Robust and sparse estimation of tensor decompositions

Visa Koivunen

Department of Signal Processing and Acoustics
Aalto University

Signal Processing
Summer School
2016

Outline

- High-dimensional Multi Aspect data
  Neuroimaging, Remote Sensing, Chemometrics, Environmetrics, Network Data, Internet Data, Data collected by mobile terminals

Agenda

1. High-Dimensional Data
2. Introduction to Tensors
3. (Sparse) Regularization of Tensor decompositions
4. Robust Tensor Estimation

Outline

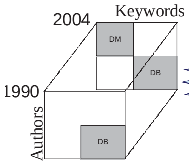
- High-dimensional Multi Aspect data
  Neuroimaging, Remote Sensing, Chemometrics, Environmetrics, Network Data, Internet Data, Data collected by mobile terminals
- Tensors accommodate such data naturally as multi-way arrays. Tensor decompositions of multilinear models provide a unifying framework for multidimensional data analysis with simplified notations and algebras.

Outline

- High-dimensional Multi Aspect data
  Neuroimaging, Remote Sensing, Chemometrics, Environmetrics, Network Data, Internet Data, Data collected by mobile terminals
- Tensors accommodate such data naturally as multi-way arrays. Tensor decompositions of multilinear models provide a unifying framework for multidimensional data analysis with simplified notations and algebras.
- Sparsity constraints can be used
  for accurate signal recovery (e.g. compressed sensing) or
  to eliminate unnecessary redundant features of modern data sets (e.g. financial data, DNA micro arrays, network traffic flows, fMRIs).

Tensors

- A tensor $\boldsymbol{\mathcal{A}}$ of order d: a d-way array with d indices $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ represents a multi-linear operator with coordinates $\boldsymbol{\mathcal{A}}_{i_1..i_d}$ (or $\boldsymbol{\mathcal{A}}(i_1, \cdots, i_d)$)
- Low-order tensors: scalar a (order-0 tensor),
  vector $\mathbf{a}$ (order-1 tensor),
  matrix $\mathbf{A}$ (order-2 tensor)
- Example: Social Network Analysis
  Data in three modes: time, author, keyword (order-3 tensor)

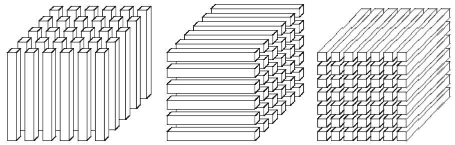


Outline

- High-dimensional Multi Aspect data
  Neuroimaging, Remote Sensing, Chemometrics, Environmetrics, Network Data, Internet Data, Data collected by mobile terminals
- Tensors accommodate such data naturally as multi-way arrays. Tensor decompositions of multilinear models provide a unifying framework for multidimensional data analysis with simplified notations and algebras.
- Sparsity constraints can be used
  for accurate signal recovery (e.g. compressed sensing) or
  to eliminate unnecessary redundant features of modern data sets (e.g. financial data, DNA micro arrays, network traffic flows, fMRIs).
- Robustness ensures the resistance to heavy-tailed errors or outliers that appear commonly in high-dimensional data, improving data analyses.

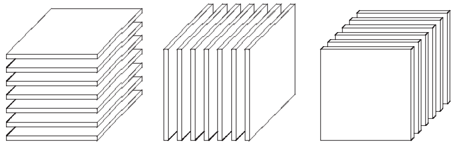
Terminologies

- Mode: number of dimensions, also known as ways or orders
- Fiber: the higher order analogue of matrix rows and columns, obtained by fixing all but one index



eg. $\boldsymbol{\mathcal{A}}(:, j, k)$, $\boldsymbol{\mathcal{A}}(i, :, k)$, $\boldsymbol{\mathcal{A}}(i, j, :)$ for a third-order tensor $\boldsymbol{\mathcal{A}}$

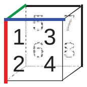
- Slice: two-dimensional sections of a tensor, defined by fixing all but two indices



eg. $\boldsymbol{\mathcal{A}}(i, :, :)$, $\boldsymbol{\mathcal{A}}(:, j, :)$, $\boldsymbol{\mathcal{A}}(:, :, k)$ for a third-order tensor $\boldsymbol{\mathcal{A}}$

Tensor Unfoldings

- Unfolding (= Flattening, Matricizing): Converting a Tensor to a Matrix
- Mode-k unfolding: mode-k fibers of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ are assembled to produce an n_k by N/n_k matrix where $N = n_1 \cdots n_d$ and $k = 1, \ldots d$.

