On the convergence of higher-order orthogonality iteration and its extension

Yangyang Xu

IMA, University of Minnesota

SIAM Conference LA15, Atlanta

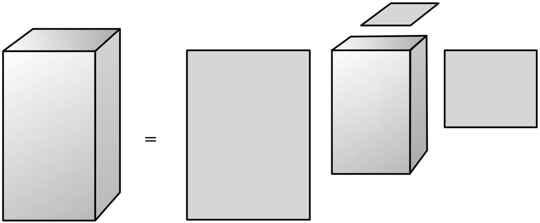
October 27, 2015

Best low-multilinear-rank approximation of tensors

$$\min_{\boldsymbol{\mathcal{C}},\mathbf{A}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_1 \mathbf{A}_1 \ldots \times_N \mathbf{A}_N\|_F^2,$$

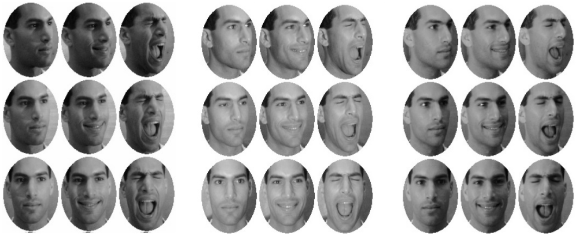
subject to $\mathbf{A}_n \in \mathcal{O}_{I_n \times r_n} \triangleq \{\mathbf{A}_n \in \mathbb{R}^{I_n \times r_n} : \mathbf{A}_n^\top \mathbf{A}_n = \mathbf{I}\}, \forall n.$

- Any tensor has a multilinear SVD [Lathauwer-Moor-Vandewalle'00]



- Truncated multilinear SVD is not the best [L-M-V'00]

Face recognition by low-rank tensor approximation



Matrix Decomposition

$$\mathbf{D} \approx \mathbf{U}\mathbf{C}$$

Tensor Decomposition

$$\boldsymbol{\mathcal{D}} \approx \boldsymbol{\mathcal{C}} \times_1 \mathbf{U}_1 \cdots \times_5 \mathbf{U}_5$$

- Tensor decomposition can improve recognition accuracy over 50% [Vasilescu-Terzopoulos'02]

Higher-order Orthogonality Iteration (HOOI)

Given $\mathbf{A}$, the optimal $\boldsymbol{\mathcal{C}}$ is

$$\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top.$$

With this $\boldsymbol{\mathcal{C}}$ plugged in, the approximation problem becomes

$$\min_{\mathbf{A}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1\mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N\mathbf{A}_N^\top\|_F^2,$$
$$\text{subject to } \mathbf{A}_n \in \mathcal{O}_{I_n \times r_n},\ \forall n,$$

which is equivalent to

$$\max_{\mathbf{A}} \|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2 = \|\mathbf{A}_n^\top \textbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i \neq n} \mathbf{A}_i^\top)\|_F^2,$$
$$\text{subject to } \mathbf{A}_n \in \mathcal{O}_{I_n \times r_n},\ \forall n.$$

Higher-order Orthogonality Iteration (HOOI)

Algorithm 1: Higher-order orthogonality iteration (HOOI)

Input: $\mathcal{X}$ and $(r_1, \ldots, r_N)$
Initialization: choose $(\mathbf{A}_1^0, \ldots, \mathbf{A}_N^0)$ with $\mathbf{A}_n^0 \in \mathcal{O}_{I_n \times r_n}, \forall n$ and set $k = 0$
while *not convergent* **do**
  for $n = 1, \ldots, N$ **do**
    Set $\mathbf{A}_n^{k+1}$ to an orthonormal basis of the dominant r_n-dimensional left singular subspace of

$$\mathbf{G}_n^k = \textbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i<n} (\mathbf{A}_i^{k+1})^\top \times_{i>n} (\mathbf{A}_i^k)^\top)$$

