An Introduction to Nonnegative Matrix Factorization

Nicolas Nadisic

26 July 2022 — ICCOPT 2022, Lehigh University

University of Mons, Belgium

What was supposed to happen

Me	Hiroyuki Kasai	Andersen Man Shun Ang
		
Introduction to NMF	NMFLibrary (toolbox in Matlab)	Accelerating algorithms for NMF

What I will talk about

- High-level overview on NMF
- Not much math, many images
- Intuitions and key ideas

A bit superficial, but I will stick around after the talk for deeper discussions

With a ~~little~~ big help from my friends

Nicolas Gillis

Arnaud Vandaele

The motivation behind Nonnegative Matrix Factorization

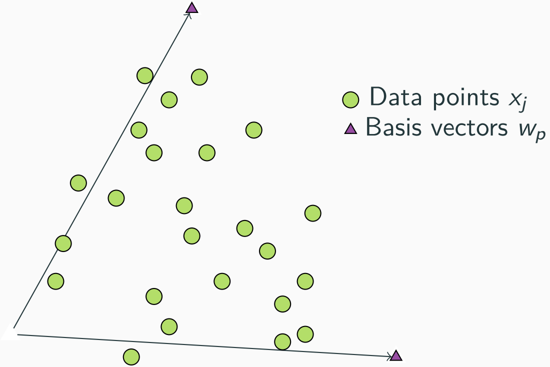
- Given a set of n data points x_j, for j in $1, 2, \ldots, n$
- We want to understand the underlying structure of the data

The motivation behind Nonnegative Matrix Factorization

- Given a set of n data points x_j, for j in $1, 2, \ldots, n$
- We want to understand the underlying structure of the data
- By finding a set of r basis vectors w_p such that for all j

$$x_j \approx \sum_{p=1}^{r} w_p h_{jp} \qquad \text{for some nonneg. weights } h_{jp}$$

This is a form of linear dimensionality reduction.



Nonnegative Matrix Factorization (NMF)

The NMF model

$$X = WH + N \in \mathbb{R}^{m \times n}$$

where X, W and H are entry-wise nonnegative, N is noise

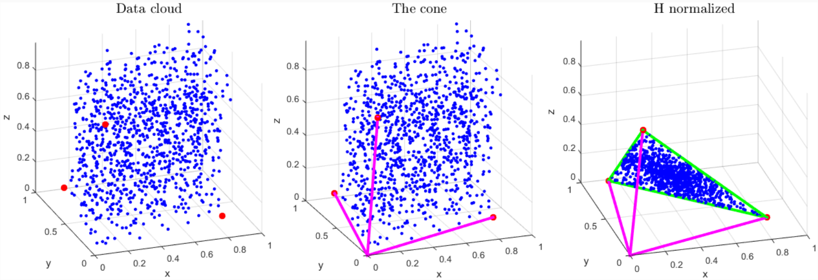
Problem:

- Given X and a rank $r \in \mathbb{N}$, $r \ll m, n$
- Estimate $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$

Geometrically:

- Columns of $W \Rightarrow$ basis vectors defining a cone
- Columns of $X \Rightarrow$ noisy data points contained in that cone

NMF = finding a cone



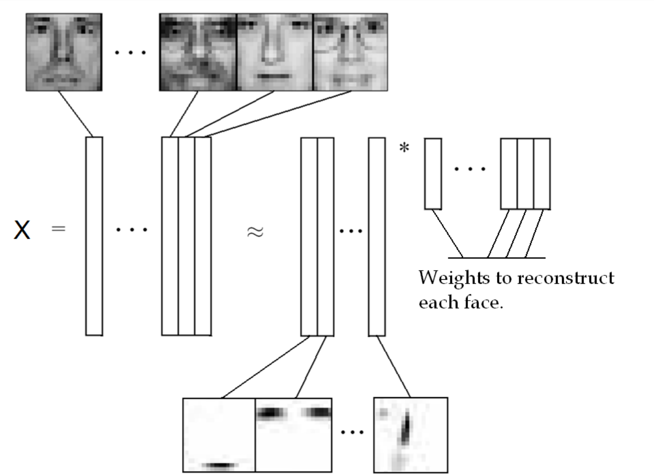
Why nonnegativity?

