Dimensionality Reduction in Higher-Order-Only ICA [1]

Lieven De Lathauwer, Bart De Moor and Joos Vandewalle[2]

Published in:

Proc. IEEE Signal Processing Workshop on Higher-Order Statistics
July 21-23, 1997
Banff, Alberta, Canada, pp. 316-320.
EURASIP Student Paper Prize

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/SISTA/delathauwer/reports/ldl-97-59.ps.Z*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SISTA/COSIC, Kardinaal Mercierlaan 94, 3001 Leuven, Belgium, Tel. 32/16/32 18 05, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista.* E-mail: *Lieven.DeLathauwer@esat.kuleuven.ac.be*, *bart.DeMoor@esat.kuleuven.ac.be*, *Joos.Vandewalle@esat.kuleuven.ac.be.* This work has partly been supported by the Flemish Government (Concerted Research Action GOA-MIPS, the F.W.O. (Fund for Scientific Research - Flanders) project G.0292.95, the F.W.O. project G.0256.97, the F.W.O. Research Communities ICCoS and Advanced Numerical Methods for Mathematical Modelling), the Belgian State (Interuniversity Poles of Attraction IUAP P4-02 and IUAP P4-24) and the European Commission (Human Capital and Mobility Network SIMONET and SCIENCE-ERNSI SC1-CT92-0779). Bart De Moor is a senior Research Associate with the F.W.O. Lieven De Lathauwer is a Research Assistant with the I.W.T. (Flemish Institute for Science and Technology in Industry). The scientific responsibility is assumed by the authors.

Abstract

Most algebraic methods for Independent Component Analysis (ICA) consist of a second-order and a higher-order stage. The former can be considered as a classical Principal Component Analysis (PCA), with a three-fold goal: (a) reduction of the parameter set of unknowns to the manifold of orthogonal matrices, (b) standardization of the unknown source signals to mutually uncorrelated unit-variance signals, and (c) determination of the number of sources. In the higher-order stage the remaining unknown orthogonal factor is determined by imposing statistical independence on the source estimates. Like all correlation-based techniques, this set-up has the disadvantage that it is affected by additive Gaussian noise. However it is possible to solve the problem, in a way that is conceptually blind to additive Gaussian noise, by resorting only to higher-order cumulants. The purpose of this paper is to explain how the dimensionality of the ICA-model can algebraically be reduced to the true number of sources in higher-order-only schemes.

Dimensionality Reduction in Higher-Order-Only ICA

Lieven De Lathauwer Bart De Moor Joos Vandewalle
K.U.Leuven - E.E. Dept.- ESAT - SISTA/COSIC
Kard. Mercierlaan 94, B-3001 Leuven (Heverlee), Belgium
{ Lieven.DeLathauwer Bart.DeMoor Joos.Vandewalle } @esat.kuleuven.ac.be

Abstract

Most algebraic methods for Independent Component Analysis (ICA) consist of a second-order and a higher-order stage. The former can be considered as a classical Principal Component Analysis (PCA), with a three-fold goal: (a) reduction of the parameter set of unknowns to the manifold of orthogonal matrices, (b) standardization of the unknown source signals to mutually uncorrelated unit-variance signals, and (c) determination of the number of sources. In the higher-order stage the remaining unknown orthogonal factor is determined by imposing statistical independence on the source estimates. Like all correlation-based techniques, this set-up has the disadvantage that it is affected by additive Gaussian noise. However it is possible to solve the problem, in a way that is conceptually blind to additive Gaussian noise, by resorting only to higher-order cumulants. The purpose of this paper is to explain how the dimensionality of the ICA-model can algebraically be reduced to the true number of sources in higher-order-only schemes.

1. Introduction

In this paper, the basic ICA-model will be denoted as follows:

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

in which the observed vector Y, the source vector X and the noise vector N are zero-mean random vectors with values in $\mathbb{R}$ or $\mathbb{C}$. The components of X are assumed to be mutually statistically independent, as well as statistically independent from the noise components; the mixing matrix $\mathbf{M}$ has linearly independent columns. As is well-known, the goal of ICA consists of the estimation of the mixing matrix and the corresponding realizations of X, given only realizations of Y.

