Appendix: Cluster Analysis

More information can be found in Everitt, Landau & Leese, 2001. ***Custer Analysis***. Oxford University Press

Similarity and Dissimilarity Metrics

- The similarity metric has to satisfy the properties:

  Symmetry: $s_{ij} = s_{ji}$ for all $i, j \in \{1,2, \dots, n\}$

  Ordering: $s_{ij} \leq s_{ii}$ similarity of itself larger than else

  Self-similarity: $s_{ii} = s_{jj}$

  - Notes:
    - Correlations can be considered similarities on the scale $-1 \leq r_{ij} \leq 1$
    - The topological spatial adjacency can be considered a similarity measure.
    - Frequently it is assumed that the similarity $s_{ij} \geq 0$.
    - There is no upper limit for the similarity unless the assumption is made that $s_{11} = \cdots = s_{nn} = 1$.
    - The similarities among all $n$objects can be pooled into a similarity matrix $\mathbf{S}_{n\times n}$.
- The dissimilarity metric has to satisfy the properties

$$\text{Symmetry: } d_{ij} = d_{ji} \text{ for all } i, j \in \{1,2,\dots,n\}$$
$$\text{Ordering: } d_{ij} \geq 0 \; and \; d_{11} = \cdots = d_{nn} = 0$$

- Notes:
  - If a dissimilarity metric also satisfies the triangle equation $d_{ij} \leq d_{ik} + d_{jk}$ then it has a geometric interpretation as a distance measure.
  - A dissimilarity metric has a natural lower bound of zero.
  - The dissimilarities among all $n$objects can be pooled into a dissimilarity matrix $\mathbf{D}_{n\times n}$.

- Possible transformations between both metrics:
  - An inverse relationship between both metrics exists $d_{ij} \approx \frac{1}{s_{ij}}$
  - For $0 \leq s_{ij} \leq 1$ the transformation becomes: $d_{ij} = 1 - s_{ij}$
  - For $-1 \leq s_{ij} \leq 1$ the transformation becomes $d_{ij} = \frac{1}{2} \cdot \left(1 - s_{ij}\right)$
  - $s_{ij} = 1 - \frac{d_{ij}}{\max(d_{ij})}$

Nominal Scaled Variables

- Objects with ***nominal*** scaled features:

- Nominal scaled features with K factor levels must be encoded [a] using either the one-hot coding with K binary variables or the dummy coding with $K-1$ binary variables (see Boehmke et al. pp 58-66). p should be k here
- Let $\mathbf{x}_i = (b_{i1}, \ldots, b_{iP})^T$ and $\mathbf{x}_j = \left(b_{j1}, \ldots, b_{jP}\right)^T$ where each object is characterized by P binary features $b_{ip} = \begin{cases} 1 & \text{if feature } p \text{ is present} \\ 0 & \text{otherwise} \end{cases}$ and analogue for b_{jp}
- Both feature vectors can be organized in a contingency table counting the concordant and discordant pairs of features:

	1	0	
1	c_{11}	c_{10}	$c_{11} + c_{10} = c_{1+}$
0	c_{01}	c_{00}	$c_{01} + c_{00} = c_{0+}$
	$c_{11} + c_{01} = c_{+1}$	$c_{10} + c_{00} = c_{+0}$	P

- The number of concordant pairs is $c_{11} + c_{00}$ and the number of discordant pairs is $c_{10} + c_{01}$
- Several similarity and dissimilarity metrics can be derived from this table. See **`help(dist)`** and the options **`binary`**:

```
> x <- c(0, 0, 1, 1, 1, 1)
> y <- c(1, 0, 1, 1, 0, 1)
> dist(rbind(x, y), method = "binary")
```

x
y 0.4

This is a ***dissimilarity*** measure because $d_{ij} = \frac{c_{01}+c_{10}}{c_{01}+c_{10}+c_{11}}$. Here missing feature pairs c_{00} are not counted because neither feature is at all present.

- A generic error rate similarity metric is

$$s_{ij=} = \frac{\alpha \cdot (c_{11} + c_{00})}{\alpha \cdot (c_{11} + c_{00}) + (1 - \alpha) \cdot (c_{10} + c_{01})}$$

  where $0 < \alpha < 1$ weights concordant and discordant pairs differently for $\alpha = 0.5$ we obtain the concordance rate $s_{ij} = (c_{11} + c_{00})/P$.

