Convergence Analysis on SVD-based Algorithms for Tensor Low Rank Approximations

Yu Guan

National University of Singapore

November 21, 2018

Outline

Outline

Outline

Outline

Topics

- Best rank-1 approximation of tensors;
- Orthogonal low rank tensor approximation;
- Convergence analysis of ADM.

What is a tensor?

- An order-k tensor can be regarded as a k-dimensional array of real or complex numbers on which algebraic operations generalizing analogous operations on matrices are defined.
- A vector is a tensor of order 1.
- A matrix is a tensor of order 2.

- A real-valued tensor of order-k can be represented by $T = [\tau_{i_1,\ldots,i_k}] \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_k}$ with elements $\tau_{i_1,\ldots,i_k}$ accessed via k indices.
- A tensor of the form

$$\bigotimes_{\ell=1}^{k} \mathbf{u}^{(\ell)} = \mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{u}^{(k)} := [u_{i_1}^{(1)} \ldots u_{i_k}^{(k)}]$$

where elements are the products of entries from vectors $\mathbf{u}^{(\ell)} \in \mathbb{R}^{I_\ell}$, $\ell = 1, \ldots, k$, is said to be of rank one.

Best Rank-1 Approximation

- Given $T \in \mathbb{R}^{I_1 \times \cdots \times I_k}$, determine
  - unit vectors $\mathbf{u}^{(\ell)} \in \mathbb{R}^{I_\ell}$, $\ell = 1, \ldots k$, and
  - scalar $\lambda \in \mathbb{R}$,

  such that

$$\left\| T - \lambda \mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{u}^{(k)} \right\|_F^2$$

  is minimized.
  - For fixed unit vectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}$, the optimal value of λ is

$$\lambda = \lambda\left(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}\right) = \left\langle T, \bigotimes_{\ell=1}^{k} \mathbf{u}^{(\ell)} \right\rangle.$$

Symmetric Tensor

An order-k square tensor T is said to be symmetric if

$$\mathcal{T}_{i_1,\ldots,i_k} = \mathcal{T}_{i_{\sigma(1)},\ldots,i_{\sigma(k)}}$$

with respect to all possible permutations σ over the integers $\{1,\ldots,k\}$.

Background of Symmetric Case

- (Qi, 2011) conjectured and (Zhang etc., 2012) proved that the best symmetric rank-1 approximation to a symmetric tensor is its best rank-1 approximation .
- The best rank-1 approximation to a symmetric tensor 'can be chosen' symmetric (Friedland, 2013).
- There might be non-symmetric best rank-1 approximations (Friedland, 2013) for a symmetric tensor.

Background of Algorithms

- The alternating least squares (ALS) method works on improving one factor $\mathbf{u}^{(\ell)}$ a time (Kroonenberg etc., 1980).
- However, the method suffers from slow convergence and easy stagnation at a local solution.
- Alternating two factors simultaneously by SVD was mentioned in (Lathauwer etc., 2000) with no particular elaboration.
- (Friedland etc., 2013) was more carefully postulated with numerical testing on some synthetic and real data sets of third-order tensors.

Comparison of Two Ideas

- SVD approach has the obvious advantage that, starting from the same point, one step of SVD-based iteration is superior to two consecutive steps of ALS iteration.
- There is no theory at present to support that the improvement by the SVD-based iteration will continue to be superior in the long run.
- Through numerical experiments, however, it has been suggested that for large scale data the SVD-based method might have better limiting behavior leading to better approximations (Friedland etc., 2013).

Convergence

- The convergence theory for the ALS method was established much later than the method had been put into practice (Comon etc., 2009), (Uschmajew, 2012) and (Wang etc., 2014).
- For the SVD-based algorithm, the convergence of the generalized Rayleigh quotients is obvious, but the convergence analysis for the iterates themselves has been elusive in the literature (Friedland etc., 2013).
- (Yang etc., 2016) investigates the convergence theory by using the Łojasiewicz gradient inequality.

Our Contributions

- We provide a rigorous mathematical proof for the convergence of iterates from specific SVD-based algorithms.
- Our approach relies on only the continuity of singular vectors and real analysis.

Lemma

Given a matrix $A \in \mathbb{R}^{m \times n}$, then the global maximum of the generalized Rayleigh quotient

$$\max_{\substack{\mathbf{y} \in \mathbb{R}^m, \|\mathbf{y}\| = 1 \\ \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\| = 1}} \mathbf{y}^\top A \mathbf{z}$$

is precisely the largest singular value σ_1 of A, where the global maximizer $(\mathbf{y}_1, \mathbf{z}_1)$ consists of precisely the corresponding left and right singular vectors. The best rank-1 approximation to A is given by $\sigma_1 \mathbf{y}_1 \mathbf{z}_1^\top$. In the event that $A \in \mathbb{R}^{m \times m}$ is symmetric and that the largest singular value of A is simple, then $\mathbf{y} = \pm \mathbf{z}$ depending on the sign of the dominant eigenvalue $\lambda_1 = \pm \sigma_1$ and, hence, the best rank-1 approximation to A is symmetric.

Linear Mapping

Given a fixed partitioning $[\![k]\!] = \boldsymbol{\alpha} \cup \boldsymbol{\beta}$, we shall regard an order-k tensor $T \in \mathbb{R}^{I_1 \times \dots \times I_k}$ as a "matrix representation" of a linear operator mapping order-s tensors to order-t tensors. Specifically, we identify T with the linear map

$$\mathscr{T}_{\boldsymbol{\beta}} : \mathbb{R}^{I_{\alpha_1} \times \dots \times I_{\alpha_s}} \to \mathbb{R}^{I_{\beta_1} \times \dots \times I_{\beta_t}},$$

such that for any $S \in \mathbb{R}^{I_{\alpha_1} \times \dots \times I_{\alpha_s}}$,

Linear Mapping

we have

$$\mathscr{T}_{\boldsymbol{\beta}}(S) := T \circledast_{\boldsymbol{\beta}} S = [\langle \mathcal{T}_{[:|\ell_1,\ldots,\ell_t]}, S \rangle] \in \mathbb{R}^{I_{\beta_1} \times \ldots \times I_{\beta_t}}$$

where

$$\langle \mathcal{T}_{[:|\ell_1,\ldots,\ell_t]}, S \rangle := \sum_{i_1=1}^{I_{\alpha_1}} \cdots \sum_{i_s=1}^{I_{\alpha_s}} \mathcal{T}_{[i_1,\ldots,i_s|\ell_1,\ldots,\ell_t]} s_{i_1,\ldots,i_s}$$

is the Frobenius inner product generalized to multi-dimensional arrays.

