Capítulo 8

CLUSTER ANALYSIS

John Tukey (1915-2000)

American statistician. Creator of the fast Fourier transform and one of the pioneers in diagnostic methods and robust estimation. He coined the terms software, bit, boxplot and many others. He was the founding director of the Department of Statistics at Princeton, and combined his teaching with work at AT&T Bell Laboratories. His contributions ushered in a new era of statistics in the second half of the 20th century.

8.1 BACKGROUND

The purpose of cluster analysis is to group items into homogeneous groups according to shared properties or similarities. Normally, observations are grouped, but cluster analysis can also be applied to the grouping of variables. The methods are also known as automatic or unsupervised classification methods and unsupervised pattern recognition. The term unsupervised is applied in order to differentiate it from discriminant analysis which we will study in Chapter 13. Cluster analysis studies three types of problems:

Data clustering. We have what we suspect is a heterogenous sample and would like to partition it into a number of groups so that:

(1) each item belongs to one, and only one, of the groups;

(2) all items are classified;

(3) each group is internally homogenous.

For example, we have a database of purchases and wish to create a typology of customers according to their standards when purchasing .

Construction of hierarchies. We wish to structure the items in a set hierarchically by similarity. For example, we have a survey on attributes of different professions and we want to order them by similarity. A hierarchical classification implies that the data is ordered by levels such that the lower levels are contained in the higher levels. This type of classification is quite frequent in biology, in the classification of animals, plants, and so on. These methods do not define groups per se, rather they define the structure of a series of associations which can exist between the items. Nevertheless, as we will see, the constructed hierarchy also provides a partition of the data into clusters.

Classification of variables. In problems with a large number of variables, it is useful to carry out an initial study in order to partition the variables into groups. This study can guide us in considering the formal models for reducing dimensions which will be studied later. Variables can be classified in groups or structured in a hierarchy.

The methods of partition use a data matrix, but the hierarchical algorithms use a distance or similarity matrix. When grouping variables, the starting point is the matrix of the relationship between variables: for continuous variables, the correlation matrix is usually used. With discrete variables, as we will see, a Chi-square distance is used.

First we will look at methods of clustering.

8.2 CLASSICAL METHODS OF DATA CLUSTERING

8.2.1 Fundamentals of the k-means algorithm

Suppose we have a sample of n items with p variables. The objective is to partition this sample into prearranged groups, G. The k-means algorithm (which with our notation should be of G-means) requires the following four steps.

(1) Choose G points as centroids of the initial groups. This can be done by:

a) randomly assigning objects to groups and taking the centroids of the groups formed in this way;

b) taking as the centroids the G points farthest from each other;

c) constructing initial groups using a priori information and calculating their centroids, or selecting their centroids a priori.

(2) Calculate the Euclidean distances of each item to the centroid of the G groups, and assign each item to the closest distance. The assignment is carried out sequentially and when a new item is introduced the coordinates are recalculated from the new centroid of the group.

(3) Define a criterion of optimality and check whether reassigning some of the items improves the criterion.

(4) When it is no longer possible to improve the optimality criterion, stop the process.

8.2.2 Implementation of the algorithm

The criterion of homogeneity used in the k-means algorithm is *the sum of squares within groups (SSW)* for all variables, which is equivalent to the weighted sum of the variances of the variables in the groups:

$$SSW=\sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_g}(x_{ijg}-\overline{x}_{jg})^2 \tag{8.1}$$

where x_{ijg} is the value of variable j in the item i of group g and $\overline{x}_{jg}$ is the mean of this variable in the group. The criterion is written as

$$\min SSW = \min \sum_{g=1}^{G}\sum_{j=1}^{p} n_g s_{jg}^2 \tag{8.2}$$

where n_g is the number of items in group g and s_{jg}^2 is the variance of variable j in said group.

