Nonnegative and low rank approximations

Jeremy E. Cohen

IRISA, Rennes, France

30 November 2020

I. INTRODUCTION

Our fil rouge

Let $y \in \mathbb{R}^3_+$ a color in RBG

Let $A \in \mathbb{R}^{3 \times d}_+$ a collection of paint pots

We can perform conical combinations of colors

$$0.5 \begin{pmatrix} 250 \\ 207 \\ 176 \end{pmatrix} + 0.5 \begin{pmatrix} 255 \\ 140 \\ 102 \end{pmatrix} = \begin{pmatrix} 252.5 \\ 173.5 \\ 139 \end{pmatrix}$$

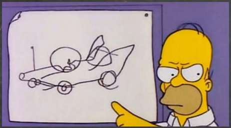
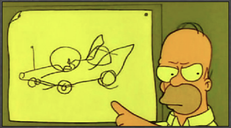
Any $\sum_i \alpha_i y_i$ with $0 \leq \alpha_i$ and $\sum_i \alpha_i y_i \leq 255$ is a color.

Our fil rouge

Set $d(y, \hat{y}) = \|y - \hat{y}\|_2^2$ as the loss.

Problem 1: paint color y as well as possible using paint pots A.

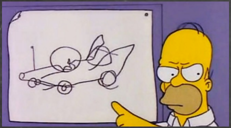
$$\text{Find } x \in \mathbb{R}_+^d \text{ such that } d(y, Ax) \text{ is minimal}$$

 $\approx \ \forall_i, \ x_{1i}\bullet + x_{2i}\bullet =$ 

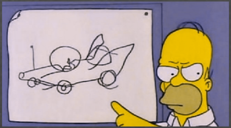
Our fil rouge

Problem 2: given a painting $\{y_i\}_{i \leq n}$, find its closest 2-color version.

Find $A \in \mathbb{R}_+^{3\times 2}$ and $x_i \in \mathbb{R}_+^2$ such that $\forall i \leq n, y_i \approx Ax_i$



$\approx \ \forall_i, \ x_{1i}\bullet + x_{2i}\bullet =$



A quick quizz!

Visit https://www.wooclap.com/ITWISTQ1

Importance of regularization

Problem 1: if $A \in \mathbb{R}^{3\times d}$ with $d \gg 3$, then

$$\min_{x\in\mathbb{R}^d} \|y - Ax\|_2^2$$

Importance of regularization

Problem 1: if $A \in \mathbb{R}^{3\times d}$ with $d \gg 3$, then

$$\min_{x\in\mathbb{R}^d} \|y - Ax\|_2^2$$

has infinitely many solution, $x_0^* + z$ with $z \in \mathrm{Ker}(A)$. Most (all?) of them are bad because of negative coefficients.

Importance of regularization

Problem 1: if $A \in \mathbb{R}^{3\times d}$ with $d \gg 3$, then

$$\min_{x\in\mathbb{R}^d} \|y - Ax\|_2^2$$

has infinitely many solution, $x_0^* + z$ with $z \in \mathrm{Ker}(A)$. Most (all?) of them are bad because of negative coefficients.

Problem 2: without nonnegativity, then

$$\min_{A\in\mathbb{R}^{3\times r},\, x_i\in\mathbb{R}^r} \sum_i \|y_i - Ax_i\|_2^2$$

Importance of regularization

Problem 1: if $A \in \mathbb{R}^{3\times d}$ with $d \gg 3$, then

$$\min_{x\in\mathbb{R}^d} \|y - Ax\|_2^2$$

has infinitely many solution, $x_0^* + z$ with $z \in \mathrm{Ker}(A)$. Most (all?) of them are bad because of negative coefficients.

Problem 2: without nonnegativity, then

$$\min_{A\in\mathbb{R}^{3\times r},\, x_i\in\mathbb{R}^r} \sum_i \|y_i - Ax_i\|_2^2$$

has again infinitely many bad (negative) solutions, even for $r = 2$ using for instance the truncated SVD of $Y = [y_1, \dots, y_n]$.

Outline

- Nonnegative Least Squares
  - Theory
  - Algorithms
- Matrix and tensor rank
  - Matrix rank
  - Nonnegative rank
  - Tensor (nonnegative) rank

- Nn. Matrix Factorization
  - Theory
  - Algorithms
  - Applications
- Nn. Tensor Factorization
  - Algorithms
  - Application

II. Nonnegative Least Squares

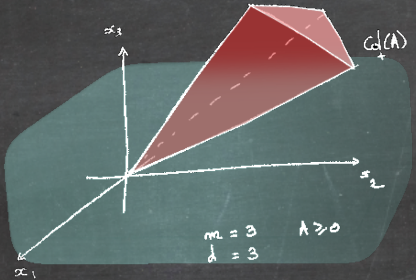
Cones

Definition:
For a given matrix $A \in \mathbb{R}^{m \times d}$, let $\text{col}_+(A) = \{Ax \mid x \geq 0\}$.

Proposition:
For any matrix A, the set $\text{col}_+(A)$ is a convex cone, *i.e.*

$$\lambda_1 x_1 + \lambda_2 x_2 \in \text{col}_+(A)$$

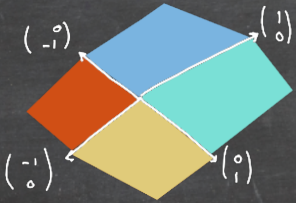
if $x_1, x_2 \in \text{col}_+(A)$ and $\lambda_1, \lambda_2 \geq 0$.



Cones

The cone $\text{col}_+(A)$ may not be have low-dimensional facets.

$$\text{col}_+\left(\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}\right) = \mathbb{R}^2$$



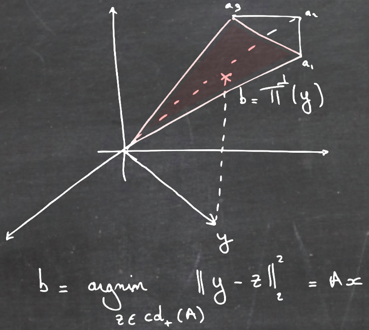
(but it's all fine when $A \geq 0$)

NNLS formulation

Definition:

The NNLS problem is equivalently defined as

1. Find $x \in \underset{x \in \mathbb{R}_+^d}{\operatorname{argmin}} \|y - Ax\|_2^2$
2. Find $b \in \text{col}_+(A)$ and $x \in \mathbb{R}_+^d$ s.t. $b = Ax, \; b = \Pi^{\perp}_{\text{col}_+(A)}(y)$



A few questions

Existence of solutions?

Uniqueness of b, of x?

Properties of a solution x?

b exists and is unique

Proposition:
For any $A \in \mathbb{R}^{m\times d}$, the map

$$y \to \Pi^{\perp}_{\mathrm{col}_{+}(A)}(y)$$

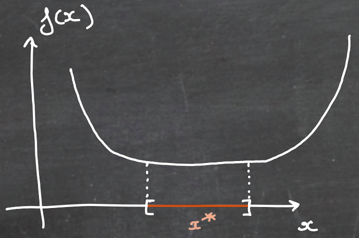
is well defined.

Proof idea:
The map $f : z \to \|y - z\|_2^2$ is coercive and continuous. Because $\mathrm{col}_{+} A$ is closed, f must attain its minimum value on $\mathrm{col}_{+}(A)$. Further, f strongly convex in $\mathbb{R}^{m\times d}$, thus in particular on its restriction to the convex set $\mathrm{col}_{+}(A)$. Strongly convex functions admit unique global minimizers when they exist.

