Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 1995-64C

Higher-Order Power Method - Application in Independent Component Analysis [1]

Lieven De Lathauwer[2] Pierre Comon[3] Bart De Moor[4]
Joos Vandewalle

December 1995

Invited paper at the special session:
Blind Separation of Sources - Brain Information Processing
1995 Int. Symposium on Nonlinear Theory and Its Applications
NOLTA'95
Caesars Palace, Las Vegas, USA
December 10-14, 1995

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/SISTA/delathauwer/reports/ldl-95-64.ps.Z*

[2]ESAT - Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, 3001 Leuven (Heverlee), Belgium, tel 32/16/32 18 05, fax 32/16/32 19 86, email: *Lieven.DeLathauwer@esat.kuleuven.ac.be*. Lieven De Lathauwer is a research assistant with the I.W.T. (Flemish Institute for Support of Scientific-Technological Research in Industry). The research reported in this paper was partially supported by the Belgian Programme on Interuniversity Poles of Attraction (IUAP-17, IUAP-50), initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. It was also supported by the European Community Research program ESPRIT, Basic Research Working Group nr. 6620 (ATHOS). The scientific responsibility rests with its authors.

[3]I3S-CNRS, 250 rue Albert Einstein, F-06560 Sophia-Antipolis, France

[4]Bart De Moor is a research associate with the N.F.W.O. (Belgian National Fund for Scientific Research).

Higher-Order Power Method - Application in Independent Component Analysis *

Lieven De Lathauwer*, Pierre Comon**, Bart De Moor* and Joos Vandewalle*

* K.U.LEUVEN, ESAT-SISTA, Kard. Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium
Tel/Fax: +32 [(0)16] 321085 / 321986, E-mail:Lieven.Delathauwer@esat.kuleuven.ac.be
** THOMSON-SINTRA, BP 157, F-06903 Sophia Antipolis Cedex, France
Tel/Fax: +33 92964244 / 92963977, E-mail: comon@asm.thomson.fr

Abstract

It is investigated how a given higher-order tensor can be approximated in least-squares sense by a rank-1 tensor. A multilinear equivalent of the power method is proposed. It is shown that the tensor approximation yields a new line of approach to the problem of Independent Component Analysis.

I. Introduction

The scientific boom in the field of higher-order statistics involves an increasing need for numerical tools in multilinear algebra: higher-order moments and cumulants of multivariate stochastic processes are higher-order tensors. The purpose of this paper is twofold: first we derive a power algorithm for the computation of the best rank-1 approximation of higher-order tensors, and secondly a link with the problem of blind source separation is explicited.

With respect to the best rank-1 approximation, the paper is organized as follows. After a short introduction of some basic definitions of multilinear algebra in Section II.A, Section II.B shows that the approximation problem leads in a very natural way to a higher-order equivalent of the well-known power method, used for the computation of the Eigenvalue Decomposition (EVD) and the Singular Value Decomposition (SVD) of matrices. In the tensor case several stationary points might exist, each with a different domain of attraction. The choice of an appropriate initial guess, in the attraction region of the global optimum, is addressed in Section II.C. The power equations have independently been proposed in [6], however without mentioning the monotonous convergence properties of the algorithm, and neither taking into account the presence of spurious stationary points.

As far as the problem of blind source separation is concerned, let us first recall the basic linear statistical model:

$$Y = \mathbf{M}X + N \tag{1}$$

In this equation *observation vector* Y, *source vector* X and *noise vector* N are zero-mean random vectors with values in $\mathbb{R}$ or $\mathbb{C}$. The *transfer matrix* $\mathbf{M}$ has linearly independent columns. The components of X are mutually statistically independent, as well as statistically independent from the noise components.

The goal of blind source separation, or Independent Component Analysis (ICA), now consists of the estimation of $\mathbf{M}$ and the corresponding realizations of X, given only realizations of Y.

Considering the fundamental character of the ICA-problem it comes as no surprise that quite a lot of algorithms have been proposed since the first publication on this subject - for an overview see [1]. The procedure derived in our paper is a variant of one basic technique presented in the same article.

