}}_X^{(N)}$ be a sample estimate of $\mathscr{C}_X^{(N)}$. Working with Z instead of X, amounts to replacing $\hat{\mathscr{C}}_X^{(N)}$ by the rank-($R, R, \ldots, R$) estimator

$$\hat{\mathscr{C}}_Y^{(N)} \overset{\text{def}}{=} \hat{\mathscr{C}}_X^{(N)} \times_1 \mathbf{P} \times_2 \mathbf{P} \cdots \times_N \mathbf{P}. \tag{44}$$

In [2] it is proved that the mean squared error (MSE) of this estimator is equal to the sum of a bias and a variance term:

$$\left\|\mathscr{C}_X^{(N)} - \hat{\mathscr{C}}_Y^{(N)}\right\|^2 = \operatorname{bias}^2\left(\hat{\mathscr{C}}_Y^{(N)}\right) + \operatorname{var}\left(\hat{\mathscr{C}}_Y^{(N)}\right), \tag{45}$$

with

$$\operatorname{bias}^2\left(\hat{\mathscr{C}}_Y^{(N)}\right) = \left\|\mathscr{C}_X^{(N)} - \mathscr{C}_Y^{(N)}\right\|^2, \tag{46}$$

$$\begin{aligned}\operatorname{var}\left(\hat{\mathscr{C}}_Y^{(N)}\right) &= \left\|\mathscr{C}_Y^{(N)} - \hat{\mathscr{C}}_Y^{(N)}\right\|^2 \\ &= \left\|\left(\mathscr{C}_X^{(N)} - \hat{\mathscr{C}}_X^{(N)}\right) \times_1 \mathbf{P} \times_2 \mathbf{P} \cdots \times_N \mathbf{P}\right\|^2.\end{aligned} \tag{47}$$

It is shown in [2] that, for low SNR, the ratio $\operatorname{var}\left(\hat{\mathscr{C}}_Y^{(N)}\right)/\operatorname{var}\left(\hat{\mathscr{C}}_X^{(N)}\right)$ is of the order of magnitude of $\left(\frac{R}{I}\right)^N$. Hence, subspace processing may not only reduce the number of statistics that have to be estimated (if $\mathbf{U}$ is known beforehand) and processed, but also decrease the variance of the estimator. An intuitive explanation is that, if the entries of $\mathscr{C}_X^{(N)} - \hat{\mathscr{C}}_X^{(N)}$ are of the same order of magnitude, the projection will lead to a rank-($R, R, \ldots, R$) tensor of which the squared Frobenius-norm is roughly proportional to the number of entries. (The same observation can be made for matrices.)

The price that has to be paid, is a possible increase of the bias (46). However, globally, the MSE of the rank-$(R, R, \dots, R)$ estimator can be significantly smaller than that of the full rank-$(I, I, \dots, I)$ estimator.

In some applications one has prior knowledge of the true cumulant $\mathscr{C}_X^{(N)}$. In [2] it is proposed to keep the bias low in such cases by choosing the columns of $\mathbf{U}$ equal to the dominant left singular vectors of the matrix unfolding $\left(\mathbf{C}_X^{(N)}\right)_{(1)}$. This corresponds in fact to a truncation of the HOSVD of $\left(\mathbf{C}_X^{(N)}\right)_{(1)}$. As we have seen earlier, this approach is suboptimal. The error is only bounded as in Eq. (9). Instead minimization of (46) corresponds to the determination of the best rank-$(R, R, \dots, R)$ approximation of $\mathscr{C}_X^{(N)}$, as discussed in Section 4.

Moreover, even when $\mathscr{C}_X^{(N)}$ is not known, it can make sense to do a dimensionality reduction. In this case, one has to compute the estimate $\hat{\mathscr{C}}_X^{(N)}$ (there is no computational gain here), and then determine its rank-$(R, R, \dots, R)$ approximation. Appropriate values of R can be chosen by inspection of the 1-mode singular value spectrum. The gain is a reduction of the variance and the fact that the actual signal-processing algorithm is based on the manipulation of a smaller number of statistics.

The results presented in this section should be seen in the right perspective. In [2] it is shown that the cumulant estimator (for different values of R) that has the smallest MSE, does not necessarily yield the smallest MSE for the output of a specific signal processing algorithm. This depends on the way the HOS are processed by the algorithm. In this context, the techniques discussed in this paper can be seen as tools that can help to improve the performance.

