Dimensionality Reduction in ICA and Rank-$(R_1, R_2, \ldots, R_N)$ Reduction in Multilinear Algebra*

Lieven De Lathauwer[1,2] and Joos Vandewalle[2]

[1] ETIS (CNRS, ENSEA, UCP), UMR 8051, Cergy-Pontoise, France
delathau@ensea.fr
[2] E.E. Dept. (ESAT) - SCD, K.U.Leuven, Leuven, Belgium
vdwalle@esat.kuleuven.ac.be

Abstract. We show that the best rank-$(R_1, R_2, \ldots, R_N)$ approximation in multilinear algebra is a powerful tool for dimensionality reduction in ICA schemes without prewhitening. We consider the application to different classes of ICA algorithms.

1 Introduction

In this paper, the ICA-model is denoted as follows:

$$X = \mathbf{M}S + E, \tag{1}$$

in which the observations $X \in \mathbb{R}^I$, the sources $S \in \mathbb{R}^R$ and the noise $E \in \mathbb{R}^I$ are zero-mean random vectors. The components of S are assumed to be mutually statistically independent, as well as statistically independent from the noise components. The mixing matrix $\mathbf{M}$ is full column rank. The goal of ICA consists of the estimation of the mixing matrix and/or the corresponding realizations of S, given only realizations of X. For notational convenience we only consider real-valued signals. The generalization to complex signals is straightforward.

Many ICA applications involve high-dimensional data in which however only few sources have significant contributions. Examples are electro-encephalography (EEG), magneto-encephalography (MEG), nuclear magnetic resonance (NMR), hyper-spectral image processing, data analysis, etc. To reduce the computational complexity and to decrease the variance of the results [1], one may wish to reduce the dimensionality of the problem from the number of observation channels, I, to the number of sources, R. In this paper we will discuss algebraic means to perform the dimensionality reduction, other than a classical prewhitening. The

* L. De Lathauwer holds a permanent research position with the French C.N.R.S.; he also holds a honorary position with the K.U.Leuven. J. Vandewalle is a Full Professor with the K.U.Leuven. Part of this research was supported by the Research Council K.U.Leuven (GOA-MEFISTO-666), the Flemish Government (F.W.O. project G.0240.99, F.W.O. Research Communities ICCoS and ANMMM, Tournesol project T2004.13) and the Belgian Federal Government (IUAP V-22).

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

Algebraic prerequisites are introduced in Sect. 2. The application to dimensionality reduction in ICA forms the subject of Sect. 3. Section 4 illustrates the approach by means of two simulations. Section 5 is the conclusion. The algebraic tools are extensively described in [7, 8, 14, 18]. The application to ICA is discussed in [9].

Notation. Scalars are denoted by lower-case letters (a, b, ...), vectors by capitals (A, B, ...), matrices by bold-face capitals ($\mathbf{A}$, $\mathbf{B}$, ...) and tensors by calligraphic letters ($\mathcal{A}, \mathcal{B}, \ldots$). In this way, the entry with row index i and column index j in a matrix $\mathbf{A}$, i.e., $(\mathbf{A})_{ij}$, is symbolized by a_{ij}. However, I, J, N, R denote the upper bounds of indices i, j, n, r. A_j stands for the jth column of $\mathbf{A}$. $\mathrm{St}(R, I)$ is standard notation for the Stiefel manifold of column-wise orthonormal $(I \times R)$ matrices.

2 Basics of Multilinear Algebra

2.1 Elementary Definitions

Consider a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \ldots \times I_N}$ and matrices $\mathbf{U}^{(n)} \in \mathbb{R}^{J_n \times I_n}$, $1 \leqslant n \leqslant N$. Then $\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \ldots \times_N \mathbf{U}^{(N)}$ is a $(J_1 \times J_2 \ldots \times J_N)$-tensor of which the entries are given by [18]

$$b_{j_1 j_2 \ldots j_N} = \sum_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N} u^{(1)}_{j_1 i_1} u^{(2)}_{j_2 i_2} \ldots u^{(N)}_{j_N i_N}.$$

