Robust Near-Seperable Nonnegative Matrix Factorization Using Linear Optimization

Nicolas Gillis

nicolas.gillis@uclouvain.be
https://sites.google.com/site/nicolasgillis/

ICTEAM Institute, Université catholique de Louvain

Joint work with Robert Luce (T.U. Berlin)

ROKS 2013

July 10, 2013
Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$\min_{U \geq 0, V \geq 0} ||M - UV||^2_F = \sum_{i,j} (M - UV)_{ij}^2. \quad \text{(NMF)}$$

NMF is a linear dimensionality reduction technique for nonnegative data:

$$M(:, i) \approx \sum_{k=1}^r U(:, k) V(k, i) \quad \text{for all } i.$$  

Why nonnegativity?

→ **Interpretability**: Nonnegativity constraints lead to a sparse and part-based representation.

→ **Many applications**: Text mining, hyperspectral unmixing, image processing, community detection, clustering, etc.
Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$
\min_{U \geq 0, V \geq 0} \| M - UV \|_F^2 = \sum_{i,j} (M - UV)_{ij}^2.
$$

(NMF)

NMF is a linear dimensionality reduction technique for nonnegative data:

$$
M(:,i) \approx \sum_{k=1}^{r} U(:,k) V(k,i) \quad \text{for all } i.
$$

Why nonnegativity?

→ **Interpretability**: Nonnegativity constraints lead to a sparse and part-based representation.

→ **Many applications.** Text mining, hyperspectral unmixing, image processing, community detection, clustering, etc.
Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$\min_{U \geq 0, V \geq 0} \| M - UV \|_F^2 = \sum_{i,j} (M - UV)_{ij}^2.$$  \hspace{1cm} \text{(NMF)}

NMF is a linear dimensionality reduction technique for nonnegative data:

$$M(:,i) \approx \sum_{k=1}^{r} U(:,k) \cdot V(k,i) \quad \text{for all } i.$$

Why nonnegativity?

$\rightarrow$ **Interpretability**: Nonnegativity constraints lead to a sparse and part-based representation.

$\rightarrow$ **Many applications.** Text mining, hyperspectral unmixing, image processing, community detection, clustering, etc.
Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$
\min_{U \geq 0, V \geq 0} \| M - UV \|_F^2 = \sum_{i,j} (M - UV)_{ij}^2.
$$

(NMF)

NMF is a linear dimensionality reduction technique for nonnegative data :

$$
M(:,i) \approx \sum_{k=1}^{r} U(:,k) V(k,i) \quad \text{for all } i.
$$

Why nonnegativity?

→ **Interpretability**: Nonnegativity constraints lead to a sparse and part-based representation.

→ **Many applications**: Text mining, hyperspectral unmixing, image processing, community detection, clustering, etc.
Text Mining

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

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

\[
M = \text{Dictionary} \Rightarrow \ldots \ast \ldots = U V
\]

Sets of words found simultaneously in different texts
Text Mining

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

\[ M = \text{Dictionary} \ast \cdots \ast \cdots = U V \]

Sets of words found simultaneously in different texts

ROKS 2013 Robust Separable NMF
Hyperspectral Unmixing

Hyperspectral data cube of Ludwigsburg (Germany) acquired with the imaging spectrometer HyMap©

Figure: Hyperspectral image.

Goal. Recover the endmembers and their abundances.

Model. Linear mixing model.
Hyperspectral Unmixing

Hyperspectral data cube of Ludwigsburg (Germany) acquired with the imaging spectrometer HyMap®

**Figure:** Hyperspectral image.

**Goal.** Recover the endmembers and their abundances.

**Model.** Linear mixing model.
Hyperspectral Unmixing

- Basis elements allow to recover the different materials;
- Weights allow to know which pixel contains which material.
Hyperspectral Unmixing

- Basis elements allow to recover the different materials;
- Weights allow to know which pixel contains which material.

\[ M = \text{wavelengths} \quad \cdots \quad \approx \quad \cdots \quad \ast \quad \cdots \quad \text{pixels} \quad = \quad U \quad V \]

Spectral signatures of each constitutive material
Hyperspectral Unmixing

- Basis elements allow to recover the different materials;
- Weights allow to know which pixel contains which material.
Hyperspectral Unmixing

Figure: Urban dataset.
Hyperspectral Unmixing

Figure: Urban dataset.
Hyperspectral Unmixing

Figure: Urban dataset.
Can we only solve NMF problems?

