

ALTERNATING MAHALANOBIS DISTANCE MINIMIZATION FOR STABLE AND ACCURATE CP DECOMPOSITION

NAVJOT SINGH* AND EDGAR SOLOMONIK*.

Abstract. CP decomposition (CPD) is prevalent in chemometrics, signal processing, data mining and many more fields. While many algorithms have been proposed to compute the CPD, alternating least squares (ALS) remains one of the most widely used algorithm for computing the decomposition. Recent works have introduced the notion of eigenvalues and singular values of a tensor and explored applications of eigenvectors and singular vectors in areas like signal processing, data analytics and in various other fields. We introduce a new formulation for deriving singular values and vectors of a tensor by considering the critical points of a function different from what is used in the previous work. Computing these critical points in an alternating manner motivates an alternating optimization algorithm which corresponds to alternating least squares algorithm in the matrix case. However, for tensors with order greater than equal to 3, it minimizes an objective function which is different from the commonly used least squares loss. Alternating optimization of this new objective leads to simple updates to the factor matrices with the same asymptotic computational cost as ALS. We show that a subsweep of this algorithm can achieve a superlinear convergence rate for exact CPD with known rank and verify it experimentally. We then view the algorithm as optimizing a Mahalanobis distance with respect to each factor with ground metric dependent on the other factors. This perspective allows us to generalize our approach to interpolate between updates corresponding to the ALS and the new algorithm to manage the tradeoff between stability and fitness of the decomposition. Our experimental results show that for approximating synthetic and real-world tensors, this algorithm and its variants converge to a better conditioned decomposition with comparable and sometimes better fitness as compared to the ALS algorithm.

Key words. tensor decomposition, CP decomposition, alternating least squares, eigenvalues, singular values, Mahalanobis Distance, condition number

AMS subject classifications. 15A69, 15A72, 65K10, 65Y20, 65Y04, 65Y05, 68W25

1. Introduction. The canonical polyadic or CANDECOMP/PARAFAC (CP) tensor decomposition [19, 22] is used for analysis and compression of multi-parameter datasets, and prevalent in tensor methods for scientific simulation [15, 38, 41, 49]. For an order 3 tensor \mathcal{T} , a rank R CP decomposition is

$$\mathcal{T} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!], \quad t_{ijk} = \sum_{r=1}^R a_{ir} b_{jr} c_{kr}.$$

Determining the CP rank or finding an approximate CP decomposition of a tensor, so as to minimize,

$$(1.1) \quad f(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \frac{1}{2} \left\| \mathcal{T} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \right\|_F^2,$$

are NP-hard problems [21]. The CP decomposition of a tensor can be computed via various optimization algorithms, such as alternating least squares [5, 20, 24, 48] which aims to minimize the objective (1.1) in an alternating manner by considering all except one factor matrix fixed. There have been several attempts to improve the performance of ALS algorithm by considering its variations [36, 40, 42, 47]. Several methods which aim to minimize (1.1) with respect to all the factor matrices use gradient-based information [1, 43, 45, 51, 54, 57] to update all the factors simultaneously. Another set of methods optimize for all the factors simultaneously by formulating (1.1) as a nonlinear least squares problem by considering the entries of all the factors as variables. In addition to gradient information, these iterative methods use second order information to compute the next step which requires a system solve [43, 56], and can be achieved by an implicit conjugate gradient algorithm [50, 51].

Tensor eigenvalue problems are relevant in the context of solving multilinear systems, simulating quantum systems, exponential data fitting and many other application areas [46]. However, the study of tensor eigenvalues and tensor singular values is at a relatively nascent stage, [32, 35] provide a definition and introduction to eigenvalues and singular values of a tensor. Computing eigenvalues of a tensor is a hard problem, and can be solved via iterative methods for special cases such as computing a subset of eigenvalues of a tensor [30] or computing the real eigenvalues pairs of a real symmetric tensor [16]. The tensor eigenvalue problem is motivated by applications like blind source separation [7], independent component analysis

*Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 (navjot2@illinois.edu, solomon2@illinois.edu).

(ICA) [23] which also motivate a closely related problem of diagonalizing a tensor. The concept of tensor diagonalization was introduced in [14], where approximate diagonalization of the tensor is considered by minimizing the sum of squares of off-diagonal entries or maximizing the sum of squares of diagonal entries of the tensor. There have been many follow up works [33, 34, 55, 59] which consider approximate diagonalization of the tensor by invertible and orthogonal transformations.

In this work, we introduce a formulation for computing the singular values and vectors of a tensor by considering a logarithmic penalty function instead of Lagrangian variables [35] and computing the critical points of this function. This formulation when generalized to computing invariant subspaces of a matrix, leads to diagonalization of the matrix and can be linked to the singular values and vectors of the matrix. When extended to tensors with order greater than or equal to 3, this formulation leads to another notion of diagonalization of the tensor which is different from the one introduced in prior work. The critical points of this new function spectrally diagonalize the tensor, i.e., the transformed equidimensional tensor of mode length R has R elementary eigenvectors with unit eigenvalues. Computing these stationary points alternatively motivates an alternating optimization algorithm for computing the CP decomposition of a tensor.

1.1. Motivation: Eigenvectors via Lagrangian Optimization. In the case of low-rank matrix approximation, the Eckart-Young-Mirsky theorem shows that the best low-rank approximation may be obtained from the singular value decomposition (SVD). This connection relates low-rank factors to critical points of the bilinear form, $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{A} \mathbf{y}$ with $\|\mathbf{x}\| \neq 0$, $\|\mathbf{y}\| \neq 0$. For tensors of order 3 and higher, tensor singular values have been similarly derived from critical points of multilinear forms. In particular, Lim [35] derives singular vectors and singular values by imposing constraints $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ and considering the critical points of the Lagrangian function. The same results may be obtained by instead considering a logarithmic interior point barrier function for the constraints, $\|\mathbf{x}\| \neq 0$, $\|\mathbf{y}\| \neq 0$, so

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{A} \mathbf{y} - \log(\|\mathbf{x}\| \|\mathbf{y}\|), \quad \nabla f(\mathbf{x}, \mathbf{y}) = \mathbf{0} \Rightarrow \mathbf{A} \mathbf{y} = \mathbf{x}/\|\mathbf{x}\|^2, \quad \mathbf{A}^T \mathbf{x} = \mathbf{y}/\|\mathbf{y}\|^2.$$

Consequently, with $\sigma = 1/(\|\mathbf{x}\| \|\mathbf{y}\|)$ and $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$, $\mathbf{v} = \mathbf{y}/\|\mathbf{y}\|$, we have $\mathbf{A} \mathbf{v} = \sigma \mathbf{u}$ and $\mathbf{A}^T \mathbf{u} = \sigma \mathbf{v}$. The use of a coefficient for the barrier function only affects the scaling of any critical point vectors, \mathbf{x} and \mathbf{y} . For tensors of order 3 and higher, tensor singular values can be similarly derived from critical points of

$$\begin{aligned} f(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) &= \sum_{i_1 \dots i_N} t_{i_1 \dots i_N} x_{i_1}^{(1)} \dots x_{i_N}^{(N)} - \log(\|\mathbf{x}^{(1)}\| \dots \|\mathbf{x}^{(N)}\|), \\ \nabla f(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = \mathbf{0} &\Rightarrow \frac{x_{i_j}^{(j)}}{\|\mathbf{x}^{(j)}\|^2} = \sum_{i_1 \dots \hat{i}_j \dots i_N} t_{i_1 \dots i_N} x_{i_1}^{(1)} \dots \hat{x}_{i_j}^{(j)} \dots x_{i_N}^{(N)}, \end{aligned}$$

where $i \dots \hat{j} \dots k$ implies j is omitted from the sequence. The use of 2-norm in the above definitions leads to l^2 singular vectors [35] and with a symmetric tensor and each $\mathbf{x}^{(i)} = \mathbf{x}^{(j)}$ it yields Z-eigenvectors [32]. Choosing another vector norm in $\{3, \dots, N\}$ leads to other notions of singular vectors and eigenvectors [32].

The only significant known correspondence between tensor singular vectors or eigenvectors and the CP decomposition, is in the case of a rank $R = 1$ CP. In this case, the singular vector with the largest singular value and the best rank-1 approximation coincide (for symmetric tensors, these also correspond to the largest eigenvalue tensor eigenvector). The rank-1 approximation problem is also NP-hard for tensors of order 3 and higher [21]. In this work, motivated by an efficient iterative scheme, we consider an extension of the variational notion of a single singular vector tuple to many.

1.2. Tensor Spectral Diagonalization via Lagrangian Optimization. We denote an inner product of matrices as $\langle \mathbf{X}, \mathbf{Y} \rangle = \langle \text{vec}(\mathbf{X}), \text{vec}(\mathbf{Y}) \rangle$, and similar for tensors \mathbf{X}, \mathbf{Y} . The invariant subspaces of a matrix \mathbf{A} may be obtained by considering the critical points of a generalization of $\mathbf{x}^T \mathbf{A} \mathbf{y} = \langle \mathbf{A}, \mathbf{x} \mathbf{y}^T \rangle$ to the matrix case,

$$f(\mathbf{X}, \mathbf{Y}) = \langle \mathbf{A}, \mathbf{X} \mathbf{Y}^T \rangle, \quad \text{s.t.} \quad \det(\mathbf{X}^T \mathbf{X}) \neq 0, \det(\mathbf{Y}^T \mathbf{Y}) \neq 0.$$

Transforming the inequality constraint into a logarithmic barrier function, we obtain

$$(1.2) \quad \mathcal{L}_f(\mathbf{X}, \mathbf{Y}) = \langle \mathbf{A}, \mathbf{XY}^T \rangle - \frac{1}{2} (\log(\det(\mathbf{X}^T \mathbf{X})) - \log(\det(\mathbf{Y}^T \mathbf{Y})))$$

$$(1.3) \quad = \text{tr}(\mathbf{X}^T \mathbf{AY}) - \frac{1}{2} \text{tr}(\log(\mathbf{X}^T \mathbf{XY}^T \mathbf{Y})).$$

The critical points of \mathcal{L}_f satisfy,

$$\mathbf{AYX}^T \cong \mathbf{I} \text{ and } \mathbf{A}^T \mathbf{XY}^T \cong \mathbf{I}.$$

At a critical point of $f(\mathbf{X}, \mathbf{Y})$, the column span of \mathbf{X} , $\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, must be an invariant subspace of \mathbf{AA}^T , while the columns of \mathbf{Y} span an invariant subspace of $\mathbf{A}^T \mathbf{A}$. These critical points diagonalize \mathbf{A} in the sense that $\mathbf{X}^T \mathbf{AY} = \mathbf{I}$. In the tensor case, a critical point $(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)})$ of

$$(1.4) \quad \begin{aligned} f(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}) &= \langle \mathbf{T}, [\![\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}]\!] \rangle, \text{ s.t. } \det(\mathbf{X}^{(n)T} \mathbf{X}^{(n)}) \neq 0, \forall n \in \{1, \dots, N\}, \\ \mathcal{L}_f(\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}) &= \sum_{r=1}^R \sum_{i_1 \dots i_N} t_{i_1 \dots i_N} x_{i_1 r}^{(1)} \cdots x_{i_N r}^{(N)} - \frac{1}{2} \text{tr}(\log(\mathbf{X}^{(1)T} \mathbf{X}^{(1)} \cdots \mathbf{X}^{(N)T} \mathbf{X}^{(N)})) \end{aligned}$$

gives invariant subspaces of the tensor in the sense that, $\forall i \in \{1, \dots, N\}$,

$$\forall \mathbf{v} \in \text{span}\{\bigotimes_{j \neq i} \mathbf{x}_1^{(j)}, \dots, \bigotimes_{j \neq i} \mathbf{x}_R^{(j)}\}, \quad \mathbf{T}_{(i)} \mathbf{v} \in \text{span}\{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_R^{(i)}\},$$

where $\mathbf{T}_{(i)}$ is the mode- i matricization (unfolding) of the tensor \mathbf{T} .

Since the reconstructed tensor $\tilde{\mathbf{T}}$, where $\tilde{\mathbf{T}} = [\![\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(N)}]\!]$, $\mathbf{Y}^{(n)} = \mathbf{X}^{(n)\dagger T}$, $\forall n \in \{1, \dots, N\}$, captures the action of $\mathbf{T}_{(i)}$ on an invariant subspace, the application of each matricization may be performed with bounded backward error. In Section 5.1.1, we show that the backward error is bounded by $\|\mathbf{T}_{(i)} \mathbf{z}^\perp\|$, where \mathbf{z}^\perp is the projection of \mathbf{z} onto the orthogonal complement of column span of $\bigodot_{j=1, j \neq n}^N \mathbf{X}^{(j)}$. We show that this bound also holds for ALS, and in general for a family of algorithms based on alternating minimization of Mahalanobis distance [12] between the input and reconstructed tensor.

Another observation regarding the critical points of (1.4) is that each matricization of the tensor reconstructed from a critical point, $\mathbf{X} = [\![\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}]\!]$ is a right inverse of the corresponding matricization of the input tensor \mathbf{T} when CP rank is equal to the mode length and more generally,

$$\mathbf{T}_{(i)} \mathbf{X}_{(i)}^T = \mathbf{\Pi}^{(i)}, \quad \forall i \in \{1, \dots, N\},$$

where each $\mathbf{\Pi}^{(i)}$ is a projector onto the column space of $\mathbf{X}^{(i)}$. This \mathbf{X} satisfies some but not all of the properties of previously proposed generalizations of the Moore-Penrose inverse to tensors [34, 52].

Further, the critical point gives a transformation that spectrally diagonalizes \mathbf{T} ,

$$z_{j_1 \dots j_N} = \sum_{i_1 \dots i_N} t_{i_1 \dots i_N} x_{i_1 j_1}^{(1)} \cdots x_{i_N j_N}^{(N)},$$

so that \mathbf{Z} has R eigenvectors that are elementary vectors with unit eigenvalues (for any tensor eigenvector/eigenvalue definition, i.e., l^p eigenvector for any choice of p [32]), since

$$z_{j_k j_k \dots j_p j_k \dots j_k} = \delta_{j_k j_p} \text{ for all } p \neq k \in \{1, \dots, N\}.$$

Beyond these properties, we show that $\mathbf{X}^{(1)\dagger T}, \dots, \mathbf{X}^{(N)\dagger T}$ may be used to obtain an exact CP decomposition or an effective low-rank approximate CP decomposition.

1.3. Alternating Optimization Method. The critical points defined above may be computed efficiently by a method similar to ALS. ALS solves a set of overdetermined linear equations at each step to minimize Frobenius norm error relative to one factor, e.g., it solves for \mathbf{A} in

$$(\mathbf{C} \odot \mathbf{B}) \mathbf{A}^T \cong \mathbf{T}_{(1)}^T.$$

The update rule for each subproblem, may be written as a product of the pseudoinverse of the Khatri-Rao product of two factors and an unfolding of the tensor, i.e.,

$$\mathbf{A} = \mathbf{T}_{(1)}(\mathbf{C} \odot \mathbf{B})^{\dagger T}.$$

Some of the major advantages of the ALS algorithm is its guaranteed monotonic decrease in residual, low per-iteration computational cost, and amenability to parallelization. It has been shown that ALS achieves linear local convergence to minima of the CP residual norm [58].

