Module 2: Cluster Analysis

2.1 Introduction

Cluster analysis is a method for separating data into clusters or groups in a situation where no prior information about a grouping structure is available (*unsupervised classification*), as opposed to classification (*supervised classification*) where prior information about the number of groups and their individual characteristics is known and used for assigning new units to groups.

The goal of a cluster analysis is that units in a cluster should be as similar as possible, and clusters should also be as different as possible. Do not expect that a cluster analysis will always divide a data set into nicely separated clusters.

The main reasons for doing a cluster analysis are

data exploration,

visualisation,

data reduction,

hypothesis generation.

Cluster analysis is not a statistical method involving any probability distributions, and therefore the resulting solution cannot be assigned any measures of uncertainty. There are two main approaches towards cluster analysis:

(1) visualisation,

(2) clustering algorithms.

The two approaches can also be used in combination. The next two sections will go into more details with these approaches using a small example data set as illustration.

2.2 Data formats

Assume that data is a collection of information about n subjects or units. There are two common formats in which the data can be given, both of which involves the notion of a *matrix*. The data format can either be a data matrix or a dissimilarity matrix.

Data consisting of measurements obtained for each unit can represented by a data matrix, denoted X, which is a rectangular array with numbers arranged in columns and rows. For instance, the data matrix containing measurements on p variables for each of the n units has the form

$$X = \begin{bmatrix} X_{11} & \dots & \dots & X_{1p} \\ \vdots & \ddots & & \vdots \\ & & & \\ \vdots & & \ddots & \vdots \\ X_{n1} & \dots & \dots & X_{np} \end{bmatrix}.$$

The element in row i and column j, at position (i, j), is denoted X_{ij}. By convention the number of subjects is equal to the number of rows (n) whereas the number of variables is equal to the number of columns (p). We only consider techniques for clustering units (not clustering of variables) for quantitative measurements.

Data consisting of measures of dissimilarity between all pairs of two units can be represented using a dissimilarity matrix D of the form

$$D = \begin{bmatrix} D_{11} & \dots & \dots & D_{1n} \\ \vdots & \ddots & & \vdots \\ & & & \\ \vdots & & \ddots & \vdots \\ D_{n1} & \dots & \dots & D_{nn} \end{bmatrix}.$$

The elements in D are called dissimilarities. Only elements outside the diagonal are of interest because they correspond to dissimilarities between pairs of different units.

If data is represented in a data matrix X and they are quantitative then the dissimilarity matrix D can be constructed by means of a distance measure, often called a *metric*. The most commonly used distance measure is the Euclidean distance which is the sum of the squared differences between pairs of measurements. For units i and j with rows $X_i = (X_{i1}, \dots, X_{ip})$ and $X_j = (X_{j1}, \dots, X_{jp})$ in X, respectively, the Euclidean distance is

$$D_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}. \tag{2.1}$$

Another distance measure is the city-block distance defined by the formula

$$D_{ij} = \sum_{k=1}^{p} |X_{ik} - X_{jk}|$$

which measures distance as travelled if one only walks in the directions of the coordinate axes.

Construction of a dissimilarity matrix from a distance measure yields a symmetric matrix: the entries D_{ij} and D_{ji} are equal for all i and j between 1 and n. Neither the Euclidean distance nor the city-block distance may be appropriate if the variables are measured in different units. Dissimilarity matrices need not always be the result of converting a data matrix using a distance measure. Sometimes the observed data is a set of dissimilarities. Examples being pairwise comparisons of a food product or measurements obtained as results of a pairs of treatments. As a consequence the dissimilarity matrices need not be symmetric, that is D_{ij} need not be equal to D_{ji}. Thus dissimilarities are not necessarily distances.