11. Conclusion

Due to their multi-way character the number of data in higher-order signal processing can quickly become very high. Together with the fact that multilinear algebraic techniques are generally more expensive than conventional matrix techniques, this may lead to an unacceptable computational load. However, important computational savings may result from performing a dimensionality reduction in a preprocessing step. This may for instance be relevant for ICA when it involves a high number of sensors and a low number of sources, which is common in several applications. Moreover, due to a lower variance low-dimensional estimators are often more accurate than high-dimensional estimators. The HOSVD and the best rank-$(R_1, R_2, \dots, R_N)$ approximation of higher-order tensors are important tools for realizing the dimensionality reduction.

Acknowledgements

Lieven De Lathauwer holds a permanent research position with the French Centre National de la Recherche Scientifique (CNRS); he also holds a honorary research position with the K.U.Leuven. Joos Vandewalle is Full Professor with the K.U. Leuven.

This work is supported by several institutions:

1. Research Council K.U.Leuven: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology).
2. Flemish Government: the F.W.O. project G.0240.99 (Multilinear Generalisations of the Singular Value Decomposition and Applications in Signal Processing and System Identification), the F.W.O. Research Communities ICCoS (Identification and Control of Complex Systems) and ANMMM (Advanced Numerical Methods for Mathematical Modelling).
3. Belgian Federal Government: Interuniversity Poles of Attraction Programme IUAP IV-02 (1997–2001): Modeling, Identification, Simulation and Control of Complex Systems, and IUAP V-22 (2002–2006): Dynamical Systems and Control: Computation, Identification & Modelling.

The scientific responsibility is assumed by the authors.

References

[1] S.-I. Amari, A. Cichocki, S. Makino, N. Murata, in: Proc. Fourth Int. Symp. on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 2003.

[2] T.F. André, R.D. Nowak, B.D. Van Veen, Low-rank estimation of higher order statistics, IEEE Trans. Signal Process. 45 (1997) 673–685.

[3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second order statistics, IEEE Trans. Signal Process. 45 (2) (1997) 434–444.

[4] A. Belouchrani, M.G. Amin, Blind source separation based on time-frequency signal representations, IEEE Trans. Signal Process. 46 (11) (1998) 2888–2897.

[5] R. Bro, PARAFAC: Tutorial & applications, in: 2nd Internet Conf. in Chemometrics (INCINC'96), Chemom. Intell. Lab. Syst., vol. 38, 1997, pp. 149–171 (special issue).

[6] J.-F. Cardoso, Iterative techniques for blind source separation using only fourth-order cumulants, in: Proc. EUSIPCO-92, Brussels, Belgium, 1992, pp. 739–742.

[7] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, IEE Proc.-F 140 (1994) 362–370.

[8] J.-F. Cardoso, On the performance of orthogonal source separation algorithms, in: Proc. EUSIPCO-94, Edinburgh, Scotland, UK, vol. 2, 1994, pp. 776–779.

[9] J.-F. Cardoso, A. Souloumiac, Jacobi angles for simultaneous diagonalization, SIAM J. Matrix Anal. Appl. 17 (1) (1996) 161–164.

[10] P. Comon, Independent component analysis, a new concept, Higher Order Statistics, Signal Process. 36 (1994) 287–314 (special issue).

[11] P. Comon, B. Mourrain, Decomposition of quantics in sums of powers of linear forms, Higher Order Statistics, Signal Process. 53 (2–3) (1996) 93–108 (special issue).

[12] P. Comon, Tensor decompositions, in: J.G. McWhirter, I.K. Proudler (Eds.), Mathematics in Signal Processing V, Clarendon Press, Oxford, UK, 2002.

[13] K. De Cock, Principal angles in system theory, information theory and signal processing, Ph.D. Thesis, K.U.Leuven, E.E. Dept. (ESAT), Belgium, 2002.

[14] L. De Lathauwer, B. De Moor, J. Vandewalle, Independent component analysis based on higher-order statistics only, in: Proc. 8th IEEE SP Workshop on Statistical Signal and Array Processing (SSAP-96), Corfu, Greece, 1996, pp. 356–359.

[15] L. De Lathauwer, B. De Moor, J. Vandewalle, Dimensionality reduction in higher-order-only ICA, in: Proc. IEEE Signal Processing Workshop on Higher-Order Statistics, Banff, Alta., Canada, 1997, pp. 316–320.

[16] L. De Lathauwer, Signal processing based on multilinear algebra, Ph.D. Thesis, K.U.Leuven, E.E. Dept. (ESAT), Belgium, 1997.

[17] L. De Lathauwer, J. Vandewalle, A residual bound for the mixing matrix in ICA, in: Proc. EUSIPCO-98, Rhodos, Greece, 1998, pp. 2065–2068.

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

[19] L. De Lathauwer, B. De Moor, J. Vandewalle, On the best rank-1 and rank-($R_1, R_2, \dots, R_N$) approximation of higher-order tensors, SIAM J. Matrix Anal. Appl. 21 (2000) 1324–1342.