To obtain a critical point in the high-order tensor function (1.4), we propose a different alternating update scheme, which finds the solution \mathbf{U} to the linear least squares problem,

$$(\mathbf{T}_{(1)}(\mathbf{W} \odot \mathbf{V}))\mathbf{U}^T \cong \mathbf{I}.$$

With $\mathbf{A}^T = \mathbf{U}^\dagger$, $\mathbf{B}^T = \mathbf{V}^\dagger$, and $\mathbf{C}^T = \mathbf{W}^\dagger$, we observe that the update is similar to that of ALS,

$$\mathbf{A} = \mathbf{T}_{(1)}(\mathbf{C}^{\dagger T} \odot \mathbf{B}^{\dagger T}).$$

A stationary point of this alternating update scheme provides a critical point of (1.4).

1.4. Convergence Results. The new alternating update scheme is highly effective at finding an exact CP decomposition, if one exists. In particular, we show that the method achieves a superlinear rate of local convergence to exact CP decompositions. In Section 4, we prove that the convergence order is α per subproblem or α^N per sweep of alternating updates, where α is the unique real root of the polynomial $x^{N-1} - \sum_{i=0}^{N-2} x^i$. For $N = 3$, $\alpha = (1 + \sqrt{5})/2$, while for higher N , α increases. A superlinear convergence rate for CP decomposition is also achievable via general optimization algorithms such as Gauss-Newton [50, 51]. However, the alternating optimization scheme we propose has a much lower per iteration cost (about the same as ALS, which achieves only linear convergence).

For a given tensor and any choice of rank, the critical points are generally not unique (as in the case of matrices). Theoretical characterization of the conditions under which critical points of (1.4) exist in this scenario remains an open problem as it requires proving existence of roots of a system of nonlinear equations that have the same number of variables and equations. Note that the problem of proving if the best CP rank approximation exists also requires existence of a solution of system of nonlinear equations. The best CP rank approximation may not exist, which has lead to the notion of border rank [29]. Consequently, establishing existence of critical points for our scenario is likely also nontrivial. Therefore, with the assumption that a critical point exists, we show in Section 4.1 that the proposed iterative scheme achieves local convergence to it (in this case, at a linear rate).

We perform numerical experiments in Section 6 to confirm the rate of convergence of the algorithm for computing CP decomposition of different tensors with known rank. The observed rate of convergence values agree with the theoretical rate of convergence with an error of about 0.2% for order 3 and an error of about 1% for order 4 tensors. The experiments also confirm the convergence analysis of the algorithm to stationary points for approximate decomposition of tensors as described in Lemma 4.4

1.5. Generalizations and Experimental Evaluation. The proposed algorithm may also be used for approximate CP decomposition, but does not minimize the Frobenius norm of the residual directly. Instead, when optimizing for \mathbf{A} , the algorithm minimizes

$$\|(\mathbf{I} \otimes \mathbf{B}^\dagger \otimes \mathbf{C}^\dagger) \text{vec}(\mathbf{T} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!])\|_F.$$

For a fixed residual error in the decomposition, the magnitude of this error metric will generally depend on the conditioning of \mathbf{A} , \mathbf{B} , and \mathbf{C} . Hence, this alternating minimization procedure tends to converge to well-conditioned factors (and well-conditioned CP decompositions [9, 60]).

We generalize this method by considering a Mahalanobis distance between the input and reconstructed tensor. The original motivation for Mahalanobis distance minimization of tensors came from the work on minimization of Wasserstein distance between tensors for nonnegative CP decomposition [2]. This generalization allows us to reformulate each update to a factor of the above introduced algorithm as a minimizer of a Mahalanobis distance [12] with the ground metric dependent on the other remaining factors. This reformulation helps extend the introduced algorithm to any CP rank by using the same ground metric. Moreover, we

are able to interpolate between the introduced algorithm and ALS by interpolating the ground metric. Our experiments in Section 6 suggest that the interpolated updates help manage the trade-off between fitness and conditioning of the decomposition. We measure conditioning of the decomposition by computing the normalized CPD condition number [9]. The condition number can be computed in an efficient manner for decompositions with small CP rank by compressing the matrix for which the smallest singular value needs to be computed and henceforth reducing the computational cost significantly as described in Appendix A. By using this efficient approach, we are able to track condition number of the decomposition at each iteration of the algorithms. For synthetic as well as real-world tensors, we observe that by utilizing hybrid updates of the introduced algorithm, we can find decompositions with a condition number lower by a factor as large as 10^4 with a change in fitness of about 0.01 only when compared to ALS.

2. Background. We introduce the notation and definitions used in the subsequent sections here along with a brief introduction to the alternating least squares algorithm for computing CP decomposition [11, 19].

2.1. Notation and Definitions. We use tensor algebra notation in both element-wise form and specialized form for tensor operations [29]. For vectors, bold lowercase Roman letters are used, e.g., \mathbf{x} . For matrices, bold uppercase Roman letters are used, e.g., \mathbf{X} . For tensors, bold calligraphic fonts are used, e.g., \mathcal{X} . An order N tensor corresponds to an N -dimensional array with dimensions $s_1 \times \dots \times s_N$. Elements of vectors, matrices, and tensors are denoted in subscript, e.g., x_i for a vector \mathbf{x} , x_{ij} for a matrix \mathbf{X} , and x_{ijkl} for an order 4 tensor \mathcal{X} . The i th column of a matrix \mathbf{X} is denoted by \mathbf{x}_i . The mode- n matrix product of a tensor $\mathcal{X} \in \mathbb{R}^{s_1 \times \dots \times s_N}$ with a matrix $\mathbf{A} \in \mathbb{R}^{J \times s_n}$ is denoted by $\mathcal{X} \times_n \mathbf{A}$, with the result having dimensions $s_1 \times \dots \times s_{n-1} \times J \times s_{n+1} \times \dots \times s_N$. Matricization is the process of reshaping a tensor into a matrix. Given a tensor \mathcal{X} the mode- n matricized version is denoted by $\mathbf{X}_{(n)} \in \mathbb{R}^{s_n \times K}$ where $K = \prod_{m=1, m \neq n}^N s_m$. We use parenthesized superscripts as labels for different tensors and matrices, e.g., $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are different matrices.

The Hadamard product of two matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{I \times J}$ resulting in matrix $\mathbf{W} \in \mathbb{R}^{I \times J}$ is denoted by $\mathbf{W} = \mathbf{U} * \mathbf{V}$, where $w_{ij} = u_{ij}v_{ij}$. The inner product of matrices \mathbf{U}, \mathbf{V} is denoted by $\langle \mathbf{U}, \mathbf{V} \rangle = \sum_{i,j} u_{ij}v_{ij}$. The outer product of K vectors $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(K)}$ of corresponding sizes s_1, \dots, s_K is denoted by $\mathcal{X} = \mathbf{u}^{(1)} \circ \dots \circ \mathbf{u}^{(K)}$ where $\mathcal{X} \in \mathbb{R}^{s_1 \times \dots \times s_K}$ is an order K tensor.

For matrices $\mathbf{A} \in \mathbb{R}^{I \times K} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$ and $\mathbf{B} \in \mathbb{R}^{J \times K} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$, their Khatri-Rao product resulting in a matrix of size $(IJ) \times K$ defined by $\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1, \dots, \mathbf{a}_K \otimes \mathbf{b}_K]$, where $\mathbf{a} \otimes \mathbf{b}$ denotes the Kronecker product of the two vectors. We define the Mahalanobis norm for a matrix \mathbf{A} with ground metric \mathbf{M} as $\|\mathbf{A}\|_{\mathbf{M}} = \sqrt{\text{vec}(\mathbf{A})^T \mathbf{M} \text{vec}(\mathbf{A})}$ and similarly for tensor \mathcal{T} , $\|\mathcal{T}\|_{\mathbf{M}} = \text{vec}(\mathcal{T})^T \mathbf{M} \text{vec}(\mathcal{T})$. To ease the notation for N khatri Rao products, we use $\bigodot_{n=1}^N = \mathbf{A}^{(N)} \odot \dots \odot \mathbf{A}^{(1)}$ and similarly $\bigotimes_{n=1}^N \mathbf{A}^{(n)} = \mathbf{A}^{(N)} \otimes \dots \otimes \mathbf{A}^{(1)} *_{n=1}^N \mathbf{A}^{(n)} = \mathbf{A}^{(N)} * \dots * \mathbf{A}^{(1)}$. We use $\sigma_{\min}(\mathbf{P})$ to denote the minimum singular value of the matrix \mathbf{P} .

2.2. Alternating least squares for CP decomposition. The CP tensor decomposition [19, 22] for an input tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ is denoted by

$$\mathcal{X} \approx [[\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]], \quad \text{where } \mathbf{A}^{(i)} = [\mathbf{a}_1^{(i)}, \dots, \mathbf{a}_r^{(i)}],$$

and serves to approximate a tensor by a sum of R tensor products of vectors,

$$\mathcal{X} \approx \sum_{r=1}^R \mathbf{a}_r^{(1)} \circ \dots \circ \mathbf{a}_r^{(N)}.$$

It is sometimes useful to normalize the factor matrices so that each column of the factors has a unit 2-norm and the weights are absorbed into a vector $\boldsymbol{\lambda} \in \mathbb{R}^R$, given as,

$$\mathcal{X} \approx \sum_{r=1}^R \mathbf{a}_r^{(1)} \circ \dots \circ \mathbf{a}_r^{(N)} = \sum_{r=1}^R \lambda_r \bar{\mathbf{a}}_r^{(1)} \circ \dots \circ \bar{\mathbf{a}}_r^{(N)},$$

where $\bar{\mathbf{A}}^{(n)}$ are column normalized for all n and denoted as

$$\mathcal{X} \approx [[\boldsymbol{\Lambda}; \bar{\mathbf{A}}^{(1)}, \dots, \bar{\mathbf{A}}^{(N)}]],$$

where Λ is a diagonal matrix with λ on the diagonal. The CP-ALS method aims to minimize the nonlinear least squares problem

$$(2.1) \quad f(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}) := \frac{1}{2} \left\| \mathbf{\mathcal{X}} - [\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!] \right\|_F^2,$$

by alternatively minimizing a sequence of least squares problems for each of the factor matrices $\mathbf{A}^{(n)}$. This results in linear least squares problems for each row,

$$\mathbf{A}_{\text{new}}^{(n)} \mathbf{P}^{(n)T} \cong \mathbf{X}_{(n)},$$

where the matrix $\mathbf{P}^{(n)} \in \mathbb{R}^{D_n \times R}$, where $D_n = \prod_{i=1, i \neq n}^N I_i$ is formed by Khatri-Rao products of the other factor matrices,

$$(2.2) \quad \mathbf{P}^{(n)} = \bigodot_{m=1, m \neq n}^N \mathbf{A}^{(m)}.$$

These linear least squares problems are often solved via the normal equations [29],

$$\mathbf{A}_{\text{new}}^{(n)} \mathbf{\Gamma}^{(n)} \leftarrow \mathbf{X}_{(n)} \mathbf{P}^{(n)},$$

where $\mathbf{\Gamma} \in \mathbb{R}^{R \times R}$ can be computed via

$$(2.3) \quad \mathbf{\Gamma}^{(n)} = \bigast_{i=1, i \neq n}^N \mathbf{S}^{(i)},$$

with each $\mathbf{S}^{(i)} = \mathbf{A}^{(i)T} \mathbf{A}^{(i)}$. The *Matricized Tensor Times Khatri-Rao Product* or MTTKRP computation $\mathbf{M}^{(n)} = \mathbf{X}_{(n)} \mathbf{P}^{(n)}$ is the main computational bottleneck of CP-ALS [4]. For a rank-R CP decomposition, this computation has the cost of $O(I^N R)$ if $I_n = I$ for all $n \in \{1, \dots, N\}$. There have been various developments to optimize computation of MTTKRP, like dimension-tree algorithm [3, 25–27, 44, 61] for dense tensors and sparse MTTKRP [13] for sparse tensors.

Algorithm 3.1 Basic description of the new alternating update scheme.

```

1: Input: Tensor  $\mathbf{\mathcal{X}} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ , rank  $R$ 
2: Initialize  $\{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}\}$  so each  $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$  is random
3: while not converged do
4:   for  $n \in \{1, \dots, N\}$  do
5:      $\mathbf{A}^{(n)} = \mathbf{X}_{(n)}^{(N)} \left( \bigodot_{m=1, m \neq n}^N \mathbf{A}^{(m)\dagger T} \right)$ 
6:   end for
7: end while
8: return factor matrices  $\{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}\}$ 
```

3. Basic Description of the New Algorithm. We first provide a complete description of the alternating update scheme proposed in the introduction. To compute the decomposition of a tensor $\mathbf{\mathcal{X}}$, the algorithm maintains a CP decomposition given by

$$[\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!],$$

and updates each $\mathbf{A}^{(n)}$ in an alternating manner,

$$(3.1) \quad \mathbf{A}^{(n)} = \mathbf{X}_{(n)} \left(\bigodot_{m=1, m \neq n}^N \mathbf{A}^{(m)\dagger T} \right).$$

This update may be written in elementwise form in terms of the pseudoinverses $\mathbf{U}^{(m)} = \mathbf{A}^{(m)\dagger}$ as

$$a_{i_n r}^{(n)} = \sum_{\substack{i_1 \dots i_N \\ i_1 \dots \hat{i}_n \dots i_N}} x_{i_1 \dots i_N} \prod_{m=1, m \neq n}^N u_{r i_m}^{(m)}.$$

Algorithm 3.1 details each sweep of such updates. Like with the ALS, it is advisable to recalibrate the norms of the columns of each factor before the subsweep corresponding to n th factor so that $\|\mathbf{a}_r^{(k)}\| = 1$ for all $k \neq n$ and r . With this recalibration, a valid convergence criteria is to check whether the magnitude of change in the factors exceeds a predefined threshold at each sweep. The method is invariant to the rescaling in exact arithmetic, but this calibration helps reduce the effects of round-off error. This calibration is also cost efficient, since pseudoinverse of only one matrix changes per subweep. This makes the algorithm accessible to all the optimizations involved in computing *Matricized Tensor Times Khatri-Rao Product* or MTTKRP in each subsweep of ALS such as the dimension tree algorithm [3, 25–27, 44, 61].

3.1. Cost Analysis. The cost of each sweep of Algorithm 3.1 corresponds to the cost of computing the pseudoinverse of each factor, as well as a set of N MTTKRP operations. A dimension tree or multi-sweep dimension tree [37] may be used to compute the set of MTTKRPs in the same way as done in the alternating least squares algorithm. The overall per-sweep cost with a multi-sweep dimension tree is then given by

$$\frac{2N}{N-1} \prod_{n=1}^N I_n R + O\left(\left(\sum_{n=1}^N I_n\right)R^2\right).$$

The cost of an ALS sweep with a multi-sweep dimension tree is

$$\frac{2N}{N-1} \prod_{n=1}^N I_n R + O(NR^3),$$

which is less expensive as solving an overdetermined system via normal equations is cheaper than computing the pseudoinverse of a matrix. If \mathbf{X} is sparse, the method can benefit from existing work on efficiently performing MTTKRP with a sparse tensor [13] and therefore has the same leading order cost per-sweep as that of ALS for sparse tensors as well.