In this notation we have $\mathbf{B} = \mathbf{U}^{(1)} \cdot \mathbf{A} \cdot \mathbf{U}^{(2)^T} = \mathbf{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}$. The *Frobenius-norm* of $\mathcal{A}$ is defined as $\|\mathcal{A}\| = (\sum_{i_1 \ldots i_N} a^2_{i_1 \ldots i_N})^{1/2}$. There are major differences between matrices and tensors when rank properties are concerned. A *rank-1 tensor* is a tensor that consists of the outer product of a number of vectors. For an Nth-order tensor $\mathcal{A}$ and N vectors $U^{(1)}$, $U^{(2)}$, ..., $U^{(N)}$, this means that $a_{i_1 i_2 \ldots i_N} = u^{(1)}_{i_1} u^{(2)}_{i_2} \ldots u^{(N)}_{i_N}$ for all index values, which will be concisely written as $\mathcal{A} = U^{(1)} \circ U^{(2)} \circ \ldots \circ U^{(N)}$. An *$n$-mode vector* of $\mathcal{A}$ is an I_n-dimensional vector obtained from $\mathcal{A}$ by varying the index i_n and keeping the other indices fixed. In the matrix case, 1-mode vectors are columns and 2-mode vectors are rows. The *n-rank* of a tensor is the obvious generalization of the column (row) rank of matrices: it is defined as the dimension of the vector space spanned by the n-mode vectors. If all the n-mode vectors of a tensor $\mathcal{A}$ are stacked in a matrix $\mathbf{A}_{(n)}$, then the n-rank of $\mathcal{A}$ is equal to the rank of $\mathbf{A}_{(n)}$. In contrast to the matrix case, the different n-ranks of a higher-order tensor are not necessarily the same. A tensor of which the n-ranks are equal to R_n $(1 \leqslant n \leqslant N)$ is called a *rank-$(R_1, R_2, \ldots, R_N)$ tensor*. A *rank-R tensor* is defined in yet an other way: it is a tensor that can be decomposed in a sum of R, but not less than R, rank-1 terms.

2.2 Best Rank-$(R_1, R_2, \ldots, R_N)$ Approximation

We consider the minimization of the least-squares cost function

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

under the constraint that $\hat{\mathcal{A}}$ is rank-$(R_1, R_2, \ldots, R_N)$ [8, 14]. The n-rank conditions imply that $\hat{\mathcal{A}}$ can be decomposed as

$$\hat{\mathcal{A}} = \mathcal{B} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \ldots \times_N \mathbf{U}^{(N)}, \tag{3}$$

in which $\mathbf{U}^{(n)} \in \mathrm{St}(R_n, I_n)$, $1 \leqslant n \leqslant N$, and $\mathcal{B} \in \mathbb{R}^{R_1 \times R_2 \times \ldots \times R_N}$.

Similarly to the second-order case, where the best approximation of a given matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ by a matrix $\hat{\mathbf{A}} = \mathbf{U}^{(1)} \cdot \mathbf{B} \cdot \mathbf{U}^{(2)^T}$, with $\mathbf{U}^{(1)} \in \mathrm{St}(R, I_1)$ and $\mathbf{U}^{(2)} \in \mathrm{St}(R, I_2)$, is equivalent to the maximization of $\|\mathbf{U}^{(1)^T} \cdot \mathbf{A} \cdot \mathbf{U}^{(2)}\|$, the minimization of f is equivalent to the maximization of

$$g(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}) = \|\mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \ldots \times_N \mathbf{U}^{(N)^T}\|^2. \tag{4}$$

The optimal tensor $\mathcal{B}$ follows from

$$\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)^T} \times_2 \mathbf{U}^{(2)^T} \ldots \times_N \mathbf{U}^{(N)^T}. \tag{5}$$

It is natural to ask whether, in analogy with the matrix case, the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of a higher-order tensor can be obtained by truncation of a multilinear generalization of the SVD. The situation turns out to be quite different for tensors. Truncation of the Higher-Order Singular Value Decomposition (HOSVD) discussed in [7, 18] generally yields a good but not the optimal approximation. The latter has to be computed by means of tensor generalizations of algorithms for the computation of the dominant subspace of a given matrix. In [8, 14] a higher-order generalization of the orthogonal iteration method is discussed. In [12] we present a higher-order Grassmann-Rayleigh quotient iteration. These algorithms can be initialized by means of the truncated HOSVD components. This means that the columns of a first estimate of $\mathbf{U}^{(n)}$ are determined as an orthonormal basis for the R_n-dimensional dominant subspace of the column space of $\mathbf{A}_{(n)}$, defined in Sect. 2.1 ($1 \leqslant n \leqslant N$). The subsequent optimization can be quite efficient too: in the higher-order Grassmann-Rayleigh quotient iteration, an iteration step consists of solving a square set of linear equations and the convergence is quadratic.