The variances of the variables in the groups are clearly a measure of heterogeneity in the classification, and by minimizing them, we obtain more homogenous groups. An alternative criterion would be to minimize the squared distances between points and their group centroids. If we measure the distances using a normal Euclidean distance, this criterion is written as:

$$\min \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\mathbf{x}_{ig}-\overline{\mathbf{x}}_g)'(\mathbf{x}_{ig}-\overline{\mathbf{x}}_g) = \min \sum_{g=1}^{G}\sum_{i=1}^{n_g} d^2(i,g)$$

where $d^2(i,g)$ is the squared Euclidean distance between item i of group g and the mean of the group. It is easy to prove that the two criteria are identical. Since a scalar is equal to its trace, we can write this latest criterion as

$$\min \sum_{g=1}^{G}\sum_{i=1}^{n_g} tr\left[d^2(i,g)\right] = \min tr\left[\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\mathbf{x}_{ig}-\overline{\mathbf{x}}_g)(\mathbf{x}_{ig}-\overline{\mathbf{x}}_g)'\right],$$

and, letting $\mathbf{W}$ be the matrix of the sum of squares between groups,

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$$

we get

$$\min tr(\mathbf{W}) = \min SSW$$

and the two criteria coincide. This criterion is also called the *trace criterion* and was proposed by Ward (1963).

Maximization of this criterion would entail calculating it for all possible partitions, a clearly impossible task, except with very small values of n. The k-means algorithm looks for an optimal partition with the constraint that in each repetition you can only move one item from one group to another. The algorithm works in the following way:

(1) Start with an initial allocation.

(2) Check whether moving an item reduces SSW.

(3) If it is possible to reduce SSW by moving an item, then do so. Recalculate the means of the two groups affected by the change and go back to (2). If it is not possible to reduce SSW, stop the process.

Consequently, the result of the algorithm can depend on the initial allocation and order of the items. The algorithm should always be repeated from different starting values and permuting the items in the sample. The effect of the order is usually small, but it is advisable to make sure in each case.

Trace criterion has two important properties. The first is that it is not invariant with changes of scale. When the variables are in different units, it is better to standardize them so that the result of the k-means does not depend on irrelevant changes in the scale of the measurement. When they are in the same units, it is usually better not to standardize as a variance which is much larger than the rest is probably due precisely to the fact that there are two observation groups in this variable, which may be hidden if standardized. For example, Figure 8.1 illustrates how standardization can obstruct the identification of the groups.

Figura 8.1: Standardization can obstruct the identification of the groups.

The second property of trace criterion is that minimizing the Euclidean distance produces approximately spherical groups. The reasons for this will be dealt with in Chapter 15. On the other hand, this criterion implies quantitative variables and, although it can be applied if there is a small number of binary variables, when there are many attributes, it is better to use the hierarchical methods which will be seen next.

8.2.3 Number of groups

In the usual application of the k-means algorithm, the number of groups G must be established. It is clear that this number cannot be estimated using a homogeneity criterion since the method for getting homogenous groups and minimizing the SSW is to make as many groups as there are observations. In this way, $SSW = 0$ always. Different methods have been suggested for selecting the number of groups. One approximate procedure which is used frequently (though without much justification), is to carry out an approximate F test of variability reduction, comparing the SSW with G and $G+1$ groups, and calculating the relative reduction of variability with the increase of an additional group. The test is:

$$F = \frac{SSW(G) - SSW(G+1)}{SSW(G+1)/(n-G-1)} \tag{8.3}$$

and compares the decrease in variability to the increase of one group using the average variance. The value of F is compared using an F with p, and $p(n-G-1)$ degrees of freedom, but there is not much justification for this rule because the data do not need to verify the necessary hypothesis in order to apply the F distribution. An empirical rule which provides reasonable results is that suggested by Hartigan (1975), and used in several computer programs, which is to introduce an additional group if this quotient is greater than 10.

Example: Figure 8.2 shows the ruspini data (ruspini.dat file) which include 75 pieces of data on two variables and has been used to compare different classification algorithms. The graph clearly shows four groups of data in two dimensions.

Figura 8.2: Ruspini data

Table 8.1 shows the result of applying the k-means algorithm to the unstandardized data for different numbers of groups using MINITAB. According to the F-criterion there are three groups in the data. Figures 8.3, 8.4, 8.5 and **??** show the groups obtained using the program.