4. Convergence Rate for Exact Decomposition. In this section, we theoretically analyze the asymptotic rate of local convergence of the Algorithm 3.1 for when an exact CP decomposition of rank R less than equal to the length of all modes of the tensor exists. To derive the rate of convergence for an exact decomposition, we relate the distance between the computed factor in each subproblem and the true factor, to the error in the other factor matrices. The following lemma states the error in computing $\mathbf{A}^{(N)}$, but can be trivially extended to any $\mathbf{A}^{(n)}$ for $n \in \{1, \dots, N\}$. We consider the error in the normalized factor and the error in the magnitude of CPD components modulo column scaling.

LEMMA 4.1. Suppose $\mathbf{X} = [\mathbf{D}; \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]$, where each $\mathbf{A}^{(i)} \in \mathbb{R}^{s_i \times R}$ with $s_i \geq R$ is full rank and has normalized columns, i.e., $\|\mathbf{a}_j^{(i)}\|_2 = 1$ for all i, j and $\bar{\mathbf{A}}^{(n)} = \mathbf{A}^{(n)} + \Delta^{(n)}$ also has normalized columns and satisfies $\|\Delta^{(n)}\|_F = \epsilon_n$ for $n = 1, \dots, N-1$, then $\exists \epsilon > 0$ such that if $\epsilon_n < \epsilon$ for $n = 1, \dots, N-1$, then

$$\tilde{\mathbf{A}}^{(N)} = \mathbf{X}_{(N)} (\bar{\mathbf{A}}^{(1)\dagger T} \odot \dots \odot \bar{\mathbf{A}}^{(N-1)\dagger T})$$

where $\tilde{\mathbf{A}}^{(N)} = \bar{\mathbf{A}}^{(N)} \bar{\mathbf{D}}$ ensures that $\bar{\mathbf{A}}^{(N)}$ are normalized, and satisfies

$$\begin{aligned} \|\bar{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\|_F &= O\left(\prod_{n=1}^N \epsilon_n^{N-1}\right), \\ \text{and } \|\bar{\mathbf{D}} - \mathbf{D}\|_F &= O(\epsilon). \end{aligned}$$

Proof. Let $\epsilon < 1$ and $\epsilon < \min_n(\sigma_{\min}(\mathbf{A}^{(n)}))$ for each n , and therefore $\bar{\mathbf{A}}^{(n)}$ is full rank for each $n = 1, \dots, N - 1$. Substituting the decomposition of \mathbf{X} into the computed solution, we obtain

$$\begin{aligned}\tilde{\mathbf{A}}^{(N)} &= \mathbf{A}^{(N)} \mathbf{D} \left((\bar{\mathbf{A}}^{(1)\dagger} (\bar{\mathbf{A}}^{(1)} - \Delta^{(1)})) * \dots * (\bar{\mathbf{A}}^{(N-1)\dagger} (\bar{\mathbf{A}}^{(N-1)} - \Delta^{(N-1)})) \right)^T \\ &= \mathbf{A}^{(N)} \mathbf{D} \left((\mathbf{I} - \bar{\mathbf{A}}^{(1)\dagger} \Delta^{(1)}) * \dots * (\mathbf{I} - \bar{\mathbf{A}}^{(N-1)\dagger} \Delta^{(N-1)}) \right)^T \\ &= \mathbf{A}^{(N)} \mathbf{D} \left(\mathbf{S} + (-1)^{N-1} \bar{\mathbf{A}}^{(1)\dagger} \Delta^{(1)} * \dots * \bar{\mathbf{A}}^{(N-1)\dagger} \Delta^{(N-1)} \right)^T.\end{aligned}$$

where \mathbf{S} includes all cross-terms of the Hadamard products, which must be diagonal since any such term includes a Hadamard product with an identity matrix. Since,

$$\|\mathbf{I} - \mathbf{S}\|_F = O(\max_n(\epsilon_n)) = O(\epsilon),$$

\mathbf{S} is full rank for sufficiently small ϵ . Let

$$\Delta = \left((-1)^{N-1} \bar{\mathbf{A}}^{(1)\dagger} \Delta^{(1)} * \dots * \bar{\mathbf{A}}^{(N-1)\dagger} \Delta^{(N-1)} \right)^T.$$

Now, the norm calibration diagonal matrix is defined so that $\bar{d}_{ii} = \|\tilde{\mathbf{a}}_i^{(N)}\|_2$. Since,

$$\tilde{\mathbf{A}}^{(N)} = \mathbf{A}^{(N)} \mathbf{D} (\mathbf{S} + \Delta) = \mathbf{A}^{(n)} \mathbf{D} + \mathbf{A}^{(n)} \mathbf{D} (\mathbf{S} + \Delta - \mathbf{I}),$$

and $\|\mathbf{S} + \Delta - \mathbf{I}\|_2 = O(\epsilon)$, we have

$$\bar{d}_{ii} = \|\tilde{\mathbf{a}}_i^{(N)}\|_2 \leq \|\mathbf{a}_i^{(N)}\|_2 d_{ii} + \|\mathbf{A}^{(N)} \mathbf{D}\|_F \|\mathbf{S} + \Delta - \mathbf{I}\|_2 = \|\mathbf{a}_i^{(N)}\|_2 d_{ii} + O(\epsilon) = d_{ii} + O(\epsilon).$$

Consequently, $\|\bar{\mathbf{D}} - \mathbf{D}\|_2 = O(\epsilon)$. Further, we can obtain a tighter bound (in terms of $O(\|\Delta\|_F)$ instead of $O(\epsilon)$) by considering, $\bar{\mathbf{A}}^{(N)} = \mathbf{A}^{(N)} \mathbf{D} \mathbf{S} (\mathbf{I} + \mathbf{S}^{-1} \Delta)$, so

$$\bar{d}_{ii} = \|\tilde{\mathbf{a}}_i^{(N)}\|_2 \leq \|\mathbf{a}_i^{(N)}\|_2 d_{ii} s_{ii} + \|\mathbf{A}^{(N)} \mathbf{D} \Delta\|_F = d_{ii} s_{ii} + O(\|\Delta\|_F).$$

This bound allows us to get the desired result for the error in the factor matrices,

$$\begin{aligned}(4.1) \quad \|\bar{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\|_F &= \|\tilde{\mathbf{A}}^{(N)} \bar{\mathbf{D}}^{-1} - \mathbf{A}^{(N)}\|_F = \|\mathbf{A}^{(N)} \mathbf{D} \mathbf{S} (\mathbf{I} + \mathbf{S}^{-1} \Delta) \bar{\mathbf{D}}^{-1} - \mathbf{A}^{(N)}\|_F \\ &= O(\|\mathbf{I} - \mathbf{D} \mathbf{S} (\mathbf{I} + \mathbf{S}^{-1} \Delta) \bar{\mathbf{D}}^{-1}\|_F).\end{aligned}$$

Since, we have that $\|\bar{\mathbf{D}} - \mathbf{D} \mathbf{S}\|_F = O(\|\Delta\|_F)$,

$$\begin{aligned}\|\mathbf{I} - \mathbf{D} \mathbf{S} (\mathbf{I} + \mathbf{S}^{-1} \Delta) \bar{\mathbf{D}}^{-1}\|_F &= \|\mathbf{I} - \mathbf{D} \mathbf{S} \bar{\mathbf{D}}^{-1} + \mathbf{D} \Delta \bar{\mathbf{D}}^{-1}\|_F \\ &= \|\bar{\mathbf{D}} \bar{\mathbf{D}}^{-1} - \mathbf{D} \mathbf{S} \bar{\mathbf{D}}^{-1} + \mathbf{D} \Delta \bar{\mathbf{D}}^{-1}\|_F \\ &= O(\|\Delta\|_F) + \|\mathbf{D} \Delta \bar{\mathbf{D}}^{-1}\|_F = O(\|\Delta\|_F).\end{aligned}$$

Since, $\|\Delta\|_F = O\left(\prod_{n=1}^N \epsilon_n^{N-1}\right)$, this completes the proof. \square

The above Lemma states that in Algorithm 3.1, the error in the updated CPD factor relative to the true CPD factor is bounded by the product of errors in the previous $N - 1$ factors. Using this error bound, we derive convergence rate for Algorithm 3.1.

LEMMA 4.2. *For any algorithm where the error in the update is of the order of product of error in previous k updates, the rate of convergence is equal to the positive root of the polynomial $\alpha^k - \sum_{i=0}^{k-1} \alpha^i$.*

Proof. Let the error at the n th iteration be given as e_n . The error at the n th iteration then satisfies the following in the worst case,

$$(4.2) \quad e_n = L \prod_{i=1}^k e_{n-i}, \quad \text{where } L \text{ is a constant.}$$

The above recurrence can be solved by assuming that the error satisfies the following asymptotic relation

$$(4.3) \quad e_n = Ce_{n-1}^\alpha,$$

where C is some constant and α is the rate of convergence. From (4.2) and (4.3),

$$\begin{aligned} Ce_{n-1}^\alpha &= L \prod_{i=1}^k e_{n-i}, \\ \frac{C^{\sum_{i=0}^{k-1} \frac{1}{\alpha^i}}}{L} &= e_{n-1}^{(-\alpha + \sum_{i=0}^{k-1} \frac{1}{\alpha^i})}. \end{aligned}$$

Since the left hand side is constant for $n \rightarrow \infty$,

$$\alpha^k - \sum_{i=0}^{k-1} \alpha^i = 0.$$

□

Now, with all the pieces together we can show that for exact CPD, Algorithm 3.1 locally converges at a rate which is given in the following theorem.

THEOREM 4.3. *Suppose $\mathbf{X} = [\mathbf{D}; \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]$, where each $\mathbf{A}^{(i)} \in \mathbb{R}^{s_i \times R}$ with $s_i \geq R$ is full rank and has normalized columns. Algorithm 3.1 for computing the exact CP decomposition of \mathbf{X} converges locally with a rate of convergence equal to α^N where α is the unique real root of the polynomial $x^{N-1} - \sum_{i=0}^{N-2} x^i$.*

Proof. Consider the computation of the CP decomposition of the tensor \mathbf{X} with exact rank R and initial guess $\bar{\mathbf{A}}^{(n)}$ with normalized columns such that $\bar{\mathbf{A}}^{(n)} = \mathbf{A}^{(n)} + \Delta^{(n)}$, with $\|\Delta^{(n)}\|_2 < \epsilon$ sufficiently small for $n = 1, \dots, N-1$ (as described in Lemma 4.1). Let the error in CPD at n th iteration be given as

$$E_n = \max\{\|\bar{\mathbf{D}} - \mathbf{D}\|_F, \|\bar{\mathbf{A}}^{(1)} - \mathbf{A}^{(1)}\|_F, \dots, \|\bar{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\|_F\}.$$

Also, let the error in k th subiteration of the n th iteration be given as ϵ_k . From Lemma 4.1, we know that the error in CPD is bounded by the maximum error in factor matrices. Since the error decreases at each subiteration, $\exists n$ such that $E_n = O(\epsilon_1)$.

From Lemma 4.1, we know that the error in a subsweep of the algorithm modulo the column scaling is of the order of product of errors in previous $N-1$ subsweeps, therefore the error at $(n+1)$ th iteration is bounded by the error in the first factor matrix, given by

$$E_{n+1} = O\left(\prod_{k=0}^{N-2} \epsilon_{N-k}\right).$$

By using Lemma 4.2, we know that the error in a subiteration is given by the following recurrence, where α is the positive root of $x^{N-1} - \sum_{i=0}^{N-2} x^i$,

$$\epsilon_{k+1} = O(\epsilon_k^\alpha) \quad \text{for all } k = 1, \dots, N.$$

Therefore, the error at $(n+1)$ th iteration of Algorithm 3.1 can be expressed as

$$E_{n+1} = O\left(\prod_{i=1}^{N-1} \epsilon_1^{\alpha^i}\right) = O(E_n^{\sum_{i=1}^{N-1} \alpha^i}).$$

Using the fact that α is a root of the polynomial $x^{N-1} - \sum_{i=0}^{N-2} x^i$, implies that $\alpha^{N-1} - \sum_{i=0}^{N-2} \alpha^i = 0$, that is, $\sum_{i=1}^{N-1} \alpha^i = \alpha^N$. Therefore,

$$E_{n+1} = O\left(E_n^{\alpha^N}\right).$$

□

This completes the proof to show that Algorithm 3.1 locally converges superlinearly for exact CP rank cases. We verify our theoretical results in the Section 6.

4.1. Convergence to Other Stationary Points. The result in Theorem 4.3 can be generalized to the case where a tensor \mathbf{X} can be represented as the sum of two tensors, \mathbf{T} and \mathbf{E} , where \mathbf{T} has an underlying CPD structure of rank R and \mathbf{E} has a CP decomposition that is mostly orthogonal to the decomposition of \mathbf{T} . For such an input tensor \mathbf{X} , we show that Algorithm 3.1 with CP rank R , locally converges to the underlying CP factors, provided that \mathbf{T} is associated with a stationary point exists. We analyze the convergence rate of the algorithm and show that this is a generalization of the previous result, since we converge to a subset of the CP factors with same convergence rate as in Theorem 4.3, if the factors of \mathbf{T} are in the orthogonal complement of the column space of corresponding CP factors of \mathbf{E} .

LEMMA 4.4. *For a given tensor \mathbf{X} , assume there exists a stationary point of Algorithm 3.1, yielding positive diagonal matrix \mathbf{D} and factors $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)})$ where each $\mathbf{A}^{(i)} \in \mathbb{R}^{s_i \times R}$ with $s_i \geq R$ is full rank and has normalized columns, i.e., $\|\mathbf{a}_j^{(i)}\|_2 = 1$ for all i, j , and their pseudoinverse-transposes $(\mathbf{U}^{(1)}, \dots, \mathbf{U}^{(N)})$, so $\mathbf{A}^{(n)\dagger} = \mathbf{U}^{(n)T}$. The stationary point conditions imply that $\forall n \in \{1, \dots, N\}$, we have*

$$\mathbf{A}^{(n)}\mathbf{D} = \mathbf{X}_{(n)} \left(\bigodot_{m=1, m \neq n}^N \mathbf{U}^{(m)} \right).$$

Further, assume that $\mathbf{X} = [\![\mathbf{D}; \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!] + [\![\hat{\mathbf{A}}^{(1)}, \dots, \hat{\mathbf{A}}^{(N)}]\!]$ with $\|\hat{\mathbf{A}}^{(n)T}\mathbf{U}^{(n)}\|_F \leq \epsilon_\perp$. Given approximations $\tilde{\mathbf{A}}^{(n)} = \mathbf{A}^{(n)} + \Delta_A^{(n)}$ and $\tilde{\mathbf{U}}^{(n)} = \tilde{\mathbf{A}}^{(n)\dagger T} = \mathbf{U}^{(n)} + \Delta_U^{(n)}$ with normalized columns, then $\exists \epsilon > 0$ such that if $\|\Delta_A^{(n)}\|_F, \|\Delta_U^{(n)}\|_F \leq \epsilon_n \leq \epsilon$ for $\forall n \in \{1, \dots, N-1\}$, then

$$\tilde{\mathbf{A}}^{(N)} = \mathbf{X}_{(N)}(\tilde{\mathbf{U}}^{(1)} \odot \dots \odot \tilde{\mathbf{U}}^{(N-1)})$$

satisfies $\|\tilde{\mathbf{A}}^{(N)}\bar{\mathbf{D}}^{-1} - \mathbf{A}^{(N)}\| = O(\epsilon\epsilon_\perp + \epsilon_1 \dots \epsilon_{N-1})$, where $\bar{\mathbf{D}}$ normalizes $\tilde{\mathbf{A}}^{(N)}$, i.e., $\bar{d}_{ii} = \|\tilde{\mathbf{a}}_i^{(N)}\|_2$. Further, $\|\bar{\mathbf{D}} - \mathbf{D}\|_F = O(\epsilon)$.