ICA-algorithms often have the form of a two-stage procedure. First, in a prewhitening step, the second-order statistics of the observation vector Y are used to "standardize" the problem: Y is transformed in a vector $\tilde{Y}$ with unit covariance. In the more-sensors-than-sources case this also implies a projection of Y onto the dominant subspace of the data. The standardization step allows to restrict the unknown mixing matrix to the manifold of orthogonal matrices. In the second step, the orthogonal factor is then obtained from a higher-order cumulant tensor of $\tilde{Y}$. Examples of algebraic prewhitening-based ICA-methods are [4, 3, 7, 10].

On the other hand, it is also possible to identify the mixing matrix by using *only* the higher-order cumulant tensor. For algebraic techniques, we refer to [2, 1, 5, 11]. Such higher-order-only methods have the interesting feature that they allow to boost signal-to-noise ratios when the noise is Gaussian. Although there is no prewhitening stage, one may still want to reduce the dimensionality of the higher-order cumulant in the more-sensors-than-sources case, as this may significantly reduce the computational efforts in the actual algorithm. In this paper, we will investigate how the dimensionality reduction can be achieved, from a multilinear algebraic point of view.

In Section 2, we will introduce some necessary background material on multilinear algebra. In Section 3 the actual core of our contribution is addressed: where the problem of dimensionality reduction in PCA can essentially be formulated as the best rank-R approximation of a given matrix, the higher-order-only problem leads to a clear multilinear equivalent, denoted as the best rank-$(R, R, \ldots, R)$ approximation (in least-squares sense) of the given higher-order cumulant tensor. In Section 4 we discuss if and how the approximation problem can be solved by a tensorial generalization of truncating the Eigenvalue Decomposition (EVD) or Singular Value Decomposition (SVD). Section 5 deals with the local minimisation of the quadratic error criterion by an Alternating Least Squares (ALS) algorithm that can be interpreted as a higher-order generalization of the technique of orthogonal iterations [12]. Quite some research on related topics has been done in the field of Psychometrics; main ideas behind Sections 4 and 5 can also be found in [13]. In Section 6 we give an explicit expression for the gradient of the objective function, for use in gradient-based optimization routines. Section 7 summarizes the conclusions.

Section 5 forms a generalisation of a previous paper [9], on the best rank-1 approximation of a given tensor.

For simplicity of notation, our discussion is restricted to the case of real-valued data. The results are readily generalized for complex tensors.

2. Basic definitions and notations

Multilinear algebra is the algebra of higher-order tensors. For the purpose of this paper, higher-order tensors can intuitively be imagined as multi-indexed arrays of numerical values. The tensors of interest in HOS (higher-order moments, cumulants, ...) are super-symmetric: in the real-valued case, this means that they remain unchanged under arbitrary index permutations. We subsequently touch on (a) some rank-related definitions, (b) the multiplication of a tensor by a matrix, (c) a standardized representation in matrix format, and (d) some geometry-related concepts, such as scalar product, orthogonality and norm.

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

Definition 2 (Rank) *The rank of an arbitrary Nth-order tensor $\mathcal{A}$, denoted by $r = \mathtt{rank}(\mathcal{A})$, is the minimal number of rank-1 tensors that yield $\mathcal{A}$ in a linear combination.*

Definition 3 (n-rank) *The 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 definition of n-rank generalizes the definitions of column and row rank of matrices. Def. 2 gives an other way to extend the concept of rank to higher-order tensors. In contrast to the matrix case, the different n-ranks of a higher-order tensor are not necessarily the same; and when they are equal (e.g. as a result of super-symmetry), they can still be different from the rank of the tensor. An Nth order tensor with n-ranks r_1, r_2, ... , r_N, will be denoted as a rank-$(r_1, r_2, \ldots, r_N)$ tensor. An interesting discussion of rank-related issues for super-symmetric tensors, with bibliographic pointers to the early literature, can be found in [5].

The product of two matrices can be generalized to the product of a tensor and a matrix as follows:

Definition 4 *The n-mode 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 $(I_1 \times I_2 \times \ldots J_n \ldots \times I_N)$-tensor given 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 n-mode product allows us to express the effect of basis transformations on a given tensor. As an example, for the matrices $\mathbf{F} \in \mathbb{R}^{I_1 \times I_2}$, $\mathbf{U} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{V} \in \mathbb{R}^{J_2 \times I_2}$ the matrix product $\mathbf{U} \cdot \mathbf{F} \cdot \mathbf{V}^T$ can be written as the "symmetric" expression $\mathbf{F} \times_1 \mathbf{U} \times_2 \mathbf{V}$.

