

Efficient Orthogonal Tensor Decomposition, with an Application to Latent Variable Model Learning

Franz J. Király *

Abstract

Decomposing tensors into orthogonal factors is a well-known task in statistics, machine learning, and signal processing. We study orthogonal outer product decompositions where the factors in the summands in the decomposition are required to be orthogonal across summands, by relating this orthogonal decomposition to the singular value decompositions of the flattenings. We show that it is a non-trivial assumption for a tensor to have such an orthogonal decomposition, and we show that it is unique (up to natural symmetries) in case it exists, in which case we also demonstrate how it can be efficiently and reliably obtained by a sequence of singular value decompositions. We demonstrate how the factoring algorithm can be applied for parameter identification in latent variable and mixture models.

1. Introduction

Decomposing a tensors into its components, and determining the number of those (= the rank) is a multidimensional generalization of the singular value decomposition and the matrix rank, and a reoccurring task in all practical sciences, appearing many times under different names; first discovered by Hitchcock [9] and then re-discovered under names such as PARAFAC [8] or CANDECOMP [4], it has been applied in many fields such as chemometrics, psychometrics, and signal processing [3, 14, 16]. An extensive survey of many applications can be found in [6, 15].

Recently, motivated by real world applications, orthogonality constraints on the decomposition have been studied in the literature, such as the orthogonal rank decomposition and the combinatorial orthogonal rank decomposition, which can be traced back to [7, 12], and the orthogonal decomposition in [13] and [10], the latter of which occurs for example in the identification of latent variable models from empirical moments, and several other statistical estimation tasks, see [2] for a survey. The orthogonality constraints imposed in these two branches of literature are not the same, as [7, 12] imposes summand-wise orthogonality, while in [2, 10, 13], factor-wise orthogonality can be deduced from the model constraints. In [13], a Jacobi-like and heuristic algorithm was described to obtain a close orthogonal decomposition via Jacobi angle optimization for general tensors; in [2], the authors describe a second order fixed point method for obtaining the decomposition.

In [11, 17], hierarchical tensor decomposition models are discussed in the context of latent tree graphical models, and algorithms for the identification of this decomposition are described. While this

*Department of Statistical Science, University College London; and MFO; f.kiraly@ucl.ac.uk

is not explicitly done in the language of orthogonal tensor decompositions, the idea of using flattenings is similar to the one presented, and, in the specific context of tree models, a specific instance orthogonal tensor decomposition, as described in [2].

In this paper, we study the orthogonal decomposition model, as it occurs in [2, 10], namely with factor-wise orthogonality constraints. We show that this kind of decomposition can be directly transformed to a set of singular value decompositions, both theoretically and practically. We give identifiability results for this kind of orthogonal decomposition, showing that it is unique¹ in case of existence, and we provide algorithms to obtain the orthogonal decomposition, by reducing it to a sequence of singular value decompositions. We apply these algorithms to a latent variable identification problem which was discussed in [2, 10], reducing it to a series of eigenvalue problems. In particular, by performing the reduction to singular value decomposition, we show that all existing theory on the singular value decomposition, concerning theoretical issues as well as numerical and algorithmical ones, can be readily applied to the orthogonal decomposition problem.

2. Theoretical Background

2.1. Tensors

2.1.1. Definition of a Tensor While tensors are common objects, their notation diverges throughout the literature. For ease of reading, we provide the basic definitions.

Definition 2.1. A real tensor of size $(n_1 \times n_2 \times \dots \times n_d)$ and of degree d is an element of the set

$$\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} = \left\{ \begin{array}{c} (a_{i_1 \dots i_d})_{1 \leq i_1 \leq n_1} \\ \vdots \\ 1 \leq i_d \leq n_d \end{array} \right\}.$$

If $n_1 = n_2 = \dots = n_d$, we also write $\mathbb{R}^{n \times d} := \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$. ■

2.1.2. Linear Transformation Let us introduce a useful shorthand notation for linearly transforming tensors.

Definition 2.2. Let $A \in \mathbb{R}^{m \times n}$ be a matrix. For a tensor $T \in \mathbb{R}^{n \times (\times d)}$, we denote by $A \circ T$ the application of A to T along all tensor dimensions, that is, the tensor $A \circ T \in \mathbb{R}^{m \times (\times d)}$ defined as

$$(A \circ T)_{i_1 \dots i_d} = \sum_{j_1=1}^n \dots \sum_{j_d=1}^n A_{i_1 j_1} \cdot \dots \cdot A_{i_d j_d} \cdot T_{j_1 \dots j_d}.$$

Remark 2.3 — For $T \in \mathbb{R}^{n \times (\times d)}$ and $A \in \mathbb{R}^{m \times n}, A' \in \mathbb{R}^{m' \times m}$, note that ■

$$A' \circ (A \circ T) = (A' \cdot A) \circ T.$$

◇

¹up to natural symmetries

2.1.3. Flattening A flattening of a tensor is the tensor obtained from regarding different indices as one index.

Definition 2.4. Denote by $[k] = \{1, 2, \dots, k\}$. A surjective map $\sigma : [d] \rightarrow [\tilde{d}]$ is called d -to- \tilde{d} flattening map. ■

