Low-rank approximation: from matrices to tensors

Konstantin Usevich, CNRS, CRAN (UMR 7039 / Univ. Lorraine), Nancy

31.05.2024, Marseille, workshop on Low-rank optimization

Overview

Matrix LRA

Tensors

Factorization

(CP) decomposition

CP approximation

Extensions

Overview

Matrix LRA

Tensors

Factorization

(CP) decomposition

CP approximation

Extensions

Matrix rank: factorization view

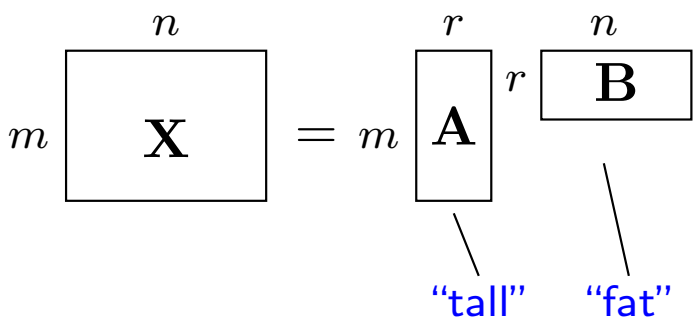
Rank of $\mathbf{X} \in \mathbb{R}^{m \times n}$ (or $\mathbb{C}^{m \times n}$)

- $\stackrel{\text{def}}{=}$ number of linearly independent columns/ rows in $\mathbf{X}$

Matrix rank: factorization view

Rank of $\mathbf{X} \in \mathbb{R}^{m \times n}$ (or $\mathbb{C}^{m \times n}$)

- $\stackrel{\text{def}}{=}$ number of linearly independent columns/ rows in $\mathbf{X}$
- $\stackrel{\text{def}}{=}$ minimal r such that $\mathbf{X}$ can be factorized as

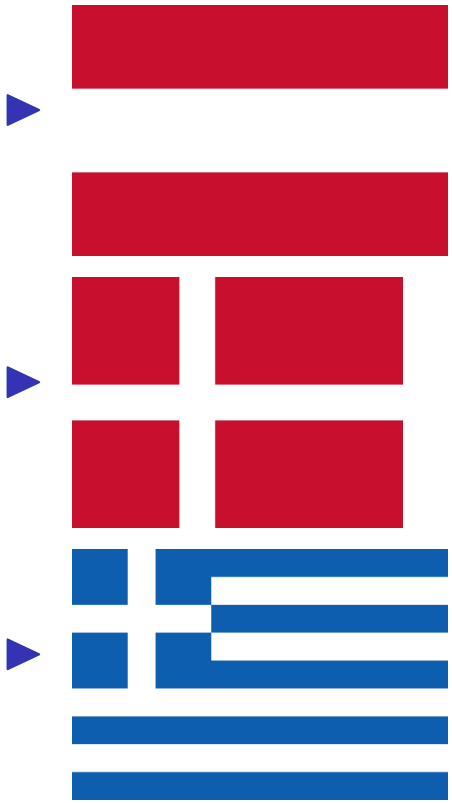


$\Rightarrow$ rank is bounded by dimensions:

$$\operatorname{rank} \mathbf{X} \leq \min(m, n)$$

How do low-rank matrices look like?

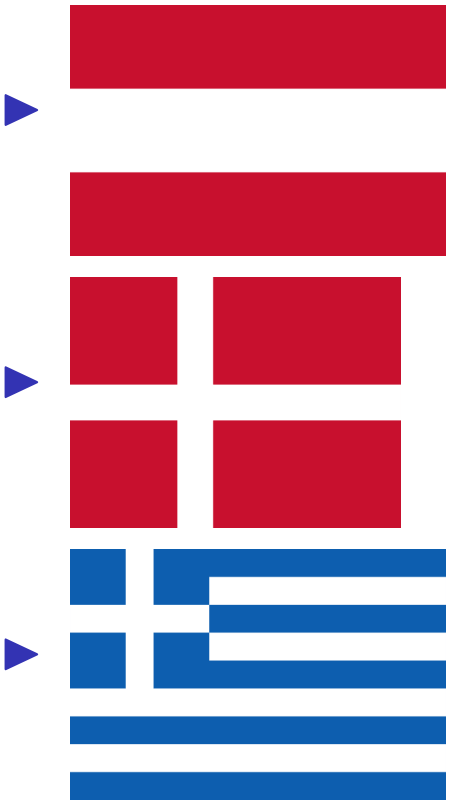
Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$):



Inspired by talks of Alex Townsend

How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$):



(Austria)

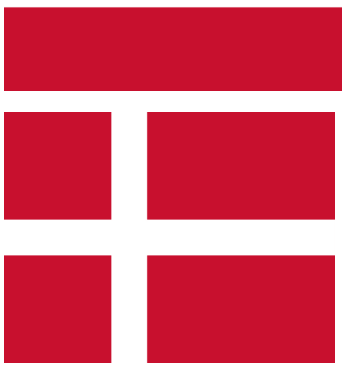
(Denmark)

(Greece)

Inspired by talks of Alex Townsend

How do low-rank matrices look like?

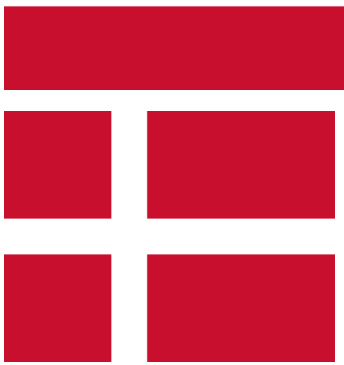
Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$):

-  (Austria) rank $= 1$
-  (Denmark)
-  (Greece)

Inspired by talks of Alex Townsend

How do low-rank matrices look like?

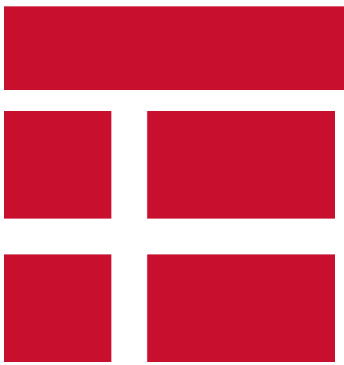
Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$):

-  (Austria) rank $= 1$
-  (Denmark) rank $= 2$
-  (Greece)

Inspired by talks of Alex Townsend

How do low-rank matrices look like?

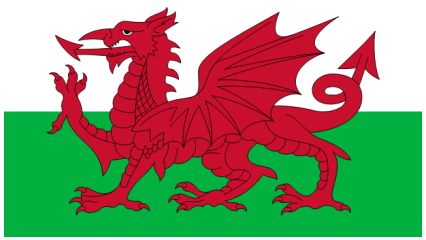
Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$):

-  (Austria) rank $= 1$
-  (Denmark) rank $= 2$
-  (Greece) rank $= 3$

Inspired by talks of Alex Townsend

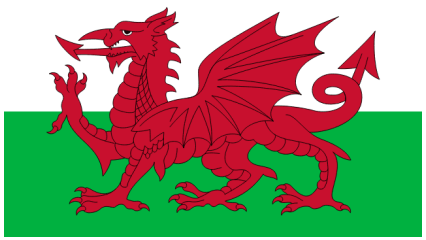
How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$), cont'd:

- 
- 

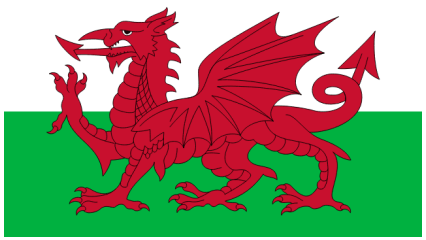
How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$), cont'd:

-  (Scotland)
-  (Wales)

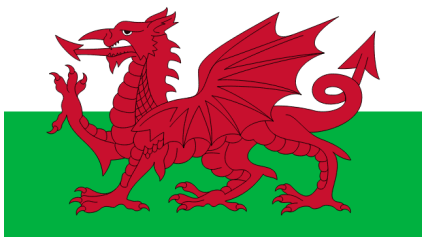
How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$), cont'd:

-  (Scotland) rank $\approx \frac{m}{2}$ (symmetry)
-  (Wales)

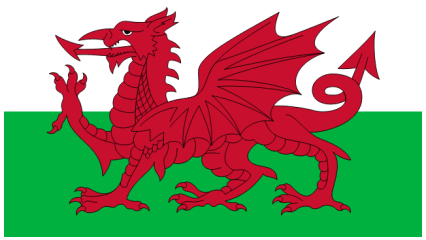
How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$), cont'd:

-  (Scotland) rank $\approx \frac{m}{2}$ (symmetry)
-  (Wales) rank $\approx \frac{5}{6}m$ (finite support)

How do low-rank matrices look like?

Ranks of flags (as $\mathbf{X} \in \mathbb{R}^{m \times n}$), cont'd:

-  (Scotland) rank $\approx \frac{m}{2}$ (symmetry)
-  (Wales) rank $\approx \frac{5}{6}m$ (finite support)

- random matrix (with a.c. probability distribution): rank $= \min(m, n)$ a.s. (with probability 1)
- many interesting matrices are [well approximated by] low-rank [Townsend, Udell, 2017]

Singular value decomposition (SVD)

a/the ("economy size") SVD ($m \leq n$):

$$\underset{m}{\overset{n}{\boxed{\boldsymbol{X}}}} = \boxed{\boldsymbol{U}} \boxed{\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m\end{matrix}} \boxed{\boldsymbol{V}^T} = \sum_{k=1}^{m} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

where

- $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ — semi-orthogonal matrices of singular vectors
  $\boldsymbol{U} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_m\end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$
- $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$ — singular values

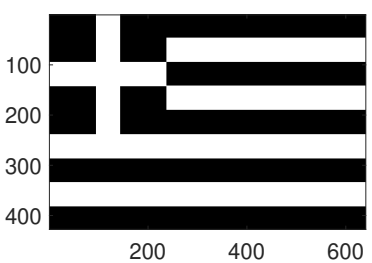
Singular value decomposition (SVD)

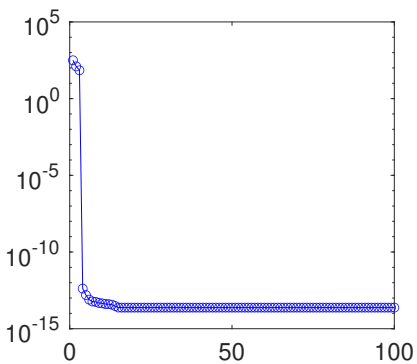
a/the ("economy size") SVD ($m \leq n$):

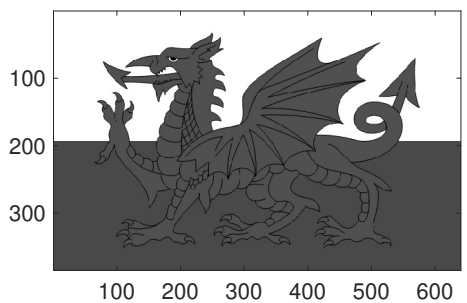
$$\underset{m}{\overset{n}{\boxed{\boldsymbol{X}}}} = \boxed{\boldsymbol{U}} \boxed{\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m\end{matrix}} \boxed{\boldsymbol{V}^T} = \sum_{k=1}^{m} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

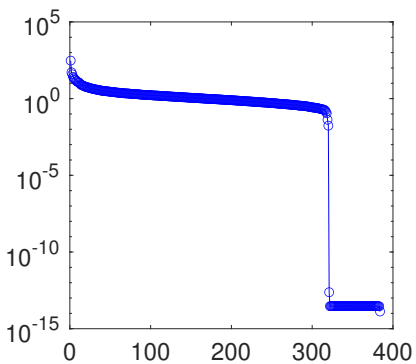
where

- $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ — semi-orthogonal matrices of singular vectors
  $\boldsymbol{U} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_m\end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_m\end{bmatrix}$
- $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$ — singular values









SVD and low-rank approximation

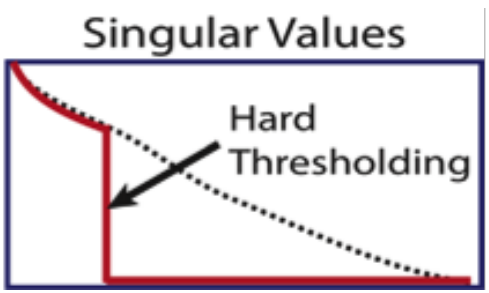
Eckart-Young(-Mirsky(-Schmidt)) theorem: best rank-r approximation

$$\min_{\widehat{\mathbf{X}}} \|\widehat{\mathbf{X}} - \mathbf{X}\| \quad \text{subject to} \quad \operatorname{rank} \widehat{\mathbf{X}} \leq r$$

in any unitarily invariant* norm $\|\cdot\|$ is given by

$$\text{tSVD}_r(\boldsymbol{X}) := \boxed{\boldsymbol{U}} \boxed{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{matrix} \quad \mathbf{0}} \boxed{\boldsymbol{V}^T}$$

truncated SVD



* Examples: Frobenius norm $\|\mathbf{X}\|_F^2 := \sum_{i,j=1}^{m,n} X_{ij}^2$, spectral norm...