3 Dimensionality Reduction in ICA

3.1 Higher-Order-Only ICA

Many ICA algorithms are prewhitening-based. In the prewhitening step the covariance matrix $\mathbf{C}_X$ of the data is diagonalized. This step has a three-fold goal: (a) determination of the number of source signals R, (b) standardization of the

sources to mutually uncorrelated unit-variance signals, and (c) reduction of the $(I \times R)$ mixing matrix to an unknown $(R \times R)$ orthogonal matrix. In a second step the remaining unknown orthogonal factor is determined from the other statistics of the data. This approach has the disadvantage that the results are affected by additive (coloured) Gaussian noise. Errors made in the prewhitening step cannot be compensated in the second step and yield a bound on the overall performance [4, 11]. However, if the sources are non-Gaussian it is possible to identify the mixing matrix by using only higher-order statistics and not the covariance matrix of the observations. Higher-order-only methods (see [10] and the references therein) have the interesting feature that they allow to boost Signal to Noise Ratios (SNR) when the noise is Gaussian.

Let $\mathcal{C}_X^{(N)}$ denote the Nth order cumulant of the observations and κ_r^S the marginal Nth order cumulant of the rth source. We suppose that all values κ_r^S are different from zero. In the absence of noise we have

$$\mathcal{C}_X^{(N)} = \sum_{r=1}^{R} \kappa_r^S \, M_r \circ M_r \circ \ldots \circ M_r. \tag{6}$$

This is a decomposition of $\mathcal{C}_X^{(N)}$ in a minimal number of rank-1 terms, as the columns of $\mathbf{M}$ are assumed to be linearly independent. As a consequence, the aim of higher-order-only ICA can be formulated as the computation of a rank-revealing decomposition of $\mathcal{C}_X^{(N)}$, taking into account that the sample cumulant equivalent of Eq. (6) may be perturbated by non-Gaussian noise components, finite datalength effects, model misfit, etc.

Tensor $\mathcal{C}_X^{(N)}$ is not only rank-R but also rank-$(R, R, \ldots, R)$. The reason is that every n-mode vector can be written as a linear combination of the R vectors M_r. So, to deal with the situation in which $I > R$, we can first project the sample cumulant on the manifold of rank-$(R, R, \ldots, R)$ tensors, using the techniques mentioned in Sect. 2.2. A subsequent step consists of the further projection on the submanifold of rank-R tensors and the actual computation of decomposition (6). The latter problem can then be solved in a lower-dimensional space.

3.2 ICA Based on Soft Whitening

In prewhitening-based ICA, on one hand, more than half of the parameters of the unknown mixing matrix are calculated from an exact decomposition of the covariance matrix $\mathbf{C}_X$. On the other hand, the complete set of other statistics is used for the estimation of less than half of the unknown parameters; here decompositions are only approximately satisfied. The ratio of the former $(R(2I - R + 1)/2)$ over the latter $(R(R-1)/2)$ number of parameters becomes bigger as I/R increases. Contrarily, in higher-order-only schemes the matrix $\mathbf{C}_X$ is not used at all. However, one may also deal with the different statistics in a more balanced way. We call this principle "soft whitening". The idea was first proposed and tested in [19]. In this section we will discuss dimensionality reduction in the context of soft whitening.

