Implicit Regularization in Regularized (Nonnegative) Low-Rank Approximations

Jérémy Cohen, Valentin Leplat
CNRS, CREATIS, Lyon

arxiv:2403.18517, under review

LORAINNE Workshop, Nancy, November 2024

The exemple of double sparse NMF

- Data matrix $Y \in \mathbb{R}^{m_1 \times m_2}$
- decomposition rank $r \in \mathbb{N}^*$
- unknown factors $X_i \in \mathbb{R}_+^{m_i \times r}$, $i \in \{1, 2\}$

$$\min_{X_1 \geq 0, X_2 \geq 0} KL(Y, X_1 X_2^T) + \mu_1 \|X_1\|_1 + \mu_2 \|X_2\|_1 \qquad (1)$$

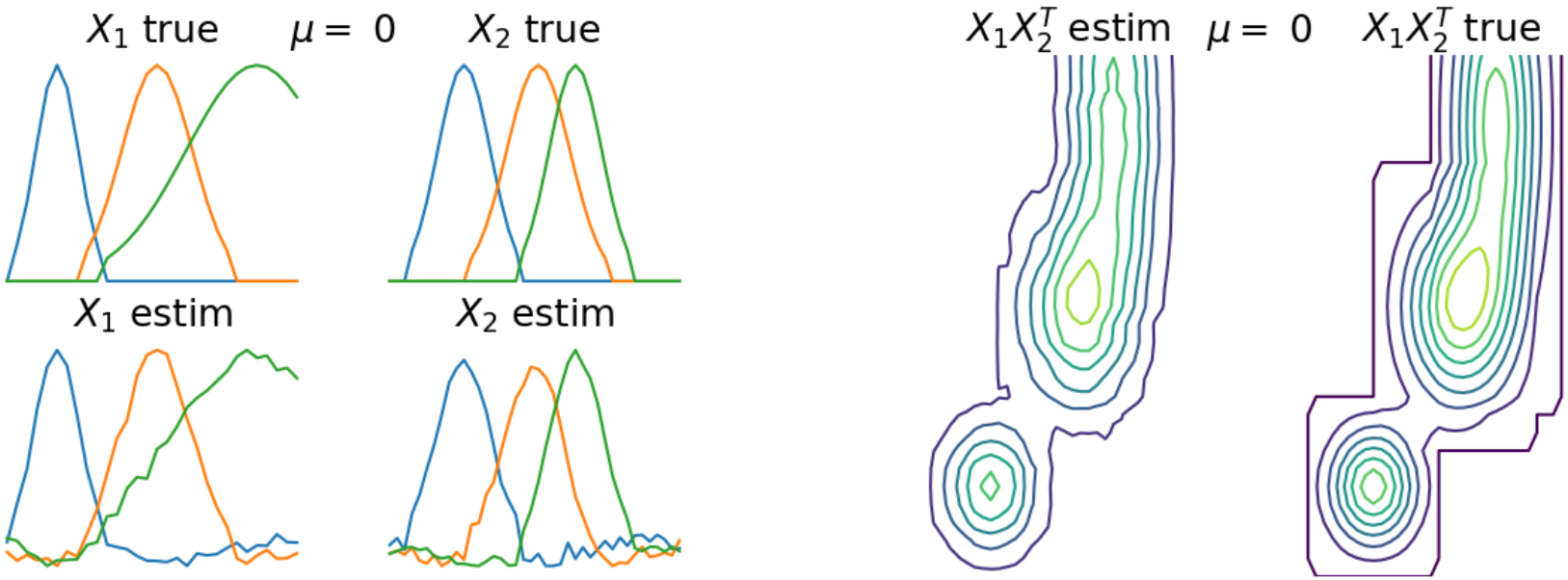


Figure: Mixture of three separable Gaussians, $\mu_1 = \mu_2 =: \mu$.

Properties of sparse NMF

No work dedicated to the characterisation of sparse NMF solutions!

We don't know how the solution behave with μ_1, μ_2.

- In which space live the solutions?
- How to choose the regularization parameters?

Some related works on dictionary learning (no nonnegativity) [Georgiev 2005, Aharon 2006, Gribonval 2015, Cohen 2018] deal with identifiability. In [Cohen 2019] we show that empirically nonnegativity can help.

Homogeneous Regularized Scale Invariant problem

We study a larger class of problems than sparse NMF, coined HRSI

$$\min_{\forall i \leq n,\, X_i \in \mathbb{R}^{m_i \times r}} f(\{X_i\}_{i \leq n}) + \sum_{i=1}^{n} \mu_i \sum_{q=1}^{r} g_i(X_i[:, q]) \tag{2}$$

- f is a **scale-invariant** differentiable cost (e.g. $\| Y - X_1 X_2^T \|_F^2$). Scaling $\{X_i\}_{i \leq n}$: $X_i \Lambda_i$ with diagonal Λ_i and $\prod_{i \leq n} \Lambda_i = I_r$.
- g_i are **homogeneous positive-definite** (e.g. ℓ_p norms) of degree p_i.
- μ_i are nonnegative regularization hyperparameters.
- X_i are the unknown factors, $X_i[:, q]$ their q-th column.

It covers sparse NMF, regularized Canonical Polyadic Decomposition, Nonnegative Tucker Decompositions...

Takewaway message

Scale invariance induces implicit penalization in HRSI.

- Better understanding of how to choose regularizations g_i, and hyperparameters μ_i.
- Better algorithms that converge faster in loss function.
- No solution characterisation yet.

Outline

1. Implicit balancing in HRSI
2. Explicit algorithmic balancing in alternating algorithms
3. Showcases on sNMF, rCPD and sNTD

Implicit balancing in Homogeneous Regularized Scale Invariant models

Our main result

Proposition [C., Leplat, in review]

Explicit HRSI

$$\min_{\forall i \leq n,\, X_i \in \mathbb{R}^{m_i \times r}} f(\{X_i\}_{i \leq n}) + \sum_{i=1}^{n} \mu_i \sum_{q=1}^{r} g_i(X_i[:, q]) \tag{3}$$

is essentially equivalent to an implicit HRSI problem

$$\min_{\forall i \leq n,\, X_i \in \mathbb{R}^{m_i \times r}} f(\{X_i\}_{i \leq n}) + \tilde{\mu} \sum_{q=1}^{r} \left(\prod_{i=1}^{n} g_i(X_i[:, q])^{\frac{1}{p_i}} \right)^{\frac{1}{\sum_{i=1}^{n} \frac{1}{p_i}}} \tag{4}$$

Also, solutions are balanced: $p_i \mu_i g_i(X_i^*[:, q])$ is constant wrt. i.