SVD and low-rank approximation

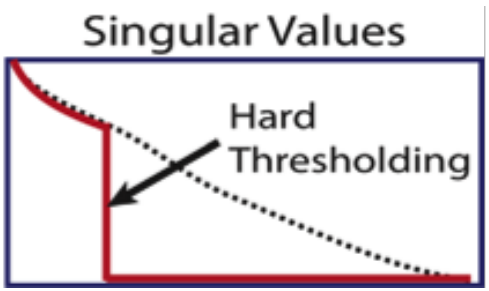
Eckart-Young(-Mirsky(-Schmidt)) theorem: best rank-r approximation

$$\min_{\widehat{\mathbf{X}}} \|\widehat{\mathbf{X}} - \mathbf{X}\| \quad \text{subject to} \quad \operatorname{rank} \widehat{\mathbf{X}} \leq r$$

in any unitarily invariant* norm $\|\cdot\|$ is given by

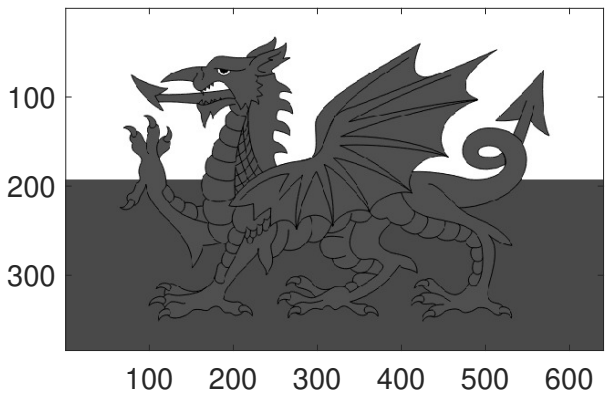
$$\text{tSVD}_r(\boldsymbol{X}) := \boxed{\boldsymbol{U}} \boxed{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{matrix} \quad \mathbf{0}} \boxed{\boldsymbol{V}^T} = \boxed{\boldsymbol{U}_{1:r,:}} \boxed{\begin{matrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{matrix}} \boxed{\boldsymbol{V}^T_{1:r,:}} = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

truncated SVD

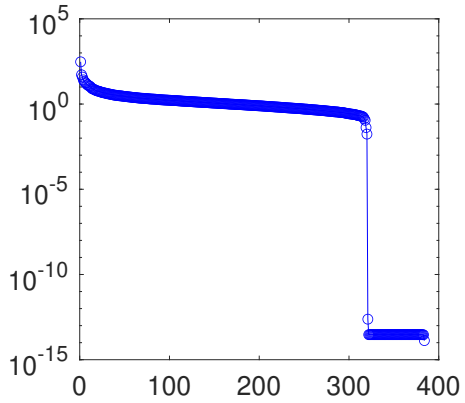


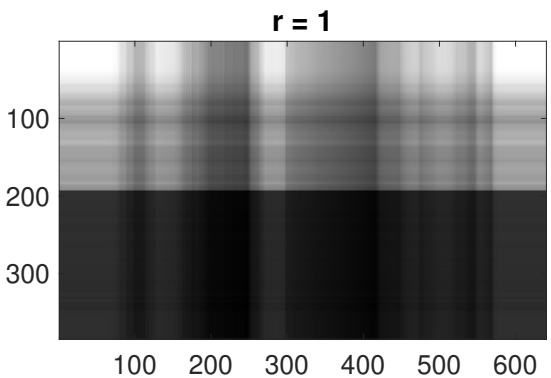
* Examples: Frobenius norm $\|\mathbf{X}\|_F^2 := \sum_{i,j=1}^{m,n} X_{ij}^2$, spectral norm...

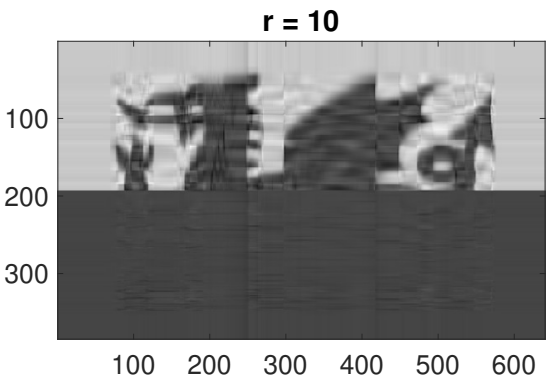
Low-rank approximations: example

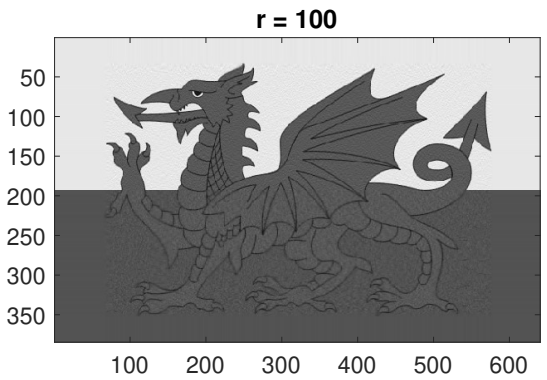


$$\approx \boxed{\boldsymbol{U}_{1:r,:}} \boxed{\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{matrix}} \boxed{\boldsymbol{V}^T_{1:r,:}}$$









Computational aspects: full and partial SVD

$$\underset{m}{\overset{n}{\boxed{\boldsymbol{X}}}} = \boxed{\boldsymbol{U}} \boxed{\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m\end{matrix}} \boxed{\boldsymbol{V}^T} = \sum_{k=1}^{m} \sigma_k \mathbf{u}_k \mathbf{v}_k^T$$

- Full SVD ($m \leq n$): $\mathcal{O}(m^2 n)$ ([Golub, Reinsh, 1970])
- $(\mathbf{u}_k, \sigma_k^2)$ — eigenvalues of $\mathbf{C} = \mathbf{X}\mathbf{X}^T$: $\mathcal{O}(m^3)$ if $m \ll n$

- Truncated SVD (r first eigenvalues/vectors):
  - Iterative (e.g., Lanzos/Arnoldi) algorithms (e.g., [Simon, 1984]) "generalization of power method"
  - Randomized SVD [Halko, Martinsson, Tropp, 2011]

  Both roughly $\boxed{\mathcal{O}(Mr)}$, where $M \leq mn$ (see next slide):

Power iteration

Goal: find the top eigenpair $\mathbf{u}_1, \lambda_1$ of $\mathbf{C}$.

- Set $\mathbf{u}^{(0)} \in \mathbb{R}^{m \times m}$ random.
- Iterate $\mathbf{u}^{(k+1)} = \frac{\mathbf{C}\mathbf{u}^{(k)}}{\|\mathbf{C}\mathbf{u}^{(k)}\|}$

Case $\mathbf{C} = \mathbf{X}\mathbf{X}^T$: cost $\boxed{\approx \mathcal{O}(Mn_{iter})}$, where

- full matrices: $M = mn$
- sparse matrices: $M = \#$non-zero entries
- structured matrices: (e.g., $M = n\log(n)$ for Hankel)

$M =$ cost of matrix-vector product (e.g., $\mathbf{X}\mathbf{v}$)

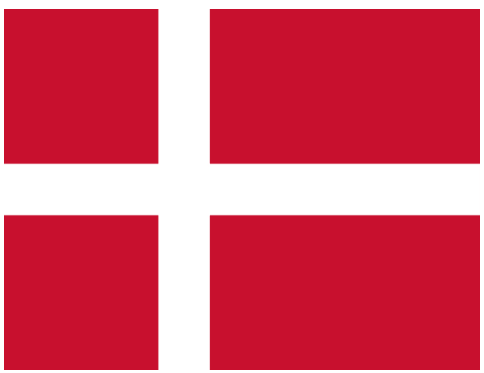
- used e.g., by Google for PageRank
- do not need to store the matrix $\mathbf{X}$
- generalizes to rank-r approximation (cost $\mathcal{O}(Mr + mr^2)$)

Cross approximation

Observation. A rank-r matrix is uniquely determined by the $r \times r$ cross if the $r \times r$ submatrix $\mathbf{X}_{\mathcal{I},\mathcal{J}}$ is nonsingular

$$\mathbf{X} = \mathbf{X}_{:,\mathcal{J}}(\mathbf{X}_{\mathcal{I},\mathcal{J}})^{-1}\mathbf{X}_{\mathcal{I},:} \qquad (*)$$

Example. $r = 2$:

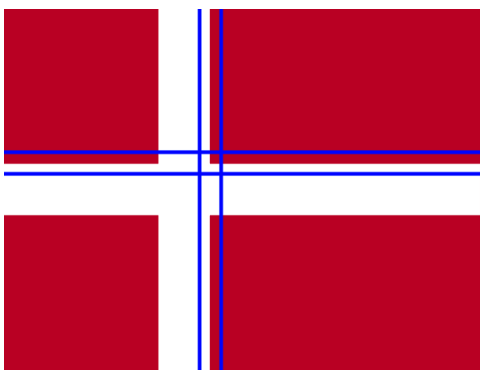


Cross approximation

Observation. A rank-r matrix is uniquely determined by the $r \times r$ cross if the $r \times r$ submatrix $\mathbf{X}_{\mathcal{I},\mathcal{J}}$ is nonsingular

$$\mathbf{X} = \mathbf{X}_{:,\mathcal{J}}(\mathbf{X}_{\mathcal{I},\mathcal{J}})^{-1}\mathbf{X}_{\mathcal{I},:} \qquad (*)$$

Example. $r = 2$:

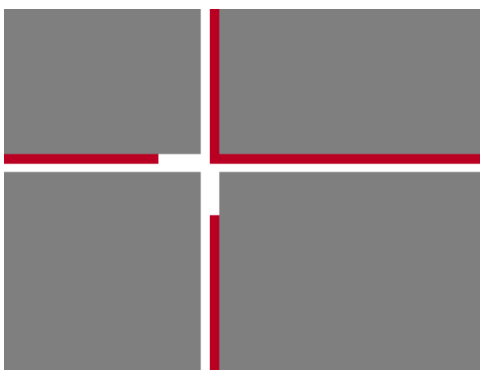


Cross approximation

Observation. A rank-r matrix is uniquely determined by the $r \times r$ cross if the $r \times r$ submatrix $\mathbf{X}_{\mathcal{I},\mathcal{J}}$ is nonsingular

$$\mathbf{X} = \mathbf{X}_{:,\mathcal{J}}(\mathbf{X}_{\mathcal{I},\mathcal{J}})^{-1}\mathbf{X}_{\mathcal{I},:} \qquad (*)$$

Example. $r = 2$:

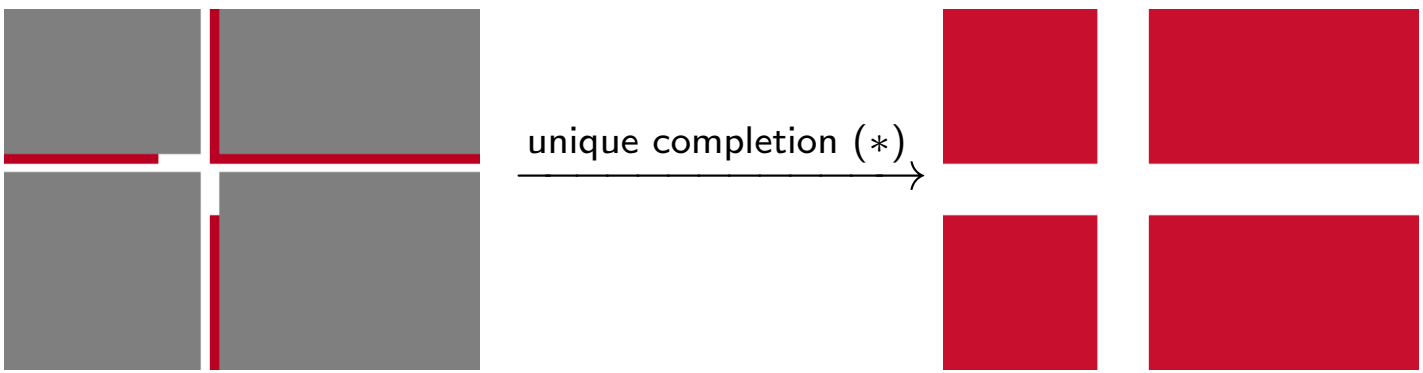


Cross approximation

Observation. A rank-r matrix is uniquely determined by the $r \times r$ cross if the $r \times r$ submatrix $\mathbf{X}_{\mathcal{I},\mathcal{J}}$ is nonsingular

$$\mathbf{X} = \mathbf{X}_{:,\mathcal{J}}(\mathbf{X}_{\mathcal{I},\mathcal{J}})^{-1}\mathbf{X}_{\mathcal{I},:} \qquad (*)$$

Example. $r = 2$:



$\Rightarrow$ $(*)$ can be used as an approximation

Cross approximation

If the matrix is huge or expensive to compute:
CUR (cross, or pseudo-skeleton) approximation (for size-r subsets $\mathcal{I}, \mathcal{J}$):

$$\widehat{\mathbf{X}}_{cross}(\mathcal{I}, \mathcal{J}) = \mathbf{X}_{:,\mathcal{J}}(\mathbf{X}_{\mathcal{I},\mathcal{J}})^{-1}\mathbf{X}_{\mathcal{I},:}$$

[Mahoney, Drineas, 2012]
Advantages:

- Need a small portion (cross) of the matrix ($O(r(m+n))$)
- Quasi-optimality (thm. in [Goreinov, Tyrtyshnikov, 2001])

$$\|\widehat{\mathbf{X}}^*_{cross} - \mathbf{X}\|_{max} \leq (r+1)\sigma_r(\mathbf{X})$$

  for $|\det(\mathbf{X}_{\mathcal{I},\mathcal{J}})| \to \max$
- iterative or randomized strategies to select $\mathcal{I}, \mathcal{J}$

Extensions of the basic problem

$$\min_{\substack{\widehat{\mathbf{X}}=\mathbf{AB} \\ \mathbf{A}\in\mathbb{R}^{m\times r},\mathbf{B}\in\mathbb{R}^{r\times n}}} \|\mathbf{X}-\widehat{\mathbf{X}}\|_F$$

- Other norms (e.g. $\|\mathbf{X}\|_W^2 = \sum_{i,j} W_{ij}X_{ij}^2$), missing data
- Constraint on the matrix:
  structured $\widehat{\mathbf{X}}$ — structured low-rank approximation
- Constraints on the factors $\mathbf{A}$, $\mathbf{B}$ (e.g., nonnegative)

Weighted (unstructured) LRA

$$\min_{\widehat{\mathbf{X}}} \|\mathbf{X} - \widehat{\mathbf{X}}\|_W^2 \quad \text{subject to} \quad \operatorname{rank}(\widehat{\mathbf{X}}) \leq r$$

where $\|\mathbf{X}\|_W^2 = \sum_{i,j=1}^{m,n} \mathbf{X}_{ij}^2 W_{ij}, \quad W_{ij} \in (0; +\infty)$ weighted norm

- $W_{ij} \equiv 1$ (or $\operatorname{rank} W = 1$) $\rightarrow$ solution by SVD
- In general case, no closed form solution:
  Gillis, Glineur, **Low-Rank Matrix Approximation with Weights or Missing Data is NP-hard**, *SIMAX*, 2011.

