ORTHOGONAL LOW RANK TENSOR APPROXIMATION: ALTERNATING LEAST SQUARES METHOD AND ITS GLOBAL CONVERGENCE

DRAFT AS OF July 5, 2014

LIQI WANG*, MOODY T. CHU†, AND BO YU‡

Abstract. With the notable exceptions of two cases — that tensors of order 2, namely, matrices, always have best approximations of arbitrary low ranks and that tensors of any order always have the best rank-one approximation, it is known that high-order tensors may fail to have best low rank approximations. When the condition of orthogonality is imposed, even under the modest assumption that only one set of components in the decomposed rank-one tensors is required to be mutually perpendicular, the situation is changed completely — orthogonal low rank approximations always exist. The purpose of this paper is to discuss the best low rank approximation subject to orthogonality. The conventional high-order power method is modified to address the orthogonality via the polar decomposition. Algebraic geometry technique is employed to show that for almost all tensors the orthogonal alternating least squares method converges globally.

Key words. orthogonal tensor decomposition, low rank approximation, alternating least squares, high-order power method, polar decomposition, global convergence, Zariski topology

AMS subject classifications. 15A29, 15A69, 41A50, 65J10, 68W25,

1. Introduction.

A real-valued tensor of order k can be represented by a k-way array

$$T = [\tau_{i_1,\ldots,i_k}] \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_k}$$

with elements $\tau_{i_1,\ldots,i_k}$ accessed via k indices. A tensor of the form

$$\mathbf{u}^{(1)}\otimes\ldots\otimes\mathbf{u}^{(k)} := \left[u^{(1)}_{i_1}\cdot u^{(2)}_{i_2}\ldots\cdot u^{(k)}_{i_k}\right], \tag{1.1}$$

where elements are the products of entries from vectors $\mathbf{u}^{(j)} \in \mathbb{R}^{I_j}$, $j = 1,\ldots,k$, is said to be of rank one. These vectors are referred to as *components* of the rank-1 tensor [25]. By a tensor decomposition we refer to the rewriting of the given tensor T as the sum of some rank-1 tensors. There are different desirable structures in the decomposition. The most general expression is the Tucker decomposition [9, 10, 20, 47]

$$T = \sum_{j_1,j_2,\ldots,j_k} g_{j_1,j_2,\ldots,j_k}\mathbf{u}^{(1)}_{j_1}\otimes\ldots\otimes\mathbf{u}^{(k)}_{j_k} \tag{1.2}$$

which includes the CANDECOMP/PARAFAC (CP) decomposition [3, 16]

$$T = \sum_j \mathbf{u}^{(1)}_j\otimes\ldots\otimes\mathbf{u}^{(k)}_j \tag{1.3}$$

as the "diagonal" case. This paper concerns only the CP decomposition, also known as the canonical decomposition [3], parallel factor [16], or canonical polyadic decomposition [20, 41]. See also [13, 23].

Tensor decompositions have important applications in a wide range of disciplines wherever multi-array data analysis is imperative. This subject, including the determination of minimum number of terms in the summation, referred to as the tensor rank, has been a major ongoing research topic with numerous partial results available in the literature. So that we may stay focused on our main point, we only mention without elaborating details the point of view that decomposing a tensor amounts to finding real-valued solutions for a high-degree polynomial system of equations. The latter is a challenging research subject in the field of the so called real algebraic geometry [1, 2, 40].

*School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, P. R. C., 116024. (lizzy@mail.dlut.edu.cn)

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. (chu@math.ncsu.edu) This research was supported in part by the National Science Foundation under grant DMS-1014666.

‡School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, P. R. C., 116024. (yubo@dlut.edu.cn)

An alternative approach in place of a complete decomposition is the low rank approximation. The basic notion consists of two parts. First, the number of terms in the summation is to be fixed a priori, whose choice is a difficult problem in itself [6, 27, 30] and affects the quality of approximation. Second, the difference between the given T and the summation is to be minimized in the sense of Frobenius norm. In contrast to the algebraic problem of finding an exact decomposition which is NP-hard [17, 19], the analytic problem of finding an approximation seems more computationally feasible. Nonetheless, it must be pointed out that, except for the case of order-2 tensors which are simply matrices or the case of rank-1 approximation for general tensors, the best low rank approximation for high-order tensors may not exist at all [11, 27, 29, 31, 42, 43]. An example of such a failure for rank-2 approximation to an order-k tensor, $k > 2$, is given in Section 2.2.

The purpose of this paper is to consider the low rank CP approximation under an extra requirement of orthogonality whose specific meaning will be explained later. The significant of this added condition is that the best orthogonal low rank tensor approximation always exists [5, 29]. On the other hand, there are applications which naturally impose orthogonality. The prerequisite of zero-mean signals being uncorrelated, for example, is essential in signal processing applications. For multi-dimensional data, orthogonal CP tensor decomposition finds applications in polarization sensitive array analysis, multiple access wireless communication systems, blind signal separation and identification, and variants of a few other independent component analysis methods. References associated with these applications can be found in [41, 49], which we shall not repeat here.

For matrices, the best low rank approximation under the Frobenius norm is given by the truncated singular value decomposition (TSVD). This is essentially the Eckart-Young theorem [14]. There are highly effective numerical algorithms for computing the TSVD of a matrix [14]. In this case, the orthogonality is an inherent property of the decomposition. For general tensors, initial efforts have been on generalizing the notion of matrix SVD to higher-dimensional data. Far from being complete, we mention [5, 9, 24, 31, 32, 36, 49] as a few references of interest in this direction. Some of these discussions also include accompanying computational schemes. See, for example, [5, 21, 32, 33]. We think the article [26] probably settles the issue that a full tensor SVD might be too much to expect for in general and, hence, a low rank approximation should be the alternative.

The"workhorse" algorithm for computing a low rank CP decomposition has been the alternating least squares (ALS) algorithm [6, 27, 46]. Even so, only its local convergence has been established to this date [48]. The special case of the best rank-1 approximation for tensors has been receiving extensive attention in the literature [4, 10, 11, 24, 44, 53], but only recently we prove for the first time that the iterates generated by the ALS method for the rank-1 approximation converge globally for almost all tensors [51]. Global convergence for general low rank approximation is still elusive.

Our main contribution in this paper is that, after proposing a modification of the conventional ALS method via the polar decomposition (PD) to address the orthogonality while maintaining a monotone ascending property, we further justify the working of our modification by a proof of global convergence. The notion of the PD is not new [5, 37], but the idea of its application to orthogonal tensor decomposition proves quite effective. The convergence analysis should be new in the field. For clarity, we summarize the chain of our arguments as follows, which will be established step by step in the subsequent discussion.

1. The iterates produced by the algorithm should maintain orthogonality.
2. The sequence has cluster points since it is bounded.
3. The sequence of increments tends to zero.
4. Cluster points are geometrically isolated for generic tensors.
5. Together, items 3 and 4 imply that there is a unique limit point of the iteration per starting point.

This paper is organized accordingly. We begin in Section 2 with some preliminary facts including the notions of adjoint and orthogonality for tensors. In Section 3, the first order optimality condition of approximation subject to orthogonality is converted into a variant of the ALS method which can also be regarded as a high-order power method. The notable addition in the algorithm is the mechanism via the polar decomposition to address the needed orthogonality. The resulting objective values are showed to enjoy an ascending property in Section 4 and hence converge. To prove the convergence of the iterates generated by the modified orthogonal ALS method requires additional mathematical insights which are presented in Section 5. In particular, we need to argue that the stationary points are isolated, which may be considered as a bonus result.

2. Preliminaries. This section introduces the notation system used in this paper and the basic concepts of adjoint and orthogonality. Expert readers may skip reading through these elementary properties and go to Section 3 directly.

2.1. Adjoint operator. The so called contraction product [7] or n-mode product [27] plays a fundamental role in tensor manipulation. For our applications, however, we are in constant need of multiple-level n-mode products. It is more convenient to introduce a special product which follows naturally from an operator point of view and brings forth the useful notion of adjoint.

An order-k tensor $T \in \mathbb{R}^{I_1 \times \ldots \times I_k}$ may be thought of as a linear operator mapping/contracting order-$(k-1)$ tensors to vectors. Because the tensor T has k "facets", it should be specified that with respect to which facet of the tensor T that the dimension contraction is taking place. For this matter, suppose ℓ is fixed, then we can identify T as the operator $\mathscr{T} \equiv \mathscr{T}_\ell$ with

$$\mathscr{T}_\ell : \mathbb{R}^{I_1 \times \ldots \times \widehat{I_\ell} \times \ldots \times I_k} \to \mathbb{R}^{I_\ell}, \tag{2.1}$$

where $\widehat{I_\ell}$ means that quantities associated with this particular index are taken out from the remaining list, and define its action at any $S \in \mathbb{R}^{I_1 \times \ldots \times \widehat{I_\ell} \times \ldots \times I_k}$ by the operation

$$\mathscr{T}_\ell(S) := T \circledast_\ell S = \begin{bmatrix} \langle \tau_{:,1,:}, S \rangle \\ \vdots \\ \langle \tau_{:,I_\ell,:}, S \rangle \end{bmatrix} \in \mathbb{R}^{I_\ell}. \tag{2.2}$$

In (2.2), the colon(:) denotes any wild indices in the tensor T, so $\tau_{:,\nu_\ell,:}$ represents the ν_ℓ-th "slice" of the tensor T in the ℓ-th direction which is an order-$(k-1)$ tensor. Also, $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product generalized to multi-dimensional arrays.

