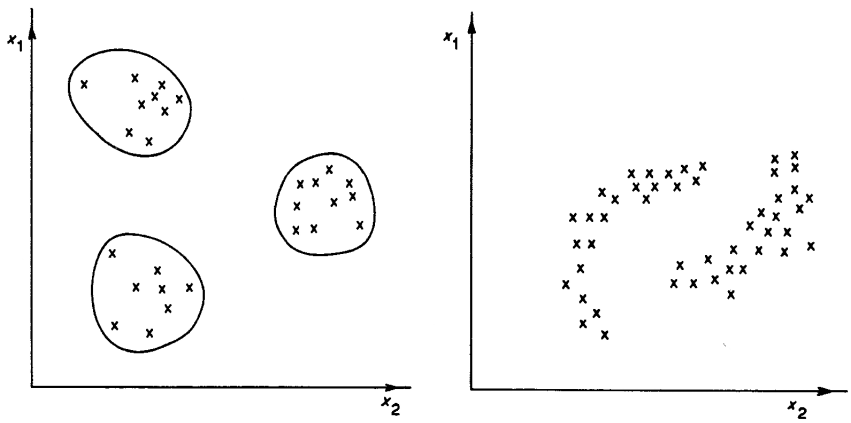
MSc II: Further Statistical Analysis (Autumn Term)
Advanced Multivariate Analysis

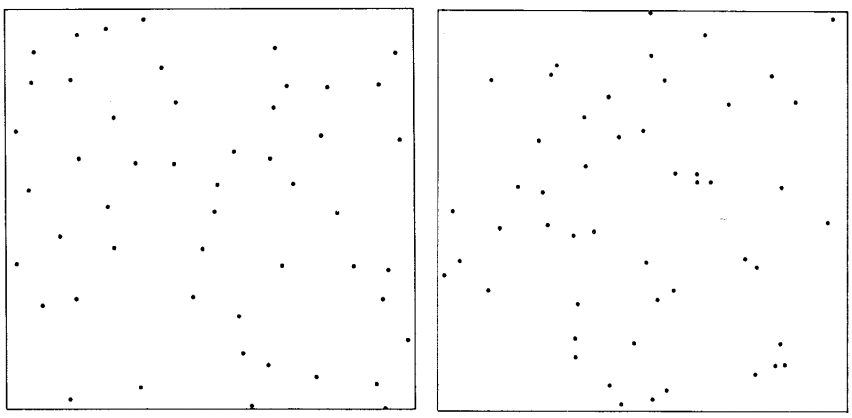
4 Cluster Analysis

The aim of Cluster Analysis is to find 'natural groupings', if any, among a set of units (objects, individuals, ...) based on data for the units. This is another exploratory, hypothesis generating technique, and there are very many techniques used in practice.

Roughly speaking we want individuals within the same cluster or group to be similar to one another while individuals in different groups are dissimilar. However we can't, in general, define a cluster. This probably accounts for the number of procedures that exist for finding them since each method used imposes some kind of structure on the data and implicitly defines what kind of grouping or clustering is being sought.



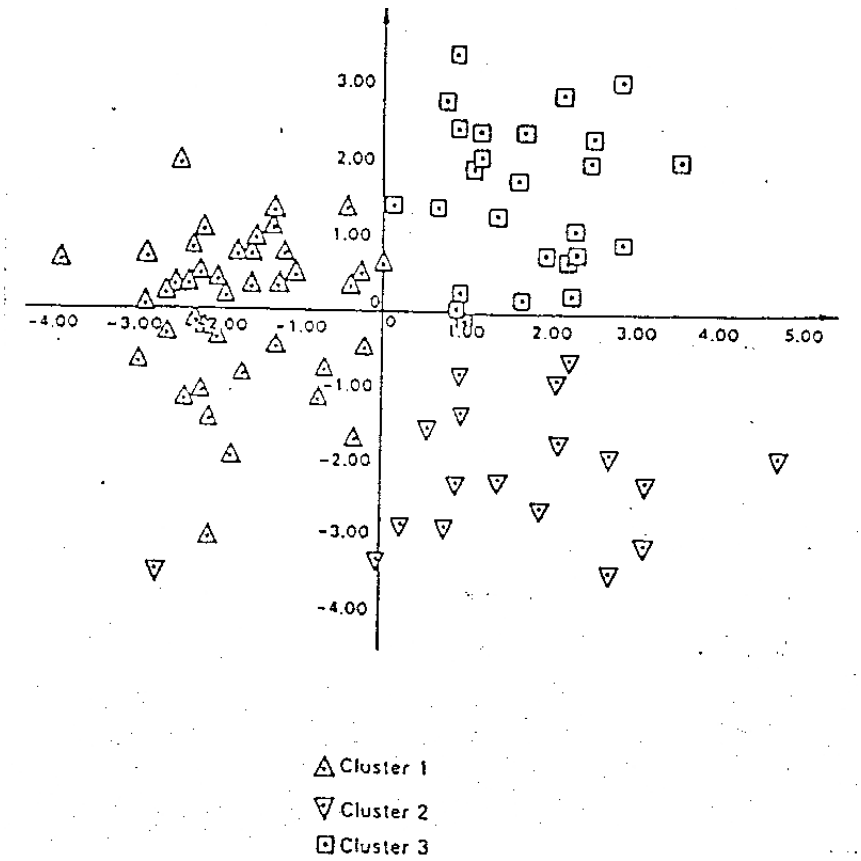
If the data can be represented graphically, the human eye is often the best detector of clustering. However this is not always so. It is hard, for example, for the human eye to detect randomness, and it may be necessary to plot and test for randomness before allowing the computer to produce clusters which may be spurious.