NNLS: convexity

$$\underset{x \geq 0}{\operatorname{argmin}} \|y - Ax\|_2^2 \qquad \text{(NNLS)}$$

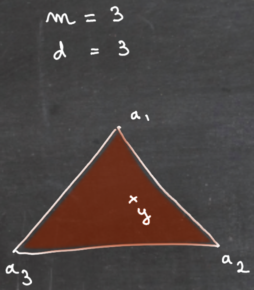
Problem (NNLS) is convex but not strictly convex unless A is fcr. Therefore, there does not exist a unique solution x in general.

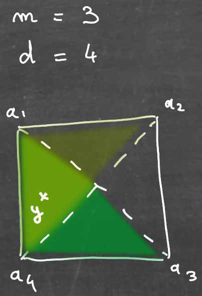


x uniqueness: exact case (interior)

Suppose that y lies in the interior of $\mathrm{col}_+(A)$. Then

- the projection b is y itself and $y = Ax$ always exists,
- if $d > m$, there is little hope for uniqueness.
- if $d \leq m$ and A is full column rank, then x is unique.

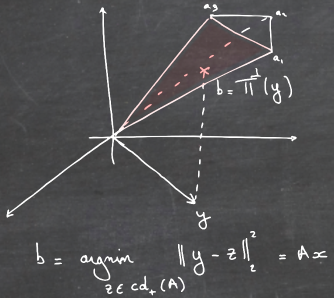




x uniqueness: exact case (border)

Informally, if y belongs to a facet of $\text{col}_+(A)$, then there exist k s.t.

$$y = Ax,\ x \geq 0,\ \|x\|_0 \leq k < m$$



Quite unlikely in practice, and similar to the approximate case. See [Donoho2005]

Illustration on Problem 1

Problem 1: paint color y as well as possible using paint pots A.

So far,

- There is always a best color approximation of y with pots A.
- When A has more than 3 colors, if $y \in \mathrm{col}_+(A)$, in general there are several solutions.

Illustration on Problem 1

Problem 1: paint color y as well as possible using paint pots A.

So far,

- There is always a best color approximation of y with pots A.
- When A has more than 3 colors, if $y \in \text{col}_+(A)$, in general there are several solutions.

What about when $y \notin \text{col}_+(A)$?

Approximate case: the main result

Theorem (Night Sky Theorem [Byrne 1981]):
Suppose that

$$y \notin \text{col}_+(A), \ \text{spark}(A) > m.$$

Then there is a unique solution the NNLS problem, which has most $m - 1$ nonzeros.

A detour by KKT

For a convex problem

$$\min_{x \in \mathbb{R}^d} f(x), \text{ s.t. } g(x) \leq 0, \text{ f, g convex}$$

with an admissible solution, considering

$$\mathcal{L}(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle,$$

x^* is a solution iff there exist λ^* s.t.

$$g(x^*) \leq 0, \quad \lambda^* \geq 0, \quad \forall i \leq d, \lambda_i^* g_i(x_i^*) = 0$$

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$$

Back to approximate NNLS

$$\nabla_x \|y - Ax\|_2^2 = 2A^T(Ax - y)$$

The KKT conditions are

$$x^* \geq 0, \quad \lambda^* \geq 0, \quad \lambda_i^* x_i^* = 0$$

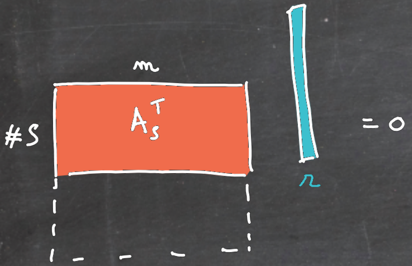
$$2A^T(Ax^* - y) - \lambda^* = 0$$

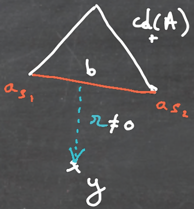
In particular, when $x_i^* > 0, \lambda_i^* = 0$, thus on the support S of x^*,

$$A_S^T(Ax^* - y) = 0$$

A marvelous equation

$$A_S^T(Ax^* - y) = 0 \iff A_S^T r = 0,\ r = y - b^*$$





As long as $r \neq 0$ and any A_S is fcr,

- $\#S < m$
- For any $i \in S$, $a_i \in \text{col}_\perp(r) := \mathcal{H}$

End of proof and computation of x^*

Any solution has its support in $S^* = \{i \leq d,\ a_i \in \mathcal{H}\}$.
Moreover, the linear system

$$A_{S^*} z = b$$

has a unique solution for fcr A_{S^*}.
Consequently, once the support of a solution S^* is known, within the hypotheses of the Night Sky Theorem, the unique solution is obtained by

$$x^* = A_{S^*}^{\dagger} y$$

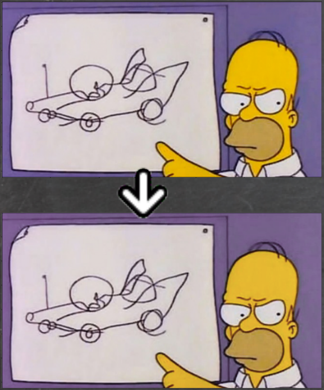
where $A^{\dagger}$ is the pseudo-inverse of A_{S^*}.

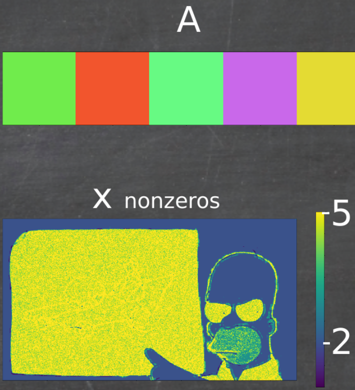
Illustration on Problem 1

Problem 1: paint color y as well as possible using paint pots A.

- There is always a best color approximation of y with pots A.
- When A has more than 3 colors, if $y \in \mathrm{col}_+(A)$, in general there are several solutions.
- If $y \notin \mathrm{col}_+(A)$, with high probability, there is a unique solution.

Illustration on Homer





How to solve NNLS??

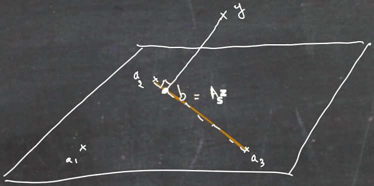
Active set in NNLS

Proposition: (admitted for exact case)
Any NNLS problem has a solution x with at most m non-zeros.

If we know the support S of that solution, then

$$\underset{z \in \mathbb{R}^{\#S}}{\operatorname{argmin}} \|y - A_S z\|_2^2$$

is solved in closed form and yields the solution (KKT).



The LH active set algorithm

Idea: (Lawson and Hanson (1974)

1. Start with empty support S
2. Add a columns of A greedily to S
3. Compute the projection on col(A_S)
4. Stop if KKT conditions are met
5. If projection has negative coefficients, move along the update until no negatives are left
6. return to 2)

The LH active set algorithm

Idea: (Lawson and Hanson (1974)

1. Start with empty support S
2. Add a columns of A greedily to S
3. Compute the projection on col(A_S)
4. Stop if KKT conditions are met
5. If projection has negative coefficients, move along the update until no negatives are left
6. return to 2)

The selection rule

$$S \leftarrow S \cup \operatorname*{argmax}_{j \notin S} \langle a_j, r \rangle$$

where $r = y - \Pi^{\perp}_{A_S}(y) =: y - Ax^S$

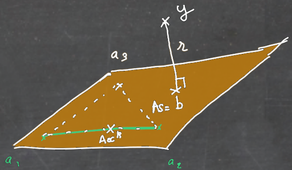
Recall KKT conditions

$$\lambda_j^* \geq 0, \quad 2\langle a_j, y - Ax^* \rangle = -\lambda_j^*$$

The column with "most negative" Lagrange multiplier is chosen.

