JOURNAL OF CHEMOMETRICS
J. Chemometrics 2000; **14**: 123–149

An introduction to independent component analysis

Lieven De Lathauwer*[†], Bart De Moor and Joos Vandewalle

KU Leuven, EE Department (ESAT), SISTA/COSIC, Kard. Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium

SUMMARY

This paper is an introduction to the concept of independent component analysis (ICA) which has recently been developed in the area of signal processing. ICA is a variant of principal component analysis (PCA) in which the components are assumed to be mutually statistically independent instead of merely uncorrelated. The stronger condition allows one to remove the rotational invariance of PCA, i.e. ICA provides a meaningful unique bilinear decomposition of two-way data that can be considered as a linear mixture of a number of independent source signals. The discipline of multilinear algebra offers some means to solve the ICA problem. In this paper we briefly discuss four orthogonal tensor decompositions that can be interpreted in terms of higher-order generalizations of the symmetric eigenvalue decomposition. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: multilinear algebra; eigenvalue decomposition; principal component analysis; independent component analysis; higher-order statistics

1. INTRODUCTION

This paper is intended to provide an introduction to a fundamental issue that has received an increasing amount of attention from the signal-processing research community in the last decade, namely the concept of independent component analysis (ICA), also known as blind source separation (BSS). Disciplines involved are statistics, neural networks, pattern recognition, information theory, system identification, etc. [1,2]. In this contribution we have to limit ourselves to the algebraic approach: in a natural way, ICA poses the question of generalizations of matrix algebraic techniques to multilinear algebra, i.e. the algebra of 'multiway matrices' or 'higher-order tensors'. A second objective of the paper is to give a brief overview of a class of orthogonal tensor decompositions that can be interpreted as higher-order counterparts of the symmetric matrix eigenvalue decomposition (EVD). Like e.g. the EVD and the singular value decomposition (SVD) of matrices, these decompositions can be considered as tools useful for a wide range of applications.

In a nutshell, the goal of ICA is the decomposition of a set of multisensor data in an *a priori* unknown linear mixture of *a priori* unknown source signals, relying on the assumption that the source signals are mutually statistically independent. This concept is in fact a fine-tuning of the well-known principal component analysis (PCA), where one aims at the decomposition in a linear mixture of

* Correspondence to: L. De Lathauwer, KU Leuven, EE Department (ESAT), SISTA/COSIC, Kard. Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium.

[†] E-mail: lieven.delathauwer@esat.kuleuven.ac.be

Contract/grant sponsor: Flemish Government; Contract/grant number: G.0240.99; G.0256.97
Contract/grant sponsor: Belgian State, Prime Minister's Office; Contract/grant number: IUAP P4-02; IUAP P4-24.

uncorrelated components—the weaker condition resulting in a rotational indeterminacy of the solution. To overcome this indeterminacy, the crucial observation is that statistical independence not only imposes constraints on the covariance of the sources, but also involves their higher-order statistics (HOS); the concept of HOS was introduced in References [3,4]. ICA and PCA are not only related from a statistical point of view, but also from a computational perspective, as they both rely on an EVD-type decomposition, in linear and multilinear algebra respectively.

From an algorithmic point of view, three approaches are possible: (i) first a PCA is carried out and subsequently the remaining degree of freedom is fixed by resorting to HOS; (ii) the solution is directly obtained from the HOS while avoiding the use of second-order statistics; or (iii) second- and higher-order statistics are exploited in a combined way. Each approach has its pros and cons; the main point of difference is that working with HOS has the advantage that Gaussian noise can be suppressed to some extent, but on the other hand it requires longer data sets than needed for second-order calculations. A second observation is that, at present, the three methods of working have not yet been studied to the same depth. For this paper we choose to focus on the first type of procedure. For bibliographic pointers related to the other approaches we refer to Reference [2].

We begin with a brief exposition of the required preliminary material of multilinear algebra and HOS in Section 2. Section 3 discusses ICA in conceptual terms. Subsequently we give a formal problem definition, analyze the mechanism of PCA-based ICA routines, discuss the issue of identifiability, provide some measures of performance and make a comparison between PCA and ICA. Section 4 illustrates the ideas with a conceptual example. Subsequently, Section 5 briefly discusses four algebraic algorithms. Their performance is illustrated by means of a number of numerical experiments at the end of the section.

Let us finally enumerate some notational conventions. Vectors are written as bold lower-case letters, matrices as bold capitals and tensors of order higher than two as bold script letters. The transpose of a matrix $\mathbf{A}$ will be written as $\mathbf{A}^{\mathrm{T}}$ and its Moore–Penrose pseudoinverse [5] as $\mathbf{A}^{\dagger}$. $\mathrm{E}\{\cdot\}$ denotes the statistical expectation. $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is the vector space of real-valued $I_1 \times I_2 \times \ldots \times I_N$ tensors.

2. BASIC DEFINITIONS

In this section we introduce some elementary notations and definitions needed in the further developments. Section 2.1 deals with some basic concepts of multilinear algebra; Section 2.2 deals with HOS.

2.1. Multilinear algebra

First the definition of an outer product generalizes expressions of the type $\mathbf{ab}^{\mathrm{T}}$ in which $\mathbf{a}$ and $\mathbf{b}$ are vectors.

Definition 1 (outer product)

The outer product $\mathcal{A} \circ \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{\mathrm{P}} \times J_1 \times J_2 \times \ldots \times J_{\mathrm{Q}}}$ of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_{\mathrm{P}}}$ and a tensor $\mathcal{B} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_{\mathrm{Q}}}$ is defined by

$$(\mathcal{A} \circ \mathcal{B})_{i_1 i_2 \ldots i_P j_1 j_2 \ldots j_Q} \stackrel{\mathrm{def}}{=} a_{i_1 i_2 \ldots i_P} b_{j_1 j_2 \ldots j_Q}$$

for all values of the indices.

For example, the entries of an Nth-order tensor $\mathcal{A}$ equal to the outer product of N vectors $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$ are given by $a_{i_1 i_2 \ldots i_N} = u_{i_1}^{(1)} u_{i_2}^{(2)} \ldots u_{i_N}^{(N)}$.

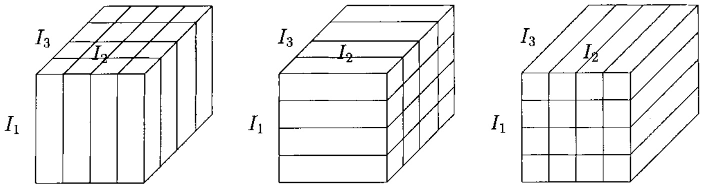


Figure 1. A 4 × 4 × 4 tensor considered as a set of column vectors, row vectors and mode-3 vectors respectively.

Next we give straightforward generalizations of the scalar product, orthogonality and the Frobenius norm.

Definition 2 (scalar product)

The scalar product $\langle \mathcal{A}, \mathcal{B} \rangle$ of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is defined as

$$\langle \mathcal{A}, \mathcal{B} \rangle \overset{\mathrm{def}}{=} \sum_{i_1} \sum_{i_2} \ldots \sum_{i_N} b_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N}$$

The tensor scalar product of two vectors **x** and **y** reduces to the well-known form $\mathbf{y}^{\mathrm{T}}\mathbf{x}$.

Definition 3 (orthogonality)

Tensors whose scalar product equals zero are mutually orthogonal.

Definition 4 (Frobenius norm)

The Frobenius norm of a tensor $\mathcal{A}$ is given by

$$\|\mathcal{A}\| \overset{\mathrm{def}}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$$

In tensor terminology, column vectors, row vectors, etc. will be called mode-1 vectors, mode-2 vectors, etc. In general, the mode-n vectors of an Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ are the I_n-dimensional vectors obtained from $\mathcal{A}$ by varying the index i_n and keeping the other indices fixed (Figure 1).

The multiplication of a higher-order tensor with a matrix can be defined as follows.

Definition 5

The mode-n product of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is an $I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \ldots \times I_N$ tensor defined by

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots j_n \ldots i_N} = \sum_{i_n} a_{i_1 i_2 \ldots i_n \ldots i_N} u_{j_n i_n}$$

for all index values.

The mode-n product allows one to express the effect of a basis transformation in $\mathbb{R}^{I_n}$ on the tensor $\mathcal{A}$.

By way of illustration, let us look at the matrix product $\mathbf{A} = \mathbf{U}^{(1)} \cdot \mathbf{B} \cdot \mathbf{U}^{(2)^{\mathrm{T}}}$ involving matrices $\mathbf{B} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{U}^{(1)} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{J_2 \times I_2}$ and $\mathbf{A} \in \mathbb{R}^{J_1 \times J_2}$. Working with 'generalized transposes' in the

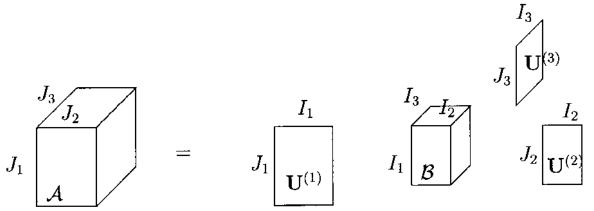


Figure 2. Visualization of the multiplication of a third-order tensor $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with matrices $\mathbf{U}^{(1)} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{J_2 \times I_2}$ and $\mathbf{U}^{(3)} \in \mathbb{R}^{J_3 \times I_3}$.

multilinear case (in which the fact that mode-1 vectors are transpose-free would not have an inherent meaning) can be avoided by observing that the relationships of $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$ (not $\mathbf{U}^{(2)^{\mathrm{T}}}$) with **B** are in fact completely similar: in the same way as $\mathbf{U}^{(1)}$ makes linear combinations of the rows of **B**, $\mathbf{U}^{(2)}$ makes linear combinations of the columns; in the same way as the columns of **B** are multiplied by $\mathbf{U}^{(1)}$, its rows are multiplied by $\mathbf{U}^{(2)}$; in the same way as the columns of $\mathbf{U}^{(1)}$ are associated with the column space of **A**, the columns of $\mathbf{U}^{(2)}$ are associated with the row space. This typical relationship is denoted by means of the $\times_n$ symbol: $\mathbf{A} = \mathbf{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}$.

Figure 2 visualizes the equation $\mathcal{A} = \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}$ for third-order tensors $\mathcal{A} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$ and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. Unlike the conventional way to visualize second-order matrix products, $\mathbf{U}^{(2)}$ has not been transposed, for reasons of symmetry. Multiplication with $\mathbf{U}^{(1)}$ involves linear combinations of the 'horizontal matrices' (index i_1 fixed) in $\mathcal{B}$. Stated otherwise, multiplication of $\mathcal{B}$ with $\mathbf{U}^{(1)}$ means that every column of $\mathcal{B}$ (indices i_2 and i_3 fixed) has to be multiplied from the left with $\mathbf{U}^{(1)}$. Multiplication with $\mathbf{U}^{(2)}$ and $\mathbf{U}^{(3)}$ can be expressed in a similar way.

2.2. Higher-order statistics

The basic quantities of HOS are higher-order moments and higher-order cumulants. First, moment tensors of a real stochastic vector are defined as follows.

Definition 6 (moment)

The *N*th-order moment tensor $\mathcal{M}_{\mathrm{x}}^{(N)} \in \mathbb{R}^{I \times I \times \ldots \times I}$ of a real stochastic vector $\mathbf{x} \in \mathbb{R}^I$ is defined by the element-wise equation

$$(\mathcal{M}_{\mathrm{x}}^{(N)})_{i_1 i_2 \ldots i_N} = \mathsf{Mom}(x_{i_1}, x_{i_2}, \ldots, x_{i_N}) \overset{\mathrm{def}}{=} \mathsf{E}\{x_{i_1} x_{i_2} \ldots x_{i_N}\} \tag{1}$$

The first-order moment is the mean of the stochastic vector. The second-order moment is the correlation matrix (following the definition adopted in e.g. Reference [6], in which the mean is not subtracted).

On the other hand, cumulants of a real stochastic vector are defined as follows.