Extended semi-norms and matrix completion

$$\min_{\widehat{\mathbf{X}}} \|\mathbf{X} - \widehat{\mathbf{X}}\|_W^2 \text{ subject to } \quad \operatorname{rank}(\widehat{\mathbf{X}}) \leq r$$

$\|\mathbf{X}\|_W^2 := \sum_{i,j=1}^{m,n} \mathbf{X}_{ij}^2 W_{ij},\ W_{ij} \in [0; +\infty]$ weighted extended semi-norm

- fixed values: $W_{i,} = +\infty \longleftrightarrow$ constraint $\mathbf{X}_{ij} = \widehat{\mathbf{X}}_{ij}$
- missing values: $W_{ij} = 0 \longleftrightarrow \widehat{\mathbf{X}}_{ij}$ is not important

Extended semi-norms and matrix completion

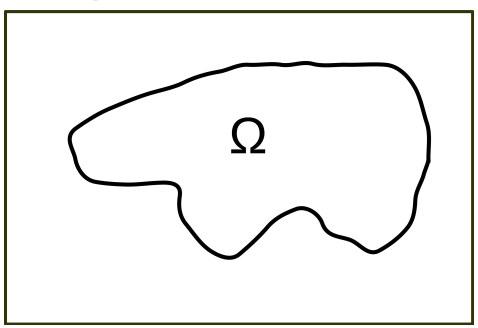
$$\min_{\widehat{\mathbf{X}}} \|\mathbf{X} - \widehat{\mathbf{X}}\|_W^2 \text{ subject to } \quad \operatorname{rank}(\widehat{\mathbf{X}}) \leq r$$

$\|\mathbf{X}\|_W^2 := \sum_{i,j=1}^{m,n} \mathbf{X}_{ij}^2 W_{ij},\ W_{ij} \in [0; +\infty]$ weighted extended semi-norm

- fixed values: $W_{i,} = +\infty \longleftrightarrow$ constraint $\mathbf{X}_{ij} = \widehat{\mathbf{X}}_{ij}$
- missing values: $W_{ij} = 0 \longleftrightarrow \widehat{\mathbf{X}}_{ij}$ is not important

Extreme case: $W_{ij} \in \{0, +\infty\}$ — exact matrix completion

$$\min \operatorname{rank}(\widehat{\mathbf{X}})$$
$$\text{subject to } (\widehat{X})_{ij} = X_{ij}, \quad \forall (i,j) \in \Omega$$

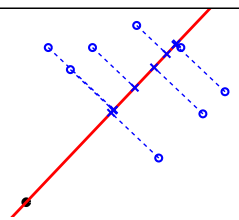


Structured low-rank approximation

[Markovsky, 2008] : **Problem (SLRA).** Given a structured matrix $\mathbf{X} \in \mathcal{S}$

$$\underset{\widehat{X}}{\text{minimize}} \ \|\mathbf{X} - \widehat{\mathbf{X}}\|_W^2 \ \text{ subject to } \ \widehat{\mathbf{X}} \in \mathcal{S} \text{ and } \operatorname{rank} \widehat{\mathbf{X}} \leq r$$

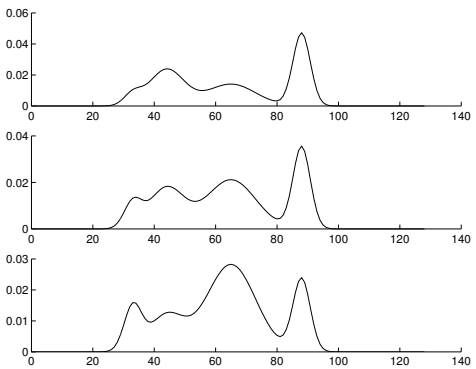
Data $\approx$ low-complexity model

structure $\mathcal{S}$	approximation problem
unstructured	fit by r-dim. subspace 
Hankel	fitting by complex exponentials
block-Hankel	linear system identification model reduction
Sylvester	approx. greatest common divisor
generalized Vandermonde	fit set of points by algebraic hypersurfaces

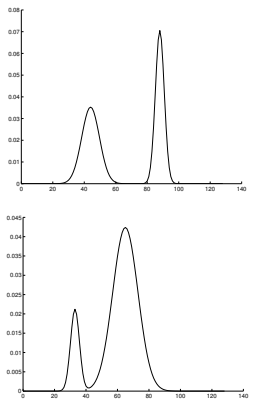
Source separation $\leftrightarrow$ matrix factorization

Example from spectroscopy:

- each observed spectrum is a linear combination of "pure spectra"
- different conditions — different coefficients.



$$= \begin{bmatrix} 2/3 & 1/3 \\ 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix}$$

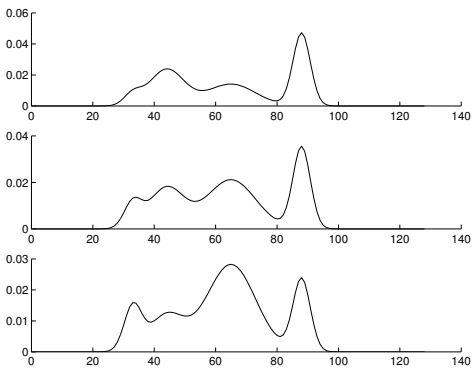


Instantaneous mixture model: $\mathbf{X}(x,\lambda) = \sum_k \mathbf{a}_k(x)\mathbf{s}_k(\lambda)$

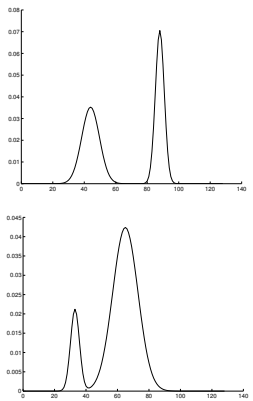
Source separation $\leftrightarrow$ matrix factorization

Example from spectroscopy:

- each observed spectrum is a linear combination of "pure spectra"
- different conditions — different coefficients.



$$= \begin{bmatrix} 2/3 & 1/3 \\ 1/2 & 1/2 \\ 1/3 & 2/3 \end{bmatrix}$$



Instantaneous mixture model: $\mathbf{X}(x, \lambda) = \sum_k \mathbf{a}_k(x)\mathbf{s}_k(\lambda)$

$$\boxed{\mathbf{X}} = \begin{matrix} | \\ \mathbf{a}_1 \\ | \end{matrix} \begin{matrix} |\text{—} \ \mathbf{s}_1^{\mathsf{T}} \ \text{—}| \end{matrix} + \cdots + \begin{matrix} | \\ \mathbf{a}_r \\ | \end{matrix} \begin{matrix} |\text{—} \ \mathbf{s}_r^{\mathsf{T}} \ \text{—}| \end{matrix} = \overset{r}{\boxed{\mathbf{A}}} \ r \ \boxed{\mathbf{S}^T}$$

matrix factorization

Can we recover $\mathbf{A}$, $\mathbf{S}$ from $\mathbf{X}$?

Matrix factorizations are non-unique

Does not happen for matrices:

$$\boxed{\mathbf{M}} = \boxed{\mathbf{A}}\boxed{\mathbf{S}^T} = \boxed{\mathbf{A}}\boxed{\mathbf{Q}}\boxed{\mathbf{Q}^{-1}}\boxed{\mathbf{S}^T}$$

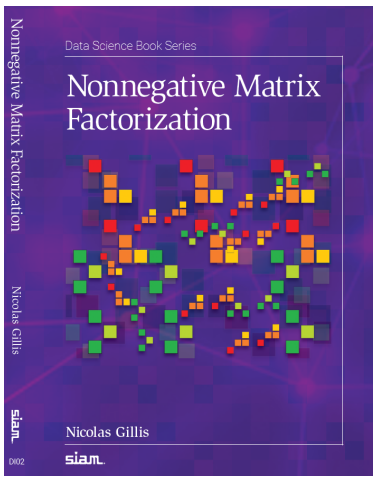
nonunique (change of basis)

However, constraints on factors can guarantee essential uniqueness

$$\mathbf{Q} = \underbrace{\mathbf{\Lambda}}_{\text{diagonal}} \cdot \underbrace{\mathbf{\Pi}}_{\text{permutation}}$$

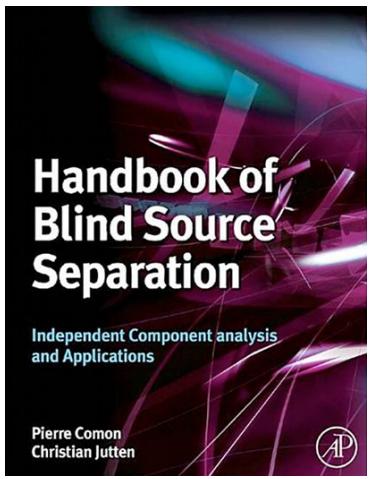
Examples of constraints:

nonnegativity
$\mathbf{A} \geq 0, \mathbf{S} \geq 0$



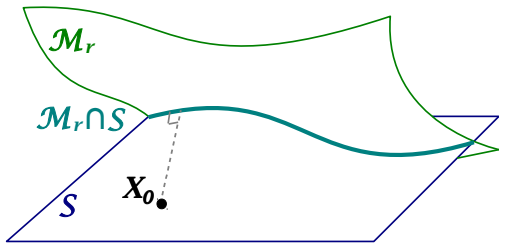
,

independence
(constraint on $\mathbf{S}$)



Other tools/link to optimization

- $\mathcal{M}_r^{m\times n} = \{\mathbf{X} \in \mathbb{R}^{m\times n} | \operatorname{rank}(\mathbf{X}) = r\}$ — smooth manifold



visualisation of SLRA

$\rightarrow$ optimization on manifolds [Absil, Mahoney, Sepulchre, 2008], [Boumal, 2023]

- $\mathcal{M}_{\leq r}^{m\times n} = \{\mathbf{X} | \operatorname{rank}(\mathbf{X}) \leq r\}$ — algebraic variety (stratified set) link to determinantal representations of algebraic varieties
- low rank $\leftrightarrow$ sparsity of singular values

$$\underbrace{\|(\sigma_1, \ldots, \sigma_m)\|_0}_{\text{rank}} \leftrightarrow \underbrace{\|(\sigma_1, \ldots, \sigma_m)\|_1}_{\text{nuclear norm}}$$

other sparsity-promoting penalties ...

- other norms/divergences/losses....
- dynamical low-rank approximation $\mathbf{X}(t) \in \mathcal{M}_r$, varies over time

Matrix factorization $\leftrightarrow$ neural networks

Matrix factorization: $\mathbf{X} = \mathbf{U}\mathbf{V}^T$,$\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_r\end{bmatrix}$, $\mathbf{V} = \begin{bmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_r\end{bmatrix}$
Decompose linear map $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$, $\mathbf{f}(\mathbf{z}) = \mathbf{X}\mathbf{z}$ as

$$\mathbf{f}(\mathbf{z}) = \mathbf{u}_1 \cdot (\mathbf{v}_1^\top \mathbf{z}) + \cdots + \mathbf{u}_r \cdot (\mathbf{v}_r^\top \mathbf{z}),$$

Matrix factorization $\leftrightarrow$ neural networks

Given nonlinear map $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$, decompose it as

$$\mathbf{f}(\mathbf{z}) = \mathbf{u}_1 g_1(\mathbf{v}_1^\top \mathbf{z}) + \cdots + \mathbf{u}_r g_r(\mathbf{v}_r^\top \mathbf{z}),$$

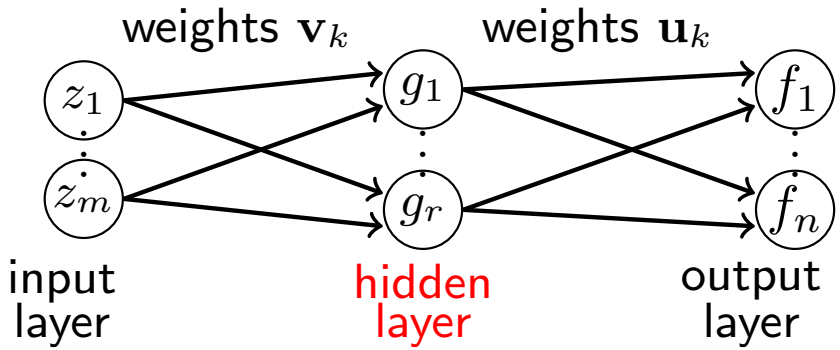
where $g_k(t)$ are univariate functions (see. e.g., [Comon,Qi,U., 2017])

Matrix factorization $\leftrightarrow$ neural networks

Given nonlinear map $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$, decompose it as

$$\mathbf{f}(\mathbf{z}) = \mathbf{u}_1 g_1(\mathbf{v}_1^\top \mathbf{z}) + \cdots + \mathbf{u}_r g_r(\mathbf{v}_r^\top \mathbf{z}),$$

where $g_k(t)$ are univariate functions (see. e.g., [Comon,Qi,U., 2017])



See for example:

- One-hidden layer model: ([Marcotte, Gribonval, Peyré, 2024])
- deep linear networks (products of matrices): [Malgouyres, 2020]
- deep NMF [Leplat et al, 2024]

Overview

Matrix LRA

Tensors

Factorization

(CP) decomposition

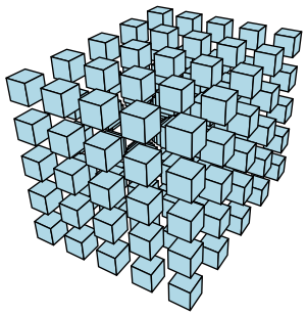
CP approximation

Extensions

Tensors: some notation

Tensor product of vector spaces (over a field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$):

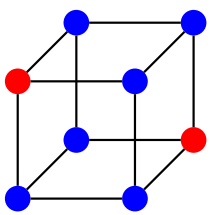
- $\mathcal{T} \in \mathbb{F}^{n_1 \times \cdots \times n_d} \stackrel{\text{def}}{=} \mathbb{F}^{n_1} \otimes \mathbb{F}^{n_2} \otimes \cdots \otimes \mathbb{F}^{n_d}$
- d-way array $\mathcal{T} = \left[\mathcal{T}_{i_1,i_2,\ldots,i_d}\right]_{i_1,i_2,\ldots,i_d=1}^{n_1,n_2,\ldots,n_d} \in \mathbb{F}^{n_1 \times \cdots \times n_d}$
- 3-rd order **tensor**: $\mathcal{T} = \left[\mathcal{T}_{ijk}\right]_{i,j,k=1}^{I,J,K} \in \mathbb{F}^{I \times J \times K}$



- tensor (outer) product: $\mathcal{T} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$: $\quad \mathcal{T}_{ijk} = a_i b_j c_k$

Examples:

- $\mathcal{T} = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right] \otimes \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right] \otimes \left[\begin{smallmatrix}2\\-3\end{smallmatrix}\right] \quad \mathcal{T}_{:,:,1} = \begin{bmatrix}0 & 2\\ 0 & 2\end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix}0 & -3\\ 0 & -3\end{bmatrix},$
- $\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2$ diagonal tensor



$$\mathcal{T}_{:,:,1} = \begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix},$$

Some history

- late 1800s–early 1900s: algebraic geometry (Sylvester, Terracini, Segre, ...)
- 1927 (Hitchcock): introduced tensor rank
- Multiway models in psychometrics: Cattell (1940s), Tucker (1960s), Harshman (1970s)
- Popular models in chemometrics (1980s)
- Theory of complexity: Strassen (1980s)
- Signal processing: Comon (1990s)
- ...

