Dictionary-based and sparse decompositions

Nicolas Gillis

Joint work with Robert Luce and Jérémy Cohen

Outline

Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

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

Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

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

NMF is a linear dimensionality reduction technique for nonnegative data :

$$\underbrace{M(:,i)}_{\geq 0} \approx \sum_{k=1}^{r} \underbrace{U(:,k)}_{\geq 0} \underbrace{V(k,i)}_{\geq 0} \qquad \text{for all } i.$$

Nonnegative Matrix Factorization (NMF)

Given a matrix $M \in \mathbb{R}_+^{p \times n}$ and a factorization rank $r \ll \min(p, n)$, find $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{r \times n}$ such that

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

NMF is a linear dimensionality reduction technique for nonnegative data :

$$\underbrace{M(:,i)}_{\geq 0} \approx \sum_{k=1}^{r} \underbrace{U(:,k)}_{\geq 0} \underbrace{V(k,i)}_{\geq 0} \qquad \text{for all } i.$$

Why nonnegativity?

$\rightarrow$ **Interpretability**: Nonnegativity constraints lead to easily interpretable factors (and a sparse and part-based representation).

$\rightarrow$ **Many applications.** image processing, text mining, hyperspectral unmixing, community detection, clustering, etc.

Example 1: Blind hyperspectral unmixing

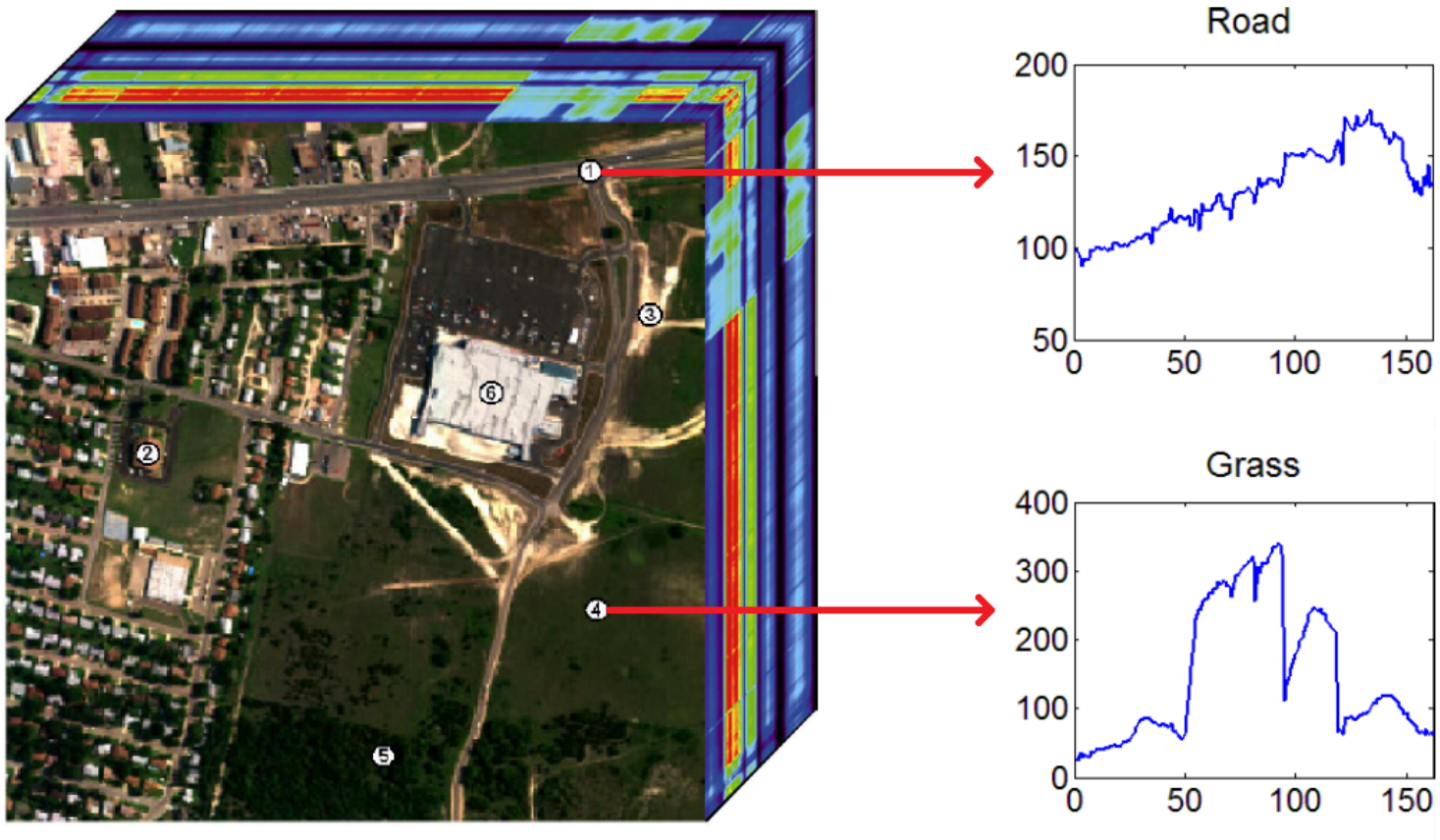


Figure: Urban hyperspectral image, 162 spectral bands and 307-by-307 pixels.

Example 1: Blind hyperspectral unmixing

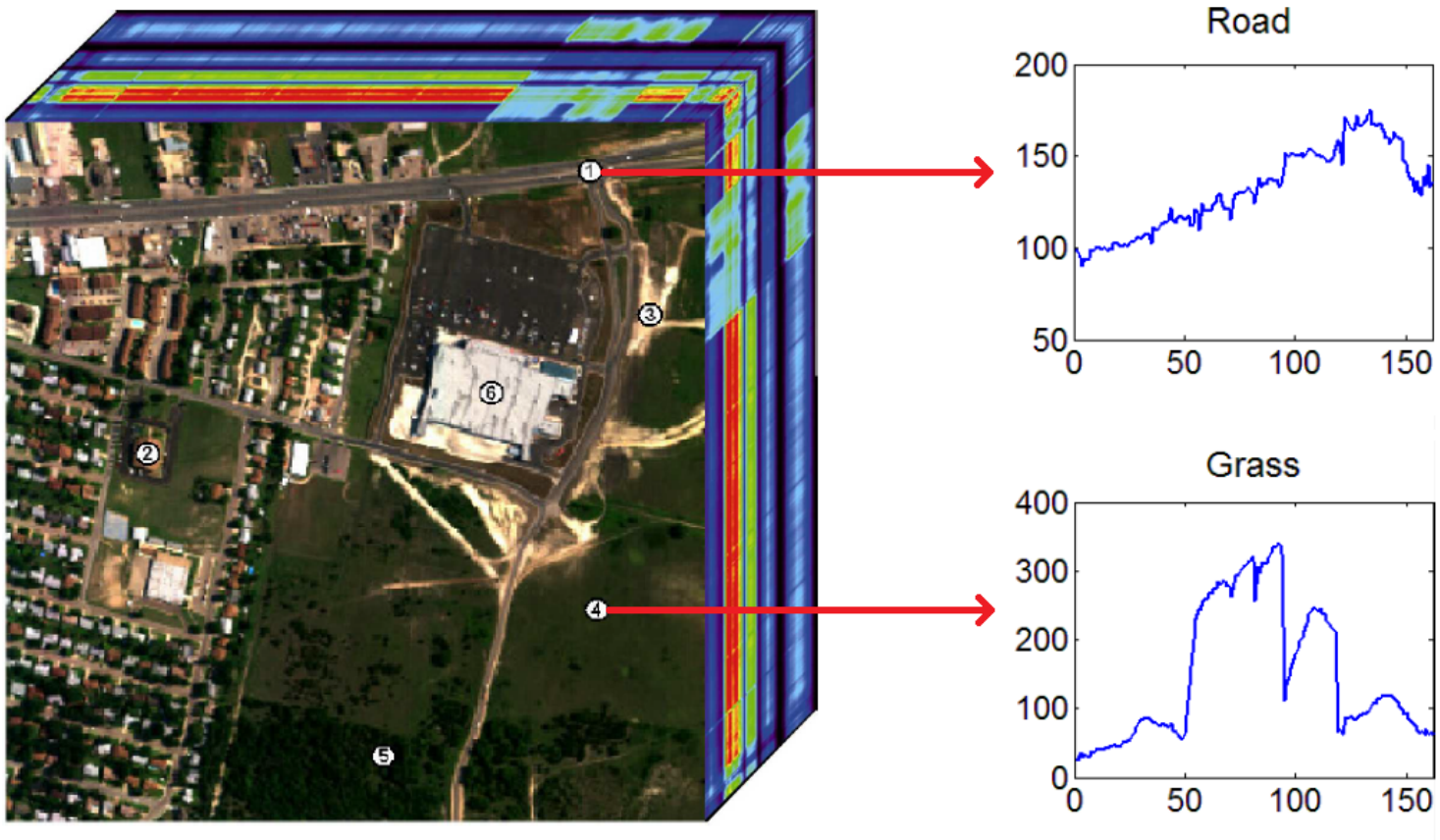
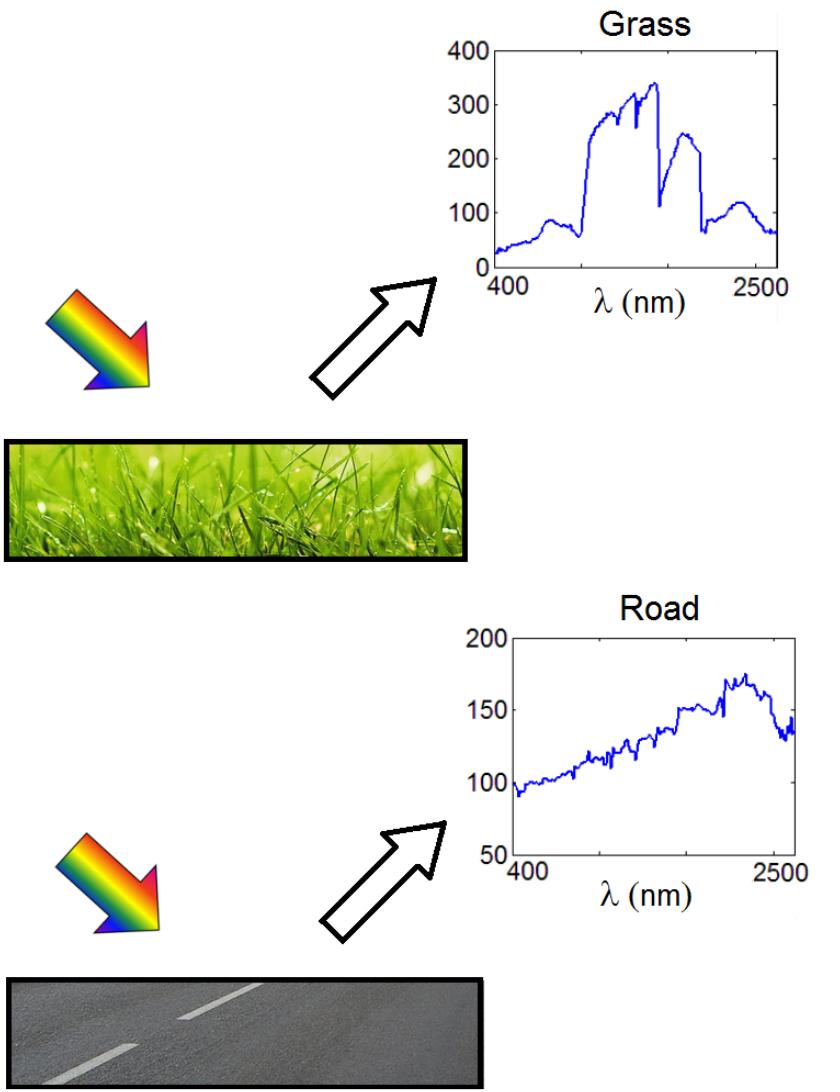
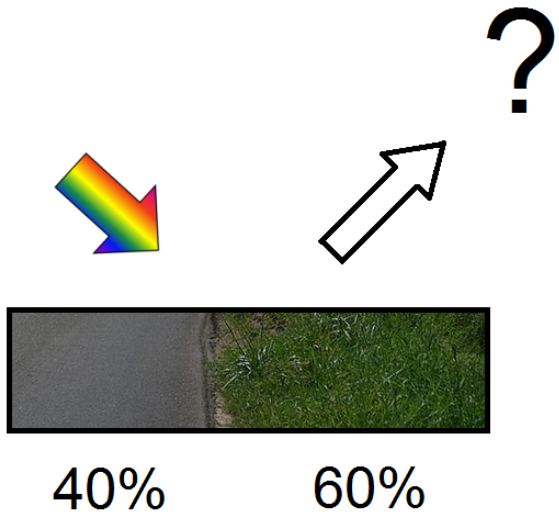


Figure: Urban hyperspectral image, 162 spectral bands and 307-by-307 pixels.

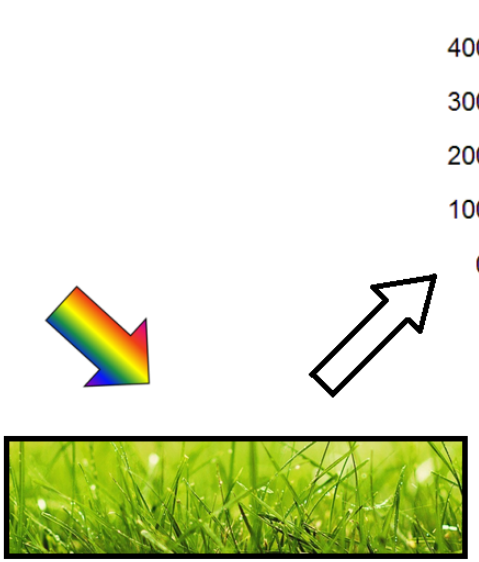
Problem. Identify the materials and classify the pixels.