Section III recapitulates the main results of [1] that will be used in this paper. For an extensive study of the basic algorithm, and for the development of the statistical framework of contrast functions in which this algorithm has been derived, the reader is referred to the original article. In a nutshell, we have the following. ICA can be performed by a two-stage procedure of which the first step consists of whitening the data by diagonalizing their (second-order) covariance matrix. In this way the transfer matrix can be estimated up to an orthogonal (unitary) transformation (Section III.A). In the next step the remaining

*The research reported in this paper was partially supported by the Belgian Program on Interuniversity Attraction Poles (IUAP-17, IUAP-50), the European Community Research program ESPRIT, Basic Research Working Group nr. 6620 (ATHOS) and the Flemish Institute for Support of Scientific-Technological Research in Industry (I.W.T.). L. De Lathauwer is a Research Assistant supported by the I.W.T.; B. De Moor is a Research Associate of the National Fund for Scientific Research (N.F.W.O.)

factor is estimated by (approximate) diagonalization of the fourth-order cumulant of the whitened data. The diagonalization procedure can be carried out as a Jacobi-type iteration on the two-dimensional subproblems (Section III.B).

As a new contribution, Section IV shows that a two-dimensional ICA-problem can be formulated in terms of the best rank-1 approximation of a $(2 \times 2 \times 2 \times 2)$-tensor.

Our presentation is restricted to the case of real-valued data.

II. Higher-order power method

A. Basic definitions and notations

We study the least-squares approximation of a higher-order tensor by a rank-1 tensor. A rank-1 tensor is defined as follows:

Definition 1 *An Nth-order tensor $\mathcal{A}$ has* rank 1 *when it equals the outer product of N vectors U, V, $\ldots Z$:*

$$\mathcal{A} = U \circ V \circ \ldots \circ Z \tag{2}$$

i.e. $\mathcal{A}_{i_1 i_2 \ldots i_N} = U_{i_1} V_{i_2} \ldots Z_{i_N}$ *for all index values.*

Although this will not further be explained, our results can be generalized to the best rank-$(r_1, r_2, \ldots, r_N)$ approximation of higher-order tensors [4], in which the rank quantities are defined according to:

Definition 2 *The* mode-n rank *of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, denoted by r_n, is defined as the dimension of the vector space generated by the $I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N$ vectors, obtained by varying the index i_n of $\mathcal{A}_{i_1 i_2 \ldots i_N}$ while keeping the other indices fixed.*

The mode-n rank of $\mathcal{A}$ corresponds to the rank of its mode-n matrix unfolding:

Definition 3 *The* mode-n matrix unfolding *of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, denoted by $\mathbf{A}^{(n)}$, is the $(I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N)$-matrix that contains the element $\mathcal{A}_{i_1 i_2 \ldots i_N}$ on the position given by row number i_n and column number $(i_1 - 1) I_2 I_3 \ldots I_{n-1} I_{n+1} \ldots I_N + (i_2 - 1) I_3 \ldots I_{n-1} I_{n+1} \ldots I_N + \ldots + i_N$.*

Before starting the actual derivation, we introduce some additional concepts of multilinear algebra:

Definition 4 *The scalar product $\langle \mathcal{A}, \mathcal{B} \rangle$ of two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$ is defined as*

$$\langle \mathcal{A}, \mathcal{B} \rangle \stackrel{\text{def}}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} \mathcal{B}_{i_1 i_2 \ldots i_N} \tag{3}$$

Due to the introduction of a scalar product it becomes possible to treat tensors in a geometrical way. We have e.g.:

Definition 5 *Tensors of which the scalar product equals 0, are mutually orthogonal.*

Definition 6 *The Frobenius-norm of a tensor $\mathcal{A}$ is given by*

$$\|\mathcal{A}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} \tag{4}$$

The Frobenius-norm can be interpreted as the "size" of the tensor. The square of this norm can be seen as the "energy" in the tensor.

