

LABORATOIRE



INFORMATIQUE, SIGNAUX ET SYSTÈMES
DE SOPHIA ANTIPOLIS
UMR 6070

TENSOR DIAGONALIZATION BY ORTHOGONAL TRANSFORMS

Pierre Comon, Mikael Sorensen

Projet ASTRE

Rapport de recherche
ISRN I3S/RR-2007-06-FR

February 2007

RÉSUMÉ :

Les techniques tensorielles sont utilisées de plus en plus en traitement du signal. En particulier, il est souvent utile de transformer un tenseur en un autre qui soit le plus diagonal possible. Nous proposons un algorithme qui fournit 3 matrices orthogonales, chacune opérant sur un des trois modes d'un tenseur d'ordre trois, de façon à maximiser sa trace. Un autre algorithme est proposé pour maximiser la somme des carrés des éléments diagonaux. De tels algorithmes ont déjà été proposés dans le cas de tenseurs symétriques, et utilisés dans le cadre de l'Analyse en Composantes Indépendantes basée sur les cumulants. Notre contribution réside dans l'extension des algorithmes existants au cas non symétrique. On prouve que la solution peut être obtenue en un nombre fini de décompositions en éléments propres de faible dimension, et qu'aucune recherche exhaustive n'est nécessaire.

MOTS CLÉS :

tenseur diagonalisation

ABSTRACT:

Tensor techniques are increasingly used in Signal Processing. In particular, it is often of interest to transform a tensor into another that is as diagonal as possible. We propose an algorithm that yields 3 orthogonal matrices, each acting on every of the three modes of a third order tensor, so that its trace is maximized. Another algorithm is proposed, which maximizes the sum of squares of diagonal entries. Such algorithms have been already proposed in the case of symmetric tensors, and used in the frame of cumulant-based Independent Component Analysis. Our contribution extends existing algorithms to the non symmetric case. It is proved that the solution can be obtained within a finite number of low-dimensional eigenvalue decompositions, and that no exhaustive search is necessary.

KEY WORDS :

tensor diagonalization

Tensor Diagonalization by Orthogonal Transforms

Report I3S-RR-2007-06

Pierre Comon, Mikael Sorensen
www.i3s.unice.fr/~pcomon/Astre/equipe.htm

February 28, 2007

Abstract

Tensor techniques are increasingly used in Signal Processing and Factor Analysis. In particular, it is often of interest to transform a tensor into another that is as diagonal as possible. We propose in this paper an algebraic algorithm that yields 3 orthogonal matrices, each acting on every of the three modes of a third order tensor, so that its trace is maximized. Another algorithm is proposed, which maximizes the sum of squares of diagonal entries. Such algorithms have been already proposed in the case of symmetric tensors, and used in the frame of cumulant-based Independent Component Analysis. Our contribution extends existing algorithms to the non symmetric case. It is proved that the solution can be obtained within a finite number of low-dimensional eigenvalue decompositions, and that no exhaustive search is necessary.

Keywords: tensor, canonical decomposition, Parafac, tensor rank, congruent diagonalization.

1 Introduction

Tensors have been used in Signal Processing for more than a decade, first more or less implicitly through High-Order Statistics [18] [11], in particular for Blind Techniques. Second, orthogonal Tensor Diagonalization has been required in Independent Component Analysis [3]; such tensors were built of cumulants and were symmetric. More recently, a deterministic Blind Identification technique has been proposed and decomposes the data tensor [19]; the decomposition was run iteratively with the help of an Alternating Least Squares algorithm.

Tensors have thus become *de facto* useful tools in various application areas including signal processing and data analysis, even if a reliable theoretical framework and associated numerical algorithms are still lacking.

After general statements related to tensors, including notation and terminology we focus our discussion on the reduction to diagonal arrays by orthogonal change of bases. In particular, our contribution concerns non symmetric tensors, defined on the tensor product of three or more different Euclidian spaces.

Outer product Let $A_{i..j}$ and $B_{k..l}$ be two arrays, of any dimensions. The outer product of these two arrays is the array \mathbf{C} whose entries are defined as:

$$C_{i..j k..l} = A_{i..j} B_{k..l}$$

If arrays \mathbf{A} and \mathbf{B} have r_A and r_B indices, respectively, then \mathbf{C} has $r_A + r_B$ indices. One denotes this outer product as:

$$\mathbf{C} = \mathbf{A} \circ \mathbf{B} \tag{1}$$

Now let \mathbf{T} be a L -way array of dimensions N_ℓ , $1 \leq \ell \leq L$. This array always admits a decomposition into a sum of outer products as:

$$\mathbf{T} = \sum_{r=1}^R \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \dots \circ \mathbf{u}_r^{(L)} \tag{2}$$

where $\mathbf{u}_r^{(\ell)}$ is a $N_\ell \times 1$ array, $\forall r$. This writing is not unique, especially if nothing is imposed to limit the value of integer R .

Tensors If \mathbf{T} takes its values in a field \mathbb{K} , which can be the real or the complex field, arrays $\mathbf{u}_r^{(\ell)}$ may be considered as vectors of the linear space \mathbb{K}^{N_ℓ} . Thus, as a combination of tensor products of vectors, \mathbf{T} may be considered as a tensor. Under a linear change of coordinate system in each space \mathbb{K}^{N_ℓ} , defined by a matrix $\mathbf{A}^{(\ell)}$, the tensor is represented by another array, obtained by the multi-linear transform $\{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \dots, \mathbf{A}^{(L)}\}$. Since it is legitimate once a basis has been defined in the space, **no distinction** will be made in the remainder between the tensor and its array representation.

Similarly, as soon as a canonical basis is fixed in the linear space $\mathbb{K}^{N_1 N_2 \dots N_L}$, no distinction needs to be made between the tensor product between vectors \mathbf{u}_r^ℓ appearing in (2), and the array obtained by making the Kronecker product between the \mathbf{u}_r^ℓ 's considered as vectors of coordinates.

The number of indices necessary to describe the tensor coordinates is called the *order* of the tensor. Thus, a tensor of order L has L *dimensions*, N_ℓ .

The outer product between two tensors \mathbf{A} and \mathbf{B} is often referred to as *tensor product*, and often denoted $\mathbf{A} \otimes \mathbf{B}$; we shall denote it as $\mathbf{A} \circ \mathbf{B}$ because of a possible confusion. In fact, an array associated with a tensor can always be stored in matrix format. Following this practice, the matrix representation of the outer product $\mathbf{A} \circ \mathbf{B}$ is given by the *Kronecker product* of their corresponding matrix representations, and denoted as $\mathbf{A} \otimes \mathbf{B}$. We found it less confusing to use different notations, especially when tensors are of order larger than 2.

Consistency of terminology In physics, the rank of an array sometimes refers to the number of indices minus 1. This is very confusing, since matrices are particular tensors, and with this terminology they would always have rank 1. Yet, the rank of a matrix refers to a totally different object, namely the number of non zero singular values. Hence, it is mandatory to avoid this inappropriate wording, and use *order* to refer to the number of indices.

Valence In some disciplines (including physics), the distinction is made between *covariant* and *contravariant* indices in array representations of tensors. This is relevant when tensors are mappings from an Euclidian space to another, and when their entries are not changed in the same way by a change of basis. The scalar product between vectors may be seen as the image of a vector of the primal space by a linear form of the dual space. A contraction is always represented as a scalar product, hence by the action of a linear form of the dual on a vector of the primal. Consequently, it may be relevant to distinguish between indices corresponding to the primal (covariant indices, appearing as subscripts) and those corresponding to the dual (contravariant indices, appearing as superscripts).

In statistics, tensors of order L are generally symmetric, and are obtained by taking L th derivatives of some scalar function, like a characteristic function. All indices are thus of same nature. For computational purposes, putting some indices as subscripts and others as superscripts may ease the calculus [15].

In Data Analysis, it is somewhat less obvious, but it appears that there is no general rule that could apply, and that could tell us that some indices should be contravariant. On the contrary, there might be more than two Euclidian spaces under consideration (*e.g.* unsymmetric tensors with non

equal dimensions) [19] [20]. Thus again, all indices should be of same nature in a given array, unless reliably justified.

There is however one exception where the distinction might be relevant in statistics, data analysis or signal processing. This occurs if variables take their values in the complex field. In fact, complex conjugation may be seen as a duality operation, which transforms a vector into a linear form in the dual. More generally for a complex tensor, covariant indices would be transformed into contravariant ones under a complex conjugation. Therefore, it may be relevant and meaningful to make this distinction [14].

In order to simplify the presentation, mainly third order real tensors will be subsequently considered. Let \mathbf{T} be a three-way tensor with entries T_{ijk} , $1 \leq i \leq I$, $1 \leq j \leq J$, $1 \leq k \leq K$. Such a 3-way tensor is of dimensions $I \times J \times K$.

