Cluster analysis

Based on H.C. Romesburg: *Cluster analysis for researchers*,
Lifetime Learning Publications, Belmont, CA, 1984
P.H.A. Sneath and R.R. Sokal: *Numerical Taxonomy*, Freeman,
San Francisco, CA, 1973

Jens C. Frisvad
BioCentrum-DTU

Biological data analysis and chemometrics
Two primary methods

• Cluster analysis (no projection)
  – Hierarchical clustering
  – Divisive clustering
  – Fuzzy clustering

• Ordination (projection)
  – Principal component analysis
  – Correspondence analysis
  – Multidimensional scaling
Advantages of cluster analysis

- Good for a quick overview of data
- Good if there are many groups in data
- Good if unusual similarity measures are needed
- Can be added on ordination plots (often as a minimum spanning tree, however)
- Good for the nearest neighbours, ordination better for the deeper relationships
Different clustering methods

- NCLAS: Agglomerative clustering by distance optimization
- HMCL: Agglomerative clustering by homogeneity optimization
- INFCL: Agglomerative clustering by information theory criteria
- MINGFC: Agglomerative clustering by global optimization
- ASSIN: Divisive monothetic clustering
- PARREL: Partitioning by global optimization
- FCM: Fuzzy c-means clustering
- MINSPAN: Minimum spanning tree
- REBLOCK: Block clustering (k-means clustering)
SAHN clustering

• Sequential agglomerative hierarchic nonoverlapping clustering
Single linkage

- Nearest neighbor, minimum method
- Close to minimum spanning tree
- Contracting space
- Chaining possible
- $\alpha_J = 0.5$, $\alpha_K = 0.5$, $\beta = 0$, $\gamma = -0.5$
- $U_{J,K} = \min U_{jk}$

$$U_{(J,K)L} = \alpha_J U_{J,L} + \alpha_K U_{K,L} + \beta U_{J,K} + \gamma |U_{J,L} - U_{K,L}|$$
FIGURE 5-3
Single linkage clustering of the data in Table 5-1. For explanation, see text.
Complete linkage

- Furthest neighbor, maximum method
- Dilating space
- $\alpha_J = 0.5, \alpha_K = 0.5, \beta = 0, \gamma = 0.5$
- $U_{J,K} = \max U_{jk}$
Step 1
Level of cohesion
(maximum link), $\Delta = 1$

Step 2
Level of cohesion
(maximum link), $\Delta \leq \sqrt{5}$

Step 3
Level of cohesion
(maximum link), $\Delta \leq \sqrt{13}$
Step 4
Levels of cohesion (maximum links)
Left cluster, \( \Delta \leq 5 \)
Right cluster, \( \Delta \leq \sqrt{26} \)

Step 5
Level of cohesion (maximum link), \( \Delta \leq \sqrt{73} \)

Phenogram A

Phenogram B
Average linkage

• Arithmetic average
  – Unweighted: UPGMA (group average)
  – Weighted: WPGMA

• Centroid
  – Unweighted centroid (Centroid)
  – Weighted centroid (Median)
FIGURE 5-11
The effects of several clustering methods (for explanation see text) on the criterion for admitting L (containing one OTU) to the cluster formed of four OTU's in J plus one in K. OTU's are indicated by solid circles. Abscissa is distance $\Delta$.
Ordinary clustering

- Obtain the data matrix
- Transform or standardize the data matrix
- Select the best resemblance or distance measure
- Compute the resemblance matrix
- Execute the clustering method (often UPGMA = average linkage)
- Rearrange the data and resemblance matrices
- Compute the cophenetic correlation coefficient
## Binary similarity coefficients
(between two objects i and j)

<table>
<thead>
<tr>
<th></th>
<th>j</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>
Matches and mismatches

- $m = a + b$ (number of matches)
- $u = c + d$ (number of mismatches)
- $n = m + u = a + b + c + d$ (total sample size)

- Similarity (often 0 to 1)
- Dissimilarity (distance) (often 0 to 1)
- Correlation (-1 to 1)
Simple matching coefficient

- $\text{SM} = \frac{a + d}{a + b + c + d} = \frac{m}{n}$

- Euclidean distance for binary data:
  - $D = 1 - \text{SM} = \frac{b + c}{a + b + c + d} = \frac{u}{n}$
Avoiding zero zero comparisons

- Jaccard = $J = \frac{a}{a + b + c}$

- Sørensen or Dice: $\text{DICE} = \frac{2a}{2a + b + c}$
Correlation coefficients

Yule: \( (ad - bc) / (ad + bc) \)

\[
PHI = \frac{(ad - bc)}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}
\]
Other binary coefficients

