Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

In order to separate a set of objects into categories according to some measure of similarity between individual items there has to be some concept of the distance between them. For instance, for two sets of coordinates $\alpha = (x_1, y_1)$ and $\beta = (x_2, y_2)$ we could use the square of the Euclidean distance between them, that is

$$(\alpha - \beta)^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$$

as this is the squared length of the hypotenuse of a right angle triangle with coordinates (x_1, y_1), (x_2, y_1) and (x_2, y_2). We could then group items together depending on such a distance measure between them or according to distances from some fixed points. Cluster analysis extend such a concept to situations involving more than two dimensions, and using alternative measures of distance.

Calculating a distance matrix

The idea is, as in data mining, where you have a n by m matrix a_{ij} of m variables (columns) for each of n cases (rows) and wish to explore clustering, that is groupings together of like entities. To do this, you choose an appropriate pre-analysis transformation of the data, a suitable distance measure, a meaningful scaling procedure, and a sensible linkage function. SimFIT will then calculate a distance matrix, or a similarity matrix, and plot the clusters as a dendrogram. As an example, from the main SimFIT menu choose [Statistics], [Multivariate], then [Distance matrix] and analyze the test file `cluster.tf1` giving the results displayed in this table.

Variables included:
1 2 3 4 5 6 7 8
Transformation: Untransformed
Distance: Euclidean distance
Scaling: Unscaled
Linkage: Group average
Weighting: [weights r not used]
Distance matrix (strict lower triangle) is:

2)	22.0								
3)	36.2	28.8							
4)	22.9	29.7	36.6						
5)	1.95	16.6	31.1	24.5					
6)	39.8	32.7	40.6	31.8	26.1				
7)	21.7	28.3	38.2	21.3	19.3	36.2			
8)	14.1	24.1	42.6	18.8	18.9	34.2	18.5		
9)	32.7	23.0	45.4	44.9	23.6	38.7	36.6	33.4	
10)	31.6	23.9	37.2	41.0	22.2	43.9	33.5	33.9	(+)
10)	24.7								
11)	32.2	24.4	39.1	41.8	20.2	41.4	31.3	33.4	(+)
11)	19.9	8.25							
12)	29.9	22.7	37.7	39.0	17.2	38.4	29.2	31.4	(+)
12)	18.1	11.4	6.24						

Note that, as a distance matrix is symmetrical with diagonals = 0, only the strict lower triangle is displayed. The header to this table indicates that all eight variables were included in the analysis using untransformed data, the Euclidean distance, no data scaling, group average linkage, and no weights. The symbol (+) merely indicates wrap round due to long lines. The meaning of the parameter settings in the table header will now be explained.

Distance matrix norms

The distance d_{jk} between objects j and k is just a chosen variant of the weighted L_p norm

$$d_{jk} = \{\sum_{i=1}^{m} w_{ijk} D(a_{ji}/s_i, a_{ki}/s_i)\}^p, \text{ for some } D, \text{ e.g.,}$$

(a) The Euclidean distance $D(\alpha, \beta) = (\alpha - \beta)^2$ with $p = 1/2$ and $w_{ijk} = 1$

(b) The Euclidean squared difference $D(\alpha, \beta) = (\alpha - \beta)^2$ with $p = 1$ and $w_{ijk} = 1$

(c) The absolute distance $D = |\alpha - \beta|$ with $p = 1$ and $w_{ijk} = 1$, otherwise known as the Manhattan or city block metric.

However, as the values of the variables may differ greatly in size, so that large values would dominate the analysis, it is usual to subject the data to a preliminary transformation or to apply a suitable weighting. Often it is best to transform the data to standardized $(0, 1)$ form before constructing the dendrogram, or at least to use some sort of scaling procedure such as:

(i) use the sample standard deviation as s_i for variable i,

(ii) use the sample range as s_i for variable i, or

(iii) supply precalculated values of s_i for variable i.

Bray-Curtis dissimilarity uses the absolute distance except that the weighting factor is given by

$$w_{ijk} = \frac{1}{\sum_{i=1}^{m}(a_{ji}/s_i + a_{ki}/s_i)}$$

which is independent of the variables i and only depends on the cases j and k, and distances are usually multiplied by 100 to represent percentage differences. Bray-Curtis similarity is the complement, i.e., 100 minus the dissimilarity.

The Canberra distance measure, like the Bray-Curtis one, also derives from the absolute distance except that the weighting factor is now

$$w_{ijk} = \frac{1}{\lambda(a_{ji}/s_i + a_{ki}/s_i)}.$$

There are various conventions for defining λ and deciding what to do when values or denominators are zero with the Bray-Curtis and Canberra distance measures, and the scheme used by SIMFIT is as follows.

- If any values are negative the calculation is terminated.
- If any Bray-Curtis denominator is zero the calculation is terminated.
- If there are no zero values, then λ is equal to the number of variables in the Canberra measure.
- If both members of a pair are zero, then λ is decreased by one for each occurrence of such a pair, and the pairs are ignored.
- If one member of a pair is zero, then it is replaced by the smallest non-zero value in the data set divided by five, then scaled if required.

Distance matrix linkage

The values in a distance matrix will affect subsequent analysis. For instance, the shape of a dendrogram depends on the choice of analytical techniques and the order of objects plotted is arbitrary: groups at a

given fixed distance can be rotated and displayed in either orientation. Another choice which will affect the dendrogram shape is the method used to recalculate distances after each merge has occurred. Suppose there are three clusters i, j, k with n_i, n_j, n_k objects in each cluster and let clusters j and k be merged to give cluster jk. Then the distance from cluster i to cluster jk can be calculated in several ways.

[1] Single link: $d_{i,jk} = \min(d_{ij}, d_{ik})$

[2] Complete link: $d_{i,jk} = \max(d_{ij}, d_{ik})$

[3] Group average: $d_{i,jk} = (n_j d_{ij} + n_k d_{ik})/(n_j + n_k)$

[4] Centroid: $d_{i,jk} = (n_j d_{ij} + n_k d_{ik} - n_j n_k d_{jk}/(n_j + n_k))/(n_j + n_k)$

[5] Median: $d_{i,jk} = (d_{ij} + d_{ik} - d_{jk}/2)/2$

[6] Minimum variance: $d_{i,jk} = \{(n_i + n_j)d_{ij} + (n_i + n_k)d_{ik} - n_i d_{jk}\}/(n_i + n_j + n_k)$

Distance matrix nearest neighbors

Once a distance matrix has been calculated, it is sometimes useful to calculate the nearest neighbors, as illustrated in the next table for the previous data.

Object	Nearest	Distance
1	8	14.1067
2	5	16.5529
3	2	28.7576
4	8	18.7617
5	2	16.5529
6	5	26.0960
7	8	18.4932
8	1	14.1067
9	12	18.1384
10	11	8.24621
11	12	6.24500
12	11	6.24500

In this table, column 1 refers to the objects in logical order, column 2 indicates the object that is closest, i.e., the nearest neighbor, while column 3 records these minimum distances. Clearly, the nearest neighbors will depend upon the parameters used to configure the calculation of the distance matrix.