Change of basis For our purposes, tensors will merely denote arrays that enjoy the so-called *multi-linearity property* by linear change of coordinates. More precisely, let \mathbf{A} , \mathbf{B} and \mathbf{C} be three matrices of size $I' \times I$, $J' \times J$, and $K' \times K$, respectively. Then a tensor \mathbf{T} is transformed into a tensor \mathbf{T}' given by:

$$T'_{ijk} = \sum_{\ell mn} A_{i\ell} B_{jm} C_{kn} T_{\ell mn} \quad (3)$$

Contraction Contraction is the operation that consists of summing over one of the indices in an expression. For instance, for given tensors \mathbf{A} and \mathbf{B} of orders α and β , having a common k th dimension, one can define the tensor $\mathbf{C} = \mathbf{A} \bullet_k \mathbf{B}$ of order $\alpha + \beta - 2$ as:

$$C_{i_1 \dots i_\alpha j_1 \dots j_\beta} = \sum_{n_k=1}^{N_k} A_{i_1 \dots n_k \dots i_\alpha} B_{j_1 \dots n_k \dots j_\beta}$$

The contraction allows to define the inner product between two tensors of same order and dimensions. For instance, for two third order tensors of dimensions $I \times J \times K$, we have:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathbf{A} \bullet_1 \bullet_2 \bullet_3 \mathbf{B}$$

This inner product induces the Frobenius norm:

$$\|\mathbf{A}\|^2 = \langle \mathbf{A}, \mathbf{A} \rangle = \sum_{ij..k} |A_{ij..k}|^2$$

Note that attention must be paid when several tensors are involved in a series of contractions, like:

$$\mathbf{A} \bullet_1 \mathbf{B} \bullet_2 \mathbf{C}$$

In fact, contraction denoted this way is not associative. As an illustration, the multi-linearity property (3) is often written as:

$$\mathbf{T}' = \mathbf{T} \bullet_1 \mathbf{A} \bullet_2 \mathbf{B} \bullet_3 \mathbf{C}$$

but it means that the index k appearing in the contraction operation \bullet_k corresponds to the k th index of tensor \mathbf{T} , being understood that the summation is always performed on the second index of matrices \mathbf{A} , \mathbf{B} and \mathbf{C} . This is just a matter of convention. The contraction operation notation is pleasant because compact, but must always be redefined every time it is used in order to avoid any ambiguity.

Tensor rank Carroll and Chang [1] and Harshman [13] independently proposed a decomposition that they named CANDECOMP and PARAFAC, respectively. More precisely, given a third order tensor \mathbf{T} of size $I \times J \times K$, this decomposition consists of writing (2) with a minimal number $R(\mathbf{T})$ of terms:

$$\mathbf{T} = \sum_{r=1}^{R(\mathbf{T})} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \quad (4)$$

In other words, there exist three matrices of size $I \times R$, $J \times R$, and $K \times R$, respectively, such that

$$T_{ijk} = \sum_{r=1}^{R(\mathbf{T})} A_{ir} B_{jr} C_{kr} \quad (5)$$

The *rank* of a given tensor \mathbf{T} (and by extension, of the array defining its coordinates in a given basis) is the minimal integer $R(\mathbf{T})$ such that decomposition (4-5) –or more generally decomposition (2)– is exactly satisfied. This minimal decomposition is referred to as the tensor Canonical Decomposition (CAND).

Symmetry A tensor is said to be *symmetric* if the value of its entries do not change by any permutation of its indices: $T_{ij..k} = T_{\sigma(ij..k)}$. It is still an open problem to prove that the rank of a symmetric tensor is the same whether the constraint of symmetry is imposed in every rank-one tensor in the CAND or not.

L	N	2	3	4	5	6	7	8	9	10
3		2	5	7	10	14	19	24	30	36
4		4	9	20	37	62	97			

Table 1: Generic rank of unconstrained arrays of dimension N and order L .

Generic rank An important fact to emphasize is that, contrary to matrices, the rank of a tensor can exceed its dimensions. In order to demonstrate this, one can consider random arrays that are generated by drawing independently the entries according to a continuous distribution. We shall call such arrays *generic arrays*. In the complex field, such arrays always have the same rank with probability one [6].

For instance, a generic matrix of size $I \times J$ has a rank $\min(I, J)$. Moreover, the generic rank of matrices is maximal. These statements do not hold true anymore for higher order tensors.

As an example, a $8 \times 8 \times 8$ tensor has generically a rank equal to 24 (cf. table 1 in the complex field. This *generic rank* is obtained by computing the CAND of a tensor whose entries are randomly drawn according to a continuous probability distribution [9] [7]. For symmetric tensors, the number of degrees of freedom is smaller, and so is the generic rank: a $8 \times 8 \times 8$ tensor has generically a rank equal to 15 in the complex field.

If the CAND is computed in the real field, then it may happen that random tensors do not always have the same rank: the generic rank does not exist, and we must talk about *typical ranks*. Typical ranks are the collection of ranks that can be obtained with non zero probability. The smallest typical rank computed in the real field is equal to the generic rank computed in the complex field [6] [8]. Table 2 and 1 report generic ranks of tensors with equal dimensions, in the symmetric [7] and unconstrained [9] cases, respectively. These values cannot be computed with the help of simple arithmetic relations, unfortunately.

Another striking fact is that the maximal value of the tensor rank is generally larger than the generic rank; it is however unknown for most values of order and dimensions.

Uniqueness Uniqueness of the CAND is to be understood up to a scaling and a permutation of the columns of each mode matrix. By counting the number of degrees of freedom in both sides of (5), one may naively think

L	N	2	3	4	5	6	7	8	9	10
3		2	4	5	8	10	12	15	19	22
4		3	6	10	15	21	30	42	55	72

Table 2: Generic rank of symmetric arrays of dimension N and order L .

that this would give *generic* uniqueness conditions. Unfortunately, this rule is often true, but there are exceptions.

On the other hand, from (5), one can tell that the decomposition cannot be unique if $IJK < R(I+J+K-2)$. More generally, the number of degrees of freedom of a rank-1 tensor of order r and dimensions N_i is:

$$F(r, \mathbf{N}) = \left(\sum_{i=1}^r N_i \right) - r + 1 \quad (6)$$

because of scale ambiguities. A rule solely based on counting the number of degrees of freedom would tell us that uniqueness of the CAND of a tensor \mathbf{T} is reached if and only if

$$R(\mathbf{T}) \left[\sum_{i=1}^r N_i - r + 1 \right] \leq \prod_{i=1}^r N_i$$

It turns out that this condition is sufficient. However, it does not give the generic rank of tensors, as can be checked out by comparing the quantity below (which is actually a lower bound) with the values reported in the tables 2 or 1 for $r \in \{3, 4\}$:

$$\bar{R} \geq \left\lceil \frac{\prod_{i=1}^r N_i}{\sum_{i=1}^r N_i - r + 1} \right\rceil$$

Congruent diagonalization If relation (5) is invertible, that is, if matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} are square and admit inverses \mathbf{A}' , \mathbf{B}' , and \mathbf{C}' , then the multi-linear transform $(\mathbf{A}', \mathbf{B}', \mathbf{C}')$ brings tensor \mathbf{T} into a diagonal tensor with ones in the diagonal.

We have thus a congruent transformation that diagonalizes \mathbf{T} .

It is clear that such a transformation may exist only if the rank of \mathbf{T} is at most equal to its smaller dimension. When it is not the case, it is still possible to define a multi-linear invertible transform that minimizes all non diagonal entries in the obtained tensor, \mathbf{T}' :

$$(\mathbf{A}', \mathbf{B}', \mathbf{C}') = \underset{\mathbf{A}, \mathbf{B}, \mathbf{C}}{\text{Arg Min}} \|\mathbf{T}' - \text{Diag}(\mathbf{T}')\|^2 \quad (7)$$

tensor \mathbf{T}' is not diagonal, but as diagonal as possible, according to some norm.

Orthogonal diagonalization We particularize now the congruent diagonalization to norm-preserving multi-linear transforms. Let \mathbf{U} , \mathbf{V} , and \mathbf{W} be three orthogonal real matrices (or unitary matrices in the complex field), of size $I \times I$, $J \times J$, and $K \times K$, respectively.