The table was constructed based on information provided by the program. When going from 2 to 3 groups there is a significant reduction in variability given by

$$F=\frac{89247-51154}{51154/(75-4)}=52.87$$

Nevertheless, the reduction is not significant when going from 3 to 4 groups

$$F=\frac{51154-50017}{50017/(75-5)}=1.59.$$

The k-means algorithm installed in MINITAB could divide the data into the three groups indicated in Figure 8.4. If we apply the algorithm to the standardized data three groups are obtained again, but they are different: the first is made up of two sets of points situated in the upper part of the graph and the other two groups from the two lower sets.

Number	size	SSW(i)	SSW	F
2	34	43238		
	40	46009	89247	
3	20	3689		
	40	46009		
	15	1456	51154	52.8
4	4	170		
	16	2381		
	15	1456		
	40	46009	50017	1.59
5	4	170		
	5	292		
	11	857		
	40	46009		
	15	1456	48784	

Tabla 8.1: Table containing the information for choosing the number of groups using the k-means algorithm.

Figura 8.3: Division of the Ruspini data into two groups using Minitab.

To study how the different programs work we have applied the same analysis to these data using the SPSS k-means program. The partition into two groups is the same with both programs, but the partition into three and four groups differs as shown in Figures 8.6, 8.7 and **??**. The SPSS program produces better results than MINITAB. This example suggests that before accepting the results of a cluster analysis using the k-means algorithm it is advisable to try different starting points and algorithms.

Example: We are going to apply the k-means algorithm to the country data using only

Figura 8.4: Division of the Ruspini data into three groups using Minitab.

Figura 8.5: Division of the Ruspini data into four groups using Minitab.

Figura 8.6: Division of the Ruspini data into three groups using SPSS.

Figura 8.7: Division of the Ruspini data into four groups using SPSS.

	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$
lem	30	20	14	15	14
lew	35	22	13	16	12
imr	509	230	129	76	83
mr	15	11	9	9	9
br	64	58	37	35	26
Total=MS(G)	653	341	202	151	144
F		82.4	61.5	30.4	6.2

Tabla 8.2: Table containing the information for choosing the number of groups using the k-means algorithm.

5 demographic variables from the MUNDODES data. We will start by commenting on the results obtained when using the k-means with SPSS. To decide the number of groups this program gives us the average variance within groups for each variable. For example, if $G = 2$, two groups, the second column of Table 8.2 indicates that the average variance or unexplained variance within the two groups for the variable *lem* (life expectancy for men) is 30, for the variable *lew* (life expectancy for women) it is 35, and so on, successively. This term is calculated as follows: for each variable we carry out a decomposition of the analysis of variance of its total sum of squares $\sum(x_{ij}-\overline{x})^2$ in the explained variance, $\sum(\overline{x}_i-\overline{x})^2$, where $\overline{x}_i$ is the mean of the variable in each group, and the unexplained, $\sum(x_{ij}-\overline{x}_i)^2$. This last term divided by its degrees of freedom, which are $n-G$, give us the average variance within the groups or the unexplained. According to the definition, the sum of these variances multiplied by $n-G$ gives the SSW statistic, as indicated in formula (8.1). Table 8.2 summarizes this information.

The table shows that, as expected, the average variances of the variables diminish as more groups are formed. The table indicates that *imr* (infant mortality rate) has much more variance than the others, and thus will have significant weight in the construction of the groups which will be principally made up of the values of this variable. The variance table shows that the number of groups is five since by increasing to six the decrease in the variances is quite small. We can test this intuition by calculating the F-test statistic given by (8.3). Letting MS(G) denote the row of totals that will be equal to SSW(G)/(n-G), this statistic is calculated as

$$F = \frac{(n-G)MS(G)-(n-G-1)MS(G+1)}{MS(G+1)}$$

where $n = 91$ and G is the number of groups indicated by columns. For example, the test for whether it is advisable to have more than two groups would be

$$F = \frac{89.653 - 88.341}{341} = 82.45$$

Thus we obtain the row of F of the table, and, according to the Hartigan criterion we would choose five groups.