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$\min_{U \geq 0, V \geq 0} ||M - UV||_F^2 = \sum_{i,j} (M - UV)_{ij}^2.$$  \hspace{1cm} (NMF)

- NMF is NP-hard [V09], and highly ill-posed.
- In practice, it is often satisfactory to use locally optimal solutions for further analysis of the data. In other words, heuristics often solve the problem efficiently with acceptable answers.
- Try to analyze this state of affairs by considering generative models and algorithms that can recover hidden data.

Can we only solve NMF problems?

Given a matrix $M \in \mathbb{R}^{m \times n}$ and a factorization rank $r \in \mathbb{N}$, find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

$$
\min_{U \geq 0, V \geq 0} \| M - UV \|_F^2 = \sum_{i,j} (M - UV)_{ij}^2. \quad \text{(NMF)}
$$

- NMF is NP-hard [V09], and highly ill-posed.
- In practice, it is often satisfactory to use locally optimal solutions for further analysis of the data. In other words, heuristics often solve the problem efficiently with acceptable answers.
- Try to analyze this state of affairs by considering generative models and algorithms that can recover hidden data.

Separability Assumption

For NMF, it is possible to compute optimal solutions in polynomial time, given that the input data matrix $M$ satisfies a (rather strong) condition: separability [AGKM12].

The nonnegative matrix $M$ is $r$-separable if and only if

\[
\text{there exists an NMF } (U, V) \geq 0 \text{ of rank } r \text{ with } M = UV \text{ where each column of } U \text{ is equal to a column of } M.
\]

Separability Assumption

For NMF, it is possible to compute optimal solutions in polynomial time, given that the input data matrix $M$ satisfies a (rather strong) condition: separability [AGKM12].

The nonnegative matrix $M$ is $r$-separable if and only if

\[
\text{there exists an NMF } (U, V) \geq 0 \text{ of rank } r \text{ with } M = UV \text{ where each column of } U \text{ is equal to a column of } M.
\]

Is separability a reasonable assumption?

◊ **Hyperspectral unmixing**: separability is particularly natural: for each constitutive material, there is a ‘pure’ pixel containing only that material. This is the so called pure-pixel assumption which is widely used in hyperspectral imaging.

◊ **Text mining**: for each topic, there is a ‘pure’ document on that topic, or, for each topic, there is a ‘pure’ word (an anchor word) used only by that topic.


Is separability a reasonable assumption?

- **Hyperspectral unmixing**: separability is particularly natural: for each constitutive material, there is a ‘pure’ pixel containing only that material. This is the so called pure-pixel assumption which is widely used in hyperspectral imaging.

- **Text mining**: for each topic, there is a ‘pure’ document on that topic, or, for each topic, there is a ‘pure’ word (an anchor word) used only by that topic.

Geometric Interpretation of Separability

After normalization, the columns of $M$, $U$ and $V$ sum to one: the columns of $U$ are the vertices of the convex hull of the columns of $M$. 

![Diagram showing geometric interpretation of separability]
Separable NMF

\[ M \text{ is } r\text{-separable} \iff M = U[I_r, V']\Pi, \]

for some \( V' \geq 0 \), and some permutation matrix \( \Pi \).
Near-Separable NMF

\[ \tilde{M} = U[I_r, V'] \Pi + N, \text{ where } N \text{ is noise.} \]
Near-Separable NMF: Noise and Conditioning

We will assume that the noise is bounded (but otherwise arbitrary):

\[ ||N(:, i)||_1 \leq \epsilon, \quad \text{for all } i, \]

and some dependence on some condition number is unavoidable:

Parameter \( \alpha = \) minimum distance of a vertex to the convex hull of other vertices.
Near-Separable NMF: Noise and Conditioning

We will assume that the noise is bounded (but otherwise arbitrary):

\[ ||N(:, i)||_1 \leq \epsilon, \quad \text{for all } i, \]

and some dependence on some condition number is unavoidable:

Parameter \( \alpha \) = minimum distance of a vertex to the convex hull of other vertices.
Near-Separable NMF: Noise and Conditioning

