A Grassmann-Rayleigh Quotient Iteration for Dimensionality Reduction in ICA*

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

[1] ETIS (CNRS, ENSEA, UCP), UMR 8051, Cergy-Pontoise, France
lieven.delathauwer@ensea.fr
[2] E.E. Dept. (ESAT) - SCD, K.U.Leuven, Leuven, Belgium
{luc.hoegaerts,joos.vandewalle}@esat.kuleuven.ac.be

Abstract. We derive a Grassmann-Rayleigh Quotient Iteration for the computation of the best rank-(R_1, R_2, R_3) approximation of higher-order tensors. We present some variants that allow for a very efficient estimation of the signal subspace in ICA schemes without prewhitening.

1 Introduction

Many ICA applications involve high-dimensional data in which however only a few sources have significant contributions. Examples are nuclear magnetic resonance (NMR), electro-encephalography (EEG), magneto-encephalography (MEG), hyper-spectral image processing, data analysis, etc. To reduce the computational complexity and to decrease the variance of the results, one may wish to reduce the dimensionality of the problem from the number of observation channels, which will be denoted by I, to the number of sources, denoted by R. If one wishes to avoid a classical prewhitening, for the reasons given in [7], then the solution can be obtained by means of a so-called best rank-(R_1, R_2, R_3) approximation of a higher-order tensor [4, 5]. (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.) Consequently, in this paper we will derive a numerical algorithm to compute this approximation. It consists of a generalization to tensors of the Rayleigh Quotient Iteration (RQI) for the computation of an invariant subspace of a given matrix [1]. It also generalizes the RQI for the best rank-1 approximation of higher-order tensors [8].

This paper primarily concerns the derivation of the numerical algorithm. Due to space limitations, the relevance of this problem in the context of ICA

* L. De Lathauwer holds a permanent research position with the French CNRS; he also holds a honorary position with the K.U.Leuven. L. Hoegaerts is a Ph.D. student supported by the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT). 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).

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 335–342, 2004.

and the link with the best rank-(R_1, R_2, R_3) approximation are discussed in the companion paper [5].

With respect to the numerical aspects of Principal Component Analysis (PCA) world-wide scientific efforts are made. This has led to powerful routines for the computation of the Eigenvalue Decomposition (EVD), Singular Value Decomposition (SVD), dominant subspaces, etc. of high-dimensional matrices. So far, no clear ICA equivalent has emerged. This paper aims to be a first step in this direction.

In Sect. 2 we introduce some basic concepts of multilinear algebra. In Sect. 3 we present our basic algorithm. The formulation is in terms of arbitrary third-order tensors because (i) this allows for the easy derivation of different variants applicable in the context of ICA (Sect. 4), and because (ii) the algorithm has important applications, apart from ICA [4]. Section 5 is the conclusion.

For notational convenience we mainly focus on real-valued third-order tensors. The generalization to complex-valued tensors and tensors of order higher than three is straightforward.

Notation. Scalars are denoted by lower-case letters (a, b, ...), vectors are written as capitals (A, B, ...), matrices correspond to bold-face capitals ($\mathbf{A}$, $\mathbf{B}$, ...) and tensors are written as calligraphic letters ($\mathcal{A}$, $\mathcal{B}$, ...). 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}. There is one exception: as we use the characters i, j and r in the meaning of indices (counters), I, J and R will be reserved to denote the index upper bounds. $\otimes$ denotes the Kronecker product. $\mathbf{I}$ is the identity matrix. $\mathbf{O}(R)$ and $\mathrm{St}(R, I)$ are standard notation for the manifold of $(R \times R)$ orthogonal matrices and the Stiefel manifold of column-wise orthonormal $(I \times R)$ matrices $(I \geqslant R)$, respectively. $\mathrm{qf}(\mathbf{X})$ denotes the orthogonal factor in a QR-decomposition of a matrix $\mathbf{X}$.