Cyclic Progression for Symmetric Case (A1)

for $p = 0, 1, \cdots,$ **do**
    for $\ell = 1, 2, \cdots, k-1,$ **do**
        $\boldsymbol{\beta}_\ell = (\ell, \ell+1)$
        $C_{[p]}^{(\ell)} = T \circledast_{\boldsymbol{\beta}_\ell} \bigotimes_{i=1}^{\ell-1} \mathbf{u}_{[p+1]}^{(i)} \otimes \bigotimes_{i=\ell+2}^{k} \mathbf{u}_{[p]}^{(i)}$
        $[\mathbf{u}, s, \mathbf{v}] = \text{svds}(C_{[p]}^{(\ell)}, 1)$ {Dominant singular value triplet via Matlab routine svds}
        if $u_1 < 0$ **then**
            $\mathbf{u} = -\mathbf{u}$ {Assume the generic case that $u_1 \neq 0$; otherwise, use another entry.}
        end if
        $\mathbf{u}_{[p+1]}^{(\ell)} := \mathbf{u}$ {If $\ell = 1$, this is $\widehat{\mathbf{u}}_{[p+1]}^{(1)}$; otherwise this is the second update $\mathbf{u}_{[p+1]}^{(\ell)}$, if $2 \leq \ell < k$.}
        $\widehat{\mathbf{u}}_{[p+1]}^{(\ell+1)} := \mathbf{u}$ {Skipping this step will not affect $C_{[p]}^{(\ell+1)}$ at Line 4.}
        $\lambda_{[p+1]}^{(\ell)} := s$
    end for
    $\boldsymbol{\beta}_k = (k, 1)$
    $C_{[p]}^{(k)} = T \circledast_{\boldsymbol{\beta}_k} \bigotimes_{i=2}^{k-1} \mathbf{u}_{[p+1]}^{(i)}$
    $[\mathbf{u}, s, \mathbf{v}] = \text{svds}(C_{[p]}^{(k)}, 1)$ {Dominant singular value triplet via Matlab routine svds}
    $\mathbf{u}_{[p+1]}^{(k)} := \mathbf{u}$ {Adjust the sign properly as in Line 6.}
    $\mathbf{u}_{[p+1]}^{(1)} := \mathbf{u}$
    $\lambda_{[p+1]}^{(k)} := s$
end for

Randomization for Symmetric Case (A2)

$t \leftarrow 0$
$\lambda_0 \leftarrow \left\langle T, \bigotimes_{\ell=1}^{k} \mathbf{u}^{(\ell)} \right\rangle$
repeat
  $t \leftarrow t + 1$
  $\sigma \leftarrow$ random permutation of $\{1, \ldots, k\}$
  $\boldsymbol{\beta}_t \leftarrow (\sigma_{k-1}, \sigma_k)$ {Randomly select two factors}
  $C_t \leftarrow T \circledast_{\boldsymbol{\beta}_t} \bigotimes_{i=1}^{k-2} \mathbf{u}^{(\sigma_i)}$
  $[\mathbf{u}_t, s_t, \mathbf{v}_t] = \text{svds}(C_t, 1)$ {Dominant singular value triplet via Matlab routine svds}
  if $(\mathbf{u}_t)_1 < 0$ **then**
    $\mathbf{u}_t = -\mathbf{u}_t$
  end if
  $\lambda_t \leftarrow s_t$
  $\mathbf{u}^{(\sigma_{k-1})}, \mathbf{u}^{(\sigma_k)} \leftarrow \mathbf{u}_t$
until λ_t meets convergence criteria

Post-randomization for Symmetric Case (A3)

$t \leftarrow 0$
$\mu_0 \leftarrow \left\langle T, \bigotimes_{\ell=1}^{k} \mathbf{u}^{(\ell)} \right\rangle$
repeat
  $t \leftarrow t + 1$
  $C_t \leftarrow T \circledast \bigotimes_{i=1}^{k-2} \mathbf{u}^{(i)}$
  $[\mathbf{u}_t, s_t, \mathbf{v}_t] = \text{svds}(C_t, 1)$ {Dominant singular value triplet via Matlab routine svds}
  $\sigma \leftarrow$ random permutation of $\{1, \ldots, k\}$
  if $(\mathbf{u}_t)_1 < 0$ **then**
    $\mathbf{u}_t = -\mathbf{u}_t$
  end if
  $\mu_t \leftarrow s_t$
  $\mathbf{u}^{(\sigma_{k-1})}, \mathbf{u}^{(\sigma_k)} \leftarrow \mathbf{u}_t$ {Randomly replace two factors}
until μ_t meets convergence criteria

Cyclic Progression for Non-symmetric Case (A4)

for $p = 0, 1, \cdots,$ **do**

  for $\ell = 1, 2, \cdots, k-1,$ **do**

    $\boldsymbol{\beta}_\ell = (\ell, \ell+1)$

    $C_{[p]}^{(\ell)} = T \circledast_{\boldsymbol{\beta}_\ell} \bigotimes_{i=1}^{\ell-1} \mathbf{u}_{[p+1]}^{(i)} \otimes \bigotimes_{i=\ell+2}^{k} \mathbf{u}_{[p]}^{(i)}$ {A matrix of size $l_\ell \times l_{\ell+1}$}