Proof. We expand the update as follows,

$$\begin{aligned} \tilde{\mathbf{A}}^{(N)} &= ([\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}\mathbf{D}]\!] + [\![\hat{\mathbf{A}}^{(1)}, \dots, \hat{\mathbf{A}}^{(N)}]\!])_{(N)}(\tilde{\mathbf{U}}^{(1)} \odot \dots \odot \tilde{\mathbf{U}}^{(N-1)}) \\ &= \mathbf{A}^{(N)}\mathbf{D}(\mathbf{A}^{(1)T}\tilde{\mathbf{U}}^{(1)} * \dots * \mathbf{A}^{(N-1)T}\tilde{\mathbf{U}}^{(N-1)}) + \hat{\mathbf{A}}^{(N)}(\hat{\mathbf{A}}^{(1)T}\tilde{\mathbf{U}}^{(1)} * \dots * \hat{\mathbf{A}}^{(N-1)T}\tilde{\mathbf{U}}^{(N-1)}) \\ &= \mathbf{A}^{(N)}\mathbf{D}((\mathbf{I} + \mathbf{A}^{(1)T}\Delta_U^{(1)}) * \dots * (\mathbf{I} + \mathbf{A}^{(N-1)T}\Delta_U^{(N-1)})) \\ &\quad + \hat{\mathbf{A}}^{(N)}((\hat{\mathbf{A}}^{(1)T}\mathbf{U}^{(1)} + \hat{\mathbf{A}}^{(1)T}\Delta_U^{(1)}) * \dots * (\hat{\mathbf{A}}^{(N-1)T}\mathbf{U}^{(N-1)} + \hat{\mathbf{A}}^{(N-1)T}\Delta_U^{(N-1)})). \end{aligned}$$

By the stationary point condition, we have that

$$\hat{\mathbf{A}}^{(N)}(\hat{\mathbf{A}}^{(1)T}\mathbf{U}^{(1)} * \dots * \hat{\mathbf{A}}^{(N-1)T}\mathbf{U}^{(N-1)}) = \mathbf{0}.$$

Consequently, the error reduces to the cross terms of the summations, i.e.,¹

$$\begin{aligned} \tilde{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\mathbf{D} &= \mathbf{A}^{(N)}\mathbf{D} \left[\underset{n=1}{\overset{N}{\ast}} \mathbf{A}^{(n)T}\Delta_U^{(n)} + \mathbf{I} * \sum_{k=1}^{N-1} \left(\underset{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}}{\underset{k}{\ast}} \underset{l=1}{\overset{k}{\ast}} \mathbf{A}^{(m_l)T}\Delta_U^{(m_l)} \right) \right] \\ &\quad + \hat{\mathbf{A}}^{(N)} \left[\sum_{k=1}^{N-1} \left(\underset{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}, \{w_1, \dots, w_{N-k}\} = \{1, \dots, N\} \setminus \{m_1, \dots, m_k\}}{\underset{k}{\ast}} \underset{l=1}{\overset{k}{\ast}} \hat{\mathbf{A}}^{(m_l)T}\mathbf{U}^{(m_l)} \underset{l=1}{\overset{N-k}{\ast}} \hat{\mathbf{A}}^{(w_l)T}\Delta_U^{(w_l)} \right) \right] \end{aligned}$$

First, since each error term has Frobenius norm $O(\|\Delta^{(n)}\|) = O(\epsilon)$, the column norms of $\tilde{\mathbf{A}}^{(N)}$ will yield $\bar{\mathbf{D}}$ with $\|\bar{\mathbf{D}} - \mathbf{D}\|_F = O(\epsilon)$. Further, in order to get the error bound on $\tilde{\mathbf{A}}^{(N)}$, we consider the diagonal matrix,

$$\mathbf{S} = \mathbf{I} + \mathbf{I} * \sum_{k=1}^{N-1} \left(\underset{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}}{\underset{k}{\ast}} \underset{l=1}{\overset{k}{\ast}} \mathbf{A}^{(m_l)T}\Delta_U^{(m_l)} \right),$$

¹For $N = 3$, the right-hand side of this formula is

$$\begin{aligned} &\mathbf{A}^{(3)}\mathbf{D}[\mathbf{A}^{(1)T}\Delta_U^{(1)} * \mathbf{A}^{(2)T}\Delta_U^{(2)} + \mathbf{I} * (\mathbf{A}^{(1)T}\Delta_U^{(1)} + \mathbf{A}^{(2)T}\Delta_U^{(2)})] \\ &\quad + \hat{\mathbf{A}}^{(3)}[\hat{\mathbf{A}}^{(1)T}\mathbf{U}^{(1)} * \hat{\mathbf{A}}^{(2)T}\Delta_U^{(2)} + \hat{\mathbf{A}}^{(2)T}\mathbf{U}^{(2)} * \hat{\mathbf{A}}^{(1)T}\Delta_U^{(1)}]. \end{aligned}$$

and show that $\|\bar{\mathbf{D}} - \mathbf{DS}\|_F = O(\epsilon_1 \dots \epsilon_{N-1} + \epsilon\epsilon_\perp)$. Since,

$$\begin{aligned}\tilde{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\mathbf{DS} &= \mathbf{A}^{(N)}\mathbf{DS} \left[S^{-1} \underset{n=1}{\ast} \mathbf{A}^{(n)T} \Delta_U^{(n)} \right] \\ &\quad + \hat{\mathbf{A}}^{(N)} \left[\sum_{k=1}^{N-1} \left(\sum_{\substack{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}, \\ \{w_1, \dots, w_{N-k}\} = \{1, \dots, N\} \setminus \{m_1, \dots, m_k\}}} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(m_l)T} \mathbf{U}^{(m_l)} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(w_l)T} \Delta_U^{(w_l)} \right) \right],\end{aligned}$$

and $\|\mathbf{S} - \mathbf{I}\|_F = O(\epsilon)$, we have

$$\|\tilde{\mathbf{A}}^{(N)} - \mathbf{A}^{(N)}\mathbf{DS}\|_F = O(\epsilon_1 \dots \epsilon_{N-1} + \epsilon\epsilon_\perp).$$

The same bound follows for each column, so $\|\bar{\mathbf{D}} - \mathbf{DS}\|_F = O(\epsilon_1 \dots \epsilon_{N-1} + \epsilon\epsilon_\perp)$. Now, using this bound on

$$\begin{aligned}\tilde{\mathbf{A}}^{(N)}\bar{\mathbf{D}}^{-1} &= \mathbf{A}^{(N)}\mathbf{DS} \left[S^{-1} \underset{n=1}{\ast} \mathbf{A}^{(n)T} \Delta_U^{(n)} + \mathbf{I} \right] \bar{\mathbf{D}}^{-1} \\ &\quad + \hat{\mathbf{A}}^{(N)} \left[\sum_{k=1}^{N-1} \left(\sum_{\substack{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}, \\ \{w_1, \dots, w_{N-k}\} = \{1, \dots, N\} \setminus \{m_1, \dots, m_k\}}} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(m_l)T} \mathbf{U}^{(m_l)} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(w_l)T} \Delta_U^{(w_l)} \bar{\mathbf{D}}^{-1} \right) \right],\end{aligned}$$

we obtain

$$\begin{aligned}\|\tilde{\mathbf{A}}^{(N)}\bar{\mathbf{D}}^{-1} - \mathbf{A}^{(N)}\|_F &= \left\| \mathbf{A}^{(N)} \mathbf{D} \left[\underset{n=1}{\ast} \mathbf{A}^{(n)T} \Delta_U^{(n)} \right] \bar{\mathbf{D}}^{-1} \right. \\ &\quad \left. + \hat{\mathbf{A}}^{(N)} \left[\sum_{k=1}^{N-1} \left(\sum_{\substack{\{m_1, \dots, m_k\} \subset \{1, \dots, N\}, \\ \{w_1, \dots, w_{N-k}\} = \{1, \dots, N\} \setminus \{m_1, \dots, m_k\}}} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(m_l)T} \mathbf{U}^{(m_l)} \underset{l=1}{\ast} \hat{\mathbf{A}}^{(w_l)T} \Delta_U^{(w_l)} \bar{\mathbf{D}}^{-1} \right) \right] \right\|_F \\ &\quad + O(\epsilon_1 \dots \epsilon_{N-1} + \epsilon\epsilon_\perp) \\ &= O(\epsilon_1 \dots \epsilon_{N-1} + \epsilon\epsilon_\perp).\end{aligned}\quad \square$$

5. Approximate Decomposition. In the above sections, we have provided a motivation for Algorithm 3.1 to compute a CP decomposition of rank R with R being less than or equal to the smallest mode length of the input tensor. We have shown in Theorem 4.3 that this algorithm exhibits a super linear local convergence rate for exact CP decomposition problems and achieves a desirable approximation for special input tensors as described in Lemma 4.4. We now focus on the case of finding a good CP approximation for an arbitrary input tensor. We show that Algorithm 3.1 can be viewed as performing coupled minimization of the residual error of the decomposition in terms of a Mahalanobis distance metric [12]. Note that this perspective of the algorithm is different from the one introduced in Section 1.2, however it allows us to formulate an alternating minimization algorithm which generalizes the Algorithm 3.1 to any CP rank and to interpolate between the updates of ALS and Algorithm 3.1.

5.1. Mahalanobis Distance Minimization. Each update of Algorithm 3.1 may be viewed as minimizing a residual error with rescaled components. For an order 3 tensor \mathbf{X} in updating the first factor, it minimizes

$$(5.1) \quad \|(\mathbf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!])(\mathbf{C}^{\dagger T} \otimes \mathbf{B}^{\dagger T})\|_F.$$

Since the column span of $\mathbf{C} \odot \mathbf{B}$ is the same as that of $\mathbf{C}^{\dagger T} \otimes \mathbf{B}^{\dagger T}$, the transformed residual preserves all components of the residual error that may be reduced in choosing \mathbf{A} with \mathbf{B} and \mathbf{C} fixed, since the residual may be written as

$$\mathbf{A}(\mathbf{I}_R \odot \mathbf{I}_R)^T - \mathbf{X}_{(1)}(\mathbf{C}^{\dagger T} \otimes \mathbf{B}^{\dagger T}).$$

We show that this may be viewed as optimizing a single overall objective function relative to each factor, while keeping the distance metric associated with that factor independent (and then updating it thereafter). This interpretation then enables us to extend Algorithm 3.1 for any CP rank and introduce methods that are a hybrid of Algorithm 3.1 and standard ALS.

5.1.1. Alternating Mahalanobis Distance Minimization. We consider a variant of Mahalanobis distance [18], which computes the distance between vectors \mathbf{x} and \mathbf{y} as $d(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{M}(\mathbf{x} - \mathbf{y})$ for a given symmetric positive definite matrix \mathbf{M} . The matrix \mathbf{M} is called the ground metric matrix. Ground metric generalizes the Euclidean distance to Mahalanobis distance by rotation and scaling of the axes along which the distance is computed. While, the underlying ground metric may already be known, it may also be learned via various metric learning techniques [6, 31]. In optimal transport applications and other applications which require computation of distance between probability distributions, a Wasserstein distance is considered instead. Wasserstein distance between tensors with a given ground metric has been considered for nonnegative CP decomposition [2]. The ground metric in Wasserstein distance may be learned via similar metric learning techniques as in Mahalanobis distance [17]. Simultaneous optimization for a ground metric and Wasserstein distance between matrices has been used for nonnegative matrix factorization [62].

We consider minimization of the Mahalanobis distance between tensors with a fixed ground metric which maybe updated later. In particular, the objective function minimized for an input tensor $\mathbf{X} \in \mathbb{R}^{I_1 \cdots I_N}$, and factors $\mathbf{A}^{(i)} \in \mathbb{R}^{I_i \times R}$ is

$$(5.2) \quad f(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}) = \frac{1}{2} \text{vec}(\mathbf{X} - \mathbf{Y})^T \mathbf{M} \text{vec}(\mathbf{X} - \mathbf{Y}),$$

where $\mathbf{Y} = [\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!]$.

We restrict the ground metric matrix \mathbf{M} to be Kronecker structured defined as

$$\mathbf{M} = \bigotimes_{k=1}^N \mathbf{M}^{(k)-1}.$$

Each $\mathbf{M}^{(k)-1}$ maybe viewed as a ground metric for each mode of the tensor. This restriction allows us to exploit the computational benefits of the structure and enables us to formulate an efficient alternating minimization algorithm. We consider the objective in (5.2) for a general (fixed) ground metric for alternating optimization, which also allows us to formulate different algorithms for CP decomposition by changing the ground metric. We derive an update for the alternating minimization with respect to n th factor matrix, given by

$$(5.3) \quad \mathbf{A}^{(n)} = \min_{\mathbf{A}^{(n)}} \frac{1}{2} \|\mathbf{X}_{(n)} - \mathbf{A}^{(n)} \mathbf{P}^{(n)T}\|_M^2,$$

where $\mathbf{P}^{(n)} = \bigodot_{m=1, m \neq n}^N \mathbf{A}^{(m)}$.