Definition 2.5. Let $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$ be a tensor, and let σ be a d -to- \tilde{d} flattening map. Then, the σ -flattening of T is the degree \tilde{d} tensor $\sigma \dashv T \in \mathbb{R}^{\tilde{n}_1 \times \dots \times \tilde{n}_{\tilde{d}}}$, with $\tilde{n}_k = \prod_{\ell \in \sigma^{-1}(k)} n_\ell$, defined as

$$(\sigma \dashv T)_{j_1 \dots j_{\tilde{d}}} := T_{i_1 \dots i_d}, \text{ where } j_k = (i_\ell : \ell \in \sigma^{-1}(k)).$$

Conversely, if $\tilde{T} = \sigma \dashv T$, then we write $T = \sigma \vdash \tilde{T}$ and call T the unflattening of \tilde{T} . ■

Note that the indices of $\sigma \dashv T$ are, as defined, tuples of indices of T ; however, this does not contradict the definition of tensor since $[n_1] \times [n_2] \times \dots \times [n_k]$ can be bijectively mapped onto $\left[\prod_{i=1}^k n_i \right]$. It is convenient to choose the lexicographical ordering for the bijection, but it is mathematically not necessary to fix any such bijection.

For unflattening, if only \tilde{T} and σ are given, it is not clear what $\sigma \vdash \tilde{T}$ should be without further specification, since the same unflattening can arise from different tensors even if σ is fixed. Therefore, we will use it only in the context where a given flattening is being reversed, or partially reversed, therefore making the unflattening well-defined.

Example 2.6. Let $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ be a tensor, let $\sigma : 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 2$. The σ -flattening of T is a $(n_1 \times n_2 n_3)$ -matrix $\tilde{T} := \sigma \dashv T$. The columns of $\sigma \dashv \tilde{T}$ are all the $n_2 n_3$ sub- $(n_1 \times 1 \times 1)$ -tensors of T where second and third index are fixed. The columns of $\sigma \dashv T$ are indexed by the pairs (k, ℓ) , or, alternatively, by bijection, by the lexicographical index number $(k-1) \cdot n_2 + \ell$. Taking any $(n'_1 \times n_2 n_3)$ -submatrix of \tilde{T} , we can unflatten to obtain a $(n'_1 \times n_2 \times n_3)$ -tensor $\sigma \vdash \tilde{T}$. ■

2.1.4. Outer Product Furthermore, we introduce notation for creating tensors of higher order out of tensors of lower order:

Definition 2.7. Let $v^{(1)} \in \mathbb{R}^{n_1}, \dots, v^{(d)} \in \mathbb{R}^{n_d}$. The *outer product* of the $v^{(k)}$ is the tensor $v^{(1)} \otimes \dots \otimes v^{(d)} \in \mathbb{R}^{n_1 \times \dots \times n_d}$ defined by

$$(v^{(1)} \otimes \dots \otimes v^{(d)})_{i_1 \dots i_d} := \prod_{k=1}^d v_{i_k}^{(k)}.$$

In case that $v = v^{(1)} = \dots = v^{(d)}$, we also write $v^{\otimes d} := v^{(1)} \otimes \dots \otimes v^{(d)}$.

Similarly, if $A \in \mathbb{R}^{n_1 \times \dots \times n_c}$ and $B \in \mathbb{R}^{n_{c+1} \times \dots \times n_d}$ are tensors, the outer product of A and B is the tensor $A \otimes B \in \mathbb{R}^{n_1 \times \dots \times n_d}$ defined as

$$(A \otimes B)_{i_1 \dots i_d} := \prod_{k=1}^c A_{i_1 \dots i_c}^{(k)} \cdot \prod_{k=c+1}^d B_{i_{c+1} \dots i_d}^{(k)}.$$

Outer products of several tensors $A_1 \otimes \dots \otimes A_k$ by induction on k , namely:

$$A_1 \otimes \dots \otimes A_k := (A_1 \otimes \dots \otimes A_{k-1}) \otimes A_k.$$

A useful calculation rule for linear transformation is the following: ■

Lemma 2.8. Let $A \in \mathbb{R}^{n \times d_1}$ and $B \in \mathbb{R}^{n \times d_2}$, let $P \in \mathbb{R}^{m \times n}$. Then,

$$P \circ (A \otimes B) = (P \circ A) \otimes (P \circ B).$$

Similarly, if $v \in \mathbb{R}^n$, then $P \circ (v^{\otimes d}) = (P \circ v)^{\otimes d}$.

Outer products are also compatible with flattenings:

Lemma 2.9. Let $A \in \mathbb{R}^{n_1 \times \dots \times n_c}$ and $B \in \mathbb{R}^{n_{c+1} \times \dots \times n_d}$. Let τ be a d -to- k -flattening, let σ_1 be the restriction of τ to $[c]$, and let σ_2 be the $(d-c)$ -to- \tilde{k} -flattening defined by $\sigma_2(i) := \tau(c+i)$. Then,

$$\tau \dashv (A \otimes B) = (\sigma_1 \dashv A) \otimes (\sigma_2 \dashv B).$$

2.2. Orthogonality and Duality

We briefly review the notions of scalar product and some results, which can also be found in [12] in slightly different formulation and slightly less generality.