Consider also the following example:

Data were collected on 86 long-term prisoners and 8 scores (quantatitive variables) were used to produce a dissimilarity matrix and a cluster analysis was carried out. Three clusters were found. In order to try and represent the clusters graphically it was assumed that the clusters represented three samples from populations with different mean vectors and a common covariance matrix - and a canonical variate analysis was carried out using the original 8 variables. The 86 prisoners were then plotted with respect to the 2 canonical axes.

(**Note**: Cannot possibly carry out a 1 way MANOVA test since groupings were determined by examining the data).



It is clear from the above plot that there is no real 'clustering'. Thus there are problems in deciding whether it makes <u>sense</u> to cluster, defining what is <u>meant</u> by a cluster, and deciding what method to use bearing in mind that each method imposes it's own structure on the data.

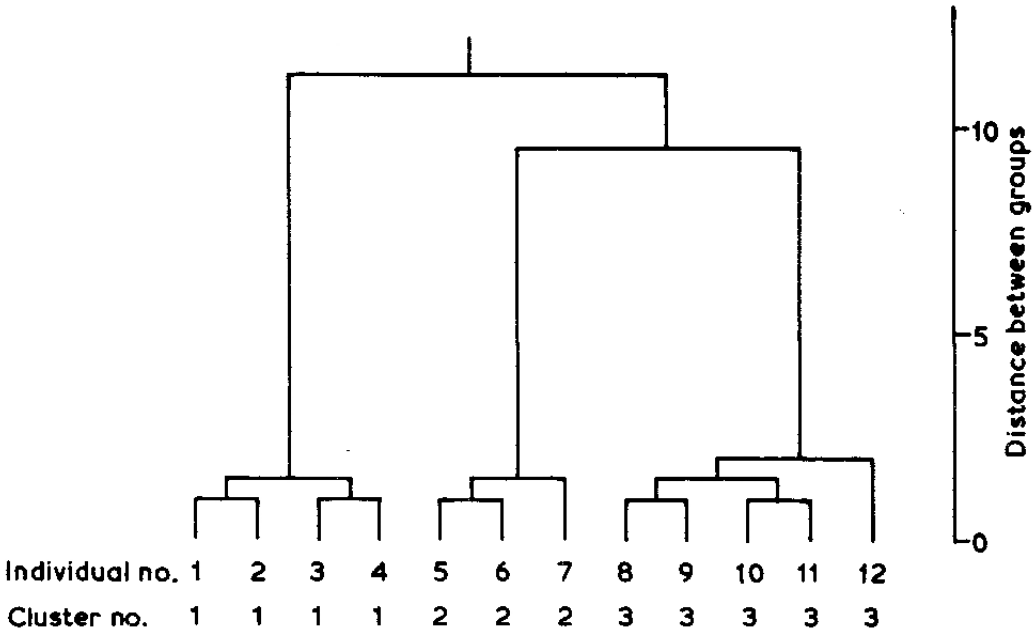
e.g. Some cluster techniques always find ball shaped clusters, others find chains, some don't seem to detect unequal sized groups, some look specifically for mixtures of multivariate normal distributions but assume a common covariance matrix. Even if this holds the methods do not always work well.

It is often sensible to use several methods of cluster analysis (giving some thought to which would be suitable) - and it is also very important to find a suitable graphical representation on which to superimpose the resulting clustering.

We will consider only hierarchical cluster methods operating on a dissimilarity matrix $\mathbf{D} = (d_{ij})$.