By the orthogonal change of bases defined by the triplet $(\mathbf{U}, \mathbf{V}, \mathbf{W})$, tensor \mathbf{G} is transformed into a tensor \mathbf{T} with entries:

$$T_{ijk} = \sum_{pqr} U_{ip} V_{jq} W_{kr} G_{pqr} \quad (8)$$

Under orthogonal transforms, the tensor Frobenius norm is invariant, so that maximizing the sum of squares of diagonal entries is equivalent to minimizing the non diagonal ones. Therefore, the criterion below is appropriate in order to obtain a tensor as diagonal as possible [3]:

$$\Upsilon_2(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \sum_i T_{iii}^2 \quad (9)$$

Again, note that contrary to matrices, for which the Singular Value Decomposition (SVD) yields an exact diagonal form via a maximization of diagonal entries, criterion (9) does not in general lead to a diagonal tensor. The reason is not due to the criterion, but to the fact that tensors have generically a rank that is larger than the smallest dimension [5]. Hence, tensor diagonalization by change of bases (general congruent, or unitary) can only be an *approximation*.

2 Jacobi sweeping

Finding the absolute maximum of criterion (9) is a complicated problem, since these criteria are trigonometric functions in many variables. However, we shall show in this paper that it is possible to solve several much simpler problems in cascade instead.

The first step is to decompose each orthogonal matrix into a product of plane rotations, the so-called Givens rotations, which is possible up to a multiplicative diagonal matrix with unit modulus entries. By doing this, we are left with a single unknown to characterize every Givens rotation. If all Givens rotations are kept fixed except in one given plane in each mode, then the optimization criterion reduces to a rational function in three variables,

$\psi(x, y, z)$, where we can decide that x , y and z denote the tangent of each angle. For instance for the first mode:

$$\mathbf{Q}[\alpha] = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \frac{1}{\sqrt{1+x^2}} \begin{pmatrix} 1 & x \\ -x & 1 \end{pmatrix}$$

where c and s denote $\cos(\alpha)$ and $\sin(\alpha)$, respectively. Of course, this procedure is iterative, even if it is not a relaxation in the strict sense because the optimization is executed over successive elements of a multiplicative group.

In order to be able to carry out such an optimization, it is necessary to solve the problem in the 2-dimensional case, as for matrices. This sweeping strategy, well known for matrices, has already been utilized for symmetric tensors [12] [3], giving birth to the so-called CoM (Contrast Maximization) algorithms. Stationary points of such Jacobi sweeping algorithms is addressed in Appendix.

3 Symmetric tensors

We recall in this section the main results that have been obtained so far for orthogonal diagonalization of symmetric tensors; most of them are reported in [3].

3.1 Maximization of the sum of squares

Invariance property First of all, it can be noticed that the Frobenius norm of a tensor does not change under the action of orthogonal transforms. For the sake of simplicity, let's prove it in the case of real 3rd order tensors, being understood that exactly the same proof can be derived for tensors of any order (possibly complex, under unitary transforms).

Let \mathbf{Q} be an orthogonal matrix, that is, $\sum_j Q_{ij}Q_{jk} = \delta_{ik}$, where δ_{ij} is null except when the two indices are equal, in which case $\delta_{ii} = 1$. Then a symmetric tensor \mathbf{G} is transformed into a tensor \mathbf{T} , whose entries can be written, according to the multi-linearity property, as:

$$T_{ijk} = \sum_{\ell mn} Q_{i\ell}Q_{jm}Q_{kn}G_{\ell mn}$$

Next, the calculation of $\|\mathbf{T}\|^2$ yields

$$\sum_{ijk} T_{ijk}^2 = \sum_{\ell mn} \sum_{\ell' m' n'} \sum_i Q_{i\ell}Q_{i\ell'} \sum_j Q_{jm}Q_{jm'} \sum_k Q_{kn}Q_{kn'} G_{\ell mn}G_{\ell' m' n'}$$

which leads eventually to:

$$\sum_{ijk} T_{ijk}^2 = \sum_{\ell mn} \sum_{\ell' m' n'} \delta_{\ell\ell'} \delta_{mm'} \delta_{nn'} G_{\ell mn} G_{\ell' m' n'}$$

by using the fact that \mathbf{Q} is orthogonal. This is nothing but $\sum_{\ell mn} G_{\ell mn}^2 = \|\mathbf{G}\|^2$.

The consequence is that, as for matrices, minimizing the sum of squares of non diagonal terms is equivalent to the maximization of the sum of squares of diagonal ones, hence the optimization criterion:

$$\Upsilon_1(\mathbf{Q}) = \sum_i |T_{iii}|^2 \quad (10)$$

Symmetry property Now consider the 2-dimensional case. First observe that

$$\mathbf{Q}[\alpha - \pi/2] = \mathbf{Q}[\alpha] \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{Q}[\alpha]$$

In other words, when changing the rotation angle α into $\alpha - \pi/2$, the tangent x is transformed into $-1/x$, so that $(T_{1..1}, T_{2..2})$ is transformed into $(-T_{2..2}, T_{1..1})$. Define the optimization criterion $\psi_2(x) = \Upsilon(\mathbf{Q}[\alpha])$, with $x = \tan(\alpha)$. Then the above symmetry property means:

$$\psi_2(-1/x) = \psi_2(x)$$

This allows not only to reduce the the domain to search for stationary points, but also allows to reduce the degree of the polynomial to root. In fact, it can be seen that $\psi_2(x)$ is a rational function in x of the form

$$\psi_2(x) = \frac{\rho(x)}{(1+x^2)^2}$$

where $\rho(x)$ is a polynomial of degree $2r$ in x , if symmetric tensors of order r are considered. The equation defining stationary points is given by $\omega(x) = (1+x^2)\rho'(x) - 4x\rho(x)$ This polynomial is of degree $2r$, and not $2r+1$ as one may think at first glance. For $r \in \{3, 4\}$ such polynomials (of degree 6 or 8) are generally not solvable algebraically. But it turns out that they are in the present case, because of the particular symmetry property that we just pointed out.

Algebraic solution Since $\omega(x) = 0$ must yield the same roots as $x^{2r}\omega(-1/x) = 0$, the roots x_k can be paired:

$$\omega(x) = x^r \prod_{k=1}^r (x - x_k) \left(x + \frac{1}{x_k}\right) = x^r \prod_{k=1}^r (x^2 - \xi_k x - 1)$$

if we define $\xi_k = x_k - 1/x_k$. So let the new variable $\xi = x - 1/x$. Then, besides possible roots at the origin, which can be easily checked out, polynomial $\omega(x)$ vanishes if and only if polynomial

$$\Omega(\xi) = \prod_{k=1}^r (\xi - \xi_k)$$

vanishes. This polynomial is now of degree r only, and can be solved algebraically. Once roots ξ_k have been calculated, roots $(x_k, -1/x_k)$ can be deduced by rooting the polynomial $x^2 - \xi_k x - 1 = 0$.

For symmetric tensors of order $r = 3$, one obtains a polynomial of degree 2 (and not 3 as expected) [3]:

$$\Omega_3(\xi; g) = d_2 \xi^2 + d_1 \xi - 4 d_2, \quad (11)$$

with

$$\begin{aligned} a_3 &= G_{111}^2 + G_{222}^2, \\ a_2 &= 6 (G_{122} G_{222} - G_{111} G_{112}), \\ a_1 &= 9 (G_{122}^2 + G_{112}^2) + 6 (G_{112} G_{222} + G_{111} G_{122}); \\ d_2 &= a_2/6 = G_{122} G_{222} - G_{111} G_{112}, \\ d_1 &= a_1/3 - a_3. \end{aligned}$$

For symmetric tensors of order $r = 4$, one obtains the polynomial of degree 4 [3] [2]:

$$\Omega_4(\xi; g) = \sum_{i=0}^4 c_i \xi^i \quad (12)$$

with

$$\begin{aligned}
b_4 &= G_{1111}^2 + G_{2222}^2, \\
b_3 &= -8(G_{1111}G_{1112} - G_{1222}G_{2222}), \\
b_2 &= 4b_4 + t + 2w, \\
b_1 &= 4b_3 - 2uv, \\
b_0 &= 2(b_4 + t + 2w + 36G_{1122}^2 + 2G_{1111}G_{2222} \\
&\quad + 32G_{1112}G_{1222}); \\
c_4 &= -b_3/8 = G_{1111}G_{1112} - G_{2222}G_{1222}, \\
c_3 &= 2b_4 - b_2/4 = b_4 - (t + 2w)/4, \\
c_2 &= 3b_3/2 - 3b_1/8 = 3uv/4, \\
c_1 &= b_2 - b_0/2, \\
c_0 &= b_1/2 = 2b_3 - uv.
\end{aligned}$$

where

$$\begin{aligned}
t &= 16(G_{1112}^2 + G_{1222}^2), \\
u &= G_{1111} + G_{2222} - 6G_{1122}, \\
v &= 4(G_{1222} - G_{1112}), \\
w &= 6G_{1122}(G_{1111} + G_{2222}).
\end{aligned}$$

As a conclusion, thanks to symmetry properties, we have been able to solve *algebraically* the search for *absolute* extrema of Υ_r , $r = 3, 4$.