Definition 2.10. A scalar product is defined on $\mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$ by

$$\begin{aligned} \langle \cdot, \cdot \rangle : \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} \times \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} &\longrightarrow \mathbb{R} \\ (A, B) &\mapsto \sum_{i_1=1}^{n_1} \dots \sum_{i_d=1}^{n_d} A_{i_1 \dots i_d} \cdot B_{i_1 \dots i_d} \end{aligned}$$

As usual, $A, B \in \mathbb{R}^{n_1 \times \dots \times n_d}$ are called orthogonal to each other if $\langle A, B \rangle = 0$, and A is called normal if $\langle A, A \rangle = 1$. A set $A_1, \dots, A_r \in \mathbb{R}^{n_1 \times \dots \times n_d}$ is called orthonormal if $\langle A_i, A_j \rangle = \delta_{ij}$, where δ_{ij} is the Kronecker-delta. \blacksquare

By identification of $\mathbb{R}^{n_1 \times \dots \times n_d}$ with \mathbb{R}^N , where $N = \prod_{i=1}^d n_i$, the scalar product on tensors inherits all properties of the real scalar product.

Remark 2.11 — It is seen by checking definitions that the scalar product on matrices is identical to the trace product, i.e., $\langle A, B \rangle = \text{Tr}(A^T B)$ for $A, B \in \mathbb{R}^{m \times n}$. \diamond

An important property of the scalar product is compatibility with flattenings:

Lemma 2.12. Let $T_1, T_2 \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, let σ be a d -to- \tilde{d} flattening map. Then,

$$\langle T_1, T_2 \rangle = \langle \sigma \dashv T_1, \sigma \dashv T_2 \rangle.$$

In particular, T_1 and T_2 are orthogonal to each other if and only if $\sigma \dashv T_1$ and $\sigma \dashv T_2$ are.

Proof. A flattening is a bijection on the set of entries, therefore the result of the entry-wise scalar product is not changed by flattening. \square

Proposition 2.13. Let $A_1^{(j)}, A_2^{(j)} \in \mathbb{R}^{n_1^{(j)} \times \dots \times n_{c_j}^{(j)}}$, for $j = 1, \dots, k$. Then,

$$\left\langle A_1^{(1)} \otimes \dots \otimes A_1^{(k)}, A_2^{(1)} \otimes \dots \otimes A_2^{(k)} \right\rangle = \prod_{j=1}^k \left\langle A_1^{(j)}, A_2^{(j)} \right\rangle.$$

In particular, if there exists j such that $A_1^{(j)}, A_2^{(j)}$ are orthogonal to each other, then the outer products $A_1^{(1)} \otimes \dots \otimes A_1^{(k)}$ and $A_2^{(1)} \otimes \dots \otimes A_2^{(k)}$ are orthogonal to each other.

Proof. By performing induction on k , it suffices to prove the statement for $k = 2$: Let $A_1, A_2 \in \mathbb{R}^{n_1 \times \dots \times n_c}$ and $B_1, B_2 \in \mathbb{R}^{n_{c+1} \times \dots \times n_d}$. Then,

$$\langle A_1 \otimes B_1, A_2 \otimes B_2 \rangle = \langle A_1, B_1 \rangle \cdot \langle A_2, B_2 \rangle.$$

We proceed to prove this statement. Let σ_1 be the c -to-1-flattening, let σ_2 be the $(d-c)$ -to-1-flattening. Let $v_i = \sigma_1 \dashv A_i$, and $w_i = \sigma_2 \dashv B_i$ for $i = 1, 2$. By Lemma 2.12, it holds that

$$\langle A_i, B_i \rangle = \langle v_i, w_i \rangle \quad \text{for } i = 1, 2.$$

Let τ be the d -to-2-flattening defined by $\tau : \{1, \dots, c\} \mapsto \{1\}, \{c+1, \dots, d\} \mapsto \{2\}$. Let $C_i = \tau \dashv (A_i \otimes B_i)$. By Lemma 2.12, it holds that

$$\langle A_1 \otimes B_1, A_2 \otimes B_2 \rangle = \langle C_1, C_2 \rangle.$$

By Lemma 2.9, it holds that

$$\langle C_1, C_2 \rangle = \langle v_1 \otimes w_1, v_2 \otimes w_2 \rangle.$$

Using that scalar product on tensors is the trace product (see 2.11), we obtain

$$\langle v_1 \otimes w_1, v_2 \otimes w_2 \rangle = \text{Tr}(v_1 w_1^\top w_2 v_2^\top).$$

The cyclic property of the trace product for matrices yields

$$\text{Tr}(v_1 w_1^\top w_2 v_2^\top) = \text{Tr}(w_1^\top w_2 v_2^\top v_1) = w_1^\top w_2 v_2^\top v_1 = \langle v_1, v_2 \rangle \cdot \langle w_1, w_2 \rangle.$$

All equalities put together yield the claim. □

Corollary 2.14. *Let $\mu_1, \mu_2 \in \mathbb{R}^n$, and $d \in \mathbb{N}$, such that $\langle \mu_1, \mu_2 \rangle = 0$. Then,*

$$\langle \mu_1^{\otimes d}, \mu_2^{\otimes d} \rangle = 0.$$

Definition 2.15. Let $T \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, let $[d] = S_1 \cup S_2 \cup \dots \cup S_k$ be a partition. A decomposition

$$T = \sum_{i=1}^r w_i \cdot A_i^{(1)} \otimes \dots \otimes A_i^{(k)}$$

with $w_i \in \mathbb{R}$, and $A_i^{(j)} \in \mathbb{R}^{\times_{\ell \in S_j} n_\ell}$, such that the set of $A_i^{(j)}$ with fixed j is orthonormal, is called *rank- r orthogonal atomic decomposition* of T , with signature (S_1, \dots, S_k) . If $k = d$ and $S_i = \{i\}$, then the decomposition is called *orthogonal CP-decomposition*. ■