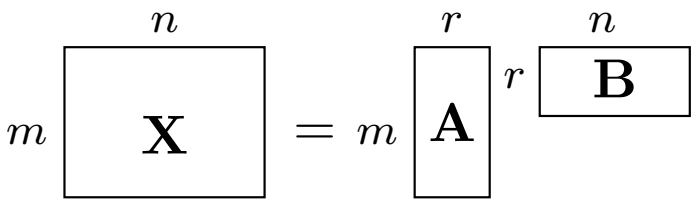
Some references

Modern references (tensor decompositions):

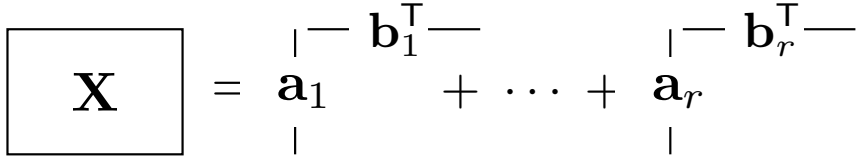
- [Kolda, Bader, 2009]: generic entry reference
- [Comon, 2009, 2014]: focus on CPD and its properties
- [Landsberg, 2012]: algebraic viewpoint
- [Grasedyck et al, 2013]: focus on approximation, scientific computing
- [Sidiropoulos et al, 2017]: more recent overview on uniqueness
- [Cichocki et al, 2016]: book on tensor networks

Reminder: matrix rank

- $\overset{\text{def}}{=}$ smallest r such that $\mathbf{X}$ can be factorized as



- $\overset{\text{def}}{=}$ minimal r such that $\mathbf{X}$ can be decomposed as



Generalization to tensors leads to **two different versions of rank**!

1. Multilinear rank (Tucker) — factorization
2. Tensor rank (CPD) — decomposition

- different decompositions for different purpose
- most other decompositions are combinations of CP and Tucker

Overview

Matrix LRA

Tensors

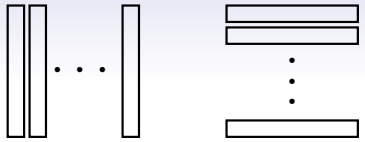
Factorization

(CP) decomposition

CP approximation

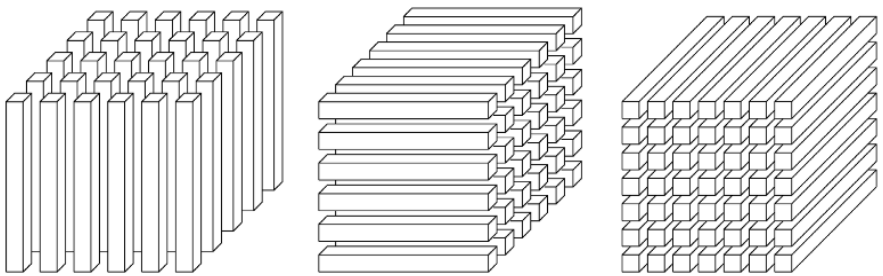
Extensions

- Matrices: columns and rows:



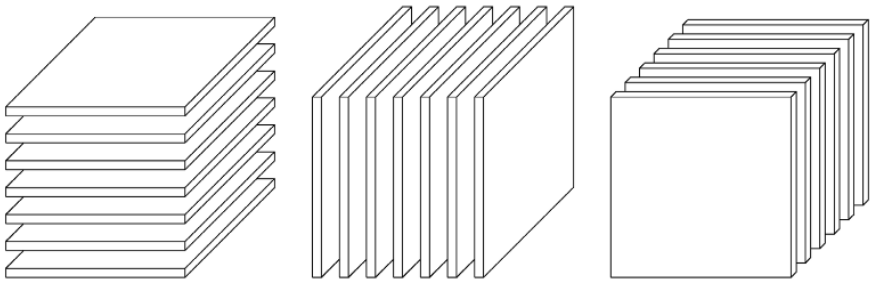
- Tensors: fibers:

columns $\mathcal{T}_{:,j,k}$ rows $\mathcal{T}_{i,:,k}$ tubes $\mathcal{T}_{i,j,:}$

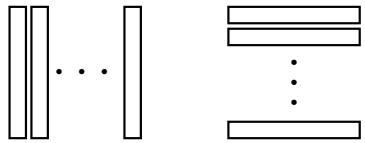


- Tensors: slices:

horizontal $\mathcal{T}_{i,:,:}$ lateral $\mathcal{T}_{:,j,:}$ frontal $\mathcal{T}_{:,:,k}$



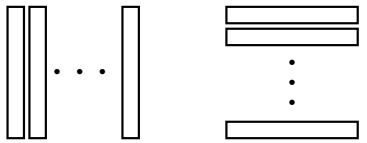
Generalizing matrix rank #1: multilinear rank

Matrix rank $\stackrel{\text{def}}{=}$ dimension of column or row span: 

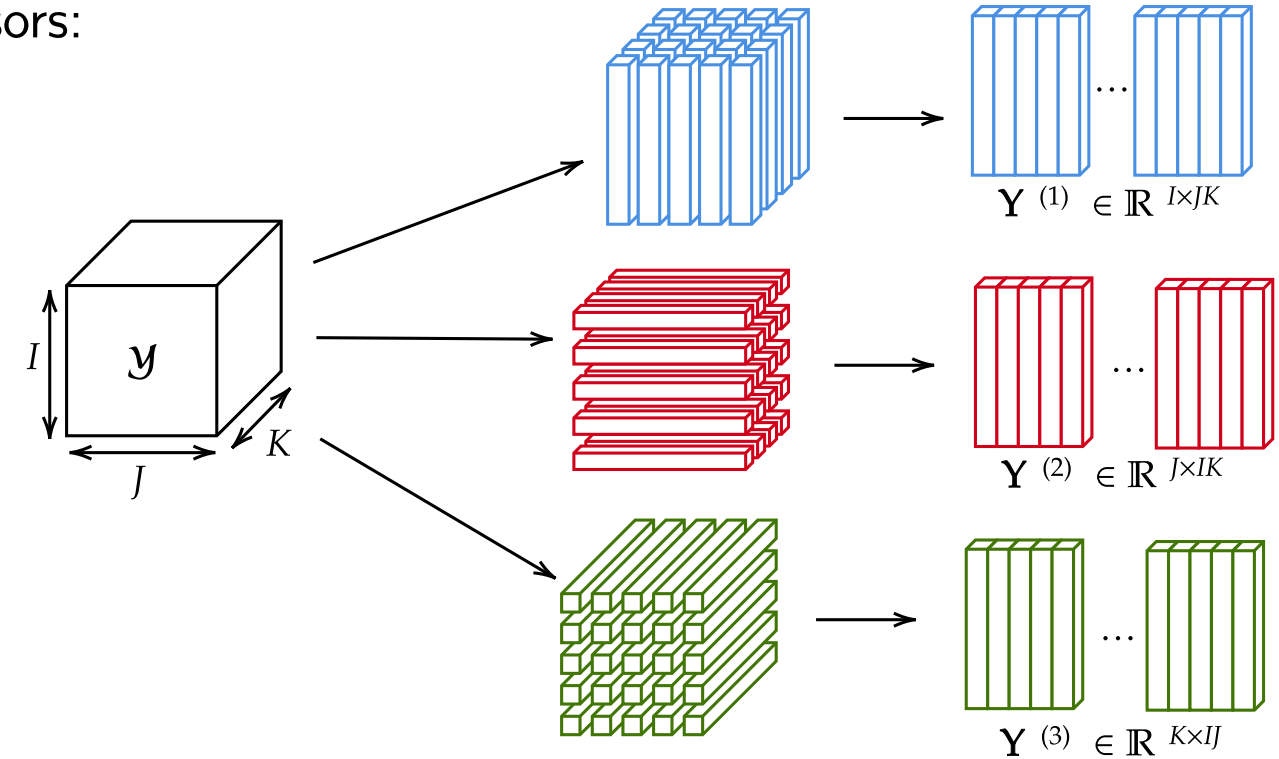
For tensors:

Generalizing matrix rank #1: multilinear rank

Matrix rank $\stackrel{\text{def}}{=}$ dimension of column or row span:



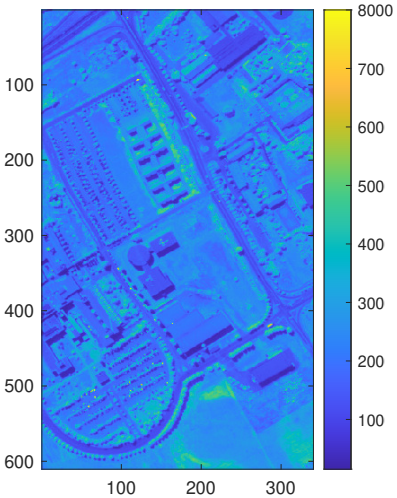
For tensors:

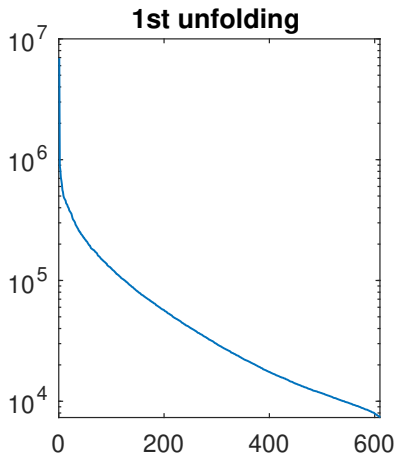


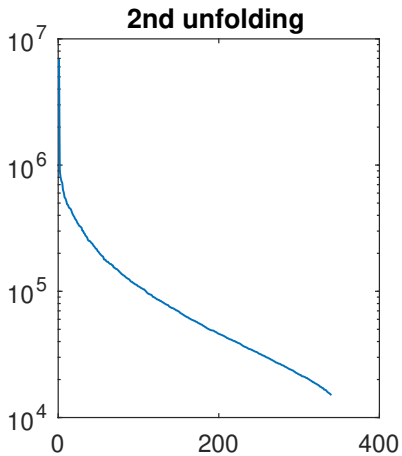
tuple of (different) multilinear ranks $(R_1, \ldots, R_d)$, $\quad R_k = \operatorname{rank}(\mathbf{Y}^{(k)})$

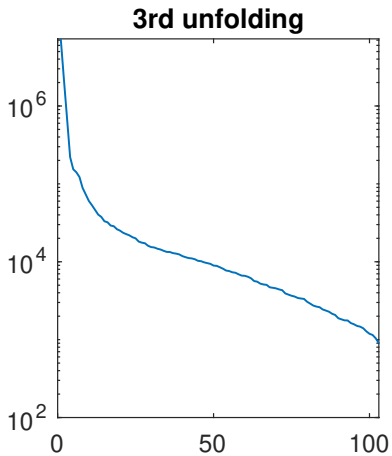
SVDs of the unfoldings

$\underbrace{\mathcal{Y} \in \mathbb{R}^{610 \times 340 \times 103}}_{\text{hyperspectral image}}$, singular values of $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{Y}^{(3)}$:









Pavia University hyperspectral dataset

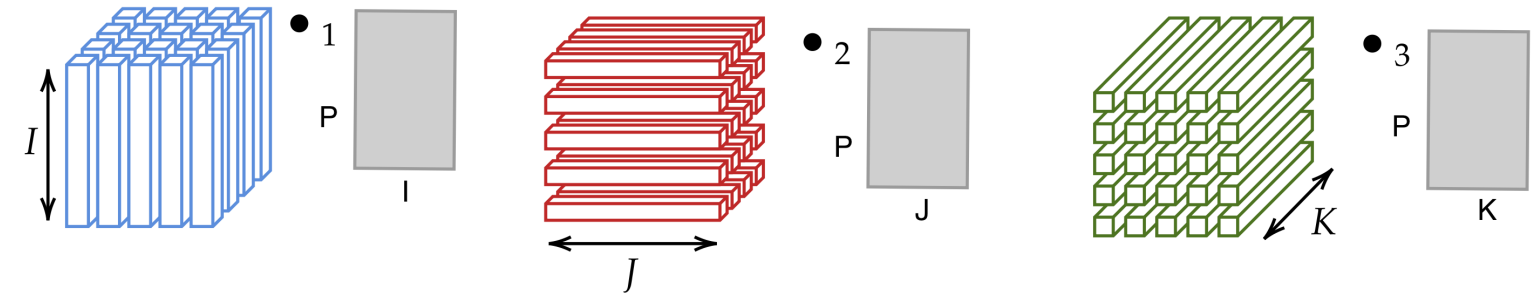
- singular values: not necessarily same distribution
- interconnected and have well-behaved geometry [Hackbusch, Kressner, Uschmajew, 2017],[Krämer, 2019]