- Implicit HRSI is fully scale-invariant.
- The nature of the regularization can change from explicit to implicit HRSI.
- Only an average $\tilde{\mu}$ of parameters μ_i matters!!

Proof sketch

Solutions to the scaling problem of HRSI

$$\min_{\forall i \leq n,\, \Lambda_i \in \mathbb{R}_+^{r \times r}} f(\{X_i \Lambda_i\}_{i \leq n}) + \sum_{i=1}^{n} \mu_i \sum_{q=1}^{r} g_i(\Lambda_i[q,q] X_i[:,q]) \tag{5}$$

where Λ_i are diagonal matrices such that $\prod_{i \leq n} \Lambda_i = I_r$ are given by

$$X_i^{(s)}[:,q] = \underbrace{\left(\frac{\beta_q}{p_i \mu_i g_i(X_i[:,q])} \right)^{1/p_i}}_{\Lambda_i[q,q]^*} X_i[:,q] \tag{6}$$

where β_q is the geometric mean of $\{p_i \mu_i g_i(X_i)[:,q], \frac{1}{p_i}\}_{i \leq n}$.
Injecting this in HRSI yields an implicit formulation of HRSI:

$$\min_{X_i^{(s)} \in \mathbb{R}^{m_i \times r}} f(\{X_i^{(s)}\}_{i \leq n}) + \tilde{\mu} \sum_{q=1}^{r} \left(\prod_{i=1}^{n} g_i(X_i^{(s)}[:,q])^{\frac{1}{p_i}} \right)^{\frac{1}{\sum_{i=1}^{n} \frac{1}{p_i}}} \tag{7}$$

with $\tilde{\mu}$ averaged from $\{\mu_i\}_{i \leq n}$.

Implicit regularization, ridge matrix

It is known [Srebro 2008] that

$$\underset{X_1, X_2 \in \mathbb{R}^{n_i \times r}}{\operatorname{argmin}} \; \| Y - X_1 X_2^T \|_F^2 + \lambda \left(\|X_1\|_F^2 + \|X_2\|_F^2 \right) \tag{8}$$

has essentially the same solutions as, setting $L = X_1 X_2^T$,

$$\underset{L \in \mathbb{R}^{n_1 \times n_2}, \operatorname{rank}(L) \leq r}{\operatorname{argmin}} \; \| Y - L \|_F^2 + \alpha \|L\|_* \tag{9}$$

From our result, we get the implicit formulation

$$\underset{\operatorname{rank}(L_q)=1}{\operatorname{argmin}} \; \| Y - L \|_F^2 + \alpha \sum_{q \leq r} \|L_q\|_F . \tag{10}$$

with $L = \sum_q L_q = \sum_q X_1[:, q] \otimes X_2[:, q]^T$.

ℓ_2 regularization in explicit HRSI induces low-rank solutions!

Mentionned in [Uschmajew 2012] for CP decomposition.

Implicit regularization, sparse NMF

$$\min_{X_1 \geq 0, X_2 \geq 0} KL(Y, X_1 X_2^T) + \mu_1 \|X_1\|_1 + \mu_2 \|X_2\|_1 \qquad (11)$$

The implicit HRSI model for sNMF writes

$$\min_{L_q \in \mathbb{R}_+^{m_1 \times m_2},\ \mathrm{rank}(L_q) \leq 1} KL(Y, \sum_{q=1}^{r} L_q) + \frac{\sqrt{\mu_1 \mu_2}}{2} \sum_{q=1}^{r} \sqrt{\|L_q\|_1}. \qquad (12)$$

- The individual sparsity levels μ_i don't matter?!
- Sparsity occurs at the level of rank-one components.

Open question

Can we use this implicit formulation to characterise solutions?

Already investigated in [Papalexakis 2013].

Explicit balancing in alternating algorithms

Alternating optimization is slow?

Consider the toy problem with $y \in \mathbb{R}$, and $0 < \lambda \leq y$.

$$\min_{x_1 \in \mathbb{R}, x_2 \in \mathbb{R}} f(x_1, x_2) \text{ where } f(x_1, x_2) = (y - x_1 x_2)^2 + \lambda(x_1^2 + x_2^2)\,. \quad (13)$$

Solutions are balanced, $x_i^* = \sqrt{y - \lambda}$

The Alternating Least Squares algorithm

$$\begin{cases} x_1^{(k+1)} = \dfrac{x_2^{(k)} y}{x_2^{2(k)} + \lambda} \\ x_2^{(k+1)} = \dfrac{x_1^{(k+1)} y}{x_1^{2(k+1)} + \lambda} \end{cases} \quad (14)$$

is provably slow!

Proposition [C. Leplat, in review]

For k large enough and $\lambda \ll y$,

$$\frac{x_1^{(k+1)} - \sqrt{y - \lambda}}{x_1^{(k)} - \sqrt{y - \lambda}} \approx 1 - 4\frac{\lambda}{y}. \quad (15)$$

Alternating optimization is slow? (2)