An orthogonal atomic decomposition does not need to exist necessarily. However, if it does, it is compatible with respect to flattenings, as Proposition 2.17 will show. We introduce notation for a more concise statement of the compatibility first:

Definition 2.16. Let (S_1, \dots, S_k) be a partition of $[d]$. We say a d -to- \vec{d} -flattening σ is *compatible* with the partition (S_1, \dots, S_k) , if it holds that $\{i, j\} \in S_\ell$ for some ℓ implies $\sigma(i) = \sigma(j)$. We say that σ is *strictly compatible* with the partition (S_1, \dots, S_k) , if it holds that $\{i, j\} \in S_\ell$ for some ℓ if and only if $\sigma(i) = \sigma(j)$. ■

Proposition 2.17. Let $T \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$. Let

$$T = \sum_{i=1}^r w_i \cdot A_i^{(1)} \otimes \dots \otimes A_i^{(k)}$$

be an orthogonal atomic decomposition with signature (S_1, \dots, S_k) , let σ be compatible with the signature. Then,

$$T = \sum_{i=1}^r w_i \cdot B_i^{(1)} \otimes \dots \otimes B_i^{(d)}, \quad \text{where } B_i^{(1)} = \sigma \dashv \left(\bigotimes_{j \in \sigma^{-1}(i)} B_j^{(1)} \right),$$

is an orthogonal atomic decomposition of $(\sigma \dashv T)$. In particular, if σ is strictly compatible with the signature, then the decomposition is also an orthogonal CP-decomposition.

Proof. This is a direct consequence of Lemma 2.12, checking compatibility of scalar product and orthogonality with the flattening at each of the sets of indices S_i . \square

2.3. Identifiability of the Orthogonal Atomic Decomposition

The orthogonal decomposition, as given in Definition 2.15, does not need to exist for a tensor, nor does it need to be unique. We will show that due to the compatibility with flattenings, if it exists, it is unique, if the rank is chosen minimal.

The main ingredient, besides flattenings, is uniqueness of singular value decomposition [18], a classical result, which we state in a convenient form:

Theorem 1. Let $A \in \mathbb{R}^{m \times n}$, let $r = \text{rank} A$. Then, there is a singular value decomposition (= orthogonal CP-decomposition)

$$A = \sum_{i=1}^r w_i \cdot u_i \cdot v_i^\top \quad \text{with } u_i \in \mathbb{R}^m, v_i \in \mathbb{R}^n, w_i \in \mathbb{R}$$

such that the u_i are orthonormal, and the v_j are orthonormal. In particular, there is no singular value decomposition of rank strictly smaller than r . Moreover, the singular value decomposition of A is unique, up to:

- (a) the sequence of summation, i.e., up to arbitrary permutation of the indices $i = 1, \dots, r$
- (b) the choice of sign of w_i, u_i, v_i , i.e., up to changing the sign in any two of w_i, u_i, v_i for fixed i
- (c) unitary transformations of the span of u_i, u_j or v_i, v_j such that $|w_i| = |w_j|$

Condition (c) includes (b) as a special case, and (c) can be removed as a condition if no two distinct w_i, w_j have the same absolute value.

Theorem 2. Let $T \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, and assume that T has an orthogonal atomic decomposition

$$T = \sum_{i=1}^r w_i \cdot A_i^{(1)} \otimes \dots \otimes A_i^{(k)}$$

of signature (S_1, \dots, S_k) , such that $w_i \neq 0$ for all i . Then:

- (i) Denote $N_j = \prod_{i \in S_j} n_i$ for $j = 1, \dots, k$. Then, $r \leq N_j$ for all j .
- (ii) There is no orthogonal atomic decomposition of T with signature (S_1, \dots, S_k) , and of rank strictly smaller than r .
- (iii) The orthogonal atomic decomposition of T of rank r is unique, up to:
 - (a) the sequence of summation, i.e., up to arbitrary permutation of the indices $i = 1, \dots, n$
 - (b) the choice of sign of $w_i, A_i^{(k)}$, i.e., up to changing the sign in any two of w_i and the $A_i^{(k)}$ for fixed i and arbitrary k
 - (c) transformations of factors $A_i^{(k)}, A_j^{(k)}$, and their respective tensor products, such that $|w_i| = |w_j|$, which induce unitary transformations in all flattenings compatible with the signature (S_1, \dots, S_k) .

Condition (c) includes (b) as a special case, and (c) can be removed as a condition if no two distinct w_i, w_j have the same absolute value.

Proof. Fix some arbitrary j . Consider the d -to-2-flattening $\sigma : S_j \mapsto \{1\}, S_i \mapsto \{2\}$ for $i \neq j$, note that σ is compatible with the signature. Let $m = N_j, n = \prod_{i \neq j} N_i$, and $A = \sigma \dashv T$. Note that A is a $(m \times n)$ -matrix. Let

$$T = \sum_{i=1}^r w_i \cdot A_i^{(1)} \otimes \dots \otimes A_i^{(k)}$$

be the orthogonal atomic decomposition of T , and let $u_i = \sigma \dashv A_i^{(j)}$, and $v_i = \sigma \dashv \bigotimes_{k \neq j} A_i^{(k)}$ for all i . Note that u_i is an m -vector, and v_i is an n -vector. By Proposition 2.17,

$$A = \sum_{i=1}^r w_i \cdot u_i \cdot v_i^\top$$

is a singular value decomposition of A .