Assume that one wants to use the sample estimates of the covariance matrix and the Nth order cumulant tensor, $\hat{\mathbf{C}}_X$ and $\hat{\mathcal{C}}_X^{(N)}$. The ICA problem now amounts to the determination of a matrix $\mathbf{M} \in \mathbb{R}^{I \times R}$ that minimizes the cost function

$$\tilde{f}(\mathbf{M}) = w_1^2 \, \|\hat{\mathbf{C}}_X - \hat{\mathbf{C}}_S \times_1 \mathbf{M} \times_2 \mathbf{M}\|^2 + w_2^2 \, \|\hat{\mathcal{C}}_X^{(N)} - \hat{\mathcal{C}}_S^{(N)} \times_1 \mathbf{M} \times_2 \mathbf{M} \ldots \times_n \mathbf{M}\|^2, \quad (7)$$

in which $\hat{\mathbf{C}}_S \in \mathbb{R}^{R \times R}$ is an unknown diagonal matrix and $\hat{\mathcal{C}}_S^{(N)} \in \mathbb{R}^{R \times R \times \ldots \times R}$ an unknown diagonal tensor; w_1 and w_2 are positive weights.

Both the column space of $w_1 \, \hat{\mathbf{C}}_X$ and the 1-mode vector space of $w_2 \, \hat{\mathcal{C}}_X^{(N)}$ are theoretically equal to the column space of $\mathbf{M}$. Hence, in comparison with Sect. 2.2, it is natural to replace the truncation of the HOSVD by the computation of the dominant subspace of the column space of a matrix in which all the columns of $w_1 \, \hat{\mathbf{C}}_X$ and the 1-mode vectors of $w_2 \, (\hat{\mathbf{C}}_X^{(N)})$ are stacked. The higher-order orthogonal iteration and the higher-order Grassmann-Rayleigh quotient iteration can be easily adapted as well [9, 12].

3.3 ICA Based on Simultaneous Matrix Diagonalization

Many ICA algorithms are based on diagonalization of a set of matrices by means of a simultaneous congruence transformation [2, 5, 16]. Given $\mathbf{A}_1, \ldots, \mathbf{A}_J \in \mathbb{R}^{I \times I}$, the aim is to find a nonsingular matrix $\mathbf{M} \in \mathbb{R}^{I \times R}$ such that, in theory,

$$\begin{aligned} \mathbf{A}_1 &= \mathbf{M} \cdot \mathbf{D}_1 \cdot \mathbf{M}^T \\ &\vdots \\ \mathbf{A}_J &= \mathbf{M} \cdot \mathbf{D}_J \cdot \mathbf{M}^T, \end{aligned} \quad (8)$$

with $\mathbf{D}_1, \ldots, \mathbf{D}_J \in \mathbb{R}^{R \times R}$ diagonal. In the presence of noise, the difference between the left- and right-hand side of Eqs. (8) has to be minimized.

The original algorithms to solve (8) are prewhitening-based. In the prewhitening step, one picks a positive (semi-)definite matrix from $\{\mathbf{A}_j\}$, say $\mathbf{A}_1$, and computes its EVD. This allows to reduce the other equations to a simultaneous orthogonal diagonalization in a possibly lower-dimensional space. On the other hand, one can also follow the soft whitening approach and solve the different equations in (8) in a more balanced way. In this section we will explain how a dimensionality reduction can be realized here, when $R < I$. For the matrix diagonalization in the lower-dimensional space, one can resort to the techniques presented and referred to in [10, 17, 19].

To see the link with Subsect. 3.1, let us stack the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_J$ in Eq. (8) in a tensor $\mathcal{A} \in \mathbb{R}^{I \times I \times J}$. Define a matrix $\mathbf{D} \in \mathbb{R}^{J \times R}$ of which the subsequent rows are given by the diagonals of $\mathbf{D}_1, \ldots, \mathbf{D}_J$. Then we have

$$\mathcal{A} = \sum_{r=1}^{R} M_r \circ M_r \circ D_r. \quad (9)$$

Let the rank of $\mathbf{D}$ be equal to R_3. Equation (9) is a decomposition of $\mathcal{A}$ in a minimal sum of rank-1 terms (if no columns of $\mathbf{D}$ are collinear [10]). This problem and the simultaneous diagonalization problem are equivalent. Hence, we can proceed in the same way as in Sect. 3.1. The dimensionality reduction can be realized by a rank-(R, R, R_3) reduction of $\mathcal{A}$. The remaining problem is the decomposition of an $(R \times R \times R_3)$-tensor in rank-1 terms.