    $[\mathbf{u}, s, \mathbf{v}] = \text{svds}(C_{[p]}^{(\ell)}, 1)$

    if $u_1 < 0$ **then**

      $\mathbf{u} = -\mathbf{u}, \mathbf{v} = -\mathbf{v}$ {Assume the generic case that $\mathbf{u}_1 \neq 0$; otherwise, use another entry.}

    end if

    $\mathbf{u}_{[p+1]}^{(\ell)} := \mathbf{u}$

    $\widehat{\mathbf{u}}_{[p+1]}^{(\ell+1)} := \mathbf{v}$ {Skipping this step will not affect $C_{[p]}^{(\ell+1)}$ at Line 4.}

    $\lambda_{[p+1]}^{(\ell)} := s$

  end for

  $\boldsymbol{\beta}_k = (1, k)$ {Not $(k, 1)$!}

  $C_{[p]}^{(k)} = T \circledast_{\boldsymbol{\beta}_k} \bigotimes_{i=2}^{k-1} \mathbf{u}_{[p+1]}^{(i)}$ {A matrix of size $l_1 \times l_k$}

  $[\mathbf{u}, s, \mathbf{v}] = \text{svds}(C_{[p]}^{(k)}, 1)$

  $\mathbf{u}_{[p+1]}^{(k)} := \mathbf{v}$ {After adjusting the signs of $\mathbf{u}$ and $\mathbf{v}$ properly as in Line 6.}

  $\mathbf{u}_{[p+1]}^{(1)} := \mathbf{u}$

  $\lambda_{[p+1]}^{(k)} := s$

end for

Randomization for Non-symmetric Case (A5)

$t \leftarrow 0$
$\lambda_0 \leftarrow \langle T, \bigotimes_{\ell=1}^{k} \mathbf{u}^{(\ell)} \rangle$
repeat
  $t \leftarrow t + 1$
  $\sigma \leftarrow$ random permutation of $\{1, \ldots, k\}$
  $\boldsymbol{\beta}_t \leftarrow (\sigma_{k-1}, \sigma_k)$
  $C_t \leftarrow T \circledast_{\boldsymbol{\beta}_t} \bigotimes_{i=1}^{k-2} \mathbf{u}^{(\sigma_i)}$
  $[\mathbf{u}_t, s_t, \mathbf{v}_t] = \text{svds}(C_t, 1)$ {Dominant singular value triplet via Matlab routine svds, assume uniqueness}
  if $(\mathbf{u}_t)_1 < 0$ **then**
    $\mathbf{u} = -\mathbf{u}_t, \mathbf{v} = -\mathbf{v}_t$ {Assume the general case that $(\mathbf{u}_t)_1 \neq 0$; otherwise, use another entry}
  end if
  $\lambda_t \leftarrow s_t$
  $\mathbf{u}^{(\sigma_{k-1})} \leftarrow \mathbf{u}_t, \mathbf{u}^{(\sigma_k)} \leftarrow \mathbf{v}_t$
until λ_t meets convergence criteria

Convergence of Objective Values

Because the SVD is involved, the generalized Rayleigh quotients are bounded and monotone increasing.

Convergence of Iterates

Theorem

For almost all order-k tensors T and arbitrary starting points, the vector sequence $\{(\mathbf{u}_t^{(1)}, \ldots, \mathbf{u}_t^{(k)})\}$ *generated by Algorithm SVD randomization converges to a local maximizer of the generalized Rayleigh quotient.*

Real Analysis

Lemma

(Moré etc., 1983) Assume that a^* is an isolated accumulation point of a sequence $\{a_t\}$ such that for every subsequence $\{a_{t_j}\}$ converging to a^*, there is an infinite subsequence $\{a_{t_{j_i}}\}$ such that $|a_{t_{j_i}+1} - a_{t_{j_i}}| \to 0$. Then the whole sequence $\{a_t\}$ converges to a^*.

Proof

- There is a subsequence $\{\mathbf{u}_{t_j}^{(\ell)}\}$ converges to the same limit point for all $\ell = 1, \ldots, k$ (symmetric case).
- For almost all tensors T, the accumulation points are geometrically isolated.
- $\|\mathbf{u}_{t_j+1}^{(\ell)} - \mathbf{u}_{t_j}^{(\ell)}\| \to 0$.

Numerical Example

All the experiments in this thesis are performed on a MacBook with 2.3 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory running MATLAB with version R2015a (8.5.0.19613).

Numerical Example for Symmetric Tensor

- Compare CPU time needed by our A1, A2, A3, symmetric SVD, conventional ALS and symmetric ALS.
- Order-3 and order-4 tensors with dimension 2^p.
- Execute each algorithm by 20 runs with random initial unit vectors.
- Stopping criteria is the iteration terminates when three consecutive generalized Rayleigh quotients do not vary more than the tolerance 10^{-8}.

CPU Time For Symmetric Case

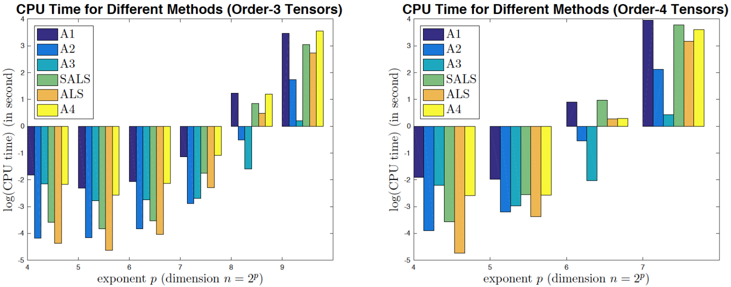


FIGURE 6.2. *Breakdown of CPU time for comparison among different methods.*

Observations

- They may converge to different limit points.
- A3 is fastest especially for large p.
- ALS and A2 perform better when p is small.
- Compared to randomise methods A2 And A3, A1 is less effective for both small and large p.

Numerical Example for Non-symmetric Tensor

- Compare CPU time required by A4, A5, ASVD, MASVD, block SVD (BSVD).
- Order-3 and order-4 tensors with dimension 2^p.
- Execute each algorithm by 20 runs with random initial unit vectors.
- Stopping criteria is the iteration terminates when three consecutive generalized Rayleigh quotients do not vary more than the tolerance 10^{-5}.

