Best Low Multilinear Rank Approximation: New Results

Lieven De Lathauwer

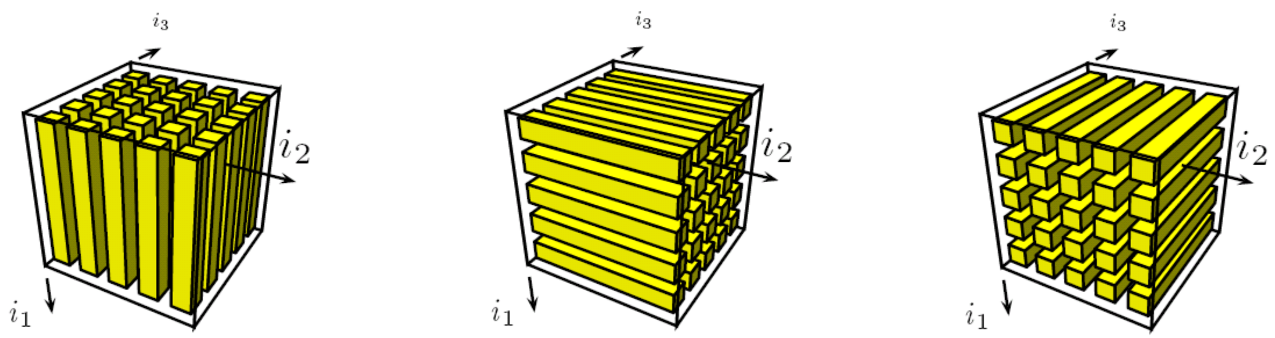
K.U.Leuven
Belgium

Overview

- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

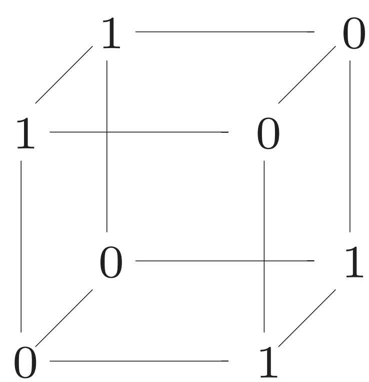
Columns, rows and mode-n vectors

Mode-n vectors of a tensor: generalization of column/row vectors of a matrix



Multilinear rank of a tensor

- The column (row) rank of a matrix $\mathbf{A}$ is equal to the maximal number of columns (rows) of $\mathbf{A}$ that form a linearly independent set
- Mode-n rank of a tensor: dimension of the vector space generated by mode-n vectors
- Mode-n ranks can be mutually different
- Rank-(R_1, R_2, R_3) tensor: $\text{rank}_1(\mathcal{A}) = R_1$, $\text{rank}_2(\mathcal{A}) = R_2$, $\text{rank}_3(\mathcal{A}) = R_3$
- Multilinear rank: (R_1, R_2, R_3)



$R_1 = R_2 = 2 \quad R_3 = 1$

Rank-1 tensor

- Rank-1 matrix: outer product of 2 vectors $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$:

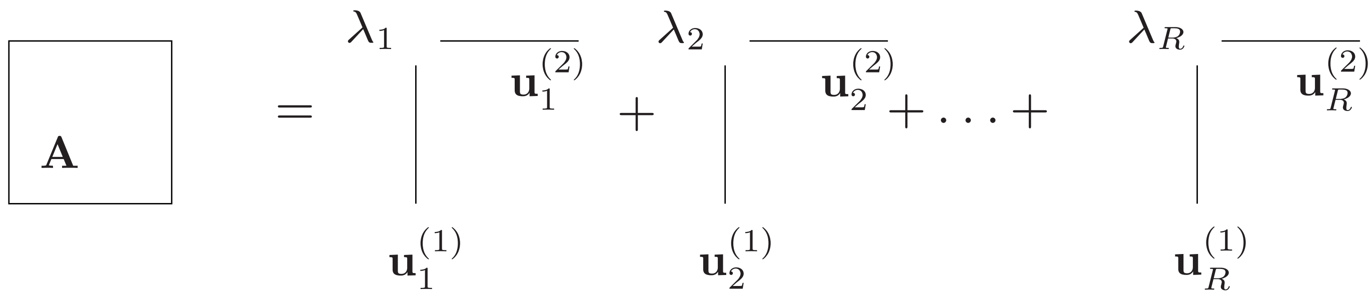
$$
\begin{aligned}
a_{i_1 i_2} &= u^{(1)}_{i_1} u^{(2)}_{i_2} \\
\mathbf{A} &= \mathbf{u}^{(1)} \cdot \mathbf{u}^{(2)^T} \equiv \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)}
\end{aligned}
$$

- Rank-1 tensor: outer product of N vectors $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$, ..., $\mathbf{u}^{(N)}$:

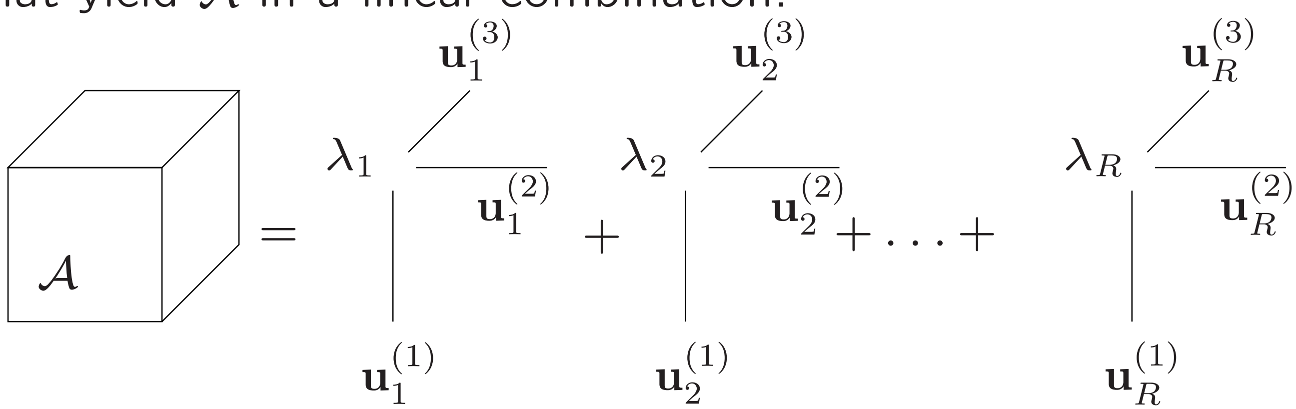
$$
\begin{aligned}
a_{i_1 i_2 \ldots i_N} &= u^{(1)}_{i_1} u^{(2)}_{i_2} \ldots u^{(N)}_{i_N} \\
\mathcal{A} &= \mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(N)}
\end{aligned}
$$

Rank of a tensor

- The rank R of a matrix $\mathbf{A}$ is minimal number of rank-1 matrices that yield $\mathbf{A}$ in a linear combination.



- The rank R of an Nth-order tensor $\mathcal{A}$ is the minimal number of rank-1 tensors that yield $\mathcal{A}$ in a linear combination.



[*Hitchcock, 1927*]

Overview

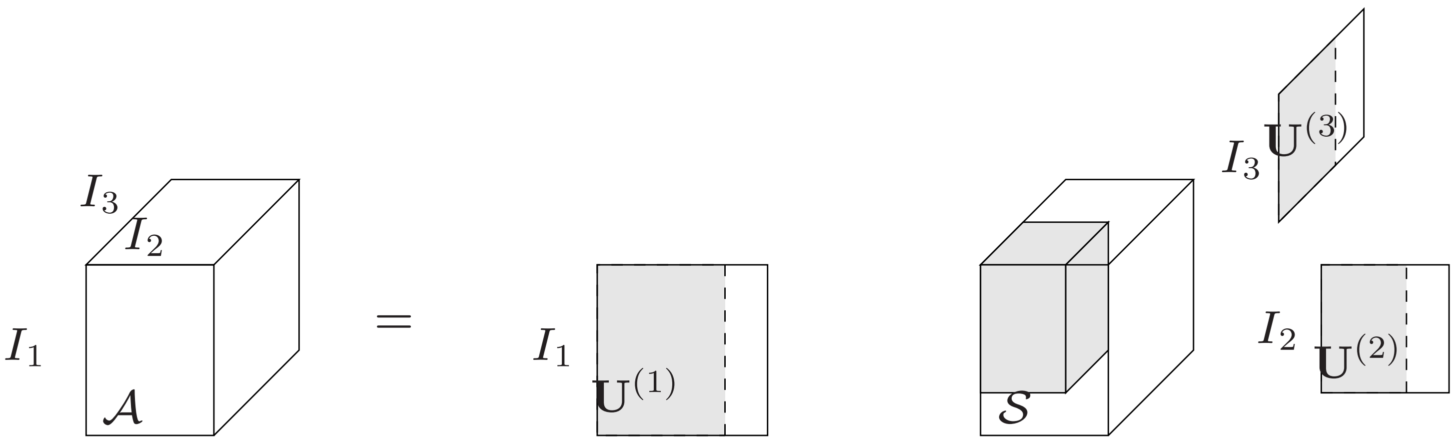
- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

Multilinear rank and associated decomposition

Definition:

$$\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \ldots \bullet_N \mathbf{U}^{(N)}$$

in which $\mathcal{S}$ is all-orthogonal and ordered
$\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}$ are orthogonal



[*Tucker '64*], [*De Lathauwer '00*]

Computation

$$\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \mathbf{U}^{(3)}$$

- $(I_1 \times I_2 I_3)$ matrix $\mathbf{A}^{(1)}$ in which all the columns are stacked

$$\text{SVD: } \mathbf{A}^{(1)} = \mathbf{U}^{(1)} \cdot \mathbf{\Sigma}^{(1)} \cdot \mathbf{U}^{(1)^T}$$

- $(I_2 \times I_3 I_1)$ matrix $\mathbf{A}^{(2)}$ in which all the row vectors are stacked

$$\text{SVD: } \mathbf{A}^{(2)} = \mathbf{U}^{(2)} \cdot \mathbf{\Sigma}^{(2)} \cdot \mathbf{U}^{(2)^T}$$

