What's happening in Nonnegative Matrix Factorization

A high level overview in 3 parts

Andersen Ang

Mathématique et recherche opérationnelle, UMONS, Belgium

Supervisor : Nicolas Gillis

Homepage: angms.science

September 27, 2018

Sparse Days 2018, Toulouse, France

Part I.
Introduction

Non-negative Matrix Factorization (NMF)

Given :

- A matrix $\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$.
- A positive integer $r \in \mathbb{N}$.

Find :

Non-negative Matrix Factorization (NMF)

Given :

- A matrix $\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$.
- A positive integer $r \in \mathbb{N}$.

Find :

- Matrices $\mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n}$ such that $\mathbf{X} = \mathbf{WH}$.

Non-negative Matrix Factorization (NMF)

Given :

- A matrix $\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$.
- A positive integer $r \in \mathbb{N}$.

Find :

- Matrices $\mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n}$ such that $\mathbf{X} = \mathbf{W}\mathbf{H}$.
- Important : everything is non-negative.

Notation : we use $\mathbf{W}\mathbf{H}$ instead of $\mathbf{W}\mathbf{H}^\top$.

Given the pair $(\mathbf{X} \in \mathbb{R}_{+}^{m \times n}, r \in \mathbb{N})$, find the pair $(\mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n})$ such that

$$\mathbf{X} = \mathbf{W}\mathbf{H}.$$

This is called *exact NMF*, NP-hard (Vavasis, 2007).

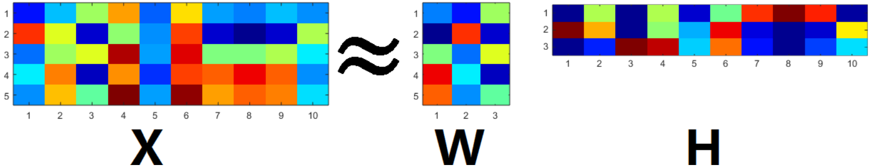
Given the pair $(\mathbf{X} \in \mathbb{R}_{+}^{m \times n}, r \in \mathbb{N})$, find the pair $(\mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n})$ such that

$$\mathbf{X} = \mathbf{W}\mathbf{H}.$$

This is called *exact NMF*, NP-hard (Vavasis, 2007).

(Low-rank) *approximate NMF* :

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}, \; 1 \leq r \leq \min\{m, n\}.$$



Vavasis, "On the complexity of nonnegative matrix factorization", SIAM J. Optim.

Find $(\mathbf{W}, \mathbf{H})$ numerically

Given $(\mathbf{X} \in \mathbb{R}_{+}^{m\times n}, r \leq \min\{m, n\})$, find $(\mathbf{W} \in \mathbb{R}_{+}^{m\times r}, \mathbf{H} \in \mathbb{R}_{+}^{r\times n})$ s.t. $\mathbf{X} \approx \mathbf{WH}$ via solving

$$[\mathbf{W},\ \mathbf{H}] = \underset{\mathbf{W}\geq\mathbf{0},\mathbf{H}\geq\mathbf{0}}{\arg\min} \|\mathbf{X} - \mathbf{WH}\|_F.$$

Find $(\mathbf{W}, \mathbf{H})$ numerically

Given $(\mathbf{X} \in \mathbb{R}_{+}^{m \times n}, r \leq \min\{m, n\})$, find $(\mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n})$ s.t. $\mathbf{X} \approx \mathbf{WH}$ via solving

$$[\mathbf{W},\ \mathbf{H}] = \underset{\mathbf{W} \geq \mathbf{0}, \mathbf{H} \geq \mathbf{0}}{\arg\min} \|\mathbf{X} - \mathbf{WH}\|_{F}.$$

- Minimizing the distance† between $\mathbf{X}$ and the approximator $\mathbf{WH}$ in F-norm.
- $\geq$ is element-wise (not positive semi-definite).
- Such non-convex minimization problem is ill-posed and also NP-hard (Vavasis, 2007).

