Measures of similarity

To identify clusters of observations we need to know how **close individuals are to each other** or **how far apart they are**.

Two individuals are 'close' when their dissimilarity of distance is small and their similarity large.

Special attention will be paid to proximity measures suitable for data consisting of repeated measures of the same variable, for example taken at different time points.

Similarity Measures for Categorical Data

Measures are generally scaled to be in the interval $[0,1]$, although occasionally they are expressed as percentages in the range $0-100\%$

Similarity value of unity indicates that both observations have identical values for all variables

Similarity value of zero indicates that the two individuals differ maximally for all variables.

Similarity Measures for Binary Data

An extensive list of similarity measures for binary data exist, the reason for such is that a large number of possible measures has to do with the apparent uncertainty as to how to **deal with the count of zero-zero matches**

In some cases, zero-zero matches are equivalent to one-one matches and therefore should be included in the calculated similarity measure

Example: Gender, where there is no preference as to which of the two categories should be coded as zero or one

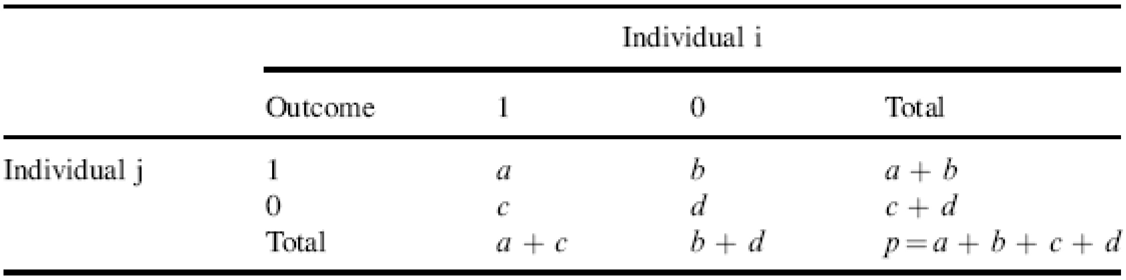
In other cases the inclusion or otherwise of the matches is more problematic

Example: When the zero category corresponds to the genuine absence of some property, such as wings in a study of insects

The question that then needs to be asked is do the co-absences contain useful information about the similarity of the two objects?

Attributing a high degree of similarity to a pair of individuals simply because they both lack a large number of attributes may not be sensible in many situations

The following table below will help when it comes to interpreting the measures

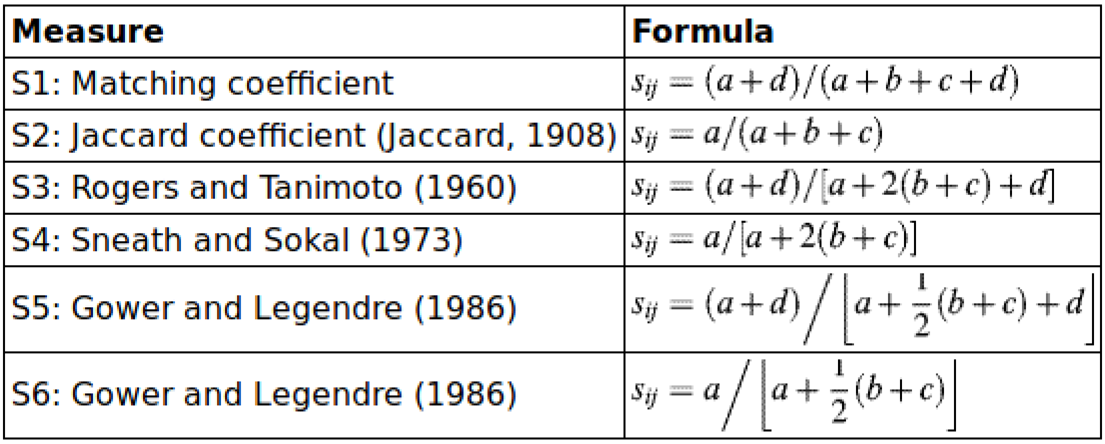


	Individual i			
	Outcome	1	0	Total
Individual j	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

Measure that ignore the co-absence (lack of both objects having a zero) are Jaccard's Coefficient (S2), Sneath and Sokal (S4)

When co-absences are considered informative, the simple matching coefficient (S1) is usually employed.

Measures S3 and S5 are further examples of symmetric coefficients that treat positive matches (a) and negative matches (d) in the same way.