2 Basic Definitions

For a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the *matrix unfoldings* $\mathbf{A}_{(1)} \in \mathbb{R}^{I_1 \times I_3 I_2}$, $\mathbf{A}_{(2)} \in \mathbb{R}^{I_2 \times I_1 I_3}$ and $\mathbf{A}_{(3)} \in \mathbb{R}^{I_3 \times I_2 I_1}$ are defined by

$$(\mathbf{A}_{(1)})_{i_1,(i_3-1)I_3+i_2} = (\mathbf{A}_{(2)})_{i_2,(i_1-1)I_1+i_3} = (\mathbf{A}_{(3)})_{i_3,(i_2-1)I_2+i_1} = a_{i_1 i_2 i_3}$$

for all index values. Straightforward generalizations apply to tensors of order higher than three. Consider $\mathbf{U}^{(1)} \in \mathbb{R}^{J_1 \times I_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{J_2 \times I_2}$, $\mathbf{U}^{(3)} \in \mathbb{R}^{J_3 \times I_3}$. Then $\mathcal{B} = \mathcal{A} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}$ is a $(J_1 \times J_2 \times J_3)$-tensor of which the entries are given by

$$b_{j_1 j_2 j_3} = \sum_{i_1 i_2 i_3} a_{i_1 i_2 i_3} u^{(1)}_{j_1 i_1} u^{(2)}_{j_2 i_2} u^{(3)}_{j_3 i_3} \ .$$

In terms of the matrix unfoldings, we have, for instance,

$$\mathbf{B}_{(1)} = \mathbf{U}^{(1)} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)})^T .$$

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. It is a column of $\mathbf{A}_{(n)}$. 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 and is equal to the rank of $\mathbf{A}_{(n)}$. An important difference with the rank of matrices, is that 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 $(1 \leqslant n \leqslant 3)$ is called a *rank-*(R_1, R_2, R_3) *tensor.* A rank-$(1,1,1)$ tensor is briefly called a *rank-1 tensor.* Real-valued tensors are called *supersymmetric* when they are invariant under arbitrary index permutations. Finally, the *Frobenius-norm* of $\mathcal{A}$ is defined as $\|\mathcal{A}\| = (\sum_{i_1 i_2 i_3} a^2_{i_1 i_2 i_3})^{1/2}$.

Now consider the minimization of the least-squares cost function

$$f(\hat{\mathcal{A}}) = \|\mathcal{A} - \hat{\mathcal{A}}\|^2 \tag{1}$$

under the constraint that $\hat{\mathcal{A}}$ is rank-(R_1, R_2, R_3). This constraint implies that $\hat{\mathcal{A}}$ can be decomposed as

$$\hat{\mathcal{A}} = \mathcal{B} \times_1 \mathbf{X}^{(1)} \times_2 \mathbf{X}^{(2)} \times_3 \mathbf{X}^{(3)}, \tag{2}$$

in which $\mathbf{X}^{(n)} \in \mathrm{St}(R_n, I_n)$, $n = 1,2,3$, and $\mathcal{B} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$. The minimization of f can be shown [3] to be equivalent to the maximization of

$$\begin{aligned} g(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}) &= \|\mathcal{A} \times_1 \mathbf{X}^{(1)^T} \times_2 \mathbf{X}^{(2)^T} \times_3 \mathbf{X}^{(3)^T}\|^2 \\ &= \|\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})\|^2 \ . \end{aligned} \tag{3}$$

For given $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}$, the optimal $\mathcal{B}$ follows from the linear equation (2).

Now assume that $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ are fixed. From (3) we see that $\mathbf{X}^{(1)}$ can only be optimal if its columns span the same subspace as the R_1 dominant left singular vectors of $\tilde{\mathbf{A}}_{(1)} = \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})$. A necessary condition is that the column space of $\mathbf{X}^{(1)}$ is an invariant subspace of $\tilde{\mathbf{A}}_{(1)} \cdot \tilde{\mathbf{A}}^T_{(1)}$. Similar conditions can be derived for the other modes. We obtain:

$$\mathbf{X}^{(1)} \cdot \mathbf{W}_1 = \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}^T_{(1)} \cdot \mathbf{X}^{(1)} \tag{4}$$

$$\mathbf{X}^{(2)} \cdot \mathbf{W}_2 = \mathbf{A}_{(2)} \cdot (\mathbf{X}^{(3)} \otimes \mathbf{X}^{(1)}) \cdot (\mathbf{X}^{(3)} \otimes \mathbf{X}^{(1)})^T \cdot \mathbf{A}^T_{(2)} \cdot \mathbf{X}^{(2)} \tag{5}$$