It is often more familiar to formulate tensor expressions in a matrix language. We define the following standard matrix representation of a higher-order tensor:

Definition 5 *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 $(i_1 - 1) I_2 I_3 \ldots I_{n-1} I_{n+1} \ldots I_N + (i_2 - 1) I_3 \ldots I_{n-1} I_{n+1} \ldots I_N + \ldots + i_N$.*

Remark that the n-mode product $\mathcal{B} = \mathcal{A} \times_n \mathbf{U}$ can be translated as $\mathbf{B}_{(n)} = \mathbf{U} \cdot \mathbf{A}_{(n)}$, and that the rank of $\mathbf{A}_{(n)}$ corresponds to the n-rank of $\mathcal{A}$, even in a numerical sense. For super-symmetric tensors, the various matrix unfoldings are equal, and will shortly be denoted as $\mathbf{A}$.

The definitions of scalar product, orthogonality and Frobenius-norm can be generalized in a trivial way:

Definition 6 *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*

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

Definition 7 *Tensors of which the scalar product equals 0, are mutually orthogonal.*

Definition 8 *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.

3. Dimensionality reduction in ICA with and without prewhitening

Let us first remind some more formal aspects of the procedure of prewhitening. The ICA-model implies that the covariance $\mathbf{C}_2^Y$ of Y is given by

$$\mathbf{C}_2^Y = \mathbf{C}_2^X \times_1 \mathbf{M} \times_2 \mathbf{M} + \mathbf{C}_2^N = \mathbf{M}\mathbf{C}_2^X\mathbf{M}^T + \mathbf{C}_2^N \quad (5)$$

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 can additionally assume that the source signals have unit variance, which at most involves an irrelevant rescaling of the columns of $\mathbf{M}$. As a consequence, a first estimate of $\mathbf{M}$, up to an orthogonal factor $\mathbf{V}$, can be obtained in the form of a square root $\mathbf{F}$ of $\mathbf{C}_2^Y$:

$$\mathbf{C}_2^Y = \mathbf{F}\mathbf{F}^T = (\mathbf{F}\mathbf{V}^T)(\mathbf{F}\mathbf{V}^T)^T \simeq \mathbf{M}\mathbf{M}^T \quad (6)$$

Substitution of the Singular Value Decomposition (SVD) $\mathbf{M} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$ of the mixing matrix, leads to a classical PCA

$$\mathbf{C}_2^Y = \mathbf{U}\mathbf{S}^2\mathbf{U}^T + \mathbf{C}_2^N = (\mathbf{U}\mathbf{S})(\mathbf{U}\mathbf{S})^T + \mathbf{C}_2^N \quad (7)$$

in which the left singular matrix $\mathbf{U}$ and the matrix of singular values $\mathbf{S}$ are estimated with an EVD of the observed covariance (or, numerically preferable, the direct SVD of the dataset), while the right singular matrix $\mathbf{V}$ remains unknown. Let us denote the dimension of Y by I and the number of sources by R. If $I > R$, then the dominant R-dimensional eigenspace of $\mathbf{C}_2^Y$ is retained. If the noise is known to be spatially white, then its variance σ_N^2 can be estimated as the mean of the smallest $I - R$ eigenvalues of $\mathbf{C}_2^Y$.

On the other hand, the fourth-order cumulant $\mathcal{C}_4^Y$ of Y can be expressed in a similar way as the covariance in Eq. (5):

$$\mathcal{C}_4^Y = \mathcal{C}_4^X \times_1 \mathbf{M} \times_2 \mathbf{M} \times_3 \mathbf{M} \times_4 \mathbf{M} + \mathcal{C}_4^N \quad (8)$$