B. Derivation of the power algorithm

The problem we want to solve, can mathematically be described as follows:

Given a real Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, find a scalar λ and unit-norm vectors $U, V, \ldots, Z$ such that the rank-1 tensor $\hat{\mathcal{A}} \stackrel{\text{def}}{=} \lambda U \circ V \circ \ldots \circ Z$ minimizes the least-squares criterion

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

This constrained optimization problem can be analyzed using the technique of Lagrange multipliers, which eventually results in the following equations:

$$\begin{aligned}
\sum_{i_2 i_3 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} V_{i_2} \ldots Z_{i_N} &= \lambda U_{i_1} \quad (6)\\
\sum_{i_1 i_3 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} U_{i_1} \ldots Z_{i_N} &= \lambda V_{i_2} \quad (7)\\
&\vdots\\
\sum_{i_1 i_2 \ldots i_{N-1}} \mathcal{A}_{i_1 i_2 \ldots i_N} U_{i_1} \ldots Y_{i_{N-1}} &= \lambda Z_{i_N} \quad (8)
\end{aligned}$$

This set of coupled non-linear equations could also have been derived for the following equivalent problem:

Given a real Nth-order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, find unit-norm vectors $U, V, \ldots, Z$ that maximize the criterion

$$g(U, V, \ldots, Z) = \Big(\sum_{i_1 i_2 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} U_{i_1} V_{i_2} \ldots Z_{i_N} \Big)^2 \tag{9}$$

The role played by λ in the second formulation is as follows: $g(U, V, \ldots, Z) = \lambda^2$, as can be seen from any of the equations (6-8).

The Lagrange equations can have several solutions. Consider the rank-1 tensor $\hat{\mathcal{A}}$, associated to an arbitrary solution. It follows from any of the equations that $\langle \mathcal{A}, \hat{\mathcal{A}} \rangle = \lambda^2$. By some formula manipulations it can be proved that the estimate is orthogonal to the residu: $\langle \hat{\mathcal{A}}, \mathcal{A} - \hat{\mathcal{A}} \rangle = 0$. It follows that the first criterion function $f(\hat{\mathcal{A}}) = \|\mathcal{A}\|^2 - \lambda^2$.

For the actual computation of the best rank-1 approximation of $\mathcal{A}$, Eqs. (6-8) can be interpreted in a constructive way. E.g. Eq. (6) shows how the optimal λU can be computed when the vectors $V, \ldots, Z$ would

Algorithm 1
Higher-Order Power Method

1. Initial values: $V^{(0)}$ is the dominant left singular vector of $\mathbf{A}^{(2)}$; $\ldots$; $Z^{(0)}$ is the dominant left singular vector of $\mathbf{A}^{(N)}$

2. Iterate until convergence:

   (a) $\tilde{U}^{(k+1)}_{i_1} = \sum_{i_2 i_3 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} V^{(k)}_{i_2} \ldots Z^{(k)}_{i_N}$;
   $\mu^{(k+1)} = \|\tilde{U}^{(k+1)}\|$;
   $U^{(k+1)} = \tilde{U}^{(k+1)}/\mu^{(k+1)}$;

   (b) $\tilde{V}^{(k+1)}_{i_2} =$
   $\sum_{i_1 i_3 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} U^{(k+1)}_{i_1} \ldots Z^{(k)}_{i_N}$;
   $\nu^{(k+1)} = \|\tilde{V}^{(k+1)}\|$;
   $V^{(k+1)} = \tilde{V}^{(k+1)}/\nu^{(k+1)}$;

   (c) $\ldots$

   (d) $\tilde{Z}^{(k+1)}_{i_N} =$
   $\sum_{i_1 i_2 \ldots i_{N-1}} \mathcal{A}_{i_1 i_2 \ldots i_N} U^{(k+1)}_{i_1} \ldots Y^{(k+1)}_{i_{N-1}}$;
   $\rho^{(k+1)} = \|\tilde{Z}^{(k+1)}\|$;
   $Z^{(k+1)} = \tilde{Z}^{(k+1)}/\rho^{(k+1)}$;

have been fixed. Hence the system of equations can be linked to a descent algorithm for the minimization of $f(\hat{\mathcal{A}})$: in each step the scalar λ and one of the vectors $U, V, \ldots, Z$ are optimized, while the other vectors are kept constant. The result is shown in step 2 of Alg. 1.

Clearly Alg. 1 is a higher-order extension of the well-known power method, by which the best rank-1 approximation of a matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ can be computed [5].