We will assume that the noise is bounded (but otherwise arbitrary):

\[ ||N(:,i)||_1 \leq \epsilon, \quad \text{for all } i, \]

and some dependence on some condition number is unavoidable:

Parameter \( \alpha = \) minimum distance of a vertex to the convex hull of other vertices.
**Hottopixx, a Linear Optimization Model**

For a normalized separable matrix $M$, we have, up to permutation,

$$M = [U, U V'] = M \begin{pmatrix} I_r & V' \\ 0_{(n-r)\times r} & 0_{(n-r)\times(n-r)} \end{pmatrix} = M X^0.$$

where $V' \leq 1_{r\times(n-r)}$. [BRRT12] proposed the following model:

$$\min_{X \in \mathbb{R}^{n \times n}} \quad p^T \text{diag}(X)$$

such that

$$\|\tilde{M} - \tilde{M} X\|_1 \leq 2\epsilon,$$
$$\text{tr}(X) = r,$$
$$0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.$$

where the entries of $p$ are distinct.

Hottopixx, a Linear Optimization Model

For a normalized separable matrix $M$, we have, up to permutation,

$$M = [U, UV'] = M \begin{pmatrix} I_r & V' \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{pmatrix} = MX^0.$$  

where $V' \leq 1_{r \times (n-r)}$. [BRRT12] proposed the following model:

$$\min_{X \in \mathbb{R}^{n \times n}} p^T \text{diag}(X)$$

such that

$$\|\tilde{M} - \tilde{M}X\|_1 \leq 2\epsilon,$$

$$\text{tr}(X) = r,$$

$$0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.$$

where the entries of $p$ are distinct.

For a normalized separable matrix $M$, we have, up to permutation,

$$M = [U, UV'] = M \begin{pmatrix} I_r & V' \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{pmatrix} = MX^0.$$

where $V' \leq 1_{r \times (n-r)}$. [BRRT12] proposed the following model:

$$\min_{X \in \mathbb{R}^{n \times n}} \text{subject to } \left\| \tilde{M} - \tilde{M}X \right\|_1 \leq 2\epsilon,$$

$$\text{tr}(X) = r,$$

$$0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.$$
Hottopixx, a Linear Optimization Model

For a normalized separable matrix $M$, we have, up to permutation,

\[
M = [U, UV'] = M \begin{pmatrix} I_r & V' \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{pmatrix}_{X^0 \in \mathbb{R}^{n \times n}} = MX^0.
\]

where $V' \leq 1_{r \times (n-r)}$. [BRRT12] proposed the following model:

\[
\begin{align*}
\min_{X \in \mathbb{R}^{n \times n}} & \quad p^T \text{diag}(X) \\
\text{such that} & \quad \|\tilde{M} - \tilde{M}X\|_1 \leq 2\epsilon, \\
& \quad \text{tr}(X) = r, \\
& \quad 0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.
\end{align*}
\]

where the entries of $p$ are distinct.

Hottopixx, a Linear Optimization Model

For a normalized separable matrix $M$, we have, up to permutation,

$$M = [U, UV'] = M \begin{pmatrix} I_{r} & V' \\ 0_{(n-r) \times r} & 0_{(n-r) \times (n-r)} \end{pmatrix} = MX^0,$$

where $V' \leq 1_{r \times (n-r)}$. [BRRT12] proposed the following model:

$$\min_{X \in \mathbb{R}^{n \times n}} \quad p^T \text{diag}(X)$$

such that

$$||\tilde{M} - \tilde{M}X||_1 \leq 2\epsilon,$$

$$\text{tr}(X) = r,$$

$$0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.$$

where the entries of $p$ are distinct.

Hottopixx, a Linear Optimization Model

\[
\min_{X \in \mathbb{R}^{n \times n}} \quad p^T \text{diag}(X)
\]

such that

\[
||\tilde{M} - \tilde{M}X||_1 \leq 2\epsilon,
\]

\[
\text{tr}(X) = r,
\]

\[
0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.
\]

**Theorem** ([G12]). If \( \epsilon \leq \mathcal{O}\left(\frac{\alpha^2}{r}\right) \), their algorithm leads to an NMF \((W, H)\) s.t.

\[
||\tilde{M} - UV||_1 \leq \mathcal{O}\left(\frac{r\epsilon}{\alpha}\right).
\]

**Drawbacks.** Requires to solve a LP in \( \mathcal{O}(n^2) \) variables, the parameters \( \epsilon \) and \( r \) have to be estimated, not very robust in practice, normalization is necessary.

