A Short Introduction to Tensor-Based Methods for Factor Analysis and Blind Source Separation

Lieven De Lathauwer[1]
Kulak — Group Science, Engineering and Technology
Katholieke Universiteit Leuven
Kortrijk, Belgium
Lieven.DeLathauwer@kuleuven-kortrijk.be

***Abstract*—In this survey we explain how the Canonical Polyadic Decomposition of higher-order tensors is connected to different types of Factor Analysis and Blind Source Separation.**

I. Factor Analysis and Blind Source Separation

Assume a data matrix $\mathbf{M} \in \mathbb{R}^{I_1 \times I_2}$ is given. The goal of Factor Analysis (FA) and Blind Source Separation (BSS) is to decompose $\mathbf{M}$ in interpretable rank-1 components:

$$\mathbf{M} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \cdot \mathbf{a}_r^{(2)^T}, \tag{1}$$

in which $\mathbf{a}_r^{(1)} \in \mathbb{R}^{I_1}$, $\mathbf{a}_r^{(2)} \in \mathbb{R}^{I_2}$, $1 \leqslant r \leqslant R$. Eq. (1) can also be written as:

$$\mathbf{M} = \mathbf{A}^{(1)} \cdot \mathbf{A}^{(2)^T}. \tag{2}$$

In FA terminology the vectors $\mathbf{a}_r^{(2)}$ are the factors and the vectors $\mathbf{a}_r^{(1)}$ are the loadings, respectively. In BSS terminology the vectors $\mathbf{a}_r^{(2)}$ are the sources and the vectors $\mathbf{a}_r^{(1)}$ are the mixing vectors, respectively, and matrix $\mathbf{A}^{(1)}$ is the mixing matrix.

It is clear that FA-BSS is a very fundamental problem. In telecommunication, the different terms in (1) could be associated to different users in the system. (In such applications, the data take complex values.) In biomedical engineering, the different terms could represent different biomedical phenomena (e.g. different parts of neuronal activity in EEG). In chemometrics, the different terms could be generated by different components in a chemical solution. Other applications can be found in econometrics, psychometrics, image processing, speech processing, vibro-acoustics, analysis of network and hyperlink data, astrophysics, and so on. Our discussion is limited to the case of linear instantaneous mixtures. Noise is considered as a perturbation of the equations and is omitted for clarity of exposition.

It is easy to see that Eqs. (1–2) by themselves do not allow us to solve the FA-BSS problem. Indeed, if (2) is a valid decomposition of $\mathbf{M}$, then so is

$$\mathbf{M} = (\mathbf{A}^{(1)}\mathbf{B}^T) \cdot (\mathbf{A}^{(2)}\mathbf{B}^{-1})^T \tag{3}$$

for any nonsingular $\mathbf{B} \in \mathbb{R}^{R \times R}$. From (3) a set of rank-1 components can be derived that is different from the set in (1) but equally valid. This corresponds to the fact that the decomposition of a matrix in rank-1 terms is not unique.

In Section II we situate Principal Component Analysis (PCA) in the context of FA-BSS. In Section III some uniqueness properties of the Canonical Polyadic Decomposition (CPD) of higher-order tensors are mentioned. We briefly explain how the decomposition can be computed and how it can be used for FA-BSS. Section IV deals with Independent Component Analysis (ICA). For different approaches we make the connection with CPD. Throughout, references are provided for further reading. We mention in particular the books and surveys [1], [2], [3], [4], [5], [6], [7]. Our presentation is in terms of real-valued data. The results can be generalized to the complex case.

Notation. Scalars are denoted by lowercase italic letters $(a, b, \ldots)$, vectors by lowercase boldface letters $(\mathbf{a}, \mathbf{b}, \ldots)$, matrices by boldface capitals $(\mathbf{A}, \mathbf{B}, \ldots)$ and tensors by calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$. Italic capitals are used to denote index upper bounds $(i = 1, 2, \ldots, I)$. 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}. Likewise, we have $(\mathcal{A})_{i_1 i_2 \ldots i_N} = a_{i_1 i_2 \ldots i_N}$. The columns of $\mathbf{A}$ are denoted by $\mathbf{a}_1$, $\mathbf{a}_2$, ...Conversely, the matrix with columns $\mathbf{a}_1$, $\mathbf{a}_2$, ...is denoted by $\mathbf{A}$. For a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, $\mathbf{T}_{:,:,i_3,\ldots,i_N}$ represents the $(I_1 \times I_2)$ slice indexed by i_3, ..., i_N and $\mathbf{T}_{i_1,\ldots,i_{N-2},:,:}$ represents the $(I_{N-1} \times I_N)$ slice indexed by i_1, ..., i_{N-2}. The $(R \times R)$ diagonal matrix holding the values a_1, a_2, ..., a_R is denoted by $\mathrm{diag}(a_1, a_2, \ldots, a_R)$. The superscripts $\cdot^T$ and $\cdot^\dagger$ denote the transpose and the Moore-Penrose pseudo-inverse, respectively. The mathematical expectation is denoted by $E\{\cdot\}$.

