A New Alternating Optimization Algorithm for CP Decomposition

Navjot Singh and Edgar Solomonik

@CS@Illinois

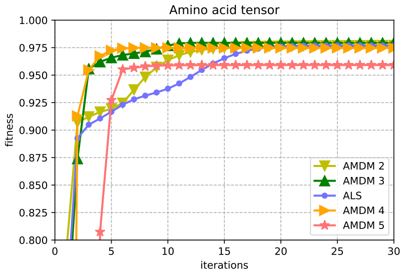
Department of Computer Science
University of Illinois at Urbana-Champaign

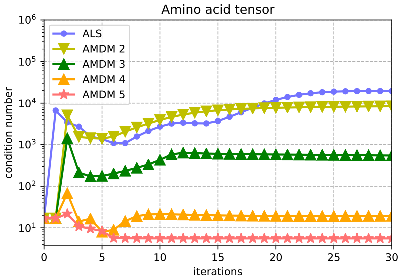
SIAM Conference on Parallel Processing for Scientific Computing (PP22)

Outline

Highlights

- 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 $\boldsymbol{\mathcal{X}}$
- Indices: $x_{i_1,i_2 \dots i_N}$ imply order$= N$

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

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

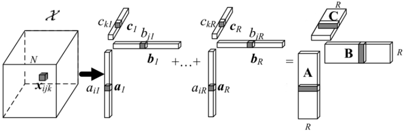


Figure: S.He et. al. Tensor Decomposition Based Electrical Data Recovery

Motivation: Singular Vectors via Variational Approach

Analogous to obtaining eigenvalues via critical points of $\mathbf{x}^T\mathbf{A}\mathbf{x}$ with unit l^2-norm constraints,

L. Lim derives singular vectors and values [2] of $\boldsymbol{A}$ via

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

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

- First order conditions yields,

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

$$\boldsymbol{A}\mathbf{v} = \sigma\mathbf{u}, \quad \boldsymbol{A}^T\mathbf{u} = \sigma\mathbf{v}$$

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

[2] Lek-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 $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \boldsymbol{A} \mathbf{y}$ with $\|\mathbf{x}\|_2 \neq 0$, $\|\mathbf{y}\|_2 \neq 0$,

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

Critical points satisfy $\boldsymbol{A}\mathbf{v} = \sigma \mathbf{u}$ and $\boldsymbol{A}^T \mathbf{u} = \sigma \mathbf{v}$ for

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

Similarly for an order 3 tensor $\boldsymbol{\mathcal{T}}$, consider

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

Critical points satsify equations

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

with $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|_2$, $\mathbf{v} = \mathbf{y}/\|\mathbf{y}\|_2$,$\mathbf{w} = \mathbf{z}/\|\mathbf{z}\|_2$, $\sigma = 1/(\|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \|\mathbf{z}\|_2)$

Motivation: Spectral Diagonalization

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

$$\mathbf{x}^T \boldsymbol{A} \mathbf{y} = \langle \boldsymbol{A}, \mathbf{x}\mathbf{y}^T \rangle$$

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

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

$$\begin{aligned} \mathcal{L}_f(\mathbf{X}, \mathbf{Y}) &= \langle \boldsymbol{A}, \boldsymbol{X}\boldsymbol{Y}^T \rangle - \frac{1}{2}(\log(\det(\boldsymbol{X}^T\boldsymbol{X})) - \log(\det(\boldsymbol{Y}^T\boldsymbol{Y}))) \\ &= \text{tr}(\mathbf{X}^T\boldsymbol{A}\mathbf{Y}) - \frac{1}{2}\text{tr}(\log(\mathbf{X}^T\mathbf{X}\mathbf{Y}^T\mathbf{Y})). \end{aligned}$$

The critical points of $\mathcal{L}_f$ satisfy $\boldsymbol{A}\boldsymbol{Y}\boldsymbol{X}^T \cong \boldsymbol{I}$ and $\boldsymbol{A}^T\boldsymbol{X}\boldsymbol{Y}^T \cong \boldsymbol{I}$

- $\boldsymbol{X} \rightarrow$ invariant subspace of $\boldsymbol{A}\boldsymbol{A}^T$
- $\boldsymbol{Y} \rightarrow$ invariant subspace of $\boldsymbol{A}^T\boldsymbol{A}$

and diagonalize $\boldsymbol{A}$ in the sense that

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{I}$$

Motivation: Spectral Diagonalization

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

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

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

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

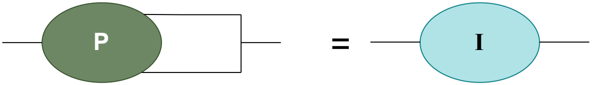


Figure: $p_{ijj} = p_{iij} = p_{iji} = \delta_{ij}$

implying $\boldsymbol{\mathcal{P}}$ has R elementary eigenvectors with unit eigenvalues, different from P. Comon's idea of diagonalizing a tensor with orthogonal matrices [3]

