Tensor models & applications

J. Henrique de M. Goulart

Toulouse INP / IRIT

henrique.goulart@irit.fr

Peyresq – June 2024

About this mini-course

Objectives & method:

1. Provide an overview of the main concepts & results pertaining to tensors & their decompositions in signal processing & machine learning.
2. Emphasis on intuitions rather than formal proofs (but still keep some).
3. Rely on application examples from these domains to motivate & llustrate the introduced tools.

Disclaimer:

- This is a very broad subject with a vast literature, spanning several communities (SP, ML, chemometrics, numerical analysis, physics, ...).
- This mini-course is my personal take on it—hence partial and biased.

Prerequisites & handout

Prerequisites:

- linear algebra
- basic probability & statistics
- basic optimization

Supporting material: An electronic handout is provided on my webpage:

`https://www.irit.fr/~Henrique.Goulart/talks/`

It contains:

- a summary of important identities and properties
- some proposed exercices
- many bibliographical pointers

Agenda

0. About this mini-course
1. Why care about tensors?
2. A jungle of tensor models
3. A mosaic of uniqueness results
4. Rank-1 approximation, tensor spectrum and power iteration
5. Low-rank approximation, in several flavors
6. Tensor PCA & asymptotic MLE performance

Why care about tensors?

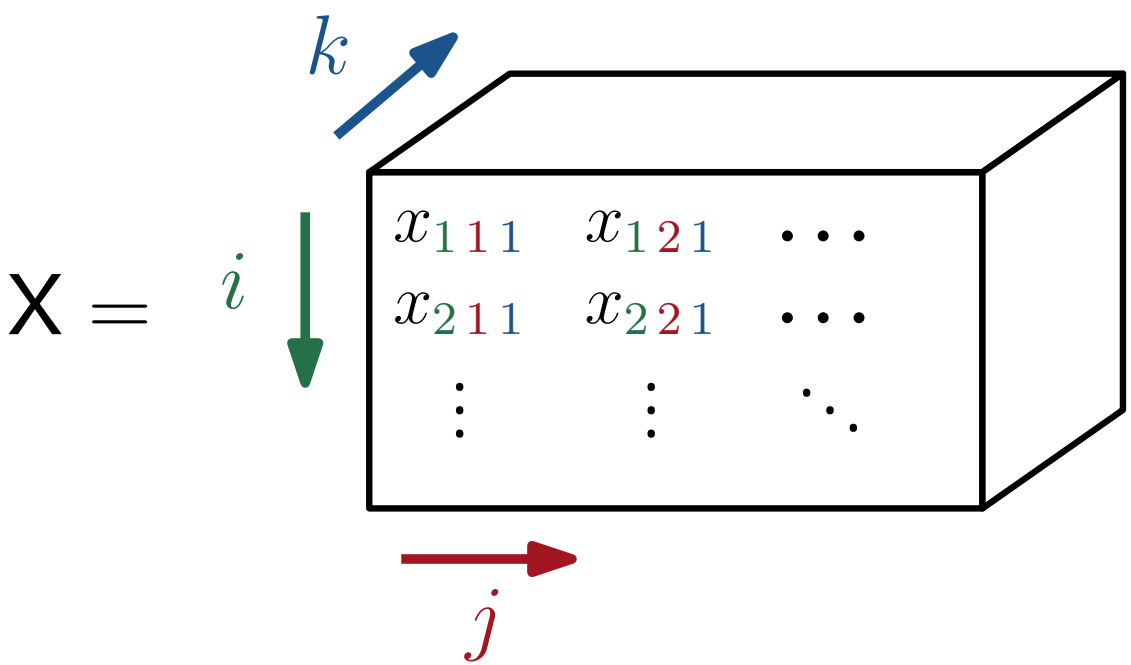
A first definition

Tensors are (for now) simply multi-way arrays (a.k.a. hypermatrices)

$$\mathsf{X} \in \mathbb{R}^{N_1 \times \cdots \times N_d}, \qquad (\mathsf{X})_{i_1 \ldots i_d} = x_{i_1 \ldots i_d},$$

where d is the order of X.

For $d = 3$, we write $\mathsf{X} = (\mathsf{X})_{ijk}$. We often say that X has d modes.



Tensor models

Numerous applications involve data (or functions thereof) or other objects that can be modelled by tensors having a low-dimensional (low-rank) structure.

Tensor models

Numerous applications involve data (or functions thereof) or other objects that can be modelled by tensors having a low-dimensional (low-rank) structure.

Such tensor models are useful for many purposes falling into two categories:

1. Information extraction
2. Complexity reduction

Let's see a few examples.

Example #1: estimating Gaussian mixtures

Take a pth-dim spherical Gaussian mixture model (GMM):

$$\boldsymbol{X} \sim \sum_{r=1}^{R} \pi_r \mathcal{N}(\boldsymbol{\mu}_r, \sigma_r^2 \mathbf{I}).$$

Goal: Given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^N$, estimate $\left\{(\pi_r, \boldsymbol{\mu}_r, \sigma_r^2)\right\}_{r=1}^{R}$.

Classical approach: expectation-minimization (EM) algorithm[1]—sensitive w.r.t. initialization, may converge very slowly.[2]

1: Dempster & al., 1977, 2: Park & Ozeki, 2009

Example #1: estimating Gaussian mixtures

Take a pth-dim spherical Gaussian mixture model (GMM):

$$\boldsymbol{X} \sim \sum_{r=1}^{R} \pi_r \mathcal{N}(\boldsymbol{\mu}_r, \sigma_r^2 \mathbf{I}).$$

Goal: Given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^N$, estimate $\left\{(\pi_r, \boldsymbol{\mu}_r, \sigma_r^2)\right\}_{r=1}^{R}$.

Classical approach: expectation-minimization (EM) algorithm[1]—sensitive w.r.t. initialization, may converge very slowly.[2]

Let's try instead the method of moments:

$$\mathbb{E}\{\boldsymbol{X}\} = \sum_{r=1}^{R} \pi_r \boldsymbol{\mu}_r,$$

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} = \sum_{r=1}^{R} \pi_r \left(\boldsymbol{\mu}_r \boldsymbol{\mu}_r^\mathsf{T} + \sigma_r^2 \mathbf{I}\right) = \sum_{r=1}^{R} \pi_r \boldsymbol{\mu}_r \boldsymbol{\mu}_r^\mathsf{T} + \bar{\sigma}^2 \mathbf{I}.$$

$\bar{\sigma}^2 = \sum_r \pi_r \sigma_r^2$

Q: Given $\mathbb{E}\{\boldsymbol{X}\}, \mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\}$, can we identify the GMM parameters?

1: Dempster & al., 1977, 2: Park & Ozeki, 2009

First try: second-order moments

Letting $\tilde{\boldsymbol{\mu}}_r := \sqrt{\pi_r}\,\boldsymbol{\mu}_r$ and $\mathbf{M} := \begin{pmatrix} \tilde{\boldsymbol{\mu}}_1 & \dots & \tilde{\boldsymbol{\mu}}_R \end{pmatrix}$, we can write:

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} = \sum_{r=1}^{R} \tilde{\boldsymbol{\mu}}_r \tilde{\boldsymbol{\mu}}_r^\mathsf{T} + \bar{\sigma}^2\mathbf{I} = \mathbf{M}\mathbf{M}^\mathsf{T} + \bar{\sigma}^2\mathbf{I}$$

First try: second-order moments

Letting $\tilde{\boldsymbol{\mu}}_r := \sqrt{\pi_r}\,\boldsymbol{\mu}_r$ and $\mathbf{M} := \begin{pmatrix} \tilde{\boldsymbol{\mu}}_1 & \dots & \tilde{\boldsymbol{\mu}}_R \end{pmatrix}$, we can write:

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} = \sum_{r=1}^{R} \tilde{\boldsymbol{\mu}}_r \tilde{\boldsymbol{\mu}}_r^\mathsf{T} + \bar{\sigma}^2\mathbf{I} = \mathbf{M}\mathbf{M}^\mathsf{T} + \bar{\sigma}^2\mathbf{I}$$

Let's focus on estimating the means $\boldsymbol{\mu}_r$.

If $R < N$, then $\bar{\sigma}^2 = \lambda_{\min}\left(\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\}\right)$, estimated by $\lambda_{\min}\left(\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{x}_m\boldsymbol{x}_m^\mathsf{T}\right)$.

Hence, we can approximately compute

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} - \bar{\sigma}^2\mathbf{I} = \mathbf{M}\mathbf{M}^\mathsf{T}.$$

Can we then recover $\{\tilde{\boldsymbol{\mu}}_r\}_{r=1}^{R}$ by matrix decomposition?

First try: second-order moments

Letting $\tilde{\boldsymbol{\mu}}_r := \sqrt{\pi_r}\,\boldsymbol{\mu}_r$ and $\mathbf{M} := \begin{pmatrix} \tilde{\boldsymbol{\mu}}_1 & \dots & \tilde{\boldsymbol{\mu}}_R \end{pmatrix}$, we can write:

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} = \sum_{r=1}^{R} \tilde{\boldsymbol{\mu}}_r \tilde{\boldsymbol{\mu}}_r^\mathsf{T} + \bar{\sigma}^2 \mathbf{I} = \mathbf{M}\mathbf{M}^\mathsf{T} + \bar{\sigma}^2 \mathbf{I}$$

Let's focus on estimating the means $\boldsymbol{\mu}_r$.

If $R < N$, then $\bar{\sigma}^2 = \lambda_{\min}\left(\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\}\right)$, estimated by $\lambda_{\min}\left(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m \boldsymbol{x}_m^\mathsf{T}\right)$.

Hence, we can approximately compute

$$\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\} - \bar{\sigma}^2 \mathbf{I} = \mathbf{M}\mathbf{M}^\mathsf{T}.$$

Can we then recover $\{\tilde{\boldsymbol{\mu}}_r\}_{r=1}^{R}$ by matrix decomposition? No:

$$\mathbf{M}\mathbf{M}^\mathsf{T} = (\mathbf{M}\mathbf{S})(\mathbf{M}\mathbf{S}^{-\mathsf{T}})^\mathsf{T}, \qquad \forall\, \mathbf{S} \in \mathrm{GL}_R(\mathbb{R}).$$

"too much freedom = not identifiable"

Going to higher orders

The third-order moments give $\binom{N+2}{3}$ distinct equations of the form:

$$\mathbb{E}\left\{X_i\, X_j\, X_k\right\} = \sum_{r=1}^{R} \pi_r\, \mathbb{E}\left\{\left[(\boldsymbol{\mu}_r)_i + (\boldsymbol{U}_r)_i\right]\left[(\boldsymbol{\mu}_r)_j + (\boldsymbol{U}_r)_j\right]\left[(\boldsymbol{\mu}_r)_k + (\boldsymbol{U}_r)_k\right]\right\}$$

where $\boldsymbol{U}_r \sim \mathcal{N}(\mathbf{0}, \sigma_r^2 \mathbf{I})$.

Going to higher orders

The third-order moments give $\binom{N+2}{3}$ distinct equations of the form:

$$\mathbb{E}\left\{X_i\, X_j\, X_k\right\} = \sum_{r=1}^{R} \pi_r\, \mathbb{E}\left\{\left[(\boldsymbol{\mu}_r)_i + (\boldsymbol{U}_r)_i\right]\left[(\boldsymbol{\mu}_r)_j + (\boldsymbol{U}_r)_j\right]\left[(\boldsymbol{\mu}_r)_k + (\boldsymbol{U}_r)_k\right]\right\}$$

where $\boldsymbol{U}_r \sim \mathcal{N}(\mathbf{0}, \sigma_r^2 \mathbf{I})$. Direct computation gives

$$\mathbb{E}\left\{X_i\, X_j\, X_k\right\} = \sum_{r=1}^{R} \pi_r \Big[(\boldsymbol{\mu}_r)_i\, (\boldsymbol{\mu}_r)_j\, (\boldsymbol{\mu}_r)_k + (\boldsymbol{\mu}_r)_i \sigma_r^2 \delta_{jk} + (\boldsymbol{\mu}_r)_j \sigma_r^2 \delta_{ik} + (\boldsymbol{\mu}_r)_k \sigma_r^2 \delta_{ij}\Big]$$

Going to higher orders

The third-order moments give $\binom{N+2}{3}$ distinct equations of the form:

$$\mathbb{E}\{X_i\, X_j\, X_k\} = \sum_{r=1}^{R} \pi_r\, \mathbb{E}\{[(\boldsymbol{\mu}_r)_i + (\boldsymbol{U}_r)_i]\,[(\boldsymbol{\mu}_r)_j + (\boldsymbol{U}_r)_j]\,[(\boldsymbol{\mu}_r)_k + (\boldsymbol{U}_r)_k]\}$$

where $\boldsymbol{U}_r \sim \mathcal{N}(\mathbf{0}, \sigma_r^2 \mathbf{I})$. Direct computation gives

$$\mathbb{E}\{X_i\, X_j\, X_k\} = \sum_{r=1}^{R} \pi_r \Big[(\boldsymbol{\mu}_r)_i\,(\boldsymbol{\mu}_r)_j\,(\boldsymbol{\mu}_r)_k + (\boldsymbol{\mu}_r)_i \sigma_r^2 \delta_{jk} + (\boldsymbol{\mu}_r)_j \sigma_r^2 \delta_{ik} + (\boldsymbol{\mu}_r)_k \sigma_r^2 \delta_{ij}\Big]$$

By introducing an appropriate symmetrization operator Sym, we get

$$\mathbb{E}\{X_i\, X_j\, X_k\} = \sum_{r=1}^{R} \pi_r (\boldsymbol{\mu}_r)_i\,(\boldsymbol{\mu}_r)_j\,(\boldsymbol{\mu}_r)_k + 3\,\mathrm{Sym}\left(\delta_{jk} \underbrace{\sum_{r=1}^{R} \pi_r \sigma_r^2 (\boldsymbol{\mu}_r)_i}_{:= (\boldsymbol{c})_i}\right)$$

A first tensor decomposition

Let's introduce some convenient (& natural) notation:

Def: elementary or rank-1 tensor

$$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2} \times \mathbb{R}^{N_3} \mapsto \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

$$(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3)_{ijk} = (\boldsymbol{x}_1)_i \, (\boldsymbol{x}_2)_j \, (\boldsymbol{x}_3)_k$$

A first tensor decomposition

Let's introduce some convenient (& natural) notation:

Def: elementary or rank-1 tensor

$$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2} \times \mathbb{R}^{N_3} \mapsto \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

$$(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3)_{ijk} = (\boldsymbol{x}_1)_i \, (\boldsymbol{x}_2)_j \, (\boldsymbol{x}_3)_k$$

This allows expressing all $\binom{N+2}{3}$ equations as:

$$\mathbb{E}\{\boldsymbol{X} \otimes \boldsymbol{X} \otimes \boldsymbol{X}\} = \sum_{r=1}^{R} \pi_r \, \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r + \mathsf{S}(\boldsymbol{c})$$

Moreover, $\boldsymbol{c}$ can be estimated from data.

A first tensor decomposition

Let's introduce some convenient (& natural) notation:

Def: elementary or rank-1 tensor

$$(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2} \times \mathbb{R}^{N_3} \mapsto \boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3 \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

$$(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \boldsymbol{x}_3)_{ijk} = (\boldsymbol{x}_1)_i \, (\boldsymbol{x}_2)_j \, (\boldsymbol{x}_3)_k$$

This allows expressing all $\binom{N+2}{3}$ equations as:

$$\mathbb{E}\{\boldsymbol{X} \otimes \boldsymbol{X} \otimes \boldsymbol{X}\} = \sum_{r=1}^{R} \pi_r \, \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r + \mathsf{S}(\boldsymbol{c})$$

Moreover, $\boldsymbol{c}$ can be estimated from data.

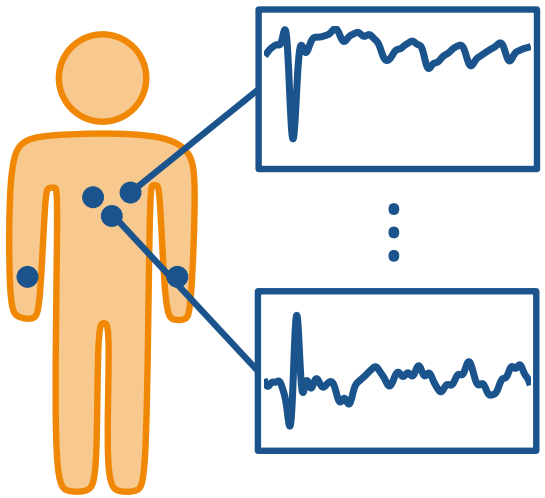
Hence, our problem $\approx$ decomposition of the third-order moment tensor:[1,2]

$$\mathbb{E}\{\boldsymbol{X} \otimes \boldsymbol{X} \otimes \boldsymbol{X}\} - \mathsf{S}(\boldsymbol{c}) = \sum_{r=1}^{R} \pi_r \, \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r \otimes \boldsymbol{\mu}_r .$$

Good news: (essentially) unique decomposition under mild constraints!

1: Hsu & Kakade, 2013, 2: Anandkumar & al., 2014

Example #2: ECG signal separation



$$x_1(n) = \sum_{r=1}^{R} (\boldsymbol{w}_r)_1 \, s_r(n)$$

$$\vdots$$

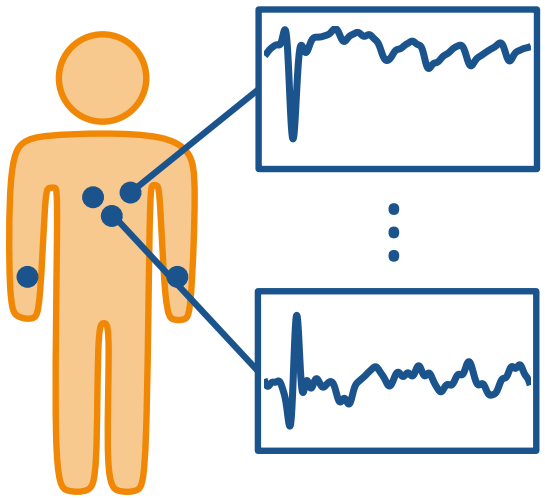
$$x_K(n) = \sum_{r=1}^{R} (\boldsymbol{w}_r)_K \, s_r(n)$$

K sensors (leads)

N observations (samples)

R sources

Example #2: ECG signal separation



$$x_1(n) = \sum_{r=1}^{R} (\boldsymbol{w}_r)_1 \, s_r(n)$$

$$\vdots$$

$$x_K(n) = \sum_{r=1}^{R} (\boldsymbol{w}_r)_K \, s_r(n)$$

K sensors (leads)

N observations (samples)

R sources

The data can be expressed as a matrix decomposition (linear mixture model):

$$\underbrace{\begin{pmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_K \end{pmatrix}}_{\mathbf{X} \in \mathbb{R}^{N \times K}} = \underbrace{\begin{pmatrix} \boldsymbol{s}_1 & \dots & \boldsymbol{s}_R \end{pmatrix}}_{\mathbf{S} \in \mathbb{R}^{N \times R}} \underbrace{\begin{pmatrix} \boldsymbol{w}_1 & \dots & \boldsymbol{w}_R \end{pmatrix}^\mathsf{T}}_{\mathbf{W}^\mathsf{T} \in \mathbb{R}^{R \times K}}$$

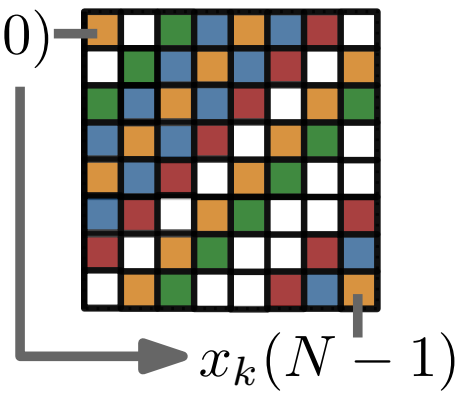
$$= \underbrace{\mathbf{S}\,\mathbf{P}}_{\tilde{\mathbf{S}}} \underbrace{\mathbf{P}^{-1}\,\mathbf{W}^\mathsf{T}}_{\tilde{\mathbf{W}}^\mathsf{T}}, \qquad \forall\, \mathbf{P} \in \mathrm{GL}_R(\mathbb{R}).$$

This model is clearly non-identifiable without further constraints.

“Tensorization” by additional diversity

Idea: “Hankelize” the observed signals to add one temporal “diversity”.[1]

$$\boldsymbol{x}_k \mapsto \mathbf{X}_k = \quad = \sum_{r=1}^{R} (\boldsymbol{w}_r)_k \, \mathbf{H}_r$$



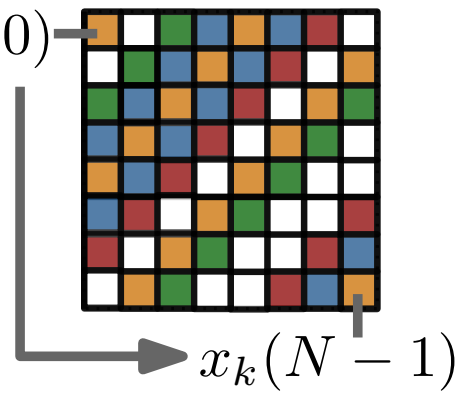
$$\boldsymbol{s}_r \mapsto \mathbf{H}_r$$

1: De Lathauwer, 2011

“Tensorization” by additional diversity

Idea: “Hankelize” the observed signals to add one temporal “diversity”.[1]

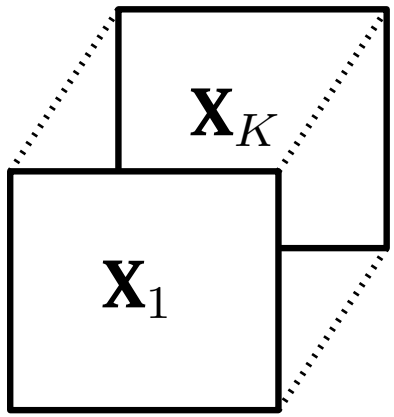
$$\boldsymbol{x}_k \mapsto \mathbf{X}_k = \quad = \sum_{r=1}^{R} (\boldsymbol{w}_r)_k \, \mathbf{H}_r$$

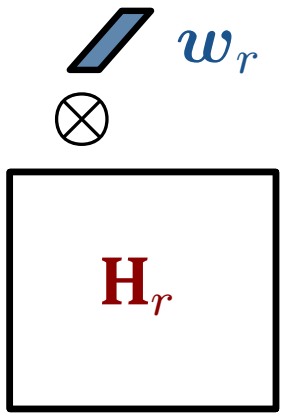


$\boldsymbol{s}_r \mapsto \mathbf{H}_r$

Stacking the $\mathbf{X}_k$ as slices of $\mathsf{X} \in \mathbb{R}^{M \times M \times K}$, with $M = \frac{N+1}{2}$, we get:

$$\mathsf{X} = \quad = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{w}_r = \sum_{r=1}^{R} \quad \boldsymbol{w}_r$$





Def: matrix-vector tensor product $\quad (\mathbf{H} \otimes \boldsymbol{w})_{ijk} = (\mathbf{H})_{ik} \, (\boldsymbol{w}_r)_k$

1: De Lathauwer, 2011

A decomposition in low-rank blocks

Without further assumptions, the model is still non-identifiable.

But we can add a reasonable one:

"Parsimony" assumption: sources are given by

$$s_r(n) = \sum_{\ell=1}^{L_r} \alpha_{\ell,r}\, z_{\ell,r}^n \quad \Rightarrow \quad \operatorname{rank} \mathbf{H}_r \leq L_r < M,$$

where $\alpha_{\ell,r}, z_{\ell,r} \in \mathbb{C}$.

A decomposition in low-rank blocks

Without further assumptions, the model is still non-identifiable.

But we can add a reasonable one:

"Parsimony" assumption: sources are given by

$$s_r(n) = \sum_{\ell=1}^{L_r} \alpha_{\ell,r}\, z_{\ell,r}^n \quad \Rightarrow \quad \operatorname{rank} \mathbf{H}_r \leq L_r < M,$$

where $\alpha_{\ell,r}, z_{\ell,r} \in \mathbb{C}$.

Hence, our source separation problem becomes that of decomposing $\mathbf{X}$ into blocks of the form (low-rank matrix) $\otimes$ vector:

$$\mathsf{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{w}_r, \qquad \forall r,\ \operatorname{rank} \mathbf{H}_r \leq L_r.$$

As before, it turns out that $\{(\mathbf{H}_r, \boldsymbol{w}_r)\}_{r=1}^R$ are (essentially) unique under mild constraints.[1]

1: De Lathauwer, 2011

Example #3: Multilinear PCA

We're given $M \times M$ face images of N_p people:

$$\mathbf{X} \in \mathbb{R}^{N \times N_\mathrm{p}}. \qquad \text{with } N = M^2.$$

Dimensionality reduction and feature extraction can be achieved by PCA (via truncated SVD of $\mathbf{X}$, after centering):

$$\mathbf{X} \approx \mathbf{U}_R \mathbf{\Sigma}_R \mathbf{V}_R^\mathsf{T} = \sum_{r=1}^{R} \sigma_r (\mathbf{U}_R)_{:r} (\mathbf{V}_R)_{:r}^\mathsf{T}, \qquad x_{ip} \approx \sum_{r=1}^{R} \sigma_r (\mathbf{U}_R)_{ir} (\mathbf{V}_R)_{pr},$$

where

- $\mathbf{U}_R \in \mathbb{R}^{N \times R}$ contains the R first left singular vectors
- $\mathbf{V}_R \in \mathbb{R}^{N_\mathrm{p} \times R}$ contains the R first right singular vectors
- $\mathbf{\Sigma}_R = \mathrm{Diag}(\sigma_1, \ldots, \sigma_R)$ contains the R first singular values

Example #3: Multilinear PCA

We're given $M \times M$ face images of N_p people:

$$\mathbf{X} \in \mathbb{R}^{N \times N_\mathrm{p}}. \qquad \text{with } N = M^2.$$

Dimensionality reduction and feature extraction can be achieved by PCA (via truncated SVD of $\mathbf{X}$, after centering):

$$\mathbf{X} \approx \mathbf{U}_R \mathbf{\Sigma}_R \mathbf{V}_R^\mathsf{T} = \sum_{r=1}^{R} \sigma_r (\mathbf{U}_R)_{:r} (\mathbf{V}_R)_{:r}^\mathsf{T}, \qquad x_{ip} \approx \sum_{r=1}^{R} \sigma_r (\mathbf{U}_R)_{ir} (\mathbf{V}_R)_{pr},$$

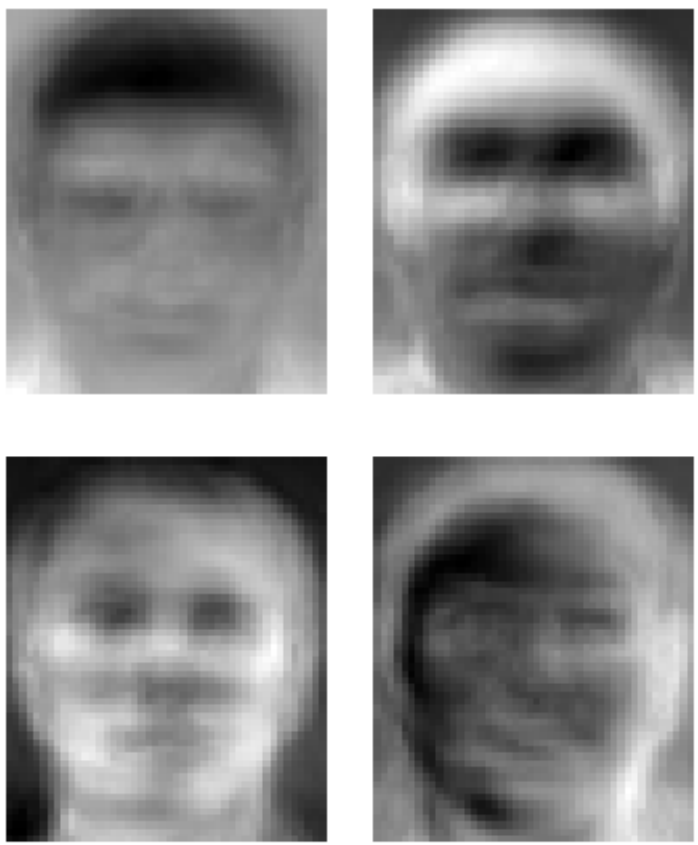
where

- $\mathbf{U}_R \in \mathbb{R}^{N \times R}$ contains the R first left singular vectors
- $\mathbf{V}_R \in \mathbb{R}^{N_\mathrm{p} \times R}$ contains the R first right singular vectors
- $\mathbf{\Sigma}_R = \mathrm{Diag}(\sigma_1, \ldots, \sigma_R)$ contains the R first singular values

Hence, $(\mathbf{\Sigma}_R \mathbf{V}_R^\mathsf{T})_{:p} \in \mathbb{R}^R$ contains the coordinates of the pth person's image w.r.t. the subspace basis $\mathbf{U}_R$.

Eigenfaces

This is the so-called eigenfaces[1] approach to dimension reduction and feature extraction.



A few eigenfaces (cols of $\mathbf{U}_R$) from the ORL database (source: Wikipedia).

1: Sirovich & Kirby, 1987

Extension to multiple diversities

Consider next N-dim face images of N_p people under N_i illumination conditions.

PCA of $\mathbf{X} \in \mathbb{R}^{N \times N_\mathrm{i} N_\mathrm{p}}$ is still possible, but two diversities "get entangled."

Extension to multiple diversities

Consider next N-dim face images of N_p people under N_i illumination conditions.

PCA of $\mathbf{X} \in \mathbb{R}^{N \times N_\mathrm{i} N_\mathrm{p}}$ is still possible, but two diversities "get entangled."

Idea: Disentagle them via a tensor structured basis

$$\mathbf{X} \approx \mathbf{U}\,\mathbf{S}\left(\mathbf{U}^{(\mathrm{i})} \boxtimes \mathbf{U}^{(\mathrm{p})}\right)^{\mathsf{T}}.$$

Def: Kronecker product $\mathbf{A} \boxtimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$

Extension to multiple diversities

Consider next N-dim face images of N_p people under N_i illumination conditions.

PCA of $\mathbf{X} \in \mathbb{R}^{N \times N_\mathrm{i} N_\mathrm{p}}$ is still possible, but two diversities "get entangled."

Idea: Disentagle them via a tensor structured basis

$$\mathbf{X} \approx \mathbf{U}\,\mathbf{S}\left(\mathbf{U}^{(\mathrm{i})} \boxtimes \mathbf{U}^{(\mathrm{p})}\right)^{\mathsf{T}}.$$

Def: Kronecker product $\mathbf{A} \boxtimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$

In tensor notation: the data $\mathsf{X} \in \mathbb{R}^{N \times N_\mathrm{p} \times N_\mathrm{i}}$ is modeled as

$$\mathsf{X} \approx (\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}) \cdot \mathsf{S} := \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} (\mathbf{U})_{:r_1} \otimes (\mathbf{U}^{(\mathrm{p})})_{:r_2} \otimes (\mathbf{U}^{(\mathrm{i})})_{:r_3}$$

Multilinear PCA

We can see $(\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}) \cdot \mathsf{S}$ as a contraction of S with $\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}$:

$$x_{npi} \approx \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\mathbf{U})_{n r_1} (\mathbf{U}^{(\mathrm{p})})_{p r_2} (\mathbf{U}^{(\mathrm{i})})_{i r_3}.$$

Multilinear PCA

We can see $(\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}) \cdot \mathsf{S}$ as a contraction of S with $\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}$:

$$x_{npi} \approx \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\mathbf{U})_{nr_1} (\mathbf{U}^{(\mathrm{p})})_{pr_2} (\mathbf{U}^{(\mathrm{i})})_{ir_3}.$$

Hence, the partial contraction $(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S} \in \mathbb{R}^{N \times R_2 \times R_3}$ given by

$$[(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S}]_{nr_2 r_3} = \sum_{r_1} s_{r_1 r_2 r_3} (\mathbf{U})_{nr_1}$$

yields a tensor having $R_2 R_3$ images which represent R_2 "people patterns" $\times$ R_3 "illumination patterns."

Multilinear PCA

We can see $(\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}) \cdot \mathsf{S}$ as a contraction of S with $\mathbf{U}, \mathbf{U}^{(\mathrm{p})}, \mathbf{U}^{(\mathrm{i})}$:

$$x_{npi} \approx \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\mathbf{U})_{nr_1} (\mathbf{U}^{(\mathrm{p})})_{pr_2} (\mathbf{U}^{(\mathrm{i})})_{ir_3}.$$

Hence, the partial contraction $(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S} \in \mathbb{R}^{N \times R_2 \times R_3}$ given by

$$[(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S}]_{nr_2 r_3} = \sum_{r_1} s_{r_1 r_2 r_3} (\mathbf{U})_{nr_1}$$

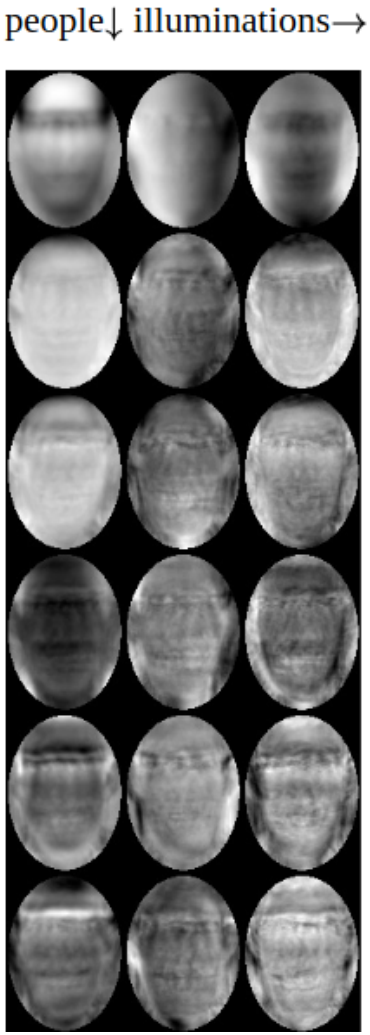
yields a tensor having $R_2 R_3$ images which represent R_2 "people patterns" $\times$ R_3 "illumination patterns."

Further contraction with the pth row of $\mathbf{U}^{(\mathrm{p})}$ and the ith row of $\mathbf{U}^{(\mathrm{i})}$ gives the approximate image of person p under illumination condition i:

$$(\mathsf{X})_{:pi} \approx \sum_{r_2 r_3} \underbrace{\left(\sum_{r_1} s_{r_1 r_2 r_3} (\mathbf{U})_{:r_1} \right)}_{\text{image pattern } (r_2, r_3)} (\mathbf{U}^{(\mathrm{p})})_{pr_2} (\mathbf{U}^{(\mathrm{i})})_{ir_3}.$$

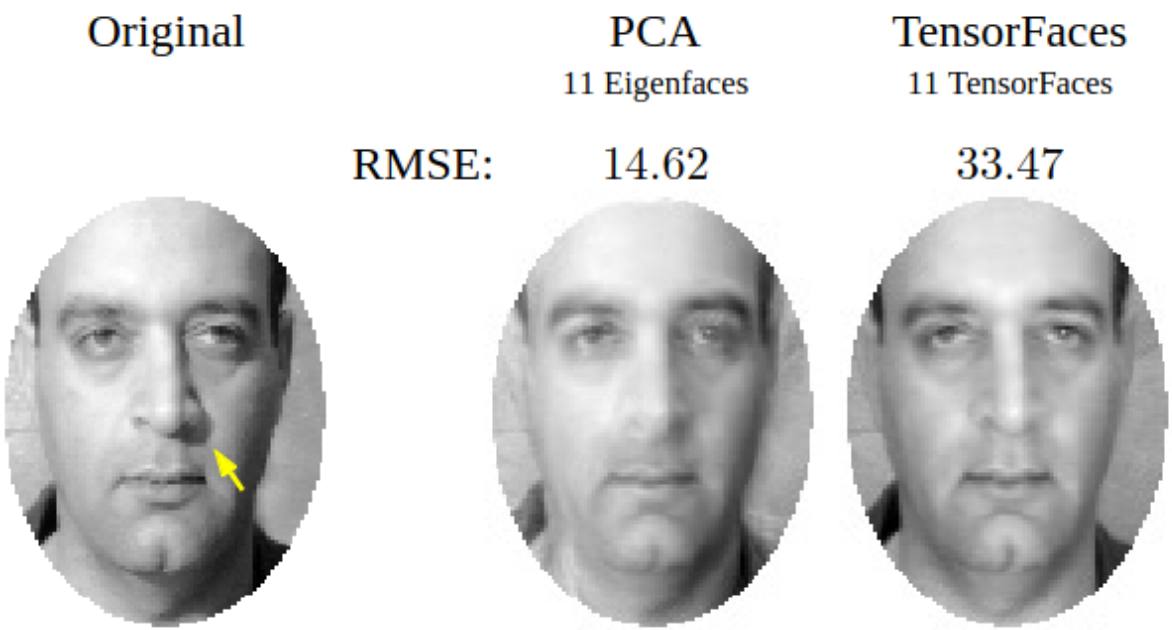
TensorFaces

This idea forms the basis of the TensorFaces approach[1].



A few "eigenmodes."[1]

It allows in particular a "strategic" dim. reduction w.r.t. some chosen diversities (e.g., illumination only, via a small R_3).