For succinct writing, let $\mathbf{M}_{(n)} = \bigotimes_{k=1, k \neq n}^N \mathbf{M}^{(k)-1}$. Since the objective function is quadratic in $\mathbf{A}^{(n)}$, a minimizer of (5.3) can be found by obtaining a gradient $\mathbf{G}^{(n)}$ with respect to the n th factor matrix and setting it to $\mathbf{0}$. The gradient is

$$\mathbf{G}^{(n)} = \mathbf{M}^{(n)-1} \mathbf{A}^{(n)} \mathbf{P}^{(n)T} \mathbf{M}_{(n)} \mathbf{P}^{(n)} - \mathbf{M}^{(n)-1} \mathbf{X}_{(n)} \mathbf{M}_{(n)} \mathbf{P}^{(n)}.$$

Setting the gradient above to be $\mathbf{0}$ and equating $\mathbf{M}^{(n)} \mathbf{M}^{(n)-1} = \mathbf{I}$, we get an update for the n th factor given as the solution of the following system,

$$(5.4) \quad \mathbf{A}^{(n)} \left(\mathbf{P}^{(n)T} \mathbf{M}_{(n)} \mathbf{P}^{(n)} \right) = \mathbf{X}_{(n)} \mathbf{M}_{(n)} \mathbf{P}^{(n)}.$$

Using the properties of Khatri-Rao products and Kronecker products, the update for the n th factor matrix reduces to the system of equations

$$(5.5) \quad \mathbf{A}^{(n)} \mathbf{Z}^{(n)} = \mathbf{X}_{(n)} \mathbf{L}^{(n)},$$

where $\mathbf{L}^{(n)} = \bigodot_{k=1, k \neq n}^N \mathbf{M}^{(k)-1} \mathbf{A}^{(k)}$,

and $\mathbf{Z}^{(n)} = \bigast_{k=1, k \neq n}^N \mathbf{A}^{(k)T} \mathbf{M}^{(k)-1} \mathbf{A}^{(k)}$.

The above update leads to the ALS algorithm if $\mathbf{M}^{(k)} = \mathbf{I}$ for all k . We can retrieve Algorithm 3.1 by defining

$$(5.6) \quad \mathbf{M}^{(k)} = \mathbf{A}^{(k)} \mathbf{A}^{(k)T} + (\mathbf{I} - \mathbf{A}^{(k)} \mathbf{A}^{(k)\dagger}), \quad \forall k \in \{1, \dots, N\}.$$

The matrix $\mathbf{I} - \mathbf{A}^{(k)} \mathbf{A}^{(k)\dagger}$ is inconsequential when applied to the factor matrices. It is included to ensure that each $\mathbf{M}^{(k)}$ is SPD. Since Algorithm 3.1 can be retrieved from the above update, we refer to Algorithm 3.1 as AMDM (Alternating Mahalanobis Distance Minimization).

Let us assume that the iteration involving the above derived alternating updates to each factor as in (5.4), converges to a critical point. We can then bound the backward error in application of each matricization of the reconstructed tensor $\mathbf{Y} = [\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!]$, since from (5.4), for each n , we have

$$\begin{aligned} \mathbf{A}^{(n)} \left(\mathbf{P}^{(n)T} \mathbf{M}_{(n)} \mathbf{P}^{(n)} \right) - \mathbf{X}_{(n)} \mathbf{M}_{(n)} \mathbf{P}^{(n)} &= \mathbf{0}, \\ \left(\mathbf{Y}_{(n)} - \mathbf{X}_{(n)} \right) \mathbf{M}_{(n)} \mathbf{P}^{(n)} &= \mathbf{0}. \end{aligned}$$

Therefore, we have that

$$\|\mathbf{Y}_{(n)} \mathbf{z} - \mathbf{X}_{(n)} \mathbf{z}\| = \|\mathbf{X}_{(n)} \mathbf{z}^\perp\|,$$

where \mathbf{z}^\perp is the projection of $\mathbf{z} \in \mathbb{R}^{\prod_{j=1, j \neq n} I_j}$ onto the orthogonal complement of column span of $\mathbf{M}_{(n)} \mathbf{P}^{(n)}$ or $\bigodot_{j=1, j \neq n}^N \mathbf{M}^{(j)} \mathbf{A}^{(j)}$. For ALS, AMDM, and the hybrid methods (introduced in Section 5.3) that interpolate between the both, it is sufficient to consider the projection onto the orthogonal complement of column span of $\bigodot_{j=1, j \neq n}^N \mathbf{A}^{(j)}$. This is because for each j , the ground metric matrices are chosen such that the column span of $\mathbf{A}^{(j)}$ is an invariant subspace of $\mathbf{M}^{(j)}$.

5.1.2. Comparison of AMDM and ALS for approximate rank-2 CPD. We use the formulation introduced in the previous subsection to generalize AMDM to the case when CP rank R is greater than the mode lengths and to derive hybrid methods that interpolate between AMDM and ALS. The residual transformation tends to equalize the weight of contribution to the objective function attributed to components of the error associated with different rank-1 parts of the CP decomposition, $[\![\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i]\!]$ without increasing the collinearity of columns of the factors. We provide an example as an intuition for this assertion.

Consider a tensor $\mathbf{X} = \lambda_1 \mathbf{X}_1 + \lambda_2 \mathbf{X}_2 + \mathbf{N}$ where \mathbf{X}_1 and \mathbf{X}_2 are normalized rank-1 tensors and \mathbf{N} is noise of small magnitude. Assume that $\lambda_1 \gg \lambda_2$ and the rank-1 tensors are highly correlated, i.e., the factors have collinear columns. Let the current CPD approximation be $\mathbf{Y} = [\![\bar{\boldsymbol{\lambda}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \bar{\lambda}_1 \mathbf{Y}_1 + \bar{\lambda}_2 \mathbf{Y}_2$. The least squares objective minimizes

$$\begin{aligned} \|\mathbf{X} - \mathbf{Y}\|_F^2 &= \text{vec}(\mathbf{X} - \mathbf{Y})^T \text{vec}(\mathbf{X} - \mathbf{Y}) \\ &= \text{vec}(\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{N})^T \text{vec}(\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{N}) \\ &= \|\mathbf{E}_1\|_F^2 + \|\mathbf{E}_2\|_F^2 + 2\text{vec}(\mathbf{E}_1)^T \text{vec}(\mathbf{E}_2) + 2\text{vec}(\mathbf{N})^T (\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{N}), \end{aligned}$$

where $\mathbf{E}_1 = \lambda_1 \bar{\mathbf{X}}_1 - \bar{\lambda}_1 \mathbf{Y}_1$ and $\mathbf{E}_2 = \lambda_2 \bar{\mathbf{X}}_2 - \bar{\lambda}_2 \mathbf{Y}_2$. Alternating least squares algorithm may reduce $\|\mathbf{E}_1\|_F$ and the component of \mathbf{E}_2 in the direction of \mathbf{E}_1 , since it leads to reduction of the terms with larger contribution in the error. This causes an increase in collinearity of the approximated factors and a more ill-conditioned decomposition. On the other hand, the objective in (5.1) can be expressed as

$$\begin{aligned} \|(\mathbf{X} - \mathbf{Y}) \times_1 \mathbf{I} \times_2 \mathbf{B}^{\dagger T} \times_3 \mathbf{C}^{\dagger T}\|_F^2 &= \text{vec}(\mathbf{X} - \mathbf{Y})^T \mathbf{M} \text{vec}(\mathbf{X} - \mathbf{Y}) \\ &= \|\mathbf{E}_1\|_{\mathbf{M}}^2 + \|\mathbf{E}_2\|_{\mathbf{M}}^2 + 2\text{vec}(\mathbf{E}_1)^T \mathbf{M} \text{vec}(\mathbf{E}_2) + 2\text{vec}(\mathbf{N})^T \mathbf{M} (\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{N}), \end{aligned}$$

where $\mathbf{M} = \mathbf{C}^{\dagger T} \mathbf{C}^{\dagger} \otimes \mathbf{B}^{\dagger T} \mathbf{B}^{\dagger} \otimes \mathbf{I}$. The matrix \mathbf{M} rescales the components of the error according to the inverse of square of singular values of the factors, since

$$\begin{aligned} \|\mathbf{E}_1\|_{\mathbf{M}}^2 &= \text{vec}(\mathbf{E}_1)^T \mathbf{M} \text{vec}(\mathbf{E}_1) \\ &= \text{vec}(\bar{\mathbf{E}}_1)^T (\Sigma_C^{-2} \otimes \Sigma_B^{-2} \otimes \mathbf{I}) \text{vec}(\bar{\mathbf{E}}_1), \end{aligned}$$

where $\bar{\mathcal{E}}_1 = \mathcal{E}_1 \times_1 \mathbf{I} \times_2 \mathbf{U}_B \times_3 \mathbf{U}_C$, with \mathbf{U}_B and \mathbf{U}_C being the left singular vectors of \mathbf{B} and \mathbf{C} respectively. Thus, the error is rotated by the left singular vectors of \mathbf{B} and \mathbf{C} , and then rescaled by square of inverse of singular values of \mathbf{B} and \mathbf{C} , i.e., Σ_B^{-2} , Σ_C^{-2} . Therefore, if the approximated factors are collinear, the contribution in the direction of the singular vectors of \mathbf{B} and \mathbf{C} with largest singular value is weighed proportionally less and similarly the contribution of error in the direction of those with smaller singular value is weighed more. This reduces the imbalance in error and leads to a better conditioned decomposition.

5.2. Generalizing AMDM to Any CP Rank. The AMDM algorithm as described in Algorithm 3.1 imposes a constraint that CP rank should be less than or equal to the smallest mode length of the tensor. We now describe how the update in (5.5) with the ground metric as defined in (5.6) leads to the definition of AMDM without conditions imposed on the rank. For each $\mathbf{M}^{(k)}$,

$$\mathbf{M}^{(k)-1} = \mathbf{A}^{(k)\dagger T} \mathbf{A}^{(k)\dagger} + (\mathbf{I} - \mathbf{A}^{(k)} \mathbf{A}^{(k)\dagger}).$$

The linear system for updating the n th factor matrix as in (5.5) can then be simplified to get

$$\begin{aligned} \mathbf{A}^{(n)} \mathbf{Z}^{(n)} &= \mathbf{X}_{(n)} \mathbf{L}^{(n)}, \\ \text{where } \mathbf{L}^{(n)} &= \bigodot_{k=1, k \neq n}^N \mathbf{M}^{(k)-1} \mathbf{A}^{(k)} = \bigodot_{k=1, k \neq n}^N \mathbf{A}^{(k)\dagger T}, \\ \text{and } \mathbf{Z}^{(n)} &= \bigast_{k=1, k \neq n}^N \mathbf{A}^{(k)T} \mathbf{M}^{(k)-1} \mathbf{A}^{(k)} = \bigast_{k=1, k \neq n}^N \mathbf{A}^{(k)\dagger} \mathbf{A}^{(k)}. \end{aligned}$$

The above update is equivalent to Algorithm 3.1 when CP rank $R \leq I_m$, $\forall m \in \{1, \dots, N\}$, since in that case $\mathbf{Z}^{(n)} = \mathbf{I}$ for all n . For the case when CP rank R is larger than the mode lengths, we get a symmetric semi-definite system of equations. Since the system is semi-definite, a pivoted Cholesky decomposition followed by a triangular solve be used for the solution to exist. Alternatively, as in ALS, a regularization term may be introduced to make the system positive definite. The cost of forming this system, $\mathbf{Z}^{(n)}$ is of the same leading order as ALS, i.e., $O(IR^2)$ per subsweep, since it requires to obtain a pseudo inverse of the factor in $(n-1)$ th iteration, matrix multiplication and Hadamard products of the factors and their previously obtained pseudoinverses. The system solve amounts to a computational cost of $O(R^3)$ to solve the system.

5.3. Interpolating Between AMDM and ALS. Performing alternating Mahalanobis distance minimization with an identity ground metric is equivalent to performing ALS. To explore methods that interpolate between AMDM and ALS, we can interpolate the ground metric between the identity matrix and the one associated with AMDM given in (5.6). We can find such ground metrics by decomposing each factor matrix into two low rank matrices, such that $\mathbf{A}^{(k)} = \mathbf{A}_1^{(k)} + \mathbf{A}_2^{(k)}$, where $\mathbf{A}_1^{(k)}$ is the best rank- t approximation of $\mathbf{A}^{(k)}$. In other words, $\mathbf{A}_1^{(k)}$ contains the largest t singular values and the corresponding singular vectors of $\mathbf{A}^{(k)}$ and $\mathbf{A}_2^{(k)}$ contains the rest. The ground metric for each mode can then be defined using only the first part as

$$(5.7) \quad \mathbf{M}^{(k)} = \mathbf{A}_1^{(k)} \mathbf{A}_1^{(k)T} + (\mathbf{I} - \mathbf{A}_1^{(k)} \mathbf{A}_1^{(k)\dagger}), \quad \forall k \in \{1, \dots, N\}.$$

By defining a ground metric based on only the first part of the singular value decomposition of the factors leads to hybrid methods, since if the first part is all of the singular value decomposition then we get back the ground metric in AMDM and if it is none of the same then we get back the identity matrix by convention as the orthogonal complement of none is everything.

Note that for each k ,

$$\mathbf{M}^{(k)-1} = \mathbf{A}_1^{(k)\dagger T} \mathbf{A}_1^{(k)\dagger} + (\mathbf{I} - \mathbf{A}_1^{(k)} \mathbf{A}_1^{(k)\dagger}).$$

The update for the n th factor matrix also becomes a combination of AMDM and ALS where the first part of the singular value decomposition of factors is treated as in AMDM and the second one as in ALS. More

precisely, the system of equations for updating the n th factor matrix is

$$(5.8) \quad \begin{aligned} \mathbf{A}^{(n)} \mathbf{Z}^{(n)} &= \mathbf{X}_{(n)} \mathbf{L}^{(n)}, \\ \text{where } \mathbf{L}^{(n)} &= \bigodot_{k=1, k \neq n}^N \mathbf{M}^{(k)-1} \mathbf{A}^{(k)} = \bigodot_{k=1, k \neq n}^N (\mathbf{A}_1^{(k)\dagger} + \mathbf{A}_2^{(k)}), \\ \text{and } \mathbf{Z}^{(n)} &= \bigast_{k=1, k \neq n}^N \mathbf{A}^{(k)T} \mathbf{M}^{(k)-1} \mathbf{A}^{(k)} = \bigast_{k=1, k \neq n}^N (\mathbf{A}_1^{(k)\dagger} \mathbf{A}_1^{(k)} + \mathbf{A}_2^{(k)T} \mathbf{A}_2^{(k)}). \end{aligned}$$

We describe the above derived hybrid algorithm in Algorithm 5.1. The algorithm starts by normalizing columns of all the factors and absorbing norms in the first factor as described in Section 3 and then computing a reduced SVD of all the factors which costs $O(\sum_{n=1}^N I_n R \min(I_n, R))$. At the n th subsweep of the algorithm, operations performed are similar in computational cost as that of ALS. Right and left hand sides of the system, \mathbf{L}_n and \mathbf{Z}_n require $O(I_n R^2)$ and $O(R^2 \min(I_n, R))$ operations respectively. The symmetric semi definite system solve requires $O(R^3)$. Computing the right hand side, i.e., performing MTTKRP is the most computationally expensive operation with a cost of $O(\prod_{n=1}^N I_n R)$ for each subsweep. In addition to this, the factor matrix obtained after the solve is normalized and a reduced SVD is obtained to update the singular value decomposition which costs $O(I_n R \min(I_n, R))$. Therefore, the asymptotic computational cost of the algorithm is the same as ALS being $O(\prod_{n=1}^N I_n R)$.