Measure	Formula
S1: Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
S2: Jaccard coefficient (Jaccard, 1908)	$s_{ij} = a/(a+b+c)$
S3: Rogers and Tanimoto (1960)	$s_{ij} = (a+d)/[a+2(b+c)+d]$
S4: Sneath and Sokal (1973)	$s_{ij} = a/[a+2(b+c)]$
S5: Gower and Legendre (1986)	$s_{ij} = (a+d)\Big/\left[a+\frac{1}{2}(b+c)+d\right]$
S6: Gower and Legendre (1986)	$s_{ij} = a\Big/\left[a+\frac{1}{2}(b+c)\right]$

Similarity Measures for Categorical Data with More Than Two Levels

Categorical data where the variables have more than two levels (for example, eye color) could be dealt with in a similar way to binary data, with each level of a variable being regarded as a single binary variable.

This is not an attractive approach, however, simply because of the large number of ‘negative’ matches which will inevitably be involved.

A superior method is to allocate a score of zero or one to each variable depending on whether the two observations are the same on that variable. These scores are then averaged over all p variables to give the required similarity coefficient as

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ik}$$

Dissimilarity and Distance Measures for Continuous Data

A **metric** on a set X is a distance function

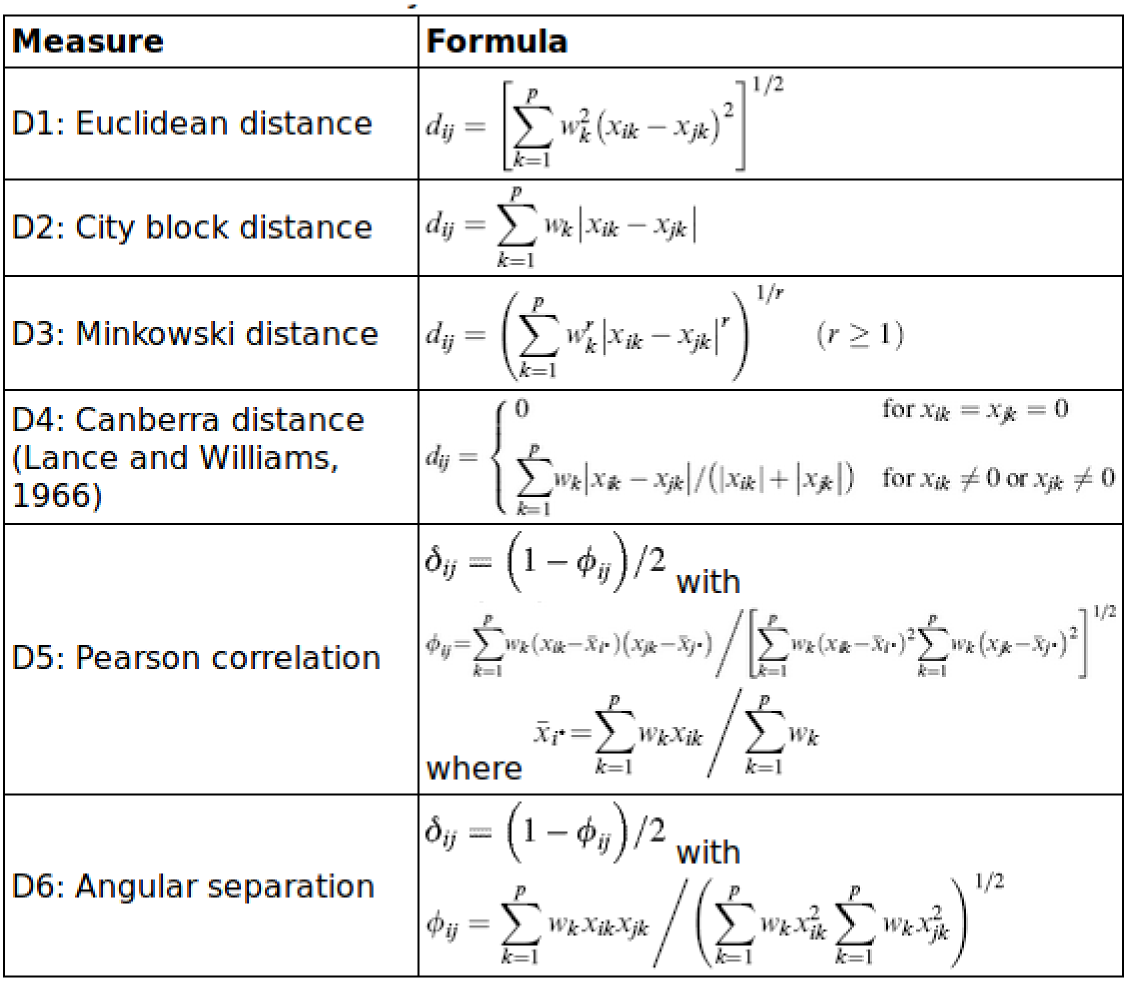
$$d : X \times X \to [0, \infty)$$

where $[0, \infty)$ is the set of non-negative real numbers and for all $x, y, z \in X$, the following conditions are satisfied