3.2 Maximization of the trace

Under some assumptions involving the signs of rank-one terms [16], it is legitimate to drop the squares. This can be easily shown under the following assumptions: (i) tensor \mathbf{T} is of even order, (ii) it is exactly diagonalizable by congruent transform, (iii) its diagonal form contains entries having the same sign ε .

Proof. Without restricting the generality, let's derive the proof in the case of a 4th order tensor. Define the optimization criterion

$$\Upsilon_1(\mathbf{U}) = \sum_i |T_{iiii}| \quad (13)$$

Then, by assumption (ii) and using the multilinearity property, there exist a diagonal tensor \mathbf{D} and an invertible matrix \mathbf{A} such that

$$T_{iiii} = \sum_j A_{ij}^4 D_{jjjj}$$

Yet, by assumption (iii), $D_{jjjj} = \varepsilon|D_{jjjj}|$. As a consequence, $T_{iiii} = \varepsilon \sum_j A_{ij}^4 |D_{jjjj}|$, which shows that $|T_{iiii}| = \varepsilon T_{iiii}$. Hence, $\Upsilon_1(\mathbf{U}) = \varepsilon \text{trace}\{\mathbf{T}\}$, which completes the proof. If $\varepsilon > 0$, one can thus maximize the trace in order to diagonalize \mathbf{T} (and minimize its trace if $\varepsilon < 0$).

This result a priori does not hold true for 3rd order tensors. We describe now an algebraic approach allowing to compute the absolute extrema of the trace of a fourth order tensor.

Consider a symmetric tensor \mathbf{G} of dimension $2 \times 2 \times 2 \times 2$. After an orthogonal transform defined by the matrix:

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

Since the expressions are quite simple in the present case, let's give them explicitly. Tensor \mathbf{G} is transformed into a tensor \mathbf{T} whose diagonal entries are:

$$\begin{bmatrix} T_{1111} \\ T_{2222} \end{bmatrix} = \frac{1}{(1+x^2)^2} \begin{bmatrix} G_{1111} + 4xG_{1112} + 6x^2G_{1122} + 4x^3G_{1222} + x^4G_{2222} \\ G_{1111}x^4 - 4x^3G_{1112} + 6x^2G_{1122} - 4xG_{1222} + G_{2222} \end{bmatrix}$$

The criterion to maximize (or minimize) is $\psi(x) = T_{1111} + T_{2222}$, that we can denote $\psi(x) = \rho(x)/(1+x^2)^2$. This time, $\rho(x)$ is of degree 4. Now stationary points of $\psi(x)$ are given by the roots of $\omega(x) = (1+x^2)\rho'(x) - 4x\rho(x)$, which is actually of degree 4. Thus its roots can be algebraically computed, by resorting to Ferrari's algorithm for instance.

This presentation of the solution hides an important property, so that it is not clear whether the maximization is still feasible algebraically in the complex case, where we have two unknowns. It turns out that it is indeed feasible, as demonstrated in [4]. Let's then present another solution in the real case following the same lines.

Notice that $\cos^2 2\alpha = (1+x^4-2x^2)(1+x^2)^{-2}$, $\sin^2 2\alpha = 4x^2(1+x^2)^{-2}$, and that $\sin 2\alpha \cos 2\alpha = 2(x-x^3)(1+x^2)^{-2}$. By plugging back these expressions in that of $\psi(x)$, one can notice that the criterion may be viewed as a quadratic form:

$$\psi(x) = \mathbf{v}^T \begin{pmatrix} G_{1111} + G_{2222} & G_{1222} - G_{1112} \\ G_{1222} - G_{1112} & 3G_{1122} + (G_{1111} + G_{2222})/2 \end{pmatrix} \mathbf{v}$$

where vector $\mathbf{v}^T = [\cos 2\alpha, \sin 2\alpha]$. The angle α of the rotation can thus be found by computing the eigenvalue decomposition of this 2×2 matrix, and then solving a simple trigonometric equation.

In fact, as in section 3.1, we have rooted two lower degree polynomials, instead of rooting $\omega(x)$ directly. Bear in mind that the computational complexity is an important issue, since in dimension N the elementary maximization will be executed $N(N-1)/2$ times per sweep, and that there will be in general several sweeps (experimental results suggest $O(\sqrt{N})$ sweeps as a rule of thumb).

4 Unsymmetric tensors

In this section, we try to derive similar results for tensors of more general form. The Jacobi sweeping technique clearly holds, and it suffices to address the 2-dimensional problem. For the sake of convenience and without restricting too much the generality, we shall limit our attention mainly to the $2 \times 2 \times 2$ tensors.

Even if, as far as diagonalization is concerned, maximizing the trace is not meaningful for 3rd order unsymmetric tensors (except perhaps for a quite restricted subset), we start with the least difficult problem: the minimization of the trace. For this purpose, consider the criterion:

$$\Upsilon_1(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \sum_i T_{iii} = \text{trace}\{\mathbf{T}\} \quad (14)$$

Remind that the goal is to devise an routine that is able to solve the 2-dimensional problem with a computational complexity as small as possible (the routine is called every time a pair is processed, that is, many times within a sweep). Therefore, it is desirable to find the *absolute maximum* in a non iterative manner.

4.1 Maximization of the trace

By using the multi-linear relation (8), and after some simple manipulations, the trace of a two-dimensional tensor can be written after an orthogonal change of bases:

$$\sqrt{(1+x^2)(1+y^2)(1+z^2)} \psi_1(x, y, z) = \mathbf{g}^T \boldsymbol{\zeta} \quad (15)$$

where $\boldsymbol{\zeta} = [xyz, xy, xz, yz, x, y, z, 1]^T$ and

$$\mathbf{g}^T = [G_{222} - G_{111}, G_{221} + G_{112}, G_{212} + G_{121}, G_{122} + G_{211}, \\ G_{211} - G_{122}, G_{121} - G_{212}, G_{112} - G_{221}, G_{111} + G_{222}]$$

Stationary points of criterion $\psi_1(x, y, z)$ are thus solutions of the polynomial system below:

$$(g_4yz + g_6y + g_7z + g_8)x - (g_1yz + g_2y + g_3z + g_5) = 0 \quad (16)$$

$$(g_3xz + g_5x + g_7z + g_8)y - (g_1xz + g_2x + g_4z + g_6) = 0 \quad (17)$$

$$(g_2xy + g_5x + g_6y + g_8)z - (g_1xy + g_3x + g_4y + g_7) = 0 \quad (18)$$

where g_i denote the entries of vector \mathbf{g} . This system can be alternatively written as

$$\mathbf{g}_x^T \boldsymbol{\zeta} = 0, \mathbf{g}_y^T \boldsymbol{\zeta} = 0, \mathbf{g}_z^T \boldsymbol{\zeta} = 0 \quad (19)$$

with obvious notation for vectors \mathbf{g}_x , \mathbf{g}_y and \mathbf{g}_z . This system could be solved by resorting to standard elimination tools [10], and would yield 27 solutions for the triplet (x, y, z) . This would be computationally intensive and would ignore the particular structure of the system, namely its sparsity.

4.1.1 Algebraic solution via resultants

The idea consists of remarking that the system (19) is multi-linear in variables x, y, z , so that one variable can be easily eliminated. In other words, $\mathbf{g}_x^T \boldsymbol{\zeta}$ can be written as $a_x(x, y)z + b_x(x, y)$ where

$$a_x(x, y) = (g_4xy + g_7x - g_1y - g_3) \quad (20)$$

$$b_x(x, y) = (g_6xy + g_8x - g_2y - g_5) \quad (21)$$

We have similar expressions for \mathbf{g}_y and \mathbf{g}_z . For instance, the elimination of z yields a system of 2 polynomial equations of degree 4 in two unknowns:

$$a_y b_x - a_x b_y = 0, \quad a_z b_x - a_x b_z = 0 \quad (22)$$

hence having at most 16 distinct solutions. This is a great progress, and this system can be solved with the help of resultants. To be more concrete, this resultant is a determinant of the form

$$\begin{vmatrix} A(y) & 0 & 0 & D(y) \\ B(y) & A(y) & D(y) & E(y) \\ C(y) & B(y) & E(y) & F(y) \\ 0 & C(y) & F(y) & 0 \end{vmatrix} = 0$$

where $A(y)$, $B(y)$, $C(y)$, $D(y)$ are polynomials of degree 2 in y , such that $a_y b_x - a_x b_y = A(y)x^2 + B(y)x + C(y)$ and $a_z b_x - a_x b_z = D(y)x^2 + E(y)x + F(y)$. Note that the system (22) was particular, since only nine monomials

of degree 4 in two variables were present over the fifteen possible ones. The system was indeed sparse, which makes it possible to have a so small resultant. The rooting of this resultant of degree 8 yields the values of y . By plugging them back in (22), one gets two possible solutions for x , for every value of y , and hence 16 solutions for (x, y) . The corresponding value of z is eventually obtained by using (18).