In order for the above distance measure to work the variables need to be on the same scale, as otherwise one or more variables (the ones having the largest numerical values) will dominate in the calculation of the distance. However, in cases where variables are not measured on comparable scales two other distance measures may be useful: The standardised Euclidean distance is defined as

$$D_{ij} = \sqrt{\sum_{k=1}^{p} \frac{(X_{ik} - X_{jk})^2}{s_k^2}}$$

where $s_1^2, \ldots, s_p^2$ are the empirical variances of the p variables. Mahalanobis distance is defined as follows in matrix notation

$$D_{ij} = (X_i - X_j)^t S^{-1} (X_i - X_j)$$

where X_i and X_j are the vectors $(X_{i1}, \ldots, X_{ip})$ and $(X_{j1}, \ldots, X_{jp})$, respectively and S the empirical variance-covariance matrix of the p variables. Both distance measures are *scale invariant*: Multiplication of all values of a variable does not change the distances.

2.2.1 Example: temperature and yield

The tiny example introduced in this sub-section will serve as illustration of the clustering techniques discussed in this chapter.

Consider a data set consisting of measurements of the variables `temperature` and `yield` for 8 different geographical locations. The data set used is represented by the 8×2 data matrix

$$\begin{bmatrix} 24.0 & 113 \\ 25.1 & 109 \\ 24.7 & 124 \\ 25.5 & 106 \\ 31.9 & 156 \\ 30.3 & 167 \\ 29.6 & 159 \\ 30.8 & 168 \end{bmatrix}.$$

Interest in this data set lies in finding groups of locations that exhibit similar features as expressed by the two variables `temperature` and `yield`.

The dissimilarity matrix derived from the data matrix using Euclidean distance is

$$\begin{bmatrix} 0 & & & & & & & \\ 4.15 & 0 & & & & & & \\ 11.02 & 15.01 & 0 & & & & & \\ 7.16 & 3.03 & 18.02 & 0 & & & & \\ 43.72 & 47.49 & 32.80 & 50.41 & 0 & & & \\ 54.37 & 58.23 & 43.36 & 61.19 & 11.12 & 0 & & \\ 46.34 & 50.20 & 35.34 & 53.16 & 3.78 & 8.03 & 0 & \\ 55.42 & 59.27 & 44.42 & 62.23 & 12.05 & 1.12 & 9.08 & 0 \end{bmatrix}.$$

The entries in the dissimilarity matrix are calculated using the formula (2.1). The diagonal entries are all 0 as they are calculated as

$$\begin{aligned} D_{ii} = & \sqrt{\sum_{k=1}^{2}(X_{ik} - X_{ik})^2} = \sqrt{(X_{i1} - X_{i1})^2 + (X_{i2} - X_{i2})^2} \\ = & \sqrt{0^2 + 0^2} = 0. \end{aligned}$$

Consider the off-diagonal entries below the diagonal. For example the entry 4.15 in row 2 and column 1 is obtained from the calculation

$$\begin{aligned} D_{21} = & \sqrt{\sum_{k=1}^{2}(X_{2k} - X_{1k})^2} = \sqrt{(X_{21} - X_{11})^2 + (X_{22} - X_{12})^2} \\ = & \sqrt{(25.1 - 24.0)^2 + (109 - 113)^2} = \sqrt{1.1^2 + 4^2} \\ = & \sqrt{17.21} \\ = & 4.15. \end{aligned}$$

The calculation of D_{12} also gives the result 4.15. Thus the equality $D_{21} = D_{12}$ holds. This is true for D_{ij} and D_{ji} for for all pairs of i and j between 1 and 8, showing that the dissimilarity matrix is symmetric. The symmetry is the reason why only entries below the diagonal are displayed; the entries above the diagonal would be the same.

2.3 Visual approach: Scatter plots

The visual approach is simply to inspect some plot of the data matrix. Even though this approach contains a subjective element in the sense that it relies on subjective assessment, it is a useful approach, and it could be used whenever a sensible display of the data matrix is possible. Depending on the form of the clusters the visual approach may be superior to the algorithmic approach, as many algorithms (as we shall se in the next section) impose some kind of structure on the clusters.

In addition to the scatter plots it may be useful to plot histograms of the individual variables to assess the ranges of the variables.