Algorithm 5.1 General-AMDM: Alternating Mahalanobis Distance Minimization with singular value thresholding

```

1: Input: Tensor  $\mathbf{X} \in \mathbb{R}^{I_1 \times \dots \times I_N}$ , threshold  $t$ , rank  $R$ 
2: Initialize  $\{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}\}$  so each  $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R}$  is random
3: for  $n \in \{2, \dots, N\}$  do
4:    $\mathbf{A}^{(n)} = \text{normalize}(\mathbf{A}^{(n)})$ 
5:    $\mathbf{U}^{(n)} = \min(I_n, R)$  left singular vectors of  $\mathbf{A}^{(n)}$ 
6:    $\mathbf{V}^{(n)} = \min(I_n, R)$  right singular vectors of  $\mathbf{A}^{(n)}$ 
7:    $\mathbf{s}^{(n)} = \min(I_n, R)$  singular values of  $\mathbf{A}^{(n)}$ 
8: end for
9: while Convergence do
10:   for  $n \in \{1, \dots, N\}$  do
11:     for  $m \in \{1, \dots, N\}, m \neq n$  do
12:        $\mathbf{s}_{ps}^{(m)} = \text{first } t \text{ values of } \mathbf{s}^{(m)}$  inverted and others as it is
13:        $\mathbf{L}_m = \mathbf{U}^{(m)} \text{diag}(\mathbf{s}_{ps}^{(m)}) \mathbf{V}^{(m)T}$ 
14:        $\mathbf{Z}_m = \mathbf{V}^{(m)} \text{diag}(\mathbf{s}_{ps}^{(m)} * \mathbf{s}^{(m)}) \mathbf{V}^{(m)T}$ 
15:     end for
16:     Solve for  $\mathbf{A}^{(n)}$  in  $\mathbf{A}^{(n)} \mathbf{Z}^{(n)} = \mathbf{X}_{(n)} \mathbf{L}^{(n)}$  as in (5.5)
17:     Check Convergence, if converged: Break
18:      $\mathbf{A}^{(n)} = \text{normalize}(\mathbf{A}^{(n)})$ 
19:     Update  $\mathbf{U}^{(n)} = \min(I_n, R)$  left singular vectors of  $\mathbf{A}^{(n)}$ 
20:     Update  $\mathbf{V}^{(n)} = \min(I_n, R)$  right singular vectors of  $\mathbf{A}^{(n)}$ 
21:     Update  $\mathbf{s}^{(n)} = \min(I_n, R)$  singular values of  $\mathbf{A}^{(n)}$ 
22:   end for
23: end while
24: return factor matrices  $\{\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}\}$ 

```

6. Numerical Experiments. We perform numerical experiments to demonstrate the convergence behaviour of the AMDM algorithm for various tensors which include synthetic examples and tensors arising in different applications. We use absolute residual and fitness of the decomposition in Frobenius norm as metrics to measure the closeness of the decomposition to the input tensor. For an input tensor \mathbf{X} , these are given as

$$r = \|\mathbf{X} - \mathbf{Y}\|_F \text{ and } f = 1 - \frac{\|\mathbf{X} - \mathbf{Y}\|_F}{\|\mathbf{X}\|_F},$$

respectively, where $\mathcal{Y} = [\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!]$ is the approximated tensor. For measuring the stability of the decomposition or the degree of overlap of CP components, we use the normalized CPD condition number [9] to measure the sensitivity or degree of overlap of rank 1 components of the decomposition.

The normalized CP condition number is given by the reciprocal of the smallest singular value of the Terracini's matrix associated with the CP decomposition. For an equidimensional tensor of order N with mode length s and CP rank R , the size of Terracini's matrix is $s^N \times (N(s-1) + 1)R$. For CP rank lower than mode lengths of the tensor, this matrix can be compressed to $R^N \times (N(R-1) + 1)R$ and the CPD condition number can be efficiently computed with a cost of $O(R^{N+4})$. The details of computation of the condition number are in Appendix A. Our experiments consider two types of synthetic tensors.

Tensor made by random matrices (*Random* tensor). We create these tensors based on known uniformly distributed randomly-generated factor matrices $\mathbf{A}^{(n)} \in (0, 1)^{s \times R}$, $\mathcal{X} = [\![\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!]$.

Tensor made by collinear random matrices (*Collinearity* tensor). We use a similar approach as used in [1] to generate factor matrices with a fixed value of collinearity, say C . That means that these tensors are created with randomly-generated factors $\mathbf{A}^{(n)} \in \mathbb{R}^{s \times R}$ with the following property,

$$\begin{aligned} \mathbf{a}_r^{(n)T} \mathbf{a}_z^{(n)} &= C, \\ \text{s.t. } \|\mathbf{a}_r^{(n)}\| &= 1, \forall r \neq z \in \{1, \dots, R\}. \end{aligned}$$

We then set $\lambda_i = i$, $\forall i \in \{1, \dots, R\}$ to create a tensor, $\mathcal{X} = [\![\lambda; \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(N)}]\!]$.

We consider four tensors from various real world applications.

Sleep-EDF tensor: This dataset has been used to identify sleeping patterns. It comprises of electroencephalogram (EEG), electromyography (EMG) data in the non-rapid eye movement (NREM) stage of sleep [28].

MGH tensor: This dataset consists of data from Massachusetts General Hospital. It includes combinations of electroencephalogram (EEG), respiratory signals, and electromyogram signals (EMG). This dataset was used to analyze sleep using deep neural networks [8].

SCF tensor. We consider the density fitting tensor (Cholesky factor of the two-electron integral tensor) arising in quantum chemistry. This tensor has been used previously in [50] to compare the efficacy of the Gauss-Newton and alternating least squares algorithm. We leverage the PySCF library [53] to generate the three dimensional compressed density fitting tensor, representing the compressed restricted Hartree-Fock wave function of water molecule chain systems with STO-3G basis set. The number of molecules in the system is set to three for this experiment.

Amino acid tensor. This data set consists of five simple laboratory-made samples. Each sample contains different amounts of tyrosine, tryptophan and phenylalanine dissolved in phosphate buffered water. The samples were measured by fluorescence [10].

The experiments are divided broadly into two categories,

Exact decomposition. We create synthetic tensors with known CP rank R and compare the convergence behaviour of the AMDM algorithm with the alternating least squares algorithm for exact CP decomposition.

Approximate decomposition. We create synthetic tensors with known CP rank and special structure such as with added noise or as described in Lemma 4.4. We then approximate these tensors with CP rank R which is lower than the underlying decomposition rank. We also consider real world tensors from different applications with unknown CP rank.

6.1. Exact CP decomposition. We compare alternating least squares and Algorithm 3.1 for computing exact CP decomposition of synthetic tensors in Figure 1 and 2 and verify our theoretical results. We create *Collinearity* and *Random* tensors of specified CP rank to analyze the convergence of these algorithms.

In Figure 1, we create equidimensional synthetic tensors of order N with each mode length s to follow $s^N = 100^3$ with fixed CP rank $R = 20$. We initialize the factors with uniformly distributed random matrices for both the algorithms and plot the absolute residual for each iteration. For the *Collinearity* tensors, collinearity value, i.e., C is set to be 0.9. We can observe superlinear convergence of Algorithm 3.1 for both the cases, while ALS appears to be slowed down by the ‘swamp’ phenomenon for the *Collinearity* tensors. In Figure 2(a), we plot the empirical rate of convergence of a subiteration by using the relative residual after subiterations. Since order 2 is just matrix decomposition, AMDM and ALS algorithm become equivalent with a linear convergence rate. We can observe that the rate of local convergence for order 3 and 4 is consistent with what we observed in Section 4 with an error of 0.2% and 1% for order 3 and 4 respectively.

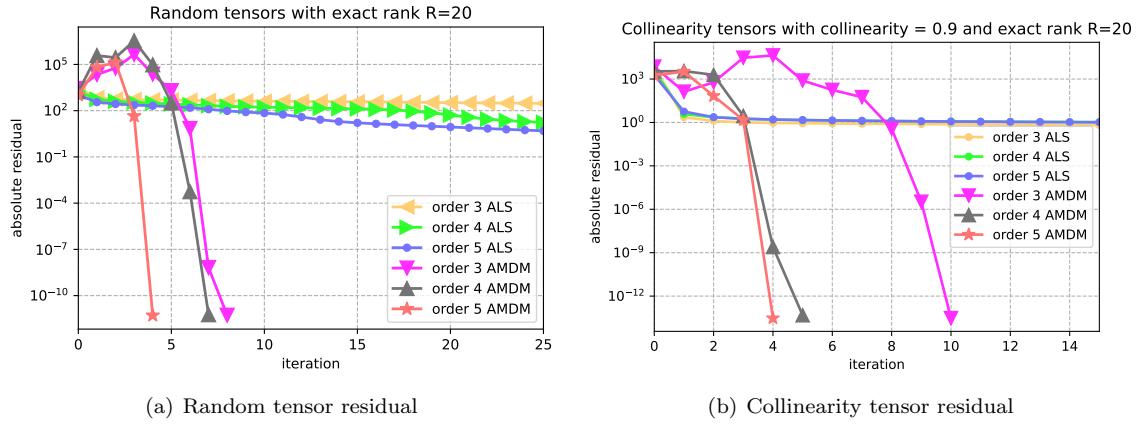


FIGURE 1. Superlinear convergence of AMDM algorithm for exact CPD

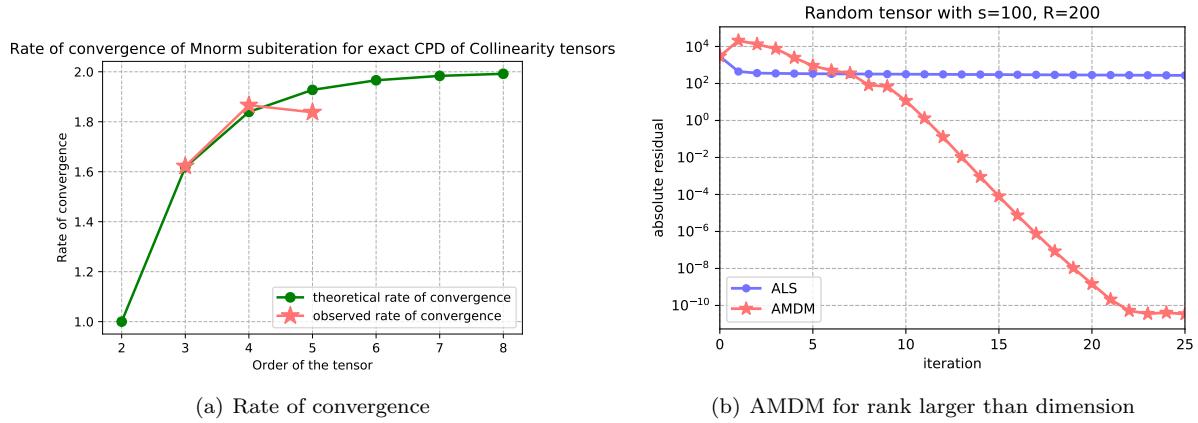
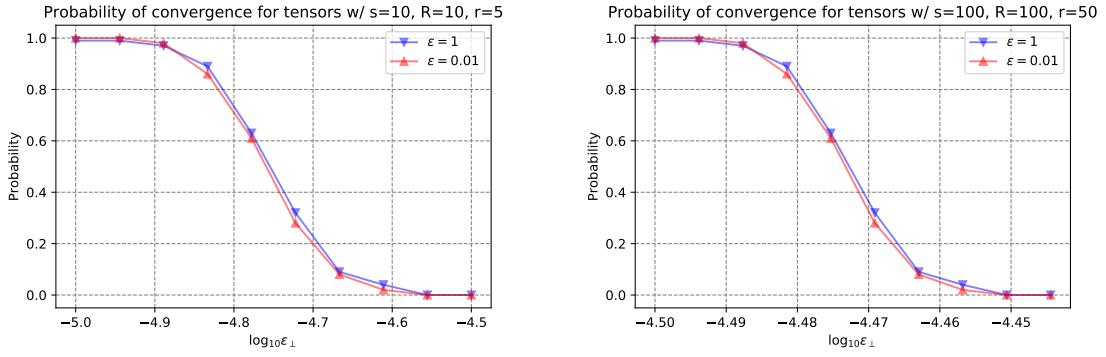


FIGURE 2. Rate of convergence of AMDM subiteration and linear convergence for large rank for exact CPD

For order 5 and above, it is difficult to verify the rate of convergence as there does not exist two data points of subiterations where we do not converge to machine precision and at the same time the assumptions for error to be small are satisfied.

In Figure 2(b), we create a *Random* equidimensional tensor of with mode length $s = 100$ and CP rank $R = 200$. We can observe that the Algorithm 5.1 converges linearly to the exact solution while taking only a few iterations to do so. Since $R > s$, we observe a linear convergence rate which is consistent with our theoretical results. ALS makes slow progress for this case. A reason for that might again be related to the collinearity of the factors, since when $R > s$, collinearity is high and ALS is more likely to experience the ‘swamp’ phenomenon [39].

6.2. Approximate CP decomposition. We plot the probability of convergence to the desired decomposition for synthetic tensors as described in 4.4 with respect to the ϵ_{\perp} to verify our theoretical results in Figure 3. We construct these tensors by constructing first $R/2$ columns of the factors with random matrices and then constructing the other half by projecting it onto the orthogonal complement of the column space of the first half and adding Gaussian noise of amplitude ϵ_{\perp} to the same. We construct 100 such tensors and for each tensor we consider 5 initial guesses which are ϵ away from the desired decomposition as described in 4.4. We plot the probability of convergence over these 100 tensors by considering if atleast 1 initial guess is within 10^{-9} of the desired factors. We observe that the probability of convergence to the desired decomposition is 1 when the ϵ_{\perp} is small, irrespective of the size and ϵ , thereby verifying that the first half of CP decomposition is a stationary point. The probability decreases as we increase ϵ_{\perp} , i.e., the decomposition converges to different stationary points for these tensors.



