Cluster Analysis Using Nonnegative Matrix Factorization

Charles D. Wessell
North Carolina State University

Mathfest
August 7, 2010
Cluster this data set

Charles D. Wessell
Cluster Analysis Using Nonnegative Matrix Factorization
Did your answer look like this?
What is a cluster?

A group of elements from a data set. Grouped elements are similar in some way, ungrouped elements are dissimilar in some way.
A group of *elements* from a data set
What is a cluster?

- A group of *elements* from a data set
- Grouped elements are *similar* in some way
What is a cluster?

- A group of *elements* from a data set

- Grouped elements are *similar* in some way

- Ungrouped elements are *dissimilar* in some way
Cluster Analysis

- For low-dimensional data sets, our eyes are excellent at clustering.
For low-dimensional data sets, our eyes are excellent at clustering.

Cluster analysis becomes much more challenging (and much more interesting) if the data set is both large and high-dimensional.
For low-dimensional data sets, our eyes are excellent at clustering.

Cluster analysis becomes much more challenging (and much more interesting) if the data set is both large and high-dimensional.

The goal of cluster analysis is to find hidden structure in a data set.
Some large, high-dimensional data sets:

<table>
<thead>
<tr>
<th>Element</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>Ratings by Netflix Customers</td>
</tr>
<tr>
<td>Netflix Customer</td>
<td>Movie Ratings</td>
</tr>
<tr>
<td>Cancer Patient</td>
<td>Gene Expression Levels</td>
</tr>
<tr>
<td>Iris Flower</td>
<td>Petal and Sepal Measurements</td>
</tr>
<tr>
<td>Voting District</td>
<td>Vote Counts for Candidates</td>
</tr>
<tr>
<td>Scotch</td>
<td>Flavor Ratings</td>
</tr>
</tbody>
</table>

Notice that in all these data sets the attributes are nonnegative.
Some of the dozens of data clustering algorithms:
- Hierarchical clustering
- $k$-means
- Self organizing maps
- Probability mixture models
- Singular value decomposition based models

Today, we will talk about a relatively new one - nonnegative matrix factorization.
Developed in the mid-1990s by Finnish researchers as Positive Matrix Factorization

Generalized as Nonnegative Matrix Factorization by Lee and Seung in a 1999 paper in *Nature*
What is Nonnegative Matrix Factorization?

- Start with a nonnegative matrix $A$ (i.e., each entry $a_{ij} \geq 0$)

- Goal: Find nonnegative matrices $W$ and $H$ such that $A \approx WH$

- If $A$ is $m \times n$, we desire $W$ to be $m \times k$ and $H$ to be $k \times n$, where $k \ll \min(m, n)$.

- In short: $A_{m \times n} \approx W_{m \times k}H_{k \times n}$
### NMF explained with a real data set

**A** is a $6 \times 8$ matrix containing the number of votes cast for each 2004 and 2008 presidential candidate in the eight states that begin with the letter *M*.

$$
A = \begin{pmatrix}
\text{ME} & \text{MD} & \text{MA} & \text{MI} & \text{MN} & \text{MS} & \text{MO} & \text{MT} \\
\text{Bush-2004} & 330201 & 1024703 & 1071109 & 2313746 & 1346695 & 684981 & 1455713 & 266063 \\
\text{Kerry-2004} & 396842 & 1334493 & 1803800 & 2479183 & 1445014 & 458094 & 1259171 & 173710 \\
\text{Others-2004} & 13709 & 27482 & 37479 & 46323 & 36678 & 9290 & 16480 & 10672 \\
\text{Obama-2008} & 421923 & 1629467 & 1904098 & 2872579 & 1573354 & 554662 & 1441911 & 232159 \\
\text{McCain-2008} & 295273 & 959862 & 1108854 & 2048639 & 1275409 & 724597 & 1445814 & 243882 \\
\text{Others-2008} & 13967 & 42267 & 68117 & 88976 & 61606 & 10606 & 41224 & 16709 
\end{pmatrix}
$$
What does each column of \( W \) represent?
How NMF helps us cluster this data set

\[ A_{6 \times 8} \approx W_{6 \times 3} H_{3 \times 8} = \]