2.3.1 Two dimensions

In this case the data matrix X is a $n \times 2$ matrix, having n rows and 2 columns, corresponding to measurements of two variables x_1 and x_2, say, for n subjects. A data matrix with only two columns can be displayed graphically by means of a scatter plot where the pairs (x_{1i}, x_{2i}) for $i = 1, \ldots, n$ constitute the points.

It may be difficult to see and convince oneself that there are no clusters in a scatter plot. Beware of spurious clustering. Having said this, sometimes the objective of a cluster analysis may be to obtain an ad hoc clustering which serves more as crude grouping of the data in order to reduce the dimensionality of the data than an actual identification of clusters.

2.3.2 More than two dimensions

Two-dimensional plots of all pairs of variables may sometimes be useful, but may sometimes also fail to reveal important features in the data, for instance in cases where the structure in the data matrix is pertaining to a coordinate system that differs from the usual coordinate system (maybe coordinate axes are not orthogonal). Alternatively, a principal components analysis (PCA) could be applied and then a scatter plot of the scores of the first two principal components could be constructed and a two-dimensional scatter plot could be used to detect clusters. This approach may be particularly useful if the the first two components explain a large amount of the total variation in the data. If more than two components are needed to capture a substantial part of the variation, an algorithmic approach could supplement the PCA: The algorithmic approach would then be based on the principal components rather than the original variables.

2.3.3 Example: temperature and yield (continued 1)

A scatter plot of the 8×2 data matrix is seen in Figure 2.1. The plot is constructed by representing the 8 points

$$(24.0, 113), \quad (25.1, 109), \quad (24.7, 124), \quad (25.5, 106),$$
$$(31.9, 156), \quad (30.3, 167), \quad (29.6, 159), \quad (30.8, 168)$$

in a two-dimensional coordinate system, letting each variable correspond to an axis.

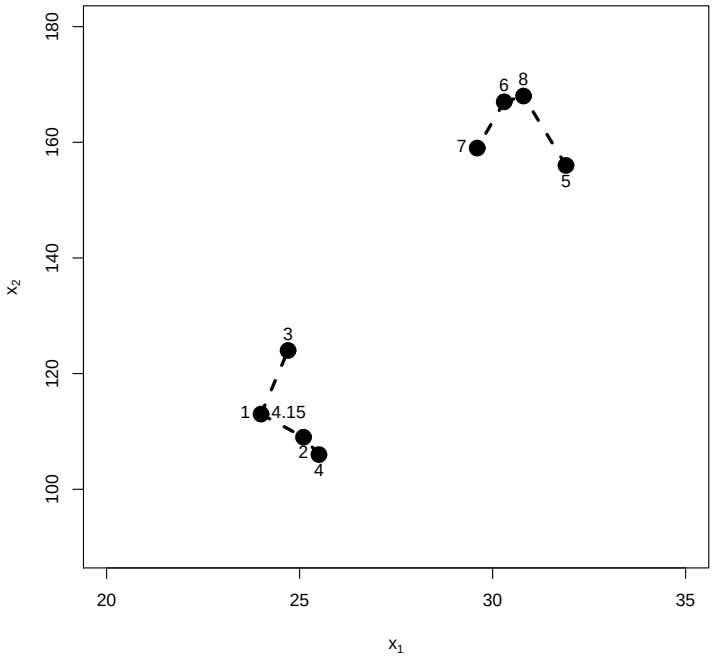

Figure 2.1: Scatter plot of data matrix. Units numbers 1–8 and the distance D_{21} are displayed.

Two clusters are visible: They are clearly separated and the units forming the two cluster are not too scattered, forming relatively homogeneous clusters. Within the two clusters in Figure 2.1 the distances between points that are closest are indicated with dashed lines. The distance between units 1 and 2, which is 4.15, is added to the plot.

2.4 Algorithmic approach

In contrast to the visual approach the algorithmic approach replaces the visual assessment of dissimilarity between subjects by a numerical algorithm which uses a measure of dissimilarity, that is dissimilarities expressed in terms of numbers.