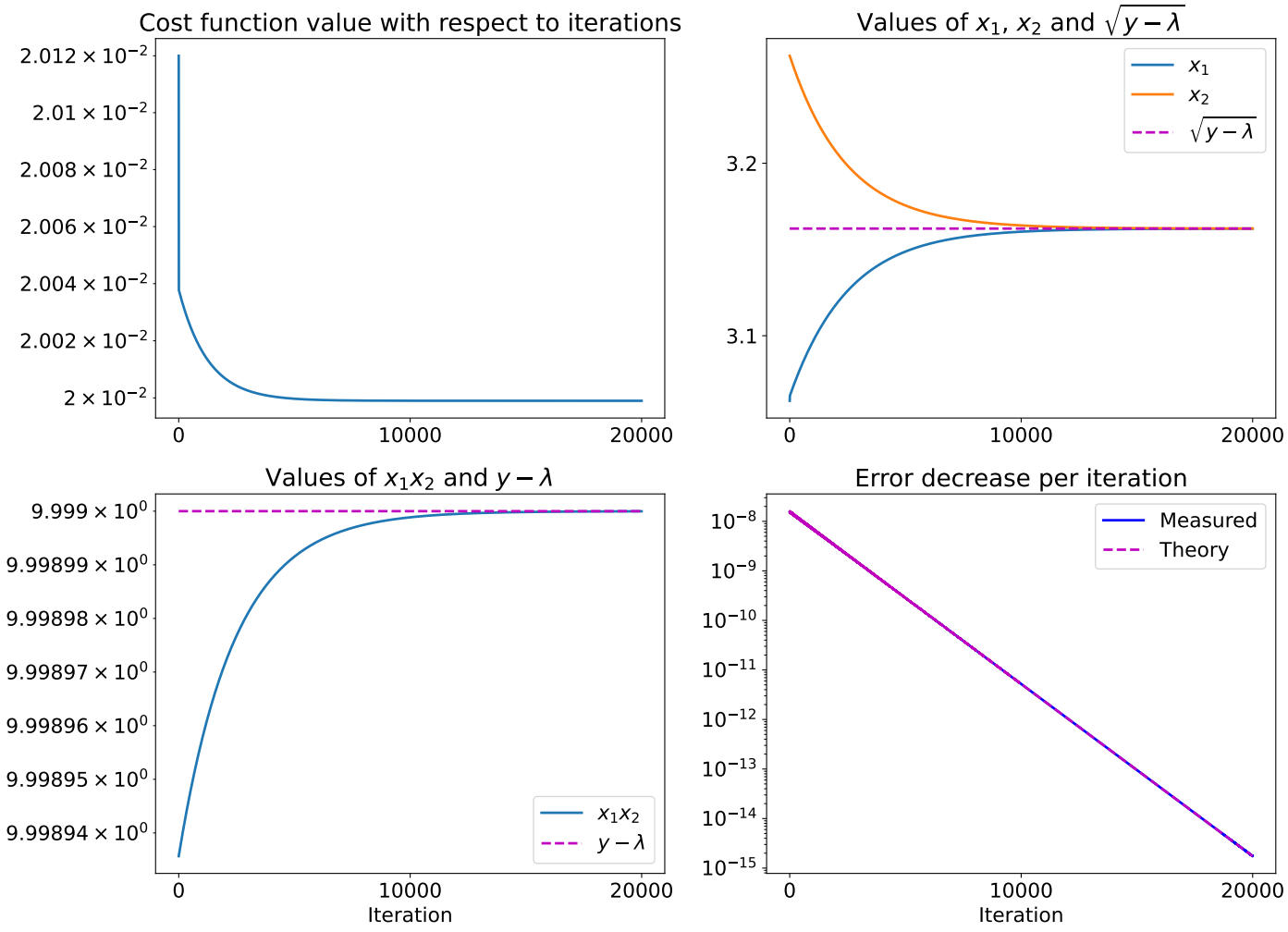


Figure: $y = 10$ and $\lambda = 10^{-3}$.

Balancing in practice

Observation: An alternating minimization algorithm can be extremely slow to converge to balanced solutions.

Idea

Explicitly normalize factors to minize the loss wrt. scalings.

The normalization for X_i is of the form

$$X_i[:, q]^* = \left(\frac{\beta}{p_i g_i(X_i[:, q])} \right)^{1/p_i} X_i[:, q]. \tag{16}$$

where β is the weighted geometric mean of the regularizations. We can perform this operation for each parameter matrix at the end of each outer loop of e.g. MU for sNMF.

A meta algorithm for regularized LRA

We propose a generic alternating Majorization Minimization algorithm with balancing, that converges in cost and iterates.

% Main optimization loop:
for $k = 1 : \text{maxiter}$ **do**
  for $i = 1 : n$ and $q = 1 : r$ **do**
    $X_i[:, q]^{(k+1)} \leftarrow \text{argmin}_{x \in \mathcal{X}_i} \bar{\phi}(x | X_i[:, q]^{(k)})$
  end for
  $\{X_i^{(k+1)}[:, q]\}_{i \leq n}$ balanced using (16) for all $q \leq r$
end for
return $\{X_i[:, q]\}_{i \leq n}$

Bonus contribution

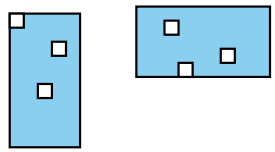
We derive (closed-form) updates for ℓ_p^p regularizations with beta-divergence loss. E.g. for KL divergence and $p = 2$,

$$x[k] \leftarrow \frac{\sqrt{c^2 + 8\mu x[k]t} - c}{4\mu} \tag{17}$$

with $\mu > 0$ and c, t depend on data and X_i.

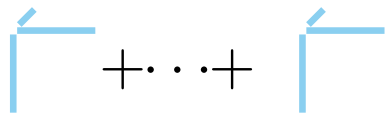
Showcases: sNMF, rCPD, sNTD

Explicit Models



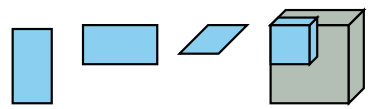
double sparse NMF (sNMF)

$$\min_{X_1 \in \mathbb{R}_+^{m_1 \times r}, X_2 \in \mathbb{R}_+^{m_2 \times r}} KL(Y, X_1 X_2^T) + \mu_1 \|X_1\|_1 + \mu_2 \|X_2\|_1$$



ridge Canonical Polyadic Decomposition (rCPD)

$$\min_{X_i \in \mathbb{R}_+^{m_i \times r}} \| Y - I_r \times_1 X_1 \times_2 X_2 \times_3 X_3 \|_F^2 + \mu \left(\|X_1\|_F^2 + \|X_2\|_F^2 + \|X_3\|_F^2 \right)$$



sparse Nonnegative Tucker Decomposition (sNTD)

in the paper!

Implicit equivalent models

implicit sNMF

$$\min_{L_q \in \mathbb{R}_+^{m_1 \times m_2},\ \mathrm{rank}(L_q) \leq 1} KL(Y, \sum_{q=1}^{r} L_q) + 2\sqrt{\mu_1 \mu_2} \sum_{q=1}^{r} \sqrt{\|L_q\|_1}. \quad (18)$$

- Empirically, tuning μ_1 vs μ_2 has an effect, right?
- Does explicit balancing in a MU algorithm really help?

implicit rCPD with $L_q = X_1[:, q] \otimes X_2[:, q] \otimes X_3[:, q]$

$$\min_{\{L_q\}_{1 \leq r},\ \mathrm{rank}(L_q) \leq 1} \|Y - \sum_{q=1}^{r} L_q\|_F^2 + 3\mu \sum_{q=1}^{r} \|L_q\|_F^{\frac{2}{3}}. \quad (19)$$

- Empirically, do we observe a biais towards low-rank solutions?
- Does explicit balancing in a HALS algorithm really help?