- (i) In particular, the u_i are a system of r orthonormal vectors in \mathbb{R}^m . Therefore, $r \leq m = N_j$. Since j was arbitrary, statement (i) follows.
- (ii) Since the w_i are non-zero, it holds that $\text{rank} A = r$. Would there be an orthogonal atomic decomposition of T with signature (S_1, \dots, S_k) of rank strictly smaller than r , there would be a singular value decomposition of A of rank strictly smaller than r , contradicting Proposition 2.17.
- (iii) Observe that the flattening by σ induces a bijection between the orthogonal atomic decompositions of T , of rank r , and the singular value decompositions of A , of rank r . The statement in (iii) then follows directly from the uniqueness asserted in Proposition 2.17 for the singular value decomposition of A . \square

Again, we would like to stress that the present orthogonal decomposition model is different from the one in [12]; ours being factor-wise orthogonal between different summands, while the orthogonal rank decomposition in [12] being summand-wise orthogonal, and the combinatorial orthogonal rank decomposition enforcing orthogonality of factors in the same summand. Therefore, Theorem 2 does not contradict Lemma 3.5 in [12].

Another result which seems to be folklore, but not available in the literature, is that it is a strong restriction for a tensor to assume that it has an orthogonal decomposition. Since it is almost implied by the identifiability Theorem 2, we state a quantitative version of this:

Proposition 2.18. *The set of tensors $T \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, with $d \geq 3$, and $n_j \geq 2$ for all j , for which T has an orthogonal CP-decomposition, is a Lebesgue zero set.*

Proof. The CP-decomposition can be viewed as an algebraic map

$$\begin{aligned} \phi : \mathbb{R} \times \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_d} &\rightarrow \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} \\ (w_i, v_i^{(j)}) &\mapsto \sum_{i=1}^r w_i \cdot v_i^{(1)} \otimes \dots \otimes v_i^{(d)}. \end{aligned}$$

Since the left hand side is an irreducible variety, the image of the map ϕ also is. The orthogonal CP-decompositions form an algebraic subset of the left hand side. Therefore the state follows from the fact that ϕ is not surjective. This follows from a degree of freedom resp. dimension count. One has

$$\begin{aligned} D_1 &:= \dim \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d} = \prod_{i=1}^d n_i, \quad \text{and} \\ D_2 &:= \dim(\mathbb{R} \times \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_d})^r = r \cdot (n_1 + \dots + n_d + 1). \end{aligned}$$

Theorem 2 (i) implies

$$D_2 \leq n_j \cdot (n_1 + \dots + n_k + 1).$$

An explicit computation shows that $D_1 \gtrsim D_2$, which proves the statement.

The proof above can be rephrased in terms of the CP-rank (see [5] for an introduction), can be obtained by observing that the generic CP-rank of the tensors in questions must be strictly larger than $\min(n_1, \dots, n_k)$, then proceeding again by arguing that the algebraic set of tensors with orthogonal CP-decompositions must be a proper subset of all tensors with that format, thus a Lebesgue zero set. \square

Proposition 2.18 can be extended to orthogonal atomic decompositions with signature (S_1, \dots, S_k) , $k \geq 3$, by considering suitable unflattenings.

2.4. Tensors and Moments

We briefly show how tensors relate to moments of multivariate real random variables:

Definition 2.19. Let X be a real n -dimensional random variable. Then, define:

$$\begin{aligned} \text{the characteristic function of } X \text{ as} & \quad \varphi_X(\tau) := \mathbb{E} [\exp(i\tau X)], \\ \text{the moment generating function of } X \text{ as} & \quad \chi_X(\tau) := \log \mathbb{E} [\exp(i\tau X)], \end{aligned}$$

where $\tau \in \mathbb{R}^{1 \times n}$ is a formal vector of variables. The d -th *moment* (or moment tensor) $\mathbf{M}_d(X) \in \mathbb{R}^{n^{(\times d)}}$ of X , and the d -th *cumulant* (or cumulant tensor) $\kappa_d(X) \in \mathbb{R}^{n^{(\times d)}}$ of X are defined² as the coefficients

²in case of convergence

in the multivariate Taylor expansions

$$\begin{aligned}\varphi_X(\tau) &= \sum_{d=1}^{\infty} (i\tau) \circ \frac{\mathbf{M}_d(X)}{d!}, \\ \chi_X(\tau) &= \sum_{d=1}^{\infty} (i\tau) \circ \frac{\kappa_d(X)}{d!}.\end{aligned}$$

■

In the following, we will always assume that the moments and cumulants in question exist.

The moments and cumulants of a linearly transformed random variable are the multilinearly transformed moments.