Fig. (1) visualizes a $(2 \times 2 \times 2)$-example: for different initial values we plotted, after each iteration, the angle between $V^{(k)}$ and the horizontal axis versus the angle between $U^{(k)}$ and the horizontal axis. The angles were normalized to the interval $[-\pi/2, +\pi/2]$. The tensor $\mathcal{A}$ was *all-symmetric* (i.e. invariant under all permutations of its indices) and defined by:

$$\begin{cases} \mathcal{A}_{111} = 1.5578; & \mathcal{A}_{222} = 1.1226; \\ \mathcal{A}_{112} = -2.4443; & \mathcal{A}_{221} = -1.0982; \end{cases}$$

We remark the following:

- Even in the case of all-symmetric tensors Alg. 1 leads to unsymmetric intermediate results. A symmetric version of the algorithm, based on the iteration

$$\tilde{U}^{(k+1)}_{i_1} = \sum_{i_2 i_3} \mathcal{A}_{i_1 i_2 i_3} U^{(k)}_{i_2} U^{(k)}_{i_3} \qquad (10)$$

  is unreliable since it does not necessarily decrease the criterion function $f(\hat{\mathcal{A}})$ in a monotonous way.

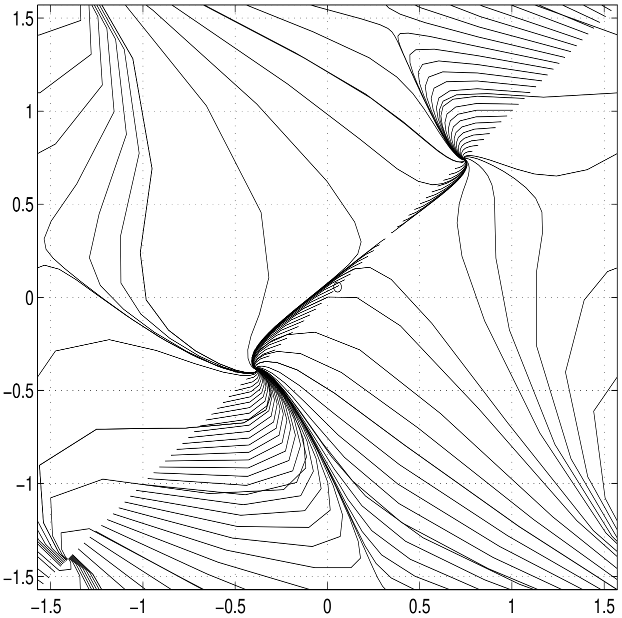

Figure 1: Visualization of Alg. 1 for a typical example in $\mathbb{R}^{2\times 2\times 2}$. Abscis: the angle in radials between $U^{(k)}$ and the horizontal axis. Ordinate: the angle between $Y^{(n)}$ and the horizontal axis. Both angles are normalized to the interval $[-\pi/2, +\pi/2]$. The circle shows the initial guess obtained by HOSVD. The global optimum is $(-0.3860, -0.3860)$.

- There are two stable and two unstable symmetric stationary solutions. Three of them correspond to the three solutions of the Lagrange equations

$$\sum_{i_2 i_3} \mathcal{A}_{i_1 i_2 i_3} U_{i_2} U_{i_3} = \lambda U_{i_1} \qquad (1 \leqslant i_1 \leqslant 2) \quad (11)$$

  on the unit circle $\|U\| = 1$. The solution (0.3347,0.3347) is more special: even after convergence the three substeps of each iteration produce unsymmetric results, but these effects compensate such that the overall power iteration is symmetric.

C. Initial value

In all our simulations we have observed that the initial value for the power algorithm, corresponding to the first higher-order singular vectors of $\mathcal{A}$, belongs to the attraction region of the best rank-1 approximation. We restrict ourselves to a brief reminder of the concept of Higher-Order Singular Value Decomposition (HOSVD); for more details we refer to e.g. [2, 3].

Theorem 1 *Every $(I_1 \times I_2 \times \ldots I_N)$-tensor $\mathcal{A}$ can be written as*

$$\mathcal{A}_{i_1 i_2 \ldots i_N} = \sum_{j_1 j_2 \ldots j_N} \mathcal{S}_{j_1 j_2 \ldots j_N} \mathbf{U}^{(1)}_{i_1 j_1} \ldots \mathbf{U}^{(N)}_{i_N j_N} \qquad (12)$$

in which:

- $\mathbf{U}^{(n)} = \left[U_1^{(n)} U_2^{(n)} \ldots U_{I_n}^{(n)}\right]$ *is an orthogonal* $(I_n \times I_n)$*-matrix*
- *the core tensor* $\mathcal{S}$ *is a* $(I_1 \times I_2 \times \ldots I_N)$*-tensor of which the subtensors* $\mathcal{S}_{i_n=\alpha}$*, obtained by fixing the nth index to* α*, have the properties of:*

  - all-orthogonality: two subtensors $\mathcal{S}_{i_n=\alpha}$ *and* $\mathcal{S}_{i_n=\beta}$ *are orthogonal for all possible values of* n*,* α *and* β *subject to* $\alpha \neq \beta$*:*

  $$\langle \mathcal{S}_{i_n=\alpha}, \mathcal{S}_{i_n=\beta} \rangle = 0 \quad \textit{when} \quad \alpha \neq \beta \tag{13}$$

  - ordering:

  $$\|\mathcal{S}_{i_n=1}\| \geq \|\mathcal{S}_{i_n=2}\| \geq \ldots \geq \|\mathcal{S}_{i_n=r_n}\| > 0 \tag{14}$$

  and

  $$\|\mathcal{S}_{i_n=r_n+1}\| = \ldots = \|\mathcal{S}_{i_n=I_N}\| = 0 \tag{15}$$

  for all possible values of n*.*

The Frobenius-norms $\|\mathcal{S}_{i_n=i}\|$*, symbolized by* $\sigma_i^{(n)}$*, are the n-mode singular values of* $\mathcal{A}$ *and the vector* $U_i^{(n)}$ *is the ith n-mode singular vector.*

Clearly the HOSVD is a formal generalization of the SVD of matrices. Moreover, it can be proved that the HOSVD of a second-order tensor boils down to its matrix SVD. Many SVD-properties, like the link with the EVD, the concept of oriented energy, etc., have already been generalized and show a strong analogy between the matrix and the higher-order case. Figure 2 visualizes the decomposition for a third-order tensor.

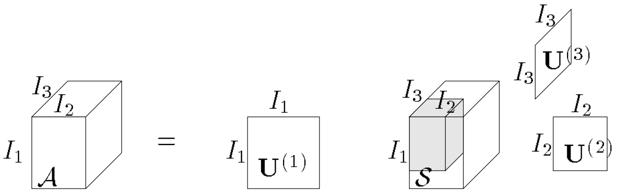

Figure 2: Visualization of the HOSVD of a third-order tensor.

The matrix $\mathbf{U}^{(n)}$ $(1 \leqslant n \leqslant N)$ can be computed as the left singular matrix of $\mathbf{A}^{(n)}$, the mode-n matrix unfolding of $\mathcal{A}$. After computation of the higher-order singular matrices, the core tensor $\mathcal{S}$ can be found as:

$$\mathcal{S}_{j_1 j_2 \ldots j_N} = \sum_{i_1 i_2 \ldots i_N} \mathcal{A}_{i_1 i_2 \ldots i_N} \mathbf{U}^{(1)}_{i_1 j_1} \ldots \mathbf{U}^{(N)}_{i_N j_N} \tag{16}$$

Our observations suggest a remarkable link between matrix and tensor algebra. On one hand it is well-known that the best rank-1 approximation of a matrix corresponds to its SVD, truncated after the first term. On the other hand, restricting the HOSVD to the term in $\mathcal{S}_{11\ldots 1}$ also produces a good rank-1 estimate for higher-order tensors, as can be explained from the properties of the higher-order singular vectors and the ordering constraint of the core tensor. We conjecture (up till now without formal proof) that this estimate is still related to the optimal one: the HOSVD-approximation belongs to the valley of the criterion function $f(\hat{\mathcal{A}})$, corresponding to the best rank-1 approximation.

III. An ICA-algorithm recapitulated

Consider model equation (1). Information on the transfer matrix can only be extracted from the second- and higher-order statistics of the observations. The second-order statistics will be used for standardization of Y, or whitening, to a zero-mean stochastic vector $\tilde{Y}$ with unit covariance matrix. The corresponding standardized transfer matrix belongs to the manifold of orthogonal matrices; it will be estimated from the fourth-order cumulant of $\tilde{Y}$.

