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

Find:
Matrices $W \in \mathbb{R}^{m \times r}$, $H \in \mathbb{R}^{r \times n}$ such that $X = WH$.

Important: everything is non-negative.

Notation: we use $WH$ instead of $WH^\top$. 
Non-negative Matrix Factorization (NMF)

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

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

Non-negative Matrix Factorization (NMF)

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

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

**Notation:** we use $WH$ instead of $WH^\top$. 

Always 😊
BE POSITIVE!
Exact and approximate NMF

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

\[ X = WH. \]

This is called exact NMF, NP-hard (Vavasis, 2007).
Exact and approximate NMF

Given the pair \( (X \in \mathbb{R}^{m \times n}, r \in \mathbb{N}) \), find the pair \( (W \in \mathbb{R}^{m \times r}, H \in \mathbb{R}^{r \times n}) \) such that
\[
X = WH.
\]

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

(Low-rank) approximate NMF:
\[
X \approx WH, \quad 1 \leq r \leq \min\{m, n\}.
\]

Find \((W, H)\) numerically

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

\[
[W, H] = \arg \min_{W \geq 0, H \geq 0} \|X - WH\|_F.
\]
Find \((W, H)\) numerically

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

\[
[W, H] = \arg \min_{W \geq 0, H \geq 0} \|X - WH\|_F.
\]

- Minimizing the distance\(^\dagger\) between \(X\) and the approximator \(WH\) 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 \(\geq 0\) will be skipped.

\(\dagger\)This talk does not consider other distance functions.
The scope of this talk

Given \((X, r)\), find \((W, H)\) via solving

\[
[W, H] = \arg \min_{W,H} ||X - WH||_F \text{ subject to } *,
\]

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

* 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).
The scope of this talk

Given \((X, r)\), find \((W, H)\) via solving

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

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)

\[ X(:,j) \approx \sum_{k=1}^{r} W(:,k) H(k,j) \]

- \( X(:,j) \): spectral signature of \( j \)th pixel
- \( W(:,k) \): spectral signature of \( k \)th endmember
- \( H(k,j) \): abundance of \( k \)th endmember in \( j \)th pixel

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

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

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 \( \text{rank}^+ := \text{smallest } r \text{ such that } \)

\[
X = \sum_{i=1}^{r} X_i, \quad : X_i \text{ rank-1 and non-negative.}
\]

How to find / estimate / bound \( \text{rank}^+ \), e.g. \( \text{rank}_{\text{psd}}(X) \leq \text{rank}^+(X) \).
- Extended formulations and combinatorics
- Log-rank Conjecture of communication system
- 3-SAT, Exponential time hypothesis, \( \mathbb{P} \neq \mathbb{NP} \)
Part II.
NMF geometry & Separable NMF
NMF tells a picture of a cone

Given $\mathbf{X}$, the NMF $\mathbf{X} = \mathbf{WH}$ tells a picture of a 
(non-negative simplicial† convex) cone.

† Assumes $\mathbf{W}$ is full rank.
NMF tells a picture of a cone

Given $X$, the NMF $X = WH$ tells a picture of a (non-negative simplicial\textsuperscript{†} convex) cone.

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

\textsuperscript{†}Assumes $W$ is full rank.
NMF tells a picture of a hull

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

NMF 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: $X = WH$,

- $W = X(:, \mathcal{J})$, $\mathcal{J}$ index set
Separable Non-negative Matrix Factorization

Algebra: \( X = WH \),
- \( W = X(:, \mathcal{J}) \), \( \mathcal{J} \) index set
- \( H = [I_r \ H'] \Pi_r \), columns of \( H' \) sum-to-1.

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

Problem: find \( W \) ⇐⇒ 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, anchor words, extreme ray, extreme point, generators.
Separable Non-negative Matrix Factorization

**Algebra:** \( X = WH \),
- \( W = X(:, J) \), \( J \) index set
- \( H = [I_r \ H']\Pi_r \), columns of \( H' \) sum-to-1.

**Geometry:** \( X \) (data points) are convex combination (described by \( H \)) of vertices (\( W \)).
Separable Non-negative Matrix Factorization

Algebra: \( X = WH, \)
- \( W = X(:, \mathcal{J}), \mathcal{J} \) index set
- \( H = [I_r \ H'] \Pi_r, \) columns of \( H' \) sum-to-1.

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

Problem: find \( W \iff \) 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, anchor words, extreme ray, extreme point, generators.
Fast and robust algorithm for separable NMF

Problem: \[ [W, H] = \arg \min_{W, H} \| X - WH \|_F \text{ s.t. } W = X(:, J), H = [I_r H'] \Pi_r, H'\top 1 \leq 1. \]

**Successive Projection Algorithm** (Gillis-Vavasis, 2014)
Fast and robust algorithm for separable NMF

Problem: \( [W, H] = \arg \min_{W, H} \|X - WH\|_F \) s.t. \( W = X(:, J), H = [I_r H'] \Pi_r, H'^T 1 \leq 1 \).

**Successive Projection Algorithm** (Gillis-Vavasis, 2014)

1. Step 1: find the column in \( X \) with the largest norm.
   - Geometry: the point furthest away has largest norm.
   - Now we have \( W = [x_1] \).