CPU Time For Non-symmetric Case

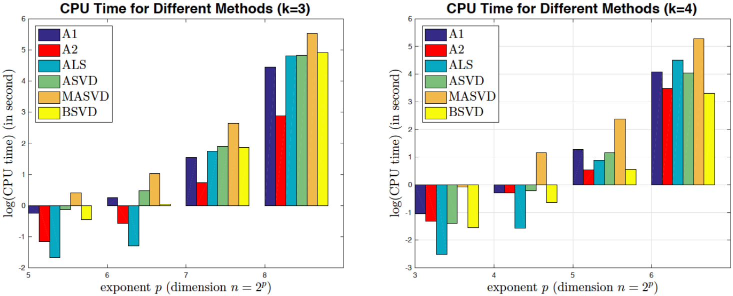


Figure 1: Comparison of CPU time among different methods.

Observations

- For problems of modest sizes, the cost of SVD computation outruns that of the high-order power method.
- For odd order tensors, the BSVD slows down.
- For order-4 tensors, A5 and the BSVD method are about equally fast.
- A4 should always be less effective than A5.
- The MASVD requires multiple ASVD calculation, so it is more expensive than ASVD.
- The ASVD checks through all possible permutations, so its performance is about the same as that of the A4.

Tensor Decompositions

- Tucker Decomposition

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

- CANDECOMP/PARAFAC (CP) Decomposition

$$T = \sum_{j} \lambda_j \mathbf{u}_j^{(1)} \otimes \ldots \otimes \mathbf{u}_j^{(k)}.$$

Applications

Tensor decomposition has been applied in a wide range of areas:

- signal processing, numerical linear algebra, computer vision, numerical analysis, data mining and analysis,
- graph analysis, neuroscience, image processing, component analysis, network analysis, scientific computing,
- telecommunications, independent component analysis (ICA) , Newton potential, stochastic PDEs.

Challenges and ill-posedness

- Best low rank approximation of a matrix ($k = 2$) always exists. (Eckart-Young Theorem)
- The rank-1 approximation is theoretically guaranteed to have a global optimum.
- Best rank-R ($R > 1$) approximation for high-order tensors may not exist .

Example

Let $\mathbf{u}_1, \mathbf{v}_1 \in \mathbb{R}^{I_1}$, $\mathbf{u}_2, \mathbf{v}_2 \in \mathbb{R}^{I_2}$, and $\mathbf{u}_3, \mathbf{v}_3 \in \mathbb{R}^{I_3}$ be vectors such that each pair $\mathbf{u}_i, \mathbf{v}_i$ is linearly independent. Define tensor

$$T := \mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \mathbf{v}_3 + \mathbf{u}_1 \otimes \mathbf{v}_2 \otimes \mathbf{u}_3 + \mathbf{v}_1 \otimes \mathbf{u}_2 \otimes \mathbf{u}_3 \in \mathbb{R}^{I_1 \times I_2 \times I_3},$$

and for each $n \in \mathbb{N}$,

$$T_n := n \left(\mathbf{u}_1 + \frac{1}{n} \mathbf{v}_1 \right) \otimes \left(\mathbf{u}_2 + \frac{1}{n} \mathbf{v}_2 \right) \otimes \left(\mathbf{u}_3 + \frac{1}{n} \mathbf{v}_3 \right) - n \mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \mathbf{u}_3.$$

Then T has rank 3 and rank of T_n is at most 2. But $\| T_n - T \| \to 0$ as $n \to \infty$. Therefore, T does not have a best rank-2 approximation.

Solution

- Orthogonality requirement ensures the existence.
  1. Complete orthogonality:
     For all $i = 1, \ldots, k$, and $1 \leq r_1 \neq r_2 \leq R$, $\langle \mathbf{u}_{r_1}^{(i)}, \mathbf{u}_{r_1}^{(i)} \rangle = 1$, and $\langle \mathbf{u}_{r_1}^{(i)}, \mathbf{u}_{r_2}^{(i)} \rangle = 0$.
  2. Semi-orthogonality:
     For all $i = 1, \ldots, k$, and $1 \leq r_1 \leq R$, $\langle \mathbf{u}_{r}^{(i)}, \mathbf{u}_{r}^{(i)} \rangle = 1$ and there is one i such that
     $$\langle \mathbf{u}_{r_1}^{(i)}, \mathbf{u}_{r_2}^{(i)} \rangle = 0, \quad \forall 1 \leq r_1 \neq r_2 \leq R.$$
  3. Orthogonality:
     For all $i = 1, \ldots, k$, and $1 \leq r \leq R$, $\langle \mathbf{u}_{r}^{(i)}, \mathbf{u}_{r}^{(i)} \rangle = 1$, and for some $1 \leq i_1 < \ldots < i_\mu \leq k$,
     $$\left\langle \mathbf{u}_{r_1}^{(i_1)}, \mathbf{u}_{r_2}^{(i_1)} \right\rangle = 0, \ldots, \left\langle \mathbf{u}_{r_1}^{(i_\mu)}, \mathbf{u}_{r_2}^{(i_\mu)} \right\rangle = 0, \quad \forall 1 \leq r_1 \neq r_2 \leq R.$$

Orthogonal Low Rank Approximation

- Given $T \in \mathbb{R}^{I_1 \times \dots \times I_k}$, determine
  - unit vectors $\mathbf{u}_r^{(i)} \in \mathbb{R}^{I_i}$, $i = 1, \dots k$,
  - scalars $\lambda_r \in \mathbb{R}$,

  such that

$$\left\| T - \sum_{r=1}^{R} \lambda_r \underbrace{\bigotimes_{i=1}^{k} \mathbf{u}_r^{(i)}}_{H_r} \right\|_F^2,$$

  is minimized subject to the mutual orthogonality condition that

$$\langle H_{r_1}, H_{r_2} \rangle = \prod_{i=1}^{k} \left\langle \mathbf{u}_{r_1}^{(i)}, \mathbf{u}_{r_2}^{(i)} \right\rangle = \delta_{r_1 r_2}, \quad \text{for all} \quad 1 \leq r_1, r_2 \leq R,$$

Open Question

- Complete orthogonal low rank approximation are studied in (Chen etc., 2008).
- Semi-orthogonal low rank approximation of tensors are studied in (Wang etc., 2015).
- It is interesting to impose orthogonality to more than one factor matrix.
    - (Wang etc., 2015) pointed that "More study is needed".
    - (Wang etc., 2015) addressed that "The question of more than one semi-orthogonal factor matrix, except for the case of complete orthogonality, remains open".

