The Why and How of Nonnegative Matrix Factorization

Group 2

In a few words

NMF is a powerful tool for the analysis of **high-dimensional** data as it automatically extracts **sparse** and **meaningful** features from a set of **nonnegative** vectors. NMF was introduced by Paatero and Tapper in 1994 and further developped by Lee and Seung in 1999.

Nonnegative Matrix Factorization : definition and properties

Nonnegative matrix factorization (NMF) is a Linear dimensionality reduction (LDR). As a reminder, a LDR computes a set of $r < min(p, n)$ basis elements $w_k \in R^p$ for $1 \leq k \leq r$ to approximate a set of data points $x_j \in R^p$ for $1 \leq j \leq n$ such that $\forall j, x_j \approx \sum_{k=1}^{r} w_k h_j(k)$, for some weights $h_j \in R^r$.
LDR is equivalent to low-rank matrix approximation where: $X \approx WH$ where

- $X \in R^{p\times n}$: $X(:, j) = x_j$ for $1 \leq j \leq n$. Each column of the matrix X is a data point.
- $W \in R^{p\times r}$: $W(:, k) = w_k$ for $1 \leq k \leq r$. Each column of the matrix W is a basis element.
- $H \in R^{r\times n}$: $H(:, j) = h_j$ for $1 \leq j \leq n$. Each column of the matrix H gives the coordinates of a data point X(:,j) in the basis W.

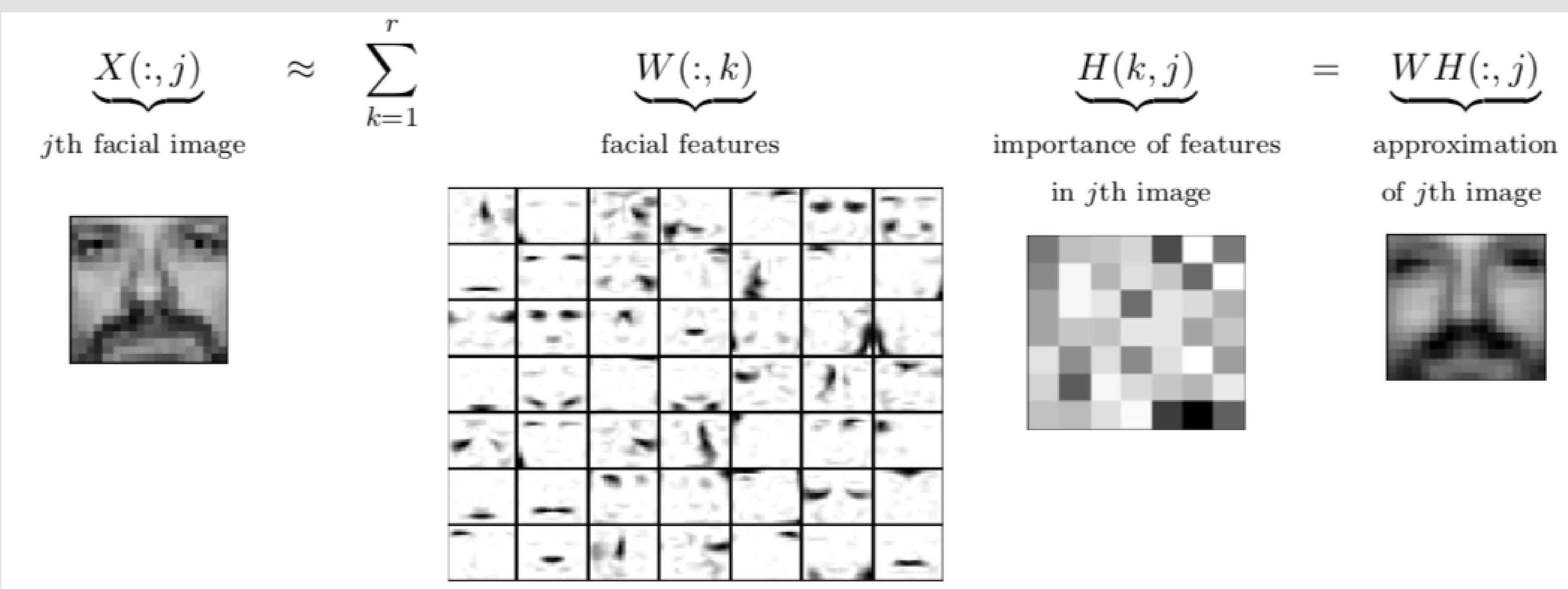
NMF : decomposing a given nonnegative data matrix X as $X \approx WH$ where $W \geq 0$ and $H \geq 0$. W and H are thus component-wise nonnegative.

Applications

Image processing

Goal: facial feature extraction
The **data matrix** $X \in_+^{p\times n}$ carries information about n face images. Each image has q pixels and therefore each column of X represents an image of a face.
$\longrightarrow$ The (i, j)th entry of X represents the gray-level of the ith pixel in the jth face.
The **nonnegative matrix factorization** can be interpreted as follows:

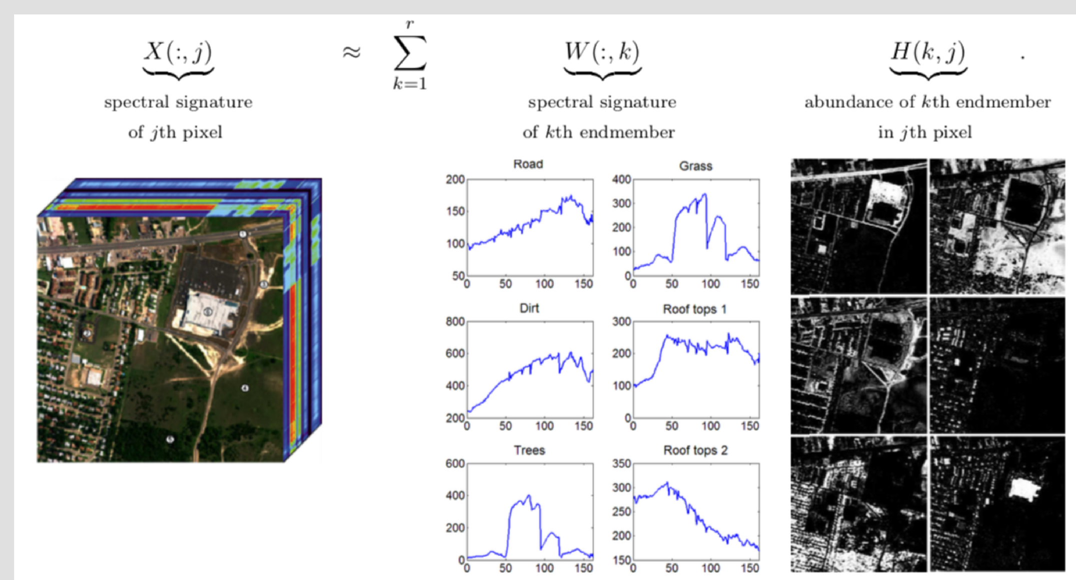
- Each column of the matrix W represents a facial feature
- The (k, j)th entry of H represents the importance of the kth feature in the jth face



Hyperspectral unmixing

Goal: identify the constitutive materials present in an image and classify the pixels according to the abundance of each constitutive material
The **data matrix** $X \in_+^{n\times n}$ encodes the spectral signatures of the pixels in a scene being imaged.
The spectral signature of a pixel is the fraction of incident light being reflected by that pixel at different wavelengths.
The **nonnegative matrix factorization** can be interpreted as follows:

- Each colum of the matrix W represents the spectral signature of a constitutive material (such as road, grass,...)
- The (k, j)th entry of H represents the abundance of the kth material in the jth pixel



Text Mining

Goal: topic recovery and document classification
The **data matrix** $X \in_+^{n\times n}$ encodes the frequency of some words in a list of documents.
$\longrightarrow$ The (i, j)th entry of X represents the number of times the ith word appears in the jth document.
The **nonnegative matrix factorization** can be interpreted as follows:

- Each colum of the matrix W represents a topic
- The (k, j)th entry of H represents the importance of the kth topic in the jth document

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

Algorithms and Difficulties

Optimization Problem

We want to solve $\min_{W\in\mathbb{R}^{p\times r}, H\in\mathbb{R}^{r\times n}} ||X - WH||_F^2$, such that $W \geqslant 0$, $H \geqslant 0$. Our use of the Frobenius norm implies the assumption that the noise is Gaussian. This is not always the best choice; in practice, depending on the application, other choices are possible too, such as:

- Kullback–Leibler divergence used in text mining;
- Itakura–Saito distance used in music analysis;
- ℓ_1 norm used to improve robustness against outliers;
- etc.

Difficulties

NMF is not a trivial task:

- NMF is **NP-hard** because of the nonnegativity constraints. In the unconstrained case, SVD can be used. In practice, assumptions and heuristics allow us to solve the problem fairly efficiently.
- NMF is **ill-posed**: if (W, H) is an NMF of X, then so is $(W', H') = (WQ, Q^{-1}H)$, where $WQ \geqslant 0$, $Q^{-1}H \geqslant 0$. This can be solved by
  - using **priors** for W and H, such as sparsity;
  - adding **regularization** terms.

  Finding application-specific solutions is an active area of research.

Connections to other problems

In this section, we will present several connections between the NMF and other mathematical problems. But first, let us introduce the nonnegative rank.

Definition (Nonnegative rank)

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

Intuitively, the columns of X can be seen as vectors that we want to generate from a basis formed by the columns of W. We want our basis to have as few vectors as possible.

Graph theory : Bipartite dimension

finding the bipartite dimension of a graph gives a lower bound for the nonnegative rank. Indeed, let $G(X) = (V_1 \cup V_2, E)$ be a bipartite graph induced by X (i.e. $(i, j) \in E \Leftrightarrow X_{ij} \neq 0$).
The **bipartite dimension** (or the minimum biclique cover) $\mathrm{bc}(G(X))$ is the minimum number of bicliques needed to cover all edges in E.
The rectangle covering bound :

$$\mathrm{bc}(G(X)) \leq \mathrm{rank}_+(X)$$

Linear Optimization : Extended Formulation

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

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

Finding the minimum size of an extended formulation of the polytope defined by the constrains amounts to calculating the nonnegative rank of 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 reason why it is interesting to find an extended formulation is that it allows to solve the (LP) in polynomial time.

Conclusion

To sum up, we have seen that the NMF does not only generate meaningful features, but also has many applications in various domains. Indeed, we can use nonnegative matrices to describe very different things, from pixel values, to word counts, graphs, linear optimization problems, but we can also find them in probability theory, communication complexity and many more. Although there are theoretical difficulties we have to deal with in this linear dimensionality reduction technique, such as its NP-hardness and ill-posedness, there still exist different ways to deal with them.

References