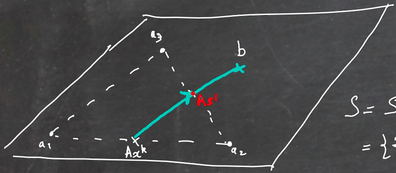
The error $\min_z \|y - A_S z\|_2^2$ can only go down in this step.

The backward step



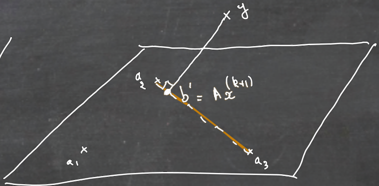
2/ $b \notin \mathrm{col}_+(A_S)$, $As' \in \mathrm{col}_+(A_{\{2,3\}})$ 1/

iteration k $\quad S = \{1,2\} \cup \{3\}$



$S = S \setminus \{1\}$
$= \{2,3\}$

3/



$t \mapsto \| y - A(x^k + t(s - x^k)) \|_2^2$

is decreasing strictly by strong convexity

Stop since $b' \in \mathrm{col}_+(A_S)$

and $\| y - b' \|_2^2 \leq \| y - As' \|_2^2$

AS algorithm pros and cons

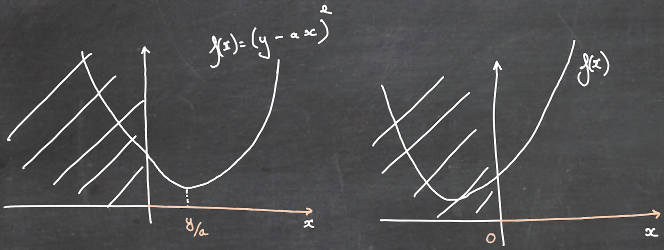
- Finite number of iterations
- Fast if warm start
- Early stop

- May test all supports
- Cold start is often slow
- No matrix version

A Block-Coordinate algorithm

Observation: The scalar problem is solved in closed form

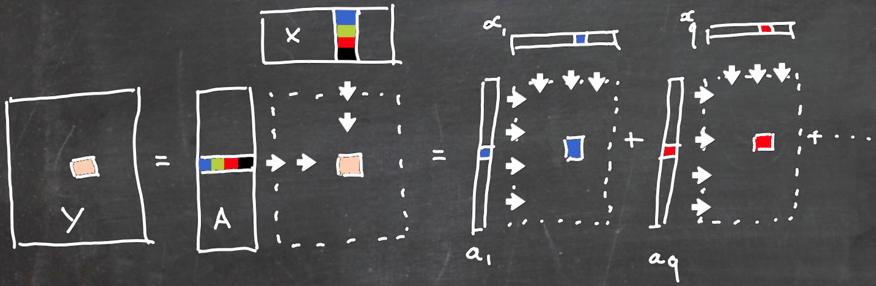
$$\underset{x \in \mathbb{R}_+}{\operatorname{argmin}} \, (y - ax)^2 = \left[\frac{y}{a}\right]^+$$



$$\underset{x \in \mathbb{R}_+}{\operatorname{argmin}} \, \|y - ax\|_2^2 = \frac{1}{\|a\|_2^2}\left[a^T y\right]^+ \qquad \underset{x^T \in \mathbb{R}_+^n}{\operatorname{argmin}} \, \|Y - ax^T\|_F^2 = \frac{1}{\|a\|_2^2}\left[a^T Y\right]^+$$

Matrix Multiplication?

$$Y = AX, \quad Y_{ji} = \sum_{q=1}^{d} A_{jq} X_{qi}, \quad Y = \sum_{q=1}^{d} a_q \otimes x_q$$

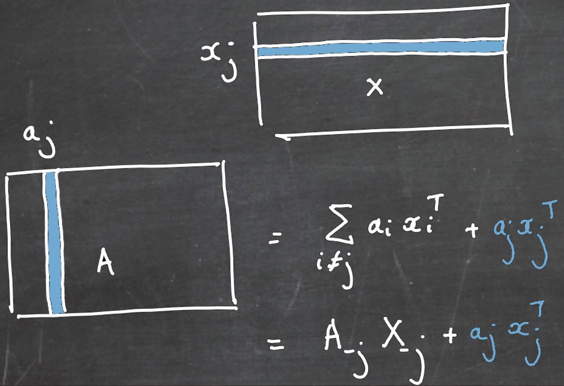


https://ncase.me/matrix/

Solving per row

We solve several NNLS problems with $Y = [y_1, \dots, y_n]$ and $X = [x_1, \dots, x_n]$, *i.e.*

$$\underset{X \in \mathbb{R}_+^{d \times n}}{\operatorname{argmin}} \|Y - AX\|_F^2$$



HALS

A block coordinate algorithm solves

$$\underset{x_j \in \mathbb{R}_+^n}{\operatorname{argmin}} \|(Y - A_{-j}X_{-j}) - a_j x_j\|_F^2$$

for each x_j alternatively until convergence.

Proposition: [Bertsekas 1995, earlier?]
As long as A has no zero column, the HALS iterates converge towards a minimizer of NNLS.

HALS pseudocode

Algorithm 1 HALS for NNLS

Inputs: Y, A
while Convergence is not met **do**
  for j in [1..d] **do**
    Compute $Z = Y - A_{-j}X_{-j}$
    If $a_j \neq 0$, set $x_j = \left[\frac{a_j^T Z}{\|a_j\|_2^2}\right]^+$
  end for
end while

Improvable by pre-allocation, see NMF section.

HALS pros and cons

- Flexible (similar problems)
- Early stop
- BLAS3 matrix version

- Infinite number of steps
- Slower than AS if very good start

A remark

$$x \leftarrow \frac{1}{\|a\|_2^2}\left[a^T Y\right]^+$$

is exactly

- a projected least squares update.
- a projected gradient step with the Lipschitz constant as inverse stepsize.
- a Gauss-Newton step.

$$\nabla_x \left[\frac{1}{2}\|Y - ax\|_F^2\right](x) = -a^T Y + \|a\|_2^2 x$$

We can use that logic to derive HALS for NNLS variants.

HALS for sparse NNLS

Let $\lambda > 0$ and consider

$$\underset{X \in \mathbb{R}_+^{d \times n}}{\operatorname{argmin}} \frac{1}{2}\|Y - AX\|_F^2 + \lambda \|X\|_1$$

To obtain the HALS update rule, consider

$$\underset{x_j \in \mathbb{R}_+^n}{\operatorname{argmin}}\ h_j(x_j) := \frac{1}{2}\|Z_j - a_j x_j\|_F^2 + \lambda \|x_j\|_1$$

By setting

$$\nabla_x h_j(x) = -a_j^T Z_j + \|a_j\|_2^2 x + \lambda \mathbb{1}$$

to zero, solving and projecting, we get

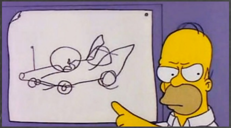
$$x_j^* = \left[\frac{a_j^T Z_j - \lambda \mathbb{1}}{\|a_j\|^2}\right]^+$$

III. Matrix and Tensor rank(s)

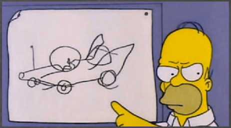
Back to the fil rouge

Problem 2: given a painting $\{y_i\}_{i \leq n}$, find its closest 2-color version.

$$\text{Find } A \in \mathbb{R}_+^{3\times 2} \text{ and } x_i \in \mathbb{R}_+^2 \text{ such that } \forall i \leq n, y_i \approx A x_i$$



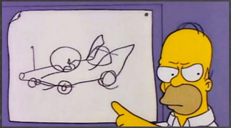
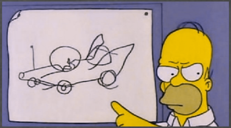
$\approx \;\forall_i,\; x_{i1}\bullet + x_{i2}\bullet =$



Back to the fil rouge

Problem 2: given a painting $\{y_i\}_{i \leq n}$, find its closest 2-color version.

$$\text{Find } A \in \mathbb{R}_+^{3\times 2} \text{ and } x_i \in \mathbb{R}_+^2 \text{ such that } \forall i \leq n, y_i \approx A x_i$$

 $\approx \ \forall_i, \ x_{i1}$ ● $+ \ x_{i2}$ ● $=$ 

A NNLS problem for each A??