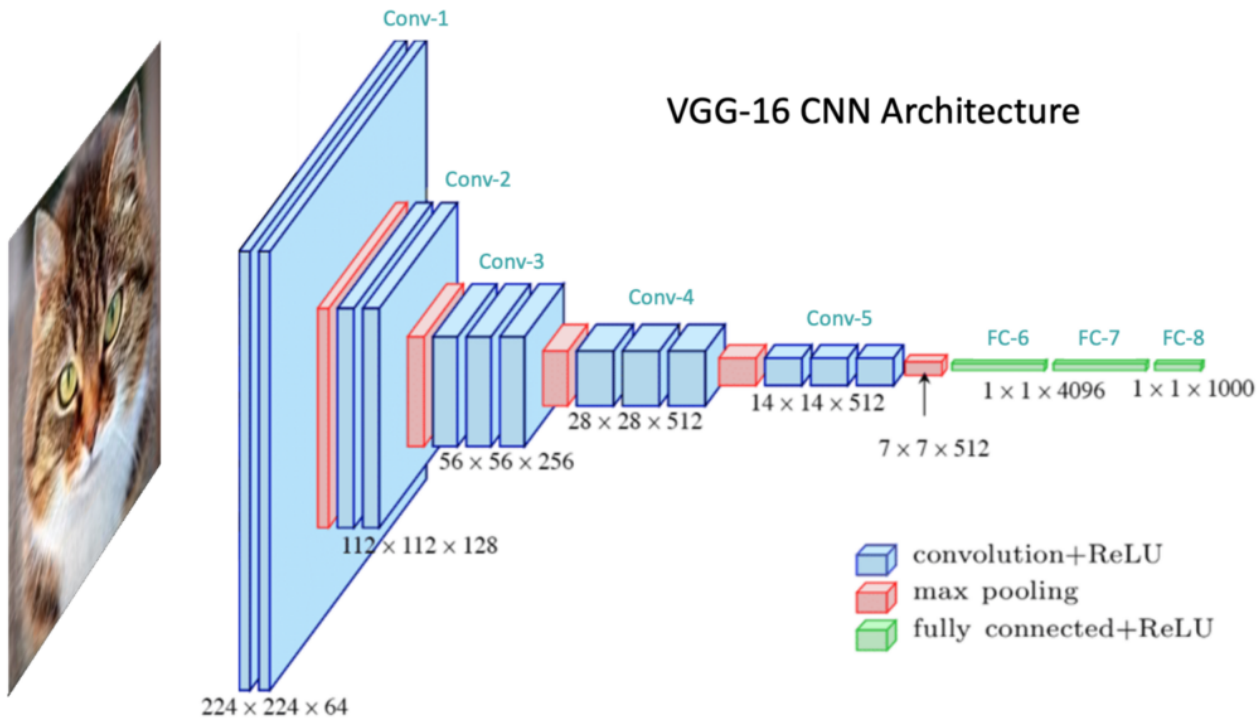
Dimension reduction:[1] *PCA vs TensorFaces with* $(R_1, R_2, R_3) = (176, 11, 1)$. *(Original size:* $7943 \times 11 \times 16$*.)*

1: Vasilescu & Terzopoulos, 2002

Example #4: Deep learning (of course!)

Numerous deep learning models are parameterized by tensors.

Typical example: CNNs parameterized by convolution kernels.



A typical CNN architecture (source: learnopencv.com).

Kernel of `conv2D` layer with $H \times W$ filters, I input channels and O output channels:

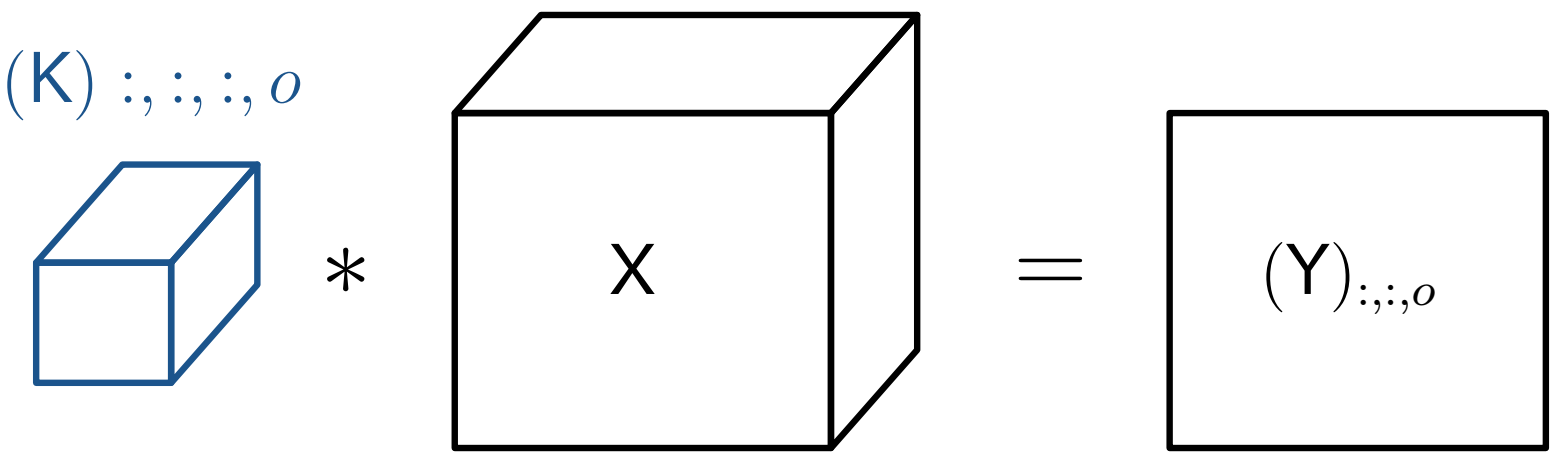
$$\mathsf{K} \in \mathbb{R}^{H \times W \times I \times O}.$$

2D convolution in CNNs

conv2D maps $\mathsf{X} \in \mathbb{R}^{M \times N \times I}$ into $\mathsf{Y} \in \mathbb{R}^{M \times N \times O}$ according to

$$y_{mno} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \sum_{i=1}^{I} (\mathsf{K})_{hwio}\, x_{m+h,n+w,i}.$$

Per output channel o:



Complexity: $\mathrm{O}(MNOHWI)$.

Most expensive stage of inference: performing the required convolutions.

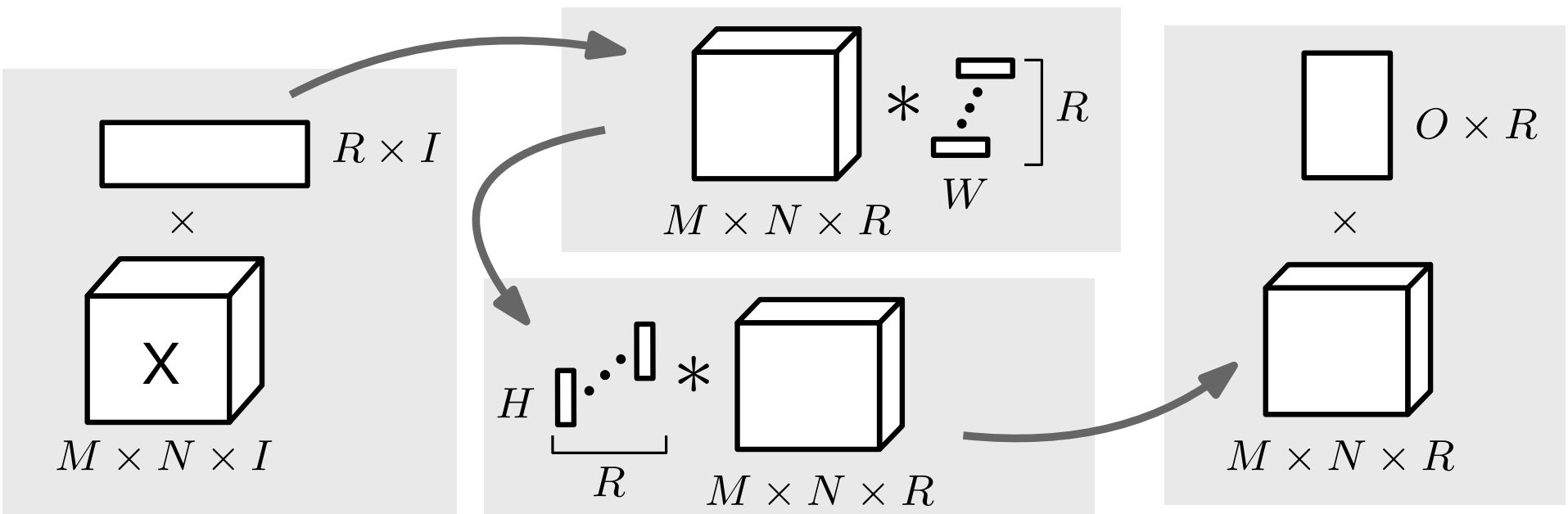
Decomposing convolution kernels

Idea: decompose K into separable terms using $R(H + W + I + O)$ params,[1]

$$(\mathsf{K})_{hwio} = \sum_{r=1}^{R} a_{hr}\, b_{wr}\, c_{ir}\, d_{or},$$

$$y_{mno} = \sum_{r=1}^{R} d_{or} \sum_{h=0}^{H-1} a_{hr} \sum_{w=0}^{W-1} b_{wr} \sum_{i=1}^{I} c_{ir}\, x_{m+h,w+n,i}.$$

Complexity: $\mathrm{O}(MNR(H + W + I + O))$

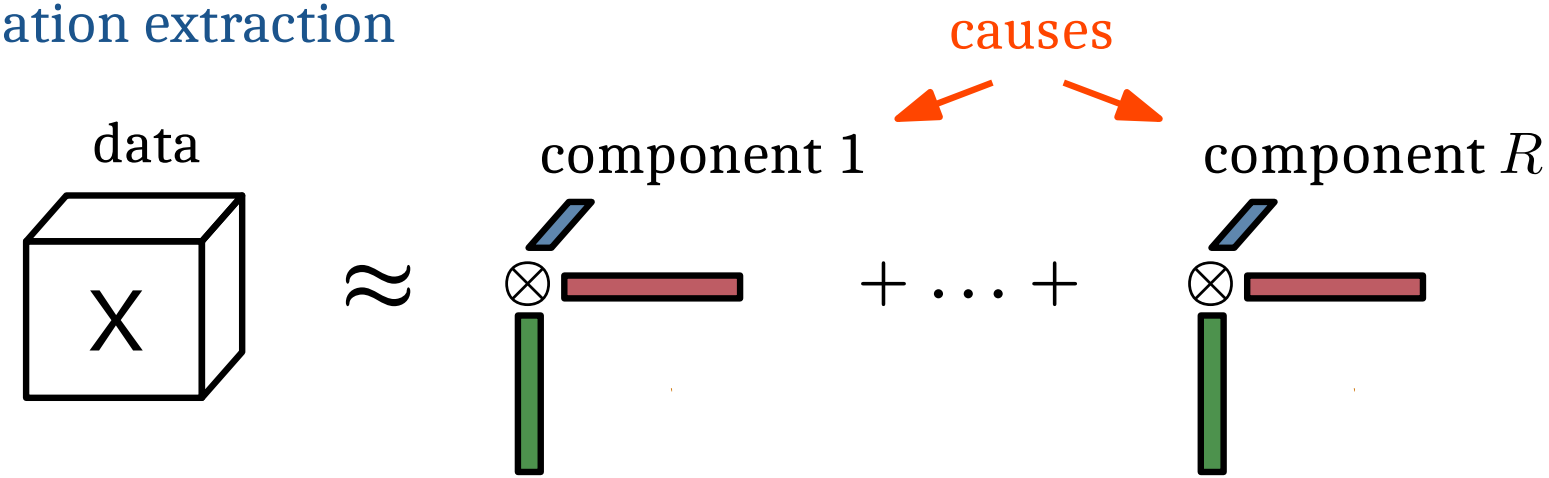


1: Lebedev & al., 2015

Summary : two main motivations (1/2)

Tensor models generally serve two main purposes:

1. Information extraction



$$(\mathsf{X})_{ijk} \approx \sum_{r=1}^{R} (\boldsymbol{a}_r)_i (\boldsymbol{b}_r)_j (\boldsymbol{c}_r)_k$$

data = observations or functions thereof, R "small", "simple" components

Ex: exploratory data analysis,[1] source separation, latent variable model estimation, ...

1: Hong & al., 2020

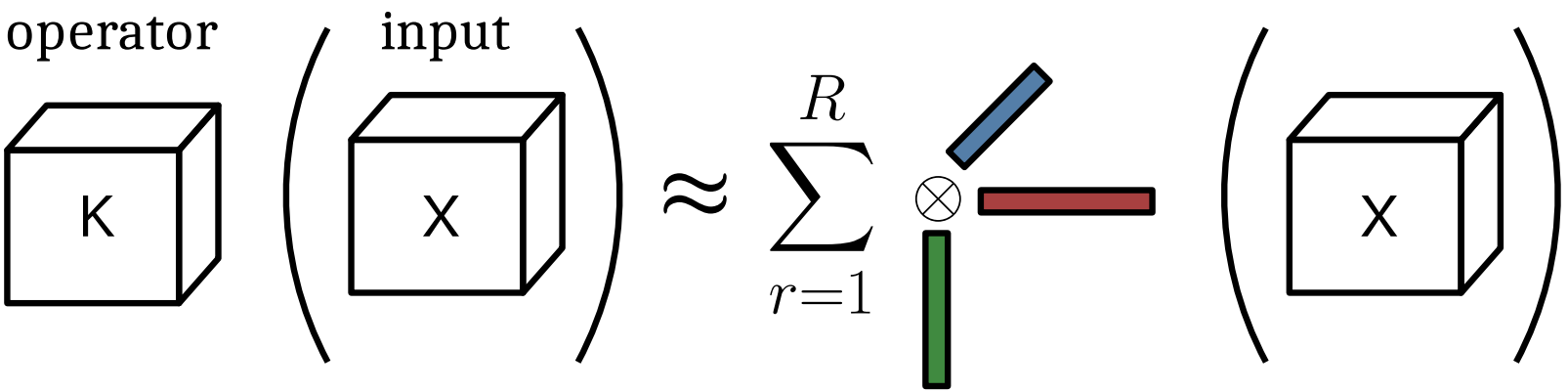
Summary : two main motivations (2/2)

2. Complexity reduction

- Dimensionality reduction with tensor structured basis

$$\mathsf{X} \approx (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \cdot \mathsf{S}$$

- Operator $\approx$ operator with "low-rank structure"



Ex: nonlinear system modelling,[1,2] deep learning, multivariate density approximation,[3] ...

1: Favier & al., 2012, 2: Goulart & Burt, 2022, 3: Dolgov & al., 2020

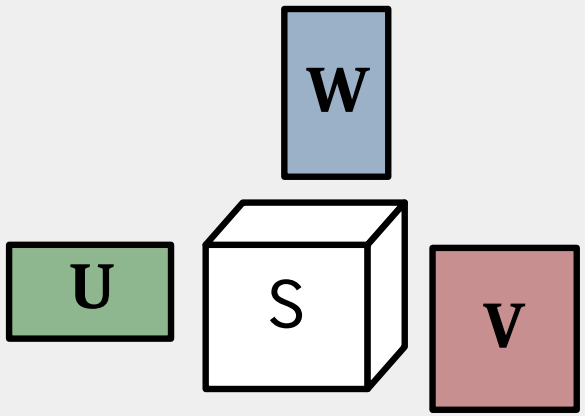
A jungle of tensor models

Let’s start with some useful notation & definitions.

Multilinear transformation

Def: Given $\mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and $\mathbf{U} \in \mathbb{R}^{N_1 \times R_1}$, $\mathbf{V} \in \mathbb{R}^{N_2 \times R_2}$ and $\mathbf{W} \in \mathbb{R}^{N_3 \times R_3}$, the multilinear transformation

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

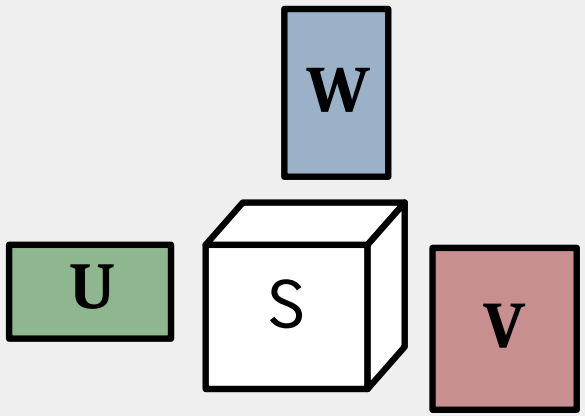


is defined as $$[(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}]_{n_1 n_2 n_3} = \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\mathbf{U})_{n_1 r_1} (\mathbf{V})_{n_2 r_2} (\mathbf{W})_{n_3 r_3}.$$

Multilinear transformation

Def: Given $\mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and $\mathbf{U} \in \mathbb{R}^{N_1 \times R_1}$, $\mathbf{V} \in \mathbb{R}^{N_2 \times R_2}$ and $\mathbf{W} \in \mathbb{R}^{N_3 \times R_3}$, the multilinear transformation

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$



is defined as $$[(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}]_{n_1 n_2 n_3} = \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\mathbf{U})_{n_1 r_1} (\mathbf{V})_{n_2 r_2} (\mathbf{W})_{n_3 r_3}.$$

Notational conventions:

(i) replacing any matrix by a dot $\cdot$ means no transformation in that mode

$$[(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S}]_{n_1 r_2 r_3} = \sum_{r_1} s_{r_1 r_2 r_3} (\mathbf{U})_{n_1 r_1}$$

(ii) the transpose is omitted when transformation by vectors are performed

$$(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \cdot \mathsf{S} = \sum_{r_1 r_2 r_3} s_{r_1 r_2 r_3} (\boldsymbol{u})_{r_1} (\boldsymbol{v})_{r_2} (\boldsymbol{w})_{r_3}.$$

Multilinear transformation: properties

(i) Expansion in terms of a tensor basis:

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} \, \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{w}_{r_3}$$

(ii) Action on rank-1 tensors:

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} = (\mathbf{U}\boldsymbol{a}) \otimes (\mathbf{V}\boldsymbol{b}) \otimes (\mathbf{W}\boldsymbol{c})$$

Hence, by linearity,

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \left(\sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r \right) = \sum_{r=1}^{R} (\mathbf{U}\boldsymbol{a}_r) \otimes (\mathbf{V}\boldsymbol{b}_r) \otimes (\mathbf{W}\boldsymbol{c}_r)$$

(iii) Composition:

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot [(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathsf{S}] = (\mathbf{UX}, \mathbf{VY}, \mathbf{WZ}) \cdot \mathsf{S}$$
$$(\mathbf{U}, \cdot, \cdot) \cdot [(\cdot, \mathbf{V}, \cdot) \cdot \mathsf{S}] = (\cdot, \mathbf{V}, \cdot) \cdot [(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{S}]$$

Scalar product and Euclidean norm

Def: Scalar product

$$\langle \mathsf{X}, \mathsf{Y} \rangle = \sum_{ijk} x_{ijk}\, y_{ijk}$$

In particular:

$$\begin{cases} \langle \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}, \mathsf{Y} \rangle = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \cdot \mathsf{Y} \\ \langle \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle = \langle \boldsymbol{a}, \boldsymbol{u} \rangle \langle \boldsymbol{b}, \boldsymbol{v} \rangle \langle \boldsymbol{c}, \boldsymbol{w} \rangle \end{cases}$$

Def: The Frobenius norm of W is defined as $\|\mathsf{W}\|_{\mathsf{F}} := \sqrt{\langle \mathsf{W}, \mathsf{W} \rangle} = \sqrt{\sum_{ijk} w_{ijk}^2}$

1: Cichocki & al., 2016

Natural (& useful) isomorphisms

Idea: "view" tensors through the lens of certain isomorphisms with matrix and vector spaces. Useful for both analysis & computation.

(Two finite-dim vector spaces are isomorphic ($\simeq$) iff they have the same dim.)

Natural (& useful) isomorphisms

Idea: "view" tensors through the lens of certain isomorphisms with matrix and vector spaces. Useful for both analysis & computation.

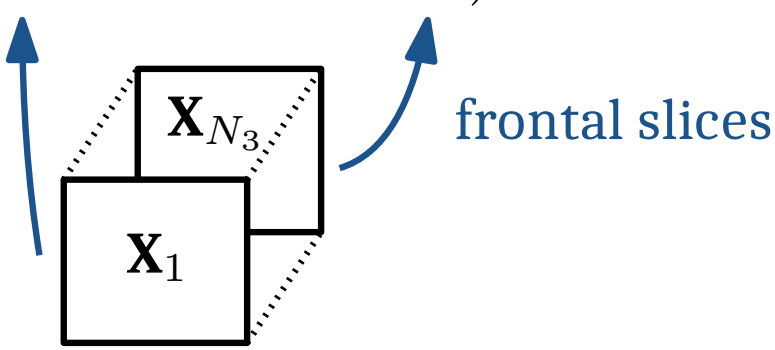
(Two finite-dim vector spaces are isomorphic ($\simeq$) iff they have the same dim.)

In particular, the isomorphism

$$\mathbb{R}^{N_1 \times N_2 \times N_3} \simeq \mathbb{R}^{N_1 \times N_3 N_2}.$$

leads to the notion of a mode-1 unfolding

$$\mathcal{X} \mapsto \mathbf{X}_{(1)} := \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_{N_3} \end{pmatrix} \in \mathbb{R}^{N_1 \times N_3 N_2}$$



Concretely, x_{ijk} is the element $(i, n(j,k))$ of $\mathbf{X}_{(1)}$, with $n(j,k) := (k-1)N_2 + j$.

Analogous definitions can be given to $\mathbf{X}_{(2)} \in \mathbb{R}^{N_2 \times N_3 N_1}$ and $\mathbf{X}_{(3)} \in \mathbb{R}^{N_3 \times N_2 N_1}$.

Natural (& useful) isomorphisms (cont'd)

Ex: If $\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$, then:

$$\mathbf{X}_{(1)} = \boldsymbol{a}\,(\boldsymbol{c} \boxtimes \boldsymbol{b})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}\,(\boldsymbol{c} \boxtimes \boldsymbol{a})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}\,(\boldsymbol{b} \boxtimes \boldsymbol{a})^{\mathsf{T}}.$$

Natural (& useful) isomorphisms (cont'd)

Ex: If $\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$, then:

$$\mathbf{X}_{(1)} = \boldsymbol{a}\,(\boldsymbol{c} \boxtimes \boldsymbol{b})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}\,(\boldsymbol{c} \boxtimes \boldsymbol{a})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}\,(\boldsymbol{b} \boxtimes \boldsymbol{a})^{\mathsf{T}}.$$

Ex: If $\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}$, then:

$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{S}_{(2)}\,(\mathbf{W} \boxtimes \mathbf{U})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{S}_{(3)}\,(\mathbf{V} \boxtimes \mathbf{U})^{\mathsf{T}}.$$

Natural (& useful) isomorphisms (cont'd)

Ex: If $\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$, then:

$$\mathbf{X}_{(1)} = \boldsymbol{a}\,(\boldsymbol{c} \boxtimes \boldsymbol{b})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}\,(\boldsymbol{c} \boxtimes \boldsymbol{a})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}\,(\boldsymbol{b} \boxtimes \boldsymbol{a})^{\mathsf{T}}.$$

Ex: If $\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}$, then:

$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{S}_{(2)}\,(\mathbf{W} \boxtimes \mathbf{U})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{S}_{(3)}\,(\mathbf{V} \boxtimes \mathbf{U})^{\mathsf{T}}.$$

Rmk: In practice, $\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}$ is computed by a sequence of matrix-matrix products and unfoldings/foldings.

Natural (& useful) isomorphisms (cont'd)

Ex: If $\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$, then:

$$\mathbf{X}_{(1)} = \boldsymbol{a}\,(\boldsymbol{c} \boxtimes \boldsymbol{b})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}\,(\boldsymbol{c} \boxtimes \boldsymbol{a})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}\,(\boldsymbol{b} \boxtimes \boldsymbol{a})^{\mathsf{T}}.$$

Ex: If $\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}$, then:

$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{S}_{(2)}\,(\mathbf{W} \boxtimes \mathbf{U})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{S}_{(3)}\,(\mathbf{V} \boxtimes \mathbf{U})^{\mathsf{T}}.$$

Rmk: In practice, $\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}$ is computed by a sequence of matrix-matrix products and unfoldings/foldings.

Another commonly used isomorphism is the vectorization operation

$$\mathbb{R}^{N_1 \times N_2 \times N_3} \ni \mathsf{X} \mapsto \boldsymbol{x} = \operatorname{vec} \mathsf{X} \in \mathbb{R}^{N_3 N_2 N_1},$$

such that X_{ijk} is "sent" to $x_{n(i,j,k)}$ with $n(i,j,k) := N_2 N_1 (k-1) + N_1 (j-1) + i$.

Ex: By vectorization,

$$\operatorname{vec}(\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}) = \boldsymbol{c} \boxtimes \boldsymbol{b} \boxtimes \boldsymbol{a},$$

$$\operatorname{vec}[(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}] = (\mathbf{W} \boxtimes \mathbf{V} \boxtimes \mathbf{U}) \operatorname{vec} \mathsf{S}.$$

...and now let's finally discuss low-rank tensor models.

What is the rank of a tensor?

For a matrix $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$,

$$\begin{aligned}\operatorname{rank} \mathbf{X} &= \dim \operatorname{colspan} \mathbf{X} = \dim \left\{ \mathbf{X}\boldsymbol{v} \ : \ \boldsymbol{v} \in \mathbb{R}^{N_2} \right\} \\ &= \dim \operatorname{rowspan} \mathbf{X} = \dim \left\{ \mathbf{X}^\mathsf{T}\boldsymbol{v} \ : \ \boldsymbol{v} \in \mathbb{R}^{N_1} \right\} \\ &= \min \left\{ R \ : \ \mathbf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \boldsymbol{b}_r^\mathsf{T} \right\}\end{aligned}$$

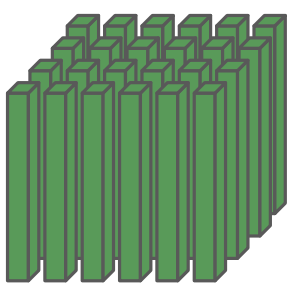
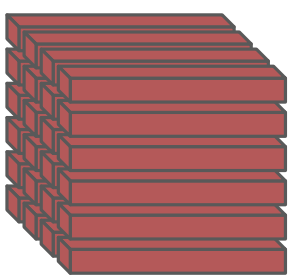
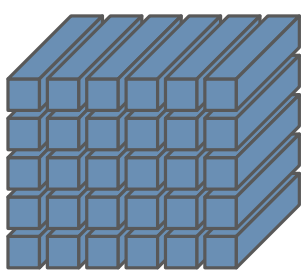
What is the rank of a tensor?

For a matrix $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$,

$$\begin{aligned} \operatorname{rank} \mathbf{X} &= \dim \operatorname{colspan} \mathbf{X} = \dim \left\{ \mathbf{X}\boldsymbol{v} \;:\; \boldsymbol{v} \in \mathbb{R}^{N_2} \right\} \\ &= \dim \operatorname{rowspan} \mathbf{X} = \dim \left\{ \mathbf{X}^\mathsf{T}\boldsymbol{v} \;:\; \boldsymbol{v} \in \mathbb{R}^{N_1} \right\} \\ &= \min \left\{ R \;:\; \mathbf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \boldsymbol{b}_r^\mathsf{T} \right\} \end{aligned}$$

This suggests (at least) two definitions for the rank of $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$:

1. Dimension(s) of subspaces spanned by the fibers of each mode:

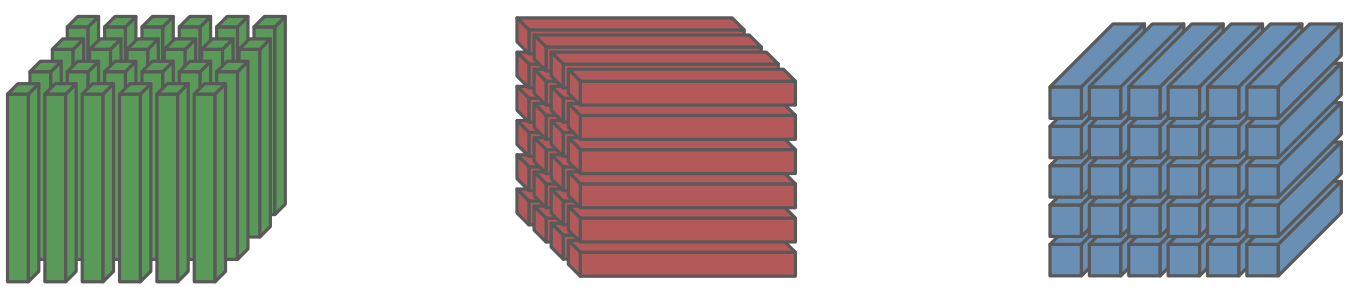
What is the rank of a tensor?

For a matrix $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$,

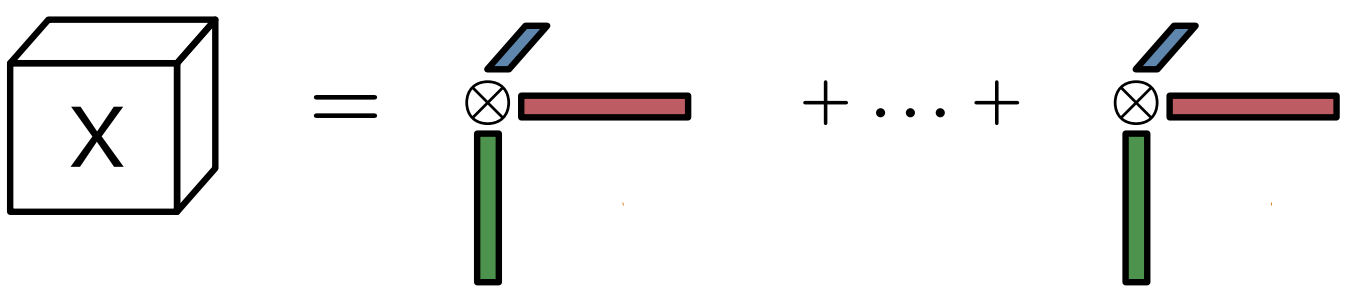
$$\begin{aligned} \operatorname{rank} \mathbf{X} &= \dim \operatorname{colspan} \mathbf{X} = \dim \left\{ \mathbf{X}\boldsymbol{v} \ : \ \boldsymbol{v} \in \mathbb{R}^{N_2} \right\} \\ &= \dim \operatorname{rowspan} \mathbf{X} = \dim \left\{ \mathbf{X}^{\mathsf{T}}\boldsymbol{v} \ : \ \boldsymbol{v} \in \mathbb{R}^{N_1} \right\} \\ &= \min \left\{ R \ : \ \mathbf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \boldsymbol{b}_r^{\mathsf{T}} \right\} \end{aligned}$$

This suggests (at least) two definitions for the rank of $\mathcal{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$:

1. Dimension(s) of subspaces spanned by the fibers of each mode:



2. Minimal number of rank-1 tensors needed to produce $\mathcal{X}$:



1. The multilinear rank

Recall that: fibers of X = columns of each $\mathbf{X}_{(i)}$.

1. The multilinear rank

Recall that: fibers of X = columns of each $\mathbf{X}_{(i)}$.

The first notion of rank can be formalized as

$$\begin{aligned}
\operatorname{rank}_1 \mathsf{X} &:= \dim \operatorname{span}\left\{(\cdot, \boldsymbol{v}, \boldsymbol{w}) \cdot \mathsf{X} \,:\, \boldsymbol{v} \in \mathbb{R}^{N_2}, \boldsymbol{w} \in \mathbb{R}^{N_3}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(1)}(\boldsymbol{w} \boxtimes \boldsymbol{v})\right\} = \operatorname{rank} \mathbf{X}_{(1)} \\
\operatorname{rank}_2 \mathsf{X} &:= \dim \operatorname{span}\left\{(\boldsymbol{u}, \cdot, \boldsymbol{w}) \cdot \mathsf{X} \,:\, \boldsymbol{u} \in \mathbb{R}^{N_1}, \boldsymbol{w} \in \mathbb{R}^{N_3}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(2)}(\boldsymbol{w} \boxtimes \boldsymbol{u})\right\} = \operatorname{rank} \mathbf{X}_{(2)} \\
\operatorname{rank}_3 \mathsf{X} &:= \dim \operatorname{span}\left\{(\boldsymbol{u}, \boldsymbol{v}, \cdot) \cdot \mathsf{X} \,:\, \boldsymbol{u} \in \mathbb{R}^{N_1}, \boldsymbol{v} \in \mathbb{R}^{N_2}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(3)}(\boldsymbol{v} \boxtimes \boldsymbol{u})\right\} = \operatorname{rank} \mathbf{X}_{(3)}
\end{aligned}$$

1. The multilinear rank

Recall that: fibers of X = columns of each $\mathbf{X}_{(i)}$.

The first notion of rank can be formalized as

$$\begin{aligned}
\operatorname{rank}_1 \mathsf{X} &:= \dim \operatorname{span}\left\{(\cdot, \boldsymbol{v}, \boldsymbol{w}) \cdot \mathsf{X} \,:\, \boldsymbol{v} \in \mathbb{R}^{N_2}, \boldsymbol{w} \in \mathbb{R}^{N_3}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(1)}(\boldsymbol{w} \boxtimes \boldsymbol{v})\right\} = \operatorname{rank} \mathbf{X}_{(1)} \\
\operatorname{rank}_2 \mathsf{X} &:= \dim \operatorname{span}\left\{(\boldsymbol{u}, \cdot, \boldsymbol{w}) \cdot \mathsf{X} \,:\, \boldsymbol{u} \in \mathbb{R}^{N_1}, \boldsymbol{w} \in \mathbb{R}^{N_3}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(2)}(\boldsymbol{w} \boxtimes \boldsymbol{u})\right\} = \operatorname{rank} \mathbf{X}_{(2)} \\
\operatorname{rank}_3 \mathsf{X} &:= \dim \operatorname{span}\left\{(\boldsymbol{u}, \boldsymbol{v}, \cdot) \cdot \mathsf{X} \,:\, \boldsymbol{u} \in \mathbb{R}^{N_1}, \boldsymbol{v} \in \mathbb{R}^{N_2}\right\} \\
&= \dim \operatorname{span}\left\{\mathbf{X}_{(3)}(\boldsymbol{v} \boxtimes \boldsymbol{u})\right\} = \operatorname{rank} \mathbf{X}_{(3)}
\end{aligned}$$

Def:[1] Multilinear rank: $\operatorname{mrank} \mathsf{X} := (\operatorname{rank}_1 \mathsf{X}, \operatorname{rank}_2 \mathsf{X}, \operatorname{rank}_3 \mathsf{X})$.

1: Hitchcock, 1928

Properties of mrank

(i) Unlike the matrix case, mrank components are generally distinct.

Ex: With

$$\mathcal{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \in \mathbb{R}^{2\times 2\times 2}, \qquad \begin{matrix} & \mathbf{X}_1 & & \mathbf{X}_2 \end{matrix}$$

$$\mathbf{X}_{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{X}_{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Properties of mrank

(i) Unlike the matrix case, mrank components are generally distinct.

Ex: With

$$\mathsf{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \in \mathbb{R}^{2\times 2\times 2}, \qquad \begin{matrix} & \mathbf{X}_1 & & \mathbf{X}_2 \end{matrix}$$

$$\mathbf{X}_{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{X}_{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

(ii) The mrank is a tensor property:

Prop: mrank X is invariant w.r.t. a change of basis

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{X}, \qquad \mathbf{U} \in \mathrm{GL}(N_1),\ \mathbf{V} \in \mathrm{GL}(N_2),\ \mathbf{W} \in \mathrm{GL}(N_3).$$

Proof: For any nonsingular $\mathbf{U} \in \mathbb{R}^{N_1 \times N_1}$ (and similarly for the other modes),

$$\operatorname{rank}_1[(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{X}] = \operatorname{rank}[\mathbf{U}\mathbf{X}_{(1)}] = \operatorname{rank} \mathbf{X}_{(1)} = \operatorname{rank}_1 \mathsf{X}.$$

■

Properties of mrank

(i) Unlike the matrix case, mrank components are generally distinct.

Ex: With

$$\mathsf{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \in \mathbb{R}^{2\times 2\times 2}, \qquad \begin{matrix} & \mathbf{X}_1 & & \mathbf{X}_2 \end{matrix}$$

$$\mathbf{X}_{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{X}_{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

(ii) The mrank is a tensor property:

Prop: mrank X is invariant w.r.t. a change of basis

$$(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{X}, \qquad \mathbf{U} \in \mathrm{GL}(N_1),\ \mathbf{V} \in \mathrm{GL}(N_2),\ \mathbf{W} \in \mathrm{GL}(N_3).$$

Proof: For any nonsingular $\mathbf{U} \in \mathbb{R}^{N_1 \times N_1}$ (and similarly for the other modes),

$$\operatorname{rank}_1[(\mathbf{U}, \cdot, \cdot) \cdot \mathsf{X}] = \operatorname{rank}[\mathbf{U}\mathbf{X}_{(1)}] = \operatorname{rank} \mathbf{X}_{(1)} = \operatorname{rank}_1 \mathsf{X}.$$

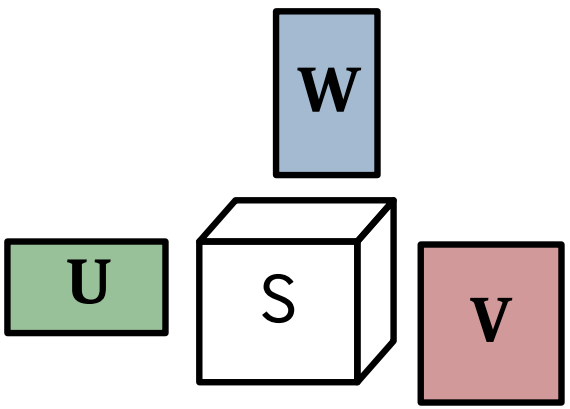
■

(iii) $\operatorname{mrank} \mathsf{X} \leq (\min\{N_1, N_2N_3\}, \min\{N_2, N_1N_3\}, \min\{N_3, N_1N_2\})$

The Tucker model

The Tucker model[1] has the form

$$\mathbb{R}^{N_1 \times N_2 \times N_3} \ni \quad \mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \quad \mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$$



with:

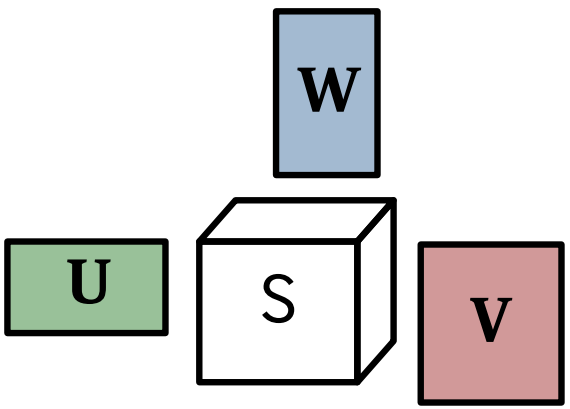
- $\mathbf{U} \in \mathbb{R}^{N_1 \times R_1}, R_1 \leq N_1$
- $\mathbf{V} \in \mathbb{R}^{N_2 \times R_2}, R_2 \leq N_2$
- $\mathbf{W} \in \mathbb{R}^{N_3 \times R_3}, R_3 \leq N_3$

1: Tucker, 1966

The Tucker model

The Tucker model[1] has the form

$$\mathbb{R}^{N_1 \times N_2 \times N_3} \ni \quad \mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \quad \mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$$



with:

- $\mathbf{U} \in \mathbb{R}^{N_1 \times R_1}, R_1 \leq N_1$
- $\mathbf{V} \in \mathbb{R}^{N_2 \times R_2}, R_2 \leq N_2$
- $\mathbf{W} \in \mathbb{R}^{N_3 \times R_3}, R_3 \leq N_3$

Recall that

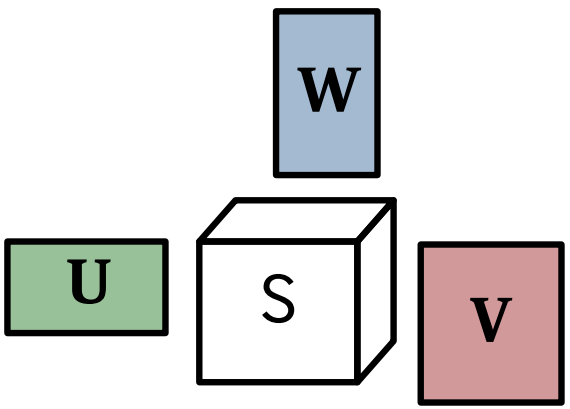
$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^\mathsf{T}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{S}_{(2)}\,(\mathbf{W} \boxtimes \mathbf{U})^\mathsf{T}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{S}_{(3)}\,(\mathbf{V} \boxtimes \mathbf{U})^\mathsf{T}$$

1: Tucker, 1966

The Tucker model

The Tucker model[1] has the form

$$\mathbb{R}^{N_1 \times N_2 \times N_3} \ni \quad \mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \quad \mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$$



with:

- $\mathbf{U} \in \mathbb{R}^{N_1 \times R_1}, R_1 \leq N_1$
- $\mathbf{V} \in \mathbb{R}^{N_2 \times R_2}, R_2 \leq N_2$
- $\mathbf{W} \in \mathbb{R}^{N_3 \times R_3}, R_3 \leq N_3$

Recall that

$$\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^{\mathsf{T}}$$
$$\mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{S}_{(2)}\,(\mathbf{W} \boxtimes \mathbf{U})^{\mathsf{T}}$$
$$\mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{S}_{(3)}\,(\mathbf{V} \boxtimes \mathbf{U})^{\mathsf{T}}$$

Clearly, $\operatorname{mrank} \mathsf{X} \leq (R_1, R_2, R_3)$ component-wise.