Our Problem

- Orthogonal low rank approximation:

$$\begin{cases} \min \left\| T - \sum_{r=1}^{R} \lambda_r \bigotimes_{i=1}^{k} \mathbf{u}_r^{(i)} \right\|_F^2, \\ \text{subject to } \textcolor{red}{\text{orthogonality constraint}}. \end{cases} \tag{1}$$

- Orthogonality constraint:

$$\left\langle \mathbf{u}_r^{(i)}, \mathbf{u}_r^{(i)} \right\rangle = 1, \text{For all } i = 1, \ldots, k, \text{ and } 1 \leq r \leq R$$

$$\left\langle \mathbf{u}_{r_1}^{(k-\mu+1)}, \mathbf{u}_{r_2}^{(k-\mu+1)} \right\rangle = 0, \ldots, \left\langle \mathbf{u}_{r_1}^{(k)}, \mathbf{u}_{r_2}^{(k)} \right\rangle = 0,$$

$$\forall 1 \leq r_1 \neq r_2 \leq R.$$

An Equivalent Formulation

- The optimal scales λ_r can also be interpreted as the length of the projection of the "vector" T onto the "unit vector" H_r under the Frobenius inner product,

$$\lambda_r = \left\langle T, \bigotimes_{i=1}^{k} \mathbf{u}_r^{(i)} \right\rangle = \left\langle T \circledast_\ell \left(\bigotimes_{i=1}^{\ell-1} \mathbf{u}_r^{(i)} \otimes \bigotimes_{i=\ell+1}^{k} \mathbf{u}_r^{(i)} \right), \mathbf{u}_r^{(\ell)} \right\rangle .$$

- The orthogonal low rank approximation problem (1) can be reformulated as

$$\begin{cases} \max \sum_{r=1}^{R} \lambda_r^2, \\ \text{subject to the orthogonality constraint.} \end{cases} \tag{2}$$

Existing Algorithms

- For matrices ($k = 2$), the best low rank approximation is TSVD (Eckart-Young theorem).
- For general tensors ($k > 2$), the "workhorse" algorithm for orthogonal low rank approximation of tensor has been alternating least squares (ALS) method.
  - (Wang etc., 2015) proved convergence globally.
  - Numerical computation of the completely orthogonal in (Chen etc., 2008).

Contributions

- We develop an SVD-based algorithm which updates two factors simultaneously.
- To address the orthogonality, we apply polar decomposition for μ factors.
- The convergence of our SVD-based algorithm is analyzed for both objective function and iterates themselves.

Algorithm Description

- The update of first $k-\mu$ factors by SVD.
  - If $k-\mu$ is even, update $\mathbf{u}_r^{(\ell)}$ and $\mathbf{u}_r^{(\ell+1)}$ simultaneously by SVDs for $\ell = 1, 3, \ldots, k-\mu-1$.
  - If $k-\mu$ is odd, update $\mathbf{u}_r^{(k-\mu-1)}$ twice.
- To address the orthogonality constraint, update $\mathbf{u}_r^{(\ell)}$ for $k-\mu+1 \leq \ell \leq k$ by polar decomposition.

Algorithm 6

Require: Starting unit vectors $\mathbf{u}_{r.[0]}^{(\ell)} \in \mathbb{R}^{I_\ell}$ and $\mathbf{u}_{i,[0]}^{(\ell)} \perp \mathbf{u}_{j,[0]}^{(\ell)}$ for $\ell = k - \mu + 1, \ldots, k$

$T = \frac{1}{\|T\|_F} T$ {Normalize T}
$\tau := k - \mu - 1$
if $k - \mu$ is odd **then**
  $\tau := k - \mu - 2$
end if
for $p = 0, 1, \ldots,$ **do**
  for $\ell = 1, 3, \ldots, \tau$ **do**
   $\beta_\ell = (\ell, \ell + 1)$ **do**
   for $r = 1, 2, \ldots, R,$
    $C_{r,[p+1]}^{(\ell)} = T \circledast_{\beta_\ell} \left(\bigotimes_{i=1}^{\ell-1} \mathbf{u}_{r,[p+1]}^{(i)} \otimes \bigotimes_{i=\ell+2}^{k} \mathbf{u}_{r,[p]}^{(i)} \right)$ {A matrix of size $I_\ell \times I_{\ell+1}$}
    $[\mathbf{u}, s, \mathbf{v}] = \mathrm{svds}(C_{r,[p+1]}^{(\ell)}, 1)$ {Dominant singular value triplet via Matlab routine svds;assume uniqueness}
    if $\mathbf{u}_1 < 0$ **then**
     $\mathbf{u} = -\mathbf{u}, \mathbf{v} = -\mathbf{v}$
    end if

$\mathbf{u}_{r,[p+1]}^{(\ell)} := \mathbf{u}$

$\mathbf{u}_{r,[p+1]}^{(\ell+1)} := \mathbf{v}$ {if $k-\mu$ is even, use $\hat{\mathbf{u}}_{r,[p+1]}^{(k-\mu-2)} := \mathbf{v}$}