4.1 Hierarchical Cluster Methods

Hierarchical clustering techniques produce a tree structure of clusters such that as we move back towards the root, new clusters form by merging old ones. Thus we have a nested sequence of partitions into g groups, starting with $g = n$ (where n is the number of individuals) and finishing with $g = 1$. The hierarchical structure is usually represented by a tree-diagram or dendrogram with a distance scale associated with it.



Notes:

1. The ordering of the branches is arbitrary to some extent. i.e. units 1 and 12 cannot be taken to be the most dissimilar.

2. For any particular value of distance on the associated scale, a horizontal line can be drawn giving the number and combination of clusters existing at that distance level.

3. We start with distances between individuals but the vertical axis represents distances between groups (i.e. the 'threshold' distances at which groups join). Inter-group or inter-cluster distances can be defined in many ways and the method used defines the method

of cluster analysis and determines the kind of clusters obtained.

We shall consider only agglomerative algorithms for producing an hierarchical structure. That is, the algorithm always starts with n clusters, the n separate individuals, and gradually merges them into fewer clusters until $g = 1$. Note that there are also divisive algorithms. Note also that some cluster methods can be implemented using either an agglomerative algorithm or a divisive algorithm, or some other algorithm. (This is true of single linkage cluster analysis - each method gives the same dendrogram).

General Agglomerative Procedure

Let d_{ij} denote the dissimilarity between the ith and jth individuals and D_{ij} denote the dissimilarity between the ith and jth groups being formed.

(**1.**) Start with n clusters ($D_{ij} = d_{ij}$).

(**2.**) Find the smallest element of (D_{ij}), say D_{rs} and unite groups r and s into a new group. Record D_{rs} as the threhold distance at which this new group formed. (*What about ties?*).

(**3.**) Calculate the dissimilarity between the new group (rs) and each other group i using

$$D_{i(rs)} = f(D_{ir}, D_{is}, D_{rs}, n_r, n_s, n_i)$$

for all $i \neq r$ or s, thus producing a reduced order matrix (D_{ij}).

(**4.**) Go to (**2.**) unless just one group remains.

We shall consider 7 different methods that differ only in the function used to calculate distances between groups.

M1: Single Linkage Cluster Analysis

$$\begin{aligned} D_{i(rs)} &= \min\{D_{ir}, D_{is}\} \\ &= \frac{1}{2}\{D_{ir} + D_{is} - |D_{ir} - D_{is}|\} \end{aligned}$$

M2: Unweighted Average Linkage

$$D_{i(rs)} = \frac{1}{2}(D_{ir} + D_{is})$$

M3: Weighted Average Linkage

$$D_{i(rs)} = \frac{n_r D_{ir} + n_s D_{is}}{n_r + n_s}$$

M4: Complete Linkage

$$\begin{aligned} D_{i(rs)} &= \max\{D_{ir}, D_{is}\} \\ &= \frac{1}{2}\{D_{ir} + D_{is} + |D_{ir} - D_{is}|\} \end{aligned}$$

Note that methods M1 and M4 use only the rank order of the distances.

The final three methods are based on geometrical ideas in which the n original units are thought of as points in some Euclidean space.

M5: Centroid Method

D_{ij} is the squared Euclidean distance between the centroids of clusters i and j.

$$D_{i(rs)} = \frac{n_r}{n_r + n_s} D_{ir} + \frac{n_s}{n_r + n_s} D_{is} - \frac{n_r n_s}{(n_r + n_s)^2} D_{rs}$$

M6: Median Method

D_{ij} is the squared Euclidean distance between the medians of clusters i and j (or weighted centroids).

$$D_{i(rs)} = \frac{1}{2} D_{ir} + \frac{1}{2} D_{is} - \frac{1}{4} D_{rs}$$

M7: Ward's Method (Minimum Sum of Squares)

D_{ij} is the increase in the sum of squares within clusters, after fusion, summed over all variables.

$$D_{i(rs)} = \frac{n_i + n_r}{n_i + n_r + n_s} D_{ir} + \frac{n_i + n_s}{n_i + n_r + n_s} D_{is} - \frac{n_i}{n_i + n_r + n_s} D_{rs}$$

These, and many other methods are special cases of a general form for calculating distances between groups. The method is due to Lance and Williams.

$$D_{i(rs)} = \alpha_r D_{ir} + \alpha_s D_{is} - \beta D_{rs} + \gamma |D_{ir} - D_{is}|$$

Thus <u>all</u> the above methods (and others) are implemented in a single algorithm.

