A New Alternating Optimization Algorithm for CP Decomposition

Navjot Singh and Edgar Solomonik

Department of Computer Science
University of Illinois at Urbana-Champaign

SIAM Conference on Parallel Processing for Scientific Computing (PP22)
Outline

1. Overview
2. Motivation
3. New Alternating Update Scheme for CPD
4. Exact CP Decomposition
5. Approximate CP Decomposition
6. Conclusion and Future Work
Introduce computation of singular vals/vecs via considering multilinear function associated with the tensor with log barrier penalty

Critical points of the above spectrally diagonalize an order $N$ tensor

Analyze local convergence of the algorithm for exact CPD of rank lesser than mode lengths

Formulation that generalizes the algorithm to perform well conditioned\(^1\) approximate CPD

\(^1\)P. Breiding and N. Vannieuwenhoven, SIMAX 2018
Overview

- Tensor: A multidimensional array $\mathcal{X}$
- Indices: $x_{i_1,i_2\ldots i_N}$ imply order $= N$

- CP tensor decomposition breaks down a tensor into sum of rank 1 components.
- CPD of an order 3 tensor $\mathcal{X}$ with rank $R$ and factor matrices $A, B, C$: $\mathcal{X} = \langle A, B, C \rangle$

$$x_{ijk} = \sum_{l=1}^{R} a_{il}b_{jl}c_{kl}$$

Figure: S.He et. al. Tensor Decomposition Based Electrical Data Recovery
Motivation: Singular Vectors via Variational Approach

Analogous to obtaining eigenvalues via critical points of $x^T A x$ with unit $l^2$-norm constraints, L. Lim derives singular vectors and values of $A$ via

- critical points of $\frac{x^T A y}{\|x\|_2 \|y\|_2}$ with unit norm constraints.
- Lagrangian is

$$L(x, y, \sigma) = x^T A y - \sigma(\|x\|_2 \|y\|_2 - 1)$$

- First order conditions yields,

$$A \frac{y}{\|y\|_2} = \sigma \frac{x}{\|x\|_2}, \quad A^T \frac{x}{\|x\|_2} = \sigma \frac{y}{\|y\|_2}, \quad \|x\|_2 \|y\|_2 = 1$$

$$Av = \sigma u, \quad A^T u = \sigma v$$

Order 3 tensor eigen/singular values and vectors can be derived similarly

---

2Lek-Heng Lim Singular values and Eigenvalues of a tensor: A variational approach
Motivation: Singular Vectors via Lagrangian Optimization

One can also obtain singular values and vectors by considering bilinear form $f(x, y) = x^T A y$ with $\|x\|_2 \neq 0, \|y\|_2 \neq 0$,

$$\mathcal{L}_f(x, y) = x^T A y - \log(\|x\|_2 \|y\|_2)$$

Critical points satisfy $A v = \sigma u$ and $A^T u = \sigma v$ for

$$u = x/\|x\|_2, \quad v = y/\|y\|_2, \quad \sigma = 1/\|x\|_2 \|y\|_2.$$

Similarly for an order 3 tensor $T$, consider

$$\mathcal{L}_f(x, y, z) = \sum_{i,j,k} t_{ijk} x_i y_j z_k - \log(\|x\|_2 \|y\|_2 \|z\|_2).$$

Critical points satisfy equations

$$\sum_{j,k} t_{ijk} v_j w_k = \sigma u, \quad \sum_{i,k} t_{ijk} u_i w_k = \sigma v, \quad \sum_{i,j} t_{ijk} u_i v_j = \sigma w$$

with $u = x/\|x\|_2, \quad v = y/\|y\|_2, \quad w = z/\|z\|_2, \quad \sigma = 1/(\|x\|_2 \|y\|_2 \|z\|_2)$.
Motivation: Spectral Diagonalization

This notion can be generalized for $R > 1$ vectors, since

$$x^T A y = \langle A, x y^T \rangle$$

consider $\langle A, B \rangle = \langle \text{vec}(A), \text{vec}(B) \rangle$,

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

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

$$= \text{tr}(X^T A Y) - \frac{1}{2} \text{tr}(\log(X^T X Y^T Y)).$$

The critical points of $\mathcal{L}_f$ satisfy $AYX^T \approx I$ and $A^TXY^T \approx I$