  increase $k \leftarrow k + 1$

- Widely used and fit to large-scale problems
- Better scalability than Newton-type methods such as Newton-Grassmann [Elden-Savas'09] and Riemannian trust region [Ishteva et. al'11]
- Convergence is still an open question

this talk addresses the open question
subspace convergence of HOOI

- Why care about convergence
  - Without convergence, HOOI can give severely different multilinear subspace by running to different numbers of iterations
- Challenges
  - Non-convexity
  - Non-uniqueness of dominant subspace
  - Non-uniqueness of orthonormal basis
- How to establish convergence
  - Greedy selection of orthonormal basis
  - Kurdyka-Łojasiewicz property

Greedy higher-order orthogonality iteration

Algorithm 2: Greedy HOOI

Input: $\boldsymbol{\mathcal{X}}$ and $(r_1, \ldots, r_N)$

Initialization: choose $(\mathbf{A}_1^0, \ldots, \mathbf{A}_N^0)$ with $\mathbf{A}_n^0 \in \mathcal{O}_{I_n \times r_n}$, $\forall n$ and set $k = 0$

while *not convergent* **do**

  for $n = 1, \ldots, N$ **do**

    Set $\mathbf{A}_n^{k+1} \in \operatorname{argmin}_{\mathbf{A}_n} \|\mathbf{A}_n - \mathbf{A}_n^k\|_F$ over all orthonormal bases of the dominant r_n-dimensional left singular subspace of

$$\mathbf{G}_n^k = \textbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i<n} (\mathbf{A}_i^{k+1})^\top \times_{i>n} (\mathbf{A}_i^k)^\top)$$

  increase $k \leftarrow k + 1$

- Iterate mapping is continuous
- Will be used to establish convergence of original HOOI

Block-nondegeneracy

$\mathbf{A}$ is a *block-nondegenerate* point if $\sigma_{r_n}(\mathbf{G}_n) > \sigma_{r_n+1}(\mathbf{G}_n)$, $\forall n$ where

$$\mathbf{G}_n = \textbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i \neq n} \mathbf{A}_i)$$

- Dominant multilinear subspace is unique
- Equivalent to negative definiteness of block Hessian on $\mathcal{O}_{I_n \times r_n}$ for a local maximizer
- Necessary condition for convergence
  - A degenerate first-order optimal point is not stationary

Convergence analysis of greedy HOOI

Let $\{\mathbf{A}^k\}_{k\geq 1}$ be the sequence from greedy HOOI. Assume there is a block-nondegenerate limit point $\bar{\mathbf{A}}$.

- $\bar{\mathbf{A}}$ is a block-wise maximizer and a critical point
- There is $\alpha > 0$ such that if $\mathbf{A}^k$ sufficiently close to $\bar{\mathbf{A}}$, then

$$\alpha\|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F^2 \leq F(\mathbf{A}^{k+1}) - F(\mathbf{A}^k).$$

  where $F(\mathbf{A}) = \|\boldsymbol{\mathcal{X}} \times_1 \mathbf{A}_1^\top \ldots \times_N \mathbf{A}_N^\top\|_F^2 - \sum_{n=1}^N \iota_{\mathcal{O}_{I_n \times r_n}}$
- If $\mathbf{A}^k$ is sufficiently close to $\bar{\mathbf{A}}$, then

$$\text{dist}(\mathbf{0}, \partial F(\mathbf{A}^k)) \leq C\|\mathbf{A}^k - \mathbf{A}^{k-1}\|_F.$$

- Further with Kurdyka-Łojasiewicz property $\Rightarrow$ $\mathbf{A}^k \to \bar{\mathbf{A}}$ as $k \to \infty$
- Local convergence to globally optimal solution

Convergence analysis of greedy HOOI

Kurdyka-Łojasiewicz Property: the following quantity is bounded near $\bar{\mathbf{A}}$ for a certain $\theta \in [0, 1)$

$$\frac{|F(\mathbf{A}) - F(\bar{\mathbf{A}})|^\theta}{\text{dist}(\mathbf{0}, \partial F(\mathbf{A}))}$$