The last step is to find one of the two absolute maxima. In order to do this, we just need to compute $\Upsilon_1(x, y, z)$ at every of the 16 solutions, and pick up the one that yields the maximum.

4.1.2 Algebraic solution via heuristic manipulations

An algebraic solution can be obtained by first eliminate one of the variables, just as in the resultant based approach, say x , then the stationary points has to satisfy the equations

$$\begin{aligned} A(z) + B(z)y + C(z)y^2 &= 0 \\ D(z) + E(z)y + F(z)y^2 &= 0 \end{aligned} \quad (23)$$

where A, B, C, D, E, F are 2nd order degree polynomials in z . This is equivalent to the equation

$$\begin{aligned} F(z) (A(z) + B(z)y + C(z)y^2) - C(z) (D(z) + E(z)y + F(z)y^2) &= 0 \Leftrightarrow \\ (F(z)A(z) - C(z)D(z)) + (F(z)B(z) - C(z)E(z))y &= 0 \Leftrightarrow \\ G(z) + H(z)y &= 0 \end{aligned}$$

where G, H are 4th order degree polynomials in z .

Assume that $H(z) \neq 0$ then

$$y = -\frac{G(z)}{H(z)}. \quad (24)$$

Substitute (24) into (23) we get

$$\begin{aligned} A(z) + B(z) \left(-\frac{G(z)}{H(z)} \right) + C(z) \left(-\frac{G(z)}{H(z)} \right)^2 &= 0 \Leftrightarrow \\ A(z)H(z)^2 - B(z)G(z)H(z) + C(z)G(z)^2 &= 0 \end{aligned}$$

which is a 10th order degree polynomial in z .

In summary to find the maximum we need to solve a 10th order degree polynomial in z next calculate two 4th order degree polynomials in z ten times in order to find the corresponding y and finally the corresponding x can be found from (16). This approach reduces the number of roots to 10.

4.1.3 Eigenvector approach

This approach, sometimes attributed to Macaulay, is based on results borrowed from Algebraic Geometry. In order to introduce them, first define the concept of ideal, in the ring \mathcal{R} of polynomials in several variables.

Ideal An ideal \mathcal{I} is a subring of \mathcal{R} , such that $\forall q \in \mathcal{I}$ and $\forall p \in \mathcal{R}$, then the product $pq \in \mathcal{I}$.

Consider a system \mathcal{P} of polynomial equations in several variables, $q_n(\mathbf{x}) = 0$, $1 \leq n \leq N$, where each polynomial $q_n(\mathbf{x})$ is of global degree d_n . Here \mathbf{x} stands for the set of unknowns $\{x_n\}_{n=1}^N$. Now denote \mathcal{I} the ideal spanned by $\langle q_n \rangle_{n=1}^N$. The problem is to find the variety associated with this ideal, assuming that it is zero-dimensional (which means that the variety is constituted of a finite number of points). We know from a Bézout theorem that if the number of solutions of \mathcal{P} is finite, then it must be at most equal to the product of the degrees, $\prod_n d_n$.

The first key result is that any ideal of \mathcal{R} is finitely generated, which means that there exists a family of polynomials $1 \leq q_n, 1 \leq n \leq N$, such that $\forall q \in \mathcal{I}, \exists p_n \in \mathcal{R}, q = \sum_n p_n q_n$. In other words, any ideal is entirely characterized by a finite generating family $\langle q_i \rangle_{i=1}^n$, often called *basis of the ideal*. This is known as Hilbert's basis theorem. An ideal may have many different bases.

Quotient Then we define the quotient ring modulo \mathcal{I} , denoted $\mathcal{A} = \mathcal{R}/\mathcal{I}$, as the set of equivalence classes as follows: two polynomials p_1 and p_2 belong to the same class, which we write $\tilde{p}_1 \equiv \tilde{p}_2$, iff $p_1 - p_2 \in \mathcal{I}$.

Dual Next, \mathcal{R} and hence \mathcal{A} are also linear spaces. We can then define the dual space $\hat{\mathcal{A}}$. For this purpose, let p_1 and p_2 be two polynomials of the same class, \tilde{p} ; we have by definition $p_1 - p_2 \in \mathcal{I}$. Let ℓ be a linear form of $\hat{\mathcal{A}}$. This form maps \tilde{p} to a number $\ell(\tilde{p})$, and hence p_1 and p_2 to the same number. By linearity, $\ell(p_1) = \ell(p_2)$ yields $\ell(p_1 - p_2) = 0$. Thus, we see that $\hat{\mathcal{A}}$ is the subspace of $\hat{\mathcal{R}}$ of linear forms vanishing on \mathcal{I} .

In particular, let l_α be the linear form that maps a polynomial to its value at a point α , $p(\alpha)$. because of what we have just seen, l_α is in $\hat{\mathcal{A}}$ iff α is a common root of \mathcal{P} , that is:

$$l_\alpha \in \hat{\mathcal{A}} \Leftrightarrow q_n(\alpha) = 0, \forall n \in \{1, \dots, N\} \quad (25)$$

Multiplication operator For any fixed polynomial a of \mathcal{A} , define now the multiplication operator \mathcal{M}_a , which maps any polynomial p of \mathcal{A} to the product $\mathcal{M}_a(p) = pa$.

The transpose operator \mathcal{M}_a^T is, by definition of the transposition, mapping any linear form ℓ of $\hat{\mathcal{A}}$ to the linear form $\mathcal{M}_a^T \ell$ defined by:

$$\forall q \in \mathcal{A}, \mathcal{M}_a^T \ell(q) = \ell(\mathcal{M}_a(q)) = \ell(qa).$$

Eigenvalue decomposition The key result on which eigenvalue techniques are based is the following: eigenvectors of \mathcal{M}_a^T are the linear forms 1_α , where α is any root of \mathcal{P} . Let's prove this result. By definition of \mathcal{M}_a^T , we have $\forall q \in \mathcal{R}$:

$$\mathcal{M}_a^T 1_\alpha(q) = 1_\alpha(\mathcal{M}_a(q)) = 1_\alpha(qa) = a(\alpha)1_\alpha(q)$$

As a consequence, $\mathcal{M}_a^T 1_\alpha = a(\alpha)1_\alpha$, which indeed shows that forms 1_α are eigenvectors of \mathcal{M}_a^T associated with eigenvalues $a(\alpha)$. Note that from (25), α needs to be a root of \mathcal{P} in order for 1_α to belong to $\hat{\mathcal{A}}$.

Basis To summarize, in order to solve polynomial system \mathcal{P} , it is sufficient to build a matrix representing operator \mathcal{M}_a^T in an appropriate basis of $\hat{\mathcal{A}}$, and then to compute all its eigenvectors. If eigenvalues are distinct (which will occur generically), each eigenvector will yield a solution.

This approach is attractive if the construction of this matrix is not too time consuming. In particular, if we have a series of similar polynomial systems to solve, in which most of the work can be done once for all symbolically, the numerical computations will be limited to a minimum, that is, mainly to the calculation of the eigenvalue decomposition.

One possibility to build the basis is suggested in [17]. It is formed of all monomials of the form $\prod_k x_k^{\beta_k}$, where $0 \leq \beta_k \leq d_k - 1$. Other bases could be thought of, but independently of the basis chosen, we know that the matrix obtained will be of size at most (and generically equal to) $\prod_n d_n$. The matrix representing the multiplication operator \mathcal{M}_a^T will always have that size. The conditioning of the EVD calculation, and also the computational effort necessary to build the matrix itself, will depend on the choice of the basis.

Algebraic maximization of the trace Let's turn now to our practical application, and assume the basis

$$\mathcal{B} = \{1, x, x^2, y, xy, x^2y, z, xz, x^2z, yz, xyz, x^2yz\}$$

The elimination of the monomial xyz is indeed possible in two of the three equations in system (16-18), so that $\prod_n d_n = 12$.

4.2 Maximization of the absolute trace

Let the contrast function be

$$\tilde{\Upsilon}_1(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \sum_i |K_{iii}| = \text{trace}\{|K|\}. \quad (26)$$

Assume that $r(K) \geq 2$ then

$$\begin{aligned} \nabla \tilde{\Upsilon}_1 &= \sum_{i=1}^2 \text{sign}(K_{iii}) \nabla K_{iii} \\ &= \sum_{i=1}^2 s_i \nabla K_i \\ &= \begin{cases} \pm (\nabla K_1 + \nabla K_2), & s_1 = s_2 \\ \pm (\nabla K_1 - \nabla K_2), & s_1 = -s_2 \end{cases} \end{aligned}$$

The minimizer of (26) can be found by evaluating the stationary points of $\nabla K_1 + \nabla K_2$ and $\nabla K_1 - \nabla K_2$.