[20] L. De Lathauwer, B. De Moor, J. Vandewalle, An introduction to independent component analysis, Multi-way Analysis, J. Chemom. 14 (2000) 123–149 (special issue).

[21] L. De Lathauwer, B. De Moor, J. Vandewalle, Independent component analysis and (simultaneous) third-order tensor diagonalization, IEEE Trans. Signal Process. 49 (2001) 2262–2271.

[22] L. De Lathauwer, B. De Moor, J. Vandewalle, Computation of the canonical decomposition by means of a simultaneous generalized Schur decompositition, SIAM J. Matrix Anal. Appl., in press.

[23] L. De Lathauwer, B. De Moor, J. Vandewalle, A prewhitening-induced bound on the identification error in independent component analysis, Tech. Report 01-12, K.U.Leuven, E.E. Dept. (ESAT), SCD-SISTA, Belgium, 2001.

[24] G.H. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, MD, 1996.

[25] R.A. Harshman, Foundations of the PARAFAC procedure: model and conditions for an "explanatory" multi-mode factor analysis, UCLA Working Papers in Phonetics 16(1) (1970).

[26] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[27] E. Kofidis, P. Regalia, On the best rank-1 approximation of higher order supersymmetric tensors, SIAM J. Matrix Anal. Appl. 23 (2002) 863–884.

[28] T. Kolda, Orthogonal tensor decompositions, SIAM J. Matrix Anal. Appl. 23 (2001) 243–255.

[29] P.M. Kroonenberg, J. de Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 45 (1980) 69–97.

[30] P.M. Kroonenberg, Three-Mode Principal Component Analysis, DSWO Press, Leiden, 1983.

[31] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra Appl. 18 (1977) 95–138.

[32] S.E. Leurgans, R.T. Ross, R.B. Abel, A decomposition for three-way arrays, SIAM J. Matrix Anal. Appl. 14 (1993) 1064–1083.

[33] C. Nikias, J. Mendel, Signal processing with higher-order spectra, IEEE Signal Proc. Mag. (1993) 10–37.

[34] C.L. Nikias, A.P. Petropulu, Higher-Order Spectra Analysis. A Nonlinear Signal Processing Framework, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[35] D.-T. Pham, J.-F. Cardoso, Blind separation of instantaneous mixtures of non-stationary sources, IEEE Trans. Signal Process. 49 (9) (2001) 1837–1848.

[36] D.-T. Pham, Joint approximate diagonalization of positive definite Hermitian matrices, SIAM J. Matrix Anal. Appl. 22 (4) (2001) 1136–1152.

[37] N. Sidiropoulos, G. Giannakis, R. Bro, Blind PARAFAC receivers for DS-CDMA systems, IEEE Trans. Signal Process. 48 (2000) 810–823.

[38] J. ten Berghe, J. de Leeuw, P.M. Kroonenberg, Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms, Psychometrika 52 (1987) 183–191.

[39] L. Tong et al., Indeterminacy and identifiability of blind identification, IEEE Trans. Circuits Systems 38 (1991) 499–509.

[40] L.R. Tucker, The extension of factor analysis to three-dimensional matrices, in: H. Gulliksen, N. Frederiksen (Eds.), Contributions to Mathematical Psychology, Holt, Rinehart & Winston, NY, 1964, pp. 109–127.

[41] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.

[42] A.-J. van der Veen, A. Paulraj, An analytical constant modulus algorithm, IEEE Trans. Signal Process. 44 (1996) 1136–1155.

[43] A.-J. van der Veen, Joint diagonalization via subspace fitting techniques, in: Proc. IEEE ICASSP, Salt Lake City, UT, 2001.

[44] J. Vandewalle, D. Callaerts, Singular value decomposition: a powerful concept and tool in signal processing, Mathematics in Signal Processing II, Clarendon Press, Oxford, 1990.

[45] J. Vandewalle, B. De Moor, On the use of the singular value decomposition in identification and signal processing, in: Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, NATO ASI Series, vol. F70, 1991, pp. 321–360.

[46] S. Van Huffel, J. Vandewalle, The Total Least Squares Problem: Computational Aspects and Analysis, Frontiers in Applied Mathematics Series, vol. 9, SIAM, Philadelphia, PA, 1991.

[47] B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. (1988) 4–24.

[48] M. Wax, T. Kailath, Detection of signals by information theoretic criteria, IEEE Trans. Acoust. Speech Signal Process. 33 (1986) 387–392.

[49] A. Yeredor, Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation, IEEE Trans. Signal Process. 50 (7) (2002) 1545–1553.