Non-negative Matrix Factorization with Correntropy Similarity Measure and its Application to Text Mining and Pattern Analysis

Dr. Jacek M. ZURADA

Dr. Jan Chorowski
Dr. Tolga Ensari
Ehsan Hosseini Asl

FEDERATED CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS
Kraków, Poland, 8 - 11 September, 2013
Outline

- Critical review of PCA technique
- Why NMF?
- NMF Algorithms
- Clustering with NMF
- Applications Areas:
  - Image, Gene/Protein, Latent Semantic Indexing
- Document Clustering Using NMF w. Correntropy
- Face Recognition Using NMF w. Correntropy
- Conclusion
Team (and other) Credits

- Dr. Tolga Ensari, Technical University of Istanbul and postdoc at the University of Louisville
- Dr. Jan Chorowski, University of Louisville
- Ehsan Hosseini Asl, University of Louisville
- Parts of this work presented at ICANN 2012 [5], and ICMLA 2012 [26]

- NMF history is like other stories (such as EBP), credits to Paatero et al.
Popular MF Methods

- We have a multi-dimensional data matrix of n samples (as columns) and need to transform it ….. for a reason

- Principal Component Analysis (PCA) finds the best decomposition preserving the most of data variance

- Singular Value Decomposition (SVD) finds the best rank-k factorization (according to the Frobenius norm)

- Independent Component Analysis (ICA) finds the factorization into statistically independent components

- Where does NMF fit in this context?
Principal Component Analysis (PCA)

- Assume $n$ data samples, $\{x\}$, of dimension $m$, and mean value $E(x) = \frac{1}{n} \sum_{k=1}^{n} x_k = 0$

- It can be proven that the ‘best’ projecting transformation of $x$ is $y = Wtx$, where $x$ and $y$ are (mx1), $W$ is (mxm),

- where rows $w_i$ of matrix $W^t$ are normalized Eigenvectors of $R_{xx}$
Principal Component Analysis (ctd)

To achieve data compression, the eigenvalues corresponding to the eigenvectors $w_1, w_2, \ldots, w_m$ need to be ordered in $W$ in the decreasing order so that:

$w_1$ corresponds to the largest eigenvalue, $w_2$ to the second largest, \ldots etc.

Inverse transformation:

$x = Wy$ for $k = m$ (lossless), or

$x \sim W'y'$ for $k \leq m$ (lossy) where $W'$ is $mxk$, $y'$ is $kx1$
Data Matrix Factorization

- Decompose into component matrices $W$ and $H$
- Simple problem, rich solutions/implications

\[
\begin{align*}
A & \rightarrow W \cdot H \\
& (m \times n) \quad (m \times k) \quad (k \times n)
\end{align*}
\]
Matrix Factorization

- Every column of \( A \) represents a sample.
- Decompose the sample matrix:
  \[ A \approx WH \]
- The columns of \( W \) are basis vectors of \( W \) (j=1,2..k).
- The columns of \( H \) are the encodings of the samples:
  \[ A^i = \sum_j W^j H_{ji} \]
Examples of $A$, $W$

Data: $A$ for $m=9$ (pixels)  Three different $W$
  $n=12$ (samples) for $k=10$,  
  10 and 6

Adapted From Hoyer [3]
Non-negative Factorization Motivation

- PCA and ICA find holistic representations of data
- All basis vectors are used to represent each data vector (sample)
- The PCA/ICA basis vectors hard to interpret because complex cancellations of terms need to be analyzed
- **NMF Solution**: constrain all matrices to contain only non-negative elements
- **NMF Result**: no cancellations, linear superposition only, considerable sparsity!
**Definition of NMF**

Given a data matrix $A (m \times n)$, we are looking for $W (m \times k) \geq 0$ and $H: (k \times n) \geq 0$, where $k \leq \min(m, n)$, for the best approximation on:

$$A \approx WHW^T$$

The objective (cost) function is

$$\min_{W,H} \| A - WH \|^2$$

- $W$: Basis vectors matrix, tend to be sparse
- $H$: Coefficient (or encoding) matrix, usually also sparse (nonnegative lower dimensional representation)
- $k$: Low-rank value

Dr. Jacek M. ZURADA Non-negative Matrix Factorization (NMF) for Pattern Analysis
Definition of NMF


- **NMF can be used for**:
  - Unsupervised Dimensionality Reduction
  - Clustering
NMF Algorithms

- Multiplicative Update Rule (Lee & Seung, 2001).
- Gradient Descent Algorithm (Hoyer, 2004).
- Alternating Least Squares (Paatero, 1994).