* From now on, the inequality notations $\geq \mathbf{0}$ will be skipped.

†This talk does not consider other distance functions.

The scope of this talk

Given $(\mathbf{X}, r)$, find $(\mathbf{W}, \mathbf{H})$ via solving

$$[\mathbf{W},\ \mathbf{H}] = \arg\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F \text{ subject to } \star \ ,$$

The scope of this talk

Given $(\mathbf{X}, r)$, find $(\mathbf{W}, \mathbf{H})$ via solving

$$[\mathbf{W},\ \mathbf{H}] = \arg\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F \text{ subject to } \star\ ,$$

where $\star$: additional constraint(s)/regularization(s) that make the problem "better".

$\star$ in this talk :

- Nothing (this part) - NMF in the original form, being a NP-hard and ill-posed problem.
- Separability (part II) - to tackle the NP-hardness.
- Minimum volume (part III) - to generalize the separability.

Q : What about sparsity regularizer ?

A : Non-negativity *induces* sparsity (sparse NMF not covered in this talk).

For non-NMF people : why NMF ?

- Interpretability

  NMF beats similar tools (PCA, SVD, ICA) due to the interpretability on non-negative data.

- Model correctness

  NMF can find ground truth (under certain conditions).

- Mathematical curiosity

  NMF is related to some serious problems in mathematics.

- ~~My boss tell me to do it.~~

Why NMF - Hyper-spectral image application (1/2)

$$\underbrace{X(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{H(k,j)}_{\substack{\text{abundance of } k\text{th endmember}\\ \text{in } j\text{th pixel}}} .$$

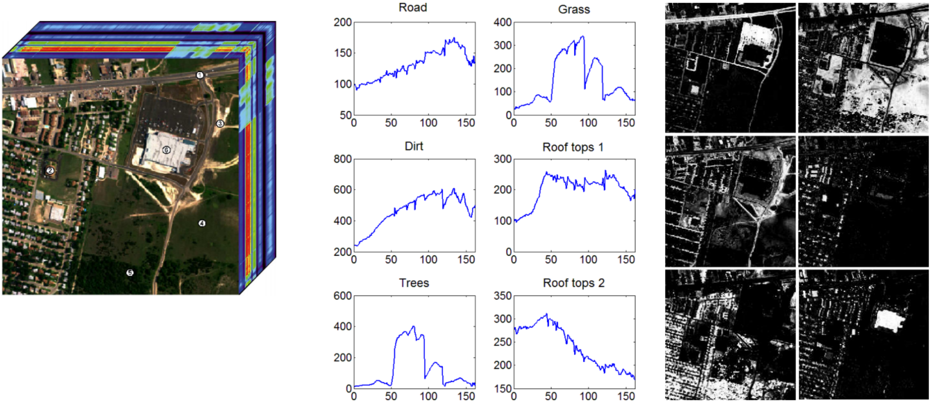


Figure: Hyper-spectral image decomposition. Figure shamelessly copied from (Gillis,2014).

N. Gillis, "The why and how of nonnegative matrix factorization", 2014

Why NMF - Hyper-spectral image application (2/2)