[3] P. Comon, M. Sorensen Tensor diagonalization with orthogonal transformation

New Alternating Update Scheme

Consider a rank R CP decomposition of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{s\times s\times s}$,

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

which maybe obtained by ALS via minimizing

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

by alternating updates such as

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

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

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

Convergence to Exact Decomposition

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

- 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 atleast quartic convergence per sweep of alternating updates
  - per subsweep, convergence order is α where α 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.

Exact Decomposition Error Analysis

The error in one factor scales with the product of errors in the other factors

Lemma

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

$$\tilde{\boldsymbol{A}}^{(N)} = \boldsymbol{\mathcal{X}}_{(N)}(\tilde{\boldsymbol{A}}^{(1)\dagger T} \odot \cdots \odot \bar{\boldsymbol{A}}^{(N-1)\dagger T})$$

satisfies

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

for some diagonal $\boldsymbol{D}^{(N)}$.

Exact Decomposition Error Analysis

A rough sketch of proof of the above Lemma follows from substituting true decomposition in the update rule

$$
\begin{aligned}
\tilde{\boldsymbol{A}}^{(N)} &= \boldsymbol{A}^{(N)} \Big((\tilde{\boldsymbol{A}}^{(1)\dagger} \boldsymbol{A}^{(1)}) * \cdots * (\tilde{\boldsymbol{A}}^{(N-1)\dagger} \boldsymbol{A}^{(N-1)}) \Big)^T \\
&= \boldsymbol{A}^{(N)} \left((\boldsymbol{D}^{(1)} - \tilde{\boldsymbol{A}}^{(1)\dagger} \boldsymbol{\Delta}^{(1)}) * \cdots * (\boldsymbol{D}^{(N-1)} - \tilde{\boldsymbol{A}}^{(N-1)\dagger} \boldsymbol{\Delta}^{(N-1)}) \right)^T \\
&= \boldsymbol{A}^{(N)} \left(\boldsymbol{D} + (-1)^{N-1} \tilde{\boldsymbol{A}}^{(1)\dagger} \boldsymbol{\Delta}^{(1)} * \cdots * \tilde{\boldsymbol{A}}^{(N-1)\dagger} \boldsymbol{\Delta}^{(N-1)} \right)^T,
\end{aligned}
$$

where $\boldsymbol{D}$ is diagonal matrix.

Exact Decomposition Experimental Performance

Rate of convergence of AMDM only depends on the (matrix) rank of underlying factors

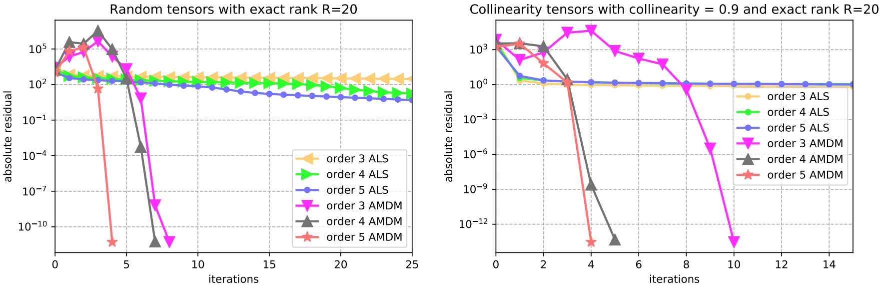


Figure: CP Decomposition of synthetic tensors with rank 20 and 100^3 entries

Approximate CP Decomposition

- The proposed update for $\boldsymbol{A}$ minimizes

$$\frac{1}{2}\|(\boldsymbol{\mathcal{X}} - [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!])_{(1)}(\boldsymbol{C}^{\dagger T} \otimes \boldsymbol{B}^{\dagger T})\|_F^2.$$

- The residual being

$$\boldsymbol{X}_{(1)}\big(\boldsymbol{C}^{\dagger T} \odot \boldsymbol{B}^{\dagger T}\big) - \boldsymbol{A}\big(\boldsymbol{I} \odot \boldsymbol{I}\big)$$