As we have seen, the variable *imr* is very important in the construction of groups and Figure 8.8 presents a histogram of this variable. We see that this variable, which is going to

Figura 8.8: Histogram of the variable infant mortality rate indicating the presence of between four and five groups of countries.

have a dominant weight in the formation of the groups, clearly indicates the heterogeneity of the sample. In the constructed groups the one with the lowest infant mortality rate is three, which includes the European countries except for Albania, and the highest mortality is two, which includes the poorest countries in Africa.

Figure 8.9 illustrates the position of the two most influential variables in 5 groups in the graph and Figure 8.10 shows the composition of the groups. We see that group 3 is mainly made up of European countries, Japan and North America, group 1 includes the poorest European countries, the wealthiest Latin American countries and other countries such as China and Egypt. Group 4 encompasses medium level developing countries in Africa (South Africa and Zaire), Latin American (Brazil), Asia (India and Indonesia) and Saudi Arabia. Finally groups 2 and 5 include less developed countries.

Figura 8.9: A scatter plot graph of the groups containing the variables infant mortality rate and birth rate.

Figura 8.10: Indication of the countries which belong to each of the groups.

We have repeated the analysis using MINITAB for five groups. This program provides the sum of squares within the groups by clusters instead of by variables, as indicated:

```
           Number of    Within cluster  Average distance Maximum distance
           observations  sum of squares   from centroid    from centroid
Cluster1        21           10855.985           20.220           58.275
Cluster2        14             833.119            7.357           10.902
```

```
Cluster3        28           960.586          5.415           9.925
Cluster4         9           864.347          8.977          15.250
Cluster5        19          3126.014         12.110          21.066
```

Example: The results for unstandardized data are similar, but not identical, as we can see in Figure 8.11, where the groups are represented in the plane of the two variables with greatest variance and which will have greater weight in the determination of the groups. Upon standardizing the variables the results change substantially by having greater weight than the rest of the variables; the groups are more homogeneous by continents and in Europe the countries are separated by East and West. The results are given in Figure 8.12 where again the two most important variables have been used.

Figura 8.11: Results of k-means with MINITAB for the unstandardized MUNDODES data. Five groups are formed. Infant mortality (C4) is shown on the y-axis and birth rate (C2) on the x-axis.

Figura 8.12: Results of the k-means for the unstandardized MUNDODES data using MINITAB. Infant mortality (C4) is found on the y-axis and birth rate (C2) on the x-axis.

8.3 HIERARCHICAL METHODS

8.3.1 Distances and Similarities

Euclidean Distances

Hierarchical methods start from a matrix of distances or similarities between items in the sample, then a hierarchy is constructed based on a distance. If all the variables are continuous, the most frequently used distance is the Euclidean distance between standardized variables. In general, it is not advisable to use Mahalanobis distances since the only covariance matrix available is that of the sample as a whole, which can show correlations that are very different from those existing between the variables within the groups. For example, Figure 8.13 was generated using two groups of independent normal variables with means of (0,0) and (5,5) and unit variance. The position of the groups generates a strong positive correlation in the set of points, which disappears when we consider each of the groups separately.

Figura 8.13: Two groups of uncorrelated variables can lead to a high correlation between variables.

In order to decide whether or not to standardize variables, before beginning the analysis it is advisable to take into account the above comments and the purpose of the study. If we do not standardize, the Euclidean distance will depend, above all, on the variables with greater values, and the results of the analysis can change completely when its scale of measurement is modified. If we do standardize, we are giving a similar weight to the variables a priori, independent of their original variability, which may not always be the most appropriate.