In other words, why don't you just apply PCA and call it a day?

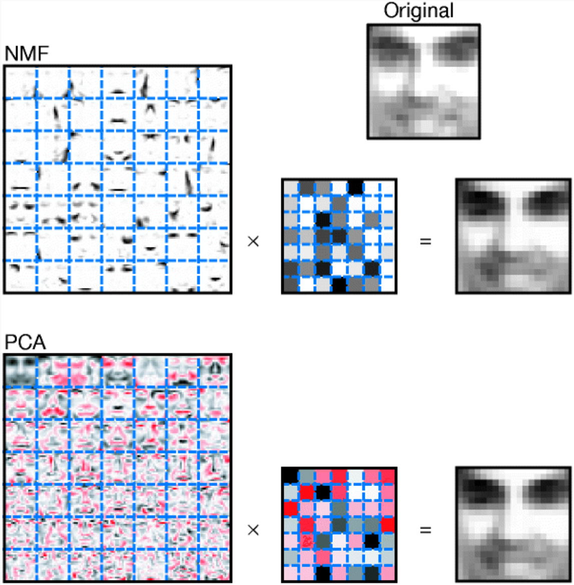
- Nonnegativity produces more interpretable solutions
- Natural constraint in many applications
- Favors the sparsity of the factors
- Curiosity: the NMF model is related to lots of interesting problems in math and machine learning

The pioneer paper on NMF + Application 1

"Learning the parts of objects by non-negative matrix factorization", Lee & Seung, 1999.



The pioneer paper on NMF + Application 1

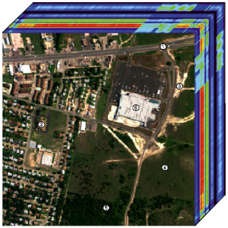


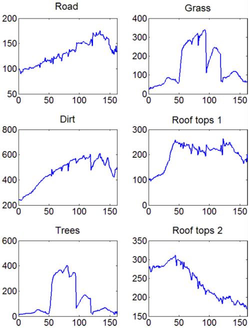
Actually, NMF has been around for a long time

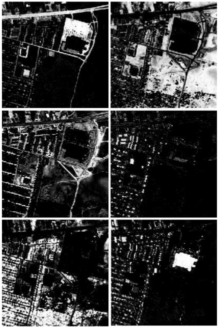
- Paper called "Positive matrix factorization" by Paatero & Tapper, 1994.
- Same model and algorithms exist under different names since the 1960's in the analytical chemistry community
- Also since the late 1980's in the geoscience and remote sensing community

Application 2: hyperspectral unmixing

$$\underbrace{X(:,j)}_{\substack{\text{spectral signature of}\\ \text{j-th pixel}}} \approx \sum_p \underbrace{W(:,p)}_{\substack{\text{spectral signature of}\\ \text{p-th material}}} \underbrace{H(p,j)}_{\substack{\text{abundance of p-th}\\ \text{material in the j-th pixel}}}$$