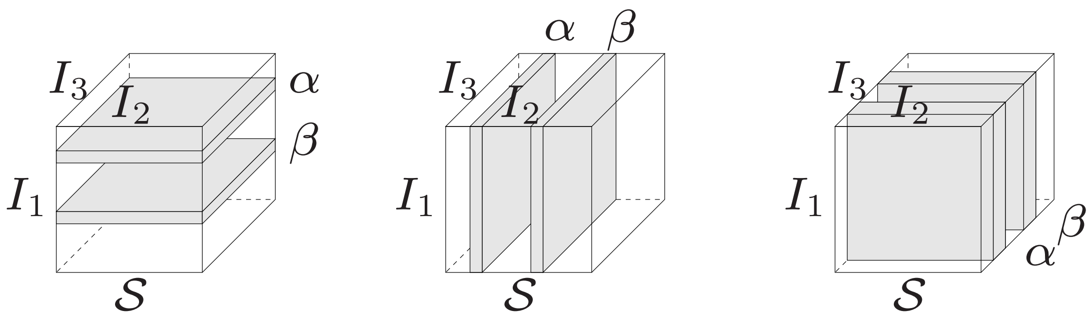
- $(I_3 \times I_1 I_2)$ matrix $\mathbf{A}^{(3)}$ in which all the mode-3 vectors are stacked

$$\text{SVD: } \mathbf{A}^{(3)} = \mathbf{U}^{(3)} \cdot \mathbf{\Sigma}^{(3)} \cdot \mathbf{U}^{(3)^T}$$

- Compute $\mathcal{S}$:

$$\mathcal{S} = \mathcal{A} \bullet_1 \mathbf{U}^{(1)^T} \bullet_2 \mathbf{U}^{(2)^T} \bullet_3 \mathbf{U}^{(3)^T}$$

All-orthogonality:

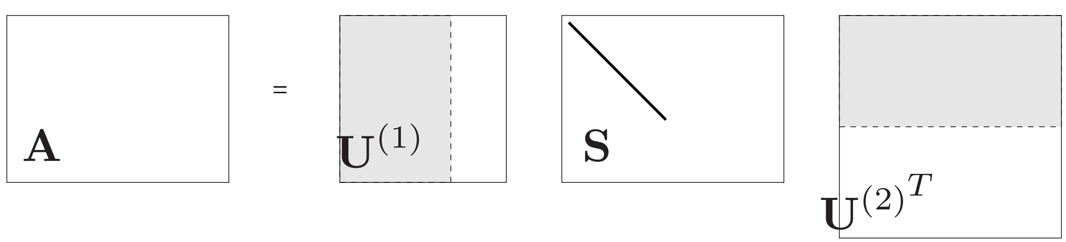


All-orthogonality is a generalization of diagonality

Ordering: slices have decreasing Frobenius norm

Norms of slices = mode-n singular values

Matrix SVD:



Definition:

$$\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \ldots \bullet_N \mathbf{U}^{(N)}$$

Properties:

- Mode-n singular values = norms of slices = singular values of $\mathbf{A}_{(n)}$
- N sets of singular values
- Mode-n rank revealing
- Number of significant $\sigma_i^{(n)}$ = numerical mode-n rank
- Similar uniqueness properties as matrix SVD
- Decomposition of 2nd order tensor = matrix SVD
- Link with matrix EVD:

$$\mathbf{A}_{(n)} \cdot {\mathbf{A}_{(n)}}^T = \mathbf{U}^{(n)} \cdot \text{diag}((\sigma_1^{(n)})^2, \ldots, (\sigma_{I_n}^{(n)})^2) \cdot {\mathbf{U}^{(n)}}^T$$

Large datasets:

[*Mahoney et al. '06*], [*Tyrtyshnikov et al. '06*], [*Oseledets et al. '08*]

Matrix formulation:

$$\mathbf{A}_{I_1 \times I_2 I_3 I_4} = \mathbf{U}^{(1)} \cdot \mathbf{S}_{R_1 \times R_2 R_3 R_4} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)} \otimes \mathbf{U}^{(4)})^T$$
$$\mathbf{A}_{I_1 I_2 \times I_3 I_4} = (\mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)}) \cdot \mathbf{S}_{R_1 R_2 \times R_3 R_4} \cdot (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(4)})^T$$

Kronecker product:

[*Van Loan '00*]

Overview

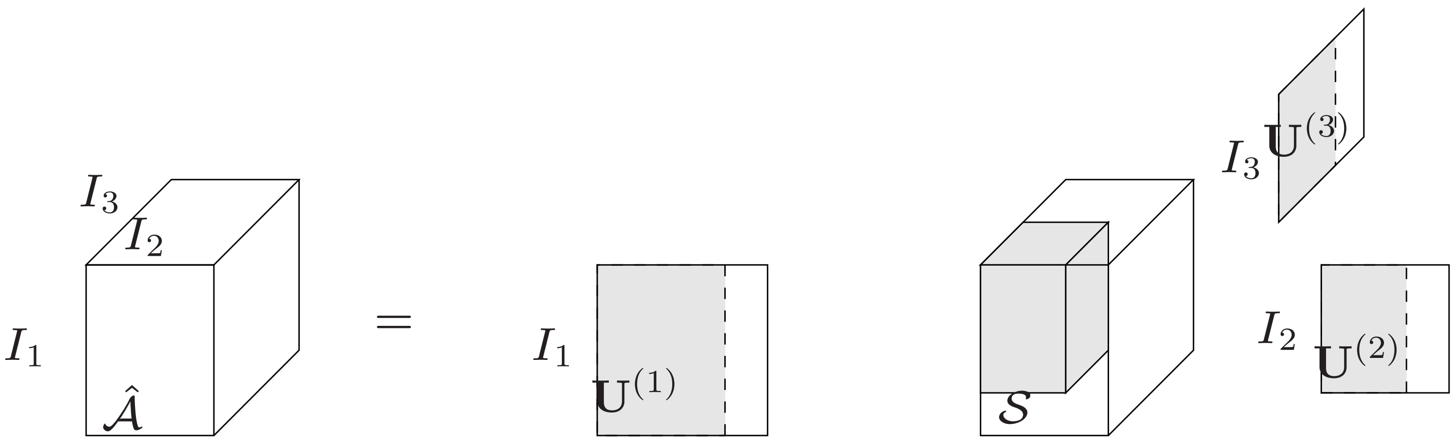
- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

