Implicit balancing in penalized low-rank approximations

with a short introduction to constrained tensor decompositions

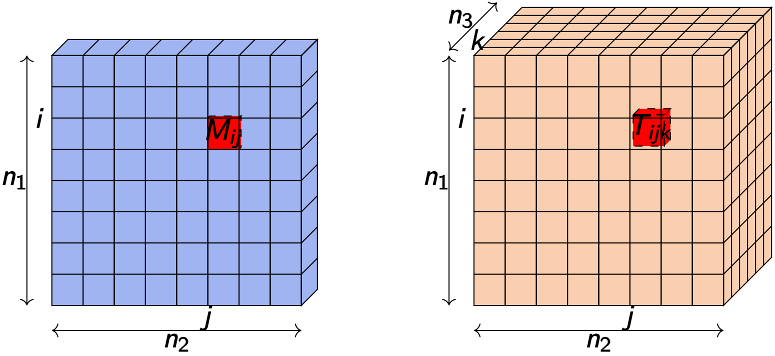
Jeremy Cohen
CNRS, CREATIS lab

Interpretable Constrained Tensor Decompositions Minisymposium
ICIAM Workshop, Tokyo, 2023.08.23

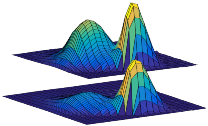
Matrices/Tensors as multiway arrays

Let $\mathcal{T}$ a tensor in $\mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$

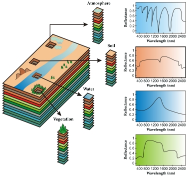
- modes: indices of the tensor from 1 to d. e.g. i is the first mode index.
- order: d. e.g. the tensor below is a third order tensor.



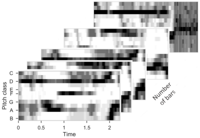
Examples of tensors in data science



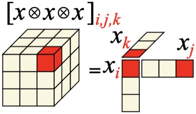
Tensor as Raw Data Excitation Emission Matrices



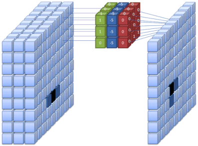
Tensor as Raw Data Hyperspectral Images
[courtesy of J Chanussot]



Tensor as Processed Data Tensor spectrogram



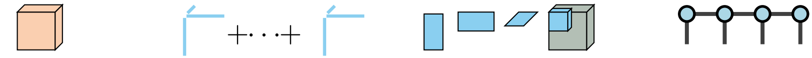
Tensor as Data Properties Data Moments



Tensor as Model Parameters Convolutional Neural Networks
[figure from commons.wikimedia.org]

Tensors and dimensionality reduction

Number of parameters:



$\mathcal{O}(n^d)$ $\quad$ $\mathcal{O}(dnr)$ $\quad$ $\mathcal{O}(dnr + r^d)$ $\quad$ $\mathcal{O}(dnr^2)$

with r the number of components.

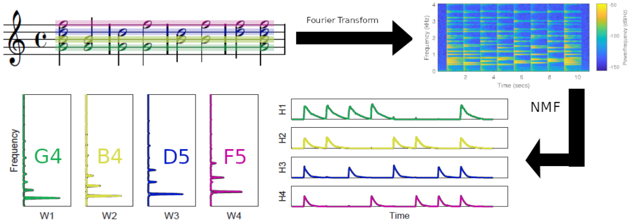
Inverse Problems

- Matrix-Tensor completion
- Blind Source separation
- Denoising, deconvolution
- Phase retrieval
- ...

Compression, Low Complexity

- Data mining
- Neural Networks
- Partial Differential Equations
- ...

Interpretability and constraints



Constraints help with interpretation by

- enhancing uniqueness.
- improving the quality of the solution.
- sometimes, making the optimization problem easier.

Examples: Nonnegative Matrix/Tucker Factorization, (Convolutional) Dictionary Learning, Principal Component Analysis. . .