In this equation, $\mathcal{C}_4^X$ and $\mathcal{C}_4^N$ symbolize the fourth-order cumulant tensors of X resp. N; the latter vanishes if the noise is Gaussian. Like $\mathbf{C}_2^X$, $\mathcal{C}_4^X$ is diagonal, as the source signals are statistically independent. Due to the diagonality property, Eq. (8) can be rewritten as

$$\mathcal{C}_4^Y \simeq \sum_{r=1}^{R} \kappa_r^X M_r \circ M_r \circ M_r \circ M_r \quad (9)$$

in which κ_r^X represents the marginal cumulant of the rth source (we assume that the sources are kurtic), and in which M_r denotes the rth column of $\mathbf{M}$. Eq. (9) is a decomposition of $\mathcal{C}_4^Y$ in a minimal number of rank-1 terms, as the columns of $\mathbf{M}$ are assumed to be linearly independent; it can even be proved that the decomposition is essentially unique, up to some trivial indeterminacies [14]. As a consequence, the aim of higher-order-only ICA can be formulated as the computation of a rank-revealing decomposition of $\mathcal{C}_4^Y$, taking into account that the sample cumulant equivalent of Eq. (9) may be perturbated by non-Gaussian noise components, finite datalength effects, model misfit, etc.

In addition, according to the definition of n-rank, every tensor satisfying the assumptions of Eq. (9) is a rank-(R, R, R, R) tensor. To deal with the situation in which $I > R$, we propose to perform an orthogonal projection of the sample cumulant $\hat{\mathcal{C}}_4^Y$ on the manifold of super-symmetric rank-(R, R, R, R) tensors, before addressing the harder problem of the further projection on the submanifold of rank-R tensors and the actual computation of decomposition (9). Notice that in the second-order case both projections coincide, as n-rank and rank are necessarily the same, which corresponds to the single best rank-R approximation problem, associated with Eq. (7).

Formally, the problem we want to solve, can be formulated as follows:
Given a sample cumulant $\hat{\mathcal{A}} \in \mathbb{R}^{I\times I\times I\times I}$, find a super-symmetric tensor $\tilde{\mathcal{A}} \in \mathbb{R}^{I\times I\times I\times I}$, with $r_1 = r_2 = r_3 = r_4 = R$, that minimizes the least-squares criterion

$$f(\tilde{\mathcal{A}}) = \|\hat{\mathcal{A}} - \tilde{\mathcal{A}}\|^2 \quad (10)$$

The conditions imply that $\tilde{\mathcal{A}}$ can be decomposed as:

$$\tilde{\mathcal{A}} = \mathcal{B} \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \mathbf{U} \times_4 \mathbf{U} \quad (11)$$

in which $\mathbf{U} \in \mathbb{R}^{I\times R}$ has orthonormal columns and $\mathcal{B} \in \mathbb{R}^{R\times R\times R\times R}$ is super-symmetric. After computation of the best rank-(R, R, R, R) approximation, the actual ICA-algorithm can be applied to $\mathcal{B}$ instead of $\hat{\mathcal{A}}$; $\mathcal{B}$ can be interpreted as the cumulant tensor of the R-dimensional stochastic vector $Z = \mathbf{U}^T Y$.

4. Best rank-(R, R, R, R) approximation and Higher-Order Eigenvalue Decomposition

In this section we discuss the tensorial equivalent of the computation of the best rank-R approximation of a symmetric matrix by truncation of its symmetric EVD. First, the decomposition itself can be generalized as follows:

Theorem 1 *Every super-symmetric Nth order tensor $\mathcal{A} \in \mathbb{R}^{I\times I\times \ldots \times I}$ can be written as*

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U} \times_2 \mathbf{U} \ldots \times_N \mathbf{U} \quad (12)$$

in which:

- $\mathbf{U} = [U_1 U_2 \ldots U_I]$ *is an orthogonal $(I \times I)$-matrix*
- *the core tensor $\mathcal{S}$ is a super-symmetric $(I \times I \times \ldots I)$-tensor of which the subtensors $\mathcal{S}_{i_n=\alpha}$, obtained by fixing one arbitrary 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

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

The Frobenius-norms $\|\mathcal{S}_{i_n=i}\|$, symbolized by s_i, are the higher-order eigenvalues of $\mathcal{A}$ and the vector U_i is the ith higher-order eigenvector.

This Higher-Order Eigenvalue Decomposition (HOEVD) is the super-symmetric case of the Higher-Order Singular Value Decomposition (HOSVD), discussed in [8]. Clearly, the HOEVD is a formal generalization of the EVD of symmetric matrices. Moreover, it can be proved that the HOEVD of a second-order symmetric tensor essentially boils down to its matrix EVD.