<u>Note 1</u>: Considering the geometric methods in particular, we think of clusters as points in space.

M5 (Centroid method):

Clusters r, s and i are represented as $\mathbf{x}_r$, $\mathbf{x}_s$ and $\mathbf{x}_i$, their mean vectors or centroids.

The centroid of (rs) is

$$\frac{n_r}{n_r + n_s}\mathbf{x}_r + \frac{n_s}{n_r + n_s}\mathbf{x}_s$$

Since D_{ij} is the squared Euclidean distance between centroids i and j, we have

$$D_{i(rs)} = (\mathbf{x}_i - \mathbf{x}_{(rs)})^T(\mathbf{x}_i - \mathbf{x}_{(rs)})$$

After some algebra this reduces to

$$D_{i(rs)} = \frac{n_r}{n_r + n_s}(\mathbf{x}_i - \mathbf{x}_r)^T(\mathbf{x}_i - \mathbf{x}_r) + \frac{n_s}{n_r + n_s}(\mathbf{x}_i - \mathbf{x}_s)^T(\mathbf{x}_i - \mathbf{x}_s) - \frac{n_r n_s}{(n_r + n_s)^2}(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s)$$

i.e.

$$D_{i(rs)} = \frac{n_r}{n_r + n_s}D_{ir} + \frac{n_s}{n_r + n_s}D_{is} - \frac{n_r n_s}{(n_r + n_s)^2}D_{rs}$$

Suppose now that we are using $d_{ij}^2 = c_{ii} + c_{jj} - 2c_{ij}$ where the c_{ij} are similarities. Suppose also that we scale the similarities such that $0 \leq c_{ij} \leq 1$, $\forall i, j$, so that $d_{ij}^2 = 2(1 - c_{ij})$. Then for centroid cluster analysis:

$$2(1 - c_{i(rs)}) = \frac{2n_r}{n_r + n_s}(1 - c_{ir}) + \frac{2n_s}{n_r + n_s}(1 - c_{is}) - \frac{2n_r n_s}{(n_r + n_s)^2}(1 - c_{rs})$$

$$\Rightarrow c_{i(rs)} = \frac{n_r}{n_r + n_s}c_{ir} + \frac{n_s}{n_r + n_s}c_{is} - \frac{n_r n_s}{(n_r + n_s)^2}(1 - c_{rs})$$

So that we obtain a simple update formula for similarities. A feature of centroid cluster analysis is that small groups tend to be swamped by large groups and their distinctive features lost.

M6 (Median method):

$$D_{i(rs)} = \frac{1}{2}D_{ir} + \frac{1}{2}D_{is} - \frac{1}{4}D_{rs}$$

(as if $n_r = n_s$ in the Centroid method. This actually implies a rather complicated weighting). A feature of Median cluster analysis is that possibly undue weight may be given to the latest arrivals in the cluster.

Note 2: The geometric methods M5 and M6 can sometimes give rise to problems since it is possible to get

$$D_{i(rs)} \leq D_{rs}(\leq \min(D_{ir}, D_{is}))$$

which would lead to an illegal dendrogram. However, it is possible that the advantage of a geometric interpretation, which helps to tie in these methods with graphical representations, could outweigh this disadvantage. (See, for example, Everitt *et al* (2001), and Gordon (1981)).

Note 3: The algorithm we have described does not say what to do in the case of ties. It is still the case that, in the presence of ties, many algorithms simply select the pair of groups to join at random.

Properties of the Dendrogram

The dendrogram resulting from the algorithm has a distance scale associated with it that is based on the method chosen for calculating distances between groups. It also defines a new set of distances between individuals, d^*_{ij}, where d^*_{ij} is the threshold distance at which i and j just become members of the same group.

The d^*_{ij} satisfy the triangle (metric) inequality and also satisfy the ultrametric inequality which says that

$$d^*_{ij} \leq \max(d^*_{ik}, d^*_{kj}), \quad \forall i, j, k$$

(check from dendrograms).

A matrix (d_{ij}) can be exactly represented by a dendrogram if and only if the d_{ij} satisfy the ultrametric inequality.

Goodness-of-Fit

The goodness-of-fit of a dendrogram to given data (d_{ij}) can be assessed using a number of measures comparing the d_{ij} and the fitted distances d^*_{ij}.

For example,

(i) $\sum_{i<j} |d_{ij} - d^*_{ij}|$

(ii) $\sum_{i<j} (d^2_{ij} - d^{*2}_{ij})$ (c/f Principal components)

(iii) $\sum_{i<j} (d_{ij} - d^*_{ij})^2$ (Least squares)

For these measures, some kind of scaling or standardisation would be required to make them comparable over different sets of data.