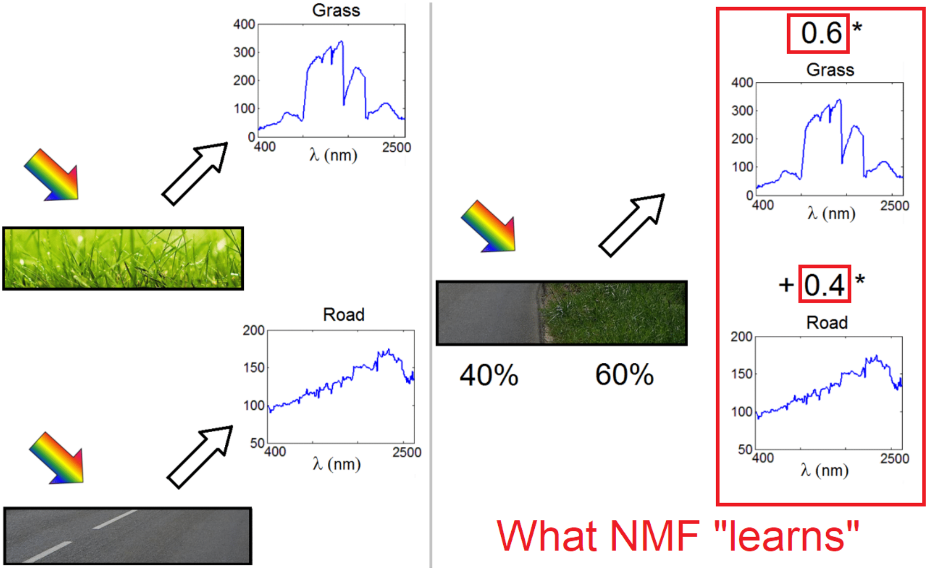


Figure: Hyper-spectral imaging. Figure modified from N. Gillis.

Why NMF - other examples

Application side

- Spectral unmixing in analytical chemistry (one of the earliest work)
- Representation learning on human face (the work that popularizes NMF)
- Topic modeling in text mining
- Probability distribution application on identification of Hidden Markov Model
- Bioinformatics : gene expression
- Time-frequency matrix decompositions for neuroinformatics
- (Non-negative) Blind source separation
- (Non-negative) Data compression
- Speech denoising
- Recommender system
- Face recognition
- Video summarization
- Forensics
- Art work conservation (identify true color used in painting)
- Medical imaging – image processing on small object
- Mid-infrared astronomy – image processing on large object
- 2 days ago : Tells whether a banana or a fish is healthy by "looking" at them

Theoretical numerical side

- A test-box for generic optimization programs : NMF is a constrained non-convex (but biconvex) problem
- Robustness analysis of algorithm
- Tensor
- Sparsity

Analytical side

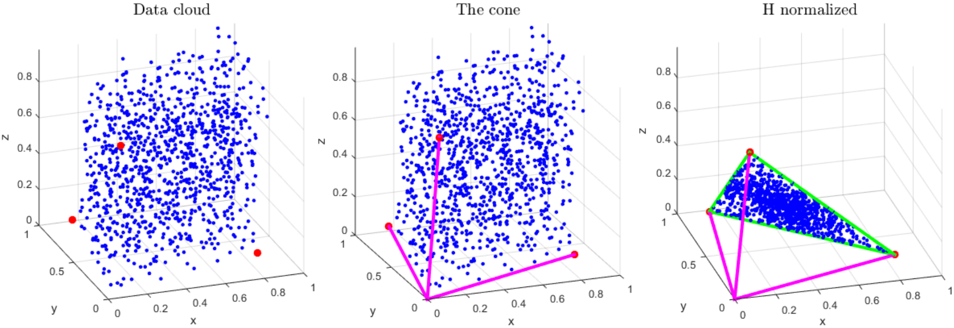
- Non-negative rank rank$^+$:= smallest r such that
$$\mathbf{X} = \sum_{i=1}^{r} \mathbf{X}_i, \quad : \ \mathbf{X}_i \text{ rank-1 and non-negative.}$$
How to find / estimate / bound rank$^+$, e.g. $\text{rank}_{\text{psd}}(\mathbf{X}) \leq \text{rank}^+(\mathbf{X})$.
- Extended formulations and combinatorics
- Log-rank Conjecture of communication system
- 3-SAT, Exponential time hypothesis, $\mathbf{P} \neq \mathbf{NP}$

Part II.
NMF geometry & Separable NMF

NMF tells a picture of a cone

Given $\mathbf{X}$, the NMF $\mathbf{X} = \mathbf{W}\mathbf{H}$ tells a picture of a