Matrix rank

Definition:
For some matrix $Y \in \mathbb{R}^{m \times n}$, a factorization

$$Y = \sum_{q=1}^{d} a_q \otimes x_q = AX$$

is called a rank-d decomposition of Y for $d \leq \min(m, n)$.

Definition:
The rank of a matrix Y is the smallest d such that Y admits a rank-d decomposition,

$$\min \left\{ d \in \mathbb{N},\ Y = \sum_{q \leq d} a_q \otimes x_q \right\}$$

Matrix rank facts

The following other definitions of rank are equivalent:

- Dimension of column space of Y
- Dimension of row-space of Y
- Largest square submatrix B of Y with $\det(B) \neq 0$
- Dimension of the Kernel of Y
- Number of positive singular values of Y

Also, it holds that

- $\text{rank}(Y) \leq \min(m, n)$
- For a "generic" Y, $\text{rank}(Y) = \min(m, n)$
- The set $\{Y, \text{ rank}(Y) \leq d\}$ is closed.

Matrix rank facts

The following other definitions of rank are equivalent:

- Dimension of column space of Y
- Dimension of row-space of Y
- Largest square submatrix B of Y with $\det(B) \neq 0$
- Dimension of the Kernel of Y
- Number of positive singular values of Y

Also, it holds that

- $\text{rank}(Y) \leq \min(m, n)$
- For a "generic" Y, $\text{rank}(Y) = \min(m, n)$
- The set $\{Y, \text{ rank}(Y) \leq d\}$ is closed.

A reformulation of Problem 2

Let us drop nonnegativity for now. Then Problem 2 boils down to

$$\underset{Z\in\mathbb{R}^{m\times n}}{\operatorname{argmin}} \|Y - Z\|_F^2 \;\text{ s.t. }\; \operatorname{rank}(Z) \leq d$$

For the 2-color best painting, we set $d = 2$. This is the projection on the set of low-rank matrices.

Proposition:
(i) A best low-rank approximation Z^* always exists.
(ii) A solution is known in closed form by considering the SVD

$$Y = U\Sigma V^T,\ U^T U = I_m,\ V^T V = I_n,\ \Sigma_{ij} = \sigma_i \delta_{ij}$$

and truncating the $\operatorname{rank}(A) - d$ smallest singular values.
(iii) If the dth singular value is simple then Z^* is unique.

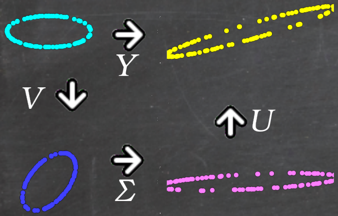
A short focus on SVD

Singular Value Decomposition:
For any $Y \in \mathbb{R}^{m \times m}$ there exist orthogonal matrices $U, V \in \mathbb{R}^{m \times m}$ and a nonnegative diagonal matrix $\Sigma \in \mathbb{R}_+^{m \times m}$ such that

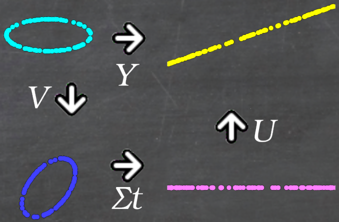
$$Y = U \Sigma V$$

A linear map is a rotation, a scaling/projection, and a rotation.

A short focus on tSVD



V, Σ, U applied sequentially



Rank-one approximation

Why truncation?

Intuition:

$$
\begin{aligned}
\|Y - Z\|_F^2 &= \|U\Sigma V^T - Z\|_F^2 \\
&= \|\Sigma - U^T Z V\|_F^2 \\
&= \|\Sigma - \tilde{Z}\|_F^2
\end{aligned}
$$

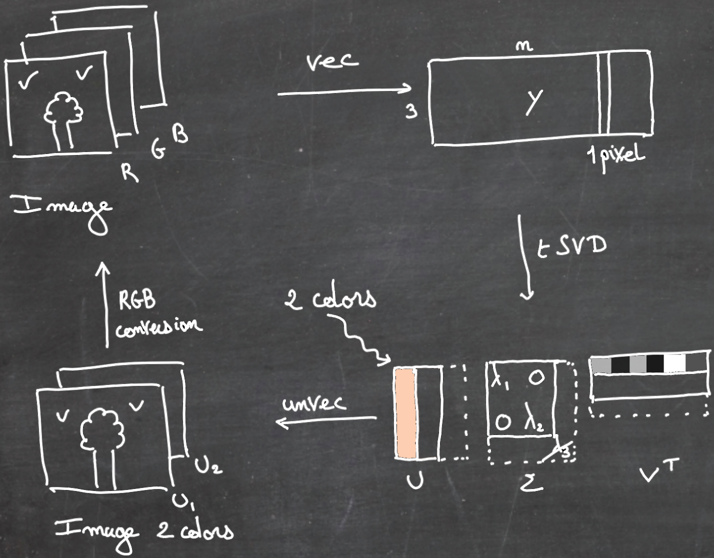
We can guess that

$$
\min_{\text{rank}(\tilde{Z}) \leq d} \|\Sigma - \tilde{Z}\|_F^2 = \min_{\|z\|_0 \leq d} \|s - z\|_2^2
$$

where $Diag(s) = \Sigma$. Finally $Z^* = U\Sigma(1:d)V^T$.

Actual proof on Wikipedia!

Did we solve Problem 2?



Did we? Your opinion.

Vote at https://www.wooclap.com/ITWISTQ2

Lunch break!!

Nonnegative Rank

The SVD rarely provides nonnegative entries for U, V except for $d = 1$, see Perron-Frobenius Theorem. We need nonnegativity constraints!!

Definition:
Let $Y \in \mathbb{R}_+^{m \times n}$ a nonnegative matrix. A nonnegative matrix factorization of Y is a factorization

$$Y = AX$$

for $A \in \mathbb{R}_+^{m \times d}$ and $X \in \mathbb{R}_+^{d \times n}$. The smallest such d is the nonnegative rank of Y, *i.e.*

$$\text{rank}_+(Y) = \min \left\{ d \in \mathbb{N},\ Y = \sum_{q=1}^{d} a_q \otimes x_q \text{ and } a_q \geq 0, x_q \geq 0 \right\}$$

Tensor rank

In this talk, tensors are multidimensional arrays $T \in \mathbb{R}^{m \times n \times p}$.