1. $d(x, y) \geq 0$ non-negativity or separation axiom
    1. $d(x, y) = 0 \iff x = y$ identity of indiscernibles
2. $d(x, y) = d(y, x)$ symmetry
3. $d(x, z) \leq d(x, y) + d(y, z)$ subadditivity or triangle inequality

Conditions 1 and 2 define a positive-definite function

All distance measures are formulated so as to allow for differential weighting of the quantitative variables w_k denotes the nonnegative weights of p variables



Measure	Formula
D1: Euclidean distance	$d_{ij} = \left[\sum_{k=1}^{p} w_k^2 \left(x_{ik} - x_{jk}\right)^2\right]^{1/2}$
D2: City block distance	$d_{ij} = \sum_{k=1}^{p} w_k \left\lvert x_{ik} - x_{jk} \right\rvert$
D3: Minkowski distance	$d_{ij} = \left(\sum_{k=1}^{p} w_k^r \left\lvert x_{ik} - x_{jk} \right\rvert^r\right)^{1/r} \quad (r \geq 1)$
D4: Canberra distance (Lance and Williams, 1966)	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk} \rvert / \left(\lvert x_{ik} \rvert + \lvert x_{jk} \rvert\right) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
D5: Pearson correlation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\bullet})(x_{jk} - \bar{x}_{j\bullet}) \Big/ \left[\sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\bullet})^2 \sum_{k=1}^{p} w_k (x_{jk} - \bar{x}_{j\bullet})^2\right]^{1/2}$ where $\bar{x}_{i\bullet} = \sum_{k=1}^{p} w_k x_{ik} \Big/ \sum_{k=1}^{p} w_k$
D6: Angular separation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2\right)^{1/2}$

Proposed dissimilarity measures can be broadly divided into distance measures and correlation-type measures.

Distance Measures

L^p Space

The Minkowski distance is a metric in normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance

$$D(X,Y) = (\sum_{i=1}^{n} w_i^p |x_i - y_i|^p)^{\frac{1}{p}}$$

This is a metric for $p > 1$

Manhattan Distance

This is the case in the Minkowski distance when $p = 1$

$$d(X,Y) = \sum_{i=1}^{n} w_i |x_i - y_i|$$

Manhattan distance depends on the rotation of the coordinate system, but does not depend on its reflection about a coordinate axis or its translation

$$d(x, y) = d(-x, -y)$$

$$d(x, y) = d(x + a, y + a)$$

Shortest paths are not unique in this metric

Euclidean Distance

This is the case in the Minkowski distance when $p = 2$. The Euclidean distance between points X and Y is the length of the line segment connection them.

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} w_i^2 (x_i - y_i)^2}$$

There is a unique path in which it has the shortest distance. This distance metric is also translation and rotation invariant

Squared Euclidean Distance

The standard Euclidean distance can be squared in order to place progressively greater weight on objects that are farther apart. In this case, the equation becomes

$$d(X, Y) = \sum_{i=1}^{n} w_i^2 (x_i - y_i)^2$$

Squared Euclidean Distance is not a metric as it does not satisfy the triangle inequality, however, it is frequently used in optimization problems in which distances only have to be compared.

Chebyshev Distance

The Chebyshev distance is where the distance between two vectors is the greatest of their differences along any coordinate dimension.

It is also known as **chessboard distance**, since in the game of chess the minimum number of moves needed by a king to go from one square on a chessboard to another equals the Chebyshev distance

$$d(X, Y) = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$= max_i(|x_i - y_i|)$$

Chebyshev distance is translation invariant

Canberra Distance Measure

The Canberra distance (D4) is a weighted version of the L_1 Manhattan distance. This measure is very sensitive to small changes close to $x_{ik} = x_{jk} = 0$.

It is often regarded as a generalization of the dissimilarity measure for binary data. In this context the measure can be divided by the number of variables, p, to ensure a dissimilarity coefficient in the interval $[0, 1]$

It can then be shown that this measure for binary variables is just one minus the matching coefficient.