When $r(K) = 1$ then $\tilde{\Upsilon}_1 = s_1 \nabla K_1 = 0 \Leftrightarrow \nabla K_1 = 0$ which implies that $\nabla \tilde{\Upsilon}_1 = \nabla \Upsilon_1 = 0$. Hence $\max \sum_i |K_{iii}| = \max \sum_i K_{iii}$ when $r(K) = 1$.

4.3 Maximization of the sum of squares

Now consider criterion $\Upsilon_2(\mathbf{U}, \mathbf{V}, \mathbf{W})$ defined in (9). Its numerator is not multi-linear in the unknowns anymore, neither the equations defining its stationary points.

Symmetry properties As in section 3.1, we can try to make use of symmetry properties. As before, observe that

$$\mathbf{Q}[\alpha_i - \pi/2] = \mathbf{Q}[\alpha_i] \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{Q}[\alpha_i]$$

In other words, when changing the rotation angles α_i in each of the three modes into $\alpha_i - \pi/2$, (x, y, z) is transformed into $(-1/x, -1/y, -1/z)$, so that (T_{111}, T_{222}) is transformed into $(-T_{222}, T_{111})$. The consequence is that

$$\psi_2(-1/x, -1/y, -1/z) = \psi_2(x, y, z)$$

which allows to reduce the search for stationary points. In fact, if (x_o, y_o, z_o) is a stationary point then so is $(-1/x_o, -1/y_o, -1/z_o)$.

Unfortunately, contrary to the symmetric case, it is not obvious to reduce the degree of the polynomial system by using this symmetry property, even at even orders. Only the domain of search is divided by half.

Algebraic solution When \mathbf{U} , \mathbf{V} and \mathbf{W} are restricted to be plane rotations the functional (9) can be expressed as the rational function

$$\Upsilon_2(x, y, z) = \sum_{i,k,l=0}^2 \frac{\alpha_{i,j,k} x^i y^j z^k}{(1+x^2)(1+y^2)(1+z^2)}, \quad \alpha_{i,j,k} \in \mathbb{R} \quad (27)$$

The stationary points of (27) are the solutions to

$$\nabla \Upsilon_2(x, y, z) = \begin{bmatrix} \frac{\partial \Upsilon_2(x,y,z)}{\partial x} \\ \frac{\partial \Upsilon_2(x,y,z)}{\partial y} \\ \frac{\partial \Upsilon_2(x,y,z)}{\partial z} \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \sum_{i,j,k=0}^2 \mu_{ijk} x^i y^j z^k \\ \sum_{i,j,k=0}^2 \beta_{ijk} x^i y^j z^k \\ \sum_{i,j,k=0}^2 \gamma_{ijk} x^i y^j z^k \end{bmatrix} = \mathbf{0}$$

where $\mu_{ijk}, \beta_{ijk}, \gamma_{ijk} \in \mathbb{R}$. Let $I = \langle p_1, p_2, p_3 \rangle \subseteq \mathbb{R}[x, y, z]$ denote the ideal generated by p_1, p_2 and p_3 and where $\mathbb{R}[x, y, z]$ denotes the set of polynomials in x, y and z with coefficients in \mathbb{R} .

Furthermore let $G = \{g_i\}_{i=1}^s \subset \mathbb{R}[x, y, z]$ be a Gröbner basis of I wrt. the lexicographic ordering $x > y > z$. If $I_{yz} = I \cap \mathbb{R}[y, z]$ and $I_z = I \cap \mathbb{R}[z]$ then according to the elimination theorem $G_{yz} = G \cap \mathbb{R}[y, z]$ and $G_z = G \cap \mathbb{R}[z]$ are gröbner bases for the ideals I_{yz} and I_z respectively.

Since $I = \langle p_1, p_2, p_3 \rangle = \langle g_1, \dots, g_s \rangle$ implies that $V(p_1, p_2, p_3) = V(g_1, \dots, g_s)$, where $V(I) = \{(x, y, z) \in \mathbb{R}^3 \mid f(x, y, z) = 0 \forall f \in I\}$, the problem to solve $p_1 = p_2 = p_3 = 0$ when given the Gröbner basis G of I can be solved by backsubstitution since the new system $g_1 = \dots = g_s = 0$ can be put into a triangular form.

Numerical solutions Due to the complexity of optimizing the functional (9) some suboptimal solutions will be considered next. First an algorithm called ALS1 will be presented. The ALS1 algorithm will estimate \mathbf{U} , \mathbf{V} and \mathbf{W} in the successive manner:

$$(\mathbf{U}^{(i)}, \mathbf{V}^{(i-1)}, \mathbf{W}^{(i-1)}) \rightarrow (\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{W}^{(i-1)}) \rightarrow (\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{W}^{(i)})$$

where the upper index denotes the iteration number.

We have the equivalence

$$\frac{\partial \Upsilon_2(x, y, z)}{\partial x} \Big|_{y=\bar{y}, z=\bar{z}} = 0 \Leftrightarrow \sum_{i=0}^2 \beta_i x^i = 0 \quad , \quad \beta_i \in \mathbb{R}, \quad (28)$$

and hence finding the stationary points is equivalent to solving a second degree polynomial. Due to symmetry of (27) similar expressions are obtained when setting the partial derivative of (27) wrt. y or z equal to zero.

Another possibility is to estimate two out of the three transform matrices simultaneously and this approach will be called ALS2. In this scheme the transform matrices will be estimated in the successive manner

$$(\mathbf{U}^{(i-1)}, \mathbf{V}^{(i)}, \mathbf{W}^{(i)}) \rightarrow (\mathbf{U}^{(i)}, \mathbf{V}^{(i)}, \mathbf{W}^{(i+1)}) \rightarrow (\mathbf{U}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{W}^{(i+1)})$$

Assume that x is fixed, then the candidates for the pair (y, z) can be found via the equivalence

$$\frac{\partial \Upsilon_2(x, y, z)}{\partial y} \Big|_{x=\bar{x}} = 0 \Leftrightarrow \sum_{i,j,k=0}^2 \alpha_{ijk} \tilde{x}^i y^j z^k = \sum_{i=0}^2 a_i(z) y^i = 0 \quad (29)$$

$$\frac{\partial \Upsilon_2(x, y, z)}{\partial z} \Big|_{x=\bar{x}} = 0 \Leftrightarrow \sum_{i,j,k=0}^2 \alpha_{ijk} \tilde{x}^i y^j z^k = \sum_{i=0}^2 b_i(z) y^i = 0 \quad (30)$$

where $a_i(z)$ and $b_i(z)$ are second degree polynomials in z . Applying the method of resultants we end up with the polynomial

$$\text{Res}(\Upsilon_{2_y}, \Upsilon_{2_z}, y) = \sum_{i=0}^8 \gamma_i z^i$$

The roots of this eight degree polynomial can be estimated by an EVD of its corresponding companion matrix. After the roots has been estimated they will be plugged back into (29) and hence estimating the corresponding y is reduced to finding the roots of a second degree polynomial. Again due to symmetry of (27) similar expressions are obtained when fixing y or z .

Other ways to find the roots will be reported in a future report.

5 Computer results

To diagonalize a tensor $\mathbf{T} \in \mathbb{R}^{N \times N \times N}$ with $N > 2$ a cyclic Jacobi-diagonalization approach is taken. In each step a $2 \times 2 \times 2$ -tensor, denoted $\hat{\mathbf{T}}$ is

approximately diagonalized and the 2×2 transform matrices that approximately diagonalizes $\tilde{\mathbf{T}}$ are denoted $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$ and $\hat{\mathbf{W}}$, and they are estimated in an iterative manner as described earlier.

