CLUSTER

Cluster Measures

Measures for Continuous Data

EUCLID

The distance between two items, *x* and *y*, is the square root of the sum of the squared differences between the values for the items.

$$\text{EUCLID}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

SEUCLID

The distance between two items is the sum of the squared differences between the values for the items.

$$\text{SEUCLID}(x, y) = \sum_i (x_i - y_i)^2$$

CORRELATION

This is a pattern similarity measure.

$$\text{CORRELATION}(x, y) = \frac{\sum_i (Z_{xi} Z_{yi})}{N}$$

where Z_{xi} is the (standardized) *Z*-score value of *x* for the *i*th case or variable, and *N* is the number of cases or variables.

COSINE

This is a pattern similarity measure.

$$\text{COSINE}(x, y) = \frac{\sum_i (x_i y_i)}{\sqrt{\left(\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)\right)}}$$

CHEBYCHEV

The distance between two items is the maximum absolute difference between the values for the items.

$$\text{CHEBYCHEV}(x, y) = \max_i |x_i - y_i|$$

BLOCK

The distance between two items is the sum of the absolute differences between the values for the items.

$$\text{BLOCK}(x, y) = \sum_i |x_i - y_i|$$

MINKOWSKI(*p*)

The distance between two items is the *p*th root of the sum of the absolute differences to the *p*th power between the values for the items.

$$\text{MINKOWSKI}(x, y) = \left(\sum_i |x_i - y_i|^p\right)^{1/p}$$

POWER (p, r)

The distance between two items is the *r*th root of the sum of the absolute differences to the *p*th power between the values for the items.

$$\text{POWER}(x,y) = \left(\sum_i |x_i - y_i|^p \right)^{1/r}$$

Measures for Frequency Count Data

CHISQ

The magnitude of this similarity measure depends on the total frequencies of the two cases or variables whose proximity is computed. Expected values are from the model of independence of cases (or variables), *x* and *y*.

$$\text{CHISQ}(x,y) = \sqrt{\sum_i \frac{\left(x_i - E(x_i)\right)^2}{E(x_i)} + \sum_i \frac{\left(y_i - E(y_i)\right)^2}{E(y_i)}}$$

PH2

This is the CHISQ measure normalized by the square root of the combined frequency. Therefore, its value does not depend on the total frequencies of the two cases or variables whose proximity is computed.

$$\text{PH2}(x,y) = \frac{\text{CHISQ}(x,y)}{\sqrt{N}}$$

Measures for Binary Data

PROXIMITIES constructs a 2×2 contingency table for each pair of items in turn. It uses this table to compute a proximity measure for the pair.

Item 2

		Present	Absent
Item 1	Present	*a*	*b*
	Absent	*c*	*d*

PROXIMITIES computes all binary measures from the values of *a*, *b*, *c*, and *d*. These values are tallies across variables (when the items are cases) or tallies across cases (when the items are variables).

Russel and Rao Similarity Measure

This is the binary dot product.

$$\mathrm{RR}(x, y) = \frac{a}{a+b+c+d}$$

Simple Matching Similarity Measure

This is the ratio of the number of matches to the total number of characteristics.

$$\mathrm{SM}(x, y) = \frac{a+d}{a+b+c+d}$$

Jaccard Similarity Measure

This is also known as the similarity ratio.

$$\mathrm{JACCARD}(x, y) = \frac{a}{a+b+c}$$

Dice or Czekanowski or Sorenson Similarity Measure

$$\mathrm{DICE}(x, y) = \frac{2a}{2a+b+c}$$

Sokal and Sneath Similarity Measure 1

$$\mathrm{SS1}(x, y) = \frac{2(a+d)}{2(a+d)+b+c}$$

Rogers and Tanimoto Similarity Measure

$$\mathrm{RT}(x, y) = \frac{a+d}{a+d+2(b+c)}$$

Sokal and Sneath Similarity Measure 2

$$\mathrm{SS2}(x, y) = \frac{a}{a+2(b+c)}$$

Kulczynski Similarity Measure 1

This measure has a minimum value of 0 and no upper limit. It is undefined when there are no nonmatches $(b = 0 \text{ and } c = 0)$. Therefore, PROXIMITIES assigns an artificial upper limit of 9999.999 to K1 when it is undefined or exceeds this value.

$$\mathrm{K1}(x, y) = \frac{a}{b+c}$$

Sokal and Sneath Similarity Measure 3

This measure has a minimum value of 0, has no upper limit, and is undefined when there are no nonmatches $(b = 0 \text{ and } c = 0)$. As with K1, PROXIMITIES assigns an artificial upper limit of 9999.999 to SS3 when it is undefined or exceeds this value.

$$\mathrm{SS3}(x, y) = \frac{a+d}{b+c}$$

Conditional Probabilities

The following three binary measures yield values that you can interpret in terms of conditional probability. All three are similarity measures.