Proposition 2.20. *Let X be a real n -dimensional random variable and let $A \in \mathbb{R}^{m \times n}$. Then,*

$$\begin{aligned}\mathbf{M}_d(A \cdot X) &= A \circ \mathbf{M}_d(X), \\ \kappa_d(A \cdot X) &= A \circ \kappa_d(X).\end{aligned}$$

Proof. We prove the statement for cumulants, the proof for moments is completely analogous. For the cumulant generating functions χ_X of X and $\chi_{A \cdot X}$ of $A \cdot X$, it holds that

$$\begin{aligned}\chi_{A \cdot X}(\tau) &= \mathbb{E} [\exp(i\tau \cdot A \cdot X)] \\ &= \mathbb{E} [\exp(i(\tau \cdot A) \cdot X)] \\ &= \sum_{d=1}^{\infty} (i\tau) \circ \left(A \circ \frac{\mathbf{M}_d(X)}{d!} \right).\end{aligned}$$

The last equality follows from the definition of $\chi_X(\tau)$. But by definition, it also holds that

$$\chi_{A \cdot X}(\tau) = \sum_{d=1}^{\infty} (i\tau) \circ \frac{\mathbf{M}_d(A \cdot X)}{d!},$$

therefore the statement follows from comparing coefficient tensors. □

3. Relation to Mixture Models

3.1. The Estimation Problem

Throughout the paper, we will consider the following independent rank 1 mixture model:

Generative Model: X_1, \dots, X_r are independent, \mathbb{R}^n -valued random variables, with $r \leq n$, and probability/mass density functions $X_i \sim p_i$. Let $w_1, \dots, w_r \in \mathbb{R}$ be arbitrary such that $\sum_{i=1}^r w_i = 1$, and let $Y \sim \sum_{i=1}^r w_i p_i$ be the corresponding mixture of the X_i . Assume that there are $\mu_1, \dots, \mu_r \in \mathbb{R}^n$ with $\|\mu_i\|_2 = 1$, and random variables $Z_i \in \mathbb{R}$, such that $X_i = \mu_i \cdot Z_i$. Assume that the μ_i are linearly independent, and $\mathbf{M}_d(Z_i) = 1$ for $d = 2, \dots, m$.

Estimation Task: Given $\mathbf{M}_2(Y), \mathbf{M}_3(Y), \dots, \mathbf{M}_m(Y)$, $m \geq 3$, or estimators thereof, determine/estimate μ_i and w_i for $i = 1, \dots, r$.

While the above scenario seems very restrictive, several important problems can be reduced to this setting, see for example [10], or chapter 3 of [2]. We recommend the interested reader to read the exposition there.

3.2. Algebraic Formulation via Moments

The estimation problem presented above can be reformulated as a purely algebraic problem, see [2]. Namely, the \mathbf{M}_i are explicitly calculable in terms of the expectations of the μ_i and w_i . Then, Proposition 2.20 implies that $\mathbf{M}_d(X_i) = \mu_i^{\otimes d}$ for all d , therefore $\mathbf{M}_d(Y) = \sum_{i=1}^r w_i \cdot \mu_i^{\otimes d}$ for all d , thus yielding the following algebraic version of the estimation problem.

Algebraic Problem: Let $r \leq n$, let $\mu_1, \dots, \mu_r \in \mathbb{R}^n$ be linearly independent and $w_1, \dots, w_r \in \mathbb{R}$ arbitrary such that $\sum_{i=1}^r w_i = 1$. Given (exact or noisy estimators for)

$$\mathbf{M}_d = \sum_{i=1}^r w_i \cdot \mu_i^{\otimes d} \quad \text{for } d = 2, \dots, m, \text{ with } m \geq 3,$$

determine the μ_i and w_i .

4. Algorithms

4.1. Orthogonal Decomposition of Tensors

A special case of orthogonal decomposition is singular value decomposition (SVD). There are a huge amount of well-studied methods for obtaining the singular value decomposition, which we will not discuss. However, we will make extensive use of the SVD algorithm, as described in Algorithm 1 as a black box.

Algorithm 1 SVD. Singular Value Decomposition of Matrices.

Input: A matrix $A \in \mathbb{R}^{m \times n}$. *Output:* The singular value decomposition $A = U \cdot \Sigma \cdot V^T$, with $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ orthogonal, $\Sigma \in \mathbb{R}^{r \times r}$ diagonal, and the rank $r = \text{rank} A$

First, for completeness, we treat the trivial case in Algorithm 2.

Algorithm 2 OTD1. Orthogonal Tensor Decomposition in one factor.

Input: A tensor $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$, a signature (S_1) . *Output:* The orthogonal atomic decomposition $T = \sum_{i=1}^r w_i \cdot A_i$.

1: Return rank $r = 1$, coefficients $w_1 = \|T\|$, factors $A_1 = \|T\|^{-1} \cdot T$.

Now we explicitly describe how to compute the orthogonal decomposition if each summand has two tensor factors. Algorithm 3 computes the decomposition by a proper reformatting of the entries, computing the singular value decomposition, then reformatting again.

Algorithm 3 OTD2. Orthogonal Tensor Decomposition in two factors.