Definition 7 (cumulant)

The *N*th-order cumulant tensor $\mathcal{C}_{\mathrm{x}}^{(N)} \in \mathbb{R}^{I \times I \times \ldots \times I}$ of a real stochastic vector $\mathbf{x} \in \mathbb{R}^I$ is defined by the element-wise equation

$$
\begin{aligned}
(\mathscr{C}_{\mathrm{x}}^{(N)})_{i_1 i_2 \ldots i_N} &= \mathsf{Cum}(x_{i_1}, x_{i_2}, \ldots, x_{i_N}) \\
&\overset{\text{def}}{=} \sum (-1)^{K-1}(K-1)!\mathsf{E}\left\{\prod_{i\in \mathsf{A}_1} x_i\right\}\mathsf{E}\left\{\prod_{i\in \mathsf{A}_2} x_i\right\}\ldots\mathsf{E}\left\{\prod_{i\in \mathsf{A}_K} x_i\right\} \qquad (2)
\end{aligned}
$$

where the summation involves all possible partitions $\{\mathsf{A}_1, \mathsf{A}_2,\ldots,\mathsf{A}_K\}$ $(1 \leqslant K \leqslant N)$ of the integers $\{i_1,i_2,\ldots,i_N\}$. For a real zero-mean stochastic vector $\mathbf{x}$ the cumulants up to order four are explicitly given by

$$(\mathbf{c}_{\mathrm{x}})_i = \mathsf{Cum}(x_i) \overset{\text{def}}{=} \mathsf{E}\{x_i\} \qquad (3)$$

$$(\mathbf{C}_{\mathrm{x}})_{i_1 i_2} = \mathsf{Cum}(x_{i_1}, x_{i_2}) \overset{\text{def}}{=} \mathsf{E}\{x_{i_1} x_{i_2}\} \qquad (4)$$

$$(\mathscr{C}_{\mathrm{x}}^{(3)})_{i_1 i_2 i_3} = \mathsf{Cum}(x_{i_1}, x_{i_2}, x_{i_3}) \overset{\text{def}}{=} \mathsf{E}\{x_{i_1} x_{i_2} x_{i_3}\} \qquad (5)$$

$$
\begin{aligned}
(\mathscr{C}_{\mathrm{x}}^{(4)})_{i_1 i_2 i_3 i_4} = \mathsf{Cum}(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}) &\overset{\text{def}}{=} \mathsf{E}\{x_{i_1} x_{i_2} x_{i_3} x_{i_4}\} - \mathsf{E}\{x_{i_1} x_{i_2}\}\mathsf{E}\{x_{i_3} x_{i_4}\} \\
&\quad - \mathsf{E}\{x_{i_1} x_{i_3}\}\mathsf{E}\{x_{i_2} x_{i_4}\} - \mathsf{E}\{x_{i_1} x_{i_4}\}\mathsf{E}\{x_{i_2} x_{i_3}\} \qquad (6)
\end{aligned}
$$

For every component x_i of $\mathbf{x}$ that has a non-zero mean, x_i has to be replaced in these formulae, except in Equations (3) and (2) when it applies to a first-order cumulant, by $x_i - \mathsf{E}\{x_i\}$.

Let us first illustrate the meaning of Equation (2) by means of the second-order case (Equation (4)). As there are two possible partitions of $\{i_1, i_2\}$, namely $\{\{i_1, i_2\}\}$ (the number of partition classes $K = 1$) and $\{\{i_1\}, \{i_2\}\}$ $(K = 2)$, Equation (2) reads as

$$(\mathbf{C}_{\mathrm{x}})_{i_1 i_2} = \mathsf{E}\{x_{i_1} x_{i_2}\} - \mathsf{E}\{x_{i_1}\}\mathsf{E}\{x_{i_2}\}$$

in which x_i (x_j) has to be replaced by $x_i - \mathsf{E}\{x_i\}$ $(x_j - \mathsf{E}\{x_j\})$ if x_i (x_j) has a non-zero mean. Since the second term drops by definition, we obtain the form of Equation (4).

It turns out that, again, the first-order cumulant is the mean of the stochastic vector. The second-order cumulant is the covariance matrix. The interpretation of a cumulant of order higher than two is not straightforward, but the powerful properties listed below will demonstrate the importance of the concept. For the moment it suffices to state that cumulants of a set of random variables give an indication of their mutual statistical dependence (completely independent variables resulting in a zero cumulant), and that higher-order cumulants of a single random variable are some measure of its non-Gaussianity (cumulants of a Gaussian variable, for $N > 2$, being equal to zero).

Table I illustrates the definitions for two important univariate probability density functions.

At first sight, higher-order moments, because of their straightforward definition, might seem more interesting quantities than higher-order cumulants. However, cumulants have a number of important properties that are not shared with higher-order moments, such that in practice cumulants are more frequently used. We enumerate some of the most interesting Properties (without proof) [7,8].

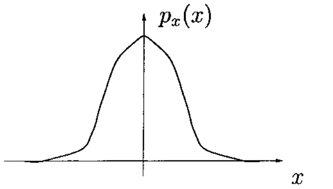
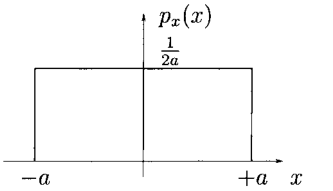
1. *Supersymmetry*. Moments and cumulants are symmetric in their arguments, i.e.

$$(\mathscr{M}_{\mathrm{x}}^{(N)})_{i_1 i_2 \ldots i_N} = (\mathscr{M}_{\mathrm{x}}^{(N)})_{\mathsf{P}(i_1 i_2 \ldots i_N)} \qquad (7)$$

$$(\mathscr{C}_{\mathrm{x}}^{(N)})_{i_1 i_2 \ldots i_N} = (\mathscr{C}_{\mathrm{x}}^{(N)})_{\mathsf{P}(i_1 i_2 \ldots i_N)} \qquad (8)$$

in which P is an arbitrary permutation of the indices.

Table I. Moments and cumulants, up to order four, of a Gaussian and a uniform distribution

Gaussian distribution			
	$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2}{2\sigma^2})$		
	n	$m_x^{(n)}$	$c_x^{(n)}$
	1	0	0
	2	σ^2	σ^2
	3	0	0
	4	$3\,\sigma^4$	0
Uniform distribution			
	$p_x(x) = \frac{1}{2a}$ $(x \in [-a, +a])$		
	n	$m_x^{(n)}$	$c_x^{(n)}$
	1	0	0
	2	$a^2/3$	$a^2/3$
	3	0	0
	4	$3a^4/5$	$-2a^4/15$

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

$$\mathcal{M}_{\tilde{\mathrm{x}}}^{(N)} = \mathcal{M}_{\mathrm{x}}^{(N)} \times_1 \mathbf{A} \times_2 \mathbf{A} \ldots \times_N \mathbf{A} \tag{9}$$

$$\mathcal{C}_{\tilde{\mathrm{x}}}^{(N)} = \mathcal{C}_{\mathrm{x}}^{(N)} \times_1 \mathbf{A} \times_2 \mathbf{A} \ldots \times_N \mathbf{A} \tag{10}$$

3. *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.
4. *Partitioning of independent variables*. If a subset of I stochastic variables $x_1, x_2, \ldots, x_I$ is independent of the other variables, then we have

$$\mathsf{Cum}(x_1, x_2, \ldots, x_I) = 0 \tag{11}$$

This property does not hold in general for moments (e.g. for two mutually independent random variables x and y we have that $\mathsf{Mom}(x,x,y,y) = \mathsf{E}\{x^2\} \cdot \mathsf{E}\{y^2\}$, which does not vanish unless one of the variables is identically equal to zero). 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 the ICA techniques that will be discussed in this paper. To clarify this, let us have a look at the second-order case. Let us assume a stochastic vector $\mathbf{x} \in \mathbb{R}^I$ with mutually independent entries that are not necessarily zero-mean. Unless the entries are zero-mean, the correlation matrix (second-order moment) of $\mathbf{x}$ is not a diagonal matrix, as the mean of the entries is not subtracted in the definition of a moment. On the other hand, the covariance matrix (second-order cumulant) is a diagonal matrix regardless of the mean of **x**. The definition of a cumulant is such that this property holds for all cumulant orders.

5. *Sum of independent variables*. If the stochastic variables x_1, x_2, …, x_I are mutually independent from the stochastic variables y_1, y_2, …, y_I, then we have

$$\mathsf{Cum}(x_1+y_1,x_2+y_2,\ldots,x_k+y_k)=\mathsf{Cum}(x_1,x_2,\ldots,x_k)+\mathsf{Cum}(y_1,y_2,\ldots,y_k) \qquad (12)$$

The cumulant tensor of a sum of independent random vectors is the sum of the individual cumulants. This property does not hold for moments either; as a matter of fact, it explains the term 'cumulant'. (One could expand $\mathsf{Mom}(x_1,y_1,x_2+y_2,\ldots,x_k+y_k)$ as a sum over all possible *x/y* combinations, but the cross-terms, containing *x* as well as *y* entries, do not necessarily vanish, as opposed to the cumulant case—see the previous property.)

6. *Non-Gaussianity*. If *y* is a Gaussian variable with the same mean and variance as a given stochastic variable *x*, then, for $N \geqslant 3$, it holds that

$$\mathcal{C}_x^{(N)}=\mathcal{M}_x^{(N)}-\mathcal{M}_y^{(N)} \qquad (13)$$

As a consequence, higher-order cumulants of a Gaussian variable are zero (see Table I). In combination with the multilinearity property, we observe that higher-order cumulants have the interesting property of being blind for additive Gaussian noise. Namely, if a stochastic variable *x* is corrupted by additive Gaussian noise *n*, i.e.

$$\hat{x}=x+n$$

then we nevertheless have that

$$\mathcal{C}_{\hat{x}}^{(N)}=\mathcal{C}_x^{(N)}+\mathcal{C}_n^{(N)}=\mathcal{C}_x^{(N)}$$

Generally speaking, it becomes harder to estimate HOS from sample data as the order increases, i.e. longer data sets are required to obtain the same accuracy [9,10]. Hence in practice the use of HOS is usually restricted to third- and fourth-order cumulants. For symmetric distributions, fourth-order cumulants are used, since the third-order cumulants vanish, as mentioned in Property 3.

3. INDEPENDENT COMPONENT ANALYSIS

3.1. The problem

Assume the basic linear statistical model

$$\mathbf{y}=\mathbf{Mx}+\mathbf{n}=\tilde{\mathbf{y}}+\mathbf{n} \qquad (14)$$

in which $\mathbf{y}\in\mathbb{R}^I$ is called the *observation vector*, $\mathbf{x}\in\mathbb{R}^J$ is the *source vector* and $\mathbf{n}\in\mathbb{R}^I$ represents additive *noise*. $\tilde{\mathbf{y}}\in\mathbb{R}^I$ is the *signal part* of the observations. $\mathbf{M}\in\mathbb{R}^{I\times J}$ is called the *mixing matrix* — its entries m_{ij} indicate to what extent the *j*th source component contributes to the *i*th observation channel ($1\leqslant i\leqslant I$, $1\leqslant j\leqslant J$), i.e. they determine how the sources are 'mixed' in the observations. The columns $\{\mathbf{m}_j\}$ of **M** are the *mixing vectors;* its range is known as the *signal subspace*.

The concept of independent component analysis can now be formulated as follows.

> The goal of independent component analysis (ICA) consists of the estimation of the mixing matrix **M** and/or the corresponding realizations of the source vector **x**, given only realizations of the observation vector **y**, under the following assumptions.

1. The mixing vectors are linearly independent.
2. The components of $\mathbf{x}$ are mutually statistically independent, as well as independent from the noise components.