Towards factorization: tensor/matrix product

k-th mode contraction :

- with $\mathbf{M} \in \mathbb{F}^{m \times n_k}$: $(\mathcal{Y} \bullet_k \mathbf{M})_{i_1 \dots i_d} \stackrel{\text{def}}{=} \sum_{j=1}^{n_k} \mathcal{Y}_{i_1 \dots i_{k-1} j i_{k+1} \dots i_d} M_{i_k, j}$

For $\mathcal{Y} \in \mathbb{R}^{I \times J \times K}$



For matrices ($d = 2$):

$$\mathbf{Y} \bullet_1 \mathbf{M} = \mathbf{M}\mathbf{Y}, \quad \mathbf{Y} \bullet_2 \mathbf{M} = \mathbf{Y}\mathbf{M}^{\mathsf{T}}$$

For tensors:

$$(\mathcal{Y} \bullet_k \mathbf{M})^{(k)} = \mathbf{M}\mathbf{Y}^{(k)}$$

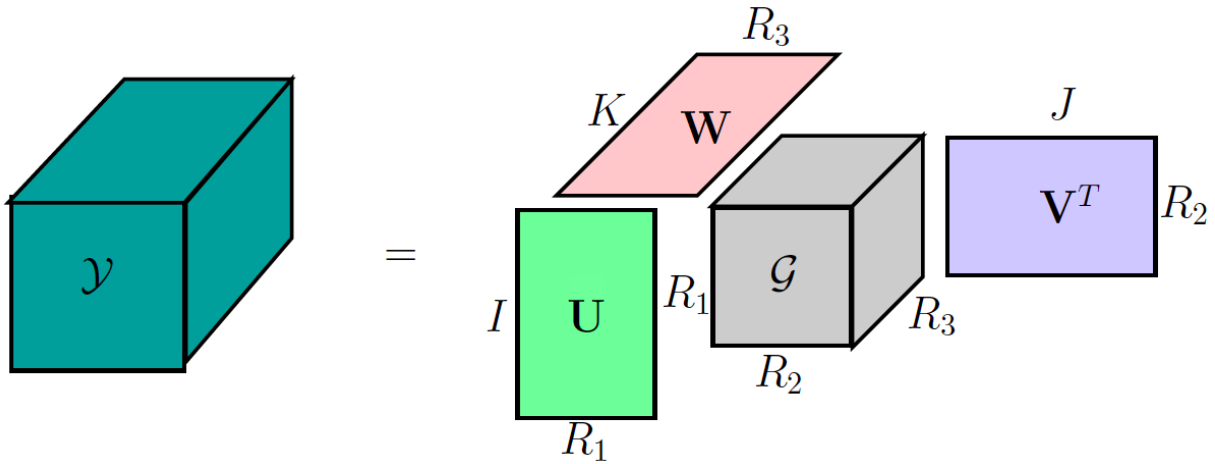
multiplication of k-th unfolding on the left

Multilinear rank and factorization

For $\mathcal{Y} \in \mathbb{R}^{I \times J \times K}$ with ML ranks (R_1, R_2, R_3):

Tucker factorization with factors $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ (*core tensor*), $\mathbf{U} \in \mathbb{R}^{I \times R_1}$, $\mathbf{V} \in \mathbb{R}^{J \times R_2}$, $\mathbf{W} \in \mathbb{R}^{K \times R_3}$

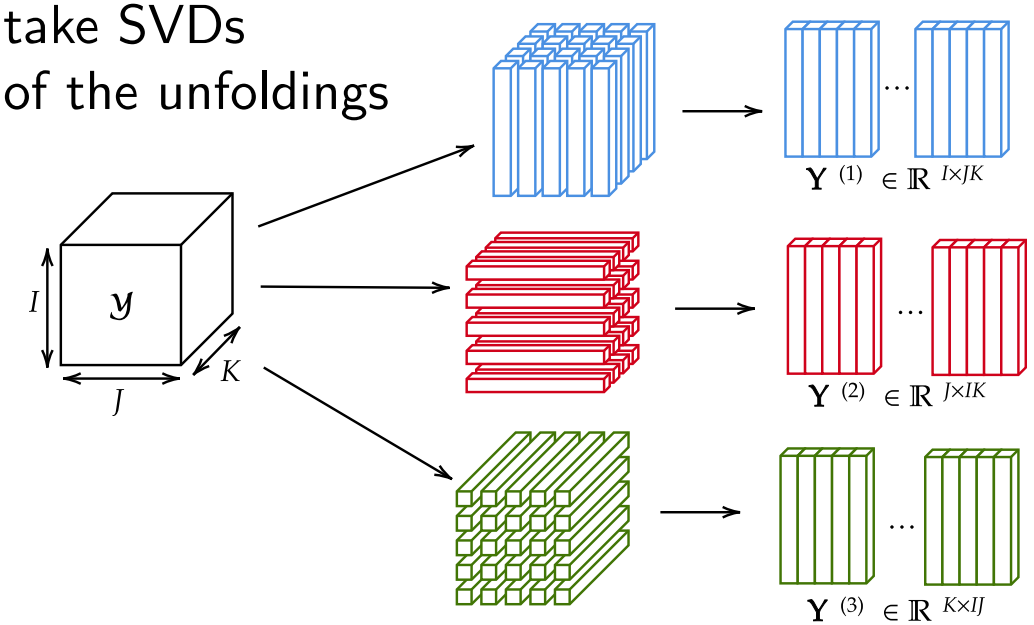
$$\mathcal{Y} = [\![\mathcal{G}; \mathbf{U}, \mathbf{V}, \mathbf{W}]\!] \stackrel{\text{def}}{=} \mathcal{G} \bullet_1 \mathbf{U} \bullet_2 \mathbf{V} \bullet_3 \mathbf{W}.$$



- non-unique (as in the matrix case): $\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{Q}$
- in general (random $\mathcal{Y}$), $(R_1, R_2, R_3) = (\min(I, JK), \min(J, IK), \min(K, IJ))$

Higher-order SVD

- take SVDs of the unfoldings



$$\mathbf{Y}_{(1)} = \mathbf{U}^{(SVD)} \mathbf{\Sigma}_1 [*]$$

$$\mathbf{Y}_{(2)} = \mathbf{V}^{(SVD)} \Sigma_2 [*]$$

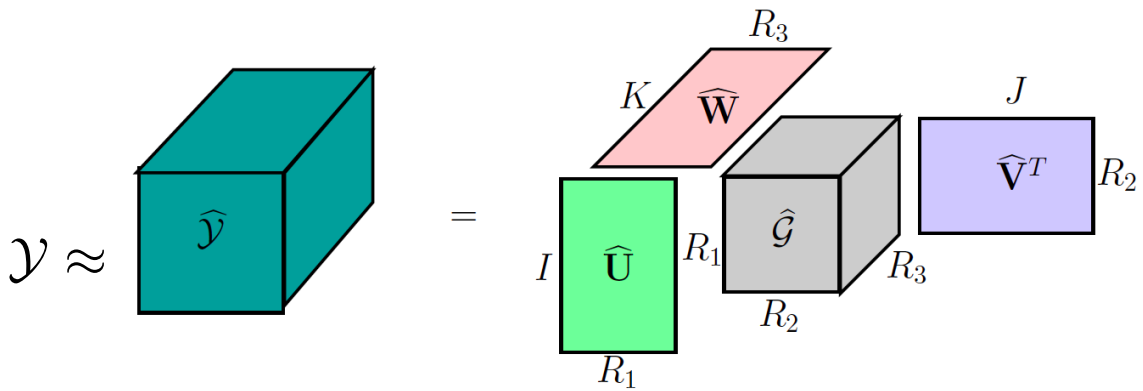
$$\mathbf{Y}_{(3)} = \mathbf{W}^{(SVD)} \Sigma_3 [*]$$

- Compute $\mathcal{G}^{(SVD)} = \mathcal{Y} \bullet_1 (\mathbf{U}^{(SVD)})^\mathsf{T} \bullet_2 (\mathbf{V}^{(SVD)})^\mathsf{T} \bullet_3 (\mathbf{W}^{(SVD)})^\mathsf{T}$

HOSVD $\boxed{\mathcal{Y} = \mathcal{G}^{(SVD)} \bullet_1 \mathbf{U}^{(SVD)} \bullet_2 \mathbf{V}^{(SVD)} \bullet_3 \mathbf{W}^{(SVD)}}$

Best low multilinear approximation

Compute best (R_1, R_2, R_3)-Tucker approximation:



$$\min_{\widehat{\mathcal{G}},\widehat{\mathbf{U}},\widehat{\mathbf{V}},\widehat{\mathbf{W}}} \|\mathcal{Y} - \widehat{\mathcal{G}} \bullet_1 \widehat{\mathbf{U}} \bullet_2 \widehat{\mathbf{V}} \bullet_3 \widehat{\mathbf{W}}\|_F$$

- a good (suboptimal) solution is given by truncating **HOSVD**:

$$\widehat{\mathbf{U}} = \mathbf{U}^{(SVD)}_{:,1:R_1}, \widehat{\mathbf{V}} = \mathbf{V}^{(SVD)}_{:,1:R_2}, \widehat{\mathbf{W}} = \mathbf{W}^{(SVD)}_{:,1:R_3}, \widehat{\mathcal{G}} = \mathcal{G}^{(SVD)}_{1:R_1,1:R_2,1:R_3}$$

- HOOI: alternating minimization over $\widehat{\mathbf{U}}$,$\widehat{\mathbf{V}}$, $\widehat{\mathbf{W}}$
  [De Lathauwer, De Moor, Vandewalle, 2000]
- optimization over the (R_1, R_2, R_3)-rank manifold
  [Kressner, Steinlechner, Vandereycken, 2013], [Kasai, Mishra, 2016]

Tucker approximation: very useful for compression, completion tasks

Overview

Matrix LRA

Tensors

Factorization

(CP) decomposition

CP approximation

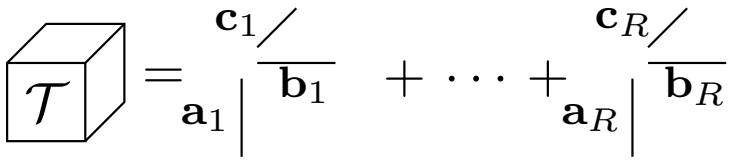
Extensions

Tensor rank and CPD

Rank-1 tensor: $\mathcal{T} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$: $\quad \mathcal{T}_{ijk} = a_i b_j c_k$

(Canonical) **Polyadic Decomposition** — sum of R rank-one tensors:

$$\mathcal{T} = \sum_{k=1}^{R} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k,$$

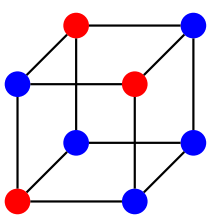


(CP) tensor rank: rank $(\mathcal{T}) \stackrel{\text{def}}{=}$ minimal such R

Earlier names: **C**ANDECOMP/**P**ARAFAC

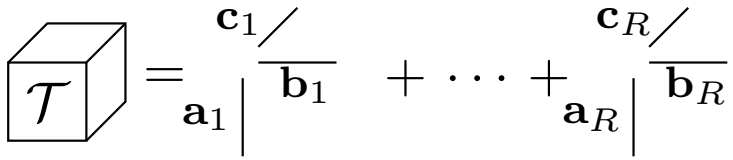
Example:

$$\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$$

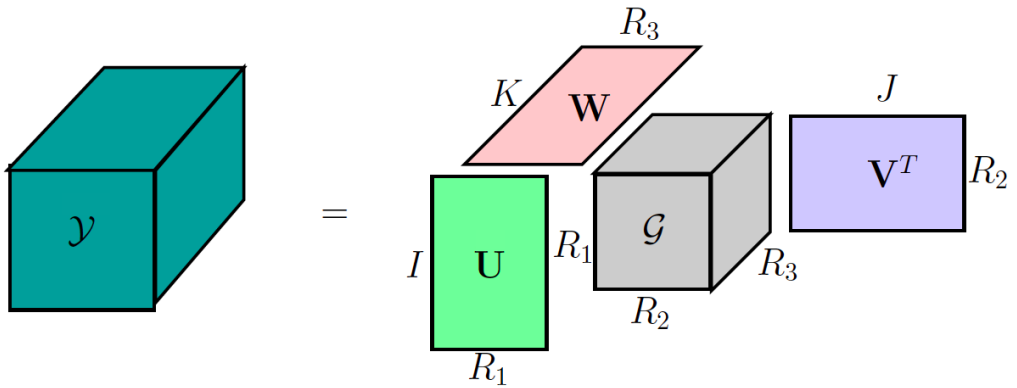


$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Tensor rank: basic properties



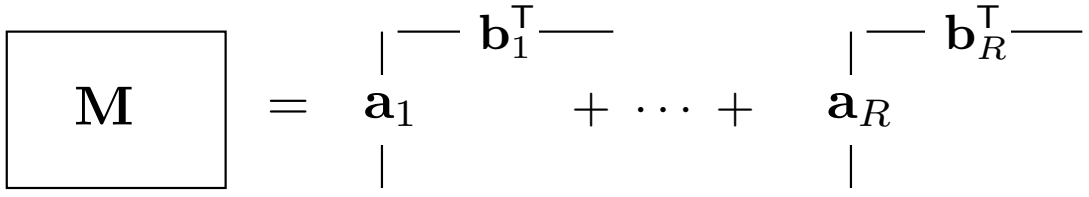
- Relation with multilinear ranks:



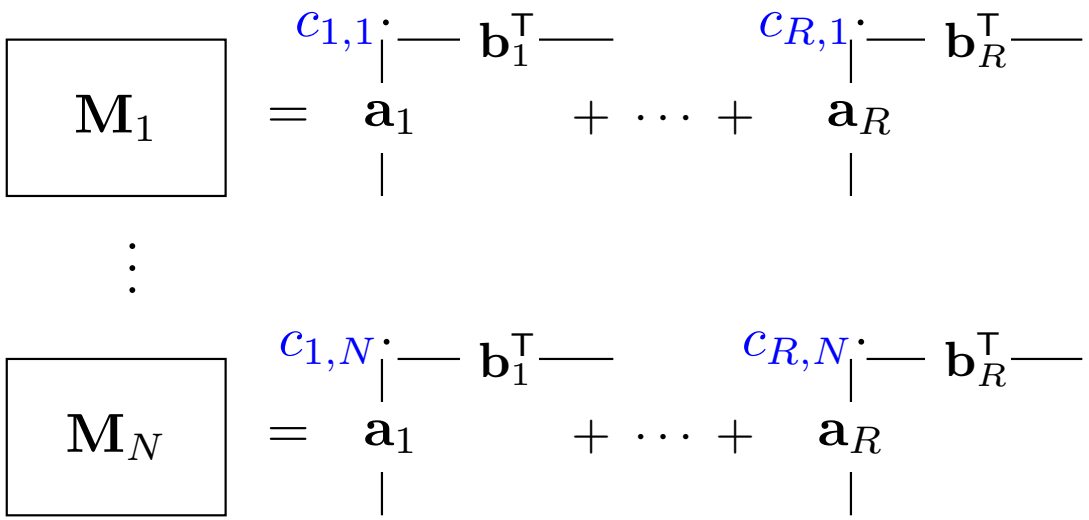
$$\max(R_1, R_2, R_3) \leq R \leq \min(R_2 R_3, R_1 R_3, R_1 R_2)$$

- NP-hard (to compute exact rank): [Hillar, Lim, 2013]
- But has many nice properties and applications

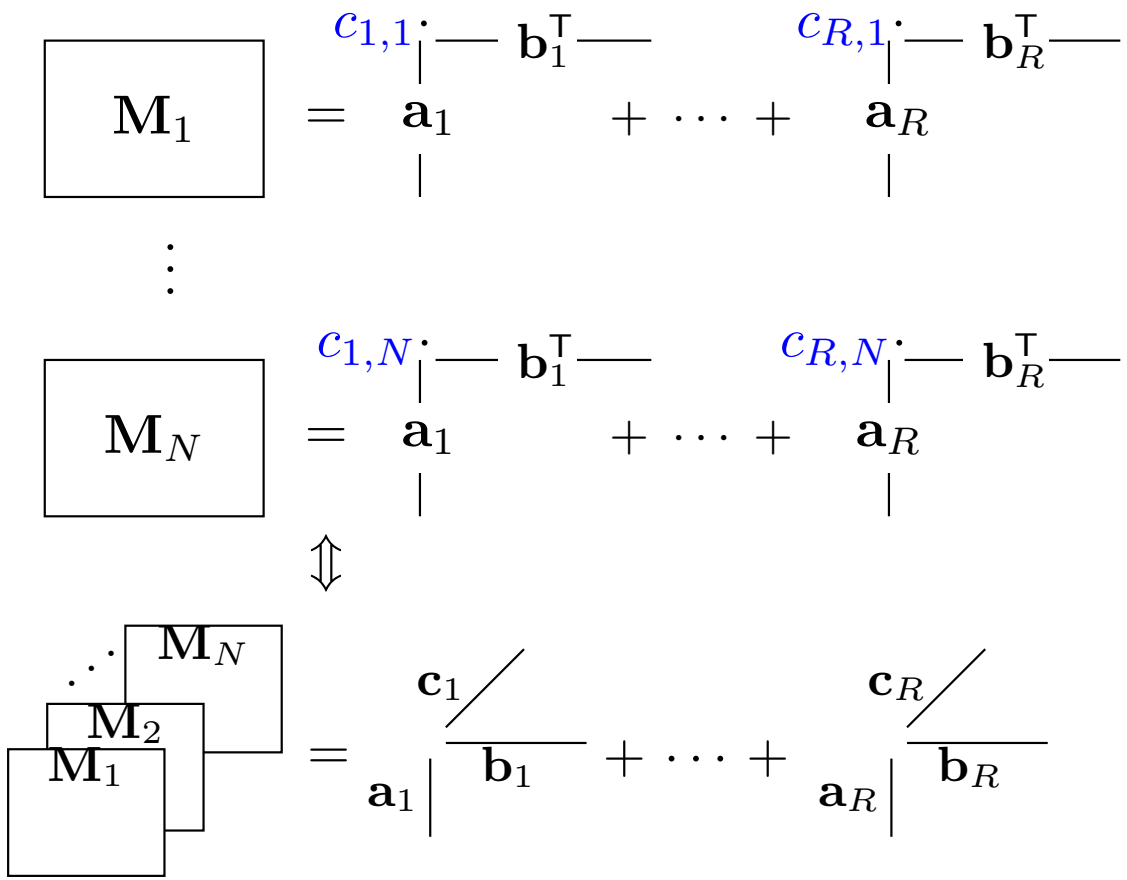
Simultaneous matrix factorization and CPD



Simultaneous matrix factorization and CPD



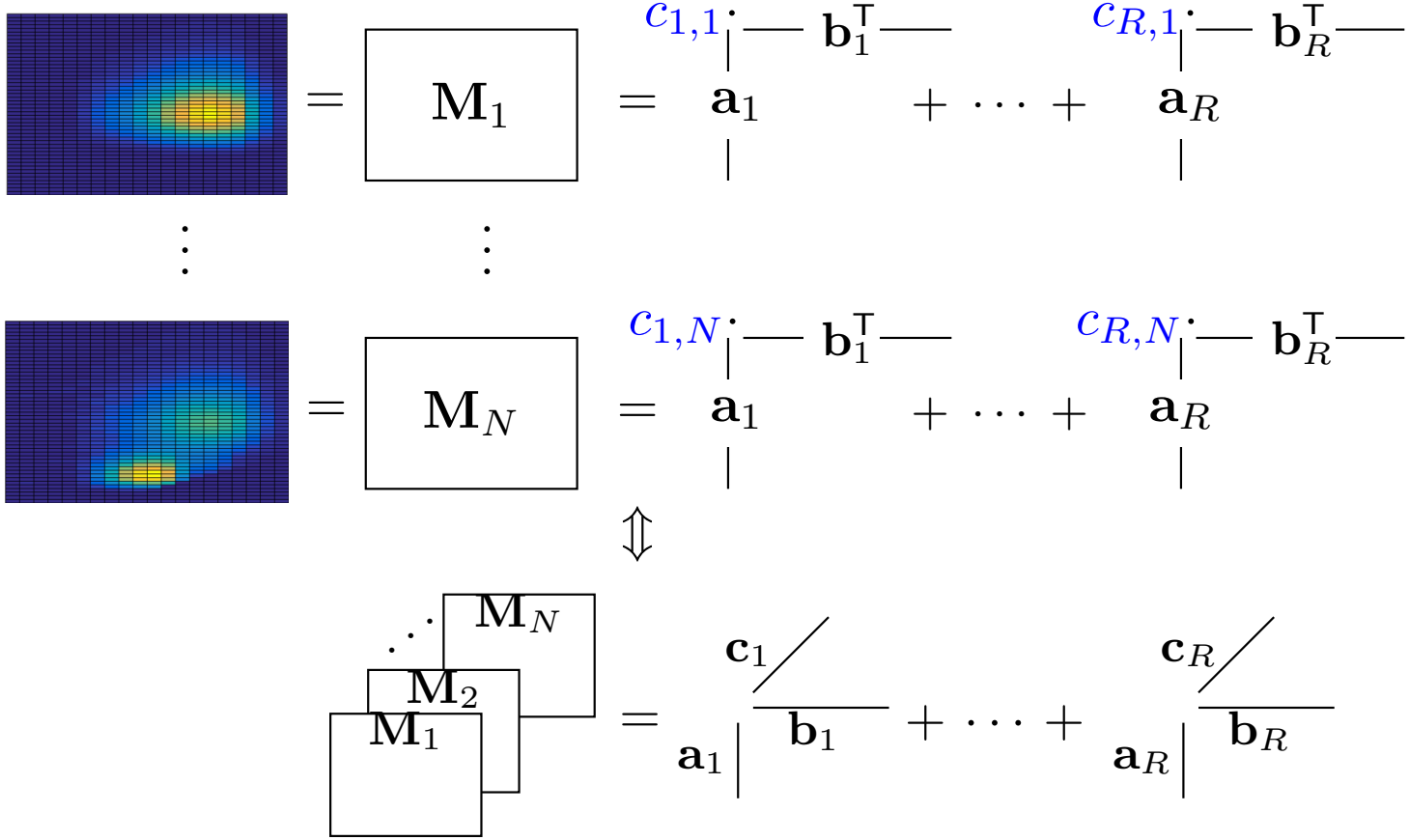
Simultaneous matrix factorization and CPD



Simultaneous matrix factorization and CPD

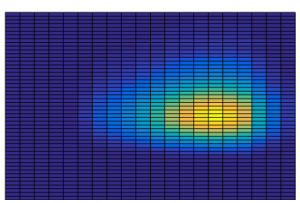
fluorescence spectroscopy
emission/excitation matrix

rank $=$ # components in a mixture
N experiments, different concentrations



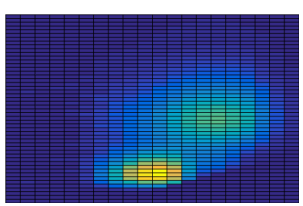
Tensor rank in applications

Application area	tensor rank R
(blind source separation) independent component analysis	# of sources
multiway factor analysis (spectroscopy, chemometrics, ...)	# of components
antenna array processing	# of transmitters
$\vdots$	$\vdots$



$$= \boxed{\mathbf{M}_1} = c_{1,1}\, \mathbf{a}_1 \mathbf{b}_1^{\mathsf{T}} + \cdots + c_{R,1}\, \mathbf{a}_R \mathbf{b}_R^{\mathsf{T}}$$

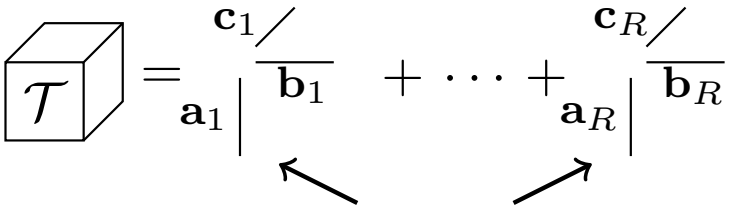
$\vdots \qquad \vdots$



$$= \boxed{\mathbf{M}_N} = c_{1,N}\, \mathbf{a}_1 \mathbf{b}_1^{\mathsf{T}} + \cdots + c_{R,N}\, \mathbf{a}_R \mathbf{b}_R^{\mathsf{T}}$$

Essential uniqueness of a CPD

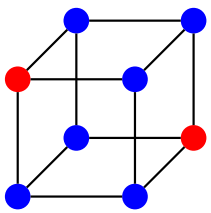
$$\mathcal{T} = \sum_{k=1}^{R} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k,$$



up to permutations and rescaling $(\mathbf{a}'_k = \alpha\mathbf{a}_k, \mathbf{b}'_k = \beta\mathbf{b}_k, \mathbf{c}'_k = \frac{1}{\alpha\beta}\mathbf{c}_k)$

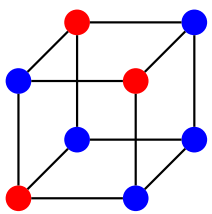
Examples:

- unique: $\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2$



$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

- non-unique: $\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$



$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

Third-order tensors: Kruskal sufficient condition

Shorthand notation: $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] := \sum_{k=1}^{R} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k$

factor matrices $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_R]$, $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_R]$, $\mathbf{C} = [\mathbf{c}_1 \cdots \mathbf{c}_R]$

Theorem *[Kruskal, 1978]*.

The decomposition $\underset{I \times J \times K}{\mathcal{T}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is unique if

$$\mathrm{kr}(\mathbf{A}) + \mathrm{kr}(\mathbf{B}) + \mathrm{kr}(\mathbf{C}) \geq 2r + 2,$$

$\underbrace{\mathrm{kr}(\mathbf{A})}_{\text{Kruskal rank}} \overset{\text{def}}{=}$ maximal k such that any k columns are linearly independent.

Example (my favourite tensor):

- $K = 2$ (2 frontal slices)
- $\mathbf{A}$, $\mathbf{B}$ — full column rank, $\mathbf{C}$ — with non-collinear columns

$2R + 2 = 2R + 2 \Rightarrow$ unique decomposition

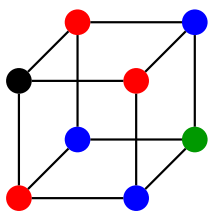
Real vs. complex rank

In general, they are different:

$$\mathsf{rank}_{\mathbb{C}}(\mathcal{T}) \leq \mathsf{rank}_{\mathbb{R}}(\mathcal{T})$$

Example (*[Kruskal, 1983]*, but also *[Sylvester, 1851]*):

$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

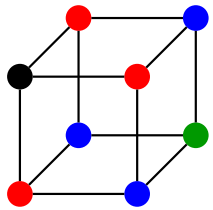


$$\mathcal{T} = \frac{i}{2}\left(\left[\begin{smallmatrix} 1 \\ -i \end{smallmatrix}\right] \otimes \left[\begin{smallmatrix} 1 \\ -i \end{smallmatrix}\right] \otimes \left[\begin{smallmatrix} 1 \\ -i \end{smallmatrix}\right] - \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right] \otimes \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right] \otimes \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]\right)$$

$\Rightarrow \mathsf{rank}_{\mathbb{C}}(\mathcal{T}) = \mathsf{rank}_{S,\mathbb{C}}(\mathcal{T}) = 2$, but $\mathsf{rank}_{\mathbb{R}}(\mathcal{T}) = \mathsf{rank}_{S,\mathbb{R}}(\mathcal{T}) = 3$

Symmetric tensor rank vs. tensor rank

- $S_n^d \stackrel{\text{def}}{=}$ vector space of symmetric tensors,



S_2^3

Symmetric rank: $\text{rank}_S(\mathcal{T}) \stackrel{\text{def}}{=}$ minimal r such that

$$\mathcal{T} = \sum_{k=1}^{r} \lambda_k \mathbf{a}_k \otimes \cdots \otimes \mathbf{a}_k, \qquad \mathbf{a}_k \in \mathbb{F}^n$$

obviously $\text{rank}(\mathcal{T}) \leq \text{rank}_S(\mathcal{T})$

Comon's conjecture: $\boxed{\forall \mathcal{T} \in S_n^d, \quad \text{rank}(\mathcal{T}) = \text{rank}_S(\mathcal{T})}$

- true for
  - matrices $d = 2$
  - ranks smaller than order/dimension *[Friedland, 2017]*, *[Zhang, Huang, Qi, 2017]*
  - generically for small ranks *[Lim, Qi, 2020]*,
- counterexample by *[Shitov, 2017]*: $n = 800$, $d = 3$, $\text{rank}(\mathcal{T}) = 903$

Maximal and generic ranks

Maximal rank

$r_{max} \stackrel{\text{def}}{=}$ maximal possible rank

- different for symmetric/non-symmetric
- different for $\mathbb{R}$ and $\mathbb{C}$

(Very few) known cases

tensor space	dimension	r_{max}	reference
$\mathbb{F}^{n\times n}$	n^2	n	matrices
$\mathbb{C}^{2\times n\times n}$	$2n^2$	$\lfloor \frac{3n}{2} \rfloor$	*[Grigoriev, 1978], [Ja'Ja', 1978]*
$\mathbb{F}^{n\times n\times n}$	n^3	$\leq \frac{(n+1)n}{2}$	*[Atkinson, Stephens, 1979]*

What if we draw tensors randomly?