A. Step 1: standardization

The covariance $\mathbf{C}_2^Y$ of Y is given by

$$\mathbf{C}_2^Y = \mathbf{M}\mathbf{C}_2^X\mathbf{M}^t + \mathbf{C}_2^N \tag{17}$$

in which $\mathbf{C}_2^X$ and $\mathbf{C}_2^N$ symbolize the covariance of X resp. N.

$\mathbf{C}_2^X$ is diagonal, since the source signals are uncorrelated. We might as well assume that the source signals have unit variance, which at most involves a scaling of the columns of $\mathbf{M}$. This is not detrimental to the method's generality: it is clear that the ICA problem set-up implies that the transfer matrix can only be estimated up to a scaling and a permutation of its columns.

We may conclude that a first estimate of $\mathbf{M}$, up to an orthogonal factor $\mathbf{F}$, can be obtained as a square root of $\mathbf{C}_2^Y$:

$$\mathbf{C}_2^Y = \mathbf{T}\mathbf{T}^t = (\mathbf{T}\mathbf{F}^t)(\mathbf{T}\mathbf{F}^t)^t \simeq \mathbf{M}\mathbf{M}^t \tag{18}$$

We note here that, without a priori knowledge, the noise effects can't be treated in an appropriate way: since the statistical distribution of N is unknown, it is impossible to distinguish between the signal and the noise term in Eq. (17).

Although any congruence transformation of $\mathbf{C}_2^Y$ provides a square root that is theoretically valid, it is preferable to compute the EVD

$$\mathbf{C}_2^Y = \mathbf{E}\mathbf{D}^2\mathbf{E}^t = (\mathbf{E}\mathbf{D})(\mathbf{E}\mathbf{D})^t = \mathbf{T}\mathbf{T}^t \tag{19}$$

since this allows to handle the case of singular or ill-conditioned covariance matrices by an easy projection. When the covariance matrix has been estimated following $\mathbf{C}_2^Y = \frac{1}{P}\mathbf{A}_Y\mathbf{A}_Y^t$ in which $\mathbf{A}_Y$ is an $(I \times P)$-dimensional dataset containing P realizations of Y, then the factor ($\mathbf{ED}$) can be obtained in a numerically more reliable way from the SVD of $\mathbf{A}_Y$ [5].

Finally the standardized observation vector $\tilde{Y}$ can be obtained as

$$\tilde{Y} = \mathbf{D}^{\dagger}\mathbf{E}^t Y \tag{20}$$

in which † denotes the Moore-Penrose pseudo-inverse. This standardization corresponds to nothing else but the Principal Component Analysis (PCA) of the dataset.

B. Step 2: determination of $\mathbf{F}$

The fourth-order cumulant $\mathcal{C}_4^Y$ of Y is given by

$$(\mathcal{C}_4^Y)_{ijkl} = \sum_{pqrs} \mathbf{M}_{ip}\mathbf{M}_{jq}\mathbf{M}_{kr}\mathbf{M}_{ls}(\mathcal{C}_4^X)_{pqrs} + (\mathcal{C}_4^N)_{ijkl} \tag{21}$$

in which $\mathcal{C}_4^X$ and $\mathcal{C}_4^N$ symbolize the fourth-order cumulants of X resp. N. $\mathcal{C}_4^X$ is diagonal, i.e. for non-identical indices the components are zero, since the source signals are statistically independent.

Define the following function over the set of orthogonal matrices:

$$h(\mathbf{Q}) \stackrel{\text{def}}{=} \sum_{i=1}^{I} \mathcal{C}_{iiii}^2 \tag{22}$$

with

$$\mathcal{C}_{ijkl} \stackrel{\text{def}}{=} \sum_{pqrs} \mathbf{Q}_{ip}\mathbf{Q}_{jq}\mathbf{Q}_{kr}\mathbf{Q}_{ls}(\mathcal{C}_4^{\tilde{Y}})_{pqrs} \tag{23}$$

where $\mathcal{C}_4^{\tilde{Y}}$ denotes the standardized fourth-order cumulant. Then the unknown factor $\mathbf{F}$ in $\mathbf{M} \simeq \mathbf{TF}^t$ will be estimated as the orthogonal matrix where $h(\mathbf{Q})$ reaches its global maximum. As in the whitening step it is impossible, without a priori knowledge, to take appropriate care of the noise term.