  E.g. Let $\boldsymbol{\mathcal{A}} =$ 

  Three mode-k unfoldings are

$$\mathbf{A}_{(1)} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix}$$

$$\mathbf{A}_{(2)} = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}$$

$$\mathbf{A}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

- The inner product of $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is

$$< \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} >= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \mathrm{a}_{ijk} \mathrm{b}_{ijk} = \mathrm{vec}(\boldsymbol{\mathcal{A}})^T \mathrm{vec}(\boldsymbol{\mathcal{B}})$$

- Frobenius norm :

$$\|\boldsymbol{\mathcal{A}}\|_F = \sqrt{< \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{A}} >}$$

Tensor Unfoldings

- Different people use different ordering of the columns for the mode-k unfolding. In general, the specific permutation of columns is not important so long as it is consistent across related calculations.
- 'Vectorization' : vec operation

  Turn matrices into vectors by stacking columns

  Turn tensors into vectors by stacking mode-1 fibers

  E.g. (continued)

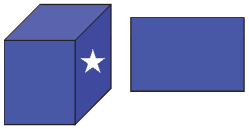
$$\mathrm{vec}(\boldsymbol{\mathcal{A}}) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{bmatrix}$$

Tensor Mode-n Multiplication

- Tensor times Matrix

  For $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{I \times J \times K}$, $\mathbf{B} \in \mathbb{R}^{M \times J}$, and $\mathbf{c} \in \mathbb{R}^{I}$,

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{A}} \times_2 \mathbf{B} \qquad (\in \mathbb{R}^{I \times M \times K})$$



$$\mathrm{x}_{imk} = \sum_j \mathrm{a}_{ijk} b_{mj}$$

$$\mathbf{X}_{(2)} = \mathbf{B}\mathbf{A}_{(2)}$$

- Tensor times Vector

$$\mathbf{X} = \boldsymbol{\mathcal{A}} \ \bar{\times}_1 \ \mathbf{c} \qquad (\in \mathbb{R}^{J \times K})$$

$$\mathrm{x}_{jk} = \sum_i \mathrm{a}_{ijk} c_i$$

Matrix Products

- $\circ$ denotes outer product: $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^T$, for $\mathbf{a} \in \mathbb{R}^I$, $\mathbf{b} \in \mathbb{R}^J$, $\mathbf{c} \in \mathbb{R}^K$.
  E.g. $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} (\in \mathbb{R}^{I \times J \times K}) : (\mathbf{a} \circ \mathbf{b} \circ \mathbf{c})_{ijk} = a_i b_j c_k$
- Kronecker Product of matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1J}B \\ a_{21}B & a_{22}B & \cdots & a_{2J}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}B & a_{I2}B & \cdots & a_{IJ}B \end{bmatrix}$$

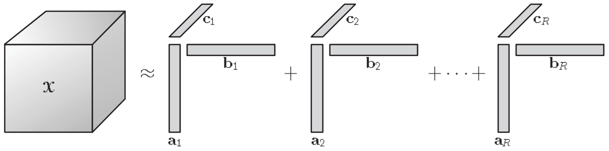
- Khatri-Rao product of matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$ and $\mathbf{B} \in \mathbb{R}^{J \times R}$ is a $IJ \times R$ matrix: $\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \cdots & \mathbf{a}_R \otimes \mathbf{b}_R \end{bmatrix}$.
  In the vector case, $\mathbf{a} \otimes \mathbf{b} = \mathrm{vec}(\mathbf{b}\mathbf{a}^T)$.

Specially Structured Tensors (3-way tensors)

- Kruskal Tensor : $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$, $\boldsymbol{\gamma} \in \mathbb{R}^R$,

$$\boldsymbol{\mathcal{X}} \equiv [\![\boldsymbol{\gamma}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r .$$

  where $\mathbf{a} \in \mathbb{R}^I$, $\mathbf{b} \in \mathbb{R}^J$ and $\mathbf{c} \in \mathbb{R}^K$ form the unit-norm column vectors of $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$



Rank of a tensor

- rank($\boldsymbol{\mathcal{X}}$) : the smallest number of rank-one tensors that sum to $\boldsymbol{\mathcal{X}}$.
  E.g. If $\mathbf{a} \in \mathbb{R}^I$, $\mathbf{b} \in \mathbb{R}^J$, $\mathbf{c} \in \mathbb{R}^K$, then $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ is a rank-1 tensor.
  E.g. If a tensor $\boldsymbol{\mathcal{X}}$ has a minimal representation as

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^{r_3} \sigma_{ijk} (\mathbf{c}_k \otimes \mathbf{b}_j \otimes \mathbf{a}_i)$$

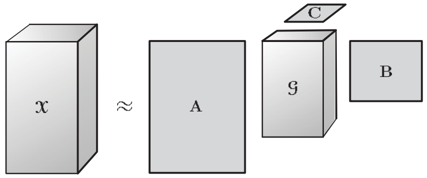
  then rank($\boldsymbol{\mathcal{X}}$) $= r_1 r_2 r_3$
- No known method to determine the rank of a specific given tensor.
- The rank of a particular tensor over the real field may be different from its rank over the complex field.

Specially Structured Tensors (3-way tensors)

- Tucker Tensor : $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I \times J \times K}$

$$\boldsymbol{\mathcal{X}} \equiv [\![\boldsymbol{\mathcal{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \boldsymbol{\mathcal{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r$$

  where $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ are the factor matrices $(\mathbf{A}^T\mathbf{A} = \mathbf{I}, \mathbf{B}^T\mathbf{B} = \mathbf{I}, \mathbf{C}^T\mathbf{C} = \mathbf{I})$ and $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core tensor.