$$\mathbf{X}^{(3)} \cdot \mathbf{W}_3 = \mathbf{A}_{(3)} \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(2)}) \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(2)})^T \cdot \mathbf{A}^T_{(3)} \cdot \mathbf{X}^{(3)} \tag{6}$$

for some $\mathbf{W}_1 \in \mathbb{R}^{R_1 \times R_1}$, $\mathbf{W}_2 \in \mathbb{R}^{R_2 \times R_2}$, $\mathbf{W}_3 \in \mathbb{R}^{R_3 \times R_3}$.

This set of equations forms the starting point for the derivation of our new algorithm. Note that only the column spaces of $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$ are of importance, and not their individual columns. This means that we are actually working on Grassmann manifolds [6].

3 Higher-Order Grassmann-Rayleigh Quotient Iteration

For $\mathbf{X}^{(1)} \in \mathrm{St}(R_1, I_1)$, $\mathbf{X}^{(2)} \in \mathrm{St}(R_2, I_2)$, $\mathbf{X}^{(3)} \in \mathrm{St}(R_3, I_3)$ and $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ we define *n-mode Rayleigh quotient matrices* as follows:

$$\mathbf{R}_1(\mathbf{X}) = \mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \tag{7}$$

$$\mathbf{R}_2(\mathbf{X}) = \mathbf{X}^{(2)^T} \cdot \mathbf{A}_{(2)} \cdot (\mathbf{X}^{(3)} \otimes \mathbf{X}^{(1)}) \tag{8}$$

$$\mathbf{R}_3(\mathbf{X}) = \mathbf{X}^{(3)^T} \cdot \mathbf{A}_{(3)} \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(2)}) \ . \tag{9}$$

This definition properly generalizes the existing definitions of Rayleigh quotients associated with an eigenvector, invariant subspace or tensor rank-1 approximation [1, 8]. The cornerstone of our algorithm is the following theorem.

Theorem 1. *Let $\mathbf{X}^{(1)} \in St(R_1, I_1)$, $\mathbf{X}^{(2)} \in St(R_2, I_2)$, $\mathbf{X}^{(3)} \in St(R_3, I_3)$ be solutions to (4–6). For small perturbations $\Delta\mathbf{X}^{(1)}$, $\Delta\mathbf{X}^{(2)}$, $\Delta\mathbf{X}^{(3)}$ satisfying*

$$\mathbf{X}^{(1)^T} \Delta\mathbf{X}^{(1)} = \mathbf{0}, \quad \mathbf{X}^{(2)^T} \Delta\mathbf{X}^{(2)} = \mathbf{0}, \quad \mathbf{X}^{(3)^T} \Delta\mathbf{X}^{(3)} = \mathbf{0}, \tag{10}$$

we have

$$\|\mathbf{R}_n(\mathbf{X})\mathbf{R}_n(\mathbf{X})^T - \mathbf{R}_n(\mathbf{X} + \Delta\mathbf{X})\mathbf{R}_n(\mathbf{X} + \Delta\mathbf{X})^T\| = O(\|\Delta\mathbf{X}\|^2) \quad n = 1, 2, 3 \ .$$

Proof. Let us consider the case $n = 1$. The cases $n = 2, 3$ are completely similar.

By definition, we have

$$\begin{aligned}
&\mathbf{R}_1(\mathbf{X} + \Delta\mathbf{X})\mathbf{R}_1(\mathbf{X} + \Delta\mathbf{X})^T = \\
&\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+(\Delta\mathbf{X}^{(1)})^T \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\Delta\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \Delta\mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\Delta\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \Delta\mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \\
&+\mathbf{X}^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)}) \cdot (\mathbf{X}^{(2)} \otimes \mathbf{X}^{(3)})^T \cdot \mathbf{A}_{(1)}^T \cdot \Delta\mathbf{X}^{(1)} + O(\|\Delta\mathbf{X}\|^2) \ .
\end{aligned}$$