Hence, useful for a low-mrank representation of a tensor (multilinear PCA).

1: Tucker, 1966

Orthogonal Tucker model

The Tucker model is a subspace representation of a tensor w.r.t. the tensor basis[1]

$$\{\boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{w}_{r_3} \ : \ r_i \in [R_i], 1 \leq i \leq 3\},$$

since

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} \, \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{u}_{r_3}.$$

Orthogonal Tucker model

The Tucker model is a subspace representation of a tensor w.r.t. the tensor basis[1]

$$\{\boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{w}_{r_3} \ : \ r_i \in [R_i], 1 \leq i \leq 3\},$$

since

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} s_{r_1 r_2 r_3} \, \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{u}_{r_3}.$$

Thus, it's natural (and w.l.o.g.) to impose $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$, and similarly for $\mathbf{V}$ and $\mathbf{W}$.

Indeed, if $\mathsf{X} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{S}'$, then we can find (semi-)orthogonal $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ s.t.

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{S}' = (\mathbf{U}\mathbf{R}_1, \mathbf{V}\mathbf{R}_2, \mathbf{W}\mathbf{R}_3) \cdot \mathsf{S}' = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \left[(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3) \cdot \mathsf{S}'\right] = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} \cdot$$

In particular, semi-orthogonality of the factors $\mathbf{U}, \mathbf{V}, \mathbf{W}$ implies

$$\mathsf{S} = (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}.$$

1: Hackbusch, 2012

High-order singular value decomp. (HOSVD)

But even under this constraint, we have a continuum of possible $(\mathbf{U}, \mathbf{V}, \mathbf{W})$, since $\forall(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3) \in \mathbb{O}(R_1) \times \mathbb{O}(R_2) \times \mathbb{O}(R_3)$,

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} = (\mathbf{U}\mathbf{Q}_1, \mathbf{V}\mathbf{Q}_2, \mathbf{W}\mathbf{Q}_3) \cdot \left[(\mathbf{Q}_1^\mathsf{T}, \mathbf{Q}_2^\mathsf{T}, \mathbf{Q}_3^\mathsf{T}) \cdot \mathsf{S}\right] = (\mathbf{U}', \mathbf{V}', \mathbf{W}') \cdot \mathsf{S}'.$$

A standard choice (useful for low-mrank approximation, as we will see) is:

High-order singular value decomp. (HOSVD)

But even under this constraint, we have a continuum of possible $(\mathbf{U}, \mathbf{V}, \mathbf{W})$, since $\forall (\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3) \in \mathbb{O}(R_1) \times \mathbb{O}(R_2) \times \mathbb{O}(R_3)$,

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} = (\mathbf{U}\mathbf{Q}_1, \mathbf{V}\mathbf{Q}_2, \mathbf{W}\mathbf{Q}_3) \cdot \left[(\mathbf{Q}_1^\mathsf{T}, \mathbf{Q}_2^\mathsf{T}, \mathbf{Q}_3^\mathsf{T}) \cdot \mathsf{S} \right] = (\mathbf{U}', \mathbf{V}', \mathbf{W}') \cdot \mathsf{S}'.$$

A standard choice (useful for low-mrank approximation, as we will see) is:

Def: The HOSVD of $\mathsf{X} \in \mathbb{R}^{N_1 \times N_3 \times N_3}$ is given by[1]

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \underbrace{\mathsf{S}}_{\text{core tensor}},$$

where

- $\mathbf{U} \in \mathbb{O}(N_1)$ contains the left singular vectors of $\mathbf{X}_{(1)}$
- $\mathbf{V} \in \mathbb{O}(N_2)$ contains the left singular vectors of $\mathbf{X}_{(2)}$
- $\mathbf{W} \in \mathbb{O}(N_3)$ contains the left singular vectors of $\mathbf{X}_{(3)}$

1: De Lathauwer & al., 2000a

HOSVD : computation & properties

Computation: Given X, one computes (possibly in parallel) the SVD of each unfolding:

$$\begin{aligned}\mathbf{X}_{(1)} &= \mathbf{U}\,\boldsymbol{\Sigma}_1\,\mathbf{Q}_1^{\mathsf{T}},\\ \mathbf{X}_{(2)} &= \mathbf{V}\,\boldsymbol{\Sigma}_2\,\mathbf{Q}_2^{\mathsf{T}},\\ \mathbf{X}_{(3)} &= \mathbf{W}\,\boldsymbol{\Sigma}_3\,\mathbf{Q}_3^{\mathsf{T}},\end{aligned}$$

and then $\mathsf{S} = (\mathbf{U}^{\mathsf{T}}, \mathbf{V}^{\mathsf{T}}, \mathbf{W}^{\mathsf{T}}) \cdot \mathsf{X}$.

HOSVD : computation & properties

Computation: Given X, one computes (possibly in parallel) the SVD of each unfolding:

$$\begin{aligned}\mathbf{X}_{(1)} &= \mathbf{U}\,\mathbf{\Sigma}_1\,\mathbf{Q}_1^\mathsf{T},\\ \mathbf{X}_{(2)} &= \mathbf{V}\,\mathbf{\Sigma}_2\,\mathbf{Q}_2^\mathsf{T},\\ \mathbf{X}_{(3)} &= \mathbf{W}\,\mathbf{\Sigma}_3\,\mathbf{Q}_3^\mathsf{T},\end{aligned}$$

and then $\mathsf{S} = (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}$.

Core slice norms: Since $\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^\mathsf{T}$, we see that

$$\mathbf{S}_{(1)} = \mathbf{\Sigma}_1\,\mathbf{Q}_1^\mathsf{T}\,(\mathbf{W} \boxtimes \mathbf{V}) \quad \Rightarrow \quad \left\|\left(\mathbf{S}_{(1)}\right)_{\ell:}\right\| = (\mathbf{\Sigma}_1)_{\ell\ell} = \|(\mathsf{S})_{\ell::}\|_\mathsf{F}$$

and thus $\|(\mathsf{S})_{\ell::}\|_\mathsf{F} \geq \|(\mathsf{S})_{\ell+1::}\|_\mathsf{F}$. Similarly, $\|(\mathsf{S})_{:\ell:}\|_\mathsf{F} \geq \|(\mathsf{S})_{:\ell+1:}\|_\mathsf{F}$ and $\|(\mathsf{S})_{::\ell}\|_\mathsf{F} \geq \|(\mathsf{S})_{::\ell+1}\|_\mathsf{F}$.

HOSVD : computation & properties

Computation: Given X, one computes (possibly in parallel) the SVD of each unfolding:

$$\begin{aligned}\mathbf{X}_{(1)} &= \mathbf{U}\,\mathbf{\Sigma}_1\,\mathbf{Q}_1^{\mathsf{T}},\\ \mathbf{X}_{(2)} &= \mathbf{V}\,\mathbf{\Sigma}_2\,\mathbf{Q}_2^{\mathsf{T}},\\ \mathbf{X}_{(3)} &= \mathbf{W}\,\mathbf{\Sigma}_3\,\mathbf{Q}_3^{\mathsf{T}},\end{aligned}$$

and then $\mathsf{S} = (\mathbf{U}^{\mathsf{T}}, \mathbf{V}^{\mathsf{T}}, \mathbf{W}^{\mathsf{T}}) \cdot \mathsf{X}$.

Core slice norms: Since $\mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{S}_{(1)}\,(\mathbf{W} \boxtimes \mathbf{V})^{\mathsf{T}}$, we see that

$$\mathbf{S}_{(1)} = \mathbf{\Sigma}_1\,\mathbf{Q}_1^{\mathsf{T}}\,(\mathbf{W} \boxtimes \mathbf{V}) \quad \Rightarrow \quad \left\|\left(\mathbf{S}_{(1)}\right)_{\ell:}\right\| = (\mathbf{\Sigma}_1)_{\ell\ell} = \|(\mathsf{S})_{\ell::}\|_{\mathsf{F}}$$

and thus $\|(\mathsf{S})_{\ell::}\|_{\mathsf{F}} \geq \|(\mathsf{S})_{\ell+1::}\|_{\mathsf{F}}$. Similarly, $\|(\mathsf{S})_{:\ell:}\|_{\mathsf{F}} \geq \|(\mathsf{S})_{:\ell+1:}\|_{\mathsf{F}}$ and $\|(\mathsf{S})_{::\ell}\|_{\mathsf{F}} \geq \|(\mathsf{S})_{::\ell+1}\|_{\mathsf{F}}$.

Uniqueness: The HOSVD is unique (up to the signs of the columns of $\mathbf{U}, \mathbf{V}, \mathbf{W}$) provided that no $\mathbf{X}_{(i)}$ has repeated singular values.

2. The tensor rank

The second notion of rank is defined as:

Def: Rank of X:

$$\operatorname{rank}\mathsf{X} := \min\left\{ R \in \mathbb{N} \,:\, \mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r,\ \boldsymbol{a}_r \in \mathbb{R}^{N_1}, \boldsymbol{b}_r \in \mathbb{R}^{N_2}, \boldsymbol{c}_r \in \mathbb{R}^{N_3} \right.$$

The above set is non-empty: a trivial upper bound is $R \leq N_1 N_2 N_3$, since

$$\mathsf{X} = \sum_{n_1,n_2,n_3} (\mathsf{X})_{n_1 n_2 n_3} \underbrace{\boldsymbol{e}_{n_1} \otimes \boldsymbol{e}_{n_2} \otimes \boldsymbol{e}_{n_3}}_{\text{nonzero } (=1) \text{ only at } (n_1,n_2,n_3)}.$$

1: Hitchcock, 1927

The tensor rank is a tensor property

Prop: rank X is invariant w.r.t. a change of basis.

Proof: If $\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$ has rank R, then

$$\operatorname{rank}\left[(\mathbf{U},\mathbf{V},\mathbf{W})\cdot\mathsf{X}\right] = \operatorname{rank}\left[\sum_{r=1}^{R}(\mathbf{U}\boldsymbol{a}_r)\otimes(\mathbf{V}\boldsymbol{b}_r)\otimes(\mathbf{W}\boldsymbol{c}_r)\right] \leq R.$$

It follows that, for $\mathbf{U}\in\mathrm{GL}(N_1), \mathbf{V}\in\mathrm{GL}(N_2), \mathbf{W}\in\mathrm{GL}(N_3)$:

$$\begin{aligned}\operatorname{rank}\mathsf{X} &= \operatorname{rank}\left\{(\mathbf{U}^{-1},\mathbf{V}^{-1},\mathbf{W}^{-1})\cdot\left[(\mathbf{U},\mathbf{V},\mathbf{W})\cdot\mathsf{X}\right]\right\}\\ &\leq \operatorname{rank}((\mathbf{U},\mathbf{V},\mathbf{W})\cdot\mathsf{X})\\ &\leq \operatorname{rank}\mathsf{X}.\end{aligned}$$

■

More properties will be seen in a moment.

The polyadic decomposition

A polyadic decomposition[1] (PD) of X of rank R has the form

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R := \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r .$$

1: Hitchcock, 1927

The polyadic decomposition

A polyadic decomposition[1] (PD) of X of rank R has the form

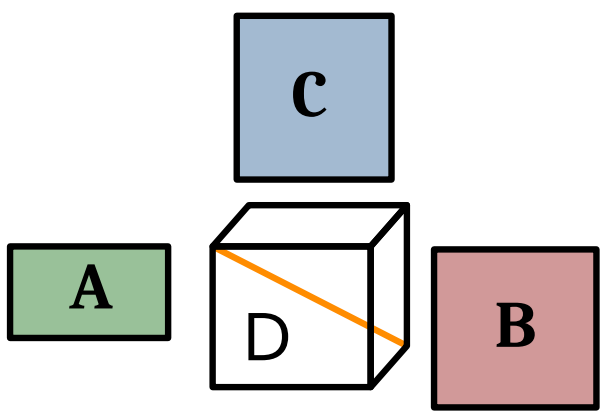
$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R := \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r .$$

It can be see as a special case of Tucker: by defining

$$\mathsf{D} \in \mathbb{R}^{R \times R \times R}, \quad (\mathsf{D})_{r_1 r_2 r_3} = \delta_{r_1 r_2 r_3},$$

we get

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D} = \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \sum_{r_3=1}^{R} (\mathsf{D})_{r_1 r_2 r_3} \, \boldsymbol{a}_{r_1} \otimes \boldsymbol{b}_{r_2} \otimes \boldsymbol{c}_{r_3} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$



1: Hitchcock, 1927

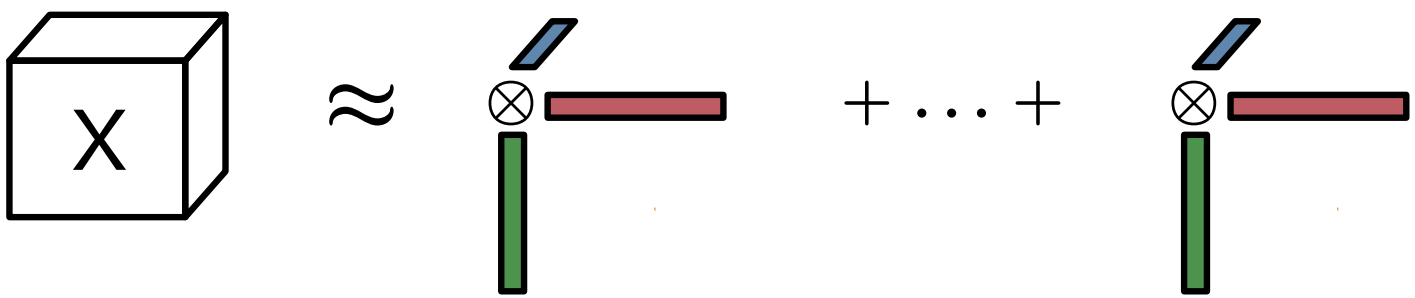
The canonical PD (CPD)

Def: If

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$

is minimal ($\operatorname{rank} \mathsf{X} = R$), then it is called the CPD of X.

It is also often called a PARAFAC (for parallel factors) decomposition.



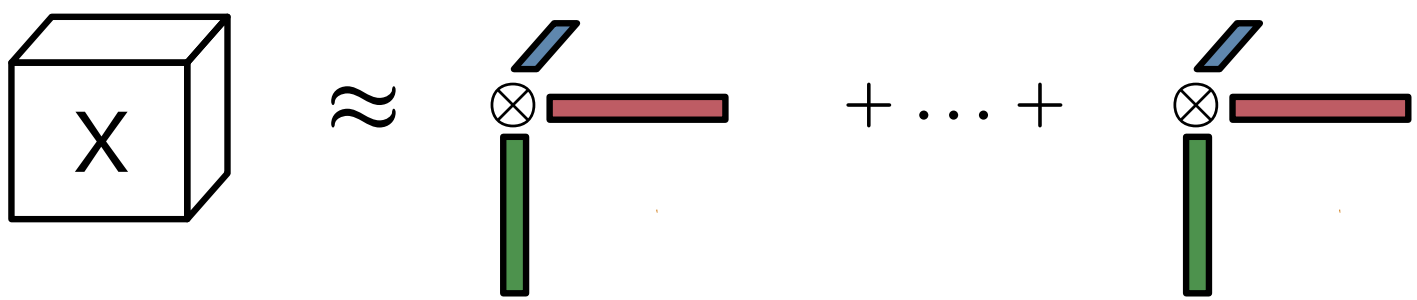
The canonical PD (CPD)

Def: If

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$

is minimal ($\operatorname{rank} \mathsf{X} = R$), then it is called the CPD of X.

It is also often called a PARAFAC (for parallel factors) decomposition.



Uniqueness: The CPD has quite strong uniqueness properties, discussed ahead.

This key feature has been a major driving force for its study, beginning in the Psychometrics litterature.[1,2,3]

1: Cattell, 1944, 2: Carroll & Chang, 1970, 3: Harshman, 1970

Slices & unfoldings of a CPD

Take $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r \quad \Leftrightarrow \quad x_{ijk} = \sum_{r=1}^{R} a_{ir}\, b_{jr}\, c_{kr}.$

Slices: Frontal slices of X can be expressed as

$$\mathbb{R}^{N_1 \times N_2} \ni \mathbf{X}_k = (\mathsf{X})_{::k} = \sum_{r=1}^{R} c_{kr}\, \boldsymbol{a}_r\, \boldsymbol{b}_r^{\mathsf{T}} = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^{\mathsf{T}},$$

with $\mathbf{D}_k(\mathbf{C}) := \mathrm{Diag}(c_{k1}, \ldots, c_{kR})$. Similarly, for horizontal and vertical slices:

$$(\mathsf{X})_{i::} = \mathbf{B}, \mathbf{D}_k(\mathbf{A})\, \mathbf{C}^{\mathsf{T}}, \qquad (\mathsf{X})_{:j:} = \mathbf{A}\, \mathbf{D}_k(\mathbf{B})\, \mathbf{C}^{\mathsf{T}}.$$

Slices & unfoldings of a CPD

Take $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r \quad \Leftrightarrow \quad x_{ijk} = \sum_{r=1}^{R} a_{ir}\, b_{jr}\, c_{kr}.$

Slices: Frontal slices of X can be expressed as

$$\mathbb{R}^{N_1 \times N_2} \ni \mathbf{X}_k = (\mathsf{X})_{::k} = \sum_{r=1}^{R} c_{kr}\, \boldsymbol{a}_r\, \boldsymbol{b}_r^{\mathsf{T}} = \mathbf{A}\mathbf{D}_k(\mathbf{C})\mathbf{B}^{\mathsf{T}},$$

with $\mathbf{D}_k(\mathbf{C}) := \mathrm{Diag}(c_{k1}, \ldots, c_{kR})$. Similarly, for horizontal and vertical slices:

$$(\mathsf{X})_{i::} = \mathbf{B}, \mathbf{D}_k(\mathbf{A})\, \mathbf{C}^{\mathsf{T}}, \qquad (\mathsf{X})_{:j:} = \mathbf{A}\, \mathbf{D}_k(\mathbf{B})\, \mathbf{C}^{\mathsf{T}}.$$

Unfoldings: By linearity of unfolding,

$$\mathbf{X}_{(1)} = \sum_{r=1}^{R} (\boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r)_{(1)} = \sum_{r=1}^{R} \boldsymbol{a}_r\, (\boldsymbol{c}_r \boxtimes \boldsymbol{b}_r)^{\mathsf{T}} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}},$$

Def: Khatri-Rao product (columnwise $\boxtimes$): $\mathbf{C} \odot \mathbf{B} := \begin{pmatrix} \boldsymbol{c}_1 \boxtimes \boldsymbol{b}_1 & \ldots & \boldsymbol{c}_R \boxtimes \boldsymbol{b}_R \end{pmatrix}$.

Unfoldings & vectorization of a CPD

Unfoldings: By the same reasoning,

$$\mathbf{X}_{(2)} = \mathbf{B}(\mathbf{C} \odot \mathbf{A})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}}.$$

Unfoldings & vectorization of a CPD

Unfoldings: By the same reasoning,

$$\mathbf{X}_{(2)} = \mathbf{B}(\mathbf{C} \odot \mathbf{A})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}}.$$

Vectorization: Again by linearity,

$$\operatorname{vec} \mathsf{X} = \sum_{r=1}^{R} \operatorname{vec}(\boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r) = \sum_{r=1}^{R} \boldsymbol{c}_r \boxtimes \boldsymbol{b}_r \boxtimes \boldsymbol{a}_r = (\mathbf{C} \odot \mathbf{B} \odot \mathbf{A})\mathbf{1}.$$

Computation of a CPD

In special cases, a CPD can be computed by non-iterative algebraic methods which rely on standard numerical linear algebra routines.

However, no general algorithm of that kind exists.

Computation of a CPD

In special cases, a CPD can be computed by non-iterative algebraic methods which rely on standard numerical linear algebra routines.

However, no general algorithm of that kind exists.

Typically, the problem is instead addressed by fitting the CPD model to the data according to some criterion or loss function.

Most usual approach: given a rank R, solve

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2 .$$

Nonconvex problem, since the model is multilinear in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.

Computation of a CPD

In special cases, a CPD can be computed by non-iterative algebraic methods which rely on standard numerical linear algebra routines.

However, no general algorithm of that kind exists.

Typically, the problem is instead addressed by fitting the CPD model to the data according to some criterion or loss function.

Most usual approach: given a rank R, solve

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2 .$$

Nonconvex problem, since the model is multilinear in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.

Clearly, this seems well-suited for approximation as well, which is the most common objective in applications.

However, low-rank approximation comes with its own pitfalls, as we'll see.

Alternating least squares (ALS)

Assume $\mathsf{X} = [\![\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star]\!]_R$. How can we address $\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathbf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2$?

Alternating least squares (ALS)

Assume $\mathsf{X} = [\![\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star]\!]_R$. How can we address $\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathbf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2$?

The simplest and best known algorithm is a BCD scheme called alternating least squares (ALS).[1,2]

Idea: While the original problem is (jointly) nonconvex, it is block-convex:

$$\begin{aligned}
\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2 &= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(i)} \\
&= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(ii)} \\
&= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(iii)}
\end{aligned}$$

1: Harshman, 1970, 2: Carroll & Chang, 1970

Alternating least squares (ALS)

Assume $\mathsf{X} = [\![\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star]\!]_R$. How can we address $\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathbf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2$?

The simplest and best known algorithm is a BCD scheme called alternating least squares (ALS).[1,2]

Idea: While the original problem is (jointly) nonconvex, it is block-convex:

$$\begin{aligned}
\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R\|_{\mathsf{F}}^2 &= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(i)} \\
&= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(ii)} \\
&= \min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathbf{X}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}}\right\|_{\mathsf{F}}^2 \quad \text{(iii)}
\end{aligned}$$

Hence, conditioned upon the current estimates $\mathbf{B}$ and $\mathbf{C}$, the solution for $\mathbf{A}$ can be explicitly derived from (i) as (assuming rank $\mathbf{C} \odot \mathbf{B} = R$):

$$\begin{aligned}
\mathbf{A} \leftarrow \mathbf{X}_{(1)} \left[(\mathbf{C} \odot \mathbf{B})^{\dagger}\right]^{\mathsf{T}} &= \mathbf{X}_{(1)} (\mathbf{C} \odot \mathbf{B})[(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}}(\mathbf{C} \odot \mathbf{B})]^{-1} \\
&= \mathbf{X}_{(1)} (\mathbf{C} \odot \mathbf{B})[(\mathbf{C}^{\mathsf{T}}\mathbf{C}) \circledast (\mathbf{B}^{\mathsf{T}}\mathbf{B})]^{-1}
\end{aligned}$$

1: Harshman, 1970, 2: Carroll & Chang, 1970

ALS: properties and variants

In its “vanilla” form, ALS is prone to facing ill-conditioning issues, slow convergence or failure to converge in reasonable time.

Nevertheless, ALS & its variants are wildly popular, as they are simple and often exhibit a good performance.

ALS: properties and variants

In its "vanilla" form, ALS is prone to facing ill-conditioning issues, slow convergence or failure to converge in reasonable time.

Nevertheless, ALS & its variants are wildly popular, as they are simple and often exhibit a good performance.

Some common variants are:

- "enhanced line search" (ELS) methods for accelerating convergence[1]
- adding a proximal term (w.r.t. current estimate $\mathbf{A}_0$) to each subproblem to circumvent ill-conditioning & guarantee well-posed updates:

$$\min_{\mathbf{A}} \left\| \mathbf{X}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}} \right\|_{\mathsf{F}}^2 + \tau \left\| \mathbf{A} - \mathbf{A}_0 \right\|_{\mathsf{F}}^2$$

(stricly convex problem, with $2\tau\mathbf{I}$ added to the Hessian)

1: Rajih & al., 2008

ALS: convergence

By construction, the objective value is nonincreasing along ALS iterations.

But what about convergence of iterates?

ALS: convergence

By construction, the objective value is nonincreasing along ALS iterations.

But what about convergence of iterates?

Local convergence: (Uschmajew, 2012) showed

- local convergence under non-degeneracy condition of local min
- local linear convergence around global minimizers of

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]_R\|_{\mathsf{F}}^2 + \tau\left(\|\mathbf{A}\|_{\mathsf{F}}^2 + \|\mathbf{B}\|_{\mathsf{F}}^2 + \|\mathbf{C}\|_{\mathsf{F}}^2\right)$$

ALS: convergence

By construction, the objective value is nonincreasing along ALS iterations.

But what about convergence of iterates?

Local convergence: (Uschmajew, 2012) showed

- local convergence under non-degeneracy condition of local min
- local linear convergence around global minimizers of

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \|\mathsf{X} - [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]_R\|_{\mathsf{F}}^2 + \tau\left(\|\mathbf{A}\|_{\mathsf{F}}^2 + \|\mathbf{B}\|_{\mathsf{F}}^2 + \|\mathbf{C}\|_{\mathsf{F}}^2\right)$$

Global convergence:

- Implied by the general framework of (Xu & al., 2013) when

$$\lambda_{\min}\left((\mathbf{C}\odot\mathbf{B})^{\mathsf{T}}(\mathbf{C}\odot\mathbf{B})\right) \geq c, \quad \text{for some } c > 0$$

  at every iteration, and idem for $(\mathbf{C}\odot\mathbf{A})^{\mathsf{T}}(\mathbf{C}\odot\mathbf{A})$ and $(\mathbf{B}\odot\mathbf{A})^{\mathsf{T}}(\mathbf{B}\odot\mathbf{A})$

- (Yang, 2023) proposed a variant which converges if we instead bound the smallest positive eigenvalues $\lambda_{\min}^{+}$ of those matrices uniformly

Algebraic sol'n by simultaneous diagonalization

A common method for the initialization of ALS (or any other iterative algorithm) is by simulatenous diagonalization (or "Jennrich's algorithm," a misnomer).[1,2]

It can be used for $d = 3$ whenever, say, $\operatorname{rank} \mathsf{X} = R \leq \min\{N_1, N_2\}$ w.l.o.g.

As we'll see, $\operatorname{mrank} \mathsf{X} \leq R$ entry-wise, thus one can project X onto $\mathbb{R}^{R \times R \times N_3}$. Hence, w.l.o.g. we assume $\mathsf{X} \in \mathbb{R}^{R \times R \times N_3}$.

1: Sanchez & Kowalski, 1990, 2: Leurgans & al., 1993

Algebraic sol'n by simultaneous diagonalization

A common method for the initialization of ALS (or any other iterative algorithm) is by simulatenous diagonalization (or "Jennrich's algorithm," a misnomer).[1,2]

It can be used for $d = 3$ whenever, say, $\operatorname{rank} \mathsf{X} = R \leq \min\{N_1, N_2\}$ w.l.o.g.

As we'll see, $\operatorname{mrank} \mathsf{X} \leq R$ entry-wise, thus one can project X onto $\mathbb{R}^{R \times R \times N_3}$. Hence, w.l.o.g. we assume $\mathsf{X} \in \mathbb{R}^{R \times R \times N_3}$.

Hypothesis: $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R$, with $\operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{B} = R$ and $\mathbf{C}$ doesn't have collinear columns.

Idea: By computing contractions $\mathbf{Y}_1 = (\cdot, \cdot, \boldsymbol{\theta}_1) \cdot \mathsf{X}$ and $\mathbf{Y}_2 = (\cdot, \cdot, \boldsymbol{\theta}_2) \cdot \mathsf{X}$, we get

$$\mathbf{Y}_i = \sum_{r=1}^{R} (\boldsymbol{c}_r^\mathsf{T} \boldsymbol{\theta}_i)\, \boldsymbol{a}_r \boldsymbol{b}_r^\mathsf{T} = \mathbf{A} \operatorname{Diag}(\boldsymbol{c}_1^\mathsf{T} \boldsymbol{\theta}_i, \ldots, \boldsymbol{c}_R^\mathsf{T} \boldsymbol{\theta}_i)\, \mathbf{B}^\mathsf{T} = \mathbf{A}\, \mathbf{D}_i \mathbf{B}^\mathsf{T}$$

Suppose $\mathbf{D}_2$ is nonsingular (generically true for random $\boldsymbol{\theta}_2$). Then,

$$\mathbf{Y} := \mathbf{Y}_1\, \mathbf{Y}_2^{-1} = \mathbf{A}\, \mathbf{D}_1 \mathbf{D}_2^{-1}\, \mathbf{A}^{-1}$$

so that the eigenvectors of $\mathbf{Y}$ are cols of $\mathbf{A}$ (with distinct eigenvalues)!

1: Sanchez & Kowalski, 1990, 2: Leurgans & al., 1993

Drawbacks, properties & extensions

Once $\mathbf{A}$ is recovered, one can compute $\mathbf{B}$ and $\mathbf{C}$ from

$$\left(\mathbf{A}^{-1}\mathbf{X}_{(1)}\right)^{\mathsf{T}} = \mathbf{C} \odot \mathbf{B},$$

by rank-1 approximation of each column.

1: Beltrán & al., 2019, 2: Bhaskara & al., 2014

Drawbacks, properties & extensions

Once $\mathbf{A}$ is recovered, one can compute $\mathbf{B}$ and $\mathbf{C}$ from

$$\left(\mathbf{A}^{-1}\mathbf{X}_{(1)}\right)^{\mathsf{T}} = \mathbf{C} \odot \mathbf{B},$$

by rank-1 approximation of each column.

Drawbacks: Possible numerical instability[1] and lack of robustness to noise.

1: Beltrán & al., 2019, 2: Bhaskara & al., 2014

Drawbacks, properties & extensions

Once $\mathbf{A}$ is recovered, one can compute $\mathbf{B}$ and $\mathbf{C}$ from

$$\left(\mathbf{A}^{-1}\mathbf{X}_{(1)}\right)^{\mathsf{T}} = \mathbf{C} \odot \mathbf{B},$$

by rank-1 approximation of each column.

Drawbacks: Possible numerical instability[1] and lack of robustness to noise.

Analysis: Error bounds were given by Bhaskara & al. (2014) under a "smoothed model," under additive noise with bounded entries (depending on the conditioning of $\mathbf{A}$, $\mathbf{B}$ and the coherence of $\mathbf{C}$).

1: Beltrán & al., 2019, 2: Bhaskara & al., 2014

Drawbacks, properties & extensions

Once $\mathbf{A}$ is recovered, one can compute $\mathbf{B}$ and $\mathbf{C}$ from

$$\left(\mathbf{A}^{-1}\mathbf{X}_{(1)}\right)^{\mathsf{T}} = \mathbf{C} \odot \mathbf{B},$$

by rank-1 approximation of each column.

Drawbacks: Possible numerical instability[1] and lack of robustness to noise.

Analysis: Error bounds were given by Bhaskara & al. (2014) under a "smoothed model," under additive noise with bounded entries (depending on the conditioning of $\mathbf{A}$, $\mathbf{B}$ and the coherence of $\mathbf{C}$).

Extensions: Higher-order tensors can be decomposed by partially ignoring the tensor structure. For instance, if $d = 2m + 1$ with $m > 1$, then one takes

$$\sum_{r=1}^{R} \boldsymbol{a}^{(1)} \otimes \cdots \otimes \boldsymbol{a}^{(d)} \mapsto \sum_{r=1}^{R} \left(\boldsymbol{a}^{(1)} \boxtimes \cdots \boxtimes \boldsymbol{a}^{(m)}\right) \otimes \left(\boldsymbol{a}^{(m+1)} \boxtimes \cdots \boxtimes \boldsymbol{a}^{(2m)}\right) \otimes \boldsymbol{a}^{(d)}.$$

Note that now R can exceed the dimensions N_i of all factors.

1: Beltrán & al., 2019, 2: Bhaskara & al., 2014

Properties of the tensor rank, part II

(i) Lower bound by mrank:

Prop: rank X is lower bounded by each component of mrank X.

Proof: Let rank $\mathsf{X} = R$ and write the CPD $\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$. Then,

$$\operatorname{rank}_1 \mathsf{X} = \operatorname{rank} \mathbf{X}_{(1)} = \operatorname{rank} \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}} \leq R,$$

and similarly for the other modes. ■

Properties of the tensor rank, part II

(i) Lower bound by mrank:

Prop: $\operatorname{rank} \mathsf{X}$ is lower bounded by each component of $\operatorname{mrank} \mathsf{X}$.

Proof: Let $\operatorname{rank} \mathsf{X} = R$ and write the CPD $\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$. Then,

$$\operatorname{rank}_1 \mathsf{X} = \operatorname{rank} \mathbf{X}_{(1)} = \operatorname{rank} \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}} \leq R,$$

and similarly for the other modes. ■

Rmk: This implies that $\operatorname{rank} \mathsf{X} \geq \operatorname{rank} \mathbf{X}_k$ for all k (similarly for horizontal and vertical slices).

Properties of the tensor rank, part II

(i) Lower bound by mrank:

Prop: rank X is lower bounded by each component of mrank X.

Proof: Let rank $\mathsf{X} = R$ and write the CPD $\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$. Then,

$$\operatorname{rank}_1 \mathsf{X} = \operatorname{rank} \mathbf{X}_{(1)} = \operatorname{rank} \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathsf{T}} \leq R,$$

and similarly for the other modes. ■

Rmk: This implies that rank $\mathsf{X} \geq$ rank $\mathbf{X}_k$ for all k (similarly for horizontal and vertical slices).

(ii) Rank invariance under embedding in larger space:

Prop: If $\mathsf{X} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{S}$ and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ have full column rank, then rank X = rank S.

Proof: See exercices.

Rmk: Often used (in an approximate fashion) when computing the CPD of a low-mrank tensor: one can work in $\mathbb{R}^{R_1 \times R_2 \times R_3}$ instead of in $\mathbb{R}^{N_1 \times N_2 \times N_3}$

Tensor rank: examples

Example: If

$$\mathsf{D} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \cdots + \boldsymbol{e}_R \otimes \boldsymbol{e}_R \otimes \boldsymbol{e}_R,$$

then $\operatorname{rank} \mathsf{D} = R$, since $\operatorname{rank} \mathsf{D} \geq \operatorname{rank}_1 \mathsf{D} = \operatorname{rank} \mathbf{I} \, (\mathbf{I} \odot \mathbf{I})^{\mathsf{T}} = R$.

□

Tensor rank: examples

Example: If

$$\mathsf{D} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \cdots + \boldsymbol{e}_R \otimes \boldsymbol{e}_R \otimes \boldsymbol{e}_R,$$

then $\operatorname{rank} \mathsf{D} = R$, since $\operatorname{rank} \mathsf{D} \geq \operatorname{rank}_1 \mathsf{D} = \operatorname{rank} \mathbf{I} \, (\mathbf{I} \odot \mathbf{I})^{\mathsf{T}} = R$. □

Example: Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ have full column rank, and define

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \textstyle\sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r.$$

We have

$$\mathsf{X} = \textstyle\sum_{r=1}^{R} (\mathbf{A}\, \boldsymbol{e}_r) \otimes (\mathbf{B}\, \boldsymbol{e}_r) \otimes (\mathbf{C}\, \boldsymbol{e}_r) = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D},$$

thus $\operatorname{rank} \mathsf{X} = R$. □

Tensor rank: examples

Example: If

$$\mathsf{D} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \cdots + \boldsymbol{e}_R \otimes \boldsymbol{e}_R \otimes \boldsymbol{e}_R,$$

then $\operatorname{rank} \mathsf{D} = R$, since $\operatorname{rank} \mathsf{D} \geq \operatorname{rank}_1 \mathsf{D} = \operatorname{rank} \mathbf{I} \, (\mathbf{I} \odot \mathbf{I})^{\mathsf{T}} = R$. □

Example: Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ have full column rank, and define

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \textstyle\sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r.$$

We have

$$\mathsf{X} = \textstyle\sum_{r=1}^{R} (\mathbf{A}\, \boldsymbol{e}_r) \otimes (\mathbf{B}\, \boldsymbol{e}_r) \otimes (\mathbf{C}\, \boldsymbol{e}_r) = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D},$$

thus $\operatorname{rank} \mathsf{X} = R$. □

Example:

$$\operatorname{rank} (\boldsymbol{a}_1 \otimes \boldsymbol{b} \otimes \boldsymbol{c} + \boldsymbol{a}_2 \otimes \boldsymbol{b} \otimes \boldsymbol{c}) = \operatorname{rank} (\boldsymbol{a}_1 + \boldsymbol{a}_2) \otimes \boldsymbol{b} \otimes \boldsymbol{c} \leq 1,$$

while

$$\operatorname{rank} (\boldsymbol{a} \otimes \boldsymbol{b}_1 \otimes \boldsymbol{c}_1 + \boldsymbol{a} \otimes \boldsymbol{b}_2 \otimes \boldsymbol{c}_2) \leq 2.$$

□

Properties of the tensor rank, part III

(iii) Upper bound by product of mrank components:

Prop: If $\operatorname{mrank} \mathsf{X} = (R_1, R_2, R_3)$, then $\operatorname{rank} \mathsf{X} \leq \min\{R_1 R_2, R_1 R_3, R_2 R_3\}$.