Taking an order-3 tensor acting on matrices as an illustration. Let $\{E_J\}$, where J is a double index, denote the standard basis $E_J = [e^J_{st}] \in \mathbb{R}^{m \times n}$ with

$$e^J_{st} = \begin{cases} 1, & \text{if } J = (s,t), \\ 0, & \text{otherwise,} \end{cases}$$

and likewise $\{\mathbf{e}_i\}$ the standard basis of $\mathbb{R}^p$. Suppose that the action of $\mathscr{T}$ at a basis matrix $E_J \in \mathbb{R}^{m \times n}$ is expressed in terms of basis $\{\mathbf{e}_i\}$ as

$$\mathscr{T}(E_J) = \sum_i \tau_{i,J} \mathbf{e}_i$$

with $\tau_{i,J} \in \mathbb{R}$. Then for any $X = [x_{ij}] \in \mathbb{R}^{m \times n}$ the linear map entails the relationship

$$\mathscr{T}(X) = \mathscr{T}\left(\sum_J x_J E_J \right) = \sum_J x_J \left(\sum_i \tau_{i,J} e_i \right) = \sum_i \left(\sum_J \tau_{i,J} x_J \right) \mathbf{e}_i. \tag{2.3}$$

By visualizing the order-3 tensor T a $p \times 1$ block matrix of which each block is an $m \times n$ matrices, the relationship (2.3) is a 2-dimensional contraction $\circledast_1$ defined by

$$\mathscr{T}(X) \equiv \underbrace{\begin{bmatrix} \boxed{\tau_{1,:}} \\ \boxed{\tau_{2,:}} \\ \vdots \\ \boxed{\tau_{p,:}} \end{bmatrix}}_{T} \circledast_1 X := [\langle \tau_{i,:}, X \rangle] \in \mathbb{R}^p.$$

It follows naturally that the adjoint operator $\mathscr{T}^* : \mathbb{R}^p \to \mathbb{R}^{m\times n}$ should be defined as

$$\mathscr{T}^*(\mathbf{v}) := \sum_i v_i \tau_{i,:} \tag{2.4}$$

which satisfies the required property that

$$\langle \mathscr{T}(X), \mathbf{v}\rangle = \langle X, \mathscr{T}^*(\mathbf{v})\rangle \tag{2.5}$$

for every $X \in \mathbb{R}^{m\times n}$ and $\mathbf{v} \in \mathbb{R}^p$.

Generalizing the spirit of (2.5) whereas the notion of "transpose" is embedded in the tensor product $\circledast_\ell$, the following result can easily be proved by rearranging terms in the summation via the associative law.

LEMMA 2.1. *For any $\ell = 1, \ldots, k$, let $\mathbf{h}^{(\ell)} \in \mathbb{R}^{I_\ell}$ be an arbitrary vector. Then it is true that*

$$\left\langle T, \mathbf{u}^{(1)}\otimes\ldots\otimes\mathbf{h}^{(\ell)}\otimes\ldots\otimes\mathbf{u}^{(k)}\right\rangle = \left\langle T\circledast_\ell \left(\mathbf{u}^{(1)}\otimes\ldots\otimes\widehat{\mathbf{u}}^{(\ell)}\otimes\ldots\otimes\mathbf{u}^{(k)}\right), \mathbf{h}^{(\ell)}\right\rangle. \tag{2.6}$$

Because (2.6) holds for any index ℓ, we have the reciprocating relationship

$$\left\langle T\circledast_i \left(\mathbf{u}^{(1)}\otimes\ldots\otimes\widehat{\mathbf{u}}^{(i)}\otimes\ldots\otimes\mathbf{u}^{(k)}\right), \mathbf{u}^{(i)}\right\rangle = \left\langle T\circledast_j \left(\mathbf{u}^{(1)}\otimes\ldots\otimes\widehat{\mathbf{u}}^{(j)}\otimes\ldots\otimes\mathbf{u}^{(k)}\right), \mathbf{u}^{(j)}\right\rangle. \tag{2.7}$$

for any two distinct indices i and j. Both identities (2.6) and (2.7) will be useful in the subsequent discussion.

2.2. Rank-2 approximation, nonexistence, and orthogonality. The problem of finding a best rank-2 approximation to a given order-k tensor $T \in \mathbb{R}^{I_1\times\ldots\times I_k}$ is to determine unit vectors $\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)} \in \mathbb{R}^{I_\ell}$, $\ell = 1, \ldots k$, and scalars $\lambda_1, \lambda_2 \in \mathbb{R}$ such that the function

$$f\left(\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_1^{(k)}, \mathbf{u}_2^{(1)}, \ldots, \mathbf{u}_2^{(k)}; \lambda_1, \lambda_2\right) := \left\| T - \lambda_1 \underbrace{\mathbf{u}_1^{(1)}\otimes\ldots\otimes\mathbf{u}_1^{(k)}}_{H_1} - \lambda_2 \underbrace{\mathbf{u}_2^{(1)}\otimes\ldots\otimes\mathbf{u}_2^{(k)}}_{H_2}\right\|_F^2 \tag{2.8}$$

is minimized. Note that all vectors are normalize to unit length. This seemingly straightforward least squares problem is far more complicated as can be seen from the following example [11].

Given linearly independent unit vectors $\mathbf{a}^{(j)}, \mathbf{b}^{(j)} \in \mathbb{R}^{I_j}$, $j = 1, \ldots, k$, consider the tensor

$$X := \mathbf{b}^{(1)}\otimes\mathbf{a}^{(2)}\otimes\ldots\otimes\mathbf{a}^{(k)} + \mathbf{a}^{(1)}\otimes\mathbf{b}^{(2)}\otimes\ldots\otimes\mathbf{a}^{(k)} + \ldots + \mathbf{a}^{(1)}\otimes\mathbf{a}^{(2)}\otimes\ldots\otimes\mathbf{b}^{(k)}$$

which is of rank k. For $\gamma \in \mathbb{R}$, define

$$Y(\gamma) := \gamma\left(\mathbf{a}^{(1)} + \frac{\mathbf{b}^{(1)}}{\gamma}\right)\otimes\left(\mathbf{a}^{(2)} + \frac{\mathbf{b}^{(2)}}{\gamma}\right)\otimes\ldots\otimes\left(\mathbf{a}^{(k)} + \frac{\mathbf{b}^{(k)}}{\gamma}\right) - \gamma\mathbf{a}^{(1)}\otimes\mathbf{a}^{(2)}\otimes\ldots\otimes\mathbf{a}^{(k)}$$

which is of rank 2. Then it is easy to see that $Y(\gamma) \to X$ as $\gamma \to \infty$, showing that the best rank-2 approximation to X does not exist. What has happened is that if we rewrite the difference $\|X - Y(\gamma)\|_F^2$ in the form of (2.8) by enforcing unit base tensors H_1 and H_2, then the two corresponding scalars λ_1 and λ_2 are necessarily going to infinity in opposite sign [11, Proposition 4.8]. This phenomenon is referred to as "degeneracy" in [29]. Such an occurrence results in longitudinal cancelation and ultimate convergence to X. In this case, the two unit base tensors H_1 and H_2 eventually become identical.

One possible remedy to avoid such a coalescing phenomenon is to impose the condition that the base tensors H_1 and H_2 remain orthogonal to each other. The reason is that orthogonality will imply boundedness of λ_1 and λ_2, as we shall explain in Section 3.1. Such a requirement of orthogonality is more in line with, but not exactly the same as, the best rank-2 approximation to a matrix. In the TSVD of a given matrix, the rank-1

base matrices are necessarily orthogonal to each other because the corresponding singular vectors are mutually orthogonal. For rank-1 tensors, however, note that the inner product can be written as

$$\langle H_1, H_2 \rangle = \prod_{\ell=1}^{k} \left\langle \mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)} \right\rangle.$$

So rank-1 base tensors H_1 and H_2 are mutually orthogonal if and only if at least one pair of its components $\left(\mathbf{u}_1^{(\ell)}, \mathbf{u}_2^{(\ell)}\right)$ are orthogonal. The implication that only one pair of orthogonality among components of rank-1 matrices is needed is very different from, and more relaxed than, the pairwise orthogonality happening in the singular value decomposition (SVD) for matrices.