When the sample contains continuous variables and attributes, the problem is more complicated. Assume that the variable x_1 is binary. The Euclidean distance between two items in the sample based on this variable is $(x_{i1} - x_{h1})^2$, which will have a value of zero if $x_{i1} = x_{h1}$, that is, when the attribute is, or is not present in both items, and one if the attribute is present in one item, but not in the other. Nevertheless, the distance between two items corresponding to a continuous standardized variable, $(x_{i1} - x_{h1})^2/s_1^2$, can be much greater than one, which means that the continuous variables are going to have greater weight, in general, than binary ones. This is acceptable in many cases but when, due to the nature of the problem, this is undesirable, the solution is to work with similarities.

Similarities The similarity coefficient according to the variable $j = 1, ..., p$ between two sample items (i, h), is defined as a non-negative and symmetric function, s_{jih}:

(1) $s_{jii} = 1$

(2) $0 \leq s_{jih} \leq 1$

(3) $s_{jih} = s_{jhi}$

If we obtain the similarities for each variable between two items we can combine them to get a global similarity coefficient between the two items. The coefficient proposed by Gower is

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih} s_{jih}}{\sum_{j=1}^{p} w_{jih}} \tag{8.4}$$

where w_{jih} is a fictitious variable equal to one if the comparison between the two items using the variable j makes sense, and will be zero if we do not include this variable in the comparison between items. For example, if the variable x_1 is whether a person has asked for a loan $(x_1 = 1)$ and $(x_1 = 0)$ if not, and the variable x_2 is for whether that person paid back the loan or not, then if a person has not asked for a loan, they are $x_1 = 0$ and it makes no sense to worry about x_2. In this case, when comparing individuals (i, j) if either of them have a value of zero in x_1, we assign to the variable w_{2ij} the value zero. The similarities between items according to the qualitative variables can be constructed individually or in blocks. The similarity between two items by a binary variable will be one if the attribute is present, and zero if not. Alternatively, we can group binary variables in homogenous groups and deal with them jointly. If we assume that all attributes carry the same weight, we can construct a measure of similarity between two items A and B in relation to these attributes taking the number of attributes present:

(1) in both (a);
(2) in A and not in B, (b);
(3) in B and not in, (c);
(4) in neither of the two, (d).

These four quantities form an *association table between items* and will serve to build measures of similarity between items. This table verifies that $n_a = a + b + c + d$, where n_a is the number of attributes.

Tabla 8.3: Data matrix when the variables are binary attributes

Items	variables (attributes)						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7
A	0	1	1	0	0	0	1
B	1	0	1	1	1	1	0
C	1	0	0	1	1	1	1
.	.	.	.	.	.	.	

For example, Table 8.3 gives a possible data matrix with seven binary attributes and from it the Association Table 8.4 was constructed. It gives the joint distribution of values 0 and 1 for items A and B. Item A contains 3 values of 1 in the set of binary variables, and, of these three cases, on one occasion item B also has a value of 1 and in the other two, it has a value of 0. Item A has a value of 0 four times, never coinciding with B since in those four cases B has a value of 1. The sum of all rows and columns must always be equal to the number of binary attributes being studied. In order to calculate a similarity coefficient between two individuals using their association table the following two main criteria are used:

	B			
		1	0	
A	1	1(a)	2(b)	3
	0	4(c)	0(d)	4
	Sum	5	2	7(n_a)

Tabla 8.4: Association table corresponding to the elements A and B

1. *Coincidence ratio.* Calculated as the total number of coincidences over the total number of attributes:
$$s_{ij} = \frac{a+d}{n_a}. \tag{8.5}$$
For example, the similarity of A and B is 1/7 and that of B and C is 5/7.

2. *Appearance ratio.* When the absence of an attribute is not relevant we can exclude the absences and calculate only the ratio of times that the attribute is present in both items. The coefficient is defined by:
$$s_{ij} = \frac{a}{a+b+c} \tag{8.6}$$

For example, with this criterion in Table 8.3 the similarity between A and B is again 1/7, and that of B and C is 4/6.