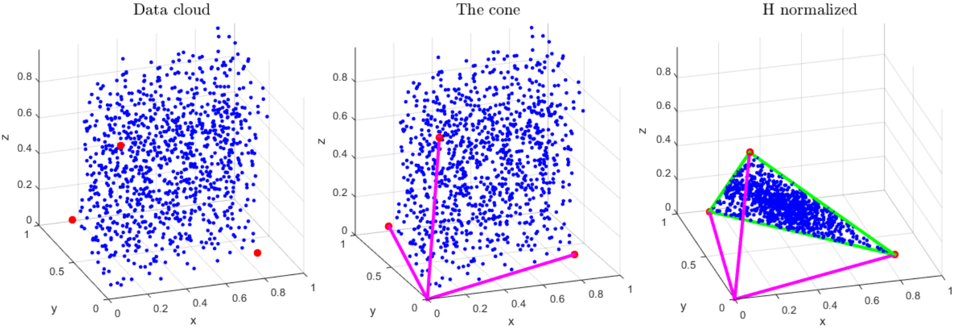
(non-negative simplicial† convex) cone.



NMF tells a picture of a cone

Given $\mathbf{X}$, the NMF $\mathbf{X} = \mathbf{W}\mathbf{H}$ tells a picture of a

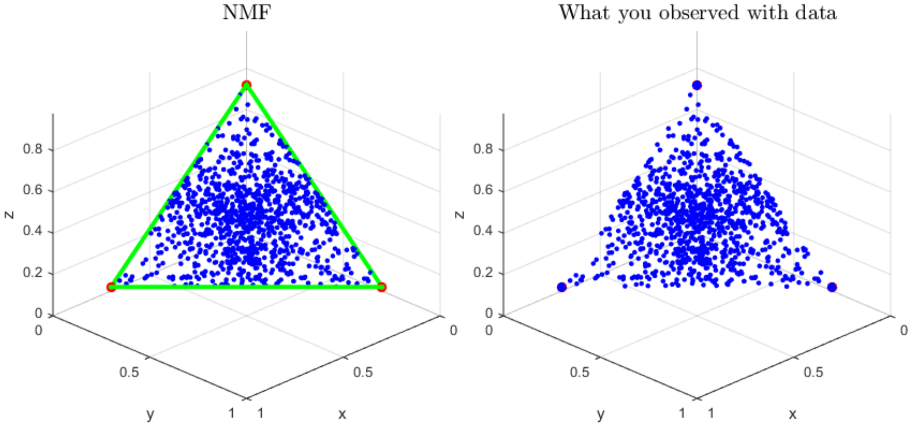
(non-negative simplicial† convex) cone.



If the columns of $\mathbf{H}$ are normalized (sum-to-1), the cone becomes (compressed into) a convex hull.

†Assumes $\mathbf{W}$ is full rank.

For $r = 3$, facing the hull we see a triangle.



NMF$_{(\mathbf{H}\text{ normalized})}$ problem geometrically means "find the vertices".

In this case, randomized NMF methods is a bad move : sub-sampling of data points remove the important points.