There are (at least) two types of methods: hierarchical and non-hierarchical. A hierarchical method produces a cluster solution which invariably contains the entire range of clusters, from n clusters to 1 cluster, and by some additional method the number of clusters has to be determined. Moreover once a subject has been assigned to a cluster, it cannot be re-assigned to another cluster later on. A non-hierarchical method need to have specified in advance how many clusters there are and then it proceeds to assign or re-assign subjects to clusters in subsequent steps; cluster membership may change from step to step. We will only discuss hierarchical clustering

The algorithmic approach is not an entirely objective approach, as it is not devoid subjective decisions as will be illustrated in subsequent subsections. On the other hand the dimensionality of the data (the number of columns in the data matrix) does not limit its applicability as is the case for the visual approach.

It is worth emphasising that algorithmic clustering methods are depending heavily on the chosen dissimilarity matrix and, consequently, the outcome may also depend on the choice of dissimilarity matrix. Therefore some care and consideration should be put into finding a reasonable dissimilarity matrix. The standardised Euclidean distance or Mahalanobis distance could be used.

2.4.1 Hierarchical clustering methods

A basic algorithm for finding the hierarchical clustering will be described. The nature of the algorithm is agglomerative: First there are as many clusters as there are subjects, then the two clusters which are closest are joined, repeatedly, and eventually all subjects have been joined in a single cluster.

In more detail the algorithm consists of the following steps

1 Define each subject as a group with dissimilarity D_{ij} between subjects i and j.

2 Find smallest element in D, D_{km} ($k \neq m$) say. Join the groups k and m into the group $\{k, m\}$ (the two closest groups are united).

3 Calculate the dissimilarity between the new group consisting of $\{k, m\}$ and the remaining groups. Form a new dissimilarity matrix.

4 Go through steps 2 and 3 until all subjects have been joined.

There are many rules for deciding how the combined units should be treated. The rules are all based on the notion *linkage*, the features of the two groups joined carried over to the union of the groups (how are the groups linked together?).

We will consider three methods for calculating the dissimilarity of a union of groups.

- single linkage
- complete linkage
- average linkage

The single-linkage method, also called nearest-neighbour method, calculates the dissimilarity between the new group $\{k, m\}$ and any other group l as the minimum of the two dissimilarities D_{kl} and D_{ml}. This form of linkage means that a single link is enough to join to groups, and this feature will allow clusters to be elongated and not necessarily spherical.

The complete-linkage method or furthest-neighbour method takes the maximum of the two dissimilarities D_{kl} and D_{ml} as the new dissimilarity. Complete-linkage implies that clusters are formed on the basis of the maximum dissimilarities between the clusters, and this tend to produce spherical clusters.

In the average linkage method, also called unweighted pair-group method using arithmetic averages (UPGMA), the dissimilarity between groups k/m and l is defined as the average of the two dissimilarities D_{kl} and D_{ml}. The average linkage method also leads to spherically-shaped clusters.

2.4.2 Dendrograms

The cluster solution from a hierarchical clustering method can be displayed in a two-dimensional plot, in a dendrogram or tree diagram. A tree is a collection of clusters such that any two clusters either are disjoint or one contains the other. The largest cluster, which is the set of all n units, contains all other clusters, and the smallest clusters are the individual units.