The second assumption is the key ingredient for ICA. It is a very strong hypothesis, but also quite natural in lots of applications: in practice, ‘mutually statistically independent’ can often be rephrased as ‘of a different nature’. ICA is therefore of interest e.g. for the separation of electromagnetic signals emitted by different users in mobile communications; for the extraction of bioelectric signals, generated by different organs, from body surface potentials; for the analysis of different sources of vibration in rotating machinery; etc. From an algebraic point of view it does not only mean that the covariance of $\mathbf{x}$ is a diagonal matrix, but also that all the higher-order cumulants of $\mathbf{x}$ are diagonal tensors. Using the properties discussed in Section 2.2, we have that

$$\mathbf{C}_{\mathrm{y}} = \mathbf{C}_{\mathrm{x}} \times_1 \mathbf{M} \times_2 \mathbf{M} + \mathbf{C}_{\mathrm{n}} \tag{15}$$

$$\mathcal{C}_{\mathrm{y}}^{(N)} = \mathcal{C}_{\mathrm{x}}^{(N)} \times_1 \mathbf{M} \times_2 \mathbf{M} \ldots \times_N \mathbf{M} + \mathcal{C}_{\mathrm{n}}^{(N)} \tag{16}$$

in which $\mathbf{C}_{\mathrm{x}}$ and $\mathcal{C}_{\mathrm{x}}^{(N)}$ are diagonal and $\mathcal{C}_{\mathrm{n}}^{(N)}$ vanishes if the noise is Gaussian.

The first assumption is, for the class of algorithms that will be discussed in this paper, required for reasons of identifiability. It holds in a generic sense when $I \geqslant J$ (regardless of the fact that an ill-conditioned mixing matrix can make the ICA problem heavier from a numerical point of view, as will be illustrated by the numerical experiments in Section 5.5). However, the identifiability constraint is not inherent to ICA as such. Under smoother conditions it is even possible to identify the mixing matrix in the situation in which there are more sources than sensors [11]; this is a topic of current investigations.

In the following subsections we will explain how the two assumptions above can be exploited to obtain an estimate $\hat{\mathbf{M}}$ of the mixing matrix. However, merely resorting to these two assumptions, it is impossible to distinguish between the signal and the noise term in Equation (14). Hence the source signals will be estimated from the observations by means of a simple matrix multiplication as follows:

$$\hat{\mathbf{x}} = \mathbf{W}^{\mathrm{T}}\mathbf{y} \tag{17}$$

$\mathbf{W}^{\mathrm{T}}$ can e.g. take the form of $\hat{\mathbf{M}}^{\dagger}$. More generally, various beamforming strategies [12] can be applied (see also Section 3.4).

The ICA problem is also addressed in the literature under the labels *blind source separation* (BSS), *signal copy*, *waveform-preserving estimation*, etc. However, the assumptions on which the solution strategies are based may sometimes differ from paper to paper.

3.2. *Prewhitening-based ICA*

The ICA problem is most often solved by a two-stage algorithm consisting of a second- and a higher-order step. In this subsection we will explain the technique in general terms. An outline of the procedure is presented as Algorithm 1.

Algorithm 1 (PCA-based ICA)

Given: T samples $\{\mathbf{y}_t\}_{1 \leqslant t \leqslant T}$ of $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{n}$ ($\mathbf{y}, \mathbf{n} \in \mathbb{R}^I, \mathbf{x} \in \mathbb{R}^J, \mathbf{M} \in \mathbb{R}^{I \times J}$). Call $\mathbf{A}_{\mathrm{y}} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T]$.

1. Prewhitening stage (PCA).

- Compute sample mean $\hat{\mathbf{m}}_{\mathrm{y}}$ from $\mathbf{A}_{\mathrm{y}}$. Define $\tilde{\mathbf{A}}_{\mathrm{y}} = [\mathbf{y}_1 - \hat{\mathbf{m}}_{\mathrm{y}}, \mathbf{y}_2 - \hat{\mathbf{m}}_{\mathrm{y}}, \ldots, \mathbf{y}_T - \hat{\mathbf{m}}_{\mathrm{y}}]/\sqrt{T-1}$.
- Truncated SVD of $\tilde{\mathbf{A}}_{\mathrm{y}}$: $\tilde{\mathbf{A}}_{\mathrm{y}} = \mathbf{U} \cdot \mathbf{S} \cdot \tilde{\mathbf{V}}^{\mathrm{T}}$, with $\mathbf{U} \in \mathbb{R}^{I \times J}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{T \times J}$ column-wise orthonormal and $\mathbf{S} \in \mathbb{R}^{I \times J}$ positive definite and diagonal.
- If $\mathbf{n}$ is spatially white and Gaussian, with variance σ_n^2, replace the diagonal entries of $\mathbf{S}$ by $\sqrt{s_{jj}^2 - \sigma_n^2}(1 \leqslant j \leqslant J)$.

(Section 3.2.1.)

2. Higher-order stage.

- Compute sample cumulant $\hat{\mathcal{C}}_{\mathrm{y}}^{(4)}$ from $\mathbf{A}_{\mathrm{y}}$.
- $\hat{\mathcal{C}}_{\mathrm{z}}^{(4)} = \hat{\mathcal{C}}_{\mathrm{y}}^{(4)} \times_1 (\mathbf{S}^\dagger \cdot \mathbf{U}^{\mathrm{T}}) \times_2 (\mathbf{S}^\dagger \cdot \mathbf{U}^{\mathrm{T}}) \times_3 (\mathbf{S}^\dagger \cdot \mathbf{U}^{\mathrm{T}}) \times_4 (\mathbf{S}^\dagger \cdot \mathbf{U}^{\mathrm{T}})$.
- (Approximate) diagonalization:

$$\hat{\mathcal{C}}_{\mathrm{z}}^{(4)} = \hat{\mathcal{C}}_{\mathrm{x}}^{(4)} \times_1 \mathbf{V}^{\mathrm{T}} \times_2 \mathbf{V}^{\mathrm{T}} \times_3 \mathbf{V}^{\mathrm{T}} \times_4 \mathbf{V}^{\mathrm{T}}$$

in which $\mathbf{V}$ is orthogonal. A class of algebraic algorithms:

- HOEVD (Section 5.1);
- MD (Section 5.2);
- JADE (Section 5.3);
- STOTD (Section 5.4).

(Section 3.2.2.)

Results: mixing matrix estimate $\hat{\mathbf{M}} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^{\mathrm{T}}$; source separation by means of MVDR, LCMV, …, beamforming.

3.2.1. Step 1: prewhitening. The prewhitening step amounts to a principal component analysis (PCA) of the observations. Briefly, the goal is to transform the observation vector $\mathbf{y}$ into another stochastic vector $\mathbf{z}$ having unit covariance. This involves the multiplication of $\mathbf{y}$ with the inverse of the square root of its covariance matrix $\mathbf{C}_{\mathrm{y}}$. When $J < I$, a projection of $\mathbf{y}$ onto the signal subspace is carried out.

Let us now discuss the prewhitening procedure in more detail. First we observe that the covariance matrix $\mathbf{C}_{\mathrm{y}}$ takes the form (for the moment the noise term in Equation (14) is neglected, for clarity)

$$\mathbf{C}_{\mathrm{y}} = \mathbf{M} \cdot \mathbf{C}_{\mathrm{x}} \cdot \mathbf{M}^{\mathrm{T}} \tag{18}$$

in which the covariance $\mathbf{C}_{\mathbf{x}}$ of $\mathbf{x}$ is diagonal, since the source signals are uncorrelated. Assuming that the source signals have unit variance (without loss of generality, as we may appropriately rescale the mixing vectors as well), we have

$$\mathbf{C}_{\mathrm{y}} = \mathbf{M} \cdot \mathbf{M}^{\mathrm{T}} \tag{19}$$

A first observation is that the number of sources can be deduced from the rank of $\mathbf{C}_{\mathbf{y}}$. Substitution of the SVD of the mixing matrix $\mathbf{M} = \mathbf{USV}^{\mathrm{T}}$ shows that the EVD of the observed covariance allows us to estimate the components $\mathbf{U}$ and $\mathbf{S}$ whilst the factor $\mathbf{V}$ remains unknown:

$$\mathbf{C}_{\mathrm{y}} = \mathbf{U} \cdot \mathbf{S}^2 \cdot \mathbf{U}^{\mathrm{T}} = (\mathbf{US}) \cdot (\mathbf{US})^{\mathrm{T}} \tag{20}$$

Hence the signal subspace can be estimated from the second-order statistics of the observations, but the actual mixing matrix remains unknown up to an orthogonal factor.

The effect of the additive noise term $\mathbf{n}$ can be neutralized by replacing $\mathbf{C}_{\mathbf{y}}$ by the noise-free covariance $\mathbf{C_y} - \mathbf{C_n}$. In the case of spatially white noise (i.e. the noise components are mutually uncorrelated and all have the same variance), $\mathbf{C}_\mathrm{n}$ takes the form $\sigma_n^2\mathbf{I}$, in which σ_n^2 is the variance of the noise on each data channel. In a more-sensors-than-sources set-up, σ_n^2 can be estimated as the mean of the 'noise eigenvalues', i.e. the smallest $I - J$ 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 Reference [13].

When the output covariance is estimated from the data as

$$\mathbf{C}_\mathrm{y} = \tilde{\mathbf{A}}_\mathrm{y}\tilde{\mathbf{A}}_\mathrm{y}^\mathrm{T} \tag{21}$$

in which $\tilde{\mathbf{A}}_\mathrm{y}$ is an $I \times T$ matrix consisting of T realizations of $\mathbf{y}$, divided by $\sqrt{T-1}$ after subtraction of the sample mean, then it is preferable to compute the factors $\mathbf{U}$ and $\mathbf{S}$ without explicit calculation of the product (21); instead they can be obtained directly from the truncated SVD of $\tilde{\mathbf{A}}_\mathrm{y}$:

$$\tilde{\mathbf{A}}_\mathrm{y} = \mathbf{U}\mathbf{S}\tilde{\mathbf{V}}^\mathrm{T} \tag{22}$$

In this way the squaring of the singular values of $\tilde{\mathbf{A}}_\mathrm{y}$, which may cause a loss of numerical accuracy [14], can be avoided.

After computation of $\mathbf{U}$ and $\mathbf{S}$, a standardized random vector $\mathbf{z}$ can be defined as

$$\mathbf{z} \overset{\mathrm{def}}{=} \mathbf{S}^\dagger \cdot \mathbf{U}^\mathrm{T} \cdot \mathbf{y} \tag{23}$$

3.2.2. Step 2: fixing the rotational degree of freedom using HOS. Here we will explain how the remaining unknown, i.e. the right singular matrix $\mathbf{V}$ of the mixing matrix $\mathbf{M}$, can be estimated. As we have already exploited the information contained in the second-order statistics of the observations, we now resort to the HOS.

Assuming that the noise is Gaussian, higher-order cumulants of the standardized random vector $\mathbf{z}$ defined in Equation (23) are given by

$$\mathcal{C}_\mathrm{z}^{(N)} = \mathcal{C}_\mathrm{x}^{(N)} \times_1 \mathbf{V}^\mathrm{T} \times_2 \mathbf{V}^\mathrm{T} \ldots \times_N \mathbf{V}^\mathrm{T} \tag{24}$$

(cf. Equation (16) with $\mathbf{z} = \mathbf{V}^\mathrm{T}\mathbf{x}$). This tensor is related to the Nth-order output cumulant by the multilinearity property:

$$\mathcal{C}_\mathrm{z}^{(N)} = \mathcal{C}_\mathrm{y}^{(N)} \times_1 (\mathbf{US})^\dagger \times_2 (\mathbf{US})^\dagger \ldots \times_N (\mathbf{US})^\dagger \tag{25}$$

The key observation is that the source cumulant $\mathcal{C}_\mathrm{x}^{(N)}$ is theoretically a diagonal tensor, since the source signals are not only uncorrelated but also higher-order independent. Hence Equation (24) is in fact a symmetric EVD-like tensor decomposition. This decomposition is unique if at most one diagonal element of $\mathcal{C}_\mathrm{x}^{(N)}$ equals zero, as will be explained in Section 3.3. However, simply counting the degrees of freedom in the decomposition model shows that in general a higher-order tensor cannot be diagonalized by means of orthogonal transformations: the supersymmetric tensor $\mathcal{C}_\mathrm{z}^{(N)}$ contains in principle $J(J+1)\ldots(J+N-1)/N!$ independent entries, whilst the decomposition allows only $J(J+1)/2$ (orthogonal factor $\mathbf{V}$) $+ J$ (diagonal of $\mathcal{C}_\mathrm{x}^{(N)}$) degrees of freedom. This means that if $\mathcal{C}_\mathrm{z}^{(N)}$ is not perfectly known (owing to a finite data length, non-Gaussian additive noise, etc.), the approximating tensor cannot be fully diagonalized in general. The way in which the estimation error is dealt with allows us to distinguish different solution strategies. Four different algebraic approaches will briefly be discussed in Section 5.