Other possiblities include

(iv) Either the Pearson product-moment correlation (coppenetic correlation) <u>or</u> a rank correlation between d_{ij} and d^*_{ij}

(v) A series of measures given by

$$\delta_\mu = \frac{[\sum_{i<j} |d_{ij} - d^*_{ij}|^{1/\mu}]^\mu}{[\sum_{i<j} d^{1/\mu}_{ij}]^\mu}, 0 \leq \mu \leq 1$$

Varying μ places greater or lesser emphasis on larger or smaller differences. As $\mu \to 0$ the largest difference dominates.

4.2 What method of Cluster analysis should we use?

Informally, it depends on the problem. It is probably sensible to try a number of cluster methods to see how stable the cluster structure is. It is also important to be aware of the characteristics of different methods and the kind of clusters they are likely to find.

For example, complete-linkage, unweighted and weighted average, centroid and Ward's methods all tend to find 'spherical' clusters, i.e. clusters with high internal affinity. Single-linkage often finds elongated, straggly clusters ('chaining'), especially in large data sets.

Complete-linkage tends to produce clusters of equal diameter, while Ward's method tends to find clusters of equal size (numbers of units). However, Ward's method is sensitive to outliers while single-linkage is sensitive to intermediate units.

Both the median and Centroid methods are subject to reversals producing illegal dendrograms. The median method weights groups towards the units most recently added, while in the centroid method the more numerous of two groups joining will dominate the merged cluster and the characteristics of small groups may be lost.

See Everitt *et al* for more properties.

We may be attracted to geometrically based methods in order to tie in with a graphical representation from a multidimensional scaling technique.

We may need to consider which techniques can handle a very large dissimilarity matrix. (A good algorithm for single-linkage can cluster thousands of units relatively easily).

Formally, Jardine and Sibson (in their book '*Mathematical Taxonomy*') propose a set of general mathematical properties that it would seem desirable for any good method to satisfy.

For example,

1. Results produced should not depend on the order of input of individuals. This can happen if ties are broken either randomly or in some systematic way.

2. The fitted dendrogram should be invariant under any monotonic transformation of the d_{ij}. Only single-linkage and complete-linkage satisfy this criterion.

3. Continuity Criterion
'Small' changes in the original data (d_{ij}) should only produce 'small' changes in the resulting tree. This seems to be a reasonable criterion since we expect observations to be subject to error and random variability.

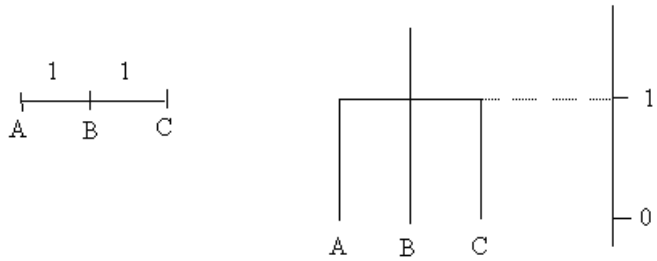
Examples:

Consider the matrix $\mathbf{D} = (d_r s) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1 \pm \epsilon & 0 & \\ 2 & 1 \mp \epsilon & 0 \end{bmatrix}$, ($\epsilon = 0$ or >0 and small).

(i)Single-linkage Cluster Analysis

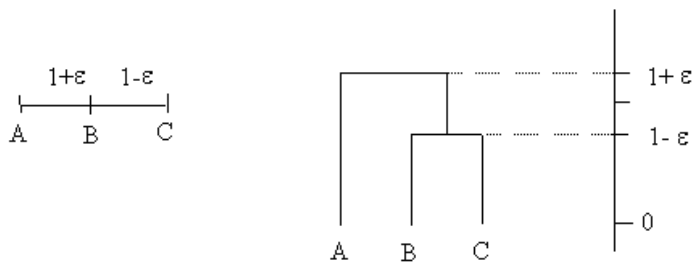
(a) $\underline{\epsilon = 0}$ (B is equidistant from A and C).

$$(d_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 2 & 1 & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 1 & 1 & 0 \end{bmatrix}$$



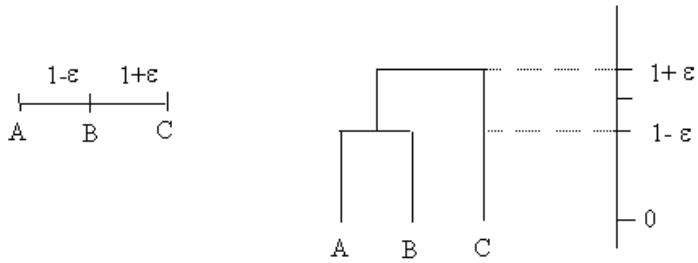
(b) .

