Log-determinant Non-Negative Matrix Factorization via Successive Trace Approximation

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

Email: manshun.ang@umons.ac.be
Homepage: angms.science

May 23, 2018

Joint work with my supervisor: Nicolas Gillis (UMONS, Belgium)
The research problem (1/2)

Non-negative Matrix Factorization (NMF) : given

- Input matrix \( \mathbf{X} \in \mathbb{R}^{m \times n} \)
- Factorization rank : \( r \), positive integer

We consider

\[
\min_{\mathbf{W} \geq 0, \mathbf{H} \geq 0} \frac{1}{2} \| \mathbf{X} - \mathbf{WH} \|_F^2.
\]

- optimization variables : \( \mathbf{W} \in \mathbb{R}_{+}^{m \times r}, \mathbf{H} \in \mathbb{R}_{+}^{r \times n} \)
- problem is : non-convex, NP-hard, ill-posed problem
- we consider low rank/complexity model \( 1 \leq r \leq \min\{m, n\} \).
- assumptions : (1) \( \mathbf{W}, \mathbf{H} \) full rank, (2) \( r \) is known.
The research problem (2/2)

**log-det Non-negative Matrix Factorization (NMF)**: given

- Input matrix: $X \in \mathbb{R}^{m \times n}$
- Factorization rank: $r$, positive integer

We consider

$$\min_{W \geq 0, H \geq 0} \frac{1}{2} \|X - WH\|_F^2 + \frac{\lambda}{2} \log \det(W^T W + \delta I_r).$$

- Optimization variables: $W \in \mathbb{R}^{m \times r}, H \in \mathbb{R}^{r \times n}$
- We consider low rank/complexity model $1 \leq r \leq \min\{m, n\}$.
- Assumptions (1) $W, H$ full rank, (2) $r$ is known.
- $\log \det(W^T W + \delta I_r)$: volume regularizer, $\delta$ is constant
- $\lambda > 0$: regularization parameter
Problem: given \( (X, r) \), solve

\[
\min_{W \geq 0, H \geq 0} \frac{1}{2} \|X - WH\|_F^2 + \frac{\lambda}{2} \log \det(W^T W + \delta I_r).
\]

**Algorithm 1** BCD framework for logdet-NMF

1: **INPUT** : \( X \in \mathbb{R}^{m \times n}, r \in \mathbb{N}_+ \)
   
   Initialization : \( W \in \mathbb{R}^{m \times r}_+ \) and \( H \in \mathbb{R}^{r \times n}_+ \)

2: **OUTPUT** : \( W \in \mathbb{R}^{m \times r}_+ \) and \( H \in \mathbb{R}^{r \times n}_+ \)

3: for \( k = 1, 2, \ldots \) do

4:   **Update** \( (W) \) via \( \arg \min_{W \geq 0} \frac{1}{2} \|X - WH\|_F^2 + \frac{\lambda}{2} \log \det(W^T W + \delta I_r). \)

5:   **Update** \( (H) \) via \( \arg \min_{H \geq 0} \frac{1}{2} \|X - WH\|_F^2. \)

6: end for
The subproblems are not symmetric.

On \( \mathbf{W} \):

\[
\underset{\mathbf{W} \geq 0}{\text{arg min}} \frac{1}{2} \| \mathbf{X} - \mathbf{WH} \|_F^2 + \frac{\lambda}{2} \log \det(\mathbf{W}^\top \mathbf{W} + \delta \mathbf{I}_r)
\]

On \( \mathbf{H} \):

\[
\underset{\mathbf{H} \geq 0}{\text{arg min}} \frac{1}{2} \| \mathbf{X} - \mathbf{WH} \|_F^2
\]

Update(\( \mathbf{H} \)) is easier, can be solved by FGM\(^\dagger\)

\[
\mathbf{H} \leftarrow \text{FGM}(\mathbf{X}, \mathbf{W}, \mathbf{H})
\]