II. Principal Component Analysis

PCA essentially amounts to computing the Singular Value Decomposition (SVD) of $\mathbf{M}$ [8]:

$$\mathbf{M} = \mathbf{U}^{(1)} \cdot \mathrm{diag}(\sigma_1, \ldots, \sigma_R) \cdot \mathbf{U}^{(2)^T} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r^{(1)} \mathbf{u}_r^{(2)^T},$$

[1]The author is also with the Electrical Engineering Department (ESAT), Research Division SCD, of the Katholieke Universiteit Leuven. Research supported by: (1) Research Council K.U.Leuven: GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), CIF1 and STRT1/08/023 (2) F.W.O.: (a) project G.0427.10N, (b) Research Communities ICCoS, ANMMM and MLDM, (3) the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), (4) EU: ERNSI.

in which $\mathbf{U}^{(1)} \in \mathbb{R}^{I_1 \times R}$ and $\mathbf{U}^{(2)} \in \mathbb{R}^{I_2 \times R}$ are column-wise orthonormal and in which the singular values σ_r are positive, $1 \leqslant r \leqslant R$. If the singular values are distinct, then the dyads $\sigma_r \mathbf{u}_r^{(1)} \mathbf{u}_r^{(2)^T}$ are unique, $1 \leqslant r \leqslant R$. Note that this uniqueness is thanks to the orthogonality constraints on $\mathbf{U}^{(1)}$ and $\mathbf{U}^{(2)}$, which were not present in the original FA-BSS problem formulation. It depends on the data whether the additional constraints make sense.

Let us give an example from chemometrics [7]. In spectroscopy, Beer-Lambert's law states that an Excitation-Emission Matrix (EEM) $\mathbf{M}$, obtained from a diluted sample, can be decomposed as in (1), where the vectors $\mathbf{a}_r^{(1)}$ are proportional to the excitation spectra, the vectors $\mathbf{a}_r^{(2)}$ proportional to the emission spectra, and R corresponds to the total number of fluorophores that are present in the solution. Since spectra are nonnegative, they can only be orthogonal if they have non-overlapping support. It is clear that here PCA is of limited use.

III. Canonical Polyadic Analysis

A. Preliminaries

Definition 1: The *outer product* $\mathcal{A} \otimes \mathcal{B}$ of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_P}$ and a tensor $\mathcal{B} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_Q}$ is the tensor defined by $(\mathcal{A} \otimes \mathcal{B})_{i_1 i_2 \ldots i_P j_1 j_2 \ldots j_Q} = a_{i_1 i_2 \ldots i_P} b_{j_1 j_2 \ldots j_Q}$, for all values of the indices.

For instance, the outer product $\mathcal{T}$ of three vectors $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$ and $\mathbf{a}^{(3)}$ is defined by $t_{i_1 i_2 i_3} = a_{i_1}^{(1)} a_{i_2}^{(2)} a_{i_3}^{(3)}$.

Definition 2: An Nth-order tensor has *rank* 1 iff it equals the outer product of N nonzero vectors.

Definition 3: The *rank* of a tensor $\mathcal{T}$ is the minimal number of rank-1 tensors that yield $\mathcal{T}$ in a linear combination.

B. Canonical Polyadic Decomposition

Definition 4: A *Canonical Polyadic Decomposition (CPD)* of a rank-R tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a decomposition of $\mathcal{T}$ in a linear combination of R rank-1 terms:

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \otimes \mathbf{a}_r^{(2)} \otimes \cdots \otimes \mathbf{a}_r^{(N)}. \tag{4}$$

The fully symmetric variant, in which $\mathbf{a}_r^{(1)} = \mathbf{a}_r^{(2)} = \ldots = \pm \mathbf{a}_r^{(N)}$, $r = 1, \ldots, R$, was studied in the nineteenth century in the context of invariant theory [9]. The unsymmetric decomposition was introduced in 1927 [10], [11]. The decomposition is also known as the Canonical Decomposition (CANDECOMP) [12] or Parallel Factor Decomposition (PARAFAC) [13]. The terms CANDECOMP and PARAFAC are sometimes used when the number of rank-1 terms is not minimal. The decomposition is visualized for third-order tensors in Fig. 1.

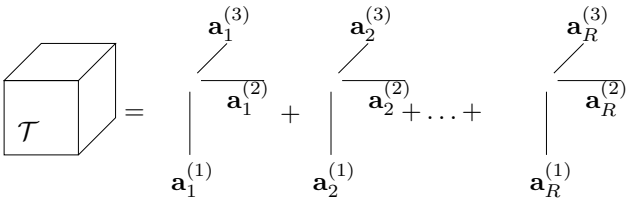


Fig. 1. Visualization of the CPD of a third-order tensor.

CPD (4) is often written in a slice-wise fashion as:

$$\begin{aligned} \mathbf{T}_{:,:,i_3,\ldots,i_N} &= \mathbf{A}^{(1)} \cdot \mathrm{diag}(a_{i_3 1}^{(3)} \ldots a_{i_N 1}^{(N)}, a_{i_3 2}^{(3)} \ldots a_{i_N 2}^{(N)}, \\ &\quad \ldots, a_{i_3 R}^{(3)} \ldots a_{i_N R}^{(N)}) \cdot \mathbf{A}^{(2)^T}, \end{aligned} \tag{5}$$

$1 \leqslant i_3 \leqslant I_3$, …, $1 \leqslant i_N \leqslant I_N$. In the case of third-order tensors, (5) reduces to

$$\mathbf{T}_{:,:,i_3} = \mathbf{A}^{(1)} \cdot \mathrm{diag}(a_{i_3 1}^{(3)}, a_{i_3 2}^{(3)}, \ldots, a_{i_3 R}^{(3)}) \cdot \mathbf{A}^{(2)^T}, \tag{6}$$

$1 \leqslant i_3 \leqslant I_3$. We see that in the third-order case the computation of the CPD of a tensor $\mathcal{T}$ is equivalent to the simultaneous diagonalization of its matrix slices.

Contrary to the decomposition of a matrix in rank-1 terms, CPD of a higher-order tensor is unique up to trivial indeterminacies under rather mild conditions. In particular, CPD can be unique for values of R that exceed the tensor dimensions. The most general conditions to date are given in [14], [15], [16], [17], [18]. In FA-BSS the following theorem often applies. It has several times been rediscovered, see e.g. [19], [20], [21].