$$(d_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1+\epsilon & 0 & \\ 2 & 1-\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1+\epsilon & 0 & \\ 1+\epsilon & 1-\epsilon & 0 \end{bmatrix}$$



(c) .

$$(d_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ 2 & 1+\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ 1+\epsilon & 1+\epsilon & 0 \end{bmatrix}$$

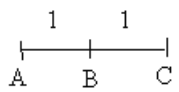


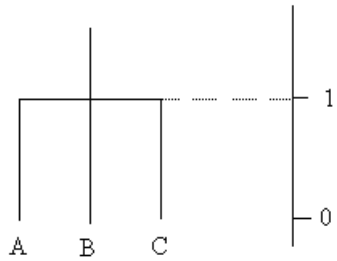
So changes in (d^*_{rs}) and the dendrogram are small. (<u>Note</u>: if ties are handled by choosing randomly which pair to join first, then for single linkage, the result is the same whichever pair are joined first and the dendrogram in (a) is eventually obtained).

(ii), (iii)<u>Weighted or unweighted average linkage</u>

(a) <u>$\epsilon = 0$</u>.

$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 2 & 1 & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 1 & 1 & 0 \end{bmatrix}$$

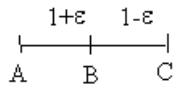


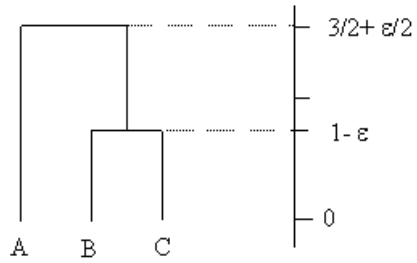


(As before, if ties are handled 'properly').

(b) .

$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1+\epsilon & 0 & \\ 2 & 1-\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ \frac{3}{2}+\frac{\epsilon}{2} & 0 & \\ \frac{3}{2}+\frac{\epsilon}{2} & 1-\epsilon & 0 \end{bmatrix}$$

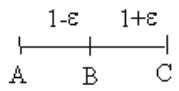


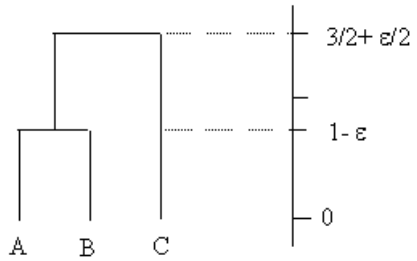


(Since $\delta_{A(BC)} = \frac{1}{2}[\delta_{AB} + \delta_{AC}] = \frac{1}{2}[1 + \epsilon + 2] = \frac{3}{2} + \frac{\epsilon}{2}$).

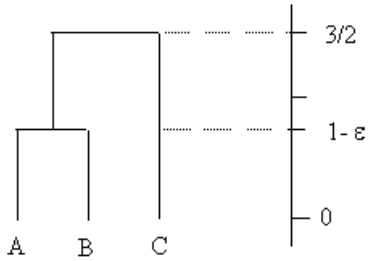
(c) .

$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ 2 & 1+\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ \frac{3}{2}+\frac{\epsilon}{2} & \frac{3}{2}+\frac{\epsilon}{2} & 0 \end{bmatrix}$$

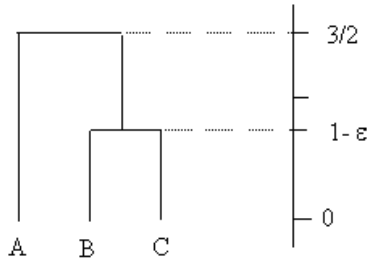




<u>Note</u>: If ties are <u>not</u> handled 'properly' and the pair to join is chosen at random, then in (a) $\epsilon = 0$ would lead to either



or

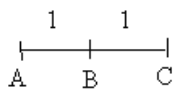


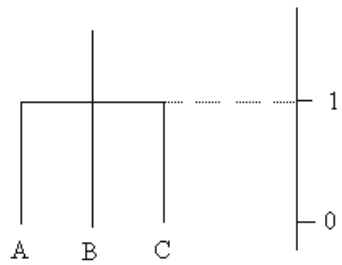
completely at random.

(iv)<u>Complete linkage</u>

(a) <u>$\epsilon = 0$</u>.