Use $(1-\theta)(a-b) \leq a^\theta(a^{1-\theta} - b^{1-\theta})$, $a, b \geq 0$ and previous inequalities to have

$$\begin{aligned}
&\|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F^2 \\
\lesssim& F(\mathbf{A}^{k+1}) - F(\mathbf{A}^k) \\
\lesssim& (F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))^\theta \big((F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))^{1-\theta} - (F(\bar{\mathbf{A}}) - F(\mathbf{A}^{k+1}))^{1-\theta}\big) \\
\lesssim& \text{dist}(\mathbf{0}, \partial F(\mathbf{A}))\big((F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))^{1-\theta} - (F(\bar{\mathbf{A}}) - F(\mathbf{A}^{k+1}))^{1-\theta}\big) \\
\lesssim& \|\mathbf{A}^k - \mathbf{A}^{k-1}\|_F \big((F(\bar{\mathbf{A}}) - F(\mathbf{A}^k))^{1-\theta} - (F(\bar{\mathbf{A}}) - F(\mathbf{A}^{k+1}))^{1-\theta}\big).
\end{aligned}$$

Together with Young's inequality gives $\sum_k \|\mathbf{A}^{k+1} - \mathbf{A}^k\|_F < \infty$

Convergence analysis of original HOOI

Starting from $\mathbf{A}^{k_0}$ sufficiently close to a block-nondegenerate limit point $\bar{\mathbf{A}}$ of $\{\mathbf{A}^k\}_{k\geq 1}$, we have

$$\mathbf{A}_n^k(\mathbf{A}_n^k)^\top = \tilde{\mathbf{A}}_n^k(\tilde{\mathbf{A}}_n^k)^\top,\ \forall n,\ \forall k \geq k_0$$

where $\{\mathbf{A}^k\}_{k\geq k_0}$ and $\{\tilde{\mathbf{A}}^k\}_{k\geq k_0}$ are sequences from the original and greedy HOOIs respectively.

- $\bar{\mathbf{A}}$ is a block-wise maximizer and a critical point
- $\mathbf{A}_n^k(\mathbf{A}_n^k)^\top \to \bar{\mathbf{A}}_n\bar{\mathbf{A}}_n^\top$ as $k \to \infty$
- Local convergence to globally optimal multilinear subspace

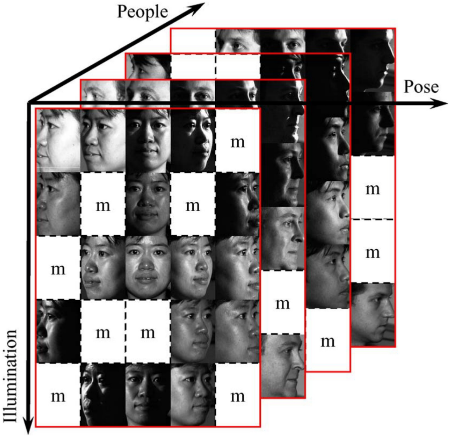
Best low-rank tensor approximation with missing values

$$\min_{\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{C}},\mathbf{A}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_1 \mathbf{A}_1 \ldots \times_N \mathbf{A}_N\|_F^2,$$

$$\text{subject to } \mathbf{A}_n \in \mathcal{O}_{I_n \times r_n}, \forall n;\ \mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}) = \mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})$$

- Reduces to the best low-rank tensor approximation if Ω contains all elements
- Can estimate the original tensor $\boldsymbol{\mathcal{M}}$ simultaneously

Best low-rank tensor approximation with missing values



Low-rank tensor approximation with missing values
[Geng-Miles-Zhou-Wang'11]

Incomplete higher-order orthogonality iteration

Algorithm 3: incomplete HOOI (iHOOI)