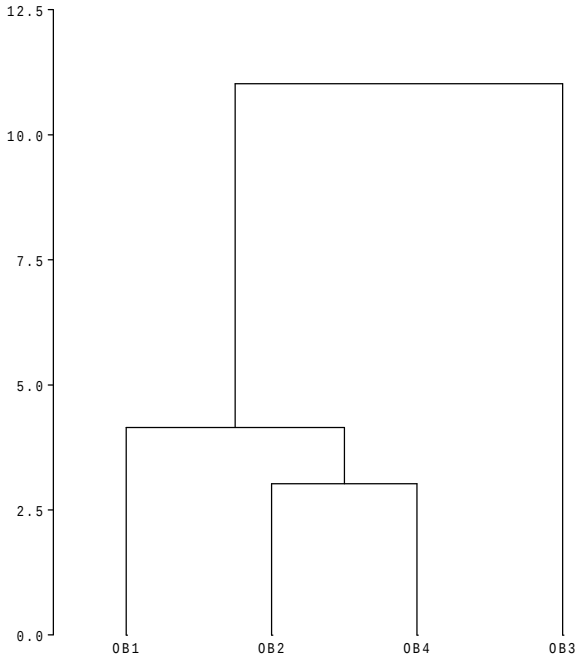

Figure 2.2: A dendrogram.

The dendrogram is the two-dimensional representation of the tree. Usually it is displayed with the largest cluster (containing all units) at the top and the smallest clusters (the units) at the bottom, being a tree up-side-down; see Figure 2.2. The dendrogram provides an overview on the cluster solution, and any horizontal cut of the dendrogram will yield a number of clusters. For instance, a cut a 7.5 on the y axis in Figure 2.2 will divide the four units into two clusters (one cluster with one unit and one cluster with three units), whereas a cut at 3.5 will result in three clusters (two clusters with one unit each and one cluster with two units). The dissimilarities used in the clustering method correspond to the lengths of the edges. Therefore the dendrogram provides a way to determine the number of clusters by looking at the heights. A particular clustering is obtained by cutting the tree at some specific height. Large heights may suggest reasonably well-separated clusters, but deciding where to cut the tree still contains an element of subjective judgment.

2.4.3 Determining the number of clusters

As already mentioned a hierarchical clustering method yields a cluster solution which contains clustering with any number of clusters between 1 and n. In addition to the visual assessment using the dendrogram some statistics are available for determining the number of clusters. We consider two statistics: The root-mean-square standard deviation (RMSSTD) and R-square (RS). RMSSTD is a measure of homogeneity within clusters, and RS indicates the extent to which clusters are different from each other.

The RMSSTD statistic is the value of

$$\sqrt{\frac{SS_1 + \ldots + SS_p}{df_1 + \ldots + df_p}}$$

that is the pooled standard deviation of all variables. The term SS_j is the within sum of squares of the jth variable and is calculated using the formula

$$SS_j = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2.$$

Large values of RMSSTD indicates that the clusters are not that homogeneous.

The RS statistics is the usual R square known from ANOVA models

$$SS_{between}/SS_{total}.$$

Both $SS_{between}$ and SS_{total} are defined in terms of sums of squares, as the total between sum of squares and total sum of squares. The total sum of squares is defined as

$$SS_{total} = \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$$

and $SS_{between}$ is the difference between SS_{total} and the total within sum of squares, that is the sum of all SS_j's for $j = 1, \ldots, p$.

The quantity $SS_{between}$ is a measure of the variation between clusters whereas SS_{total} is a measure of the total variation. The values of RS lies between 0 and 1 with values close to 1 indicating high difference between clusters.

There are no general rules available for assessing whether or not values of the statistics RMSSTD and RS are small or large, but the relative changes in the values of the statistics as the number of clusters increase can be useful in determining the number of clusters. Calculation of the statistics at each stage in the clustering algorithm, that is for each numbers

of clusters, allows plotting the values against the number of clusters. A marked decrease or increase for RMSSTD and RS ("an elbow" in the plot), respectively, may indicate that a satisfactory number of clusters have been reached.

2.4.4 Example: temperature and yield (continued 2)

Consider the single linkage method to the example data set. This initial dissimilarity matrix (given in Section2.2) is

Group	1	2	3	4	5	6	7	8
1	0							
2	4.15	0						
3	11.02	15.01	0					
4	7.16	3.03	18.02	0				
5	43.72	47.49	32.80	50.41	0			
6	54.37	58.23	43.36	61.19	11.12	0		
7	46.34	50.20	35.34	53.16	3.78	8.03	0	
8	55.42	59.27	44.42	62.23	12.05	1.12	9.08	0

The steps in the algorithm are as follows: The dissimilarity matrix for the individual units is already calculated, hence Step 1 of the algorithm is done. The single linkage method uses the minimum distance, and therefore Step 2 amounts to inspection of the dissimilarity matrix in search for the smallest positive number (not 0). This number is $D_{86} = 1.12$, the distance between units 6 and 8. Consequently we join the units 6 and 8 into one group $\{6, 8\}$.