It can be shown that h, interpreted as a mapping from the set of standardized probability densities to $\mathbb{R}$, is a contrast function. Moreover, if it can be assumed that at most one source component is non-kurtic, the contrast is discriminating over the set of standardized observation vectors. This implies that - at least under noise-free conditions - the solution $\mathbf{F}$ is unique up to the ordering and the sign of its columns.

From an algebraic point of view the optimization can be considered as an EVD-like tensor problem: we look for an orthogonal transformation of $\mathcal{C}$, according to Eq. (23), such that the result is as diagonal as possible in least-squares sense. The diagonalization procedure is carried out as a Jacobi-type iteration, where $\mathbf{F}$ is obtained as a product of elementary Givens rotations. Each elementary rotation diagonalizes as far as possible the $(2 \times 2 \times 2 \times 2)$-subtensor, associated with the marginal cumulants of the estimates of two different source components.

For the elementary subproblem an arbitrary Givens rotation can be denoted as

$$\mathbf{Q} = \frac{1}{\sqrt{1+\theta^2}} \begin{pmatrix} 1 & -\theta \\ \theta & 1 \end{pmatrix} \tag{24}$$

where θ symbolizes the tangent of the rotation angle: $\theta = \tan\alpha$. The optimal rotation can be found by expressing $h(Q)$ as a function of θ, putting its derivative to zero, and selecting the root that corresponds to the global optimum of the contrast function. In [1] it has been derived that the rooting of the derivative can basically be written as the rooting of a polynomial of degree four. We now propose an other formulation.

IV. Alternative formulation of two-dimensional ICA

First we introduce some new notations. For simplicity of the exposition, we use the standard representation of Givens rotations:

$$\mathbf{Q} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \tag{25}$$

Further we denote by $\tilde{\mathcal{C}}$ the $(2 \times 2 \times 2 \times 2)$-cumulant tensor to be diagonalized. We also construct the all-symmetric $(2 \times 2 \times 2 \times 2)$-tensor $\mathcal{T}$, defined by:

$$\begin{aligned}
\mathcal{T}_{1111} &= a && (26)\\
\mathcal{T}_{1112} &= d && (27)\\
\mathcal{T}_{1122} &= 2b/3 + t/2 + a/6 && (28)\\
\mathcal{T}_{1222} &= 3d/4 - e/4 - 3u/2 && (29)\\
\mathcal{T}_{2222} &= 2b + 3t/2 + 9a/8 + 9\,\tilde{\mathcal{C}}_{1122}^2/2 \\
&\quad + \tilde{\mathcal{C}}_{1111}\tilde{\mathcal{C}}_{2222}/4 + 4\,\tilde{\mathcal{C}}_{1112}\tilde{\mathcal{C}}_{1222} && (30)
\end{aligned}$$

in which the auxiliary variables a, b, d, e, t, u are given by:

$$\begin{aligned}
a &= \tilde{\mathcal{C}}_{1111}^2 + \tilde{\mathcal{C}}_{2222}^2 && (31)\\
b &= \tilde{\mathcal{C}}_{1112}^2 + \tilde{\mathcal{C}}_{1222}^2 && (32)\\
d &= \tilde{\mathcal{C}}_{1222}\tilde{\mathcal{C}}_{2222} - \tilde{\mathcal{C}}_{1111}\tilde{\mathcal{C}}_{1112} && (33)\\
e &= \tilde{\mathcal{C}}_{1111}\tilde{\mathcal{C}}_{1222} - \tilde{\mathcal{C}}_{1112}\tilde{\mathcal{C}}_{2222} && (34)\\
t &= \tilde{\mathcal{C}}_{1122}\,(\tilde{\mathcal{C}}_{1111} + \tilde{\mathcal{C}}_{2222}) && (35)\\
u &= \tilde{\mathcal{C}}_{1122}\,(\tilde{\mathcal{C}}_{1112} - \tilde{\mathcal{C}}_{1222}) && (36)
\end{aligned}$$

Using these notations, it is merely a matter of goniometric manipulations to show that the contrast criterion (22) takes the form

$$h(\mathbf{Q}) = \sum_{i,j,k,l=1}^{2} \tilde{\mathcal{T}}_{ijkl} U_i U_j U_k U_l \tag{37}$$

in which $U \stackrel{\text{def}}{=} (\cos(2\alpha) \;\; \sin(2\alpha))^t$. It follows that the contrast function is periodic in the rotation angle, with period $\pi/2$. By limitation of α to the interval