(a) Probability of convergence for equidimensional tensors with mode length = 10, exact CP rank = 10 and approximate rank = 5
(b) Probability of convergence for equidimensional tensors with mode length = 100, exact CP rank = 100 and approximate rank = 50

FIGURE 3. Probability of convergence for 100 tensors with 5 initial guesses in the setting as described in Lemma 4.4

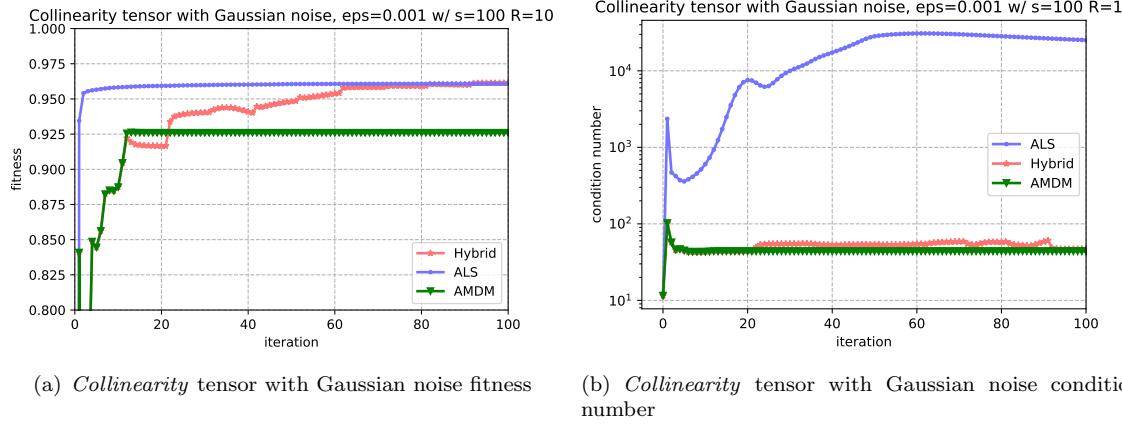


FIGURE 4. Collinearity tensor of size $100 \times 100 \times 100$ with exact CP rank $R = 10$ with added Gaussian noise with each entry distributed with mean $\mu = 0$ and standard deviation $\sigma = 0.001$

We compare ALS and different variants of AMDM for computing approximate CP decomposition of synthetic tensors in Figure 4 and application tensors that admit approximations with a low CP rank in Figure 5, Figure 6, Figure 7, and Figure 8. We plot the fitness and the condition number of the CP decomposition to compare ALS and variants of AMDM. The integer associated with AMDM $t = \#$ corresponds to the number of singular values inverted for each factor or best rank- t approximation \mathbf{A}_1 in Equation (5.7). The hybrid algorithm starts by using threshold $t = R$ in Algorithm 5.1, i.e., starts with Algorithm 3.1 and gradually decreases the threshold to 0 to recover ALS algorithm.

In Figure 4, we compute a rank 10 CP decomposition of the *Collinearity* tensor with collinearity $C = 0.9$ and exact CP rank $R = 10$ with added Gaussian noise tensor. Each entry of the noise tensor is distributed normally with mean $\mu = 0$ and standard deviation $\sigma = 0.001$. We observe that the AMDM algorithm maintains a low condition number while reaching a high fitness for both the input tensor and the underlying tensor, while ALS algorithm reaches a higher fitness at the cost of highly ill conditioned decomposition. For the hybrid algorithm, one less singular value is inverted after every 10 iterations, leading to a decomposition with fitness as high as ALS and conditioning as good as the AMDM algorithm. We observe a similar behaviour for different dimensions and CP rank approximations of the tensor, suggesting that there maybe multiple optimal decompositions for such problems.

In Figure 5 and Figure 6, we compute the CP decomposition of the SLEEP-EDF tensor and MGH tensor with CP rank $R = 10$. We consider several variants of hybrid Algorithm 5.1. For the hybrid algorithm,

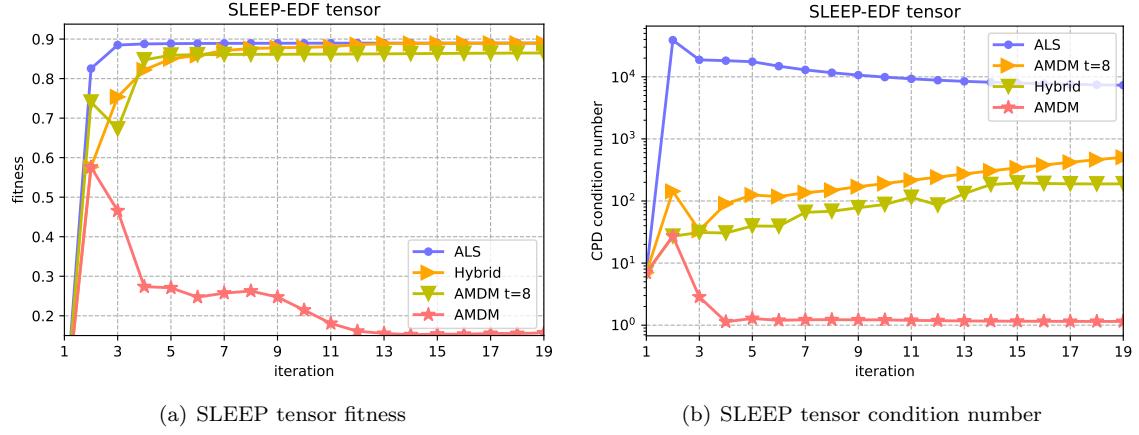


FIGURE 5. *SLEEP-EDF tensor of dimensions $2048 \times 14 \times 129 \times 86$ approximated with CP rank $R = 10$, where t is the singular value threshold in Algorithm 5.1.*

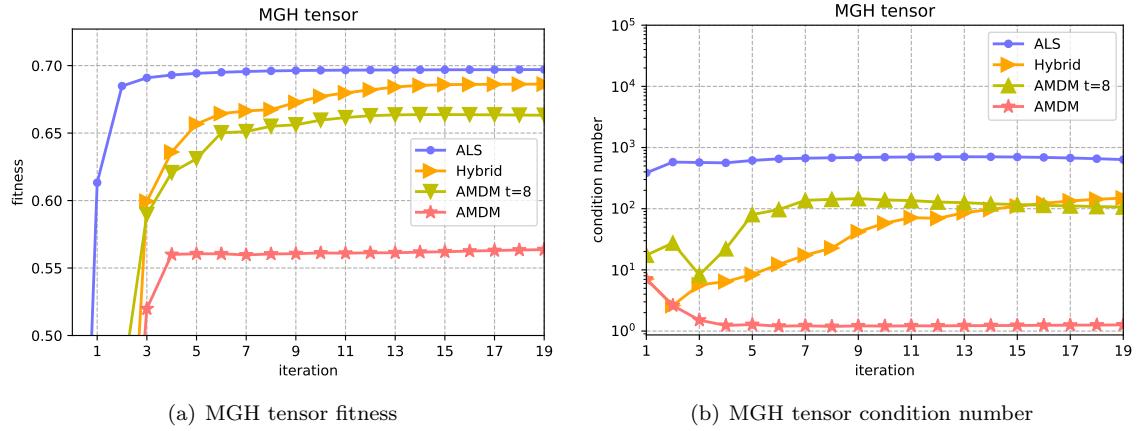


FIGURE 6. *MGH tensor of dimensions $2048 \times 12 \times 257 \times 43$ approximated with CP rank $R = 10$, where t is the singular value threshold in Algorithm 5.1*

one less singular value is inverted after every iteration. We clearly see a pattern in both the tensors that if lesser singular values are inverted then the fitness is higher and the CPD condition number is larger. We also see that the hybrid algorithm is able to achieve a fitness almost as high as ALS while maintaining a lower condition number. The condition number of decomposition with the hybrid algorithm is about 3.2x lower for the MGH tensor with an absolute difference in fitness being 0.01 or 0.015%, whereas the condition number is about 34.6x lower for the SLEEP tensor with an absolute difference fitness 0.0004 with hybrid algorithm being more accurate.

In Figure 7, we compute the CP decomposition of Amino acid tensor with rank $R = 5$. We have similar observations for the fitness and condition numbers of variants of the AMDM algorithm and ALS. In this case, the hybrid algorithm and AMDM with $t = 2$ achieve a better fitness than ALS. The maximum fitness for hybrid algorithm is 0.982 whereas the maximum fitness for ALS is 0.977. The condition number of ALS is about 69 times higher than that of the hybrid algorithm. Note that the fitness for AMDM (all singular values inverted) is 0.959 with a condition number equal to 5.56 indicating that the factor matrices have almost orthogonal columns.

In Figure 8, we compute CP decomposition of the SCF tensor with rank $R = 200$ (exceeding 2 of the 3 tensor dimensions). We use a relative tolerance criteria for computing $t = \text{argmin}_i (\frac{\sigma_{\max}}{\sigma_i} < 100)$ in the AMDM algorithm, i.e., singular values σ_i are inverted only if $\frac{\sigma_{\max}}{\sigma_i} < 100$, where σ_{\max} is the maximum singular value. The hybrid algorithm outperforms ALS in terms of fitness by reaching 0.993 fitness in 150

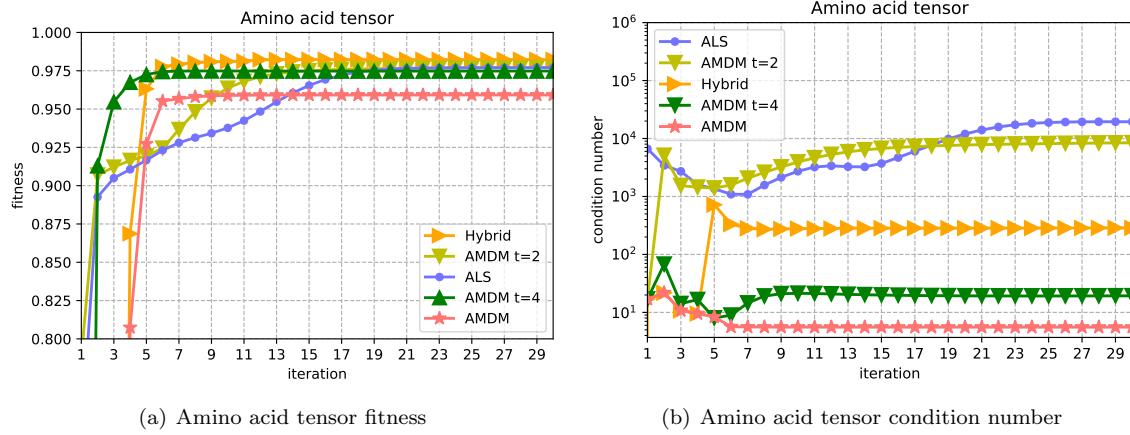


FIGURE 7. Amino acid tensor of dimensions $5 \times 61 \times 201$ approximated with CP rank $R = 5$, where t is the singular value threshold in Algorithm 5.1

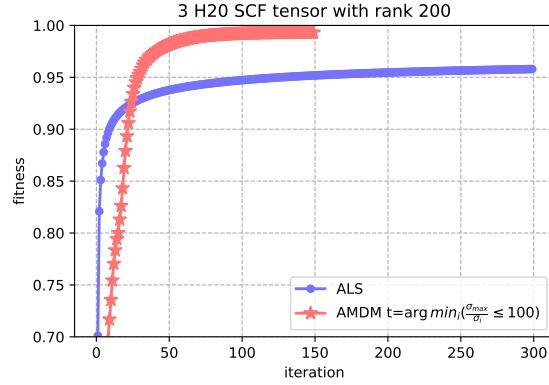


FIGURE 8. SCF tensor of size $339 \times 21 \times 21$ approximated with CP rank $R = 200$, where t is the singular value threshold in Algorithm 5.1.

iterations whereas ALS reaches 0.96 in 300 iterations.

7. Conclusion. In this work, we have proposed an alternative optimization algorithm, AMDM, to compute a CP decomposition of the tensor. This algorithm achieves superlinear local convergence for exact CP rank problems when CP rank is smaller than or equal to all the mode lengths of the tensor with the same asymptotic computational cost as that of ALS. For approximating a tensor via CP decomposition, we theoretically show that the algorithm locally converges to the stationary points of (1.4) for tensors with special CP structure. Although, the existence of these stationary points for any tensor is an open problem, we empirically confirm that the AMDM algorithm converges to these stationary points for various tensors. Viewing the algorithm as minimizing a Mahalanobis distance helps in generalization of the method for CP rank larger than the mode lengths and interpolate between AMDM and ALS algorithms. We also formulate an efficient way to compute the CPD condition number to track the condition of the decomposition throughout the algorithm. Our numerical experiments confirm that interpolation of algorithms between AMDM and ALS leads to a better conditioned decomposition without significant difference in fitness as compared to ALS for synthetic as well as most of the tested real world tensors. We provide an intuitive reasoning of this phenomenon and leave the detailed analysis as a future direction of research.

8. Acknowledgments. The authors would like to thank Ardavan (Ari) Afshar for detailed discussions about his work on minimizing Wasserstein distance between tensors from which this work is derived. The authors would also like to thank Jimeng Sun, Cheng Qian and Chaoqi Yang for having fruitful discussions and providing datasets which motivated this work. Navjot Singh and Edgar Solomonik were supported by

the US NSF OAC SSI program, award No. 1931258.

Appendix A. Computing the Condition Number of a CP Decomposition. It has been shown that the CPD condition number is the reciprocal of the smallest singular value of a matrix called Terracini's matrix. This matrix consists of the orthogonal basis for the tangent space of each of the rank-1 components of the reconstructed tensor. We will refer to the normalized condition number as the condition number of CPD and we refer the reader to [9, 60] for details about how a notion of condition number of a CP decomposition is defined and derived. Consider an equidimensional order 3 real tensor \mathcal{X} with mode length s . Let the CPD approximation of rank R be given by $\llbracket \lambda; \mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket$, then the Terracini's matrix \mathbf{U} is $\mathbf{U} = [\mathbf{U}_1 \dots \mathbf{U}_R]$, where $\forall i \in \{1, \dots, R\}$,

$$\mathbf{U}_i = [\mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i \quad \mathbf{Q}_{\mathbf{a}_i}^\perp \otimes \mathbf{b}_i \otimes \mathbf{c}_i \quad \mathbf{a}_i \otimes \mathbf{Q}_{\mathbf{b}_i}^\perp \otimes \mathbf{c}_i \quad \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{Q}_{\mathbf{c}_i}^\perp],$$

and $\mathbf{Q}_{\mathbf{a}_i}^\perp \in \mathbb{R}^{s \times (s-1)}$ is an orthogonal basis of the orthogonal complement of \mathbf{a}_i , and $\mathbf{Q}_{\mathbf{b}_i}^\perp$, and $\mathbf{Q}_{\mathbf{c}_i}^\perp$ are defined similarly. Consequently, the Terrracini's matrix is of size $s^3 \times R(3s - 2)$, and the computational cost of computing the smallest singular value via a Krylov subspace method is $O(s^5 R^2)$. For an order N tensor, this cost is $O(N^2 s^{N+2} R^2)$ and therefore expensive to compute for decompositions with moderately large mode lengths.