The matrix $\mathbf{U}$ can be computed as the left singular matrix of the matrix unfolding $\mathbf{A}$. The core tensor $\mathcal{S}$ can then be found from the following form of Eq. (12):

$$\mathcal{S} = \mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \ldots \times_N \mathbf{U}^T \tag{16}$$

The higher-order eigenvalues correspond to the singular values of $\mathbf{A}$; as such, they give numerical information on the n-rank of $\mathcal{A}$. If Eq. (9) is exactly satisfied, then the number of non-zero higher-order eigenvalues corresponds to the number of kurtic sources. In practice, the number of sources will be estimated by determining the hypothesized gap between "signal" and "noise" higher-order eigenvalues.

The properties of the higher-order eigenvectors and eigenvalues, together with the ordering convention, suggest that setting $s_{R+1} = s_{R+2} = s_N = 0$, will result in a good rank-(R, R, R, R) approximation of $\mathcal{A}$. Remarkably enough, it turns out that this approximation is in general not the globally optimal one. A second important difference with matrices, is that the criterion function $f(\hat{\mathcal{A}})$ can exhibit spurious local optima. On the other hand, we have observed that the HOEVD-approximation generally belongs to the valley of f that contains the global optimum. There is not an absolute guarantee: we have been able to generate ill-conditioned cases in which gradient-descent, starting from the HOEVD-truncate, eventually leads to a local optimum with a close to optimal fit, but nevertheless different from the global optimum [6]. In practice however, no problems of this kind have been observed.

5. Higher-order orthogonal iterations

In this section we briefly discuss an ALS algorithm for the local optimisation of f, that can be interpreted as a higher-order generalization of the method of orthogonal iterations [12].

First, observe that the minimisation of f is equivalent to the maximisation of the following function $g(\mathbf{U})$ over the matrices with mutually orthonormal columns:

$$g(\mathbf{U}) = \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 \mathbf{U}^T \times_4 \mathbf{U}^T\|^2 \tag{17}$$

Proving this relationship falls outside the scope of this article, but as a motivation we can remark that, similarly, the computation of the best rank-R approximation of a symmetric matrix $\mathbf{H}$ is equivalent to the determination of a column-wise orthonormal matrix $\mathbf{U}$ that maximizes the Frobenius-norm of $\mathbf{U} \cdot \mathbf{H} \cdot \mathbf{U}^T$.

Next, let us embed the maximisation of g in the maximisation, over the matrices $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}, \mathbf{U}^{(4)}$ with orthonormal columns, of the function $\tilde{g}$, defined as:

$$\tilde{g}(\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(4)}) = \|\mathcal{A} \times_1 {\mathbf{U}^{(1)}}^T \ldots \times_4 {\mathbf{U}^{(4)}}^T\|^2 \tag{18}$$

Because of the super-symmetry of $\mathcal{A}$, the global optimum will satisfy $\mathbf{U}^{(1)} = \ldots = \mathbf{U}^{(4)}$.

Now imagine that e.g. the matrices $\mathbf{U}^{(2)}$, $\mathbf{U}^{(3)}$ and $\mathbf{U}^{(4)}$ are fixed and that $\tilde{g}$ in Eq. (18) is merely a quadratic expression in the components of the unknown matrix $\mathbf{U}^{(1)}$. In matrix notation, we have:

$$\tilde{g} = \|{\mathbf{U}^{(1)}}^T \cdot [\mathbf{A}_{(1)} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)} \otimes \mathbf{U}^{(4)})]\|^2 \tag{19}$$

in which $\otimes$ denotes the Kronecker product. Hence the columns of $\mathbf{U}^{(1)}$ can be found as an orthonormal basis for the dominant subspace of the column space of the matrix between square brackets. Repeating this procedure for different mode numbers leads to an ALS algorithm for the (local) maximization of $f(\tilde{\mathcal{A}})$: in each step the estimate of one of the matrices $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(4)}$ is optimized, while the other matrix estimates are kept constant.

Clearly this technique is a higher-order extension of the orthogonal iteration method for matrices. A major difference is that each iteration step does not involve a linear but a multilinear transformation. A second important difference is that an iteration step also implies the estimation of a dominant subspace, instead of a mere orthonormalisation. It is inherent to the ALS-mechanism that higher-order orthogonal iterations yield unsymmetric intermediate results.