Kulczynski Similarity Measure 2

This yields the average conditional probability that a characteristic is present in one item given that the characteristic is present in the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1.

$$\mathrm{K2}(x, y) = \frac{a/(a+b) + a/(a+c)}{2}$$

Sokal and Sneath Similarity Measure 4

This yields the conditional probability that a characteristic of one item is in the same state (present or absent) as the characteristic of the other item. The measure is an average over both items acting as predictors. It has a range of 0 to 1.

$$\mathrm{SS4}(x, y) = \frac{a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)}{4}$$

Hamann Similarity Measure

This measure gives the probability that a characteristic has the same state in both items (present in both or absent from both) minus the probability that a characteristic has different states in the two items (present in one and absent from the other). HAMANN has a range of –1 to +1 and is monotonically related to SM, SS1, and RT.

$$\mathrm{HAMANN}(x, y) = \frac{(a+d) - (b+c)}{a+b+c+d}$$

Predictability Measures

The following four binary measures assess the association between items as the predictability of one given the other. All four measures yield similarities.

Goodman and Kruskal Lambda (Similarity)

This coefficient assesses the predictability of the state of a characteristic on one item (presence or absence) given the state on the other item. Specifically, lambda measures the proportional reduction in error using one item to predict the other, when the directions of prediction are of equal importance. Lambda has a range of 0 to 1.

$$t_1 = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$$
$$t_2 = \max(a + c, b + d) + \max(a + b, c + d)$$
$$\text{LAMBDA}(x, y) = \frac{t_1 - t_2}{2(a + b + c + d) - t_2}$$

Anderberg's *D* (Similarity)

This coefficient assesses the predictability of the state of a characteristic on one item (presence or absence) given the state on the other. *D* measures the actual reduction in the error probability when one item is used to predict the other. The range of *D* is 0 to 1.

$$t_1 = \max(a, b) + \max(c, d) + \max(a, c) + \max(b, d)$$
$$t_2 = \max(a + c, b + d) + \max(a + b, c + d)$$
$$\text{D}(x, y) = \frac{t_1 - t_2}{2(a + b + c + d)}$$

Yule's *Y* Coefficient of Colligation (Similarity)

This is a function of the cross-product ratio for a 2×2 table. It has a range of –1 to +1.

$$\text{Y}(x, y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Yule's *Q* (Similarity)

This is the 2×2 version of Goodman and Kruskal's ordinal measure *gamma*. Like Yule's *Y*, *Q* is a function of the cross-product ratio for a 2×2 table and has a range of –1 to +1.

$$\mathrm{Q}(x, y) = \frac{ad - bc}{ad + bc}$$

Other Binary Measures

The remaining binary measures available in PROXIMITIES are either binary equivalents of association measures for continuous variables or measures of special properties of the relation between items.

Ochiai Similarity Measure

This is the binary form of the cosine. It has a range of 0 to 1 and is a similarity measure.

$$\mathrm{OCHIAI}(x, y) = \sqrt{\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)}$$

Sokal and Sneath Similarity Measure 5

This is a similarity measure. Its range is 0 to 1.

$$\mathrm{SS5}(x, y) = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Fourfold Point Correlation (Similarity)

This is the binary form of the Pearson product-moment correlation coefficient. Phi is a similarity measure, and its range is 0 to 1.

$$\mathrm{PHI}(x, y) = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Binary Euclidean Distance

This is a distance measure. Its minimum value is 0, and it has no upper limit.

$$\text{BEUCLID}(x, y) = \sqrt{b + c}$$

Binary Squared Euclidean Distance

This is also a distance measure. Its minimum value is 0, and it has no upper limit.

$$\text{BSEUCLID}(x, y) = b + c$$

Size Difference

This is a dissimilarity measure with a minimum value of 0 and no upper limit.

$$\text{SIZE}(x, y) = \frac{(b - c)^2}{(a + b + c + d)^2}$$

Pattern Difference

This is also a dissimilarity measure. Its range is 0 to 1.

$$\text{PATTERN}(x, y) = \frac{bc}{(a + b + c + d)^2}$$

Binary Shape Difference

This dissimilarity measure has no upper or lower limit.