It is worth mentioning that (24) is in fact a CANDECOMP/PARAFAC model of $\mathcal{C}_{\mathrm{z}}^{(N)}$ [15]. If we represent the rows of $\mathbf{V}^{\mathrm{T}}$ by $\{\mathbf{v}_j\}$ and the source cumulants by $\{\kappa_{x_j}\}$ $(1 \leqslant j \leqslant J)$, then Equation (24) can be rewritten as

$$\mathcal{C}_{\mathrm{z}}^{(N)} = \sum_j \kappa_{x_j} \mathbf{v}_j^{\mathrm{T}} \circ \mathbf{v}_j^{\mathrm{T}} \circ \ldots \circ \mathbf{v}_j^{\mathrm{T}} \tag{26}$$

This is indeed an expansion of $\mathcal{C}_{\mathrm{z}}^{(N)}$ as a linear combination of rank-1 tensors (i.e. tensors that are given by an outer product of vectors). The rank-1 terms have the special property that they are supersymmetric and mutually orthogonal. Again, there may not be a complete match of both sides of Equation (26) in the case where $\mathcal{C}_{\mathrm{z}}^{(N)}$ is not perfectly known.

3.3. *Identifiability*

In this subsection we will explain to what extent the ICA solution is inherently unique, apart from the concrete algorithm that one might want to use.

First we observe that it is impossible to determine the norm of the columns of **M** in Equation (14), since a rescaling of these vectors can be compensated by the inverse scaling of the source signal values; the same holds for their sign. 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 an ICA solution is not unique [16–18]. Formally, for source vectors of which at most one component is Gaussian, we can apply the following theorem (see Reference [18], pp. 127–128).

Theorem 1

Let the *N*th-order supersymmetric tensor $\mathcal{C} \in \mathbb{R}^{J \times J \times \cdots \times J}$ be given by

$$\mathcal{C} = \mathfrak{D} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \ldots \times_N \mathbf{Q} \tag{27}$$

in which $\mathfrak{D} \in \mathbb{R}^{J \times J \times \cdots \times J}$ is diagonal, containing at most one zero on the diagonal, and $\mathbf{Q} \in \mathbb{R}^{J \times J}$ is orthogonal. $\mathcal{C}$ can be decomposed by the same model in terms of $\mathfrak{D}'$ and $\mathbf{Q}'$ iff:

- $\mathbf{Q}' = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{P}$, in which Λ is a diagonal matrix whose entries are ± 1 and **P** is a permutation; and
- $\mathfrak{D}'$ is related to $\mathfrak{D}$ in the inverse way

$$\mathfrak{D}' = \mathfrak{D} \times_1 (\mathbf{P}^{\mathrm{T}}\Lambda) \times_2 (\mathbf{P}^{\mathrm{T}}\Lambda) \ldots \times_N (\mathbf{P}^{\mathrm{T}}\Lambda) \tag{28}$$

The higher-order cumulants of Gaussian components vanish; hence these components cannot be separated in an essentially unique way. By claiming that both the covariance matrices of **y** and **x** are diagonal, it is easy to prove the theorem that makes explicit the indeterminacy [16].

Theorem 2

Let **x** be a *J*-dimensional Gaussian random vector. Let $\mathbf{M} \in \mathbb{R}^{I \times J}$ have linearly independent columns, and consider $\mathbf{y} = \mathbf{Mx}$. Then the components y_i are mutually independent iff $\mathbf{M} = \boldsymbol{\Lambda}_1\mathbf{Q}\boldsymbol{\Lambda}_2$, with $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ diagonal and **Q** orthogonal.

The fact that we assume that the sources are non-Gaussian is less restrictive than it may seem. Many signals of interest are non-Gaussian (e.g. modulated electromagnetic signals in mobile communications, rotation-induced vibrations in machine testing, reflected signals in seismic prospection, etc., to quote but a few examples in signal processing). On the other hand, the assumption of Gaussianity often applies to the noise term. Consequently, non-Gaussianity may be seen as a property that discriminates between the signals of interest and the noise.

In sum, we can state that ICA does not allow us to estimate the transfer matrix **M** as such, but that, for source vectors of which at most one component is Gaussian, it allow us to determine the appropriate rest class of the quotient set defined by the equivalence relation $\mathbf{M} \sim \mathbf{M}' \Leftrightarrow \mathbf{M}' = \mathbf{M}\mathbf{\Lambda}\mathbf{P}$. Generically a unique representative of this rest class can be obtained by normalizing the solution such that e.g.

- the source components have unit variance;
- the mixing vectors are ordered by decreasing Frobenius norm;
- for each mixing vector the entry which is largest in absolute value is positive.

3.4. Measures of performance

In this subsection we will explain how one can evaluate the quality of an estimate of the ICA solution. Actually, it is impossible to quantify *the* quality, as quality may be perceived against a background of different criteria. Namely, the goal of the ICA procedure may consist of an optimal mutual separation of the source signals, or an optimal recovery of the source signals from the noise, or an optimal estimation of the mixing matrix. For each of these viewpoints we will discuss appropriate measures of performance.

The quality of source separation and reconstruction is naturally formulated in terms of beamforming performance. In general terms the idea of beamforming consists of the construction of the matrix **W** in Equation (17) from an estimate of the mixing matrix in such a way that the performance measure of interest is maximized. A detailed discussion of beamformers falls outside the scope of this paper; we refer to Reference [12].

In terms of Equation (14) and (17), and assuming that the sources are estimated in the right order (for notational convenience), we define the following indices of performance:

$$\mathrm{SNR}_i \overset{\mathrm{def}}{=} \frac{\sigma_{x_i}^2 (\mathbf{w}_i^{\mathrm{T}} \mathbf{m}_i)^2}{\mathbf{w}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{n}} \mathbf{w}_i} \tag{29}$$

$$\mathrm{SIR}_{ij} \overset{\mathrm{def}}{=} \frac{\sigma_{x_i}^2 (\mathbf{w}_i^{\mathrm{T}} \mathbf{m}_i)^2}{\sigma_{x_j}^2 (\mathbf{w}_i^{\mathrm{T}} \mathbf{m}_j)^2} \tag{30}$$

$$\mathrm{SINR}_i \overset{\mathrm{def}}{=} \frac{\sigma_{x_i}^2 (\mathbf{w}_i^{\mathrm{T}} \mathbf{m}_i)^2}{\mathbf{w}_i^{\mathrm{T}} (\mathbf{C}_{\mathrm{y}} - \sigma_{x_i}^2 \mathbf{m}_i \mathbf{m}_i^{\mathrm{T}}) \mathbf{w}_i} \tag{31}$$

in which $\mathbf{w}_i$ is the *i*th column of **W** and $\sigma_{x_i}^2$ is the variance of the *i*th source (formulated for a single ICA problem; in a series of Monte Carlo runs, averaged values are considered). The first index is the signal/noise ratio (SNR) of the estimate of the *i*th source. It consists of the ratio of the variance of the actual contribution of the *i*th source in its estimate, over the variance of the noise contribution in the *i*th source estimate. The second index is a signal/interference ratio (SIR). It is defined as the ratio of the variance of the actual contribution of the *i*th source in its estimate, over the variance of the contribution of the *j*th source in the estimate of the *i*th source. This index quantifies the contamination by the *j*th source of the *i*th source estimate. Finally, the third index is the signal/interference-plus-noise ratio (SINR) of the *i*th source estimate. It consists of the ratio of the variance of the contribution of the *i*th source in its estimate, over the variance of all the other contributions (other sources and noise) to the *i*th source estimate. Often only the numerator of SIR_{ij}, indicating to what extent $\mathbf{W}^{\mathrm{T}}$

approximates $\mathbf{M}^\dagger$, is considered. Note that for a good estimate the denominator approximates unity, which allows us to define an approximate interference/signal ratio (ISR) as $\mathrm{ISR}_{ij} \overset{\mathrm{def}}{=} \sigma_{x_j}^2 (\mathbf{w}_i^{\mathrm{T}} \mathbf{m}_j)^2$.

The SINR is optimized by a *minimum variance distortionless response* (MVDR) filter given by

$$\mathbf{W} = (\mathbf{C}_{\mathrm{y}})^\dagger \cdot \mathbf{M} \tag{32}$$

On the other hand, the mutual interference of the sources can be cancelled by implementing a *linear constrained minimum variance* (LCMV) filter

$$\mathbf{W} = (\mathbf{C}_{\tilde{\mathrm{y}}})^\dagger \cdot \mathbf{M} \tag{33}$$

in which $\mathbf{C}_{\tilde{\mathrm{y}}}$ is the covariance matrix of the signal part of the observations. In practice, these filters will be approximated in terms of sample statistics and the estimate of the mixing matrix.

In the case where not the separation of the sources but the estimation of the mixing matrix is of primary importance, it is natural to express the ICA performance in terms of the Frobenius norm of the difference between $\mathbf{M}$ and its estimate $\hat{\mathbf{M}}$. We implicitly assume that the columns of $\hat{\mathbf{M}}$ are optimally ordered and scaled. In a Monte Carlo experiment the root mean square error (RMSE) $\sqrt{\mathsf{E}\|\mathbf{M} - \widehat{\mathbf{M}}\|^2}$ is considered.

3.5. *PCA versus ICA*

From the preceding discussion it is clear that ICA is the natural way to 'fine-tune' PCA. Both statistical techniques are linked with the algebraic concepts of EVD and SVD in a remarkable way.

In the 'classical' *second*-order statistical problem of PCA the problem of interest is to remove the correlation from data obtained as a linear mixture of independent source signals. The key tool to realize this comes from 'classical' *linear* algebra: it is the EVD of the observed covariance or, numerically, the SVD of the data set.

In this way the signal subspace, or more precisely the factors $\mathbf{U}$ and $\mathbf{S}$ in the SVD of the mixing matrix $\mathbf{M} = \mathbf{USV}^{\mathrm{T}}$, can be identified. The fact that the mixing vectors and the source signals can only be found up to an orthogonal transformation is known as the *rotational invariance* property of PCA. Uniqueness is usually obtained by adding (often artificial) constraints, e.g. mutual orthogonality of the columns of the mixing matrix estimate.

In the more recent problem of ICA one also aims at the removal of higher-order dependence, which additionally involves the use of *higher*-order statistics. It appears that from an algebraic point of view this leads to a *multilinear* EVD. By inverting the mixing matrix, the original source signals can be estimated.

Although any data set can be decorrelated by a linear transformation, it is not always possible to express it as a linear combination of independent contributions. This corresponds to the fact that Equation (24), with $\mathcal{C}_{\mathrm{x}}^{(N)}$ diagonal, is overdetermined for an arbitrary given supersymmetric higher-order tensor. Whether ICA is useful depends on the context, i.e. one should have the prior knowledge that the data set under consideration indeed consists of linear contributions of an independent nature. On the other hand, the degree to which the cumulant tensor can be diagonalized gives an indication to what extent the estimated source signals can be regarded as independent (see Property 4 of Section 2.2).

In the scheme of Algorithm 1 the prewhitening has the disadvantage, compared to the higher-order step, that the calculations are directly affected by additive Gaussian noise. It turns out that the error introduced in the PCA stage cannot be compensated by the higher-order step; it introduces an upper bound to the overall performance. Writing the SVDs of the true mixing matrix and its estimate as $\mathbf{M} = \mathbf{USV}^{\mathrm{T}}$ and $\hat{\mathbf{M}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}}$, performance bounds are given by the following theorems [19,20].