- Hamann = $H = (a + d - b - c) / (a + b + c + d)$
- Rogers and Tanimoto = $RT = (a + d) / (a + 2b + 2c + d)$
- Russel and Rao = $RR = a / (a + b + c + d)$
- Kulzynski 1 = $K1 = a / (b + c)$
- UN1 = $(2a + 2d) / (2a + b + c + 2d)$
- UN2 = $a / (a + 2b + 2c)$
- UN3 = $(a + d) / (b + c)$
Distances for quantitative (interval) data
Euclidean and taxonomic distance

\[ EUCLID = E_{ij} = \sqrt{\sum_k (x_{ki} + x_{kj})^2} \]

\[ DIST = d_{ij} = \sqrt{\frac{1}{n} \sum_k (x_{ki} + x_{kj})^2} \]
Bray-Curtis and Canberra distance

\[ \text{BRAYCURT} = d_{ij} = \sum_k |x_{ki} - x_{kj}| / \sum_k (x_{ki} + x_{kj}) \]

\[ \text{CANBERRA} = \frac{1}{n} \sum_k |x_{ki} - x_{kj}| / \sum_k (x_{ki} + x_{kj}) \]
Average Manhattan distance (city block)

\[ \text{MANHAT} = M_{ij} = \frac{1}{n} \sum_{k} |x_{ki} - x_{kj}| \]
Chi-squared distance

\[ CHISQ = d_{ij} = \sqrt{\sum_k \left( \frac{x_{ki} - x_{kj}}{x_{.i}/x_{.j}} \right)^2} \]
Cosine coefficient

\[ \text{COSINE} = c_{ij} = \frac{\sum_k x_{ki} x_{kj}}{\sqrt{\sum_k x_{ki}^2 \sum_k x_{kj}^2}} \]
Step 1. Obtain the data matrix

<table>
<thead>
<tr>
<th>Feature</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>20</td>
<td>10</td>
<td>15</td>
<td>10</td>
</tr>
</tbody>
</table>
Objects and features

• The five objects are plots of farm land
• The features are
  – 1. Water-holding capacity (%)
  – 2. Weight % soil organic matter
• Objective: find the two most similar plots
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>18.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>20.6</td>
<td>14.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>22.4</td>
<td>11.2</td>
<td>5.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>7.07</td>
<td>18.0</td>
<td>25.0</td>
<td>25.5</td>
<td>-</td>
</tr>
</tbody>
</table>
### Revised resemblance matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>(34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>18.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>7.07</td>
<td>18.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(34)</td>
<td>21.5</td>
<td>12.7</td>
<td>25.3</td>
<td>-</td>
</tr>
</tbody>
</table>
Revised resemblance matrix

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>(34)</th>
<th>(15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(34)</td>
<td>12.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(15)</td>
<td>18.0</td>
<td>23.4</td>
<td>-</td>
</tr>
</tbody>
</table>
Rvised resemblance matrix

<table>
<thead>
<tr>
<th></th>
<th>(15)</th>
<th>(234)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(15)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(234)</td>
<td>21.6</td>
<td>-</td>
</tr>
</tbody>
</table>
Attribute space

Cluster (12345)

Organic matter (percent of soil by weight)

Water-holding capacity (cm$^3$ of H$_2$O/cm$^3$ of soil)
### Rearranged data matrix

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Object 1</th>
<th>Object 2</th>
<th>Object 3</th>
<th>Object 4</th>
<th>Object 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

### Rearranged resemblance matrix

<table>
<thead>
<tr>
<th></th>
<th>Object 1</th>
<th>Object 2</th>
<th>Object 3</th>
<th>Object 4</th>
<th>Object 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td></td>
<td></td>
<td>14.1</td>
<td>11.2</td>
<td></td>
</tr>
<tr>
<td>Object 2</td>
<td>20.6</td>
<td>22.4</td>
<td>18.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object 3</td>
<td>25.0</td>
<td>25.5</td>
<td>18.0</td>
<td>7.07</td>
<td></td>
</tr>
</tbody>
</table>
Cophenetic correlation coefficient
(Pearson product-moment correlation coefficient)

- A comparison of the similarities according to the similarity matrix and the similarities according to the dendrogram

\[
r_{X,Y} = \frac{\sum xy - (1/n)(\sum x)(\sum y)}{\sqrt{(\sum x^2 - (1/n)(\sum x)^2)(\sum y^2 - (1/n)(\sum y)^2)}}
\]
NTSYS

- Import matrix
-Transpose matrix if objects are rows (they are supposed to be columns in NTSYS) (transp in transformation / general)
- Consider log1 or autoscaling (standardization)
- Select similarity or distance measure (similarity)
- Produce similarity matrix
NTSYS (continued)

• Select clustering procedure (often UPGMA) (clustering)
• Calculate cophenetic matrix (clustering)
• Compare similarity matrix with cophenetic matrix (made from the dendrogram) and write down the cophenetic correlation (graphics, matrix comparison)
• Write dendrogram (graphics, treeplot)