In Step 3 the dissimilarities between the group $\{6, 8\}$ and the remaining units are calculated as the minimum of the distances between the group $\{6, 8\}$ and the remaining units

$$
\begin{aligned}
D_{\{6,8\}1} &= \min\{D_{61}, D_{81}\} = D_{61} = 54.37, \\
D_{\{6,8\}2} &= \min\{D_{62}, D_{82}\} = D_{62} = 58.23, \\
D_{\{6,8\}3} &= \min\{D_{63}, D_{83}\} = D_{63} = 43.36, \\
D_{\{6,8\}4} &= \min\{D_{64}, D_{84}\} = D_{64} = 61.19, \\
D_{\{6,8\}5} &= \min\{D_{65}, D_{85}\} = D_{65} = 11.12, \\
D_{\{6,8\}7} &= \min\{D_{67}, D_{87}\} = D_{67} = 8.03.
\end{aligned}
$$

Dissimilarities between the remaining units are left unchanged. The new dissimilarity matrix has the following form

Group	1	2	3	4	5	$\{6,8\}$	7
1	0						
2	4.15	0					
3	11.02	15.01	0				
4	7.16	3.03	18.02	0			
5	43.72	47.49	32.80	50.41	0		
$\{6,8\}$	54.37	58.23	43.36	61.19	11.12	0	
7	46.34	50.20	35.34	53.16	3.78	8.03	0

Step 4 leads back to Step 2 and Step 3, but now using the new dissimilarity matrix (which two groups are joined next?). Once all units have been joined into a single group the clustering algorithm stops.

The minimum distances used in the algorithm to join groups (single linkage) are

$$1.12, 3.03, 3.78, 4.15, 8.03, 11.02, 32.8.$$

These minimum distances are the heights of the clusters in the corresponding dendrogram which is given in Figure 2.3. The scale on the y-axis is minimum distances.

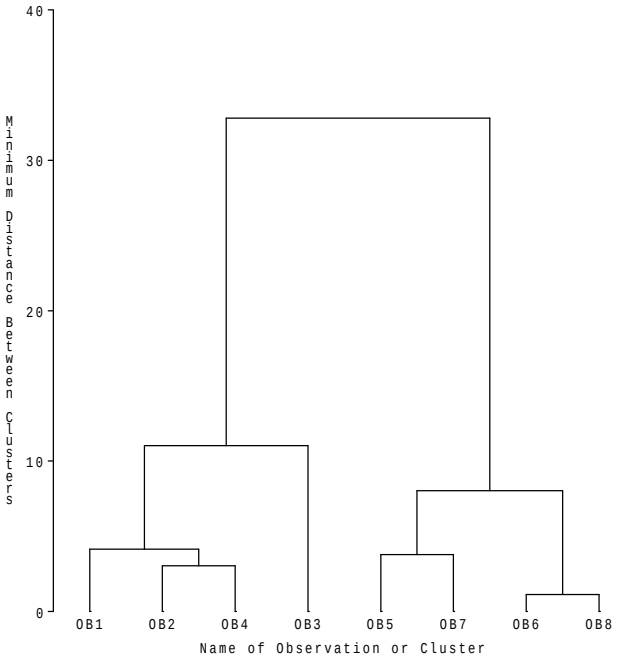

Figure 2.3: Dendrogram of a single-linkage Euclidean distance based cluster solution.

For example at height 1.12 we find a horizontal line segment joining the two units 6 and 8, and at height 32.8 we find the cluster containing all units. The two clusters previously identified on the the scatter plot are clearly visible as there are two edges with large heights.