Theorem 3

Assuming two sources, the quality of separation for the global ICA algorithm is bounded by the quality of the prewhitening in the following way:

$$\mathbf{ISR}_{12} + \mathbf{ISR}_{21} \geqslant \frac{\left((\hat{\mathbf{S}}^{\dagger} \cdot \hat{\mathbf{U}}^{\mathrm{T}} \cdot \mathbf{M})^{\mathrm{T}}(\hat{\mathbf{S}}^{\dagger} \cdot \hat{\mathbf{U}}^{\mathrm{T}} \cdot \mathbf{M})\right)^2_{12}}{\|\hat{\mathbf{S}}^{\dagger} \cdot \hat{\mathbf{U}}^{\mathrm{T}} \cdot \mathbf{M}\|^2} \tag{34}$$

The equality sign holds only for a perfect reconstruction of the mixing matrix, in which case both sides vanish.

It can be proved that errors in the prewhitening stage cause the right-hand side of Equation (34) to be non-vanishing.

Theorem 4

The accuracy of the mixing matrix estimate is bounded by the accuracy of the prewhitening as follows (assuming that $\mathbf{M}$ and $\hat{\mathbf{M}}$ are normalized in the same way):

$$\|\mathbf{M} - \hat{\mathbf{M}}\|^2 \geqslant \sum_i (s_{ii}^2 + \hat{s}_{ii}^2 - 2\sigma_{ii}) \tag{35}$$

in which σ_{ii} is the ith singular value of $\hat{\mathbf{S}}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{US}$. The inequality reduces to an equality for an optimal choice of the orthogonal factor in the higher-order ICA step.

It can be proved that the right-hand side of Equation (35) vanishes iff no estimation error is introduced in the prewhitening stage.

The bounds given by Theorem 3 and 4 can be used as a reference indicating the ultimate performance that can be achieved.

4. EXAMPLE

In this section we will illustrate the general concept of ICA by means of an example. Because this is nice for visualization, we consider deterministic instead of stochastic signals. The only difference is that the expectation operator $\mathsf{E}\{\cdot\}$ should be replaced by averaging over the interval over which the signals are considered, i.e. the covariance matrix and the higher-order cumulant tensor are computed by averaging over an interval instead of averaging over samples.

We assume the two source signals depicted in Figure 3. The first source is a sine wave, the second one is a block wave:

$$x_1(t) = \sqrt{2}\sin t$$

$$x_2(t) = \begin{cases} 1 & \text{iff} \quad k\pi \leqslant t < k\pi + \pi/2 \\ -1 & \text{iff} \quad k\pi + \pi/2 \leqslant t < (k+1)\pi, \quad k \in \mathbb{Z} \end{cases}$$

Both signals are considered over the interval $[0, 4\pi]$. They are zero-mean and their covariance matrix equals the identity (hence the scaling factor in the definition of $x_1(t)$). The third-order cumulants

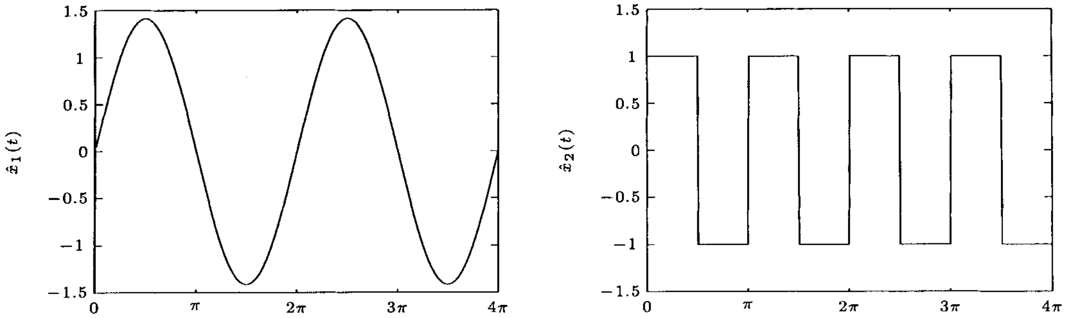


Figure 3. The two source signals considered in the example of Section 4.

vanish, since the signals are symmetric about the axis $x = 0$, which is the deterministic counterpart of Property 3 in Section 2.2. In correspondence with Property 4, the fourth-order cumulant tensor is diagonal; it contains the entries

$$c_{x_1}^{(4)} = \frac{1}{4\pi}\int_0^{4\pi} (\sqrt{2}\sin t)^4 \mathrm{d}t - 3\left(\frac{1}{4\pi}\int_0^{4\pi} (\sqrt{2}\sin t)^2 \mathrm{d}t\right)^2 = -\frac{3}{2}$$

$$c_{x_2}^{(4)} = \frac{1}{4\pi}\int_0^{4\pi} x_2^4(t)\mathrm{d}t - 3\left(\frac{1}{4\pi}\int_0^{4\pi} x_2^2(t)\mathrm{d}t\right)^2 = -2$$

Note that the fourth-order moment tensor is not diagonal:

$$(\mathcal{M}_{\mathrm{x}}^{(4)})_{1122} = \frac{1}{4\pi}\int_0^{4\pi} (\sqrt{2}\sin t)^2 \mathrm{d}t = 1$$

In the definition (6) of the fourth-order cumulant, on the other hand, the same term is subtracted again.

We assume the following mixing matrix:

$$\mathbf{M} = \frac{1}{4}\begin{pmatrix} -1 & -3\sqrt{3} \\ 3\sqrt{3} & -5 \end{pmatrix}$$

The SVD of this matrix is given by

$$\mathbf{M} = \mathbf{U}\cdot\mathbf{S}\cdot\mathbf{V}^{\mathrm{T}} = \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{pmatrix}$$

in which **U** and **V** are orthogonal and **S** is diagonal.

The observations

$$\begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \mathsf{M} \cdot \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$$

are displayed in Figure 4. These signals are clearly mixtures of a sine and a block wave. We do not consider an additive noise term, for clarity.

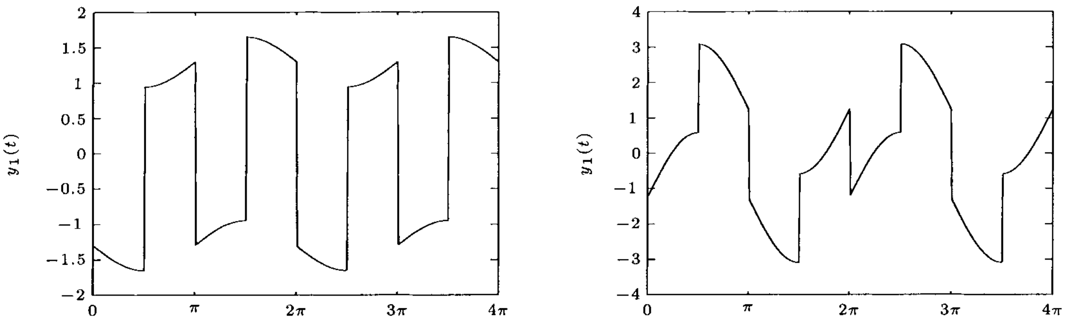


Figure 4. The observation signals considered in the example of Section 4.

A PCA yields the two signals shown in Figure 5. They are uncorrelated, i.e.

$$\frac{1}{4\pi}\int_0^{4\pi} z_1(t)z_2(t)\mathrm{d}t = 0$$

but far from equal to the original two sources $x_1(t)$ and $x_2(t)$. The problem is that any orthogonal rotation of $\mathbf{Z}(t) = (z_1(t)\ z_2(t))^{\mathrm{T}}$ yields signals that are mutually uncorrelated. Stated otherwise, $\mathbf{Z}(t)$ is the result of an orthogonal rotation of $\mathbf{X}(t)$: $\mathbf{Z}(t) = \mathbf{V}^{\mathrm{T}} \cdot \mathbf{X}(t)$.

The orthogonal factor can be found from the fourth-order cumulant $\mathcal{C}_{\mathrm{z}}^{(4)}$, of which the entries are given by

$$\begin{aligned}
(\mathcal{C}_{\mathrm{z}}^{(4)})_{1111} &= -39/32\\
(\mathcal{C}_{\mathrm{z}}^{(4)})_{1112} &= 9\sqrt{3}/32\\
(\mathcal{C}_{\mathrm{z}}^{(4)})_{1122} &= -21/32\\
(\mathcal{C}_{\mathrm{z}}^{(4)})_{1222} &= -5\sqrt{3}/32\\
(\mathcal{C}_{\mathrm{z}}^{(4)})_{2222} &= -31/32
\end{aligned}$$

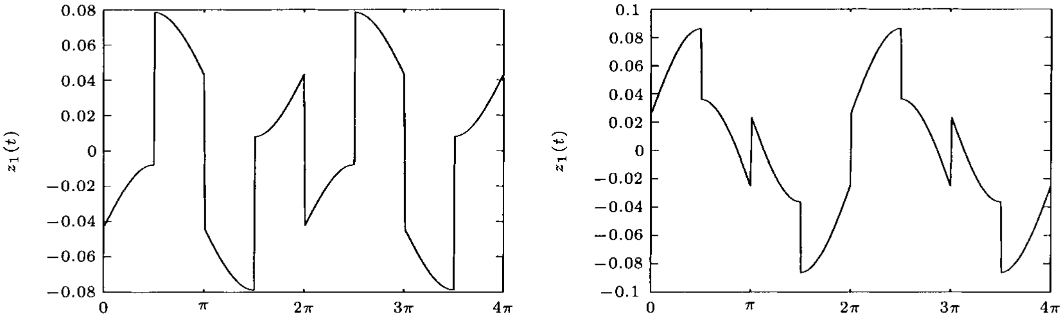


Figure 5. The PCA components of the signals in Figure 4.

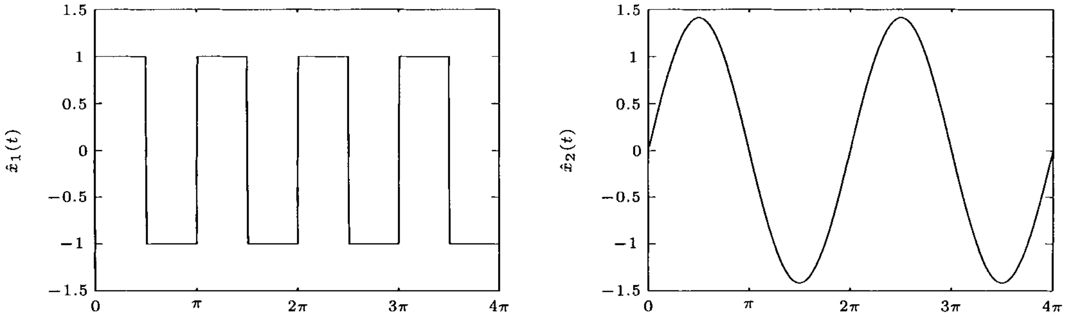


Figure 6. The ICA components of the signals in Figure 4.

It can be verified that **V** satisfies

$$\mathcal{C}_{\mathrm{z}}^{(4)} = \mathcal{C}_{\mathrm{x}}^{(4)} \times_1 \mathbf{V}^{\mathrm{T}} \times_2 \mathbf{V}^{\mathrm{T}} \times_3 \mathbf{V}^{\mathrm{T}} \times_4 \mathbf{V}^{\mathrm{T}}$$

(see Equation (24) and Property 2 of Section 2.2). Moreover, in Section 3.3 we explained that, apart from some trivial indeterminacies, the orthogonal diagonalizer of $\mathcal{C}_{\mathrm{z}}^{(4)}$ is unique. We will discuss four concrete computational procedures in the next section. After the calculation of **V**, one can achieve the source separation. The result is shown in Figure 6.