$\tilde{\mathbf{T}}$ is extracted from \mathbf{T} as follows:

$$\tilde{\mathbf{T}}(:, :, 1) = \begin{bmatrix} \mathbf{T}_{p,p,p} & \mathbf{T}_{p,q,p} \\ \mathbf{T}_{q,p,p} & \mathbf{T}_{q,q,p} \end{bmatrix}, \quad \tilde{\mathbf{T}}(:, :, 2) = \begin{bmatrix} \mathbf{T}_{p,p,q} & \mathbf{T}_{p,q,q} \\ \mathbf{T}_{q,p,q} & \mathbf{T}_{q,q,q} \end{bmatrix}$$

Let $\mathbf{K} \in \mathbb{R}^{N \times N \times N}$ be the transformed tensor after each step then

$$\mathbf{K} = \mathbf{T} \bullet_1 \mathbf{U} \bullet_2 \mathbf{V} \bullet_3 \mathbf{W}$$

where

$$\begin{aligned} \mathbf{U} &= \delta_{i,n} \delta_{n,j} + \delta_{i,p} \delta_{p,j} (\hat{\mathbf{U}}_{1,1} - 1) + \delta_{i,q} \delta_{q,j} (\hat{\mathbf{U}}_{2,2} - 1) + \delta_{i,p} \delta_{q,j} \hat{\mathbf{U}}_{1,2} + \delta_{i,q} \delta_{p,j} \hat{\mathbf{U}}_{2,1} \\ \mathbf{V} &= \delta_{i,n} \delta_{n,j} + \delta_{i,p} \delta_{p,j} (\hat{\mathbf{V}}_{1,1} - 1) + \delta_{i,q} \delta_{q,j} (\hat{\mathbf{V}}_{2,2} - 1) + \delta_{i,p} \delta_{q,j} \hat{\mathbf{V}}_{1,2} + \delta_{i,q} \delta_{p,j} \hat{\mathbf{V}}_{2,1} \\ \mathbf{W} &= \delta_{i,n} \delta_{n,j} + \delta_{i,p} \delta_{p,j} (\hat{\mathbf{W}}_{1,1} - 1) + \delta_{i,q} \delta_{q,j} (\hat{\mathbf{W}}_{2,2} - 1) + \delta_{i,p} \delta_{q,j} \hat{\mathbf{W}}_{1,2} + \delta_{i,q} \delta_{p,j} \hat{\mathbf{W}}_{2,1} \end{aligned}$$

It should be noted that this algorithm is monotonic since $\|\mathbf{K}\|^2 = \|\mathbf{T}\|^2$, $\mathbf{K}_{ppp}^2 + \mathbf{K}_{qqq}^2 \geq \mathbf{T}_{ppp}^2 + \mathbf{T}_{qqq}^2$ and $\mathbf{K}_{iii}^2 = \mathbf{T}_{iii}^2$, $i \neq p, q$.

The outline of the cyclic Jacobi diagonalization algorithm can be seen in algorithm 1, where the function $\text{best}_{\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{W}}}(\tilde{\mathbf{T}})$ tries to diagonalize the subtensor based on either the ALS1 algorithm or the ALS2 algorithm.

To see if the cyclic Jacobi diagonalization algorithm is capable of approximately diagonalizing an arbitrary tensor some simulations will be conducted. A measure of how diagonal a tensor is, is the following

$$\gamma = \frac{\sum_{n=1}^N \mathbf{T}_{nnn}^2}{\sum_{i,j,k=1}^N \mathbf{T}_{ijk}^2}$$

In the first simulation, a tensor $\mathbf{T} \in \mathbb{R}^{8 \times 8 \times 8}$ with elements randomly drawn from a uniform distribution is diagonalized. In the first case a tensor were $\mathbf{T}_{ijk} \in U(-100, 100)$ and in the second case a tensor with $\mathbf{T}_{ijk} \in U(0, 100)$ are diagonalized and the results can be seen in figure 1(a) and 1(b) respectively.

In the second simulation the objective is to try to diagonalize the tensor $\mathbf{T} = \sum_{r=1}^8 \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$ where $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i \in U(-100, 100)$ in the first case and $\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i \in U(0, 100)$ in the second case. The results can be seen in figure 2(a) and 2(b) respectively.

Finally a diagonalization of an orthogonally diagonalizable tensor has been carried out where \mathbf{U} , \mathbf{V} and \mathbf{W} consists of orthonormal bases of the subspace

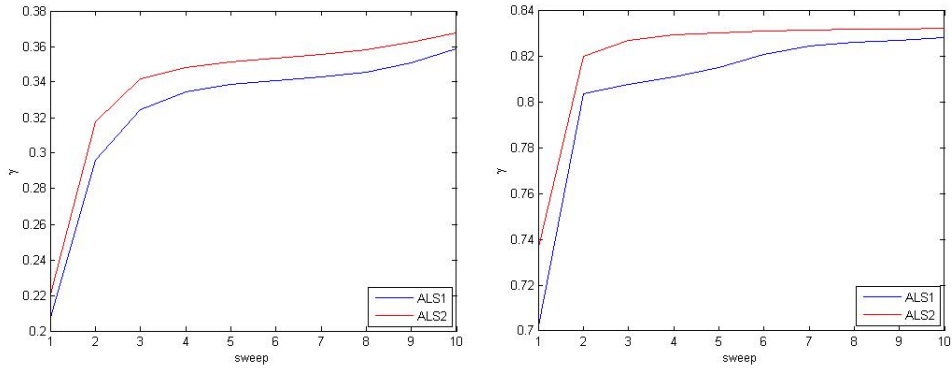
Algorithm 1 Cyclic Jacobi Diagonalization

```

for  $p = 1$  to  $N - 1$  do
  for  $q = p + 1$  to  $N$  do
     $\tilde{\mathbf{T}}(:, :, 1) = \begin{bmatrix} \mathbf{T}_{p,p,p} & \mathbf{T}_{p,q,p} \\ \mathbf{T}_{q,p,p} & \mathbf{T}_{q,q,p} \end{bmatrix}$ ,  $\tilde{\mathbf{T}}(:, :, 2) = \begin{bmatrix} \mathbf{T}_{p,p,q} & \mathbf{T}_{p,q,q} \\ \mathbf{T}_{q,p,q} & \mathbf{T}_{q,q,q} \end{bmatrix}$ 
     $[\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{W}}] = \text{best}_{\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{W}}}(\tilde{\mathbf{T}})$ 
     $\mathbf{U} = \delta_{i,n}\delta_{n,j} + \delta_{i,p}\delta_{p,j}(\hat{\mathbf{U}}_{1,1} - 1) + \delta_{i,q}\delta_{q,j}(\hat{\mathbf{U}}_{2,2} - 1) + \delta_{i,p}\delta_{q,j}\hat{\mathbf{U}}_{1,2} +$ 
 $\delta_{i,q}\delta_{p,j}\hat{\mathbf{U}}_{2,1}$ 
     $\mathbf{V} = \delta_{i,n}\delta_{n,j} + \delta_{i,p}\delta_{p,j}(\hat{\mathbf{V}}_{1,1} - 1) + \delta_{i,q}\delta_{q,j}(\hat{\mathbf{V}}_{2,2} - 1) + \delta_{i,p}\delta_{q,j}\hat{\mathbf{V}}_{1,2} +$ 
 $\delta_{i,q}\delta_{p,j}\hat{\mathbf{V}}_{2,1}$ 
     $\mathbf{W} = \delta_{i,n}\delta_{n,j} + \delta_{i,p}\delta_{p,j}(\hat{\mathbf{W}}_{1,1} - 1) + \delta_{i,q}\delta_{q,j}(\hat{\mathbf{W}}_{2,2} - 1) + \delta_{i,p}\delta_{q,j}\hat{\mathbf{W}}_{1,2} +$ 
 $\delta_{i,q}\delta_{p,j}\hat{\mathbf{W}}_{2,1}$ 
     $\mathbf{T} = \mathbf{T} \underset{1}{\bullet} \mathbf{U} \underset{2}{\bullet} \mathbf{V} \underset{3}{\bullet} \mathbf{W}$ 
     $\mathbf{X} = \mathbf{U}\mathbf{X}$ 
     $\mathbf{Y} = \mathbf{V}\mathbf{Y}$ 
     $\mathbf{Z} = \mathbf{W}\mathbf{Z}$ 
  end for
end for

```

spanned by the columns of random matrices contained in $\mathbb{R}^{8 \times 8}$ and their entries are elements in $U(-100, 100)$ in the first case and $U(0, 100)$ in the second case. The results can be seen in figure 3(a) and 3(b) respectively.



(a) $\mathbf{T}_{ijk} \in U(-100, 100)$

(b) $\mathbf{T}_{ijk} \in U(0, 100)$

Figure 1: Diagonalization of a random tensor.

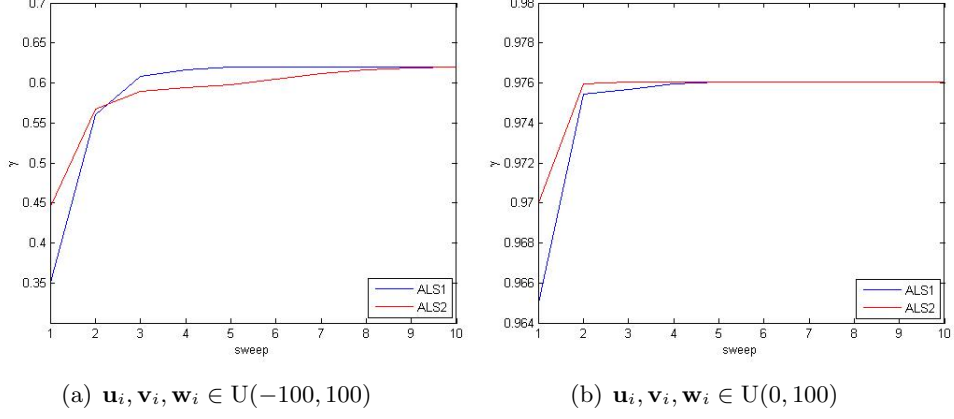


Figure 2: Diagonalization of the random tensor of size $8 \times 8 \times 8$, $\mathbf{T} = \sum_{r=1}^8 \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$.