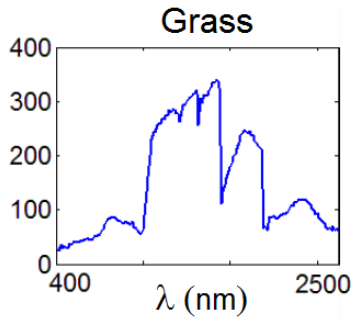
Linear mixing model

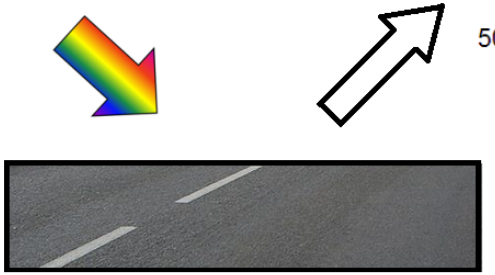


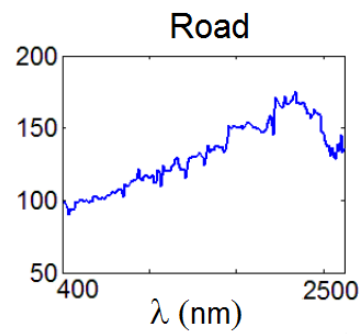


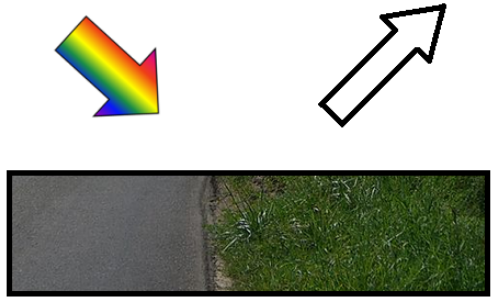
Linear mixing model





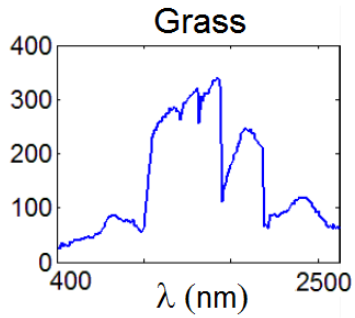




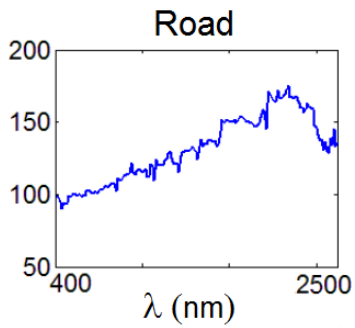


40% 60%

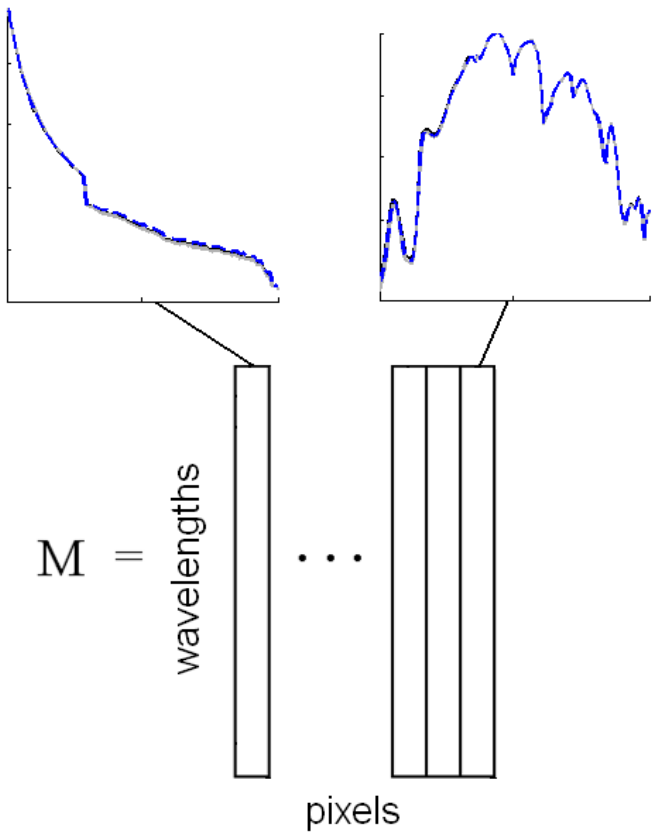
0.6 *



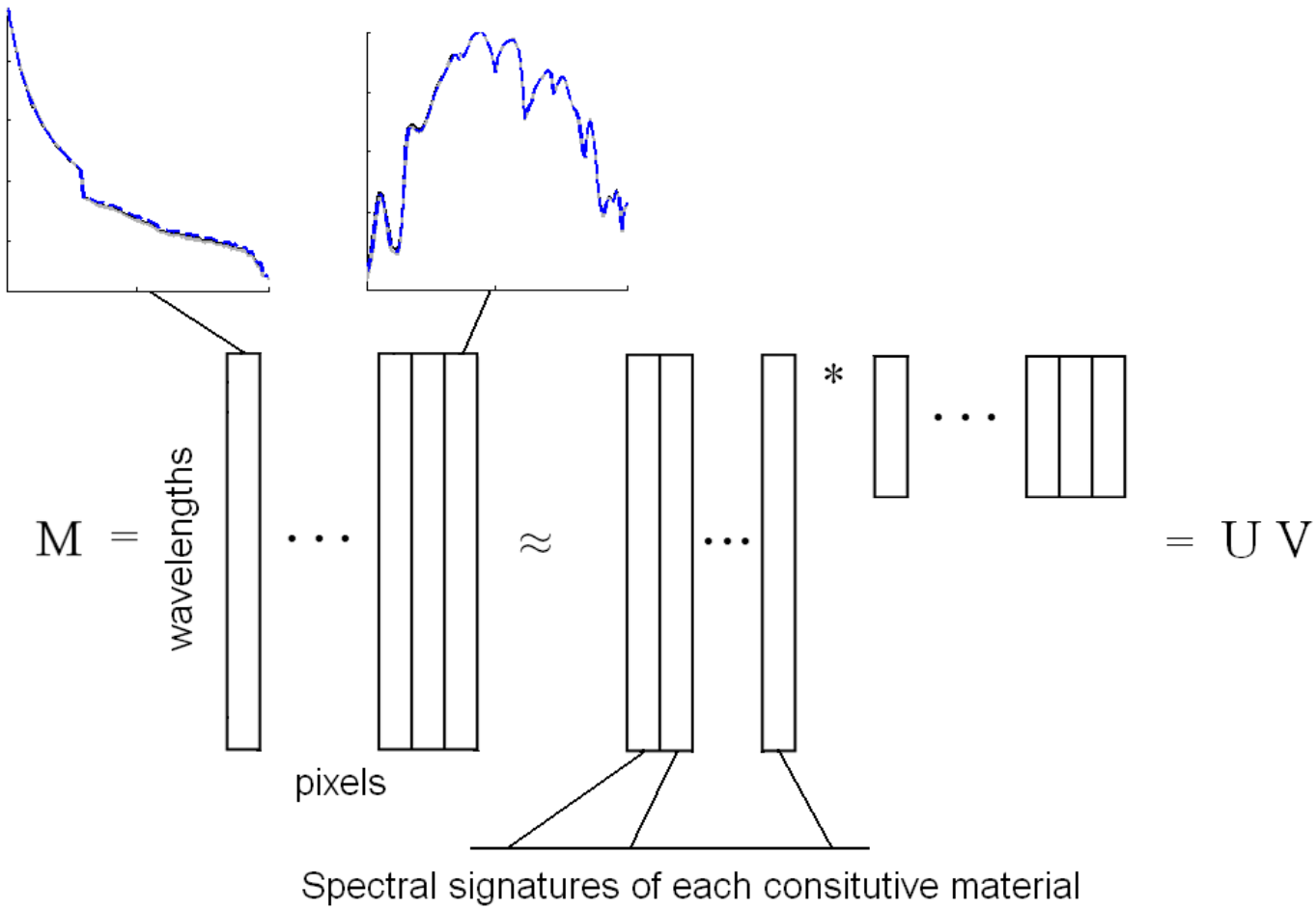
+ 0.4 *



Example 1: Blind hyperspectral unmixing with NMF

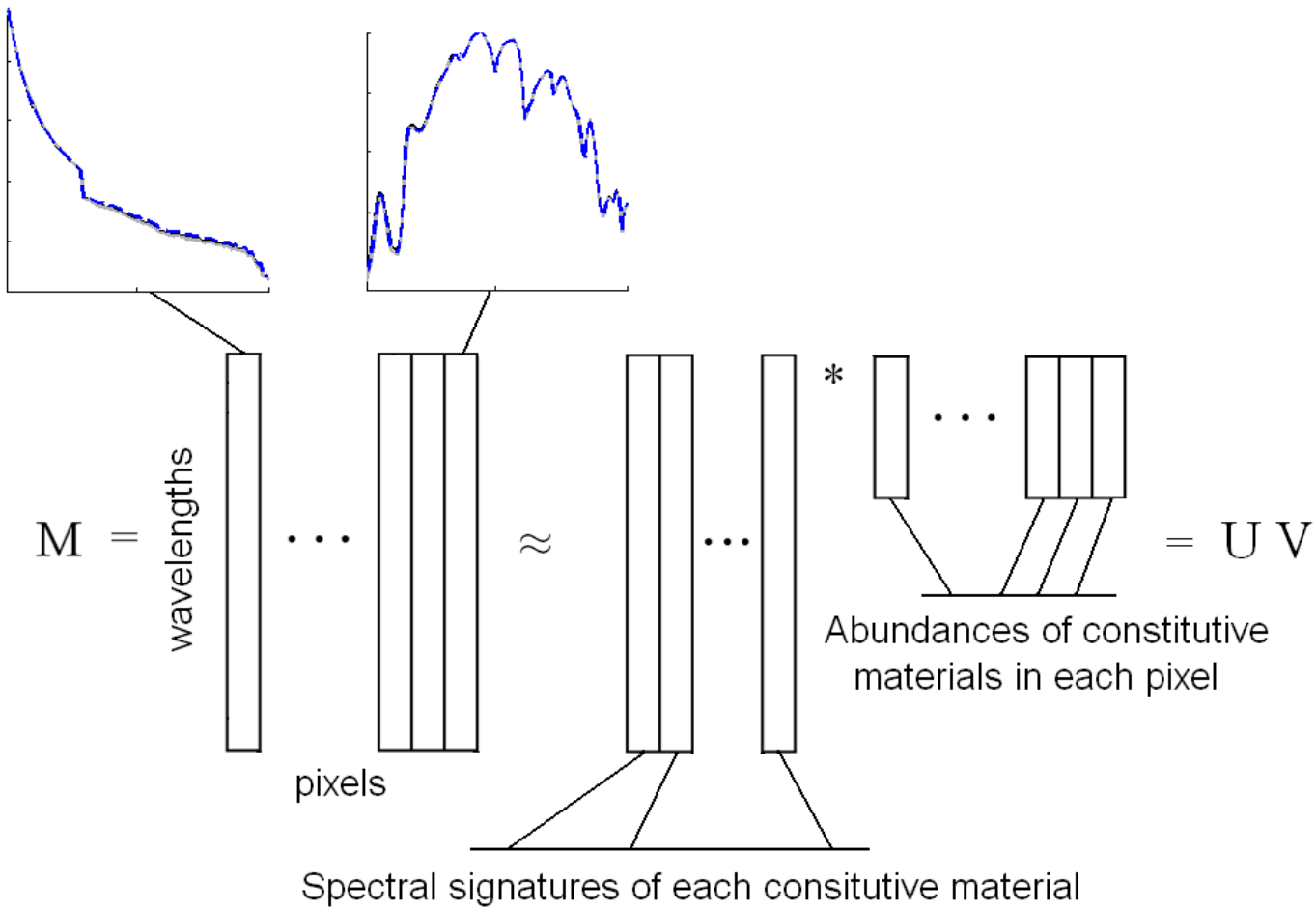


Example 1: Blind hyperspectral unmixing with NMF



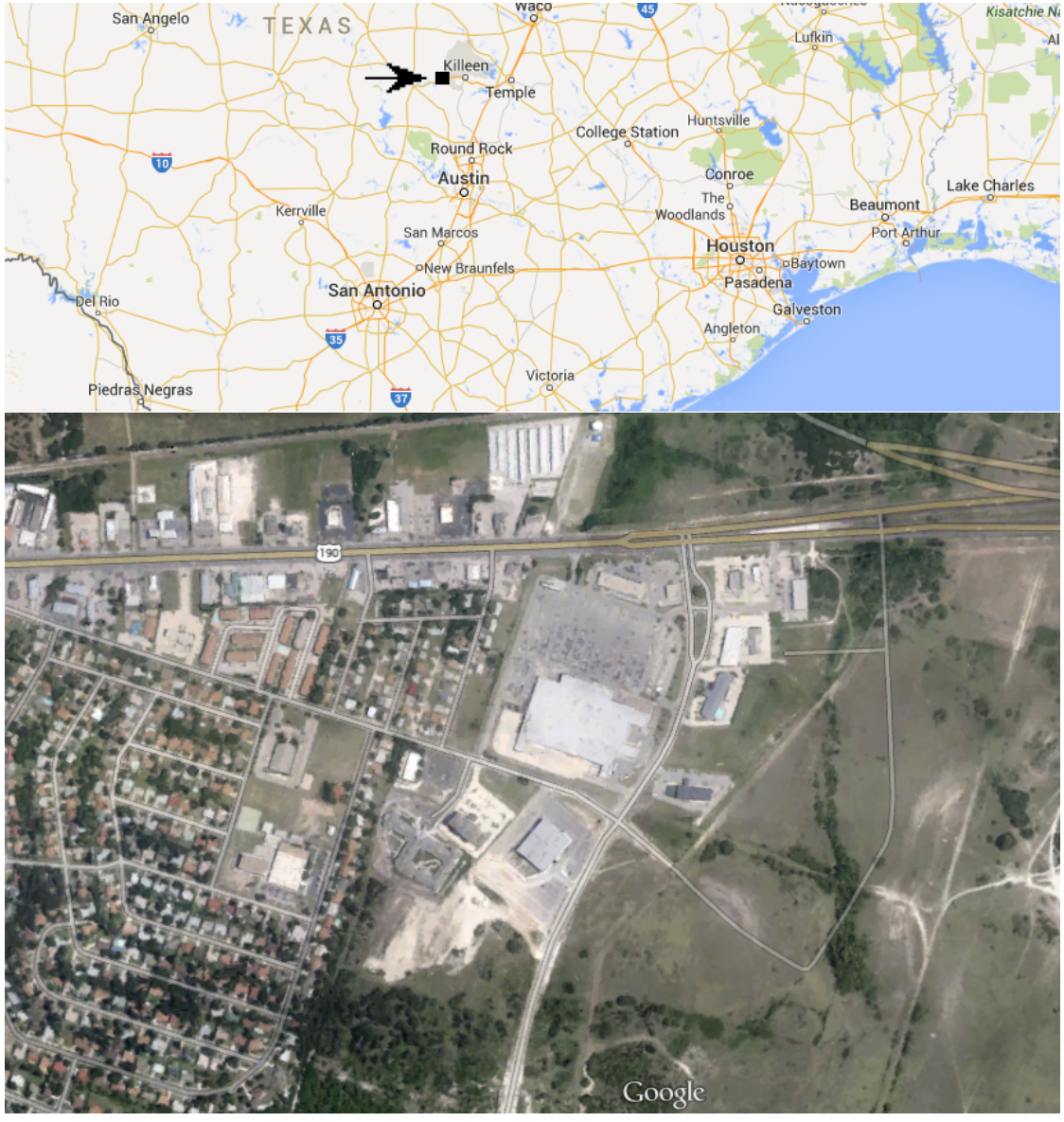
- Basis elements allow to recover the different endmembers: $U \geq 0$;

Example 1: Blind hyperspectral unmixing with NMF



- Basis elements allow to recover the different endmembers: $U \geq 0$;
- Abundances of the endmembers in each pixel: $V \geq 0$.