Objective (Cost) Function:

Square of the Euclidean distance between $A$ and $WH$:

$$
\| A - WH \|_F^2 = \sum_{ij} (A_{ij} - (WH)_{ij})^2
$$

Dr. Jacek M. ZURADA
Non-negative Matrix Factorization (NMF) for Pattern Analysis
NMF Algorithms

- **Multiplicative Update Rule** for $W$ and $H$ matrices:

  $H_{i,j} = H_{i,j} \frac{(W^T A)_{i,j}}{(W^T WH)_{i,j} + \epsilon}$

  $W_{i,j} = W_{i,j} \frac{(A H^T)_{i,j}}{(WHH^T)_{i,j} + \epsilon}$

- Iteratively update until $\|A - WH\|_F^2$ is below some threshold.
- ‘$\epsilon$’ is a sufficiently small positive number to avoid division by 0.
Why to Use NMF? (Advantages)

- Nonnegative constraints are physically meaningful:
  - Pixels in digital image → Image Processing
    In linear combination, NMF does not result in cancellations between positive and negative numbers.
  - Molecule concentration in bioinformatics (e.g. RNA, protein analysis, etc.) → Microarray Analysis
  - Signal intensities in mass spectrometry → Computational Proteomics
  - Gene/Protein Microarray Data Analysis
  - Text/Data Mining
Why not to Use NMF? (Disadvantages)

- The NMF problem has no unique solution!
  So, $W$ and $H$ are not unique!
  Depends on the factorization algorithm, initialization and parameters chosen for numerical decomposition.

- The minimization problem defining NMF is non-convex. All algorithms converge to a local optimum only!
Applications in Image Processing

- In Lee & Seung’s experiment, the image database has $n=2,429$ facial images, each consisting of $m=361$ elements (19x19 pixels).
- NMF learns a set of $k=49$ basis images (in a 7x7 grid).

Applications in Image Processing

- W **basis** and H **encodings** contain a large fraction of coefficients close to 0, so both the **basis images** and **image encodings** are sparse.

- There exist several versions of mouths, noses and other facial parts.

Same Decompositions with PCA

- The original database has $n=2,429$ facial images
- $k=49$ eigenvectors (eigenfaces, left), each arranged as $19\times19$ pixels per square
- PCA learns pos/neg representations (on a $7\times7$ grid).

Applications in Image Processing

- What don’t we like about PCA here?:
  - PCA involves adding up some basis images, then subtracting others.
  - Basis images/vectors not physically intuitive.
  - Subtracting doesn’t make sense in context of some applications.

- How do you subtract eigenfaces?

- What does subtraction mean in the context of document classification, or in other app-s which have only terms >0?
Results for the experiment:

- NMF learns a **parts-based** representation of faces, whereas PCA learns **holistic** representation.

- A face image can be composed as a linear combination of multiple basis images (with non-negativity constraints).

- In linear combination, NMF **does not perform cancellations** between positive and negative numbers.
Clustering with NMF

- Some areas where NMF is used for clustering and classification:
  - Document Clustering (Text Mining)
  - Cancer and Tumor Clustering/Classification
  - Microarray/Gene Expression Data Clustering/Classification
  - 3D Data Clustering
  - Protein Fold Recognition
  - Face Recognition
  - Digit Recognition
  - Audio Source Separation
Applications in Gene/Protein Data

- **Goal:** Discover hidden patterns in large quantities of data produced from microarray experiments.
- Explore data to identify structure without supervision.
- Data are represented in non-negative matrix (genes × samples).
Applications in Gene/Protein Data

‘$k$’ is low-rank value (or number of clusters)
These `$k$’ genes are called meta-genes.