Experimental Setup

XP	**sNMF**	**rCPD**
sizes $(n_i, r, \hat{r})$	(30, 4, 4)	(30, 4, 6)
data generation	$X_i \sim \mathcal{P}(\alpha X_1 X_2^T)$	$X_i \sim \mathcal{U}[0,1]$
factors sparsity	30%	None
SNR	40	40
epsilon	$1e-16$	$1e-16$

sNMF: Two cases $\mu_1 = \mu_2$ and $\mu_1 = 1$

Both: Comparing balancing, no balancing and balancing at initialization only.

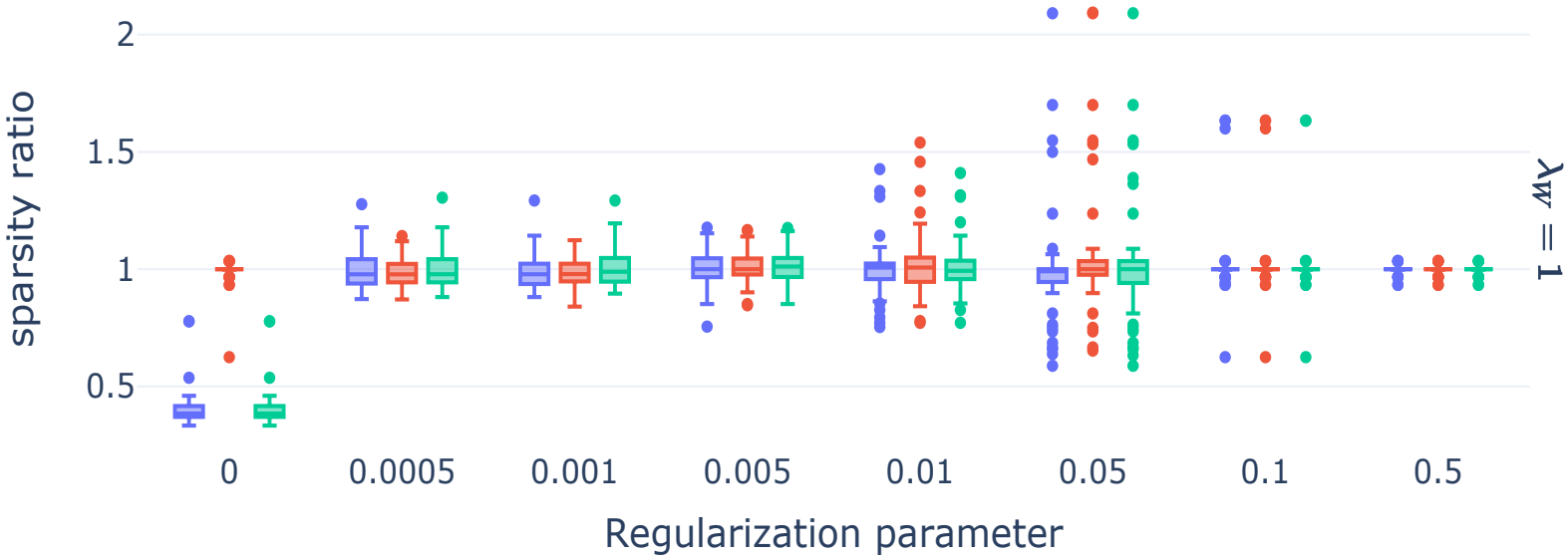
Evaluation with loss function, sparsity and Factor Match Score

$$\text{Tr}\left(\prod_i \hat{X}_i^T X_i\right) \tag{20}$$

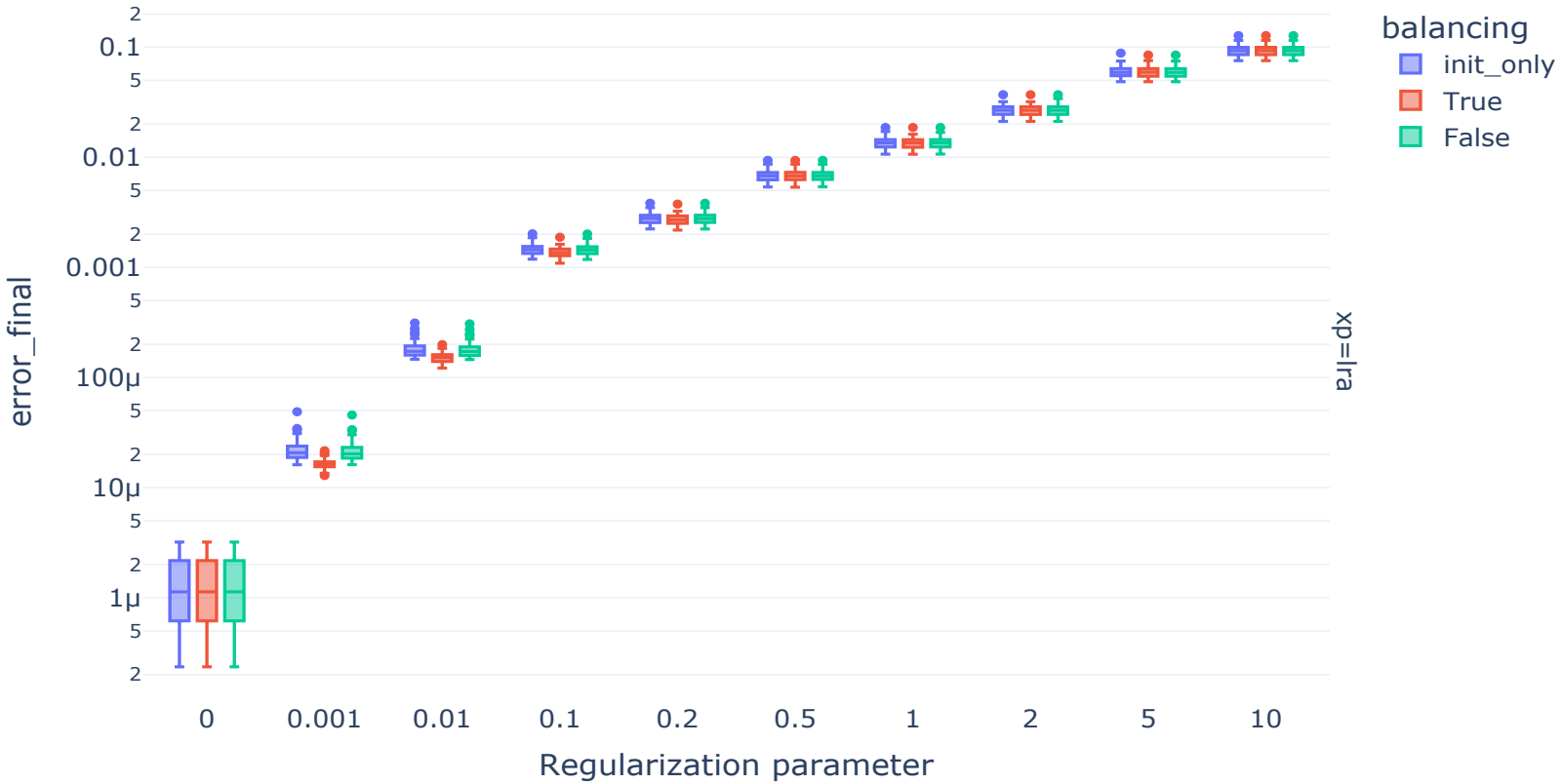
(after columnwise normalization and permutation).