$\longrightarrow \text{rank}\,(\mathcal{T}) = ?$

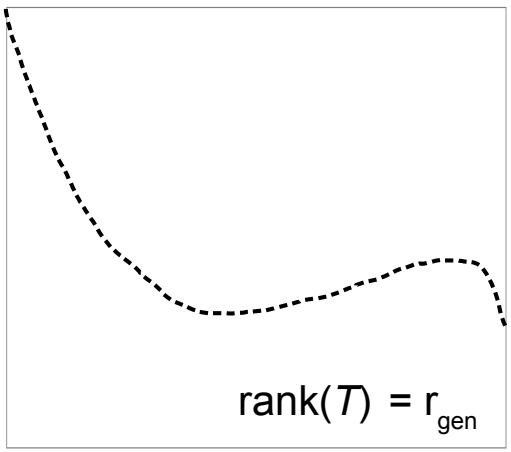
random = from absolutely continuous probability distribution

- important in practice (noise, numerical errors)
- often easier to find (than r_{max})

Complex tensors: generic rank

With probablility 1 a complex tensor has rank r_{gen} (generic rank)

(i.e. other tensors have measure zero)



- matrices ($\mathbb{C}^{n\times n}$): $r_{max} = r_{gen} = n$
- cubic tensors ($\mathbb{C}^{n\times n\times n}$, $n > 3$) [Lickteig, 1985]

$$r_{gen} = \left\lceil \frac{n^3}{3n-2} \right\rceil \approx \frac{n^2}{3}$$

Generic ranks

- symmetric tensors (S^d_n, $d \geq 3$) [Alexander, Hirschowitz, 1996]

$$r_{gen} = \left\lceil \frac{\binom{n+d-1}{n-1}}{n} \right\rceil,$$

  with few exceptions: $(d, n) = (3, 5)$ or $d = 4, n \in \{3, 4, 5\}$

- general case ($\mathbb{C}^{I_1 \times \ldots \times I_d}$, $d \geq 3$):

$$r_{gen} \stackrel{?}{=} \left\lceil \frac{I_1 \cdots I_d}{I_1 + \cdots + I_d - d + 1} \right\rceil, \quad \text{with few exceptions}$$

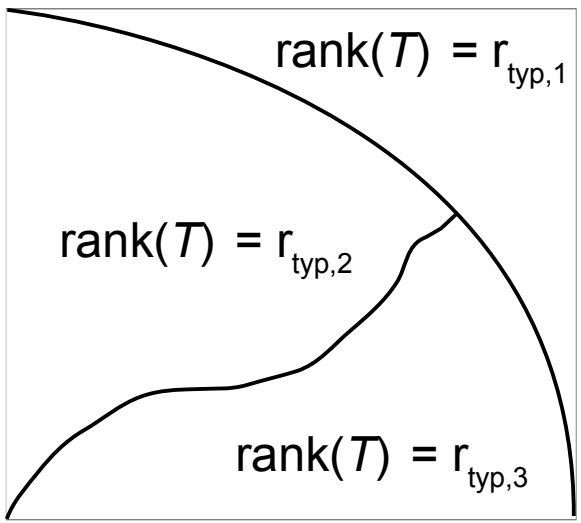
  - conjectured *[Abo, Ottaviani, Peterson, 2009]*
  - computer proof *[Chiantini, Ottaviani, Vannieuwenhoven, 2014]*

- [Blekherman, Teitler, 2014]: $r_{max} \leq 2r_{gen}$

Proofs: algebraic geometry (dimensions of secant varieties)

Real tensors: typical ranks

For real tensors, several typical ranks may appear with nonzero probability.



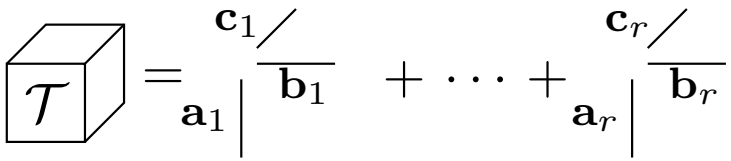
Example *[Bergqvist, 2013]*: $\mathcal{T} \in \mathbb{R}^{2\times 2\times 2}$ with i.i.d. Gaussian elements has:

- rank 2 with probability $\frac{\pi}{4}$
- rank 3 with probability $1 - \frac{\pi}{4}$

Some numerical consequences

1. noise, numerical errors $\Rightarrow$ $\text{rank}(\mathcal{T}) = r_{gen}$ (or a typical rank in $\mathbb{R}$)
2. very difficult to find tensors with higher ranks:

If we generate



with $\mathbf{a}_k$, $\mathbf{b}_k$, $\mathbf{c}_k$ random, has

$$\text{rank}(\mathcal{T}) = \begin{cases} r, & r \leq r_{gen}, \\ r_{gen}, & r_{gen} < r \leq r_{max}. \end{cases}$$

Example. $n \times n \times n$ tensors with rank $\left\lceil \frac{n^3}{3n-2} \right\rceil < r \leq \frac{(n+1)n}{2}$.

Generic uniqueness (identifiability)

For a fixed rank r: whether "almost all" decompositions are unique

- Kruskal-type conditions give weak bounds
- Study the properties secant algebraic variety ($\sigma_r \stackrel{\text{def}}{=}$: (Zariski) closure of tensors of rank r)
- Generic uniqueness: uniqueness for all tensors in σ_r except a set of Lebesgue measure 0

Most recent results:

1. *[Chiantini, Ottaviani, 2012]*: CPD of $\mathcal{T} \in \mathbb{C}^{I \times J \times K}$, with $I \geq J \geq K$ is generically unique if $r \leq 2^{\lfloor \log_2 J \rfloor + \lfloor \log_2 K \rfloor - 2}$.

2. *[Chiantini, Ottaviani, Vannieuwenhoven, 2014]* (computer proof): complex identifiability holds for all subgeneric ranks $r < r_{gen} = \left\lceil \frac{IJK}{I+J+K-2} \right\rceil$
   no identifiability for $r > r_{gen}$

3. *[Qi, Comon, Lim, 2016]*, *[Chiantini, Ottaviani, Vanniueuwenhoven, 2017]*:
   all identifiability results are valid for real-valued tensors

Overview

Matrix LRA

Tensors

Factorization

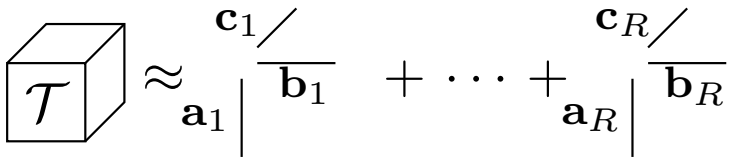
(CP) decomposition

CP approximation

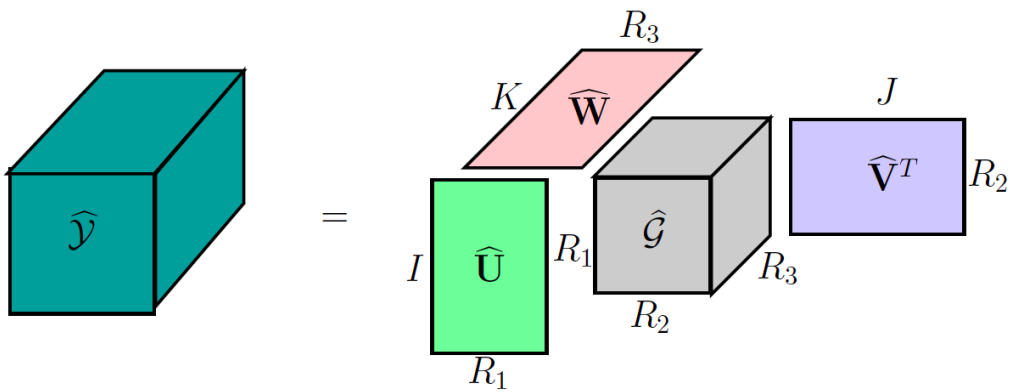
Extensions

CP vs. Tucker

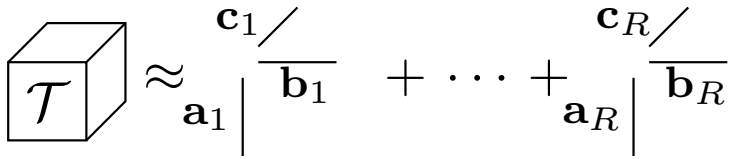
- CPD: $(I + J + K - 2)R$



- Tucker: $IR_1 + JR_2 + KR_3 + R_1R_2R_3 - R_1^2 - R_2^2 - R_3^2$

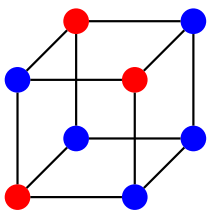


CP approximation is ill-posed



Best r-rank approximation ($r > 1$) may not exist

$$\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$$



$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

rank $\mathcal{T} = 3$, but approximated by rank-2 tensor to any accuracy

$$\mathcal{T} = \frac{1}{2\varepsilon}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2)^{\otimes 3} - \frac{1}{2\varepsilon}(\mathbf{e}_1 - \varepsilon \mathbf{e}_2)^{\otimes 3} + \mathcal{O}(\varepsilon)$$

Set of rank-$\leq r$ tensors is not closed for $r > 1$

Rank-one tensor approximation

$$\min_{\mathbf{a},\mathbf{b},\mathbf{c}} \|\mathcal{T} - \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}\|_F^2$$

- well-posed (minimum exists)
- block-coordinate descent (ALS, non-symmetric power method) converges globally
  (to a stationary point) [Uschmajew, 2015]
- related to the notion of singular vectors/eigenvectors of a tensor [Qi, Luo, 2017]
- number of stationary points is known [Freidland, Ottaviani, 2014]

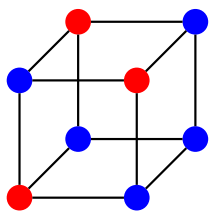
Successive approximation (deflation)

Subtracting rank-one approximation may increase tensor rank:

$$\mathcal{X}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{X}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

$\operatorname{rank}(\mathcal{X}) = 2$, but $\mathcal{X} - \widehat{\mathcal{X}}_1 = \mathcal{T}$, $\operatorname{rank}(\mathcal{T}) = 3$

$$\mathcal{T} = \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_1 + \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1$$



$$\mathcal{T}_{:,:,1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{T}_{:,:,2} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

But when does it work then?

Successive rank-1 approximation

- Orthogonally decomposable tensors [Zhang, Golub, 2001]:

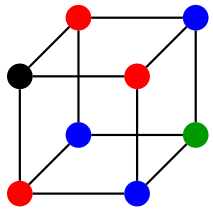
$$\mathcal{T} = \sum_{k=1}^{R} \mathbf{a}_k \otimes \mathbf{b}_k \otimes \mathbf{c}_k, \quad \mathbf{a}_k \perp \mathbf{a}_j, \quad \mathbf{b}_k \perp \mathbf{b}_j, \quad \mathbf{c}_k \perp \mathbf{c}_j,$$

  The successive best rank-1 approximation returns the components in the sum

- Cyclic (sequential) rank-one approximation [da Silva, Comon, de Almeida, 2015]:

$$\min \|\mathcal{T} - (\widehat{\mathcal{T}}_1 + \ldots + \widehat{\mathcal{T}}_R)\| \text{ subject to } \operatorname{rank}(\widehat{\mathcal{T}}_k) = 1.$$

Rank-one approximation of symmetric tensors

- $\mathcal{T} \in S^d_n$ — symmetric tensors,  S^3_2
- Best non-symmetric approximation can be chosen symmetric (and is unique a.s.) [Friedland, Ottaviani, 2014]
- Best symmetric approximation $\leftrightarrow$ maximization of a polynomial

$$\min_{\lambda,\mathbf{v}} \|\mathcal{T} - \lambda \mathbf{v} \otimes \cdots \otimes \mathbf{v}\|_F^2 \leftrightarrow \max_{\|\mathbf{v}\|_2=1} |\mathcal{T} \bullet_1 \mathbf{v} \cdots \bullet_d \mathbf{v}|$$

- stationary points: eigenvectors of the tensor

$$\mathcal{T} \cdot \mathbf{v}^{d-1} = \mu \mathbf{v}$$

- In total, $\frac{(d-1)^n-1}{d-2}$ (complex) eigenvectors [Cartwright, Sturmfels, 2013]
- There are cases when the power method diverges [Chen, Saad, 2009]
- Orthogonally decomposable tensors: power method converges, deflation works decomposition [Anandkumar et al, 2013], [Robeva,2016]

Algorithms for CP approximation

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathcal{Y} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]\|_F^2 + \ldots$$

Constraints/regularization:

- Orthogonality: [Comon, 1993], [Robeva, 2016]
- Coherence: [Lim, Comon, 2014]
- Nonnegativity: [Qi, Comon, Lim, 2016]