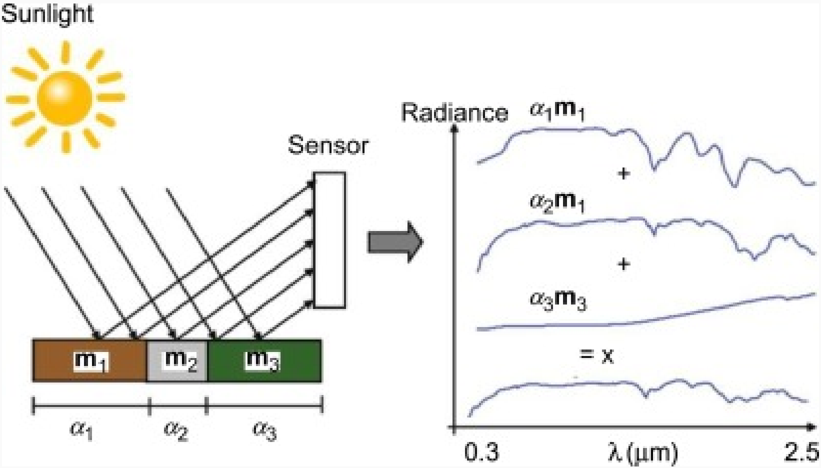




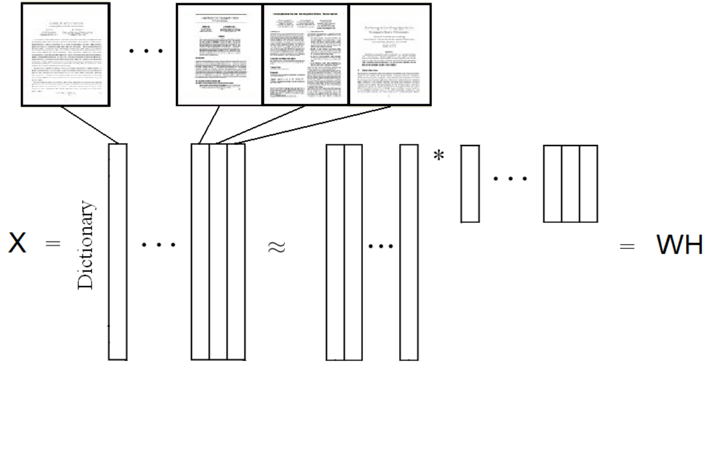


Images from J. Bioucas Dias and N. Gillis.

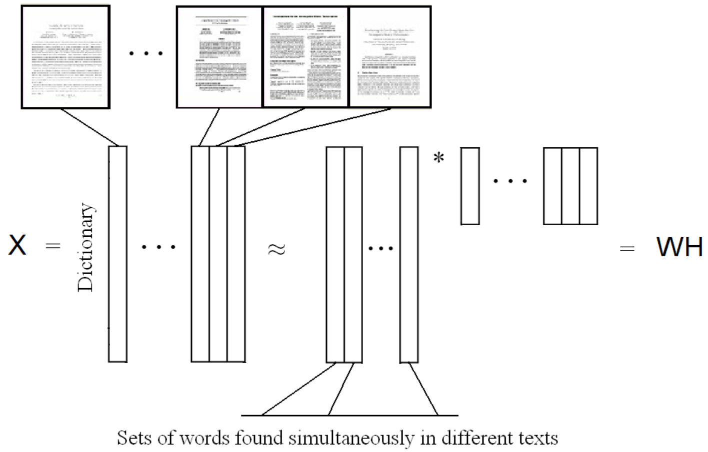
Linear mixing model



Application 3: topic modeling and document classification

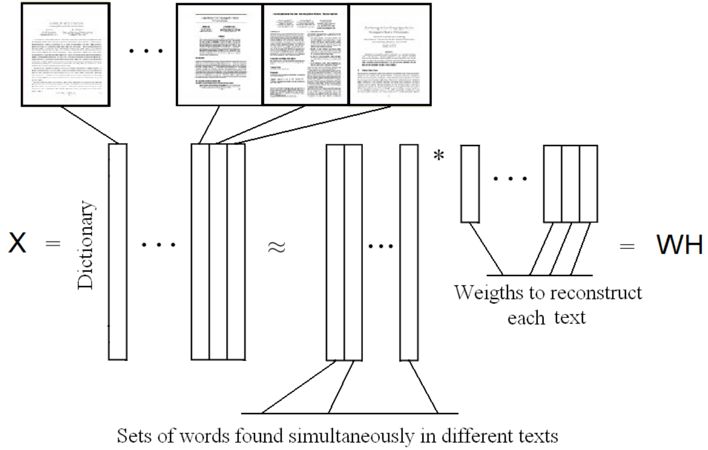


Application 3: topic modeling and document classification



- Basis elements allow to recover the different topics;

Application 3: topic modeling and document classification



- Basis elements allow to recover the different topics;
- Weights allow to assign each text to its corresponding topics.