Challenges for regularized tensor decompositions

Regularized Canonical Polyadic Decomposition (CPD):

$$\underset{A,B,C\in\mathbb{R}^{n_i\times r}}{\operatorname{argmin}} \|\mathcal{T}-\sum_{q=1}^{r} A[:,q]B[:,q]C[:,q]\|_F^2 + g_A(A) + g_B(B) + g_C(C)$$

- nonsmooth
- nonconvex

Nonnegative CPD $g_A = g_B = g_C = \eta_{\mathbb{R}_+}$
Sparse CPD $g_A = g_B = g_C = \|\cdot\|_1$

Challenges

- Numerical Optimization
- Efficient implementations
- Identifiability properties
- Multimodality

Program of the Minisymposium

co-organized with Daniel Dunlavy and Axel Marmoret.

1. Wednesday 13:20

Jeremy Cohen (Implicit balancing in penalized LRA)
Jamie Haddock (Hierarchical and neural tensor factorizations)
Derek DeSantis (Nonnegative canonical tensor decompositions with linear constraints: nnCANDELINC)
Clémence Prévost (Joint Data Fusion and Blind Unmixing using Nonnegative Tensor Decomposition)

2. Wednesday 15:30

Nico Vervliet (A quadratically convergent proximal algorithm for nonnegative tensor decomposition)
Carla Schenker (PARAFAC2-based coupled matrix and tensor factorization with constraints)
Daniel Dunlavy (Constrained Tucker Decompositions and Conservation Principles for Direct Numerical Simulation Data Compression)
Rafal Zdunek (Incremental Nonnegative Tucker Decomposition with Block-coordinate Descent and Recursive Approaches)

Program of the Minisymposium

co-organized with Daniel Dunlavy and Axel Marmoret.

3. Friday 10:40

Koby Hayashi (Speeding up Nonnegative Low-rank Approximations with Parallelism and Randomization)

Neriman Tockan (A probabilistic nonnegative tensor factorization method for tumor microenvironment analysis)

Ruhui Jin (Scalable symmetric Tucker tensor decomposition)

Izabel Aguiar (A tensor factorization model of mulitlayer network interdependence)

Discuss with other participants, onsite and online! Ask questions!

Implicit balancing in penalized low-rank approximations

Implicit regularization in matrix LRA

Let M some data matrix, and r a factorization rank.

$$\underset{W,H\in\mathbb{R}^{n_i\times r}}{\text{argmin}} \; \|M - WH^T\|_F^2 + \lambda\left(\|W\|_F^2 + \|H\|_F^2\right)$$

has balanced solutions $\|W^*[:,q]\|_2 = \|H^*[:,q]\|_2$. Its solutions are equivalent up to scaling to the solutions of

$$\underset{\text{rank}(L_q)=1}{\text{argmin}} \; \|M - \sum_{q\leq r} L_q\|_F^2 + \alpha \sum_{q\leq r} \|L_q\|_F$$

for some $\alpha > 0$, which is a modified Group-LASSO [Srebro 2008]. Remarkably, this can be reformulated as

$$\underset{L\in\mathbb{R}^{n_1\times n_2},\text{rank}(L)\leq r}{\text{argmin}} \; \|M - L\|_F^2 + \alpha\|L\|_*$$

Ridge penalties induce low-rank regularizations!

Also mentionned in [Uschmajew 2012] for CPD.

Penalized CPD framework

$$\underset{A,B,C\in\mathbb{R}^{n_i\times r}}{\text{argmin}} \|\mathcal{T}-[\![A,B,C]\!]\|_F^2+\sum_{q=1}^{r} g_A(A[:,q])+g_B(B[:,q])+g_C(C[:,q])$$

where g are homogeneous functions of degree p_A, p_B, p_C.

Observation

Since the CPD is scale invariant, solutions must minimize the scale of the regularization terms!