Algorithms for CP approximation

- Alternating minimization (ALS, AO-ADMM)
  (Global) convergence properties in the regularized case [Xu, Yin, 2013]
- Nonlinear least squares
- Riemannian optimization
- Algebraic algorithms
  - Generalized eigenvalue decomposition (non-symmetric tensors)
  - Structured matrix approximation (symmetric tensors)

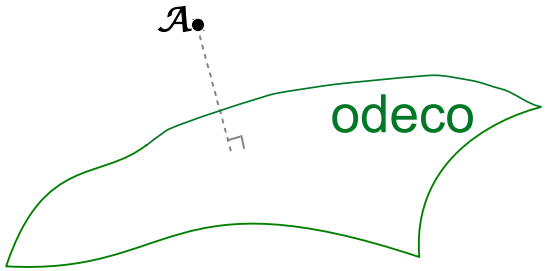
Tensor diagonalization as orthogonal approximation

By duality:

$$\max_{\mathbf{Q}\in\mathcal{O}_n} \| \operatorname{diag} \mathcal{A} \bullet_1 \mathbf{Q}^\mathsf{T} \cdots \bullet_d \mathbf{Q}^\mathsf{T} \|_2^2$$

$$= \|\mathcal{A}\|_F^2 - \underbrace{\min_{\substack{\mathbf{Q}\in\mathcal{O}_n \\ \mathbf{Q}=[\,\mathbf{u}_1 \,\cdots\, \mathbf{u}_n\,]}} \|\mathcal{A} - \sum_{k=1}^{n} \mu_k \, \mathbf{u}_k \otimes \cdots \otimes \mathbf{u}_k\|_F^2}_{\text{best } n\text{-rank symmetric orthogonal approximation}}$$

- Euclidean distance to odeco *[Robeva, 2016]* variety



- Best non-symmetric approximation $\neq$ best symmetric approximation *[Li, U., Comon, 2019]* (a variant of Comon's conjecture is false)

Optimization on orthogonal/unitary group

$$\max_{\mathbf{Q}\in\mathcal{O}_n} f(\mathbf{Q}) \quad \text{or} \quad \max_{\mathbf{U}\in\mathcal{U}_n} f(\mathbf{U}), \quad f \text{ is a low-order polynomial}$$

Algebraic orthogonal tensor decompositions:

- Deflation (successive rank-one approximation) *[Delfosse, Loubaton, 1995]*, *[Anandkumar, 2013]*, *[Robeva, 2016]*
- EVD of tensor slice(s) *[De Lathauwer, 2006]*, *[Kolda, 2015]*

Optimization on the manifold:

- Riemannian optimization (CG, SD, BFGS, RTR) *[Absil et al., 2008]*
- Jacobi-type algorithms [Comon, 1993], [Li, U., Comon, 2020] (convergence)

What do we know about CP approximation?

Given $\mathcal{T} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] + \mathcal{E}$ and

$$(\widehat{\mathbf{A}}^*, \widehat{\mathbf{B}}^*, \widehat{\mathbf{C}}^*) = \arg \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}} \|\mathcal{T} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]\|,$$

- Approaches using random matrix theory [Goulart, Couillet, Comon, 2022] (mostly rank-1)
- Perturbation bounds for some algebraic algorithms [Evert, De Lathauwer, 2022] (very small ranks)
- Uniqueness of best approximation in a small neighorhood [Friedland, Stawiska, 2016] (non-constructive)

Overview

Matrix LRA

Tensors

Factorization

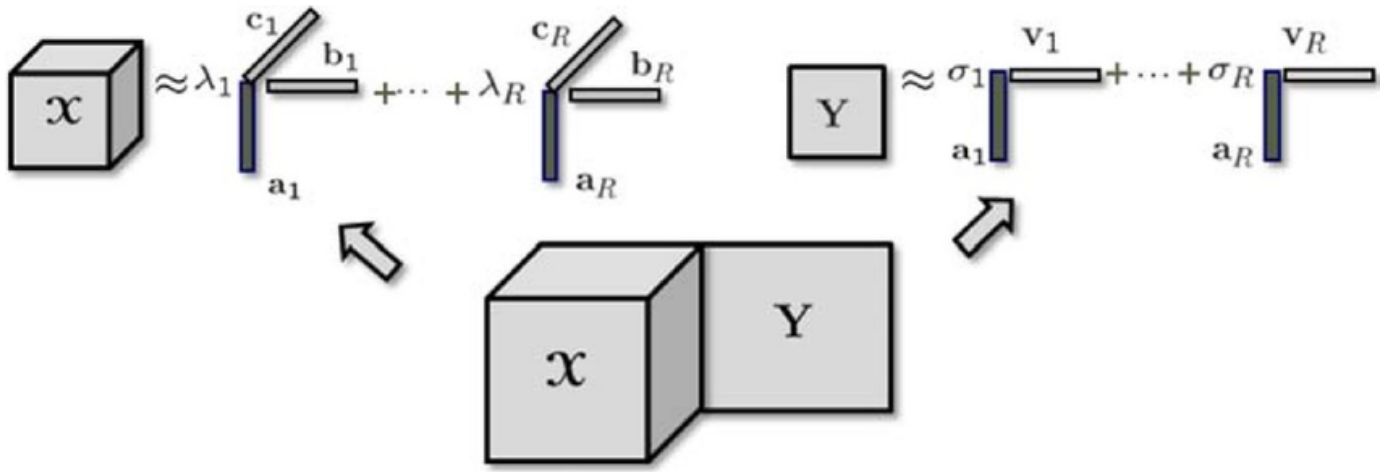
(CP) decomposition

CP approximation

Extensions

Coupled factorizations

Factors of different tensor/matrix decompositions may be shared:



[Acar, Kolda, Dunlavy, 2011]

Joint CPD of symmetric tensors

ICA model: $\mathbf{x} = A\mathbf{s}, \quad A = \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_r \end{bmatrix} \in \mathbb{R}^{n \times r}$,

cumulants of $\mathbf{x}$ up to order d:

$$\begin{aligned} \mathcal{C}_{\mathbf{x}}^{(1)} &= c_{1,1}\mathbf{a}_1 + \cdots + c_{1,r}\mathbf{a}_r, \\ \mathcal{C}_{\mathbf{x}}^{(2)} &= c_{2,1}\mathbf{a}_1 \otimes \mathbf{a}_1 + \cdots + c_{2,r}\mathbf{a}_r \otimes \mathbf{a}_r, \\ &\vdots \\ \mathcal{C}_{\mathbf{x}}^{(d)} &= c_{d,1}\mathbf{a}_1 \otimes \cdots \otimes \mathbf{a}_1 + \cdots + c_{d,r}\mathbf{a}_r \otimes \cdots \otimes \mathbf{a}_r, \end{aligned} \tag{1}$$

where $c_{j,k}$ is the j-th cumulant of s_k.

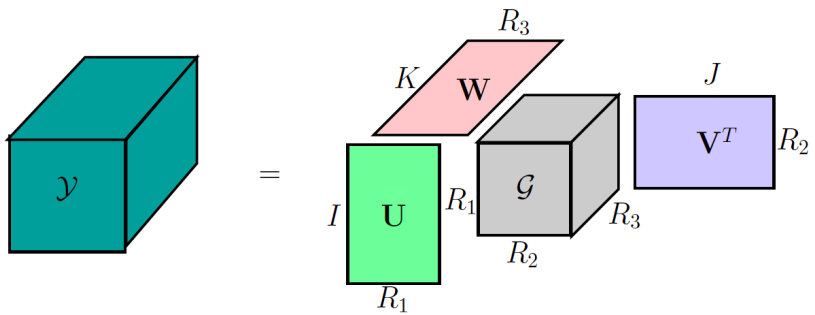
Problem. Given $\mathcal{C}_{\mathbf{x}}^{(j)}$, find $\mathbf{a}_k$ and $c_{j,k}$
(diagonalize all the cumulants simultaneously)

Sum of Tucker: block-term decomposition

[De Lathauwer, 2008]:
For fixed (R_1, R_2, R_3):

$$\mathcal{T} = \sum_{k=1}^{r} \mathcal{G}_k \bullet_1 \mathbf{U}_k \bullet_2 \mathbf{V}_k \bullet_3 \mathbf{W}_k$$

each term in the sum has ML-rank (R_1, R_2, R_3):



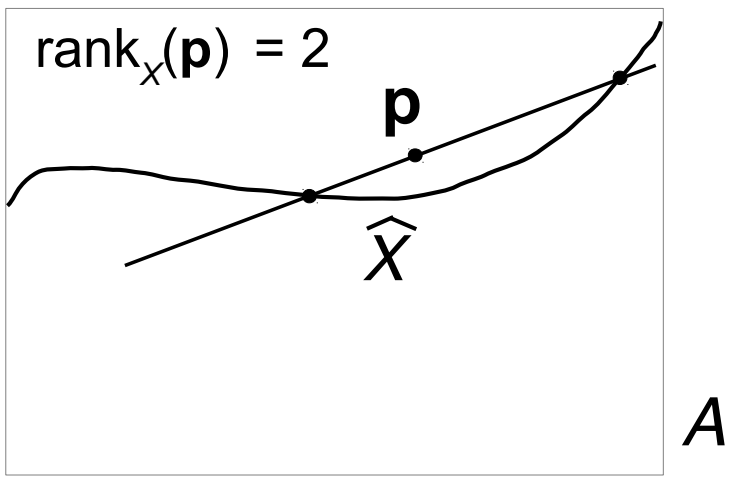
Special case: $(R_1, R_2, R_3) = (L, L, 1)$: very useful in signal processing, see e.g. [Goulart, et al. 2020]

Additive decompositions

X-rank (join) decomposition [Zak, 2004], [Landsberg, 2012], [Comon, Qi, U., 2017] (sparse algebraic decomposition)

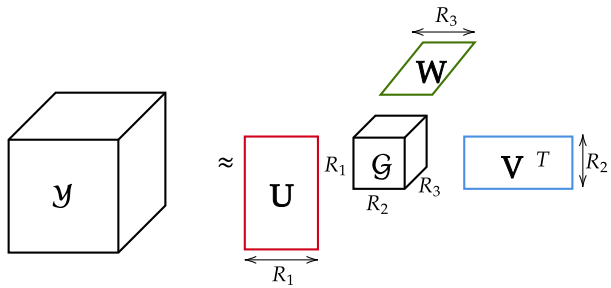
$$\mathbf{p} = \mathbf{x}_1 + \cdots + \mathbf{x}_r, \quad \mathbf{x}_k \in \widehat{X}$$

- $A \stackrel{\text{def}}{=}$ ambient tensor space (e.g. $A = \mathbb{C}^{I \times J \times K}$)
- $\widehat{X} \stackrel{\text{def}}{=}$ (variety of "simple"' terms) (e.g., rank-one $\widehat{X} = \{\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}\}$)
- study the properties of secant varieties $\sigma_r(\widehat{X})$

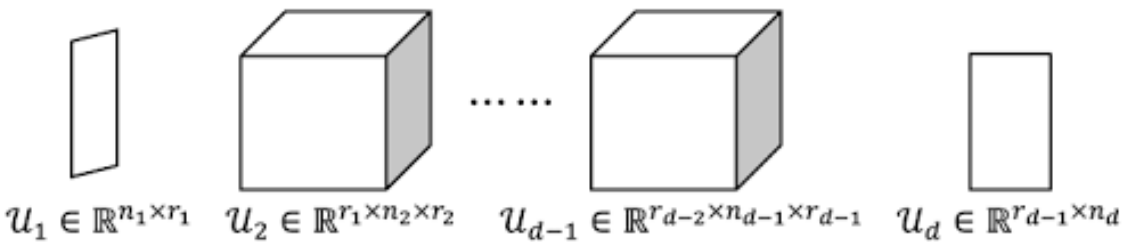


Approximation: higher-order tensors

- Tucker: $O(dIR + R^d)$: curse of dimensionality



- tensor trains, hierarchical Tucker: linear-in-d storage complexity



Decompositions: flexible/multilayer

[Harshman, Lundy, 1996], [Roald et al, 2022]

- CPD (PARAFAC)

$$\mathcal{T}_{:,:,k} = \underset{I}{\boxed{\mathbf{A}}} \cdot \overset{R}{\boxed{\mathbf{D}_{\mathbf{C}}^{(k)}}} \cdot \underset{R}{\boxed{\mathbf{B}^{\mathsf{T}}}}, \quad k = 1, \ldots, K$$

- PARAFAC-2

$$\mathcal{T}_{:,:,k} = \underset{I}{\boxed{\mathbf{A}}} \cdot \overset{R}{\boxed{\mathbf{D}_{\mathbf{C}}^{(k)}}} \cdot \underset{R}{\boxed{\mathbf{B}_k^{\mathsf{T}}}}, \quad k = 1, \ldots, K$$

- ParaTuck-2

$$\mathcal{T}_{:,:,k} = \underset{I}{\boxed{\mathbf{A}}} \cdot \overset{R}{\boxed{\mathbf{D}_{\mathbf{G}}^{(k)}}} \cdot \underset{R}{\overset{S}{\boxed{\mathbf{F}}}} \cdot \overset{S}{\boxed{\mathbf{D}_{\mathbf{H}}^{(k)}}} \cdot \underset{S}{\overset{J}{\boxed{\mathbf{B}^{\mathsf{T}}}}}, \quad k = 1, \ldots, K$$

Share the uniqueness features of CPD!

Conclusion

Which tensor decomposition (format) to use?

- factorization (Tucker, tensor trains) — compression
- decomposition (CPD, BTD) — identification of components

Challenges (personal choice):

- Guarantees (approximation bounds) on CP approximation
- Multi-layer/multilevel tensor decompositions (e.g., ParaTuck-2)

Thank you!

Conclusion

Which tensor decomposition (format) to use?

- factorization (Tucker, tensor trains) — compression
- decomposition (CPD, BTD) — identification of components

Challenges (personal choice):

- Guarantees (approximation bounds) on CP approximation
- Multi-layer/multilevel tensor decompositions (e.g., ParaTuck-2)

Advertisement (not mine):

- A number of postdoc/PhD positions in the SiMul team (Nancy)
  `https://cran-simul.github.io`
- Summer school on low-rank approximation (23-29.06.24) in Peyresq
  Organized by N. Gillis and J. Cohen

Thank you!