Part 3.
Nonnegative Matrix Factorization ⇔ K-means and Spectral Clustering

Nonnegative Matrix Factorization (NMF)

Data Matrix: n points in p-dim:

$$X = (x_1, x_2, \cdots, x_n)$$

x_i is an image, document, webpage, etc

Decomposition (low-rank approximation) $X \approx FG^T$

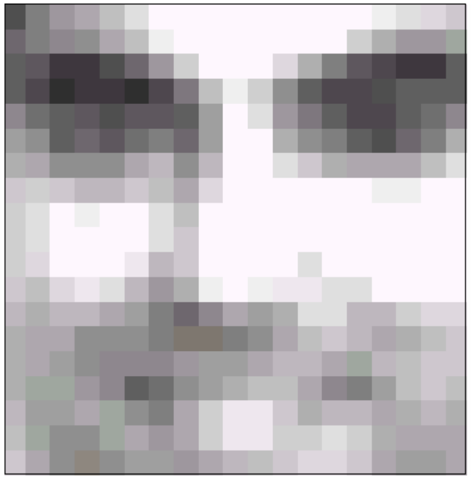
Nonnegative Matrices

$$X_{ij} \geq 0, F_{ij} \geq 0, G_{ij} \geq 0$$

$$F = (f_1, f_2, \cdots, f_k) \qquad G = (g_1, g_2, \cdots, g_k)$$

Some historical notes

- Earlier work by statistics people
- P. Paatero (1994) Environmetrices
- Lee and Seung (1999, 2000)
  - Parts of whole (no cancellation)
  - A multiplicative update algorithm

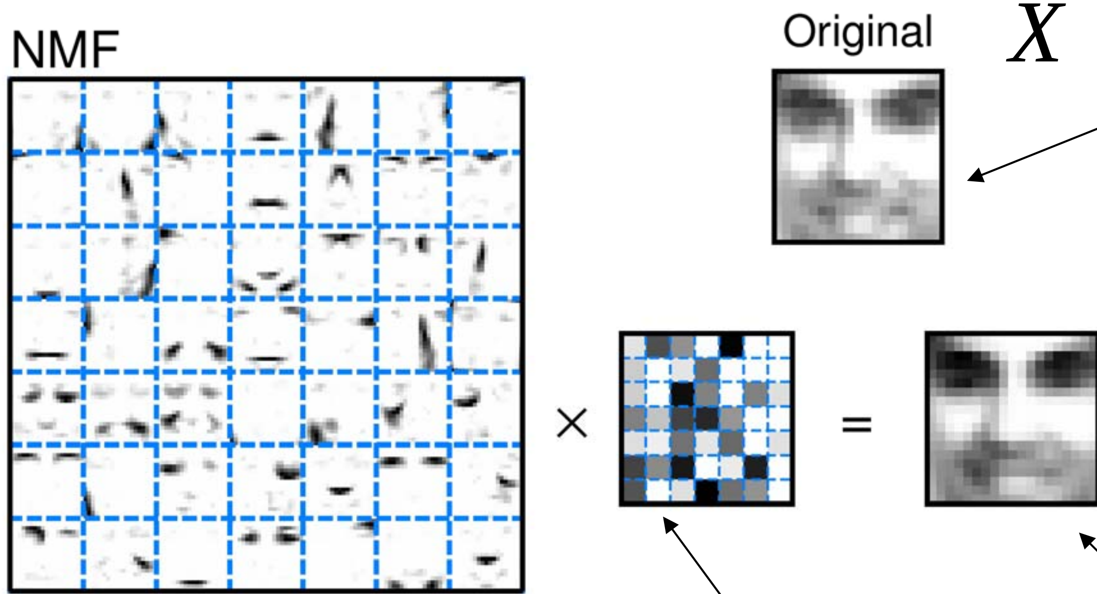


$$\longrightarrow \begin{bmatrix} 0.0 \\ 0.5 \\ 0.7 \\ 1.0 \\ \vdots \\ 0.8 \\ 0.2 \\ 0.0 \end{bmatrix}$$

Pixel vector

Parts-based perspective

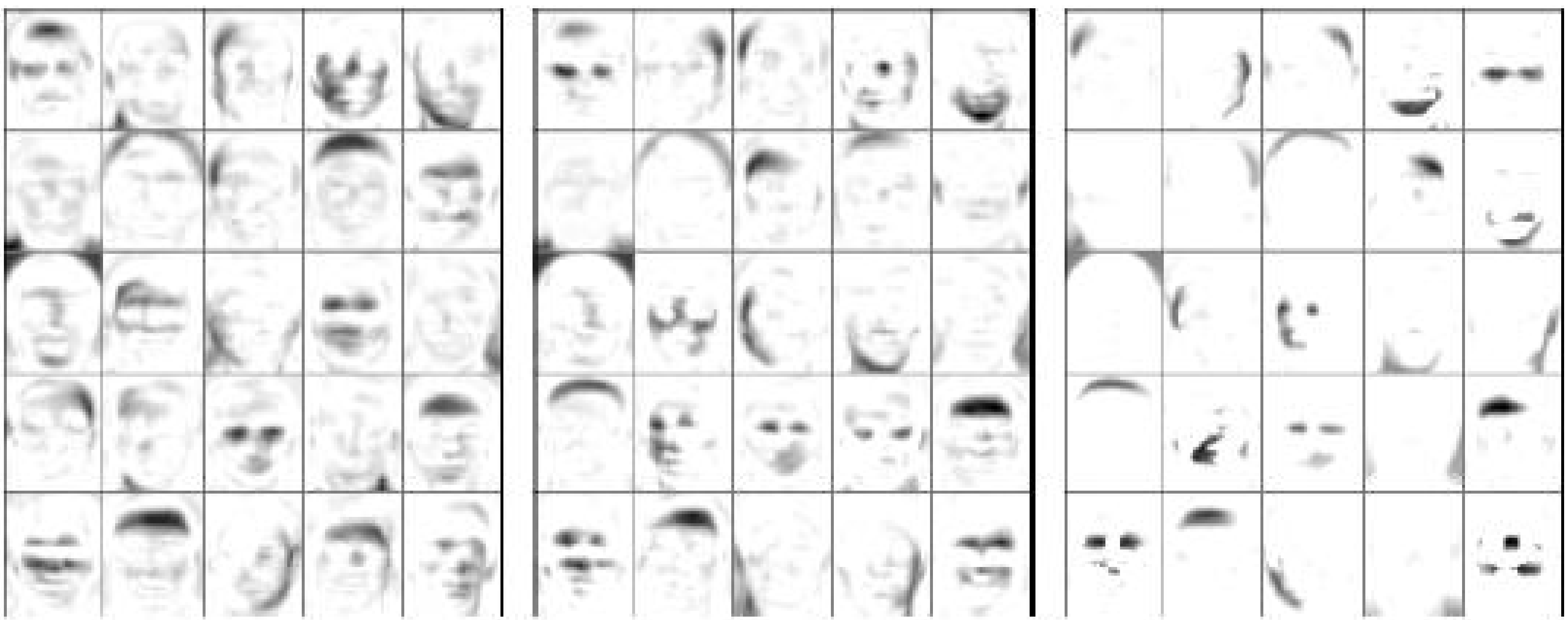
$$X = (x_1, x_2, \cdots, x_n)$$



$$F = (f_1, f_2, \cdots, f_k) \quad G = (g_1, g_2, \cdots, g_k)$$

$$X \approx FG^T$$

Sparsify F to get parts-based picture



$$X \approx FG^T \quad F = (f_1, f_2, \cdots, f_k)$$

Li, et al, 200; Hoyer 2003

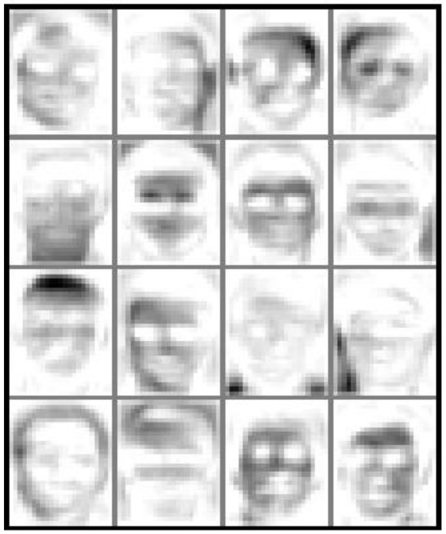
Theorem.
NMF = kernel *K*-means clustering

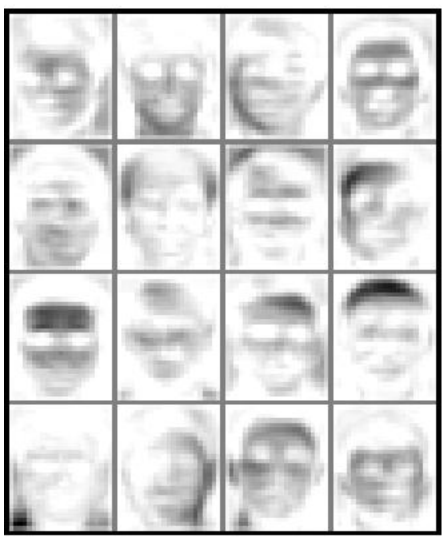
NMF produces holistic modeling of the data

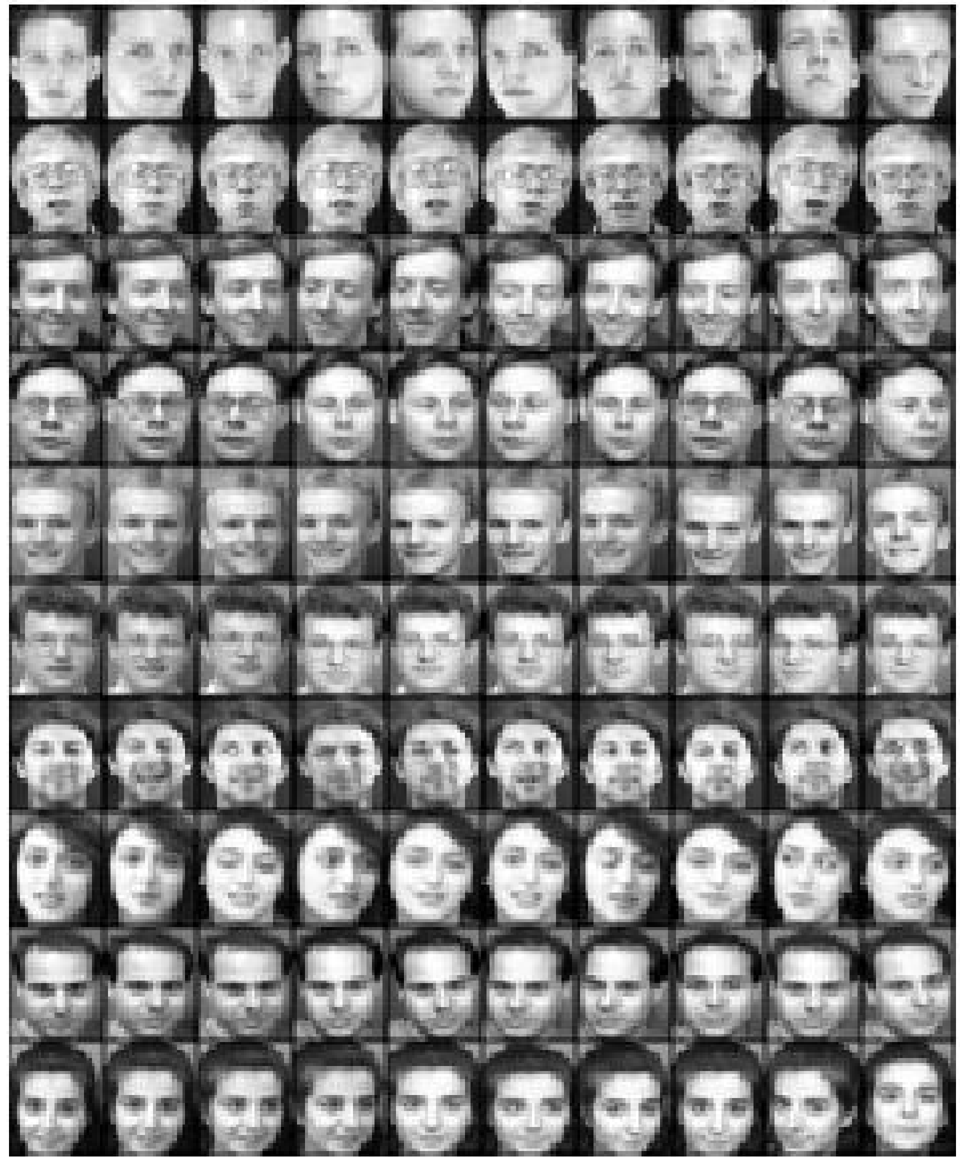
Theoretical results and experiments verification

(Ding, He, Simon, 2005)

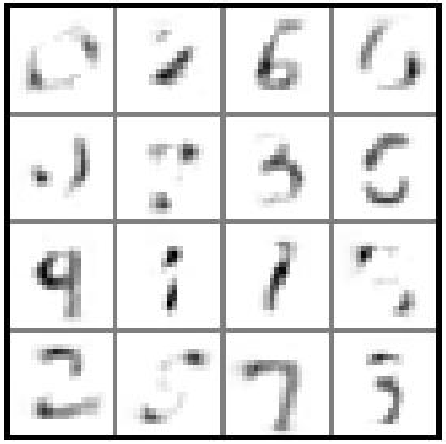
Our Results: NMF = Data Clustering

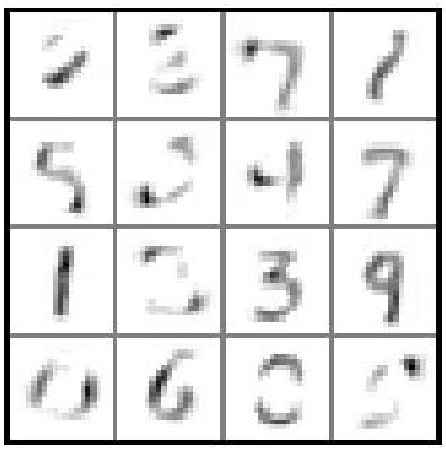


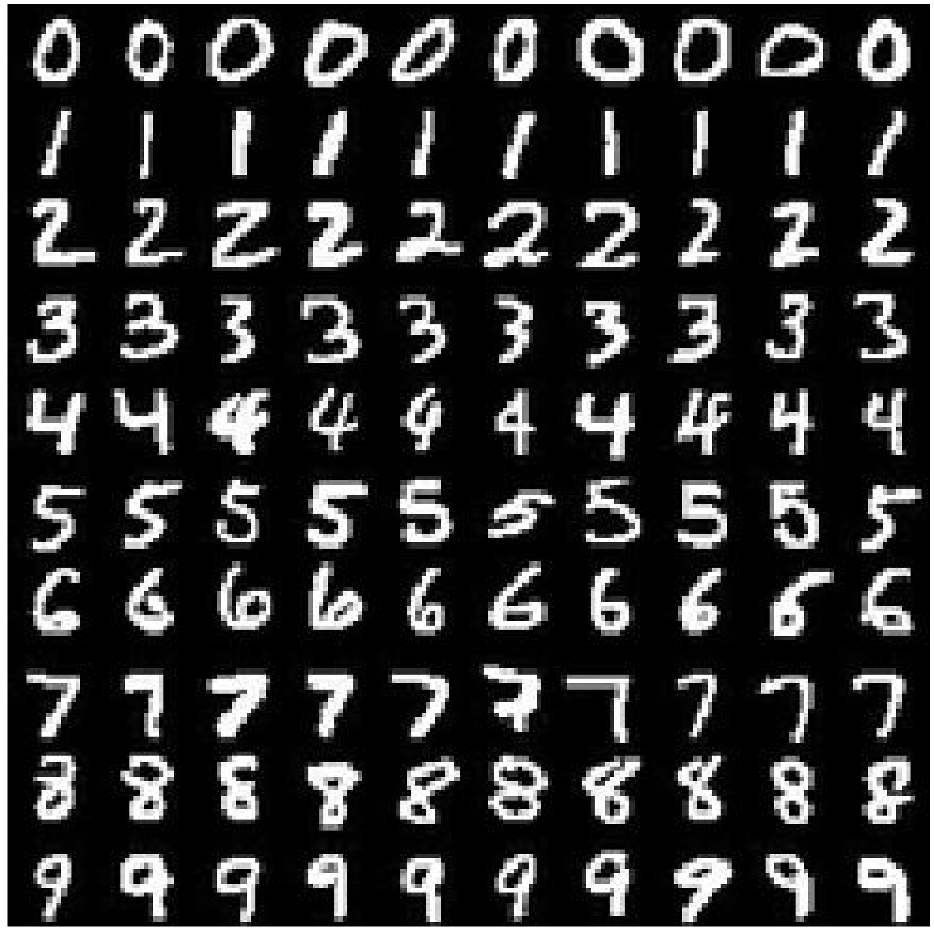




Our Results: NMF = Data Clustering







Theorem: K-means = NMF

- Reformulate K-means and Kernel K-means

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T W H)$$

- Show equivalence

$$\min_{H^T H = I, H \geq 0} \| W - H H^T \|^2$$

NMF = K-means

$$H = \arg\max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T W H)$$

$$= \arg\min_{H^T H = I, H \geq 0} [-2\mathrm{Tr}(H^T W H)]$$

$$= \arg\min_{H^T H = I, H \geq 0} [\|\mathrm{W}\|^2 \text{ -}2\mathrm{Tr}(H^T W H)] + \|H^T H\|^2$$

$$= \arg\min_{H^T H = I, H \geq 0} \|W - HH^T\|^2$$

$$\Rightarrow \arg\min_{H \geq 0} \|W - HH^T\|^2$$

Spectral Clustering = NMF

Normalized Cut: $$J_{\mathbf{Ncut}} = \sum_{<k,l>} \left(\frac{s(C_k, C_l)}{d_k} + \frac{s(C_k, C_l)}{d_l} \right) = \sum_k \frac{s(C_k, G - C_k)}{d_k}$$

$$= \frac{h_1^T (D-W) h_1}{h_1^T D h_1} + \cdots + \frac{h_k^T (D-W) h_k}{h_k^T D h_k}$$

Unsigned cluster indicators: $y_k = D^{1/2} (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / \| D^{1/2} h_k \|$

Re-write:

$$J_{\mathbf{Ncut}}(y_1, \cdots, y_k) = y_1^T (I - \tilde{W}) y_1 + \cdots + y_k^T (I - \tilde{W}) y_k$$

$$= \mathbf{Tr}(Y^T (I - \tilde{W}) Y) \qquad \tilde{W} = D^{-1/2} W D^{-1/2}$$

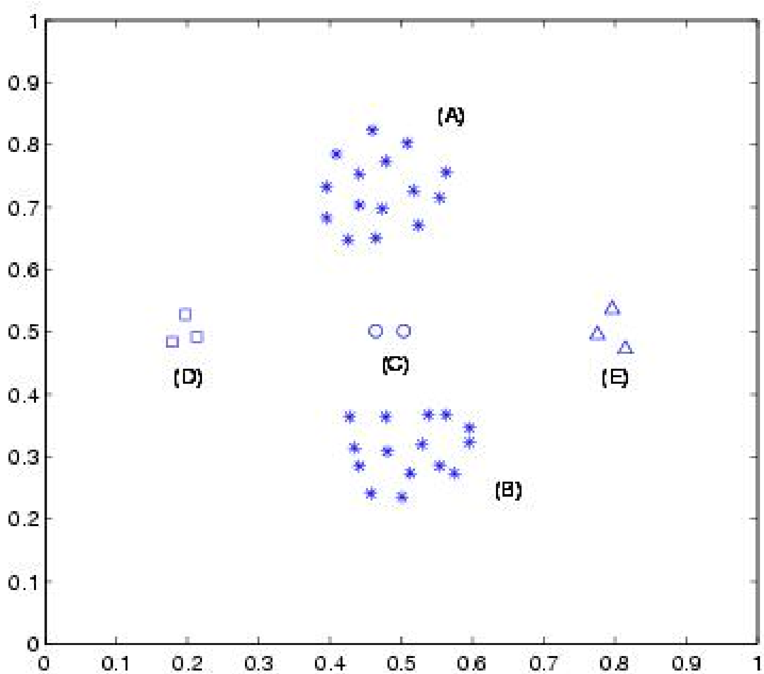
Optimize: $\max_Y \mathbf{Tr}(Y^T \tilde{W} Y)$, **subject to** $Y^T Y = I$

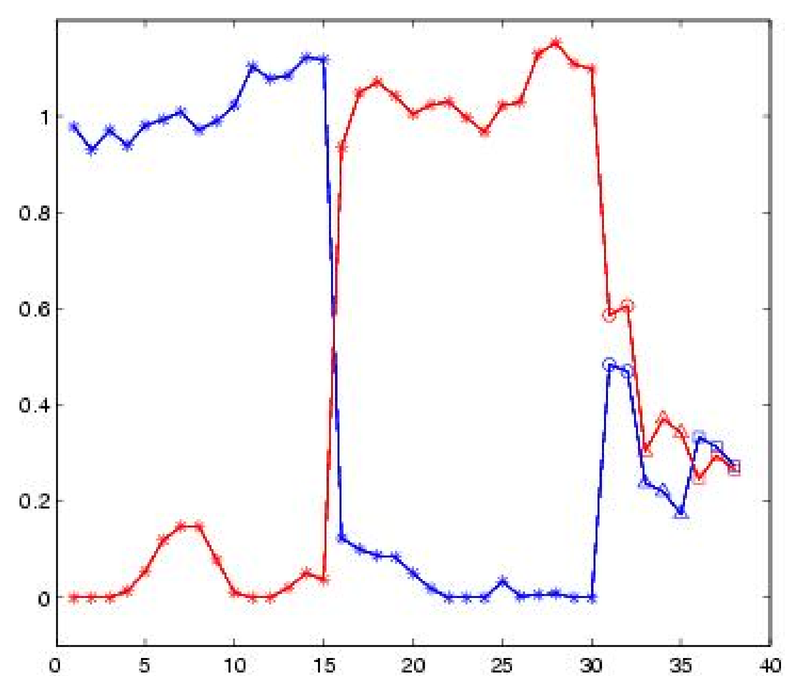
Normalized Cut $\Longrightarrow \min_{H^T H = I, H \ge 0} \| \tilde{W} - H H^T \|^2$

(Gu , et al, 2001)

Advantages of NMF over standard K-means

Soft clustering



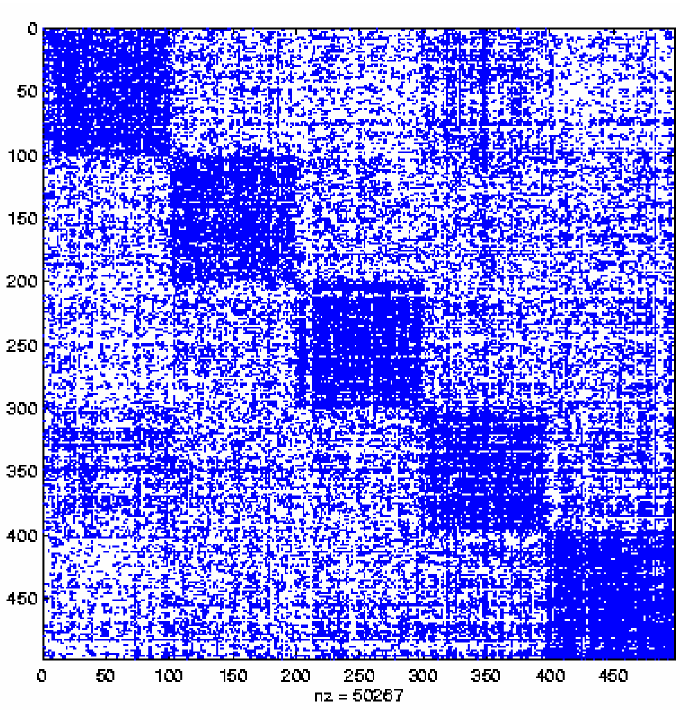


Experiments on Internet Newsgroups

NG2: comp.graphics
NG9: rec.motorcycles
NG10: rec.sport.baseball
NG15: sci.space
NG18: talk.politics.mideast

100 articles from each group.
1000 words
Tf.idf weight. Cosine similarity

cosine similarity



Accuracy of clustering results

K-means	W=HH’
0.531	0.612
0.491	0.590
0.576	0.608
0.632	0.652
0.697	0.711

Summary for Symmetric NMF

- K-means , Kernel K-means
- Spectral clustering

$$\max_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T W H)$$

- Equivalence to

$$\min_{H^T H = I, H \geq 0} \| W - HH^T \|^2$$

Nonsymmetric NMF

- K-means , Kernel K-means
- Spectral clustering

$$\max_{H^T H=I, H\geq 0} \mathrm{Tr}(H^T W H)$$

- Equivalence to

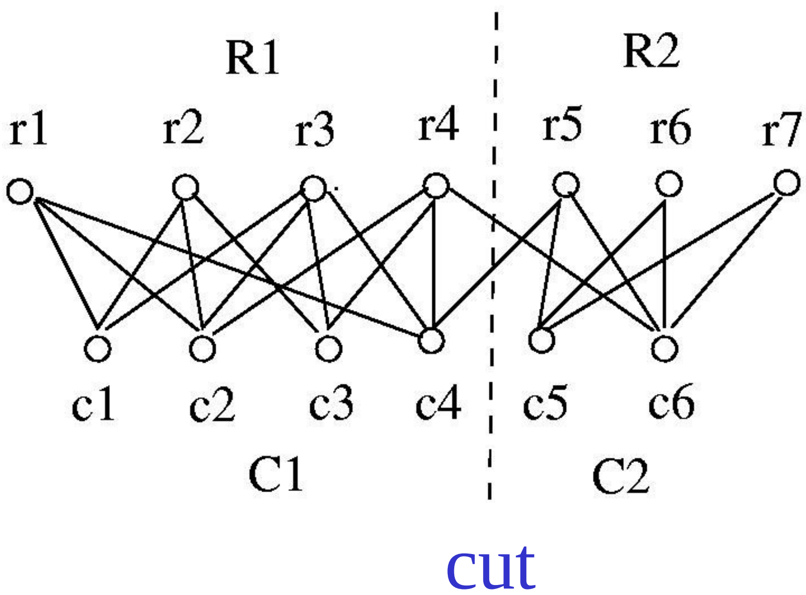
$$\min_{H^T H=I, H\geq 0} \| W - HH^T \|^2$$

Non-symmetric NMF
Rectangular Data Matrix
Bipartite Graph

- Information Retrieval: word-to-document
- DNA gene expressions
- Image pixels
- Supermarket transaction data

K-means Clustering of Bipartite Graphs

Simultaneous clustering of rows and columns



row indicators $\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = (f_1, f_2, f_3) = F$

column indicators $\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = (g_1, g_2, g_3) = G$

$$J_{Kmeans} = \sum_{j=1}^{k} \frac{s(B_{R_j, C_j})}{\sqrt{|R_j||C_j|}} = \mathrm{Tr}(F^T B G)$$

NMF = K-means clustering

$$H = \underset{\substack{F^T F = I, F \geq 0 \\ G^T G = I, G \geq 0}}{\arg\max} \operatorname{Tr}(F^T B G)$$

$$= \underset{\substack{F^T F = I, F \geq 0 \\ G^T G = I, G \geq 0}}{\arg\min} \operatorname{Tr}(-2 F^T B G)$$

$$= \underset{\substack{F^T F = I, F \geq 0 \\ G^T G = I, G \geq 0}}{\arg\min} \operatorname{Tr}(\| B \|^2 - 2 F^T B G + F^T F G^T G)$$

$$= \underset{\substack{F^T F = I, F \geq 0 \\ G^T G = I, G \geq 0}}{\arg\min} \| B - F G^T \|^2$$

Solving NMF with non-negative least square

$$J = \| X - FG^T \|^2, F \geq 0, G \geq 0$$

Fix F, solve for G; Fix G, solve for F

$$J = \sum_{i=1}^{n} \| x_i - F\tilde{g}_i \|^2, G = \begin{pmatrix} \tilde{g}_1 \\ \vdots \\ \tilde{g}_n \end{pmatrix}$$

Iterate, converge to a local minima

Solving NMF with multiplicative updating

$$J = \| X - FG^T \|^2, F \geq 0, G \geq 0$$

Fix F, solve for G; Fix G, solve for F

Lee & Seung (2000) propose

$$F_{ik} \leftarrow F_{ik} \frac{(XG)_{ik}}{(FG^T G)_{ik}} \qquad G_{jk} \leftarrow G_{jk} \frac{(X^T F)_{jk}}{(GF^T F)_{jk}}$$

Symmetric NMF

$$J = \| W - HH^T \|^2, H \geq 0$$

Constraint Optimization. KKT 1st condition

Complementarity slackness condition

$$0 = \left(\frac{\partial J}{\partial H} \right)_{ik} H_{ik} = (-4WH + 4HH^T H)_{ik} H_{ik}$$

Gradient decent

$$H_{ik} \leftarrow H_{ik} - \varepsilon_{ik} \frac{\partial J}{\partial H_{ik}} \qquad \varepsilon_{ik} = \frac{H_{ik}}{4(HH^T H)_{ik}}$$

$$H_{ik} \leftarrow H_{ik}(1 - \beta + \beta \frac{(WH)_{ik}}{(HH^T H)_{ik}})$$

Summary

- NMF is a new low-rank approximation
- The holistic picture (vs. parts-based)
- NMF is equivalent to spectral clustering
- Main advantage: soft clustering