Taken from: Saez P. C. et. al., Biclustering of Gene Expression Data by Non-Smooth Non-negative Matrix Factorization, BMC Bioinformatics, 7:78, 2006.
Applications in Gene/Protein Data

\[
\begin{bmatrix}
2 & 0 & 3 \\
4 & 1 & 3 \\
0 & 2 & 0 \\
0 & 4 & 1 \\
\end{bmatrix}
\begin{bmatrix}
0.6 & 0.0 \\
0.8 & 0.2 \\
0.0 & 0.4 \\
0.0 & 0.9 \\
\end{bmatrix}
\begin{bmatrix}
4.4 & 0.0 & 4.0 \\
0.1 & 4.6 & 0.8 \\
\end{bmatrix}
\]

Cluster 1
Cluster 2
Applications in Gene/Protein Data

Two types of leukemia

\[ A \approx W \times H \]

38 samples  2 classes  38 samples
7129 genes  7129 genes  2 classes

Class 1: 27 samples
Class 2: 11 samples

Taken from: Brunet et al., Metagenes and Molecular Pattern Discovery Using Matrix Factorization, Proc. of the National Acad. of USA, vol. 101, no. 12, March 2004.
Applications in Text Data (Intro)

- **Document Clustering**: Grouping of text documents into meaningful clusters in an unsupervised manner.

![Diagram of document clustering with categories: Government, Science, Arts]
Applications in Text Data (Intro)

- **Interpretation of Basis Vectors:**
  
  We can see the linear combination of Document #5.
  
  Implementation on MED data set (1033 medical abstracts), with $k = 10$.

  ![Diagram of Document 5's linear combination](chart)

  Dominates other encoding entries in col. 5

  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  

  *Taken from: Langville A., Meyer C., Text Mining Using NMF, North Caroline State University, 2005.*
Comparison with Orthogonal-based Factorizations

Directions found by NMF

Directions found by LSI
EXAMPLES:

Three examples to follow about document clustering using NMF:

- **Example 1** → Very small artificial data set, toy example. (contains only 4 documents)
- **Example 2** → Small set of “20-Newsgroups” data set. (contains 900 documents)
- **Example 3** → Training set of “20-Newsgroups” data set. (contains 11,000 documents)
Example 1:

- Clustering a small artificial data set of doc-s using NMF.
- The data set has 4 documents, 6 terms of interest and 3 topics.
- The data matrix is $A$ is 6x4
  (each column shows the term count of a document)

(Topics: ANIMAL d1 and d4, RELIGION d2, FOOD d3)

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>bible</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pizza</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>cat</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>tomato</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>god</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$A = \begin{bmatrix}
5 & 1 & 1 & 6 \\
0 & 4 & 1 & 0 \\
1 & 1 & 5 & 2 \\
6 & 0 & 1 & 5 \\
1 & 1 & 6 & 1 \\
0 & 4 & 0 & 0
\end{bmatrix}$
Example 1:

Results for $W$ and $H$ are ($k=3$ set as a number of topics/clusters):

- The base vectors are columns of $W$ and correspond to ‘a topic’:

<table>
<thead>
<tr>
<th>top_id</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>7.72</td>
<td>0.64</td>
<td>0.99</td>
</tr>
<tr>
<td>bible</td>
<td>0</td>
<td>0.88</td>
<td>3.98</td>
</tr>
<tr>
<td>pizza</td>
<td>1.67</td>
<td>4.95</td>
<td>0.93</td>
</tr>
<tr>
<td>cat</td>
<td>7.74</td>
<td>0.58</td>
<td>0</td>
</tr>
<tr>
<td>tomato</td>
<td>0.88</td>
<td>5.97</td>
<td>0.91</td>
</tr>
<tr>
<td>god</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

- Documents of $A$ encoded in $H$ as their topic/cluster membership values

Maximum in column indicates a dominating topic

Cluster 1 (Topic: ANIMAL)  
Cluster 2 (Topic: FOOD)  
Cluster 3 (Topic: RELIGION)
**Example 1 (ctd):**

Max. of rows in $W$ can be interpreted as representative of clustering terms:

Check the maxima: *dog* and *cat* are dominant in Topic 1, *pizza* and *tomato* in Topic 2, ...