Correlation Measures

It has often been suggested that the correlation between two observations can be used to quantify the similarity between them.

Since for correlation coefficients we have that $-1 \leq \phi_{ij} \leq 1$ with the value '1' reflecting the strongest possible positive relationship and the value '-1' the strongest possible negative relationship, these coefficients can be transformed into dissimilarities, d_{ij}, within the interval $[0, 1]$

The use of correlation coefficients in this context is far more contentious than its noncontroversial role in assessing the linear relationship between two variables based on n observations.

When correlations between two individuals are used to quantify their similarity the rows of the data matrix are standardized, not its columns.

Disadvantages

When variables are measured on different scales the notion of a difference between variable values and consequently that of a mean variable value or a variance is meaningless.

In addition, the correlation coefficient is unable to measure the difference in size between two observations.

Advantages

However, the use of a correlation coefficient can be justified for situations where all of the variables have been measured on the same scale and precise values taken are important only to the extent that they provide information about the subject's relative profile

Example: In classifying animals or plants, the absolute size of the organisms or their parts are often less important than their shapes. In such studies the investigator requires a dissimilarity coefficient that takes the value zero if and only if two individuals' profiles are multiples of each other. The angular separation dissimilarity measure has this property.

Further considerations

The Pearson correlation is sensitive to outliers. This has prompted a number of suggestions for modifying correlation coefficients when used as similarity measures; for example, robust versions of correlation coefficients such as *jackknife correlation* or altogether more general association coefficients such as *mutual information distance measure*

Mahalanobis (Maximum) Distance [Not between 2 observations]

Mahalanobis distance is a measure of distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D

Mahalanobis distance is unitless and scale-invariant and takes into account the correlations of the data set

$$D(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

Where μ is a set of mean observations and S is the covariance matrix

If the covariance matrix is diagonal then the resulting distance measure is called a normalized Euclidean distance.

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N}\frac{(x_i - y_i)^2}{s_i^2}}$$

Where s_i is the standard deviation of the x_i and y_i over the sample set

Discrete Metric

This metric describes whether or not two observations are equivalent

$$\rho(x,y) = \begin{cases} 1 & x \neq y \\ 0 & x = y \end{cases}$$

Similarity Measures for Data Containing both Continuous and Categorical Variables

There are a number of approaches to constructing proximities for mixed-mode data, that is, data in which some variables are continuous and some categorical.

1. Dichotomize all variables and use a similarity measure for binary data
2. Rescale all the variables so that they are on the same scale by replacing variable values by their ranks among the objects and then using a measure for continuous data
3. Construct a dissimilarity measure for each type of variable and combine these, either with or without differential weighting into a single coefficient.

Most general-purpose statistical software implement a number of measurs for converting two-mode data matrix into a one-mode dissimilarity matrix.

R has `cluster`, `clusterSim`, or `proxy`

Proximity Measures for Structured Data

We'll be looking at data that consists of repeated measures of the same outcome variable but under different conditions.

The simplest and perhaps most commonly used approach to exploiting the reference variable is in the construction of a reduced set of relevant summaries per object which are then used as the basis for defining object similarity.

Here we will look at some approaches for choosing summary measures and resulting proximity measures for the most frequently encountered reference vectors (e.g. time, experimental condition, and underlying factor)

Structured data arise when the variables can be assumed to follow a known *factor model*. Under *confirmatory factor analysis model* each variable or item can be allocated to one of a set of underlying factors or concepts. The factors cannot be observed directly but are 'indicated' by a number of items that are all measured on the same scale.

Note that the summary approach, while typically used with continuous variables, is not limited to variables on an interval scale. The same principles can be applied to dealing with categorical data. The difference is that summary measures now need to capture relevant aspects of the distribution of categorical variables over repeated measures.

Rows of X which represent ordered lists of elements, that is all the variables provide a categorical outcome and these variables can be aligned in one dimension, are more generally referred to as *sequences*. *Sequence analysis* is an area of research that centers on problems of events and actions in their temporal context and includes the measurements of similarities between sequences.

The most popular measure of dissimilarity between two sequences is the Levenhstein distance and counts the minimum number of operations needed to transform one sequence of categories into another, where an operation is an insertion, a deletion, or a substitution of a single category. Each operation may be assigned a penalty weight (a typical choice would be to give double the penalty to a substitution as opposed to an insertion or deletion. The measure is sometimes called the 'edit distance' due to its application in spell checkers.