2. Step 2: project the remaining columns in \( X \) onto the subspace of the orthogonal complement of the selected columns.
   - Projection matrix: \( I - x_1 x_1^T \)

3. Step 3, 4, ...: repeat step 1-2, until \( W \) has \( r \) columns.

How to get \( H \): with \( (X, W) \), do a non-negative least squares.
Fast and robust algorithm for separable NMF

Problem : \[ [W, H] = \arg \min_{W, H} \|X - WH\|_F \text{ s.t. } W = X(:, J), H = [I_r H'] \Pi_r, H'^\top 1 \leq 1. \]

Successive Projection Algorithm (Gillis-Vavasis, 2014)

- Step 1 : find the column in \( X \) with the largest norm.
  
  Geometry : the point furthest away has largest norm. Now we have \( W = [x_1] \).

- Step 2 : project the remaining columns in \( X \) onto the subspace of the orthogonal complement of the selected columns.

  Projection matrix : \( I - \frac{x_1 x_1^\top}{x_1^\top x_1} \)
Fast and robust algorithm for separable NMF

Problem: \([W, H] = \arg \min_{W,H} \|X - WH\|_F \text{ s.t. } W = X(:,J), H = [I_r H']\Pi_r, H'^\top 1 \leq 1.\]

**Successive Projection Algorithm** (Gillis-Vavasis, 2014)

- **Step 1:** find the column in \(X\) with the largest norm.
  
  Geometry: the point furthest away has largest norm.
  
  Now we have \(W = [x_1]\).

- **Step 2:** project the remaining columns in \(X\) onto the subspace of the orthogonal complement of the selected columns.
  
  Projection matrix: \(I - \frac{x_1x_1^\top}{x_1^\top x_1}\)

- **Step 3, 4, ...:** repeat step 1-2, until \(W\) has \(r\) columns.

- How to get \(H\): with \((X, W)\), do a non-negative least squares.

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)

\[
\epsilon \leq O(\sigma_{\min} W \sqrt{r\kappa W})
\]

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

- Fast
  - Computing $W$:
    - Just a modified Gram-Schmidt with column pivoting
  - Computing $H$:
    - 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 $W$ are presented in observed data $X$"

What if this is false?

† Two examples: SNPA and preconditioned SPA by Gillis et al.
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 $\epsilon \leq O\left(\frac{\sigma_{\min} W}{\sqrt{r} \kappa^2 W}\right)$, SPA satisfies

$$\max_k \|W(:, k) - X(:, J(k))\| \leq O(\epsilon \kappa^2 W).$$

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 \( \epsilon \leq O\left( \frac{\sigma_{\min} W}{\sqrt{r\kappa_2 W}} \right) \), SPA satisfies
    \[
    \max_k \| W(:, k) - X(:, J(k)) \| \leq O(\epsilon \kappa_2 W).
    \]
    In English: if noise is bounded, then the worse case fitting error is bounded.

- **Fast**
  - Computing \( W \): just a modified Gram-Schmidt with column pivoting
  - Computing \( H \): 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 \( \epsilon \leq O\left(\frac{\sigma_{\min}^W}{\sqrt{r\kappa^2_W}}\right) \), SPA satisfies

\[
\max_k \|W(:, k) - X(:, J(k))\| \leq O(\epsilon \kappa^2_W).
\]

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

- **Fast**
  - Computing \( W \): just a modified Gram-Schmidt with column pivoting
  - Computing \( H \): 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 \( W \) are *presented* in observed data \( X \)”

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 $W$ is extract as the col. of $X$ with largest norm.

How to solve it??
SPA fails when separability is false

Why fail: recall the first col. of $W$ is extract as the col. of $X$ 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: \([W, H] = \arg \min_{W,H} \|X - WH\|_F + \lambda \mathcal{V}(W),\)

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

Idea: fitting with minimum volume.

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

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

- determinant of Gramian\(^\dagger\) \(\det(W^\top W)\)
- log-determinant of Gramian\(^\dagger\) \(\log \det(W^\top W + \delta I_r)\)
- rectangular box\(^\dagger\) \(\prod_{i}^{r} / \sum_{i=1}^{r} \|w_i\|_2^2\)
- nuclear norm ball \(\|W\|_*\)

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


\(^\dagger\) On which \(\mathcal{V}(.)\) is "computationally" better: (A.-Gillis, 2018)

"Volume regularized non-negative matrix factorizations", IEEE WHISPERS18, Sep23-26, 2018, Amsterdam, NL.
Volume regularized NMF

Idea: fitting with minimum volume.

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

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

- determinant of Gramian $\dagger$ $\det(W^\top W)$
- log-determinant of Gramian $\dagger$ $\log \det(W^\top W + \delta I_r)$
- rectangular box $\dagger$ $\prod_i^r / \sum_{i=1}^r \|w_i\|_2^2$
- nuclear norm ball $\|W\|_*$

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

$\dagger$ 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

- **Tuning of the regularization parameter \( \lambda \)**
For volume regularization, \( \lambda \) 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

\[
X = \sum_{i=1}^{r} X_i, \quad : \quad \text{rank-1 and non-negative.}
\]

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