6. Gradient-based optimisation

As an alternative for ALS-iterations, one could consider the use of more general-purpose optimisation routines, in which the unknown matrix $\mathbf{U}$ is situated on the Stiefel manifold of column-wise orthonormal $(I \times R)$-matrices. The symmetry can be retained during iteration by simply treating all the modes in the same way. For the optimisation procedure it is possible to exploit knowledge of the gradient. More specifically, the gradient of g over the Stiefel manifold takes the following form:

$$\nabla_{\mathbf{U}} g = 8\,(\mathbf{I} - \mathbf{U}\mathbf{U}^T) \cdot \tilde{\mathbf{U}} \cdot \tilde{\mathbf{U}}^T \cdot \mathbf{U} \qquad (20)$$

in which

$$\tilde{\mathbf{U}} \stackrel{\text{def}}{=} \mathbf{A} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)} \otimes \mathbf{U}^{(4)}) \qquad (21)$$

Expression (20) consists of a sum of 4 equal terms, as Eq. (17) involves the matrix $\mathbf{U}$ 4 times, in a super-symmetric way. For each of the terms, the symmetric version of Eq. (19) can be interpreted as a classical subspace tracking problem, with fixed matrix $\tilde{\mathbf{U}}$.

7. Conclusions

In an ICA set-up with more observation channels than source components, the number of sources can conceptually be estimated without resorting to second-order statistics. The corresponding reduction of the dimensionality of the problem takes the form of a multilinear generalization of the best rank-R approximation of matrices. A first approximative solution can be obtained by truncation of a higher-order EVD/SVD-type decomposition, in much the same way as for matrices. The determination of the optimal result generally requires some additional fine-tuning. With this respect, we have paid some attention to an ALS-type descent technique and to gradient-based optimisation over the Stiefel manifold.

Acknowledgements

This work has partly been supported by the Flemish Government (Concerted Research Action GOA-MIPS, the F.W.O. (Fund for Scientific Research - Flanders) project G.0292.95, the F.W.O. project G.0256.97, the F.W.O. Research Communities ICCoS and Advanced Numerical Methods for Mathematical Modelling), the Belgian State (Interuniversity Poles of Attraction IUAP P4-02 and IUAP P4-24) and the European Commission (Human Capital and Mobility Network SIMONET and SCIENCE-ERNSI SC1-CT92-0779). Bart De Moor is a senior Research Associate with the F.W.O. Lieven De Lathauwer is a Research Assistant with the I.W.T. (Flemish Institute for Science and Technology in Industry). The scientific responsibility is assumed by the authors.

References

[1] J.-F. Cardoso. Super-symmetric decomposition of the fourth-order cumulant tensor; blind identification of more sources than sensors. In *Proc. ICASSP-91*, pages 3109–3112, 1991.

[2] J.-F. Cardoso. Iterative techniques for blind source separation using only fourth-order cumulants. In *Proc. EUSIPCO-92*, pages 739–742, Brussels, Belgium, August 1992.

[3] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1994.

[4] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[5] P. Comon and B. Mourrain. Decomposition of quantics in sums of powers of linear forms. *Signal Processing*, 53(2-3):93–108, September 1996.

[6] L. De Lathauwer. *Signal Processing by Multilinear Algebra*. PhD thesis, K.U.Leuven, E.E. Dept., October 1997.

[7] L. De Lathauwer, B. De Moor, and J. Vandewalle. Blind source separation by higher-order singular value decomposition. In *Proc. EUSIPCO-94*, pages 175–178, Edinburgh, Scotland, U.K., September 1994.

[8] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. ESAT/SISTA report 94-31, K.U.Leuven, December 1994. Submitted to: *SIAM J. Mat. Anal. Appl.*

[9] L. De Lathauwer, B. De Moor, and J. Vandewalle. Higher-order power method - application in independent component analysis. In *Proc. NOLTA'95*, pages 91–96, Vegas, USA, December 1995.

[10] L. De Lathauwer, B. De Moor, and J. Vandewalle. Blind source separation by simultaneous third-order tensor diagonalization. In *Proc. EUSIPCO-96*, pages 2089–2092, Trieste, Italy, September 1996.

[11] L. De Lathauwer, B. De Moor, and J. Vandewalle. Independent component analysis based on higher-order statistics only. In *Proc. SSAP-96*, pages 356–359, Corfu, Greece, June 1996.

[12] G. Golub and C. Van Loan. *Matrix computations*. Johns Hopkins University Press, second edition, 1991.

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

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