<table>
<thead>
<tr>
<th></th>
<th>&quot;State1&quot;</th>
<th>&quot;State2&quot;</th>
<th>&quot;State3&quot;</th>
<th>ME</th>
<th>MD</th>
<th>MA</th>
<th>MI</th>
<th>MN</th>
<th>MS</th>
<th>MO</th>
<th>MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bush04</td>
<td>2896700</td>
<td>0</td>
<td>836300</td>
<td>0.0868</td>
<td>0.3651</td>
<td>0.1957</td>
<td>0.7488</td>
<td>0.3301</td>
<td>0.1466</td>
<td>0.3486</td>
<td>0.0900</td>
</tr>
<tr>
<td>Kerry04</td>
<td>2244300</td>
<td>1743100</td>
<td>168600</td>
<td>0.1015</td>
<td>0.3168</td>
<td>0.7261</td>
<td>0.4317</td>
<td>0.3494</td>
<td>0.0527</td>
<td>0.2258</td>
<td>0.0000</td>
</tr>
<tr>
<td>Others04</td>
<td>39000</td>
<td>43900</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obama08</td>
<td>2802100</td>
<td>1796600</td>
<td>88100</td>
<td>0.0824</td>
<td>0.0000</td>
<td>0.6035</td>
<td>0.1546</td>
<td>0.4460</td>
<td>0.3320</td>
<td>0.5439</td>
<td>0.0110</td>
</tr>
<tr>
<td>McCain08</td>
<td>2546100</td>
<td>0</td>
<td>1011000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others08</td>
<td>75900</td>
<td>69100</td>
<td>6100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The columns of \( W \) represent three *meta-states*, while the columns of \( H \) contain the coordinates for each real state in this *meta-state* space.
A \approx WH =

\[
\begin{bmatrix}
2896700 & 0 & 836300 \\
2244300 & 1743100 & 168600 \\
39000 & 43900 & 0 \\
2802100 & 1796600 & 88100 \\
2546100 & 0 & 1011000 \\
75900 & 69100 & 6100
\end{bmatrix}
\begin{bmatrix}
.0868 & .1015 & .0824 \\
.3651 & .3168 & .0000 \\
.1957 & .7261 & .6035 \\
.7488 & .4317 & .1546 \\
.3301 & .3494 & .1546 \\
.1466 & .0527 & .4460 \\
.3486 & .2258 & .3320 \\
.0900 & .0000 & .5439 \\
.0868 & .1015 & .0824
\end{bmatrix}
\]

In other words, each state can be expressed as a linear combination of meta-states. For example,

Maine = .0868 \times "State1" + .1015 \times "State2" + .0824 \times "State3"
How NMF helps us cluster this data set

\[ A \approx W_{6 \times 3} H_{3 \times 8} = \]

Each state is assigned a cluster depending on which coordinate is largest. In this example, the clustering would be \{Maryland, Michigan, Montana\}, \{Maine, Massachusetts\}, \{Minnesota, Mississippi, Missouri\}. 
And now the bad news

Do you see a problem with the way we used values in $H$ to cluster the states?

$H = \begin{pmatrix}
.0868 & .3651 & .1957 & .7488 & .3301 & .1466 & .3486 & .0900 \\
.1015 & .3168 & .7261 & .4317 & .3494 & .0527 & .2258 & .0000 \\
.0824 & .0000 & .6035 & .1546 & .4460 & .3320 & .5439 & .0110
\end{pmatrix}$
And now the bad news

\[
H = \begin{pmatrix}
.0868 & .3651 & .1957 & .7488 & .3301 & .1466 & .3486 & .0900 \\
.1015 & .3168 & .7261 & .4317 & .3494 & .0527 & .2258 & .0000 \\
.0824 & .0000 & .6035 & .1546 & .4460 & .3320 & .5439 & .0110 
\end{pmatrix}
\]

- Do you see a problem with the way we used values in \( H \) to cluster the states?

- Also, \( W \) and \( H \) are not unique. So different runs of NMF can lead to different clusters!
Is it really bad news?

- Is there a way around these two "problems"?
Is it really bad news?

- Is there a way around these two "problems"?

- Perhaps, gathering multiple results from NMF clustering can provide us with a better solution than any single clustering.