Update(\( \mathbf{W} \)) is harder as \( \log \det(\mathbf{W}^\top \mathbf{W} + \delta \mathbf{I}_r) \) is:

- non-convex
- column-coupled
- non-proximable

Theme of the presentation

To handle

\[
\arg\min_{W \geq 0} \frac{1}{2} \|X - WH\|_F^2 + \frac{\lambda}{2} \log \det(W^T W + \delta I_r)
\]

in

**Algorithm 2** BCD framework for logdet-NMF

1: **INPUT**: \(X \in \mathbb{R}^{m \times n}, r \in \mathbb{N}_+\)
   
   Initialization: \(W \in \mathbb{R}_+^{m \times r}\) and \(H \in \mathbb{R}_+^{r \times n}\)

2: **OUTPUT**: \(W \in \mathbb{R}_+^{m \times r}\) and \(H \in \mathbb{R}_+^{r \times n}\)

3: **for** \(k = 1, 2, \ldots\) **do**
4: \(
\text{Update}(W) \text{ via } \arg\min_{W \geq 0} \frac{1}{2} \|X - WH\|_F^2 + \frac{\lambda}{2} \log \det(W^T W + \delta I_r).
\)
5: \(H \leftarrow \text{FGM}(X, W, H)\).
6: **end for**
Why: Mathematician don’t ask why, just want to solve it
- Nobody solve it *effectively* yet

Application from data science: for $X = WH$
- $X = \text{data}$
- $W = \text{basis}$
- $H = \text{coefficient, weighting, encoding, membership}$
- $r = \text{model complexity}$
- $m = \text{dimension of data}$
- $n = \# \text{ of data points}$
Application from data science: for $\mathbf{X} = \mathbf{WH}$

- $\mathbf{X}$ = data
- $\mathbf{W}$ = basis
- $\mathbf{H}$ = coefficient, weighting, encoding, membership
- $m$ = dimension of data
- $n$ = # of data points
- $r$ = model complexity

Matrix $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{H}$ with $m = 5$, $n = 7$, $r = 2$
Application from data science: for $X = WH$

- $X =$ data
- $W =$ basis
- $H =$ coefficient, weighting, encoding, membership
- $m =$ dimension of data
- $n =$ # of data points
- $r =$ model complexity

Consider 3rd column.
Motivation (2/4)

Application from data science: for \( \mathbf{X} = \mathbf{WH} \)

- \( \mathbf{X} = \) data
- \( \mathbf{W} = \) basis
- \( \mathbf{H} = \) coefficient, weighting, encoding, membership
- \( m = \) dimension of data
- \( n = \# \) of data points
- \( r = \) model complexity

Consider 3rd column
Application from data science: for $X = WH$

- $X = \text{data}$
- $W = \text{basis}$
- $H = \text{coefficient, weighting, encoding, membership}$
- $m = \text{dimension of data}$
- $n = \# \text{ of data points}$
- $r = \text{model complexity}$

Linear combination:
Motivation (2/4)

Application from data science: for $X = WH$

- $X = \text{data}$
- $W = \text{basis}$
- $H = \text{coefficient, weighting, encoding, membership}$
- $m = \text{dimension of data}$
- $n = \# \text{ of data points}$
- $r = \text{model complexity}$

Non-negativity: conic combination
Motivation (2/4)

Application from data science: for $X = WH$

- $X = \text{data}$
- $W = \text{basis}$
- $H = \text{coefficient, weighting, encoding, membership}$
- $m = \text{dimension of data}$
- $n = \# \text{ of data points}$
- $r = \text{model complexity}$

Non-negativity + normalization: convex combination
Application from data science: for $X = WH$

- $X =$ data
- $W =$ vertex / generator
- $H =$ membership
- $m =$ dimension of data
- $n =$ # of data points
- $r =$ number of basis

Geometrically: given data points, fit a convex hull
Application from data science: for $X = WH$