Theorem 1: Consider a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ that admits the CPD (4), where $N = 3$. If (i) the columns of $\mathbf{A}^{(1)}$ are linearly independent, (ii) the columns of $\mathbf{A}^{(2)}$ are linearly independent, and (iii) $\mathbf{A}^{(3)}$ does not have proportional columns, then the CPD is unique modulo permutation of the rank-1 terms and modulo scaling / counterscaling of factors within the same term.

Under the conditions in Theorem 1, CPD can be computed by means of a matrix generalized Eigenvalue Decomposition [19], [20], [21]. In [15] a more general decomposition, in which only $\mathbf{A}^{(1)}$ is required to have linearly independent columns (possibly $R > I_2, I_3$), is reworked to a decomposition of the form in Theorem 1. In [22] the same is done for a decomposition in which none of the matrices $\mathbf{A}^{(n)}$ is required to have linearly independent columns (possibly $R > I_n$, $1 \leqslant n \leqslant N$), where the working assumptions include that the tensor is (at least) of order four. Note that, in several cases, an exact decomposition can be computed by means of conventional linear algebra.

A case where CPD is not unique, is the following. Assume in (4) that, for $r \neq r'$, $\mathbf{a}_r^{(3)} \otimes \cdots \otimes \mathbf{a}_r^{(N)}$ and $\mathbf{a}_{r'}^{(3)} \otimes \cdots \otimes \mathbf{a}_{r'}^{(N)}$ are proportional. Then (5) implies that the r-th and r'-th term cannot be separated.

In the context of FA-BSS, CPD allows in principle only for an approximation of the data. One then fits the decomposition to the given tensor, for instance in least-squares sense. The solution obtained under the assumption of an exact decomposition may be used to initialize the algorithm. The most popular algorithm is Alternating Least Squares (ALS) [12], [13]. Finding (say) $\mathbf{a}_r^{(1)}$, $1 \leqslant r \leqslant R$, from (4), assuming that the factors in modes 2–N are fixed, is just a linear least-squares problem. The algorithm alternates between such conditional updates. ALS is easy and it often works well, but it can also be (very) slow, especially in the case of ill-conditioned problems. There is no guarantee that the algorithm will find the global optimum. Another disadvantage is that, in

the case of (partially) symmetric tensors, the algorithm breaks the symmetry.

Line search schemes form an interesting alternative. As noticed in [23], the multilinearity of the problem allows one to compute the optimal step along a line by polynomial rooting. Expressions for tensors of arbitrary order are given in [24]. The fact that the step is optimal, may significantly increase the convergence speed and reduce the risk of convergence to non-global optima. On the other hand, in the N-th order case, the cost of determining the polynomial coefficients is roughly equal to the cost of $2N-1$ evaluations of the cost function along the line. An additional advantage of (even suboptimal) line search is that it naturally preserves the symmetry [25].

Conjugate gradient optimization has about the same computational complexity as ALS, but converges faster in a neighbourhood of the solution and seems to be more robust to over-factoring [26], [27]. Levenberg-Marquardt optimization is quite expensive but has fast convergence and seems to work well in relatively ill-conditioned cases [28]. Under the conditions in Theorem 1, CPD is reformulated in terms of a Simultaneous Generalized Schur Decomposition, with orthogonal unknowns, in [29]. Least absolute error minimization is discussed in [30]. A number of methods are compared in [28].

Although CPD is often unique as such, it may make sense to impose orthogonality constraints if they are known to apply. This may improve the convergence, increase the accuracy and even allow uniqueness for higher values of R, see [31] and references therein. CPD with orthogonality constraints of (partially) symmetric tensors is briefly discussed in Section IV-D.

The cost function associated with the rank-R approximation of a higher-order tensor does not always have a minimum; sometimes it only has an infimum. Algorithms will then show the following behavior. Decreasing the cost function will eventually make some terms go to infinity, these terms almost completely cancel each other and the overall sum yields a better and better approximation of the given tensor. See [32], [33] and references therein. The problem cannot occur if the decomposition is exact. In FA-BSS practice, diverging components indicate that the noise level is too high, that the underlying components cannot be accurately represented by rank-1 tensors, etc. In some cases one may prevent terms from going to infinity as the cost function is decreased by imposing proper constraints. One could for instance impose nonnegativity constraints [34], orthogonality constraints [33] or bound the condition number of the factor matrices [35]. If it is impossible to impose meaningful constraints, then one has to conclude that CPD does not provide the right model for the data. A more general decomposition was introduced in [36], [37], [38]. This decomposition was used as an alternative representation in [39]; the study was limited to tensors consisting of two matrix slices.

C. Analysis

Let us resume the chemometrics example from Section II [7]. Let us assume that we dispose of $I_3 > 1$ samples, which contain the same fluorophores in different concentrations. For the different samples EEMs $\mathbf{T}_{:,:,i_3} \in \mathbb{R}^{I_1 \times I_2}$ are measured and stacked in an excitation-emission tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. The tensor $\mathcal{T}$ can be decomposed as in (4), with $N = 3$, where the vectors $\mathbf{a}_r^{(1)}$, $\mathbf{a}_r^{(2)}$ and $\mathbf{a}_r^{(3)}$ are proportional to the pure excitation spectra, emission spectra and concentration profiles, respectively, and R corresponds to the total number of fluorophores. If CPD is unique, then it allows a mathematical (as opposed to "chemical") analysis of the samples. Note that we did not impose inappropriate orthogonality constraints.