This notion of H_1 being orthogonal to H_2 corresponds to the so called *orthogonal rank approximation* which is the most general case among three definitions of orthogonality for two rank-1 tensors characterized in [25, Section 3.1]. The more restricted case of the so called *complete orthogonality* requires pairwise orthogonality of all components of rank-1 matrices, of which an example is the SVD, and is not within our consideration. It was suggested in [25, Section 4] that the completely orthogonal decomposition might be easier to manipulate than the orthogonal decomposition. This paper concerns the orthogonal rank approximation only. For the completely orthogonal decomposition, we do suggest [53, Section 3] for a theoretical discussion, and [5] and [33] for some numerical algorithms. A general treatment of the various aspects of orthogonality and associated problems can be found in [25]. See also Section 6 in this paper for a brief discussion on matrices and other references for works generalizing the Eckart-Young theorem to tensors.

3. Orthogonal low rank approximation. Generalizing (2.8), we are interested in the orthogonal rank-R CP approximation to $\mathcal{T}$ by minimizing the objective function

$$f\left(\mathbf{u}_1^{(1)}, \ldots, \mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_R^{(1)}, \ldots, \mathbf{u}_R^{(k)}; \lambda_1, \ldots, \lambda_R\right) := \left\| T - \sum_{r=1}^{R} \lambda_r \underbrace{\mathbf{u}_r^{(1)} \otimes \ldots \otimes \mathbf{u}_r^{(k)}}_{H_r} \right\|_F^2, \tag{3.1}$$

subject to the mutual orthonormality condition that

$$\langle H_i, H_j \rangle = \prod_{\ell=1}^{k} \left\langle \mathbf{u}_i^{(\ell)}, \mathbf{u}_j^{(\ell)} \right\rangle = \delta_{ij}, \quad \text{for all } 1 \le i, j \le R. \tag{3.2}$$

Obviously, (3.2) constitutes a far more complicate set of constraints among the components of the rank-1 base tensors. We shall further refine the constraints (3.2) by imposing two additional conditions, which will be assumed henceforth in this paper. First, all components $\mathbf{u}_r^{(\ell)}$, $\ell = 1, \ldots, k$, are of unit length, which automatically implies that the base tensors H_r, $r = 1, \ldots, R$, are of unit length, but not the converse. Second, there exists an integer s, $1 \le s \le k$, such the columns of the factor matrix

$$A^{(s)} := \left[\mathbf{u}_1^{(s)}, \mathbf{u}_2^{(s)}, \ldots, \mathbf{u}_R^{(s)}\right] \in \mathbb{R}^{I_s \times R} \tag{3.3}$$

are mutually orthonormal, which is sufficient for the orthogonality specified by (3.2) but significantly limits the overall structure. The condition (3.2) under the special case of (3.3), is referred to as the *semi-orthogonality* [1] [41]. To fix the idea in the following discussion, we may rearrange the indices if necessary and assume without loss of generality that the semi-orthogonality happens at $A^{(k)}$, that is, $\langle \mathbf{u}_i^{(k)}, \mathbf{u}_j^{(k)} \rangle = \delta_{ij}$ for $1 \le i, j \le R$. We stress that this work concerns the semi-orthogonality of one factor matrix $A^{(k)}$ only. The question of more than one semi-orthogonal factor matrix, except for the case of complete orthogonality [5], remains open. (See Section 5.4 for a brief discussion.)

[1] We think that this term is a somewhat misnamed connotation since now semi-orthogonality implies orthogonality.

In order to minimize (3.1) subject to the semi-orthogonality condition, two numerical approaches have been proposed in a recent paper [41]. These are techniques exploiting the simultaneous matrix diagonalization (SD-CPO) and the ALS applied to the combined mode matrices (ALS-CPO), respectively. The former can be interpreted as a simultaneous Takagi factorization and requires the calculation of a specific orthogonal matrix F. The latter maintains orthogonality by computing the SVD of a certain flattened matrix representation of the tensor, though the combined mode matrices can cleverly be made implicit. Results from extensive numerical experiments are reported, but no proof of convergence is ever given. This lack of proof is unfortunately falling in the same vein as other ALS variants toward which it was exclaimed in [6] that "Their [the ALS algorithms] extensive use is thus unexplainable."

In contrast, we apply our modified ALS method to the given tensor as it is without combining modes; we ensure the orthogonality with the much easier and straightforward polar decomposition circumventing the sophisticated procedures described in [41]; and, more importantly, we prove global convergence which should have addressed some critical questions raised in [25]. The tactics developed in the subsequent discussion might be useful for tackling more general problems.

3.1. Generalized Rayleigh quotient. One immediate consequence of orthogonality is that, for each fixed set of mutual orthogonal unit tensors H_r, $r = 1, \ldots, R$, the optimal scalars in (3.1) are necessarily given by

$$\lambda_r = \lambda_r \left(\mathbf{u}_r^{(1)}, \ldots, \mathbf{u}_r^{(k)} \right) = \left\langle T, \mathbf{u}_r^{(1)} \otimes \ldots \otimes \mathbf{u}_r^{(k)} \right\rangle, \quad r = 1, \ldots, R. \tag{3.4}$$

It effectively implies that $|\lambda_r| \leq \|T\|_F$, making a drastic difference from that without the orthogonality.

As the base tensors are mutually orthonormal, the expression for λ_i in (3.4) can also be interpreted as the Fourier coefficient of, or the length of the projection of the "vector" T onto, the the "unit vector" H_r under the Frobenius inner product. Thus, the minimization of (3.1) is equivalent to the problem of

$$\max \quad \sum_{r=1}^{R} \lambda_r^2 \tag{3.5}$$

subject to the same constraints. In the subsequent discussion, the set of variables over which the maximization is conducted will be limited to the refined constraints mentioned above.

3.2. Optimality condition. The Lagrangian for the optimization problem (3.5) is

$$\mathcal{L} := \sum_{r=1}^{R} \lambda_r^2 - \sum_{\ell=1}^{k} \sum_{r=1}^{R} \rho_r^{(\ell)} \left(\left\langle \mathbf{u}_r^{(\ell)}, \mathbf{u}_r^{(\ell)} \right\rangle - 1 \right) - \sum_{1 \leq s < t \leq R} \mu_{st} \left\langle \mathbf{u}_s^{(k)}, \mathbf{u}_t^{(k)} \right\rangle,$$

where λ_r is given by (3.4) and $\rho_r^{(\ell)}, \mu_{st}$ are Lagrange multipliers which will be identified below. Accordingly, the first order optimality condition for a stationary point is to satisfy the system of equations defined by [32]

$$\lambda_r T \circledast_\ell \left(\mathbf{u}_r^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}_r^{(\ell)} \otimes \ldots \otimes \mathbf{u}_r^{(k)} \right) = \rho_r^{(\ell)} \mathbf{u}_r^{(\ell)}, \quad \ell = 1, \ldots, k-1; \quad r = 1, \ldots, R, \tag{3.6}$$

and

$$\lambda_r \underbrace{T \circledast_k \left(\mathbf{u}_r^{(1)} \otimes \ldots \otimes \mathbf{u}_r^{(k-1)} \right)}_{\mathbf{v}_r^{(k)}} = \rho_r^{(k)} \mathbf{u}_r^{(k)} + \sum_{r<t} \frac{\mu_{rt}}{2} \mathbf{u}_t^{(k)} + \sum_{s<r} \frac{\mu_{sr}}{2} \mathbf{u}_s^{(k)}, \quad r = 1, \ldots, R. \tag{3.7}$$

Taking inner product on each side of (3.6) with respect to $\mathbf{u}_r^{(\ell)}$ and taking the relationship (3.4) into account, we see that the Lagrange multipliers $\rho_r^{(\ell)}$ must satisfy the relationships

$$\rho_r^{(\ell)} = \lambda_r^2, \quad \ell = 1, \ldots, k; \quad r = 1, \ldots, R, \tag{3.8}$$

which then reduces the systems (3.6) to

$$T \circledast_\ell \left(\mathbf{u}_r^{(1)} \otimes \ldots \otimes \widehat{\mathbf{u}}_r^{(\ell)} \otimes \ldots \otimes \mathbf{u}_r^{(k)} \right) = \lambda_r \mathbf{u}_r^{(\ell)}, \quad \ell = 1, \ldots, k-1; \quad r = 1, \ldots, R. \tag{3.9}$$

Likewise, the relationships

$$\begin{cases} \frac{\mu_{rt}}{2} & = \lambda_r \left\langle \mathbf{v}_r^{(k)}, \mathbf{u}_t^{(k)} \right\rangle, \quad \text{if } r < t \\ \frac{\mu_{sr}}{2} & = \lambda_r \left\langle \mathbf{v}_r^{(k)}, \mathbf{u}_s^{(k)} \right\rangle, \quad \text{if } s < r \end{cases} \tag{3.10}$$

also hold. So we can rewrite (3.7) as

$$\mathbf{v}_r^{(k)} = \lambda_r \mathbf{u}_r^{(k)} + \sum_{r<t} \left\langle \mathbf{v}_r^{(k)}, \mathbf{u}_t^{(k)} \right\rangle \mathbf{u}_t^{(k)} + \sum_{s<r} \left\langle \mathbf{v}_r^{(k)}, \mathbf{u}_s^{(k)} \right\rangle \mathbf{u}_s^{(k)}, \quad r = 1, \ldots, R. \tag{3.11}$$

In all, (3.9) and (3.11) together represent a system of $R \sum_{\ell=1}^{k} I_\ell$ nonlinear equations for the unknowns $\mathbf{u}_r^{(\ell)}$, $\ell = 1, \ldots, k$, $r = 1, \ldots, R$. A solution to this system gives rise to a stationary point of the problem (3.5).