Separable Non-negative Matrix Factorization

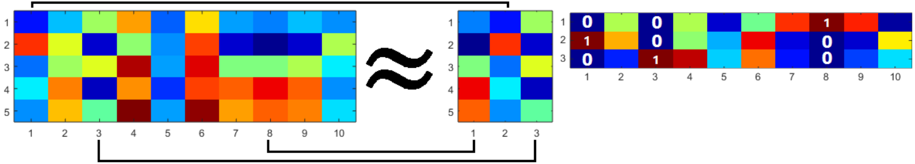
Algebra : $\mathbf{X} = \mathbf{WH}$,

- $\mathbf{W} = \mathbf{X}(:, \mathcal{J})$, $\mathcal{J}$ index set

Separable Non-negative Matrix Factorization

Algebra : $\mathbf{X} = \mathbf{W}\mathbf{H}$,

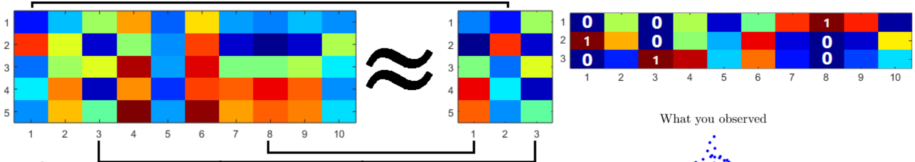
- $\mathbf{W} = \mathbf{X}(:, \mathcal{J})$, $\mathcal{J}$ index set
- $\mathbf{H} = [\mathbf{I}_r \ \mathbf{H}']\mathbf{\Pi}_r$, columns of $\mathbf{H}'$ sum-to-1.



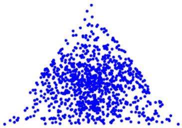
Separable Non-negative Matrix Factorization

Algebra : $\mathbf{X} = \mathbf{W}\mathbf{H}$,

- $\mathbf{W} = \mathbf{X}(:, \mathcal{J})$, $\mathcal{J}$ index set
- $\mathbf{H} = [\mathbf{I}_r \ \mathbf{H}']\mathbf{\Pi}_r$, columns of $\mathbf{H}'$ sum-to-1.



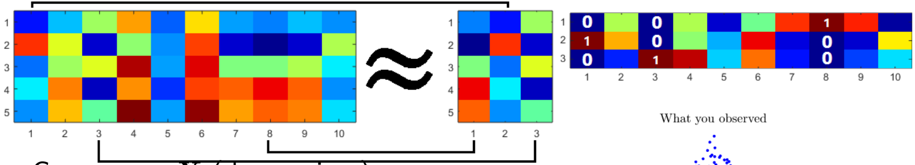
Geometry : $\mathbf{X}$ (data points) are convex combination (described by $\mathbf{H}$) of vertices ($\mathbf{W}$).



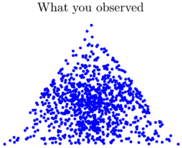
Separable Non-negative Matrix Factorization

Algebra : $\mathbf{X} = \mathbf{WH}$,

- $\mathbf{W} = \mathbf{X}(:, \mathcal{J})$, $\mathcal{J}$ index set
- $\mathbf{H} = [\mathbf{I}_r \ \mathbf{H}']\mathbf{\Pi}_r$, columns of $\mathbf{H}'$ sum-to-1.



Geometry : $\mathbf{X}$ (data points) are convex combination (described by $\mathbf{H}$) of vertices ($\mathbf{W}$).



Problem : find $\mathbf{W} \iff$ find vertices from data cloud.

- Not NP-hard anymore, solvable
- Algorithm : LP, SPA, X-ray, SNPA, ...

Separability (Donoho-Stodden, 2004)

"When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts", NIPS, 2014

Other names : pure pixel, anchord words, extreme ray, extreme point, generators.

Fast and robust algorithm for separable NMF

Problem : $[\mathbf{W},\mathbf{H}] = \underset{\mathbf{W},\mathbf{H}}{\arg\min} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F$ s.t. $\mathbf{W} = \mathbf{X}(:,\mathcal{J}), \mathbf{H} = [\mathbf{I}_r \mathbf{H}']\mathbf{\Pi}_r, \mathbf{H}'^\top \mathbf{1} \leq \mathbf{1}$.

Successive Projection Algorithm (Gillis-Vavasis, 2014)

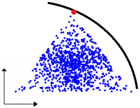
Fast and robust algorithm for separable NMF

Problem : $[\mathbf{W}, \mathbf{H}] = \underset{\mathbf{W},\mathbf{H}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F$ s.t. $\mathbf{W} = \mathbf{X}(:, \mathcal{J}), \mathbf{H} = [\mathbf{I}_r \mathbf{H}']\mathbf{\Pi}_r, \mathbf{H}'^\top \mathbf{1} \leq \mathbf{1}$.