Results for sNMF

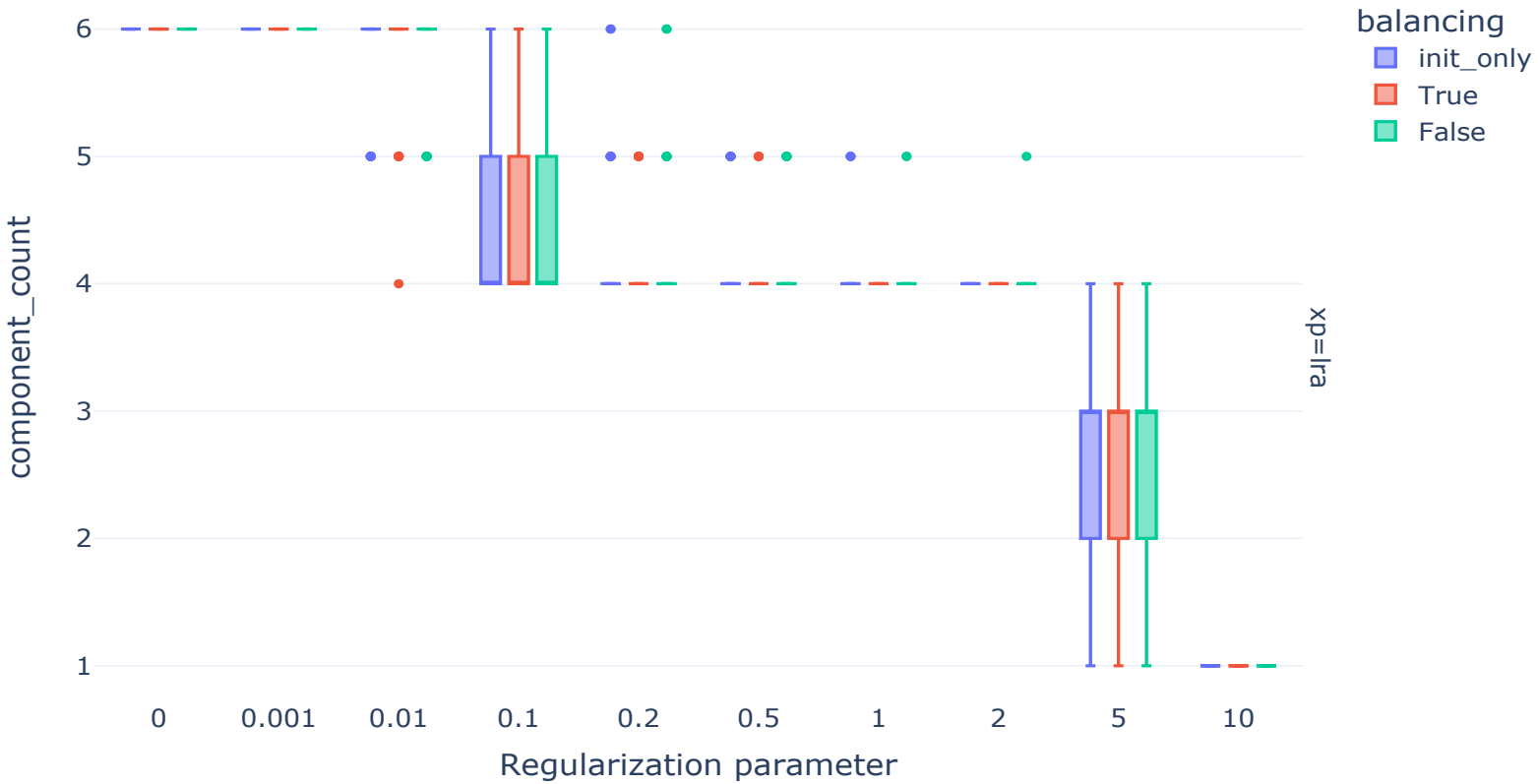
Results for sNMF



Results for rCPD



Results for rCPD



Conclusions

Take-aways

- Explicit regularizations may behave unexpectedly because of scale invariance.
- Don't tune the hyperparameters independently!
- Balance your solutions in the algorithm, or at least at first and last iteration.

Perspectives

- What about non-homogeneous/non positive definite regularizations?

Thank you for your attention!

Updates for sNMF

MU factor update:

$$\hat{X}_1 = \max\left(X_1 \odot \frac{X_2 \frac{M}{X_2^T X_1}}{\mu_1 e_{m_1 \times r} + e_{m_1} \otimes X_2^T e_{m_2}}, \epsilon\right) \tag{21}$$

$$\hat{X}_2 = \max\left(X_2 \odot \frac{\frac{M^T}{X_1 X_2^T} X_1^T}{\mu_2 e_{m_2 \times r} + e_{m_2} \otimes X_1 e_r}, \epsilon\right) \tag{22}$$