Although the above two proposals are the most frequently used, there are other alternatives. For example, if we want to apply double weight to the coincidences, so that $s_{ij} = 2(a+d)/(2(a+d)+b+c)$, or only take into consideration the coincidences and take $s_{ij} = a/(b+c)$. Finally, the similarity coefficients for a continuous variable are constructed by taking
$$s_{jih} = 1 - \frac{|x_{ij} - x_{hj}|}{\text{rank}(x_j)}$$
In this way, the resulting coefficient will always be between zero and one. When we have various variables these coefficients can be combined as indicated in (8.4).

Once we have obtained the global similarity between items, we can transform the coefficients into distances. The simplest way is to define the distance using $d_{ij} = 1 - s_{ij}$, but this relationship may not verify the triangular property. It can be proven that if the similarity matrix is positive definite (what would happen if we calculate the similarities by (8.5) or (8.6)), and we define the distance by:
$$d_{ij} = \sqrt{2(1-s_{ij})}$$
then we would verify the triangular property (see exercise 6.5).

8.3.2 Hierarchical Algorithms

Supposing a matrix of distances or similarities, we would like to classify the items in a hierarchy. The existing algorithms work in such a way that the items are assigned successively to groups. There is no provision, however, for reallocation of items which have been incorrectly grouped earlier. In other words, once it has been done, it cannot be changed later. The algorithms are of the two following types:

1. *Agglomerative.* Starts with individual items, which are then grouped.
2. *Divisive.* Starts with a grouping of items which are then divided successively until individual items are reached.

Agglomerative algorithms require less calculation time and are more frequently used. For further reading on divisive algorithms, the reader can consult Seber (1984).

8.3.3 Agglomerative Methods

The agglomerative algorithms used always have the same structure and only differ in the way the distances between groups are calculated. The steps are as follows.

1. Start with as many groups as items, n. The distances between groups are the distances between the original items.
2. Choose the two most similar items in the distance matrix, and with these form a new group.
3. Replace the two items used in (2) with the new group. The distances between this new group and the earlier ones are calculated with the criterion given below.
4. Go back to (2) and repeat steps (2) and (3) until all of the groups are fused into a single cluster.

Criteria for defining distances between groups

Suppose we have group A with n_a items, and group B with n_b items, and they merge to form group (AB) with $n_a + n_b$ items. The distance from the new group, (AB) to another group C with n_c items, is usually calculated using one of the following five methods:

1. *Simple linkage or nearest neighbor.* The distance between the two new groups is the smallest distance between groups before being fused. Thus:

   $$d(C; AB) = min\,(d_{CA}, d_{CB})$$

   A simple way to calculate the minimum between the two distances with a computer is to take

   $$min\,(d_{CA}, d_{CB}) = 1/2\,(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|)$$

   Indeed, if $d_{CB} > d_{CA}$ the term in absolute value is $d_{CB} - d_{CA}$ and the result of the operation is d_{CA}, which is the smallest distance. If $d_{CA} > d_{CB}$ then the second term is $d_{CA} - d_{CB}$ and the result is d_{CB}.

As this criterion depends only on the order of the distances, it does not change with monotonic transformations; we obtain the same hierarchy even when the distances are numerically different. It has been shown that this criterion tends to produce very long, drawn out clusters which can contain quite dissimilar items at the opposite ends.

2. *Complete linkage or farthest neighbor.* The distance between the two new clusters is the greatest of the distances between groups before merging. That is:

$$d(C;AB) = \max\left(d_{CA}, d_{CB}\right)$$

and we can prove that

$$\max\left(d_{CA}, d_{CB}\right) = 1/2\left(d_{CA} + d_{CB} + \left|d_{CA} - d_{CB}\right|\right).$$

This criterion is also unchanging with monotonic transformations of distances as it depends, like the former, on the order of the distances. It tends to produce spherical clusters.

3. *Average linkage.* The distance between the two new groups is the weighted average between the distances between groups before being fused. That is:

$$d(C;AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

As the values of the distances are weighted, this criterion is not invariant to monotonic transformations of distances.