Optimal matching algorithms (OMAs) need to be employed to find the minimum number of operations required to match one sequence to another. One such algorithm for aligning sequences is the Needleman-Wunsch algorithm, which is commonly used in bioinformatics to align proteins.

The *Jary similarity measure* is a related measure of similarity between sequences of categories often used to delete duplicates in the area of record linkage.

Inter-group Proximity Measures

In clustering applications, it becomes necessary to consider how to measure the proximity between groups of individuals.

1. The proximity between two groups might be defined by a suitable summary of the proximities between individuals from either group
2. Each group might be described by a representative observation by choosing a suitable summary statistic for each variable, and the inter group proximity defined as the proximity between the representative observations.

Inter-group Proximity Derived from the Proximity Matrix

For deriving inter-group proximities from a matrix of inter-individual proximities, there are a variety of possibilities

- Take the smallest dissimilarity between any two individuals, one from each group. This is referred to as *nearest-neighbor distance* and is the basis of the clustering technique known as *single linkage*
- Define hte intergroup distance as the largest distance between any two individuals, one from each group. This is known as the *furthest-neighbour distance* and constitute the basis of *complete linkage* cluster method.
- Define as the average dissimiliarity between individuals from both groups. Such a measure is used in *group average clustering*

Inter-group Proximity Based on Group Summaries for Continuous Data

One method for constructing inter-group dissimilarity measures for continuous data is to simply substitute group means (also known as the centroid) for the variable values in the formulae for inter-individual measures

More appropriate, however, might be measures which incorporate in one way or another, knowledge of within-group variation. One possibility is to use Mahallanobis's generalized distance.

Mahalanobis (Maximum) Distance

Mahalanobis distance is a measure of distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D

Mahalanobis distance is unitless and scale-invariant and takes into account the correlations of the data set

$$D(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

Where μ is a set of mean observations and S is the covariance matrix

If the covariance matrix is diagonal then the resulting distance measure is called a normalized Euclidean distance.

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}}$$

Where s_i is the standard deviation of the x_i and y_i over the sample set

Thus, the Mahalanobis distance incraeses with increasing distances between the two group centers and with decreasing within-group variation.

By also employing within-group correlations, the Mahalanobis distance takes account the possibly non-spherical shapes of the groups.

The use of Mahalanobis implies that the investigator is willing to **assume** that the covariance matrices are at least approximately the same in the two groups. When this is not so, this measure is an inappropriate inter-group measure. Other alternatives exist such as the one proposed by Anderson and Bahadur

$$\delta_{\mathrm{AB}} = \max_{t} \frac{2\mathbf{b}_t'\mathbf{d}}{\left(\mathbf{b}_t'\mathbf{W}_{\mathrm{A}}\mathbf{b}_t\right)^{1/2} + \left(\mathbf{b}_t'\mathbf{W}_{\mathrm{B}}\mathbf{b}_t\right)^{1/2}},$$

Another alternative is the *normal information radius* suggested by Jardine and Sibson

$$\mathrm{NIR} = \frac{1}{2}\log_2\left\{\frac{\det\left[\frac{1}{2}\left(\mathbf{W}_{\mathrm{A}} + \mathbf{W}_{\mathrm{B}}\right)\right] + \frac{1}{4}\left(\bar{\mathbf{x}}_{\mathrm{A}} - \bar{\mathbf{x}}_{\mathrm{B}}\right)'\left(\bar{\mathbf{x}}_{\mathrm{A}} - \bar{\mathbf{x}}_{\mathrm{B}}\right)}{\det\left(\mathbf{W}_{\mathrm{A}}\right)^{1/2}\det\left(\mathbf{W}_{\mathrm{B}}\right)^{1/2}}\right\}.$$

Inter-group Proximity Based on Group Summaries for Categorical Data

Approaches for measuring inter-group dissimilarities between groups of individuals for which categorical variables have been observed have been considered by a number of authors. Balakrishnan and Sanghvi (1968), for example, proposed a dissimilarity index of the form

$$G^2 = \sum_{k=1}^{p} \sum_{l=1}^{c_k+1} \frac{(p_{\mathrm{A}kl} - p_{\mathrm{B}kl})^2}{p_{kl}},$$

where p_{Akl} and p_{Bkl} are the proportions of the lth category of the kth variable in group A and B respectively, $p_{kl} = \frac{1}{2}(p_{\mathrm{A}kl} + p_{\mathrm{B}kl})$, ck + 1 is the number of categories for the kth variable and p is the number of variables.