Note that the dendrogram is not unique as there are several ways to arrange the clusters. In the above dendrogram subject 1 and subjects 2 and 4 could change places without changing the clustering structure.

In order to have some measure to help deciding on or help confirming our impression about the number of clusters, the RMSSTD and RS statistics can be employed.

Consider again the cluster obtained by the first step of the clustering algorithm (joining units 6 and 8, yielding 7 clusters). The value of RMSSTD for 7 clusters is calculated using the formula

$$\text{RMSSTD}(7) = \sqrt{\frac{SS_{temperature} + SS_{yield}}{df_{temperature} + df_{yield}}}$$

with sums of squares based on the units joined (units 6 and 8) only. Insertion the relevant numbers yields

$$= \sqrt{\frac{[(30.3 - 30.55)^2 + (30.8 - 30.55)^2] + [(167 - 167.5)^2 + (168 - 167.5)^2]}{1 + 1}}$$

$$= \sqrt{(0.125 + 0.5)/2}$$

$$= 0.559$$

where 30.55 is the average of 30.3 and 30.8 and 167.5 the average of 167 and 168. Figure 2.4 displays the values of RMSSTD plotted against number of clusters.

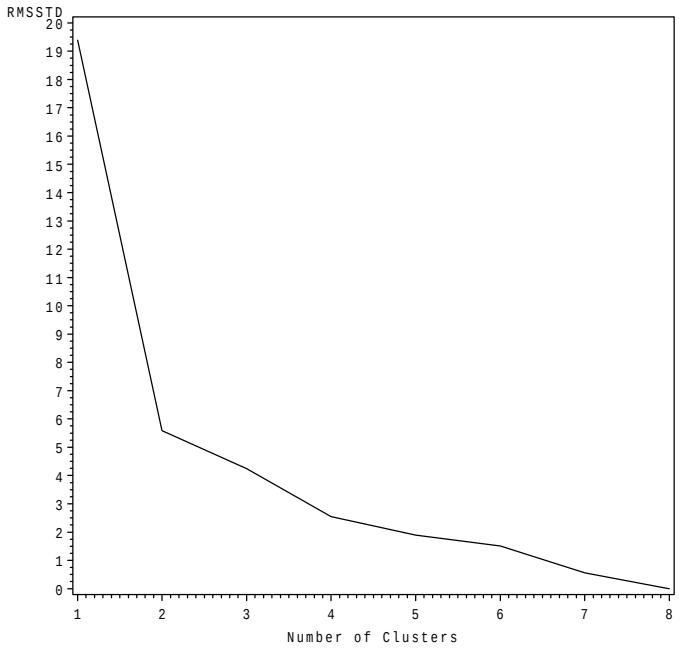

Figure 2.4: RMSSTD as a function of the number of clusters.

The great change in RMSSTD occurs when going from 1 to 2 clusters, indicating that there are two clusters because the additional decrease from having 3 clusters is not that large compared to the decrease from 1 to 2 clusters.

The value of the RS statistic based on the first step in the clustering algorithm is calculated using the (within) sums of squares for each of the two variables for cluster consisting of units 6 and 8. The sums of squares were already calculated when calculating the RMSSTD statistic for 7 clusters above; the sums of squares are $SS_{temperature} = 0.125$ and $SS_{yield} = 0.5$. The total within sum of squares for the two clusters is obtained by adding the 2 sums of

squares, giving 0.625. The total sum of squares SS_{Total} is equal to 5263.4. The total between sum of squares is obtained by subtracting SS_{within} from SS_{total}.

$$\text{RS}(7) = \frac{SS_{between}}{SS_{total}} = \frac{5263.4 - 0.625}{5263.4} = 0.99988$$

The plot of RS against number of clusters (Figure 2.5) shows that going from 1 to 2 clusters yields a large gain as regards how well the clusters are separated. Additional clustering does not produce any marked increase in R.