$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 2 & 1 & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1 & 0 & \\ 1 & 1 & 0 \end{bmatrix}$$

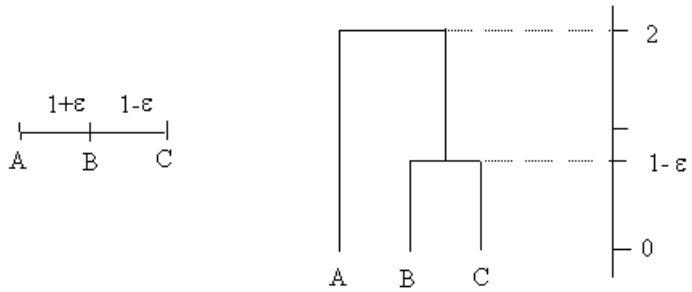




(Again, if ties are handled 'properly').

(b) .

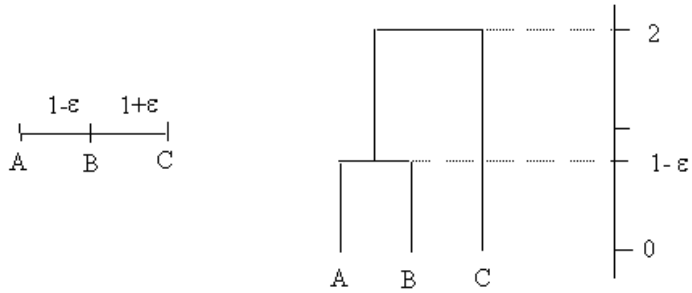
$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1+\epsilon & 0 & \\ 2 & 1-\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 2 & 0 & \\ 2 & 1-\epsilon & 0 \end{bmatrix}$$



(Since $\delta_{A(BC)} = \max\{\delta_{AB}, \delta_{AC}\} = 2$).

(c) .

$$(d_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ 2 & 1+\epsilon & 0 \end{bmatrix} \qquad (d^*_{rs}) = \begin{array}{c} \mathrm{A} \\ \mathrm{B} \\ \mathrm{C} \end{array} \begin{bmatrix} 0 & & \\ 1-\epsilon & 0 & \\ 2 & 2 & 0 \end{bmatrix}$$



Thus all these methods, except single-linkage, have large changes in output associated with small changes in input. That is, in this example, we have shown that there is a discontinuity associated with distances that are either very close or tied.

4. Fitting-together conditions

Roughly speaking these say that adding or subtracting an individual from the original data set should make the structure of the dendrogram change relatively little. In this case clustering will be robust with respect to outliers. Jardine and Sibson find a precise mathematical formulation of this.

A number of other conditions that should be satisfied are also suggested by Jardine and Sibson, but single-linkage turns out to be the only method that satisfies them all. It also has other advantages:

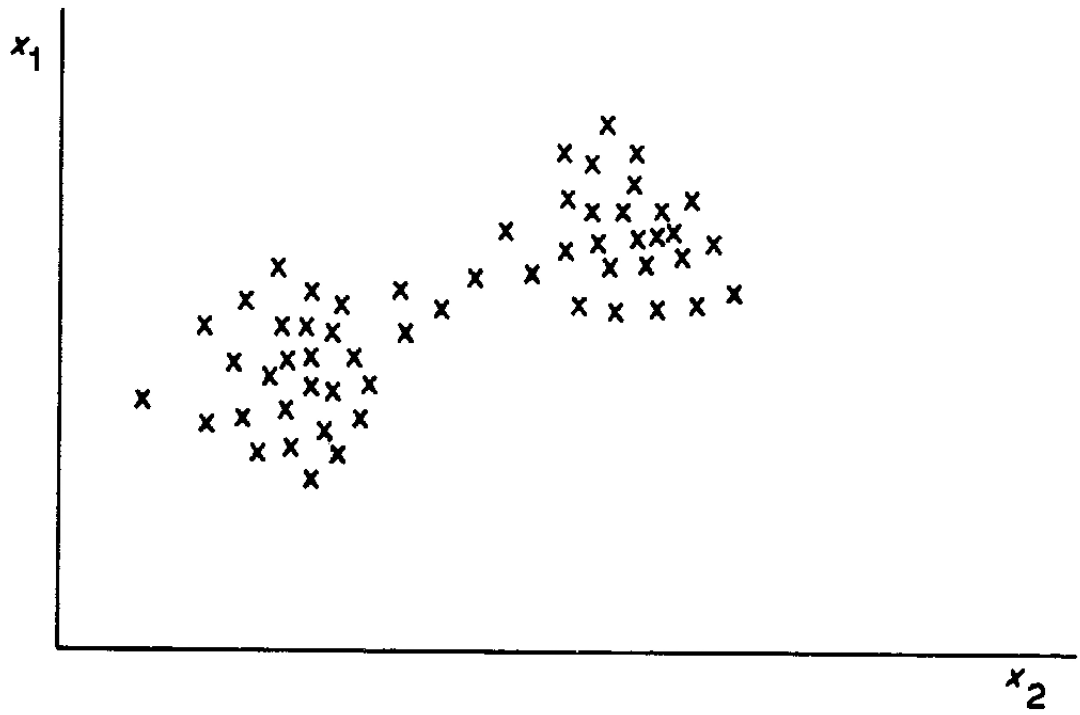
(i) Single-linkage cluster analysis is very fast and efficient to implement. There is a good algorithm - neither agglomerative nor divisive - which can cope with large amounts of data.