Proof: From an mrank–(R_1, R_2, R_3) Tucker representation of X, we get

$$\mathsf{X} = \sum_{r_1, r_2, r_3} s_{r_1 r_2 r_3} \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{w}_{r_3} = \sum_{r_1, r_2} \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \left(\sum_{r_3} s_{r_1 r_2 r_3} \boldsymbol{w}_{r_3} \right),$$

which shows that $\operatorname{rank} \mathsf{X} \leq R_1 R_2$. The other bounds are shown similarly. ■

Properties of the tensor rank, part III

(iii) Upper bound by product of mrank components:

Prop: If mrank $\mathsf{X} = (R_1, R_2, R_3)$, then rank $\mathsf{X} \leq \min\{R_1R_2, R_1R_3, R_2R_3\}$.

Proof: From an mrank–(R_1, R_2, R_3) Tucker representation of X, we get

$$\mathsf{X} = \sum_{r_1,r_2,r_3} s_{r_1r_2r_3}\, \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \boldsymbol{w}_{r_3} = \sum_{r_1,r_2} \boldsymbol{u}_{r_1} \otimes \boldsymbol{v}_{r_2} \otimes \left(\sum_{r_3} s_{r_1r_2r_3} \boldsymbol{w}_{r_3} \right),$$

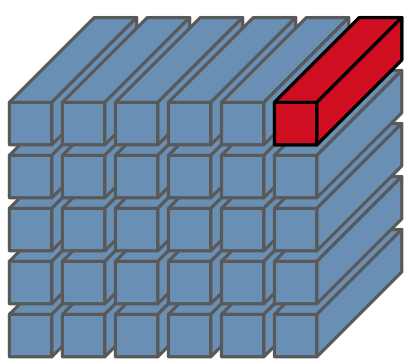
which shows that rank $\mathsf{X} \leq R_1R_2$. The other bounds are shown similarly. ■

A more intuitive way of seeing this: assume w.l.o.g. that mrank $\mathsf{X} = (N_1, N_2, N_3)$ and "synthetize" X as

$$\mathsf{X} = \sum\nolimits_{n_1,n_2,n_3} x_{n_1n_2n_3} \boldsymbol{e}_{n_1} \otimes \boldsymbol{e}_{n_2} \otimes \boldsymbol{e}_{n_3}$$

$$= \sum\nolimits_{n_1,n_2} \boldsymbol{e}_{n_1} \otimes \boldsymbol{e}_{n_2} \otimes \underbrace{\left(\sum_{n_3} x_{n_1n_2n_3}\, \boldsymbol{e}_{n_3} \right)}_{\text{"tube" at } (n_1,n_2)}.$$

tube at $(1, N_2)$



Properties of the tensor rank, part III cont'd

(iv) rank X depends on the underlying field

Example: Take

$$x_{ijk} = \cos(i+j+k) = \frac{1}{2}\left(e^{\mathrm{i}(i+j+k)} + e^{-\mathrm{i}(i+j+k)}\right)$$
$$= \frac{1}{2}\, e^{\mathrm{i}\, i} e^{\mathrm{i}\, j} e^{\mathrm{i}\, k} + \frac{1}{2}\, \overline{e^{\mathrm{i}\, i} e^{\mathrm{i}\, j} e^{\mathrm{i}\, k}}.$$

Hence, rank $\mathsf{X} = 2$ over $\mathbb{C}$.

Properties of the tensor rank, part III cont'd

(iv) rank X depends on the underlying field

Example: Take

$$x_{ijk} = \cos(i+j+k) = \frac{1}{2}\left(e^{\mathrm{i}(i+j+k)} + e^{-\mathrm{i}(i+j+k)}\right)$$
$$= \frac{1}{2}\, e^{\mathrm{i}\, i} e^{\mathrm{i}\, j} e^{\mathrm{i}\, k} + \frac{1}{2}\, \overline{e^{\mathrm{i}\, i} e^{\mathrm{i}\, j} e^{\mathrm{i}\, k}}.$$

Hence, rank X $= 2$ over $\mathbb{C}$.

By using again Euler's identity $e^{\mathrm{i}\ell} = a_\ell + \mathrm{i} b_\ell$ with $a_\ell := \cos \ell$ and $b_\ell := \sin \ell$,

$$x_{ijk} = a_i a_j a_k - a_i b_j b_k - b_i a_j b_k - b_i b_j a_k \qquad \text{(0 or 2 factors } b\text{)}$$
$$= \left[(\mathbf{H}, \mathbf{H}, \mathbf{H}) \cdot \underbrace{\left(\begin{array}{cc|cc} 1 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 \end{array}\right)}_{\text{rank}=3} \right]_{ijk}, \qquad \text{with } \mathbf{H} := \begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} \end{pmatrix}$$
$$= -(a_i + b_i) b_j b_k + (a_i - b_i) a_j a_k + b_i (a_j - b_j)(a_k - b_k).$$

Hence, rank X $= 3$ over $\mathbb{R}$!

□

Properties of the tensor rank, part III cont'd

(v) rank X can exceed all tensor dimensions

Example: Let

$$\mathsf{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right) = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_1 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2.$$

Is this minimal?

Properties of the tensor rank, part III cont'd

(v) $\operatorname{rank} \mathsf{X}$ can exceed all tensor dimensions

Example: Let

$$\mathsf{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right) = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_1 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2.$$

Is this minimal?

First, note that $\operatorname{rank} \mathsf{X} \geq \operatorname{rank}_1 \mathsf{X} = 2$. Suppose that $\operatorname{rank} \mathsf{X} = 2$. Then, we can write $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{2\times 2}$ and

$$\mathbf{X}_1 = \mathbf{I} = \mathbf{A}\mathbf{D}_1(\mathbf{C})\mathbf{B}^{\mathsf{T}}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \mathbf{A}\mathbf{D}_2(\mathbf{C})\mathbf{B}^{\mathsf{T}}.$$

Properties of the tensor rank, part III cont'd

(v) $\operatorname{rank} \mathsf{X}$ can exceed all tensor dimensions

Example: Let

$$\mathsf{X} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right) = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_1 + \boldsymbol{e}_1 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2.$$

Is this minimal?

First, note that $\operatorname{rank} \mathsf{X} \geq \operatorname{rank}_1 \mathsf{X} = 2$. Suppose that $\operatorname{rank} \mathsf{X} = 2$. Then, we can write $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ with $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{2\times 2}$ and

$$\mathbf{X}_1 = \mathbf{I} = \mathbf{A}\mathbf{D}_1(\mathbf{C})\mathbf{B}^{\mathsf{T}}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \mathbf{A}\mathbf{D}_2(\mathbf{C})\mathbf{B}^{\mathsf{T}}.$$

As $\mathbf{X}_1$ is nonsingular, it follows that

$$\mathbf{X}_2\,\mathbf{X}_1^{-1} = \mathbf{X}_2 = \mathbf{X}_2\mathbf{B}^{-\mathsf{T}}(\mathbf{D}_1(\mathbf{C}))^{-1}\mathbf{A}^{-1} = \mathbf{A}\,\underbrace{\mathbf{D}_2(\mathbf{C})(\mathbf{D}_1(\mathbf{C}))^{-1}}_{:=\mathbf{D}}\,\mathbf{A}^{-1},$$

implying $\mathbf{X}_2$ is diagonalizable—a contradiction!

Conclusion: $\operatorname{rank} \mathsf{X} = 3$. This is actually the maximal rank on $\mathbb{R}^{2\times 2\times 2}$.

□

Maximal rank

Q: What is the maximal rank of a tensor from $\mathbb{R}^{N_1 \times N_2 \times N_3}$?

Many results have been derived for more or less particular cases, e.g.:

- maximal rank over $\mathbb{R}^{2\times2\times2}$ or $\mathbb{C}^{2\times2\times2}$ is[1] 3
- maximal rank over $\mathbb{R}^{3\times3\times3}$ or $\mathbb{C}^{3\times3\times3}$ is[1,2] 5
- maximal rank over $\mathbb{R}^{N_1 \times N_2 \times N_3}$ is[3] at most $N_1 + \lfloor N_3/2 \rfloor \, N_2$

See the references given in the handout for more results.

This question is not of much concern for the applications that we consider here: both uniqueness and compressibility require a much smaller rank.

1: Kruskal, 1989, 2: Bremner & Hu, 2013, 3: Sumi & al., 2010

Generic and typical ranks

Recap for matrices:

Let $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$ be such that $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. **Q**: What is rank **X**?

Generic and typical ranks

Recap for matrices:

Let $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$ be such that $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. **Q**: What is rank $\mathbf{X}$?

Answer: $\mathbb{P}(\operatorname{rank} \mathbf{X} = \min\{N_1, N_2\}) = 1, \;\; \mathbb{P}(\operatorname{rank} \mathbf{X} < \min\{N_1, N_2\}) = 0$

The set of rank-defective matrices has null Lebesgue measure $\Rightarrow$ any absolutely continuous distribution (w.r.t. Lebesgue) will assign zero mass to it.

What about tensors?

Generic and typical ranks

Recap for matrices:

Let $\mathbf{X} \in \mathbb{R}^{N_1 \times N_2}$ be such that $x_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. **Q**: What is rank $\mathbf{X}$?

Answer: $\mathbb{P}(\text{rank}\,\mathbf{X} = \min\{N_1, N_2\}) = 1,\ \ \mathbb{P}(\text{rank}\,\mathbf{X} < \min\{N_1, N_2\}) = 0$

The set of rank-defective matrices has null Lebesgue measure $\Rightarrow$ any absolutely continuous distribution (w.r.t. Lebesgue) will assign zero mass to it.

What about tensors?

Example[1]: Let's do a similar experiment: $\mathsf{X} \in \mathbb{R}^{2\times 2\times 2}$, with $x_{ijk} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then[2]:

- over $\mathbb{R}$, $\mathbb{P}(\text{rank}\,\mathsf{X} = 2) = \frac{\pi}{4},\ \ \mathbb{P}(\text{rank}\,\mathsf{X} = 3) = 1 - \frac{\pi}{4}$
- over $\mathbb{C}$, $\mathbb{P}(\text{rank}\,\mathsf{X} = 2) = 1$

What is going on?

1: Sidiropoulos & al., 2017, 2: Bergqvist, 2013

Generic and typical ranks, cont'd

Write $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ the CPD of X. Since $\operatorname{rank} \mathbf{X}_1 = \operatorname{rank} \mathbf{X}_2 = 2$ a.s., if $\operatorname{rank} \mathsf{X} = 2$ over $\mathbb{R}$, then all (real-valued) matrices in

$$\mathbf{X}_1 = \mathbf{A}\,\mathbf{D}_1(\mathbf{C})\mathbf{B}^{\mathsf{T}}, \quad \mathbf{X}_2 = \mathbf{A}\,\mathbf{D}_2(\mathbf{C})\mathbf{B}^{\mathsf{T}}$$

must be nonsingular.

Generic and typical ranks, cont'd

Write $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ the CPD of X. Since $\operatorname{rank} \mathbf{X}_1 = \operatorname{rank} \mathbf{X}_2 = 2$ a.s., if $\operatorname{rank} \mathsf{X} = 2$ over $\mathbb{R}$, then all (real-valued) matrices in

$$\mathbf{X}_1 = \mathbf{A}\,\mathbf{D}_1(\mathbf{C})\mathbf{B}^\mathsf{T}, \quad \mathbf{X}_2 = \mathbf{A}\,\mathbf{D}_2(\mathbf{C})\mathbf{B}^\mathsf{T}$$

must be nonsingular. Next, we compute (does this look familiar?):

$$\mathbf{X}_1\mathbf{X}_2^{-1} = \mathbf{A}\,\mathbf{D}_1(\mathbf{C})\,(\mathbf{D}_2(\mathbf{C}))^{-1}\,\mathbf{A}^{-1} = \mathbf{A}\,\mathbf{D}\,\mathbf{A}^{-1}.$$

But real-valued matrices can have complex-valued eigenvalues with positive probability, in which case we would have a contradiction.

By contrast, this works with probability one if $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{C}^{2\times 2}$, hence $\operatorname{rank} \mathsf{X} = 2$ a.s. over $\mathbb{C}$.

□

Generic and typical ranks, cont'd

Write $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ the CPD of X. Since $\operatorname{rank} \mathbf{X}_1 = \operatorname{rank} \mathbf{X}_2 = 2$ a.s., if $\operatorname{rank} \mathsf{X} = 2$ over $\mathbb{R}$, then all (real-valued) matrices in

$$\mathbf{X}_1 = \mathbf{A}\,\mathbf{D}_1(\mathbf{C})\mathbf{B}^\mathsf{T}, \quad \mathbf{X}_2 = \mathbf{A}\,\mathbf{D}_2(\mathbf{C})\mathbf{B}^\mathsf{T}$$

must be nonsingular. Next, we compute (does this look familiar?):

$$\mathbf{X}_1\mathbf{X}_2^{-1} = \mathbf{A}\,\mathbf{D}_1(\mathbf{C})\,(\mathbf{D}_2(\mathbf{C}))^{-1}\,\mathbf{A}^{-1} = \mathbf{A}\,\mathbf{D}\,\mathbf{A}^{-1}.$$

But real-valued matrices can have complex-valued eigenvalues with positive probability, in which case we would have a contradiction.

By contrast, this works with probability one if $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{C}^{2\times 2}$, hence $\operatorname{rank} \mathsf{X} = 2$ a.s. over $\mathbb{C}$. □

In general: For tensors in $\mathbb{R}^{N_1 \times \cdots \times N_d}$,

- there is a unique[1] generic rank over $\mathbb{C}$, smaller than the maximal rank
- there are multiple typical ranks in $\mathbb{R}$ (the smallest one is generic[1] over $\mathbb{C}$).

1: Blekherman & Teitler, 2015

How large is the generic rank?

For an $N_1 \times \cdots \times N_d$ tensor over $\mathbb{C}$, the generic rank is, according to the Abo–Ottaviani–Peterson conjecture[1,2]

$$\text{grank}(N_1, \ldots, N_d) = \left\lceil \frac{\prod_{i=1}^{d} N_i}{1 + \sum_{i=1}^{d} (N_i - 1)} \right\rceil,$$

with some exceptions.

As we will discuss ahead, CPD uniqueness generically holds up to grank -1 (but not beyond).

1: Abo & al., 2009, 2: Vannieuwenhoven, 2015

The block-term decomposition (BTD) model

What happens if two or more terms of a PD share a vector? For example:

$$\begin{aligned}\mathsf{X} &= \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{c} + \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{c} + \dots \\ &= (\boldsymbol{a}_1 \otimes \boldsymbol{b}_1 + \boldsymbol{a}_2 \otimes \boldsymbol{b}_2) \otimes \boldsymbol{c} + \dots \\ &= (\mathbf{A}\mathbf{B}^\mathsf{T}) \otimes \boldsymbol{c} + \dots\end{aligned}$$

"Full" identifiability breaks down, since $\mathbf{A}$ and $\mathbf{B}$ are not identifiable from $\mathbf{A}\mathbf{B}^\mathsf{T}$.

But can at least $\mathbf{A}\mathbf{B}^\mathsf{T}$ and $\boldsymbol{c}$ be identifiable?

The block-term decomposition (BTD) model

What happens if two or more terms of a PD share a vector? For example:

$$\begin{aligned}\mathsf{X} &= \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 \otimes \boldsymbol{c} + \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 \otimes \boldsymbol{c} + \dots \\ &= (\boldsymbol{a}_1 \otimes \boldsymbol{b}_1 + \boldsymbol{a}_2 \otimes \boldsymbol{b}_2) \otimes \boldsymbol{c} + \dots \\ &= (\mathbf{A}\mathbf{B}^\mathsf{T}) \otimes \boldsymbol{c} + \dots\end{aligned}$$

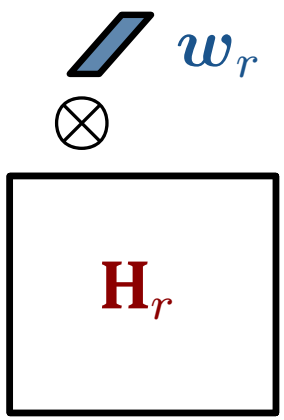
"Full" identifiability breaks down, since $\mathbf{A}$ and $\mathbf{B}$ are not identifiable from $\mathbf{A}\mathbf{B}^\mathsf{T}$.

But can at least $\mathbf{A}\mathbf{B}^\mathsf{T}$ and $\boldsymbol{c}$ be identifiable?

Yes, under a low-rank constraint.

This is the point of the BTD model:[1]

$$\mathsf{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{c}_r, \quad \operatorname{rank} \mathbf{H}_r \leq L_r$$



Various uniqueness results also exist, some of which we will discuss ahead.

1: De Lathauwer, 2008

BTD as a sum of low-mrank tensors

Q: If $\mathsf{X} = \mathbf{H} \otimes \boldsymbol{c}$ with $\operatorname{rank} \mathbf{H} = L$, then $\operatorname{mrank} \mathsf{X} = ?$

Write

$$\mathbf{X}_{(1)} = \begin{pmatrix} c_1 \mathbf{H} & c_2 \mathbf{H} & \dots & c_{N_3} \mathbf{H} \end{pmatrix} = (\mathbf{1}_{N_3}^{\mathsf{T}} \boxtimes \mathbf{H}) \operatorname{Diag}(\boldsymbol{c} \boxtimes \mathbf{1}_{N_2})$$

to see that $\operatorname{rank}_1 \mathsf{X} = L$ (similarly for rank_2). Also,

$$\mathbf{X}_{(3)} = \boldsymbol{c} \operatorname{vec}(\mathbf{H})^{\mathsf{T}} \quad \Rightarrow \quad \operatorname{rank}_3 \mathsf{X} = 1.$$

Hence, $\operatorname{mrank} \mathsf{X} = (L, L, 1)$ & we can see the BTD as a sum of low-mrank blocks.

BTD as a sum of low-mrank tensors

Q: If $\mathsf{X} = \mathbf{H} \otimes \boldsymbol{c}$ with $\operatorname{rank} \mathbf{H} = L$, then $\operatorname{mrank} \mathsf{X} = ?$

Write

$$\mathbf{X}_{(1)} = \begin{pmatrix} c_1 \mathbf{H} & c_2 \mathbf{H} & \dots & c_{N_3} \mathbf{H} \end{pmatrix} = (\mathbf{1}_{N_3}^{\mathsf{T}} \boxtimes \mathbf{H}) \operatorname{Diag}(\boldsymbol{c} \boxtimes \mathbf{1}_{N_2})$$

to see that $\operatorname{rank}_1 \mathsf{X} = L$ (similarly for rank_2). Also,

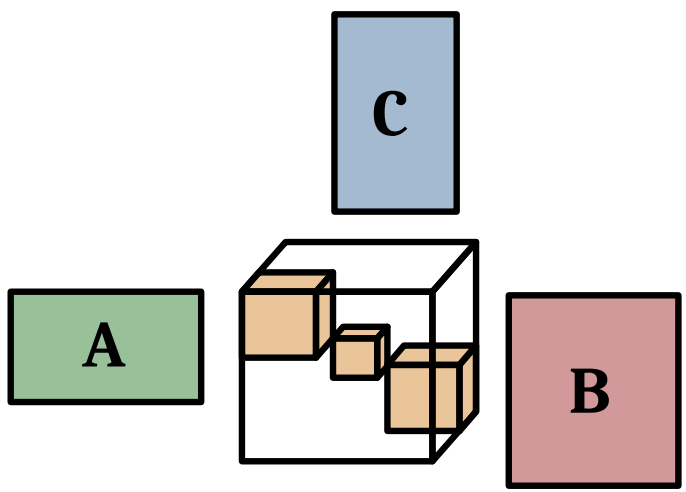
$$\mathbf{X}_{(3)} = \boldsymbol{c} \operatorname{vec}(\mathbf{H})^{\mathsf{T}} \quad \Rightarrow \quad \operatorname{rank}_3 \mathsf{X} = 1.$$

Hence, $\operatorname{mrank} \mathsf{X} = (L, L, 1)$ & we can see the BTD as a sum of low-mrank blocks.

More generally, it can be defined that way:[1]

$$\mathsf{X} = \sum_{r=1}^{R} (\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r) \cdot \mathsf{S}_r$$

Some uniqueness results also exist for this more general model.[1]



1: De Lathauwer, 2008

BTD computation

Just like the CPD, if X is given by a BTD

$$\mathsf{X} = \sum_{r=1}^{R} [\mathbf{A}_r^\star (\mathbf{B}_r^\star)^\mathsf{T}] \otimes \boldsymbol{c}_r^\star,$$

with $\mathbf{A}_r^\star \in \mathbb{R}^{N_1 \times L_r}$, $\mathbf{B}_r^\star \in \mathbb{R}^{N_2 \times L_r}$ and $\boldsymbol{c}_r^\star \in \mathbb{R}^{N_3}$, then its computation can be carried out via

$$\min_{\{\mathbf{A}_r, \mathbf{B}_r, \boldsymbol{c}_r\}_{r=1}^R} \left\| \mathsf{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r \, \mathbf{B}_r^\mathsf{T} \right) \otimes \boldsymbol{c}_r \right\|_\mathsf{F}^2,$$

with same dimensions.

BTD computation

Just like the CPD, if X is given by a BTD

$$\mathsf{X} = \sum_{r=1}^{R} [\mathbf{A}_r^\star (\mathbf{B}_r^\star)^\mathsf{T}] \otimes \boldsymbol{c}_r^\star,$$

with $\mathbf{A}_r^\star \in \mathbb{R}^{N_1 \times L_r}$, $\mathbf{B}_r^\star \in \mathbb{R}^{N_2 \times L_r}$ and $\boldsymbol{c}_r^\star \in \mathbb{R}^{N_3}$, then its computation can be carried out via

$$\min_{\{\mathbf{A}_r, \mathbf{B}_r, \boldsymbol{c}_r\}_{r=1}^R} \left\| \mathsf{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r \, \mathbf{B}_r^\mathsf{T} \right) \otimes \boldsymbol{c}_r \right\|_\mathsf{F}^2,$$

with same dimensions.

In particular, an ALS algorithm can also be easily derived from the form of $\mathbf{X}_{(1)}$, $\mathbf{X}_{(2)}$ and $\mathbf{X}_{(3)}$.

Again, this approach allows approximate computation but comes with some difficulties, as discussed ahead.

A mosaic of uniqueness results

Inherent (or trivial) ambiguities

Typically, uniqueness of parameters of a tensor model can be only shown up to two inherent (or trivial) ambiguities:

(i) permutation, since the indexing of terms in a sum is (usually) arbitrary;

(ii) scaling, due to the multilinear nature of the model.

Inherent (or trivial) ambiguities

Typically, uniqueness of parameters of a tensor model can be only shown up to two inherent (or trivial) ambiguities:

(i) permutation, since the indexing of terms in a sum is (usually) arbitrary;

(ii) scaling, due to the multilinear nature of the model.

Example: Take $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ of size R. For any permutation $\pi \in \mathfrak{S}_R$, and any nonzero scalars α_r, β_r we have

$$\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} (\alpha_r \boldsymbol{a}_{\pi_r}) \otimes (\beta_r \boldsymbol{b}_{\pi_r}) \otimes ((\alpha_r \beta_r)^{-1} \boldsymbol{c}_{\pi_r}).$$

Inherent (or trivial) ambiguities

Typically, uniqueness of parameters of a tensor model can be only shown up to two inherent (or trivial) ambiguities:

(i) permutation, since the indexing of terms in a sum is (usually) arbitrary;

(ii) scaling, due to the multilinear nature of the model.

Example: Take $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ of size R. For any permutation $\pi \in \mathfrak{S}_R$, and any nonzero scalars α_r, β_r we have

$$\mathsf{X} = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} (\alpha_r \boldsymbol{a}_{\pi_r}) \otimes (\beta_r \boldsymbol{b}_{\pi_r}) \otimes ((\alpha_r \beta_r)^{-1} \boldsymbol{c}_{\pi_r}).$$

Partial remedy: "extract" norms λ_r and sort the terms according to them:

$$\mathsf{X} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r, \quad \|\boldsymbol{a}_r\| = \|\boldsymbol{b}_r\| = \|\boldsymbol{c}_r\| = 1, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R \geq 0.$$

However, sign (or rotation, in $\mathbb{C}$) ambiguities remain.

□

Essential uniqueness

Def: The parameters of a tensor decomposition are essentially unique if they are unique up to trivial (permutation and scaling) ambiguities.

Essential uniqueness

Def: The parameters of a tensor decomposition are essentially unique if they are unique up to trivial (permutation and scaling) ambiguities.

Example: Essential uniqueness of a rank-R CPD $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is equivalent to:

$$\forall (\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}) \in \mathbb{R}^{I \times R} \times \mathbb{R}^{J \times R} \times \mathbb{R}^{K \times R},$$

$$\mathsf{X} = [\![\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}]\!] \quad \Leftrightarrow \quad \tilde{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1, \ \tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2, \ \tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3,$$

where $\mathbf{\Pi} \in \mathbb{R}^{R \times R}$ is a permutation matrix and $\mathbf{\Lambda}_i \in \mathbb{R}^{R \times R}$ are diagonal scaling matrices satisfying

$$\mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \mathbf{\Lambda}_3 = \mathbf{I}.$$

□

Essential uniqueness

Def: The parameters of a tensor decomposition are essentially unique if they are unique up to trivial (permutation and scaling) ambiguities.

Example: Essential uniqueness of a rank-R CPD $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is equivalent to:

$$\forall(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}) \in \mathbb{R}^{I\times R} \times \mathbb{R}^{J\times R} \times \mathbb{R}^{K\times R},$$
$$\mathsf{X} = [\![\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}]\!] \quad \Leftrightarrow \quad \tilde{\mathbf{A}} = \mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1,\ \tilde{\mathbf{B}} = \mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2,\ \tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3,$$

where $\mathbf{\Pi} \in \mathbb{R}^{R\times R}$ is a permutation matrix and $\mathbf{\Lambda}_i \in \mathbb{R}^{R\times R}$ are diagonal scaling matrices satisfying

$$\mathbf{\Lambda}_1\mathbf{\Lambda}_2\mathbf{\Lambda}_3 = \mathbf{I}.$$

□

Example: Essential uniqueness of a R-block BTD

$$\mathsf{X} = \sum_{r=1}^{R} (\mathbf{A}_r\mathbf{B}_r)^{\mathsf{T}} \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} \mathbf{H}_r^{\mathsf{T}} \otimes \boldsymbol{c}_r$$

means uniqueness up to permutation of blocks and to rescaling of $(\mathbf{H}_r, \boldsymbol{c}_r)$:

$$(\mathbf{H}_r, \boldsymbol{c}_r) \mapsto (\alpha\mathbf{H}_r, \alpha^{-1}\boldsymbol{c}_r), \quad \alpha > 0.$$

□

A first intuition

Recall: lack of uniqueness = “too much freedom”

$$\mathbf{A}\mathbf{B}^{\mathsf{T}} = \mathbf{A}\mathbf{P}\,\mathbf{P}^{-1}\mathbf{B} = (\mathbf{A}\mathbf{P})\,(\mathbf{B}\mathbf{P}^{-\mathsf{T}})^{\mathsf{T}} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^{\mathsf{T}}, \qquad \forall \mathbf{P} \in \mathrm{GL}_R.$$

A first intuition

Recall: lack of uniqueness = "too much freedom"

$$\mathbf{A}\mathbf{B}^\mathsf{T} = \mathbf{A}\mathbf{P}\,\mathbf{P}^{-1}\mathbf{B} = (\mathbf{A}\mathbf{P})\,(\mathbf{B}\mathbf{P}^{-\mathsf{T}})^\mathsf{T} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^\mathsf{T}, \qquad \forall \mathbf{P} \in \mathtt{GL}_R.$$

Now, take a PD of size R, say $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$, and write

$$\mathbf{X}_{(1)} = \mathbf{A}\,(\mathbf{C} \odot \mathbf{B})^\mathsf{T}.$$

If we try the same trick as above, namely

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{P}\,\mathbf{P}^{-1}(\mathbf{C} \odot \mathbf{B})^\mathsf{T} = \tilde{\mathbf{A}}\left((\mathbf{C} \odot \mathbf{B})\mathbf{P}^{-\mathsf{T}}\right)^\mathsf{T},$$

then for which $\mathbf{P}$ can we write

$$(\mathbf{C} \odot \mathbf{B})\mathbf{P}^{-\mathsf{T}} = \tilde{\mathbf{C}} \odot \tilde{\mathbf{B}} \quad ?$$

A first intuition

Recall: lack of uniqueness = "too much freedom"

$$\mathbf{A}\mathbf{B}^\mathsf{T} = \mathbf{A}\mathbf{P}\,\mathbf{P}^{-1}\mathbf{B} = (\mathbf{A}\mathbf{P})\,(\mathbf{B}\mathbf{P}^{-\mathsf{T}})^\mathsf{T} = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^\mathsf{T}, \qquad \forall \mathbf{P} \in \mathrm{GL}_R.$$

Now, take a PD of size R, say $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$, and write

$$\mathbf{X}_{(1)} = \mathbf{A}\,(\mathbf{C} \odot \mathbf{B})^\mathsf{T}.$$

If we try the same trick as above, namely

$$\mathbf{X}_{(1)} = \mathbf{A}\mathbf{P}\,\mathbf{P}^{-1}(\mathbf{C} \odot \mathbf{B})^\mathsf{T} = \tilde{\mathbf{A}}\left((\mathbf{C} \odot \mathbf{B})\mathbf{P}^{-\mathsf{T}}\right)^\mathsf{T},$$

then for which $\mathbf{P}$ can we write

$$(\mathbf{C} \odot \mathbf{B})\mathbf{P}^{-\mathsf{T}} = \tilde{\mathbf{C}} \odot \tilde{\mathbf{B}} \quad ?$$

Rigidity: In general, $\mathbf{P}^{-\mathsf{T}}$ can only permute and rescale the columns of $\mathbf{C} \odot \mathbf{B}$, as these are vectorized rank-1 matrices, which do not form a linear space.

(Hence, $\mathbf{P}$ only permutes and rescales the columns of $\mathbf{A}$ as well.)

Warmup: a CPD with full rank factors

Let's prove (essential) uniqueness of the CPD when $\mathbf{A}, \mathbf{B}, \mathbf{C}$ all have full column rank. Write

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D}, \quad \text{where} \quad \mathsf{D} = \sum_{r=1}^{R} \boldsymbol{e}_r^{\otimes 3} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!] .$$

Claim: The CPD of X is unique iff the CPD of D is unique.

(Uniqueness of $\mathsf{D} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$ means that every CPD of D has the form $\mathsf{D} = [\![\mathbf{\Pi\Lambda}_1, \mathbf{\Pi\Lambda}_2, \mathbf{\Pi\Lambda}_3]\!]$.)

Warmup: a CPD with full rank factors

Let's prove (essential) uniqueness of the CPD when $\mathbf{A}, \mathbf{B}, \mathbf{C}$ all have full column rank. Write

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D}, \quad \text{where} \quad \mathsf{D} = \sum_{r=1}^{R} \boldsymbol{e}_r^{\otimes 3} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!] .$$

Claim: The CPD of X is unique iff the CPD of D is unique.

(Uniqueness of $\mathsf{D} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$ means that every CPD of D has the form $\mathsf{D} = [\![\mathbf{\Pi\Lambda}_1, \mathbf{\Pi\Lambda}_2, \mathbf{\Pi\Lambda}_3]\!]$.)

Proof: $\Rightarrow$) Writing a CPD $\mathsf{D} = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$ of rank R, we have

$$\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \mathsf{D} = (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot [(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{D}] = [\![\mathbf{AU}, \mathbf{BV}, \mathbf{CW}]\!] .$$

By uniqueness of the CPD of X, we must have

$$\mathbf{AU} = \mathbf{A\Pi\Lambda}_1, \quad \mathbf{BV} = \mathbf{B\Pi\Lambda}_2, \quad \mathbf{CW} = \mathbf{C\Pi\Lambda}_3$$

with $\mathbf{\Lambda}_1\mathbf{\Lambda}_2\mathbf{\Lambda}_3 = \mathbf{I}$. The result then follows from all factors admiting left inverses.

$\Leftarrow$) See exercices.

■

Warmup: uniqueness of $[\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$

Proposition: The CPD $\mathsf{D} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$ is unique.

Proof sketch: Take $\mathsf{D} = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$. From

$$\mathbf{I}(\mathbf{I} \odot \mathbf{I})^{\mathsf{T}} = \mathbf{U}(\mathbf{W} \odot \mathbf{V})^{\mathsf{T}}$$

we see that $\operatorname{rank} \mathbf{U} = R$. By symmetry of roles, $\operatorname{rank} \mathbf{U} = \operatorname{rank} \mathbf{V} = \operatorname{rank} \mathbf{W} = R$.

Warmup: uniqueness of $[\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$

Proposition: The CPD $\mathsf{D} = [\![\mathbf{I}, \mathbf{I}, \mathbf{I}]\!]$ is unique.

Proof sketch: Take $\mathsf{D} = [\![\mathbf{U}, \mathbf{V}, \mathbf{W}]\!]$. From

$$\mathbf{I}(\mathbf{I} \odot \mathbf{I})^{\mathsf{T}} = \mathbf{U}(\mathbf{W} \odot \mathbf{V})^{\mathsf{T}}$$

we see that $\operatorname{rank} \mathbf{U} = R$. By symmetry of roles, $\operatorname{rank} \mathbf{U} = \operatorname{rank} \mathbf{V} = \operatorname{rank} \mathbf{W} = R$.

Now, write the jth slice of D

$$\boldsymbol{e}_j \boldsymbol{e}_j^{\mathsf{T}} = \mathbf{U} \operatorname{Diag}(\bar{\boldsymbol{v}}_j) \mathbf{W}^{\mathsf{T}},$$

where $\bar{\boldsymbol{v}}_j$ is the jth row of $\mathbf{V}$. As $\mathbf{U}, \mathbf{W}$ have rank R, $\bar{\boldsymbol{v}}_j$ can only have one nonzero entry. Suppose $\bar{\boldsymbol{v}}_j = c_j \boldsymbol{e}_{\ell_j}$, with $c_j \neq 0$. It follows that

$$\boldsymbol{e}_j \boldsymbol{e}_j^{\mathsf{T}} = c_j \boldsymbol{u}_{\ell_j} \boldsymbol{w}_{\ell_j}^{\mathsf{T}},$$

and we are forced to take both $\boldsymbol{u}_{\ell_j}$ and $\boldsymbol{w}_{\ell_j}$ proportional to $\boldsymbol{e}_j$. Furthermore, the ℓ_j are all distinct (since $\operatorname{rank} \mathbf{V} = R$). Hence,

$$\mathbf{U} = \mathbf{\Pi} \mathbf{\Lambda}_1, \quad \mathbf{V} = \mathbf{\Pi} \mathbf{\Lambda}_2, \quad \mathbf{U} = \mathbf{\Pi} \mathbf{\Lambda}_3,$$

with $\mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \mathbf{\Lambda}_3 = \mathbf{I}$ and $\mathbf{\Pi} = \begin{pmatrix} \boldsymbol{e}_{\ell_1} & \dots & \boldsymbol{e}_{\ell_R} \end{pmatrix}$.

■

Uniqueness under relaxed assumptions

Previous result: uniqueness of CPD is already stronger than that of $\mathbf{A}\mathbf{B}^{\mathsf{T}}$.

Uniqueness under relaxed assumptions

Previous result: uniqueness of CPD is already stronger than that of $\mathbf{A}\mathbf{B}^\mathsf{T}$.

In particular, it imples:

Corollary: If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sampled from absolutely continuous distributions (w.r.t. Lebesgue) and $\min\{I, J, K\} \geq R$, then $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is essentially unique.

Uniqueness under relaxed assumptions

Previous result: uniqueness of CPD is already stronger than that of $\mathbf{A}\mathbf{B}^\mathsf{T}$.

In particular, it imples:

Corollary: If $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are sampled from absolutely continuous distributions (w.r.t. Lebesgue) and $\min\{I, J, K\} \geq R$, then $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is essentially unique.

But it turns out that uniqueness holds much more generally!

The Kruskal rank

Def: The Kruskal rank of $\mathbf{A}$ is the largest number $k_{\mathbf{A}}$ such that every subset of $k_{\mathbf{A}}$ columns of $\mathbf{A}$ is linearly independent ($=$ spark $\mathbf{A} - 1$ in compressive sensing).

In particular, $k_{\mathbf{A}} = 0$ iff $\mathbf{A}$ has a null col.

Intuition: measures the "degree" of (in)dependence of the cols of $\mathbf{A} \in \mathbb{R}^{I \times R}$.

In particular,

$$k_{\mathbf{A}} \leq \operatorname{rank} \mathbf{A}, \quad \text{with equality if} \quad \operatorname{rank} \mathbf{A} = R.$$

Plays a central role in Kruskal's celebrated uniqueness result[1] & extensions.

(This is Joseph Kruskal, not to be confused with his brothers Willam Kruskal, statistician & author of the Kruskal-Wallis test, and Martin Kruskal, physicist.)

1: Kruskal, 1977

Kruskal rank: examples

In all examples below, $\{\boldsymbol{a}_i\}$ forms a linearly independent family.

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_1 + \boldsymbol{a}_2 \end{pmatrix}$.

Kruskal rank: examples

In all examples below, $\{\boldsymbol{a}_i\}$ forms a linearly independent family.

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_1 + \boldsymbol{a}_2 \end{pmatrix}$.

$k_{\mathbf{A}} = 2 = \operatorname{rank} \mathbf{A} < R$. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & -2\boldsymbol{a}_1 \end{pmatrix}$.

Kruskal rank: examples

In all examples below, $\{\boldsymbol{a}_i\}$ forms a linearly independent family.