4. *Centroid method or K-means* . This method is generally used only with continuous variables. We make the distances between two groups equal to the Euclidean distance from the group centroids, where the centroid is the mean of the observations belonging to the group. When the two groups are merged, we can calculate the new distances between them without using the original items. It can be demonstrated (see exercise 8.5) that the squared Euclidean distance from a group C to the merged groups A, with n_a items, and B, with n_b, is

$$d^2(C;AB) = \frac{n_a}{n_a + n_b} d^2_{CA} + \frac{n_b}{n_a + n_b} d^2_{CB} - \frac{n_a n_b}{(n_a + n_b)^2} d^2_{AB}$$

Ward's method

A somewhat different way of constructing hierarchical groupings was proposed by Ward and Wishart. It differs from the above methods in that it begins directly from the items, instead of using a distance matrix, and that a global measure of heterogeneity in the grouping of observations in the clusters is defined. This measure, already seen in section 8.2, is $\mathbf{W}$,and is the sum of Euclidean distances between each item and the mean of its group:

$$\mathbf{W} = \sum_g \sum_{i \in g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g) \tag{8.7}$$

where $\overline{\mathbf{x}}_g$ is the mean of group g. The criterion starts with the assumption that each item forms a group, $g = n$, and therefore $\mathbf{W}$ (8.7) is zero. Next, the elements which produce the minimum increment of $\mathbf{W}$ are merged. This obviously implies taking the closest ones with the Euclidean distance. In the next step we have $n-1$ groups, $n-2$ of one element and one of the two elements. Again, we decide which groups to merge so that $\mathbf{W}$ grows as little as possible, thus we end up with $n-2$ groups, and continue in the same way until we have one single group. The values of $\mathbf{W}$ indicate the growth of the criterion as groups are formed and can be used to decide how many natural groups are contained in our data.

In each step it can be proved that the groups which must be merged in order to minimize $\mathbf{W}$ are those such that:

$$\min \frac{n_a n_b}{n_a + n_b} (\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)'(\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)$$

Comparison

It is difficult to give general guidelines which justify the use of one criterion over another, although the last three are the most frequently used. We recommend determining which criterion is the most reasonable for the data to be grouped and, in the case of doubt, to try various ones and compare the results.

The dendogram

The dendogram, or hierarchical tree diagram, graphically displays the results of the grouping process in the form of a tree. All the criteria for defining distances which we have given have the property that if we consider three groups A, B, C:

$$d(A, C) \leq \max\{d(A, B), d(B, C)\}$$

and a measure of distance which has this property is called *ultrametric.* This property is stronger than the triangular, as an ultrametric is always a distance. Indeed, if $d(A, C)$ is less than or equal to the maximum of $d(A, B)$ then $d(B, C)$ must then be less than or equal to the sum $d(A, B) + d(B, C)$. The dendogram is the representation of an ultrametric, and is constructed in the following way:

1. The n initial items are displayed in the lower part of the graph.

2. The fusions between the items are represented by three straight lines. These come together, or merge, at nodes whose position along a distance axis indicate the level at which the fusions occur.

3. The process is repeated until all items are connected by straight lines.

If we cut the dendogram at any given distance, we obtain a classification of the number of existing groups at this level and the items which form them.

The dendogram is useful when the points have a clearly hierarchical structure, but they can be deceiving if applied blindly as the two points might seem similar when in fact, they are not or vice versa.

Example: We are going to apply the algorithms we have studied to the following initial distance matrix between items.

$$\begin{array}{c|cccc|c|cccc|} & A & B & C & D & & & & & \\ A & 0 & 1 & 4 & 2,5 & & 0 & 1 & 4 & 2{,}5 \\ B & 1 & 0 & 2 & 3 & = & & 0 & 2 & 3 \\ C & 2 & 2 & 0 & 4 & & & & 0 & 4 \\ D & 2,5 & 3 & 4 & 0 & & & & & 0 \end{array}$$

Method 1 Simple linkage or nearest neighbor. The minimum value outside the diagonal of the distance matrix is 1 and it corresponds to the distance between items A and B. We merge them to form a cluster and then, calculate the new distances from an item to the cluster (AB) as the minimum of the distances of this item to A and B. That is:

$$\begin{aligned} d(AB,C) &= min(4;2) = 2; \\ d(AB,D) &= min(2,5;3) = 2,5. \end{aligned}$$