Urban hyperspectral image



Urban hyperspectral image

$$\underbrace{\mathbf{M}(:,j)}_{\substack{\text{spectral signature} \\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathbf{U}(:,k)}_{} \underbrace{\mathbf{V}(k,j)}_{} .$$

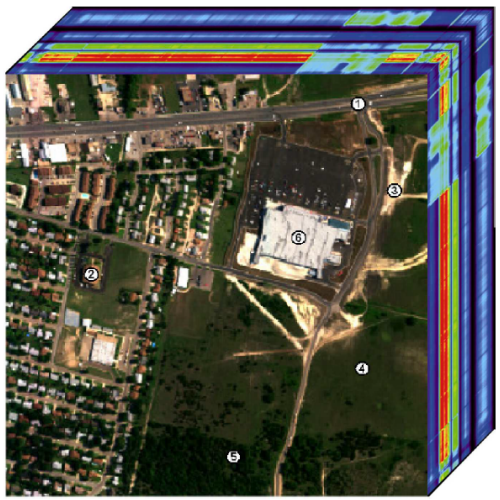
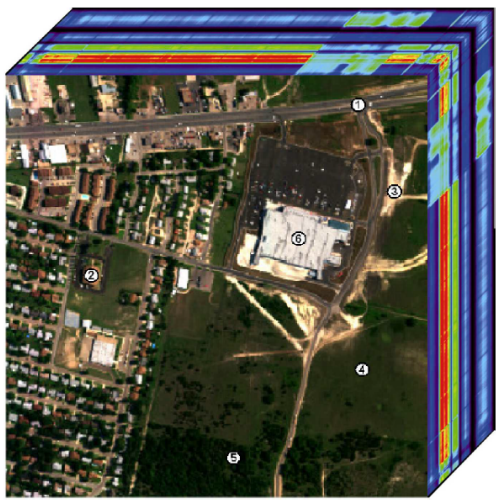


Figure: Decomposition of the Urban dataset.

Urban hyperspectral image

$$\underbrace{\mathbf{M}(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathbf{U}(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{\mathbf{V}(k,j)}_{\text{}} .$$



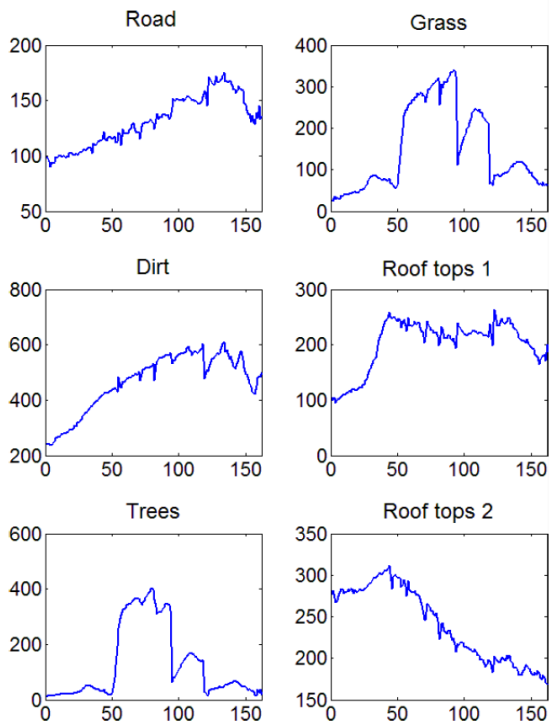
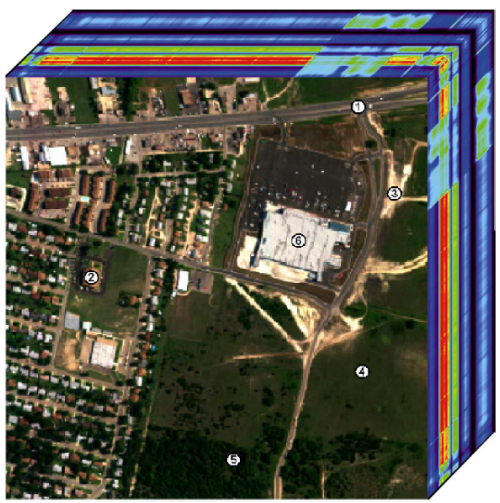
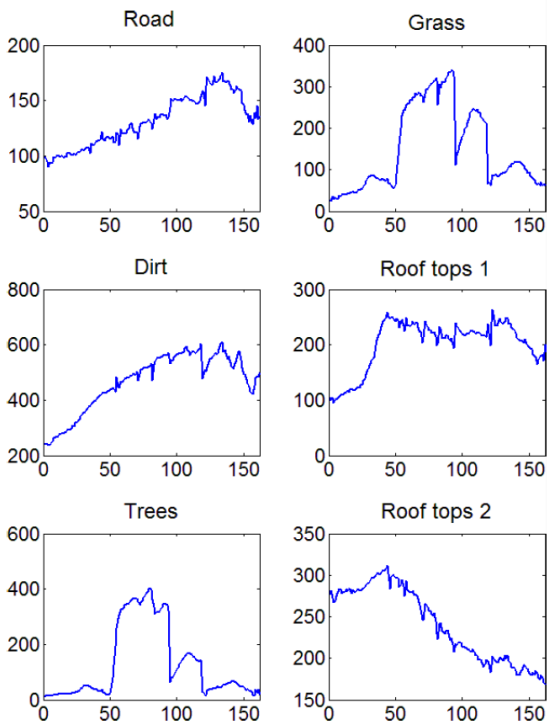


Figure: Decomposition of the Urban dataset.

Urban hyperspectral image

$$\underbrace{\mathbf{M}(:,j)}_{\substack{\text{spectral signature}\\ \text{of } j\text{th pixel}}} \approx \sum_{k=1} \underbrace{\mathbf{U}(:,k)}_{\substack{\text{spectral signature}\\ \text{of } k\text{th endmember}}} \underbrace{\mathbf{V}(k,j)}_{\substack{\text{abundance of } k\text{th endmember}\\ \text{in } j\text{th pixel}}} .$$





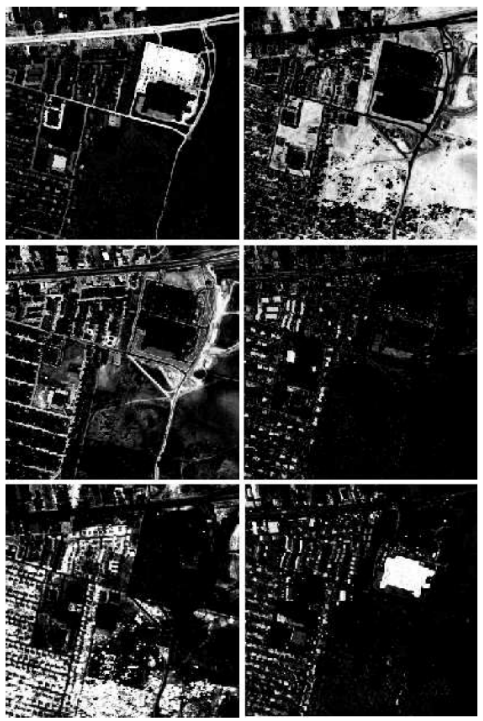
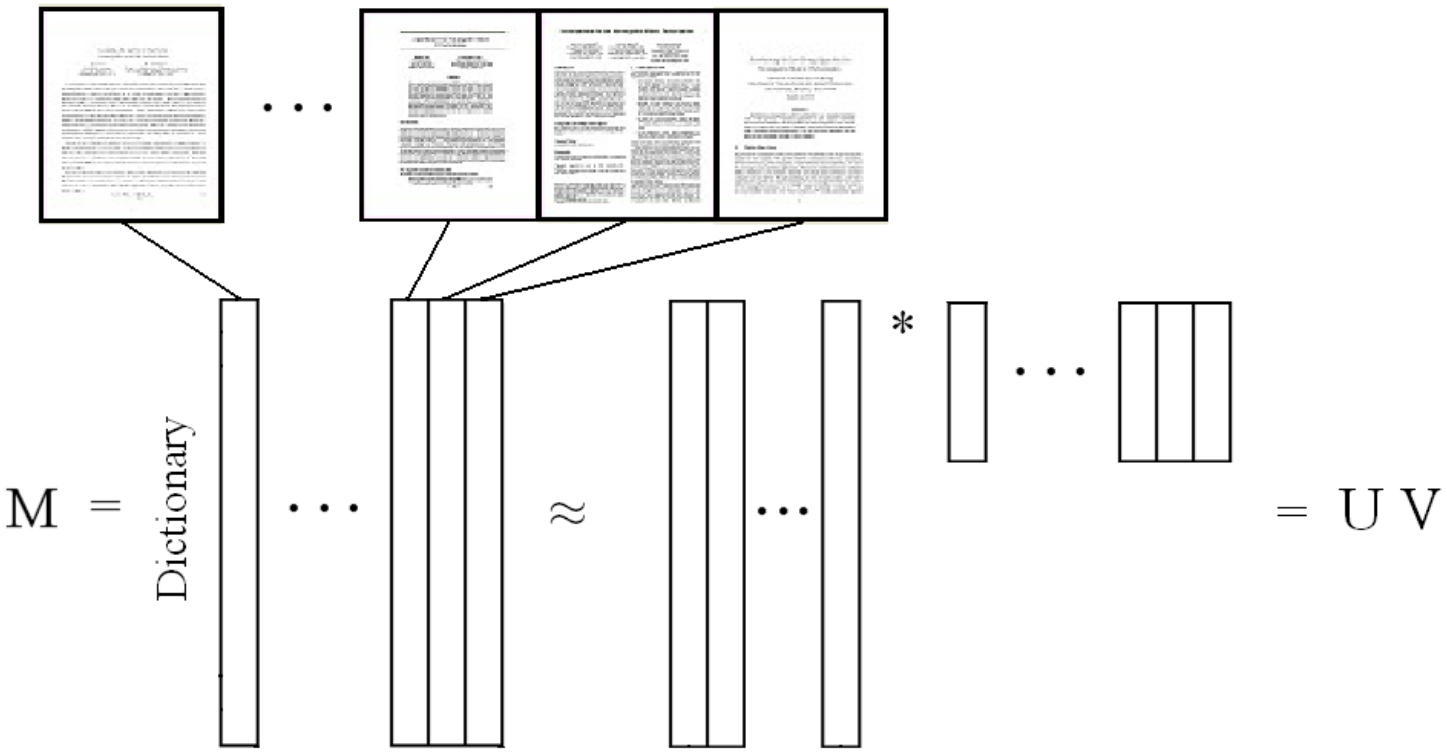
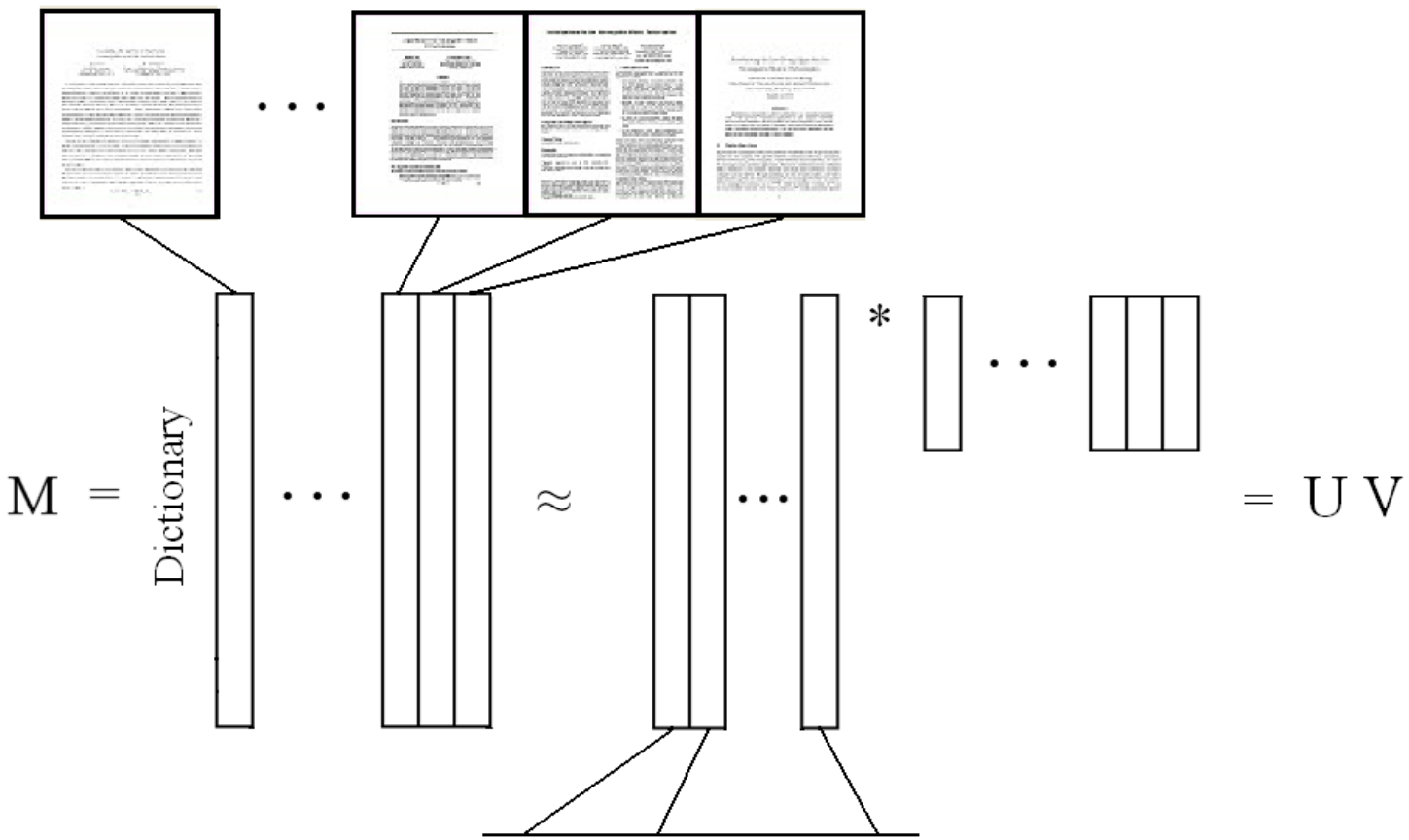


Figure: Decomposition of the Urban dataset.