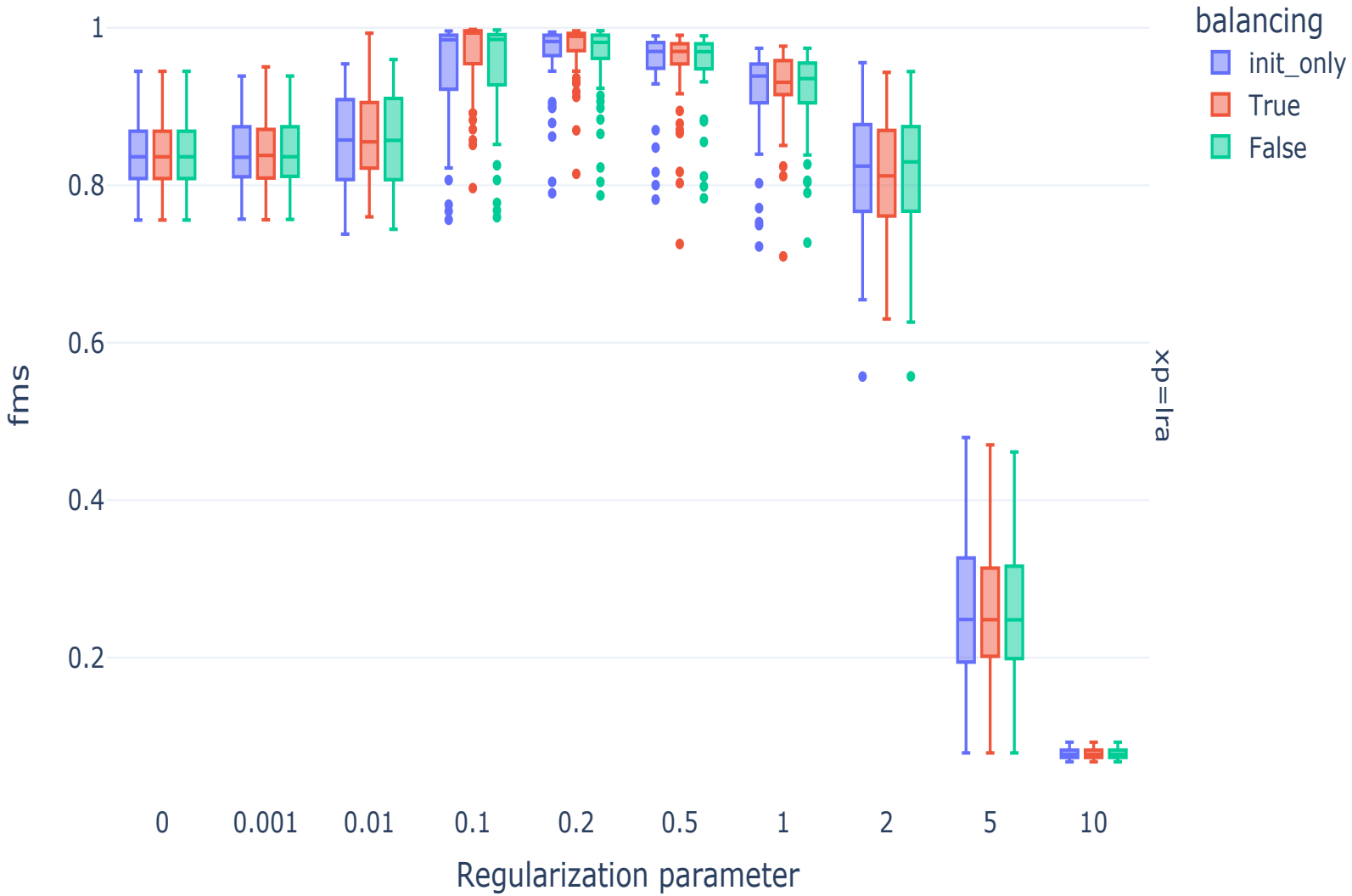
Balancing:

$$X_1[:, q] \leftarrow \frac{\beta_q}{\mu_1 \|X_1[:, q]\|_1} X_1[:, q] \tag{23}$$

$$X_2[:, q] \leftarrow \frac{\beta_q}{\mu_2 \|X_2[:, q]\|_1} X_2[:, q] \tag{24}$$

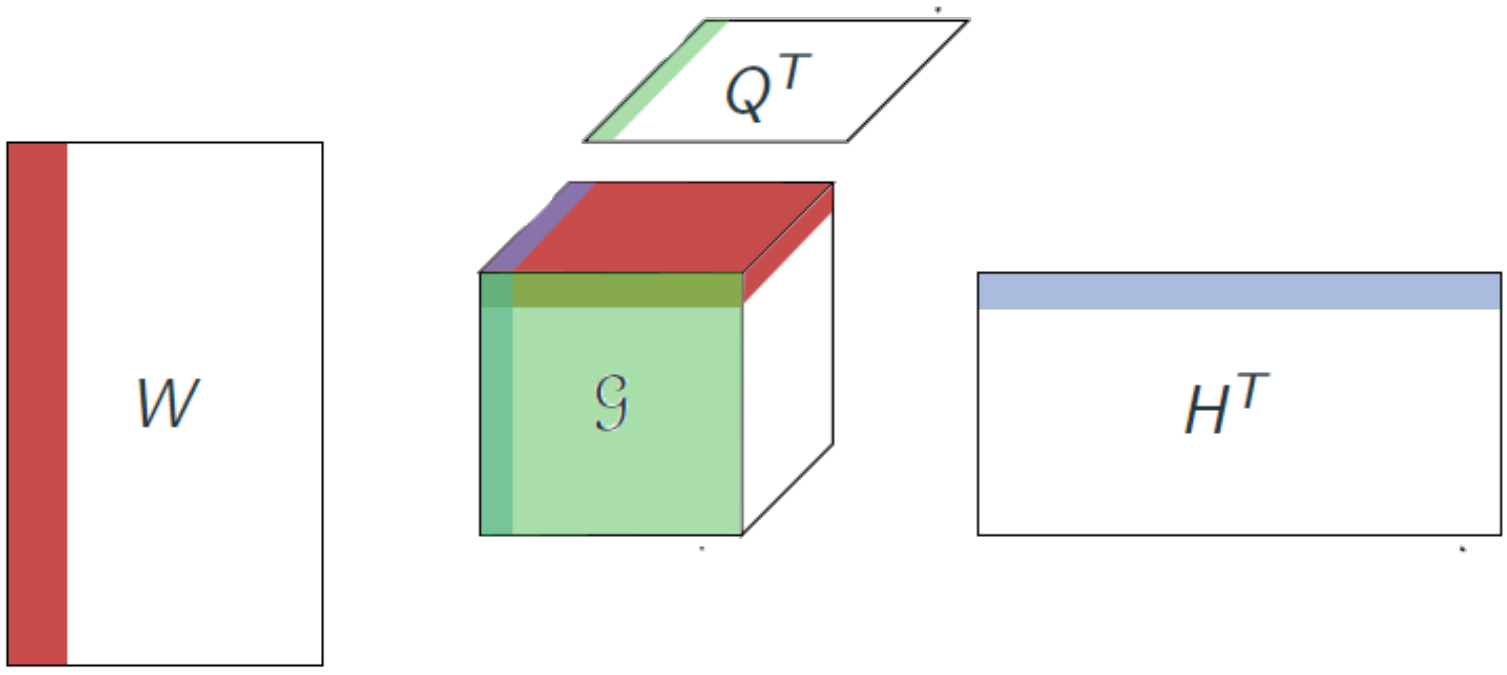
where $\beta_q = \sqrt{\mu_1 \mu_2 \|X_1[:, q]\|_1 \|X_2[:, q]\|_1}$.