- $X \rightarrow$ invariant subspace of $AA^T$
- $Y \rightarrow$ invariant subspace of $A^T A$

and diagonalize $A$ in the sense that

$$X^T A Y = I$$
Motivation: Spectral Diagonalization

Similarly for an order 3 tensor $\mathcal{T}$, $\langle \mathcal{T}, \mathcal{Y} \rangle = \langle \text{vec}(\mathcal{T}), \text{vec}(\mathcal{Y}) \rangle$

$$\mathcal{L}_f(X, Y, Z) = \langle \mathcal{T}, [X, Y, Z] \rangle - \frac{1}{2} \text{tr}(\log(X^TXY^TYZ^T))$$

The critical points of $\mathcal{L}_f$ diagonalize the $\mathcal{T}$ such that

$$\mathcal{P} = \mathcal{T} \times_1 X \times_2 Y \times_3 Z,$$

implying $\mathcal{P}$ has $R$ elementary eigenvectors with unit eigenvalues, different from $\mathcal{P}$. Comon’s idea of diagonalizing a tensor with orthogonal matrices

3P. Comon, M. Sorensen Tensor diagonalization with orthogonal transformation
New Alternating Update Scheme

Consider a rank $R$ CP decomposition of a tensor $\mathbf{x} \in \mathbb{R}^{s \times s \times s}$,

$$\mathbf{x} = [A, B, C], \text{ i.e. } x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr},$$

which may be obtained by ALS via minimizing

$$f(A, B, C) = \frac{1}{2} \| \mathbf{x} - [A, B, C] \|_F^2$$

by alternating updates such as

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

We propose a different update, which for $R \leq s$ is,

$$A = X_{(1)}(C^{\dagger T} \odot B^{\dagger T})$$
Convergence to Exact Decomposition

When seeking an exact CP decomposition of rank $R \leq s$

- ALS achieves a linear convergence rate \(^4\)
- High order convergence possible via optimizing all factors, eg. using Gauss-Newton \(^5,6,7\), but is expensive
- The proposed algorithm achieves at least quartic convergence per sweep of alternating updates
  - per subsweep, convergence order is $\alpha$ where $\alpha$ is the real positive root of $x^{N-1} - \sum_{i=0}^{N-2} x^i$ for order $N$ tensor, i.e., $(1 + \sqrt{5})/2$ for order 3.
  - cost per iteration roughly the same as ALS (dominated by MTTKRP) and therefore easily parallelizable

---

\(^4\) A. Uschmajew, SIMAX 2012

\(^5\) P. Paatero, Chemometrics and Intelligent Laboratory Systems 1997

\(^6\) A.H. Phan, P. Tichavsky, A. Cichocki, SIMAX 2013

\(^7\) N. Singh, L. Ma, H. Yang, E.S., SISC 2021.
The error in one factor scales with the product of errors in the other factors.

**Lemma**

Suppose $\mathbf{X} = [A^{(1)}, \ldots, A^{(N)}]$, where each $A^{(i)} \in \mathbb{R}^{s_i \times R}$ is full rank with $s_i \geq R$ and $\tilde{A}^{(n)} = A^{(n)} D^{(n)} + \Delta^{(n)}$ and satisfies $\|\Delta^{(n)}\|_F = \epsilon_n$ for $n = 1, \ldots, N - 1$, then $\exists \epsilon > 0$ such that if $\epsilon_n < \epsilon$ for $n = 1, \ldots, N - 1$,

$$\tilde{A}^{(N)} = \mathbf{X}^{(N)} (\tilde{A}^{(1)}^T \odot \cdots \odot \tilde{A}^{(N-1)}^T)$$

satisfies

$$\|\tilde{A}^{(N)} D^{(N)} - A^{(N)}\|_F = O \left( \prod_{n=1}^{N} \epsilon_{n-1} \right),$$

for some diagonal $D^{(N)}$. 
A rough sketch of proof of the above Lemma follows from substituting true decomposition in the update rule