Large $wij$: base vector (or topic) $w_i$ is strongly represented by term $j$

<table>
<thead>
<tr>
<th>top_id</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>7.72</td>
<td>0.64</td>
<td>0.99</td>
</tr>
<tr>
<td>bible</td>
<td>0</td>
<td>0.88</td>
<td>3.98</td>
</tr>
<tr>
<td>(6x3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pizza</td>
<td>1.67</td>
<td>4.95</td>
<td>0.93</td>
</tr>
<tr>
<td>cat</td>
<td>7.74</td>
<td>0.58</td>
<td>0</td>
</tr>
<tr>
<td>tomato</td>
<td>0.88</td>
<td>5.97</td>
<td>0.91</td>
</tr>
<tr>
<td>god</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>doc_id</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.71</td>
<td>0.0</td>
<td>0.04</td>
<td>0.7</td>
</tr>
<tr>
<td>t2</td>
<td>0.02</td>
<td>0.01</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>t3</td>
<td>0</td>
<td>0.99</td>
<td>0.02</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Document Clustering Using NMF

Example 1 in graphics mode:

$$W \times H = A$$

white: 0, black: 1 (or highest positive entry)
**Example 1 revisited, only 3 eigenvalues non-zero:**

**Centered Data Matrix** $A_c$

<table>
<thead>
<tr>
<th>doc id</th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>1.75</td>
<td>-2.25</td>
<td>-2.25</td>
<td>2.75</td>
</tr>
<tr>
<td>bible</td>
<td>-1.25</td>
<td>2.75</td>
<td>-0.25</td>
<td>-1.25</td>
</tr>
<tr>
<td>pizza</td>
<td>-1.25</td>
<td>-1.25</td>
<td>2.75</td>
<td>-0.25</td>
</tr>
<tr>
<td>cat</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
</tr>
<tr>
<td>tomato</td>
<td>-1.25</td>
<td>-1.25</td>
<td>3.75</td>
<td>-1.25</td>
</tr>
<tr>
<td>god</td>
<td>-1</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

**Eigenvalues:**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-8.64E-17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>5.28E-16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2.72E-15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.3487</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8.3777</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16.0235</td>
<td>0</td>
</tr>
</tbody>
</table>

Dr. Jacek M. ZURADA

Non-negative Matrix Factorization (NMF) for Pattern Analysis
Example 1 revisited, eigenvectors shown:

**W (full size)**

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0.2650</td>
<td>0.3102</td>
<td>0.2155</td>
<td>0.6879</td>
<td>0.0282</td>
<td>-0.5594</td>
</tr>
<tr>
<td>bible</td>
<td>0.8438</td>
<td>-0.0858</td>
<td>-0.2328</td>
<td>0.0356</td>
<td>0.3466</td>
<td>0.3238</td>
</tr>
<tr>
<td>pizza</td>
<td>0.0557</td>
<td>-0.7111</td>
<td>0.0227</td>
<td>0.4536</td>
<td>-0.5049</td>
<td>0.1730</td>
</tr>
<tr>
<td>cat</td>
<td>0.3207</td>
<td>-0.4009</td>
<td>0.2383</td>
<td>-0.5263</td>
<td>-0.0779</td>
<td>-0.6296</td>
</tr>
<tr>
<td>tomato</td>
<td>0.3194</td>
<td>0.4333</td>
<td>0.4390</td>
<td>-0.1928</td>
<td>-0.6284</td>
<td>0.2919</td>
</tr>
</tbody>
</table>

**W^T (transformation matrix for A_c to PC space, full size)**

<table>
<thead>
<tr>
<th>V1</th>
<th>dog</th>
<th>bible</th>
<th>pizza</th>
<th>cat</th>
<th>Tomato</th>
<th>god</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0.2650</td>
<td>0.8438</td>
<td>0.0557</td>
<td>0.3208</td>
<td>0.3195</td>
<td>-0.0976</td>
</tr>
<tr>
<td>bible</td>
<td>0.3102</td>
<td>-0.0858</td>
<td>-0.7111</td>
<td>-0.4009</td>
<td>0.4334</td>
<td>-0.2052</td>
</tr>
<tr>
<td>pizza</td>
<td>0.2155</td>
<td>-0.2328</td>
<td>0.0227</td>
<td>0.2383</td>
<td>0.4390</td>
<td>0.8058</td>
</tr>
<tr>
<td>cat</td>
<td>0.6879</td>
<td>0.0356</td>
<td>0.4536</td>
<td>-0.5263</td>
<td>-0.1928</td>
<td>0.0742</td>
</tr>
<tr>
<td>Tomato</td>
<td>0.0282</td>
<td>0.3466</td>
<td>-0.5049</td>
<td>-0.0779</td>
<td>-0.6284</td>
<td>0.4723</td>
</tr>
<tr>
<td>god</td>
<td>-0.5594</td>
<td>0.3238</td>
<td>0.1731</td>
<td>-0.6296</td>
<td>0.2919</td>
<td>0.2654</td>
</tr>
</tbody>
</table>
### Document Clustering Using PCA