(ii) The results of single-linkage are invariant under monotone transformation since only the rank order of the distances counts.

(iii) Ties are easily handled. Can choose which pair of units to join arbitrarily and it makes no difference to the final dendrogram.

<u>However</u>, single-linkage cluster analysis has one major defect - that it has a tendency to produce long strung-out clusters, the so-called chaining effect. Such clusters do not have much connectivity associated with them.

With single-linkage it is also easy for clusters to fail to be identified because of the presence of 'intermediates', as in the example below.



4.3 The Minimum Spanning Tree (MST) and it's connection with Single-Linkage Cluster Analysis

Definition: A spanning tree is a network whose nodes are the n units, whose links are straight lines joining selected pairs of units with lengths equal to the corresponding distances (dissimilarities) between units and which satisfies:

(i) All pairs of nodes are connected, either directly or via other node.

(ii) There are no circuits.

The length of a spanning tree is defined as the sum of the lengths of all its links and the minimum spanning tree (MST) is the spanning tree with minimum length.

The single-linkage dendrogram is closely related to the MST. The link lengths are the threshold distances in the dendrogram and to find clusters at any distance level δ break links whose lengths are $> \delta$.

The MST is characterised by the following property.

$$d_{rs} \geq \text{maximum of link lengths on the unique path from } r \text{ to } s$$

Use of the MST in graphical ways of depicting MV data

It is important to try to display data graphically to assess clusterings obtained. We have looked at ways of finding coordinates to represent given distances (or could use fitted distances such as the ultrametric distances from the dendrogram). However, these can only be plotted in two dimensions, and therefore it is important to find graphical methods for displaying full (or further) multivariate information for use with classical scaling, ordinal scaling, canonical variates analysis, principal components, cluster analysis etc.

Adding the MST to a 2-dimensional plot is useful in highlighting any distortions inherent in the low dimensional representation of higher order distances.

Note: There are a great many techniques in use for displaying multivariate data. For example, adding arrows to represent 3rd and 4th dimensional information is sometimes helpful.

4.4 Andrews Plots (or Harmonic Curves)

Suppose each individual r is measured on p variables. Then,

$$\mathbf{x}_r^T = (x_{r1}, x_{r2}, ..., x_{rp})$$

For each $\mathbf{x}_r$ we can define a function $f_{\mathbf{x}_r}(t)$ as

$$f_{\mathbf{x}_r}(t) = \frac{x_{r1}}{\sqrt{2}} + x_{r2}\sin(t) + x_{r3}\cos(t) + x_{r4}\sin(2t) + ...$$

Note that $\frac{1}{\sqrt{2}}$, $\sin(t)$, $\cos(t)$, $\sin(2t)$, $\cos(2t)$, ... are orthogonal functions over [-π, π]. For each individual r we can plot the function $t \rightarrow f_{\mathbf{x}_r}(t)$ over $-\pi \leq t \leq \pi$, and can put a number of curves on the same plot and compare them.

Note: Curve depends on the <u>order</u> of the variables. Low frequency differences are most easily noticeable so the 'most important' variables should be put first. e.g. use principal component scores, canonical variate scores, classical scaling axes etc.

These curves contain <u>all</u> the information in the multivariate sample.

also, note that

$$\int_{-\pi}^{\pi} [f_{\mathbf{x}_r}(t) - f_{\mathbf{x}_s}(t)]^2 dt = \pi(\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s) = \pi \times \text{squared Euclidean distance in all } p \text{ dimensions}$$

Note: For each t we have a different linear combination of the original variables and plots of the $f_{\mathbf{x}_r}(t)$ show how the individuals differ for each combination.