The example in the previous paragraph is different from the one in Section II in the sense that we dispose of a data tensor instead of just a data matrix. The tensorization relies on the measurement of different EEMs, for samples that contain the same fluorophores in different concentrations. Such a set of samples is not always available.

CPD can also be useful in cases where only a single data matrix is available. Obviously, the data need to be tensorized first. One could for instance compute a two-way wavelet representation for all the rows of the data matrix, as in [40], [41]. Stacking the different wavelet matrices yields a data tensor. The crucial assumption here is that, after tensorization, the components are approximately rank-1 tensors.

IV. Independent Component Analysis

A. Preliminaries

Definition 5: Let $\mathbf{x} \in \mathbb{R}^I$ denote a random vector and $p_{\mathbf{x}}(\mathbf{x})$ its probability density function. The *second characteristic function* $\Psi_{\mathbf{x}}(\boldsymbol{\omega})$ of $\mathbf{x}$ is the natural logarithm of the complex conjugated Fourier transform of $p_{\mathbf{x}}(\mathbf{x})$:

$$\Psi_{\mathbf{x}}(\boldsymbol{\omega}) = \ln(E\{e^{j\boldsymbol{\omega}^T\mathbf{x}}\}). \tag{7}$$

Note that $p_{\mathbf{x}}(\mathbf{x})$ and $\Psi_{\mathbf{x}}(\boldsymbol{\omega})$ contain the same information. The second characteristic function of a Gaussian random vector is a multivariate polynomial of degree two. Cumulants of a random vector are derived from the Taylor series of $\Psi_{\mathbf{x}}(\boldsymbol{\omega})$ about zero [42].

Definition 6: Let $\mathbf{x} \in \mathbb{R}^I$ denote a random vector and $\Psi_{\mathbf{x}}(\boldsymbol{\omega})$ its second characteristic function. Then the *N-th order cumulant* $\mathcal{C}_{\mathbf{x}}^{(N)} \in \mathbb{R}^{I \times I \times \ldots \times I}$ of $\mathbf{x}$ is the fully symmetric N-th order tensor of which the entries are given by:

$$(\mathcal{C}_{\mathbf{x}}^{(N)})_{i_1 i_2 \ldots i_N} = (-j)^N \left. \frac{\partial^N \Psi_{\mathbf{x}}(\boldsymbol{\omega})}{\partial\omega_{i_1}\partial\omega_{i_2}\ldots\partial\omega_{i_N}} \right|_{\boldsymbol{\omega}=\mathbf{0}}. \tag{8}$$

For Gaussian random vectors only first- and second-order cumulants can be different from zero. This implies that cumulants of order higher than two do not sense the "Gaussian part" of a random vector. On the other hand, they are also blind to additive Gaussian noise.

Let $\tilde{\mathbf{x}} = \mathbf{x} - E\{\mathbf{x}\}$ denote the centered version of random vector $\mathbf{x} \in \mathbb{R}^I$. Explicit expressions for the cumulants of $\mathbf{x}$ up

to order four are:

$$
\begin{aligned}
(\mathbf{c}_{\mathbf{x}})_{i_1} &= E\{x_{i_1}\}, && (9)\\
(\mathbf{C}_{\mathbf{x}})_{i_1 i_2} &= E\{\tilde{x}_{i_1}\tilde{x}_{i_2}\}, && (10)\\
(\mathcal{C}^{(3)}_{\mathbf{x}})_{i_1 i_2 i_3} &= E\{\tilde{x}_{i_1}\tilde{x}_{i_2}\tilde{x}_{i_3}\}, && (11)\\
(\mathcal{C}^{(4)}_{\mathbf{x}})_{i_1 i_2 i_3 i_4} &= E\{\tilde{x}_{i_1}\tilde{x}_{i_2}\tilde{x}_{i_3}\tilde{x}_{i_4}\}\\
&\quad -E\{\tilde{x}_{i_1}\tilde{x}_{i_2}\}E\{\tilde{x}_{i_3}\tilde{x}_{i_4}\}\\
&\quad -E\{\tilde{x}_{i_1}\tilde{x}_{i_3}\}E\{\tilde{x}_{i_2}\tilde{x}_{i_4}\}\\
&\quad -E\{\tilde{x}_{i_1}\tilde{x}_{i_4}\}E\{\tilde{x}_{i_2}\tilde{x}_{i_3}\}, && (12)
\end{aligned}
$$

for all index entries. An explicit expression for cumulants of arbitrary order is given by the Leonov-Shiryaev formula [43, (2.5)].

Cumulants of odd order are zero if $p_{\mathbf{x}}(\mathbf{x})$ is symmetric about the origin [43]. For increasing order and/or dimension, the number of entries to be estimated increases as $O(I^N)$. (To be precise, it is equal to $\binom{I+N-1}{N}$.) Further, the higher the order of a cumulant, the larger is the variance of its sample estimate. On the other hand, the higher the cumulant order, the more different are the rank-1 terms in (15) below [44].

B. Analysis

ICA starts from decomposition (1) where it is additionally assumed that the rows of $\mathbf{A}^{(2)}$ represent samples of a random vector, say $\mathbf{x}$, of which the components are mutually statistically independent [1], [2], [45]. Let us denote $i = i_1$, $I = I_1$ and $\mathbf{A} = \mathbf{A}^{(1)}$. The columns of $\mathbf{M}$ represent samples of an observed random vector, say $\mathbf{y}$, in which the independent components of $\mathbf{x}$ are linearly mixed: $\mathbf{y} = \mathbf{A}\mathbf{x}$. We focus on the estimation of $\mathbf{A}$. If the mixing vectors are linearly independent, then the sources may subsequently be found as $\mathbf{A}^{(2)} = \mathbf{A}^{\dagger} \cdot \mathbf{M}$. This is not possible if the mixing vectors are linearly dependent, in particular in the case $I < R$ (i.e., in the "more-sources-than-sensors" scenario). Under certain conditions the estimation of the sources nevertheless remains possible [46].