- Kruskal tensor is a special case of Tucker tensor where the core tensor is superdiagonal and $P = Q = R$.

Rank-1 Tensors

Recall that $(\mathbf{a}\circ\mathbf{b}\circ\mathbf{c})_{ijk} = a_i b_j c_k$ for $\mathbf{a}\in\mathbb{R}^I$, $\mathbf{b}\in\mathbb{R}^J$, and $\mathbf{c}\in\mathbb{R}^K$.

$$\boldsymbol{\mathcal{Y}} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \circ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \circ \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \Leftrightarrow \mathrm{vec}(\boldsymbol{\mathcal{Y}}) = \begin{bmatrix} y_{111} \\ y_{211} \\ y_{121} \\ \vdots \\ y_{132} \\ y_{232} \end{bmatrix} = \begin{bmatrix} a_1 b_1 c_1 \\ a_2 b_1 c_1 \\ a_1 b_2 c_1 \\ \vdots \\ a_1 b_3 c_2 \\ a_2 b_3 c_2 \end{bmatrix} = \mathbf{c}\otimes\mathbf{b}\otimes\mathbf{a}$$

$$\begin{aligned} \mathbf{Y}_{(1)} &= \begin{bmatrix} a_1b_1c_1 & a_1b_2c_1 & a_1b_3c_1 & a_1b_1c_2 & a_1b_2c_2 & a_1b_3c_2 \\ a_2b_1c_1 & a_2b_2c_1 & a_2b_3c_1 & a_2b_1c_2 & a_2b_2c_2 & a_2b_3c_2 \end{bmatrix} \\ &= \begin{bmatrix} a_1\cdot(\mathbf{c}\otimes\mathbf{b})^T \\ a_2\cdot(\mathbf{c}\otimes\mathbf{b})^T \end{bmatrix} = \mathbf{a}\otimes(\mathbf{c}\otimes\mathbf{b})^T \end{aligned}$$

Tensor decompositions

approximate a tensor by a low-rank set of factors along each tensor mode

- CANDECOMP = Canonical Decomposition (Carrol and Chang, 1970)
- PARAFAC = Parallel Factors (Harshman, 1970)
- CANDECOMP / PARAFAC (CP) decomposition

$$\boldsymbol{\mathcal{X}} \approx [\![\boldsymbol{\gamma};\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r\circ\mathbf{b}_r\circ\mathbf{c}_r.$$

- Tucker decomposition (Tucker, 1966) : Three-mode factor analysis, Three-mode PCA, or Orthogonal array decomposition

$$\boldsymbol{\mathcal{X}} \approx [\![\boldsymbol{\mathcal{G}};\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\mathbf{a}_p\circ\mathbf{b}_q\circ\mathbf{c}_r$$

  Not Unique!
- HOSVD, HOOI, HOPM,...

Tensors in matrix form

- Kruskal Tensor : $\boldsymbol{\mathcal{X}}\in\mathbb{R}^{I\times J\times K}$

$$\begin{aligned} \mathbf{X}_{(1)} &= \sum_{r=1}^{R} \gamma_r \mathbf{a}_r\otimes(\mathbf{c}_r\otimes\mathbf{b}_r)^T = \mathbf{A}\mathrm{diag}(\gamma_r)(\mathbf{C}\odot\mathbf{B})^T \\ \mathbf{X}_{(2)} &= \mathbf{B}\boldsymbol{\Gamma}(\mathbf{C}\odot\mathbf{A})^T \\ \mathbf{X}_{(3)} &= \mathbf{C}\boldsymbol{\Gamma}(\mathbf{B}\odot\mathbf{A})^T \quad \text{where } \boldsymbol{\Gamma} \text{ is } \mathrm{diag}(\boldsymbol{\gamma}) \end{aligned}$$

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) = (\mathbf{C}\odot\mathbf{B}\odot\mathbf{A})\boldsymbol{\gamma}$$

- Tucker Tensor : $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}}\times_1\mathbf{A}\times_2\mathbf{B}\times_3\mathbf{C}\in\mathbb{R}^{I\times J\times K}$

$$\begin{aligned} \mathbf{X}_{(1)} &= \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})^T \\ \mathbf{X}_{(2)} &= \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C}\otimes\mathbf{A})^T \\ \mathbf{X}_{(3)} &= \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B}\otimes\mathbf{A})^T \end{aligned}$$

$$\mathrm{vec}(\boldsymbol{\mathcal{X}}) = (\mathbf{C}\otimes\mathbf{B}\otimes\mathbf{A})\mathrm{vec}(\boldsymbol{\mathcal{G}})$$

Uniqueness of decompositions

- Tucker is NOT unique. Let $\mathbf{U}$ be an $P\times P$ orthogonal matrix.