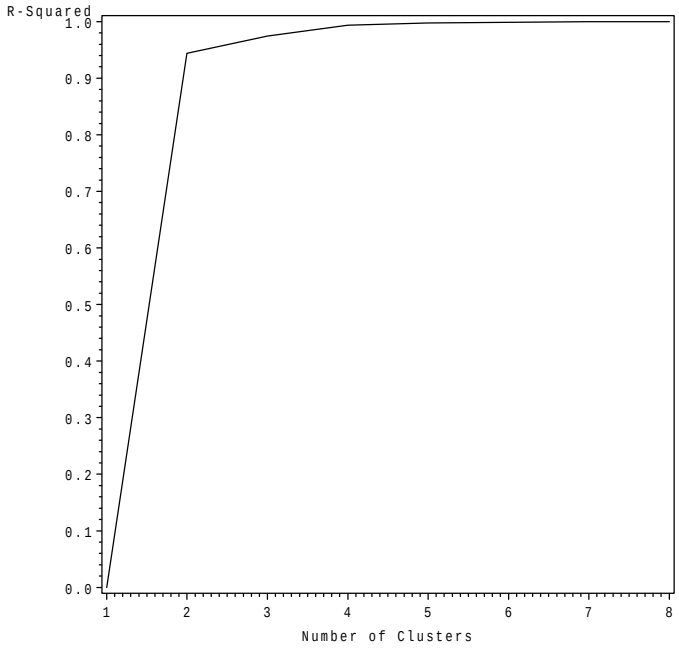

Figure 2.5: RS as a function of the number of clusters.

2.4.5 Choice of method

The features of the three methods introduced in this section (single, complete and average linkage) may serve as a guideline for choosing the right method: What kind of clusters do you expect? Spherical or non-spherical clusters? Are clusters well-separated or is chaining possible? In absence of such expectations one approach would be to try several methods and compare the results. The dendrogram and the RMSSTD and RS statistics may be helpful in determining the numbers of clusters. For a given number of clusters comparisons can be based on both visual approaches and algorithmic approaches by investigating whether or not clusters are identical with respect to the units they contain.

There are various other hierarchical clustering methods that also could be applied and compared to the results of the above methods; see the references in the next section for more details.

One way of comparing the results of clustering methods is to examine whether one method is giving a more interpretable result than the another methods. External validity of the result is obtained by relating the result to some external, independent source of information. This could be relevant expert assessments, previous analyses of similar data or subject-specific theoretical considerations.

Another way of validating the cluster solution is to permute the rows and/or the columns of the data matrix, to obtain cluster solutions of the permuted data matrices and compare these to the original cluster solution. See Eisen *et al.* (1998) for an example.

2.5 More on cluster analysis

The books by Chatfield and Collins (1980) and Krzanowski (1988) both cover cluster analysis in some detail, from a statistical perspective. Sharma (1996) provides a detailed discussion of practical aspects of cluster analysis with examples in SAS.

This chapter only dealt with visualisation given in a data matrix. Data in the form of a dissimilarity matrix can also be visualised using a technique called multidimensional scaling where a configuration of points whose mutual positions reflect the dissimilarities between units. In other words we want to construct a map from the distances between a set of locations. For this to work the dissimilarities should be interpretable as Euclidean distances, at least approximately.

Non-hierarchical clustering methods require a priori specification of the number of clusters and the corresponding cluster means and then iteratively assigns units to the closest clusters according to some reassignment rule. This process continues until clusters only change minimally. Used in itself non-hierarchical methods perform poorly (how to choose cluster means), but used in combination with a hierarchical method which is used to provide the initial number of clusters and the cluster means the methods are useful. Hierarchical and non-hierarchical methods should be seen as complementary approaches.

Recent advances in computer-aided graphical methods have produced several ways to visualise high-dimensional data, by displaying data from different perspectives in a three-dimensional space Swayne *et al.* (1998). These methods extend the visual approach described in 2.2 and they may sometimes be better than the classical clustering techniques discussed in this chapter, but their use requires some experience.