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_1 + \boldsymbol{a}_2 \end{pmatrix}$.

$k_{\mathbf{A}} = 2 = \operatorname{rank} \mathbf{A} < R$. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & -2\boldsymbol{a}_1 \end{pmatrix}$.

$k_{\mathbf{A}} = 1 < \operatorname{rank} \mathbf{A} = 2 < R$, since $\mathbf{A}$ holds a pair of collinear columns. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 + \boldsymbol{a}_2 & \boldsymbol{a}_2 + \boldsymbol{a}_3 & \boldsymbol{a}_3 + \boldsymbol{a}_1 & \boldsymbol{a}_1 + 2\boldsymbol{a}_2 + \boldsymbol{a}_3 \end{pmatrix}$.

Kruskal rank: examples

In all examples below, $\{\boldsymbol{a}_i\}$ forms a linearly independent family.

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_1 + \boldsymbol{a}_2 \end{pmatrix}$.

$k_{\mathbf{A}} = 2 = \operatorname{rank} \mathbf{A} < R$. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & -2\boldsymbol{a}_1 \end{pmatrix}$.

$k_{\mathbf{A}} = 1 < \operatorname{rank} \mathbf{A} = 2 < R$, since $\mathbf{A}$ holds a pair of collinear columns. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 + \boldsymbol{a}_2 & \boldsymbol{a}_2 + \boldsymbol{a}_3 & \boldsymbol{a}_3 + \boldsymbol{a}_1 & \boldsymbol{a}_1 + 2\boldsymbol{a}_2 + \boldsymbol{a}_3 \end{pmatrix}$.

$k_{\mathbf{A}} = 2 < \operatorname{rank} \mathbf{A} = 3 < R$. □

Example: $\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_3 & \mathbf{0} \end{pmatrix}$.

The zero column implies $k_{\mathbf{A}} = 0$. □

Some *necessary* conditions

Claim: Uniqueness of $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ cannot hold if $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}\} < 2$.

Remark: This condition can be relaxed for higher orders.

Some *necessary* conditions

Claim: Uniqueness of $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ cannot hold if $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}\} < 2$.

Remark: This condition can be relaxed for higher orders.

Claim: If $\operatorname{rank} \mathsf{X} < R$, then no rank-$R$ PD of X can be unique.

Proof: Any term $\boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$ of an $(R-1)$-rank PD can always be split into

$$(\boldsymbol{a}_r - \boldsymbol{w}) \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r + \boldsymbol{w} \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r, \qquad \forall \boldsymbol{w} \in \mathbb{R}^I.$$

■

Some *necessary* conditions

Claim: Uniqueness of $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ cannot hold if $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}\} < 2$.

Remark: This condition can be relaxed for higher orders.

Claim: If $\operatorname{rank}\mathsf{X} < R$, then no rank-$R$ PD of X can be unique.

Proof: Any term $\boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$ of an $(R-1)$-rank PD can always be split into

$$(\boldsymbol{a}_r - \boldsymbol{w}) \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r + \boldsymbol{w} \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r, \qquad \forall \boldsymbol{w} \in \mathbb{R}^I.$$

■

Claim: If $\min\{\operatorname{rank}\mathbf{B} \odot \mathbf{A}, \operatorname{rank}\mathbf{C} \odot \mathbf{B}, \operatorname{rank}\mathbf{C} \odot \mathbf{A}\} < R$, then $\operatorname{rank}\mathsf{X} < R$.

Proof: If, say, $\operatorname{rank}\mathbf{B} \odot \mathbf{A} < R$, then take any nonzero $\boldsymbol{v} \in \ker \mathbf{B} \odot \mathbf{A}$ and choose some $\boldsymbol{z}$ such that $\mathbf{C} + \boldsymbol{z}\boldsymbol{v}^\mathsf{T}$ has a null column to write a rank-$(R-1)$ PD:

$$\begin{aligned}
\mathsf{X}_{(3)} &= \mathbf{C}\,(\mathbf{B} \odot \mathbf{A})^\mathsf{T} \\
&= \mathbf{C}\,(\mathbf{B} \odot \mathbf{A})^\mathsf{T} + \left[(\mathbf{B} \odot \mathbf{A})\boldsymbol{v}\boldsymbol{z}^\mathsf{T}\right]^\mathsf{T} \\
&= (\mathbf{C} + \boldsymbol{z}\boldsymbol{v}^\mathsf{T})(\mathbf{B} \odot \mathbf{A})^\mathsf{T}.
\end{aligned}$$

■

Uniqueness under relaxed assumptions (cont'd)

Thm (Sidiropoulos & al., 2017): Let $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{R \times R}$ and $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$. If

$$\operatorname{rank} \mathbf{C} = R \quad \text{and} \quad k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2,$$

then $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are essentially unique.

Uniqueness under relaxed assumptions (cont'd)

Thm (Sidiropoulos & al., 2017): Let $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{R \times R}$ and $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$. If

$$\operatorname{rank} \mathbf{C} = R \quad \text{and} \quad k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2,$$

then $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are essentially unique.

The proof relies on a special case of Kruskal's permutation lemma: the main tool used for Kruskal's 1977 proof of the most general known result.

Lemma (Sidiropoulos & al., 2017): Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathbb{R}^{R \times R}$ be two nonsingular matrices. If

$$\forall \boldsymbol{v} \in \mathbb{R}^{K} \text{ such that } \left\| \boldsymbol{v}^{\mathsf{T}} \tilde{\mathbf{C}} \right\|_0 = 1 \quad \text{we have} \quad \left\| \boldsymbol{v}^{\mathsf{T}} \mathbf{C} \right\|_0 = 1,$$

then $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3$, where $\mathbf{\Pi}, \mathbf{\Lambda}_3$ are as before.

Uniqueness under relaxed assumptions (cont'd)

Thm (Sidiropoulos & al., 2017): Let $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, $\mathbf{C} \in \mathbb{R}^{R \times R}$ and $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$. If

$$\operatorname{rank} \mathbf{C} = R \quad \text{and} \quad k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2,$$

then $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are essentially unique.

The proof relies on a special case of Kruskal's permutation lemma: the main tool used for Kruskal's 1977 proof of the most general known result.

Lemma (Sidiropoulos & al., 2017): Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathbb{R}^{R \times R}$ be two nonsingular matrices. If

$$\forall \boldsymbol{v} \in \mathbb{R}^{K} \text{ such that } \left\| \boldsymbol{v}^{\mathsf{T}} \tilde{\mathbf{C}} \right\|_0 = 1 \quad \text{we have} \quad \left\| \boldsymbol{v}^{\mathsf{T}} \mathbf{C} \right\|_0 = 1,$$

then $\tilde{\mathbf{C}} = \mathbf{C} \boldsymbol{\Pi} \boldsymbol{\Lambda}_3$, where $\boldsymbol{\Pi}, \boldsymbol{\Lambda}_3$ are as before.

Proof: By the above condition & the fact that $\operatorname{rank} \tilde{\mathbf{C}}^{-1} \mathbf{C} = R$,

$$\tilde{\mathbf{C}}^{-1} \tilde{\mathbf{C}} = \mathbf{I} \quad \Rightarrow \quad \tilde{\mathbf{C}}^{-1} \mathbf{C} = \boldsymbol{\Lambda}_3^{-1} \boldsymbol{\Pi}^{\mathsf{T}}.$$

■

Uniqueness under relaxed assumptions (cont'd)

Proof (of the Thm): We will shortly see that $k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2$ implies $k_{\mathbf{B}\odot\mathbf{A}} = R = \operatorname{rank}(\mathbf{B} \odot \mathbf{A})$. Hence, combined with

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}}$$

we get $\operatorname{rank}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = \operatorname{rank} \tilde{\mathbf{C}} = R$.

Uniqueness under relaxed assumptions (cont'd)

Proof (of the Thm): We will shortly see that $k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2$ implies $k_{\mathbf{B}\odot\mathbf{A}} = R = \text{rank}(\mathbf{B} \odot \mathbf{A})$. Hence, combined with

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}}$$

we get $\text{rank}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = \text{rank}\,\tilde{\mathbf{C}} = R$. Next, take a linear combination of slices

$$(\cdot, \cdot, \boldsymbol{v}) \cdot \mathsf{X} = \sum_{r=1}^{R} (\boldsymbol{v})_r \mathbf{X}_r = \mathbf{A}\,\text{Diag}\,(\boldsymbol{v}^{\mathsf{T}}\mathbf{C})\mathbf{B}^{\mathsf{T}} = \tilde{\mathbf{A}}\,\text{Diag}\,(\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{C}})\tilde{\mathbf{B}}^{\mathsf{T}}.$$

But, $\text{rank}\,\mathbf{A}\,\text{Diag}(\boldsymbol{v}^{\mathsf{T}}\mathbf{C})\mathbf{B}^{\mathsf{T}} = \text{rank}\,\tilde{\mathbf{A}}\,\text{Diag}(\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{C}})\tilde{\mathbf{B}}^{\mathsf{T}} \leq \text{rank}\,\text{Diag}(\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{C}}) = \left\|\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{C}}\right\|_0$.

Suppose now $\left\|\boldsymbol{v}^{\mathsf{T}}\tilde{\mathbf{C}}\right\|_0 = 1$. Then $\text{rank}\,\mathbf{A}\,\text{Diag}(\boldsymbol{v}^{\mathsf{T}}\mathbf{C})\mathbf{B}^{\mathsf{T}} \leq 1$ and we want to show $\left\|\boldsymbol{v}^{\mathsf{T}}\mathbf{C}\right\|_0 = 1$ (to use the Lemma).

Uniqueness under relaxed assumptions (cont'd)

Proof (of the Thm): We will shortly see that $k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2$ implies $k_{\mathbf{B}\odot\mathbf{A}} = R = \text{rank}(\mathbf{B} \odot \mathbf{A})$. Hence, combined with

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^\mathsf{T} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^\mathsf{T}$$

we get $\text{rank}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = \text{rank}\,\tilde{\mathbf{C}} = R$. Next, take a linear combination of slices

$$(\cdot, \cdot, \boldsymbol{v}) \cdot \mathsf{X} = \sum_{r=1}^{R} (\boldsymbol{v})_r \mathbf{X}_r = \mathbf{A}\,\text{Diag}\,(\boldsymbol{v}^\mathsf{T}\mathbf{C})\mathbf{B}^\mathsf{T} = \tilde{\mathbf{A}}\,\text{Diag}\,(\boldsymbol{v}^\mathsf{T}\tilde{\mathbf{C}})\tilde{\mathbf{B}}^\mathsf{T}.$$

But, $\text{rank}\,\mathbf{A}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\mathbf{C})\mathbf{B}^\mathsf{T} = \text{rank}\,\tilde{\mathbf{A}}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\tilde{\mathbf{C}})\tilde{\mathbf{B}}^\mathsf{T} \leq \text{rank}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\tilde{\mathbf{C}}) = \left\|\boldsymbol{v}^\mathsf{T}\tilde{\mathbf{C}}\right\|_0$.

Suppose now $\left\|\boldsymbol{v}^\mathsf{T}\tilde{\mathbf{C}}\right\|_0 = 1$. Then $\text{rank}\,\mathbf{A}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\mathbf{C})\mathbf{B}^\mathsf{T} \leq 1$ and we want to show $\left\|\boldsymbol{v}^\mathsf{T}\mathbf{C}\right\|_0 = 1$ (to use the Lemma).

Let $\bar{\mathbf{A}} \in \mathbb{R}^{I\times\bar{R}}$ hold the subset of $\bar{R} := \left\|\boldsymbol{v}^\mathsf{T}\mathbf{C}\right\|_0$ cols $\boldsymbol{a}_i$ of $\mathbf{A}$ such that $(\boldsymbol{v}^\mathsf{T}\mathbf{C})_i \neq 0$, similarly for $\bar{\mathbf{B}} \in \mathbb{R}^{J\times\bar{R}}$, $\bar{\mathbf{C}} \in \mathbb{R}^{R\times\bar{R}}$.

Then,

$$\mathbf{A}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\mathbf{C})\mathbf{B}^\mathsf{T} = \bar{\mathbf{A}}\,\text{Diag}(\boldsymbol{v}^\mathsf{T}\bar{\mathbf{C}})\bar{\mathbf{B}}^\mathsf{T}.$$

Uniqueness under relaxed assumptions (cont'd)

By the Sylvester inequality (see handout),

$$\operatorname{rank} \bar{\mathbf{A}} \operatorname{Diag}(v^{\mathsf{T}} \bar{\mathbf{C}}) \bar{\mathbf{B}}^{\mathsf{T}} \geq \operatorname{rank} \bar{\mathbf{A}} + \operatorname{rank} \bar{\mathbf{B}} - \bar{R} \geq \min\left\{k_{\mathbf{A}}, \bar{R}\right\} + \min\left\{k_{\mathbf{B}}, \bar{R}\right\} - \bar{R}.$$

Uniqueness under relaxed assumptions (cont'd)

By the Sylvester inequality (see handout),

$$\operatorname{rank} \bar{\mathbf{A}} \operatorname{Diag}(\boldsymbol{v}^{\mathsf{T}} \bar{\mathbf{C}}) \bar{\mathbf{B}}^{\mathsf{T}} \geq \operatorname{rank} \bar{\mathbf{A}} + \operatorname{rank} \bar{\mathbf{B}} - \bar{R} \geq \min\left\{k_{\mathbf{A}}, \bar{R}\right\} + \min\left\{k_{\mathbf{B}}, \bar{R}\right\} - \bar{R}.$$

Hence, since $\operatorname{rank} \bar{\mathbf{A}} \operatorname{Diag}(\boldsymbol{v}^{\mathsf{T}} \bar{\mathbf{C}}) \bar{\mathbf{B}}^{\mathsf{T}} = \operatorname{rank} \mathbf{A} \operatorname{Diag}(\boldsymbol{v}^{\mathsf{T}} \mathbf{C}) \mathbf{B}^{\mathsf{T}} \leq 1$ by assumption,

$$\min\left\{k_{\mathbf{A}}, \bar{R}\right\} + \min\left\{k_{\mathbf{B}}, \bar{R}\right\} - \bar{R} \leq 1.$$

Uniqueness under relaxed assumptions (cont'd)

By the Sylvester inequality (see handout),

$$\operatorname{rank} \bar{\mathbf{A}} \operatorname{Diag}(\boldsymbol{v}^\mathsf{T}\bar{\mathbf{C}})\bar{\mathbf{B}}^\mathsf{T} \geq \operatorname{rank} \bar{\mathbf{A}} + \operatorname{rank} \bar{\mathbf{B}} - \bar{R} \geq \min\left\{k_{\mathbf{A}}, \bar{R}\right\} + \min\left\{k_{\mathbf{B}}, \bar{R}\right\} - \bar{R}.$$

Hence, since $\operatorname{rank} \bar{\mathbf{A}} \operatorname{Diag}(\boldsymbol{v}^\mathsf{T}\bar{\mathbf{C}})\bar{\mathbf{B}}^\mathsf{T} = \operatorname{rank} \mathbf{A} \operatorname{Diag}(\boldsymbol{v}^\mathsf{T}\mathbf{C})\mathbf{B}^\mathsf{T} \leq 1$ by assumption,

$$\min\left\{k_{\mathbf{A}}, \bar{R}\right\} + \min\left\{k_{\mathbf{B}}, \bar{R}\right\} - \bar{R} \leq 1.$$

Three cases arise:

(i) If $\bar{R} \leq \min\{k_{\mathbf{A}}, k_{\mathbf{B}}\}$, then $\bar{R} \leq 1$, hence $\bar{R} = 1$ since $\boldsymbol{v}^\mathsf{T}\mathbf{C}$ cannot vanish.

(ii) The case $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}\} \leq \bar{R} \leq \max\{k_{\mathbf{A}}, k_{\mathbf{B}}\}$ is impossible, as it implies $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}\} \leq 1$ (violating $k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2$).

(iii) Finally, $\max\{k_{\mathbf{A}}, k_{\mathbf{B}}\} \leq \bar{R}$ is equally impossible, since $k_{\mathbf{A}} + k_{\mathbf{B}} \geq R + 2$ by assumption, yielding

$$R + 1 \leq k_{\mathbf{A}} + k_{\mathbf{B}} - 1 \leq \bar{R} \leq R.$$

Conclusion: $\bar{R} = \left\|\boldsymbol{v}^\mathsf{T}\mathbf{C}\right\|_0 = 1$ as per the Lemma, and therefore $\tilde{\mathbf{C}} = \mathbf{C}\boldsymbol{\Pi}\boldsymbol{\Lambda}_3$.

Uniqueness under relaxed assumptions (cont'd)

Now, from

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}},$$

we pre-multiply by $\mathbf{C}^{-1}$ to get

$$(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = (\mathbf{B} \odot \mathbf{A})\mathbf{\Pi}\mathbf{\Lambda}_3^{-1} = (\mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2) \odot (\mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1),$$

where $\mathbf{\Lambda}_1\mathbf{\Lambda}_2 = \mathbf{\Lambda}_3^{-1}$, as claimed. ■

Uniqueness under relaxed assumptions (cont'd)

Now, from

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}},$$

we pre-multiply by $\mathbf{C}^{-1}$ to get

$$(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = (\mathbf{B} \odot \mathbf{A})\mathbf{\Pi}\mathbf{\Lambda}_3^{-1} = (\mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2) \odot (\mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1),$$

where $\mathbf{\Lambda}_1\mathbf{\Lambda}_2 = \mathbf{\Lambda}_3^{-1}$, as claimed. ■

Remark: The extension to tall $\mathbf{C} \in \mathbb{R}^{K \times R}$ is easy: if rank $\mathbf{C} = R$, then it contains a nonsingular $R \times R$ submatrix. Uniqueness and full rank of $\mathbf{A} \odot \mathbf{B}$ then yields uniqueness of the whole $\mathbf{C}$:

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}} = \tilde{\mathbf{C}}\left[(\mathbf{B} \odot \mathbf{A})\mathbf{\Pi}\mathbf{\Lambda}_3^{-1}\right]^{\mathsf{T}}$$

Uniqueness under relaxed assumptions (cont'd)

Now, from

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}},$$

we pre-multiply by $\mathbf{C}^{-1}$ to get

$$(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}}) = (\mathbf{B} \odot \mathbf{A})\mathbf{\Pi}\mathbf{\Lambda}_3^{-1} = (\mathbf{B}\mathbf{\Pi}\mathbf{\Lambda}_2) \odot (\mathbf{A}\mathbf{\Pi}\mathbf{\Lambda}_1),$$

where $\mathbf{\Lambda}_1\mathbf{\Lambda}_2 = \mathbf{\Lambda}_3^{-1}$, as claimed. ■

Remark: The extension to tall $\mathbf{C} \in \mathbb{R}^{K \times R}$ is easy: if rank $\mathbf{C} = R$, then it contains a nonsingular $R \times R$ submatrix. Uniqueness and full rank of $\mathbf{A} \odot \mathbf{B}$ then yields uniqueness of the whole $\mathbf{C}$:

$$\mathbf{C}(\mathbf{B} \odot \mathbf{A})^{\mathsf{T}} = \tilde{\mathbf{C}}(\tilde{\mathbf{B}} \odot \tilde{\mathbf{A}})^{\mathsf{T}} = \tilde{\mathbf{C}}\left[(\mathbf{B} \odot \mathbf{A})\mathbf{\Pi}\mathbf{\Lambda}_3^{-1}\right]^{\mathsf{T}}$$

Example: If $R = 3$ and

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_3 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \boldsymbol{b}_1 + \boldsymbol{b}_2 \end{pmatrix}, \ \mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 & \boldsymbol{c}_3 \end{pmatrix},$$

where distinct vectors are independent, then rank $\mathbf{C} = 3 = R$ and $k_{\mathbf{A}} + k_{\mathbf{B}} = 5 = R + 2$, hence $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ is unique. □

Kruskal's uniqueness theorem

Jiang & Sidiropoulos (2004) gave another, more accessible proof of Kruskal's permutation lemma in 2004. A simplified statement is:

Lemma (Stegeman & Sidiropoulos, 2007): Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathbb{R}^{K \times R}$ and $k_{\mathbf{C}} \geq 2$. If

$$\forall \boldsymbol{v} \in \mathbb{R}^{K}, \ \left\| \boldsymbol{v}^{\mathsf{T}} \mathbf{C} \right\|_0 \leq \left\| \boldsymbol{v}^{\mathsf{T}} \tilde{\mathbf{C}} \right\|_0,$$

then $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3$, where $\mathbf{\Pi}, \mathbf{\Lambda}_3$ are as before.

Kruskal's uniqueness theorem

Jiang & Sidiropoulos (2004) gave another, more accessible proof of Kruskal's permutation lemma in 2004. A simplified statement is:

Lemma (Stegeman & Sidiropoulos, 2007): Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathbb{R}^{K \times R}$ and $k_{\mathbf{C}} \geq 2$. If

$$\forall \boldsymbol{v} \in \mathbb{R}^{K}, \ \left\| \boldsymbol{v}^{\mathsf{T}} \mathbf{C} \right\|_0 \leq \left\| \boldsymbol{v}^{\mathsf{T}} \tilde{\mathbf{C}} \right\|_0,$$

then $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3$, where $\mathbf{\Pi}, \mathbf{\Lambda}_3$ are as before.

This is the main tool used in the proof of Kruskal's famous uniqueness result:

Thm (Kruskal, 1977): Let $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ of size $R > 1$, such that

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2.$$

Then, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are essentially unique (and $\operatorname{rank} \mathsf{X} = R$).

Kruskal's uniqueness theorem

Jiang & Sidiropoulos (2004) gave another, more accessible proof of Kruskal's permutation lemma in 2004. A simplified statement is:

Lemma (Stegeman & Sidiropoulos, 2007): Let $\mathbf{C}, \tilde{\mathbf{C}} \in \mathbb{R}^{K \times R}$ and $k_{\mathbf{C}} \geq 2$. If

$$\forall \boldsymbol{v} \in \mathbb{R}^{K}, \ \left\| \boldsymbol{v}^{\mathsf{T}} \mathbf{C} \right\|_0 \leq \left\| \boldsymbol{v}^{\mathsf{T}} \tilde{\mathbf{C}} \right\|_0,$$

then $\tilde{\mathbf{C}} = \mathbf{C}\mathbf{\Pi}\mathbf{\Lambda}_3$, where $\mathbf{\Pi}, \mathbf{\Lambda}_3$ are as before.

This is the main tool used in the proof of Kruskal's famous uniqueness result:

Thm (Kruskal, 1977): Let $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ of size $R > 1$, such that

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2.$$

Then, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are essentially unique (and $\operatorname{rank} \mathsf{X} = R$).

Generically, $k_{\mathbf{A}} = \min\{I, R\}$, $k_{\mathbf{B}} = \min\{J, R\}$ and $k_{\mathbf{C}} = \min\{K, R\}$, so that uniqueness holds whenever $R \leq \frac{1}{2}(\min\{I, R\} + \min\{J, R\} + \min\{K, R\} - 2)$.

Is Kruskal's condition also necessary?

For $R = 2$, the condition $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2 = 6$ is also necessary, since otherwise $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}\} < 2$.

Is Kruskal's condition also necessary?

For $R = 2$, the condition $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2 = 6$ is also necessary, since otherwise $\min\{k_{\mathbf{A}}, k_{\mathbf{B}}, k_{\mathbf{C}}\} < 2$.

ten Berge and Sidiropoulos (2002) examined necessity for higher R:

- Still necessary for $R = 3$ (proof by enumeration of cases).
- Necessary also for $R = 4$ if ranks = k-ranks. Otherwise, no (see exercises).
- Not necessary in general for $R > 5$. Finding necessary and sufficient conditions is complicated:

> "It also has been shown that, in cases of small k-rank, the particular pattern of zeros, after pretransformation to have identity submatrices in $\mathbf{A}, \mathbf{B}, \mathbf{C}$, may have a decisive impact on uniqueness. This implies that attempts to derive necessary and sufficient conditions for uniqueness are doomed unless they take that very pattern into account."

Extension to higher orders

Kruskal's theorem was extended by Sidiropoulos & Bro (2000) by reduction to the third-order case ($d = 3$) via the following additivity lemma:

Lemma (Sidiropoulos & Bro, 2000): Let $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$. If $k_{\mathbf{A}} k_{\mathbf{B}} \neq 0$, then

$$R \geq k_{\mathbf{A} \odot \mathbf{B}} \geq \min \left\{ R, k_{\mathbf{A}} + k_{\mathbf{B}} - 1 \right\}.$$

(Otherwise, $k_{\mathbf{A} \odot \mathbf{B}} = 0$.)

Extension to higher orders

Kruskal's theorem was extended by Sidiropoulos & Bro (2000) by reduction to the third-order case ($d = 3$) via the following additivity lemma:

Lemma (Sidiropoulos & Bro, 2000): Let $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$. If $k_{\mathbf{A}} k_{\mathbf{B}} \neq 0$, then

$$R \geq k_{\mathbf{A} \odot \mathbf{B}} \geq \min\left\{R, k_{\mathbf{A}} + k_{\mathbf{B}} - 1\right\}.$$

(Otherwise, $k_{\mathbf{A} \odot \mathbf{B}} = 0$.)

For instance, for $d = 4$ and $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{G}, \mathbf{H}]\!]_R$:

$$x_{ijk\ell} = \sum_{r=1}^{R} (\boldsymbol{a}_r)_i \, (\boldsymbol{b}_r)_j \, (\boldsymbol{g}_r)_k \, (\boldsymbol{h}_r)_\ell, \quad \text{with } k_{\mathbf{A}} \geq k_{\mathbf{B}} \geq k_{\mathbf{G}} \geq k_{\mathbf{H}},$$

define $\mathsf{Y} \in \mathbb{R}^{I \times J \times M}$ with $M = KL$ via the bijection $(k, \ell) \leftrightarrow m$:

$$y_{ijm(k,\ell)} = x_{ijk\ell}, \qquad m(k, \ell) = (k-1)L + \ell$$

Extension to higher orders (cont'd)

This implies

$$y_{ijm} = \sum_{r=1}^{R} (\boldsymbol{a}_r)_i \, (\boldsymbol{b}_r)_j \, (\boldsymbol{c}_r)_m, \quad \text{with} \quad \mathbf{C} = \mathbf{G} \odot \mathbf{H}.$$

Extension to higher orders (cont'd)

This implies

$$y_{ijm} = \sum_{r=1}^{R} (\boldsymbol{a}_r)_i \, (\boldsymbol{b}_r)_j \, (\boldsymbol{c}_r)_m, \quad \text{with} \quad \mathbf{C} = \mathbf{G} \odot \mathbf{H}.$$

If this decomposition is unique, so is that of X. By Kruskal's thm, this holds when

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} = k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{G} \odot \mathbf{H}} \geq 2R + 2.$$

We have two cases:

(i) If $k_{\mathbf{G}} + k_{\mathbf{H}} - 1 \leq R$, then by the lemma $k_{\mathbf{C}} \geq k_{\mathbf{G}} + k_{\mathbf{H}} - 1$. In that case, if

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{G}} + k_{\mathbf{H}} \geq 2R + 3,$$

then Kruskal's bound is met.

Extension to higher orders (cont'd)

This implies

$$y_{ijm} = \sum_{r=1}^{R} (\boldsymbol{a}_r)_i \, (\boldsymbol{b}_r)_j \, (\boldsymbol{c}_r)_m, \quad \text{with} \quad \mathbf{C} = \mathbf{G} \odot \mathbf{H}.$$

If this decomposition is unique, so is that of X. By Kruskal's thm, this holds when

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} = k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{G} \odot \mathbf{H}} \geq 2R + 2.$$

We have two cases:

(i) If $k_{\mathbf{G}} + k_{\mathbf{H}} - 1 \leq R$, then by the lemma $k_{\mathbf{C}} \geq k_{\mathbf{G}} + k_{\mathbf{H}} - 1$. In that case, if

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{G}} + k_{\mathbf{H}} \geq 2R + 3,$$

then Kruskal's bound is met.

(ii) If $k_{\mathbf{G}} + k_{\mathbf{H}} - 1 > R$, then $k_{\mathbf{C}} = R$. Combined with $k_{\mathbf{A}} + k_{\mathbf{B}} \geq k_{\mathbf{G}} + k_{\mathbf{H}} > R + 1$, we get

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2.$$

Extension to higher orders (cont'd)

This argument can be generalized, yielding:

Thm (Sidiropoulos & Bro, 2000): If

$$\sum_{i=1}^{d} k_{\mathbf{A}^{(i)}} \geq 2R + d - 1,$$

then the decomposition $\mathsf{X} = \left[\!\left[\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(d)} \right]\!\right]_R$ is essentially unique.

Extension to higher orders (cont'd)

This argument can be generalized, yielding:

Thm (Sidiropoulos & Bro, 2000): If

$$\sum_{i=1}^{d} k_{\mathbf{A}^{(i)}} \geq 2R + d - 1,$$

then the decomposition $\mathsf{X} = \left[\!\left[\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(d)} \right]\!\right]_R$ is essentially unique.

Remarks:

- Uniqueness now does not need $\min_i k_{\mathbf{A}^{(i)}} > 1$. Example: $R = 2$, $d = 4$, $k_{\mathbf{A}^{(1)}} = k_{\mathbf{A}^{(2)}} = k_{\mathbf{A}^{(3)}} = R = 2$ and $k_{\mathbf{A}^{(4)}} = 1$.
- The condition becomes less restrictive as d grows (since it "distributes" the term $2R - 1$ among d factors).

Generic uniqueness

Generic factors: If the columns of each $\mathbf{A}^{(i)} \in \mathbb{R}^{N_i \times R}$ are independently drawn from an abs. continuous distribution, then $\mathbb{P}(k_{\mathbf{A}^{(i)}} = \min\{N_i, R\}) = 1$.

In this case, the sufficient condition reduces to

$$\sum_{i=1}^{d} \min\{N_i, R\} \geq 2R + d - 1.$$

In particular, if d grows by one, the LHS grows by at least two (since $N_i \geq 2$).

Hence, the bound is eventually satisfied as d grows.

Generic uniqueness

Generic factors: If the columns of each $\mathbf{A}^{(i)} \in \mathbb{R}^{N_i \times R}$ are independently drawn from an abs. continuous distribution, then $\mathbb{P}(k_{\mathbf{A}^{(i)}} = \min\{N_i, R\}) = 1$.

In this case, the sufficient condition reduces to

$$\sum_{i=1}^{d} \min\{N_i, R\} \geq 2R + d - 1.$$

In particular, if d grows by one, the LHS grows by at least two (since $N_i \geq 2$).

Hence, the bound is eventually satisfied as d grows.

Generic tensor of subgeneric rank: A generic tensor with rank bounded as

$$R \leq \operatorname{grank}(N_1, \ldots, N_d) - 1 = \left\lceil \frac{\prod_{i=1}^{d} N_i}{1 + \sum_{i=1}^{d}(N_i - 1)} \right\rceil - 1$$

admits a unique CPD almost surely if $\prod_{i=1}^{d} N_i \leq 15000$ (w/ some exceptions).[1]

(Compare the above bounds on R.)

1: Chiantini & al., 2014

Uniqueness of the block-term decomposition

Model:

$$\mathsf{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} (\underbrace{\mathbf{A}_r}_{N_1 \times L_r} \underbrace{\mathbf{B}_r^{\mathsf{T}}}_{N_2 \times L_r}) \otimes \boldsymbol{c}_r, \qquad \operatorname{rank} \mathbf{H}_r \leq L_r$$

Factors: $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \dots & \mathbf{A}_R \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \dots & \mathbf{B}_R \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_R \end{pmatrix}$.

Uniqueness of the block-term decomposition

Model:

$$\mathsf{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} (\underbrace{\mathbf{A}_r}_{N_1 \times L_r} \underbrace{\mathbf{B}_r^{\mathsf{T}}}_{N_2 \times L_r}) \otimes \boldsymbol{c}_r, \qquad \operatorname{rank} \mathbf{H}_r \leq L_r$$

Factors: $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \dots & \mathbf{A}_R \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \dots & \mathbf{B}_R \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_R \end{pmatrix}$.

Thm (De Lathauwer, 2008): If $k_{\mathbf{C}} > 1$ and $\mathbf{A}, \mathbf{B}$ are full column rank, then the above BTD essentially unique.

Remark: This requires (and holds generically when) $\sum_r L_r \leq \min\{N_1, N_2\}$.

Uniqueness of the block-term decomposition

Model:

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{c}_r = \sum_{r=1}^{R} (\underbrace{\mathbf{A}_r}_{N_1 \times L_r} \underbrace{\mathbf{B}_r^{\mathsf{T}}}_{N_2 \times L_r}) \otimes \boldsymbol{c}_r, \qquad \operatorname{rank} \mathbf{H}_r \leq L_r$$

Factors: $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \dots & \mathbf{A}_R \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \dots & \mathbf{B}_R \end{pmatrix}$ and $\mathbf{C} = \begin{pmatrix} \boldsymbol{c}_1 & \dots & \boldsymbol{c}_R \end{pmatrix}$.

Thm (De Lathauwer, 2008): If $k_{\mathbf{C}} > 1$ and $\mathbf{A}, \mathbf{B}$ are full column rank, then the above BTD essentially unique.

Remark: This requires (and holds generically when) $\sum_r L_r \leq \min\{N_1, N_2\}$.

Thm (De Lathauwer, 2011): If $\mathbf{C}$ has full column rank and every nontrivial combination of $\mathbf{H}_r$ yields a matrix of higher rank:

$$\operatorname{rank} \textstyle\sum_{\ell=1}^{p} w_\ell \mathbf{H}_{r_\ell} > \max_\ell \operatorname{rank} \mathbf{H}_{r_\ell}, \quad w_\ell \neq 0,$$

then the BTD is unique. The latter condition is necessary.

Example: $\mathbf{H}_1 \otimes \boldsymbol{c}_1 + \mathbf{H}_2 \otimes \boldsymbol{c}_2 = (\mathbf{H}_1 + w\mathbf{H}_2) \otimes \boldsymbol{c}_1 + \mathbf{H}_2 \otimes (\boldsymbol{c}_2 - w\boldsymbol{c}_1)$. □

Rank-1 approx., tensor spectrum & power iteration

Low-rank decomposition

Up to now: focus on the properties of various (exact) tensor decompositions.

Hereafter: how to compute such decompositions?

Low-rank decomposition

Up to now: focus on the properties of various (exact) tensor decompositions.

Hereafter: how to compute such decompositions?

Warmup: Recovering the vectors that make up a rank-1 tensor

$$\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

Low-rank decomposition

Up to now: focus on the properties of various (exact) tensor decompositions.

Hereafter: how to compute such decompositions?

Warmup: Recovering the vectors that make up a rank-1 tensor

$$\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

That's easy: compute the dominant left singular vectors of

$$\mathbf{X}_{(1)} = \boldsymbol{a}(\boldsymbol{c} \boxtimes \boldsymbol{b})^\mathsf{T}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}(\boldsymbol{c} \boxtimes \boldsymbol{a})^\mathsf{T}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}(\boldsymbol{b} \boxtimes \boldsymbol{a})^\mathsf{T}$$

Low-rank decomposition

Up to now: focus on the properties of various (exact) tensor decompositions.

Hereafter: how to compute such decompositions?

Warmup: Recovering the vectors that make up a rank-1 tensor

$$\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$$

That's easy: compute the dominant left singular vectors of

$$\mathbf{X}_{(1)} = \boldsymbol{a}(\boldsymbol{c} \boxtimes \boldsymbol{b})^{\mathsf{T}}, \quad \mathbf{X}_{(2)} = \boldsymbol{b}(\boldsymbol{c} \boxtimes \boldsymbol{a})^{\mathsf{T}}, \quad \mathbf{X}_{(3)} = \boldsymbol{c}(\boldsymbol{b} \boxtimes \boldsymbol{a})^{\mathsf{T}}$$

Now let's make things more interesting: recover $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ from

$$\mathsf{X} = \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} + \mathsf{W},$$

where W is a noise term ($\mathbb{R}^{N_1 \times N_2 \times N_3}$-valued realization of a random tensor).

What about using the same solution?

Low-rank approximation (LRA)

Computing the dominant singular vectors of X now amounts to looking for a rank-one approximation of it.

It is a natural way of estimating the low-rank signal $\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$ "planted" in X (though not optimal, as we'll see).

More generally, low-rank models are virtually always computed by means of low-rank approximation (LRA) algorithms.

Low-rank approximation (LRA)

Computing the dominant singular vectors of X now amounts to looking for a rank-one approximation of it.

It is a natural way of estimating the low-rank signal $\boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}$ "planted" in X (though not optimal, as we'll see).

More generally, low-rank models are virtually always computed by means of low-rank approximation (LRA) algorithms.

Several natural questions arise:

- In which sense should one approximate X?
- How should one formulate the approximation problem?
- Which algorithms can be deployed, and how do they perform?
- What is the best possible performance among all possible algorithms?
- Under which conditions can LRA recover the sought signal/information?