- $X =$ data
- $W =$ vertex / generator
- $H =$ membership
- $m =$ dimension of data
- $n =$ # of data points
- $r =$ number of basis

Geometrically: given data points, fit a **minimum volume** convex hull
Motivation (4/4)

Optimization problem

\[
\begin{align*}
\min_{W \geq 0, H \geq 0, 1^T H \leq 1_n} & \quad \frac{1}{2} \| X - WH \|^2_F \\
+ & \quad \frac{\lambda}{2} \log \det(W^T W + \delta I_r)
\end{align*}
\]

Geometry: fit a min. vol. convex hull

Many applications
Application 1: Hyper-spectral imaging

Geoinformatic application: decomposition of hyper spectral images

Output: $W$ spectral basis, $H$ weighting maps

Input: hyper-spectral images data cube
Application 2: Source Separation (e.g. music)

Music input

**Input Audio Signal x**

**Spectrogram: Representation Frequency vs Time of Input Signal**
Application 2: Source Separation (e.g. music)

Decomposition output

Basis Vectors

\( W \)

\( H \)
On solving the problem

Optimization problem

$$\begin{align*}
\min_{W \succeq 0, \quad H \succeq 0} & \quad \frac{1}{2} \|X - WH\|_F^2 \\
& \quad \frac{\lambda}{2} \log \det(W^T W + \delta I_r) \\
& \quad \frac{1}{r} H \leq 1_n
\end{align*}$$

The key to solve such problem is to use majorization-minimization

Geometry: fit a min. vol. convex hull
How to solve The key inequality : logdet-trace inequality

Given a positive definite matrix $A \in \mathbb{R}^{r \times r}$, we have
$log \det A \leq \text{tr}(A - I_r)$.

So we now have an upper bound : put $A = W^\top W + \delta I_r$

$$
\log \det(W^tW + \delta I_r) \leq \text{tr}(W^\top W + (\delta - 1)I_r)
$$

$$
\|X - WH\|_F^2 + \lambda \log \det(W^\top W + \delta I_r) \leq \|X - WH\|_F^2 + \lambda \text{tr}(W^\top W + (\delta - 1)I_r)
$$

- $\log \det(W^\top W + \delta I_r)$ is not convex w.r.t. $W$ but the trace is.
How to solve The key inequality: logdet-trace inequality

Given a positive definite matrix $A \in \mathbb{R}^{r \times r}$, we have
$$\log \det A \leq \text{tr}(A - I_r).$$
So we now have an upper bound: put $A = W^T W + \delta I_r$

$$\log \det(W^t W + \delta I_r) \leq \text{tr}(W^T W + (\delta - 1)I_r)$$
$$\|X - WH\|_F^2 + \lambda \log \det(W^T W + \delta I_r) \leq \|X - WH\|_F^2 + \lambda \text{tr}(W^T W + (\delta - 1)I_r)$$

- $\log \det(W^T W + \delta I_r)$ is not convex w.r.t. $W$ but the trace is.
- Algorithm that minimizes this upper bound:

1. for $k = 1$ to itermax do
2. $W \leftarrow \arg \min_{W \geq 0} \|X - WH\|_F^2 + \lambda \text{tr}(W^T W + (\delta - 1)I_r)$.
3. $H \leftarrow \text{FGM}(X, W, H)$.
4. end for

- Don’t stop here, it can be better !!
A closer look on the logdet-trace inequality

- Given a positive definite matrix $A \in \mathbb{R}^{r \times r}$, we have

$$\log \det A \leq \text{tr}(A - I_r).$$
A closer look on the logdet-trace inequality

- Given a positive definite matrix $A \in \mathbb{R}^{r \times r}$, we have
  \[
  \log \det A \leq \text{tr}(A - I_r).
  \]

- Let $\mu$ denotes eigenvalues, we have
  \[
  \det A = \prod_{i} \mu_i \quad \text{and} \quad \text{tr} A = \sum_{i} \mu_i.
  \]