5. A CLASS OF ALGEBRAIC TECHNIQUES

In Section 3.2.2 we explained that the ICA problem can be solved by means of an appropriate multilinear generalization of the symmetric EVD. Actually, there are various ways in which such a generalization could be defined. In Section 3.3 we explained that in theory the solution is essentially unique (if the higher-order cumulant of the sources has at most one zero on its diagonal), but in practice different approaches may not produce the same result. The reason is that the multilinear generalizations should be defined for arbitrary supersymmetric higher-order tensors, and not merely for higher-order tensors that can be diagonalized by means of an orthogonal transformation, as in Equation (24), since the latter property is generically lost when noise is present. As such, different multilinear generalizations have their own identifiability conditions, perturbation properties, etc. This is particularly relevant when the noise level is significant.

In this section we will discuss a class of four multilinear EVD generalizations. The rationale behind these approaches is explained and the main theoretical results are stated. It is briefly explained how the orthogonal factor **V** in Equation (24) can be calculated in each of the four cases. However, for detailed computational procedures the reader is referred to the extended report [21] or to the original references. Also for details about the derivations the reader is referred to the literature.

The exposition requires that the reader is familar with some basic concepts of linear algebra and related numerical issues. We refer to References [5,14].

5.1. ICA by means of higher-order eigenvalue decomposition (HOEVD)

As will be explained below, the decomposition defined in the following theorem fits the form required in Equation (24) [22].

Theorem 5 (Nth-order supersymmetric eigenvalue decomposition)

Every *N*th-order supersymmetric $J \times J \times \ldots \times J$ tensor $\mathcal{A}$ can be written as the product

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U} \times_2 \mathbf{U} \ldots \times_N \mathbf{U} \tag{36}$$

in which:

- $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_J]$ is an orthogonal $J \times J$ matrix;
- $\mathcal{S}$ is an *N*th-order supersymmetric $J \times J \times \ldots \times J$ tensor of which the subtensors $\mathcal{S}_{i_n = \alpha}$, obtained by fixing the *n*th index to α, have the property of all-orthogonality, i.e. two subtensors $\mathcal{S}_{i_n = \alpha}$ and $\mathcal{S}_{j_n = \beta}$ are orthogonal for all possible values of *n*, α and β subject to $\alpha \neq \beta$:

$$\langle \mathcal{S}_{j_n=\alpha}, \mathcal{S}_{j_n=\beta} \rangle = 0 \quad \text{when} \quad \alpha \neq \beta \tag{37}$$

The unsymmetric variant of this decomposition is also known as the Tucker model in psychometrics [23,24].

Applied to a supersymmetric third-order tensor $\mathcal{A}$, Theorem 5 says that it is always possible to find an orthogonal transformation of the column, row and mode-3 space such that the supersymmetric tensor $\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^{\mathrm{T}} \times_2 \mathbf{U}^{\mathrm{T}} \times_3 \mathbf{U}^{\mathrm{T}}$ is all-orthogonal. This means that the different 'horizontal matrices' of $\mathcal{S}$ (the first index i_1 is kept fixed, whilst the other two 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, and because of the symmetry, the different 'frontal' matrices (i_2 fixed) and the different 'vertical' matrices (i_3 fixed) should be mutually orthogonal as well. This is illustrated in Figure 7.

It is clear that Theorem 5 is a multilinear generalization of the EVD of symmetric matrices, as diagonality is a special case of all-orthogonality. Relaxation of the condition of diagonality to all-orthogonality is required to ensure that the decomposition always exists. It can even be shown that the decomposition exhibits essentially the same uniqueness properties as the matrix EVD. Moreover, it is a true generalization of the matrix decomposition in the sense that, when Theorem 5 is applied to matrices (second-order tensors), it leads to the classical matrix EVD. In other words, in the definition of the EVD of symmetric matrices the constraint of diagonality may be replaced by the condition of all-orthogonality—for matrices the result is the same, up to some trivial normalization conventions.

There are many striking analogies between the matrix EVD and the multilinear generalization of Theorem 5. In this respect we use the term *higher-order eigenvalue decomposition* (HOEVD) in this paper, for convenience. Note at this point that the existence of different types of multilinear EVD extensions may not be excluded—as a matter of fact, focusing on different properties of the matrix EVD does lead to the definition of different (perhaps formally less striking) multilinear

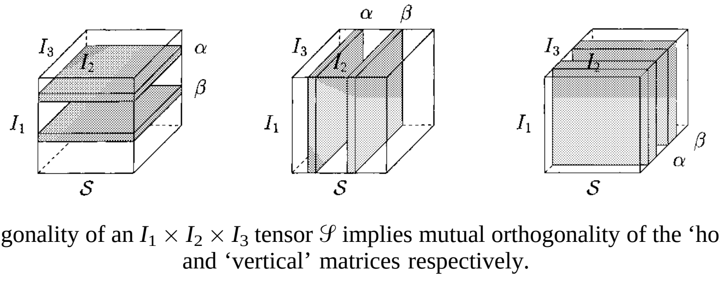


Figure 7. All-orthogonality of an $I_1 \times I_2 \times I_3$ tensor $\mathcal{S}$ implies mutual orthogonality of the 'horizontal', 'frontal' and 'vertical' matrices respectively.

generalizations; e.g. the techniques described in the following subsections can be considered as alternatives.

The HOEVD can be obtained as follows. The eigenmatrix **U** can be computed as the left singular matrix of a matricized version of $\mathscr{A}$ (see the terminology proposed in Reference [25]), denoted by $\mathbf{A}_{(n)}$, in which all the mode-n vectors are stacked as columns (the ordering is of no importance); the tensor $\mathscr{S}$ then follows from Equation (36).

As tensor diagonality is a special case of all-orthogonality, Equation (24) shows an HOEVD of the cumulant tensor $\mathscr{C}_{\mathrm{z}}^{(N)}$. Hence the higher-order step in the scheme of Section 3.2.2 can be interpreted as an HOEVD. To obtain a unique solution, slightly stronger conditions are required than the ones discussed in Section 3.3: to make the SVD of the matrix unfolding of $\widehat{\mathscr{C}}_{\mathrm{z}}^{(N)}$ essentially unique, it is required that the singular values are mutually different. It can be shown that this implies that the Nth-order cumulants of the components of **z** should be mutually different (N even), or mutually different in absolute value (N odd) [18].

5.2. *ICA by means of maximal diagonality (MD)*

We already stressed that a generic higher-order tensor cannot be diagonalized by means of orthogonal transformations. In the previous subsection this problem was intercepted by replacing the condition of diagonality of the matrix of eigenvalues, in the definition of the matrix EVD, by the condition of all-orthogonality in a multilinear EVD equivalent. An interesting alternative definition of a tensorial EVD could involve the optimal diagonalization, in a least squares sense, of the higher-order tensor. This criterion will be called *maximal diagonality* (MD). Formally, the factor **V** in Equation (24) will be estimated as the orthogonal matrix **U** that maximizes the function

$$f(\mathbf{U}) = \sum_{j} c'^{2}_{jj\ldots j} \tag{38}$$

in which the $J \times J \times \ldots \times J$ tensor $\mathscr{C}'$ is defined as

$$\mathscr{C}' \stackrel{\mathrm{def}}{=} \mathscr{C}_{\mathrm{z}}^{(N)} \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \ldots \times_N \mathbf{U} \tag{39}$$

A computational procedure was proposed in Reference [16]. The idea is to base the optimal diagonalization of the standardized higher-order cumulant $\mathscr{C}_{\mathrm{z}}^{(N)}$ in Equation (24) on a multilinear generalization of the Jacobi technique for the computation of the EVD of a symmetric matrix [14]. Hence the unknown factor **V** in Equation (24) is estimated as a product of elementary Jacobi rotations, where each elementary rotation maximally diagonalizes the $2 \times 2 \times \ldots \times 2$ subtensor associated with the marginal cumulants of the estimates of two different source components. For example, if one aims to address the j_1th and j_2th source component, then the part of the global cumulant tensor consisting of the entries of which each index is equal to either j_1 or j_2 is considered. All the possible source combinations (i.e. all the different choices of the pair j_1/j_2) are addressed one after the other in a fixed order, and one iterates over such ‘sweeps’. For example, in an ICA problem with three sources, one may sweep over the pairs (1,2), (1,3) and (2,3).

For a detailed description of the computational procedure we refer to References [16,21].

5.3. *ICA by means of joint approximate diagonalization of eigenmatrices (JADE)*

Consider again the fourth-order cumulant $\mathscr{C}_{\mathrm{z}}^{(4)}$ of the standardized random vector **z** in the higher-order ICA stage, as in Equation (24). A very efficient technique can be derived by interpreting this fourth-order tensor as the representation of a matrix-to-matrix mapping, in the same way as a matrix represents a vector-to-vector mapping, and examining the structure of this transformation [26].

Formally, a linear mapping is associated with $\mathcal{C}_z^{(4)}$ in the following way:

$$\mathbf{A}' = \mathcal{C}_z^{(4)}(\mathbf{A}) \quad \Longleftrightarrow \quad a'_{ij} = \sum_{kl} (\mathcal{C}_z^{(4)})_{ijkl} a_{kl} \tag{40}$$

for all index values. In analogy with the EVD of a classical vector-to-vector mapping, one can look for matrices that are simply rescaled by the mapping (40). These matrices are called 'eigenmatrices'; the scaling factors are generalized eigenvalues. In analogy with the EVD of a symmetric matrix, eigenmatrices corresponding to different eigenvalues are mutually orthogonal, owing to the symmetry of the mapping. Taking Equation (24) into account, one can derive that the EVD has the form

$$\mathcal{C}_z^{(4)} = \sum_{j}^{J} \kappa_{x_j} (\mathbf{v}_j^{\mathrm{T}} \mathbf{v}_j) \circ (\mathbf{v}_j^{\mathrm{T}} \mathbf{v}_j) \tag{41}$$

in which:

- the eigenvalues κ_{xj} $(1 \leqslant j \leqslant J)$ correspond to the source cumulants;
- the eigenmatrices $\mathbf{v}_j^{\mathrm{T}} \mathbf{v}_j (1 \leqslant j \leqslant J)$ are equal to the outer products of the rows of $\mathbf{V}$, represented by $\{\mathbf{v}_j\}$, with themselves—note that the eigenmatrices are indeed mutually orthogonal if $\{\mathbf{v}_j\}$ are mutually orthogonal, i.e. if $\mathbf{V}$ is orthogonal then Equation (41) defines an EVD structure.

We conclude that, in the absence of noise, the unknown $\mathbf{V}$ can directly be obtained from the EVD (41). Moreover, we remark that all the matrices in the range of $\mathcal{C}_z^{(4)}$ can be written as a linear combination of the eigenmatrices (in the same way as any vector in the range of a vector-to-vector mapping can be written as a linear combination of the eigenvectors associated with the eigenvalues that are different from zero), such that they can be diagonalized by $\mathbf{V}$. That is, if a matrix $\mathbf{T}$ is in the range of (40), it can be expanded as

$$\mathbf{T} = \sum_j d_j \mathbf{v}_j^{\mathrm{T}} \mathbf{v}_j$$

for a certain set of scalars $\{d_j\}$; this can be rewritten as the EVD

$$\mathbf{T} = \mathbf{V}^{\mathrm{T}} \cdot \mathbf{D} \cdot \mathbf{V}$$

in which $\mathbf{D}$ is a diagonal matrix containing $\{d_j\}$ on the diagonal.