**Example 1 revisited:**

#### $H_{pca}$ (3x4)

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1.53E-16</td>
<td>-3.33E-16</td>
<td>3.19E-16</td>
<td>-3.47E-16</td>
</tr>
<tr>
<td>d2</td>
<td>3.05E-16</td>
<td>2.22E-16</td>
<td>-5.83E-16</td>
<td>-1.39E-16</td>
</tr>
<tr>
<td>d3</td>
<td>-7.77E-16</td>
<td>1.33E-15</td>
<td>-1.22E-15</td>
<td>7.77E-16</td>
</tr>
<tr>
<td>d4</td>
<td>-0.8199</td>
<td>0.0258</td>
<td>-0.0538</td>
<td>0.8479</td>
</tr>
<tr>
<td>d5</td>
<td>0.3267</td>
<td>3.9570</td>
<td>-4.2119</td>
<td>-0.0719</td>
</tr>
<tr>
<td>d6</td>
<td>-4.1195</td>
<td>4.2533</td>
<td>3.7425</td>
<td>-3.8763</td>
</tr>
</tbody>
</table>

#### $W_{pca}$ (6x3)

<table>
<thead>
<tr>
<th>basis id</th>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>t4</th>
<th>t5</th>
<th>t6</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
<td>0.2651</td>
<td>0.3102</td>
<td>0.2155</td>
<td>0.6879</td>
<td>0.0282</td>
<td>-0.5595</td>
</tr>
<tr>
<td>bible</td>
<td>0.8438</td>
<td>-0.0858</td>
<td>-0.2328</td>
<td>0.0356</td>
<td>0.3466</td>
<td>0.3238</td>
</tr>
<tr>
<td>pizza</td>
<td>0.0557</td>
<td>-0.7111</td>
<td>0.0228</td>
<td>0.4536</td>
<td>-0.5049</td>
<td>0.1731</td>
</tr>
<tr>
<td>cat</td>
<td>0.3208</td>
<td>-0.4009</td>
<td>0.2383</td>
<td>-0.5263</td>
<td>-0.0779</td>
<td>-0.6296</td>
</tr>
<tr>
<td>tomato</td>
<td>0.3194</td>
<td>0.4333</td>
<td>0.4391</td>
<td>-0.1928</td>
<td>-0.6284</td>
<td>0.2919</td>
</tr>
<tr>
<td>god</td>
<td>-0.0976</td>
<td>-0.2052</td>
<td>0.8058</td>
<td>0.0742</td>
<td>0.4723</td>
<td>0.2654</td>
</tr>
</tbody>
</table>
Document Clustering Using PCA

Example 1 revisited in color:

\[ W_{pca} \times H_{pca} = A_c \]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.75</td>
<td>-2.25</td>
<td>-2.25</td>
<td>2.75</td>
<td></td>
</tr>
<tr>
<td>-1.25</td>
<td>2.75</td>
<td>-0.25</td>
<td>-1.25</td>
<td></td>
</tr>
<tr>
<td>-1.25</td>
<td>-1.25</td>
<td>2.75</td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-1.25</td>
<td>-1.25</td>
<td>3.75</td>
<td>-1.25</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

Black/grey – positive entries
Dark red/red – negative entries

\[
\begin{array}{c|c|c|c|c}
\text{dog} & \text{bible} & \text{pizza} & \text{cat} & \text{tomato} \\
\text{bible} & \text{pizza} & \text{cat} & \text{tomato} & \text{god} \\
\text{pizza} & \text{cat} & \text{tomato} & \text{god} & \text{god} \\
\text{cat} & \text{tomato} & \text{god} & \text{god} & \text{god} \\
\text{tomato} & \text{god} & \text{god} & \text{god} & \text{god} \\
\end{array}
\]