Definition:
Let $Y \in \mathbb{R}^{m \times n \times p}$ a tensor. A rank d decomposition of Y is a factorization

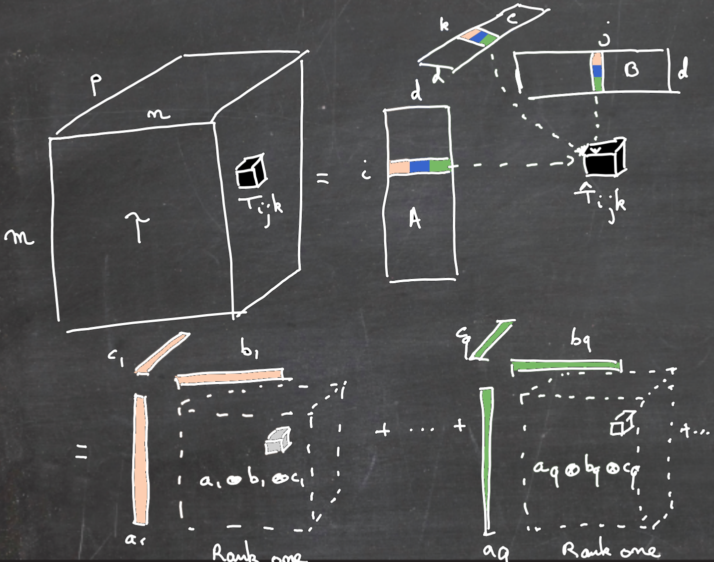
$$Y_{ijk} = \sum_{q=1}^{d} A_{iq} B_{jq} C_{kq}$$

for $A \in \mathbb{R}^{m \times d}$, $B \in \mathbb{R}^{n \times d}$ and $C \in \mathbb{R}^{p \times d}$. The smallest such d is the rank of Y, *i.e.*

$$\text{rank}(Y) = \min \left\{ d \in \mathbb{N},\ Y = \sum_{q=1}^{d} a_q \otimes b_q \otimes c_q \right\}$$

Rank decomposition

Others names: CPD, PARAFAC, CANDECOMP...



Computing the rank?

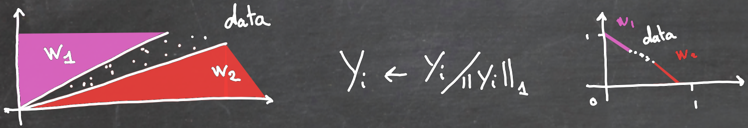
Computing or guessing the rank is extremely difficult in general, except for the matrix rank.

IV. Nonnegative Matrix Factorization

Exact and approximate NMF

Exact NMF (known rank d):

$$\text{Find } W \in \mathbb{R}_+^{m\times d}, H \in \mathbb{R}_+^{d\times n} \text{ s.t. } Y = WH$$



Approximate NMF (fixed approx. rank d, Frobenius loss):

$$\text{Solve} \quad \underset{W\in\mathbb{R}_+^{m\times d}, H\in\mathbb{R}_+^{d\times n}}{\operatorname{argmin}} \quad \|Y - WH\|_F^2$$

Nonconvex problem! But convex constraints!

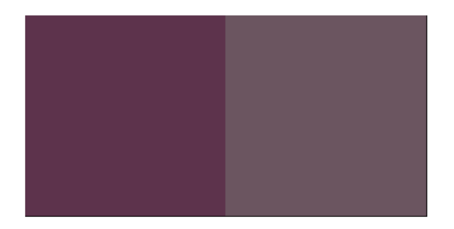
A little experiment

We have painted Homer with 2 colors using NNLS.

This image has nonnegative rank 2.

A little experiment

Goal:
Recover the two colors that were used.

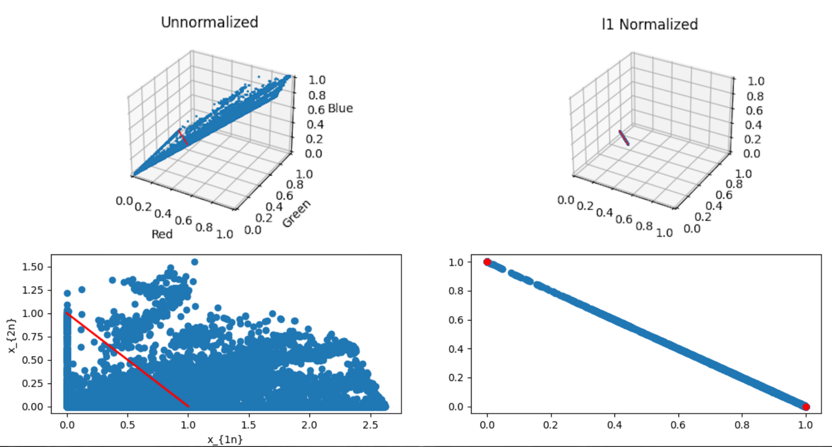


Procedure:
Compute 9 times a rank-2 NMF of the matrix $Y \in \mathbb{R}_+^{3\times d}$ with an alternating HALS algorithm (see later).
Initialized with $W_{ij} \sim \text{abs}\,(\mathcal{N}(0,1))$

A little experiment

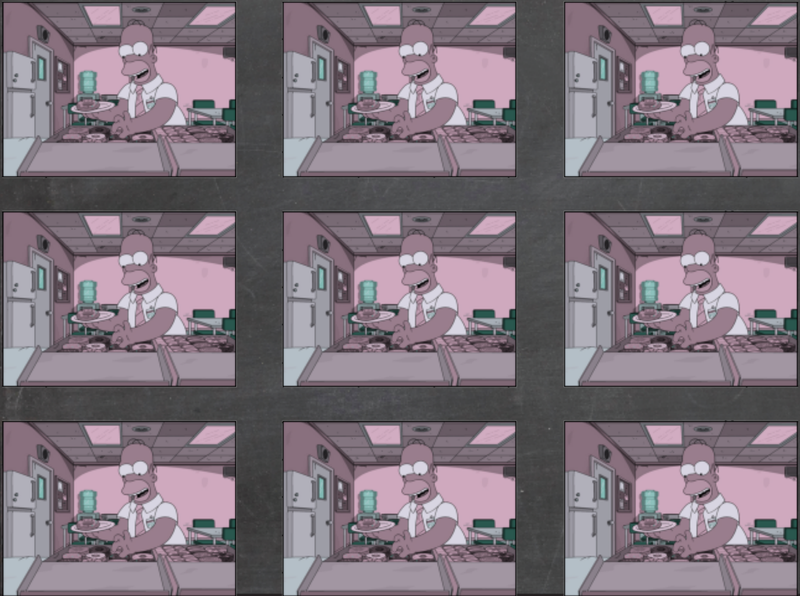
What will happen? Vote: https://www.wooclap.com/ITWISTQ3

A little experiment: data

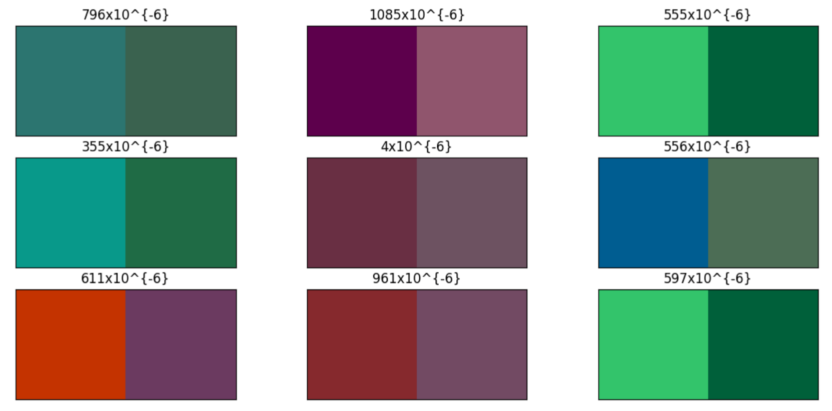


red: ground truth 2-colors

A little experiment: reconstructions



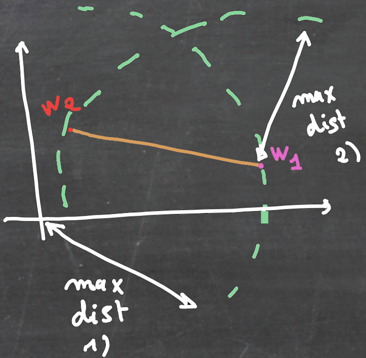
A little experiment: colors



Wait a minute...