The system (3.11) represents a Fourier series expansion of $\mathbf{v}_r^{(k)}$ with respect to the basis $\left\{ \mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_R^{(k)} \right\}$. Having the linear combination for each $\mathbf{v}_r^{(k)}$ rewritten in the outer product form, we can express the subsystem (3.11) collectively in the matrix equation[2]

$$\underbrace{\left[\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_R^{(k)} \right]}_{V^{(k)}} = \underbrace{\left[\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_R^{(k)} \right]}_{U^{(k)}} \underbrace{\left[\left\langle \mathbf{v}_i^{(k)}, \mathbf{u}_j^{(k)} \right\rangle \right]}_{P^{(k)}}, \tag{3.12}$$

where $V^{(k)}, U^{(k)} \in \mathbb{R}^{I_k \times R}$ and $P^{(k)} \in \mathbb{R}^{R \times R}$. Note that the scalar λ_r in (3.11) has gone into the diagonal of $P^{(k)}$ because the semi-orthogonality at $A^{(k)}$ implies that $\lambda_r = \langle \mathbf{v}_r^{(k)}, \mathbf{u}_r^{(k)} \rangle$.

At first glance, the right side of (3.12) denotes only a factorization of $V^{(k)}$. As $P^{(k)}$ is square matrix, we would think that $U^{(k)}$ might be expressed in terms of $V^{(k)}$. The trouble is that, in order to satisfy (3.11), the matrix $P^{(k)}$ itself also involves the unknown vectors $\mathbf{u}_r^{(k)}$. Among the various possible factorizations of $V^{(k)}$, our first demand is that columns of $U^{(k)}$ should be mutually orthonormal. Furthermore, note that the way $\mathbf{v}_r^{(k)}$ is defined in (3.7) is similar to those $\mathbf{u}_r^{(\ell)}$ defined in (3.9), except that the former is yet to be orthonormalized. We thus are motivated that $\mathbf{u}_r^{(k)}$ should stay in close proximity to the normalized $\mathbf{v}_r^{(k)}$ for each $r = 1, \ldots, R$. Judging from these requirements, we propose using the polar decomposition of the column-wise normalized $V^{(k)}$ as the solution for $U^{(k)}$ because of the following known results [18, Theorem 1.1 and Corollary 2.3].

THEOREM 3.1. *Assume $I_k \geq R$. Let $D = \operatorname{diag}\left\{ \|\mathbf{v}_1^{(k)}\|_2, \ldots, \|\mathbf{v}_R^{(k)}\|_2 \right\}$. Then there exist a matrix $U^{(k)} \in \mathbb{R}^{I_k \times R}$ and a unique symmetric positive semi-definite matrix $S^{(k)} \in \mathbb{R}^{R \times R}$ such that the equation (3.12) holds for $P^{(k)} = S^{(k)} D$ and ${U^{(k)}}^\top U^{(k)} = I_R$. Furthermore,*

1. *If $\operatorname{rank}(V^{(k)}) = R$, then $S^{(k)}$ is positive definite and $U^{(k)}$ is uniquely determined.*
2. *$U^{(k)}$ is the nearest orthogonal approximation to the normalized $V^{(k)}$ in the sense that*

$$\|V^{(k)} D^{-1} - U^{(k)}\|_F \leq \|V^{(k)} D^{-1} - Z\|_F$$

for any $Z \in \mathbb{R}^{I_k \times R}$ satisfying $Z^\top Z = I_R$.

The polar decomposition of a given matrix can easily be computed by means of the SVD[3]. It serves as the key mechanism in our algorithm to maintain the orthogonality.

[2] A similar setting with emphasis on symmetry has been considered in [5, Section 5.3] which also leads to the polar decomposition as we shall describe below. The difference is that the setting in [5] is for complete orthogonality and, hence, symmetry follows. Our work concerns only semi-orthogonality.

[3] Actually, the update in the algorithm ALS1-CPO in [41] is computing precisely the PD of ${A^{(3)}}^\top$ by means of the SVD, whose motivation is explained in a different way via [41, Section 4.3].

3.3. Modified high-order power iteration. A commonly used tactic for solving the nonlinear system (3.9) and (3.11) is to alternate directions among the variables in a way similar to the classical Gauss-Seidel method [15, 33]. This is also known as the high-order power method [9, 32, 34]. The added challenge is to maintain the required semi-orthogonality. We propose to handle this condition through the polar decomposition and justify its working in Sections 4 and 5. The iterative procedure as is summarized in Algorithm 1.

Some remarks on the notation are due. For clarity, we have included an extra subscript $\cdot_{[p]}$ to denote the advance in iteration. So, $\cdot^{(\ell)}_{r,[p]}$ indicates that the associated quantity $\cdot^{(\ell)}_{r}$ is resulted from the p-th iteration, though in a real coding such a reference to p is not needed. Note that for each fixed p and ℓ, the calculation of $\mathbf{u}^{(\ell)}_{r,[p+1]}$ for $r = 1, \ldots, R$ in the innermost loop can be done in parallel. The sequence $\left\{\lambda^{(\ell)}_{r,[p]}\right\}_{p=1}^{\infty}$ will prove useful for approximating the generalized Rayleigh quotient λ_r. In particular, we shall use the second property in Theorem 3.1 to maintain a specific property of monotonicity which will be explained in the next section.

Algorithm 1 modifies the classical alternating least squares algorithm [6, 27, 48] with the regression step for $A^{(k)}$ replaced by orthogonal regression[4]. The main contribution of the paper is to prove the global convergence of the iterates $\left\{\mathbf{u}^{(\ell)}_{r,[p]}\right\}$.

Algorithm 1 (High-order power method for orthogonal low rank decomposition.)

Require: Given $\mathbf{u}^{(\ell)}_{r,[0]} \in \mathbb{R}^{I_\ell}$, $\ell = 1, \ldots, k$, $r = 1, \ldots, R$, and $\mathbf{u}^{(k)}_{i,[0]} \perp \mathbf{u}^{(k)}_{j,[0]}$ for $1 \le i \ne j \le R$,
  for $p = 0, 1, \cdots,$ **do**
    for $\ell = 1, 2, \cdots, k-1,$ **do**
      for $r = 1, 2, \cdots, R,$ **do**
        {*This loop can be done in parallel*}
        $\mathbf{u}^{(\ell)}_{r,[p+1]} = T \circledast_\ell \left(\mathbf{u}^{(1)}_{r,[p+1]} \otimes \ldots \otimes \mathbf{u}^{(\ell-1)}_{r,[p+1]} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \mathbf{u}^{(\ell+1)}_{r,[p]} \ldots \otimes \mathbf{u}^{(k)}_{r,[p]}\right)$
        $\lambda^{(\ell)}_{r,[p+1]} := \|\mathbf{u}^{(\ell)}_{r,[p+1]}\|_2$
        $\mathbf{u}^{(\ell)}_{r,[p+1]} := \frac{\mathbf{u}^{(\ell)}_{r,[p+1]}}{\lambda^{(\ell)}_{r,[p+1]}}$
      end for
    end for
    for $r = 1, 2, \cdots, R,$ **do**
      {*This loop can be done in parallel*}
      $\mathbf{v}^{(k)}_{r,[p+1]} = T \circledast_k \left(\mathbf{u}^{(1)}_{r,[p+1]} \otimes \ldots \otimes \mathbf{u}^{(k-1)}_{r,[p+1]}\right)$
      $\lambda^{(k)}_{r,[p+1]} := \|\mathbf{v}^{(k)}_{r,[p+1]}\|_2$
      $\mathbf{v}^{(k)}_{r,[p+1]} := \frac{\mathbf{v}^{(k)}_{r,[p+1]}}{\lambda^{(k)}_{r,[p+1]}}$ {*define columns of* V *here*}
    end for
    $[U, P] = \texttt{poldec}(V)$ {*polar decomposition is called here*}
    for $r = 1, 2, \cdots, R,$ **do**
      $\mathbf{u}^{(k)}_{r,[p+1]} := U(:, r)$
    end for
  end for

[4] All ALS methods are conceptually similar — updating one set of variables a time via least squares while keeping all other variables fixed. The only difference is at the formulation of the objective functions. See footnote #5 for a few variations. Both approaches are dealing with the CP formulation, so the ALS-CPO in [41] and our Algorithm 1 are mathematically equivalent. The subtle difference is at the fact that the innermost loop (for the index r) of Algorithm 1 can be processed in parallel is precisely where the ALS-CPO comes in to put forth in matrix form.

