

The Why and How of Nonnegative Matrix Factorization

Topic Presentation

Group 02

LINMA2380 — Matrix computations

December 10, 2020

Summary

- 1 Introduction
- 2 Applications
- 3 Algorithms
- 4 Connections to other problems

Agenda

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

- 1 What : Definitions and properties
- 2 Why : Applications
- 3 How : Algorithms
- 4 What next : Connections with Problems in Mathematics and Computer Science
- 5 Conclusion

What : Definitions and properties

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

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** : $X \approx WH$

- $X \in R^{p \times n}$: $X(:, j) = x_j$ for $1 \leq j \leq n$
- $W \in R^{p \times r}$: $W(:, k) = w_k$ for $1 \leq k \leq r$
- $H \in R^{r \times n}$: $H(:, j) = h_j$ for $1 \leq j \leq n$

NMF : decomposing a given nonnegative data matrix X as $X \approx WH$

Applications - Image processing


Goal : Facial Feature Extraction





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

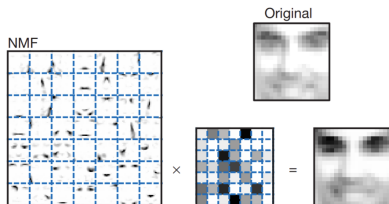
- p : total number of pixels
- n : number of faces
- $X(i, j)$: the gray-level of the i -th pixel in the j -th face

Applications - Image processing

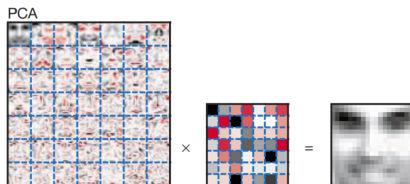
$$\underbrace{X(:,j)}_{\text{jth facial image}} \approx \sum_{k=1}^r \underbrace{W(:,k)}_{\text{facial features}}$$


$$\underbrace{H(k,j)}_{\text{importance of features in jth image}} = \underbrace{WH(:,j)}_{\text{approximation of jth image}}$$



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

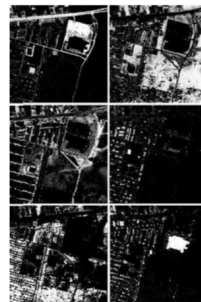
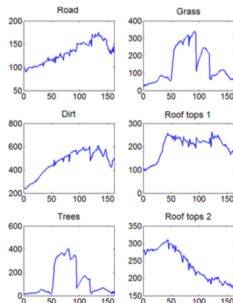
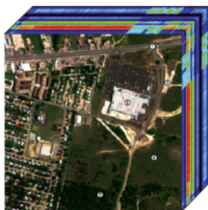
Spectral signature of a pixel: fraction of incident light being reflected by that pixel at different wavelengths

Applications - Hyperspectral Unmixing

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

- 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:** $\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2$, such that $W \geq 0, H \geq 0$.

Optimization Problem

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

Optimization Problem

- **Mathematical formulation:** $\min_{W \in \mathbb{R}^{p \times r}, H \in \mathbb{R}^{r \times n}} \|X - WH\|_F^2$, such that $W \geq 0, H \geq 0$.
- 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** (but in practice, this is rarely a problem).

Issues

- NMF is **NP-hard** (but in practice, this is rarely a problem).
- 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!

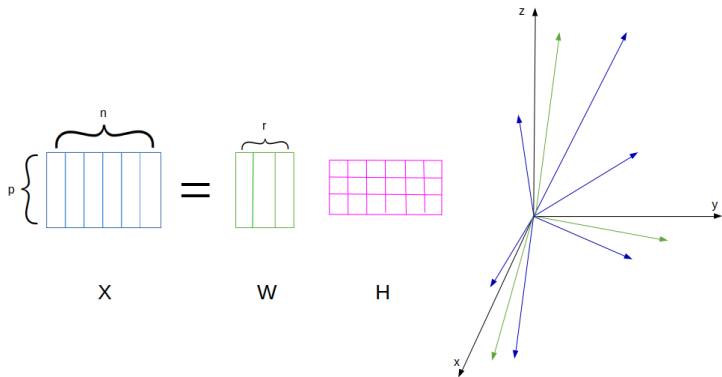
Issues

- NMF is **NP-hard** (but in practice, this is rarely a problem).
- 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 .

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 .*

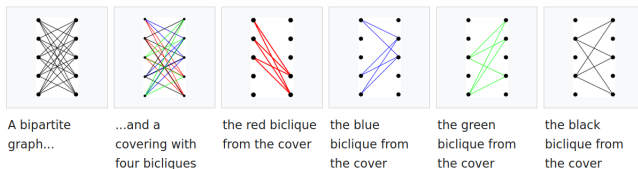


Figure: Example for biclique edge cover [**biclique**]

For any $(W, H) \geq 0$ s.t. $X = WH = \sum_{k=1}^r W_{:k} H_{k:} := \sum_{k=1}^r X_k$, we have

$$G(X) = \cup_{k=1}^r G(W_{:k} H_{k:})$$

where $G(W_{:k} H_{k:})$ are complete bipartite subgraphs
 $(bc(G(W_{:k} H_{k:})) = 1 \forall k)$.

Theorem (Rectangle covering bound)

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

Linear Optimization : Extended formulation

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

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 $\{x \in \mathbb{R}^n \mid b_i - A_i x \geq 0\}$ 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 $\text{rank}_+(X)$.

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