$$\boldsymbol{\mathcal{X}} \approx [\![\boldsymbol{\mathcal{G}};\mathbf{A},\mathbf{B},\mathbf{C}]\!] = \boldsymbol{\mathcal{G}}\times_1\mathbf{A}\times_2\mathbf{B}\times_3\mathbf{C} = (\boldsymbol{\mathcal{G}}\times_1\mathbf{U}^T)\times_1(\mathbf{A}\mathbf{U})\times_2\mathbf{B}\times_3\mathbf{C}$$

$$\mathbf{X}_{(1)} \approx \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})^T = \mathbf{A}\mathbf{U}\mathbf{U}^T\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})^T$$

- CP is often unique.
  Assume that CP decomposition is exact.
  Sufficient condition for uniqueness (Kruskal, 1977):

$$2R+2 \le \kappa_A + \kappa_B + \kappa_C$$

  where κ_A = k-rank of a matrix $\mathbf{A}$ = max number k such that any k columns are linearly independent.

Solving for Tucker

$$\text{For } \mathcal{X} \approx [\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$$

Given that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are orthonormal, the optimal core is

$$\mathcal{G} = [\![\mathcal{X}; \mathbf{A}^T, \mathbf{B}^T, \mathbf{C}^T]\!]$$

$$\begin{aligned} \|\mathcal{X} - [\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|^2 &= \|\mathcal{X}\|^2 - 2 < \mathcal{X}, [\![\mathcal{G}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] > + \|\mathcal{G}\|^2 \\ &= \|\mathcal{X}\|^2 - \|\mathcal{G}\|^2 \end{aligned}$$

If $\mathbf{B}$ and $\mathbf{C}$ are fixed, then we can solve for $\mathbf{A}$ as follows:

$$\|[\![\mathcal{X}; \mathbf{A}^T, \mathbf{B}^T, \mathbf{C}^T]\!]\| = \|[\![\mathbf{A}^T \mathbf{X}_{(1)}(\mathbf{C} \otimes \mathbf{B})]\!]\|$$

Optimal $\mathbf{A}$ is P left leading singular vectors for $\mathbf{X}_{(1)}(\mathbf{C} \otimes \mathbf{B})$.

Higher Order Orthogonal Iterations (HOOI)

Tucker-Alternating Least Squares : Successively solve for each component while fixing others.

- Initialize P, Q, R. Calculate $\mathbf{A}, \mathbf{B}, \mathbf{C}$ via HOSVD.
- Repeat until converged...
  -$\mathbf{A}$ is P left leading singular vectors for $\mathbf{X}_{(1)}(\mathbf{C} \otimes \mathbf{B})$.
  -$\mathbf{B}$ is Q left leading singular vectors for $\mathbf{X}_{(2)}(\mathbf{C} \otimes \mathbf{A})$.
  -$\mathbf{C}$ is R left leading singular vectors for $\mathbf{X}_{(3)}(\mathbf{B} \otimes \mathbf{A})$.
- Solve for the core:
  $$\mathcal{G} = [\![\mathcal{X}; \mathbf{A}^T, \mathbf{B}^T, \mathbf{C}^T]\!]$$

Kroonenberg & De Leeuw, Psychometrika, 1980

HOSVD (Higher Order SVD)

De Lathauwer, De Moor, & Vandewalle, SIMAX, 1980
The HOSVD of a tensor $\mathcal{X}$ involves computing the matrix SVDs of its modal unfoldings $\mathcal{X}_{(1)}, ..., \mathcal{X}_{(d)}$.
For $\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$

1. $\mathbf{A}$ is P left leading singular vectors for $\mathbf{X}_{(1)}$.
2. $\mathbf{B}$ is Q left leading singular vectors for $\mathbf{X}_{(2)}$.
3. $\mathbf{C}$ is R left leading singular vectors for $\mathbf{X}_{(3)}$.
4. $$\mathcal{G} = [\![\mathcal{X}; \mathbf{A}^T, \mathbf{B}^T, \mathbf{C}^T]\!]$$

HOSVD: Not optimal but often used as an initialization for Tucker-ALS algorithm
The core is NOT in general diagonal.
Unlike the matrix SVD, HOSVD cannot be expressed as a sum of a few orthogonal outer-product terms.

CANDECOMP / PARAFAC (CP) decomposition

approximates a tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ by a predicted tensor $\hat{\mathcal{X}}$ consisting of a sum of rank-1 tensors:

$$\hat{\mathcal{X}} \equiv [\![\gamma; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \triangleq \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$

Thus, we model $\mathcal{X}$ as

$$\mathcal{X} = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \mathcal{E} \qquad (1)$$

where $\mathbf{a}_r \in \mathbb{R}^I$, $\mathbf{b}_r \in \mathbb{R}^J$ and $\mathbf{c}_r \in \mathbb{R}^K$ for $r = 1, \dots, R$ form the unit-norm column vectors of $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ and the tensor $\mathcal{E} \in \mathbb{R}^{I \times J \times K}$ contains the *error terms.*
GOAL: *to minimize*

$$\|\mathcal{X} - \hat{\mathcal{X}}\|_F = \|\mathbf{X}_{(k)} - \hat{\mathbf{X}}_{(k)}\|_F$$