\[
\begin{array}{c|c|c|c}
\text{d1} & \text{d2} & \text{d3} & \text{d4} \\
\text{Dog} & \text{dog} & \text{dog} & \text{dog} \\
\text{Bible} & \text{bible} & \text{bible} & \text{bible} \\
\text{Pizza} & \text{pizza} & \text{pizza} & \text{pizza} \\
\text{Cat} & \text{cat} & \text{cat} & \text{cat} \\
\text{Tomato} & \text{tomato} & \text{tomato} & \text{tomato} \\
\text{God} & \text{god} & \text{god} & \text{god} \\
\end{array}
\]

\[
X = \begin{array}{cc}
\begin{array}{cc}
\begin{array}{cc}
\text{dog} & \text{bible} \\
\text{bible} & \text{pizza} \\
\text{pizza} & \text{cat} \\
\text{cat} & \text{tomato} \\
\text{tomato} & \text{god} \\
\end{array} &
\begin{array}{cc}
\text{dog} & \text{bible} \\
\text{bible} & \text{pizza} \\
\text{pizza} & \text{cat} \\
\text{cat} & \text{tomato} \\
\text{tomato} & \text{god} \\
\end{array}
\end{array}
\end{array}
\]
Recommendation Systems with NMF

Extension of Example 1:

- Compressing a purchase history data using NMF.
- Table of purchases of m items by n users
- The data matrix is $A$ is mxn, with column j showing the purchase record of user j

\[
\begin{array}{cccccc}
\text{purchases} & \text{user1} & \text{user2} & \ldots & \text{usern} \\
\text{item 1} & 5 & 0 & \ldots & 6 \\
\text{item 2} & 0 & 4 & \ldots & 0 \\
\text{item m} & 1 & 3 & \ldots & 0 \\
\end{array}
\]
Example 2:
Initial experiments for small set of “20-Newsgroups” data set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Documents (Small set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-Newsgroups</td>
<td>900</td>
</tr>
</tbody>
</table>
Document Clustering Using NMF

- We propose to minimize the negative of the Correntropy distance measure to improve the factorization quality.

\[ \text{“Correntropy”} = \text{“Correlation”} + \text{“Entropy”} \]

- The entropy-based criterion is defined as a unusual nonlinear distance measure, as opposed to the usual quadratic norm.
Properties of Correntropy

\[
Loss(A, WH) = - \text{Corr}(A, WH) = - \sum_{i,j} \exp \left( \frac{- (A_{ij} - (WH)_{ij})^2}{2 \sigma^2} \right)
\]

**minimize:**

\[
Loss(A, WH)
\]

**subject to:**

\[
W_{ik} \geq 0 \text{ and } H_{kj} \geq 0, \forall \ i, k, j
\]
Properties of Correntropy

- It is **insensitive** to outliers:
  
  Unlike the Euclidean norm, the function $1 + \text{Loss}$ **reaches 1 and flattens** for large error values.

- Therefore, it can be used as an outlier-insensitive similarity measure.
Example 2: MATLAB Output after using NMF Clustering (k=10 clusters) for 20-Newsgroups Data Set:

```
religion christian poepl god line detail valu moral server scienc talk object jesus saw mac built arab frank dwyer configur
RELIGION

name uk mathew shall folk tree righteous pin speed ram ps mb meg isa centri slot ns simm
COMPUTER

mail drive help pleas info anybodi video manufactur monitor vga

name com server file help sandvik newton appl kent ignor spread window

stein brad kill guess imagin final water org reveal river sourc israel isra arab civilian ncsu mb alan norton lebanon hasan
nysernet hernlem lebanes

net know object option thank summar advanc compil righteous anybodi driver latest site ftp ati window bio

avail price street charg card uk mathew cost sorri plus display fix driver super vga ati ultra mb ship diamond beast armenian

atheism version atheist exist god stein edu answer cs keith ve ac charley wingat mango umd contradictori imag ultb isc rit
mozumd il

version word god rather man brad keep shall said hear turkish org tree heart righteous receiv luke bless davidian ps isa
turkey armenian sdpa armenia urartu POLITICS

```

Example 3:

We use the $l$-Corr loss function to cluster the set of 11,000 documents and compare the clustering performance with the Euclidean Distance (EucD)-based NMF.

$$Loss_{EucD}(A, WH) = EucD(A, WH) = \sum_{i,j} 0.5 \left( A_{ij} - (WH)_{ij} \right)^2$$
Example 3:
- The comparison is illustrated with 20-Newsgroups data set.