The cost of computing the condition number can be decreased significantly for when rank of the CP decomposition is lesser than all the mode lengths of the input tensor, i.e., if $R < s$. Assume that $R \leq s$, then since the condition number is invariant to orthogonal transformations [9], for a CPD of an order 3 tensor,

$$\kappa(\llbracket \lambda; \mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket) = \kappa(\llbracket \lambda; \mathbf{Q}_\mathbf{A}^T \mathbf{A}, \mathbf{Q}_\mathbf{B}^T \mathbf{B}, \mathbf{Q}_\mathbf{C}^T \mathbf{C} \rrbracket),$$

where $\mathbf{Q}_\mathbf{A} \in \mathbb{R}^{s \times s} = [\mathbf{Q}_\mathbf{A}^{(1)} \mathbf{Q}_\mathbf{A}^{(2)}]$, and the columns of $\mathbf{Q}_\mathbf{A}^{(1)} \in \mathbb{R}^{s \times R}$ are an orthogonal basis of the column space of \mathbf{A} , while the columns of $\mathbf{Q}_\mathbf{A}^{(2)} \in \mathbb{R}^{s \times (s-R)}$ are an orthogonal basis for the orthogonal complement of the column space of \mathbf{A} . We define $\mathbf{Q}_\mathbf{B} = [\mathbf{Q}_\mathbf{B}^{(1)} \mathbf{Q}_\mathbf{B}^{(2)}]$ and $\mathbf{Q}_\mathbf{C} = [\mathbf{Q}_\mathbf{C}^{(1)} \mathbf{Q}_\mathbf{C}^{(2)}]$ similarly. The transformed Terracini's matrix $\bar{\mathbf{U}} = [\bar{\mathbf{U}}_1 \dots \bar{\mathbf{U}}_R]$, where $\forall i \in \{1, \dots, R\}$,

$$\bar{\mathbf{U}}_i = [\underbrace{\mathbf{Q}_\mathbf{A}^T \mathbf{a}_i \otimes \mathbf{Q}_\mathbf{B}^T \mathbf{b}_i \otimes \mathbf{Q}_\mathbf{C}^T \mathbf{c}_i}_{\bar{\mathbf{U}}_i^{(1)}} \quad \underbrace{\bar{\mathbf{Q}}_{\mathbf{a}_i}^\perp \otimes \mathbf{Q}_\mathbf{B}^T \mathbf{b}_i \otimes \mathbf{Q}_\mathbf{C}^T \mathbf{c}_i}_{\bar{\mathbf{U}}_i^{(2)}} \quad \underbrace{\mathbf{Q}_\mathbf{A}^T \mathbf{a}_i \otimes \bar{\mathbf{Q}}_{\mathbf{b}_i}^\perp \otimes \mathbf{Q}_\mathbf{C}^T \mathbf{c}_i}_{\bar{\mathbf{U}}_i^{(3)}} \quad \underbrace{\mathbf{Q}_\mathbf{A}^T \mathbf{a}_i \otimes \mathbf{Q}_\mathbf{B}^T \mathbf{b}_i \otimes \bar{\mathbf{Q}}_{\mathbf{c}_i}^\perp}_{\bar{\mathbf{U}}_i^{(4)}}],$$

where $\bar{\mathbf{Q}}_{\mathbf{a}_i}^\perp = \mathbf{Q}_\mathbf{A}^T \mathbf{Q}_{\mathbf{a}_i}^\perp$ is an orthogonal basis of the orthogonal complement of $\mathbf{Q}_\mathbf{A}^T \mathbf{a}_i$, and $\bar{\mathbf{Q}}_{\mathbf{b}_i}^\perp$, and $\bar{\mathbf{Q}}_{\mathbf{c}_i}^\perp$ are defined similarly. Note that $\bar{\mathbf{U}}_i^{(j)} \bar{\mathbf{U}}_i^{(k)} = \mathbf{0}$ for $j \neq k$, since $\bar{\mathbf{Q}}^{\perp T} \mathbf{a}_i \mathbf{Q}_\mathbf{A}^T \mathbf{a}_i = 0$. Consequently,

$$\sigma_{\min}(\mathbf{U}) = \min_{j \in \{1, 2, 3, 4\}} \sigma_{\min}([\mathbf{U}_1^{(j)} \quad \dots \quad \mathbf{U}_R^{(j)}]).$$

After this transformation, we can obtain a reduced form of smaller dimensions each of the four matrices to compute the condition number more efficiently. Note that, $\mathbf{Q}_\mathbf{A}^T \mathbf{a}_i = \begin{bmatrix} \mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_i \\ \mathbf{0} \end{bmatrix}$ and similar for \mathbf{B} and \mathbf{C} , we have that

$$\sigma_{\min}([\mathbf{U}_1^{(1)} \quad \dots \quad \mathbf{U}_R^{(1)}]) = \sigma_{\min}([\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_1 \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_1 \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_1 \quad \dots \quad \mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_R \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_R \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_R]).$$

The reduced matrix above is of dimension $R^3 \times R$ instead of $s^3 \times R$. Further, we can choose the columns $\mathbf{Q}_{\mathbf{a}_i}^\perp$ so that $\bar{\mathbf{Q}}_{\mathbf{a}_i}^\perp = \mathbf{Q}_\mathbf{A}^T \mathbf{Q}_{\mathbf{a}_i}^\perp = \begin{bmatrix} \mathbf{Q}_{\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_1}^\perp & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$, and similar for \mathbf{B} and \mathbf{C} . Consequently, for $j = 2$,

$$\begin{aligned} \sigma_{\min}([\mathbf{U}_1^{(2)} \quad \dots \quad \mathbf{U}_R^{(2)}]) &= \sigma_{\min}\left(\left[\begin{bmatrix} \mathbf{Q}_{\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_1}^\perp & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_1 \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_1 \quad \dots \end{bmatrix}\right]\right) \\ &= \min\{\sigma_{\min}([\mathbf{Q}_{\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_1}^\perp \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_1 \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_1 \quad \dots]), \sigma_{\min}([\mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_1 \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_1 \quad \dots])\} \\ &= \sigma_{\min}([\mathbf{Q}_{\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_1}^\perp \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_1 \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_1 \quad \dots \quad \mathbf{Q}_{\mathbf{Q}_\mathbf{A}^{(1)T} \mathbf{a}_R}^\perp \otimes \mathbf{Q}_\mathbf{B}^{(1)T} \mathbf{b}_R \otimes \mathbf{Q}_\mathbf{C}^{(1)T} \mathbf{c}_R]), \end{aligned}$$

and similar for $j = 3, 4$. The dimensions of the above reduced matrix are $R^3 \times R(R - 1)$, hence a direct computation of the singular value decomposition can be used to compute the condition number with cost $O(R^7)$.

All the above arguments can be generalized to an order N non equidimensional tensor. Therefore, we showed that the condition number of CPD is invariant to the following transformation

$$\kappa([\![\lambda; \mathbf{A}_1 \dots, \mathbf{A}_N]\!]) = \kappa([\![\lambda; \mathbf{Q}_{\mathbf{A}_1}^{(1)T} \mathbf{A}_1, \dots, \mathbf{Q}_{\mathbf{A}_N}^{(N)T} \mathbf{A}_N]\!]),$$

where $\forall i \in \{1, \dots, N\}$, columns of $\mathbf{Q}_{\mathbf{A}_i}^{(1)}$ are an orthonormal basis of the column space of \mathbf{A}_i .

REFERENCES

- [1] E. Acar, D. M. Dunlavy, and T. G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, 2011.
- [2] A. Afshar, K. Yin, S. Yan, C. Qian, J. C. Ho, H. Park, and J. Sun. Swift: Scalable wasserstein factorization for sparse nonnegative tensors. In *Proceedings of the AAAI Conference*, 2021.
- [3] G. Ballard, K. Hayashi, and R. Kannan. Parallel nonnegative CP decomposition of dense tensors. *arXiv preprint arXiv:1806.07985*, 2018.
- [4] G. Ballard, N. Knight, and K. Rouse. Communication lower bounds for matricized tensor times Khatri-Rao product. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 557–567. IEEE, 2018.
- [5] C. Battaglino, G. Ballard, and T. G. Kolda. A practical randomized CP tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.
- [6] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.
- [7] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.
- [8] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650, 2018.
- [9] P. Breiding and N. Vannieuwenhoven. The condition number of join decompositions. *SIAM Journal on Matrix Analysis and Applications*, 39(1):287–309, 2018.
- [10] R. Bro. PARAFAC tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.
- [11] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35(3):283–319, 1970.
- [12] M. P. Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.
- [13] J. Choi, X. Liu, S. Smith, and T. Simon. Blocking optimization techniques for sparse tensor computation. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 568–577. IEEE, 2018.
- [14] P. Comon. Tensor diagonalization, a useful tool in signal processing. *IFAC Proceedings Volumes*, 27(8):77–82, 1994.
- [15] F. Cong, Q.-H. Lin, L.-D. Kuang, X.-F. Gong, P. Astikainen, and T. Ristaniemi. Tensor decomposition of EEG signals: A brief review. *Journal of neuroscience methods*, 248:59–69, 2015.
- [16] C.-F. Cui, Y.-H. Dai, and J. Nie. All real eigenvalues of symmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 35(4):1582–1601, 2014.
- [17] M. Cuturi and D. Avis. Ground metric learning. *The Journal of Machine Learning Research*, 15(1):533–564, 2014.
- [18] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.
- [19] R. A. Harshman. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. 1970.
- [20] K. Hayashi, G. Ballard, J. Jiang, and M. Tobia. Shared memory parallelization of MTTKRP for dense tensors. *arXiv preprint arXiv:1708.08976*, 2017.
- [21] C. J. Hillar and L.-H. Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):45:1–45:39, Nov. 2013.
- [22] F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Studies in Applied Mathematics*, 6(1-4):164–189, 1927.
- [23] A. Hyvärinen. Survey on independent component analysis. 1999.
- [24] L. Karlsson, D. Kressner, and A. Uschmajew. Parallel algorithms for tensor completion in the CP format. *Parallel Computing*, 57:222–234, 2016.
- [25] O. Kaya. *High performance parallel algorithms for tensor decompositions*. PhD thesis, 2017.
- [26] O. Kaya and Y. Robert. Computing dense tensor decompositions with optimal dimension trees. *Algorithmica*, 81(5):2092–2121, 2019.
- [27] O. Kaya and B. Uçar. *Parallel CP decomposition of sparse tensors using dimension trees*. PhD thesis, Inria-Research Centre Grenoble–Rhône-Alpes, 2016.
- [28] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.
- [29] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
- [30] T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

- [31] B. Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.
- [32] G. Li, L. Qi, and G. Yu. The z-eigenvalues of a symmetric tensor and its application to spectral hypergraph theory. *Numerical Linear Algebra with Applications*, 20(6):1001–1029, 2013.
- [33] J. Li, K. Usevich, and P. Comon. Globally convergent jacobi-type algorithms for simultaneous orthogonal symmetric tensor diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 39(1):1–22, 2018.
- [34] M. Liang and B. Zheng. Further results on moore–penrose inverses of tensors with application to tensor nearness problems. *Computers & Mathematics with Applications*, 77(5):1282–1293, 2019.
- [35] 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*, pages 129–132. IEEE, 2005.
- [36] L. Ma and E. Solomonik. Accelerating alternating least squares for tensor decomposition by pairwise perturbation. *arXiv preprint arXiv:1811.10573*, 2018.
- [37] L. Ma and E. Solomonik. Efficient parallel cp decomposition with pairwise perturbation and multi-sweep dimension tree. In *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 412–421. IEEE, 2021.
- [38] K. Maruhashi, F. Guo, and C. Faloutsos. Multiaspectforensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 203–210. IEEE, 2011.
- [39] B. C. Mitchell and D. S. Burdick. Slowly converging PARAFAC sequences: swamps and two-factor degeneracies. *Journal of Chemometrics*, 8(2):155–168, 1994.
- [40] D. Mitchell, N. Ye, and H. De Sterck. Nesterov acceleration of alternating least squares for canonical tensor decomposition. *arXiv preprint arXiv:1810.05846*, 2018.
- [41] K. R. Murphy, C. A. Stedmon, D. Graeber, and R. Bro. Fluorescence spectroscopy and multi-way techniques. PARAFAC. *Analytical Methods*, 5(23):6557–6566, 2013.
- [42] D. Nion and L. De Lathauwer. An enhanced line search scheme for complex-valued tensor decompositions. Application in DS-CDMA. *Signal Processing*, 88(3):749–755, 2008.
- [43] P. Paatero. A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 38(2):223–242, 1997.
- [44] A.-H. Phan, P. Tichavský, and A. Cichocki. Fast alternating LS algorithms for high order CANDECOMP/PARAFAC tensor factorizations. *IEEE Transactions on Signal Processing*, 61(19):4834–4846, 2013.
- [45] A.-H. Phan, P. Tichavsky, and A. Cichocki. Low complexity damped Gauss–Newton algorithms for CANDECOMP-/PARAFAC. *SIAM Journal on Matrix Analysis and Applications*, 34(1):126–147, 2013.
- [46] L. Qi, H. Chen, and Y. Chen. *Tensor eigenvalues and their applications*, volume 39. Springer, 2018.
- [47] M. Rajih, P. Comon, and R. A. Harshman. Enhanced line search: A novel method to accelerate PARAFAC. *SIAM journal on matrix analysis and applications*, 30(3):1128–1147, 2008.
- [48] M. D. Schatz, T. M. Low, R. A. van de Geijn, and T. G. Kolda. Exploiting symmetry in tensors for high performance: Multiplication with symmetric tensors. *SIAM Journal on Scientific Computing*, 36(5):C453–C479, 2014.
- [49] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582.
- [50] N. Singh, L. Ma, H. Yang, and E. Solomonik. Comparison of accuracy and scalability of Gauss–Newton and alternating least squares for CANDECOMC/PARAFAC decomposition. *SIAM Journal on Scientific Computing*, 43(4):C290–C311, 2021.
- [51] L. Sorber, M. Van Barel, and L. De Lathauwer. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank- $(l_r, l_r, 1)$ terms, and a new generalization. *SIAM Journal on Optimization*, 23(2):695–720, 2013.
- [52] L. Sun, B. Zheng, C. Bu, and Y. Wei. Moore–penrose inverse of tensors via einstein product. *Linear and Multilinear Algebra*, 64(4):686–698, 2016.
- [53] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, et al. PySCF: The Python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018.
- [54] P. Tichavský, A. H. Phan, and A. Cichocki. A further improvement of a fast damped Gauss–Newton algorithm for CANDECOMP-PARAFAC tensor decomposition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5964–5968. IEEE, 2013.
- [55] P. Tichavský, A. H. Phan, and A. Cichocki. Non-orthogonal tensor diagonalization, 2016.
- [56] G. Tomasi and R. Bro. PARAFAC and missing values. *Chemometrics and Intelligent Laboratory Systems*, 75(2):163–180, 2005.
- [57] G. Tomasi and R. Bro. A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis*, 50(7):1700–1734, 2006.
- [58] A. Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM Journal on Matrix Analysis and Applications*, 33(2):639–652, 2012.
- [59] K. Usevich, J. Li, and P. Comon. Approximate matrix and tensor diagonalization by unitary transformations: convergence of Jacobi-type algorithms. *SIAM Journal on Optimization*, 30(4):2998–3028, 2020.
- [60] N. Vannieuwenhoven. Condition numbers for the tensor rank decomposition. *Linear Algebra and Its Applications*, 535:35–86, 2017.
- [61] N. Vannieuwenhoven, K. Meerbergen, and R. Vandebril. Computing the gradient in optimization algorithms for the CP decomposition in constant memory through tensor blocking. *SIAM Journal on Scientific Computing*, 37(3):C415–C438, 2015.
- [62] G. Zen, E. Ricci, and N. Sebe. Simultaneous ground metric learning and matrix factorization with earth mover’s distance. In *2014 22nd International Conference on Pattern Recognition*, pages 3690–3695. IEEE, 2014.