- 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 $\boldsymbol{\mathcal{X}}$ and Mahalanobis distance objective

$$f(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) = \frac{1}{2}\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{Y}}\|_{\boldsymbol{M}} = \frac{1}{2}\text{vec}\big(\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{Y}}\big)^T\boldsymbol{M}\text{vec}\big(\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{Y}}\big),$$
$$\text{where } \boldsymbol{\mathcal{Y}} = [\![\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}]\!],$$
$$\text{with } \boldsymbol{M} = \bigotimes_{k=1}^{3}\boldsymbol{M}^{(k)-1} \text{ being SPD.}$$

Minimization with respect to $\boldsymbol{A}$ results in the following update

$$\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{X}_{(1)}\boldsymbol{L},$$
$$\text{where } \boldsymbol{L} = \left(\boldsymbol{M}^{(3)-1}\boldsymbol{C}\right)\odot\left(\boldsymbol{M}^{(2)-1}\boldsymbol{B}\right),$$
$$\text{and } \boldsymbol{Z} = \left(\boldsymbol{B}^T\boldsymbol{M}^{(2)-1}\boldsymbol{B}\right)*\left(\boldsymbol{C}^T\boldsymbol{M}^{(3)-1}\boldsymbol{C}\right).$$

[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 θ singular values and vectors for each factor to construct $\boldsymbol{M}^{(k)}$,

$$\begin{aligned}
\boldsymbol{M}^{(1)} &= \boldsymbol{A}_1\boldsymbol{A}_1^T + (\boldsymbol{I} - \boldsymbol{A}_1{\boldsymbol{A}_1}^{\dagger}),\\
\boldsymbol{M}^{(2)} &= \boldsymbol{B}_1\boldsymbol{B}_1^T + (\boldsymbol{I} - \boldsymbol{B}_1{\boldsymbol{B}_1}^{\dagger}),\\
\boldsymbol{M}^{(3)} &= \boldsymbol{C}_1\boldsymbol{C}_1^T + (\boldsymbol{I} - \boldsymbol{C}_1{\boldsymbol{C}_1}^{\dagger}).
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{Z} &= \boldsymbol{X}_{(1)}\boldsymbol{L},\\
\text{where } \boldsymbol{L} &= \Big((\boldsymbol{C}_1^{\dagger T} + \boldsymbol{C}_2) \odot (\boldsymbol{B}_1^{\dagger T} + \boldsymbol{B}_2)\Big),\\
\text{and } \boldsymbol{Z} &= \Big((\boldsymbol{C}_1^{\dagger}\boldsymbol{C}_1 + \boldsymbol{C}_2^T\boldsymbol{C}_2) * (\boldsymbol{B}_1^{\dagger}\boldsymbol{B}_1 + \boldsymbol{B}_2^T\boldsymbol{B}_2)\Big).
\end{aligned}$$

Generalizing AMDM for All CP Ranks

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

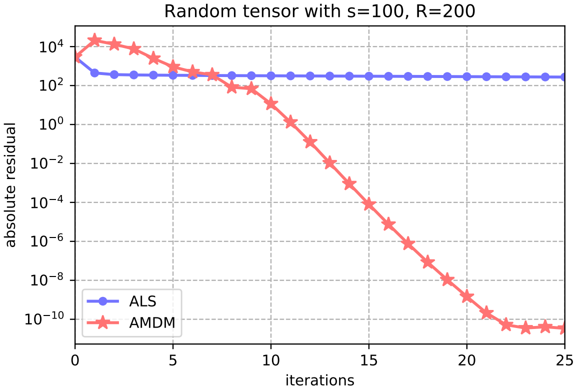
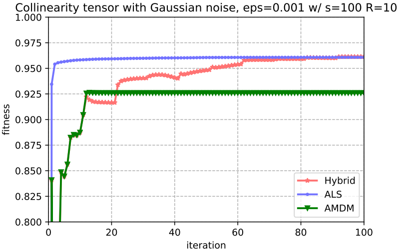
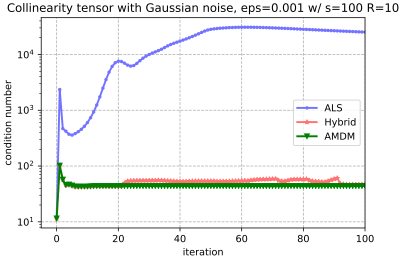


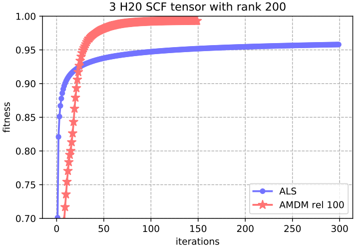
Figure: Linear convergence for exact CPD of a $100 \times 100 \times 100$ tensor with CP rank $R = 200$

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