In this expansion the first term equals $\mathbf{R}_1(\mathbf{X})\mathbf{R}_1(\mathbf{X})^T$. The first-order terms vanish, because of (4–6) and (10). This proves the theorem. □

Consider perturbations $\Delta\mathbf{X}^{(1)}$, $\Delta\mathbf{X}^{(2)}$, $\Delta\mathbf{X}^{(3)}$ satisfying (10). Using Theorem 1, saying that $\mathbf{W}_1 = \mathbf{R}_1(\mathbf{X}) \cdot \mathbf{R}_1(\mathbf{X})^T$ is only subject to second-order perturbations, we have the following linear expansion of (4):

$$\begin{aligned}
&(\mathbf{X}^{(1)} + \Delta\mathbf{X}^{(1)}) \cdot \mathbf{R}_1(\mathbf{X}) \cdot \mathbf{R}_1(\mathbf{X})^T = \\
&\mathbf{A}_{(1)} \cdot \left[(\mathbf{X}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T})\right] \cdot \mathbf{A}_{(1)}^T \cdot (\mathbf{X}^{(1)} + \Delta\mathbf{X}^{(1)}) + \\
&\mathbf{A}_{(1)} \cdot \Big[(\Delta\mathbf{X}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T}) + (\mathbf{X}^{(2)} \cdot \Delta\mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T}) + \\
&(\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}) \otimes (\Delta\mathbf{X}^{(3)}\mathbf{X}^{(3)^T}) + (\mathbf{X}^{(2)}\mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)}\Delta\mathbf{X}^{(3)^T})\Big] \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)}.
\end{aligned} \tag{11}$$

Now, let the (approximate) true solution be given by $\overline{\mathbf{X}}^{(n)} = \mathbf{X}^{(n)} + \Delta\mathbf{X}^{(n)}$, $n = 1, 2, 3$. First we will justify conditions (10). It is well-known [6] that, for $\overline{\mathbf{X}}^{(n)}$ to be on the Stiefel manifold, the perturbation can up to first order terms be decomposed as in

$$\overline{\mathbf{X}}^{(n)} = \mathbf{X}^{(n)}(\mathbf{I} + \Delta\mathbf{E}_1^{(n)}) + (\mathbf{X}^{\perp})^{(n)}\Delta\mathbf{E}_2^{(n)},$$

in which $\Delta\mathbf{E}_1^{(n)} \in \mathbb{R}^{R_n \times R_n}$ is skew-symmetric and $(\mathbf{X}^{\perp})^{(n)} \in \mathrm{St}(I_n - R_n, I_n)$ perpendicular to $\mathbf{X}^{(n)}$. As a first order approximation we have now

$$\overline{\mathbf{X}}^{(n)} \cdot (\mathbf{I} - \Delta\mathbf{E}_1^{(n)}) = \mathbf{X}^{(n)} + (\mathbf{X}^{\perp})^{(n)}\Delta\mathbf{E}_2^{(n)} \ . \tag{12}$$

Because of the skew symmetry of $\Delta\mathbf{E}_1^{(n)}$, the matrix $\overline{\mathbf{X}}^{(n)} \cdot (\mathbf{I} - \Delta\mathbf{E}_1^{(n)})$ is in first order column-wise orthonormal, and it has the same column space as $\overline{\mathbf{X}}^{(n)}$. Because only this column space is of importance (and not the individual columns), (12) implies that we can limit ourselves to perturbations satisfying (10).

From (11) we have

$$\begin{gathered}\overline{\mathbf{X}}^{(1)} \cdot \mathbf{R}_1(\mathbf{X}) \cdot \mathbf{R}_1(\mathbf{X})^T = \\ \mathbf{A}_{(1)} \cdot (\mathbf{X}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T}) \cdot \mathbf{A}_{(1)}^T \cdot (\overline{\mathbf{X}}^{(1)} - 4\mathbf{X}^{(1)}) + \\ \mathbf{A}_{(1)} \cdot \Big[(\overline{\mathbf{X}}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T}) + (\mathbf{X}^{(2)} \cdot \overline{\mathbf{X}}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \mathbf{X}^{(3)^T}) + \\ (\mathbf{X}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\overline{\mathbf{X}}^{(3)} \cdot \mathbf{X}^{(3)^T}) + (\mathbf{X}^{(2)} \cdot \mathbf{X}^{(2)^T}) \otimes (\mathbf{X}^{(3)} \cdot \overline{\mathbf{X}}^{(3)^T})\Big] \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)}\end{gathered} \tag{13}$$