4. Convergence of objective values. In the following discussion, all vectors with indices in the form $\cdot_{r,[p]}^{(\ell)}$ refer to those defined in Algorithm 1. For $r = 1, \ldots, R$, $p = 0, 1, \ldots$, evaluate the generalized Rayleigh quotient at $\left(\mathbf{u}_{r,[p]}^{(1)}, \ldots, \mathbf{u}_{r,[p]}^{(k)}\right)$ and denote the scalar by

$$\lambda_{r,[p]} := \lambda_r \left(\mathbf{u}_{r,[p]}^{(1)}, \ldots, \mathbf{u}_{r,[p]}^{(k)}\right).$$

Changing the signs of the starting points if necessary, we may assume that $\lambda_{r,[0]} > 0$ for all $r = 1, \ldots, R$. The progressive way of defining the first $\mathbf{u}_{r,[p+1]}^{(\ell)}$ described in Algorithm 1 ensures an inherent ascending property on its iterates in the sense that [34, 51]

$$\lambda_{r,[p]} \leq \lambda_{r,[p+1]}^{(1)} \leq \lambda_{r,[p+1]}^{(2)} \leq \ldots \leq \lambda_{r,[p+1]}^{(k-1)} \leq \lambda_{r,[p+1]}^{(k)}, \quad r = 1, \ldots, R. \tag{4.1}$$

Since $\mathbf{u}_{r,[p+1]}^{(k)}$ is defined differently, we need to argue that the cascade of inequalities in (4.1) continues to hold if $\lambda_{r,[p+1]}^{(k)}$ is replaced by $\lambda_r \left(\mathbf{u}_{r,[p+1]}^{(1)}, \ldots, \mathbf{u}_{r,[p+1]}^{(k)}\right)$. This is one major watershed we have to mark clearly when the conventional ALS method is modified in order to ensure the semi-orthogonality. For the conventional ALS where orthogonality is not an issue, see the general discussion in [34, Section 4] and [51].

Recall that the ultimate goal is to maximize $\sum_{r=1}^{R} \lambda_r^2$ subject to the refined semi-orthogonality constraints. During the intermediate steps of Algorithm 1, we have already generated unit vectors $\mathbf{u}_{r,[p+1]}^{(\ell)}$, $r = 1, \ldots, R$, $\ell = 1, \ldots, k-1$, and the inequalities (4.1). Consider the linear functional $\pi_r^{(k)} : \mathbb{R}^{I_k} \to \mathbb{R}$, $r = 1, \ldots, R$, defined by

$$\pi_r^{(k)} \left(\mathbf{y}_r^{(k)}\right) := \left\langle T, \left(\mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(k-1)}\right) \otimes \mathbf{y}_r^{(k)} \right\rangle \tag{4.2}$$

over the unit sphere in $\mathbb{R}^{I_k}$. It is readily seen that $\pi_r^{(k)} \left(\mathbf{u}_{r,[p]}^{(k)}\right) = \lambda_{r,[p+1]}^{(k-1)} \leq \pi_r^{(k)} \left(\mathbf{v}_{r,[p+1]}^{(k)}\right) = \lambda_{r,[p+1]}^{(k)}$ because (the normalized) $\mathbf{v}_{r,[p+1]}^{(k)}$ is the global maximizer for the functional $\pi_r^{(k)}$ over all unit vectors $\mathbf{y}_r^{(k)} \in \mathbb{R}^{I_k}$ and $\mathbf{u}_{r,[p]}^{(k)}$, $r = 1, \ldots, R$, are some orthonormal vectors obtained from the previous iteration. The concern is that we need to make a comparison with $\mathbf{u}_{r,[p+1]}^{(k)}$, $r = 1, \ldots, R$.

In the same way as we have derived (3.11), the optimality condition for maximizing $\sum_{r=1}^{R} \left(\pi_r^{(k)}\right)^2$ subject to the semi-orthogonality condition among $\mathbf{y}_1^{(k)}, \ldots, \mathbf{y}_R^{(k)}$ can be described in terms of the polar decomposition

$$\left[\mathbf{v}_{1,[p+1]}^{(k)}, \ldots, \mathbf{v}_{R,[p+1]}^{(k)}\right] = \left[\mathbf{u}_{1,[p+1]}^{(k)}, \ldots, \mathbf{u}_{R,[p+1]}^{(k)}\right] P, \tag{4.3}$$

with $P = [p_{ij}] \in \mathbb{R}^{R \times R}$ being symmetric and positive semi-definite. The polar decomposition ensures not only that $\mathbf{u}_{r,[p+1]}^{(k)}$, $r = 1, \ldots, R$, are mutually orthonormal, but also that

$$\left\langle \mathbf{u}_{r,[p+1]}^{(k)}, \mathbf{v}_{r,[p+1]}^{(k)} \right\rangle = p_{rr} \geq 0, \quad r = 1, \ldots, R.$$

It follows that, for each $r = 1, \ldots, R$, the pair of unit vectors $\mathbf{u}_{r,[p+1]}^{(k)}$ and $\mathbf{v}_{r,[p+1]}^{(k)}$ maintain an acute angle. Since $\mathbf{v}_{r,[p+1]}^{(k)}$ points to the steepest ascend direction for $\pi_r^{(k)}$, so does $\mathbf{u}_{r,[p+1]}^{(k)}$ increase the value of the functional $\pi_r^{(k)}$. Collectively, the updated set $\left\{\mathbf{u}_{1,[p+1]}^{(k)}, \ldots, \mathbf{u}_{R,[p+1]}^{(k)}\right\}$ is generically the unique set of vectors representing the minimum change of $\left\{\mathbf{v}_{1,[p+1]}^{(k)}, \ldots, \mathbf{v}_{R,[p+1]}^{(k)}\right\}$ while satisfying both the optimality condition

(4.3) and the orthonormal constraint. Note that the old set $\left\{\mathbf{u}_{1,[p]}^{(k)},\ldots,\mathbf{u}_{R,[p]}^{(k)}\right\}$ satisfies only the condition of orthogonality. We thus establish the inequalities

$$\sum_{r=1}^{R}\lambda_{r,[p]}^{2}\leq\sum_{r=1}^{R}\left(\lambda_{r,[p+1]}^{(k-1)}\right)^{2}=\sum_{r=1}^{R}\left(\pi_{r}^{(k)}\left(\mathbf{u}_{r,[p]}^{(k)}\right)\right)^{2}\leq\sum_{r=1}^{R}\left(\pi_{r}^{(k)}\left(\mathbf{u}_{r,[p+1]}^{(k)}\right)\right)^{2}=\sum_{r=1}^{R}\lambda_{r,[p+1]}^{2}. \tag{4.4}$$

Different from the conventional ALS without orthogonality, it is worth noting that the last inequality in

$$\lambda_{r,[p]}\leq\lambda_{r,[p+1]}^{(1)}\leq\lambda_{r,[p+1]}^{(2)}\leq\ldots\leq\lambda_{r,[p+1]}^{(k-1)}\leq\lambda_{r,[p+1]},\quad r=1,\ldots,R \tag{4.5}$$

may not hold. However, each sweep in the outer loop for p in Algorithm 1 has the effect of increasing the objective value of $\sum_{r=1}^{R}\lambda_{r}^{2}$. We have already pointed out that $\lambda_{r}\leq\|T\|_{F}$, due to the orthogonality. The sequence $\left\{\sum_{r=1}^{R}\lambda_{r,[p]}^{2}\right\}$ of objective values generated from iterates by Algorithm 1 must converge.

5. Convergence of iterates. The argument in the preceding section shows only the convergence of the objective values $\sum_{r=1}^{R}\lambda_{r,[p]}^{2}$, which is not enough yet to guarantee the convergence of iterates themselves to a maximizer. More need be done. In fact, the lack of convergence analysis has always been a major issue in the ALS applications[5] [8, 27, 34]. Only recently, under some additional conditions on the Hessian matrix, was a local convergence analysis of the conventional ALS method for the low-rank approximation finally accomplished in [48]. Global convergence for general low rank approximation is still elusive. Using algebraic geometry argument, we have made a small progress recently in proving the global convergence of the ALS iterates for the rank-1 approximations in [51].

Our work here is different in two aspects. First, we need to address the added orthogonality. Second, no convergence proof is ever given for orthogonal low-rank approximation, despite the two methods proposed in [41] for the calculation. Proving the global convergence of the orthogonal tensor iterates generated by Algorithm 1 will be a new advance in the field, which is the primal contribution of this paper. Our approach consists of three separate parts which we present below.

5.1. Geometrically isolated stationary points. Algorithm 1 generates a sequence of vectors which we group as

$$\left\{\left(\mathbf{u}_{1,[p]}^{(1)},\ldots,\mathbf{u}_{1,[p]}^{(k)}\right),\ldots,\left(\mathbf{u}_{R,[p]}^{(1)},\ldots,\mathbf{u}_{R,[p]}^{(k)}\right)\right\}.$$

We see from the above discussion that any accumulation point of this sequence necessarily satisfies the system of nonlinear equations

$$\begin{cases} T\circledast_{\ell}\left(\mathbf{u}_{r}^{(1)}\otimes\ldots\otimes\widehat{\mathbf{u}}_{r}^{(\ell)}\otimes\ldots\otimes\mathbf{u}_{r}^{(k)}\right)=\left\langle T,\mathbf{u}_{r}^{(1)}\otimes\ldots\otimes\mathbf{u}_{r}^{(k)}\right\rangle\mathbf{u}_{r}^{(\ell)}, & \ell=1,\ldots,k-1;\\ T\circledast_{k}\left(\mathbf{u}_{r}^{(1)}\otimes\ldots\otimes\mathbf{u}_{r}^{(k-1)}\right)=\sum_{t=1}^{R}\left\langle T,\mathbf{u}_{r}^{(1)}\otimes\ldots\otimes\mathbf{u}_{r}^{(k-1)}\otimes\mathbf{u}_{t}^{(k)}\right\rangle\mathbf{u}_{t}^{(k)}, & r=1,\ldots,R. \end{cases} \tag{5.1}$$

We want to argue that generically an accumulation point is geometrically isolated.