For the q-th set of columns and fixed estimates A, B, C, the optimal scaling may be computed as

$$\underset{\lambda_A\lambda_B\lambda_C=1}{\text{argmin}} \lambda_A^{p_A} g_A(A[:,q]) + \lambda_B^{p_B} g_B(B[:,q]) + \lambda_C^{p_C} g_C(C[:,q])$$

A little trick about means

Balancing occurs because of these two equivalent phenomena:

$$\min_{a \geq 0, b \geq 0} a + b \text{ such that } ab = p \tag{1}$$

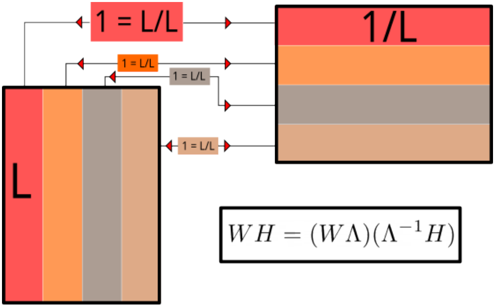
for a given $p \geq 0$ has solution $a = b = \sqrt{p}$.

$$\max_{a \geq 0, b \geq 0} ab \text{ such that } a + b = s \tag{2}$$

for a given $s \geq 0$ has solution $a = b = \frac{s}{2}$.

We can compute the optimal scaling in closed form easily!

Regularization and scale invariance



Scale invariance has an implicit balancing effect!

$$\forall q \leq r, \; \|W^*[:, q]\| \propto \|H^*[:, q]\|$$

Some equivalent reformulations

Explicit Reg.	Invariance	Implicit Reg
$\ell_2(W)^2 + \ell_2(H)^2$	col. scale, rotation	$\\|\sum_q L_q\\|_*$ or $\sum_q \\|L_q\\|_F$
$\ell_p(W)$	col. scale	ill-posed
$\ell_1(W) + \ell_1(H)$	col. scale	$\sum_q \sqrt{\\|L_q\\|_1}$
$\ell_1(W)^2 + \ell_1(H)^2$	scale	$\\|W \otimes H\\|_1^2$
$\sum_q \ell_1(W[:,q])^2 + \ell_1(H[:,q])^2$	col. scale	$\\|L\\|_1$
$\ell_1(W) + \ell_2(H)^2$	col. scale	$\sum_q \left(\\|W[:,q]\\|_1 \\|H[:,q]\\|_2^2\right)^{\frac{2}{3}}$
$\ell_1(W) + \sum_j \ell_2(H[:,q])$	col. scale	$\sum_q \left(\sum_j \\|L_q[j,:]\\|_2\right)^{1/4}$
$\sum_q \ell_1(W[:,q])^2 + \ell_2(H)^2$	col. scale	$\sum_q \sum_j \\|L_q[j,:]\\|_2$

Table: $L_q = W[:,q]H[:,q]^T$

Using rescaling in an optimization algorithm

Penalized LRA models may converge slowly to a local minimum because the regularization terms are flat!

Idea

Explicitly normalize the columns of the factors to minimize the penalization terms with respect to scaling.

The normalization formula for A with fixed B, C:

$$A[:,q]^* = \left(\frac{\beta}{p_A g_A(A[:,q])}\right)^{1/p_A} A[:,q]$$

with β some known constant of A, B, C.

Experiment setup

Comparing HALS (alternating algorithm) for nonnegative CPD (nnCPD) with/without scaling at each outer iteration.