A rank-1 model with Gaussian noise

Take the rank-1-Gaussian model (symmetric, for simplicity) of size $N \times \cdots \times N$:

$$\mathsf{X} = \lambda\, \boldsymbol{a}^{\otimes 3} + \mathsf{W}, \qquad \boldsymbol{a}^{\otimes 3} := \boldsymbol{a} \otimes \boldsymbol{a} \otimes \boldsymbol{a},$$

where $\|\boldsymbol{a}\| = 1$, $\lambda \in \mathbb{R}$ is an SNR parameter and W is Gaussian & symmetric:

$$p(\mathsf{W}) = \frac{1}{Z_3(N)} \exp\left(-\frac{N}{2} \|\mathsf{W}\|_{\mathsf{F}}^2\right)$$

A rank-1 model with Gaussian noise

Take the rank-1-Gaussian model (symmetric, for simplicity) of size $N \times \cdots \times N$:

$$\mathsf{X} = \lambda\, \boldsymbol{a}^{\otimes 3} + \mathsf{W}, \qquad \boldsymbol{a}^{\otimes 3} := \boldsymbol{a} \otimes \boldsymbol{a} \otimes \boldsymbol{a},$$

where $\|\boldsymbol{a}\| = 1$, $\lambda \in \mathbb{R}$ is an SNR parameter and W is Gaussian & symmetric:

$$p(\mathsf{W}) \;=\; \frac{1}{Z_3(N)}\, \exp\left(-\frac{N}{2}\, \|\mathsf{W}\|_{\mathsf{F}}^2\right)$$

Consequently, $\quad p(\mathsf{X} \,|\, \lambda, \boldsymbol{a}) \;\sim\; \exp\left(-\frac{N}{2}\, \left\|\mathsf{X} - \lambda\, \boldsymbol{a}^{\otimes 3}\right\|_{\mathsf{F}}^2\right) \quad$ and:

Maximum likelihood estimator (MLE):

$$(\hat{\lambda}, \hat{\boldsymbol{a}}) \;:=\; \underset{\mu, \|\boldsymbol{u}\|=1}{\arg\min} \left\|\mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2$$

Best rank-1 approx. & spectral tensor norm

Well-known equivalence:

$$\|\mathsf{X}\| := \max_{\|\boldsymbol{u}\|=1} \left| \sum_{ijk} x_{ijk}\, u_i u_j u_k \right| \quad \Leftrightarrow \quad \min_{\mu,\, \|\boldsymbol{u}\|=1} \sum_{ijk} \overbrace{\left\| \mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3} \right\|_{\mathsf{F}}^2}^{(x_{ijk} - \mu\, u_i u_j u_k)^2}$$

Best rank-1 approx. & spectral tensor norm

Well-known equivalence:

$$\|\mathsf{X}\| := \max_{\|\boldsymbol{u}\|=1} \left| \sum_{ijk} x_{ijk}\, u_i u_j u_k \right| \quad \Leftrightarrow \quad \min_{\mu,\, \|\boldsymbol{u}\|=1} \sum_{ijk} \overbrace{\left(x_{ijk} - \mu\, u_i u_j u_k\right)^2}^{\left\|\mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2}$$

Proof:

$$F(\mu, u) = \left\|\mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2 = \|\mathsf{X}\|_{\mathsf{F}}^2 - 2\mu \underbrace{\left\langle \mathsf{X}, \boldsymbol{u}^{\otimes 3} \right\rangle}_{= (\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}} + \mu^2 \underbrace{\left\|\boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2}_{= 1}$$

$$= \|\mathsf{X}\|_{\mathsf{F}}^2 + \mu\left(\mu - 2\,(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}\right)$$

Setting $\frac{\partial F}{\partial \mu}(\mu, u) = 0$ gives $\mu = (\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}$, leading to:

$$\min_{\|\boldsymbol{u}\|=1} -\left((\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}\right)^2 \quad \Leftrightarrow \quad \max_{\|\boldsymbol{u}\|=1} \left|(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}\right|$$

■

Best rank-1 approx. & spectral tensor norm

Well-known equivalence:

$$\|\mathsf{X}\| := \max_{\|\boldsymbol{u}\|=1} \left| \sum_{ijk} x_{ijk}\, u_i u_j u_k \right| \quad \Leftrightarrow \quad \min_{\mu,\, \|\boldsymbol{u}\|=1} \overbrace{\sum_{ijk} (x_{ijk} - \mu\, u_i u_j u_k)^2}^{\left\|\mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2}$$

Proof:

$$F(\mu, u) = \left\|\mathsf{X} - \mu\, \boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2 = \|\mathsf{X}\|_{\mathsf{F}}^2 - 2\mu \underbrace{\left\langle \mathsf{X}, \boldsymbol{u}^{\otimes 3} \right\rangle}_{= (\boldsymbol{u},\boldsymbol{u},\boldsymbol{u}) \cdot \mathsf{X}} + \mu^2 \underbrace{\left\|\boldsymbol{u}^{\otimes 3}\right\|_{\mathsf{F}}^2}_{= 1}$$

$$= \|\mathsf{X}\|_{\mathsf{F}}^2 + \mu\, (\mu - 2\, (\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X})$$

Setting $\frac{\partial F}{\partial \mu}(\mu, u) = 0$ gives $\mu = (\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}$, leading to:

$$\min_{\|\boldsymbol{u}\|=1} -((\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X})^2 \quad \Leftrightarrow \quad \max_{\|\boldsymbol{u}\|=1} \left|(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}\right|$$

$$(-\boldsymbol{u}, -\boldsymbol{u}, -\boldsymbol{u}) \cdot \mathsf{X} = -(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}$$

Moreover:

$$\max_{\|\boldsymbol{u}\|=1} \left|(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}\right| \quad \Leftrightarrow \quad \max_{\|\boldsymbol{u}\|=1} (\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) \cdot \mathsf{X}$$

■

Spectral norm & tensor eigenpairs

For convenience, denote $\mathsf{X} \cdot \boldsymbol{a}^3 := (\boldsymbol{a}, \boldsymbol{a}, \boldsymbol{a}) \cdot \mathsf{X}$ and $\mathsf{X} \cdot \boldsymbol{a}^2 := (\boldsymbol{a}, \boldsymbol{a}, \cdot) \cdot \mathsf{X}$.

MLE problem

$$\max_{\|\boldsymbol{u}\|=1} \mathsf{X} \cdot \boldsymbol{u}^3$$

Lagrangian

$$L(\mu, \boldsymbol{u}) = \frac{1}{3} \mathsf{X} \cdot \boldsymbol{u}^3 - \frac{\mu}{2} \left(\|\boldsymbol{u}\|^2 - 1 \right)$$

Critical points satisfy: $\dfrac{\partial}{\partial \boldsymbol{u}} L(\mu, \boldsymbol{u}) = \mathsf{X} \cdot \boldsymbol{u}^2 - \mu \, \boldsymbol{u} = 0, \qquad \|\boldsymbol{u}\| = 1$

Spectral norm & tensor eigenpairs

For convenience, denote $\mathsf{X} \cdot \boldsymbol{a}^3 := (\boldsymbol{a}, \boldsymbol{a}, \boldsymbol{a}) \cdot \mathsf{X}$ and $\mathsf{X} \cdot \boldsymbol{a}^2 := (\boldsymbol{a}, \boldsymbol{a}, \cdot) \cdot \mathsf{X}$.

MLE problem

$$\max_{\|\boldsymbol{u}\|=1} \mathsf{X} \cdot \boldsymbol{u}^3$$

Lagrangian

$$L(\mu, \boldsymbol{u}) = \frac{1}{3} \mathsf{X} \cdot \boldsymbol{u}^3 - \frac{\mu}{2} \left(\|\boldsymbol{u}\|^2 - 1 \right)$$

Critical points satisfy: $\dfrac{\partial}{\partial \boldsymbol{u}} L(\mu, \boldsymbol{u}) = \mathsf{X} \cdot \boldsymbol{u}^2 - \mu \, \boldsymbol{u} = 0, \qquad \|\boldsymbol{u}\| = 1$

Def: Tensor ℓ_2-eigenvalue equations[1] : $\mathsf{X} \cdot \boldsymbol{u}^2 = \mu \, \boldsymbol{u}, \qquad \|\boldsymbol{u}\| = 1$

In particular, the MLE $\hat{\boldsymbol{a}}$ (any global maximizer) verifies $\mathsf{X} \cdot \hat{\boldsymbol{a}}^2 = \mu_{\max} \, \hat{\boldsymbol{a}}$.

But how many critical points should one expect to exist?

1: Lim, 2005

Spectral norm & tensor eigenpairs

For convenience, denote $\mathsf{X} \cdot \boldsymbol{a}^3 := (\boldsymbol{a}, \boldsymbol{a}, \boldsymbol{a}) \cdot \mathsf{X}$ and $\mathsf{X} \cdot \boldsymbol{a}^2 := (\boldsymbol{a}, \boldsymbol{a}, \cdot) \cdot \mathsf{X}$.

MLE problem

$$\max_{\|\boldsymbol{u}\|=1} \mathsf{X} \cdot \boldsymbol{u}^3$$

Lagrangian

$$L(\mu, \boldsymbol{u}) = \frac{1}{3} \mathsf{X} \cdot \boldsymbol{u}^3 - \frac{\mu}{2} \left(\|\boldsymbol{u}\|^2 - 1 \right)$$

Critical points satisfy: $\dfrac{\partial}{\partial \boldsymbol{u}} L(\mu, \boldsymbol{u}) = \mathsf{X} \cdot \boldsymbol{u}^2 - \mu \, \boldsymbol{u} = 0, \qquad \|\boldsymbol{u}\| = 1$

Def: Tensor ℓ_2-eigenvalue equations[1] : $\mathsf{X} \cdot \boldsymbol{u}^2 = \mu \, \boldsymbol{u}, \qquad \|\boldsymbol{u}\| = 1$

In particular, the MLE $\hat{\boldsymbol{a}}$ (any global maximizer) verifies $\mathsf{X} \cdot \hat{\boldsymbol{a}}^2 = \mu_{\max} \, \hat{\boldsymbol{a}}$.

But how many critical points should one expect to exist?

Thm (Cartwright & Sturmfels, 2013): A symmetric tensor of dims N has up to $((d-1)^N - 1)/(d-2)$ distinct eigenvalues (the bound is generically attained).

1: Lim, 2005

Algorithm: tensor power iteration

The tensor eigenvalue eqns naturally suggest a fixed point iteration:

$$\hat{\boldsymbol{u}} \leftarrow (\mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1})/\hat{\mu}, \qquad \text{with} \quad \hat{\mu} = \left\| \mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1} \right\|$$

This is known as tensor power iteration.[1,2]

1: De Lathauwer & al., 2000b, 2: Kofidis & Regaila, 2002, 3: Robeva, 2016, 4: Kolda, 2001

Algorithm: tensor power iteration

The tensor eigenvalue eqns naturally suggest a fixed point iteration:

$$\hat{\boldsymbol{u}} \leftarrow (\mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1})/\hat{\mu}, \qquad \text{with} \quad \hat{\mu} = \left\| \mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1} \right\|$$

This is known as tensor power iteration.[1,2]

It succeeds (with proper initialization) in finding all rank-1 terms of an orthogonal decomposition (odeco[3,4]) $\mathsf{X} = \sum_{r=1}^{R} \lambda_r \boldsymbol{v}_r^{\otimes d}$, where $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}$:

Thm (Anandkumar & al., 2014): Let $\mathsf{X} = \sum_{r=1}^{R} \lambda_r \, \boldsymbol{v}_r^{\otimes d}$ be an odeco of X. Then, with random initialization, power iteration converges to one $\boldsymbol{v}_r$ a.s.

Furthermore, $\{\boldsymbol{v}_r\}$ are the only robust eigenvectors of X: those having a neighborhood where power iteration converges to them.

But what about the non-orthogonal case?

1: De Lathauwer & al., 2000b, 2: Kofidis & Regaila, 2002, 3: Robeva, 2016, 4: Kolda, 2001

Shifted tensor power iteration

Most symmetric tensors are not odeco (generically, they aren't).

Still, power iteration is useful for rank-1 approximation (ex: tensor PCA).

However, the basic iteration (previous slide) is not generally convergent.[1]

1: Kolda & Mayo, 2011

Shifted tensor power iteration

Most symmetric tensors are not odeco (generically, they aren't).

Still, power iteration is useful for rank-1 approximation (ex: tensor PCA).

However, the basic iteration (previous slide) is not generally convergent.[1]

Kolda & Mayo (2011) have thus proposed a shifted power iteration:

$$\hat{\boldsymbol{u}} \leftarrow (\mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1} + \alpha \hat{\boldsymbol{u}})/\hat{\mu}, \qquad \text{with} \quad \hat{\mu} = \left\| \mathsf{X} \cdot \hat{\boldsymbol{u}}^{d-1} + \alpha \hat{\boldsymbol{u}} \right\|$$

They established the following sufficient condition for convergence:

$$|\alpha| > (d-1) \max_{\|\boldsymbol{v}\|=1} \left\| \mathsf{X} \cdot \boldsymbol{v}^{d-2} \right\|.$$

This allows computing all stable eigenpairs $(\mu, \boldsymbol{u})$: those at which $\nabla^2_{\boldsymbol{u}\boldsymbol{u}} L(\mu, \boldsymbol{u})$ is definite when projected onto $\boldsymbol{u}^\perp$. (By contrast, only the dominant eigenpair can be computed in the matrix case.)

1: Kolda & Mayo, 2011

Low-rank approximation, in several flavors

Low-rank tensor approximation

We turn now to low-rank approximation of a tensor.

In the matrix case, a central result is:

Thm (Eckart–Young, 1936): The solution to

$$\min_{\operatorname{rank} \hat{\mathbf{X}} \leq R} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_{\mathsf{F}}$$

corresponds to truncating the SVD of $\mathbf{X}$ at rank R.

Low-rank tensor approximation

We turn now to low-rank approximation of a tensor.

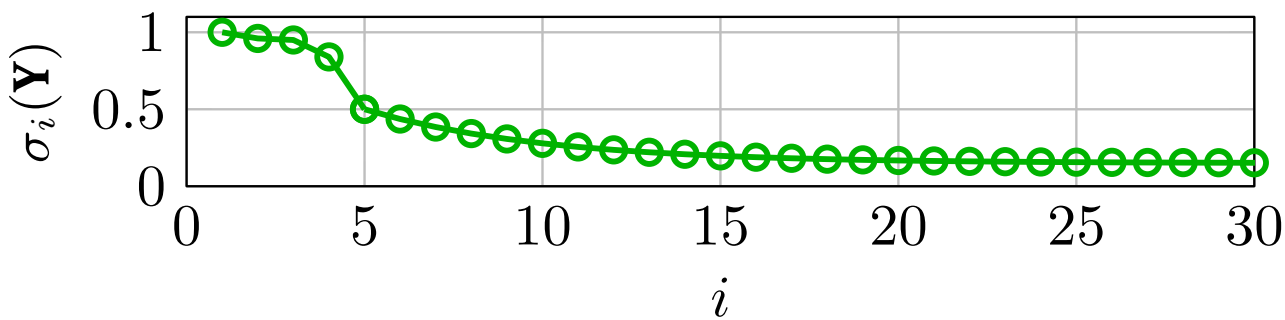
In the matrix case, a central result is:

Thm (Eckart–Young, 1936): The solution to

$$\min_{\operatorname{rank}\hat{\mathbf{X}} \leq R} \left\| \mathbf{X} - \hat{\mathbf{X}} \right\|_{\mathsf{F}}$$

corresponds to truncating the SVD of $\mathbf{X}$ at rank R.

$$\mathbf{Y} = \sum_i \sigma_i \boldsymbol{u}_i \, \boldsymbol{v}_i^T$$



Low-rank tensor approximation

We turn now to low-rank approximation of a tensor.

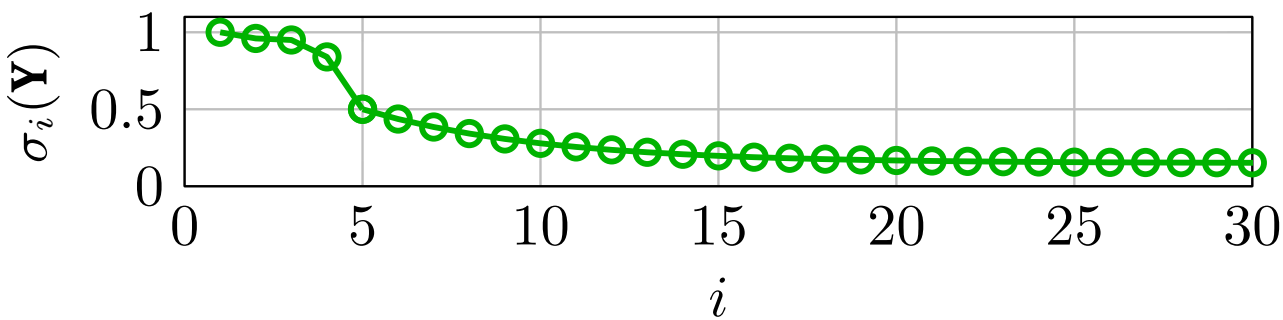
In the matrix case, a central result is:

Thm (Eckart–Young, 1936): The solution to

$$\min_{\operatorname{rank}\hat{\mathbf{X}}\leq R} \left\|\mathbf{X}-\hat{\mathbf{X}}\right\|_{\mathsf{F}}$$

corresponds to truncating the SVD of $\mathbf{X}$ at rank R.

$$\mathbf{Y} = \sum_i \sigma_i \boldsymbol{u}_i \, \boldsymbol{v}_i^T$$



No such a general result exists for tensors, for the usual notions of rank.

Moreover, for rank > 1, all the discussion on results & algorithms is conditioned upon the definition of rank which is relevant to a certain end.

Low-mrank approximation

Goal: Approximate $\mathsf{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ by $\hat{\mathsf{X}}$ such that $\operatorname{mrank} \hat{\mathsf{X}} \leq (R_1, R_2, R_3)$

Low-mrank approximation

Goal: Approximate $\mathsf{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ by $\hat{\mathsf{X}}$ such that $\operatorname{mrank} \hat{\mathsf{X}} \leq (R_1, R_2, R_3)$

Recall that $\hat{\mathsf{X}}$ must admit a Tucker decomposition

$$\hat{\mathsf{X}} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \qquad \mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3},$$

where $\mathbf{U}, \mathbf{V}, \mathbf{W}$ are semi-orthogonal w.l.o.g. Hence, a first, natural approach is:

$$\min_{\mathsf{S}, \mathbf{U}, \mathbf{V}, \mathbf{W}} \|\mathsf{X} - (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}\|_{\mathsf{F}}^2 .$$

Low-mrank approximation

Goal: Approximate $\mathsf{X} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ by $\hat{\mathsf{X}}$ such that mrank $\hat{\mathsf{X}} \leq (R_1, R_2, R_3)$

Recall that $\hat{\mathsf{X}}$ must admit a Tucker decomposition

$$\hat{\mathsf{X}} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \qquad \mathsf{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3},$$

where $\mathbf{U}, \mathbf{V}, \mathbf{W}$ are semi-orthogonal w.l.o.g. Hence, a first, natural approach is:

$$\min_{\mathsf{S},\mathbf{U},\mathbf{V},\mathbf{W}} \|\mathsf{X} - (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}\|_{\mathsf{F}}^2 .$$

Since

$$\begin{aligned} \|\mathsf{X} - (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}\|_{\mathsf{F}}^2 &= \|\mathsf{X}\|_{\mathsf{F}}^2 - 2\langle \mathsf{X}, (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S} \rangle + \|(\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}\|_{\mathsf{F}}^2 \\ &= \|\mathsf{X}\|_{\mathsf{F}}^2 - 2\left\langle (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}, \mathsf{S} \right\rangle + \|\mathsf{S}\|_{\mathsf{F}}^2 , \end{aligned}$$

it is easy to show that every mimimizer must satisfy $\mathsf{S} = (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}$, and thus the problem is equivalent to

$$\arg\min_{\mathbf{U},\mathbf{V},\mathbf{W}} \left\| \mathsf{X} - (\mathbf{U}\mathbf{U}^\mathsf{T}, \mathbf{V}\mathbf{V}^\mathsf{T}, \mathbf{W}\mathbf{W}^\mathsf{T}) \cdot \mathsf{X} \right\|_{\mathsf{F}}^2 = \arg\max_{\mathbf{U},\mathbf{V},\mathbf{W}} \left\| \overbrace{(\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}}^{\mathsf{S}} \right\|_{\mathsf{F}}^2$$

Truncated HOSVD (THOSVD)

Recall that the HOSVD of X is built by computing the SVD of each unfolding:

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \qquad \text{where } \begin{cases} \mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{\Sigma}_1\,\mathbf{Q}_1^\mathsf{T} \\ \mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{\Sigma}_2\,\mathbf{Q}_2^\mathsf{T} \\ \mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{\Sigma}_3\,\mathbf{Q}_3^\mathsf{T} \end{cases}$$

Consequence: $\|(\mathsf{S})_{\ell::}\|_\mathsf{F} \geq \|(\mathsf{S})_{\ell+1::}\|_\mathsf{F}$, and similarly, $\|(\mathsf{S})_{:\ell:}\|_\mathsf{F} \geq \|(\mathsf{S})_{:\ell+1:}\|_\mathsf{F}$ and $\|(\mathsf{S})_{::\ell}\|_\mathsf{F} \geq \|(\mathsf{S})_{::\ell+1}\|_\mathsf{F}$.

Truncated HOSVD (THOSVD)

Recall that the HOSVD of X is built by computing the SVD of each unfolding:

$$\mathsf{X} = (\mathbf{U}, \mathbf{V}, \mathbf{W}) \cdot \mathsf{S}, \qquad \text{where } \begin{cases} \mathbf{X}_{(1)} = \mathbf{U}\,\mathbf{\Sigma}_1\,\mathbf{Q}_1^\mathsf{T} \\ \mathbf{X}_{(2)} = \mathbf{V}\,\mathbf{\Sigma}_2\,\mathbf{Q}_2^\mathsf{T} \\ \mathbf{X}_{(3)} = \mathbf{W}\,\mathbf{\Sigma}_3\,\mathbf{Q}_3^\mathsf{T} \end{cases}$$

Consequence: $\|(\mathsf{S})_{\ell::}\|_\mathsf{F} \geq \|(\mathsf{S})_{\ell+1::}\|_\mathsf{F}$, and similarly, $\|(\mathsf{S})_{:\ell:}\|_\mathsf{F} \geq \|(\mathsf{S})_{:\ell+1:}\|_\mathsf{F}$ and $\|(\mathsf{S})_{::\ell}\|_\mathsf{F} \geq \|(\mathsf{S})_{::\ell+1}\|_\mathsf{F}$.

This suggests truncating[1] $\mathbf{U}$ at R_1 cols, $\mathbf{V}$ at R_2 cols & $\mathbf{W}$ at R_3 cols:

$$\mathsf{X} \approx \hat{\mathsf{X}} = (\mathbf{U}_{:,1:R_1}, \mathbf{V}_{:,1:R_2}, \mathbf{W}_{:,1:R_3}) \cdot \mathsf{S}_{1:R_1,1:R_2,1:R_3}$$

(Special case: $R_1 = R_2 = R_3 = 1$, we get $\mathsf{X} \approx \sigma\, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w}$ built from the dominant left singular vectors of each unfolding.)

But is it optimal?

1: De Lathauwer & al., 2000a

Quasi-optimality of truncated HOSVD

Unfortunately, no! Nevertheless, it is said to be quasi-optimal in the sense:

Thm (Vannieuwenhoven & al., 2012): Let $\mathsf{X}^\star$ be a best mrank-$(R_1, \ldots, R_d)$ approx. of X, and $\hat{\mathsf{X}}$ be obtained by truncating its HOSVD at $(R_1, \ldots, R_d)$. Then,

$$\left\| \mathsf{X} - \hat{\mathsf{X}} \right\|_{\mathsf{F}}^2 \leq d \, \| \mathsf{X} - \mathsf{X}^\star \|_{\mathsf{F}}^2 .$$

Quasi-optimality of truncated HOSVD

Unfortunately, no! Nevertheless, it is said to be quasi-optimal in the sense:

Thm (Vannieuwenhoven & al., 2012): Let $\mathsf{X}^\star$ be a best mrank-$(R_1, \ldots, R_d)$ approx. of X, and $\hat{\mathsf{X}}$ be obtained by truncating its HOSVD at $(R_1, \ldots, R_d)$. Then,

$$\left\| \mathsf{X} - \hat{\mathsf{X}} \right\|_{\mathsf{F}}^2 \leq d \, \| \mathsf{X} - \mathsf{X}^\star \|_{\mathsf{F}}^2 .$$

Proof ($d = 3$): Define the orthogonal projector $\mathbf{P}_U := \mathbf{U}_{:,1:R_1} \mathbf{U}_{:,1:R_1}^\mathsf{T}$ onto the dominant subspace of dim R_1 of $\mathbf{X}_{(1)}$, and similarly for $\mathbf{P}_V$ and $\mathbf{P}_W$. Write

$$\begin{aligned}
\mathsf{E} := \mathsf{X} - \hat{\mathsf{X}} &= \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \mathbf{P}_W) \cdot \mathsf{X} = \mathsf{X} + (\mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} \\
&\quad + (\mathbf{P}_U, \mathbf{P}_V, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \mathbf{P}_W) \cdot \mathsf{X} \\
&= (\mathbf{I} - \mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} + (\mathbf{P}_U, \mathbf{I} - \mathbf{P}_V, \cdot) \cdot \mathsf{X} + (\mathbf{P}_U, \mathbf{P}_V, \mathbf{I} - \mathbf{P}_W) \cdot \mathsf{X}
\end{aligned}$$

Quasi-optimality of truncated HOSVD

Unfortunately, no! Nevertheless, it is said to be quasi-optimal in the sense:

Thm (Vannieuwenhoven & al., 2012): Let $\mathsf{X}^\star$ be a best mrank-$(R_1, \ldots, R_d)$ approx. of X, and $\hat{\mathsf{X}}$ be obtained by truncating its HOSVD at $(R_1, \ldots, R_d)$. Then,

$$\left\|\mathsf{X} - \hat{\mathsf{X}}\right\|_{\mathsf{F}}^2 \leq d \left\|\mathsf{X} - \mathsf{X}^\star\right\|_{\mathsf{F}}^2.$$

Proof ($d = 3$): Define the orthogonal projector $\mathbf{P}_U := \mathbf{U}_{:,1:R_1}\mathbf{U}_{:,1:R_1}^\mathsf{T}$ onto the dominant subspace of dim R_1 of $\mathbf{X}_{(1)}$, and similarly for $\mathbf{P}_V$ and $\mathbf{P}_W$. Write

$$\begin{aligned}
\mathsf{E} := \mathsf{X} - \hat{\mathsf{X}} &= \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \mathbf{P}_W) \cdot \mathsf{X} = \mathsf{X} + (\mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} \\
&\quad + (\mathbf{P}_U, \mathbf{P}_V, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \cdot) \cdot \mathsf{X} - (\mathbf{P}_U, \mathbf{P}_V, \mathbf{P}_W) \cdot \mathsf{X} \\
&= (\mathbf{I} - \mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X} + (\mathbf{P}_U, \mathbf{I} - \mathbf{P}_V, \cdot) \cdot \mathsf{X} + (\mathbf{P}_U, \mathbf{P}_V, \mathbf{I} - \mathbf{P}_W) \cdot \mathsf{X}
\end{aligned}$$

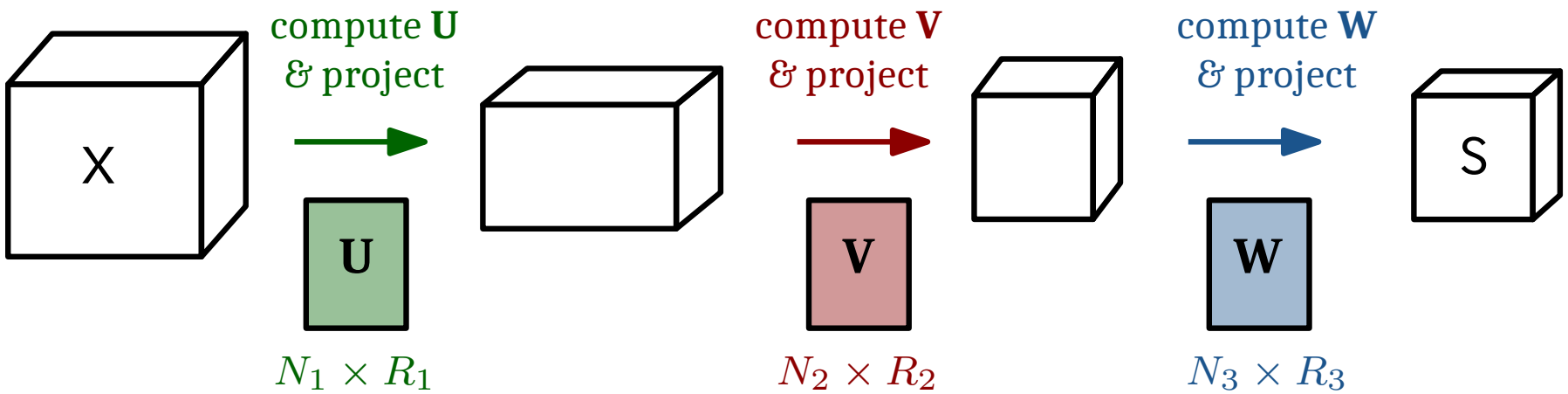
By orthogonality, non-expansiveness and the Eckart–Young thm,

$$\begin{aligned}
\|\mathsf{E}\|_{\mathsf{F}}^2 &\leq \|(\mathbf{I} - \mathbf{P}_U, \cdot, \cdot) \cdot \mathsf{X}\|_{\mathsf{F}}^2 + \|(\cdot, \mathbf{I} - \mathbf{P}_V, \cdot) \cdot \mathsf{X}\|_{\mathsf{F}}^2 + \|(\cdot, \cdot, \mathbf{I} - \mathbf{P}_W) \cdot \mathsf{X}\|_{\mathsf{F}}^2 \\
&\leq \sum_{i=1}^3 \left\|\mathbf{X}_{(i)} - \mathbf{X}_{(i)}^\star\right\|_{\mathsf{F}}^2 = 3 \left\|\mathsf{X} - \mathsf{X}^\star\right\|_{\mathsf{F}})^2
\end{aligned}$$

■

Variants of THOSVD

Variants of THOSVD exist,[1,2,3] for instance by sequentially performing (optimal) modal projections:



They have lower complexity, often display superior empirical performance & are provably never worse than THOSVD in special cases.

But in general, they're all subject to the same quasi-optimality bound.

1: Vannieuwenhoven & al., 2012, 2: da Silva & al., 2016, 3: Goulart & Comon, 2017

High-order orthogonal iteration (HOOI)

The previous algebraic (non-iterative) soln's can be refined by means of an iterative algorithm.

A popular one is HOOI,[1] which amounts to an alternating opt scheme for

$$\max_{\mathbf{U},\mathbf{V},\mathbf{W}} \left\| (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X} \right\|_\mathsf{F}^2, \quad \text{subj. to } \begin{cases} \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_{R_1} \\ \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_{R_2} \\ \mathbf{W}^\mathsf{T}\mathbf{W} = \mathbf{I}_{R_3} \end{cases}$$

1: De Lathauwer & al., 2000b, 2: Xu, 2018

High-order orthogonal iteration (HOOI)

The previous algebraic (non-iterative) soln's can be refined by means of an iterative algorithm.

A popular one is HOOI,[1] which amounts to an alternating opt scheme for

$$\max_{\mathbf{U},\mathbf{V},\mathbf{W}} \left\| (\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X} \right\|_\mathsf{F}^2, \quad \text{subj. to } \begin{cases} \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}_{R_1} \\ \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_{R_2} \\ \mathbf{W}^\mathsf{T}\mathbf{W} = \mathbf{I}_{R_3} \end{cases}$$

For instance, if $\mathbf{V}, \mathbf{W}$ are held fixed, then $\mathbf{U}$ easily follows from a (truncated) SVD. One thus performs at each iteration:

(i) Recompute $\mathbf{U}$ from the SVD of $\left((\cdot, \mathbf{V}^\mathsf{T}, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}\right)_{(1)}$

(ii) Recompute $\mathbf{V}$ from the SVD of $\left((\mathbf{U}^\mathsf{T}, \cdot, \mathbf{W}^\mathsf{T}) \cdot \mathsf{X}\right)_{(2)}$

(iii) Recompute $\mathbf{W}$ from the SVD of $\left((\mathbf{U}^\mathsf{T}, \mathbf{V}^\mathsf{T}, \cdot) \cdot \mathsf{X}\right)_{(3)}$

Converges to a local sol'n whose unfoldings above have distinct dominating singular values).[2]

1: De Lathauwer & al., 2000b, 2: Xu, 2018

Best rank-R approximation

Consider now the best rank-R approximation of X:

$$\inf_{\operatorname{rank} \hat{\mathsf{X}} \leq R} \left\| \mathsf{X} - \hat{\mathsf{X}} \right\|_{\mathsf{F}} = \inf_{\mathbf{A},\mathbf{B},\mathbf{C}} \| \mathsf{X} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R \|_{\mathsf{F}}$$

Can we replace inf by min? In general, no: a minimizer might not exist if $R > 1$.

Best rank-R approximation

Consider now the best rank-R approximation of X:

$$\inf_{\operatorname{rank}\hat{\mathsf{X}}\leq R}\left\|\mathsf{X}-\hat{\mathsf{X}}\right\|_{\mathsf{F}}=\inf_{\mathbf{A},\mathbf{B},\mathbf{C}}\|\mathsf{X}-[\![\mathbf{A},\mathbf{B},\mathbf{C}]\!]_R\|_{\mathsf{F}}$$

Can we replace inf by min? In general, no: a minimizer might not exist if $R>1$.

Example: Previosly, we have shown that

$$\mathsf{X}=\left(\begin{array}{cc|cc}1&0&0&1\\0&1&0&0\end{array}\right)=\boldsymbol{e}_1\otimes\boldsymbol{e}_1\otimes\boldsymbol{e}_1+\boldsymbol{e}_2\otimes\boldsymbol{e}_2\otimes\boldsymbol{e}_1+\boldsymbol{e}_1\otimes\boldsymbol{e}_2\otimes\boldsymbol{e}_2.$$

has rank 3. More generally, if $\mathbf{A}\in\mathbb{R}^{N_1\times 2}$, $\mathbf{B}\in\mathbb{R}^{N_2\times 2}$ and $\mathbf{C}\in\mathbb{R}^{N_3\times 2}$ all have rank 2, then

$$\mathsf{Y}=(\mathbf{A},\mathbf{B},\mathbf{C})\cdot\mathsf{X}=\boldsymbol{a}_1\otimes\boldsymbol{b}_1\otimes\boldsymbol{c}_1+\boldsymbol{a}_2\otimes\boldsymbol{b}_2\otimes\boldsymbol{c}_1+\boldsymbol{a}_1\otimes\boldsymbol{b}_2\otimes\boldsymbol{c}_2$$

has rank 3, but does not admit a best rank-2 approximation:

$$\begin{aligned}\hat{\mathsf{Y}}_m&:=m(\boldsymbol{a}_1+m^{-1}\boldsymbol{a}_2)\otimes(\boldsymbol{b}_2+m^{-1}\boldsymbol{b}_1)\otimes(\boldsymbol{c}_1+m^{-1}\boldsymbol{c}_2)-m\boldsymbol{a}_1\otimes\boldsymbol{b}_2\otimes\boldsymbol{c}_1\\&=\mathsf{Y}+\mathsf{O}(1/m)\ \to\ \mathsf{Y}\end{aligned}$$

□

Ill-posedness of best rank-R approximation

A celebrated paper by de Silva & Lim (2008) discusses this issue in depth.

It argues that there also exist examples of any orders $d \geq 3$ for R in $\{2, \ldots, \min_d N_d\}$.

Ill-posedness of best rank-R approximation

A celebrated paper by de Silva & Lim (2008) discusses this issue in depth.

It argues that there also exist examples of any orders $d \geq 3$ for R in $\{2, \ldots, \min_d N_d\}$.

Furthermore, by a complete classification of $2 \times 2 \times 2$ tensors by orbit type, they are able to show that in $\mathbb{R}^{N_1 \times N_2 \times N_3}$, no element from the nonempty open set

$$\left\{ (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \left(\begin{array}{cc|cc} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{array} \right) \; : \; \operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{B} = \operatorname{rank} \mathbf{C} = 2 \right\}$$

admits a best rank-2 approximation.

Ill-posedness of best rank-R approximation

A celebrated paper by de Silva & Lim (2008) discusses this issue in depth.

It argues that there also exist examples of any orders $d \geq 3$ for R in $\{2, \ldots, \min_d N_d\}$.

Furthermore, by a complete classification of $2 \times 2 \times 2$ tensors by orbit type, they are able to show that in $\mathbb{R}^{N_1 \times N_2 \times N_3}$, no element from the nonempty open set

$$\left\{ (\mathbf{A}, \mathbf{B}, \mathbf{C}) \cdot \left(\begin{array}{cc|cc} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \end{array} \right) \,:\, \operatorname{rank} \mathbf{A} = \operatorname{rank} \mathbf{B} = \operatorname{rank} \mathbf{C} = 2 \right\}$$

admits a best rank-2 approximation.

By contrast, a best rank-R approximation generically exists[1] over $\mathbb{C}$.

1: Qi & al., 2020

What if we try, anyway?

One might want to dismiss this issue by focusing on a "reasonable rank-R approximation" instead.

1: de Silva & Lim, 2008, 2: Paatero, 2000

What if we try, anyway?

One might want to dismiss this issue by focusing on a "reasonable rank-R approximation" instead.

Yet, if a solution does not exist, then at least some of the terms in the decomposition must diverge in norm,[1] so the "approximate solution" is typically useless.

1: de Silva & Lim, 2008, 2: Paatero, 2000

What if we try, anyway?

One might want to dismiss this issue by focusing on a "reasonable rank-R approximation" instead.

Yet, if a solution does not exist, then at least some of the terms in the decomposition must diverge in norm,[1] so the "approximate solution" is typically useless.

Also, failing to account for it can bring serious difficulties even when a best approximation with the sought rank exists:

> "A well-posed problem in the neighborhood of an ill-posed one is ill-conditioned."

For instance, a numerical algorithm can traverse regions close to tensors who do not admit a best approximation.[2]

1: de Silva & Lim, 2008, 2: Paatero, 2000

CPD condition number

Goal: What is the relation between the rank-1 terms of

$$\mathsf{X} = \sum_{r=1}^{R} \mathsf{X}_r = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$

and those obtained by rank-R approximation $\hat{\mathsf{X}}$ of $\mathsf{Y} \approx \mathsf{X}$ (e.g., $\mathsf{Y} = \mathsf{X} + \mathsf{N}$)?

CPD condition number

Goal: What is the relation between the rank-1 terms of

$$\mathsf{X} = \sum_{r=1}^{R} \mathsf{X}_r = \sum_{r=1}^{R} \boldsymbol{a}_r \otimes \boldsymbol{b}_r \otimes \boldsymbol{c}_r$$

and those obtained by rank-R approximation $\hat{\mathsf{X}}$ of $\mathsf{Y} \approx \mathsf{X}$ (e.g., $\mathsf{Y} = \mathsf{X} + \mathsf{N}$)?