Input: $\mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})$ and $(r_1, \ldots, r_N)$
Initialization: choose $\boldsymbol{\mathcal{X}}^0$ and $(\mathbf{A}_1^0, \ldots, \mathbf{A}_N^0)$ and set $k = 0$
while *not convergent* **do**
  for $n = 1, \ldots, N$ **do**
    Set $\mathbf{A}_n^{k+1}$ to an orthonormal basis of the dominant r_n-dimensional left singular subspace of

$$\mathbf{G}_n^k = \textbf{unfold}_n(\boldsymbol{\mathcal{X}} \times_{i<n} (\mathbf{A}_i^{k+1})^\top \times_{i>n} (\mathbf{A}_i^k)^\top)$$

  $\boldsymbol{\mathcal{X}}^{k+1} \leftarrow \mathcal{P}_\Omega(\boldsymbol{\mathcal{M}}) + \mathcal{P}_{\Omega^c}\big(\boldsymbol{\mathcal{X}}^k \times_{n=1}^N \mathbf{A}_n^{k+1}(\mathbf{A}_n^{k+1})^\top\big)$
  increase $k \leftarrow k + 1$

- $\boldsymbol{\mathcal{X}}$-update is one step of the alternating minimization for

$$\min_{\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{X}}} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_{i=1}^N \mathbf{A}_i^{k+1}(\mathbf{A}_i^{k+1})^\top\|_F^2, \text{ subject to } \mathcal{P}_\Omega(\boldsymbol{\mathcal{X}}) = \mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})$$

Experimental results

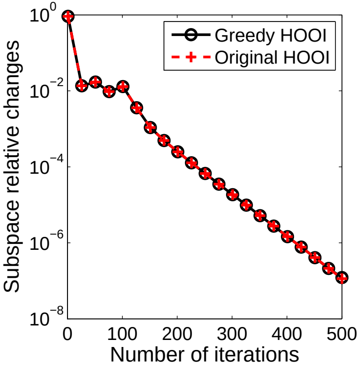
Low-multilinear-rank tensor approximation

- Relation between iterate sequences by original and greedy HOOIs

Low-rank tensor completion

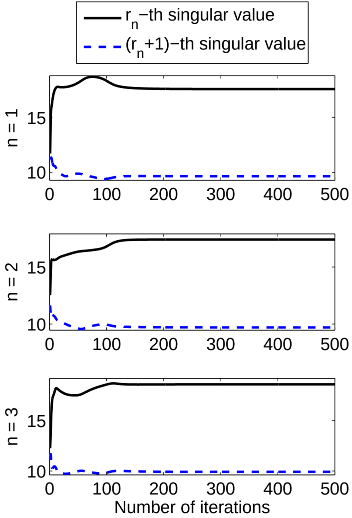
- Phase transition plots (simulate how many samples guarantee exact reconstruction)

Comparing HOOIs on subspace learning

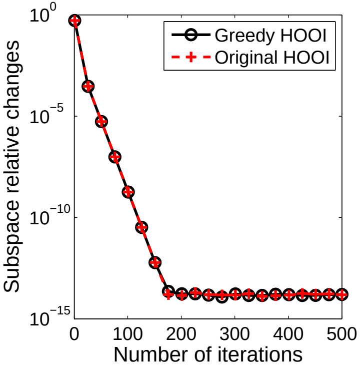


Gaussian random tensor
full size: $50 \times 50 \times 50$
core size: $5 \times 5 \times 5$

$$\mathbf{A}_n^k(\mathbf{A}_n^k)^\top = \tilde{\mathbf{A}}_n^k(\tilde{\mathbf{A}}_n^k)^\top, \forall n, k$$