Input: A tensor $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$, a signature (S_1, S_2) . *Output:* The orthogonal atomic decomposition $T = \sum_{i=1}^r w_i \cdot A_i \otimes B_i$ (assumed to exist), including the rank r

- 1: Define $\sigma : [d] \rightarrow [2], S_i \mapsto \{i\}$.
 - 2: Set $A \leftarrow (\sigma \dashv T)$. Note that $A \in \mathbb{R}^{m \times n}$, with $m = \prod_{i \in S_1} n_i, n = \prod_{i \in S_2} n_i$.
 - 3: Compute the SVD of $A = U \cdot \Sigma \cdot V^\top$, see Algorithm 1.
 - 4: Return rank $r = \text{rank} A$.
 - 5: Return coefficients $w_i = \Sigma_{ii}$ for $i = 1, \dots, r$.
 - 6: For all i , let U_i be the i -th column of U , let V_i be the i -th columns of V .
 - 7: Return factors $A_i = \sigma \vdash U_i, B_i = \sigma \vdash V_i$ for $i = 1, \dots, r$.
-

The algorithm for the general case, Algorithm 4, consists as well of repeated applications of reindexing and singular value decomposition. Variants of singular value decomposition exist with adjustable noise tolerance or singular value thresholding, and can therefore be employed to obtain thresholding and numerically stable variants of Algorithm 4. Furthermore, step 1 allows for an arbitrary choice of k -to-2-flattening, in each recursion. Since in the presence of noise, the results might differ when taking a different sequence flattenings, the numerical stability can be improved by clustering the results of all possible choices, then averaging.

Algorithm 4 OTD. Orthogonal Tensor Decomposition.

Input: A tensor $T \in \mathbb{R}^{n_1 \times \dots \times n_d}$, a signature (S_1, \dots, S_k) . *Output:* The orthogonal atomic decomposition $T = \sum_{i=1}^r w_i \cdot A_i^{(1)} \otimes \dots \otimes A_i^{(k)}$ (assumed to exist), including the rank r

- 1: Choose any k -to-2-flattening map τ .
 - 2: Set $\tilde{S}_j \leftarrow \cup_{i \in \tau^{-1}(j)} S_i$ for $j = 1, 2$.
 - 3: Set $\tilde{T} \leftarrow \tau \dashv T$.
 - 4: Use OTD2, Algorithm 3, to compute the orthogonal atomic decomposition $\tilde{T} = \sum_{i=1}^r w_i \cdot A_i \otimes B_i$ with signature $(\tilde{S}_1, \tilde{S}_2)$.
 - 5: Return the w_i as coefficients and r as the rank for the decomposition of T .
 - 6: For $i = 1, \dots, r$, use the suitable one of OTD1, OTD2, OTD, i.e., Algorithm 2, 3, or 4, to compute the orthogonal atomic decomposition $(\tau \vdash A_i) = \sum_{i=1}^1 1 \cdot \otimes_{\tau(j) \in S_1} A_i^{(j)}$, noting that rank is one, and using the signature $(S_j : \tau(j) \in \tilde{S}_1)$.
 - 7: For $i = 1, \dots, r$, use the suitable one of OTD1, OTD2, OTD, i.e., Algorithm 2, 3, or 4, to compute the orthogonal atomic decomposition $(\tau \vdash B_i) = \sum_{i=1}^1 1 \cdot \otimes_{\tau(j) \in S_2} A_i^{(j)}$, noting that rank is one, and using the signature $(S_j : \tau(j) \in \tilde{S}_2)$.
 - 8: Return the $A_i^{(j)}$ as factors for T .
-

Termination of Algorithm 4 is implied by the observation that in each recursion, the partition of $[d]$ is made strictly finer. Since $[d]$ has finite cardinality, there is only a finite number of recursions. The fact that the decompositions in steps 6 and 7 have rank one, and coefficients 1, follows from the uniqueness of the orthogonal decomposition guaranteed in Theorem 2. Correctness of Algorithm 4 follows from repeated application of Proposition 2.17, and the uniqueness of singular value decomposition.

4.2. An Estimator for the Mixture Model

For illustrative purposes, we write out Algorithm 4 for the problem introduced in 3, which has also extensively been studied in [2]:

Example: Let $r \leq n$, let $\mu_1, \dots, \mu_r \in \mathbb{R}^n$ be linearly independent and $w_1, \dots, w_r \in \mathbb{R}$ arbitrary such that $\sum_{i=1}^r w_i = 1$. Given (exact or noisy estimators for)

$$\mathbf{M}_d = \sum_{i=1}^r w_i \cdot \mu_i^{\otimes d} \quad \text{for } d = 2, 3,$$

determine the μ_i and w_i .

Algorithm 5 solves the problem, by reducing it to

Algorithm 5 Model identification.

Input: $\mathbf{M}_2, \mathbf{M}_3$ *Output:* $w_1, \dots, w_r, \mu_1, \dots, \mu_r$.

- 1: Set $r \leftarrow \text{rank}(\mathbf{M}_2)$.
 - 2: Compute the SVD³ $\mathbf{M}_2 = U \cdot \Sigma \cdot U^\top$.
 - 3: Set $W \leftarrow U \cdot \Sigma^{-\frac{1}{2}}$.
 - 4: Set $T := W^\top \circ \mathbf{M}_3$.
 - 5: Define the flattening map $\sigma : 1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 2$.
 - 6: Set $\tilde{T} := \sigma \dashv T$.
 - 7: Compute the rank r SVD $\tilde{T} = \sum_{i=1}^r \tilde{w}_i \cdot \tilde{\mu}_i^{(1)} \cdot v_i^\top$.
 - 8: Return $w_i = \tilde{w}_i^{-2}$ for $i = 1, \dots, r$.
 - 9: Set $\tilde{A}_i = (\sigma \vdash v_i)$ for $i = 1, \dots, r$.
 - 10: Compute the rank 1 SVD $\tilde{A}_i = \tilde{\mu}_i^{(2)} \cdot (\tilde{\mu}_i^{(3)})^\top$.
 - 11: Set $\tilde{\mu}_i \leftarrow \frac{1}{3} (\tilde{\mu}_i^{(1)} + \tilde{\mu}_i^{(2)} + \tilde{\mu}_i^{(3)})$, for $i = 1, \dots, r$.
 - 12: Compute the pseudo-inverse B of W' . Return $\mu_i = B \cdot \tilde{\mu}_i \cdot \tilde{w}_i$, for $i = 1, \dots, r$.
-