Hottopixx, a Linear Optimization Model

\[
\min_{X \in \mathbb{R}^{n \times n}} \quad p^T \text{diag}(X)
\]

such that
\[
\|\tilde{M} - \tilde{M}X\|_1 \leq 2\epsilon,
\]
\[
\text{tr}(X) = r,
\]
\[
0 \leq X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j.
\]

**Theorem** ([G12]). If \(\epsilon \leq O\left(\frac{\alpha^2}{r}\right)\), their algorithm leads to an NMF \((W, H)\) s.t.
\[
\|\tilde{M} - UV\|_1 \leq O\left(\frac{r\epsilon}{\alpha}\right).
\]

**Drawbacks.** Requires to solve a LP in \(O(n^2)\) variables, the parameters \(\epsilon\) and \(r\) have to be estimated, not very robust in practice, normalization is necessary.

An Improved Linear Optimization Model

\[
\begin{align*}
\min_{X \in \mathbb{R}_{+}^{n \times n}} & \quad p^T \text{diag}(X) \\
\text{such that} & \quad ||\tilde{M} - \tilde{M}X||_1 \leq 2\epsilon, \\
& \quad \text{tr}(X) = r, \\
& \quad X_{ij} \leq X_{ii} \text{ for all } i, j, \\
& \quad X_{ii} \leq 1 \text{ for all } i,
\end{align*}
\]

where \( p \) is a vector with positive entries.

The new model detects the factorization rank \( r \) automatically.

Same robustness analysis as Hottopixx applies for any \( \rho > 0 \).

Does not require column normalization.

If the columns of \( U \) are isolated: \( \epsilon \leq \mathcal{O}(\alpha) \Rightarrow ||\tilde{M} - UV||_1 \leq \mathcal{O}(\epsilon) \), which is provably more robust than Hottopixx for which \( \epsilon \leq \mathcal{O}\left(\frac{\alpha}{r}\right) \).

An Improved Linear Optimization Model

\[
\min_{X \in \mathbb{R}^{n \times n}_+} \quad p^T \text{diag}(X)
\]

such that

\[
\|\tilde{M} - \tilde{M}X\|_1 \leq 2\epsilon,
\]

\[
\text{tr}(X) = r,
\]

\[
X_{ij} \leq X_{ii} \text{ for all } i, j,
\]

\[
X_{ii} \leq 1 \text{ for all } i,
\]

where \(p\) is a vector with positive entries.

The new model detects the factorization rank \(r\) automatically.

Same robustness analysis as Hottopixx applies for any \(\rho > 0\).

Does not require column normalization.

If the columns of \(U\) are isolated: \(\epsilon \leq O(\alpha) \Rightarrow \|\tilde{M} - UV\|_1 \leq O(\epsilon)\),

which is provably more robust than Hottopixx for which \(\epsilon \leq O\left(\frac{\alpha}{r}\right)\).

An Improved Linear Optimization Model

\[ \min_{X \in \mathbb{R}^{n \times n}_+} p^T \text{diag}(X) \]

such that

\[ \| \tilde{M} - \tilde{M}X \|_1 \leq \rho \epsilon, \]

\[ \text{tr}(X) = r, \]

\[ X_{ij} \leq X_{ii} \text{ for all } i, j, \]

\[ X_{ii} \leq 1 \text{ for all } i, \]

where \( p \) is a vector with positive entries.

The new model detects the factorization rank \( r \) automatically.

Same robustness analysis as Hottopixx applies for any \( \rho > 0 \).

Does not require column normalization.

If the columns of \( U \) are isolated:

\[ \epsilon \leq \mathcal{O}(\alpha) \Rightarrow \| \tilde{M} - UV \|_1 \leq \mathcal{O}(\epsilon), \]

which is provably more robust than Hottopixx for which \( \epsilon \leq \mathcal{O}\left(\frac{\alpha}{r}\right)\).