Example 2: topic recovery and document classification

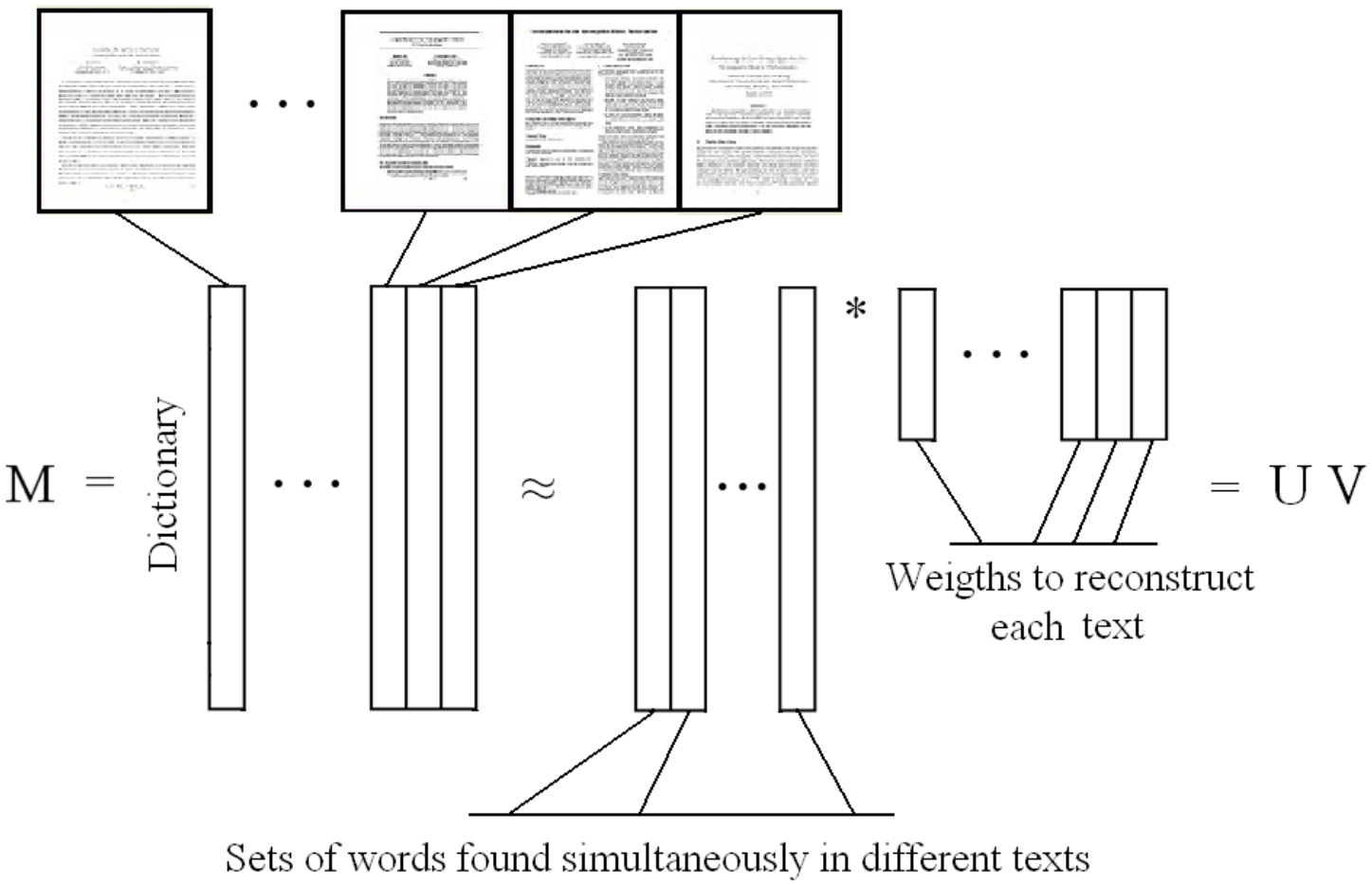


Example 2: topic recovery and document classification



- Basis elements allow to recover the different topics;

Example 2: topic recovery and document classification



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

NMF Algorithms

Given a matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank $r \in \mathbb{N}$:

$$\min_{U \in \mathbb{R}_+^{m \times r}, V \in \mathbb{R}_+^{r \times n}} ||M - UV||_F^2 = \sum_{i,j} (M - UV)_{ij}^2. \qquad \text{(NMF)}$$

NMF is NP-hard (Vavasis, 2009).

NMF Algorithms

Given a matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank $r \in \mathbb{N}$:

$$\min_{U \in \mathbb{R}_+^{m \times r}, V \in \mathbb{R}_+^{r \times n}} ||M - UV||_F^2 = \sum_{i,j} (M - UV)_{ij}^2. \qquad \text{(NMF)}$$

NMF is NP-hard (Vavasis, 2009).

Standard framework:

0. Initialize (U, V). Then, alternatively update U and V:
    1. Update $V \approx \text{argmin}_{X \geq 0} ||M - UX||_F^2$. (NNLS)
    2. Update $U \approx \text{argmin}_{Y \geq 0} ||M - YV||_F^2$. (NNLS)

NMF Algorithms

Given a matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank $r \in \mathbb{N}$:

$$\min_{U \in \mathbb{R}_+^{m \times r}, V \in \mathbb{R}_+^{r \times n}} ||M - UV||_F^2 = \sum_{i,j} (M - UV)_{ij}^2. \qquad \text{(NMF)}$$

NMF is NP-hard (Vavasis, 2009).

Standard framework:

0. Initialize (U, V). Then, alternatively update U and V:
    1. Update $V \approx \text{argmin}_{X \geq 0} ||M - UX||_F^2$. (NNLS)
    2. Update $U \approx \text{argmin}_{Y \geq 0} ||M - YV||_F^2$. (NNLS)

Most NMF algorithms come with **no guarantees** (except convergence to stationary points).

NMF Algorithms

Given a matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank $r \in \mathbb{N}$:

$$\min_{U \in \mathbb{R}_+^{m \times r}, V \in \mathbb{R}_+^{r \times n}} ||M - UV||_F^2 = \sum_{i,j} (M - UV)_{ij}^2. \qquad \text{(NMF)}$$

NMF is NP-hard (Vavasis, 2009).

Standard framework:

0. Initialize (U, V). Then, alternatively update U and V:
    1. Update $V \approx \text{argmin}_{X \geq 0} ||M - UX||_F^2$. (NNLS)
    2. Update $U \approx \text{argmin}_{Y \geq 0} ||M - YV||_F^2$. (NNLS)

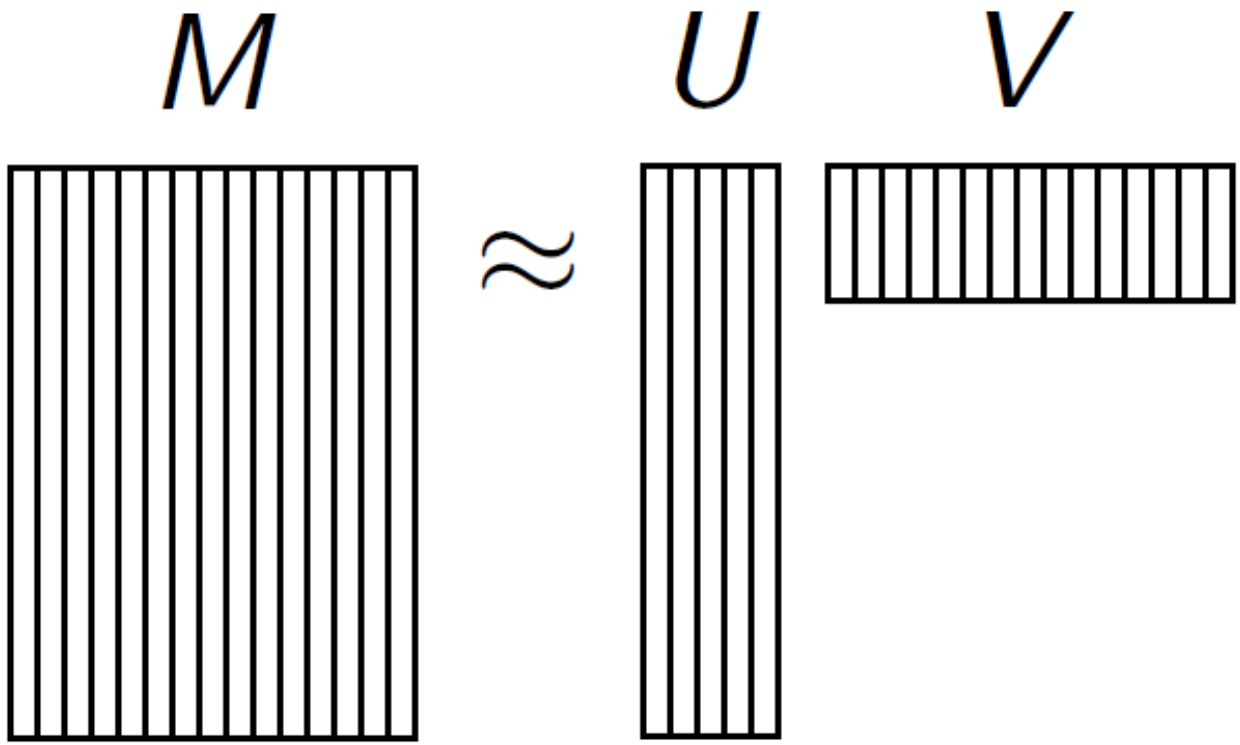
Most NMF algorithms come with **no guarantees** (except convergence to stationary points).

Solution is in general **highly non-unique**: indentifiability issues.

NMF under the separability assumption

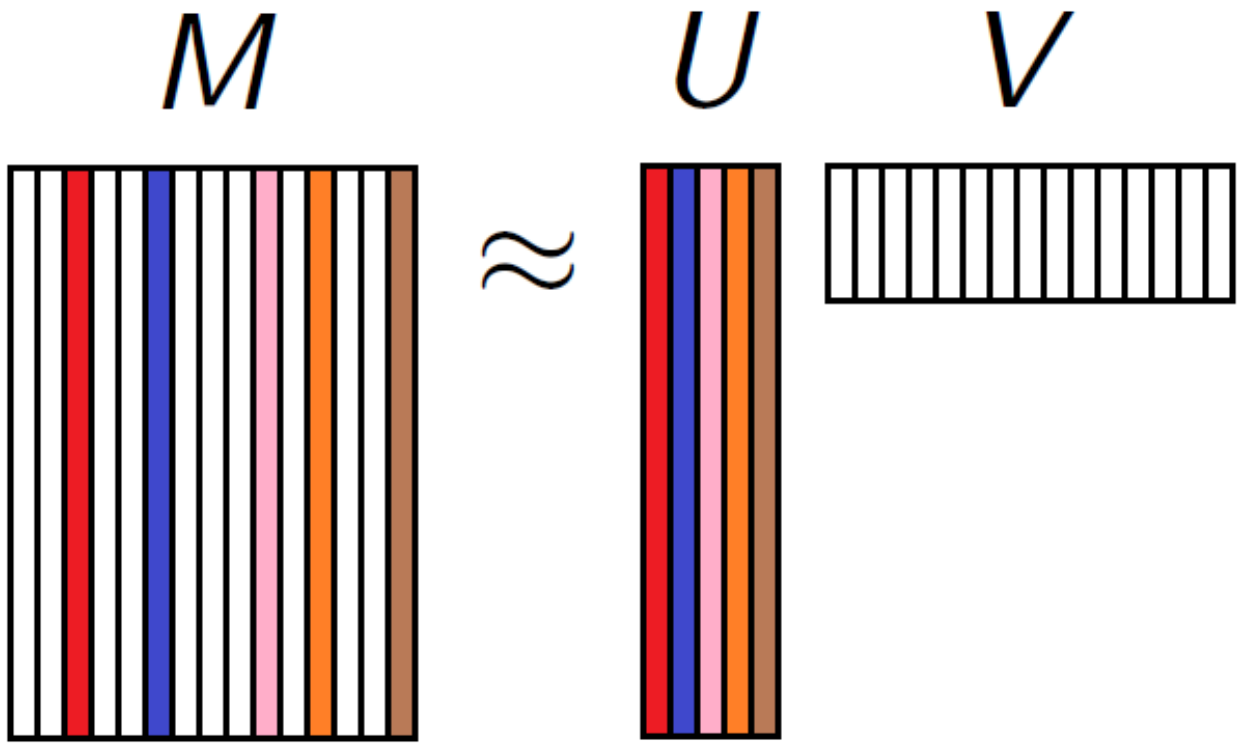
Separability Assumption

Separability of M: there exists an index set $\mathcal{K}$ and $V \geq 0$ with $M = \underbrace{M(:, \mathcal{K})}_{U} V$, with $|\mathcal{K}| = r$.



Separability Assumption

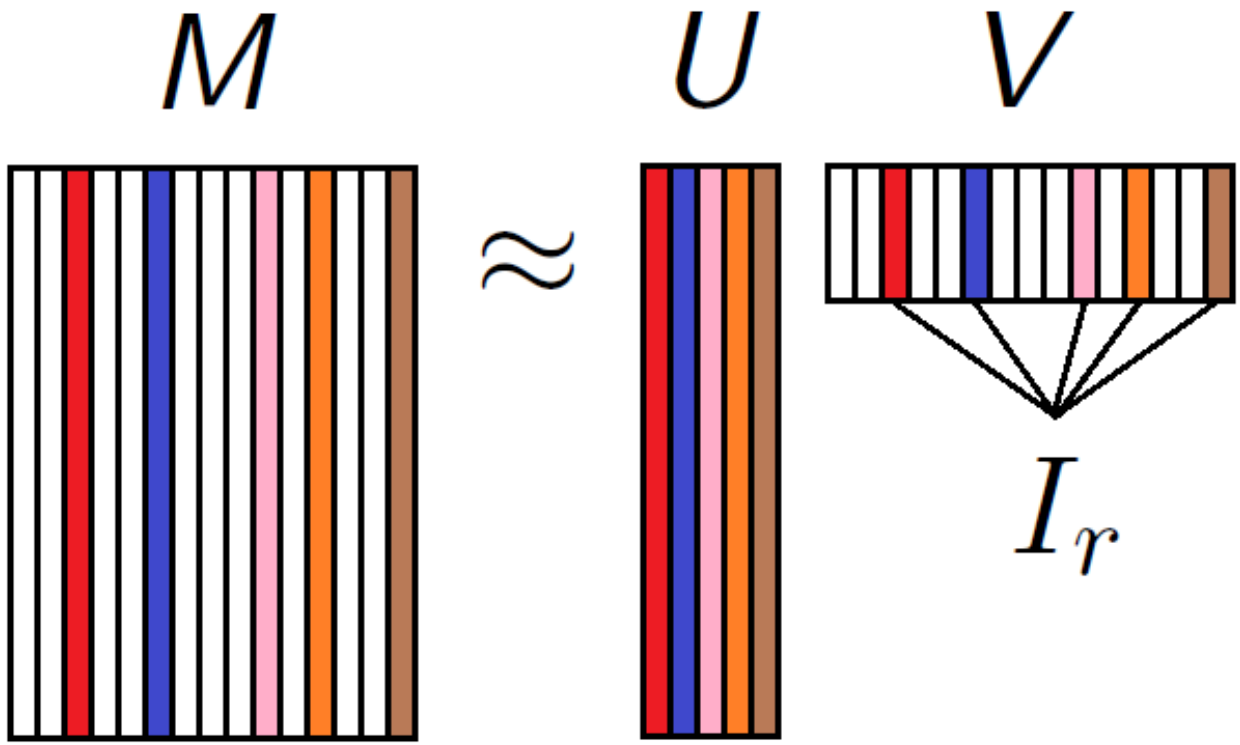
Separability of M: there exists an index set $\mathcal{K}$ and $V \geq 0$ with $M = \underbrace{M(:, \mathcal{K})}_{U} V$, with $|\mathcal{K}| = r$.



[AGKM12] Arora, Ge, Kannan, Moitra, *Computing a Nonnegative Matrix Factorization – Provably*, STOC 2012.

Separability Assumption

Separability of M: there exists an index set $\mathcal{K}$ and $V \geq 0$ with $M = \underbrace{M(:, \mathcal{K})}_{U} V$, with $|\mathcal{K}| = r$.



[AGKM12] Arora, Ge, Kannan, Moitra, *Computing a Nonnegative Matrix Factorization – Provably*, STOC 2012.

Applications

- In hyperspectral imaging, this is the pure-pixel assumption: for each material, there is a 'pure' pixel containing only that material.

  [M+14] Ma et al., *A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing*, IEEE Signal Processing Magazine 31(1):67-81, 2014.

Applications

- In hyperspectral imaging, this is the pure-pixel assumption: for each material, there is a 'pure' pixel containing only that material.

  [M+14] Ma et al., *A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing*, IEEE Signal Processing Magazine 31(1):67-81, 2014.

- In document classification: for each topic, there is a 'pure' word used only by that topic (an 'anchor' word).

  [A+13] Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees*, ICML 2013.

Applications

- In hyperspectral imaging, this is the pure-pixel assumption: for each material, there is a 'pure' pixel containing only that material.

  [M+14] Ma et al., *A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing*, IEEE Signal Processing Magazine 31(1):67-81, 2014.

- In document classification: for each topic, there is a 'pure' word used only by that topic (an 'anchor' word).

  [A+13] Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees*, ICML 2013.

- Time-resolved Raman spectra analysis: each substance has a peak in its spectrum while the other spectra are (close) to zero.

  [L+16] Luce et al., *Using Separable Nonnegative Matrix Factorization for the Analysis of Time-Resolved Raman Spectra*, Appl Spectrosc. 2016.

Applications

- In hyperspectral imaging, this is the pure-pixel assumption: for each material, there is a 'pure' pixel containing only that material.

  [M+14] Ma et al., *A Signal Processing Perspective on Hyperspectral Unmixing: Insights from Remote Sensing*, IEEE Signal Processing Magazine 31(1):67-81, 2014.

- In document classification: for each topic, there is a 'pure' word used only by that topic (an 'anchor' word).

  [A+13] Arora et al., *A Practical Algorithm for Topic Modeling with Provable Guarantees*, ICML 2013.

- Time-resolved Raman spectra analysis: each substance has a peak in its spectrum while the other spectra are (close) to zero.

  [L+16] Luce et al., *Using Separable Nonnegative Matrix Factorization for the Analysis of Time-Resolved Raman Spectra*, Appl Spectrosc. 2016.

- Others: video summarization, foreground-background separation.

  [ESV12] Elhamifar, Sapiro, Vidal, *See all by looking at a few: Sparse modeling for finding representative objects*, CVPR 2012.
  [KSK13] Kumar, Sindhwani, *Near-separable Non-negative Matrix Factorization with ℓ_1- and Bregman Loss Functions*, SIAM data mining 2015.

Combinatorial formulation for separable NMF

We want to find the index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$M \quad = \quad M(:, \mathcal{K})\, V.$$

Combinatorial formulation for separable NMF

We want to find the index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$M \quad = \quad M(:, \mathcal{K})\, V.$$

This is equivalent to finding $X \in \mathbb{R}^{n \times n}$ with r non-zero rows such that

$$M \quad = \quad M\, X.$$

Combinatorial formulation for separable NMF

We want to find the index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$M \quad = \quad M(:, \mathcal{K})\, V.$$

This is equivalent to finding $X \in \mathbb{R}^{n \times n}$ with r non-zero rows such that

$$M \quad = \quad M\, X.$$

A combinatorial formulation:

$$\min_X ||X||_{\text{row},0} \quad \text{such that } M = MX \;\; \text{or} \;\; ||M - MX|| \leq \epsilon.$$

Combinatorial formulation for separable NMF

We want to find the index set $\mathcal{K}$ with $|\mathcal{K}| = r$ such that

$$M \quad = \quad M(:, \mathcal{K})\, V.$$

This is equivalent to finding $X \in \mathbb{R}^{n \times n}$ with r non-zero rows such that

$$M \quad = \quad M\, X.$$

A combinatorial formulation:

$$\min_X ||X||_{\text{row},0} \quad \text{such that } M = MX \;\; \text{or} \;\; ||M - MX|| \leq \epsilon.$$

How to make X row sparse?

A Linear Optimization Model

$$\begin{aligned} \min_{X \in \mathbb{R}_+^{n \times n}} \quad & \text{trace}(X) = ||\,\text{diag}(X)||_1 \\ \text{such that} \quad & ||M - MX|| \leq \epsilon, \\ & X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j. \end{aligned}$$

A Linear Optimization Model

$$\begin{aligned} \min_{X \in \mathbb{R}_+^{n \times n}} \quad & \text{trace}(X) = ||\,\text{diag}(X)||_1 \\ \text{such that} \quad & ||M - MX|| \leq \epsilon, \\ & X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j. \end{aligned}$$

Robustness: noise $\leq \mathcal{O}\left(\kappa(U)^{-1}\right) \Rightarrow$ error $\leq \mathcal{O}\left(\frac{r\epsilon}{\kappa}\right)$ [GL14].

[GL14] G., Luce, *Robust Near-Separable NMF Using Linear Optimization*, JMLR 2014.

A Linear Optimization Model

$$\begin{aligned} \min_{X \in \mathbb{R}_+^{n \times n}} \quad & \text{trace}(X) = ||\,\text{diag}(X)||_1 \\ \text{such that} \quad & ||M - MX|| \leq \epsilon, \\ & X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j. \end{aligned}$$

Robustness: noise $\leq \mathcal{O}\left(\kappa(U)^{-1}\right) \Rightarrow$ error $\leq \mathcal{O}\left(\frac{r\epsilon}{\kappa}\right)$ [GL14].

This model is an improvement over [B+12]: more robust and detects the factorization rank r automatically.

[GL14] G., Luce, *Robust Near-Separable NMF Using Linear Optimization*, JMLR 2014.

[B+12] Bittorf, Recht, Ré, Tropp, *Factoring nonnegative matrices with LPs*, NIPS 2012.

A Linear Optimization Model

$$\begin{aligned} \min_{X \in \mathbb{R}_+^{n \times n}} \quad & \text{trace}(X) = ||\,\text{diag}(X)||_1 \\ \text{such that} \quad & ||M - MX|| \leq \epsilon, \\ & X_{ij} \leq X_{ii} \leq 1 \text{ for all } i, j. \end{aligned}$$

Robustness: noise $\leq \mathcal{O}\left(\kappa(U)^{-1}\right) \Rightarrow$ error $\leq \mathcal{O}\left(\frac{r\epsilon}{\kappa}\right)$ [GL14].

This model is an improvement over [B+12]: more robust and detects the factorization rank r automatically.

It is equivalent [GL16] to using $||X||_{1,\infty} = \sum_{i=1}^{d} ||X(i,:)||_\infty$ as a convex surrogate for $||X||_{\text{row},0}$ [E+12].

[GL14] G., Luce, *Robust Near-Separable NMF Using Linear Optimization*, JMLR 2014.

[B+12] Bittorf, Recht, Ré, Tropp, *Factoring nonnegative matrices with LPs*, NIPS 2012.

[E+12] Esser et al., A convex model for NMF and dimensionality reduction on physical space, IEEE Trans. Image Processing, 2012.

[GL16] G. and Luce, A Fast Gradient Method for Nonnegative Sparse Regression with Self Dictionary, IEEE Trans. Image Processing, 2018.

Practical Model and Algorithm

$$\min_{X \in \Omega} \; ||M - MX||_F^2 + \mu \, \mathrm{tr}(X),$$

$$\Omega = \{X \in \mathbb{R}^{n,n} \mid X_{ii} \leq 1, w_i X_{ij} \leq w_j X_{ii} \forall i, j\}.$$

Practical Model and Algorithm

$$\min_{X \in \Omega} \; ||M - MX||_F^2 + \mu \, \mathrm{tr}(X),$$

$$\Omega = \{X \in \mathbb{R}^{n,n} \mid X_{ii} \leq 1, w_i X_{ij} \leq w_j X_{ii} \forall i, j\}.$$

We used a fast gradient method (optimal 1st order):

1. Choose an initial point $X^{(0)}$, $Y = X^{(0)}$, $\alpha_1 \in (0, 1)$.
2. $k = 1, 2, \ldots$
   1. $X^{(k)} = \mathcal{P}_\Omega \left(Y - \frac{1}{L} \nabla f(Y) \right)$.
   2. $Y = X^{(k)} + \beta_k \left(X^{(k)} - X^{(k-1)} \right)$,
      where $\beta_k = \frac{\alpha_k (1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ with $\alpha_{k+1} \geq 0$ t.q. $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2$.

[GL16] G. and Luce, A **F**ast **G**radient Method for **N**onnegative **S**parse **R**egression with Self Dictionary, IEEE Trans. Image Processing, 2018.

Practical Model and Algorithm

$$\min_{X \in \Omega} \; ||M - MX||_F^2 + \mu \, \mathrm{tr}(X),$$

$$\Omega = \{X \in \mathbb{R}^{n,n} \mid X_{ii} \leq 1, w_i X_{ij} \leq w_j X_{ii} \forall i, j\}.$$

We used a fast gradient method (optimal 1st order):

1. Choose an initial point $X^{(0)}$, $Y = X^{(0)}$, $\alpha_1 \in (0, 1)$.
2. $k = 1, 2, \ldots$
   1. $X^{(k)} = \mathcal{P}_\Omega \left(Y - \frac{1}{L} \nabla f(Y) \right)$.
   2. $Y = X^{(k)} + \beta_k \left(X^{(k)} - X^{(k-1)} \right)$,
      where $\beta_k = \frac{\alpha_k (1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ with $\alpha_{k+1} \geq 0$ t.q. $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2$.

Projection onto Ω can be done effectively in $\mathcal{O}(n^2 \log(n))$ operations.

[GL16] G. and Luce, A **F**ast **G**radient Method for **N**onnegative **S**parse **R**egression with Self Dictionary, IEEE Trans. Image Processing, 2018.

Practical Model and Algorithm

$$\min_{X \in \Omega} \; ||M - MX||_F^2 + \mu \, \mathrm{tr}(X),$$

$$\Omega = \{X \in \mathbb{R}^{n,n} \mid X_{ii} \leq 1, w_i X_{ij} \leq w_j X_{ii} \forall i, j\}.$$

We used a fast gradient method (optimal 1st order):

1. Choose an initial point $X^{(0)}$, $Y = X^{(0)}$, $\alpha_1 \in (0, 1)$.
2. $k = 1, 2, \ldots$
   1. $X^{(k)} = \mathcal{P}_\Omega \left(Y - \frac{1}{L} \nabla f(Y) \right)$.
   2. $Y = X^{(k)} + \beta_k \left(X^{(k)} - X^{(k-1)} \right)$,
      where $\beta_k = \frac{\alpha_k (1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ with $\alpha_{k+1} \geq 0$ t.q. $\alpha_{k+1}^2 = (1 - \alpha_{k+1}) \alpha_k^2$.

Projection onto Ω can be done effectively in $\mathcal{O}(n^2 \log(n))$ operations.

The total computational cost is $\mathcal{O}(pn^2)$ operations.

[GL16] G. and Luce, A **F**ast **G**radient Method for **N**onnegative **S**parse **R**egression with Self Dictionary, IEEE Trans. Image Processing, 2018.

Hyperspectral unmixing

	$r = 6$		$r = 8$	
	Time (s.)	Rel. err. (%)	Time (s.)	Rel. err. (%)
VCA	1.02	18.05	1.05	22.68
VCA-500	0.03	7.19	0.09	7.25
SPA	0.26	9.58	0.32	9.45
SPA-500	$<$0.01	10.05	$<$0.01	8.86
SNPA	13.60	9.63	23.02	5.64
SNPA-500	0.15	10.05	0.25	8.86
XRAY	28.17	7.50	95.34	6.82
XRAY-500	0.15	8.07	0.28	7.36
H2NMF	12.20	5.81	14.92	5.47
H2NMF-500	0.27	5.87	0.37	5.68
FGNSR-500	40.11	**5.07**	39.49	**4.08**

Table: Numerical results for the Urban HSI (the best result is highlighted in bold).

Figure: Abundance maps extracted by FGNSR-500.

Why only self-dictionary?

We can generalize the previous model to any dictionary:

$$M \quad \approx \quad \underbrace{D(:, \mathcal{K})}_{=U} \; V,$$

where $\mathcal{K}$ selects atoms in the dictionary $D \in \mathbb{R}^{p \times d}$ (e.g., a hyperspectral library).

Why only self-dictionary?

We can generalize the previous model to any dictionary:

$$M \quad \approx \quad \underbrace{D(:,\mathcal{K})}_{=U} \; V,$$

where $\mathcal{K}$ selects atoms in the dictionary $D \in \mathbb{R}^{p \times d}$ (e.g., a hyperspectral library). As before, this is equivalent to

$$\min_X ||X||_{\text{row},0} \quad \text{such that} \quad M = DX,$$

for which there exists convex relaxations, e.g., replace $||X||_{\text{row},0}$ with $||X||_{1,q} = \sum_{i=1}^{d} ||X(i,:)||_q$.
Again, this is computationally expensive for d large (d^2 variables).

[ESV12] Elhamifar, Sapiro, Vidal, See all by looking at a few: Sparse modeling for finding representative objects, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012.
[E+12] Esser et al., A convex model for nonnegative matrix factorization and dimensionality reduction on physical space, IEEE Trans. Image Process. 2012.

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

Introduce U and solve

$$\min_{\mathcal{K},U,V} ||M - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

Introduce U and solve

$$\min_{\mathcal{K},U,V} ||M - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

via alternating optimization:

✓ Optimal update for each variable U, V and $\mathcal{K}$.

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

Introduce U and solve

$$\min_{\mathcal{K},U,V} ||M - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

via alternating optimization:

- ✓ Optimal update for each variable U, V and $\mathcal{K}$.
- ✓ Computationally cheap, $O(pnr)$ if using a first-order method to update U and V.

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

Introduce U and solve

$$\min_{\mathcal{K},U,V} ||M - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

via alternating optimization:

- ✓ Optimal update for each variable U, V and $\mathcal{K}$.
- ✓ Computationally cheap, $O(pnr)$ if using a first-order method to update U and V.
- ✓ In practice, it is always able to improve the provided initial solution.

A combinatorial model and algorithm

We want to solve

$$\min_{\mathcal{K},V} ||M - D(:,\mathcal{K})V||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

Introduce U and solve

$$\min_{\mathcal{K},U,V} ||M - UV||_F^2 + \delta||U - D(:,\mathcal{K})||_F^2 \quad \text{such that} \quad |\mathcal{K}| = r \text{ and } V \geq 0.$$

via alternating optimization:

- ✓ Optimal update for each variable U, V and $\mathcal{K}$.
- ✓ Computationally cheap, $O(pnr)$ if using a first-order method to update U and V.
- ✓ In practice, it is always able to improve the provided initial solution.
- ✗ No optimality/recovery guarantee.

Numerical results for the Urban data set

	$r = 6$		$r = 8$	
	Time (s.)	Rel. err.	Time (s.)	Rel. err.
RAND-be	0.00	13.77	0.00	5.54
d-RAND-be	22.01 (11)	4.36	36.18 (19)	4.16
SPA	0.30	9.58	0.30	9.45
d-SPA	24.37 (13)	4.67	28.61 (14)	4.62
SNPA	24.34	9.63	36.72	5.64
d-SNPA	23.04 (13)	4.94	27.94 (13)	**3.97**
H2NMF	19.02	5.81	22.35	5.47
d-H2NMF	26.66 (15)	**4.05**	28.92 (14)	4.24
FGNSR-100	2.73	5.58	2.55	4.62
d-FGNSR-100	26.72 (14)	4.36	20.81 (8)	4.04
FGNSR-500	40.11	<u>5.07</u>	39.49	<u>4.08</u>
d-FGNSR-500	25.07 (13)	4.40	26.83 (12)	4.13

Figure: Abundance maps extracted by d-H2NMF.

Dictionary-based decompositions for tensors

CPD and the problem of interpretability

$$\mathcal{T} = (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}) \qquad \mathcal{I}_R$$

$$\mathcal{T} = [|\mathbf{A} \; ; \; \mathbf{B} \; ; \; \mathbf{C}|]$$

CPD: if unique, estimated factors should match true factors.

CPD and the problem of interpretability

$$\mathcal{T} = (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}) \qquad \mathcal{I}_R$$

$$\mathcal{T} = [|\mathbf{A} \; ; \; \mathbf{B} \; ; \; \mathbf{C}|]$$

CPD: if unique, estimated factors should match true factors.

Interpretability problems

- ✗ Presence of noise deteriorates solutions.
- ✗ Uniqueness may not be ensured.
- ✗ Finding the best low-rank approximation is NP-hard in general.

Consequence: results may not have any physical meaning.

CPD and the problem of interpretability

$$\mathcal{T} = (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}) \qquad \mathcal{I}_R$$

$$\mathcal{T} = [|\mathbf{A} \; ; \; \mathbf{B} \; ; \; \mathbf{C}|]$$

CPD: if unique, estimated factors should match true factors.

Interpretability problems

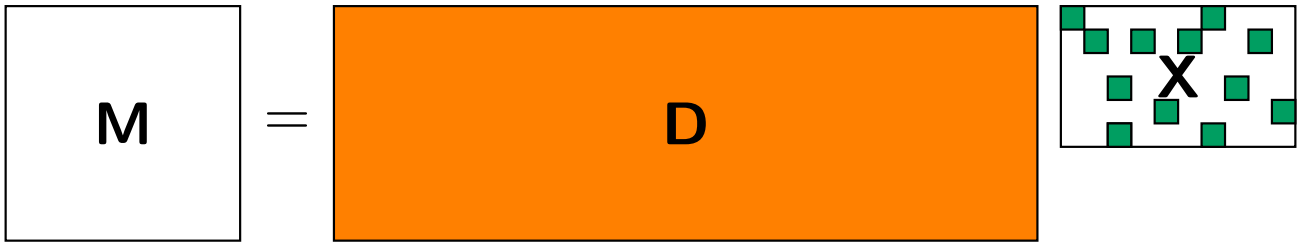
- ✗ Presence of noise deteriorates solutions.
- ✗ Uniqueness may not be ensured.
- ✗ Finding the best low-rank approximation is NP-hard in general.

Consequence: results may not have any physical meaning.

- ✓ Using dictionaries guarantees interpretability.

Dictionaries in signal processing (in a nutshell)

$$\mathbf{M} = \mathbf{DX}, \quad \|\mathbf{X}\|_0 < \delta$$



Fixed Dictionary (Sparse Coding)

minimize $\sum_i \|\mathbf{m}_i - \mathbf{D}(:, \mathcal{K}_i)\mathbf{x}_i\|_2^2$
over $\mathcal{K}_i, \mathbf{x}_i$.
[Donoho 03, Tropp 04]

Learnt Dictionary (DL,SCA)

minimize $\sum_i \|\mathbf{m}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_1$
over $\mathbf{D}, \mathbf{x}_i$.
Algorithms: [Olshausen 97, Elad 06, Mairal 10],

- Other forms of regularization (low rank, group sparsity, separability, nonnegativity, tensor structure).
- Dictionary learning and sparse coding are among the key topics in both signal processing and machine learning.

Sparse coding for tensors?

Put the constraints on $\mathbf{A}$.

$$\mathcal{T} = (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\, \mathcal{I}_R$$

where $|\mathcal{K}| = R$.

Sparse coding for tensors?

Put the constraints on **A**.

$$\mathcal{T} = (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\, \mathcal{I}_R$$

where $|\mathcal{K}| = R$.

- ✓ Generalizes easily to any order.
- ✓ Alternating algorithms can be adapted easily. Low memory requirement.
- ✓ Can be adapted for multiple atom selection (future work).

Theoretical gain

Theorem (Matrix factorization is identifiable)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$, $|\mathcal{K}| = R$, *and if there exists* $\mathbf{M} = \mathbf{D}(:, \mathcal{K})\mathbf{B}$, *then this factorization is essentially unique.*

Theoretical gain

Theorem (Matrix factorization is identifiable)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$, $|\mathcal{K}| = R$, *and if there exists* $\mathbf{M} = \mathbf{D}(:, \mathcal{K})\mathbf{B}$, *then this factorization is essentially unique.*

Theorem (Tensor factorization is often identifiable)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$, $|\mathcal{K}| = R$, *and if there exists* $\boldsymbol{\mathcal{T}} = (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\boldsymbol{\mathcal{I}}_R$, *then the following holds:*

$$(\mathbf{B} \odot \mathbf{C}) \text{ is full rank} \Rightarrow \text{the factorization is unique.}$$

Theoretical gain

Theorem (Matrix factorization is identifiable)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$, $|\mathcal{K}| = R$, *and if there exists* $\mathbf{M} = \mathbf{D}(:, \mathcal{K})\mathbf{B}$, *then this factorization is essentially unique.*

Theorem (Tensor factorization is often identifiable)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$, $|\mathcal{K}| = R$, *and if there exists* $\boldsymbol{\mathcal{T}} = (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\boldsymbol{\mathcal{I}}_R$, *then the following holds:*

$$(\mathbf{B} \odot \mathbf{C}) \text{ is full rank} \Rightarrow \text{the factorization is unique.}$$

Theorem (3d order tensor best approximation exists)

If $spark(\mathbf{D}) > R$, $R = rank(\mathbf{M})$ *and* $\#\mathcal{K} = R$, *then the minimum of*

$$f(\mathcal{K}, \mathbf{B}, \mathbf{C}) = \|\boldsymbol{\mathcal{T}} - (\mathbf{D}(:, \mathcal{K}) \otimes \mathbf{B} \otimes \mathbf{C})\boldsymbol{\mathcal{I}}_R\|_F^2$$

over the variables $(\mathcal{K}, \mathbf{B}, \mathbf{C})$ *always exists.*

Yet another alternating algorithm

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathcal{K}} \quad \|\mathcal{T} - (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\,\mathcal{I}_R\|_F^2 + \lambda \|\mathbf{A} - \mathbf{D}(:, \mathcal{K})\|_F^2.$$

Iterate until convergence:

1. Factors are updated by any well-known algorithm (ALS, gradient-based methods...).
2. $\mathcal{K}$ is obtained by finding the closest atom in $\mathbf{D}$ for each column of $\mathbf{A}$.
3. Increase λ if necessary.

Yet another alternating algorithm

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\mathcal{K}} \quad \|\mathcal{T} - (\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C})\,\mathcal{I}_R\|_F^2 + \lambda \|\mathbf{A} - \mathbf{D}(:,\mathcal{K})\|_F^2.$$

Iterate until convergence:

1. Factors are updated by any well-known algorithm (ALS, gradient-based methods...).
2. $\mathcal{K}$ is obtained by finding the closest atom in $\mathbf{D}$ for each column of $\mathbf{A}$.
3. Increase λ if necessary.

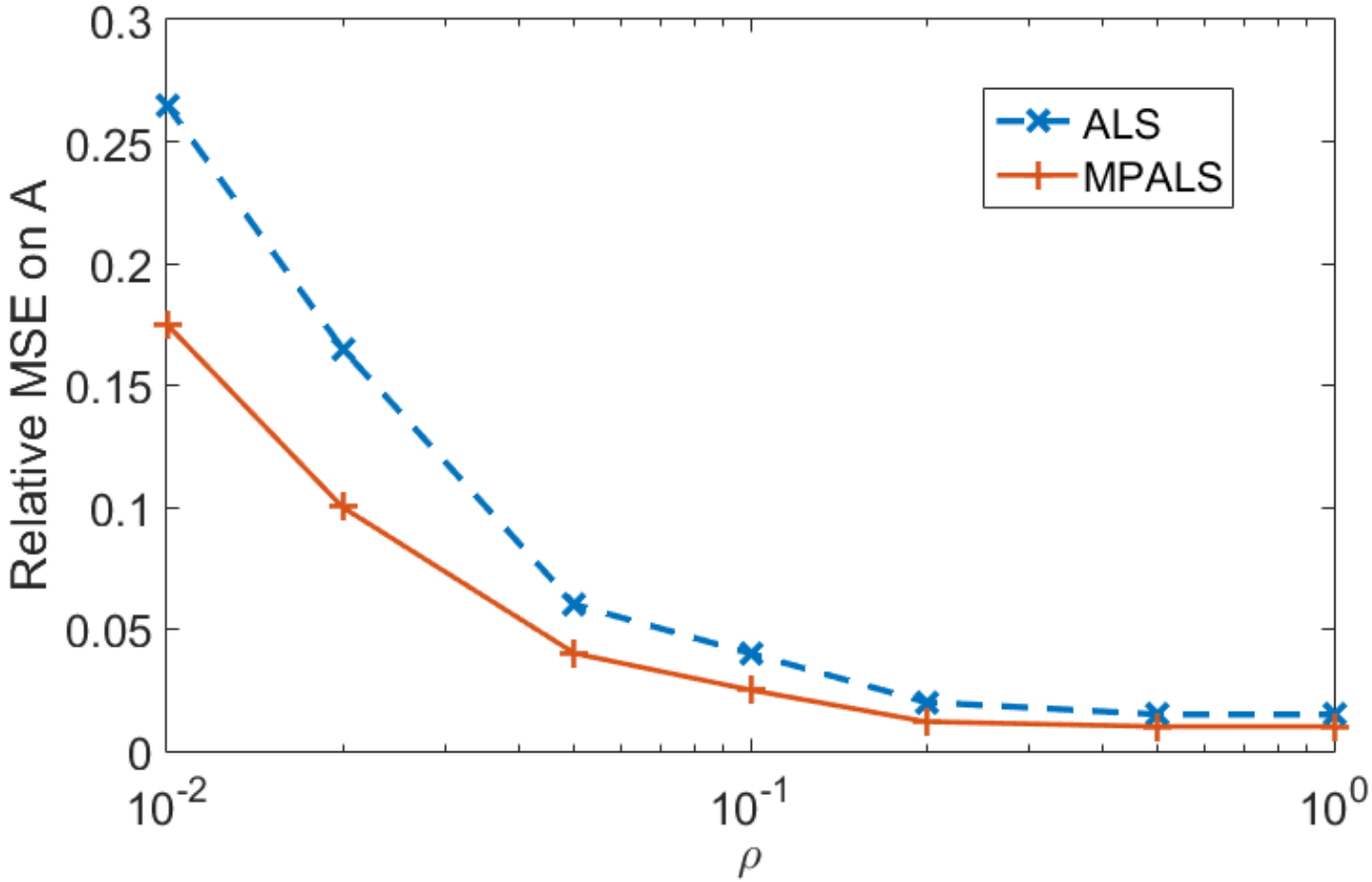
Tricks:

- To impose that no atom is selected twice, solve an assignment problem.
- If factors are constrained, simply use any off-the-shelf solver.
- Parameter λ may be tuned for naive flexible dictionary constraint.

[CG18] Cohen, G., Dictionary-based Tensor Canonical Polyadic Decomposition, IEEE Trans. on Signal Processing, 2018.

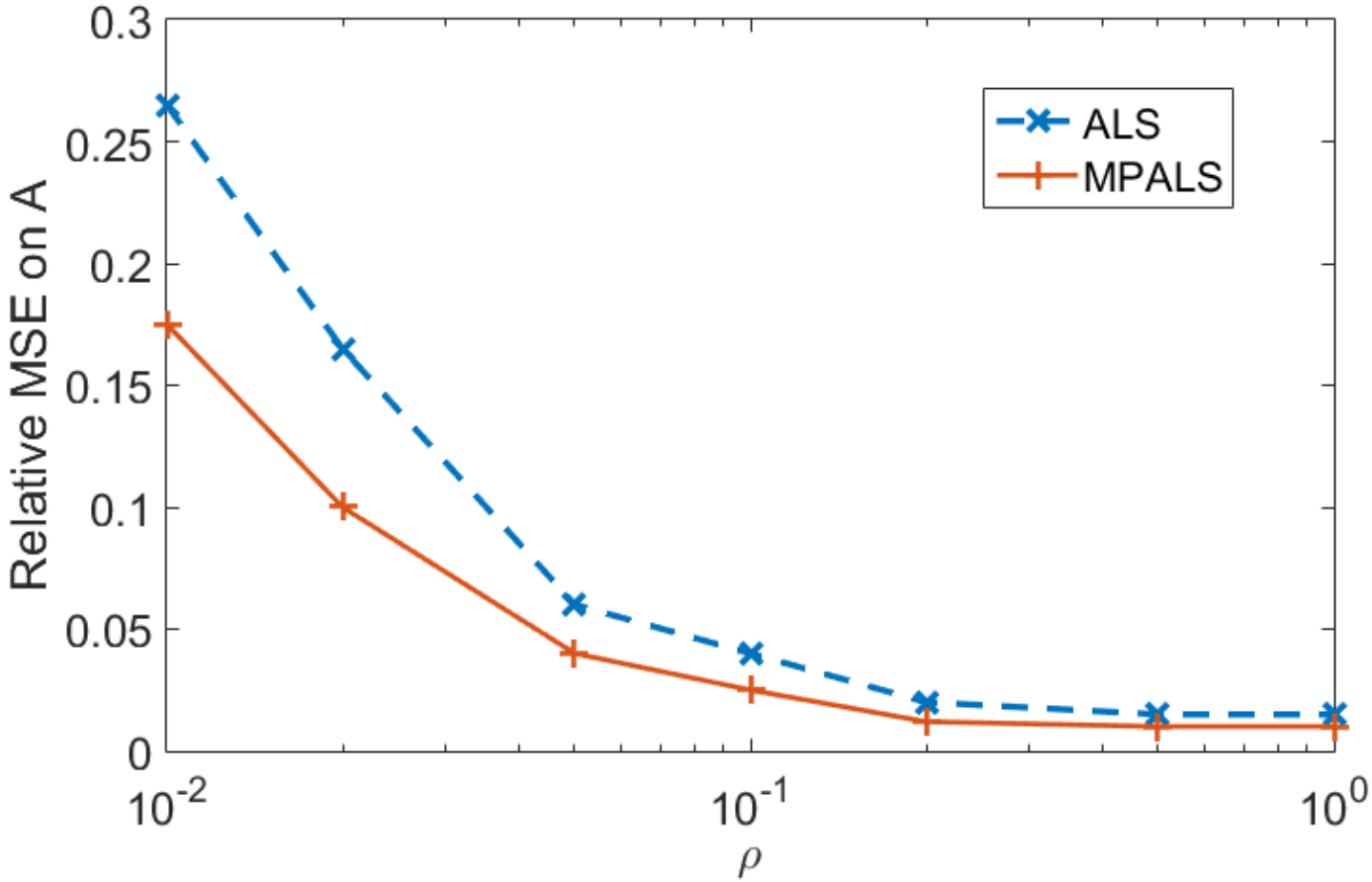
Numerical results: ALS vs. Dictionary-ALS (MPALS)

$\mathbf{A}^{(0)} = \mathbf{D}(:, \mathcal{K}), \mathbf{B}^{(0)}, \mathbf{C}^{(0)}$ randomly generated, $\mathbf{D}$ ill-conditioned, SNR$\sim$11.5dB, and use $\mathbf{C} \leftarrow \mathbf{C}^{(0)} \left(\rho \mathbf{I}_R + \frac{(1-\rho)}{R} \mathbf{1}_{R \times R} \right)$ (ill-conditioned).



