Nonnegative Matrix Factorization (NMF) and Its Extensions

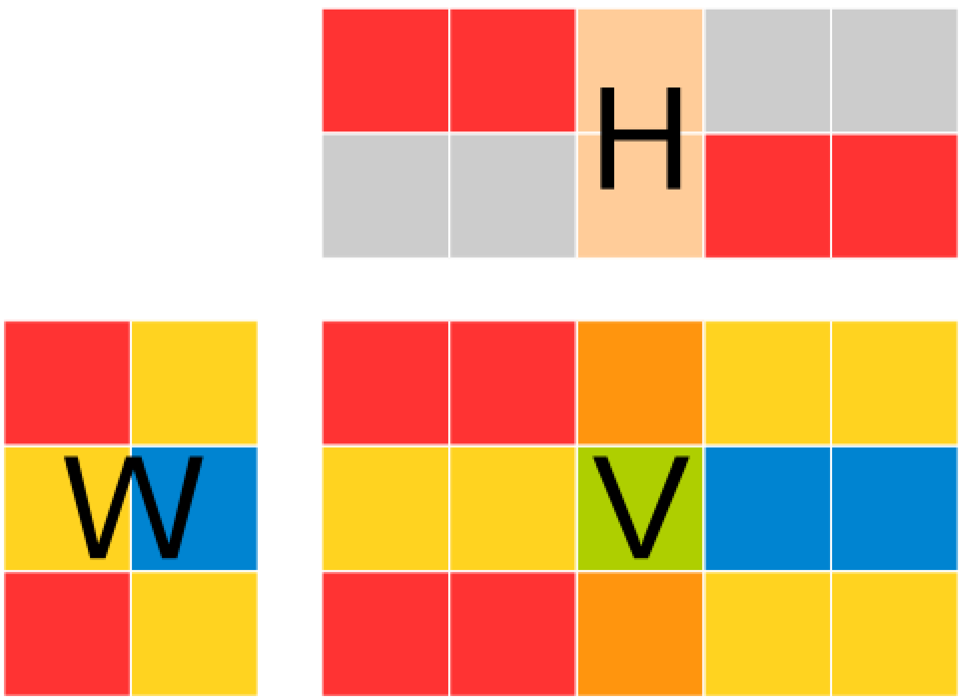
Nonnegative Matrix Factorization and Its Extensions

1. D. D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. NIPS*, 2000, pp. 556–562.
2. P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn*., vol. 5, pp. 1457-1469, 2004.
3. Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," *Proc. 14th Scandinavian Conf. Image Anal.*, 2005, pp. 333–342.
4. Chris Ding, Tao Li, and Michael I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," *IEEE trans. p*attern analysis and machine intelligence, VOL. 32, 2010.
5. X. Liu, H. Lu, and H. Gu, "Group Sparse Non-negative Matrix Factorization for Multi-Manifold Learning," Proc. the British Machine Vision Conference (BMVC), 2011, pp. 56.1-56.11.
6. Y. Liu, C. Jia, B. Li, S. Pang, and Z. Yu, "Graph Regularized Projective Non-negative Matrix Factorization for Face Recognition," *J. Comput. Inf. Sys.*, vol 9, no 5, pp. 2047-2055, 2013.
7. J. Wang, J. Z. Huang, Y. Sun, and X. Gao, "Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization," *J. Expert Systems with Applications 42* (3), 1278-1286, 2015.
8. C. Lin, and M. Pang, "Graph Regularized Nonnegative Matrix Factorization with Sparse Coding," *J. Mathematical Problems in Engineering*, 2015.