To prove that a solution of a nonlinear system is geometrically isolated, it suffices to prove that the corresponding Jacobian of the system is of full rank at such a point [22, Section 2]. This will be a difficult task for a general nonlinear system, but the theory of algebraic geometry becomes particular useful in our case.

The system (5.1) can be regarded as a square system of polynomials in $R\sum_{\ell=1}^{k}I_{\ell}$ variables, i.e., entries of $\mathbf{u}_{r}^{(\ell)}$, $\ell=1,\ldots,k$, $r=1,\ldots,R$. Each equation in the top subsystem of (5.1) has $\prod_{\ell=1}^{k}I_{\ell}$ many "leading"

[5]We need to stress that the difficulty mentioned here is attributed to the CP formulation. For other models such as DEDICOM, SCA, Dynamals, Parafac-2, INDSCAL, and INDORT where the objective functions assume different forms, variants of ALS-type algorithms have been developed and their dynamics might behave differently. For instance, under some mild assumptions, the infimum of the objective function might be attainable, which deviates from what we have demonstrated in Section 2.2 and might somewhat facilitate the convergence analysis. For a general overview on this subject, see [28].

terms of degree $k+1$ with coefficients from entries of T, while each equation in the bottom subsystem has $R\prod_{\ell=1}^{k} I_\ell$ many leading terms of degree $k+1$. Regarding the tensor T as the parameters of the polynomial system, the follow result is known from the theory of parameter continuation [40, Theorem 7.1.1].

THEOREM 5.1. *Let $\mathcal{N}(T)$ denote the number of geometrically isolated solutions to the corresponding (5.1) over the algebraically closed complex space. Then*

1. *$\mathcal{N}(T)$ is the same, say $\mathcal{N}$, for almost all $T \in \mathbb{C}^{I_1\times\ldots\times I_k}$.*
2. *For all $T \in \mathbb{C}^{I_1\times\ldots\times I_k}$, $\mathcal{N}(T) \leq \mathcal{N}$.*
3. *The subset of $\mathbb{C}^{I_1\times\ldots\times I_k}$ where $\mathcal{N}(T) = \mathcal{N}$ is a Zariski open set, that is, the exceptional subset of tensors $T \in \mathbb{C}^{I_1\times\ldots\times I_k}$ where $\mathcal{N}(T) < \mathcal{N}$ is an affine algebraic set[6] contained within an algebraic set of codimension one.*

We are interested only in real-valued tensors. Since $\mathbb{R}^n$ is Zariski dense in $\mathbb{C}^n$, the above statements hold for almost all $T \in \mathbb{R}^{I_1\times\ldots\times I_k}$, except that the number of real-valued isolated solutions varies as a function of T and is no longer a constant. The latter is an interesting topic in the active and ongoing research area called real algebraic geometry. For our application, we only need the fact of isolation.

COROLLARY 5.2. *For almost all tensors $T \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_k}$, the accumulation points of any sequence generated by Algorithm 1 are necessarily isolated. That is, around any accumulate point there is a neighborhood such that the sequence of iterates from Algorithm 1 has no other accumulate point.*

We will also make use of the following lemma from real analysis [51, Lemma 3.5].

LEMMA 5.3. *Let $\{a_n\}$ be a bounded sequence of real numbers with the property that the difference $a_{n+1} - a_n$ converges to zero as n goes to infinity. If the accumulation points for the sequence are isolated, then the sequence $\{a_n\}$ converges to a unique accumulate point.*

5.2. Diminishing increments. An array of critical questions concerning the convergence of the conventional ALS method applied to tensors has been raised and partially answered in the interesting paper [34, Section 4.3], one of which is that the magnitude of the change

$$\left\| \sum_{r=1}^{R} \lambda_{r,[p]} \mathbf{u}_{r,[p]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p]}^{(k)} - \sum_{r=1}^{R} \lambda_{r,[p+1]} \mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(k)} \right\| \to 0 \tag{5.2}$$

as p goes to infinity. We must point out that such a result was proved for the conventional ALS method only [34, Section 4.3.2]. Our orthogonal ALS scheme is somewhat different. We must reconsider the situation with further details.

Define the individual increments among factors by

$$\delta\mathbf{u}_{r,[p]}^{(\ell)} := \mathbf{u}_{r,[p+1]}^{(\ell)} - \mathbf{u}_{r,[p]}^{(\ell)}, \quad \ell = 1,\ldots,k; \quad r = 1,\ldots,R. \tag{5.3}$$

The identities

$$\left\langle \mathbf{u}_{r,[p+1]}^{(\ell)}, \delta\mathbf{u}_{r,[p]}^{(\ell)} \right\rangle = 1 - \left\langle \mathbf{u}_{r,[p+1]}^{(\ell)}, \mathbf{u}_{r,[p]}^{(\ell)} \right\rangle = \frac{1}{2} \left\| \delta\mathbf{u}_{r,[p]}^{(\ell)} \right\|^2 \tag{5.4}$$

hold for all $\ell = 1,\ldots,k$ and $r = 1,\ldots,R$. Denoting $\lambda_{r,[p+1]}^{(0)} := \lambda_{r,[p]}$, we observe that

$$\begin{aligned}
\lambda_{r,[p+1]}^{(\ell)} &= \left\langle T, \mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(\ell-1)} \otimes \left(\mathbf{u}_{r,[p]}^{(\ell)} + \delta\mathbf{u}_{r,[p]}^{(\ell)} \right) \otimes \mathbf{u}_{r,[p]}^{(\ell+1)} \otimes \ldots \otimes \mathbf{u}_{r,[p]}^{(k)} \right\rangle \\
&= \lambda_{r,[p+1]}^{(\ell-1)} + \left\langle \delta\mathbf{u}_{r,[p]}^{(\ell)}, \underbrace{T \circledast_\ell \mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(\ell-1)} \otimes \widehat{\mathbf{u}}^{(\ell)} \otimes \mathbf{u}_{r,[p]}^{(\ell+1)} \ldots \otimes \mathbf{u}_{r,[p]}^{(k)}}_{\lambda_{r,[p+1]}^{(\ell)} \mathbf{u}_{r,[p+1]}^{(\ell)}} \right\rangle \\
&= \lambda_{r,[p+1]}^{(\ell-1)} + \frac{\lambda_{r,[p+1]}^{(\ell)}}{2} \left\| \delta\mathbf{u}_{r,[p]}^{(\ell)} \right\|^2, \quad \ell = 1,\ldots,k-1; \quad r = 1,\ldots,R.
\end{aligned} \tag{5.5}$$

[6] A subset of affine n-space $\mathbb{A}^n$ over an algebraically closed field k is called an affine algebraic set if it can be written as the zero locus of a set of polynomials. By the Hilbert basis theorem, this set of polynomials can be assumed to be finite. The Zariski topology on $\mathbb{A}^n$ is simply a topology where the closed sets are precisely the algebraic sets in $\mathbb{A}^n$.

By the monotone convergence established in Section 4, we conclude that the sequence $\left\{\delta\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ for each $\ell = 1, \ldots, k-1$ and $r = 1, \ldots, R$ converges to zero as p goes to infinity.

It must be pointed out that the identity (5.5) does not work for the case $\ell = k$. Thus far, nothing is known yet for the sequence $\left\{\delta\mathbf{u}_{r,[p]}^{(k)}\right\}$ because $\mathbf{u}_{r,[p]}^{(k)}$ is generated via the additional step of the polar decomposition, which make it more complicated to estimate $\delta\mathbf{u}_{r,[p]}^{(k)}$ directly. Nonetheless, we have enough information to show that $\delta\mathbf{u}_{r,[p]}^{(k)}$ also converges to zero by the argument in the following section.

