(Semi-)Nonnegative Matrix Factorization and K-mean Clustering

Chris Ding
Lawrence Berkeley National Laboratory

Xiaofeng He	Lawrence Berkeley Nat'l Lab
Horst Simon	Lawrence Berkeley Nat'l Lab
Tao Li	Florida Int'l Univ.
Michael Jordan	UC Berkeley
Haesun Park	Georgia Tech

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)

$$X \approx FG^T$$

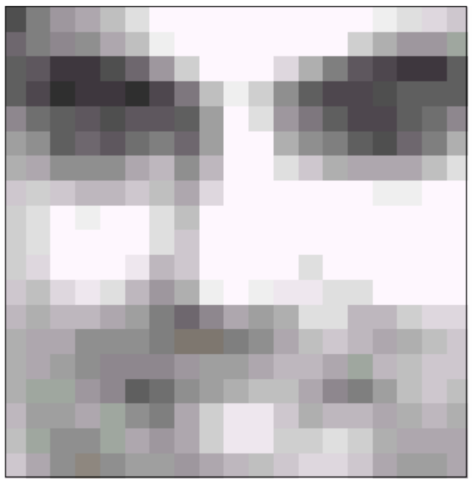
Nonnegative Matrices

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

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

Some historical notes

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

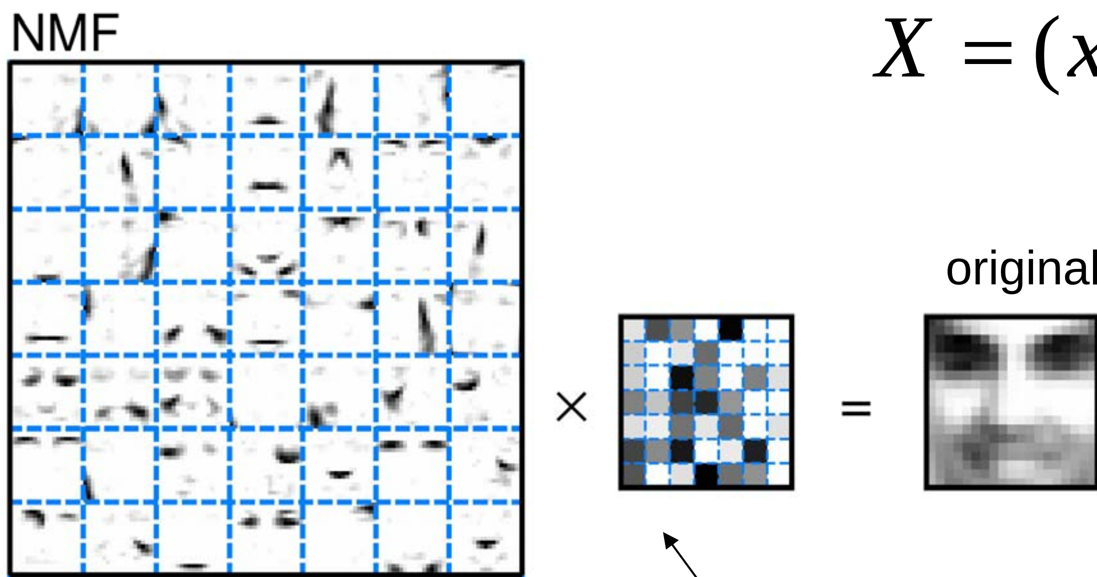


$$\longrightarrow \begin{bmatrix} 0.0 \\ 0.5 \\ 0.7 \\ 1.0 \\ \vdots \\ 0.8 \\ 0.2 \\ 0.0 \end{bmatrix}$$

Pixel vector

Lee and Seung (1999): Parts-based Perspective

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



$$FG^T \approx X$$

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

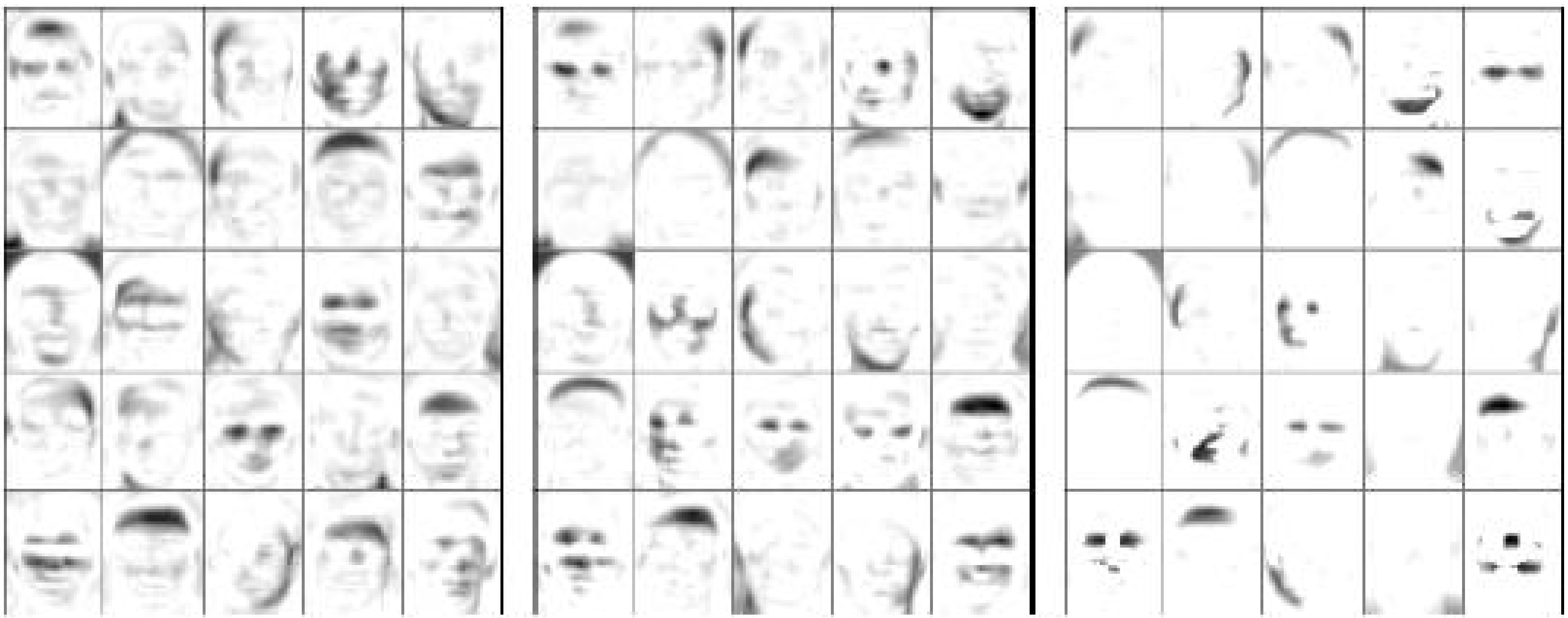
"Parts of Whole" Picture

(Li, et al, 2001; Hoyer 2003)

Straightforward NMF doesn't get parts-based picture

Several People explicitly sparsify F to get parts-based picture

Donono & Stodden (2003) study condition for parts-of-whole



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

Meanwhile …….

A number of studies empirically show the usefulness of NMF for pattern discovery/clustering:

Xu et al (SIGIR'03)

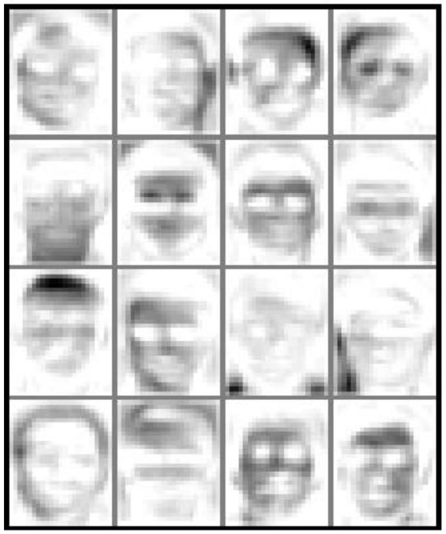
Brunet et al (PNAS'04)

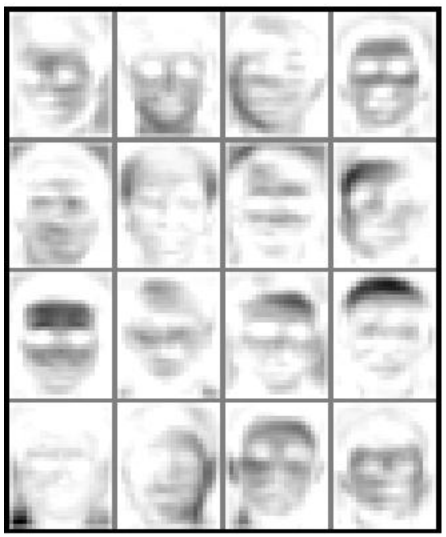
Many others

We claim:

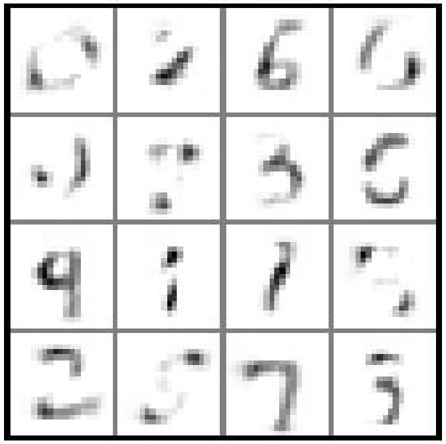
NMF factors give holistic pictures of the data

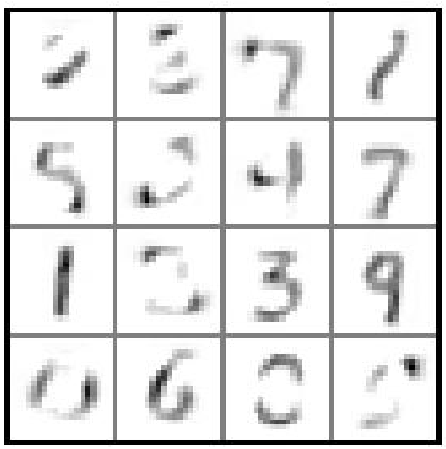
Our Experiments: NMF gives holistic pictures

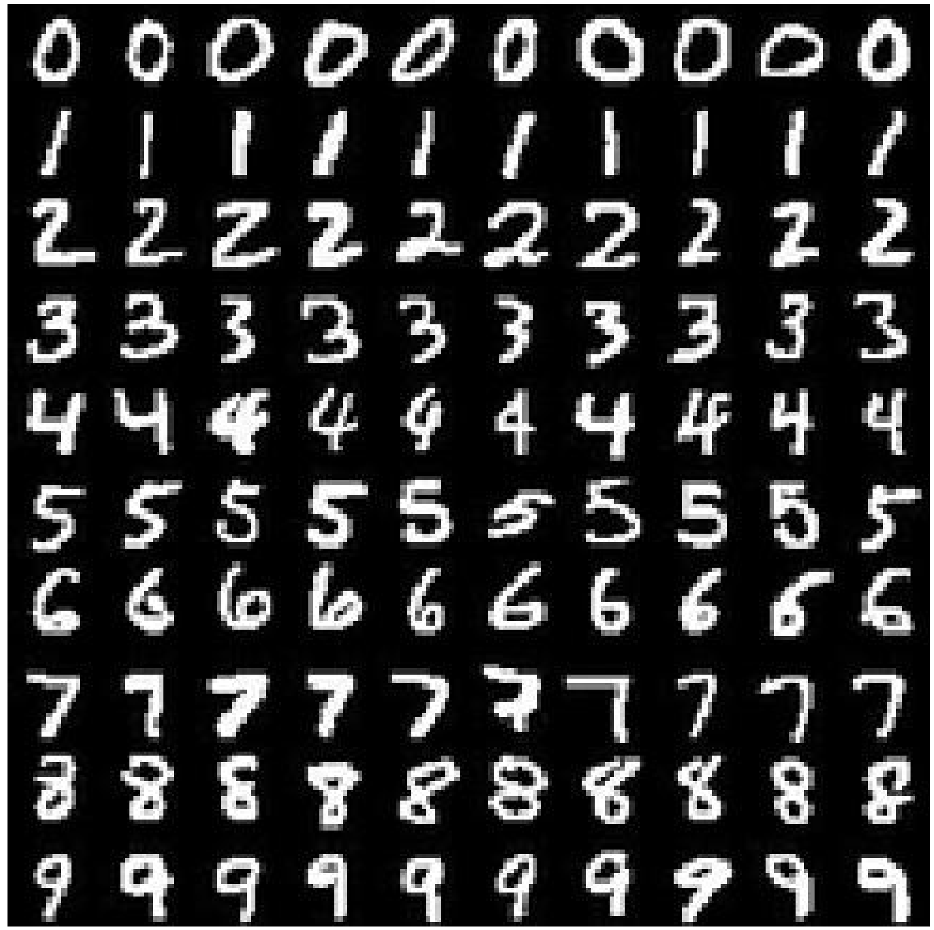




Our Experiments: NMF gives holistic pictures







Task:

Prove NMF is doing “Data Clustering”

NMF => K-means Clustering

NMF-Kmeans Theorem

G -orthogonal NMF is equivalent to relaxed K-means clustering.

Proof.

$$\min_{\substack{F \geq 0 \\ G^T G = I, G \geq 0}} \| X - FG^T \|^2$$

$$\min_{G^T G = I, G \geq 0} \mathrm{Tr}(X^T X - G^T X^T X G)$$

(Ding, He, Simon, SDM 2005)

K-means clustering

- Computationally Efficient (order-mN)
- Most widely used in practice
  - Benchmark to evaluate other algorithms

Given n points in m-dim: $X = (x_1, x_2, \cdots, x_n)^T$

K-means objective $\min J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2$

- Also called "isodata", "vector quantization"
- Developed in 1960's (Lloyd, MacQueen, Hartigan, etc)

Reformulate K-means Clustering

$$J_K = \sum_i \|x_i\|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} x_i^T x_j$$

Cluster membership indicators: $H = (h_1, \cdots, h_K)$

$$h_k = (0 \cdots 0, \overbrace{1 \cdots 1}^{n_k}, 0 \cdots 0)^T / n_k^{1/2}$$

$$J_K = \sum_i x_i^2 - \sum_{k=1}^{K} h_k^T X^T X h_k$$

Solving K-mean => $\max\limits_{H^T H = I, H \geq 0} \mathrm{Tr}(H^T X^T X H)$

(Zha, Ding, Gu, He, Simon, NIPS 2001)
(Ding & He, ICML 2004)

Reformulate K-means Clustering

Cluster membership indicators :

$$
\begin{array}{ccc}
C_1 & C_2 & C_3
\end{array}
$$

$$
\begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix} = (h_1, h_2, h_3) = H
$$

NMF-Kmeans Theorem

G -orthogonal NMF is equivalent to relaxed K-means clustering.

Proof.

$$\min_{\substack{F\geq 0 \\ G^T G=I, G\geq 0}} \| \mathrm{X} - FG^T \|^2$$

$$\min_{G^T G=I, G\geq 0} \mathrm{Tr}(X^T X - G^T X^T X G)$$

(Ding, He, Simon, SDM 2005)

Kernel *K*-means Clustering

Map feature vector to higher-dim space

$$x_i \rightarrow \phi(x_i)$$

Kernel ***K***-means objective:

$$\min J_K^\phi = \sum_{k=1}^{K} \sum_{i \in C_k} || \phi(x_i) - \phi(c_k) ||^2 \qquad \phi(c_k) \equiv \frac{1}{n_k} \sum_{i \in C_k} \phi(x_i)$$

$$J_K^\phi = \sum_i |\phi(x_i)|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \phi(x_i)^T \phi(x_j)$$

Kernal ***K***-means optimization:

$$\max J_K^\phi = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \left\langle \phi(x_i), \phi(x_j) \right\rangle = \mathrm{Tr}(H^T W H)$$

Matrix of pairwise similarities

Symmetric NMF: $W \approx HH^T$

Symmetric Nonnegative matrix

Orthogonal symmetric NMF is equivalent to Kernel K-means clustering.

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

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

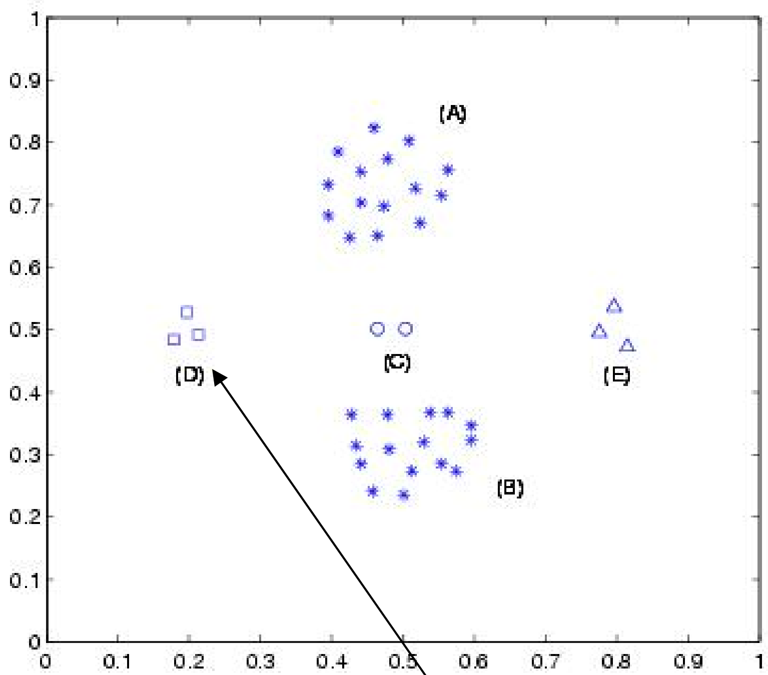
Orthogonality in NMF

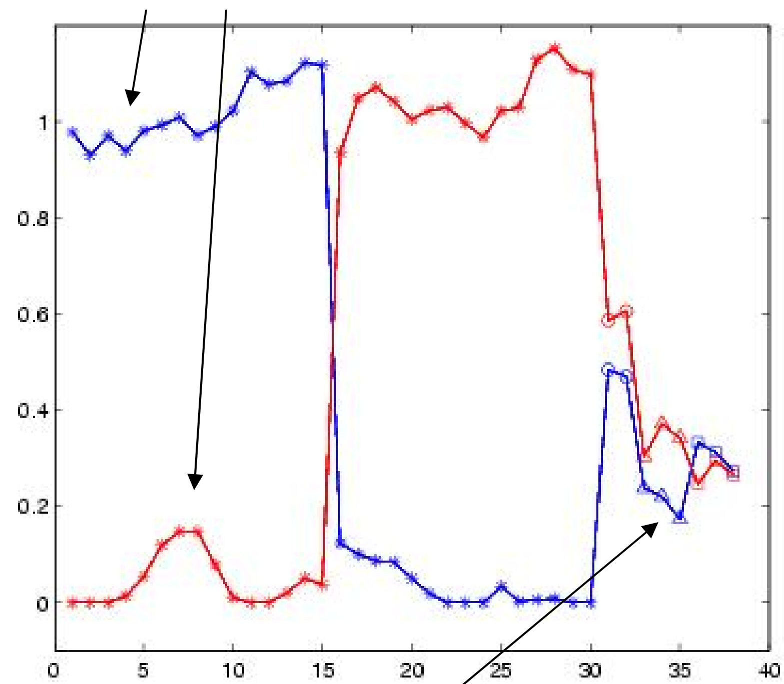
Strict orthogonal G: hard clustering

Non-orthogonal G: soft clustering

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

$H = (h_1, h_2)$





Ambiguous/outlier points

K-means Clustering Theorem

G -orthogonal NMF is equivalent to relaxed K-means clustering.

$$\min_{G^T G = I, G \geq 0} \| X_{\pm} - F_{\pm} G_{+}^{T} \|^2$$

Proof requires only G-orthogonality and nonnegativity

$$F = (f_1, f_2, \cdots, f_k) \quad \Rightarrow \quad \text{cluster centroids}$$

$$G = (g_1, g_2, \cdots, g_k) \quad \Rightarrow \quad \text{cluster indicators}$$

(Ding, Li, Jordan, 2006)

NMF Generalizations

SVD: $X_{\pm} = F_{\pm} G_{\pm}^T = U \Sigma V^T$

Semi-NMF: $X_{\pm} = F_{\pm} G_{+}^T$ (Ding, Li, Jordan, 2006)

Convex-NMF: $X_{\pm} = X_{\pm} W_{+} G_{+}^T$

Kernel-NMF: $\phi(X_{\pm}) = \phi(X_{\pm}) W_{+} G_{+}^T$

Tri-NMF: $X_{\pm} = F_{+} S_{\pm} G_{+}^T$

(Ding, Li, Peng, Park, KDD 2006)

Semi-NMF: $X_{\pm} = F_{\pm} G_{+}^{T}$

- For any mixed-sign input data (centered data)
- Clustrering and Low-rank approximation

$$\min \| X - FG^T \|$$

Update F: $$F = XG(G^TG)^{-1}$$

Update G: $$G_{ik} \leftarrow G_{ik} \sqrt{\frac{(F^TX)_{ik}^{+} + [G(FF)^{-}]_{ik}}{(F^TX)_{ik}^{-} + [G(FF)^{+}]_{ik}}}$$

(Ding, Li, Jordan, 2006)

(Ding, Li, Jordan, 2006)

Convex-NMF

In NMF $X_+ = F_+ G_+^T$ $F = (f_1, f_2, \cdots, f_k)$

In Semi-NMF $X_\pm = F_\pm G_+^T$ is in a large space

For f_k factor to capture the notion of cluster centroid,
Require f_k to be a convex combination of input data

$$f_k = w_{1k} x_1 + \cdots + w_{1n} x_n, \; F = XW_+$$

For F interpretability, (Affine combination $F = XW_\pm$)

$$X_\pm = X_\pm W_+ G_+^T$$

Convex-NMF: $X_{\pm} = X_{\pm} W_{+} G_{+}^{T}$

Computing algorithm

$$\min \| X - XWG^T \|$$

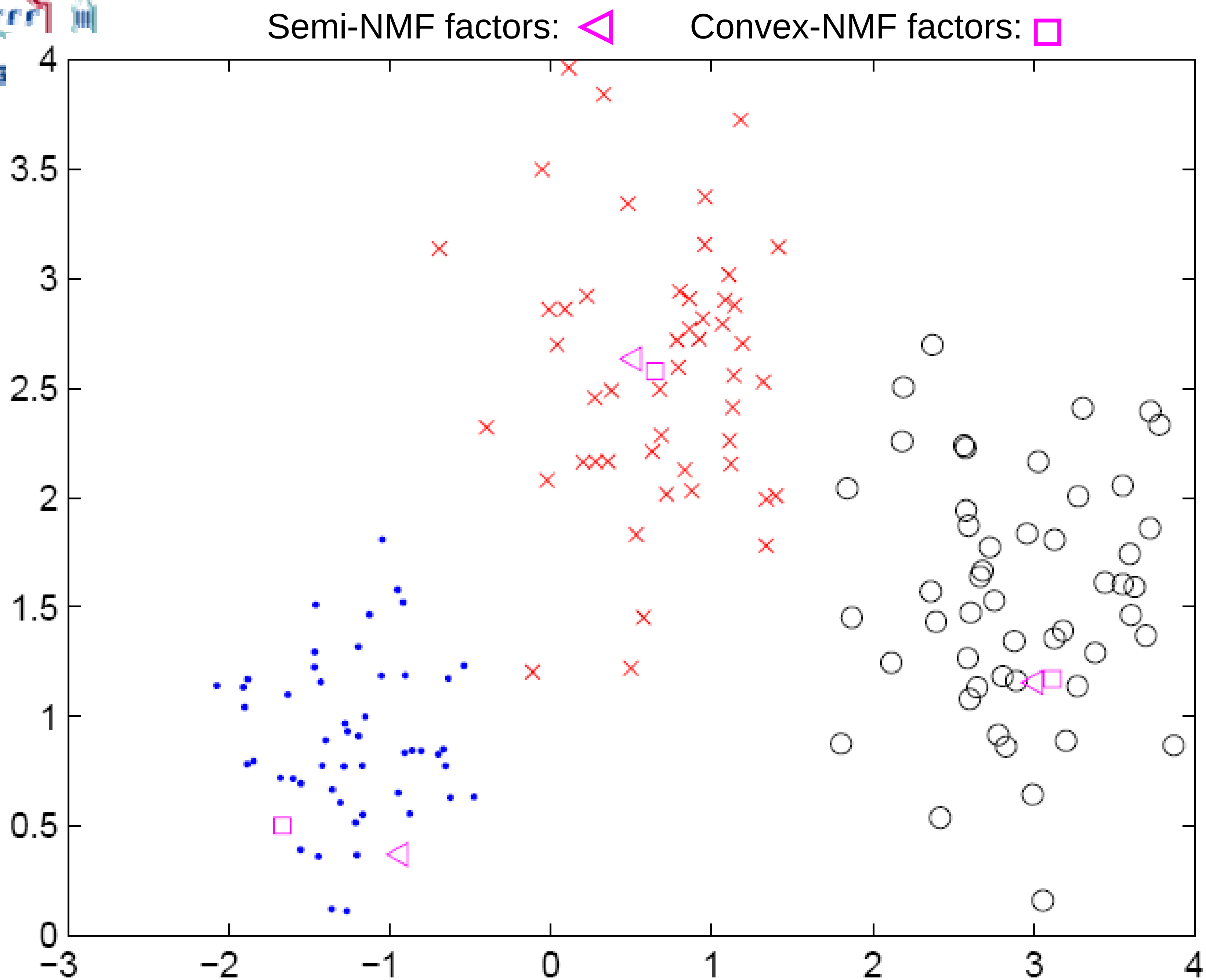
Update F:

$$W_{ik} \leftarrow W_{ik} \sqrt{\frac{[(X^TX)^+G]_{ik} + [(X^TX)^-WG^TG]_{ik}}{[(X^TX)^-G]_{ik} + [(X^TX)^+WG^TG]_{ik}}}$$

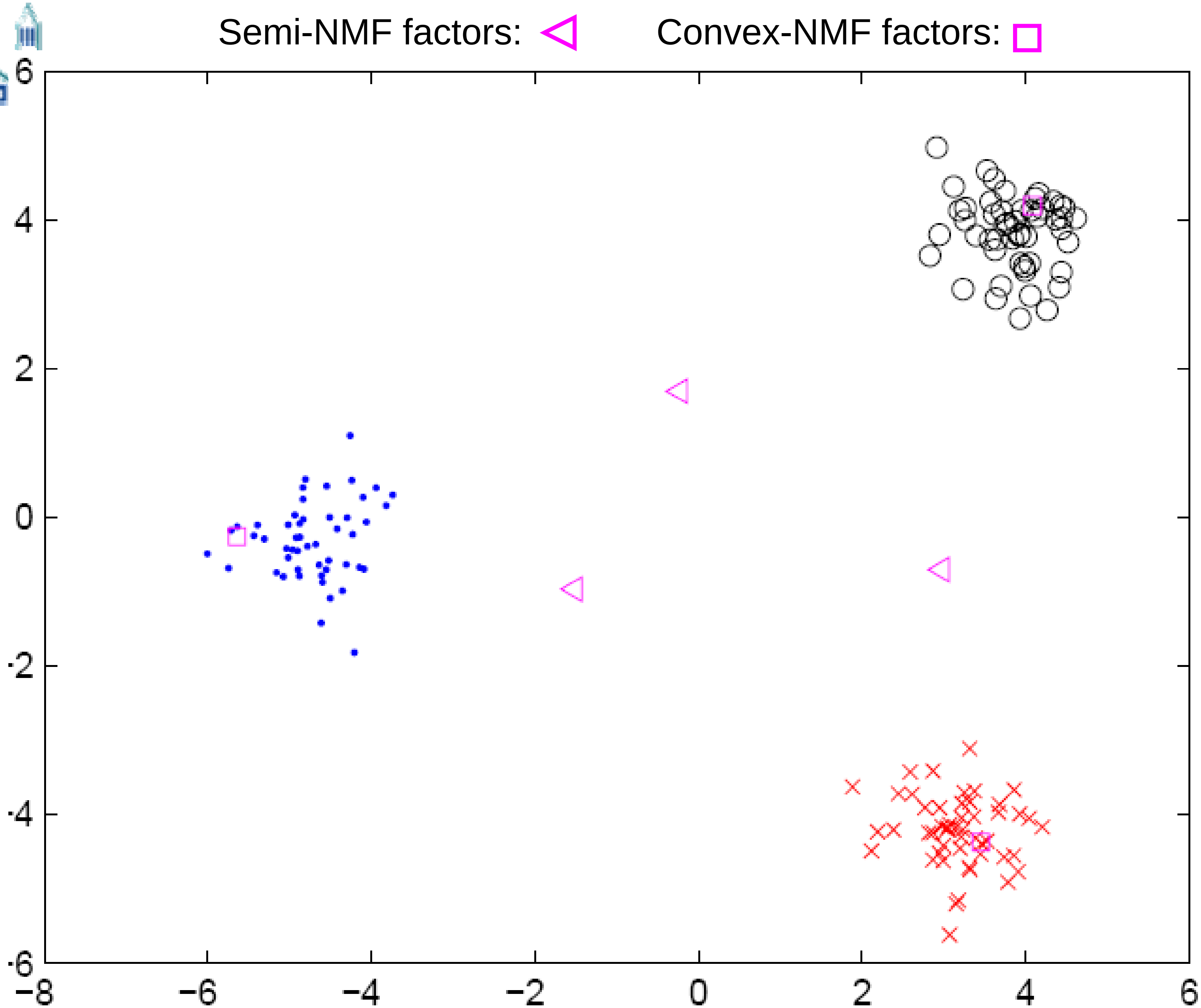
Update G:

$$G_{ik} \leftarrow G_{ik} \sqrt{\frac{[W^T(X^TX)^+]_{ik} + [GW^T(X^TX)^-W]_{ik}}{[W^T(X^TX)^-]_{ik} + [GW^T(X^TX)^+W]_{ik}}}$$

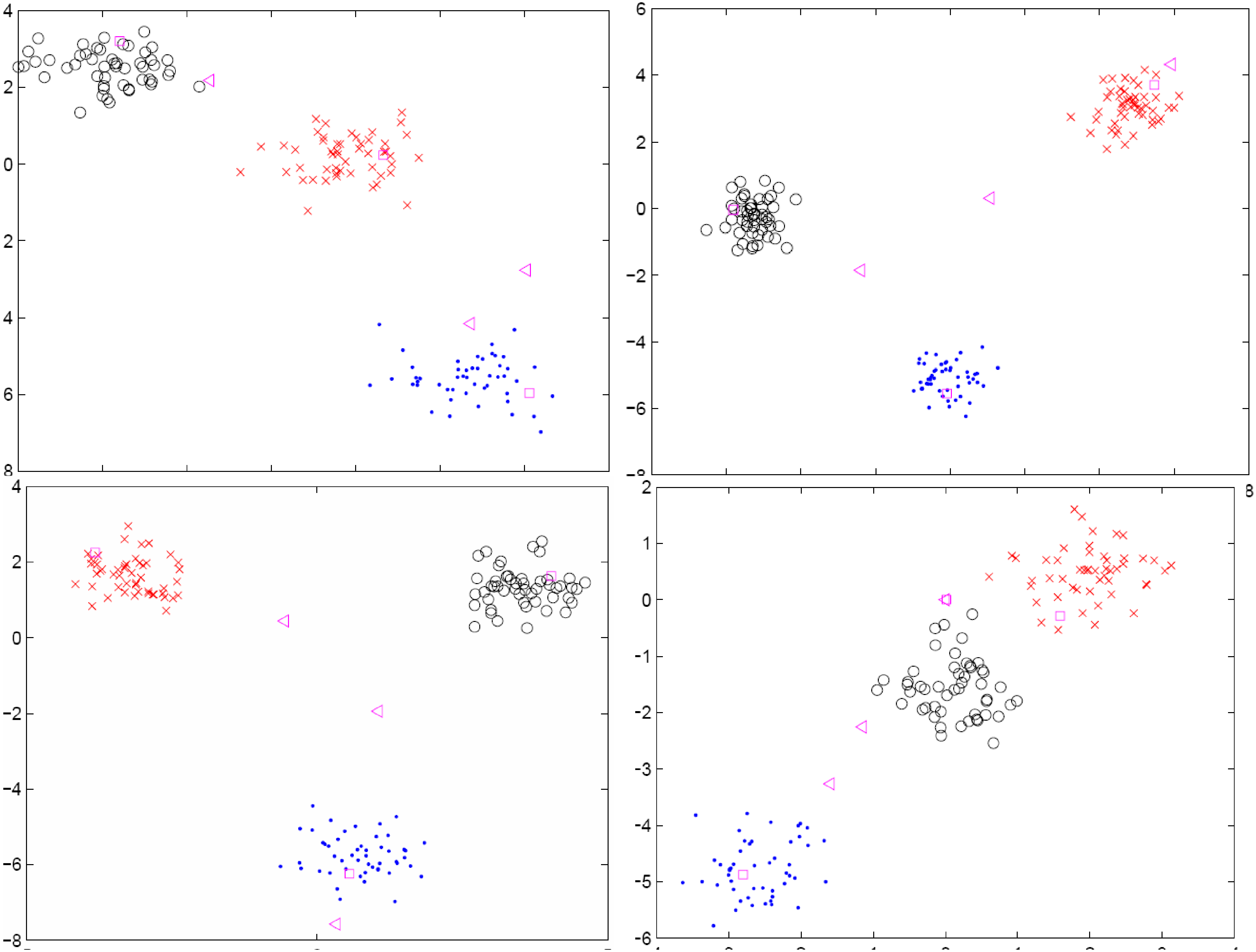
(Ding, Li, Jordan, 2006)



Semi-NMF factors:
Convex-NMF factors:
4
3.5
3
2.5
2
1.5
1
0.5
0
−3
−2
−1
0
1
2
3
4



Semi-NMF factors:
Convex-NMF factors:
6
4
2
0
-2
-4
-6
−8
−6
−4
−2
0
2
4
6



4
2
0
2
4
6
8
6
4
2
0
-2
-4
-6
4
2
0
-2
-4
-6
-8
2
1
0
-1
-2
-3
-4
-5
-6
8

Sparsity of Convex-NMF

- Sparse factorization is a recent trend.
- Sparsity is usually explicitly enforced
- Convex-NMF factors are naturally sparse

$$\| X - XWG^T \|_F^2 = \| I - WG^T \|_{X^T X}^2 = \sum_k \sigma_k^2 \| v_k^T (I - WG^T) \|^2$$

Consider $$\| I - WG^T \|^2 = \sum_k \| e_k^T (I - WG^T) \|^2$$

Solution is

$$G = W = (e_1, \cdots, e_k) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From this we infer convex NMF factors are naturally sparse

A Simple Example

$$X = \begin{pmatrix} \overbrace{\begin{matrix} 1.3 & 1.8 & 4.8 \\ 1.5 & 6.9 & 3.9 \\ 6.5 & 1.6 & 8.2 \\ 3.8 & 8.3 & 4.7 \\ -7.3 & -1.8 & -2.1 \end{matrix}}^{cluster\,1} & \overbrace{\begin{matrix} 7.1 & 5.0 & 5.2 & 8.0 \\ -5.5 & -8.5 & -3.9 & -5.5 \\ -7.2 & -8.7 & -7.9 & -5.2 \\ 6.4 & 7.5 & 3.2 & 7.4 \\ 2.7 & 6.8 & 4.8 & 6.2 \end{matrix}}^{cluster\,2} \end{pmatrix}$$

$$F_{\text{svd}} = \begin{pmatrix} -0.41 & 0.50 \\ 0.35 & 0.21 \\ 0.66 & 0.32 \\ -0.28 & 0.72 \\ -0.43 & -0.28 \end{pmatrix}, \quad F_{\text{semi}} = \begin{pmatrix} 0.05 & 0.27 \\ 0.40 & -0.40 \\ 0.70 & -0.72 \\ 0.30 & 0.08 \\ -0.51 & 0.49 \end{pmatrix}, \quad F_{\text{cnvx}} = \begin{pmatrix} 0.31 & 0.53 \\ 0.42 & -0.30 \\ 0.56 & -0.57 \\ 0.49 & 0.41 \\ -0.41 & 0.36 \end{pmatrix}, \quad C_{\text{Kmeans}} = \begin{pmatrix} 0.29 & 0.52 \\ 0.45 & -0.32 \\ 0.59 & -0.60 \\ 0.46 & 0.36 \\ -0.41 & 0.37 \end{pmatrix}$$

$$\| F_{convex} - C_{Kmeans} \| = 0.08$$

$$\| F_{semi} - C_{Kmeans} \| = 0.53$$

$$G^T_{\text{svd}} = \begin{pmatrix} 0.25 & 0.05 & 0.22 & -.45 & -.44 & -.46 & -.52 \\ 0.50 & 0.60 & 0.43 & 0.30 & -0.12 & 0.01 & 0.31 \end{pmatrix}$$

$$G^T_{\text{semi}} = \begin{pmatrix} 0.61 & 0.89 & 0.54 & 0.77 & 0.14 & 0.36 & 0.84 \\ 0.12 & 0.53 & 0.11 & 1.03 & 0.60 & 0.77 & 1.16 \end{pmatrix}$$

$$G^T_{\text{cnvx}} = \begin{pmatrix} 0.31 & 0.31 & 0.29 & 0.02 & 0 & 0 & 0.02 \\ 0 & 0.06 & 0 & 0.31 & 0.27 & 0.30 & 0.36 \end{pmatrix}$$

$$\| X - FG^T \| = 0.27940,\ 0.27944,\ 0.30877$$

SVD Semi Convex

Experiments on 7 datasets

	Reuters	URCS	WebKB4	Log	WebAce	Ionosphere	Wave
data sign	+	+	+	+	+	±	±
# instance	2900	476	4199	1367	2340	351	5000
# class	10	4	4	9	20	2	2
	Clustering Accuracy						
K-means	0.4448	0.4250	0.3888	0.6876	0.4001	0.4217	0.5018
NMF	0.4947	0.5713	0.4218	0.7805	0.4761	-	-
Semi-NMF	0.4867	0.5628	0.4378	0.7385	0.4162	0.5947	0.5896
Convex-NMF	0.4789	0.5340	0.4358	0.7257	0.4086	0.5470	0.5738
	Sparsity (percentage of nonzeros)						
Semi-NMF	0.9720	0.9688	0.9993	0.9104	0.9543	0.8177	0.9747
Convex-NMF	0.6152	0.6448	0.5976	0.5070	0.6427	0.4986	0.4861
	Orthogonality						
Semi-NMF	0.6578	0.5527	0.7785	0.5924	0.7253	0.9069	0.5461
Convex-NMF	0.1979	0.1948	0.1146	0.4815	0.5072	0.1604	0.2793

NMF variants always perform better than K-means

Kernel NMF -- Generalized Convex NMF

Map feature vector to higher-dim space

$$x_i \rightarrow \phi(x_i) \qquad \phi(X) = [\phi(x_1), \phi(x_2), \cdots, \phi(x_n)]$$

NMF/semi-NMF $\phi(X) = FG^T$ depends on the explicit mapping function $\phi(\bullet)$

Kernel NMF: $\phi(X) = [\phi(X)W]G^T$

Minimization objective depends on kernel only:

$$\| \phi(X) - \phi(X)WG^T \|^2 = \mathrm{Tr}(I - GW^T)\langle \phi(X), \phi(X) \rangle (I - WG^T)$$

(Ding & He, ICML 2004)

Kernel *K*-means Clustering

Map feature vector to higher-dim space

$$x_i \longrightarrow \phi(x_i)$$

Kernel ***K***-means objective:

$$\min J_K^\phi = \sum_{k=1}^{K} \sum_{i \in C_k} || \phi(x_i) - \phi(c_k) ||^2 \qquad \phi(c_k) \equiv \frac{1}{n_k} \sum_{i \in C_k} \phi(x_i)$$

$$J_K^\phi = \sum_i | \phi(x_i) |^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \phi(x_i)^T \phi(x_j)$$

Kernal ***K***-means optimization:

$$\max J_K^\phi = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \left\langle \phi(x_i), \phi(x_j) \right\rangle = \mathrm{Tr}(H^T W H)$$

Matrix of pairwise similarities

NMF and PLSI : Equivalence

So far we only use the Frobenius norm as the NMF objective function. Another objective is the KL divergence

$$x_i \longrightarrow \phi(x_i)$$

Kernel ***K***-means objective:

$$\min J_K^\phi = \sum_{k=1}^{K} \sum_{i \in C_k} \| \phi(x_i) - \phi(c_k) \|^2 \qquad \phi(c_k) \equiv \frac{1}{n_k} \sum_{i \in C_k} \phi(x_i)$$

$$J_K^\phi = \sum_i |\phi(x_i)|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \phi(x_i)^T \phi(x_j)$$

Kernal ***K***-means optimization:

$$\max J_K^\phi = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in C_k} \left\langle \phi(x_i), \phi(x_j) \right\rangle = \mathrm{Tr}(H^T W H)$$

Matrix of pairwise similarities

Kernel-NMF Algorithm

Computing algorithm depends only on the kernel

$$\langle \phi(X), \phi(X) \rangle$$

Update F:

$$W_{ik} \leftarrow W_{ik} \sqrt{\frac{[\boxed{(X^T X)}^+ G]_{ik} + [(X^T X)^- W G^T G]_{ik}}{[(X^T X)^- G]_{ik} + [(X^T X)^+ W G^T G]_{ik}}}$$

Update G:

$$G_{ik} \leftarrow G_{ik} \sqrt{\frac{[W^T (X^T X)^+]_{ik} + [G W^T (X^T X)^- W]_{ik}}{[W^T (X^T X)^-]_{ik} + [G W^T (X^T X)^+ W]_{ik}}}$$

(Ding, Li, Jordan, 2006)

Orthogonal Nonnegative Tri-Factorization

3-factor NMF with explicit orthogonality constraints

$$\min_{\substack{F^T F = I, F \geq 0 \\ G^T G = I, G \geq 0}} \| X_{\pm} - F_{+} S_{\pm} G_{+}^T \|^2$$

1. Solution is unique
2. Can't reduce to NMF

Simultaneous K-means clustering of rows and columns

$F = (f_1, f_2, \cdots, f_k)$ => Row cluster indicators

$G = (g_1, g_2, \cdots, g_k)$ => Column cluster indicators

(Ding, Li, Peng, Park, KDD 2006)

K-means clustering objective function

$$X = (x_1, x_2, \cdots, x_n) \;=\; \text{input data}$$

$$F = (f_1, f_2, \cdots, f_k) \;=\; \text{cluster centroids}$$

$$G = (g_1, g_2, \cdots, g_k) \;=\; \text{cluster indicators}$$

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - f_k \|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} g_{ik} \| x_i - f_k \|^2 = \| X - FG^T \|^2$$

NMF-like algorithms are different ways to relax F , G !

$$f_k = Xg_k / n_k, \;\; F = XGD_n^{-1}, D_n = diag(n_1, \cdots, n_k)$$

$$J_K = \| X - XGD_n^{-1}G^T \|^2 = \| X - X\tilde{G}\tilde{G}^T \|^2, \tilde{G}^T\tilde{G} = I$$

NMF ⇔ PLSI

NMF objective functions

- Frobenius norm
- KL-divergence:

$$J_{KL} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \log \frac{x_{ij}}{(FG^T)_{ij}} - x_{ij} + (FG^T)_{ij}$$

Probabilistic LSI (Hoffman, 1999) is a latent variable model for clustering:

$$J_{PLSI} = \sum_{i=1}^{m} \sum_{j=1}^{n} x(w_i, d_j) \log p(w_i, d_j)$$

$$p(w_i, d_j) = \sum_{k} p(w_i \mid z_k) p(z_k) p(d_j \mid z_k)$$

We can show $J_{PLSI} = -J_{NMF-KL} + \text{constant}$

(Ding, Li, Peng, AAAI 2006)

Summary

- NMF is doing K-means clustering (or PLSI)
- **Interpretability** is key to motivate new NMF-like factorizations
  - Semi-NMF, Convex-NMF, Kernel-NMF, Tri-NMF
- NMF-like algorithms always outperform K-means clustering
- Advantage: hard/soft clustering
- **Convex-NMF** enforces notion of cluster centroids and is naturally sparse

NMF: A new/rich paradigm for unsupervised learning

References

- On the Equivalence of Nonnegative Matrix Factorization and K-means /Spectral clustering, Chris Ding, Xiaofeng He, Horst Simon, SDM 2005.
- Convex and Semi-Nonnegative Matrix Factorization, Chris Ding, Tao Li, Michael Jordan, submitted
- Orthogonal Non-negative Matrix Tri-Factorization for clustering, Chris Ding, Tao Li, Wei Peng, Haesun Park, KDD 2006.
- Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-square and a Hybrid Algorithm, Chris Ding, Tao Li, Wei Peng, AAAI 2006.

Data Clustering: NMF and PCA

NMF is useful due to nonnegativity.

$$\min_{G^T G = I, G \geq 0} \| X_\pm - F_\pm G_+^T \|^2$$

G-orthogonality and nonnegativity

$$F = (f_1, f_2, \cdots, f_k) \quad \Rightarrow \quad \text{cluster centroids}$$

$$G = (g_1, g_2, \cdots, g_k) \quad \Rightarrow \quad \text{cluster indicators}$$

What happens if we ignore nonnegativity?

K-means clustering ⇔ PCA

Ignore nonnegativity => orth. transform *R*

$$\min_{G^T G=I, G\geq 0} \| X_{\pm} - (F_{\pm}R)(G_{+}R)^T \|^2$$

Equivelevant to $\max_{GR} \mathrm{Tr}\,[(GR)^T X^T X (GR)]$

(Ding & He, ICML 2004)

Solution is given by SVD: $X = U\Sigma V^T, U = FR, V = GR$

Cluster indicator projection: $GG^T = (GR)(GR)^T = VV^T$

Centroid subspace projection: $FF^T = (FR)(FR)^T = UU^T$

PCA/SVD is automatically doing K-means clustering

A Simple Example

$$X = \begin{pmatrix} \overbrace{\begin{matrix} 1.3 & 1.8 & 4.8 \\ 1.5 & 6.9 & 3.9 \\ 6.5 & 1.6 & 8.2 \\ 3.8 & 8.3 & 4.7 \\ -7.3 & -1.8 & -2.1 \end{matrix}}^{cluster\,1} & \overbrace{\begin{matrix} 7.1 & 5.0 & 5.2 & 8.0 \\ -5.5 & -8.5 & -3.9 & -5.5 \\ -7.2 & -8.7 & -7.9 & -5.2 \\ 6.4 & 7.5 & 3.2 & 7.4 \\ 2.7 & 6.8 & 4.8 & 6.2 \end{matrix}}^{cluster\,2} \end{pmatrix}$$

$$F_{\mathrm{svd}} = \begin{pmatrix} -0.41 & 0.50 \\ 0.35 & 0.21 \\ 0.66 & 0.32 \\ -0.28 & 0.72 \\ -0.43 & -0.28 \end{pmatrix}, \quad F_{\mathrm{semi}} = \begin{pmatrix} 0.05 & 0.27 \\ 0.40 & -0.40 \\ 0.70 & -0.72 \\ 0.30 & 0.08 \\ -0.51 & 0.49 \end{pmatrix}, \quad F_{\mathrm{cnvx}} = \begin{pmatrix} 0.31 & 0.53 \\ 0.42 & -0.30 \\ 0.56 & -0.57 \\ 0.49 & 0.41 \\ -0.41 & 0.36 \end{pmatrix}, \quad C_{\mathrm{Kmeans}} = \begin{pmatrix} 0.29 & 0.52 \\ 0.45 & -0.32 \\ 0.59 & -0.60 \\ 0.46 & 0.36 \\ -0.41 & 0.37 \end{pmatrix}$$

$$\| F_{convex} - C_{Kmeans} \| = 0.08$$

$$\| F_{semi} - C_{Kmeans} \| = 0.53$$

$$G^T_{\mathrm{svd}} = \begin{pmatrix} 0.25 & 0.05 & 0.22 & -.45 & -.44 & -.46 & -.52 \\ 0.50 & 0.60 & 0.43 & 0.30 & -0.12 & 0.01 & 0.31 \end{pmatrix}$$

$$G^T_{\mathrm{semi}} = \begin{pmatrix} 0.61 & 0.89 & 0.54 & 0.77 & 0.14 & 0.36 & 0.84 \\ 0.12 & 0.53 & 0.11 & 1.03 & 0.60 & 0.77 & 1.16 \end{pmatrix}$$

$$G^T_{\mathrm{cnvx}} = \begin{pmatrix} 0.31 & 0.31 & 0.29 & 0.02 & 0 & 0 & 0.02 \\ 0 & 0.06 & 0 & 0.31 & 0.27 & 0.30 & 0.36 \end{pmatrix}$$

$$\| X - FG^T \| = 0.27940,\ 0.27944,\ 0.30877$$

SVD Semi Convex

NMF = Spectral Clustering (Normalized Cut)

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

(Gu , et al, 2001)

cluster indicators:

$$y_k = D^{1/2}(0\cdots0,\overbrace{1\cdots1}^{n_k},0\cdots0)^T / \| D^{1/2} h_k \|$$

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