$[-\pi/4, +\pi/4]$, $\mathbf{Q}$ can be restricted to the set of inner rotations. Furthermore Eq. (37) establishes the link with the best rank-1 approximation. Namely, comparison to Eq. (9) proves the following theorem:

Theorem 2 *The best rank-1 approximation of $\mathcal{T}$ is given by*

$$\hat{\mathcal{T}} = \lambda \, U \circ U \circ U \circ U \tag{38}$$

in which $U = (\cos(2\alpha) \;\; \sin(2\alpha))^t$ corresponds to the rotation angle for contrast maximization and λ is the global contrast maximum.

After computation of U by the higher-order power algorithm, the actual elements of an inner Givens rotation can be obtained from the basic goniometric relations $\cos\alpha = (1 + \cos(2\alpha)/2)^{1/2}$ and $\sin\alpha = \sin(2\alpha)/(2\cos\alpha)$.

As a small comparison with the technique proposed in [1] we remark the following:

- In the original paper the computation of the Givens-rotation basically involves the rooting of a polynomial of degree four. The roots can be found via *analytical* expressions. However these formulas are quite cumbersome and involve cube roots of sums which contain square roots. In the alternative formulation the polynomial rooting is replaced by a power *iteration*. This iteration is based on a sequence of simple sums and products. To give a rough idea of the computational cost: on the average, to reduce the relative error on the elements of the Givens rotation to 10^{-5}, 10-25 power iterations seem to be sufficient.
- In the original method all the four roots have to be computed before the global optimum can be selected. The best rank-1 procedure has the advantage that it immediately looks for the global optimum. The price is fairly low: it is the computation of the dominant left singular vector of a (2×8)-matrix, as explained in Section II.C.
- From Eq. (37) it follows that during the power iteration the optimization of the contrast function can be monitored by the evolution of $\mu^{(k)}$, $\nu^{(k)}$, ...(see Alg. 1). This property allows to implement the termination criterion in terms of the value of the contrast. This is interesting, since in Jacobi-type methods it is seldom necessary to compute the rotation angles with very high accuracy: for the global optimization it is more important that the criterion function is optimized in a monotonous way at a sufficient convergence speed. It has even been observed that a reasonable estimate (1-2 digits accuracy) of the transfer matrix can be obtained when each power iteration is reduced to the computation of the HOSVD initial value.

V. Conclusion

We studied the best rank-1 approximation of arbitrary higher-order tensors. From the Lagrange equations of the least-squares criterion function a computation scheme was derived. This algorithm can be interpreted as the higher-order equivalent of the power method for matrices. We proposed to start the iteration at a rank-1 estimate, obtained by truncation of the HOSVD of the original tensor: it is conjectured that this initial value is situated in the attraction domain of the global optimum. The higher-order power algorithm is still a topic of our research, e.g. the relations between the different stationary points are currently being explicited.

We showed that, for a set-up with two sources, blind source separation of a whitened dataset can be formulated in terms of the best rank-1 approximation of an all-symmetric $(2 \times 2 \times 2 \times 2)$-tensor. This leads to a new conceptual approach to ICA.

References

[1] P. Comon, "Independent Component Analysis, A New Concept?", Signal Processing, Special Issue Higher Order Statistics, 36, 3, pp. 287-314, April 1994.

[2] L. De Lathauwer, B. De Moor, J. Vandewalle, "A Singular Value Decomposition for Higher-Order Tensors", Proc. ATHOS Workshop on System Identification and High-Order Statistics, Nice, France, Sept. 20-21, 1993.

[3] L. De Lathauwer, B. De Moor, J. Vandewalle, The Higher-Order Singular Value Decomposition, int. report 94-31, ESAT/SISTA, K.U. Leuven, 1994.

[4] L. De Lathauwer, Signaalverwerking m.b.v. Hogere-Orde Tensoren, report for I.W.T., Sept. 1995, pp. 5-9.

[5] G.H. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, 1991.

[6] M. Schmutz, "Optimal and suboptimal separable expansions for 3D-signal processing", Patt. Recogn. Letters, 8, pp. 217-220, Elsevier Science Publishers B.V., North-Holland, Nov. 1988.