Breiding & Vannieuwenhoven (2018) studied the condition number $\kappa(\mathscr{X})$ of the local inverse of

$$\Phi : \mathcal{M} \times \cdots \times \mathcal{M} \to \mathbb{R}^{N_1 \times N_2 \times N_3}$$
$$\mathscr{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_R) \mapsto \textstyle\sum_{r=1}^{R} \mathsf{X}_r$$

$\mathcal{M}$ = manifold of rank-1 tensors

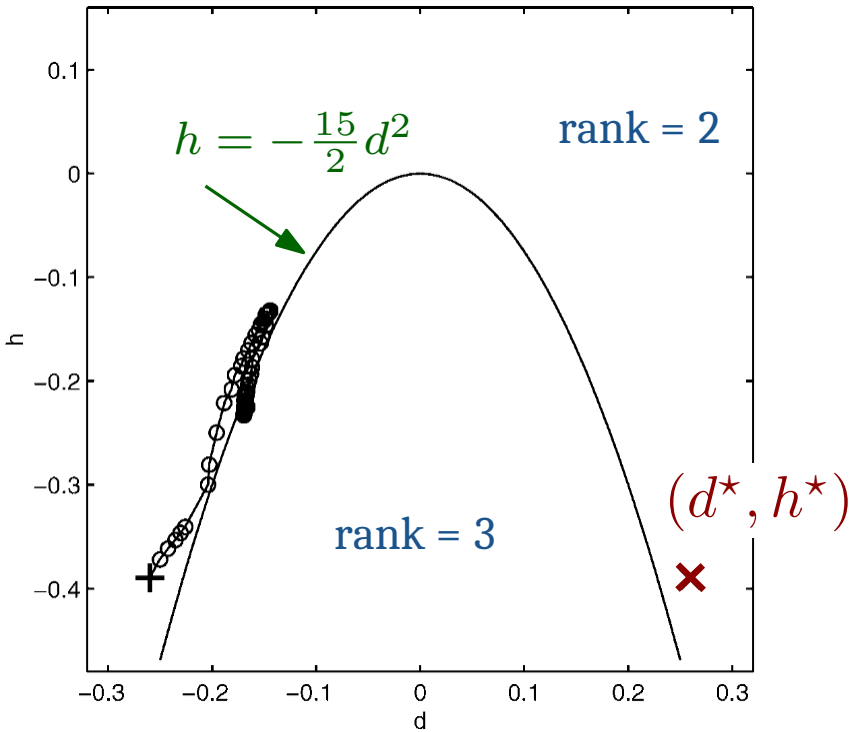
in a neighborhood of $\mathscr{X}$, denoted $\Phi^{-1}_{\mathscr{X}}$, which allows estimating

$$\underbrace{\left\| \mathscr{X} - \Phi^{-1}_{\mathscr{X}}(\hat{\mathsf{X}}) \right\|}_{\text{error over parameters}} \lesssim \kappa(\mathscr{X}) \underbrace{\left\| \Phi(\mathscr{X}) - \hat{\mathsf{X}} \right\|}_{\text{reconstruction error}},$$

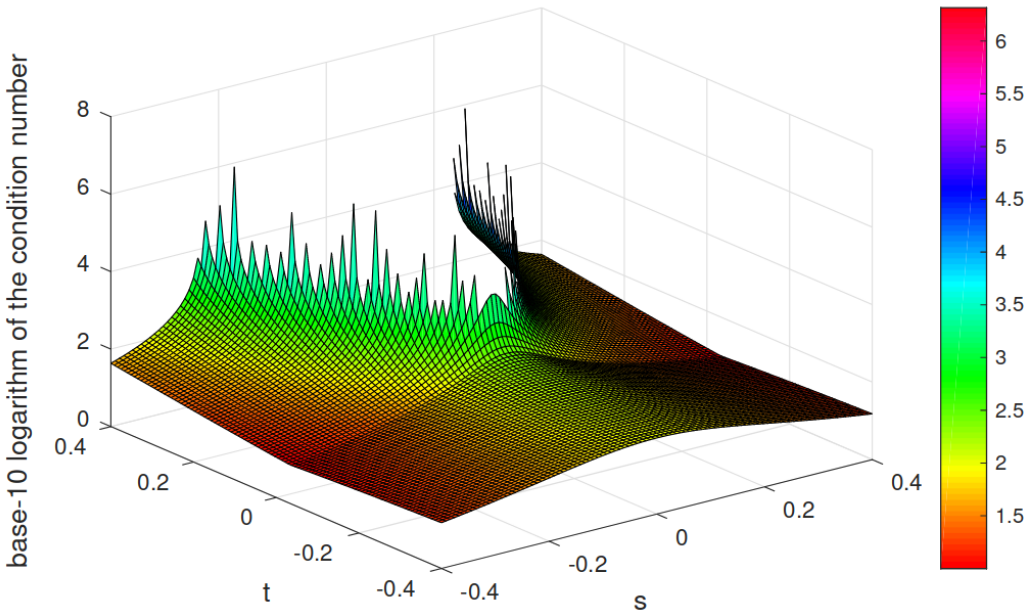
for $\hat{\mathsf{X}}$ a rank-R tensor in the neighborhood of $\mathsf{X} = \Phi(\mathscr{X})$.

Paatero's example

They also showed that $\kappa(\mathscr{X})$ must diverge as $\Phi(\mathscr{X})$ tends to a tensor of rank higher than R.



Paatero's example:[1] *trajectory of ALS algorithm trying to compute a rank-2 PD of* $\mathsf{X} = \left(\begin{array}{cc|cc} 0 & 1 & 30 & 0 \\ 1 & d^\star & 0 & h^\star \end{array}\right)$.



Corresponding behavior of the condition number (in log scale).

1: Paatero, 2000, 2: Breiding & Vannieuwenhoven, 2018

A positive result

In applications, by assumption $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R + \mathsf{W}$, even if X does not admit a best rank-R approximation.

1: Evert & De Lathauwer, 2022

A positive result

In applications, by assumption $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R + \mathsf{W}$, even if X does not admit a best rank-R approximation.

For "sufficiently small" W, the problem is well posed:[1]

If $R \leq \min\{N_1, N_2\}$, then there exists $\epsilon > 0$ such that X admits a best rank-R approximation for all W satisfying $\|\mathsf{W}\|_{\mathsf{F}} \leq \epsilon$.

Roughly, ϵ depends on the conditioning of "partial" CPDs of $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$.

1: Evert & De Lathauwer, 2022

A positive result

In applications, by assumption $\mathsf{X} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_R + \mathsf{W}$, even if X does not admit a best rank-R approximation.

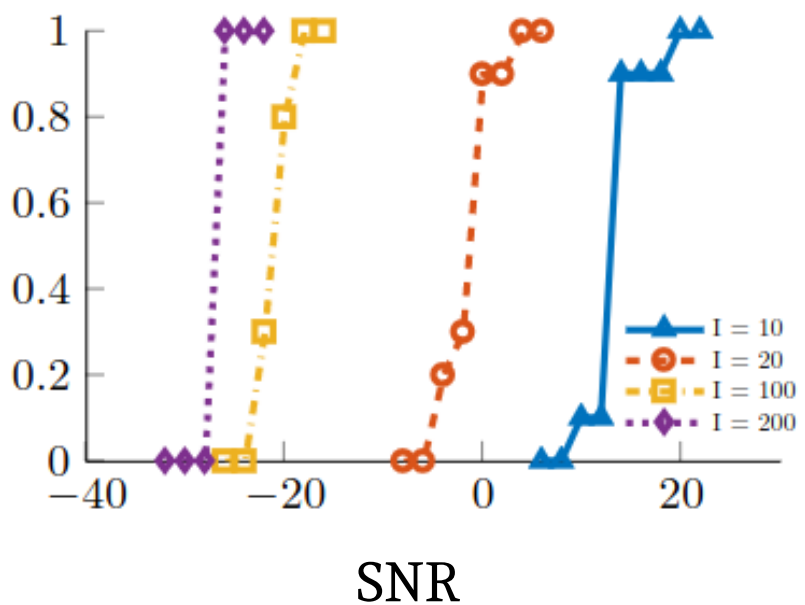
For "sufficiently small" W, the problem is well posed:[1]

If $R \leq \min\{N_1, N_2\}$, then there exists $\epsilon > 0$ such that X admits a best rank-R approximation for all W satisfying $\|\mathsf{W}\|_{\mathsf{F}} \leq \epsilon$.

Roughly, ϵ depends on the conditioning of "partial" CPDs of $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$.

Another, similar result by the same authors allows asserting that a given tensor has a best rank-R approx.

Example:[1] proportion of random tensors $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]_4 + \mathsf{W} \in \mathbb{R}^{I \times I \times I}$ guaranteed to have a best rank-4 approx., as a function of the SNR



1: Evert & De Lathauwer, 2022

Approximate CPD computation

Numerous algorithms exist for computing an approximate CPD, including

- alternating optimization (block coordinate descent)
- algebraic methods
- classical optimization schemes
- stochastic gradient
- distributed schemes
- ...

Several bibliographical pointers are given in the handout.

Best BTD approximation

Recall the BTD model

$$\begin{aligned}\mathcal{X} &\approx \sum_{r=1}^{R} \mathbf{H}_r \otimes \boldsymbol{c}_r, \quad \operatorname{rank} \mathbf{H}_r \leq L_r, \\ &= \sum_{r=1}^{R} \left(\mathbf{A}_r \, \mathbf{B}_r^{\mathsf{T}}\right) \otimes \boldsymbol{c}_r,\end{aligned}$$

where $\mathbf{A}_r \in \mathbb{R}^{N_1 \times L_r}$, $\mathbf{B}_r \in \mathbb{R}^{N_2 \times L_r}$ and $\boldsymbol{c}_r \in \mathbb{R}^{N_3}$.

A first, natural formulation would thus be

$$\inf_{\{\mathbf{A}_r, \mathbf{B}_r, \boldsymbol{c}_r\}_{r=1}^{R}} \left\| \mathcal{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r \, \mathbf{B}_r^{\mathsf{T}}\right) \otimes \boldsymbol{c}_r \right\|_{\mathsf{F}}^2,$$

with a fixed (given) choice of R and $L_1, \ldots, L_R$.

Does a minimizer always exist? Not always...

Sets of BTDs

To study BTD approximations, we need sets analogous to $\{\mathsf{X} : \operatorname{rank} \mathsf{X} \leq R\}$.

For $L_1 \geq L_2 \geq \cdots \geq L_{N_3} \geq 0$, we define:

$$\mathcal{B}_{L_1,\ldots,L_{N_3}} \triangleq \left\{ \sum_{r=1}^{N_3} \mathbf{H}_r \otimes \boldsymbol{w}_r \;\middle|\; \operatorname{rank} \mathbf{H}_r \leq L_r \text{ and } \boldsymbol{w}_1, \ldots, \boldsymbol{w}_R \text{ are l. i.} \right\}$$

If $L_{R+1} = \cdots = L_{N_3} = 0$, we can simply write $\mathcal{B}_{L_1,\ldots,L_R} = \mathcal{B}_{L_1,\ldots,L_{N_3}}$

1: Goulart & Comon, 2019.

Sets of BTDs

To study BTD approximations, we need sets analogous to $\{\mathsf{X} : \operatorname{rank}\mathsf{X} \leq R\}$.

For $L_1 \geq L_2 \geq \cdots \geq L_{N_3} \geq 0$, we define:

$$\mathcal{B}_{L_1,\ldots,L_{N_3}} \triangleq \left\{ \sum_{r=1}^{N_3} \mathbf{H}_r \otimes \boldsymbol{w}_r \;\middle|\; \operatorname{rank}\mathbf{H}_r \leq L_r \text{ and } \boldsymbol{w}_1,\ldots,\boldsymbol{w}_R \text{ are l. i.} \right\}$$

If $L_{R+1} = \cdots = L_{N_3} = 0$, we can simply write $\mathcal{B}_{L_1,\ldots,L_R} = \mathcal{B}_{L_1,\ldots,L_{N_3}}$

Example: If

$$\mathsf{X} = \left(\sum_{i=1}^{3} \boldsymbol{u}_i \otimes \boldsymbol{v}_i \right) \otimes \boldsymbol{w}_1 + \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{w}_2,$$

then $\mathsf{X} \in \mathcal{B}_{3,1} \subset \mathcal{B}_{3,2} \subset \mathcal{B}_{3,2,1} \subset \mathcal{B}_{4,4,2}$ □

1: Goulart & Comon, 2019.

Sets of BTDs

To study BTD approximations, we need sets analogous to $\{\mathsf{X} : \operatorname{rank}\mathsf{X} \leq R\}$.

For $L_1 \geq L_2 \geq \cdots \geq L_{N_3} \geq 0$, we define:

$$\mathcal{B}_{L_1,\ldots,L_{N_3}} \triangleq \left\{ \sum_{r=1}^{N_3} \mathbf{H}_r \otimes \boldsymbol{w}_r \;\middle|\; \operatorname{rank}\mathbf{H}_r \leq L_r \text{ and } \boldsymbol{w}_1,\ldots,\boldsymbol{w}_R \text{ are l. i.} \right\}$$

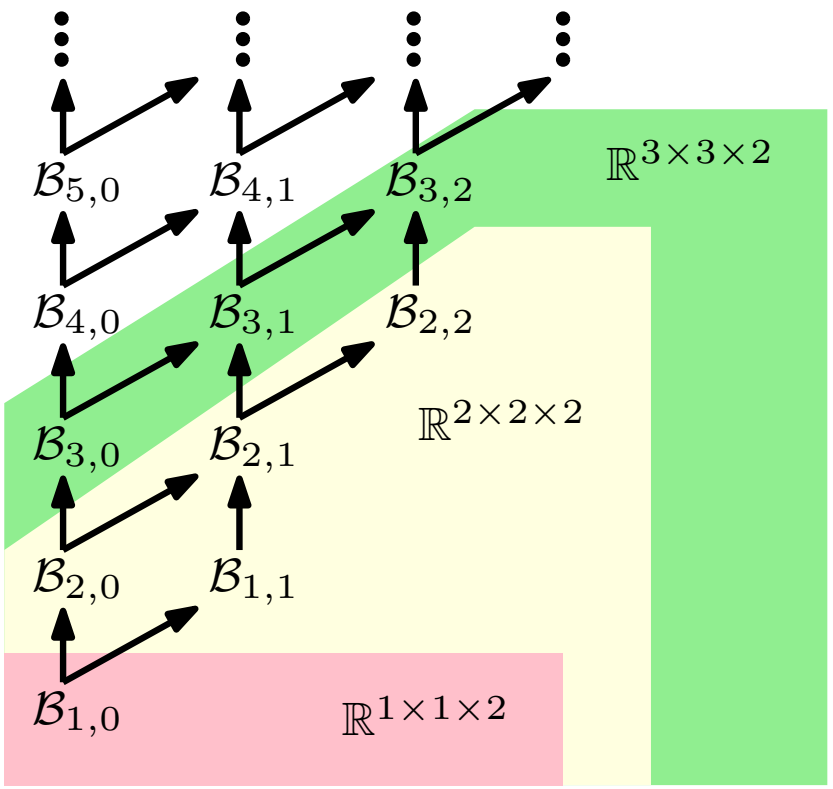
If $L_{R+1} = \cdots = L_{N_3} = 0$, we can simply write $\mathcal{B}_{L_1,\ldots,L_R} = \mathcal{B}_{L_1,\ldots,L_{N_3}}$

Example: If

$$\mathsf{X} = \left(\sum_{i=1}^{3} \boldsymbol{u}_i \otimes \boldsymbol{v}_i \right) \otimes \boldsymbol{w}_1 + \boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{w}_2,$$

then $\mathsf{X} \in \mathcal{B}_{3,1} \subset \mathcal{B}_{3,2} \subset \mathcal{B}_{3,2,1} \subset \mathcal{B}_{4,4,2}$ □

This induces a richer hierarchy than the one induced by the tensor rank.



Hierarchy of BTD sets

1: Goulart & Comon, 2019.

Minimal ranks and counter-example

Def: Minimal (BTD) ranks[1] of X

$$\rho(\mathsf{X}) = (L_1, \ldots, L_{N_3}) \quad \text{when} \quad \mathsf{X} \in \mathcal{B}_{S_1,\ldots,S_{N_3}} \Leftrightarrow \forall r, S_r \geq L_r.$$

1: Goulart & Comon, 2019.

Minimal ranks and counter-example

Def: Minimal (BTD) ranks[1] of X

$$\rho(\mathsf{X}) = (L_1, \ldots, L_{N_3}) \quad \text{when} \quad \mathsf{X} \in \mathcal{B}_{S_1,\ldots,S_{N_3}} \Leftrightarrow \forall r, S_r \geq L_r.$$

Example: $\mathsf{X} \in \mathcal{B}_{4,2} \cap \mathcal{B}_{3,3} \implies \rho(\mathsf{X}) \neq (4,2)$ and $\rho(\mathsf{X}) \neq (3,3)$. □

Now we're ready to generalize the example given for the CPD.

1: Goulart & Comon, 2019.

Minimal ranks and counter-example

Def: Minimal (BTD) ranks[1] of X

$$\rho(\mathsf{X}) = (L_1, \ldots, L_{N_3}) \quad \text{when} \quad \mathsf{X} \in \mathcal{B}_{S_1,\ldots,S_{N_3}} \;\Leftrightarrow\; \forall r, S_r \geq L_r.$$

Example: $\mathsf{X} \in \mathcal{B}_{4,2} \cap \mathcal{B}_{3,3} \implies \rho(\mathsf{X}) \neq (4,2)$ and $\rho(\mathsf{X}) \neq (3,3)$. □

Now we're ready to generalize the example given for the CPD.

Example: Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N_1 \times S}$, $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{N_2 \times S}$ and l.i. vectors $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{N_3}$,

$$\mathsf{X} := \left(\mathbf{A}\mathbf{C}^\mathsf{T} + \mathbf{B}\mathbf{D}^\mathsf{T}\right) \otimes \boldsymbol{v} + (\mathbf{B}\mathbf{C}^\mathsf{T}) \otimes \boldsymbol{w},$$

$$\mathsf{X}_n := n\left[\left(\mathbf{B} + n^{-1}\mathbf{A}\right)\left(\mathbf{C} + n^{-1}\mathbf{D}\right)^\mathsf{T}\right] \otimes (\boldsymbol{v} + n^{-1}\boldsymbol{w}) - n(\mathbf{B}\mathbf{C}^\mathsf{T}) \otimes \boldsymbol{v} = \mathsf{X} + o(1).$$

$\hookrightarrow \in \mathcal{B}_{S,S}$

If $\min\left\{\text{rank}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix}, \text{rank}\begin{pmatrix}\mathbf{C} & \mathbf{D}\end{pmatrix}\right\} > R := \frac{3}{2}S$, then $\mathsf{X} \notin \mathcal{B}_{S,S}$, by Sylvester's inequality. Hence:

$$\arg\min_{\hat{\mathsf{X}} \in \mathcal{B}_{S,S}} \left\|\mathsf{X} - \hat{\mathsf{X}}\right\|_\mathsf{F} = \emptyset.$$

□

1: Goulart & Comon, 2019.

Non-existence of best BTD approximation

As in the CPD case, this can happen with positive probability for real tensors:

Thm (G. & Comon): No $2K \times 2K \times 2$ real-valued tensor X such that $\rho(\mathsf{X}) = (2K, 2K)$ admits a best approximation in $\mathcal{B}_{2K-1,2K-1}$. These tensors form a non-empty open set.

Non-existence of best BTD approximation

As in the CPD case, this can happen with positive probability for real tensors:

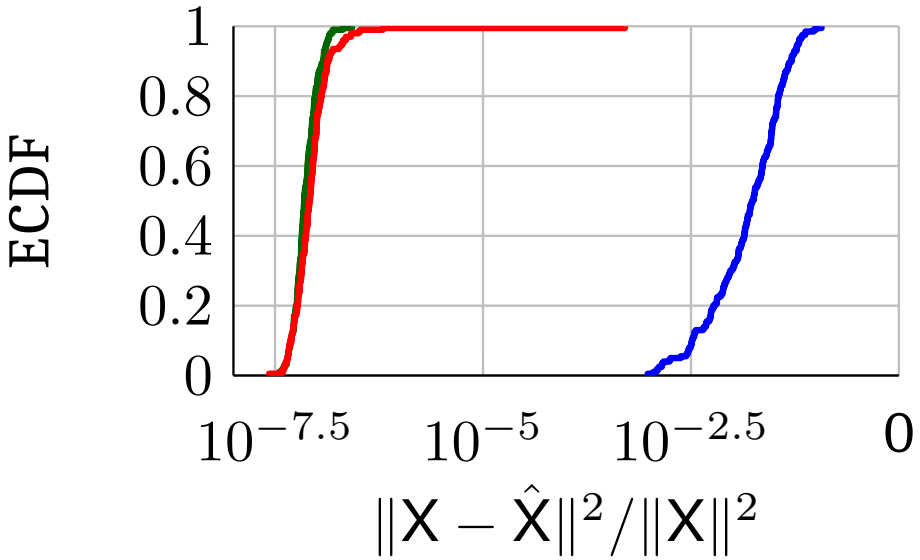
Thm (G. & Comon): No $2K \times 2K \times 2$ real-valued tensor X such that $\rho(\mathsf{X}) = (2K, 2K)$ admits a best approximation in $\mathcal{B}_{2K-1,2K-1}$. These tensors form a non-empty open set.

What happens if we try to compute it anyway?

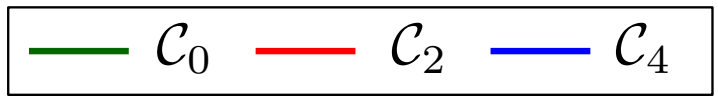
- Optimization algorithm never seems to converge
- Keeps improving error but ultimately gives nearly collinear blocks
- Norms of these blocks "blow up," but overall error stays bounded

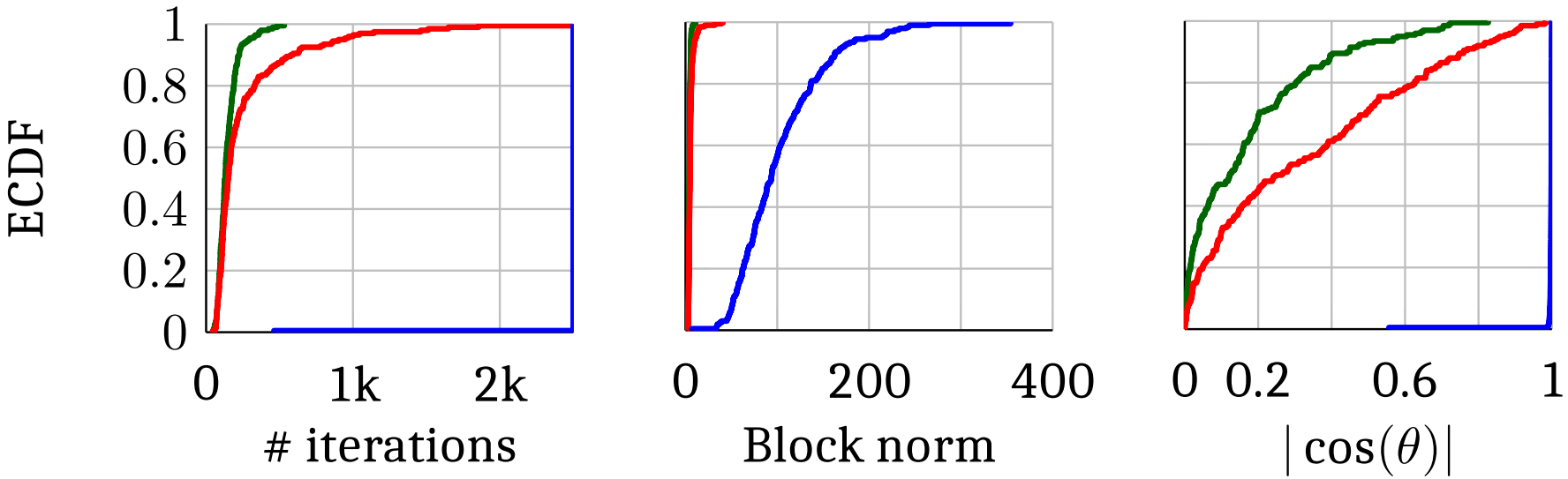
Numerical example

Approximation in $\mathcal{B}_{3,3}$ of $4 \times 4 \times 2$ tensor with opt. algorithm (best of 50 init.):



Classes of tensors, as per # of eigenvalues $\in \mathbb{C}$ of associated pencil:





Computation of approximate BTD

Several algorithms exist (see handout), including:

- alternating optimization (BCD)
- algebraic methods
- standard optimization schemes

Computation of approximate BTD

Several algorithms exist (see handout), including:

- alternating optimization (BCD)
- algebraic methods
- standard optimization schemes

Some of them are based on the standard least-squares formulation

$$\inf_{\{\mathbf{A}_r, \mathbf{B}_r, \boldsymbol{c}_r\}_{r=1}^{R}} \left\| \mathsf{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r \, \mathbf{B}_r^{\mathsf{T}} \right) \otimes \boldsymbol{c}_r \right\|_{\mathsf{F}}^2 .$$

Problems with this approach:

- the structure $(L_1, \ldots, L_R)$ must be fixed a priori
- a solution might not exist
- traversing regions of ill-conditioned BTDs (& thus slow progress)
- poor local minima due to block rank inversion

Joint estimation of BTD parameters & ranks

Alternative: automatic rank selection by regularization

$$\min_{\{\mathbf{A}_r,\mathbf{B}_r,\boldsymbol{c}_r\}_{r=1}^{R}} \left\| \mathcal{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r\, \mathbf{B}_r{}^{\mathsf{T}}\right) \otimes \boldsymbol{c}_r \right\|_{\mathsf{F}}^2 + \lambda \underbrace{\sum_{r=1}^{R} \left(\sum_{m=1}^{L_m} (\|(\mathbf{A}_r)_{:m}\| + \|(\mathbf{B}_r)_{:m}\|) + \|\boldsymbol{c}_r\| \right)}_{= \|\mathbf{A}\|_{2,1} + \|\mathbf{B}\|_{2,1} + \|\mathbf{C}\|_{2,1}}$$

Solutions always exist (coercive objective).

Joint estimation of BTD parameters & ranks

Alternative: automatic rank selection by regularization

$$\min_{\{\mathbf{A}_r,\mathbf{B}_r,\boldsymbol{c}_r\}_{r=1}^{R}} \left\| \mathsf{X} - \sum_{r=1}^{R} \left(\mathbf{A}_r \mathbf{B}_r^{\mathsf{T}}\right) \otimes \boldsymbol{c}_r \right\|_{\mathsf{F}}^2 + \lambda \underbrace{\sum_{r=1}^{R} \left(\sum_{m=1}^{L_m} (\|(\mathbf{A}_r)_{:m}\| + \|(\mathbf{B}_r)_{:m}\|) + \|\boldsymbol{c}_r\| \right)}_{= \|\mathbf{A}\|_{2,1} + \|\mathbf{B}\|_{2,1} + \|\mathbf{C}\|_{2,1}}$$

Solutions always exist (coercive objective).

Alternating group lasso (AGL) algorithm[1]:

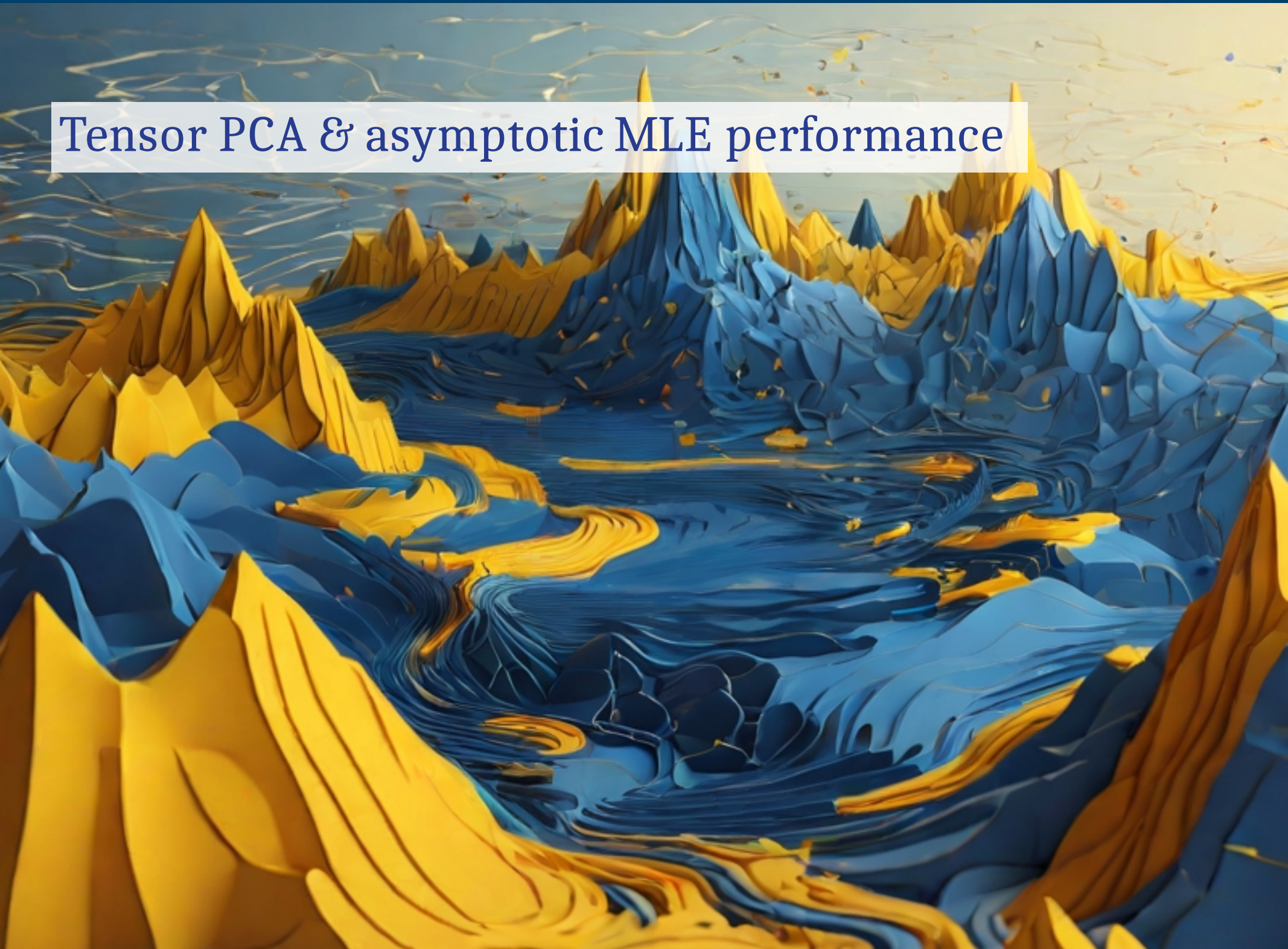
- block-coordinate descent algorithm
- each subproblem is a (convex) group lasso problem
- by adding a proximal term, all limit points are stationary points[2]

Subproblem in $\mathbf{A}$ (those in $\mathbf{B}$ and $\mathbf{C}$ are similar):

$$\min_{\mathbf{A}} \|\mathrm{vec}(\mathsf{X}) - \mathbf{W}_{\mathbf{B},\mathbf{C}} \,\mathrm{vec}(\mathbf{A})\|^2 + \lambda \|\mathbf{A}\|_{2,1} + \tau \|\mathbf{A} - \mathbf{A}_0\|_{\mathsf{F}}^2$$

1: Goulart & al., 2020, 2: Razaviyayn & al., 2013

Tensor PCA & asymptotic MLE performance



Tensor PCA & large-dimensional regime

Large body of recent work on the tensor PCA problem[1]

$$\max_{\|\boldsymbol{u}\|=1} \sum_{ijk} x_{ijk}\, u_i u_j u_k = \max_{\|\boldsymbol{u}\|=1} \lambda \langle \boldsymbol{a}, \boldsymbol{u} \rangle^3 + \sum_{ijk} w_{ijk}\, u_i u_j u_k$$

since the introduction of the spiked (symmetric) rank-1 model (same as before)

$$\mathsf{X} = \lambda\, \boldsymbol{a}^{\otimes 3} + \mathsf{W}.$$

Special attention is paid to the large-dimensional regime, as $N \to \infty$.

1: Montanari & Richard, 2014

Tensor PCA & large-dimensional regime

Large body of recent work on the tensor PCA problem[1]

$$\max_{\|\boldsymbol{u}\|=1} \sum_{ijk} x_{ijk}\, u_i u_j u_k = \max_{\|\boldsymbol{u}\|=1} \lambda \langle \boldsymbol{a}, \boldsymbol{u} \rangle^3 + \sum_{ijk} w_{ijk}\, u_i u_j u_k$$

since the introduction of the spiked (symmetric) rank-1 model (same as before)

$$\mathsf{X} = \lambda\, \boldsymbol{a}^{\otimes 3} + \mathsf{W}.$$

Special attention is paid to the large-dimensional regime, as $N \to \infty$.

Connections to many fields & exciting hot topics (see refs. on handout):

- study of disordered systems, spin glasses & statistical physics
- random optimization landscape
- high-dimensional probability & statistics
- random matrix theory

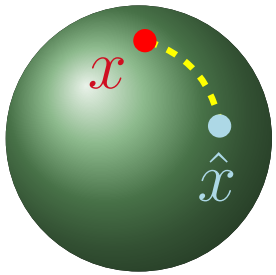
1: Montanari & Richard, 2014

Asymptotic performance limits?

Given any estimator $\hat{\boldsymbol{a}} : \mathcal{S}^3(N) \to \mathbb{S}^{N-1}$, a natural performance measure is:

Def: alignment (or overlap)

$$\alpha_{3,N}(\lambda) := \langle \boldsymbol{a}, \hat{\boldsymbol{a}}(\mathsf{X}) \rangle \quad \in [-1, 1]$$



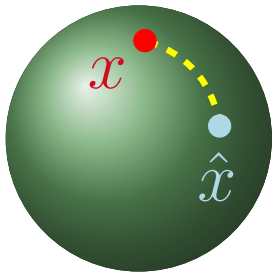
Example: If $\hat{\boldsymbol{a}} \sim \mathcal{U}(\mathbb{S}^{N-1})$, then asymptotically $\boldsymbol{a} \perp \hat{\boldsymbol{a}}$ almost surely.

Asymptotic performance limits?

Given any estimator $\hat{\boldsymbol{a}} : \mathcal{S}^3(N) \to \mathbb{S}^{N-1}$, a natural performance measure is:

Def: alignment (or overlap)

$$\alpha_{3,N}(\lambda) := \langle \boldsymbol{a}, \hat{\boldsymbol{a}}(\mathsf{X}) \rangle \quad \in [-1, 1]$$



Example: If $\hat{\boldsymbol{a}} \sim \mathcal{U}(\mathbb{S}^{N-1})$, then asymptotically $\boldsymbol{a} \perp \hat{\boldsymbol{a}}$ almost surely.

Central questions:

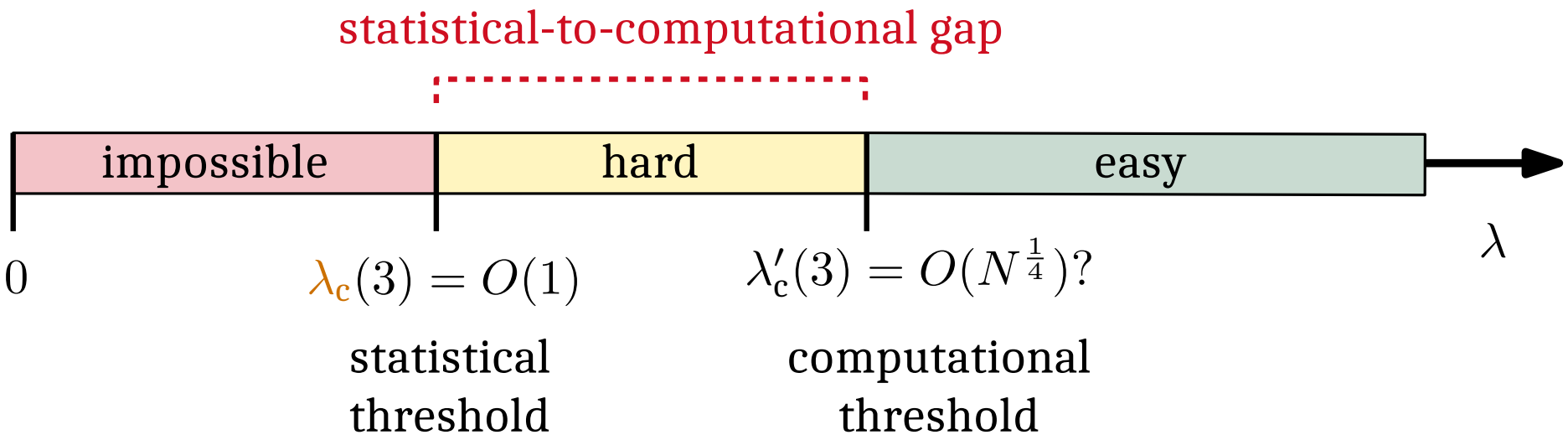
1. Weak recovery: for which range of λ is there a $\hat{\boldsymbol{a}}$ such that

$$\limsup_{N\to\infty} \mathbb{E}\{\alpha_{3,N}(\lambda)\} > 0 \ ?$$

2. Best asymptotic alignment: what is the largest attainable value of $\limsup_{N\to\infty} \mathbb{E}\{\alpha_{3,N}(\lambda)\}$ for each λ?

Answers and conjectured gap

1. Regimes of weak recovery:



2. Maximum likelihood estimation (MLE) attains the information-theoretic bound on the alignment for all λ.

Richard & Montanari, 2014, Montanari & al., 2017, Ben Arous & al., 2019, Jagannath & al., 2020, Perry & al., 2020, Ros & al., 2019

MLE performance

Settled by Jagannath–Lopatto–Miolane (2020), thanks to spin glass theory:

$$\mu^{\star}_{d,N}(\lambda) = \max_{\|\boldsymbol{u}\|=1} \left\{ \lambda \langle \boldsymbol{a}, \boldsymbol{u} \rangle^d + \mathsf{W} \cdot \boldsymbol{u}^d \right\} \quad \xrightarrow{\text{a.s.}} \quad \mathsf{GS}_d + \int_0^{\lambda} q_d^*(t)^{d/2} \, dt$$

$$|\alpha_{d,N}(\lambda)| = |\langle \boldsymbol{a}, \hat{\boldsymbol{a}} \rangle| \quad \xrightarrow{\text{a.s.}} \quad \sqrt{q_d^*(\lambda)}$$

Explicit expressions exist for $d = 3, 4, 5$.