4 Simulation Results

Our first simulation illustrates the technique described in Subsect. 3.1. Data are generated according to the following model:

$$X = \mathbf{M}_1 S + \mathbf{M}_2 \, \sigma_{\tilde{E}} \, \tilde{E},$$

in which the entries of $S \in \mathbb{R}^4$ are ± 1, with equal probability, and in which $\tilde{E} \in \mathbb{R}^{12}$ is zero-mean unit-variance Gaussian noise. $\mathbf{M}_1 \in \mathbb{R}^{12\times 4}$ and $\mathbf{M}_2 \in \mathbb{R}^{12\times 12}$ are random matrices of which the columns have been normalized to unit length. The data length is 500. A Monte Carlo experiment consisting of 500 runs is carried out for different SNR values (controlled by $\sigma_{\tilde{E}}$). Because the observations are corrupted by noise with an unknown colour, the independent components are estimated from the fourth-order cumulant of X. First the dimensionality of the problem is reduced from 12 to 4, and subsequently both sides of (6) are matched in the least-squares sense by means of the technique described in [3]. This algorithm was initialized with the starting value proposed in [15] and with 9 random starting values. The best result was retained. The dimensionality reduction was achieved by means of the algorithm described in [8].

Let the estimate of $\mathbf{M}_1$ be represented by $\hat{\mathbf{M}}_1$ and let the columns of $\hat{\mathbf{M}}_1$ be normalized to unit length and optimally ordered. Then our error measure is defined as the mean of the squared off-diagonal entries of $\hat{\mathbf{M}}_1^{\dagger} \cdot \mathbf{M}_1$. This error measure can be interpreted as an approximate average Interference to Signal Ratio (ISR). Only the results for which the ISR is smaller than 0.04 are retained; the other results are considered as failures. A failure means that 10 initializations were not enough or that the estimate of the low-dimensional cumulant was simply too bad to get sufficiently accurate results. The results are shown in Fig. 1. The plots show that after an inexpensive dimensionality reduction, a very accurate source separation is possible for moderate SNR values.

In a second simulation we consider a simultaneous matrix diagonalization. The experiment illustrates that is sometimes possible to extract the subspace associated with a particular class of sources [6]. In this simulation, the sources of interest are the signals obtained by passing 4 mutually independent zero-mean unit-variance Gaussian i.i.d. sequences through the filters $h_1(z^{-1}) = (1 + 0.9z^{-1})^{-1}$, $h_2(z^{-1}) = (1 - 0.9z^{-1})^{-1}$, $h_3(z^{-1}) = (1 + 0.8z^{-1})^{-1}$, $h_4(z^{-1}) = (1 - 0.8z^{-1})^{-1}$. A second set of 4 independent sources is uniform over $[-\sqrt{3}, \sqrt{3}]$ and i.i.d. The rest of the set-up is the same as in the first simulation. By considering only covariance matrices for nonzero time-lag $\tau = 1, 2, 3, 4$, we are able

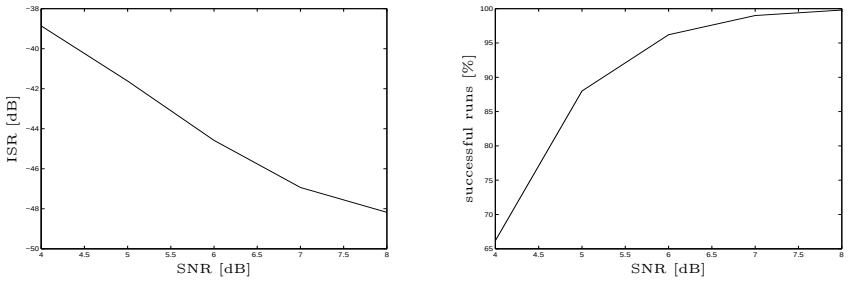


Fig. 1. Average ISR (left) and number of succesful runs (right) in the first simulation.

to estimate the subspace associated to the first class of source signals. The estimate is obtained by computing the best rank-$(4, 4, 4)$ approximation of the $(12 \times 12 \times 4)$ tensor $\mathcal{A}$ in which the covariance matrices are stacked, as explained in Subsect. 3.3. Here we used the higher-order Grassmann-Rayleigh quotient iteration algorithm. After projection, the sources of interest can be separated in a vector space of dimension 4 instead of 12.