Application 3: topic modeling and document classification

Five of the topics extracted by NMF on tdt2_top30 (1998 USA news):

Lewinsky scandal	Israeli-Palestinian conflict	Stock Market	Winter olympics in Nagano	Sports
lewinsky	israel	percent	olympic	game
mrs	israeli	stock	games	denver
jones*	netanyahu	market	olympics	team
lawyers	palestinian	stocks	nagano	super
clinton	peace	points	gold	bowl
president	palestinians	investors	medal	packers
sexual	arafat	prices	team	jordan
jordan**	bank	index	japan	play
relationship	minister	companies	winter	green
told	talks	quarter	won	bulls

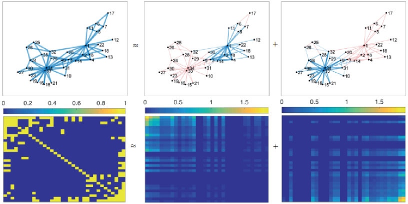
*Paula Jones sued Bill Clinton for an earlier sexual harassment affair.

**Vernon Jordan, a friend and political adviser to Bill Clinton, helped Monica Lewinsky after she left the White House.

Toolbox: This example can be run with https://gitlab.com/ngillis/nmfbook/

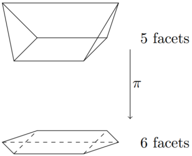
Other applications

Community detection



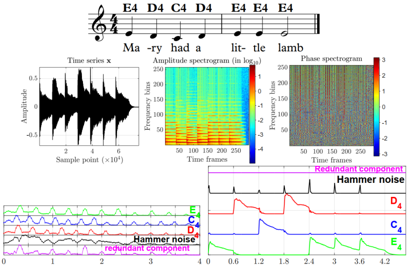
Yang, Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, ACM Web search and data mining, 2013.

Representing polytopes compactly



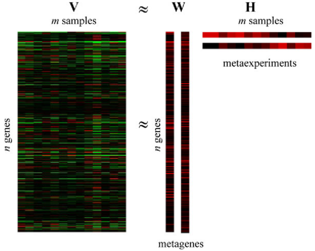
Extended formulations in combinatorial optimization, Kaibel, Optima, 2011.

Audio source separation



Févotte, Bertin, Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, Neural computation, 2009

Microarray data analysis



Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Kim and Park, Bioinformatics, 2007.

When are you gonna talk about optimization?

- In real-world applications, the model is slightly wrong and the data is noisy
- May be impossible to find W and H such that $X = WH$
- Therefore, we look for the best approximation

When are you gonna talk about optimization?

- In real-world applications, the model is slightly wrong and the data is noisy
- May be impossible to find W and H such that $X = WH$
- Therefore, we look for the best approximation

Approximate NMF

$$\min_{W \geq 0, H \geq 0} \|X - WH\|$$

where $\|.\|$ is some error measure serving as objective function.

- Different assumptions lead to different objectives
- We can also add regularizers or constraints

Some objective functions

- The most standard one is squared Frobenius norm, corresponds to assumption of Gaussian noise, work well in hyperspectral unmixing

$$\min \|X - WH\|_F^2$$

- ℓ_1-norm $\Rightarrow$ Laplace noise, more robust to outliers
- β-divergence $\Rightarrow$ Poisson noise
- Itakura–Saito distance, used for audio
- ...