Results for rCPD



Sparse Nonnegative Tucker is harder

NTD does not fit HRSI because scaling ambiguity is not separable.



Two possible scalings

Scalar scaling

$$\min_{\substack{w\geq 0, h\geq 0,\\ q\geq 0, g\geq 0}} f(\{w,h,q,g\}) + \mu(\|g\|_1 + \|w\|_F^2 + \|h\|_F^2 + \|q\|_F^2) \quad (25)$$

where $w = \mathrm{vec}(W)$, $h = \mathrm{vec}(H)$, $q = \mathrm{vec}(Q)$, $g = \mathrm{vec}(\mathcal{G})$.

Sinkhorn scaling

$$\underset{\Lambda_W \text{ diagonal}}{\operatorname{argmin}} \ \|\mathcal{G} \times_1 \Lambda_W \times_2 \Lambda_H \times_3 \Lambda_Q\|_1 + \|W\Lambda_W^{-1}|_F^2 \quad (26)$$

$$\underset{\Lambda_H \text{ diagonal}}{\operatorname{argmin}} \ \|\mathcal{G} \times_1 \Lambda_W \times_2 \Lambda_H \times_3 \Lambda_Q\|_1 + \|H\Lambda_H^{-1}\|_F^2 \quad (27)$$

$$\underset{\Lambda_Q \text{ diagonal}}{\operatorname{argmin}} \ \|\mathcal{G} \times_1 \Lambda_W \times_2 \Lambda_H \times_3 \Lambda_Q\|_1 + \|Q\Lambda_Q^{-1}\|_F^2 \ . \quad (28)$$

Ridge regularized factors estimation with KL

The problem

$$\underset{W \geq 0}{\operatorname{argmin}}\ KL(V|WU) + \mu \|W\|_F^2 \tag{29}$$

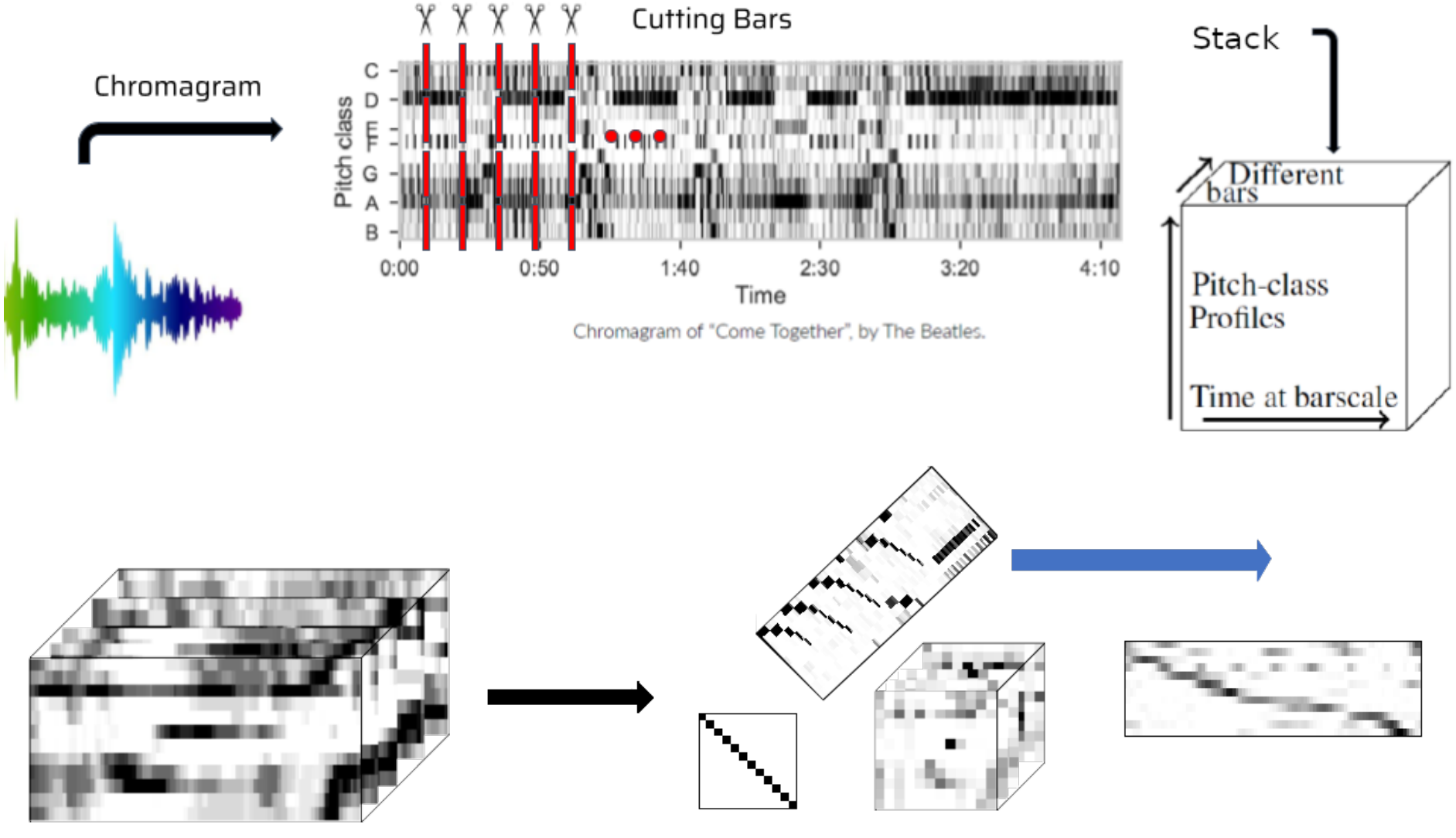
can be solved by iterating

$$\hat{W} = \max\left(\frac{\left[C^{\cdot 2} + S\right]^{\cdot \frac{1}{2}} - C}{2\mu}, \epsilon\right) \tag{30}$$

where $C = EU^T$ with E is a all-one matrix of size m_1-by-$m_2 m_3$ and $S = 4\mu\tilde{W} \odot \left(\frac{[V]}{[\tilde{W}U]} U^T\right)$.