CP decomposition

The model (1) can be expressed in a *matrix form by unfolding the tensor into a matrix along any of the three modes.*
Unfolding the tensor $\mathcal{X}$ *along the first mode yields a* $I \times JK$*-matrix denoted as* $\mathbf{X}_{(1)}$ *so that the equivalent representation of* (1) *is*

$$\mathbf{X}_{(1)} = \mathbf{A}\Gamma(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E}_{(1)}, \qquad (2)$$

where $\Gamma = \mathrm{diag}(\gamma)$ *and* $\mathbf{E}_{(1)}$ *denotes the unfolded* $I \times JK$ *matrix of* $\mathcal{E}$*.*
Similarly,

$$\mathbf{X}_{(2)} = \mathbf{B}\Gamma(\mathbf{C} \odot \mathbf{A})^T + \mathbf{E}_{(2)},$$
$$\mathbf{X}_{(3)} = \mathbf{C}\Gamma(\mathbf{B} \odot \mathbf{A})^T + \mathbf{E}_{(3)}$$

ALS for CP decompositions

1. Initialize $\mathbf{B}$ and $\mathbf{C}$ by $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$
2. $\hat{\mathbf{A}} = \mathbf{X}_{(1)}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \hat{\mathbf{C}} \odot \hat{\mathbf{B}}$
3. $\hat{\mathbf{B}} = \mathbf{X}_{(2)}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \hat{\mathbf{C}} \odot \hat{\mathbf{A}}$
4. $\hat{\mathbf{C}} = \mathbf{X}_{(3)}\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}$, where $\mathbf{Z} = \hat{\mathbf{B}} \odot \hat{\mathbf{A}}$
5. Repeat steps 2–4 until the relative change in fit is small.

The Alternating Least Squares (ALS) for CP decompositions

Consider the case that $\mathbf{B}$ and $\mathbf{C}$ are fixed and that γ_r's are the scales of the columns of $\mathbf{A}$, i.e., $\mathbf{a}_r$'s are no-longer unit vectors, but $\gamma_r = \|\mathbf{a}_r\|$.

$$\min_{\mathbf{A}} \|\mathcal{X} - \llbracket\gamma; \mathbf{A}, \mathbf{B}, \mathbf{C}\rrbracket\|^2 = \min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|^2 \qquad (3)$$

of which the LS solution is

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

with † denoting the Moore-Penrose inverse.
Note that $(\mathbf{C} \odot \mathbf{B})^T(\mathbf{C} \odot \mathbf{B}) = (\mathbf{B}^T\mathbf{B}) * (\mathbf{C}^T\mathbf{C})$ where $*$ denotes pointwise multiplication.

- ALS idea: Solve for each factor in turn, leaving all the others fixed.

Sparsity for Tensors

generally means that only a few entries are non-zero.

Two notions of Sparsity:

1. The considerable number of data elements are zero or close to zero in their relative magnitude.
2. In regularization methods (e.g. ridge regression, LASSO), **sparsity** is used for the estimated regression parameters that are either shrunk towards zero or put to zero by increasing the penalty of model complexity.

Relation of two notions: The underlying sparsity of tensor data naturally implies that factor matrices of a decomposed tensor are sparse as well. Regularization methods successfully estimate tensor factors compared to the usual tensor estimation based on the least squares.

Regularization methods

- Ridge Regression (Tikhonov regularization): A. E. Hoerl and R. W. Kennard, 1970
  Let $\mathbf{Z} = \mathbf{C} \odot \mathbf{B}$ for brevity in updating $\hat{\mathbf{A}}$ fixing the others fixed.
  The minimization in (3) simplifies to $\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}\mathbf{Z}^T\|_2^2$.
  The Ridge Regression in our context can be formulated

$$\hat{\mathbf{A}} \equiv \mathrm{RR}(\mathbf{X}_{(1)}, \mathbf{Z}, \lambda) = \arg\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}\mathbf{Z}^T\|_2^2 + \lambda\|\mathbf{A}\|_2^2. \qquad (4)$$

- LASSO (Least Absolute Shrinkage and Selection Operator): R. Tibshirani, 1996
  To obtain sparse solutions, we solve for $\mathbf{A}$ using $\ell_2 - \ell_1$ criterion function instead of ℓ_2 criterion in (4):

$$\hat{\mathbf{A}} \equiv \mathrm{LASSO}(\mathbf{X}_{(1)}, \mathbf{Z}, \lambda) = \arg\min_{\mathbf{A}} \|\mathbf{X}_{(1)} - \mathbf{A}\mathbf{Z}^T\|_2^2 + \lambda\|\mathbf{A}\|_1$$

Sparse Regularization methods

We introduce (sparse) regularization methods for tensor decompositions which are useful for dimensionality reduction, feature selection, data compression, data visualization as well as signal recovery.