An Improved Linear Optimization Model

\[
\min_{X \in \mathbb{R}^{n \times n}} \quad p^T \text{diag}(X)
\]

such that
\[
\|\tilde{M} - \tilde{M}X\|_1 \leq \rho \epsilon,
\]
\[
\text{tr}(X) = r,
\]
\[
\|\tilde{M}(:, i)\|_1 X_{ij} \leq \|\tilde{M}(:, j)\|_1 X_{ii} \quad \text{for all } i, j,
\]
\[
X_{ii} \leq 1 \quad \text{for all } i,
\]

where \( p \) is a vector with positive entries.

The new model detects the factorization rank \( r \) automatically.

Same robustness analysis as Hottopixx applies for any \( \rho > 0 \).

Does not require column normalization.

If the columns of \( U \) are isolated: \( \epsilon \leq O(\alpha) \Rightarrow \|\tilde{M} - UV\|_1 \leq O(\epsilon) \),

which is provably more robust than Hottopixx for which \( \epsilon \leq O\left(\frac{\alpha}{r}\right) \).

An Improved Linear Optimization Model

\[
\min_{X \in \mathbb{R}^{n \times n}_+} \quad p^T \text{diag}(X)
\]

such that

\[
||\tilde{M} - \tilde{M}X||_1 \leq \rho \epsilon,
\]

\[
\text{tr}(X) = r,
\]

\[
||\tilde{M}(:,i)||_1X_{ij} \leq ||\tilde{M}(:,j)||_1X_{ii} \text{ for all } i, j,
\]

\[
X_{ii} \leq 1 \text{ for all } i,
\]

where \( p \) is a vector with positive entries.

The new model detects the factorization rank \( r \) automatically.

Same robustness analysis as Hottopixx applies for any \( \rho > 0 \).

Does not require column normalization.

If the columns of \( U \) are isolated: \( \epsilon \leq O(\alpha) \Rightarrow ||\tilde{M} - UV||_1 \leq O(\epsilon) \),

which is provably more robust than Hottopixx for which \( \epsilon \leq O\left(\frac{\alpha}{r}\right) \).

Numerical Experiments

- Each entry of $U \in \mathbb{R}^{50 \times 10}_+$ uniform in $[0, 1]$; each column normalized.
- Each of the 90 columns of $V' \in \mathbb{R}^{10}_+$, Dirichlet.

**Figure:** Noise is Gaussian.
Numerical Experiments

◊ Each entry of $U \in \mathbb{R}_{+}^{50 \times 10}$ uniform in $[0, 1]$; each column normalized.
◊ Each of the 90 columns of $V' \in \mathbb{R}_{+}^{10}$, Dirichlet.

**Figure:** Noise is sparse (75%), non-zero entries are Gaussian.
Numerical Experiments

- Each entry of $U \in \mathbb{R}_{+}^{50 \times 10}$ uniform in $[0, 1]$; each column normalized.
- Each of the 90 columns of $V' \in \mathbb{R}_{+}^{10}$, Dirichlet.

Figure: Noise is very sparse: one non-zero entry per column.
Conclusion

1. Nonnegative matrix factorization (NMF)
   - Easily interpretable linear dimensionality reduction technique for nonnegative data, with *many* applications

2. Separable NMF
   - Separability makes NMF problems efficiently solvable
   - Need for fast, practical and robust algorithms

3. A new LP model for near-separable NMF
   - More robust, more flexible, always feasible, no normalization
   - but ... computationally expensive.
   (Possible fix: preselect a ‘good’ subset of columns.)
Conclusion

1. Nonnegative matrix factorization (NMF)
   - Easily interpretable linear dimensionality reduction technique for nonnegative data, with many applications

2. Separable NMF
   - Separability makes NMF problems efficiently solvable
   - Need for fast, practical and robust algorithms

3. A new LP model for near-separable NMF
   - More robust, more flexible, always feasible, no normalization
   - but ... computationally expensive.
   (Possible fix: preselect a ‘good’ subset of columns.)
Conclusion

1. Nonnegative matrix factorization (NMF)
   - Easily interpretable linear dimensionality reduction technique for nonnegative data, with many applications

2. Separable NMF
   - Separability makes NMF problems efficiently solvable
   - Need for fast, practical and robust algorithms

3. A new LP model for near-separable NMF
   - More robust, more flexible, always feasible, no normalization
   - but ... computationally expensive.
   (Possible fix: preselect a 'good' subset of columns.)
Reference.

Code available on https://sites.google.com/site/nicolasgillis/.

Thank you for your attention!