Comparing HOOIs on subspace learning

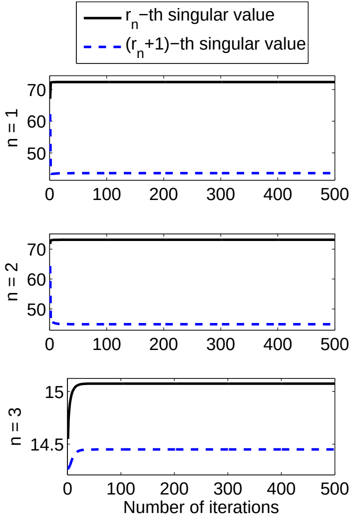


Yale Face Database

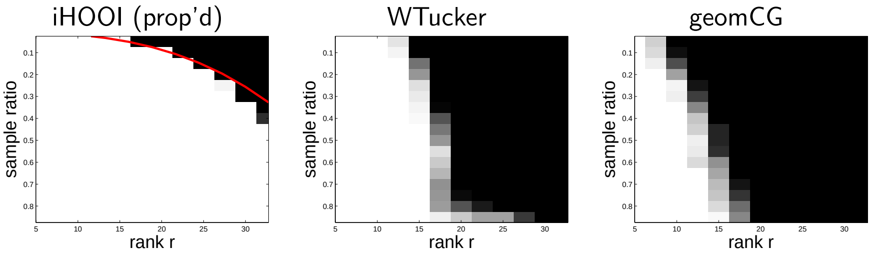
full size: $38 \times 64 \times 2958$

core size: $5 \times 5 \times 20$

$$\mathbf{A}_n^k(\mathbf{A}_n^k)^\top = \tilde{\mathbf{A}}_n^k(\tilde{\mathbf{A}}_n^k)^\top, \forall n, k$$



Phase transition plots



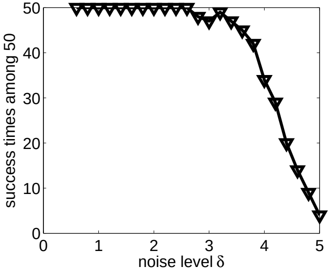
- $50 \times 50 \times 50$ randomly generated tensor with rank (r, r, r)
- True rank provided (adaptive rank estimate can be applied)
  - WTucker [Filipovic-Jukic'13]: alternating minimization to $\min_{\mathbf{A},\boldsymbol{\mathcal{C}}} \|\mathcal{P}_\Omega(\boldsymbol{\mathcal{C}} \times_{i=1}^N \mathbf{A}_i) - \mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})\|_F^2$
  - geomCG [Kressner-Steinlechner-Vandereycken'13]: Riemannian optimization method

Proposed method achieves near-optimal reconstruction

Convergence to globally optimal solution

Observation: $\frac{\|\mathcal{P}_\Omega(\boldsymbol{\mathcal{C}}\times_{n=1}^N \mathbf{A}_n)-\mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})\|_F}{\|\mathcal{P}_\Omega(\boldsymbol{\mathcal{M}})\|_F}$ always small

- Low sampling ratio $\Rightarrow$ possibly more than one feasible low-rank solutions
- $\mathbf{A}^0$ sufficiently close to bases of $\boldsymbol{\mathcal{M}} \Rightarrow \boldsymbol{\mathcal{C}}^k \times_{n=1}^N \mathbf{A}_n^k \rightarrow \boldsymbol{\mathcal{M}}$



tensor size: (50,50,50); rank: (20,20,20); sample ratio: 10%

$\mathbf{A}^0 \leftarrow$ truncated HOSVD of $\boldsymbol{\mathcal{M}} + \delta\|\boldsymbol{\mathcal{M}}\|_F \frac{\boldsymbol{\mathcal{N}}}{\|\boldsymbol{\mathcal{N}}\|_F}$

Conclusions

- Give subspace sequence convergence of the HOOI method
  - first result on convergence of HOOI
  - via the convergence of a greedy HOOI and Kurdyka-Łojasiewicz inequality
  - block-nondegeneracy is sufficient and necessary
- Propose an incomplete HOOI method for applications with missing values
  - Subspace learning and tensor completion simultaneously
  - Empirically, near-optimal recovery on low-rank tensors

References

- Y. Xu. On the convergence of higher-order orthogonality iteration. arXiv, 2015.
- Y. Xu. On higher-order singular value decomposition from incomplete data. arXiv, 2014.