Exploiting the symmetry of the problem, we obtain similar expressions for the 2-mode and 3-mode Rayleigh quotient matrices. The global set consists of linear equations in $\overline{\mathbf{X}}^{(1)}$, $\overline{\mathbf{X}}^{(2)}$, $\overline{\mathbf{X}}^{(3)}$. This means that it can be written in the form

$$\mathbf{M}_{\mathcal{A},\mathbf{X}}\overline{X} = B_{\mathcal{A},\mathbf{X}}, \tag{14}$$

in which the coefficients of $\mathbf{M}_{\mathcal{A},\mathbf{X}} \in \mathbb{R}^{(I_1R_1+I_2R_2+I_3R_3)\times(I_1R_1+I_2R_2+I_3R_3)}$ and $B_{\mathcal{A},\mathbf{X}} \in \mathbb{R}^{I_1R_1+I_2R_2+I_3R_3}$ depend on $\mathcal{A}$ and $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{X}^{(3)}$ and in which the coefficients of $\overline{\mathbf{X}}^{(1)}$, $\overline{\mathbf{X}}^{(2)}$, $\overline{\mathbf{X}}^{(3)}$ are stacked in $\overline{X}$. (Explicit expressions for $\mathbf{M}_{\mathcal{A},\mathbf{X}}$ and $B_{\mathcal{A},\mathbf{X}}$ are not given due to space limitations.) Hence, given $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{X}^{(3)}$ and the associated n-mode Rayleigh quotient matrices, $\overline{\mathbf{X}}^{(1)}$, $\overline{\mathbf{X}}^{(2)}$, $\overline{\mathbf{X}}^{(3)}$ can be estimated by solving a square linear set of equations in $I_1R_1 + I_2R_2 + I_3R_3$ unknowns.

The resulting algorithm is summarized in Table 1. The algorithm can be initialized with the truncated components of the Higher-Order Singular Value Decomposition [2]. This means that the columns of $\overline{\mathbf{X}}_0^{(n)}$ are taken equal to the dominant left singular vectors of $\mathbf{A}_{(n)}$, $n = 1, 2, 3$. See [3, 4] for more details.

The convergence of Alg. 1 is quadratic:

Table 1. GRQI for the computation of the best rank-(R_1, R_2, R_3) approximation of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$.

Given initial estimates $\overline{\mathbf{X}}_0^{(1)} \in \mathbb{R}^{I_1 \times R_1}$, $\overline{\mathbf{X}}_0^{(2)} \in \mathbb{R}^{I_2 \times R_2}$, $\overline{\mathbf{X}}_0^{(3)} \in \mathbb{R}^{I_3 \times R_3}$
Iterate until convergence:

1. Normalize to matrices on Stiefel manifold:

$$\mathbf{X}_k^{(1)} = \mathrm{qf}(\overline{\mathbf{X}}_k^{(1)}) \qquad \mathbf{X}_k^{(2)} = \mathrm{qf}(\overline{\mathbf{X}}_k^{(2)}) \qquad \mathbf{X}_k^{(3)} = \mathrm{qf}(\overline{\mathbf{X}}_k^{(3)})$$

2. Compute n-mode Rayleigh quotient matrices:

$$\begin{aligned}\mathbf{R}_1(\mathbf{X}_k) &= \mathbf{X}_k^{(1)^T} \cdot \mathbf{A}_{(1)} \cdot (\mathbf{X}_k^{(2)} \otimes \mathbf{X}_k^{(3)}) \\ \mathbf{R}_2(\mathbf{X}_k) &= \mathbf{X}_k^{(2)^T} \cdot \mathbf{A}_{(2)} \cdot (\mathbf{X}_k^{(3)} \otimes \mathbf{X}_k^{(1)}) \\ \mathbf{R}_3(\mathbf{X}_k) &= \mathbf{X}_k^{(3)^T} \cdot \mathbf{A}_{(3)} \cdot (\mathbf{X}_k^{(1)} \otimes \mathbf{X}_k^{(2)})\end{aligned}$$