Best Multilinear Rank-$(R_1, R_2, \ldots, R_N)$ Approximation

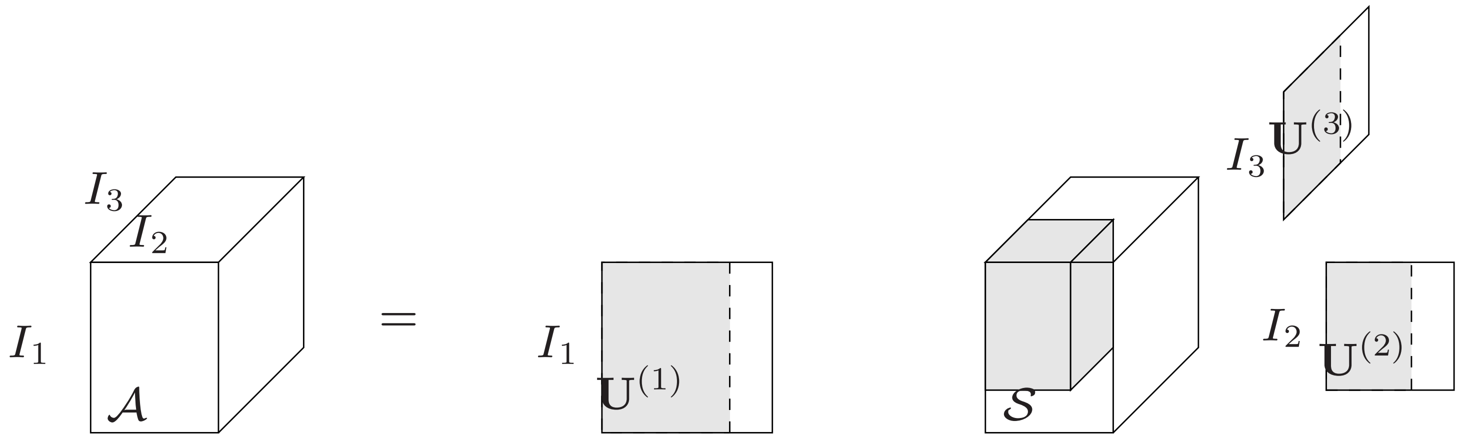
Problem: minimize $\|\mathcal{A} - \hat{\mathcal{A}}\|$ subject to

$$\text{rank}_1(\hat{\mathcal{A}}) \leqslant R_1 \qquad \text{rank}_2(\hat{\mathcal{A}}) \leqslant R_2 \qquad \text{rank}_3(\hat{\mathcal{A}}) \leqslant R_3$$

Parameterization: $\mathcal{A} = \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \mathbf{U}^{(3)}$



Truncation of multilinear SVD



Gives good but not optimal approximation

Further optimization required

Error bound:

$$\|\mathcal{A} - \hat{\mathcal{A}}\|^2 \leqslant \sum_{i_1=R_1+1}^{I_1} \sigma_{i_1}^{(1)^2} + \sum_{i_2=R_2+1}^{I_2} \sigma_{i_2}^{(2)^2} + \sum_{i_3=R_3+1}^{I_3} \sigma_{i_3}^{(3)^2}$$

Problem reformulation

Matrix case:

$$\min \|\mathbf{A} - \mathbf{U}^{(1)} \cdot \mathbf{\Sigma} \cdot {\mathbf{U}^{(2)}}^T\|$$

is equivalent with

$$\max \|{\mathbf{U}^{(1)}}^T \cdot \mathbf{A} \cdot \mathbf{U}^{(2)}\|$$

Tensor case:

$$\min \|\mathcal{A} - \mathcal{S} \bullet_1 \mathbf{U}^{(1)} \bullet_2 \mathbf{U}^{(2)} \bullet_3 \mathbf{U}^{(3)}\|$$

is equivalent with

$$\max \|\mathcal{A} \bullet_1 {\mathbf{U}^{(1)}}^T \bullet_2 {\mathbf{U}^{(2)}}^T \bullet_3 {\mathbf{U}^{(3)}}^T\|$$

$$= \max \|{\mathbf{U}^{(1)}}^T \cdot \mathbf{A}^{(1)} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)})^T\|$$

$$= \max \|{\mathbf{U}^{(2)}}^T \cdot \mathbf{A}^{(2)} \cdot (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(1)})^T\|$$

$$= \max \|{\mathbf{U}^{(3)}}^T \cdot \mathbf{A}^{(3)} \cdot (\mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)})^T\|$$

Suggests alternating least squares algorithm

(Higher-order) orthogonal iteration

Matrix case: Iterate over:

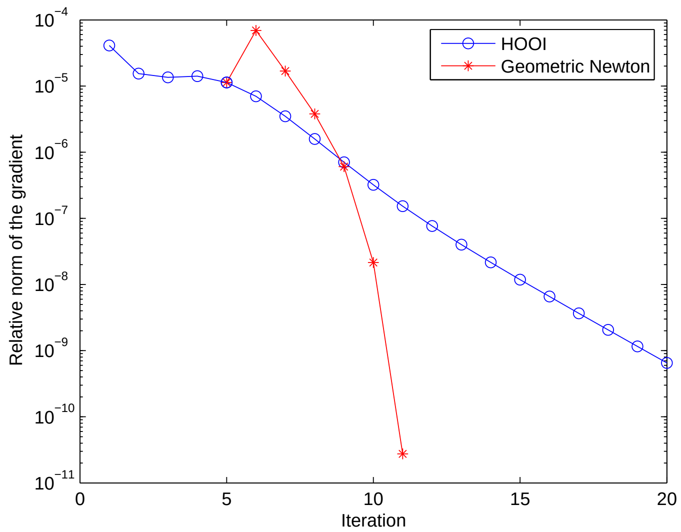
1. Compute $\mathbf{B}^{(1)} = \mathbf{A} \cdot \mathbf{U}^{(2)}$; $\mathbf{U}^{(1)} = \text{qf}(\mathbf{B}^{(1)})$
2. Compute $\mathbf{B}^{(2)} = \mathbf{A}^T \cdot \mathbf{U}^{(1)}$; $\mathbf{U}^{(2)} = \text{qf}(\mathbf{B}^{(2)})$

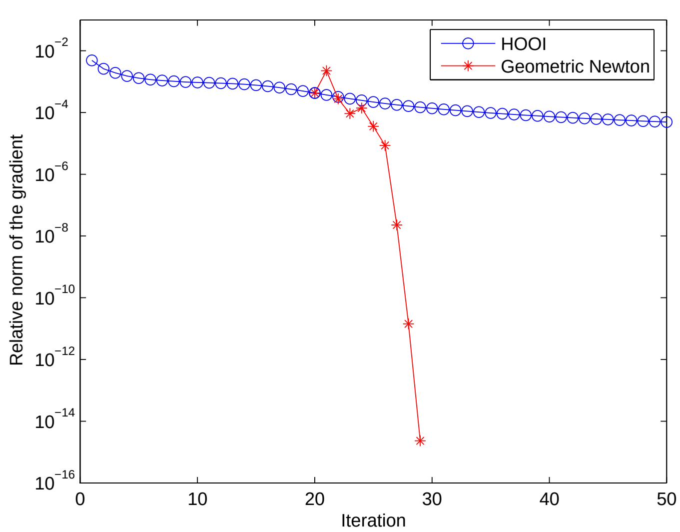
Tensor case: Iterate over:

1. Compute $\mathbf{B}^{(1)} = \mathbf{A}^{(1)} \cdot (\mathbf{U}^{(2)} \otimes \mathbf{U}^{(3)})^T$; $\mathbf{U}^{(1)}$ maximizes $\|{\mathbf{U}^{(1)}}^T \cdot \mathbf{B}^{(1)}\|$
2. Compute $\mathbf{B}^{(2)} = \mathbf{A}^{(2)} \cdot (\mathbf{U}^{(3)} \otimes \mathbf{U}^{(1)})^T$; $\mathbf{U}^{(2)}$ maximizes $\|{\mathbf{U}^{(2)}}^T \cdot \mathbf{B}^{(2)}\|$
3. Compute $\mathbf{B}^{(3)} = \mathbf{A}^{(3)} \cdot (\mathbf{U}^{(1)} \otimes \mathbf{U}^{(2)})^T$; $\mathbf{U}^{(3)}$ maximizes $\|{\mathbf{U}^{(3)}}^T \cdot \mathbf{B}^{(3)}\|$

[*Kroonenberg '83*], [*De Lathauwer '00*]

Orthogonal iterations can be slow [*Zhang and Golub '01*], [*Eldén and Savas '06*]





dominant rank-(R_1, R_2, R_3) part

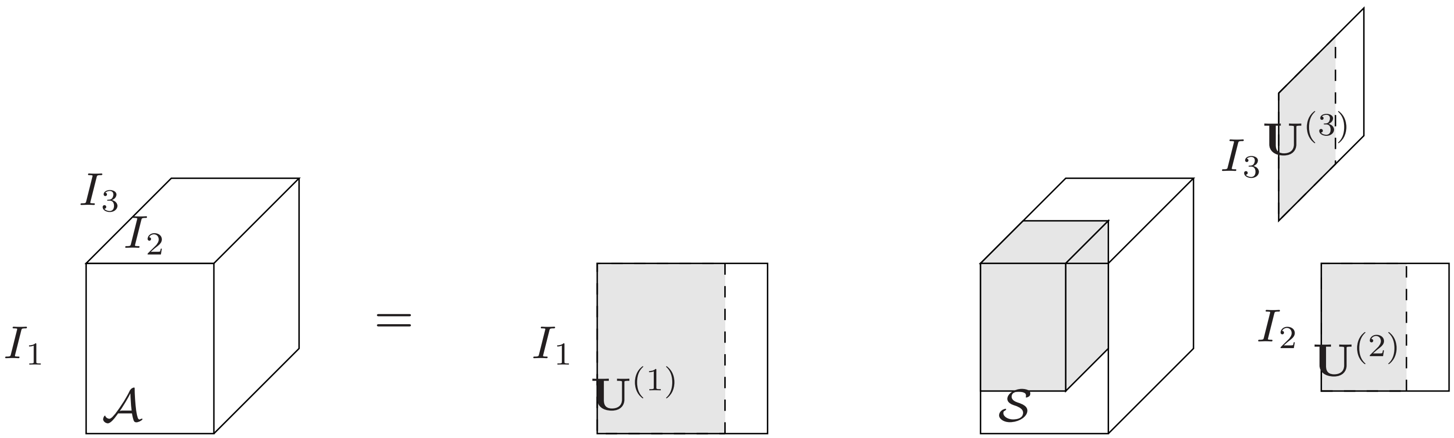
random tensor

Overview

- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

Only subspaces are of interest

$\longrightarrow$ work on products of quotient spaces (Grassmann)