Frobenius NMF

$$\min_{W\geq 0, H\geq 0} \|X - WH\|_F^2$$

- Non-convex
- NP-hard
- Ill-posed, non-unique solution $\Leftrightarrow$ identifiability issue
- Lots of local minima

Frobenius NMF

$$\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2$$

- Non-convex
- NP-hard
- Ill-posed,
  non-unique solution $\Leftrightarrow$
  identifiability issue
- Lots of local minima

In pratice: alternate optimization

1. Initialize W and H, then loop
   1.1 Fix W and optimize
   $H \approx \text{argmin}_{H \geq 0} \|X - WH\|_F^2$
   1.2 Fix H and optimize
   $W \approx \text{argmin}_{W \geq 0} \|X - WH\|_F^2$

Frobenius NMF

$$\min_{W \geq 0, H \geq 0} \|X - WH\|_F^2$$

- Non-convex
- NP-hard
- Ill-posed,
  non-unique solution $\Leftrightarrow$
  identifiability issue
- Lots of local minima

In pratice: alternate optimization

1. Initialize W and H, then loop
   1.1 Fix W and optimize
   $H \approx \text{argmin}_{H \geq 0} \|X - WH\|_F^2$
   1.2 Fix H and optimize
   $W \approx \text{argmin}_{W \geq 0} \|X - WH\|_F^2$

Subproblems are convex!

Some regularizations and constraints

Enrich the model with regularizers and constraints to:

- Leverage a-priori knowledge about the data at hand
- Improve solutions in a specific application
- Make the problem better-posed, have some *guarantees* about identifiability

Some regularizations and constraints

Enrich the model with regularizers and constraints to:

- Leverage a-priori knowledge about the data at hand
- Improve solutions in a specific application
- Make the problem better-posed, have some *guarantees* about identifiability

Examples:

- Sparse NMF
  - $\min \|X - WH\|_F^2 + \lambda \|H\|_1$
  - $\min \|X - WH\|_F^2$ s.t. $\|H(:, j)\|_0 \leq k$ for all j
- Separable NMF (details next slide)
- Minimum-volume NMF
- ...

One variant: Separable NMF

Separability assumption

There exists an index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$X = X(:, \mathcal{K})H + N$$

One variant: Separable NMF

Separability assumption

There exists an index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$X = X(:, \mathcal{K})H + N$$

Interpretation: for each vertex, there exist at least one data point equal to this vertex $\Longrightarrow$ pure-pixel assumption in hyperspectral unmixing