Under the independence assumption the following equation holds for any $\boldsymbol{\omega}$ in a neighborhood of the origin:

$$\Psi_{\mathbf{y}}(\boldsymbol{\omega}) = \sum_{r=1}^{R} \Psi_{x_r}\left(\sum_{i=1}^{I} \omega_i\, a_{ir}\right). \qquad (13)$$

Note that $\Psi_{\mathbf{y}}(\boldsymbol{\omega})$ is written as a sum of contributions of the individual sources, i.e., the sources have been separated. This remains true when we apply a linear transformation to both sides of the equation. For instance, computation of second- and higher-order derivatives at the origin leads to the following expressions for the covariance and higher-order cumulants of the data:

$$
\begin{aligned}
\mathbf{C}_{\mathbf{y}} &= \mathbf{A} \cdot \boldsymbol{\Sigma}^2 \cdot \mathbf{A}^T, && (14)\\
\mathcal{C}^{(N)}_{\mathbf{y}} &= \sum_{r=1}^{R} c^{(N)}_{x_r}\, \mathbf{a}_r \otimes \mathbf{a}_r \otimes \cdots \otimes \mathbf{a}_r, && (15)
\end{aligned}
$$

in which $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_R)$ is positive, with σ_r^2 the variance of x_r. Again, there are no cross-terms because of the statistical independence.

Eq. (14) is the symmetric equivalent of (2). This matrix decomposition is not unique. However, because of the symmetry and because the observed covariance $\mathbf{C}_{\mathbf{y}}$ is positive (semi)definite, the square root of the latter can be found up to an orthogonal factor $\mathbf{Q} \in \mathbb{R}^{R\times R}$:

$$\mathbf{C}_{\mathbf{y}} = (\mathbf{A}\boldsymbol{\Sigma}\mathbf{Q}^T) \cdot (\mathbf{A}\boldsymbol{\Sigma}\mathbf{Q}^T)^T. \qquad (16)$$

Eq. (16) is the symmetric equivalent of (3). On the other hand, (15) is a CPD of a symmetric tensor and this decomposition is unique up to trivial indeterminacies under rather mild conditions. Variants of this basic fact underly all algebraic methods for ICA. In (15) the FA-BSS problem is tensorized by means of higher-order cumulants. This is possible under the assumption that the sources are non-Gaussian. (Jointly considering (14) and (15), we can in fact admit one Gaussian source.) Note that (15) allows one to estimate the mixing matrix in cases where $I < R$ [22].

Instead of computing derivatives at the origin, one could compute derivatives at different points in a neighbourhood of the origin. Eq. (13) implies that, if we take N-th order derivatives at $\tilde{I}$ points, these can be stacked in an $(N+1)$-th order tensor $\tilde{\mathcal{C}}^{(N)}_{\mathbf{y}} \in \mathbb{C}^{I\times I\times\cdots\times I\times\tilde{I}}$ that admits the decomposition

$$\tilde{\mathcal{C}}^{(N)}_{\mathbf{y}} = \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{a}_r \otimes \cdots \otimes \mathbf{a}_r \otimes \tilde{\mathbf{a}}_r, \qquad (17)$$

in which $\tilde{\mathbf{a}}_r$ depends on the points for which the derivatives are computed, $1 \leqslant r \leqslant R$. This is a CPD of a tensor with partial symmetry. In the case of second-order derivatives, $\tilde{\mathcal{C}}^{(2)}_{\mathbf{y}}$ is a third-order tensor. The computation of the CPD of the latter is equivalent with the simultaneous diagonalization of its matrix slices $(\tilde{\mathcal{C}}^{(2)}_{\mathbf{y}})_{:,:,\tilde{i}}$, $1 \leqslant \tilde{i} \leqslant \tilde{I}$, see Section III-B. This is the approach taken in [47]. In [48] derivatives of order $N > 2$ are computed in order to deal with situations involving $I = 2$ observation channels and $R > 2$ sources. More general cases are discussed in [2, Chapter 9] and references therein. Again, for $N > 2$, (17) contains only contributions from non-Gaussian sources. For $N = 2$, Gaussian sources contribute by means of vectors $\tilde{\mathbf{a}}_r$ that are proportional to $(1\ 1\ \ldots\ 1)^T$. Hence, following the explanation in Section III-B, Gaussian sources cannot be mutually separated.

C. Covariance-based approaches

Under the assumptions in the previous section at most one Gaussian source can be dealt with. In the current section we work under related but different assumptions that allow several or all sources to be Gaussian. The techniques were introduced in [49], [50].