5.3. Global convergence. The following theorem completes the last stage of convergence analysis we have been aiming for.

THEOREM 5.4. *For almost all tensors $T \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_k}$, , the sequences $\left\{\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ generated by the orthogonal ALS methods converges for each $\ell = 1, \ldots, k$ and $r = 1, \ldots, R$.*

Proof. Consider first the case when $\ell = 1, \ldots, k-1$ and $r = 1, \ldots, R$. Starting with any initial values $\mathbf{u}_{r,[0]}^{(\ell)}$, it is clear that all iterates $\left\{\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ are bounded. So accumulation points exist. By definition, any accumulation point necessarily satisfies the system of nonlinear equations (5.1) whose solutions are geometrically isolated. We have already argued in Section 5.2 that $\left\{\delta\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ converges to zero as p goes to infinity. It follows that the sequence $\left\{\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ must converge to a single accumulation point because, by Lemma 5.3, each of its components converges. In this way, we establish the global convergence for every element in the subgroup

$$\left\{\left(\mathbf{u}_{1,[p]}^{(1)}, \ldots, \mathbf{u}_{1,[p]}^{(k-1)}\right), \ldots, \left(\mathbf{u}_{R,[p]}^{(1)}, \ldots, \mathbf{u}_{R,[p]}^{(k-1)}\right)\right\}. \tag{5.6}$$

We now argue for the case $\ell = k$. By continuity and the definition of $\mathbf{v}_{r,[p]}^{(k)}$, the convergence of every element in (5.6) implies the global convergence of the sequences $\left\{\mathbf{v}_{r,[p]}^{(k)}\right\}$. Under the generic assumption, the polar decomposition of $\left[\mathbf{v}_{1,[p+1]}^{(k)}, \ldots, \mathbf{v}_{R,[p+1]}^{(k)}\right]$ is unique. The convergence of $\left\{\mathbf{u}_{r,[p]}^{(k)}\right\}$, $r = 1, \ldots, R$ therefore also follows. The global convergence of the orthogonal ALS method for generic tensors is now complete. □

We conclude this section with the following remark. Although we have just established the fact that the orthogonal ALS method converges globally, such a global convergence should not be confused with the possibility that the iterates may converge to a local minimizer of the objective function f defined in (2.8). A limit point obtained by the ALS method depends on the starting point and is not guaranteed to be the best rank-R approximation to T [44]. Questions such as the number of achievable minimizers and starting strategies to achieve these points should be an interesting research topic by themselves.

5.4. Generalization. It is natural to ask the question of generalizing the notion of semi-orthogonality to more than one factor matrix. We have mentioned that the semi-orthogonality of one orthogonal factor matrix is already enough for maintaining the orthogonal condition (3.2). The more restricted constraint by adding extra semi-orthogonal factor matrix might be for other purposes. See, for example, the next section for the SVD.

Consider the scenario where we require that $\langle \mathbf{u}_i^{(s)}, \mathbf{u}_j^{(s)} \rangle = \delta_{ij}$ for $1 \leq i, j \leq R$ and $s = k-1, k$. In this case, it should not be difficult to modify the Lagrangian and re-derive the optimality condition as we have done in Section 3.2. In particular, it might be possible to modify Algorithm 1 by first applying a polar decomposition to the normalized vector of

$$\mathbf{v}_{r,[p+1]}^{(k-1)} = T \circledast_{k-1} \left(\mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(k-2)} \otimes \widehat{\mathbf{u}}^{(k-1)} \otimes \mathbf{u}_{r,[p]}^{(k)}\right)$$

to obtain $\mathbf{u}_{r,[p+1]}^{(k-1)}$ followed by a second polar decomposition applied to the normalized vector of

$$\mathbf{v}_{r,[p+1]}^{(k)} = T \circledast_{k} \left(\mathbf{u}_{r,[p+1]}^{(1)} \otimes \ldots \otimes \mathbf{u}_{r,[p+1]}^{(k-1)}\right)$$

to obtain $\mathbf{u}_{r,[p+1]}^{(k)}$. We conjecture that the global convergence still holds. However, the technique employed in the preceding section might not be immediately generalizable because we need to prove the convergence of both sequences $\left\{\mathbf{v}_{r,[p]}^{(k-1)}\right\}$ and $\left\{\mathbf{v}_{r,[p]}^{(k)}\right\}$ simultaneously. More study is needed.

6. An example by matrices. It should be of interest to compare how the ALS algorithm performs for the case when T is a matrix with or without orthogonality. The fundamental problem is to find the best rank-R approximation

$$\min \left\| T - \sum_{r=1}^{R} \lambda_r \mathbf{u}_r^{(1)} \otimes \mathbf{u}_r^{(2)} \right\|, \tag{6.1}$$

subject to the constraint that variables $\mathbf{u}_r^{(1)}$ and $\mathbf{u}_r^{(2)}$, $r = 1, \ldots, R$, are of unit length[7]. We already know from the Eckart-Young theorem that the TSVD of T is a stationary point and is the absolute best rank-R approximation and that the complete orthogonality is an inherent property. In our present setting, we require that only the matrix $\left[\mathbf{u}_1^{(2)}, \ldots, \mathbf{u}_R^{(2)}\right]$ has orthogonal columns.

Suppose that we proceed to update the first components of the rank-1 matrices via

$$\mathbf{u}_{r,[p+1]}^{(1)} = \frac{1}{\lambda_{r,[p+1]}^{(1)}} T \mathbf{u}_{r,[p]}^{(2)}, \quad r = 1, \ldots, R, \tag{6.2}$$

followed by the intermediate vectors

$$\mathbf{v}_{r,[p+1]}^{(2)} = T^\top \mathbf{u}_{r,[p+1]}^{(1)}, \quad r = 1, \ldots R. \tag{6.3}$$

There are three possible alternatives to proceed from this point on.

1. Had we simply regarded $\mathbf{u}_{r,[p+1]}^{(2)} = \mathbf{v}_{r,[p+1]}^{(2)}$ and repeated steps (6.2) and (6.3) recursively without orthogonalization, then this process would have been equivalent to the simultaneous subspace iteration of the power method applied to the matrices $TT^\top$ and $T^\top T$ on the 2-dimensional subspaces $span\left\{\mathbf{u}_{1,[0]}^{(1)}, \ldots, \mathbf{u}_{R,[0]}^{(1)}\right\}$ and $span\left\{\mathbf{u}_{1,[0]}^{(2)}, \ldots, \mathbf{u}_{R,[0]}^{(2)}\right\}$, respectively. The subspaces eventually would have coalesced into the dominant singular vectors of T and that is not what we prefer to see.
2. In Algorithm 1, we calculate the polar decomposition and define the second components of rank-1 matrices via

$$\left[\mathbf{v}_{1,[p+1]}^{(2)}, \ldots, \mathbf{v}_{R,[p+1]}^{(2)}\right] = \left[\mathbf{u}_{1,[p+1]}^{(2)}, \ldots, \mathbf{u}_{R,[p+1]}^{(2)}\right] P, \tag{6.4}$$

   which orthogonalizes the basis vectors $\left\{\mathbf{v}_{1,[p+1]}^{(2)}, \ldots, \mathbf{v}_{R,[p+1]}^{(2)}\right\}$ and, hence, prevents them from coalescence. Even so, note that our orthogonal ALS method differs from the SVD in that it takes no action to orthogonalize the vectors $\left\{\mathbf{u}_{1,[p+1]}^{(1)}, \ldots, \mathbf{u}_{R,[p+1]}^{(1)}\right\}$. At the end, Algorithm 1 does not guarantee that these $\left\{\mathbf{u}_{r,[p+1]}^{(1)}\right\}$ are mutually orthogonal in general. So the limit point is a local minimizer of (6.1), but is not necessarily the SVD.
3. If we regard that (6.2) only generates $\mathbf{v}_{r,[p+1]}^{(1)}$ and apply the polar decomposition to obtain $\mathbf{u}_{r,[p+1]}^{(1)}$ before proceeding to (6.3), then we generate iterates of mutually orthonormal columns $\mathbf{u}_{1,[p]}^{(1)}, \ldots, \mathbf{u}_{R,[p]}^{(1)}$ and $\mathbf{u}_{1,[p]}^{(2)}, \ldots, \mathbf{u}_{R,[p]}^{(2)}$. This echoes the generalization raised in Section 5.4. The limit of the sequence

[7] A subtle difference between the Kronecker product and the more general tensor product should be clarified. The former emphasizes a specific matrix structure while the latter emphasizes a collection of mixed product (See (1.1)). In particular, the tensor product $\mathbf{u} \otimes \mathbf{v}$ of two column vectors should be identified as the vectorized 2-dimensional array $\mathbf{v}\mathbf{u}^\top$. Throughout this note, we use $\otimes$ to denote the tensor product (1.1). Therefore, we interpret $T \circledast_1 \mathbf{v} = T\mathbf{v}$.

$\sum_{r=1}^{R} \lambda_{r,[p]} \mathbf{u}_{r,[p]}^{(1)} \mathbf{u}_{r,[p]}^{(2)\top}$ is an orthogonal low rank approximation to T, but we see from numerical experiments that it hardly converges to the TSVD of T. The problem of low rank approximation generally has multiple local solutions.

The feasible set of only the second components being mutually orthogonal is larger than the set of complete orthogonality of all components. The more restricted optimization problem generally should give rise to an inferior objective value. That is not the case of low-rank approximation to matrices — the more demanding constraints, i.e., the complete orthogonality, produces the SVD of T which turns out to be the absolute optimal solution. To put it differently, this absolute optimal approximation automatically entails the completely orthogonal decomposition. This remarkable result cannot be easily generalized to high-order tensors.