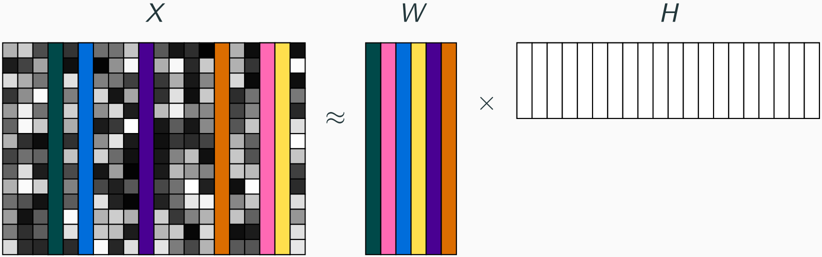
One variant: Separable NMF

Separability assumption

There exists an index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

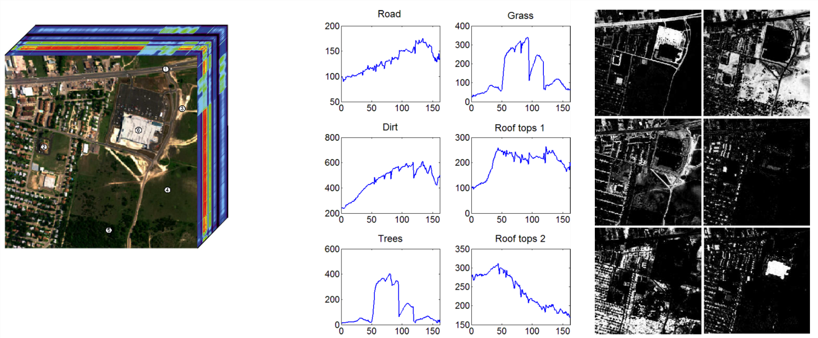
$$X = X(:, \mathcal{K})H + N$$

Interpretation: for each vertex, there exist at least one data point equal to this vertex $\Longrightarrow$ pure-pixel assumption in hyperspectral unmixing



Separability in hyperspectral unmixing: pure-pixel

$$\underbrace{X(:,j)}_{\text{spectral signature of } j\text{-th pixel}} \approx \sum_p \underbrace{W(:,p)}_{\text{spectral signature of } p\text{-th material}} \underbrace{H(p,j)}_{\text{abundance of } p\text{-th material in the } j\text{-th pixel}}$$



Separable NMF

- NMF is NP-hard in general
- Under the separability assumption, it is solvable in polynomial time
- Identifiability of the solution

4 years of my life in 5 slides

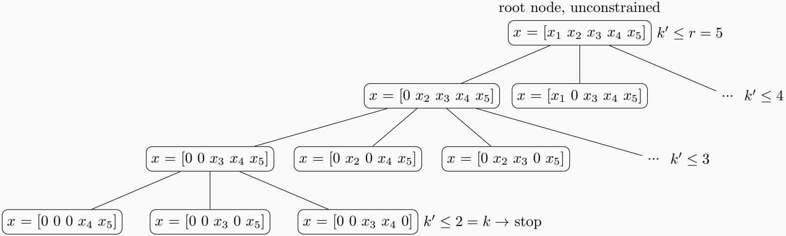
- Focus on ℓ_0-sparsity constraints $\Longrightarrow$ combinatorial problems
- Exact algorithms
- Combine sparse optimization and NMF

Column-wise k-sparse NMF: exact algorithm

$$\min \|X - WH\|_F^2 \text{ s.t. } \|H(:, j)\|_0 \leq k \text{ for all } j$$

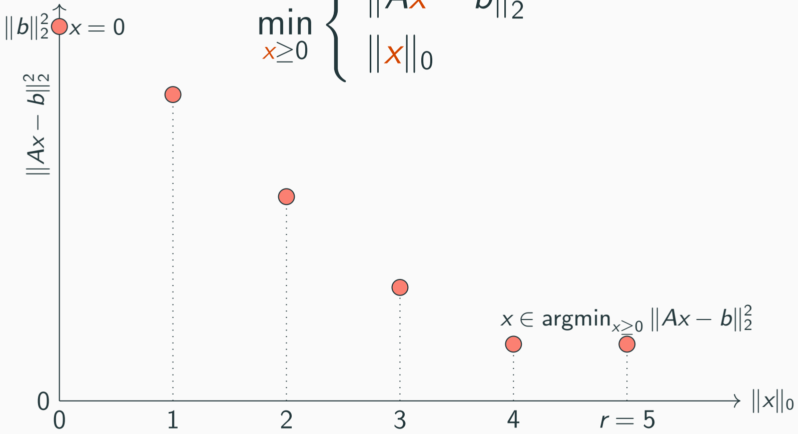
Subproblem is k-sparse NNLS:

$$\min_{x \geq 0} \|Ax - b\|_2^2 \text{ s.t. } \|x\|_0 \leq k$$



Bi-objective extension

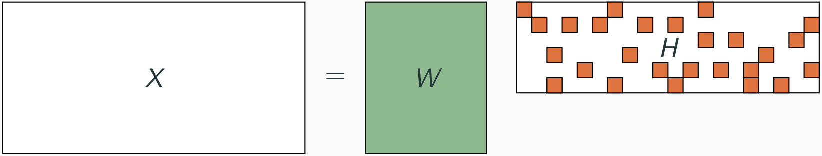
$$\min_{x \geq 0} \begin{cases} \|Ax - b\|_2^2 \\ \|x\|_0 \end{cases}$$