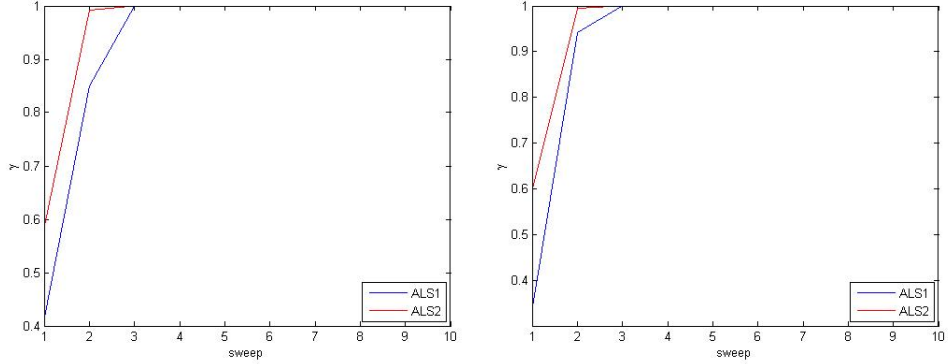


Figure 3: Diagonalization of an orthogonally diagonalizable tensor.

Another –equivalent– diagonality measure is the following

$$\text{off}(\mathbf{T}) = \frac{\sum_{i,j,k=1}^N \mathbf{T}_{ijk}^2 - \sum_{n=1}^N \mathbf{T}_{nnn}^2}{\sum_{i,j,k=1}^N \mathbf{T}_{ijk}^2}.$$

Furthermore the performance of the algorithms will be measured on the

complexity of the algorithms which is measured by the number of multiplications used by the given algorithm.

To compare the different $2 \times 2 \times 2$ -tensor diagonalization algorithms, the performance will be measured on tensors in $\mathbb{R}^{5 \times 5 \times 5} \ni \mathbf{T} = \mathbf{D} + \mathbf{E}$, where \mathbf{D} is an orthogonally diagonalizable tensor and $\mathbf{E} \in \mathbb{R}^{5 \times 5 \times 5}$ is a random tensor where $\mathbf{E}_{ijk} \in \rho U(-100, 100)$, and where the real positive parameter ρ controls the noise level. \mathbf{D} is generated as $\mathbf{D} = \sum_{r=1}^5 \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$, where \mathbf{U}, \mathbf{V} and \mathbf{W} consists of the orthonormal bases of the subspaces spanned by the columns of random matrices contained in $\mathbb{R}^{5 \times 5}$ and which entries are randomly drawn elements in $U(-100, 100)$.

In the first simulation the algorithms will sweep 5 times. The performance of the algorithm as a function of $\frac{\|\mathbf{E}\|^2}{\|\mathbf{D}\|^2}$ can be seen at figure 4 where the plotted curves are the average values over 5 runs.

ALS1 refers to the sum of squares tensor diagonalization algorithm where only one plane rotation are estimated at the time whereas the ALS2res and ALS2heu refers to the sum of squares ALS2 tensor diagonalization algorithms where the resultant and the heuristic solution has been used respectively and in both case two plane rotations are estimated at the time. Similary TraceRes and TraceHeu refers to the absolute trace tensor diagonalization algorithms where the resultant and the heuristic solution has been used respectively.

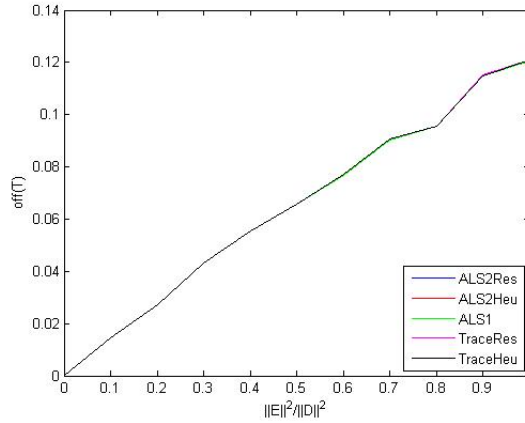


Figure 4: The average diagonalization of random tensors $\mathbf{T} = \mathbf{D} + \mathbf{E}$ as a function of $\frac{\|\mathbf{E}\|^2}{\|\mathbf{D}\|^2}$.

The computational complexity of various algorithms will be studied in a

forthcoming report, as well as stationary values of the Jacobi CoM algorithm in the unsymmetric case (the symmetric case has already reported in [3]).

References

- [1] J. D. CARROLL, J. J. CHANG, “Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition”, *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sept. 1970.
- [2] P. COMON, “Independent Component Analysis, a new concept ?”, *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, Apr. 1994, Special issue on Higher-Order Statistics.
- [3] P. COMON, “Tensor diagonalization, a useful tool in signal processing”, in *IFAC-SYSID, 10th IFAC Symposium on System Identification*, M. Blanke, T. Soderstrom, Eds., Copenhagen, Denmark, July 4-6 1994, vol. 1, pp. 77–82, invited session.
- [4] P. COMON, “From source separation to blind equalization, contrast-based approaches”, in *Int. Conf. on Image and Signal Processing (ICISP’01)*, Agadir, Morocco, May 3-5, 2001, invited plenary.
- [5] P. COMON, “Tensor decompositions”, in *Mathematics in Signal Processing V*, J. G. McWhirter, I. K. Proudler, Eds., pp. 1–24. Clarendon Press, Oxford, UK, 2002.
- [6] P. COMON, G. GOLUB, L-H. LIM, B. MOURRAIN, “Symmetric tensors and symmetric tensor rank”, *SIAM J. matrix Ana. Appl.*, 2007, to appear.
- [7] P. COMON, B. MOURRAIN, “Decomposition of quantics in sums of powers of linear forms”, *Signal Processing, Elsevier*, vol. 53, no. 2, pp. 93–107, Sept. 1996, special issue on High-Order Statistics.
- [8] P. COMON, B. MOURRAIN, L-H. LIM, G. GOLUB, “Genericity and rank deficiency of high order symmetric tensors”, in *ICASSP’06*, Toulouse, May 14-19 2006.
- [9] P. COMON, J. ten BERGE, “Generic and typical ranks of three-way arrays”, Research Report ISRN I3S/RR-2006-29-FR, I3S, Sophia-Antipolis, France, Sept. 4 2006, submitted for publication.

- [10] D. COX, J. LITTLE, D. O'SHEA, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Undergraduate Texts in Mathematics. Springer Verlag, New York, 1992, 2nd ed. in 1996.
- [11] L. de LATHAUWER, B. de MOOR, J. VANDEWALLE, "A multilinear singular value decomposition", *SIAM Jour. Matrix Ana. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.
- [12] L. de LATHAUWER, B. de MOOR, J. VANDEWALLE, "Independent Component Analysis and (simultaneous) third-order tensor diagonalization", *IEEE Trans. Sig. Proc.*, pp. 2262–2271, Oct. 2001.
- [13] R. A. HARSHMAN, "Foundations of the Parafac procedure: Models and conditions for an explanatory multimodal factor analysis", *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970, <http://publish.uwo.ca/~harshman>.
- [14] J. L. LACOUME, P. O. AMBLARD, P. COMON, *Statistiques d'ordre supérieur pour le traitement du signal*, Collection Sciences de l'Ingénieur. Masson, 1997, freely downloadable from <http://www.i3s.unice.fr/~comon/livreSOS.html>.
- [15] P. McCULLAGH, *Tensor Methods in Statistics*, Monographs on Statistics and Applied Probability. Chapman and Hall, 1987.
- [16] E. MOREAU, O. MACCHI, "High order contrasts for self-adaptive source separation", *Int. J. of Adaptive Control and Signal Processing*, vol. 10, no. 1, pp. 19–46, Jan. 1996.
- [17] B. MOURRAIN, P. TREBUCHET, "Solving projective complete intersection faster", in *Proc. Intern. Symp. on Symbolic and Algebraic Computation*, C. Traverso, Ed. 2000, pp. 231–238, New York, ACM Press.
- [18] C. L. NIKIAS, A. P. PETROPULU, *Higher-Order Spectra Analysis*, Signal Processing Series. Prentice-Hall, Englewood Cliffs, 1993.
- [19] N. D. SIDIROPOULOS, R. BRO, G. B. GIANNAKIS, "Parallel factor analysis in sensor array processing", *IEEE Trans. Sig. Proc.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.
- [20] A. SMILDE, R. BRO, P. GELADI, *Multi-Way Analysis*, Wiley, 2004.