$\longrightarrow$ Optimization on manifolds

[*Edelman, Arias and Smith '98*], [*Absil, Mahoney and Sepulchre '08*]

Faster solutions

Optimization on manifolds:

- Newton [*Eldén and Savas '06*], [*Ishteva et al. '08*]
- Quasi-Newton [*Savas and Lim '08*]
- Trust region [*Ishteva et al. '09*]
- Conjugate gradient [*Ishteva et al. '09*]

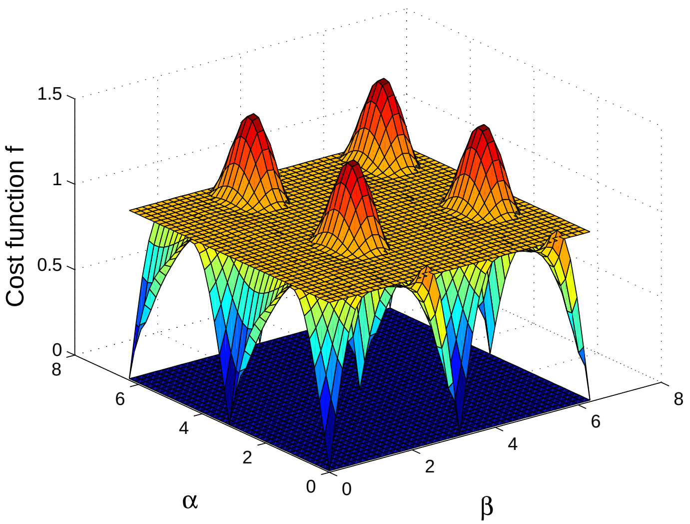
Krylov method: [*Savas '08*]

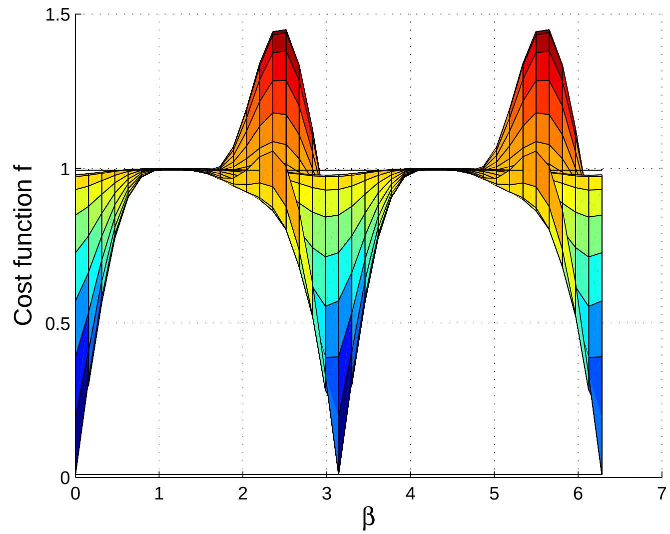
Overview

- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

Local optima (1)

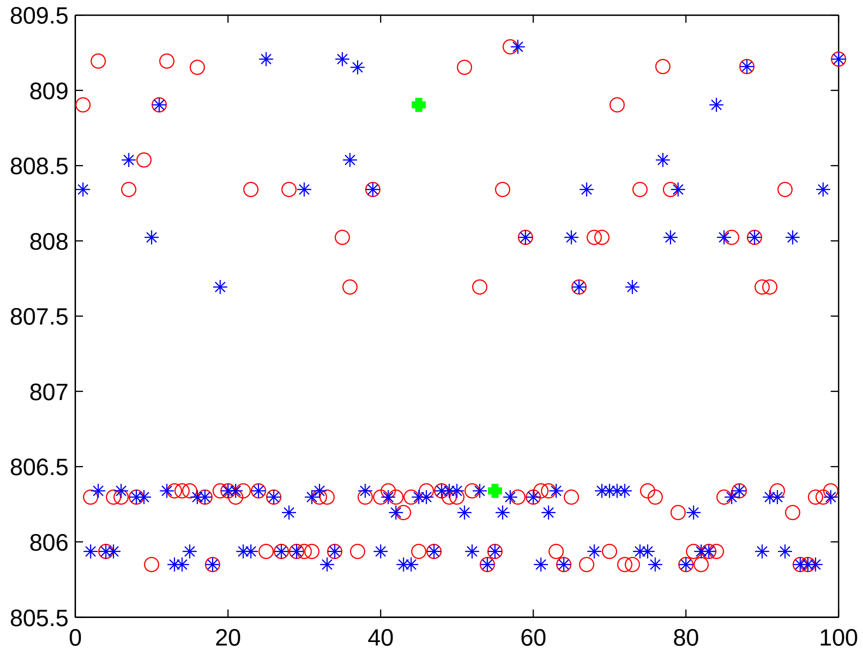
Tensor with dominant rank-(R_1, R_2, R_3) part:

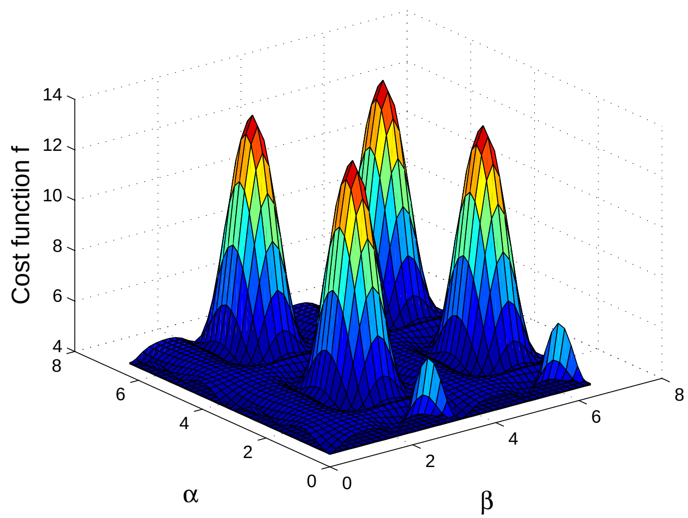


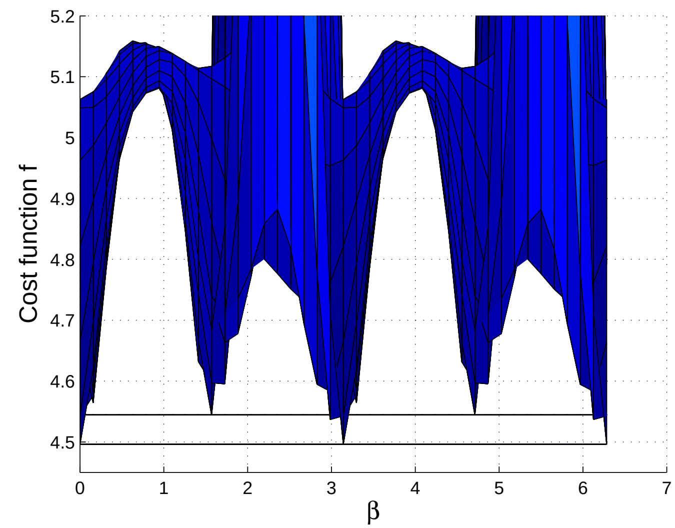


Local optima (2)

Random tensor:







$\longrightarrow$ inspection of multilinear singular values is important

Overview

- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications

Hierarchical Tucker compression

Work in several steps:

- Computation multilinear singular values and vectors
- Standard Tucker compression. Compression ratio dictated by

$$\|\mathcal{A} - \hat{\mathcal{A}}\|^2 \leqslant \sum_{i_1=R_1+1}^{I_1} {\sigma_{i_1}^{(1)}}^2 + \sum_{i_2=R_2+1}^{I_2} {\sigma_{i_2}^{(2)}}^2 + \sum_{i_3=R_3+1}^{I_3} {\sigma_{i_3}^{(3)}}^2$$

- Best rank-$(\tilde{R}_1, \tilde{R}_2, \tilde{R}_3)$ approximation

Overview

- Preliminaries
- Tucker decomposition / Multilinear SVD
- Best rank-(R_1, R_2, R_3) approximation
- Numerical algorithms
- Local optima
- Hierarchical Tucker compression
- Applications: dimensionality reduction
  estimation dominant subspace

Preprocessing for CANDECOMP/PARAFAC decomposition

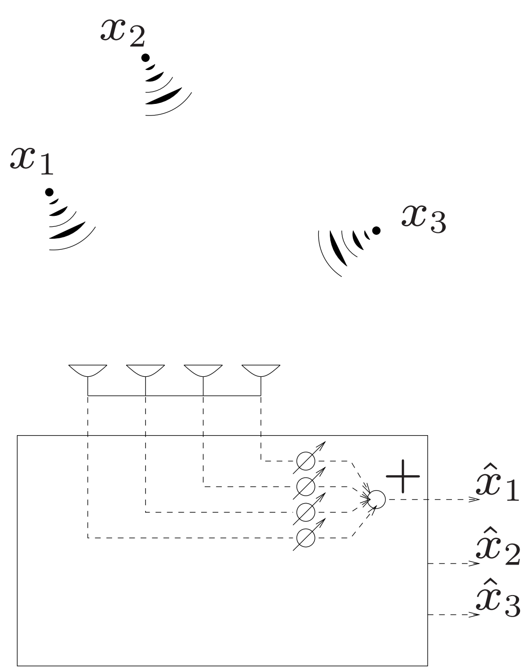
General principle:

- reduce dimensionality by means of multilinear approximation
- separate the signals in low-dimensional space
- dimensions do not have to be minimal
- transformation matrices are important
- inspection of multilinear singular values

Independent Component Analysis (ICA)

Model:

$$Y = \mathbf{M}X$$



statistical independence implies: - the variables are uncorrelated
- additional conditions on the HOS

Applications

- Speech and audio
- Image processing

  feature extraction, image reconstruction, video
- Telecommunications

  OFDM, CDMA, . . .
- Biomedical applications

  functional Magnetic Resonance Imaging, electromyogram, electro-encephalogram, (fetal) electrocardiogram, mammography, pulse oximetry, (fetal) magnetocardiogram, . . .
- Other applications

  text classification, vibratory signals generated by termites (!), electron energy loss spectra, astrophysics, . . .