- We choose number of clusters, $k$ values from 2 to 20.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-Newsgroups</td>
<td>11,000</td>
</tr>
</tbody>
</table>
Example 3 (ctd):

- Evaluation of clustering quality is via the Entropy, $e$

- Partial entropies $e_i$ of each cluster $i$ are defined as:

$$ e_i = - \sum_{j=1}^{L} p_{ij} \log_2(p_{ij}) $$

$p_{ij}$ is the probability that a member of cluster $i$ belongs to class $j$, here we use simple counts and ratios $p_{ij} = m_{ij}/m_i$

where $m_i$ is the total number of objects in cluster $i$, and $m_{ij}$ is the number of objects of class $j$ in this cluster.
Example 3 (ctd):

Entropy for the full data set is the sum of the cluster entropies weighted by the size of each cluster:

\[ e = \sum_{i=1}^{k} \frac{m_i}{m} e_i \]

where \( k \) is the total number of clusters and \( m \) is the total number of objects.
Example 3 (ctd):

- Entropy comparison table for NMF-PGD (EucD) and NMF-Corr.

- Better clustering means lower entropy. k is variable.

<table>
<thead>
<tr>
<th>Number of Clusters (k)</th>
<th>NMF-PGD (EucD)</th>
<th>NMF-Corr (σ = 1)</th>
<th>NMF-Corr (σ = 0.5)</th>
<th>NMF-Corr (σ = 0.01)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k = 2</td>
<td>3.841439</td>
<td>3.856538</td>
<td>3.850511</td>
<td>4.298769</td>
</tr>
<tr>
<td>k = 3</td>
<td>3.856632</td>
<td>3.791665</td>
<td>3.577477</td>
<td>4.269878</td>
</tr>
<tr>
<td>k = 4</td>
<td>3.779671</td>
<td>3.494205</td>
<td>3.499493</td>
<td>4.271371</td>
</tr>
<tr>
<td>k = 5</td>
<td>3.743832</td>
<td>3.603604</td>
<td>3.383459</td>
<td>4.241927</td>
</tr>
<tr>
<td>k = 6</td>
<td>3.490545</td>
<td>3.356921</td>
<td>3.301818</td>
<td>4.230222</td>
</tr>
<tr>
<td>k = 7</td>
<td>3.436018</td>
<td>3.281271</td>
<td>3.265505</td>
<td>4.201747</td>
</tr>
<tr>
<td>k = 8</td>
<td>3.303913</td>
<td>3.260363</td>
<td>2.939666</td>
<td>4.191849</td>
</tr>
<tr>
<td>k = 9</td>
<td>3.305351</td>
<td>3.336344</td>
<td>3.131309</td>
<td>4.182656</td>
</tr>
<tr>
<td>k = 10</td>
<td>3.162412</td>
<td>3.226913</td>
<td>2.935471</td>
<td>4.202036</td>
</tr>
</tbody>
</table>
Document Clustering Using NMF

- Entropy comparison between NMF-PGD (EucD) and NMF-Corr.

![Graph showing entropy comparison between NMF-PGD (EucD) and NMF-Corr for different numbers of clusters.](image)
**Document Clustering Using NMF**