The exact rank-2 NMF problem looks actually easy.

- Normalize data
- Select column of maximal l2 normal $\rightarrow W_1$
- Find its furthest column $\rightarrow W_2$
- Solve the strongly convex resulting NNLS problem $\rightarrow H$



Proposition:
The exact rank-2 NMF problem is in PTIME(n).

Rank > 3

Let's build a harder instance of Exact NMF.
Let $Y \in \mathbb{R}_+^{4\times n}$ with no zero column.

Normalization:

$$Y = WH \equiv YD_Y^{-1} = WD_W^{-1} D_W HD_Y^{-1}$$

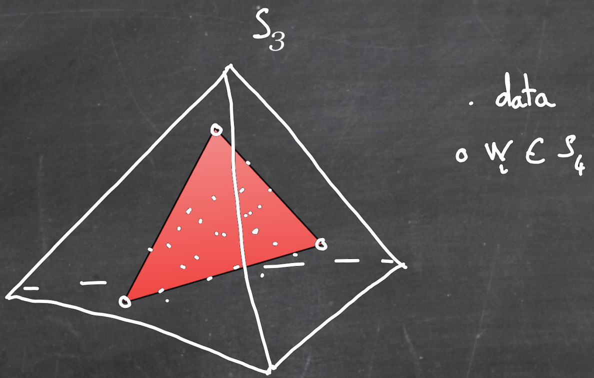
so that Y and W may wlog belong to the simplex $\mathcal{S}_3$.
Furthermore,

$$\|Y_i\|_1 = 1 = \left\| \sum_q W_{:q} H_{qi} \right\|_1 = \ldots = \|H_{:i}\|_1$$

so that H is also normalized.

Rank > 3

Proposition: [Vavasis2007]
Exact NMF with rank $3 < d < m$ as part of the input is NP-hard.



In fact this problem is still in P [Silio 1979, Agrawal 1989], nontrivially. More in the NMF book [Gillis 2020].

A comparison with sparse coding

Sparse coding

$$\min_{x \in \mathbb{R}^d} \|x\|_0 \;\; \text{s.t.} \;\; y = Ax$$

is NP-hard(d), but when fixing some sparsity $k < d$,

$$\text{Find } x \in \mathbb{R}^d, \; \|x\|_0 = k \;\; \text{s.t.} \;\; y = Ax$$

is in P(d), since it is enough to test all $\binom{d}{k} \sim \mathcal{O}(d^k)$ supports.

Approximate NMF is hard

Even rank 2 approximate NMF of a rank $d \geq 3$ matrix is hard!

And even rank 1 approximate NMF of a matrix with negative entries is NP-hard.

It's all nice, but how to compute (approximate) NMF?

Alternating Algorithms

Algorithm 2 A general alternating algorithm for NMF

1. **Inputs:** Y, d, W^0
2. Set $k = 0$
3. **while** Stopping criterion is not met **do**
4. Update H^{k+1} with fixed W^k
5. Update W^{k+1} with fixed H^{k+1}
6. **end while**

Convergence as a BCD algorithm [Bertsekas] if each NNLS has a unique solution (hard to check).

Alternating Algorithms

Algorithm 3 HALS algorithm for NMF

Inputs: Y, d, W^0
Set $k = 0$
while Stopping criterion is not met **do**
  Update H^{k+1} with fixed $W^k \leftarrow$ NNLS HALS solver
  Update W^{k+1} with fixed $H^{k+1} \leftarrow$ NNLS HALS solver
end while

Convergence guarantied by the PALM framework [Bolte 2014] when no columns of W, H^T are null through the iterations. Indeed HALS is exactly an alternating proximal gradient with Lipschitz step.

A second look at NNLS HALS

Algorithm 4 HALS for NNLS, solving for H

Inputs: Y, W, H^0
while convergence criterion is not met **do**
  for q in [1..d] **do**
    Compute $Z = Y - W_{-q}H_{-q}$
    If $W_q \neq 0$, set $H_q = \left[\frac{W_q^T Z}{\|W_q\|_2^2}\right]^+$
  end for
end while

A second look at NNLS HALS

Algorithm 4 HALS for NNLS, solving for H

Inputs: Y, W, H^0

while convergence criterion is not met **do**

  for q in [1..d] **do**

    Compute $Z = Y - W_{-q}H_{-q}$

    If $W_q \neq 0$, set $H_q = \left[\frac{W_q^T Z}{\|W_q\|_2^2}\right]^+$

  end for

end while

Important tweaks:

- Precompute $WtW := W^T W, WtY := W^T Y$
- Early stop, e.g. when $\|Y - WH\|_F^2 < 10^{-4}\|Y - WH^0\|_F^2$
- Warm start H^0 from the previous outer loop in NMF HALS

Application 1: automatic transcription

Data:

- An audio recording.

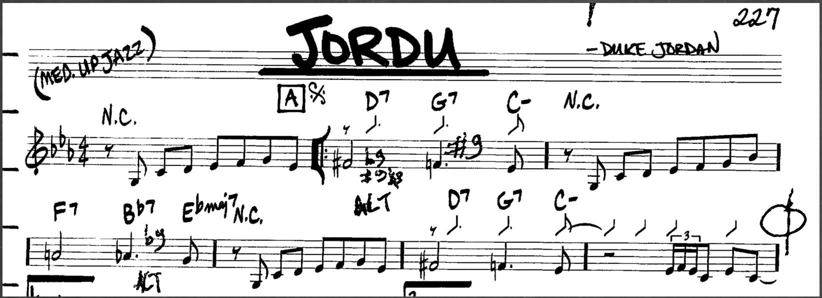
Procedure:

- Form a time-frequency matrix $Y \in \mathbb{R}_+^{n \times m}$
- Perform a rank d NMF of Y.
- In principle, identify notes and activations to produce MIDI

Goals:

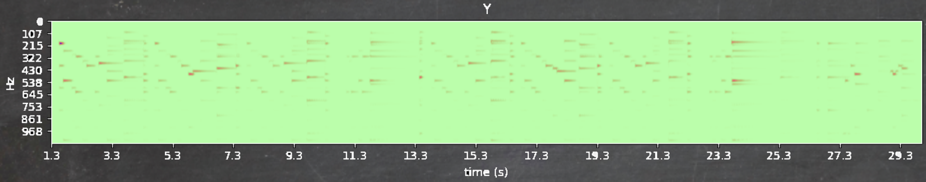
- Recover the music sheet solely from the audio