Successive Projection Algorithm (Gillis-Vavasis, 2014)

- Step 1 : find the column in $\mathbf{X}$ with the largest norm.

  Geometry : the point furthest away has largest norm.
  Now we have $\mathbf{W} = [\mathbf{x}_1]$.



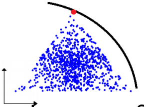
Fast and robust algorithm for separable NMF

Problem : $[\mathbf{W},\mathbf{H}] = \underset{\mathbf{W},\mathbf{H}}{\arg\min} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F$ s.t. $\mathbf{W} = \mathbf{X}(:,\mathcal{J}), \mathbf{H} = [\mathbf{I}_r \mathbf{H}']\mathbf{\Pi}_r, \mathbf{H}'^\top \mathbf{1} \leq \mathbf{1}$.

Successive Projection Algorithm (Gillis-Vavasis, 2014)

- Step 1 : find the column in $\mathbf{X}$ with the largest norm.

  Geometry : the point furthest away has largest norm.
  Now we have $\mathbf{W} = [\mathbf{x}_1]$.



- Step 2 : project the remaining columns in $\mathbf{X}$ onto the subspace of the orthogonal complement of the selected columns.

  Projection matrix : $\mathbf{I} - \dfrac{\mathbf{x}_1\mathbf{x}_1^\top}{\mathbf{x}_1^\top\mathbf{x}_1}$

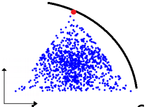
Fast and robust algorithm for separable NMF

Problem : $[\mathbf{W},\mathbf{H}] = \arg\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F$ s.t. $\mathbf{W}=\mathbf{X}(:,\mathcal{J}), \mathbf{H}=[\mathbf{I}_r\mathbf{H}']\mathbf{\Pi}_r, \mathbf{H}'^\top\mathbf{1}\leq\mathbf{1}$.

Successive Projection Algorithm (Gillis-Vavasis, 2014)

- Step 1 : find the column in $\mathbf{X}$ with the largest norm.

  Geometry : the point furthest away has largest norm.
  Now we have $\mathbf{W}=[\mathbf{x}_1]$.



- Step 2 : project the remaining columns in $\mathbf{X}$ onto the subspace of the orthogonal complement of the selected columns.

  Projection matrix : $\mathbf{I}-\dfrac{\mathbf{x}_1\mathbf{x}_1^\top}{\mathbf{x}_1^\top\mathbf{x}_1}$
- Step 3, 4, ... : repeat step 1-2, until $\mathbf{W}$ has r columns.
- How to get $\mathbf{H}$: with $(\mathbf{X},\mathbf{W})$, do a non-negative least sqaures.

Gillis-Vavasis, Fast and Robust Recursive Algorithmsfor Separable Nonnegative Matrix Factorization, 2014.

Successive Projection Algorithm (SPA)

Probably the "best" method for this kind of problem because :

Successive Projection Algorithm (SPA)

Probably the "best" method for this kind of problem because :

- **Robust**
  - It can find the vertices under bounded additive noise.
  - **Theorem.** (Gillis-Vavasis,14) If $\epsilon \leq \mathcal{O}\left(\frac{\sigma_{\mathbf{W}}^{\min}}{\sqrt{r}\kappa_{\mathbf{W}}^2}\right)$, SPA satisfies

$$\max_k \|\mathbf{W}(:,k) - \mathbf{X}(:,\mathcal{J}(k))\| \leq \mathcal{O}(\epsilon\kappa_{\mathbf{W}}^2).$$

In English : if noise is bounded, then the worse case fitting error is bounded.