Theorem 4.3 in [2] implies that the tensor T obtained in step 4 has an orthogonal CP-decomposition, and it implies the correctness of steps 8, and 12. The fact that \tilde{A}_i in step 8 has rank one, and the coefficients are 1, follow from the uniqueness of the decomposition guaranteed in Theorem 2.

Note that explicit presentation of the algorithm could be substantially abbreviated by applying ODT directly to \tilde{T} in step 4, with signature $(\{1\}, \{2, 3\})$, and then performing the analogues of steps 8 and 12. Furthermore, the accuracy of the estimator in step 11 can be improved, by repeating the procedure for the three possible signatures $(\{1\}, \{2, 3\})$, $(\{2\}, \{1, 3\})$, and $(\{3\}, \{1, 3\})$, then averaging, or weighted averaging, over the nine estimates for each $\tilde{\mu}_i$, making use of the symmetry of the problem.

Also, similar to Algorithm 4, the presented Algorithm 5, while already numerically stable, can be modified to cope better with noise by, e.g., introducing thresholding to the singular value decomposition and rank computations. The numerical stability with respect to noise is governed by the numerical stability of the SVDs performed, and the pseudo-inversion of W' in step 12.

Algorithm 5 is also related to Algorithm 1 proposed in [1]. Namely, $\text{Triples}(\eta)$, as defined in section 4.1, is a degree 2-projection of the tensor T , and therefore can be also understood as a random

projection of the flattening $\sigma \dashv T$.

Furthermore, an estimator for the hierarchical models described in [11, 17] can be constructed in a similar way.

5. Conclusion

We have demonstrated that computing the orthogonal decomposition of an arbitrary degree tensor, symmetric or not, can be reduced to a series of singular value decompositions, and we have described efficient algorithms to do so. This makes orthogonal tensor decomposition approachable by the wealth of theoretical results and existing methods for eigenvalue problems and singular value decomposition. Moreover, we have exemplified our method in the case of identifying components in a low-rank mixture model.

Acknowledgments

I thank Arthur Gretton, Zoltán Szabó, and Andreas Ziehe for interesting discussions. I thank the Mathematisches Forschungsinstitut Oberwolfach for support.

References

- [1] Anima Anandkumar, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Yi-Kai Lu. A spectral algorithm for latent dirichlet allocation. *ArXiv e-print*, 2012.
- [2] Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decomposition for learning latent variable models. *ArXiv e-print*, 2012.
- [3] Rasmus Bro. PARAFAC. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149 – 171, 1997.
- [4] J. Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'eckart–young' decomposition. *Psychometrika*, 35:283–319, 1970.
- [5] Maria V. Catalisano, Anthony V. Geramita, and Allesandro Gimigliano. Rank of tensors, secant varieties of Segre varieties and fat points. *Linear Algebra and its Applications*, pages 263–285, 2002.
- [6] Lieven De Lathauwer, Pierre Comon, and Nicola Mastronardi. Special issue on tensor decompositions and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(3):.7–.7, September 2008. ISSN 0895-4798.
- [7] J. B. Denis and T. Dhorne. Orthogonal tensor decomposition of 3-way tables. In R. Coppi and S. Bolasco, editors, *Multiway data analysis.*, pages 31–37. Elsevier, Amsterdam, 1989.
- [8] Richard A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16(84):1–84, 1970.

- [9] Frank L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6:164–189, 1927.
- [10] Daniel Hsu and Sham M. Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. *ArXiv e-print*, 2012.
- [11] Maria Ishteva, Haeson Park, and Le Song. Unfolding latent tree structures using 4th order tensors. *JMLR Workshop and Conference Proceedings, ICML 2013*, pages 316–324, 2013.
- [12] Tamara G. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.
- [13] Carla D. Moravitz Martin and Charles F. Van Loan. A jacobi-type method for computing orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, pages 1219–1232, 2006.
- [14] Dimitri Nion and Nikos D. Sidiropoulos. A PARAFAC-based technique for detection and localization of multiple targets in a mimo radar system. In *Proc. ICASSP '09*, pages 2077–2080, 2009.
- [15] Nikos D. Sidiropoulos. Low-rank decomposition of multi-way arrays: a signal processing perspective. In *Sensor Array and Multichannel Signal Processing Workshop Proceedings*, pages 52 – 58, 2004.
- [16] Nikos D. Sidiropoulos, Rasmus Bro, and Georgios B. Giannakis. Parallel factor analysis in sensor array processing. *IEEE Transactions on Signal Processing*, 48:2377–2388, August 2000.
- [17] Le Song, Maria Ishteva, Ankur Parikh, Eric Xing, and Haeson Park. Hierarchical tensor decomposition of latent tree graphical models. *JMLR Workshop and Conference Proceedings, ICML 2013*, pages 334–342, 2013.
- [18] G. Young and C. Eckart. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.