MLE performance

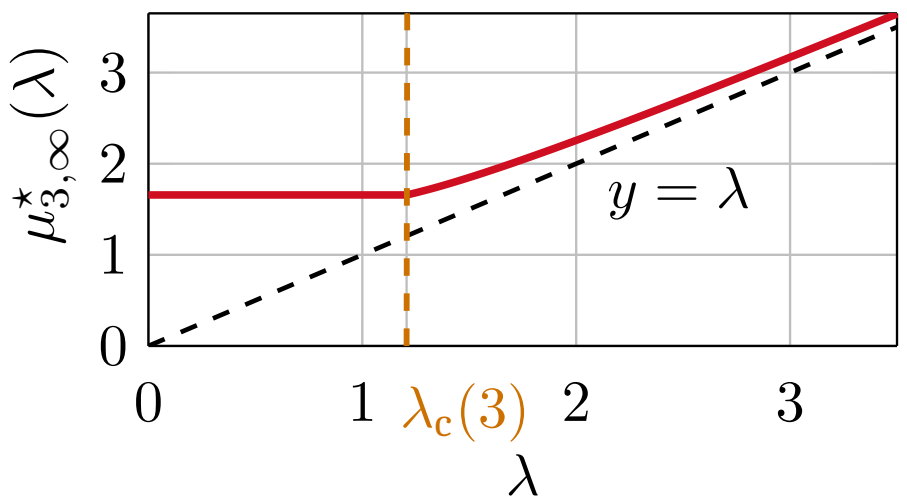
Settled by Jagannath–Lopatto–Miolane (2020), thanks to spin glass theory:

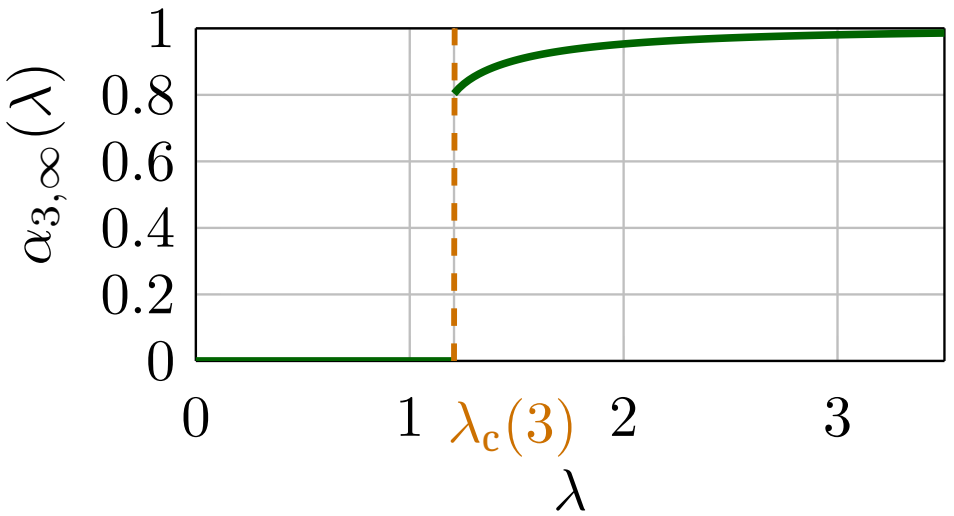
$$\mu^\star_{d,N}(\lambda) = \max_{\|\boldsymbol{u}\|=1} \left\{ \lambda \langle \boldsymbol{a}, \boldsymbol{u} \rangle^d + \mathsf{W} \cdot \boldsymbol{u}^d \right\} \xrightarrow{\text{a.s.}} \mathsf{GS}_d + \int_0^\lambda q_d^*(t)^{d/2}\, dt$$

$$|\alpha_{d,N}(\lambda)| = |\langle \boldsymbol{a}, \hat{\boldsymbol{a}} \rangle| \xrightarrow{\text{a.s.}} \sqrt{q_d^*(\lambda)}$$

Explicit expressions exist for $d = 3, 4, 5$.

For all d, these quantites undergo a phase transition at a threshold $\lambda_{\mathrm{c}}(d) = O(1)$.





Furthermore, the MLE attains the bound $\limsup_N \mathbb{E}\{|\langle \boldsymbol{a}, \hat{\boldsymbol{a}} \rangle|\} \le \sqrt{q_d^*(\lambda)}$

Extension to other spiked models?

However, it is not obvious how to use these tools to handle other, more general, models.

This motivated our recent contribution[1] where we carry out a similar analysis using tools from random matrix theory.

Using that approach, the estimation of other models has been adressed, notably in the asymmetric case:[2]

$$\mathsf{X} = \lambda\, \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} + \mathsf{W}$$

1: Goulart & al., 2022, 2: Seddik & al., 2024, 3: Lebeau & al., 2024b

Extension to other spiked models?

However, it is not obvious how to use these tools to handle other, more general, models.

This motivated our recent contribution[1] where we carry out a similar analysis using tools from random matrix theory.

Using that approach, the estimation of other models has been adressed, notably in the asymmetric case:[2]

$$\mathsf{X} = \lambda\, \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c} + \mathsf{W}$$

and for a nested matrix-tensor model[3] which applies to a simplified multi-view clustering model:

$$\mathsf{X} = (\boldsymbol{\mu}\boldsymbol{b}^{\mathsf{T}} + \mathbf{Z}) \otimes \boldsymbol{c} + \mathsf{W},$$

where $\boldsymbol{\mu} \in R^p$ models cluster means, $\boldsymbol{b} \in \{-1, 1\}^N$ models cluster assignments, $\mathbf{Z}$ is a Gaussian matrix modeling data dispersion, W models measurement noise and $\boldsymbol{c} \in \mathbb{R}^M$ models varying SNR conditions.

1: Goulart & al., 2022, 2: Seddik & al., 2024, 3: Lebeau & al., 2024b

Tensor and matrix eigenpairs

Another characterization of tensor eigenpairs (assuming $\|\boldsymbol{u}\| = 1$):

$$(\mu, \boldsymbol{u}) \text{ eigenpair of } \mathsf{Y} \quad \Leftrightarrow \quad (\mu, \boldsymbol{u}) \text{ eigenpair of } \mathsf{Y} \cdot \boldsymbol{u}^{d-2}$$

Proof: $\mu\, \boldsymbol{u} = \mathsf{Y} \cdot \boldsymbol{u}^{d-1} = (\mathsf{Y} \cdot \boldsymbol{u}^{d-2})\, \boldsymbol{u}$ ■

Tensor and matrix eigenpairs

Another characterization of tensor eigenpairs (assuming $\|\boldsymbol{u}\| = 1$):

$$(\mu, \boldsymbol{u}) \text{ eigenpair of } \mathsf{Y} \quad \Leftrightarrow \quad (\mu, \boldsymbol{u}) \text{ eigenpair of } \mathsf{Y} \cdot \boldsymbol{u}^{d-2}$$

Proof: $\mu\, \boldsymbol{u} = \mathsf{Y} \cdot \boldsymbol{u}^{d-1} = (\mathsf{Y} \cdot \boldsymbol{u}^{d-2})\, \boldsymbol{u}$ ■

Observed performance of algorithms

The performance of power iteration evidently depends on its initialization.

In tensor PCA, randomly initialized power iteration has[1,2] a (conjectured) algorithmic threshold of $\lambda_{\text{alg}} = O(N^{\frac{d-1}{2}})$. This is the SNR required to "beat entropy."

With a spectral initialization (unfolding SVD), its threshold is[3] $\lambda'_{\text{alg}} = O(N^{\frac{d-2}{4}})$.

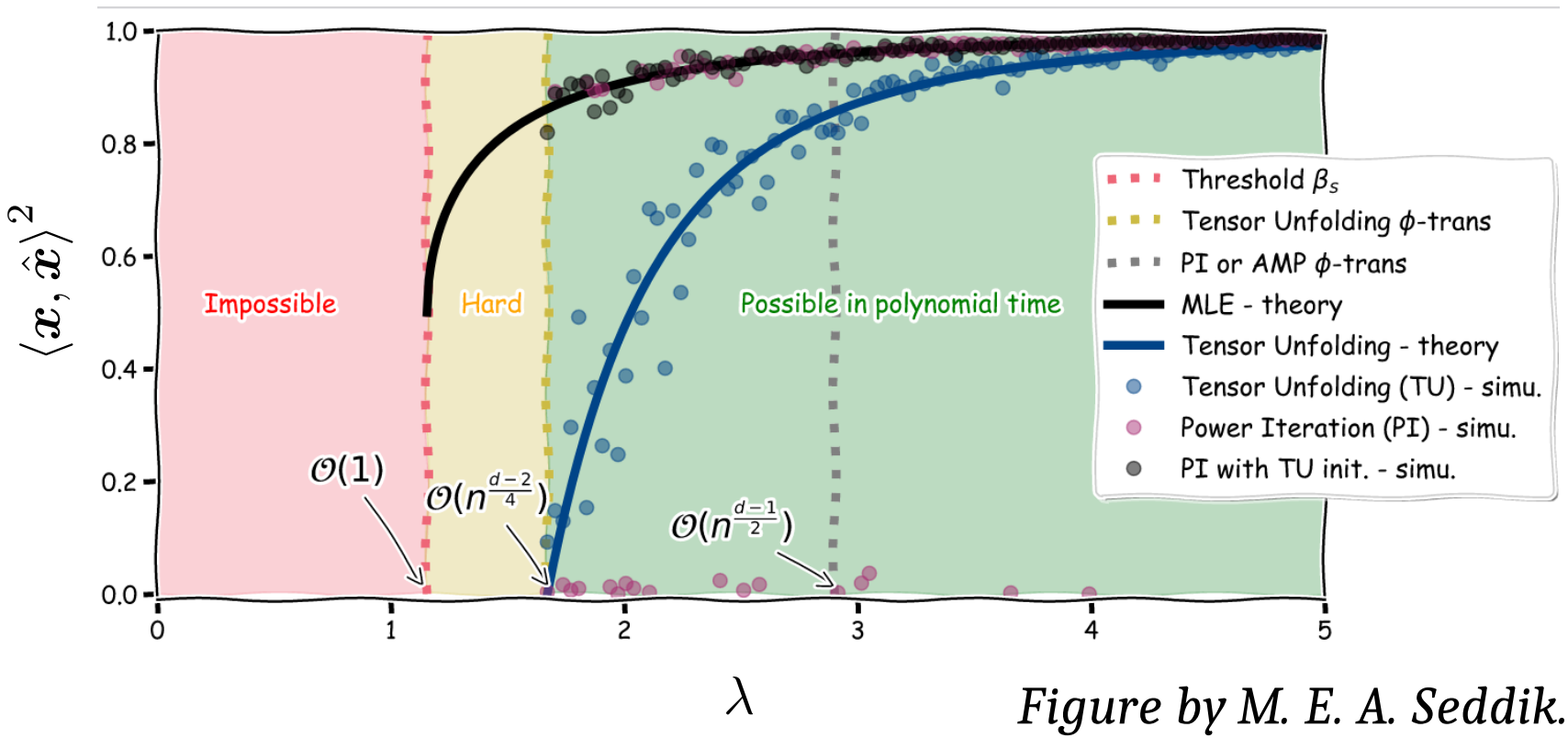


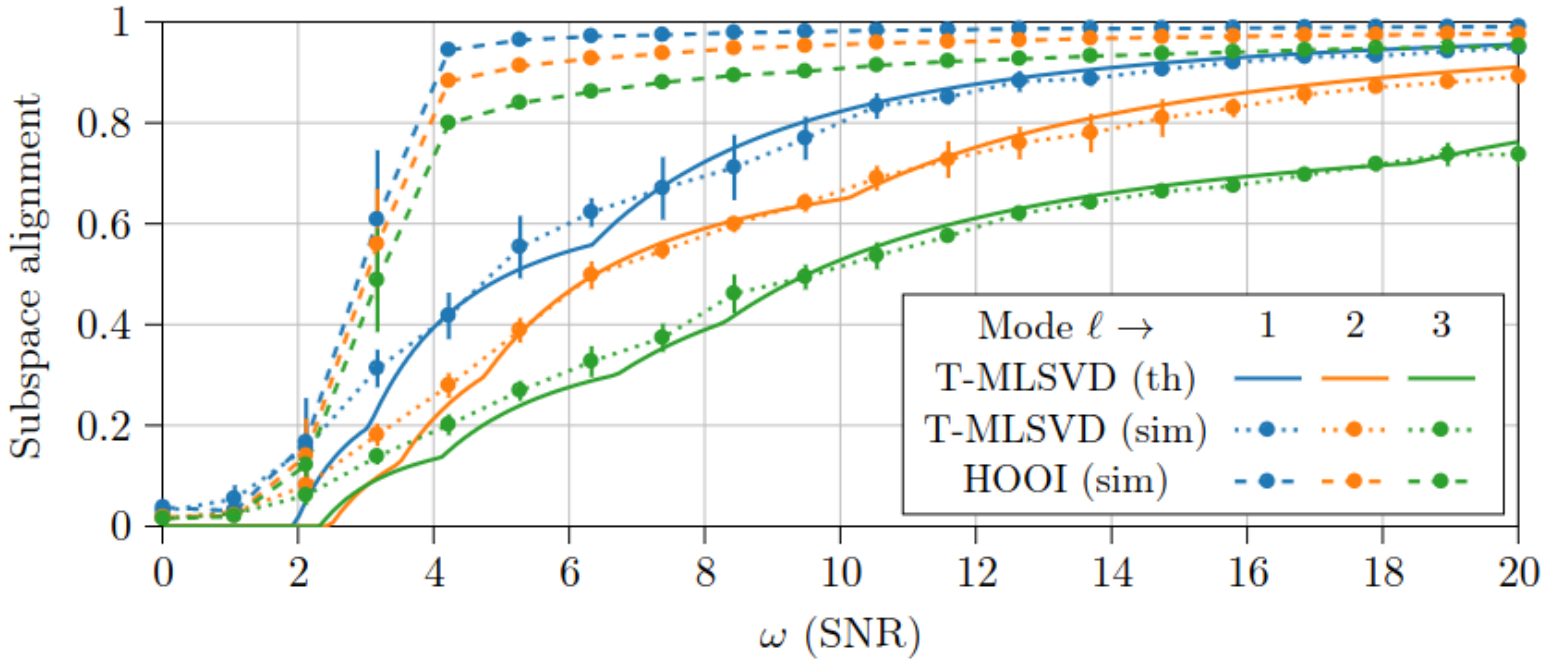
Figure by M. E. A. Seddik.

1: Montanari & Richard, 2014, 2: Huang & al., 2022, 3: Ben Arous & al., 2023

Performance of THOSVD & HOOI

The large-N performance of THOSVD was analyzed under Gaussian noise by Lebeau & al. (2024a), unveiling a phase transition w.r.t. the SNR at $O(N^{\frac{d-2}{4}})$.

Similar theoretical results for the MLE (best low-mrank approximation) do not yet exist, only empirical ones, eg:



Performances of THOSVD and HOOI[1] *for* $(N_1, N_2, N_3) = (100, 200, 300)$ *and* $(R_1, R_2, R_3) = (3, 4, 5)$.

1: Lebeau & al., 2024a

See you next fall !

Thematic Trimester

Beyond classical regimes in statistical inference and machine learning

September to November 2024

Institut de Mathématiques de Toulouse - France

1 Colloquium
2 Thematic Schools &
1 Workshop,

- **Opening Colloquium**
  September 11th, 2024
- **Thematic School: Optimization & algorithms for high-dimensional machine learning and inference**
  October 7th to 11th, 2024
- **Thematic School: Models & methods for high-dimensional machine learning and inference**
  October 14th to 18th, 2024
- **Workshop**
  November 4th to 8th, 2024

CIMI
Centre International de Mathématiques et d'Informatique
TOULOUSE

Organizers: Henrique Goulart (IRIT/Toulouse INP), Vanessa Kientz (CEA List), Vincent Lahoche (CEA List), Xiaoyi Mai (IMT/UT2J), Mohamed Tamaazousti (CEA List)

https://indico.math.cnrs.fr/category/682

Supplementary slides

Tensors: (sketch of a) formal definition

The idea is to impose the multilinear structure by means of a quotient space.

Take, for instance, two vector spaces $\mathcal{U} = \mathbb{R}^{N_1}$ and $\mathcal{V} = \mathbb{R}^{N_2}$. For any pair $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{U} \times \mathcal{V}$, we would like that

$$\boldsymbol{u} \otimes \boldsymbol{v} = (\alpha \boldsymbol{u}) \otimes (\alpha^{-1} \boldsymbol{v}), \quad \forall \alpha \neq 0.$$

Hence, $\boldsymbol{u} \otimes \boldsymbol{v}$ can be seen as an equivalence class defined on the free vector space $F(\mathcal{U} \times \mathcal{V})$ containing all elements $f_{(\alpha \boldsymbol{u}, \alpha^{-1} \boldsymbol{v})} \in F(\mathcal{U} \times \mathcal{V})$.

This leads to the definition: $\mathcal{U} \otimes \mathcal{V} := F(\mathcal{U} \times \mathcal{V})/\mathcal{N}$, where $\mathcal{N}$ is the "null subspace"

$$\mathcal{N} := \text{span} \left\{ \sum_{i_1} \sum_{i_2} \alpha_{i_1} \beta_{i_2} f_{(\boldsymbol{u}_{i_1}, \boldsymbol{v}_{i_2})} - f_{\left(\sum_{i_1} \alpha_{i_1} \boldsymbol{u}_{i_1}, \sum_{i_2} \beta_{i_2} \boldsymbol{v}_{i_2}\right)} \; : \; \boldsymbol{u}_{i_1} \in \mathcal{U}, \boldsymbol{v}_{i_2} \in \mathcal{V} \right\}.$$

Example #5: High-dim density approximation

Recent works have relied on low-rank tensor approximation for representing (& learning) densities in high-dim spaces, such as:[1]

$$q(\boldsymbol{x}) = \sum_{n_1=1}^{N} \cdots \sum_{n_d=1}^{N} a_{n_1,\ldots,n_d} f_{n_1}(x_1) \ldots f_{n_d}(x_d) = \langle \mathsf{A}, \mathsf{F}(\boldsymbol{x}) \rangle \,,$$

with $\mathsf{F}(\boldsymbol{x}) = \boldsymbol{f}(x_1) \otimes \cdots \otimes \boldsymbol{f}(x_d)$ and $\boldsymbol{f}(x_i) = \begin{pmatrix} f_1(x_i) & \ldots & f_N(x_i) \end{pmatrix}^\mathsf{T}$ a vector of chosen basis functions evaluated at x_i.

1: Novikov & al., 2021

Example #5: High-dim density approximation

Recent works have relied on low-rank tensor approximation for representing (& learning) densities in high-dim spaces, such as:[1]

$$q(\boldsymbol{x}) = \sum_{n_1=1}^{N} \cdots \sum_{n_d=1}^{N} a_{n_1,\ldots,n_d} \, f_{n_1}(x_1) \ldots f_{n_d}(x_d) = \langle \mathsf{A}, \mathsf{F}(\boldsymbol{x}) \rangle \,,$$

with $\mathsf{F}(\boldsymbol{x}) = \boldsymbol{f}(x_1) \otimes \cdots \otimes \boldsymbol{f}(x_d)$ and $\boldsymbol{f}(x_i) = \begin{pmatrix} f_1(x_i) & \ldots & f_N(x_i) \end{pmatrix}^\mathsf{T}$ a vector of chosen basis functions evaluated at x_i.

The curse of dimensionality is broken by imposing a low-rank structure on A, namely, a tensor-train structure:

$$a_{n_1,\ldots,n_d} = \underset{R_1}{\boldsymbol{g}^{(1)}(n_1)^\mathsf{T}} \underset{R_1 \times R_2}{\mathbf{G}^{(2)}(n_2)} \ldots \underset{R_{d-2} \times R_{d-1}}{\mathbf{G}^{(d-1)}(n_{d-1})} \underset{R_{d-1}}{\boldsymbol{g}^{(d)}(n_d)}.$$

The coordinate tensor A (of dim N^d) is parameterized by the above vectors & matrices, whose sizes control the model complexity.

1: Novikov & al., 2021

Pointwise evaluation, marginalizing & sampling

The separability in $x_1, \ldots, x_d$ is key to complexity reduction:

$$q(\boldsymbol{x}) = \left(\sum_{n_1} f_{n_1}(x_1) \boldsymbol{g}^{(1)}(n_1) \right)^{\mathsf{T}} \left(\sum_{n_2} f_{n_2}(x_2) \mathbf{G}^{(2)}(n_2) \right) \ldots \left(\sum_{n_d} f_{n_d}(x_d) \boldsymbol{g}^{(d)}(n_d) \right)$$

which takes $O(dNR^2)$ operations if $R_i = R$ for all i, instead of N^d!

Novikov & al., 2021

Pointwise evaluation, marginalizing & sampling

The separability in $x_1, \ldots, x_d$ is key to complexity reduction:

$$q(\boldsymbol{x}) = \left(\sum_{n_1} f_{n_1}(x_1) \boldsymbol{g}^{(1)}(n_1) \right)^{\mathsf{T}} \left(\sum_{n_2} f_{n_2}(x_2) \mathbf{G}^{(2)}(n_2) \right) \ldots \left(\sum_{n_d} f_{n_d}(x_d) \boldsymbol{g}^{(d)}(n_d) \right)$$

which takes $O(dNR^2)$ operations if $R_i = R$ for all i, instead of N^d!

Similarly, marginals can be computed using

$$\begin{aligned} q(x_1, \ldots, x_{k-1}) &= \int q(\boldsymbol{x}) \, \mathrm{d}x_k \ldots \mathrm{d}x_d \\ &= \left\langle \mathsf{A}, \boldsymbol{f}(x_1) \otimes \cdots \otimes \boldsymbol{f}(x_{k-1}) \otimes \left(\int \boldsymbol{f}(x_k) \, \mathrm{d}x_k \right) \otimes \cdots \otimes \left(\int \boldsymbol{f}(x_d) \, \mathrm{d}x_d \right) \right\rangle \end{aligned}$$

Novikov & al., 2021

Pointwise evaluation, marginalizing & sampling

The separability in $x_1, \ldots, x_d$ is key to complexity reduction:

$$q(\boldsymbol{x}) = \left(\sum_{n_1} f_{n_1}(x_1) \boldsymbol{g}^{(1)}(n_1) \right)^{\mathsf{T}} \left(\sum_{n_2} f_{n_2}(x_2) \mathbf{G}^{(2)}(n_2) \right) \ldots \left(\sum_{n_d} f_{n_d}(x_d) \boldsymbol{g}^{(d)}(n_d) \right)$$

which takes $O(dNR^2)$ operations if $R_i = R$ for all i, instead of N^d!

Similarly, marginals can be computed using

$$\begin{aligned} q(x_1, \ldots, x_{k-1}) &= \int q(\boldsymbol{x}) \, \mathrm{d}x_k \ldots \mathrm{d}x_d \\ &= \left\langle \mathsf{A}, \boldsymbol{f}(x_1) \otimes \cdots \otimes \boldsymbol{f}(x_{k-1}) \otimes \left(\int \boldsymbol{f}(x_k) \, \mathrm{d}x_k \right) \otimes \cdots \otimes \left(\int \boldsymbol{f}(x_d) \, \mathrm{d}x_d \right) \right\rangle \end{aligned}$$

(Conditional) CDFs can also be computed with 1D integration and summations/matrix-vector products, allowing in particular efficient sampling from $q(\boldsymbol{x})$.

Novikov & al., 2021

The tensor train (TT) model

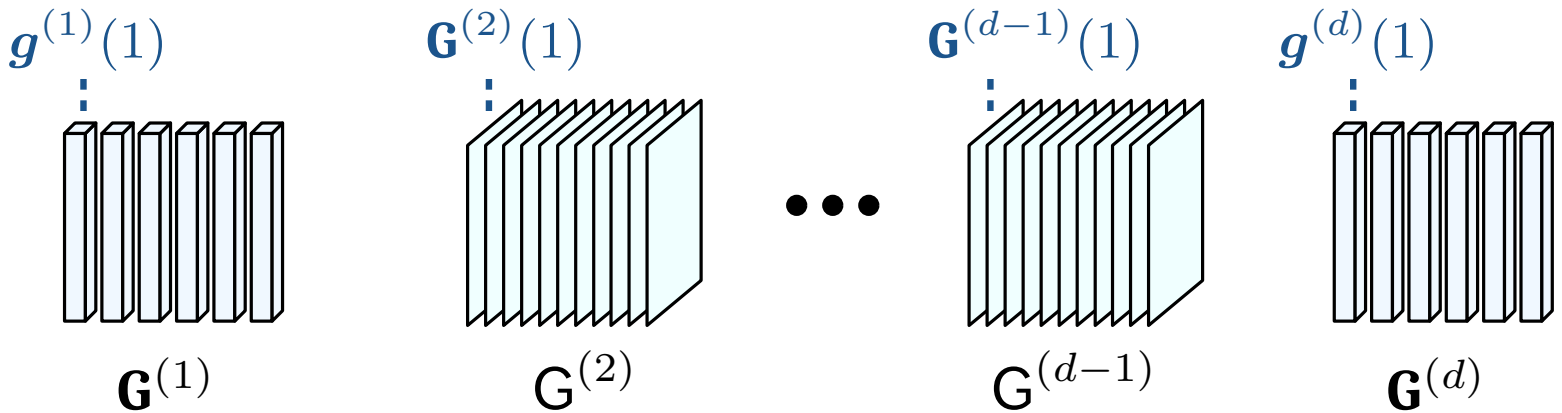
Def: Tensor train decomposition[1] of $\mathsf{X} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$:

$$x_{n_1,\ldots,n_d} = \underset{R_1}{\boldsymbol{g}^{(1)}(n_1)^\mathsf{T}} \underset{R_1 \times R_2}{\mathbf{G}^{(2)}(n_2)} \ldots \underset{R_{d-2} \times R_{d-1}}{\mathbf{G}^{(d-1)}(n_{d-1})} \underset{R_{d-1}}{\boldsymbol{g}^{(d)}(n_d)}$$

parameterized by core tensors $\mathbf{G}^{(1)} \in \mathbb{R}^{R_1 \times N_1}$, $\mathsf{G}^{(i)} \in \mathbb{R}^{R_{i-1} \times N_i \times R_i}$, $i = 2, \ldots, d-1$, and $\mathbf{G}^{(d)} \in \mathbb{R}^{R_{d-1} \times N_d}$.

The TT-rank of X is $(R_1, \ldots, R_{d-1})$.

TT cores:



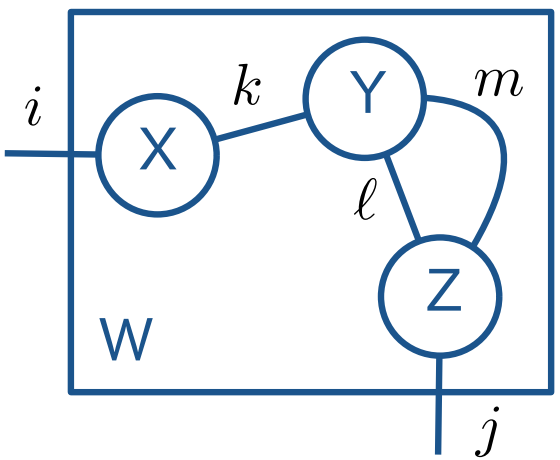
1: Oseledets, 2011

Tensor contractions

Contraction among tensors = summation over some dimensions.

Useful pictorial notation: tensor networks[1] displaying tensors as nodes and edges as their indices. A connection then means a contraction.

Ex:



$$w_{ij} = \sum_k \sum_\ell \sum_m x_{ik}\, y_{k\ell m}\, z_{\ell m j}$$

Def: Scalar product

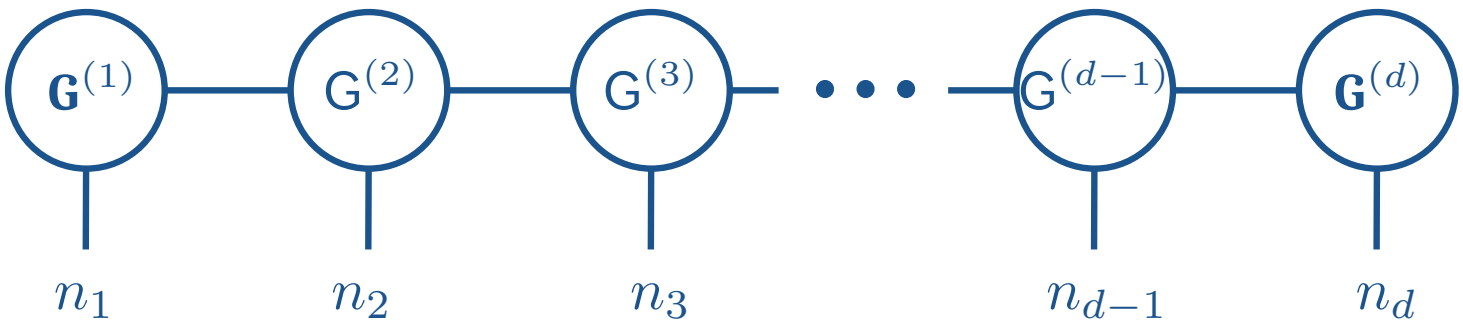
$$\langle \mathsf{X}, \mathsf{Y} \rangle = \sum_{ijk} x_{ijk}\, y_{ijk}$$

In particular:

$$\begin{cases} \langle \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}, \mathsf{Y} \rangle = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \cdot \mathsf{Y} \\ \langle \boldsymbol{a} \otimes \boldsymbol{b} \otimes \boldsymbol{c}, \boldsymbol{u} \otimes \boldsymbol{v} \otimes \boldsymbol{w} \rangle = \langle \boldsymbol{a}, \boldsymbol{u} \rangle \langle \boldsymbol{b}, \boldsymbol{v} \rangle \langle \boldsymbol{c}, \boldsymbol{w} \rangle \end{cases}$$

Properties of TT

The TT belongs to a larger class of hierarchical Tucker models[1], and can be seen as a tensor network of the form

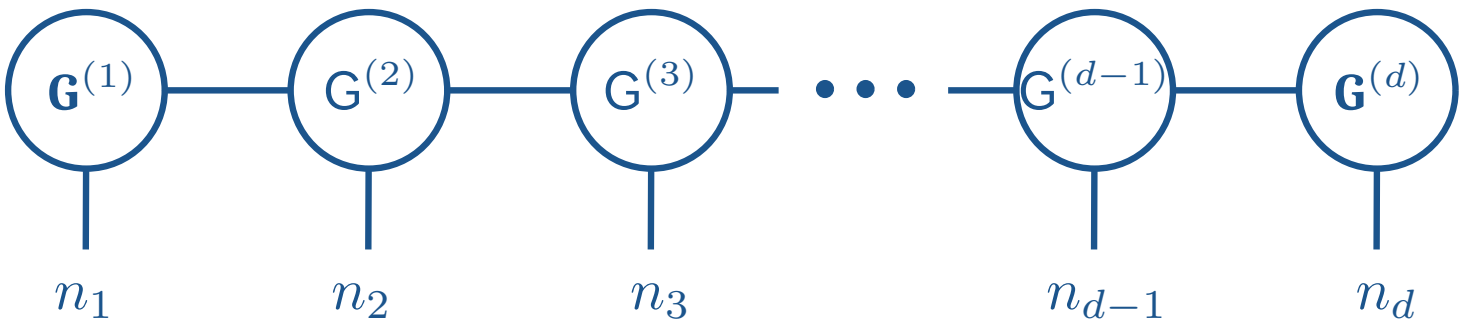


It is also known as the matrix-product state (MPS) model in the physics community.[2]

1: Hackbusch, 2012, 2: Vidal, 2003

Properties of TT

The TT belongs to a larger class of hierarchical Tucker models[1], and can be seen as a tensor network of the form



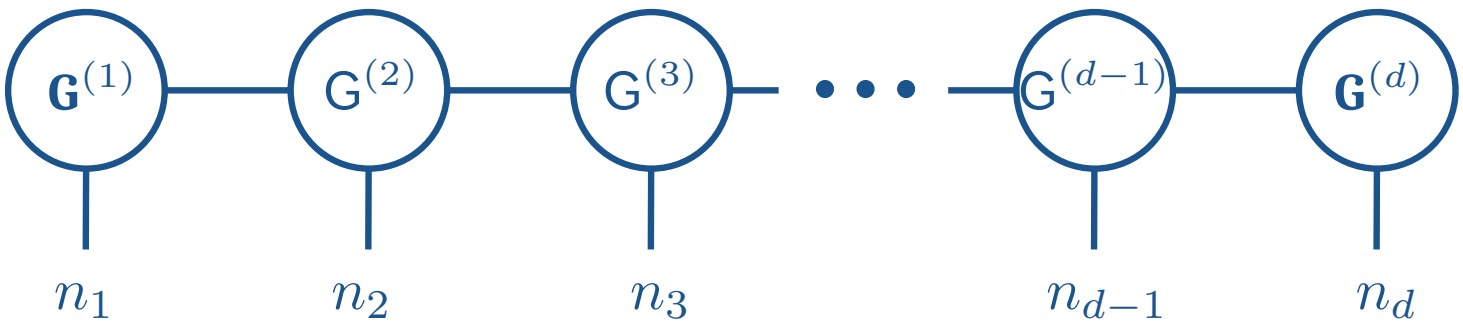
It is also known as the matrix-product state (MPS) model in the physics community.[2]

Motivation: complexity reduction. In particular, if $N_i = N$ and $R_i = R$, reduces storage cost from N^d to $O(dN^2R)$.

1: Hackbusch, 2012, 2: Vidal, 2003

Properties of TT

The TT belongs to a larger class of hierarchical Tucker models[1], and can be seen as a tensor network of the form



It is also known as the matrix-product state (MPS) model in the physics community.[2]

Motivation: complexity reduction. In particular, if $N_i = N$ and $R_i = R$, reduces storage cost from N^d to $O(dN^2R)$.

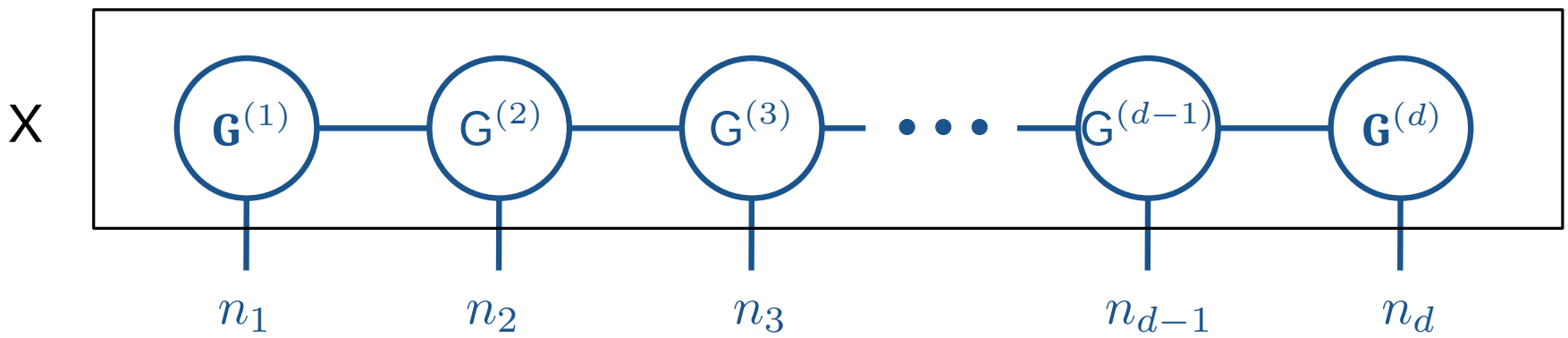
The TT is non-unique: for any nonsingular $\mathbf{S} \in \mathbb{R}^{R_i \times R_i}$,

$$\mathbf{G}^{(i)}(n_i)\,\mathbf{G}^{(i+1)}(n_{i+1}) = \mathbf{G}^{(i)}(n_i)\,\mathbf{S}\,\mathbf{S}^{-1}\,\mathbf{G}^{(i+1)}(n_{i+1}) = \tilde{\mathbf{G}}^{(i)}(n_i)\,\tilde{\mathbf{G}}^{(i+1)}(n_{i+1})$$

1: Hackbusch, 2012, 2: Vidal, 2003

Computation of approximate TT

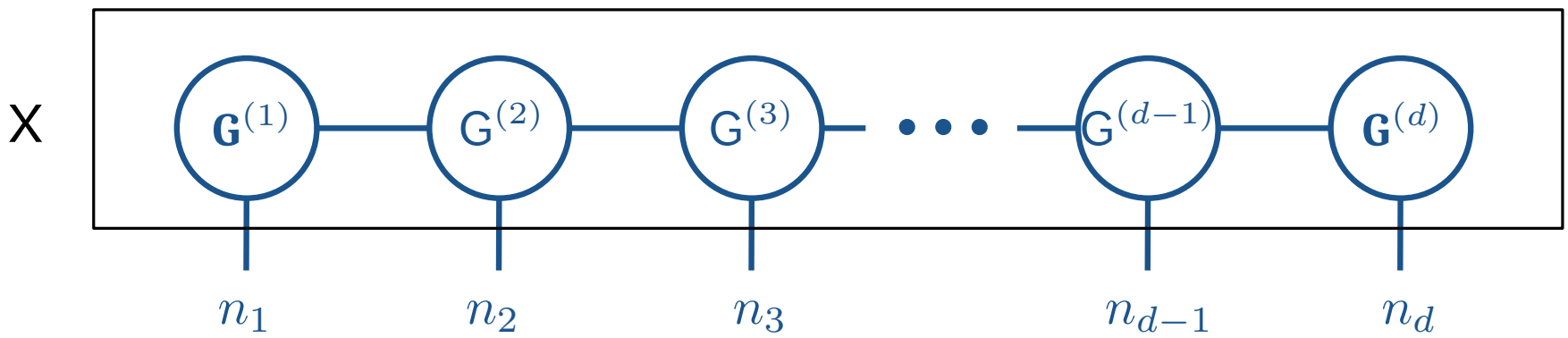
Standard (algebraic) algorithm: TT-SVD,[1] which performs a sequence of reshapings and truncated SVD, given target TT-ranks $(R_1, \ldots, R_{d-1})$.



1: Oseledets, 2011

Computation of approximate TT

Standard (algebraic) algorithm: TT-SVD,[1] which performs a sequence of reshapings and truncated SVD, given target TT-ranks $(R_1, \ldots, R_{d-1})$.



It is quasi-optimal, similarly to the THOSVD:

Thm (Oseledets, 2011): The TT-SVD algorithm satisfies

$$\|\mathsf{X} - \mathsf{X}_{\text{TT-SVD}}\|_{\mathsf{F}}^2 \leq (d-1) \left\|\mathsf{X} - \mathsf{X}_{\text{best-TT-}(R_1,\ldots,R_{d-1})}\right\|_{\mathsf{F}}^2$$

1: Oseledets, 2011