- CP - Alternating Ridge Regression (CP-ARR)
- CP - Alternating LASSO
- Sparse HOSVD
- Sparse HOOI

Estimation of the penalty parameter λ

- The high level shrinkage (HLS) estimator: $\hat{\lambda} = \frac{1}{R}\sum_{j=1}^{R} d_j^2$, where $d_1 \geq d_2 \geq \cdots \geq d_R \geq 0$ are the singular values of $\mathbf{Z}$.
  In the ridge regression $\hat{\lambda} = JK$
- Bayesian information criteria (BIC):

$$\mathrm{BIC}(\lambda) = N \ln \hat{\sigma}^2 + \mathrm{df}(\lambda) \cdot \ln N \qquad (5)$$

  where $N = JK$ is the number of columns in $\mathbf{X}_{(1)}$,
  $\hat{\sigma}^2 = \frac{1}{I}\|\mathbf{X}_{(1)} - \mathbf{A}\mathbf{Z}^T\|_2^2$ is the average squared residuals.

  Degrees of freedom : $\mathrm{df}(\lambda) = I \cdot \mathrm{Tr}\{\mathbf{H}_\lambda\} = I \cdot \sum_{j=1}^{R} d_j^2(d_j^2 + \lambda)^{-1}$
  where $\mathbf{H}_\lambda = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T$ denotes the "hat matrix".

The BIC penalty parameter estimate of λ:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda_n} \mathrm{BIC}(\lambda) \qquad (6)$$

where Λ_n is a grid of n values $\lambda_{n-1} < \lambda_{n-2} < \cdots < \lambda_1 < \lambda_0$.

Initializations for tensor decomposition algorithms

Most of the tensor decomposition algorithms heavily depend on good initializations. (Kroonenberg, 2008)

- We propose the CP alternating ridge regression (CP-ARR) to provide good starting values taking advantage of sparsity.
  The CP-ARR can be a stand-alone method when the underlying structure of tensor data demands shrinkage with nonzero values instead of sparsity with many zero values.

- With such good starting values the CP alternating LASSO method shows highly improved performance compared to the conventional decomposition algorithms.

Robust tensor decompositions

- Outliers often occurring in high-dimensional data indicate some deviation from the model assumptions and add difficulty in data analysis.
- Tensor decompositions based on least squares are highly sensitive to outliers or heavy-tailed errors resulting in biased estimates.
- Surprisingly, the use of robust estimators has been largely neglected in the tensor community.

  Other than some work in the medical imaging,
  robust tensor factorization studies are found in
  Vorobyov et al., 2005; Pang and Yuan, 2010; Chi and Kolda, 2011.

Robust CP decompositions

- Vorobyov et al., 2005; Chi & Kolda, 2011: (LAD) Let $\mathbf{X} = \mathbf{X}_{(1)}$ in (3).

$$Q_{L_1}(\mathbf{A},\mathbf{B},\mathbf{C}) = \|\mathbf{X} - \mathbf{A}\mathbf{Z}^\top\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |x_{ij} - \mathbf{z}_j^\top \mathbf{a}_i|.$$

  Note: L_1-loss is not bounded!
- We propose an M-estimation type objective function:

$$Q_\rho(\mathbf{A},\mathbf{B},\mathbf{C}) = \sum_{i=1}^{m} \hat{\sigma}_i^2 \sum_{j=1}^{n} \rho\left(\frac{x_{ij} - \mathbf{z}_j^\top \mathbf{a}_i}{\hat{\sigma}_i}\right), \tag{7}$$

  where $\hat{\sigma}_i$ is a preliminary robust scale estimate and Huber's ρ-function defined as

$$\rho_k(e) = \begin{cases} \frac{1}{2}e^2, & \text{for } |e| \le k \\ k|e| - \frac{1}{2}k^2, & \text{for } |e| > k \end{cases} \tag{8}$$

  with a *tuning constant* k.

Robust Error Measure

- Least Absolute Deviations (LAD): For the regression parameter $\beta = (\beta_1, ..., \beta_p)$, the absolute value loss function is

$$l(r(\beta)) = \sum_{j=1}^{h} |r(\beta)|$$

- Least Trimmed Squares Regression (LTS): Rousseeuw (1984) Let $\{|r_{(j)}(\beta)|\}$ denote the set of ordered absolute values of the residuals. The LTS estimator is found by minimizing the sum of squared residuals over a subset of h observations

$$\hat{\beta}_{\mathrm{LTS}} = \min \sum_{j=1}^{h} |r_{(j)}(\beta)|^2$$

- Tukey Loss Function

$$\rho(r(\beta)) = \min\{1, (1 - (r(\beta)/c)^2)^3\}$$

  where $c = 3.4437$ attaining 85% of efficiency at normal distribution.

Robust and Sparse(regularized) Tensor Decompositions

We propose novel tensor decomposition methods that enjoy both the properties of sparsity and robustness to outliers.