3. Solve the linear set of equations

$$\mathbf{M}_{\mathcal{A},\mathbf{X}_k} \overline{X}_{k+1} = B_{\mathcal{A},\mathbf{X}_k}$$

Theorem 2. *Let* $\overline{\mathbf{X}}^{(1)}, \overline{\mathbf{X}}^{(2)}, \overline{\mathbf{X}}^{(3)}, \mathbf{R}_1(\overline{\mathbf{X}}), \mathbf{R}_2(\overline{\mathbf{X}}), \mathbf{R}_3(\overline{\mathbf{X}})$ *correspond to a nonzero solution to (4–6). If* $\mathbf{M}_{\mathcal{A},\overline{\mathbf{X}}}$ *is nonsingular, then Alg. 1 converges to* $(\overline{\mathbf{X}}^{(1)}\mathbf{Q}_1, \overline{\mathbf{X}}^{(2)}\mathbf{Q}_2, \overline{\mathbf{X}}^{(3)}\mathbf{Q}_3)$*, with* $\mathbf{Q}_1 \in \mathbf{O}(R_1)$*,* $\mathbf{Q}_2 \in \mathbf{O}(R_2)$*,* $\mathbf{Q}_3 \in \mathbf{O}(R_3)$*, quadratically in a neighbourhood of* $(\overline{\mathbf{X}}^{(1)}, \overline{\mathbf{X}}^{(2)}, \overline{\mathbf{X}}^{(3)})$*.*

Proof. Because $\overline{\mathbf{X}}^{(1)}, \overline{\mathbf{X}}^{(2)}, \overline{\mathbf{X}}^{(3)}, \mathbf{R}_1(\overline{\mathbf{X}}), \mathbf{R}_2(\overline{\mathbf{X}}), \mathbf{R}_3(\overline{\mathbf{X}})$ give a solution to (4–6), we have

$$\mathbf{M}_{\mathcal{A},\overline{\mathbf{X}}} \overline{X} - B_{\mathcal{A},\overline{\mathbf{X}}} = \mathbf{0} \ .$$

Consider $\mathbf{X}^{(1)} = \overline{\mathbf{X}}^{(1)} - \Delta\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)} = \overline{\mathbf{X}}^{(2)} - \Delta\mathbf{X}^{(2)}$, $\mathbf{X}^{(3)} = \overline{\mathbf{X}}^{(3)} - \Delta\mathbf{X}^{(3)}$, with $\Delta\mathbf{X}^{(1)}$, $\Delta\mathbf{X}^{(2)}$, $\Delta\mathbf{X}^{(3)}$ satisfying (10). Because of Theorem 1 and (13) we have

$$\mathbf{M}_{\mathcal{A},\mathbf{X}} \overline{X} - B_{\mathcal{A},\mathbf{X}} = O(\|\Delta\mathbf{X}\|^2) \ .$$

Because $\mathbf{M}_{\mathcal{A},\overline{\mathbf{X}}}$ is nonsingular, we can write:

$$\begin{aligned}(\|\Delta\overline{\mathbf{X}}_{k+1}^{(1)}\|^2 + \|\Delta\overline{\mathbf{X}}_{k+1}^{(2)}\|^2 + \|\Delta\overline{\mathbf{X}}_{k+1}^{(3)}\|^2)^{1/2} = \|\overline{X} - \overline{X}_{k+1}\| \\ = \|\overline{X} - \mathbf{M}_{\mathcal{A},\mathbf{X}_k}^{-1} B_{\mathcal{A},\mathbf{X}_k}\| = O(\|\mathbf{M}_{\mathcal{A},\mathbf{X}_k} B_{\mathcal{A},\mathbf{X}_k} - \overline{X}\|) = O(\|\Delta\mathbf{X}_k\|^2) \ . \quad (15)\end{aligned}$$