Kurczynski (1969) suggested adapting the generalized Mahalanobis distance, with categorical variables replacing quantitative variables. In its most general form, this measure for inter-group distance is given by

$$D_{\mathrm{p}}^2 = (\mathbf{p}_{\mathrm{A}} - \mathbf{p}_{\mathrm{B}})' \mathbf{W}_{\mathrm{p}}^{-1} (\mathbf{p}_{\mathrm{A}} - \mathbf{p}_{\mathrm{B}}),$$

where $\mathbf{p}_{\mathrm{A}} = (p_{\mathrm{A}11}, p_{\mathrm{A}12}, \ldots, p_{\mathrm{A}1c_1}, p_{\mathrm{A}21}, p_{\mathrm{A}22},, \ldots, p_{\mathrm{A}2c_2}, \ldots, p_{\mathrm{A}k1}, p_{\mathrm{A}k2},, \ldots, p_{\mathrm{A}kc_k})'$ contains sample proportions in group A and $\mathbf{p}_{\mathrm{B}}$ is defined in a similar manner, and $\mathbf{W}_{\mathrm{p}}$ is the m × m common sample covariance matrix, where $m = \sum_{k=1}^{p} c_k$.

Weighting Variables

To weight a variable means to give it greater or lesser importance than other variables in determining the proximity between two objects.

The question is 'How should the weights be chosen?' Before we discuss this question, it is important to realize that the selection of variables for inclusion into the study already presents a form of weighting, since the variables not included are effectively being given the weight 0.

The weights chosen for the variables reflect the importance that the investigator assigns the variables for the classification task.

There are several approaches to this

- Authors obtain perceived dissimilarities between selected objects, they then model the dissimilarities using the underlying variables and weights that indicate their relative importance. The weights that best fit the perceived dissimilarities are then chosen.
- Define the weights to be inversely proportion to some measure of variability in this variable. This choice of weights implies that the importance of a variable decreases when its variability increases.
    - For a continous variable, the most commonly emplyed weight is either the reciprocal of its standard deviation or the reciprocal of its range
    - Employing variability weights is equivalent to what is commonly referred to as *standardizing* the variables.
- Construct weights from the data matrix using *variable section*. In essence, such procedures proceed in an iterative fashion to identify variables which, when contributing to a cluster algorithm, lead to internally cohesive and externally isolated clusters and, when clustered singly, produce reasonable agreement.

The second approach assumed the importance of a variable to be inversely proportional to the total variability of that variable. The total variability of a variable comprises variation both within and between groups which may exist within the set of individuals. The aim of cluster analysis is typically to identify such groups. Hence it can be argued that the importance of a variable should not be reduced because of between-group variation (on the contrary, one might wish to assign more importance to a variable that shows larger between-group variation.)

Gnanadesikan et al. (1995) assessed the ability of squared distance functions based on data-determined weights, both those described above and others, to recover groups in eight simulated and real continuous data sets in a subsequent cluster analysis. Their main findings were:

1. Equal weights, (total) standard deviation weights, and range weights were generally ineffective, but range weights were preferable to standard deviation weights.
2. Weighting based on estimates of within-cluster variability worked well overall.
3. Weighting aimed at emphasizing variables with the most potential for identifying clusters did enhance clustering when some variables had a strong cluster structure.
4. Weighting to optimize the fitting of a hierarchical tree was often even less effective than equal weighting or weighting based on (total) standard deviations.
5. Forward variable selection was often among the better performers. (Note that all-subsets variable selection was not assessed at the time.)

Standardization

In many clustering applications, the variables describing the objects to be clustered will not be measured in the same units. A number of variability measures have been used for this purpose

- When standard deviations calculated from the complete set of objects to be clustered are used, the technique is often referred to as *auto-scaling, standard scoring, or z-scoring*.
- Division by the median absolute deviations or by the ranges.

The second is shown to outperform auto-scaling in many clustering applications. As pointed out in the previous section, standardization of variables to unit variance can be viewed as a special case of weighting.

Choice of Proximity Measure

Firstly, the nature of the data should strongly influence the choice of the proximity measure.

Next, the choice of measure should depend on the scale of the data. Similarity coefficients should be used when the data is binary. For continuous data, distance of correlation-type dissimilarity measure should be used according to whether 'size' or 'shape' of the objects is of interest.

Finally, the clustering method to be used might have some implications for the choice of the coefficient. For example, making a choice between several proximity coefficients with similar properties which are also known to be monotonically related can be avoided by employing a cluster method that depends only on the ranking of the proximities, not their absolute values.