### Clustering results for $k$NMF($\text{corr} (\sigma = 0.5)$)

<table>
<thead>
<tr>
<th>Cluster ID Number</th>
<th>comp.graphics</th>
<th>comp.os.ms-windows</th>
<th>comp.windows.x</th>
<th>misc.forsale</th>
<th>rec.autos</th>
<th>rec.motorcycles</th>
<th>rec.sport.baseball</th>
<th>rec.sport.hockey</th>
<th>sci.crypt</th>
<th>sci.electronics</th>
<th>sci.med</th>
<th>sci.space</th>
<th>alt.atheism</th>
<th>soc.religion.christian</th>
<th>talk.politics.guns</th>
<th>talk.politics.mideast</th>
<th>talk.politics.misc</th>
<th>talk.r eligion.misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>15</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>48</td>
<td>15</td>
<td>140</td>
<td>25</td>
<td>434</td>
<td>553</td>
<td></td>
<td>212</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>73</td>
<td>147</td>
<td>113</td>
<td>16</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>25</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>21</td>
<td>196</td>
<td>164</td>
<td>2</td>
<td>111</td>
<td>19</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>32</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>176</td>
<td>114</td>
<td>13</td>
<td>17</td>
<td>84</td>
<td>9</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>19</td>
<td>11</td>
<td>1</td>
<td>23</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>162</td>
<td>64</td>
<td>107</td>
<td>113</td>
<td>149</td>
<td>55</td>
<td>39</td>
<td>35</td>
<td>23</td>
<td>9</td>
<td>131</td>
<td>82</td>
<td>63</td>
<td>23</td>
<td>12</td>
<td>11</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>9</td>
<td>16</td>
<td>31</td>
<td>28</td>
<td>5</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>441</td>
<td>56</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>247</td>
<td>49</td>
<td>14</td>
<td>235</td>
<td>27</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>104</td>
<td>41</td>
<td>54</td>
<td>118</td>
<td>63</td>
<td>222</td>
<td>486</td>
<td>518</td>
<td>544</td>
<td>571</td>
<td>65</td>
<td>316</td>
<td>366</td>
<td>476</td>
<td>34</td>
<td>20</td>
<td>276</td>
<td>74</td>
</tr>
</tbody>
</table>
Face Recognition Using NMF

- We use ORL face data set, it has 10 poses for 40 different persons.

  - We have reduced the size of face images from 112x92 pixels to 56x46.
  - Data set has been split into two parts: Training (50%) and Testing (50%).
We use correntropy-based NMF for training and 1-Nearest neighbor for testing (recognition) phase and this is holistic recognition graph:
We test ORL face data set with **patch occlusion** from 10% to 50% square patches for each face.

- Each experiment has been repeated **10 times** and the **average of recognition rate** has been calculated.
- Occluded faces have been used for training and testing phase.
- Recognition rate graphs to follow vs. “Patch Size (from 10% to 50%)” and “Number of Basis Images”.

**Example**: % 10 patched occlusion
64 PCA Base Vectors of Occluded Faces
64 Base NMF Vectors of Occluded Faces

b) NMF basis
Occluded Face Recognition Using NMF

- Results for “k = 36” basis images:
Occluded Face Recognition Using NMF

- Results for “k = 64” basis images:

![Graph showing recognition rate vs. patch size for different methods.](image)
Occluded Face Recognition Using NMF

Results for “k = 100” basis images:

- PCA
- NMF
- NMF-PGD
- NMF-Corr(Sigma=0.5)
- NMF-Corr(Sigma=0.1)
Occluded Face Recognition Using NMF

“Recognition Rate” vs “Number of Basis Images” for 50% patch occlusion:

- Recognition Rate (%) vs Number of Basis Images

Number of Basis Images

Recognition Rate (%)

25 36 49 64 81 100 121 144

PCA  NMF  NMF-PGD  NMF-Corr(Sigma=0.5)  NMF-Corr(Sigma=0.1)
Results show that correntropy-based NMF has better recognition rate than PCA, standard NMF and Projected Gradient Descent NMF (NMF-PGD).

It means that correntropy-based NMF learns features better than the other methods tested.

The correntropy-based NMF for faces superior for Sigma=0.5, and this choice may be data-dependent.
On-going Studies

- Using “Sparse Coding” with NMF-Corr in order to obtain better face recognition rate, especially for occluded faces.

- “Blurred” face recognition rate comparisons with other methods.

- Clustering “Microarray Gene Expression Patterns” with NMF-Corr method.
Conclusion

- We can use NMF for:
  - Dimensionality Reduction
  - Clustering / Classification

- Current NMF research is focused on:
  - Algorithms enhancing interpretability
  - Alternative objective functions
  - Suitable convergence criteria
  - Updating and initializing NMF
  - Choosing ‘k’ is still open problem!
Discussion of NMF

- Lower-rank factorizations improve performance on various data mining tasks
- Interpretation of latent features of data easier
- Orthogonality restriction removed
- More efficient, compressed storage of only low-rank basis vectors and factors
- Variability of factorization criteria (sparsity!)
- But also lack of uniqueness
- Quality of factorization can be insufficient
References

[22] Greene D., Matrix Factorization Methods for Clustering Gene Expression Data, Machine Learning Group, UCD School of Computer Science and Informatics.
Thank you…

and

Questions?