Let us assume that the data samples depend on a parameter t (say, time). We now assume that the sources are mutually uncorrelated and individually correlated in time, as in [49]. Recall from Section IV-A that $\tilde{\mathbf{x}} = \mathbf{x} - E\{\mathbf{x}\}$ denotes the centered version of random vector $\mathbf{x} \in \mathbb{R}^I$. The covariance matrix $\mathbf{C}_{\mathbf{x}}(\tau) \in \mathbb{R}^{I\times I}$ at lag τ is defined by:

$$(\mathbf{C}_{\mathbf{x}}(\tau))_{i_1 i_2} = E\{\tilde{x}_{i_1}(t)\tilde{x}_{i_2}(t+\tau)\}, \qquad (18)$$

for all index entries. Now let us consider such covariance matrices for a set of lags $\tau_1, \tau_2, \ldots, \tau_{\tilde{I}}$, where possibly $\tau_1 = 0$. We have:

$$\mathbf{C}_{\mathbf{y}}(\tau_{\tilde{i}}) = \mathbf{A} \cdot \mathbf{C}_{\mathbf{x}}(\tau_{\tilde{i}}) \cdot \mathbf{A}^T, \tag{19}$$

in which $\mathbf{C}_{\mathbf{x}}(\tau_{\tilde{i}})$ is diagonal, $1 \leqslant \tilde{i} \leqslant \tilde{I}$. Let us stack $\mathbf{C}_{\mathbf{y}}(\tau_1)$, $\mathbf{C}_{\mathbf{y}}(\tau_2)$, $\ldots$, $\mathbf{C}_{\mathbf{y}}(\tau_{\tilde{I}})$ in a tensor $\mathcal{C}_{\mathbf{y}}(\boldsymbol{\tau}) \in \mathbb{R}^{I \times I \times \tilde{I}}$. As explained in Section III-B, the simultaneous matrix diagonalization in (19) is equivalent with the computation of the CPD

$$\mathcal{C}_{\mathbf{y}}(\boldsymbol{\tau}) = \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{a}_r \otimes \tilde{\mathbf{a}}_r, \tag{20}$$

in which $\tilde{a}_{\tilde{i}r} = (\mathbf{C}_{\mathbf{x}}(\tau_{\tilde{i}}))_{rr}$, $1 \leqslant \tilde{i} \leqslant \tilde{I}$, $1 \leqslant r \leqslant R$. There are no cross-terms because the sources are mutually uncorrelated.

As a second approach, let us assume that the sources are mutually uncorrelated and nonstationary, as in [50]. The covariance matrix $\tilde{\mathbf{C}}_{\mathbf{x}}(t) \in \mathbb{R}^{I \times I}$ at time t is defined by:

$$(\tilde{\mathbf{C}}_{\mathbf{x}}(t))_{i_1 i_2} = E\{\tilde{x}_{i_1}(t)\tilde{x}_{i_2}(t)\}, \tag{21}$$

for all index entries. The tensor $\tilde{\mathcal{C}}_{\mathbf{y}}(\mathbf{t}) \in \mathbb{R}^{I \times I \times \tilde{I}}$, in which covariance matrices for $t = t_1, t_2, \ldots, t_{\tilde{I}}$ are stacked, admits the CPD

$$\tilde{\mathcal{C}}_{\mathbf{y}}(\mathbf{t}) = \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{a}_r \otimes \tilde{\mathbf{a}}_r, \tag{22}$$

in which $\tilde{a}_{\tilde{i}r} = (\tilde{\mathbf{C}}_{\mathbf{x}}(t_{\tilde{i}}))_{rr}$, $1 \leqslant \tilde{i} \leqslant \tilde{I}$, $1 \leqslant r \leqslant R$. Again, there are no cross-terms because the sources are mutually uncorrelated. Eqs. (20) and (22) can in fact be combined. Note that in this section the FA-BSS problem is tensorized through the computation of sets of covariance matrices. This is possible under the assumption that the sources are individually autocorrelated and/or nonstationary. Note also that (20) and (22) allow one to estimate the mixing matrix in cases where $I < R$ [51].

D. Prewhitening

Note that $\mathbf{C}_{\mathbf{y}}$ in (14) contains contributions from all, also Gaussian, sources. Like-wise, $\mathbf{C}_{\mathbf{y}}(0)$ in (19) contains contributions from all, also temporally white, sources. We further assume without loss of generality that also $\tilde{\mathbf{C}}_{\mathbf{y}}(t_1)$ contains contributions from all sources. (Otherwise, we work with a generic linear combination of $\tilde{\mathbf{C}}_{\mathbf{y}}(t_{\tilde{i}})$, $1 \leqslant \tilde{i} \leqslant \tilde{I}$, assuming strictly positive coefficients.)

A large class of algorithms extracts as much information as possible from $\mathbf{C}_{\mathbf{y}}$ (resp. $\mathbf{C}_{\mathbf{y}}(0)$ or $\tilde{\mathbf{C}}_{\mathbf{y}}(t_1)$) and then fixes the remaining degrees of freedom by resorting to $\mathcal{C}_{\mathbf{y}}^{(N)}$ (resp. $\mathbf{C}_{\mathbf{y}}(\tau_{\tilde{i}})$ or $\tilde{\mathbf{C}}_{\mathbf{y}}(t_{\tilde{i}})$, $2 \leqslant \tilde{i} \leqslant \tilde{I}$). Such algorithms are called prewhitening-based. Here we assume that $I \geqslant R$.

Let $\mathbf{P} \in \mathbb{R}^{I \times R}$ be a square root of $\mathbf{C}_{\mathbf{y}}$ in (14), i.e., $\mathbf{C}_{\mathbf{y}} = \mathbf{P} \cdot \mathbf{P}^T$. We know that $\mathbf{P} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{Q}^T$ for some orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{R \times R}$. Let $\mathbf{z} = \mathbf{P}^{\dagger} \cdot \mathbf{y}$. Substitution in (15) yields:

$$\mathcal{C}_{\mathbf{z}}^{(N)} = \sum_{r=1}^{R} \frac{c_{x_r}^{(N)}}{\sigma_{x_r}^N} \mathbf{q}_r \otimes \mathbf{q}_r \otimes \cdots \otimes \mathbf{q}_r. \tag{23}$$

Eq. (23) is a fully symmetric CPD of a fully symmetric tensor in which the factor matrix is orthogonality-constrained. The Jacobi iteration scheme for the computation of the Eigenvalue Decomposition of a symmetric matrix was generalized to this type of tensor decomposition in [52]. Variants were presented in [53], [54], [55]. An overview is given in [2, Chapter 5].