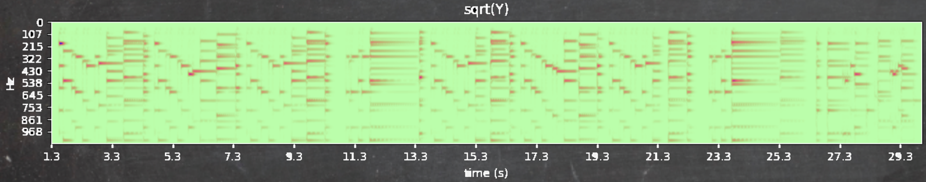
Application 1: automatic transcription



Jordu.wav

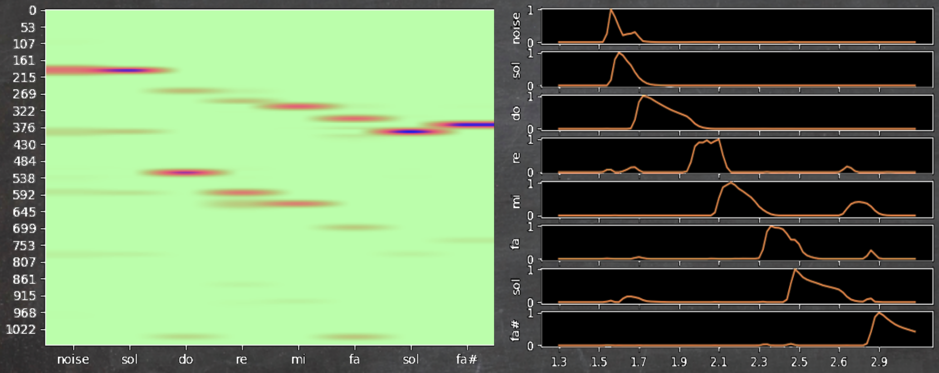
Application 1: data



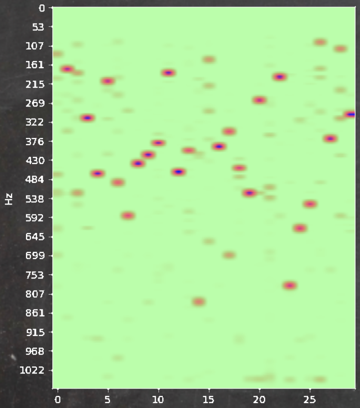


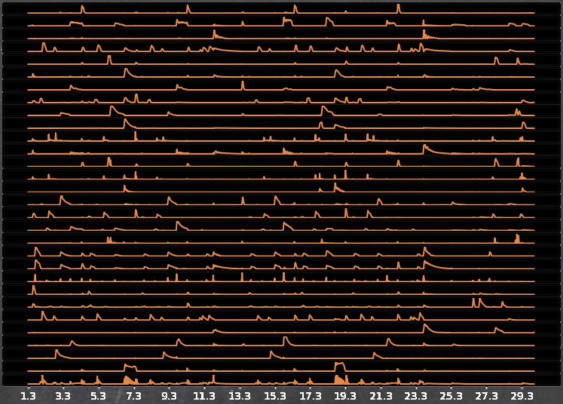
Application 1: easy case

Only the first 3 seconds, isolated notes!



Application 1: hard case





Application 2: Text mining for newbies

Data:

- A collection of $m = 8$ text files, collected from web articles.
- A dictionary of semantically useless words (from sklearn).

Procedure:

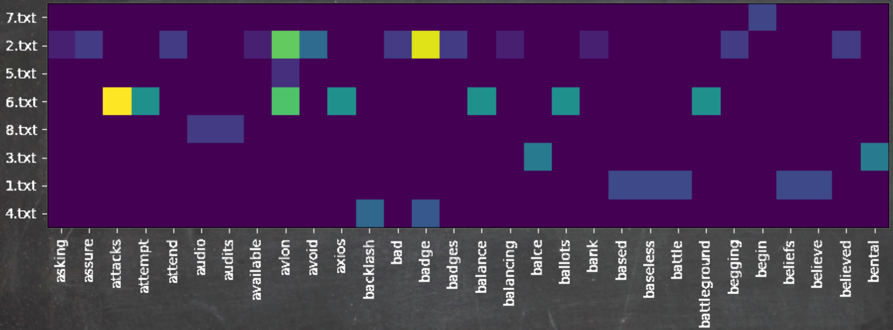
- Form a frequency matrix $Y \in \mathbb{R}_+^{8 \times n}$ (with sklearn)
- n is the number of different words in the files.
- Perform a rank 3 NMF of Y.

Goals:

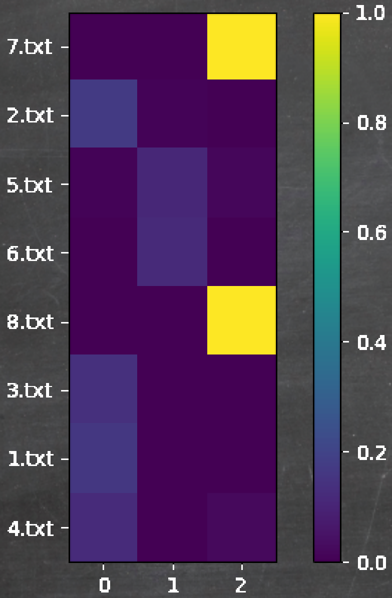
- Classify articles automatically
- Uncover hidden patterns in articles
- Generally speaking, extract information

Application 2: data

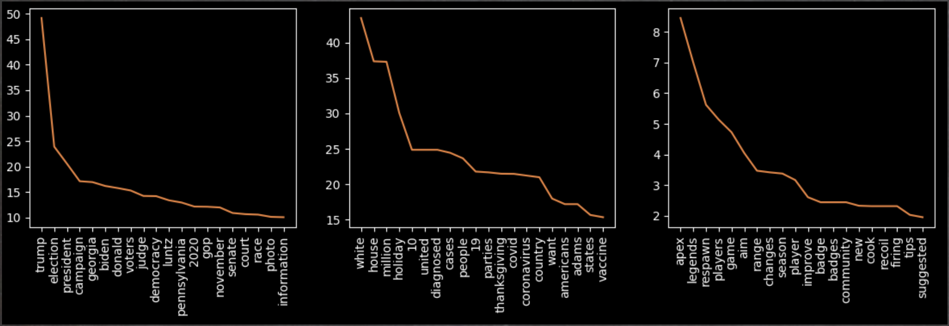
A few columns of the Y matrix



Application 2: Estimated W



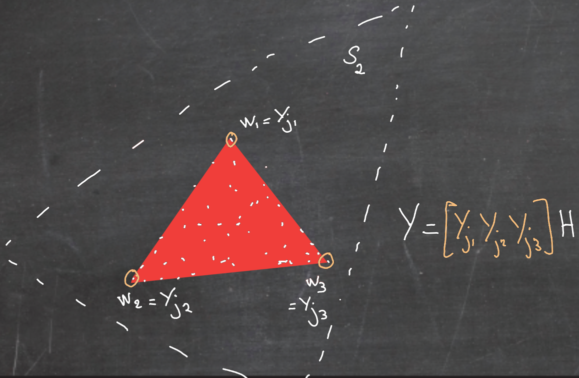
Application 2: Estimated H



Other NMF concepts

Separable NMF:[Arora 2012, Gillis 2013, ...]
Columns of W are in the data. Exact separable NMF is in P, but near-separable NMF is NP-hard.

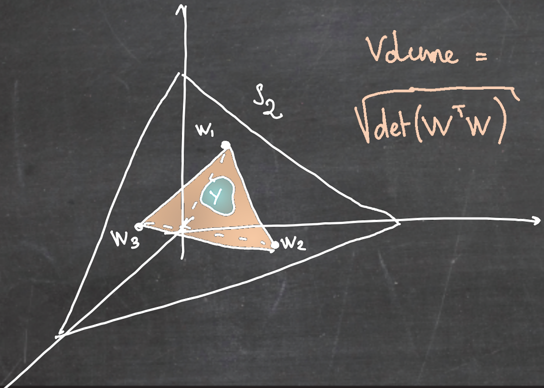
$$\underset{S \in \mathcal{P}_d([1,n]),\ H \geq 0}{\text{argmin}} \|Y - Y_S H\|_F^2$$



Other NMF concepts

Minimum volume NMF:[Fu and Huang 2016]
Penalize the volume of Conv(W). May lead to unique W and H!

$$\underset{W\geq 0,\ W^T\mathbb{1}_m=\mathbb{1}_d\ H\geq 0}{\operatorname{argmin}} \|Y - WH\|_F^2 + \lambda \log\det(W^T W + \delta I_d)$$