$\lambda_{r,[p+1]}^{(\ell)} := s, \quad \lambda_{r,[p+1]}^{(\ell+1)} := s$ {if $k-\mu$ is odd, use $\hat{\lambda}_{r,[p+1]}^{(k-\mu-2)} := s$}

end for

end for

if $\tau = k-\mu-2$ **then**

$\beta_{k-\mu-1} = (k-\mu-1,\ k-\mu)$

for $r = 1, 2, \ldots, R$, **do**

$C_{r,[p+1]}^{(k-\mu-1)} = T\circledast_{\beta_{k-\mu-1}} \left(\bigotimes_{i=1}^{k-\mu-2} \mathbf{u}_{r,[p+1]}^{(i)} \otimes \bigotimes_{i=k-\mu+1}^{k} \mathbf{u}_{r,[p]}^{(i)}\right)$ {A matrix of size $I_{k-\mu-1} \times I_{k-\mu}$}

$[\mathbf{u}, s, \mathbf{v}] = \mathrm{svds}(C_{r,[p+1]}^{(k-\mu-1)}, 1)$ {Dominant singular value triplet via Matlab routine svds;assume uniqueness}

if $\mathbf{u}_1 < 0$ **then**

$\mathbf{u} = -\mathbf{u}, \mathbf{v} = -\mathbf{v}$

end if

$\mathbf{u}_{r,[p+1]}^{(k-\mu-1)} := \mathbf{u}, \quad \mathbf{u}_{r,[p+1]}^{(k-\mu)} := \mathbf{v}$
$\lambda_{r,[p+1]}^{(k-\mu-1)} := s, \; \lambda_{r,[p+1]}^{(k-\mu)} := s$
end for
end if
for $\ell = k-\mu+1, \ldots, k$ **do**
for $r = 1, 2, \ldots, R$, **do**
$\mathbf{v}_{r,[p+1]}^{(\ell)} = T \circledast_\ell \left(\bigotimes_{i=1}^{\ell-1} \mathbf{u}_{r,[p+1]}^{(i)} \otimes \bigotimes_{i=\ell+1}^{k} \mathbf{u}_{r,[p]}^{(i)} \right)$ {define columns of $V_{[p+1]}^{(\ell)}$}
$\hat{\lambda}_{r,[p+1]}^{(\ell)} := \langle \mathbf{v}_{r,[p+1]}^{(\ell)}, \mathbf{u}_{r,[p]}^{(\ell)} \rangle$ {define diagonals of $\Lambda_{[p+1]}^{(\ell)}$}
end for
$[U_{[p+1]}^{(\ell)}, S_{[p+1]}^{(\ell)}] = \text{poldec}(V_{[p+1]}^{(\ell)} \Lambda_{[p+1]}^{(\ell)})$
for $r = 1, 2, \ldots, R$, **do**
$\mathbf{u}_{r,[p+1]}^{(\ell)} := U_{[p+1]}^{(\ell)}(:, r)$
$\lambda_{r,[p+1]}^{(\ell)} := S_{[p+1]}^{(\ell)}(r, r) (= \langle \mathbf{v}_{r,[p+1]}^{(\ell)}, \mathbf{u}_{r,[p+1]}^{(\ell)} \rangle)$
end for
end for

end for

Trace Maximizing Property

Lemma

Let matrix $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ have polar decomposition

$$A = QS,$$

where $Q \in \mathbb{R}^{m \times n}$ is the column orthogonal polar factor and $S \in \mathbb{R}^{n \times n}$ is the symmetric positive semi-definite factor. Then

$$Q = \arg \max_{P \in \mathbb{R}^{m \times n},\ P^T P = I} \mathrm{Trace}\left(P^T A\right).$$

Moreover, if A is of full column rank, then Q above is unique.

Convergence of Objective Values

- As SVD is involved for the first $k-\mu$ factors, the generalized Rayleigh quotients are bounded and monotone increasing,

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

- Polar decomposition is applied for last μ factors, by trace maximizing property,

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

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

Theorem

For almost all tensors T, the sequence $\left\{\mathbf{u}_{r,[p]}^{(\ell)}\right\}$ generated in Algorithm 6 converges for $\ell = 1, \ldots, k$, $r = 1, \ldots, R$.

- Accumulation points are isolated.
- If subsequences $\left\{\mathbf{u}_{r,[p_j]}^{(\ell)}\right\}$ generated by Algorithm 6 converge simultaneously, then subsequences $\left\{\mathbf{u}_{r,[p_j+1]}^{(\ell)}\right\}$ also converge simultaneously.
- $\left\{\mathbf{u}_{r,[p_j]}^{(\ell)}\right\}$ and $\left\{\mathbf{u}_{r,[p_j+1]}^{(\ell)}\right\}$ converge to the same limiting point.

Numerical Example

Test Algorithm 6

- $\mu = 2$ and $R = 5$;
- First 150 steps.

Comparison: by measuring

- Objective value $\sum_{r=1}^{R} \lambda_r^2$;
- Iterate error $\sum_{\ell=1}^{k} \sum_{r=1}^{R} \|\mathbf{u}_{r,[p+1]}^{(\ell)} - \mathbf{u}_{r,[p]}^{(\ell)}\|_2^2$.

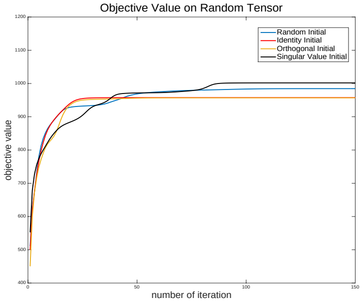
Test tensors $R^{20\times16\times10\times32}$:

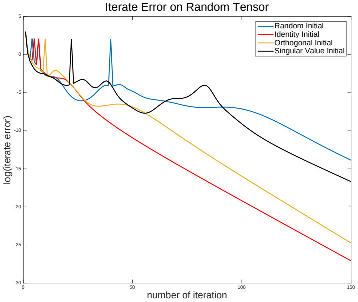
- Random tensor: randomly generate.
- Stochastic tensor:
$$\tau_{i_1,i_2,i_3,i_4} = \begin{cases} c & i_1 \neq i_2, i_2 \neq i_3, i_3 \neq i_4 \\ 0 & i_1 = i_2, i_2 \neq i_3, i_3 \neq i_4 \\ 1/20 & \text{otherwise} \end{cases}$$
, where c is randomly in $(0, 1)$ by the homogenous distribution such as $\sum_{i_1\in[\![20]\!]} \tau_{i_1,i_2,i_3,i_4} = 1$ with $i_j \neq i_{j+1}, j = 1, 2, 3$.
- Cauchy tensor: $\tau_{i_1,i_2,i_3,i_4} = \frac{1}{c(i_1)+c(i_2)+c(i_3)+c(i_4)}$, where c is a random vector with size 32.
- Hilbert tensor: $\tau_{i_1,i_2,i_3,i_4} = \frac{1}{i_1+i_2+i_3+i_4-3}$.
- Toeplitz tensor: $\tau_{i_1+j,i_2+j,i_3+j,i_4+j} = \tau_{i_1,i_2,i_3,i_4}$ for $j \in [\![min(20-i_1, 16-i_2, 10-i_3, 32-i_4)]\!]$.