Successive Projection Algorithm (SPA)

Probably the "best" method for this kind of problem because :

- **Robust**
  - It can find the vertices under bounded additive noise.
  - **Theorem.** (Gillis-Vavasis,14) If $\epsilon \leq \mathcal{O}\left(\frac{\sigma_{\mathbf{W}}^{\min}}{\sqrt{r}\kappa_{\mathbf{W}}^2}\right)$, SPA satisfies

$$\max_k \|\mathbf{W}(:,k) - \mathbf{X}(:,\mathcal{J}(k))\| \leq \mathcal{O}(\epsilon\kappa_{\mathbf{W}}^2).$$

  In English : if noise is bounded, then the worse case fitting error is bounded.
- **Fast**
  - Computing $\mathbf{W}$: just a modified Gramn-Schmidt with column pivoting
  - Computing $\mathbf{H}$: a 1st-order optimization method with Nesterov's acceleration.
- Few methods† exist that achieve both of these goals, many only one of the two."

However,

Successive Projection Algorithm (SPA)

Probably the "best" method for this kind of problem because :

- **Robust**
  - It can find the vertices under bounded additive noise.
  - **Theorem.** (Gillis-Vavasis,14) If $\epsilon \leq \mathcal{O}\left(\frac{\sigma_{\mathbf{W}}^{\min}}{\sqrt{r}\kappa_{\mathbf{W}}^2}\right)$, SPA satisfies

$$\max_k \|\mathbf{W}(:,k) - \mathbf{X}(:,\mathcal{J}(k))\| \leq \mathcal{O}(\epsilon\kappa_{\mathbf{W}}^2).$$

    In English : if noise is bounded, then the worse case fitting error is bounded.
- **Fast**
  - Computing $\mathbf{W}$: just a modified Gramn-Schmidt with column pivoting
  - Computing $\mathbf{H}$: a 1st-order optimization method with Nesterov's acceleration.
- Few methods† exist that achieve both of these goals, many only one of the two."

However, the success of SPA is based on the separability assumption :

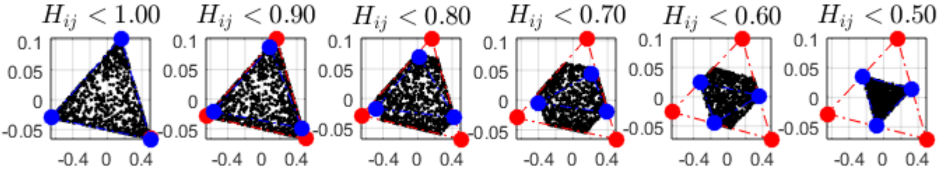
"Vertices $\mathbf{W}$ are *presented* in observed data $\mathbf{X}$"

What if this is false ?

†Two examples : SNPA and preconditioned SPA by Gillis et al.

Part III.
Volume regularized NMFs

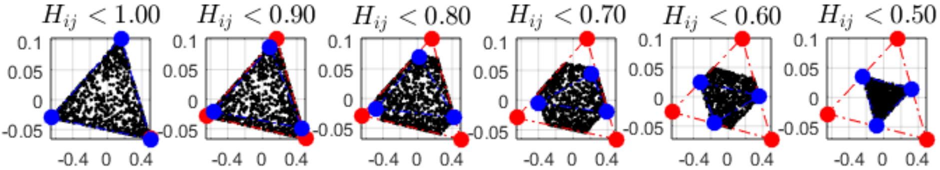
SPA fails when separability is false



Why fail : recall the first col. of $\mathbf{W}$ is extract as the col. of $\mathbf{X}$ with largest norm.

How to solve it ??

SPA fails when separability is false



Why fail : recall the first col. of $\mathbf{W}$ is extract as the col. of $\mathbf{X}$ with largest norm.

How to solve it ??

Idea : minimum volume hull fitting : Click me.

(URL : http://angms.science/eg "underscore" SNPA "underscore" ini "dot" gif)

Volume regularized NMF

Idea : fitting with minimum volume.