Basics NMF

The Nonnegative Matrix Factorisation model

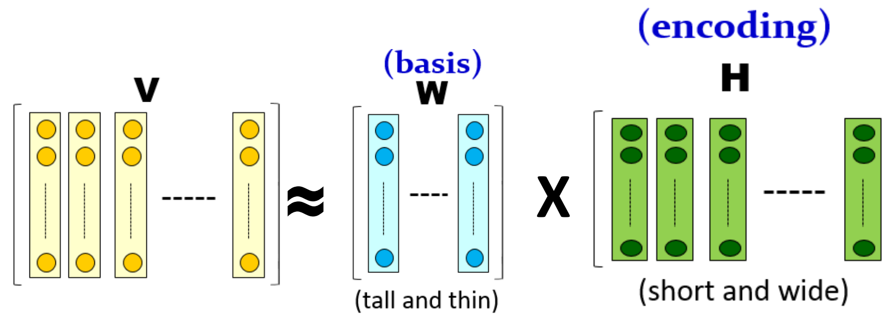
NMF provides an unsupervised linear representation of the data:



$$\mathbf{V} \approx \mathbf{W}\mathbf{H};$$

- $\mathbf{W} = [w_{fk}]$ s.t. $w_{fk} \geq 0$ and
- $\mathbf{H} = [h_{kn}]$ s.t. $h_{kn} \geq 0$.

$$\mathbf{V}_{(F\times N)} \approx \mathbf{W}_{(F\times K)} \times \mathbf{H}_{(K\times N)}$$



- $\mathbf{V}$: the $F \times N$ **data matrix**:
  - F features (rows),
  - N observations/examples/feature vectors (columns);
- $\mathbf{v}_n = (v_{1n}, \cdots, v_{Fn})^T$: the n-th **feature vector** observation among a collection of N observations $\mathbf{v}_1, \cdots, \mathbf{v}_N$;
- $\mathbf{v}_n$ is a column vector in $\mathbb{R}_+^F$; $\mathbf{v}_n$ is a row vector;

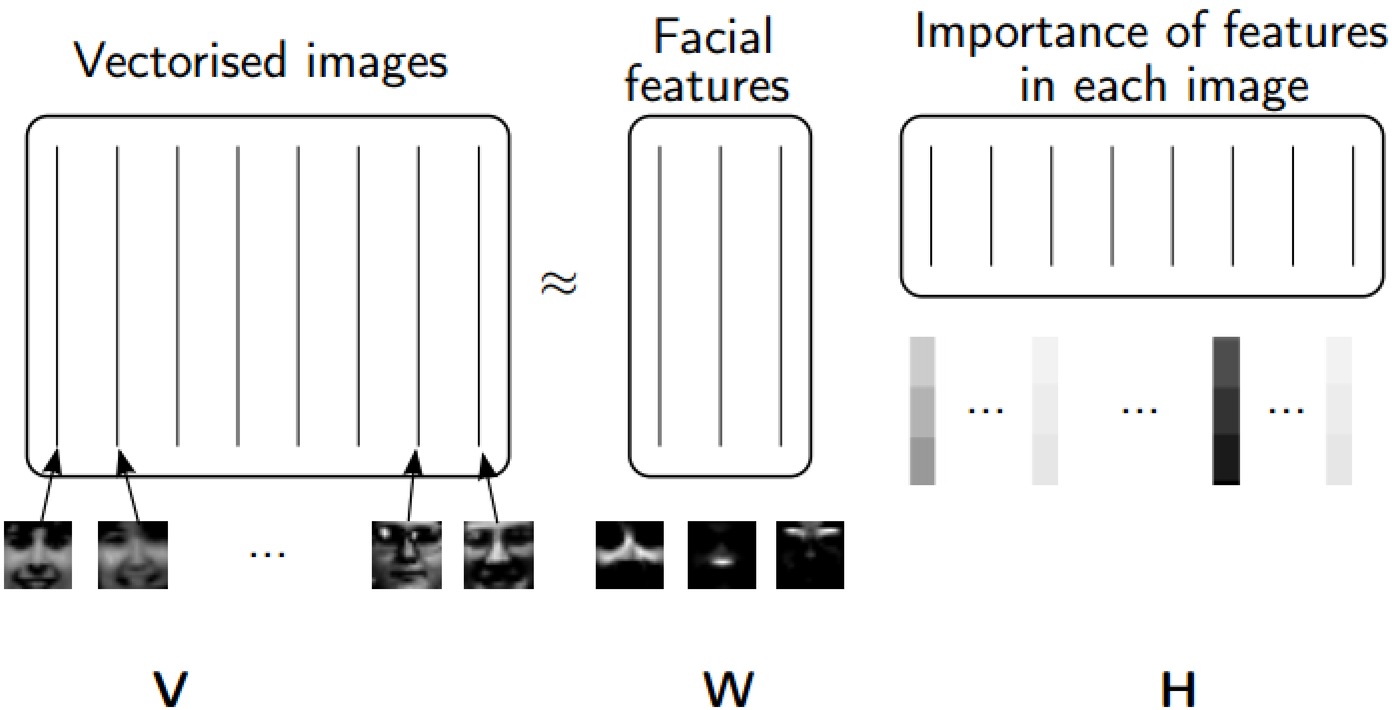
- $\mathbf{W}$: the $F \times K$ **dictionary matrix**:
  - w_{fk} is one of its coefficients,
  - $\mathbf{w}_k$ a dictionary/basis vector among K elements;
- $\mathbf{H}$: the $K \times N$ **activation/expansion matrix:**
  - $\mathbf{h}_n$: the **column vector** of activation coefficients for observation $\mathbf{v}_n$:

$$\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k \ ;$$

  - $\mathbf{h}_{k:}$: the **row vector** of activation coefficients relating to basis vector $\mathbf{w}_k$.

Usages of NMF

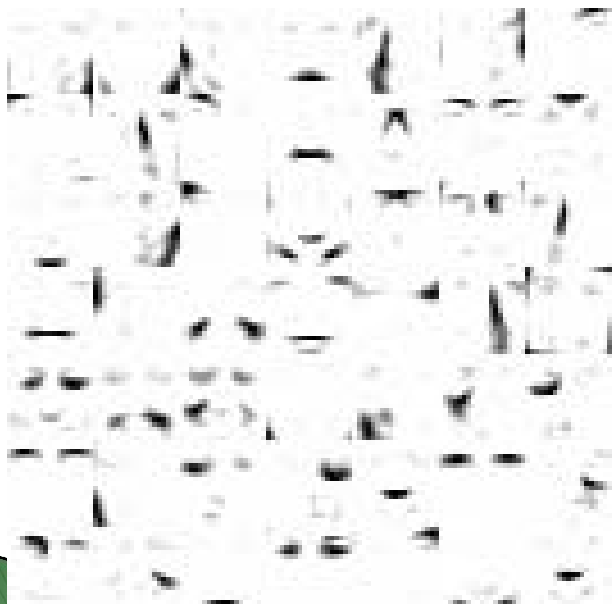
- **Learning feature**
  - learn NMF on training dataset $\mathbf{V}_{train} \rightarrow$ dictionary $\mathbf{W}$
  - exploit $\mathbf{W}$ to decompose new test examples $\mathbf{v}_n$: $\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$; $h_{kn} \geq 0$
  - use $\mathbf{h}_n$ as **feature vector** for example n.



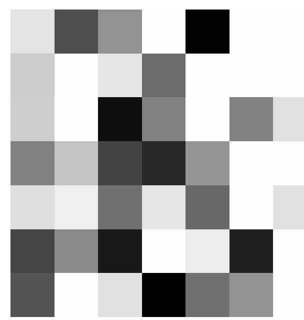
Usages of NMF

- **Learning feature**
  - learn NMF on training dataset $\mathbf{V}_{train} \rightarrow$ dictionary $\mathbf{W}$
  - exploit $\mathbf{W}$ to decompose new test examples $\mathbf{v}_n$:
    $\mathbf{v}_n \approx \sum_{k=1}^{K} h_{kn}\mathbf{w}_k$; $h_{kn} \geq 0$
  - use $\mathbf{h}_n$ as **feature vector** for example n.

Original Image



X

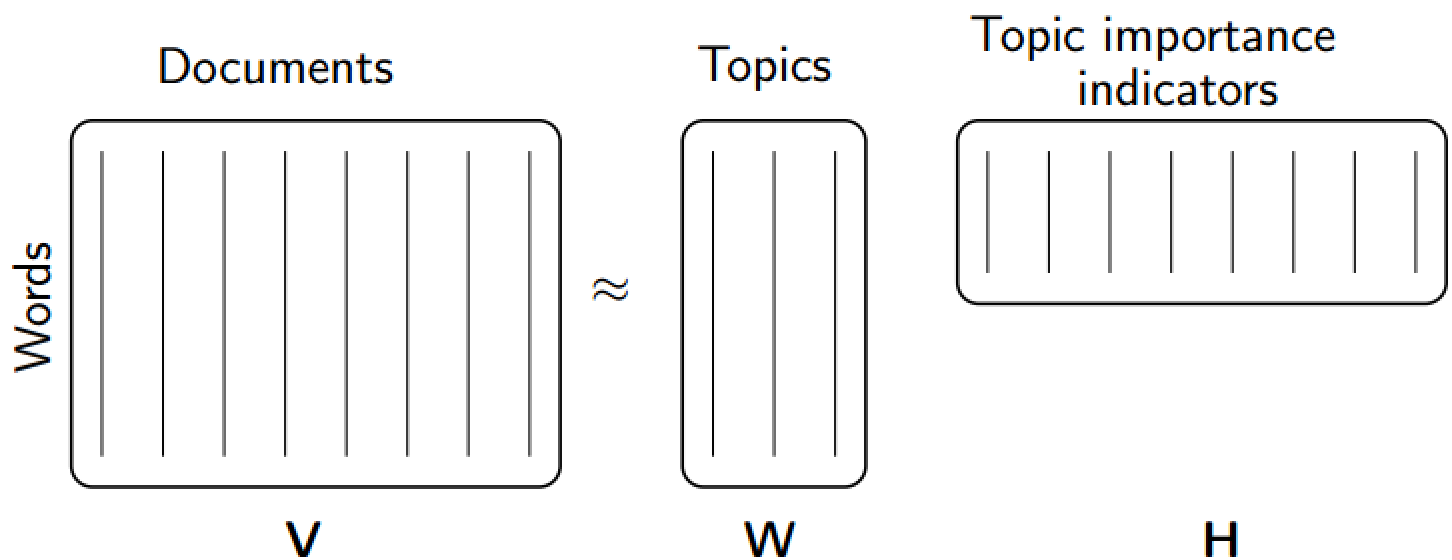


=

Usages of NMF

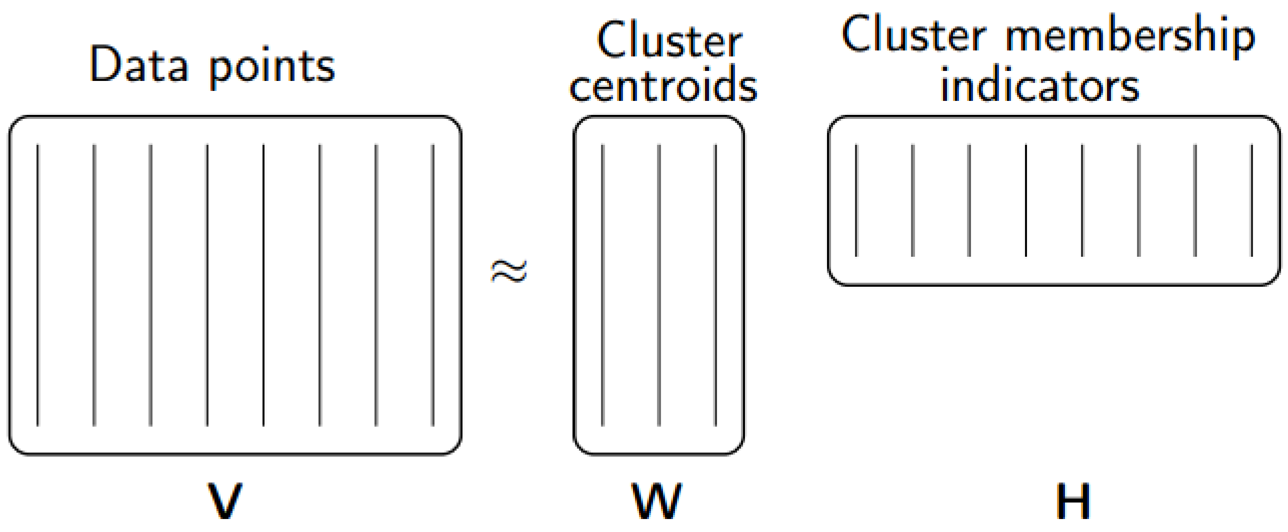
- **topics recovery**:

  assume $\mathbf{V} = [v_{fn}]$ is a (scaled) **term-document** co-occurrence matrix: v_{fn} is the frequency of occurrences of word m_f in document d_n;



Usages of NMF

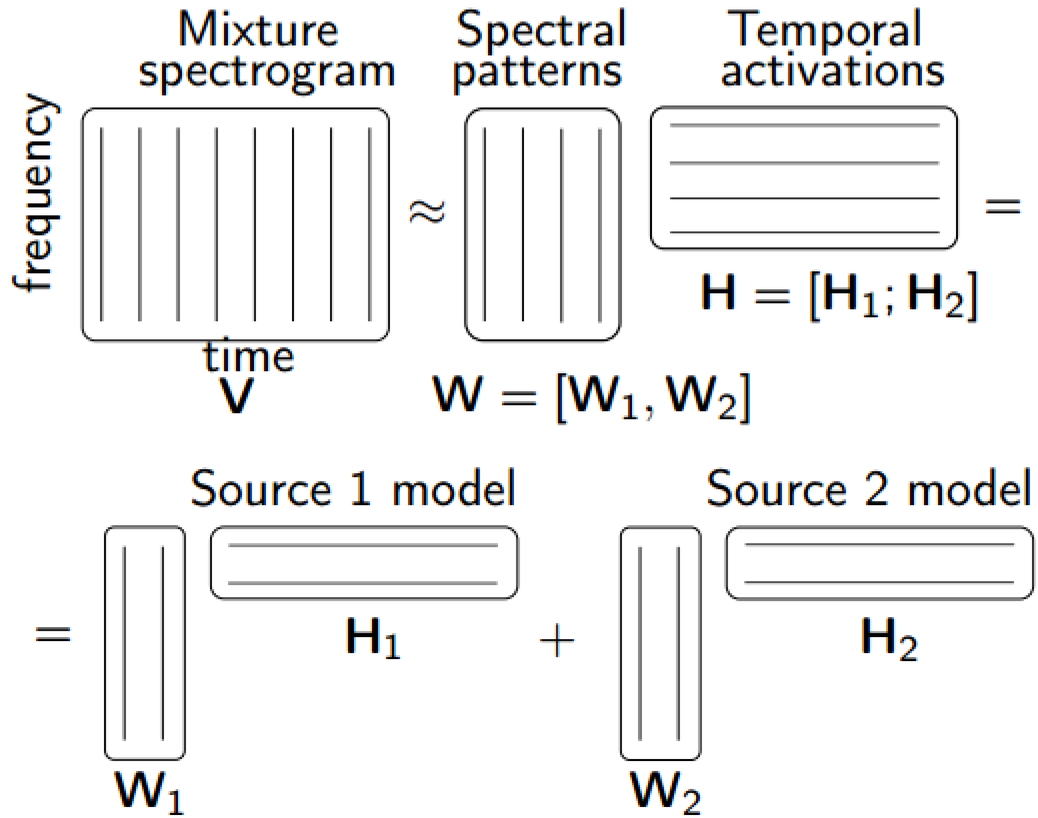
- **clustering**: like K-means (Ding et al., 2005, 2010; Xu et al., 2003):



- NMF can handle overlapping clusters and provides *soft* cluster membership indications.

Usages of NMF

- **filtering and source separation**: as with Independent Component Analysis (ICA):



Given data matrix V.

How to find W and H???

NMF optimization criteria

NMF approximation $\mathbf{V} \approx \mathbf{WH}$ is usually obtained through:

$$\min_{\mathbf{W},\mathbf{H} \geq 0} D(\mathbf{V}|\mathbf{WH}),$$

where $D(\mathbf{V}|\widehat{\mathbf{V}})$ is a *separable matrix divergence*:

$$D(\mathbf{V}|\widehat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}),$$

and $d(x|y)$ defined for all $x, y \geq 0$ is a *scalar divergence* such that:

- $d(x|y)$ is continuous over x and y;
- $d(x|y) \geq 0$ for all $x, y \geq 0$;
- $d(x|y) = 0$ if and only if $x = y$.

Popular (scalar) divergences

$$E(\mathbf{W},\mathbf{H}) = \|\mathbf{V} - \mathbf{WH}\|^2 = \sum_{i,j}(V_{ij} - (\mathbf{WH})_{ij})^2.$$

Euclidean (EUC) distance (Lee and Seung, 1999)

$$d_{EUC}(x,y) = (x-y)^2$$

Kullback-Leibler (KL) divergence (Lee and Seung, 1999)

$$d_{KL}(x,y) = x \log \frac{x}{y} - x + y$$

Itakura-Saito (IS) divergence (Févotte et al., 2009)

$$d_{IS}(x,y) = \frac{x}{y} - \log \frac{x}{y} - 1$$

$$\mathrm{D(V,W\ H)} = \sum_{i=1}^{n} \sum_{j=1}^{m} [V_{ij} \log \frac{V_{ij}}{(\mathrm{WH})_{ij}} - V_{ij} + (WH)_{ij}]$$

Optimization difficulties

An efficient solution of the NMF optimization problem

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})\,; \;\; C(\boldsymbol{\theta}) \stackrel{\text{def}}{=} D(\mathbf{V}|\mathbf{WH})$$

(where $\boldsymbol{\theta} \stackrel{\text{def}}{=} \{\mathbf{W},\mathbf{H}\}$ denotes the NMF parameters) must cope with the following difficulties:

- the **nonnegativity constraints** must be taken into account;
- **no uniqueness** of the solution is guaranteed in general;
- the optimization problem has usually a **multitude of local and global minima**.

$$\mathbf{WH} = \mathbf{WDD}^{-1}\mathbf{W}$$

Alternating optimization strategy

The problem is usually easier to optimize over one matrix (say **H**) given the other matrix (say **W**) is known and fixed.

Alternating optimization a.k.a block-coordinate descent (one iteration):

- update **W**, given **H** fixed,
- update **H**, given **W** fixed.

Gradient descent

$$h_{kn} \leftarrow h_{kn} - \mu_{kn} \nabla_{h_{kn}} C(\boldsymbol{\theta}) \, ,$$

Multiplicative update rules

A heuristic approach introduced by (Lee and Seung, 2001) to solve $\min_{\boldsymbol{\theta}} C(\boldsymbol{\theta})$

Multiplicative update (MU) rule for $\mathbf{H}$ (similarly for $\mathbf{W}$) is defined as:

$$h_{kn} \leftarrow h_{kn} \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_- / \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_+ ,$$

where

$$\nabla_{h_{kn}} C(\boldsymbol{\theta}) = \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_+ - \left[\nabla_{h_{kn}} C(\boldsymbol{\theta})\right]_- ,$$

and the summands are both nonnegative.