The performance is evaluated in terms of the cosine of the largest principal angle [13] between the subspace generated by the first 4 mixing vectors and the 1-mode vector space of the best rank-$(4, 4, 4)$ approximation of $\mathcal{A}$. The results are shown in Fig. 2. The figure shows that even for low SNR the estimate of the signal subspace is very accurate. For high SNR, the accuracy is only bounded by the precision with which the covariance matrices are estimated, given the fact that only 500 snapshots are available.

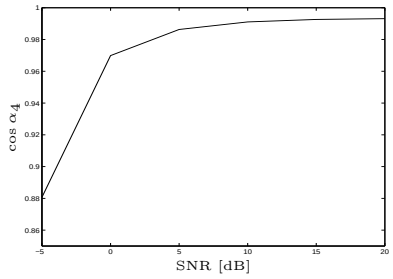


Fig. 2. Average cosine of the largest principal angle in the second simulation.

5 Conclusion

In several ICA applications the number of observation channels exceeds the number of source signals. It is not always appropriate to reduce the dimensionality

of the problem by means of a classical prewhitening, for instance because the noise has an unknown colour. In such a case, the dimensionality may be reduced by computing the best rank-$(R_1, R_2, \ldots, R_N)$ approximation of a higher-order tensor.

References

1. André, T.F., Nowak, R.D., Van Veen, B.D.: Low-rank estimation of higher order statistics. IEEE Trans. Signal Processing. **45** (1997) 673–685.
2. Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E.: A blind source separation technique using second order statistics. IEEE Trans. Signal Processing. **45** (1997) 434–444.
3. Bro, R.: PARAFAC. Tutorial & applications. Chemom. Intell. Lab. Syst. **38** (1997) 149–171.
4. Cardoso, J.-F.: On the performance of orthogonal source separation algorithms. Proc. EUSIPCO-94, Edinburgh, Scotland, U.K. 776–779.
5. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. IEE Proc.-F **140** (1994) 362–370.
6. Cruces, S., Cichocki, A., De Lathauwer, L.: Thin QR and SVD factorizations for simultaneous blind signal extraction. Proc. EUSIPCO 2004, Vienna, Austria.
7. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21** (2000) 1253–1278.
8. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. SIAM J. Matrix Anal. Appl. **21** (2000) 1324–1342.
9. De Lathauwer, L., Vandewalle, J.: Dimensionality reduction in higher-order signal processing and rank-$(R_1, R_2, \ldots, R_N)$ reduction in multilinear algebra. Lin. Alg. Appl. (to appear).
10. De Lathauwer, L., De Moor, B., Vandewalle, J.: Computation of the canonical decomposition by means of a simultaneous generalized Schur decompositition. SIAM J. Matrix Anal. Appl. (to appear).
11. De Lathauwer, L., De Moor, B., Vandewalle, J.: A prewhitening-induced bound on the identification error in independent component analysis, Tech. report.
12. De Lathauwer, L., Hoegaerts, L., Vandewalle, J.: A Grassmann-Rayleigh quotient iteration for dimensionality reduction in ICA. Proc. ICA 2004, Granada, Spain.
13. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore, Maryland (1996).
14. Kroonenberg, P.M.: Three-mode principal component analysis. DSWO Press, Leiden (1983).
15. Leurgans, S.E., Ross, R.T., Abel, R.B.: A decomposition for three-way arrays. SIAM J. Matrix Anal. Appl. **14** (1993) 1064–1083.
16. Pham, D.-T., Cardoso, J.-F.: Blind separation of instantaneous mixtures of non-stationary sources. IEEE Trans. Signal Processing. **49** (2001) 1837–1848.
17. Pham, D.-T. 2001. Joint approximate diagonalization of positive definite Hermitian matrices. SIAM J. Matrix Anal. Appl. **22** (2001) 1136–1152.
18. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika, **31** (1966) 279–311.
19. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. Signal Processing **50** (2002) 1545–1553.