  $\implies$

  \[
  \log \det A \leq \text{tr}(A - I_r) \iff \sum_{i} \log \mu_i \leq \sum_{i} (\mu_i - 1).
  \]

- $\log \mu_i$ means matrix $A$ has to be positive definite ($\mu_i > 0 \ \forall \ i$), which is satisfied for $A = W^\top W + \delta I_r$. 
A closer look on the logdet-trace inequality

- Given a positive definite matrix $A \in \mathbb{R}^{r \times r}$, we have

  \[ \log \det A \leq \text{tr}(A - I_r). \]

- Let $\mu$ denotes eigenvalues, we have

  \[ \det A = \prod_{i} \mu_i \quad \text{and} \quad \text{tr} A = \sum_{i} \mu_i. \]

- $\Rightarrow$

  \[ \log \det A \leq \text{tr}(A - I_r) \iff \sum_{i} \log \mu_i \leq \sum_{i} (\mu_i - 1). \]

- $\log \mu_i$ means matrix $A$ has to be positive definite ($\mu_i > 0 \forall i$), which is satisfied for $A = W^T W + \delta I_r$.

- $\sum_{i} \log \mu_i \leq \sum_{i} (\mu_i - 1) \iff \log \mu_i \leq \mu_i - 1 \forall i$. We can focus on the inequality $\log \mu_i \leq \mu_i - 1$ with $\mu_i \geq 0$. 


On $\log x \leq x - 1$, $x \geq 0$

- $\log x$ is concave.
- $x - 1$ is the first order Taylor approximation of $\log x$ at $x = 1$.
- $x - 1$ is the only **convex-tight** upper bound of $\log x$.
- Tight: $x - 1$ touch $\log x$ at the point $x = 1$.

† Higher order Taylor approximation of $\log x$ is tight, more accurate but not convex.
On $\log x \leq x - 1$, $x \geq 0$

- $\log x$ is concave.
- $x - 1$ is the first order Taylor approximation of $\log x$ at $x = 1$.
- $x - 1$ is the only **convex-tight** upper bound of $\log x$.†
- Tight: $x - 1$ touch $\log x$ at the point $x = 1$.
- Generalize to point $x_0$: $\log x \leq g(x|x_0) = a_1(x_0)x + a_0(x_0)$ is

$$
\log x \leq \frac{1}{x_0}x + \log x_0 - 1.
$$

† Higher order Taylor approximation of $\log x$ is tight, more accurate but not convex.
A parametric trace upper bound for $\log \det A$

\[
\log \det A = \sum \log \mu_i \\
\leq \sum \frac{1}{\mu_i} \mu_i + \log \mu_i^- - 1 \\
\leq \sum \frac{1}{\mu_{\min}} \mu_i + \log \mu_i^- - 1 \\
= \text{tr}(D^1 A + D^0)
\]

$D^1 = \frac{1}{\mu_{\min}} I_r$, $D^0 = \text{Diag}(\log \mu_i^- - 1)$, $\mu_i^-$ is $\mu_i$ of the previous step

Put $A = W^\top W + \delta I_r$, we have

\[
\log \det(W^\top W + \delta I_r) \leq \text{tr}(D^1 W^\top W + \delta D^1 + D^0) \\
\text{(ignore constants)} = \text{tr} D^1 W^\top W
\]
Comparing upper bounds for $\log \det A$

The original function $F(W) = \log \det(W^\top W + \delta I_r)$ is upper bounded by:

- Eigen bound $(B_1)$: $\text{tr } D^1 W^\top W + \text{constants}$.
- Taylor bound $(B_2)$: $\text{tr } D^{\text{Taylor}} W^\top W + \text{constants}$.

- Constants $D^1$, $D^0$ are defined as before, and constant $D^{\text{Taylor}} = (W_{-1}^\top W_{-1} + \delta I_r)^{-1}$.