This equation indicates that the convergence is quadratic. Finally, we verify that

$$\|\Delta\overline{\mathbf{X}}_{k+1}^{(n)}\|^2 = O(\min_{\mathbf{Q} \in \mathbf{O}(R_n)} \|\mathrm{qf}(\overline{\mathbf{X}}_{k+1}^{(n)}) - \mathbf{X}^{(n)}\mathbf{Q}\|^2), \quad n = 1, 2, 3 \ .$$

This means that the normalization in step 1 of Alg. 1 does not decrease the convergence rate. □

4 Variants for Dimensionality Reduction in ICA

Variant 1. Several ICA-methods are based on the joint diagonalization of a set of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_J \in \mathbb{R}^{I \times I}$. In the absence of noise, these matrices satisfy

$$\mathbf{A}_j = \mathbf{M} \cdot \mathbf{D}_j \cdot \mathbf{M}^T, \qquad j = 1, \ldots, J$$

in which $\mathbf{M}$ is the mixing matrix and $\mathbf{D}_j \in \mathbb{R}^{R \times R}$ are diagonal. These matrices can be stacked in a tensor $\mathcal{A} \in \mathbb{R}^{I \times I \times J}$. Because the columns of all $\mathbf{A}_j$ are linear combinations of the columns of $\mathbf{M}$, the 1-mode vector space of $\mathcal{A}$ is the column space of $\mathbf{M}$ and its 1-mode rank equals R. Because of the symmetry, the 2-mode vector space also coincides with the column space of $\mathbf{M}$ and the 2-mode rank is also equal to R. It can be verified that the 3-mode vectors are linear combinations of the vectors $(\mathbf{D}_1(r,r), \ldots, \mathbf{D}_J(r,r))^T$, $r = 1, \ldots, R$. This is shown in detail in [5]. Hence the 3-mode rank is bounded by R.

A dimensionality reduction can thus be achieved by computing the best rank-(R, R, R) approximation of $\mathcal{A}$. A difference with Sect. 3 is that now $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$, $\overline{\mathbf{X}}^{(1)} = \overline{\mathbf{X}}^{(2)}$, $\mathbf{R}_1(\mathbf{X}) = \mathbf{R}_2(\mathbf{X})$, because of the symmetry. When $R < J$, this can simply be inserted in (13). Equation (14) then becomes a square set in $(I + J)R$ unknowns.

Variant 2. When $R \geqslant J$, the computation can further be simplified. In this case, no dimensionality reduction in the third mode is needed, and $\mathbf{X}^{(3)}$ can be fixed to the identity matrix. Equation (13) reduces to

$$\begin{aligned}
&\overline{\mathbf{X}}^{(1)} \cdot \mathbf{R}_1(\mathbf{X}) \cdot \mathbf{R}_1(\mathbf{X})^T = \\
&\mathbf{A}_{(1)} \cdot \left[(\mathbf{X}^{(1)} \cdot \mathbf{X}^{(1)^T}) \otimes \mathbf{I} \right] \cdot \mathbf{A}_{(1)}^T \cdot (\overline{\mathbf{X}}^{(1)} - 2\mathbf{X}^{(1)}) \\
&+ \mathbf{A}_{(1)} \cdot \left[(\overline{\mathbf{X}}^{(1)} \cdot \mathbf{X}^{(1)^T} + \mathbf{X}^{(1)^T} \cdot \overline{\mathbf{X}}^{(1)}) \otimes \mathbf{I} \right] \cdot \mathbf{A}_{(1)}^T \cdot \mathbf{X}^{(1)} \ .
\end{aligned} \tag{16}$$

(Note that the factor 4 in (13) has been replaced by a factor 2, because two of the terms in (11) vanish.) Equation (14) now becomes a square set in IR unknowns.