Numerical results: ALS vs. Dictionary-ALS (MPALS)

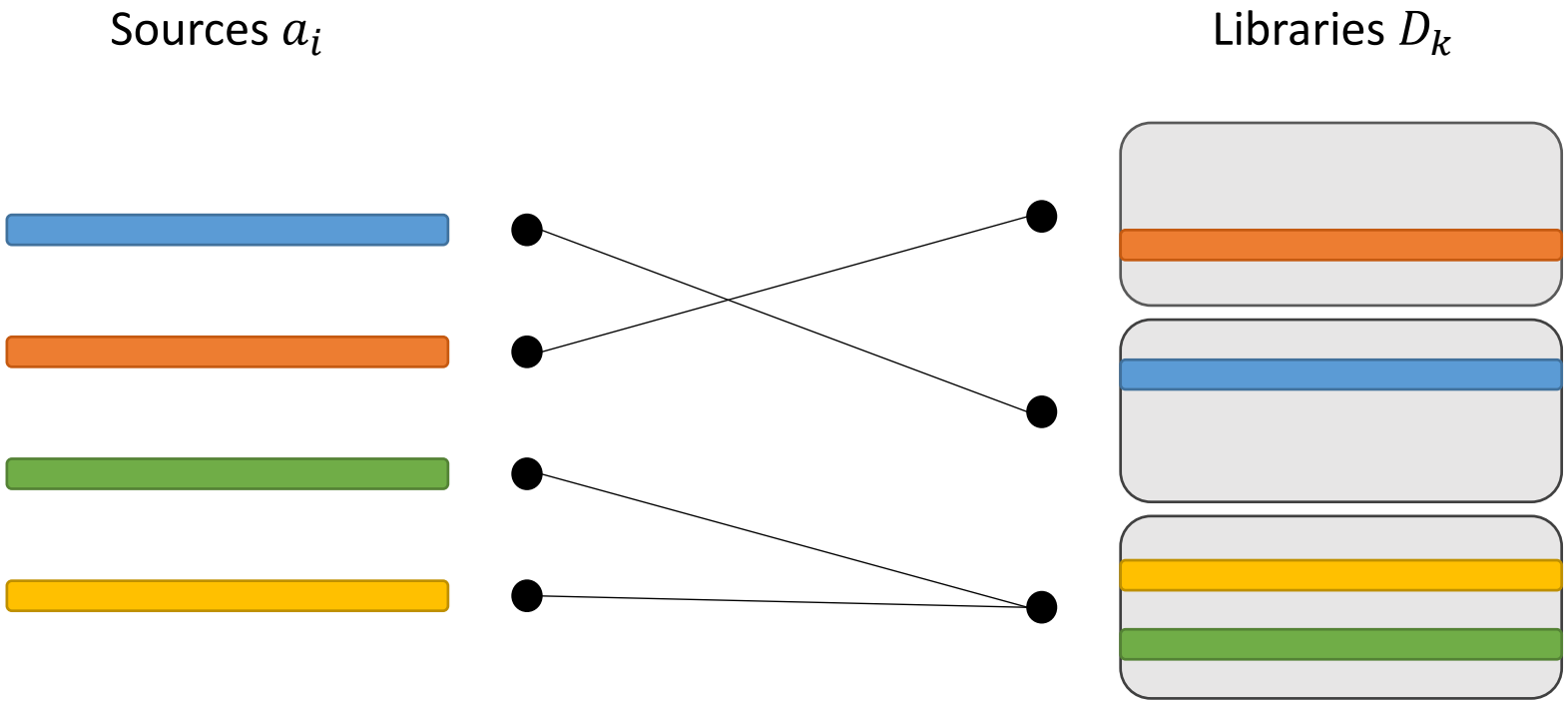
$\mathbf{A}^{(0)} = \mathbf{D}(:, \mathcal{K}), \mathbf{B}^{(0)}, \mathbf{C}^{(0)}$ randomly generated, $\mathbf{D}$ ill-conditioned, SNR$\sim$11.5dB, and use $\mathbf{C} \leftarrow \mathbf{C}^{(0)} \left(\rho \mathbf{I}_R + \frac{(1-\rho)}{R} \mathbf{1}_{R \times R} \right)$ (ill-conditioned).



We need real-world tensor data to validate MPALS further.

Extensions to multiple dictionaries

$$\mathbf{A} = [\mathbf{D}_1(:, \mathcal{K}_1), \ldots, \mathbf{D}_N(:, \mathcal{K}_n)] \quad \text{with } \underline{d_i} \leq |\mathcal{K}|_i \leq \bar{d}_i \text{ and } \sum_i |\mathcal{K}|_i = R.$$



[CG18] Cohen, G., Spectral Unmixing with Multiple Dictionaries, IEEE Geoscience and Remote Sensing Letters, 2018.

Low-rank sparse component analysis

Problem formulation: Low-rank SCA

Decompose a low rank matrix/tensor with known coefficients sparsity.

$$\text{LRSCA}: \begin{cases} \mathbf{M} = \mathbf{AB}, \quad \mathbf{A} \in \mathbb{R}^{d \times R}, \mathbf{B} \in \mathbb{R}^{R \times n}, \\ rank(M) = rank(A) = R, \\ \|\mathbf{B}(:, i)\|_0 \leq k < R. \end{cases}$$

Problem formulation: Low-rank SCA

Decompose a low rank matrix/tensor with known coefficients sparsity.

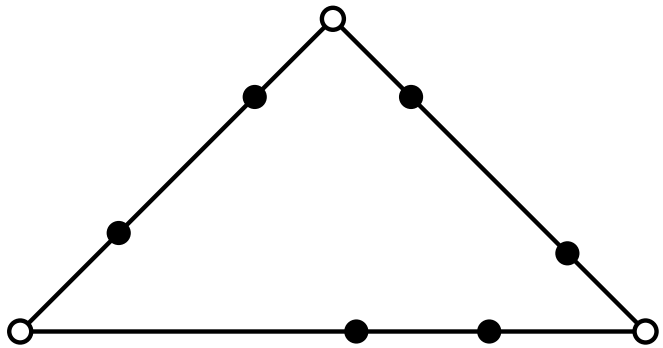
$$\text{LRSCA}: \begin{cases} \mathbf{M} = \mathbf{AB}, \quad \mathbf{A} \in \mathbb{R}^{d \times R}, \mathbf{B} \in \mathbb{R}^{R \times n}, \\ rank(M) = rank(A) = R, \\ \|\mathbf{B}(:, i)\|_0 \leq k < R. \end{cases}$$

Many existing theoretical results (see e.g. [Gribonval 16]) and algorithms (Dictionary Learning). But:

- ✗ Not many results specific to the low-rank case
- ✗ Only two deterministic identifiability results [Elad 06, Georgiev 05]
- ✗ Not much in the tensor case except ℓ_1 regularization?

Geometric intuition

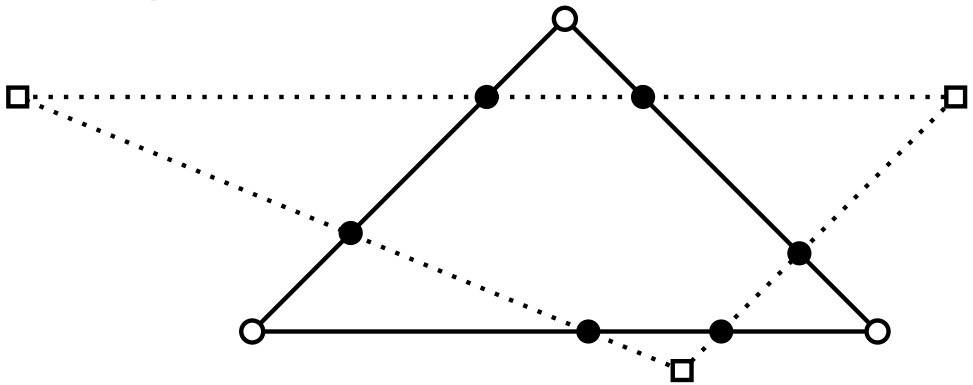
Example: $d = 3, R = 3, k = 2, n = 6$.



- ● data points
- ○ first decomposition

Geometric intuition

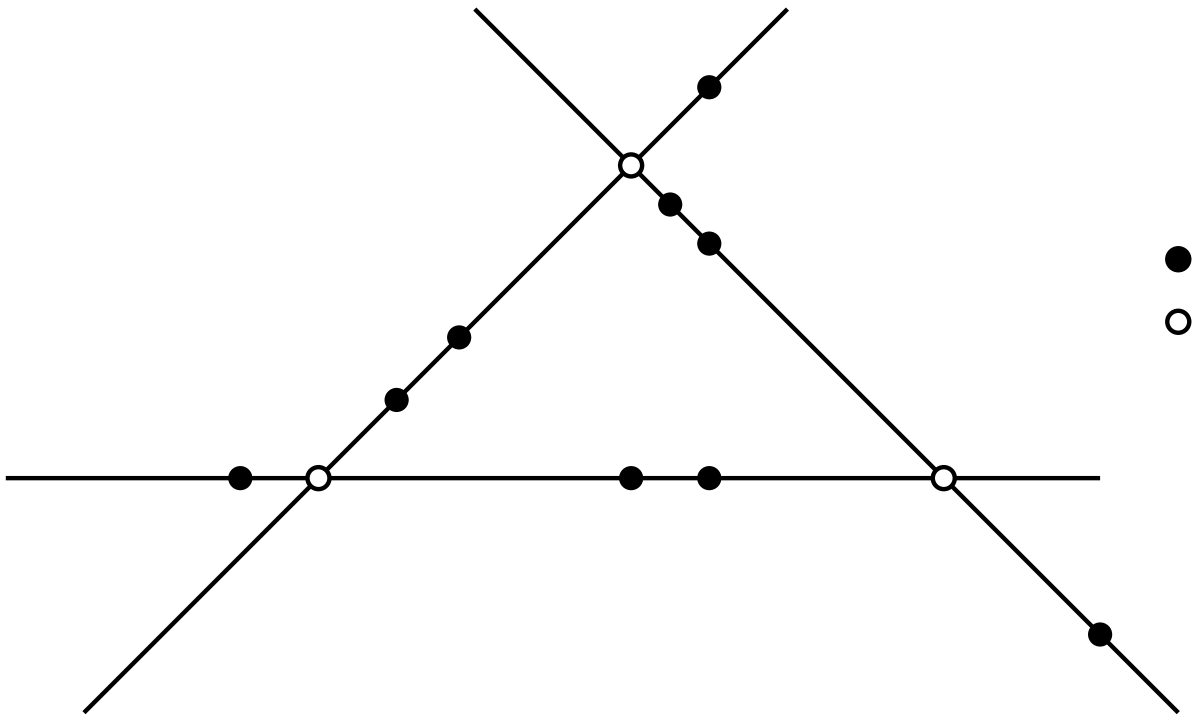
Example: $d = 3, R = 3, k = 2, n = 6$.



- ● data points
- ○ first decomposition
- □ second decomposition

Geometric intuition

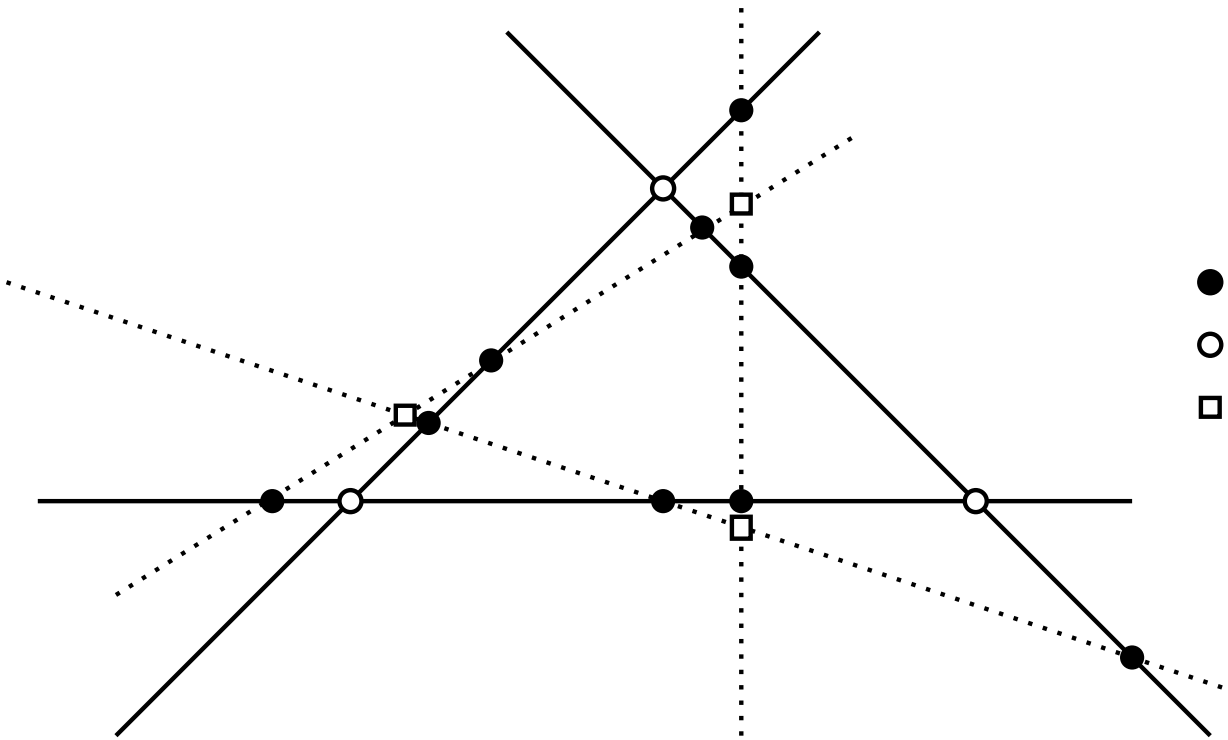
Example: $d = 3, R = 3, k = 2, n = 6$.



- ● data points
- ○ first decomposition

Geometric intuition

Example: $d = 3, R = 3, k = 2, n = 6$.



- ● data points
- ○ first decomposition
- □ second decomposition

Our identifiability results

Theorem

Let $M = AB$ satisfy the LRSCA model. The factorization (A, B) is essentially unique if on each hyperplane spanned by all but one column of $\mathbf{A}$, there are $\left\lfloor \frac{R(R-2)}{R-k} \right\rfloor + 1$ data points with spark R.

[CG18] Cohen, G., Identifiability of Low-Rank Sparse Component Analysis, arXiv:1808.08765.

Our identifiability results

Theorem

Let $M = AB$ satisfy the LRSCA model. The factorization (A, B) is essentially unique if on each hyperplane spanned by all but one column of $\mathbf{A}$, there are $\left\lfloor \frac{R(R-2)}{R-k} \right\rfloor + 1$ data points with spark R.

✓ For $k = R - 1$, this requires $R^3 - 2R^2 + R$ data points and it is tight up to the constant R (counter examples for any $n = R^3 - 2R^2$).

[CG18] Cohen, G., Identifiability of Low-Rank Sparse Component Analysis, arXiv:1808.08765.

Our identifiability results

Theorem

Let $M = AB$ satisfy the LRSCA model. The factorization (A, B) is essentially unique if on each hyperplane spanned by all but one column of $\mathbf{A}$, there are $\left\lfloor \frac{R(R-2)}{R-k} \right\rfloor + 1$ data points with spark R.

- ✓ For $k = R - 1$, this requires $R^3 - 2R^2 + R$ data points and it is tight up to the constant R (counter examples for any $n = R^3 - 2R^2$).
- ✓ For $k = 1$, this requires R data points and it is tight (one on each intersection of $R - 1$ hyperplanes).

[CG18] Cohen, G., Identifiability of Low-Rank Sparse Component Analysis, arXiv:1808.08765.

Our identifiability results

Theorem

Let $M = AB$ satisfy the LRSCA model. The factorization (A, B) is essentially unique if on each hyperplane spanned by all but one column of $\mathbf{A}$, there are $\left\lfloor \frac{R(R-2)}{R-k} \right\rfloor + 1$ data points with spark R.

