The Why and How of Nonnegative Matrix Factorization

Topic Presentation

Group 2 — Aurlie de Borman, Jennifer Leclipteur, Gilles Peiffer, and Minh-Phuong Tran

LINMA2380 — Matrix computations

December 10, 2020

Summary

Use : Analysis of high-dimensional data by automatically extracts sparse and meaningful features from a set of nonnegative data vectors

1. What : Definitions as introduction
2. Why : Applications
3. How : Formal view and algorithmic difficulties
4. What next : Connections to other problems
5. Conclusion

What : Definitions and properties

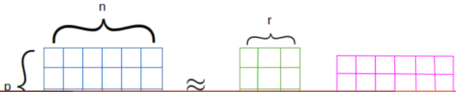
Nonnegative matrix factorization (NMF) is a Linear dimensionality reduction (LDR)

NMF : decomposing a given nonnegative data matrix X as $X \approx WH$ where $W \geq 0$ and $H \geq 0$

LDR :

- From a set of data points $x_j \in R^p$ for $1 \leq j \leq n$
- To a set of dimension $r < min(p, n)$
- Thanks to $w_k \in R^p$ for $1 \leq k \leq r$
- Such that : $\forall j, x_j \approx \sum_{k=1}^{r} w_k h_j(k)$, for some weights $h_j \in R^r$

Equivalent to **low-rank matrix approximation** : $X \approx WH$



Applications - Image processing

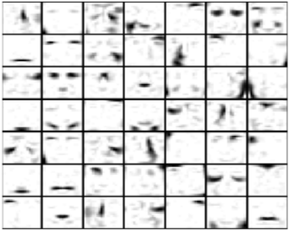
Goal : Facial Feature Extraction

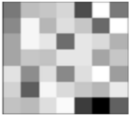
Data matrix : $X \in \mathbb{R}_{+}^{p \times n}$

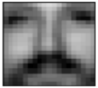
- p : total number of pixels
- n : number of faces
- $X(i,j)$: the gray-level of the i-th pixel in the j-th face

Applications - Image processing

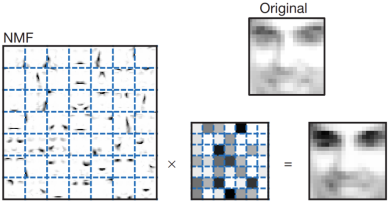
$$\underbrace{X(:,j)}_{j\text{th facial image}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{\text{facial features}} \underbrace{H(k,j)}_{\text{importance of features in } j\text{th image}} = \underbrace{WH(:,j)}_{\text{approximation of } j\text{th image}}$$



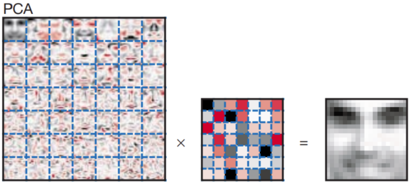




Applications - Image processing



NMF decomposition



PCA decomposition

Applications - Text Mining

Goal : Topic Recovery and Document Classification

Data matrix : $X \in \mathbb{R}_+^{n \times m}$

- each column : a document
- each line : a word
- $X(i,j)$: number of times the i-th word appears in the j-th document

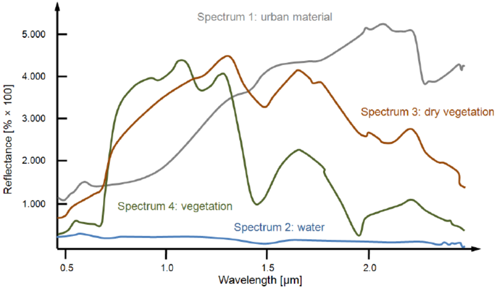
$$\underbrace{X(:,j)}_{j\text{th document}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{k\text{th topic}} \underbrace{H(k,j)}_{\substack{\text{importance of } k\text{th topic} \\ \text{in } j\text{th document}}}, \quad \text{with } W \geq 0 \text{ and } H \geq 0.$$

Applications - Hyperspectral Unmixing

Goal :

1. Identify the constitutive materials present in an image
2. Classify the pixels according to their constitutive materials

Spectral signature of a pixel: fraction of incident light being reflected by that pixel at different wavelengths

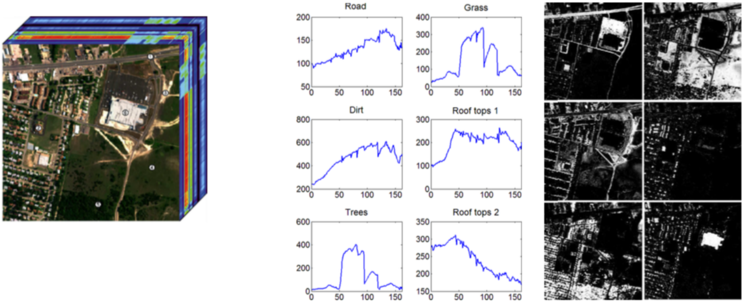


Applications - Hyperspectral Unmixing

Data matrix : $X \in \mathbb{R}^{n \times p}$

- each column : spectral signature of a pixel

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



Optimization Problem

- **Mathematical formulation**: $\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} \|X - WH\|_{\mathsf{F}}^2$, such that $W \geqslant 0$, $H \geqslant 0$.

Optimization Problem

- **Mathematical formulation**: $\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} \|X - WH\|_{\mathsf{F}}^2$, such that $W \geqslant 0$, $H \geqslant 0$.
- Frobenius norm **assumption**: noise is *Gaussian*.

Optimization Problem

- **Mathematical formulation**: $\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} \|X - WH\|_{\mathsf{F}}^2$, such that $W \geqslant 0$, $H \geqslant 0$.
- Frobenius norm **assumption**: noise is *Gaussian*.
- Other possibilities:
  - Kullback–Leibler divergence, used in text mining;
  - Itakura–Saito distance, used in music analysis;
  - ℓ_1 norm to improve robustness against outliers;
  - etc.

Issues

- NMF is **NP-hard**.
  - Because of nonnegativity *constraints*.
  - Unconstrained case can be solved efficiently using SVD.
  - Usually, NMF algorithms make certain *assumptions* and use *heuristics* to be faster.

Issues

- NMF is **NP-hard**.
  - Because of nonnegativity *constraints*.
  - Unconstrained case can be solved efficiently using SVD.
  - Usually, NMF algorithms make certain *assumptions* and use *heuristics* to be faster.
- NMF is **ill-posed**. Several “solutions” exist:
  - Using *priors* on the factors W and H (e.g. sparsity).
  - Appropriate *regularization* in the objective function.
  - Finding *application-specific solutions* is a very active area of research!

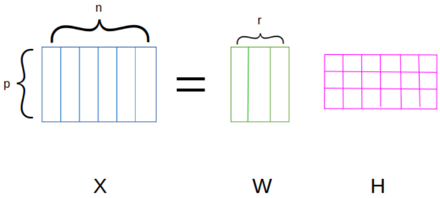
Issues

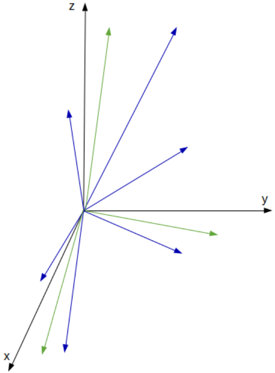
- NMF is **NP-hard**.
  - Because of nonnegativity *constraints*.
  - Unconstrained case can be solved efficiently using SVD.
  - Usually, NMF algorithms make certain *assumptions* and use *heuristics* to be faster.
- NMF is **ill-posed**. Several "solutions" exist:
  - Using *priors* on the factors W and H (e.g. sparsity).
  - Appropriate *regularization* in the objective function.
  - Finding *application-specific solutions* is a very active area of research!
- Choice of **factorization rank** r.
  - *Trial and error*.
  - Based on *SVD*.
  - Using *expert knowledge* (e.g. number of endmembers for HU application).

Nonnegative rank

Definition (Nonnegative rank)

Given $X \in \mathbb{R}_+^{p\times n}$*, the nonnegative rank of X, denoted* $\text{rank}_+(X)$ *is the minimum* r *s.t.* $\exists W \in \mathbb{R}_+^{p\times r}, H \in \mathbb{R}_+^{r\times n}$ *with* $X = WH$*.*



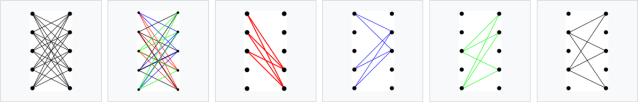


Graph Theory : Bipartite dimension

Let $G(X) = (V_1 \cup V_2, E)$ be a bipartite graph induced by X (i.e. $(i, j) \in E \Leftrightarrow X_{ij} \neq 0$).

Definition (Biclique and bipartite dimension)

- *A biclique (or a complete bipartite graph) is a bipartite graph s.t. every vertex in V_1 is connected to every vertex in V_2.*
- *The bipartite dimension (or the minimum biclique cover) bc$(G(X))$ is the minimum number of bicliques needed to cover all edges in E.*



A bipartite graph... | ...and a covering with four bicliques | the red biclique from the cover | the blue biclique from the cover | the green biclique from the cover | the black biclique from the cover

Figure: Example for biclique edge cover [**biclique**]

Theorem (Rectangle covering bound)

$$bc(G(X)) \leq rank_+(X)$$

Linear Optimization : Extended formulation

$$\begin{aligned} \text{(LP)}\ \max \quad & c^T x \\ s.t. \quad & Ax \leq b \\ & x \in \mathbb{R}^n \geq 0 \end{aligned}$$

Definition (Extended formulation)

The extended formulation of a polytope P is a higher dimensional polytope Q and a linear projection π *s.t.* $\pi(Q) = P$.

In our LP problem, an extended formulation of the polytope $P \subset \mathbb{R}^n$ defined by the constraints $Ax \leq b$, is a polytope $Q \subset \mathbb{R}^{n+r}$ defined by $Cx + Dy \leq d$ with $y \in \mathbb{R}^r$, s.t. $\pi(Q) = P$.

The slack matrix $X(i,j) = b_i - A_i v_j$.
With v_j, the j^{th} vertex of P and $\{x \in \mathbb{R}^n | b_i - A_i x \geq 0\}$ its i^{th} facet.
The (i,j) entry measures the slack of the i^{th} inequality for the j^{th} vertex.

Theorem (Yannakis)

The minimum size of an extended formulation Q of P is equal to rank$_+(X)$.

When P has exponentially many facets, finding extended formulations allows to solve the LP in polynomial time.

Thanks for listening!

Any questions?