Other NMF concepts: β-divergence NMF

Change the cost to

$$d_\beta(x,y) = \begin{cases} \frac{1}{\beta(\beta-1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{(\beta-1)}) & \text{if } \beta \notin \{0,1\} \\ x \log \frac{x}{y} - x + y & \text{if } \beta = 1 \text{ (KL div)} \\ \frac{x}{y} - \log \frac{x}{y} - 1 & \text{if } \beta = 0 \text{ (IS div)} \end{cases}$$

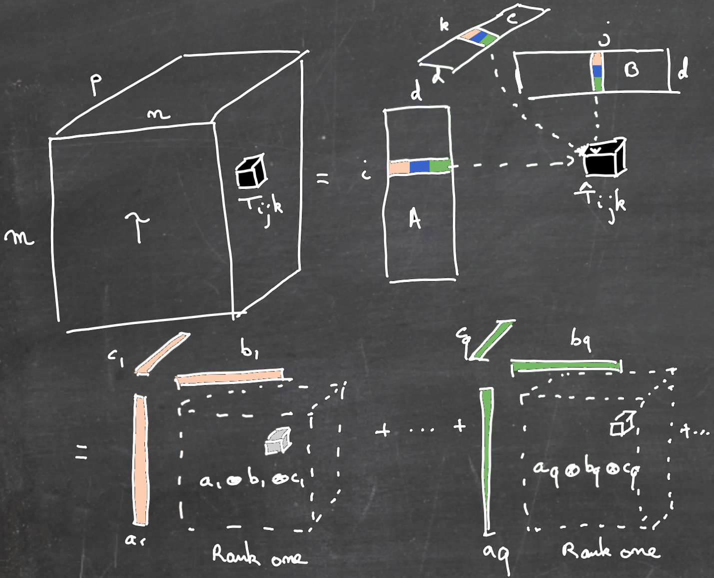
and solve

$$\operatorname*{argmin}_{W,H\geq 0} \sum_{ij} d_\beta\left(Y_{ij}, [WH]_{ij}\right)$$

typically with multiplicative updates [Fevotte Idier 2011].

IV. Nonnegative Tensor Factorization

NTF illustrated



NTF: similarities with NMF

A few equivalent formulations of exact NTF:

$$T_{ijk} = \sum_{q=1}^{d} W_{iq} H_{jq} C_{kq} = \sum_{q=1}^{d} w_q \otimes h_q \otimes c_q$$

$$Y_k := T_{::k} = W \mathrm{Diag}(C_{k:}) H^T$$

NTF can be seen as a collection of NMFs with the same W, H up to nonnegative scaling!

Moreover,

$$\underset{W,H,C \geq 0}{\mathrm{argmin}} \, \| T - \sum_{q=1}^{d} w_q \otimes h_q \otimes c_q \|_F^2$$

is still a NNLS problem with respect to one factor, *e.g.* H. This problem always has a solution, which is generically unique [Qi 2016]. **Factors W, H, C are often unique too!**.

Complexity recap

Low Rank Approximation:

$$\underset{Z \in \mathbb{R}_{(+)}^{m \times n (\times p)}}{\operatorname{argmin}} \|Y - Z\|_F^2 \text{ s.t. } \operatorname{rank}_{(+)}(Z) \leq d$$

Table: Properties of ranks [Lim2013, Vavasis2007, Friedland2013, Qi2016]

	mat. rank	mat. rank_+	ten. rank	ten. rank_+
exact	P	NP-hard	?	?
approx	P	NP-hard	NP-h., ill-posed	?
unique Z	Generic	Generic	ill-posed	Generic
unique A, X	No	No	Generic	Generic
algorithm	tSVD	Heuristics	∞	Heuristics

NTF: applying HALS

Computing the gradient:
One can check that

$$\nabla_c \left[w \otimes h \otimes c\right](w, h, c) = w^* \otimes h^* \otimes I_p$$

and therefore

$$\begin{aligned} \frac{1}{2}\nabla_{c_1} &= -\left(w_1^* \otimes h_1^* \otimes I_p\right)\left(T - \sum_{q=1}^{d} w_q \otimes h_q \otimes c_q\right) \\ &= -w_1^T T h_1 + \sum_{q=2}^{d} \langle w_1, w_q\rangle \langle h_1, h_q\rangle c_q + \|w_1\|_2^2 \|h_1\|_2^2 c_1 \end{aligned}$$

One should precompute $w_q^T T h_q \ \forall q \leq d$, $W^T W$ and $H^T H$.

NNLS for NTF

Algorithm 5 HALS for NNLS for NTF

Inputs: T, W, H, C

while Convergence is not met **do**

  for j in [1..d] **do**

    Compute $Z = T - \sum_{q \neq j} w_q \otimes h_q \otimes c_q$

    If $w_j \neq 0$ and h_j, set $c_j = \left[\frac{w_j^T Z h_j}{\|w_j\|_2^2 \|h_j\|_2^2} \right]^+$

  end for

end while

An application of NTD to chemometrics

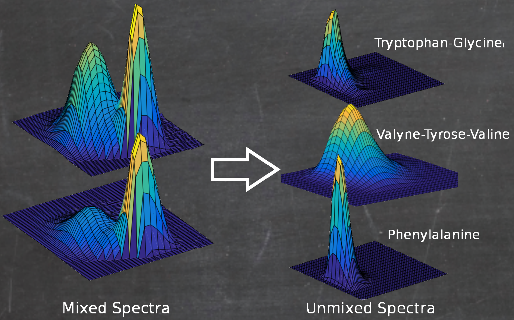
Material:

- Several mixtures of 3 fluorescent chemicals, in various concentrations.

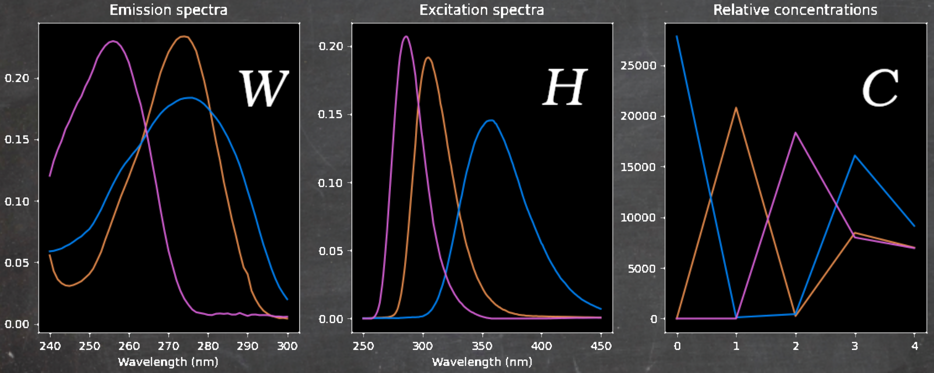
Procedure:

- Measure excitation-emission for each sample, stack in a tensor Y.
- Perform a rank 3 approximate NTF of Y.

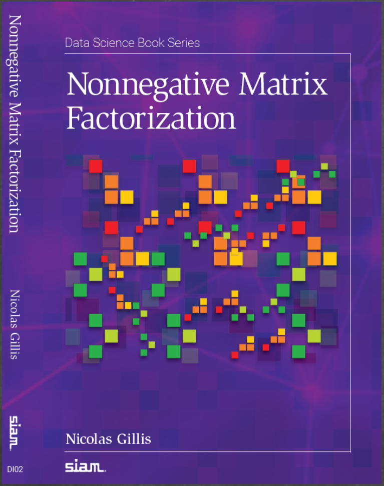
Goals:



An application of NTD to chemometrics



To go (much) further



Take home message: Stay ~~Positive~~ Nonnegative!