- A measure analog to a correlation coefficient between the pairs $\mathbf{x}_i$ and $\mathbf{x}_j$ can be obtained by

$$s_{ij} = \frac{c_{11} \cdot c_{00} - c_{10} \cdot c_{01}}{\sqrt{c_{1+} \cdot c_{0+} \cdot c_{+1} \cdot c_{+0}}}$$

Ordinal Scaled Variables

- For ***ordinal*** scaled features p similar coefficients can be calculated by recognizing the ranking of features includes lower ranked features. Let $p_1 < p_2 < p_3$ then

$$p_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\ p_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \text{ and } p_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Metric Variables

- For metrically scaled variables the generic ***Minkowski* dissimilarity** metric is

$$d_{ij}^{[q]} = \left(\sum_{p=1}^{P} \left|x_{ip} - x_{jp}\right|^q\right)^{1/q}$$

  - For $q = 1$ the Manhattan (city block) distance is obtained
  - For $q = 2$ the Euclidian distance is obtained
  - For $q = \infty$ the maximum distance $\max_p \left|x_{ip} - x_{jp}\right|$ is obtained.
  - Only the Euclidian distance are rotation invariant.
  - The squared Euclidian distance d_{ij}^2 is closely related to variance around a centroid, such the within class heterogeneity measure.
- The ***Mahalanobis* measures** the Euclidian distance between objects based on their uniformly scaled principle component scores.

$$d_{ij}^{[M]} = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \cdot \mathbf{S}^{-1} \cdot \left(\mathbf{x}_i - \mathbf{x}_j\right)$$

with $\mathbf{S}^{-1}$ being the inverse covariance matrix an $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iP})^T$ being the vector of the P features for the i^{th} observation and analogue for the j^{th} observation, or

$$d_{ij}^{[M]} = \left(\mathbf{z}_i - \mathbf{z}_j\right)^T \cdot \mathbf{R}^{-1} \cdot \left(\mathbf{z}_i - \mathbf{z}_j\right)$$

where $\mathbf{z}_i$ and $\mathbf{z}_j$ are based on z-transformed variables with $\mathbf{z}_i = (z(x_{i1}), z(x_{i2}), \dots, z(x_{iP}))^T$ and $\mathbf{R}^{-1}$ is the inverse correlation matrix among all variables (see script **`Mahalanobis.R`**).

- Notes:
  - It standardized the variance of each variable.
  - It controls for the correlation among the features and removes data redundancy.
  - The Mahalanobis distance is based on all observations. Therefore, within a cluster the features may very well be correlated.

Mixture of Measurement Levels

- For objects based on a ***mixture*** of feature types, a similarity metric based on a linear combination can be derived

$$s_{ij} = \frac{|p^{nominal}| \cdot s_{ij}^{nominal} + |p^{ordinal}| \cdot s_{ij}^{ordinal} + |p^{metric}| \cdot s_{ij}^{metric}}{P}$$

with $P = |p^{nominal}| + |p^{ordinal}| + |p^{metric}|$ where $|\quad|$ is the number of features contributing to each similarity measure.

- The Gower option in **`cluster::daisy( , metric='gower')`** calculates a mixture dissimilarity measure for metric and nominal scaled features:
  - For nominal scaled features it reports discordant pairs as 1 and concordant pairs as zero.

  - For metric or ordinal scaled features it reports the absolute difference for each feature and then it scales each d_{ij}^{p} for the feature p to d_{ij}^{p} in 0 to 1.
  - Finally, it adds the distances of all features together either as a mean or weighted mean:

$$\mathbf{D} = \frac{\sum_{p=1}^{P} w_p \cdot \mathbf{D}_p}{\sum_{p=1}^{P} w_p}$$

- Notes:
  - Hierarchical cluster analysis has the tendency to group objects with similar levels of nominal scaled features together.
  - Principal component analysis and k nearest neighbors, kMeans are not well tailored using binary features. For instance, performing a *z*-transformation on binary features does not dissolve the binning around just two values.

Generic Hierarchically Cluster Analysis Equation

- If two classes $C_k \cup C_l$ are merged then the heterogeneity increase to any of the remaining classes C_s can be calculated ***recursively*** by

$$d(C_k \cup C_l, C_s) = \alpha_k \cdot d(C_k, C_s) + \alpha_l \cdot d(C_l, C_s) + \beta \cdot d(C_k, C_l) + \gamma \cdot |d(C_k, C_s) - d(C_l, C_s)|$$

- Depending on the selected parameters values $\alpha_k, \alpha_l, \beta$ and γ the different heterogeneity update methods can be derived.

- Only if $\alpha_k, \alpha_l \geq 0$ and $\alpha_k + \alpha_l + \beta \geq 1$ as well as $|\gamma| \leq \alpha_k, \alpha_l$ then the heterogeneity of the clustering method is monotonically increasing and does not exhibit inversions.

- Parameters of recursive agglomerative cluster algorithms

Linkage Methods	α_k	α_l	β	γ
Single	$1/2$	$1/2$	0	$-1/2$
Complete	$1/2$	$1/2$	0	$1/2$
Average	$n_k/(n_k+n_s)$	$n_l/(n_l+n_s)$	0	0
Centroid	$n_k/(n_k+n_s)$	$n_l/(n_l+n_s)$	$-\frac{n_k \cdot n_l}{(n_k+n_l)^2}$	0
Ward	$\frac{n_k+n_s}{n_k+n_l+n_s}$	$\frac{n_l+n_s}{n_k+n_l+n_s}$	$-\frac{n_s}{n_k+n_l+n_s}$	0
Median	$1/2$	$1/2$	$-1/4$	0
Flexible strategy	$\varphi > 0$	φ	$1-2\cdot\varphi$	0