Now let $\mathbf{P} \in \mathbb{R}^{I \times R}$ be a square root of $\mathbf{C}_{\mathbf{y}}(0)$ in (19). With $\mathbf{P} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{Q}^T$ and $\mathbf{z} = \mathbf{P}^{\dagger} \cdot \mathbf{y}$ we obtain from (20):

$$\mathcal{C}_{\mathbf{z}}(\boldsymbol{\tau}) = \sum_{r=1}^{R} \sigma_{x_r}^{-2} \mathbf{q}_r \otimes \mathbf{q}_r \otimes \tilde{\mathbf{a}}_r. \tag{24}$$

Eq. (24) is a partially symmetric CPD of a partially symmetric tensor in which the factor matrix common to the first two modes is orthogonality-constrained. A Jacobi iteration scheme for the computation of this decomposition was presented in [49], [53]. For $\tilde{\mathcal{C}}_{\mathbf{y}}(\mathbf{t})$ in (22) we can work by analogy.

If all sources are non-Gaussian, then the mixing matrix can in principle be obtained from (15) without making use of (14). The estimate of the cumulant $\mathcal{C}_{\mathbf{y}}^{(N)}$ has the disadvantage that it asymptotically only depends on the non-Gaussian part of the data. Analogous, if there are no temporally white sources, then the mixing matrix can in principle be obtained from (19) without making use of $\mathbf{C}_{\mathbf{y}}(0)$. The matrices $\mathbf{C}_{\mathbf{y}}(\tau)$, for $\tau \neq 0$, have the disadvantage that the signal part of their estimates may be weaker than the signal part of the estimate of $\mathbf{C}_{\mathbf{y}}(0)$, since $|c_{x_r}(\tau)| \leqslant c_{x_r}(0)$. On the other hand, $\mathbf{C}_{\mathbf{y}}$ (resp. $\mathbf{C}_{\mathbf{y}}(0)$) has the disadvantage that it is affected by Gaussian (resp. temporally white) noise. We have not discussed the optimal weighting of the equations. Since we are leaving the purely algebraic framework here, suffice it to refer to [56], [57], [58] and references therein.

REFERENCES

[1] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, 2009.

[2] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[3] L. De Lathauwer, "A survey of tensor methods," in *Proc. 2009 IEEE Int. Symp. on Circuits and Systems (ISCAS 2009)*, May 24–27, 2009, pp. 2773–2776.

[4] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sep. 2009.

[5] P. M. Kroonenberg, *Applied Multiway Data Analysis*. Wiley, 2008.

[6] J. M. Landsberg, *Tensors: Geometry and Applications*. AMS, 2011, in press.

[7] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences*. Chichester, U.K.: John Wiley & Sons, 2004.

[8] I. Joliffe, *Principal Component Analysis*. John Wiley & Sons, 2005.

[9] P. Comon and B. Mourrain, "Decomposition of quantics in sums of powers of linear forms," *Signal Processing*, vol. 53, no. 2, pp. 93–107, Sep. 1996.

[10] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *J. Math. and Phys.*, vol. 6, no. 1, pp. 165–189, 1927.

[11] ——, "Multiple invariants and generalized rank of a p-way matrix or tensor," *J. Math. and Phys.*, vol. 7, no. 1, pp. 40–79, 1927.

[12] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart–Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sep. 1970.

[13] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, 1970.

[14] L. Chiantini and G. Ottaviani. (2011) On generic identifiability of 3-tensors of small rank. [Online]. Available: arXiv:1103.2696v1

[15] L. De Lathauwer, "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM J. Matrix Anal. Appl.*, vol. 28, no. 3, pp. 642–666, 2006.

[16] T. Jiang and N. D. Sidiropoulos, "Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models," *IEEE Trans. Signal Processing*, vol. 52, no. 9, pp. 2625–2636, Sep. 2004.

[17] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics," *Linear Alg. Appl.*, vol. 18, no. 2, pp. 95–138, 1977.

[18] A. Stegeman, "On uniqueness of the canonical tensor decomposition with some form of symmetry," *SIAM J. Matrix Anal. Appl.*, 2011, in press.

[19] N. M. Faber, L. M. C. Buydens, and G. Kateman, "Generalized rank annihilation method: I. derivation of eigenvalue problems," *J. Chemometrics*, vol. 8, no. 2, pp. 147–154, Mar. 1994.

[20] S. Leurgans, R. T. Ross, and R. B. Abel, "A decomposition for three-way arrays," *SIAM J. Matrix Anal. Appl.*, vol. 14, no. 4, pp. 1064–1083, Oct. 1993.

[21] E. Sanchez and B. R. Kowalski, "Tensorial resolution: a direct trilinear decomposition," *J. Chemometrics*, vol. 4, no. 1, pp. 29–45, Jan. 1990.

[22] L. De Lathauwer, J. Castaing, and J.-F. Cardoso, "Fourth-order cumulant-based identification of underdetermined mixtures," *IEEE Trans. Signal Processing*, vol. 55, no. 6, pp. 2965–2973, Jun. 2007.

[23] M. Rajih, P. Comon, and R. A. Harshman, "Enhanced line search: A novel method to accelerate PARAFAC," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1148–1171, Sep. 2008.

[24] A. Karfoul, L. Albera, and L. De Lathauwer, "Iterative methods for the canonical decomposition of multi-way arrays: Application to blind underdetermined mixture identification," *Signal Processing*, vol. 91, no. 8, pp. 1789–1802, Aug. 2011.