When noise is present and/or when the statistics of z are only known with a limited precision, the derivation above is only approximately valid. Namely, the eigenmatrices of $\mathcal{C}_z^{(4)}$ are not exactly rank-1 matrices and a matrix $\mathbf{T}$ in the range of the mapping cannot be exactly diagonalized by $\mathbf{V}$. Here it makes sense to estimate $\mathbf{V}$ as the orthogonal matrix that simultaneously maximally diagonalizes (in a least squares sense) a set of matrices that form a basis of the range. Formally, if the set to be diagonalized is given by $\{\mathbf{T}_p\}$, $\mathbf{V}$ is estimated as the orthogonal matrix U that maximizes the function

$$f(\mathbf{U}) = \sum_p \left(\sum_j |(\mathbf{T}'_p)_{jj}|^2 \right) \tag{42}$$

in which

$$\mathbf{T}'_p \overset{\text{def}}{=} \mathbf{U} \cdot \mathbf{T}_p \cdot \mathbf{U}^{\mathrm{T}} \tag{43}$$

An orthogonal basis for the range of the linear mapping can be obtained from the EVD in Equation (41), together with a first estimate of **V**—'JADE' stands for 'joint approximate diagonalization of eigenmatrices'. A variant is the simultaneous EVD of the 'matrix slices' $(\mathcal{C}_{\mathrm{z}}^{(4)})_{k,l}$, obtained by keeping the indices *k, l* in $(\mathcal{C}_{\mathrm{z}}^{(4)})_{ijkl}$ fixed while varying the indices *i, j;* these matrices span the range of the mapping as they are the image of the matrices that contain a single 'one' entry and zeros elsewhere. In the latter case, J^2 matrices are jointly diagonalized; in the former case, *J* matrices are considered (if all the sources have a non-vanishing cumulant; for identifiability, at most one vanishing cumulant can be allowed). Of course, an estimate of **V** can also be obtained from a limited set of matrices in the span of the mapping, without exploiting all the information available in $\mathcal{C}_{\mathrm{z}}^{(4)}$.

Let us now explain how this procedure of simultaneously diagonalizing a set of matrices relates to the problem we started from, namely the diagonalization of $\mathcal{C}_{\mathrm{z}}^{(4)}$ itself in Equation (24). Implicitly or explicitly, the simultaneous diagonalization of a basis of the range of (40) amounts to the approximate diagonalization of the matrix slices $(\mathcal{C}_{\mathrm{z}}^{(4)})_{k,l}$, defined in the preceding paragraph, i.e. the orthogonal factor **V** is determined in such a way that the entries of $\mathcal{C}_{\mathrm{z}}^{(4)}$ of which the first and the second index are different are minimized. However, in (24) we see that **V** is applied in a symmetric way over all modes. Hence also the entries with distinct first and third index, distinct first and fourth index, distinct second and third index, etc. are approximately minimized. It is intuitively clear that this can be considered as an approximate diagonalization of $\mathcal{C}_{\mathrm{z}}^{(4)}$ itself.

As far as the computation of the result is concerned, a set of symmetric matrices can simultaneously be diagonalized by means of a Jacobi iteration; we refer to References [21,27].

5.4. *ICA by means of simultaneous third-order tensor diagonalization (STOTD)*

The technique of simultaneous third-order tensor diagonalization (STOTD) is very similar in spirit to JADE. Instead of linking a matrix-to-matrix mapping to $\mathcal{C}_{\mathrm{z}}^{(4)}$, we now associate with it a linear transformation of $\mathbb{R}^J$ to the vector space of third-order tensors, $\mathbb{R}^{J \times J \times J}$, in the following way:

$$\mathcal{A}' = \mathcal{C}_{\mathrm{z}}^{(4)}(\mathbf{a}) \iff a'_{ijk} = \sum_l (\mathcal{C}_{\mathrm{z}}^{(4)})_{ijkl} a_l \tag{44}$$

for all index values. The SVD of this mapping is given by

$$\mathcal{C}_{\mathrm{z}}^{(4)} = \sum_j^J \sigma_j \mathcal{V}_j \circ \mathbf{v}_j^{\mathrm{T}} \tag{45}$$

in which:

- the singular values are given by sign $(\kappa_{x_j})\kappa_{x_j}$ $(1 \leqslant j \leqslant J)$, in which κ_{x_j} symbolizes the cumulant of the *j*th source;
- the corresponding right singular vectors $\mathbf{v}_j^{\mathrm{T}}$ are the columns of $\mathbf{V}^{\mathrm{T}}$ $(1 \leqslant j \leqslant J)$;
- the corresponding 'left singular tensors' $\mathcal{V}_j$ $(1 \leqslant j \leqslant J)$ are given by

$$\mathcal{V}_j = \text{sign } (\kappa_{x_j}) \mathbf{v}_j^{\mathrm{T}} \circ \mathbf{v}_j^{\mathrm{T}} \circ \mathbf{v}_j^{\mathrm{T}} \tag{46}$$

All the third-order tensors in the range of $\mathcal{C}_{\mathrm{z}}^{(4)}$ can be written as a linear combination of the left singular tensors, such that they can be diagonalized by **V**. In the STOTD algorithm, **V** is estimated as the orthogonal matrix that simultaneously diagonalizes, in a least squares sense, a set of third-order tensors that form a basis for the range of $\mathscr{C}_z^{(4)}$. Formally, if the set to be diagonalized is given by $\{\mathscr{T}_p\}$, **V** is estimated as the orthogonal matrix **U** that maximizes the function

$$f(\mathbf{U}) = \sum_p \left(\sum_j ((\mathscr{T}'_l)_{jjj})^2 \right) \tag{47}$$

in which $\mathscr{T}'_p$ equals the tensor $\mathscr{T}_p$ after multiplication with **U**:

$$\mathscr{T}'_p = \mathscr{T}_p \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \mathbf{U} \tag{48}$$

An orthogonal basis for the range of the linear mapping can be obtained from the SVD in Equation (45), together with a first estimate of **V**. It is also possible to resort to an ordinary basis by simple transformation under $\mathscr{C}_z^{(4)}$ of J linearly independent vectors. For example, transformation of the canonical unit vectors corresponds to choosing the ‘third-order tensor slices’ $(\mathscr{C}_z^{(4)})_l$, obtained by keeping the index l in $(\mathscr{C}_z^{(4)})_{ijkl}$ fixed while varying the other three indices.

Reference [28] shows how the simultaneous diagonalization of third-order tensors can be computed by means of a Jacobi iteration; a summary of the results can be found in Reference [21].

5.5. *Numerical experiments*

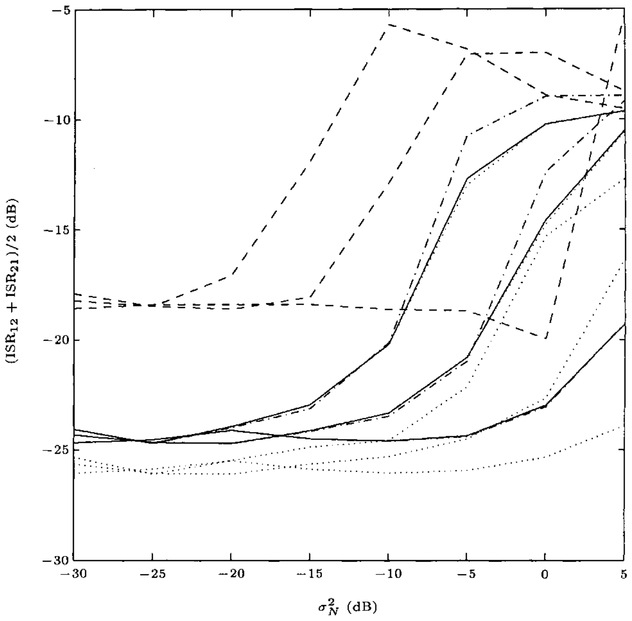
In this subsection we illustrate the performance of the methods that have been discussed, by means of some numerical results.

We consider an ICA problem with two sources and five observation channels; the overestimation of the number of sources serves to limit the influence of noise on the PCA stage, as explained in Section 3.2.1, such that the results reflect the performance of the different approaches for the higher-order stage. The first source distribution is binary (±1), with an equal probability of both values; the second distribution is uniform over the interval $[-\sqrt{3}, \sqrt{3}]$. Both sources are zero-mean and have unit variance. Data sets under consideration consist of 100 samples. For each data set the mixing matrix **M** was generated as follows: the left and right singular matrices were obtained from a QR factorization of a 5×2 and a 2×2 matrix of which the entries had been drawn from a zero-mean Gaussian distribution, and the matrix of singular values was equal to

$$\mathbf{S} = \frac{5}{\sqrt{1+k^2}} \begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}$$

in which k was the condition number. The noise is spatially white Gaussian with variance σ_N^2. Owing to the way **M** was constructed, σ_N^2 can also be interpreted as the inverse of the SNR of the observations (the Frobenius norm of the covariance of the signal part of the observations equals 5; the Frobenius norm of the noise covariance is equal to 5 σ_N). For this problem we conduct a Monte Carlo experiment consisting of 500 runs. Since both source distributions are even, the higher-order stage of the different ICA algorithms is based on fourth-order cumulants.

In Figure 8 we plot the mean ISR of the estimated LCMV filters (Equation (33)). The dotted lines below correspond to the performance bound specified in Equation (34). Figure 9 shows the INSR of the first source estimate obtained from an MVDR filter (Equation (32)). The dotted lines in the middle correspond to the MVDR filter in which the true value for **M** is used; note that the MD, JADE and STOTD filters outperform this informed beamformer. In Figure 10 the Frobenius RMSE of the



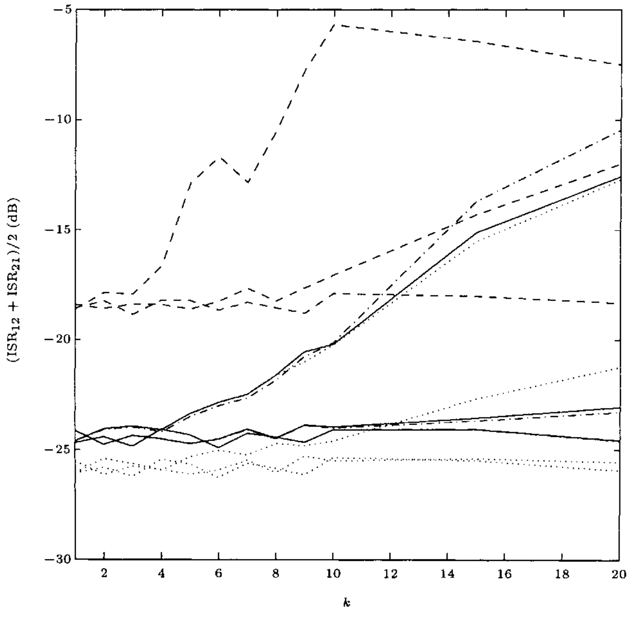


Figure 8. Mean ISR of the LCMV beamformer obtained by means of HOEVD (broken), MD (full), JADE (chain) and STOTD (dotted). Dotted below: performance bound. Top: effect of the SNR on the quality of separation for $k = 1$ (lower curves), 5 (middle curves) and 10 (upper curves). Bottom: effect of the condition number on the quality of separation for SNR = 30 dB (lower curves), 20 dB (middle curves) and 10 dB (upper curves).