Initial vectors:

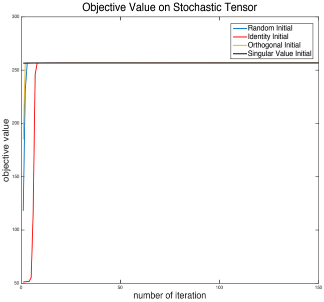
- 'Random Initial'–unit vectors $\mathbf{u}_r^{(\ell)}$ for $\ell = 1, \ldots, k$ and $r = 1, \ldots, R$ are generated randomly to satisfy orthogonality constrain with $\mu = 2$.
- 'Identity Initial'–each $[\mathbf{u}_1^{(\ell)}, \ldots, \mathbf{u}_R^{(\ell)}]$ for $\ell = 1, \ldots, k$ are taken as the first R columns of identity matrices.
- 'Orthogonal Initial'–each $[\mathbf{u}_1^{(\ell)}, \ldots, \mathbf{u}_R^{(\ell)}]$ for $\ell = 1, \ldots, k$ are taken as the first R columns of random orthonormal matrices.
- 'Singular Value Initial'–the major left singular vectors of the unfoldings of the tensors are used as initials.

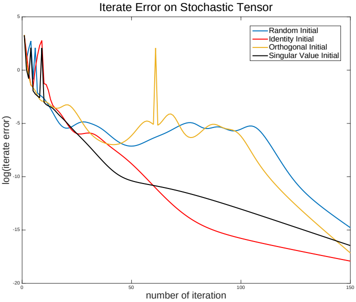
Comparison on Random Tensor



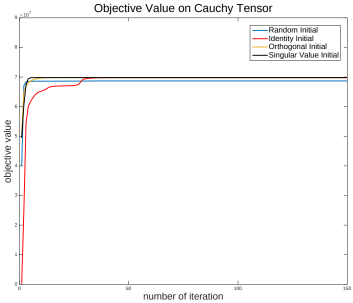


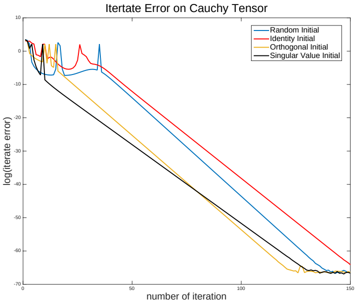
Comparison on Stochastic Tensor



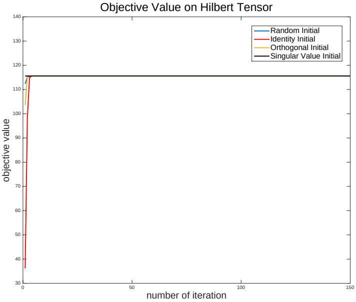


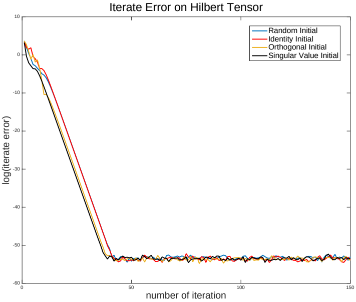
Comparison on Cauchy Tensor



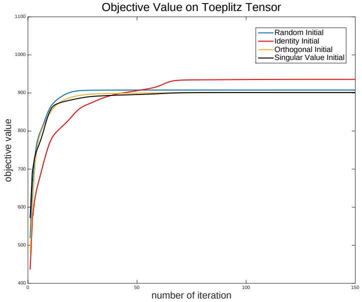


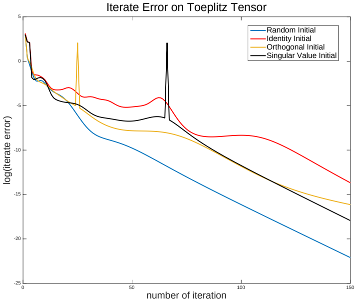
Comparison on Hilbert Tensor





Comparison on Toeplitz Tensor





Observations

Objective value:

- Objective value satisfies the monotone increasing property for each iteration;
- For different initial vectors, the approximated objective values may be different for the same test tensor, that is, iterates may converge to different limit points.
  - It is interesting to study for what tensors or what initial guesses Algorithm 6 converges to the global optimum (Chen etc., 2008).

Observations

Irerates error:

- Iterates converge, but they are not monotone in each step.
- Iterates converge but slower than that of objective values.
- When it comes to the qualities of the final approximation, among four different initial vectors, no any one does offer obvious advantage.

Definition of ADM

Alternating Direction Methods

Fixing all but one variable a time and alternating among the variables.

General Form

Many algorithms can be cast in the abstract form

$$\begin{cases} \mathbf{x}_{k+1} = f(\mathbf{y}_k), \\ \mathbf{y}_{k+1} = g(\mathbf{x}_{k+1}), \end{cases} \quad k = 0, 1, \ldots,$$

where $f : U \to V$ and $g : V \to U$.