\[
\tilde{A}^{(N)} = A^{(N)} \left( (\tilde{A}^{(1)\dagger} A^{(1)}) \star \cdots \star (\tilde{A}^{(N-1)\dagger} A^{(N-1)}) \right)^T
\]

\[
= A^{(N)} \left( (D^{(1)} - \tilde{A}^{(1)\dagger} \Delta^{(1)}) \star \cdots \star (D^{(N-1)} - \tilde{A}^{(N-1)\dagger} \Delta^{(N-1)}) \right)^T
\]

\[
= A^{(N)} \left( D + (-1)^{N-1} \tilde{A}^{(1)\dagger} \Delta^{(1)} \star \cdots \star \tilde{A}^{(N-1)\dagger} \Delta^{(N-1)} \right)^T,
\]

where \( D \) is diagonal matrix.
Rate of convergence of AMDM only depends on the (matrix) rank of underlying factors.

**Figure:** CP Decomposition of synthetic tensors with rank 20 and $100^3$ entries.
Approximate CP Decomposition

- The proposed update for $A$ minimizes
  \[
  \frac{1}{2} \| (\mathbf{X} - [A, B, C])_{(1)}(C^T \otimes B^T) \|_F^2.
  \]

- The residual being
  \[
  X_{(1)}(C^T \otimes B^T) - A(I \otimes I)
  \]

- Residual transformation tends to equalize the weight of contribution of the error associated with different rank-1 parts of the CP decompositions.

- Similar property observed when Mahalanobis distance metric is considered
Mahalanobis Distance Objective

- Original motivation for the method came from optimizing CPD with general distance metrics with Ardavan Afshar, C. Qian, and J. Sun.\(^8\)
- Consider an order 3 tensor \(X\) and Mahalanobis distance objective

\[
f(A, B, C) = \frac{1}{2}\|X - Y\|_M = \frac{1}{2}\text{vec}(X - Y)^T M \text{vec}(X - Y),
\]

where \(Y = [A, B, C]\),

with \(M = \bigotimes_{k=1}^3 M^{(k)-1}\) being SPD.

Minimization with respect to \(A\) results in the following update

\[
AZ = X^{(1)} L,
\]

where \(L = \left( M^{(3)-1} C \right) \bigodot \left( M^{(2)-1} B \right)\),

and \(Z = \left( B^T M^{(2)-1} B \right) \ast \left( C^T M^{(3)-1} C \right)\).

\(^8\) A. Ardavan, K. Yin, S. Yan, C. Qian, J.C. Ho, H. Park, and J. Sun, AAAI 2021
Generalizing AMDM to Hybrid Algorithms

Decompose factors into sum of two matrices and using first $\theta$ singular values and vectors for each factor to construct $M^{(k)}$,

\[
M^{(1)} = A_1 A_1^T + (I - A_1 A_1^\dagger),
M^{(2)} = B_1 B_1^T + (I - B_1 B_1^\dagger),
M^{(3)} = C_1 C_1^T + (I - C_1 C_1^\dagger).
\]

leads to an update that is a hybrid of AMDM and ALS, since

\[
AZ = X_{(1)} L,
\]

where \( L = \left( (C_1^\dagger T + C_2) \odot (B_1^\dagger T + B_2) \right), \)

and \( Z = \left( (C_1^\dagger C_1 + C_2^T C_2) \ast (B_1^\dagger B_1 + B_2^T B_2) \right). \)
Generalizing AMDM for All CP Ranks

It can be theoretically shown that AMDM converges linearly for CP rank $R > s$

![Graph showing linear convergence](image)

**Figure:** Linear convergence for exact CPD of a $100 \times 100 \times 100$ tensor with CP rank $R = 200$
Approximate Decomposition Results with AMDM

Using Hybrid algorithms leads to better conditioned and accurate decompositions.
Open Questions about AMDM

- Relation of AMDM with eigenvectors or singular vectors of a tensor
- Other views of the method (other than Mahalanobis Distance minimization)
- Existence of stationary points of AMDM for rank lesser than mode lengths case
- Quantifying conditioning of the alternating update in AMDM