Matrix-wise ℓ_0 constraints

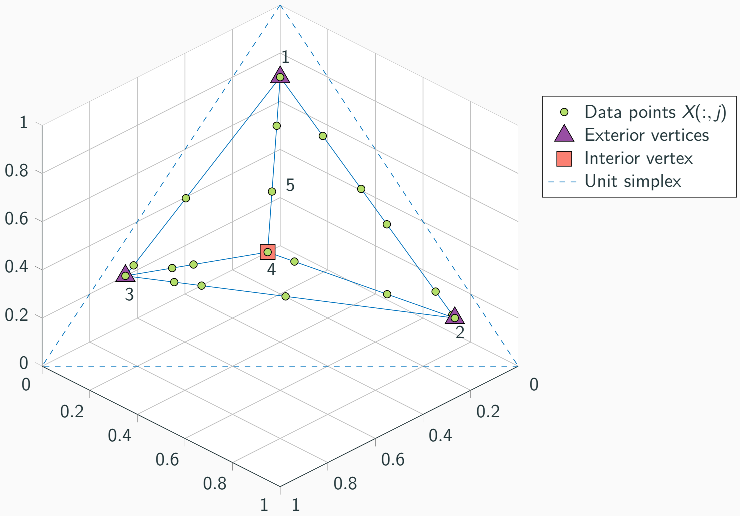
$$\min_{H \geq 0} \|X - WH\|_2^2 \text{ s.t. } \|H\|_0 \leq q$$

- Can be seen as a global sparsity budget
- If $q = k \times n$, this enforces an average k-sparsity on the columns of H

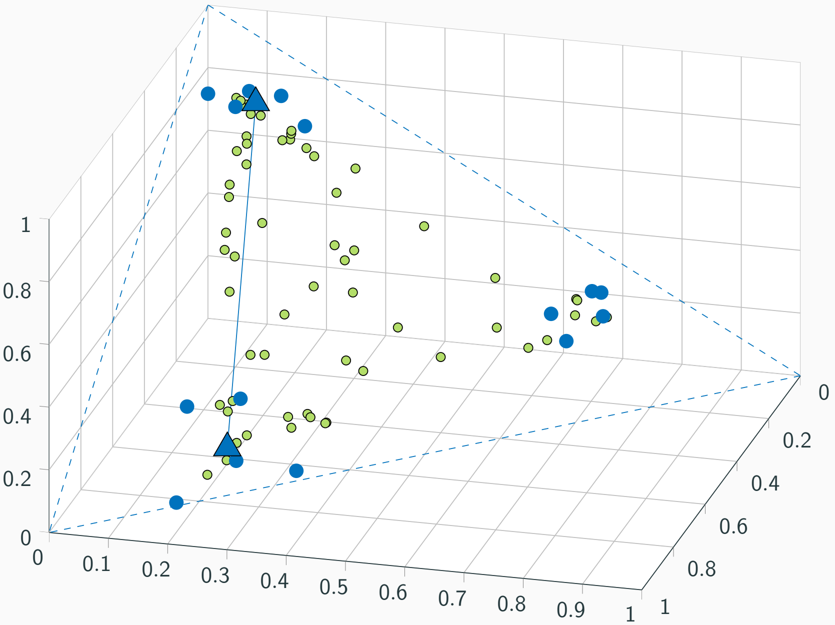


Sparse Separable NMF

$$X = X(:, \mathcal{K})H \quad \text{such that for all } j, \quad \|H(:,j)\|_0 \leq k$$



Smoothed Separable NMF



Thanks !

Contact: nicolas.nadisic@umons.ac.be

Website: http://nicolasnadisic.xyz

My supervisor's book: