Nonnegative Matrix Factorizations for data clustering and combinatorial optimizations

Chris Ding  Computer Science & Engineering Dept
University of Texas at Arlington

In collaboration with

Tao Li  Florida International University
Michael Jordan  University of California at Berkeley
Nonnegative Matrix Factorization (NMF)

Data Matrix: $n$ points in $p$-dimensional space:

$$X = (x_1, x_2, \cdots, x_n) \quad x_i \text{ is an image, document, webpage, etc}$$

Factorization (low-rank approximation)

$$X \approx FG^T$$

Nonnegative Matrices

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

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

- Earlier work by statistics people (cf. G. Golub)
- P. Paatero (1994) Environmetrics
  - Parts of whole (no cancellation)
  - A multiplicative update algorithm
Application. Image Compression/Representations

Pixel vector
Converting a set of images into a big matrix

A big matrix of linearized images

<table>
<thead>
<tr>
<th></th>
<th>0.0</th>
<th>0.5</th>
<th>0.7</th>
<th>1.0</th>
<th>0.8</th>
<th>0.2</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
<td>1.8</td>
<td>0.2</td>
<td>0.6</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.7</td>
<td>0.3</td>
<td>1.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>0.3</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>1.0</td>
<td>0.9</td>
<td>0.3</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>1.0</td>
<td>0.9</td>
<td>0.3</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

... ...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
...
NMF of a set of images

\[ X = (x_1, x_2, \cdots, x_n) \]

Matrix of images

\[ F = (f_1, f_2, \cdots, f_k) \quad G = (g_1, g_2, \cdots, g_k) \]
“Parts of Whole” Picture


Straightforward NMF doesn’t get parts-of-whole

Several People explicitly sparsify $F$ to get parts-of-whole

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

$$F = (f_1, f_2, \ldots, f_k)$$
At first glance, NMF is matrix factorization. It has nothing to do with data clustering.
Experiments: NMF gives holistic pictures (2005)

Even though we set $K = 16$ incorrectly (should be 40)
Experiments: NMF gives holistic pictures (2005)

Even though we set $K = 16$ incorrectly (should be 10)
Enhance NMF Interpretability (Ding, Li, Jordan, 2006)

Original images

Convex-NMF factors  NMF factors
NMF is interesting because it does data clustering
Theorem: NMF = K-means Clustering

The proof is based on our earlier work in proving

K-means clustering = PCA
Outline

• Introduction

• **K-means clustering = PCA**

• NMF = K-means clustering

• Various generalizations, convex-NMF for enhanced cluster centroid interpretation.

• Constrained (semi-supervised) clustering

• Using NMF to solve NP-hard problems
  – multi-way Normalized Cut spectral clustering
  – Compute maximum cliques and bicliques
$K$-means clustering

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

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

$K$-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: \[ H = (h_1, \cdots, h_K) \]

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

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

Solving K-mean \[ \Rightarrow \max_{H^T H = I, H \geq 0} \text{Tr}(H^T X^T X H) \]

(Zha, Ding, Gu, He, Simon, NIPS 2001)
(Ding & He, ICML 2004)
Reformulate K-means Clustering

Cluster membership indicators:

\[
\begin{array}{ccc}
C_1 & C_2 & C_3 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{array}
\]

\[= (h_1, h_2, h_3) = H\]
K-means Clustering = PCA

\[
\max_{Q_{k-1} \geq 0, Q_{k-1}^T Q_{k-1} = I} \quad \text{Tr}[Q_{k-1}^T (X^T X) Q_{k-1}]
\]

\[
Q_{k-1} = (q_2, \ldots, q_k)
\]

Optimal solutions of \( q_2 \cdots q_k \) are given by principal components \((v_2 \cdots v_k)\). (by Ky Fan Theorem, 1955)

\( J_K \) is bounded below by total variance minus sum of \( K \) eigenvalues of covariance:

\[
n x^2 - \sum_{k=1}^{K-1} \lambda_k < \min J_K < n x^2
\]
Simple Illustration

(A) Data points

(B) Principal Component

\[ A = \{ i \mid v_1(i) < 0 \}, \ B = \{ i \mid v_1(i) \geq 0 \} \]
PCA + Kmeans

• A long-hold practice in dealing with high dimensional data:
  – Use PCA to project to low-dimension subspace
  – Do K-means clustering in the subspace
• Popular procedure, successful in finding right clusters.
  – But no one understand why.
  – Generally explained that as noise reduction
  – However, \( X+\delta X \) change the subspace =eigenvector\([(X+\delta X)^T(X+\delta X)]\)
• K-means – PCA Equivalence Theory shed new insight
  – PCA guides us towards global solution, while K-means is easily trapped in local minima at high dimensions
Cluster Subspace = PCA Subspace

Cluster Subspace is spanned by $K$ cluster centroids $c_k$

We represent it between-class scatter matrix:

$$S_b = \sum_k n_k c_k c_k^T$$

$S_b$ project $x$ into the cluster subspace

We can prove:

$$\sum_k n_k c_k c_k^T = \sum_k \lambda_k u_k u_k^T$$

This is identical to PCA projection

$$\sum_k u_k u_k^T$$

PCA automatically projects into cluster subspace
Global K-means solution in PCA Subspace

20 Newsgroup Dataset

C. Ding, NMF for data clustering and combinatorial optimization
Global K-means solution in PCA Subspace

AT&T Face Dataset
Global K-means solution in PCA Subspace

CAR Dataset

C. Ding, NMF for data clustering and combinatorial optimization
Outline

• Introduction
• K-means clustering = PCA
• **NMF = K-means clustering**
• Various generalizations, convex-NMF for enhanced cluster centroid interpretation.
• Constrained (semi-supervised) clustering
• Using NMF to solve NP-hard problems
  – multi-way Normalized Cut spectral clustering
  – Compute maximum cliques and bicliques
K-means Clustering Theorem

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

\[
\min_{G^T G = I, G \geq 0; F} \| X_{\pm} - F_{\pm} G^T \|^2
\]

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

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

(Ding, Li, Jordan, 2006)
NMF-Kmeans Theorem Proof

(Ding, He, Simon, SDM 2005)

\[ J = \| X - FG^T \|^2 = Tr((X - FG^T)^T (X - FG^T)) \]
\[ = Tr(X^T X - 2F^T XG - FF^T) \]
\[ \frac{\partial J}{\partial F} = 2XG - 2F = 0 \]

\[
\min_{G^T G = I, G \geq 0} \text{Tr}(X^T X - G^T X^T XG)
\]
\[
\max_{G^T G = I, G \geq 0} \text{Tr}(G^T X^T XG)
\]
K-means Clustering $\Rightarrow$ Spectral Relaxation and Nonnegative matrix Factorization

$$\max_{Q \geq 0, Q^T Q = I} \text{Tr}[Q^T (X^T X) Q]$$

Nonnegative constraint: $Q \geq 0$

Orthogonality constraint: $Q^T Q = I$

Keep nonnegativity and ignore orthogonality $\Rightarrow$ NMF

Ignore nonnegativity and keep orthogonality $\Rightarrow$ Spectral clustering

(Ding, ICML2005 Tutorial)
Many **unsupervised learning methods** are closely related in a simple way.

Data Clustering = Matrix Factorizations

Tensor Decompositions: ParaFac, HOSVD do clustering automatically
K-means Clustering = PCA

\[
\max_{Q_{k-1} \geq 0, Q_{k-1}^T Q_{k-1} = I} \quad \text{Tr}[Q_{k-1}^T (X^T X) Q_{k-1}]
\]

Optimal solutions of \( q_2 \cdots q_k \) are given by principal components \( (v_2 \cdots v_k) \).

\( J_K \) bounds:

\[
n x^2 - \sum_{k=1}^{K-1} \lambda_k < \min J_K < nx^2
\]

This Eckart-Young type error bounds are derived for tensor decompositions ParaFac & HOSVD

(Ding et al, CVPR’08, ICML’09)
NMF Objective Functions

There are two objective function for $X=FG^T$

**Least Squares:**

$$J_{Least-square} = ||X - FG^T||^2 = \sum_{ij} (X - FG^T)_{ij}$$

**KL-divergence:**

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

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

\[ p(w_i, d_j) = \sum_b p(w_i \mid z_k) p(z_k) p(d_j \mid z_k) \]
\[ \sum_{i=1}^{m} p(w_i \mid z_k) = 1, \sum_{j=1}^{n} p(d_j \mid z_k) = 1, \sum_{k=1}^{K} p(z_k) = 1 \]

maximize the likelihood:

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

PLSI assumes the class conditional independence:

\[ P(w, d \mid z) = P(w \mid z)P(d \mid z) \]
NMF using KL-divergence

NMF objective function minimize the KL-divergence:

\[ X = CH^T = \tilde{C}S\tilde{H}^T \]

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

We prove NMF and PLSI are equivalent:

(a) \[ J_{PLSI} = -J_{NMF-KL} + \text{constant} \]

(b) \[ \sum_{i=1}^{m} \tilde{C}_{ik} = 1, \sum_{j=1}^{n} \tilde{H}_{jk} = 1, \sum_{k=1}^{n} S_{kk} = 1 \]  

(Ding, Li, Peng, AAAI’06)
NMF and PLSI objectives

= Chi-square

$J_{NMF} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{[(CH^T)_{ij} - F_{ij}]^2}{2F_{ij}} - \frac{[(CH^T)_{ij} - F_{ij}]^3}{3F_{ij}^2} + \ldots$

This can be used to give confidence levels
Hybrid Algorithm

NMF and PLSI optimize same objective function, but differ in details.

• Use NMF to jump out of the local minima reached by PLSI.

• Use PLSI to jump out of the local minima reached by NMF

<table>
<thead>
<tr>
<th></th>
<th>Reuters</th>
<th>WebKB</th>
<th>CSTR</th>
<th>WebAce</th>
<th>Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>0.316</td>
<td>0.410</td>
<td>0.617</td>
<td>0.416</td>
<td>0.775</td>
</tr>
<tr>
<td>NMF-only</td>
<td>0.454</td>
<td>0.619</td>
<td>0.666</td>
<td>0.520</td>
<td>0.778</td>
</tr>
<tr>
<td>PLSI-only</td>
<td>0.487</td>
<td>0.510</td>
<td>0.668</td>
<td>0.519</td>
<td>0.779</td>
</tr>
<tr>
<td>Hybrid</td>
<td>0.521</td>
<td>0.644</td>
<td>0.878</td>
<td>0.523</td>
<td>0.781</td>
</tr>
</tbody>
</table>

Table 3: Clustering Accuracy.
NMF differs from PLSI: Example

\[
X = \begin{pmatrix}
.048 & .042 & .047 & .024 & .029 & .026 \\
.035 & .040 & .045 & .016 & .023 & .029 \\
.031 & .019 & .031 & .040 & .045 & .042 \\
.027 & .023 & .031 & .032 & .039 & .045 \\
.047 & .043 & .035 & .026 & .021 & .019
\end{pmatrix}.
\]

The initial \( C_0, S_0, H_0 \) are

\[
C_0 S_0 H_0^T = \begin{pmatrix}
.24 & .20 \\
.02 & .27 \\
.31 & .16 \\
.07 & .26 \\
.36 & .11
\end{pmatrix} \begin{pmatrix}
.34 & 0 \\
0 & .66
\end{pmatrix} \begin{pmatrix}
.18 & .19 \\
.15 & .18 \\
.15 & .21 \\
.18 & .12 \\
.18 & .14 \\
.16 & .16
\end{pmatrix}^T
\]

\( \tilde{C} \tilde{S} \tilde{H}^T = \begin{pmatrix}
.33 & .14 \\
.29 & .12 \\
.02 & .33 \\
.05 & .29 \\
.32 & .11
\end{pmatrix} \begin{pmatrix}
.39 & 0 \\
0 & .61
\end{pmatrix} \begin{pmatrix}
.27 & .14 \\
.28 & .09 \\
.25 & .15 \\
.07 & .18 \\
.06 & .22 \\
.06 & .23
\end{pmatrix}^T
\]

\( CSH^T = \begin{pmatrix}
.12 & .31 \\
.10 & .28 \\
.38 & .04 \\
.33 & .07 \\
.08 & .31
\end{pmatrix} \begin{pmatrix}
.50 & 0 \\
0 & .50
\end{pmatrix} \begin{pmatrix}
.13 & .25 \\
.09 & .25 \\
.14 & .24 \\
.19 & .09 \\
.22 & .09 \\
.23 & .09
\end{pmatrix}^T
\]
Symmetric NMF: \[ W \approx HH^T \]

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: \[
\max_{H^T H = I, H \geq 0} \text{Tr}(H^T WH)
\]
Orthogonality in NMF

Strict orthogonal $G$: hard clustering

Non-orthogonal $G$: soft clustering

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

$G = (g_1, g_2)$
NMF is more flexible than K-means

K-means: model spherical distributions

NMF: model more complex distributions
NMF Normalization

(Li, Ding, ICDM’06)

If \((F, G)\) is a solution for NMF.

For any diagonal \(D > 0\), then \((FD), (GD^{-1})\) is also a solution.

The next theorem provides a natural normalization
NMF Boundedness

(Zhang, Li, Ding, ICDM’06)

Assume $X = FG^T$ holds rigoursly.

If $X$ is bounded: $X_{ij} \leq 1$.

Then the solution is bounded, i.e.

There is a unique diagonal $D>0$, such that

$$(FD)_{ik} \leq 1, \ (GD^{-1})_{ik} \leq 1$$

Note. This property does not hold for SVD
Outline

• Introduction
• K-means clustering = PCA
• NMF = K-means clustering
• Generalization, convex-NMF for enhanced cluster centroid interpretation.
• Constrained (semi-supervised) clustering
• Using NMF to solve NP-hard problems
  – multi-way Normalized Cut spectral clustering
  – Compute maximum cliques and bicliques
NMF Variations & Usages

NMF: \( X = FG^T \)  
Simplest, LeastSquare and KL versions

Semi-NMF: \( X_\pm = F_\pm G_+^T \)  
Mixed sign input data

Convex-NMF: \( X_\pm = X_\pm W + G_+^T \)  
Enhance clustering interpretation

Kernel-NMF: \( \phi(X_\pm) = \phi(X_\pm)W + G_+^T \)  
Feature & data simultaneous clustering

Tri-NMF: \( X_\pm = F S_\pm G_+^T \)  

Semi-supervised clustering \( \max \| (W + \alpha A - \beta B) - HH^T \| \)
Generalize to semi-NMF

NMF requires data are nonnegative

In statistics, data are typically centered, i.e., data have mixed signs.

We generalize NMF to \( X_{\pm} \approx F_{\pm} G_T \)

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

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

\[ \min \| X - FG^T \| \]

Update F: \[ F = XG(G^T G)^{-1} \]

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

(Ding, Li, Jordan, 2006)
Convex-NMF

(Ding, Li, Jordan, 2006)

In NMF\[ X_+ = F_+ G_T \]

In Semi-NMF\[ X_- = F_- G_T \]

is in a large space

For \( f_k \) factor to capture the notion of cluster centroid, we require \( f_k \) to be convex combination of input data

\[
 f_k = w_{1k} x_1 + \cdots + w_{1n} x_n, \quad F = XW_+
\]

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

\[
 X_\pm = X_\pm W + G_T
\]
A Simple Example

\[
\begin{array}{c}
\text{cluster 1} \\
\begin{pmatrix}
1.3 & 1.8 & 4.8 \\
1.5 & 6.9 & 3.9 \\
6.5 & 1.6 & 8.2 \\
3.8 & 8.3 & 4.7 \\
-7.3 & -1.8 & -2.1
\end{pmatrix}
\end{array}
\begin{array}{c}
\text{cluster 2} \\
\begin{pmatrix}
7.1 & 5.0 & 5.2 & 8.0 \\
-5.5 & -8.5 & -3.9 & -5.5 \\
-7.2 & -8.7 & -7.9 & -5.2 \\
6.4 & 7.5 & 3.2 & 7.4 \\
2.7 & 6.8 & 4.8 & 6.2
\end{pmatrix}
\end{array}
\]

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

\[
\begin{align*}
\| F_{convx} - C_{Kmeans} \| &= 0.08 \\
\| F_{semi} - C_{Kmeans} \| &= 0.53
\end{align*}
\]

\[
G_{svd}^T = \begin{pmatrix}
0.25 & 0.05 & 0.22 & -0.45 & -0.44 & -0.46 & -0.52 \\
0.50 & 0.60 & 0.43 & 0.30 & -0.12 & 0.01 & 0.31
\end{pmatrix},
\quad G_{semi}^T = \begin{pmatrix}
0.61 & 0.89 & 0.54 & 0.77 & 0.14 & 0.36 & 0.84 \\
0.12 & 0.53 & 0.11 & 1.03 & 0.60 & 0.77 & 1.16
\end{pmatrix},
\quad G_{convx}^T = \begin{pmatrix}
0.31 & 0.31 & 0.29 & 0.02 & 0 & 0 & 0.02 \\
0 & 0.06 & 0 & 0.31 & 0.27 & 0.30 & 0.36
\end{pmatrix}
\]

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

SVD  Semi  Convex
Original images

Convex-NMF factors
Two trials with random init

NMF factors
Two trials with random init
Kernel NMF -- Generalized Convex NMF

Map data vector to higher-dim feature space

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

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:

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

(Ding, Li, Jordan, 2006)
Orthogonal Nonnegative Tri-Factorization

3-factor NMF with explicit orthogonality constraints

\[
\min_{F^T F = I, F \geq 0, G^T G = I, G \geq 0} \| X - F S G^T \|^2
\]

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

Simultaneous K-means clustering of rows and columns

\[
F = (f_1, f_2, \cdots, f_k) \quad \Rightarrow \quad \text{Row cluster indicators}
\]

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

(Ding, Li, Peng, Park, KDD 2006)
<table>
<thead>
<tr>
<th>Factorizations</th>
<th>Updating Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF</td>
<td>$F_{ik} \leftarrow F_{ik} \frac{(XG)<em>{ik}}{(FG^T G)</em>{ik}}$</td>
</tr>
<tr>
<td></td>
<td>$G_{jk} \leftarrow G_{jk} \frac{(X^T F)<em>{jk}}{(GF^T F)</em>{jk}}$</td>
</tr>
<tr>
<td>Semi-NMF</td>
<td>$F = XG(G^T G)^{-1}$</td>
</tr>
<tr>
<td></td>
<td>$G_{ik} \leftarrow G_{ik} \sqrt{\frac{(X^T F)<em>{ik}^+ + [G(F^T F)]</em>{ik}}{(X^T F)<em>{ik}^- + [G(F^T F)]</em>{ik}^+}}$</td>
</tr>
<tr>
<td>Convex-NMF</td>
<td>$G_{ik} \leftarrow G_{ik} \sqrt{\frac{[(X^T X)^+ W]<em>{ik} + [GW^T (X^T X)^- W]</em>{ik}}{[(X^T X)^- W]<em>{ik} + [GW^T (X^T X)^+ W]</em>{ik}}}$</td>
</tr>
<tr>
<td></td>
<td>$W_{ik} \leftarrow W_{ik} \sqrt{\frac{(X^T X)^+ G]<em>{ik} + [(X^T X)^- W G^T G]</em>{ik}}{(X^T X)^- G]<em>{ik} + [(X^T X)^+ W G^T G]</em>{ik}}}$</td>
</tr>
<tr>
<td>Tri-Factorization</td>
<td>$G_{jk} \leftarrow G_{jk} \sqrt{\frac{(X^T FS)<em>{jk}}{(GG^T X^T FS)</em>{jk}}}$</td>
</tr>
<tr>
<td></td>
<td>$F_{ik} \leftarrow F_{ik} \sqrt{\frac{(XGS^T)<em>{ik}}{(FF^T XGS^T)</em>{ik}}}$</td>
</tr>
<tr>
<td></td>
<td>$S_{ik} \leftarrow S_{ik} \sqrt{\frac{(F^T XG)<em>{ik}}{(F^T FSG^T G)</em>{ik}}}$</td>
</tr>
<tr>
<td>Kernel-NMF</td>
<td>replace $X^T X$ by $\langle \phi(X)^T \phi(X) \rangle$ in Convex-NMF</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• K-means clustering = PCA
• NMF = K-means clustering
• Various generalizations, convex-NMF for enhanced cluster centroid interpretation.
• **Constrained (semi-supervised) clustering**
• Using NMF to solve NP-hard problems
  – multi-way Normalized Cut spectral clustering
  – Compute maximum cliques and bicliques
Outline

• Introduction
• K-means clustering = PCA
• NMF = K-means clustering
• Generalizations: convex-NMF
• Using NMF to solve NP-hard problems
  – Multi-way Normalized Cut spectral clustering
  – Compute maximum cliques and bicliques
Using NMF to solve Multi-way Normalized Cut spectral clustering

Normalized cut objective function (Shi & Malik, 2000)

\[ J_{nc} = \sum_{1 \leq p < q \leq K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = K - \sum_{\ell=1}^{K} \frac{h_{\ell}^T W h_{\ell}}{h_{\ell}^T D h_{\ell}} \]

\[ s(C_k, C_\ell) = \sum_{i \in C_k} \sum_{j \in C_\ell} w_{ij} \quad \rho(C_k) = \sum_{i \in C_k} d_i \]

\[ h_k = \{0, 1\}^n \text{ be an indicator vector for cluster } C_k \]

Let \( H = (h_1/\|D^{1/2}h_1\|, \ldots, h_K/\|D^{1/2}h_K\|) \)

Constrained optimization

\[ \max_{H^T D H = I, H \geq 0} \text{Tr}(H^T W H) \]

(M. Gu, et al, 2001)
Multi-way Normalized Cut spectral clustering (con’t)

Algorithm: given initial guess of $H$, iteratively update

$$H_{ij} \leftarrow H_{ij} \sqrt{\frac{(WH)_{ij}}{(DH \alpha)_{ij}}}, \quad \alpha \equiv H^T WH.$$

Correctness. If the algorithm converges, the converged solution satisfies the KKT optimality condition.

Convergence. The algorithm converges.

(Ding, Li, Jordan, 2008)
Correctness Proof: The solution converge to a local optima of KKT condition

We begin with the Lagrangian

\[ L = \text{Tr} H^T W H - \text{Tr} \alpha (H^T D H - I). \]

The KKT complementary slackness condition is

\[ (W H - D H \alpha)_{ij} H_{ij} = 0. \]

Some over \( ij \), we obtain

\[ \alpha = H^T W H \]

At convergence, the update rule satisfies

\[ (W H - D H \alpha)_{ij} H_{ij}^2 = 0. \]
Convergence Proof

Using Auxiliary Function (from Machine Learning)

$G(x,x')$ is an auxiliary function of $L(x)$ if

$$G(H,H') \leq L(H), \quad G(H,H) = L(H)$$

We maximize a lower-bound.

set $H^{(t+1)} = \arg \max_H G(H,H^{(t)})$

$$L(H^{(t)}) = G(H^{(t)},H^{(t)}) \leq G(H^{(t+1)},H^{(t)}) \leq L(H^{(t+1)})$$

$$L(H^{(1)}) \leq L(H^{(2)}) \leq L(H^{(3)}) \leq \cdots$$

$L(H)$ is monotonically increasing and is bounded from up. Thus the algorithm converges
Multi-way Normalized Cut spectral clustering (con’t)

Key steps

- Find the auxiliary function \( G(H,H') \geq L(H) \)
- Find the global maxima of \( G(H,H') \) w.r.t. \( H \).

\[
L = \text{Tr} H^T W H - \text{Tr} \alpha (H^T D H - I),
\]

The auxiliary function is

\[
G(H, \tilde{H}) = \sum_k \sum_{i,j} W_{ij} \tilde{H}_{ik} \tilde{H}_{jk} (1 + \log \frac{H_{ik} H_{jk}}{\tilde{H}_{ik} \tilde{H}_{jk}}) - \sum_p \sum_{i,k,l} \left( D \tilde{H}_{ik} H_{ik}^2 \right) \tilde{H}_{ik}
\]

\[
z \geq 1 + \log z
\]

\[
z = \frac{H_{ik} H_{jk}}{\tilde{H}_{ik} \tilde{H}_{jk}}
\]

Derive relevant inequalities

\[
\sum_{i=1}^n \sum_{p=1}^k \frac{(A S' B)_{ip} S_{ip}^2}{S_{ip}'} \geq \text{Tr}(S^T A S B)
\]

C. Ding, NMF for data clustering and combinatorial optimization
Multi-way Normalized Cut spectral clustering (con’t)

Finding global maxima

1st derivatives (gradient)

\[
\frac{\partial G(H, \tilde{H})}{\partial H_{ik}} = 2 \frac{(W \tilde{H})_{ik} \tilde{H}_{ik}}{H_{ik}} - 2 \frac{(D \tilde{H} \alpha)_{ik} H_{ik}}{\tilde{H}_{ik}}
\]

2nd derivatives (Hessian)

\[
\frac{\partial^2 G(H, \tilde{H})}{\partial H_{ik} \partial H_{j\ell}} = -2 \left[ \frac{(W \tilde{H})_{ik} \tilde{H}_{ik}}{H^2_{ik}} + \frac{(D \tilde{H} \alpha)_{ik}}{\tilde{H}_{ik}} \right] \delta_{i,j} \delta_{k,\ell}
\]

Negative semi-definite => concave function => global maxima

Global maxima:

\[
H^2_{ik} = \tilde{H}^2_{ik} \frac{(W \tilde{H})_{ik}}{(D \tilde{H} \alpha)_{ik}}
\]
Multi-way Normalized Cut spectral clustering (con’t)

Experiments

1. Initial clustering using K-means in spectral embedding
2. Update the solution until convergence

Clustering accuracy

wine

soybean
Outline

• Introduction
• K-means clustering = PCA
• NMF = K-means clustering
• Generalizations, convex-NMF for enhanced cluster centroid interpretation.
• Constrained (semi-supervised) clustering
• Using NMF to solve NP-hard problems
  – Multi-way Normalized Cut spectral clustering
  – Directed/undirected graph matching
  – Compute maximum cliques and bicliques

(Ding, Li, Jordan, 2008)
Protein Interactions

A genome has 5000 proteins. Each interacts ~ 5 others.
Protein Interaction Module: densely connected subgraphs
Maximal Cliques in a bipartite Graph
Biclusters in a 2D Dataset
Cliques in a Bipartite Graph

- Finding a complete block in the adjacency matrix
- Two types of maximal bi-cliques:
  - Maximum Node Bicliques: $\max |R|+|C|$ (perimeter)
  - Maximum Edge Bicliques: $\max |R|*|C|$ (area)

- Similarly to bi-clustering, widely used in bioinformatics
- Example. Gene expression profiles: a gene is relevant only for certain subset of cellular processes, not all processes.
Generalized Motzkin-Strauss Theorem for maximal edge biclique

Given bipartite graph with adjacency matrix $B$.
Compute maximal edge bi-clique.

**Generalized Motzkin-Strauss Theorem.**

Vector $x = (x_1, \cdots, x_n)$ on row nodes of the graph
Vector $y = (y_1, \cdots, y_m)$ on column nodes of the graph

$$\max_{x \geq 0, y \geq 0} x^T B y \quad \text{s.t.} \quad x_1^\alpha + \cdots + x_n^\alpha = 1, \alpha > 1$$
$$y_1^\beta + \cdots + y_m^\beta = 1, \beta > 1$$

Non-zero entries define the biclique
Generalized Motzkin-Strauss Theorem for maximal edge biclique

Theorem 8 Let $\beta = 1 + \epsilon, 0 < \epsilon \ll 1$. For an optimal solution $(x^*, y^*)$, if nonzero elements of $x^*$ have the same values, and if nonzero elements of $y^*$ have the same values, then $(R_1, C_1)$ is a maximal edge biclique in $B$. The objective function has the optimal value $J = (|R_1||C_1|)^{1 - 1/\beta}$. 
Algorithm for computing bicliques

\[
\begin{align*}
\max_{x \geq 0, y \geq 0} & \quad x^T A y \\
\text{s.t.} & \quad x_1^\alpha + \ldots + x_n^\alpha = 1, \alpha > 1 \\
& \quad y_1^\beta + \ldots + y_n^\beta = 1, \beta > 1
\end{align*}
\]

Update: 
\[
\begin{align*}
x_i & \leftarrow \left( x_i \frac{(By)_i}{x^T By} \right)^{1/\alpha} \\
y_j & \leftarrow \left( y_j \frac{(B^T x)_j}{x^T By} \right)^{1/\beta}
\end{align*}
\]

Theorem 1. Feasibility 
\[
\sum_i \left[ x_i^{(t+1)} \right]^\beta = \sum_i \frac{x_i^{(t)}}{[x^{(t)}]^T S x^{(t)}} = 1
\]

Theorem 2. At convergence, solution satisfies KKT condition.

Theorem 3. The iterative algorithm converges.
Biclique Example

Solution vector $\mathbf{x}$

Solution vector $\mathbf{y}$

C. Ding, NMF for data clustering and combinatorial optimization
Biclique Example
Biclique Example

There are 6 overlapping maximal bicliques. The algorithm correctly picks up the maximum-edge biclique.
NMF 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 can solve **NP-hard Combinatorial problems**
  - Normalized-Cut, Graph matching, Clique/biclique
- NMF-like algorithms are **extremely simple to implement.**
  - Optimality and convergence proved rigorously.

**NMF: A rich paradigm for unsupervised learning and combinatorial optimization problems**

C. Ding, NMF for data clustering and combinatorial optimization