Problem : $[\mathbf{W}, \mathbf{H}] = \arg\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F + \lambda \mathcal{V}(\mathbf{W})$,

where $\mathcal{V}(.)$ is a prox function that measures the vol. of the cvx hull of $\mathbf{W}$.

Volume regularized NMF

Idea : fitting with minimum volume.

Problem : $[\mathbf{W}, \mathbf{H}] = \underset{\mathbf{W},\mathbf{H}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F + \lambda \mathcal{V}(\mathbf{W})$,

where $\mathcal{V}(.)$ is a prox function that measures the vol. of the cvx hull of $\mathbf{W}$.

- determinant of Gramian† $\det(\mathbf{W}^\top \mathbf{W})$
- log-determinant of Gramian† $\log\det(\mathbf{W}^\top \mathbf{W} + \delta \mathbf{I}_r)$
- rectangular box† $\prod_i^r / \sum_{i=1}^r \|\mathbf{w}_i\|_2^2$
- nuclear norm ball $\|\mathbf{W}\|_*$

Volume regularized NMF

Idea : fitting with minimum volume.

Problem : $[\mathbf{W},\mathbf{H}] = \underset{\mathbf{W},\mathbf{H}}{\arg\min} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F + \lambda \mathcal{V}(\mathbf{W})$,

where $\mathcal{V}(.)$ is a prox function that measures the vol. of the cvx hull of $\mathbf{W}$.

- determinant of Gramian† $\det(\mathbf{W}^\top\mathbf{W})$
- log-determinant of Gramian† $\log\det(\mathbf{W}^\top\mathbf{W} + \delta\mathbf{I}_r)$
- rectangular box† $\prod_i^r / \sum_{i=1}^r \|\mathbf{w}_i\|_2^2$
- nuclear norm ball $\|\mathbf{W}\|_*$

Theoretical ground on recoverability : (Lin-Ma-Chi-Ambikapathi, 2015)

"Identifiability of the Simplex Volume Minimization Criterion for Blind Hyperspectral Unmixing: The No-Pure-Pixel Case", IEEE trans. Geosci. Remote Sensing, 2015.

† On which $\mathcal{V}$ is "computationally" better : (**A**.-Gillis, 2018)
"Volume regularized non-negative matrix factorizations",

IEEE WHISPERS18, Sep23-26, 2018, Amsterdam, NL.

Open problem : **fast** and **robust** algorithm for volume regularized NMF.

What are not discussed & open problems

- **How to actually solve NMF (and solve it fast)** - algorithm design

  e.g. People now still keep using the **slow** multiplicative update

  Gillis-Glineur, "Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization", **2011**.

  A.-Gillis, "Accelerating Non-negative matrix factorization by extrapolation", to appear in *Neural Computation*, 2018.

- **Tuning of the regularization parameter λ**

  For volume regularization, λ should be small and becoming smaller.

- **Other ideas**
  - Non-negative tensor factorizations
  - NMF + Sparsity (Cohen-Gillis, 2018, submitted)
  - Non-negative rank $\text{rank}^+ :=$ smallest r such that

    $$\mathbf{X} = \sum_{i=1}^{r} \mathbf{X}_i, \quad : \ \mathbf{X}_i \text{ rank-1 and non-negative.}$$

    How to find / estimate / bound rank^+, e.g. $\text{rank}_{\text{psd}}(\mathbf{X}) \leq \text{rank}^+(\mathbf{X})$.
  - Combinatorial optimzation, extended formulations.
  - Log-rank Conjecture, Exponential time hypothesis, $\mathbf{P} \neq \mathbf{NP}$.

Last page - summary

- Non-negative Matrix Factorization.
- Why NMF.
- Geometry of NMF.
- Separable NMF.
- When separability fails : minimum volume NMF.

Ideas are simple, devils in details.

END OF PRESENTATION.

slide in *angms.science*

ACK : my boss Nicolas Gillis, European Research Council Grant #679515.