- CP Alternating LAD + LASSO (CPA-LAD LASSO)
- CP Alternating Ridge M-Regression (CPA-RMR)
- CP Alternating LTS + LASSO
- CP Alternating Tukey + LASSO

CP Alternating LAD-LASSO METHOD

- Objective function for Robustness and Regularization:

$$\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} |x_{ij} - \mathbf{z}_j^\top \mathbf{a}_i| + \lambda_1 \|\mathbf{a}_i\|_1 \right\} + \lambda_2 \|\mathbf{B}\|_1 + \lambda_3 \|\mathbf{C}\|_1.$$

The minimum $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1 \ \cdots \ \hat{\mathbf{a}}_m)^\top$ can be found by

$$\hat{\mathbf{a}}_i = \min_{\mathbf{a}} \left\{ \sum_{j=1}^{n} |x_{ij} - \mathbf{z}_j^\top \mathbf{a}| + \lambda_1 \|\mathbf{a}\|_1 \right\} \qquad (9)$$

for $i = 1, \ldots, m$, when $\mathbf{B}$ and $\mathbf{C}$ are fixed.

Selection of the shrinkage parameter

- Bayesian information criteria (BIC):

$$\mathrm{BIC}(\boldsymbol{\lambda}) = 2N \ln \hat{\sigma} + w \cdot \mathrm{df}(\boldsymbol{\lambda}) \cdot \ln N \qquad (11)$$

where $N = I \cdot J \cdot K$, $w(= \sqrt{2})$ a weight assigned by the user, and $\hat{\sigma}$ is a scale estimate of the residuals.

$\hat{\sigma}^2 = \mathrm{ave}_{i,j}\{r_{i,}^2\}$ for CPA-LASSO,
$\hat{\sigma} = \mathrm{ave}_{i.j}\{|r_{ij}|\}$ for CPA-LADLASSO,
$\hat{\sigma}^2 = 1.4286 \cdot \mathrm{median}_{i,j}\{|r_{ij}|\}$ for CPA-RMR.

- Degrees of freedom of the model $\mathrm{df}(\lambda)$:
sum of the number of non-zero elements in factor matrices $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$.

CP Alternating Ridge M-Regression (CPA-RMR)

- Objective function:

$$\sum_{i=1}^{m} \left\{ \hat{\sigma}_i^2 \sum_{j=1}^{n} \rho\left(\frac{x_{ij} - \mathbf{z}_j^\top \mathbf{a}_i}{\hat{\sigma}_i} \right) + \lambda_1 \|\mathbf{a}_i\|_2^2 \right\} + \lambda_2 \|\mathbf{B}\|_2^2 + \lambda_3 \|\mathbf{C}\|_2^2.$$

The minimum $\hat{\mathbf{A}} = (\hat{\mathbf{a}}_1 \ \cdots \ \hat{\mathbf{a}}_m)^\top$ can be obtained by

$$\hat{\mathbf{a}}_i = \min_{\mathbf{a}} \left\{ \hat{\sigma}_i^2 \sum_{j=1}^{n} \rho\left(\frac{x_{ij} - \mathbf{z}_j^\top \mathbf{a}}{\hat{\sigma}_i} \right) + \lambda_1 \|\mathbf{a}\|_2^2 \right\} \qquad (10)$$

for $i = 1, \ldots, m$, while $\mathbf{B}$ and $\mathbf{C}$ are fixed.

- In order to keep the results invariant, we center $\mathbf{x}_i$ and columns of $\mathbf{Z}$ to have median zero.

Simulations

- Model
The observed three-way tensor is generated as $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_0 + \boldsymbol{\mathcal{E}}$, where $\boldsymbol{\mathcal{X}}_0 = \sum_{r=1}^{R} \gamma_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$ is the Kruskal tensor, $\boldsymbol{\mathcal{E}}$ is the noise tensor and the rank R is assumed to be known.
The factor matrices and the true noise-free three-way tensor $\boldsymbol{\mathcal{X}}_0$ is sparse.
- The accuracy of the obtained estimate $\hat{\boldsymbol{\mathcal{X}}}$ can be calculated by the normalized mean squared error (NMSE)

$$\mathrm{NMSE}(\hat{\boldsymbol{\mathcal{X}}}) = \frac{\|\boldsymbol{\mathcal{X}}_0 - \hat{\boldsymbol{\mathcal{X}}}\|_2^2}{\|\boldsymbol{\mathcal{X}}_0\|_2^2}.$$

Simulations - Measure of performance

- 2×2 contingency table

		Estimate of $\mathbf{A}$		
		0	$\neq 0$	sum
True $\mathbf{A}$	0	n_{1C}	n_{1M}	n_1
	$\neq 0$	n_{2M}	n_{2C}	n_2
	sum	n'_1	n'_2	$I \cdot R$

where n_{1C} (resp. n_{2C}) is the number of entries in the estimate $\hat{\mathbf{A}}$ "correctly classified" as being zero (resp. non-zero) and n_{1M} (resp. n_{2M}) is the number of entries in $\hat{\mathbf{A}}$ "misclassified" as being non-zero (resp. zero).

- The *classification error rate*: $\mathrm{CER}(\hat{\mathbf{A}}) = (n_{1M} + n_{2M})/(I \cdot R)$ or *recovery rate*: $\mathrm{RER}(\hat{\mathbf{A}}) = 1 - \mathrm{CER}(\hat{\mathbf{A}})$.