Variant 3. Now assume that one wants to avoid the use of second-order statistics (e.g. because the observations are corrupted by additive coloured Gaussian noise). We consider the case where the dimensionality reduction is based on the observed fourth-order cumulant $\mathcal{K}^Y \in \mathbb{R}^{I \times I \times I \times I}$ instead. In the absence of noise we have

$$\mathcal{K}^Y = \mathcal{K}^S \times_1 \mathbf{M} \times_2 \mathbf{M} \times_3 \mathbf{M} \times_4 \mathbf{M},$$

in which $\mathcal{K}^S \in \mathbb{R}^{R \times R \times R \times R}$ is the source cumulant. This equation implies that all n-mode vectors, for arbitrary n, are linear combinations of the R mixing vectors. In other words, $\mathcal{K}^Y$ is a supersymmetric rank-(R, R, R, R) tensor. Hence it is natural to look for a matrix $\mathbf{X}^{(1)} \in \mathrm{St}(R, I)$ that maximizes

$$g(\mathbf{X}^{(1)}) = \| \mathcal{K}^Y \times_1 \mathbf{X}^{(1)^T} \times_2 \mathbf{X}^{(1)^T} \times_3 \mathbf{X}^{(1)^T} \times_4 \mathbf{X}^{(1)^T} \|^2 \ .$$

A necessary condition is that $\mathbf{X}^{(1)}$ maximizes

$$h(\mathbf{U}) = \| \mathbf{U}^T \cdot \mathbf{K}_{(1)}^Y \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)}) \|^2, \quad \mathbf{U} \in \mathrm{St}(R, I) \ . \tag{17}$$

The matrix $\mathbf{K}^Y_{(1)} \in \mathbb{R}^{I \times I^3}$ is a matrix unfolding of $\mathcal{K}^Y$. Given (17), we can proceed as in Sect. 2 and 3.

Variant 4. Finally, we consider the mixed use of second- *and* fourth-order statistics. In this case, it is natural to consider the maximization of the function

$$g(\mathbf{X}^{(1)}) = \|\mathbf{X}^{(1)^T} \cdot \mathbf{C}^Y \cdot \mathbf{X}^{(1)}\|^2 + \|\mathbf{X}^{(1)^T} \cdot \mathbf{K}^Y_{(1)} \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)})\|^2, \quad (18)$$

in which the two terms are possibly weighted. The optimal $\mathbf{X}^{(1)}$ has to maximize $h(\mathbf{U}) = \|\mathbf{U}^T \cdot \mathbf{F}^Y(\mathbf{X}^{(1)})\|^2$, with

$$\mathbf{F}^Y(\mathbf{X}^{(1)}) = \left(\mathbf{C}^Y \cdot \mathbf{X}^{(1)} \quad \mathbf{K}^Y_{(1)} \cdot (\mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)} \otimes \mathbf{X}^{(1)})\right) .$$

A necessary condition is that

$$\mathbf{X}^{(1)} \cdot \mathbf{W}_1 = \mathbf{F}^Y(\mathbf{X}^{(1)}) \cdot (\mathbf{F}^Y(\mathbf{X}^{(1)}))^T \cdot \mathbf{X}^{(1)} \quad (19)$$

for some $\mathbf{W}_1 \in \mathbb{R}^{R \times R}$. From here, we can proceed as in Sect. 3. The role of $\mathbf{R}_1(\mathbf{X}) \cdot \mathbf{R}_1(\mathbf{X})^T$ is played by $\mathbf{W}_1$.

5 Conclusion

We have derived a higher-order Grassmann-Rayleigh Quotient Iteration, which can be used for dimensionality reduction in ICA without prewhitening. The convergence is quadratic and each iteration step merely involves solving a square set of linear equations. This is a big improvement over the algorithm discussed in [3], of which the convergence is at most linear and of which each iteration involves the partial computation of a number of SVDs. The relevance to ICA is further substantiated in [5], which also contains some simulation results.

References

1. Absil, P.-A., Mahony, R., Sepulchre, R., Van Dooren, P.: A Grassmann-Rayleigh quotient iteration for computing invariant subspaces. SIAM Rev. **44** (2002) 57–73.
2. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21** (2000) 1253–1278.
3. 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.
4. 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).
5. De Lathauwer, L., Vandewalle, J.: Dimensionality Reduction in ICA and Rank-$(R_1, R_2, \ldots, R_N)$ Reduction in Multilinear Algebra. Proc. ICA 2004.
6. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20** (1998) 303–353.
7. 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.
8. Zhang, T., Golub, G.H.: Rank-one approximation to high order tensors. SIAM J. Matrix Anal. Appl. **23** (2001) 534–550.