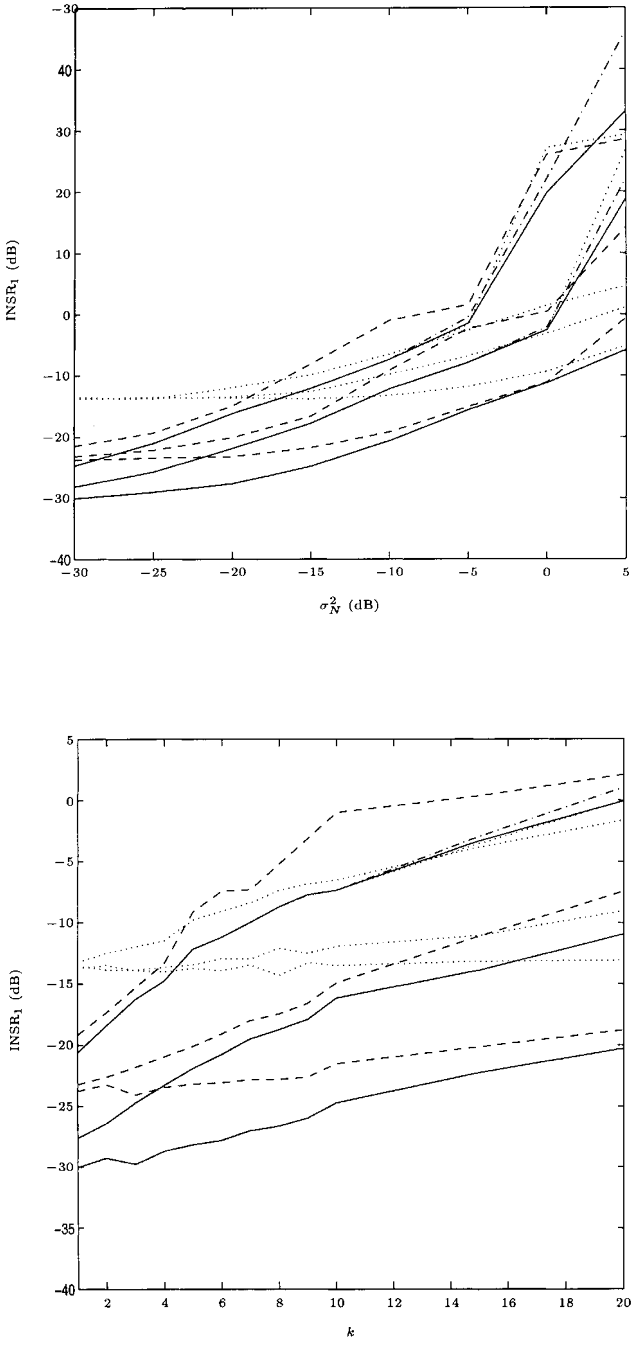
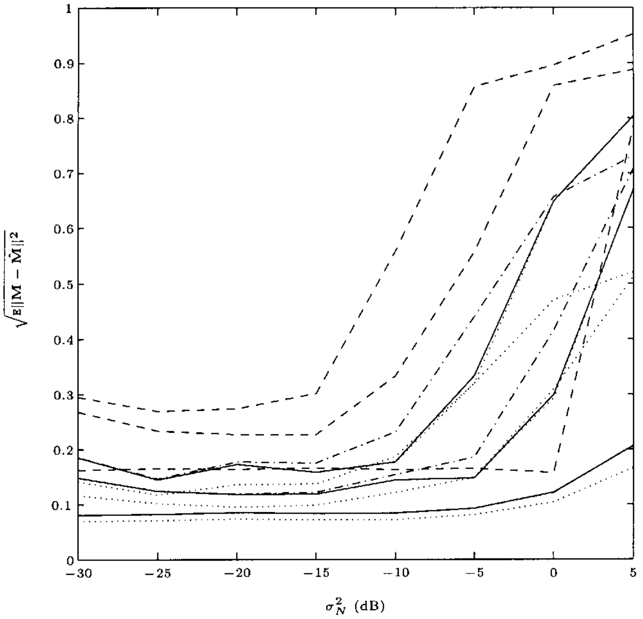


Figure 9. Mean INSR of the first source for the MVDR beamformer obtained by means of HOEVD (broken), MD (full), JADE (chain) and STOTD (dotted). Dotted middle: informed beamformer. Top: effect of the SNR on the quality of separation for $k = 1$ (lower curves), 5 (middle curves) and 10 (upper curves). Bottom: effect of the condition number on the quality of separation for SNR = 30 dB (lower curves), 20 dB (middle curves) and 10 dB (upper curves).



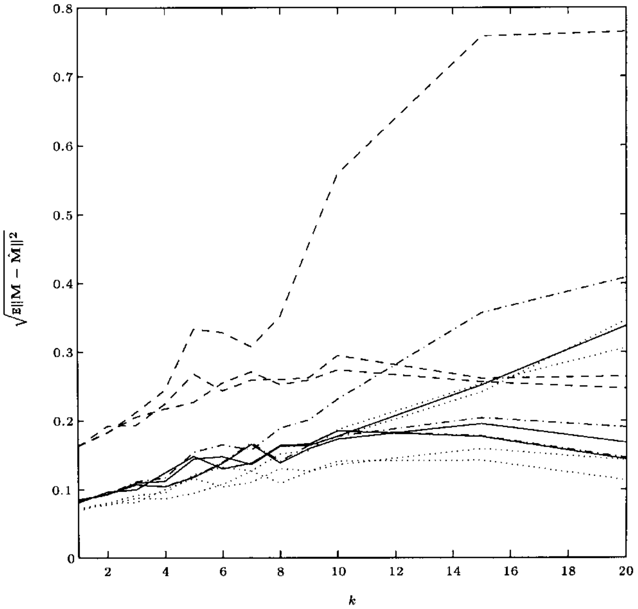


Figure 10. Frobenius RMSE for the mixing matrix estimate obtained by means of HOEVD (broken), MD (full), JADE (chain) and STOTD (dotted). Dotted below: performance bound. Top: effect of the SNR on the quality of separation for $k = 1$ (lower curves), 5 (middle curves) and 10 (upper curves). Bottom: effect of the condition number on the quality of separation for SNR = 30 dB (lower curves), 20 dB (middle curves) and 10 dB (upper curves).

mixing matrix estimate is plotted; to facilitate the comparison of the different curves, the mixing vectors are normalized to unit length. The dotted lines below show the performance bound of Equation (35). In each figure the left plot shows the performance as a function of the noise level for three different values of k, while the right plot shows the influence of how close the mixing vectors are for three different values of the SNR.

The figures show that the MD, JADE and STOTD algorithms have approximately the same accuracy; moreover, for sufficiently high SNRs and sufficiently low condition numbers the results are close to the performance bound. This is theoretically founded in Reference [29]. Because of the heavier computational load of the MD technique, the JADE and STOTD algorithms are preferable. STOTD has the slight advantage over JADE that the J tensors to be diagonalized are readily available, whereas JADE requires the computation of the dominant J-dimensional eigenspace of a $J^2 \times J^2$ matrix to reduce the number of matrices to J; in addition, JADE seems to be a bit more vulnerable w.r.t. a bad conditioning of the mixing matrix. The HOEVD technique is the least accurate of the four approaches, but on the other hand it is also by far the cheapest. Intuitively the difference in accuracy can be explained by taking in mind that in the HOEVD approach diagonality of the source cumulant tensor is a premise, while in the MD technique it is explicitly forced, i.e. in the MD technique the source estimates are explicitly made ‘as independent as possible’, while in the HOEVD technique statistical independence is merely presupposed; the same argument applies to JADE and STOTD.

6. CONCLUSION

In many applications, ICA is the natural way to set off the rotational invariance of PCA. From an algebraic point of view, ICA amounts to multilinear generalizations of the symmetric EVD. We briefly discussed a framework of four types of orthogonal transformation, in which the condition of diagonality of the matrix of eigenvalues was replaced by (i) all-orthogonality, (ii) maximal diagonality, (iii) maximal joint diagonality of a set of matrices and (iv) maximal joint diagonality of a set of third-order tensors.

ACKNOWLEDGEMENTS

This work is supported by several institutions:

(1) the Flemish Government:

  (a) Research Council KU Leuven: Concerted Research Action GOA-MIPS (Model-based Information Processing Systems) and GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology);
  (b) FWO projects G.0240.99 (Multilinear Generalisations of the Singular Value Decomposition and Applications in Signal Processing and System Identification) and G.0256.97 (Numerical Algorithms for Subspace System Identification, Extension to Special Cases);
  (c) FWO Research Communities ICCoS (Identification and Control of Complex Systems) and ANMMM (Advanced Numerical Methods for Mathematical Modelling);

(2) the Belgian State, Prime Minister’s Office—Federal Office for Scientific, Technical and Cultural Affairs—Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997–2001): Modeling, Identification, Simulation and Control of Complex Systems; and IUAP P4-24 (1997–2001): Intelligent Mechatronic Systems (IMechS)).

The scientific responsibility is assumed by its authors.

REFERENCES

1. De Lathauwer L, Comon P. (eds). *Signal Processing* 1999; **73**: 1–2.
2. Cardoso J-F, Jutten C, Loubaton P. (eds). *Proc. ICA'99. First Int. Workshop on Independent Component Analysis and Signal Separation,* Aussois, 1999.
3. Bennett JH (ed.). *Collected Papers of R. A. Fisher*, vol. 2. University of Adelaide Press: Adelaide, 1972; 351–354.
4. Thiele TN. The theory of observations. *Ann. Math. Statist.* 1931; **2**: 165–308.
5. Horn RA, Johnson CR. *Topics in Matrix Analysis*. Cambridge University Press: New York, 1991.
6. Gardner WA. *Introduction to Random Processes*. McGraw-Hill: New York, 1990.
7. Nikias CL, Mendel JM. Signal processing with higher-order spectra. *IEEE Signal Procsessing Mag.* 1993; 10–37.
8. Nikias CL, Petropulu AP. *Higher-order Spectra Analysis. A Nonlinear Signal Processing Framework*. Prentice-Hall: Englewood Cliffs, NJ, 1993.
9. Kendall MG, Stuart A. *The Advanced Theory of Statistics*, vol. 1. Griffin: London, 1977.
10. Bourin C, Bondon P. Efficiency of high-order moment estimates. *Proc. IEEE Signal Processing/ATHOS Workshop on Higher-order Statistics*, Girona, 1995; 186–190.
11. Comon P, Mourrain B. Decomposition of quantics in sums of powers of linear forms. *Signal Processing* 1996; **53**: 93–108.
12. Van Veen BD, Buckley KM. Beamforming: a versatile approach to spatial filtering. *IEEE ASSP Mag.* 1988; 4–24.
13. Wax M, Kailath T. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust., Speech, Signal Processing* 1986; **33**: 387–392.
14. Golub GH, Van Loan CF. *Matrix Computations* (3rd edn). Johns Hopkins University Press: Baltimore, MD, 1996.
15. Bro R. PARAFAC. Tutorial and applications. *Chemomometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
16. Comon P. Independent component analysis, a new concept? *Signal Processing* 1994; **36**: 287–314.
17. Tong L, Liu R, Soon V, Huang Y-F. Indeterminacy and identifiability of blind identification. *IEEE Trans. Circ. Syst.* 1991; **38**: 499–509.
18. De Lathauwer L. Signal processing based on multilinear algebra. PhD Thesis, KU Leuven, 1997.
19. Cardoso J-F. On the performance of orthogonal source separation algorithms. *Proc. EUSIPCO-94*, Edinburgh, 1994; vol. 2, 776–779.
20. De Lathauwer L, De Moor B, Vandewalle J. A residual bound for the mixing matrix in ICA. *Proc. EUSIPCO-98*, Rhodes, 1998; 2065–2068.
21. De Lathauwer L, De Moor B, Vandewalle J. Orthogonal super-symmetric tensor decompositions and independent component analysis. *Tech. Rep. 99-27*, ESAT/SISTA, KU Leuven, 1999.
22. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *Tech. Rep. 94-31*, ESAT/SISTA, KU Leuven, 1994.
23. Tucker LR. The extension of factor analysis to three-dimensional matrices. In *Contributions to Mathematical Psychology*, Gulliksen H, Frederiksen N. (eds). Holt, Rinehart and Winston: New York, 1964; 109–127.
24. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
25. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
26. Cardoso J-F, Souloumiac A. Blind beamforming for non-Gaussian signals. *IEE Proc. F* 1994; **140**: 362–370.
27. Souloumiac A, Cardoso J-F. Jacobi angles for simultaneous diagonalization. *SIAM J. Matrix Anal. Appl.* 1996; **17**: 161–164.
28. De Lathauwer L, De Moor B, Vandewalle J. Blind source separation by simultaneous third-order tensor diagonalization. *Proc. EUSIPCO-96*, Trieste, 1996; vol. 3, 2089–2092.
29. Souloumiac A, Cardoso J-F. Performances en séparation de sources. *Proc. GRETSI*, Juan les Pins, 1993; 321–324.
30. Abramowitz M, Stegun I (eds). *Handbook of Mathematical Functions*. Dover: New York, 1968; 17–18.
31. Kay DC. *Theory and Problems of Tensor Calculus*. McGraw-Hill: New York, 1988.