- ✓ For $k = R - 1$, this requires $R^3 - 2R^2 + R$ data points and it is tight up to the constant R (counter examples for any $n = R^3 - 2R^2$).
- ✓ For $k = 1$, this requires R data points and it is tight (one on each intersection of $R - 1$ hyperplanes).
- ✓ It is tight up to constant factors for any $k = \beta R$ for any fixed constant β.

[CG18] Cohen, G., Identifiability of Low-Rank Sparse Component Analysis, arXiv:1808.08765.

Our identifiability results

Theorem

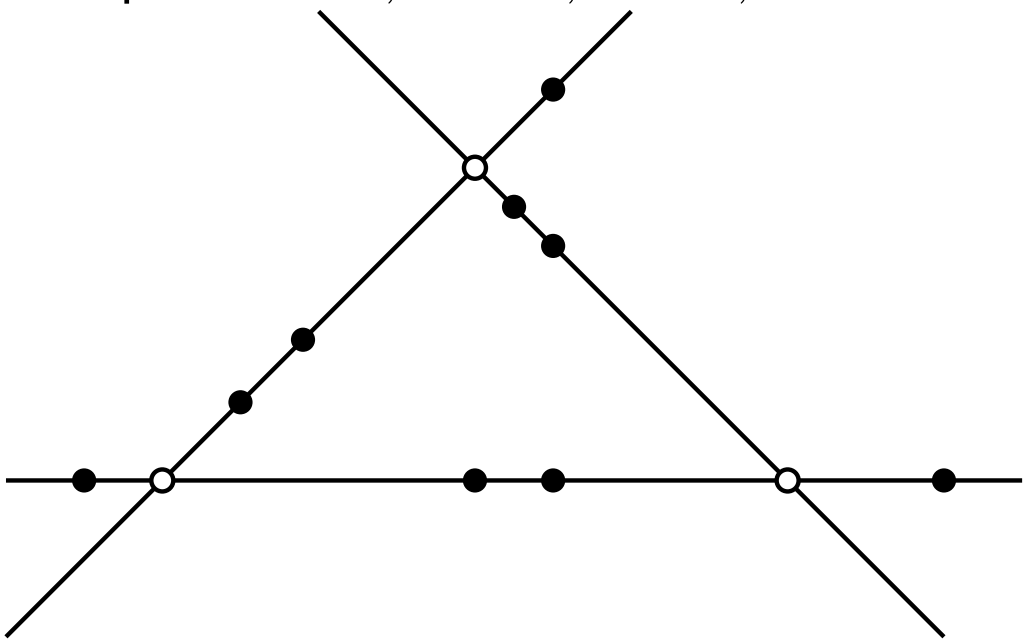
Let $M = AB$ satisfy the LRSCA model. The factorization (A, B) is essentially unique if on each hyperplane spanned by all but one column of $\mathbf{A}$, there are $\left\lfloor \frac{R(R-2)}{R-k} \right\rfloor + 1$ data points with spark R.

- ✓ For $k = R - 1$, this requires $R^3 - 2R^2 + R$ data points and it is tight up to the constant R (counter examples for any $n = R^3 - 2R^2$).
- ✓ For $k = 1$, this requires R data points and it is tight (one on each intersection of $R - 1$ hyperplanes).
- ✓ It is tight up to constant factors for any $k = \beta R$ for any fixed constant β.
- ✓ Nonnegativity helps both in theory and in practice: further work.

[CG18] Cohen, G., Identifiability of Low-Rank Sparse Component Analysis, arXiv:1808.08765.

Geometric intuition

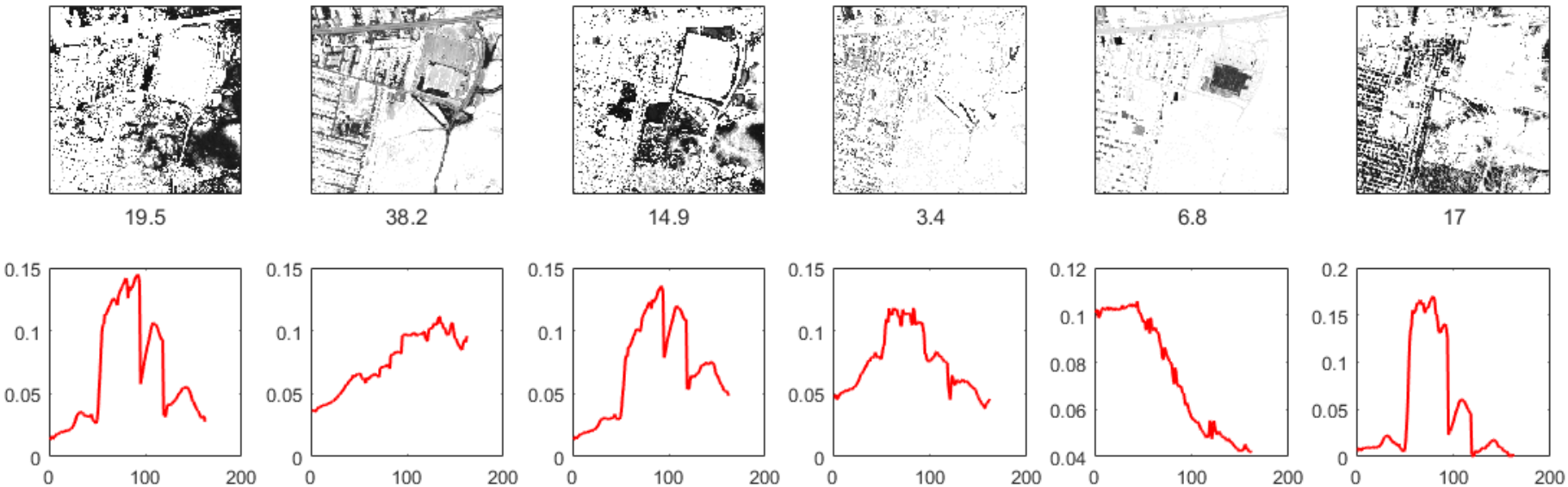
Example: $d = 3, R = 3, k = 2, n = 4 + 3 + 2$.



- ● data points
- ○ unique decomposition

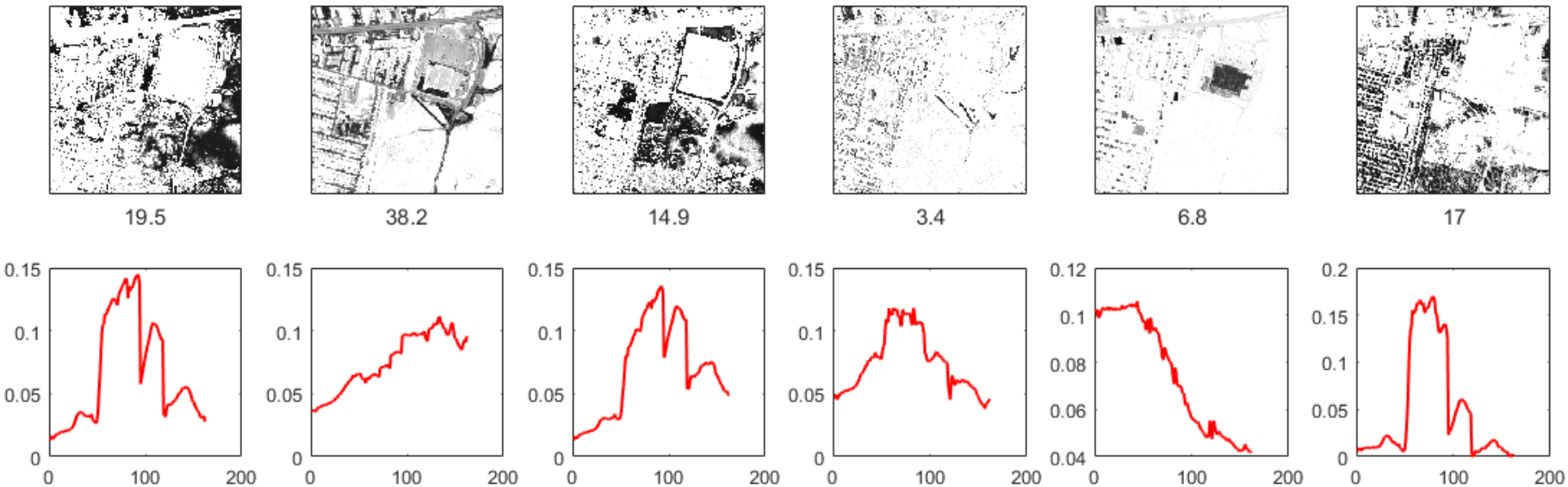
Sparsity in action

Spectral unmixing, $R = 6, k = 2$



Sparsity in action

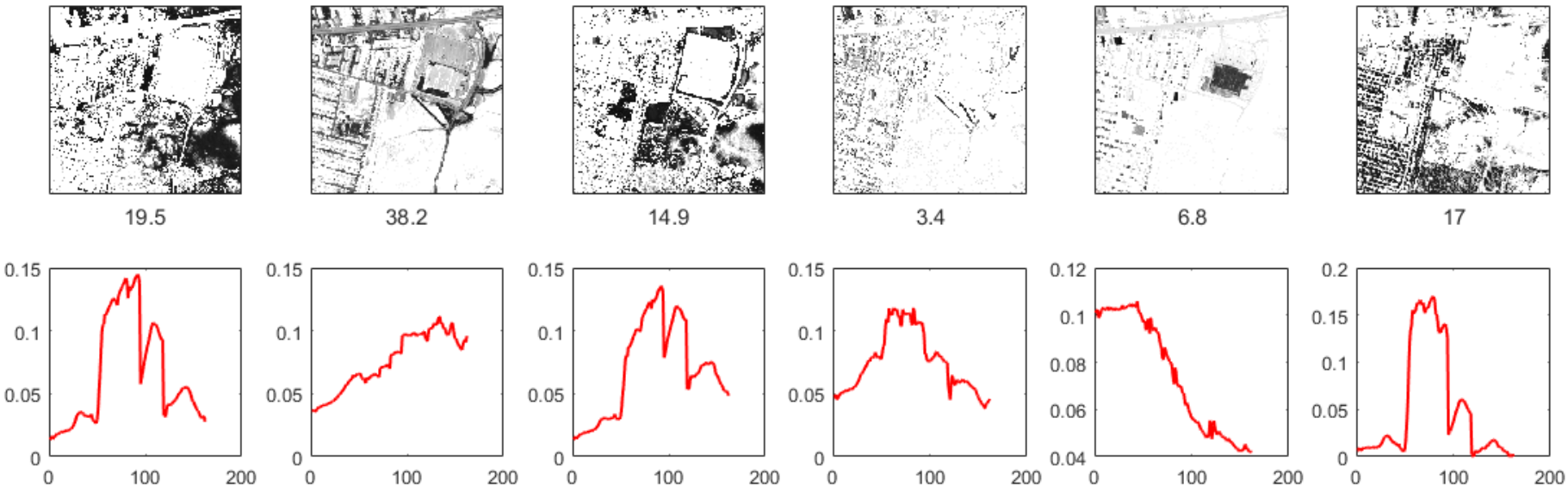
Spectral unmixing, $R = 6, k = 2$



✓ Sparsity is another way to obtain **identifiability** for matrix decompositions.

Sparsity in action

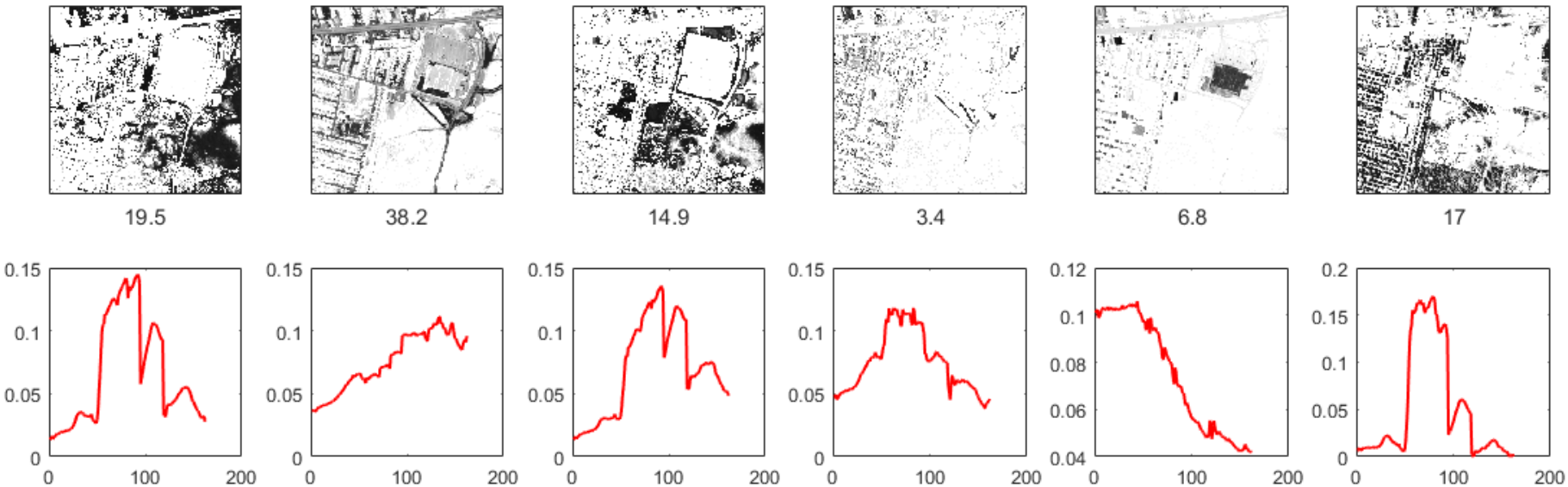
Spectral unmixing, $R = 6, k = 2$



- ✓ Sparsity is another way to obtain **identifiability** for matrix decompositions.
- ✓ Need to explore the extension to tensors.

Sparsity in action

Spectral unmixing, $R = 6, k = 2$



- ✓ Sparsity is another way to obtain **identifiability** for matrix decompositions.
- ✓ Need to explore the extension to tensors.
- ✗ Hard combinatorial problems to solve. . .

Take-home messages

1. NMF is a useful and widely used linear model in data analysis and machine learning.

Take-home messages

1. NMF is a useful and widely used linear model in data analysis and machine learning.
2. NMF is difficult (NP-hard) and ill-posed (non-uniqueness).

Take-home messages

1. NMF is a useful and widely used linear model in data analysis and machine learning.
2. NMF is difficult (NP-hard) and ill-posed (non-uniqueness).
3. NMF with (self-)dictionary is tractable and well-posed (separable NMF).

Take-home messages

1. NMF is a useful and widely used linear model in data analysis and machine learning.
2. NMF is difficult (NP-hard) and ill-posed (non-uniqueness).
3. NMF with (self-)dictionary is tractable and well-posed (separable NMF).
4. Dictionary-based decompositions lead to identifiability both for matrices and tensors. We have proposed a simple alternating scheme that works well in practice.

Take-home messages

1. NMF is a useful and widely used linear model in data analysis and machine learning.
2. NMF is difficult (NP-hard) and ill-posed (non-uniqueness).
3. NMF with (self-)dictionary is tractable and well-posed (separable NMF).
4. Dictionary-based decompositions lead to identifiability both for matrices and tensors. We have proposed a simple alternating scheme that works well in practice.
5. Another way to get identifiability is sparsity.

Thank you for your attention!

Code and papers available from
https://sites.google.com/site/nicolasgillis