Simulations - Results

- CP-ALS
  - The average NMSE (standard deviations) : 0.0652 (0.0814)
  - The average of confusion matrices for estimating $\mathbf{A} \in \mathbb{R}^{1000\times3}$:

$$\begin{pmatrix} 0 & 1507 \\ 0 & 1493 \end{pmatrix}$$

  - The CP-ALS method does not set any of the elements of $\hat{\mathbf{A}}$ to zero. Both CER and RER are about 50% and 50% indicating that the method serves as a random guess classifier.
- CP alternating LASSO (Our method)
  - The average NMSE : 0.0088 (0.0218) and average confusion matrix

$$\begin{pmatrix} 1290 & 216 \\ 83 & 1411 \end{pmatrix}$$

  - The rate of of correctly selecting zero features: $1290/1507 \approx 85.6\%$ and $\mathrm{RER}(\hat{\mathbf{A}}) = (1290 + 1411)/3000 = 90.3\%$.

Simulations (for Regularization)

- Simulation Setting I : $I = 1000, J = 20$ and $K = 20$, $M = 50$ tensors

  The Kruskal tensor $\mathcal{X}_0$ has rank $R = 3$, and only the factor matrix $\mathbf{A} \in \mathbb{R}^{1000\times3}$ is sparse where A_{ij} is either equal to a zero or an independent random deviate from $N(0, 1)$ with equal probability 1/2.

  The entries of $\mathbf{B} \in \mathbb{R}^{20\times3}$, $\mathbf{C} \in \mathbb{R}^{20\times3}$: independent $\sim N(0, 1)$. The columns of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are normalized to have unit length.

  The noise tensor: $\mathcal{E} \in \mathbb{R}^{1000\times20\times20} \sim N(0, 1)$, independent. $\gamma_1 = 1000, \gamma_2 = 500$ and $\gamma_3 = 500$.

- The *sparsity factor* (SF), the average number (based on M Monte Carlo trials) of zero elements in $\mathcal{X}_0$: $\mathrm{SF} = 12.6\%$
- The *signal to noise ratio (SNR)*, the average value of $\|\mathcal{X}_0\|^2/\|\tilde{\mathcal{E}}\|^2$: $\mathrm{SNR} = 4.2894$ (linear scale).

Simulations - Boxplots I

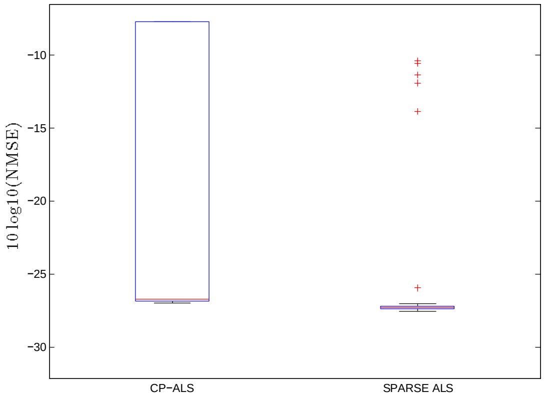


Figure: Boxplots of the $10\log_{10}(\mathrm{NMSE})$ values for simulation Setting I for ALS and CP alternating LASSO (Sparse ALS) methods

Simulations (for Sparsity and Robustness)

- Simulation Setting II:
  The heavy-tailed noise tensor $\boldsymbol{\mathcal{E}} \in \mathbb{R}^{1000\times 20\times 20}$ from the Cauchy distribution with symmetry center 0 and scale parameter 1/2 is added in place of the normal noise tensor to $\boldsymbol{\mathcal{X}}_0$ generated in Simulation setting I.
- The penalty parameter is selected by minimizing the BIC with the weight $w = \sqrt{2}$ over a grid of $\lambda = \lambda_1 = \lambda_2 = \lambda_3$ values for the computational feasibility.

Conclusions

- Multi-linear techniques using tensor decompositions provide a unifying framework for the high-dimensional data analysis.
- Sparsity enables us to extract some essential features from a big data that are easily interpretable.
- Robust (regularized) tensor decompositions clearly improves the analysis and inference of multi-dimensional data.
- We proposed a reliable method to provide good starting values based on the ridge regression/ridge M-regression.
- Combined with such initializations our robust regularization methods show highly improved performance over the conventional methods.

Simulations Results

Table: Simulation results for Cauchy noise

CP Alternating method	Classifying zeros		RER	Average NMSE (std)
	correct	incorrect		
CPA-LS	0	0	50 %	$1.0{\cdot}10^7$ $(7 \cdot 10^7)$
CPA-LASSO	61.2 %	53.2%	54.0%	$1.0 \cdot 10^7$ $(7 \cdot 10^7)$
CPA-RMR	0	0	50 %	0.0487 (0.073)
CPA-LADLASSO	93.40 %	11.7 %	90.9 %	0.0232 (0.044)

Our robust sparse methods, LAD-LASSO and RMR show excellent performance, whereas CP-ALS and the CP alternating LASSO methods yield poor estimates.