Data Sciences – CentraleSupelec
Advance Machine Learning
Course VI - Nonnegative matrix factorization

Emilie Chouzenoux

Center for Visual Computing
CentraleSupelec
emilie.chouzenoux@centralesupelec.fr

Motivation

Matrix factorization: Given a set of data entries $x_j \in \mathbb{R}^p$, $1 \leq j \leq n$, and a dimension $r < \min(p, n)$, we search for r basis elements w_k, $1 \leq k \leq r$ such that

$$x_j \approx \sum_{k=1}^{r} w_k h_j(k)$$

with some weights $h_j \in \mathbb{R}^r$.

Equivalent form:

$$\boxed{X \approx WH}$$

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

Motivation

$$\boxed{X \approx WH}$$

$\Rightarrow$ low-rank approximation / linear dimensionality reduction

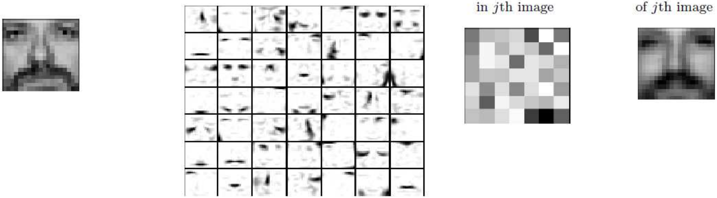
Two key aspects:

1. Which loss function to assess the quality of the approximation ?
   Typical examples: Frobenius norm, KL-divergence, logistic, Itakura-Saito.
2. Which assumptions on the structure of the factors W and H ?
   Typical examples: Independency, sparsity, normalization, non-negativity.

$$\boxed{\text{NMF:} \quad \text{find} \quad (W, H) \quad \text{s.t.} \quad X \approx WH, \quad W \geq 0, H \geq 0.}$$

Example: Facial feature extraction

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

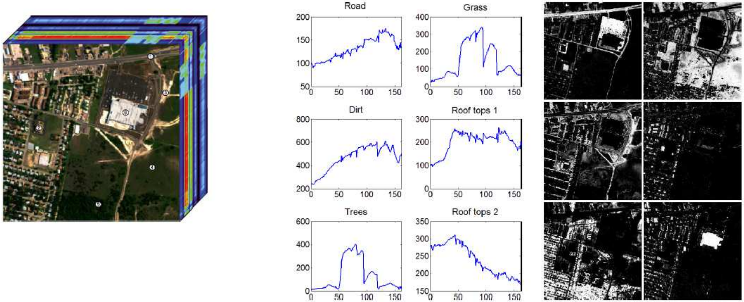


Decomposition of the CBCL face database [Lee and Seung, 1999]

⇒ Some of the features look like parts of nose or eye. Decomposition of a face as having a certain weight of a certain nose type, a certain amount of some eye type, etc.

Example: Spectral unmixing

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



Decomposition of the Urban hyperspectral image [Ma *et al.*, 2014]

$\Rightarrow$ NMF is able to compute the spectral signatures of the endmembers and simultaneously the abundance of each endmember in each pixel.

Example: Topic modeling in text mining

Goal: Decompose a term-document matrix, where each column represents a document, and each element in the document represents the weight of a certain word (e.g., term frequency - inverse document frequency). The ordering of the words in the documents is not taken into account (= bag-of-words).

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

Topic decomposition model [Blei, 2012]

⇒ The NMF decomposition of the term-document matrix yields components that could be considered as "topics", and decomposes each document into a weighted sum of topics.

Multiplicative algorithms for NMF

Challenges: NMF is NP-hard and ill-posed. Most algorithms are only guaranteed to converge to stationary point, and may be sensitive to initialization.

We present here a popular class of methods introduced in [Lee and Seung, 1999], relying on simple multiplicative updates. (Assumption: $X \geq 0$).