In fact, for high-order tensors, a completely orthogonal decomposition does not always exist. The complete orthogonalization and the absolute optimal approximation do not necessarily correlate to each other as the SVD for matrices does. We mention in passing that considerable endeavors have been devoted to generalizing Eckart-Young theorem to tensors [5, 9, 12, 21, 24, 25, 26, 31, 32, 33] with a variety of applications [35, 36, 38, 39, 45, 50, 52] and the research is still ongoing.

7. Conclusion. A higher-order tensor does not necessarily have an optimal low rank approximation. By imposing a minimum requirement of mutual orthogonality among just one component of the rank-1 base tensors, the situation is changed to an optimization over a compact set and guarantees the existence of the best orthogonal low rank approximation.

In this paper, we propose a modified alternating least squares algorithm to compute an orthogonal rank-R approximation. Two contributions are new in this work. First, we bring forth the idea of using the polar decomposition to ensure the orthogonality condition while performing the high-order power method. Second, we prove that the orthogonal ALS method converges globally for almost all general tensors.

REFERENCES

[1] V. I. ARNOLD, *Real algebraic geometry*, vol. 66 of Unitext. Translated from the 2009 Russian original by Gerald G. Gould and David Kramer.

[2] S. BASU, R. POLLACK, AND M.-F. ROY, *Algorithms in real algebraic geometry*, vol. 10 of Algorithms and Computation in Mathematics, Springer-Verlag, Berlin, second ed., 2006.

[3] J. CARROLL AND J.-J. CHANG, *Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition*, Psychometrika, 35 (1970), pp. 283–319. 10.1007/BF02310791.

[4] J. CHANG, W. SUN, AND Y. CHEN, *A modified newton's method for best rank-one approximation to tensors*, Applied Mathematics and Computation, 216 (2010), pp. 1859 – 1867. Workshop on Iterative Methods and Preconditioning Techniques.

[5] J. CHEN AND Y. SAAD, *On the tensor SVD and the optimal low rank orthogonal approximation of tensors*, SIAM J. Matrix Anal. Appl., 30 (2008/09), pp. 1709–1734.

[6] L. X. COMON, P. AND A. L. F. DE ALMEIDA, *Tensor decompositions, alternating least squares and other tales*, J. Chemometrics, 23 (2009), pp. 393–405.

[7] P. COMON, G. GOLUB, L.-H. LIM, AND B. MOURRAIN, *Symmetric tensors and symmetric tensor rank*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1254–1279.

[8] P. COMON, J. M. F. TEN BERGE, L. DE LATHAUWER, AND J. CASTAING, *Generic and typical ranks of multi-way arrays*, Linear Algebra Appl., 430 (2009), pp. 2997–3007.

[9] L. DE LATHAUWER, B. DE MOOR, AND J. VANDEWALLE, *A multilinear singular value decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278 (electronic).

[10] ———, *On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342 (electronic).

[11] V. DE SILVA AND L.-H. LIM, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.

[12] P. DRINEAS AND M. W. MAHONEY, *A randomized algorithm for a tensor-based generalization of the singular value decomposition*, Linear Algebra Appl., 420 (2007), pp. 553–571.

[13] N. K. M. FABER, R. BRO, AND P. K. HOPKE, *Recent developments in candecomp/parafac algorithms: a critical review*, Chemometrics and Intelligent Laboratory Systems, 65 (2003), pp. 119 – 137.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.

[15] L. GRIPPO AND M. SCIANDRONE, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136.

[16] R. HARSHMAN, *Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics, 16 (1970).

[17] J. HÅSTAD, *Tensor rank is NP-complete*, J. Algorithms, 11 (1990), pp. 644–654.

[18] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[19] C. J. HILLAR AND L.-H. LIM, *Most tensor problems are NP-hard*, J. ACM, 60 (2013), pp. Art. 45, 39.

[20] F. L. HITCHCOCK, *The expression of a tensor or a polyadic as a sum of products*, J. Mathematics and Physics, (1927), pp. 164–189.

[21] M. ISHTEVA, P.-A. ABSIL, S. VAN HUFFEL, AND L. DE LATHAUWER, *Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 115–135.

[22] H. B. KELLER, *Geometrically isolated nonisolated solutions and their approximation*, SIAM J. Numer. Anal., 18 (1981), pp. 822–838.

[23] H. A. L. KIERS, *Towards a standardized notation and terminology in multiway analysis*, Journal of Chemometrics, 14 (2000), pp. 105–122.

[24] E. KOFIDIS AND P. A. REGALIA, *On the best rank-1 approximation of higher-order supersymmetric tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 863–884.

[25] T. G. KOLDA, *Orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 243–255 (electronic).

[26] ———, *A counterexample to the possibility of an extension of the eckart-young low-rank approximation theorem for the orthogonal rank tensor decomposition*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 762–767.

[27] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[28] W. P. KRIJNEN, *Convergence of the sequence of parameters generated by alternating least squares algorithms*, Comput. Stat. Data Anal., 51 (2006), pp. 481–489.

[29] W. P. KRIJNEN, T. K. DIJKSTRA, AND A. STEGEMAN, *On the non-existence of optimal solutions and the occurrence of "degeneracy" in the CANDECOMP/PARAFAC model*, Psychometrika, 73 (2008), pp. 431–439.

[30] J. B. KRUSKAL, *Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95 – 138.

[31] D. LEIBOVICI AND R. SABATIER, *A singular value decomposition of a k-way array for a principal component analysis of multiway data, pta-k*, Linear Algebra Appl., 269 (1998), pp. 307 – 329.

[32] L.-H. LIM, *Singular values and eigenvalues of tensors: a variational approach*, in Computational Advances in Multi-Sensor Adaptive Processing, 2005 1st IEEE International Workshop on, dec. 2005, pp. 129 –132.

[33] C. D. M. MARTIN AND C. F. VAN LOAN, *A jacobi-type method for computing orthogonal tensor decompositions*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1219–1232.

[34] M. J. MOHLENKAMP, *Musings on multilinear fitting*, Linear Algebra Appl., 438 (2013), pp. 834–852.

[35] D. MUTI AND S. BOURENNANE, *Multidimensional filtering based on a tensor approach*, Signal Process., 85 (2005), pp. 2338–2353.

[36] P. PAATERO, *The multilinear engine: A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model*, Journal of Computational and Graphical Statistics, 8, pp. pp. 854–888.

[37] P. A. REGALIA, *Monotonically convergent algorithms for symmetric tensor approximation*, Linear Algebra Appl., 438 (2013), pp. 875–890.

[38] B. SAVAS AND L. ELDÉN, *Handwritten digit classification using higher order singular value decomposition*, Pattern Recognition, 40 (2007), pp. 993 – 1003.

[39] A. SHASHUA AND A. LEVIN, *Linear image coding for regression and classification using the tensor-rank principle*, in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, vol. 1, 2001, pp. I42–49.

[40] A. J. SOMMESE AND C. W. WAMPLER, II, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.

[41] M. SØRENSEN, L. D. LATHAUWER, P. COMON, S. ICART, AND L. DENEIRE, *Canonical Polyadic Decomposition with a Columnwise Orthonormal Factor Matrix*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 1190–1213.

[42] A. STEGEMAN, *Degeneracy in candecomp/parafac explained for $p \times p \times 2$ arrays of rank $p + 1$ or higher*, Psychometrika, 71 (2006), pp. 483–501.

[43] ———, *Low-rank approximation of generic $p \times q \times 2$ arrays and diverging components in the Candecomp/Parafac model*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 988–1007.

[44] A. STEGEMAN AND P. COMON, *Subtracting a best rank-1 approximation may increase tensor rank*, Linear Algebra Appl., 433 (2010), pp. 1276–1300.

[45] J.-T. SUN, H.-J. ZENG, H. LIU, Y. LU, AND Z. CHEN, *Cubesvd: a novel approach to personalized web search*, in Proceedings of the 14th international conference on World Wide Web, WWW '05, New York, NY, USA, 2005, ACM, pp. 382–390.

[46] J. M. F. TEN BERGE, *Least squares optimization in multivariate analysis*, vol. 25 of M&T Series, D.S.W.O. Press, Leiden, 1993.

[47] L. TUCKER, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311. 10.1007/BF02289464.

[48] A. USCHMAJEW, *Local convergence of the alternating least squares algorithm for canonical tensor approximation*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 639–652.

[49] C. F. VAN LOAN, *The ubiquitous kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.

[50] M. A. O. VASILESCU AND D. TERZOPOULOS, *Multilinear analysis of image ensembles: Tensorfaces*, in Proceedings of the 7th European Conference on Computer Vision-Part I, ECCV '02, London, UK, UK, 2002, Springer-Verlag, pp. 447–460.

[51] L. WANG AND M. T. CHU, *On the global convergence of the alternating least squares method for rank-one tensor approximation*, preprint, North Carolina State University, 2013.

[52] M. WELLING AND M. WEBER, *Positive tensor factorization*, Pattern Recogn. Lett., 22 (2001), pp. 1255–1261.
[53] T. ZHANG AND G. H. GOLUB, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550.