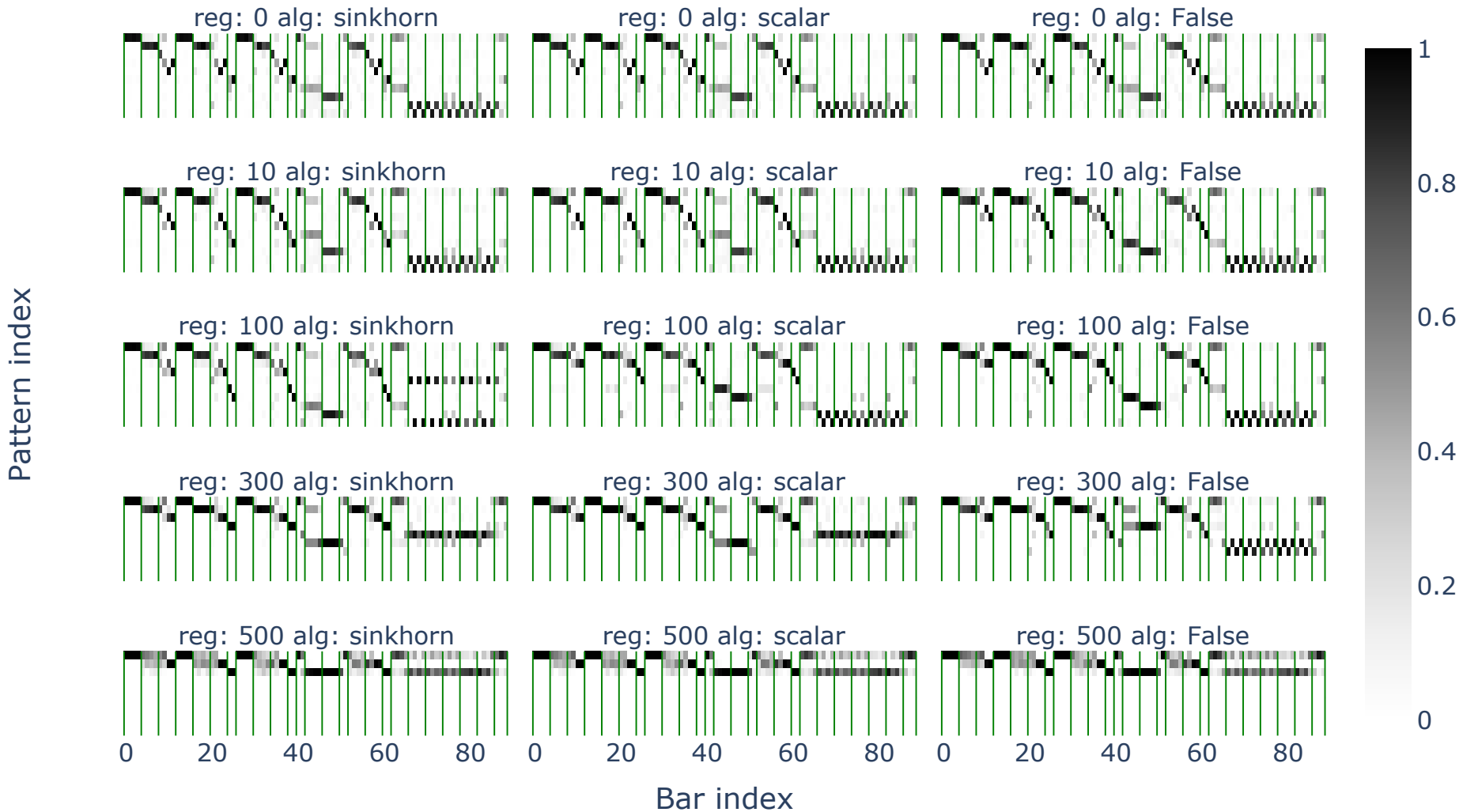
An audio redundancy detection experiment

We perform sNTD on a tensor spectrogramm, factor Q holds the song arrangement.



Note: sparsity is imposed on Q not on the core.

An audio redundancy detection experiment



Optimizing scale in HRSI

$$\inf_{\forall i \leq n,\, X_i \in \mathbb{R}^{m_i \times r}} \phi(\{X_i\}_{i \leq n}) = \inf_{\forall i \leq n,\, X_i \in \mathbb{R}^{m_i \times r},\, \prod_{i \leq n} \Lambda_i = 1} \phi(\{X_i \Lambda_i\}_{i \leq n})$$

with ϕ the HRSI cost function.

- The loss in HRSI is separable with respect to scales of factors. We consider the case $r = 1$ wlog. $(X_i \to x_i)$
- Since f is scale invariant, scaling the factors, i.e.

$$\prod_{i \leq n} \min_{\lambda_i = 1,\, \lambda_i > 0} f(\{\lambda_i x_i\}_{i \leq n}) + \sum_{i \leq n} \mu_i g_i(\lambda_i x_i) \tag{31}$$

  for fixed values of $\{x_i\}_{i \leq n}$ means finding the optimal scales to minimize the penalties

$$\prod_{i \leq n} \min_{\lambda_i = 1,\, \lambda_i > 0} \sum_{i \leq n} \lambda_i^{p_i} \underbrace{\mu_i g_i(x_i)}_{a_i} \tag{32}$$

  where p_i is the homogeneity degree of g_i.

A geometric mean identity

$$\min_{\forall i \leq n,\ \lambda_i > 0} \sum_{i \leq n} \lambda_i \text{ tel que } \prod_{i \leq n} \lambda_i = p \tag{33}$$

for $p \geq 0$ is solved uniquely by $\lambda_i^* = p^{1/n}$ for all $i \leq n$.

We can prove similarly that

$$\min_{\forall i \leq n,\ \lambda_i \geq 0} \sum_{i=1}^{n} \lambda_i^{p_i} a_i \text{ such that } \prod_{i=1}^{n} \lambda_i = 1 \tag{34}$$

has solutions

$$\lambda_i^* = \frac{\beta}{p_i a_i} \tag{35}$$

where β is the geometric mean of $\{p_i a_i, \frac{1}{p_i}\}_{i \leq n}$

$$\beta = \left(\prod_{i \leq n} (p_i a_i)^{\frac{1}{p_i}} \right)^{\frac{1}{\sum_{i \leq n} \frac{1}{p_i}}}. \tag{36}$$