* *Frobenius norm:* $\|X - WH\|_F^2$

$$W \leftarrow W \circ \frac{XH^\top}{WHH^\top}$$

$$H \leftarrow H \circ \frac{W^\top X}{W^\top WH}$$

* *KL-divergence:* $\mathcal{KL}(X, WH)$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{\ell=1}^{n}(H_{k\ell}X_{i\ell}/[WH]_{i\ell})}{\sum_{\ell=1}^{n} H_{k\ell}}$$

$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^{p}(W_{ik}X_{ij}/[WH]_{ij})}{\sum_{i=1}^{p} W_{ik}}$$

Sketch of proof

The multiplicative schemes rely on the use of separable surrogate functions, majorizing the loss w.r.t. W and H, respectively:

* *Frobenius norm:* For every $(X, W, H, \bar{H}) \geq 0$, and $1 \leq j \leq n$,

$$\|Wh_j - x_j\|_2^2 \leq \sum_{i=1}^{p} \frac{1}{[W\bar{h}_j]_i} \sum_{k=1}^{r} W_{ik}\bar{H}_{kj} \left(X_{ij} - \frac{H_{kj}}{\bar{H}_{kj}} [W\bar{h}_j]_i \right)^2$$

* *KL-divergence:* For every $(X, W, H, \bar{H}) \geq 0$, and $1 \leq j \leq n$,

$$\begin{aligned} \mathcal{KL}(x_j, Wh_j) \leq \sum_{i=1}^{p} \Bigg(& X_{ij} \log X_{ij} - X_{ij} + [Wh_j]_i \\ & - \frac{X_{ij}}{[W\bar{h}_j]_i} \sum_{k=1}^{r} W_{ik}\bar{H}_{kj} \log \left(\frac{H_{kj}}{\bar{H}_{kj}} [W\bar{h}_j]_i \right) \Bigg) \end{aligned}$$

Weighted NMF

$*$ *Weigthed Frobenius norm:* $\|\Sigma \circ (X - WH)\|_F^2$

$$W \leftarrow W \circ \frac{(\Sigma \circ X)H^\top}{(\Sigma \circ WH)H^\top}$$

$$H \leftarrow H \circ \frac{W^\top(\Sigma \circ X)}{W^\top(\Sigma \circ (WH))}$$

$*$ *Weigthed KL-divergence:* $\mathcal{KL}(X, \mathrm{Diag}(p)WH\mathrm{Diag}(q))$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{\ell=1}^{n}(H_{k\ell}X_{i\ell}/(p_i[WH]_{i\ell}))}{\sum_{\ell=1}^{n} q_\ell H_{k\ell}}$$

$$H_{kj} \leftarrow H_{kj} \frac{\sum_{i=1}^{p}(W_{ik}X_{ij}/(q_j[WH]_{ij}))}{\sum_{i=1}^{p} p_i W_{ik}}$$

✓ A typical application is matrix completion to predict unobserved data, for instance in user-rating matrices. In that case, binary weights are used, signaling the position of the available entries in X.

Regularized NMF

$\ast$ *Regularized Frobenius norm:*

$$\frac{1}{2}\|X - WH\|_F^2 + \frac{\mu}{2}\|H\|_F^2 + \lambda\|H\|_1 + \frac{\nu}{2}\|W\|_F^2$$

$$W \leftarrow W \circ \frac{XH^\top}{W(HH^\top + \nu I_r)}$$

$$H \leftarrow H \circ \frac{W^\top X - \lambda \mathbf{1}_{r\times n}}{(W^\top W + \mu I_r)H}$$

✓ The ambiguity due to rescaling of (W, H) and to rotation is frozen by the penalty terms.

Other NMF algorithms

Multiplicative updates (MU) are simple to implement but they can be slow to converge, and are sensitive to initialization. Other strategies are listed below (for the Least-Squares case):

- Alternating Least Squares: First compute the unconstrained solution w.r.t. W or H and project onto nonnegative orthant. Easy to implement but oscillations can arise (no convergence guarantee). Rather powerful for initialization purposes.
- Alternating Nonnegative Least Squares: Solve constrained problem exactly, w.r.t. W and H, in alternate manner, using inner solver (e.g., projected gradient, Quasi-Newton, active set). Expensive. Useful as refinement step of a cheap MU.
- Hierarchical Alternative Least Squares: Exact coordinate descent method, updating one column of W (resp. one line of H) at a time. Simple to implement, and similar performance than MU.