Background

$$\mathbf{y}_{k+1} = g(f(\mathbf{y}_k)), \quad k = 0, 1, \ldots. \tag{3}$$

- If $g \circ f$ is a contraction map, then the Banach fixed-point theorem asserts that the iterates from (3) converge to a unique fixed-point point.
- If $g \circ f$ is continuous and maps a convex compact set into itself, then the Brouwer fixed-point theorem asserts that there is a fixed-point $\mathbf{y}_*$ such that $g \circ f(\mathbf{y}_*) = \mathbf{y}_*$.

General Form

For more complicated problems involving n variables $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, a similar alternating iteration can be written in this form

$$\left\{\begin{array}{lcl} \mathbf{x}_{k+1}^{(1)} & = & f^{(1)}(\mathbf{x}_k^{(2)}, \mathbf{x}_k^{(3)}, \ldots, \mathbf{x}_k^{(n)}), \\ \mathbf{x}_{k+1}^{(2)} & = & f^{(2)}(\mathbf{x}_{k+1}^{(1)}, \mathbf{x}_k^{(3)}, \ldots, \mathbf{x}_k^{(n)}), \\ & \vdots & \\ \mathbf{x}_{k+1}^{(n)} & = & f^{(n)}(\mathbf{x}_{k+1}^{(1)}, \mathbf{x}_{k+1}^{(2)}, \ldots, \mathbf{x}_{k+1}^{(n-1)}). \end{array}\right. \qquad k = 0, 1, \ldots.$$

Our Work

- We propose a general framework that can be applied to many types of alternating direction methods for proving convergence.
- The conditions entailed by this framework are mild and easy to satisfy, so the theory should be of fundamental significance to many algorithms.

Lemma

Let $F : U \longrightarrow U$ be a continuous map over a closed subset $U \subset \mathbb{R}^n$. Suppose that the sequence $\{\mathbf{z}_k\}$ generated by iterative scheme $\mathbf{z}_{k+1} = F(\mathbf{z}_k)$ is well defined, bounded, and has finitely many isolated accumulation points. Then

1. *Either the sequence $\{\mathbf{z}_k\}$ converges, or*
2. *There are disjoint neighborhoods of the accumulation points such that, for k large enough, the consecutive elements $\mathbf{z}_k, \mathbf{z}_{k+1}, \ldots$ visit each neighborhood in a cyclic order.*

Main Theory

Theorem

Suppose that an alternating optimization method can be cast in the general form. Write $\mathbf{z} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ *where* $\mathbf{x}^{(\ell)} \in U^{(\ell)}$ *and* $U^{(\ell)} \subset \mathbb{R}^{I_\ell}$ *. Assume that*

- *a) The conditions in previous lemma are satisfied where* $F(\mathbf{z})$ *denotes the transition function of one complete sweep of the alternating optimization,* $\mathbf{z}_{k+1} = F(\mathbf{z}_k)$*.*

Theorem

- *b) Each* $f^{(\ell)}$ *representing the optimization mechanism in the* ℓ*-th direction is continuously differentiable and returns the unique global minimizer* $\mathbf{x}_{k+1}^{(\ell)}$ *of the restricted objective function*

$$h_\ell(\mathbf{w}) := h(\mathbf{x}_{k+1}^{(1)}, \ldots, \mathbf{x}_{k+1}^{(\ell-1)}, \mathbf{w}, \mathbf{x}_{k}^{(\ell+1)}, \ldots, \mathbf{x}_{k}^{(n)}).$$

- *c) The objective function* $h(\mathbf{z})$ *is second order continuously differentiable.*

Theorem

- *d) One of the accumulation points* $\mathbf{z}_0^*$ *of* $\{\mathbf{z}_k\}$ *is a local minimizer of* $h(\mathbf{z})$ *at which the Hessian* $\nabla^2 h(\mathbf{z}_0^*)$ *is symmetric and positive definite.*

Then the sequence $\{\mathbf{z}_k\}$ *converge.*

Applications to Some Known Cases

- The Gauss-Seidel method for solving a system of linear equations.
- The power method for finding the dominant eigenvector.
- The alternating least squares method for computing the QR decomposition.
- The alternating projection method for finding structured low rank matrices.
- Best rank-one tensor approximation.
- Tucker nearest problem.
- Structured Kronecker approximation.

Future Topics

- High order SVD;
- Quantum entanglement;
- Orthogonal symmetric tensor diagonalization;
- Segment CP approximation;
- Segment Tucker approximation.

References

[1] J. Chen and Y. Saad, On the tensor SVD and the optimal low rank orthogonal approximation of tensors, SIAM J. Matrix Anal. Appl., 30 (2008/09), pp. 1709–1734.
[2] P. Comon, X. Luciani, and A. L. De Almeida, Tensor decompositions, alternating least squares and other tales, J. Chemometrics, 23 (2009), pp. 393–405.
[3] L. De Lathauwer, B. De Moor, and J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253–1278.
[4] S. Friedland, Best rank one approximation of real symmetric tensors can be chosen symmetric, Front. Math. China, 8 (2013), pp. 19–40.

[5] S. Friedland, V. Mehrmann, R. Pajarola, and S. K. Suter, On best rank one approximation of tensors, Numer. Linear Algebra Appl., 20 (2013), pp. 942–955.
[6] P. r. Kroonenberg and J. Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms, Psychometrika, 45 (1980), pp. 69–97.
[7] J. J. More and D. C. Sorensen, Computing a trust region step, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[8] L. Qi, The best rank-one approximation ratio of a tensor space, SIAM J. Matrix Analysis Applications, (2011), pp. 430–442.

[9] A. Uschmajew, Local convergence of the alternating least squares algorithm for canonical tensor approximation, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 639–652.
[10] L. Wang and M. T. Chu, On the global convergence of the alternating least squares method for rank-one approximation to generic tensors, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1058–1072.
[11] L. Wang, M. T. Chu, and B. Yu, Orthogonal low rank tensor approximation: alternating least squares method and its global convergence, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1–19.
[12] Y. Yang, S. Hu, L. De Lathauwer, and J. A. Suykens, Convergence study of block singular value maximization methods for rank-1 approximation to higher order tensors, tech. rep., Internal Report 16-149, ESAT-SISTA, KU Leuven, 2016.

Thank you very much!