$$\text{BSHAPE}(x, y) = \frac{(a + b + c + d)(b + c) - (b - c)^2}{(a + b + c + d)^2}$$

Dispersion Similarity Measure

This similarity measure has a range of –1 to +1.

$$\text{DISPER}(x,y) = \frac{ad - bc}{(a+b+c+d)^2}$$

Variance Dissimilarity Measure

This dissimilarity measure has a minimum value of 0 and no upper limit.

$$\text{VARIANCE}(x,y) = \frac{b+c}{4(a+b+c+d)}$$

Binary Lance-and-Williams Nonmetric Dissimilarity Measure

Also known as the Bray-Curtis nonmetric coefficient, this dissimilarity measure has a range of 0 to 1.

$$\text{BLWMN}(x,y) = \frac{b+c}{2a+b+c}$$

Clustering Methods

Notation

The following notation is used unless otherwise specified:

S	Matrix of similarity or dissimilarity measures
s_{ij}	Similarity or dissimilarity measure between cluster i and cluster j
N_i	Number of cases in cluster i

General Procedure

Begin with *N* clusters each containing one case. Denote the clusters 1 through *N*.

- Find the most similar pair of clusters p and q $(p > q)$. Denote this similarity s_{pq}. If a dissimilarity measure is used, large values indicate dissimilarity. If a similarity measure is used, small values indicate dissimilarity.

- Reduce the number of clusters by one through merger of clusters p and q. Label the new cluster t $(= q)$ and update similarity matrix (by the method specified) to reflect revised similarities or dissimilarities between cluster t and all other clusters. Delete the row and column of S corresponding to cluster p.

- Perform the previous two steps until all entities are in one cluster.

- For each of the following methods, the similarity or dissimilarity matrix S is updated to reflect revised similarities or dissimilarities (s_{tr}) between the new cluster t and all other clusters r as given below.

Average Linkage between Groups

Before the first merge, let $N_i = 1$ for $i = 1$ to N. Update s_{tr} by

$$s_{tr} = s_{pr} + s_{qr}$$

Update N_t by

$$N_t = N_p + N_q$$

and then choose the most similar pair based on the value

$$s_{ij} \Big/ \left(N_i N_j \right)$$

Average Linkage within Groups

Before the first merge, let $SUM_i = 0$ and $N_i = 1$ for $i = 1$ to N. Update s_{tr} by

$$s_{tr} = s_{pr} + s_{qr}$$

Update SUM_t and N_t by

$$SUM_t = SUM_p + SUM_q + s_{pq}$$

$$N_t = N_p + N_q$$

and choose the most similar pair based on

$$\frac{SUM_i + SUM_j + s_{ij}}{\left(\left(N_i + N_j\right)\left(N_i + N_j - 1\right)\right)/2}$$

Single Linkage

Update s_{tr} by

$$s_{tr} = \begin{cases} \min\left(s_{pr}, s_{qr}\right) & \text{if } S \text{ is a dissimilarity matrix} \\ \max\left(s_{pr}, s_{qr}\right) & \text{if } S \text{ is a similarity matrix} \end{cases}$$

Complete Linkage

Update s_{tr} by

$$s_{tr} = \begin{cases} \max\left(s_{pr}, s_{qr}\right) & \text{if } S \text{ is a dissimilarity matrix} \\ \min\left(s_{pr}, s_{qr}\right) & \text{if } S \text{ is a similarity matrix} \end{cases}$$

Centroid Method

Update s_{tr} by

$$s_{tr} = \frac{N_p}{N_p + N_q} s_{pr} + \frac{N_q}{N_p + N_q} s_{qr} - \frac{N_p N_q}{\left(N_p + N_q\right)^2} s_{pq}$$

Median Method

Update s_{tr} by

$$s_{tr} = \left(s_{pr} + s_{qr}\right)/2 - s_{pq}/4$$

Ward's Method

Update s_{tr} by

$$s_{tr} = \frac{1}{\left(N_t + N_r\right)}\left[\left(N_r + N_p\right)s_{rp} + \left(N_r + N_q\right)s_{rq} - N_r s_{pq}\right]$$

Update the coefficient W by

$$W = W + .5 s_{pq}$$

Note that for Ward's method, the coefficient given in the agglomeration schedule is really the within-cluster sum of squares at that step. For all other methods, this coefficient represents the distance at which the clusters *p* and *q* were joined.

References

Anderberg, M. R. 1973. *Cluster analysis for applications*. New York: Academic Press.