Model 1: Frobenius-regularized nnCPD

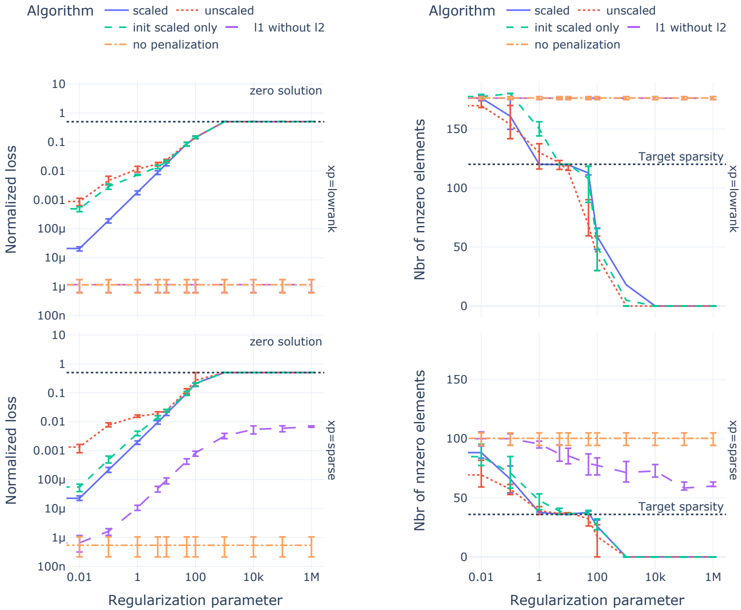
$$\underset{A,B,C\in\mathbb{R}_+^{n_i\times r}}{\operatorname{argmin}} \|\mathcal{T} - [\![A,B,C]\!]\|_F^2 + \lambda\left(\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2\right)$$

Model 2: nnCPD with sparse factor A

$$\underset{A,B,C\in\mathbb{R}_+^{n_i\times r}}{\operatorname{argmin}} \|\mathcal{T} - [\![A,B,C]\!]\|_F^2 + \lambda\left(\|A\|_1 + \|B\|_F^2 + \|C\|_F^2\right)$$

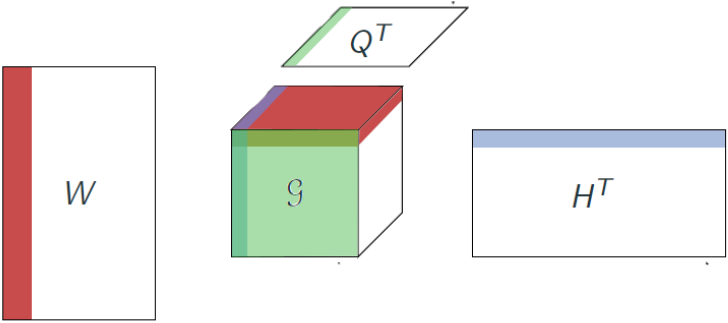
Settings: $n_i = 30$, $r = 4$, $\hat{r} = 6$, Uniform factors, 30% sparsity for ℓ_1, scaled init, grid on λ. **30 outer iterations** (early stop).

Explicit normalization effect in tensor decomposition



Work in Progress: Tucker decomposition

Adaptation for Tucker decomposition: Tensor Sinkhorn algorithm!



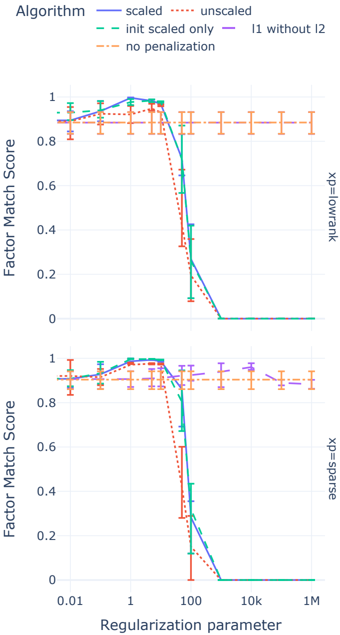
Conclusions

- Constrained Tensor factorization models are unsupervised learning techniques with interpretable outputs.
- Many practical and theoretical problems remain regarding optimization algorithms.
- I discussed the effect of scale invariance on regularized factorization problems.

Paper in progress, *French* version available for now.

Thank you for your attention! 

FMS



Algorithm
scaled
unscaled
init scaled only
l1 without l2
no penalization
Factor Match Score
1
0.8
0.6
0.4
0.2
0
xp=lowrank
xp=sparse
0.01
1
100
10k
1M
Regularization parameter