[25] M. De Vos, D. Nion, S. Van Huffel, and L. De Lathauwer, "A combination of parallel factor and independent component analysis," ESAT-SISTA, K.U.Leuven (Leuven, Belgium), Tech. Rep. 2009-04, 2009.

[26] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *J. Chemometrics*, vol. 25, no. 2, pp. 67–86, Feb. 2011.

[27] P. Paatero, "The multilinear engine: A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model," *J. Comput. Graphical Statist.*, vol. 8, no. 4, pp. 854–888, Dec. 1999.

[28] G. Tomasi and R. Bro, "A comparison of algorithms for fitting the PARAFAC model," *Comp. Stat. & Data Anal.*, vol. 50, no. 7, pp. 1700–1734, 2006.

[29] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 2, pp. 295–327, 2004.

[30] S. A. Vorobyov, Y. Rong, N. D. Sidiropoulos, and A. B. Gershman, "Robust iterative fitting of multilinear models," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2678–2689, Aug. 2005.

[31] M. Sørensen, L. De Lathauwer, P. Comon, S. Icart, and L. Deneire, "Computation of a canonical polyadic decomposition with a semi-unitary matrix factor," ESAT-SISTA, K.U.Leuven (Leuven, Belgium), Tech. Rep. 2011-95, May 2011.

[32] V. De Silva and L.-H. Lim, "Tensor rank and the ill-posedness of the best low-rank approximation problem," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1084–1127, 2008.

[33] W. P. Krijnen, T. K. Dijkstra, and A. Stegeman, "On the non-existence of optimal solutions and the occurrence of "degeneracy" in the Candecomp/Parafac model," *Psychometrika*, vol. 73, pp. 431–439, 2008.

[34] L.-H. Lim and P. Comon, "Nonnegative approximations of nonnegative tensors," *J. Chemometrics*, vol. 23, no. 7–8, pp. 432–441, Jul. 2009.

[35] L. De Lathauwer, "On avoiding diverging components in the computation of the best low rank approximation of higher-order tensors," ESAT-SISTA, K.U.Leuven (Leuven, Belgium), Tech. Rep. 2005-269, 2005.

[36] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms — Part I: Lemmas for partitioned matrices," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1022–1032, 2008.

[37] ——, "Decompositions of a higher-order tensor in block terms — Part II: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1033–1066, 2008.

[38] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms — Part III: Alternating least squares algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 3, pp. 1067–1083, 2008.

[39] A. Stegeman and L. De Lathauwer, "A method to avoid diverging components in the Candecomp/Parafac model for generic $I \times J \times 2$ arrays," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 4, pp. 1614–1638, 2009.

[40] E. Acar, C. Aykut-Bingol, H. Bingol, R. Bro, and B. Yener, "Multiway analysis of epilepsy tensors," *Bioinformatics*, vol. 23, no. 13, pp. i10–i18, Jul. 2007.

[41] M. De Vos, A. Vergult, L. De Lathauwer, W. De Clercq, S. Van Huffel, P. Dupont, A. Palmini, and W. Van Paesschen, "Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone," *Neuroimage*, vol. 37, no. 3, pp. 844–854, Sep. 2007.

[42] P. McCullagh, *Tensor Methods in Statistics*, ser. Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1987.

[43] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis. A Nonlinear Signal Processing Framework*. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[44] L. Albera, A. Ferréol, P. Comon, and P. Chevalier, "Blind identification of overcomplete mixtures of sources (BIOME)," *Linear Alg. Appl.*, vol. 391, pp. 3–30, Nov. 2004.

[45] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[46] J. Eriksson and V. Koivunen, "Identifiability, separability, and uniqueness of linear ICA models," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 601–604, Jul. 2004.

[47] A. Yeredor, "Blind source separation via the second characteristic function," *Signal Processing*, vol. 80, no. 5, pp. 897–902, May 2000.

[48] A. Taleb, "An algorithm for the blind identification of N independent signals with 2 sensors," in *Sixth Int. Symp. on Signal Processing and its Applications (ISSPA'01)*, Aug. 13–16, 2001, pp. 5–8.

[49] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[50] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non-stationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, Sep. 2001.

[51] L. De Lathauwer and J. Castaing, "Blind identification of underdetermined mixtures by simultaneous matrix diagonalization," *IEEE Trans. Signal Processing*, vol. 56, no. 3, pp. 1096–1105, Mar. 2008.

[52] P. Comon, "Independent Component Analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[53] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc.-F*, vol. 140, no. 6, pp. 362–370, Dec. 1993.

[54] P. Comon and E. Moreau, "Improved contrast dedicated to blind separation in communications," in *Proc. ICASSP*, Munich, Germany, Apr. 1997, pp. 3453–3456.

[55] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Independent component analysis and (simultaneous) third-order tensor diagonalization," *IEEE Trans. Signal Processing*, vol. 49, no. 10, pp. 2262–2271, Oct. 2001.

[56] J.-F. Cardoso, S. Bose, and B. Friedlander, "Output cumulant matching for source separation," in *Proc. IEEE Signal Processing / ATHOS Workshop on Higher-Order Statistics*, Girona, Spain, Jun. 12–14, 1995, pp. 44–48.

[57] P. Tichavský and A. Yeredor, "Fast approximate joint diagonalization incorporating weight matrices," *IEEE Trans. Signal Processing*, vol. 57, no. 3, pp. 878–891, Mar. 2009.

[58] A. Yeredor, "Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting," *IEEE Signal Process. Lett.*, vol. 7, no. 7, pp. 197–200, Jul. 2000.