ICA: basic equations

Model:

$$Y = \mathbf{M}X$$

Second order:

$$\begin{aligned}
\mathbf{C}_2^Y &= E\{YY^T\} \\
&= \mathbf{M} \cdot \mathbf{C}_2^X \cdot \mathbf{M}^{\mathrm{T}} \\
&= \mathbf{C}_2^X \bullet_1 \mathbf{M} \bullet_2 \mathbf{M}
\end{aligned}$$

uncorrelated sources: $\mathbf{C}_2^X$ is diagonal

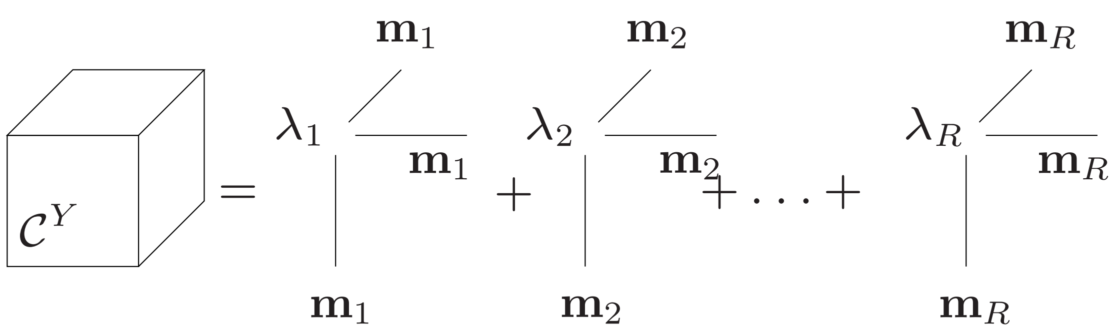
"diagonalization by congruence"

$$\boxed{\mathbf{C}_2^Y} = \sigma_1^2 \, \mathbf{m}_1 \, \mathbf{m}_1 + \sigma_2^2 \, \mathbf{m}_2 \, \mathbf{m}_2 + \ldots + \sigma_R^2 \, \mathbf{m}_R \, \mathbf{m}_R$$

Higher order:

$$\mathcal{C}_4^Y = \mathcal{C}_4^X \bullet_1 \mathbf{M} \bullet_2 \mathbf{M} \bullet_3 \mathbf{M} \bullet_4 \mathbf{M}$$

independent sources: $\mathcal{C}_4^X$ is diagonal



- This is a CANDECOMP/PARAFAC decomposition
- Hence, similar conclusions
- Often dimensionality reduction in prewhitening step (PCA):
  - Okay if noise is white
  - Not okay if noise is coloured

Harmonic retrieval

Discrete-time model:

$$x(n) = \sum_{r=1}^{R} a_r \exp\{j\varphi_r\} \exp\{(\alpha_r + 2j\pi\nu_r)n\Delta t\} \qquad n = 0, \ldots, N$$

or

$$x(n) = \sum_{r=1}^{R} c_r z_r^n \qquad n = 0, \ldots, N$$

- $c_r = a_r \exp\{j\varphi_r\}$: complex amplitudes,
- $z_r = \exp\{(\alpha_r + 2j\pi\nu_r)\Delta t\}$: signal poles

Applications:

- magnetic resonance spectroscopy
- direction-of-arrival estimation
- wireless communications
- audio and speech signal processing
- structural monitoring
- seismology
- radar
- underwater acoustics
- shape from moments
- ...

Subspace based computation

Vandermonde decomposition

$$\mathbf{H} \triangleq \begin{pmatrix} x(0) & x(1) & \cdots & x(N-L+1) \\ x(1) & x(2) & \cdots & x(N-L+2) \\ \vdots & \vdots & & \vdots \\ x(L-1) & x(L) & \cdots & x(N) \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_R \\ \vdots & \vdots & & \vdots \\ z_1^{L-1} & z_2^{L-1} & \ldots & z_R^{L-1} \end{pmatrix} \cdot \begin{pmatrix} c_1 & & \\ & \ddots & \\ & & c_R \end{pmatrix} \cdot \begin{pmatrix} 1 & z_1 & \ldots & z_1^{N-L+1} \\ 1 & z_2 & \ldots & z_2^{N-L+1} \\ \vdots & \vdots & & \vdots \\ 1 & z_R & \ldots & z_R^{N-L+1} \end{pmatrix} \triangleq \mathbf{S} \cdot \mathbf{C} \cdot \mathbf{T}^T$$

Truncated singular value decomposition

$$\mathbf{H} = \hat{\mathbf{U}} \cdot \hat{\mathbf{\Sigma}} \cdot \hat{\mathbf{V}}^H$$

Construction of a third-order tensor

- Construct a third-order Hankel-type tensor
- This tensor admits a higher-order Vandermonde decomposition
- Compute mode-n subspace generated by the signal poles

[*Papy, De Lathauwer, Van Huffel '05*], [*Haardt, Roemer, Del Galdo '08*], [*Papy, De Lathauwer, Van Huffel '09*]

Announcement

International Conference on
Tensor Decompositions and Applications (TDA 2010)
Sept. 13–17, 2010
Monopoli (Bari), Italy