- Both bounds are trace functional with an relaxation gap:
  - $(B_1)$ has eigen gap $\mu_i \geq \mu_{\text{min}}$
  - $(B_2)$ has convexification gap
  - $D^1$ is diagonal but $D^{\text{Taylor}}$ is not (it is dense) $\implies$ column-wise decomposition is possible
Algorithm 3 Successive Trace Approximation

1: INPUT: $X \in \mathbb{R}_+^{m \times n}$, $r \in \mathbb{N}_+$, $\lambda > 0$, $\delta > 0$.
2: OUTPUT: $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$.
3: INITIALIZATION: $W \in \mathbb{R}_+^{m \times r}$, $H \in \mathbb{R}_+^{r \times n}$, $D^1 = I_r$.
4: for $K = 1$ to itermax do
5:  for $k = 1$ to itermax do
6:     $W \leftarrow \arg\min_{W \succeq 0} \|X - WH\|_F^2 + \lambda \operatorname{tr} D^1 W^\top W.$
7:     $H \leftarrow \operatorname{FGM}(X, W, H)$.
8:  end for
9:  $\mu_i \leftarrow \operatorname{svd}(W^\top W + \delta I_r)$, $D^1 = \operatorname{Diag}(\mu_{\min}^{-1})$
10: end for
Improving STA

The STA algorithm with decomposed $f(w_i)$

**Algorithm 4 Successive Trace Approximation**

1. **INPUT:** $X \in \mathbb{R}^{m \times n}$, $r \in \mathbb{N}_+$, $\lambda > 0$ and $\delta > 0$
2. **OUTPUT:** $W \in \mathbb{R}_{+}^{m \times r}$ and $H \in \mathbb{R}_{+}^{r \times n}$
3. **INITIALIZATION:** $W \in \mathbb{R}_{+}^{m \times r}$, $H \in \mathbb{R}_{+}^{r \times n}$ and $D^1 = I_r$, $\gamma = 10^{-6}$
4. **for** $K = 1$ to itermax **do**
5.     **for** $k = 1$ to itermax **do**
6.         **for** $i = 1$ to $r$ **do**
7.             $w_i = \arg\min_{w_i \geq 0} f(w_i) = \|X_i - w_i h_i\|_F^2 + \lambda D^1_{ii}\|w_i\|_2^2 + \frac{\gamma}{2}\|w_i - w_i^-\|_2^2$
8.             $H \leftarrow \text{FGM}(X, W, H)$.
9.         **end for**
10. **end for**
11. $\mu_i \leftarrow \text{svd}(W^\top W + \delta I)$ and $D^1 = \text{Diag}(\mu_{\min}^{-1})$
12. **end for**
Finally we have

\textbf{Algorithm 5 STA, Final form}

1: INPUT: $X \in \mathbb{R}^{m \times n}$, $r \in \mathbb{N}_+$, $\lambda > 0$ and $\delta > 0$

2: OUTPUT: $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$

3: INITIALIZATION: $W \in \mathbb{R}^{m \times r}$, $H \in \mathbb{R}^{r \times n}$ and $D^1 = I_r$, $\gamma = 10^{-6}$

4: for $K = 1$ to itermax do

5: for $k = 1$ to itermax do

6: $P = XH^\top$ and $Q = HH^\top$.

7: for $i = 1$ to $r$ do

8: $w_i = \left[ P_i - \sum_{j=1}^{i-1} w_j Q_{ji} - \sum_{j=i+1}^{r} w_j^- Q_{ji} + \gamma w_i^- \right]_+$

9: $\mu_i \leftarrow \text{svd}(W^\top W + \delta I_r)$ and $D^1 = \text{Diag}(\mu_{\min}^{-1})$

10: end for

11: $H \leftarrow \text{FGM}(X, W, H)$.

12: end for

13: end for
Experimental result: comparing the two logdet inequality
Experimental result: comparing the two logdet inequality
- Nonegative Matrix Factorization with logdet regularizer
- Algorithmic development of solving the logdet NMF

Slides (and code) available at angms.science

– END OF PRESENTATION –