The new table of distances is obtained from the previous one by cancelling out the rows and columns of A and B and adding a new column and row corresponding to the cluster AB which contains the new distances. The result is:

$$\begin{array}{c|ccc} & AB & C & D \\ AB & 0 & 2 & 2{,}5 \\ C & 2 & 0 & 4 \\ D & 2{,}5 & 4 & 0 \end{array}$$

The minimum value outside the diagonal of the table is now 2, which corresponds to the distance between AB and C. We merge these two groups and calculate the distances to the new group:

$$d(ABC,D) = min(2,5;4) = 2,5.$$

and finally, the last groups ABC and D are merged. This process is shown in the dendogram in Figure 8.14

The dendogram shows that first items A and B are merged at distance one, this cluster is then merged to C at distance 2 and ABC to D at distance 2.5.

Method 2. Complete linkage or farthest neighbor. The first fusion is carried out as in the above case between A and B at distance one. Nevertheless, the new distances are now:

$$\begin{aligned} d(AB,C) &= \max(4;2) = 4; \\ d(AB,D) &= \max(2,5;3) = 3 \end{aligned}$$

Figura 8.14: Denndogram of simple linkage method

Figura 8.15: Dendrogram of complete linkage method

Figura 8.16: Dendrogram of average linkage method

and the next fusion will be between AB and D at distance 3. The distance from C to cluster ABD is 4 and this will be the next fusion. Figure 8.15 summarizes the process.

Method 3 . The process begins, as with the above methods, with the fusion of the closest items, AB. The new distances are d(AB,C)=3; d(AB,D)=2.75. Thus, the next fusion will be between AB and D at distance 2.75. This group ABD will be merged with C at distance d(ABC,D) = 1/2(4+2.75) = 3.375. Figure 8.16 summarizes the process.

Method 4 . The initial process is, again, the same as the above methods. The new distances are calculated as $d^2(C;AB) = \frac{1}{2}d^2_{CA} + \frac{1}{2}d^2_{CB} - \frac{1}{4}d^2_{AB} = 8 + 2 - 0,25 = 9,75$. Similarly, $d^2(D;AB) = 2,5^2/2 + 9/2 - 1/4 = 7,375$. The merger with D will be at distance $\sqrt{7,375} = 2.72$. The distance from C to the new group will be $d^2(C;ABD) = \frac{1}{3}9,75 + \frac{1}{2}16 - \frac{1}{4}7,375 = 3.16^2$, and C will be merged to the cluster at distance 3.16. Figure 8.17 shows the dendogram.

Example: Figure 8.18 presents the dendogram made with MINITAB for the MUNDODES countries using the minimum sum of squares method (Ward). The graph suggests the presence of four or five groups of countries.

Figura 8.17: Dendrogram of the centroid method

Figura 8.18: Results of a hierarchical grouping of the MUNDODES countries by birth rate variables.

The figure shows the result of the simple linkage, which is much more confusing.

Figura 8.19: Results of a hierarchical grouping for the MUNDODES countries using simple linking.

In order to compare the results of the hierarchical grouping and that of partition Figure 8.20 shows the groups obtained for the standardized data and using Ward's criterion in the graph of birth rate and infant mortality rate.

Figura 8.20: Result of the hierarchical grouping separated into five groups for standardized MUNDODES variables.

8.4 CLUSTERING BY VARIABLES

Cluster analysis of variables is an exploratory procedure which can suggest procedures for dimension reduction, such as factor analysis, or canonical correlation methods which will be studied in the second part of this book. The idea is to construct a matrix of distances or similarities between variables and apply a hierarchical classification algorithm to this matrix.

8.4.1 Distance and similarity measurements between variables

The usual measures of association between continuous variables are covariance and correlation. These measurements take into account only linear relationships. Alternatively, we could build a measure of distance between two variables $\mathbf{x}_j$ and $\mathbf{x}_h$ representing each variable as a point in $\Re^n$ and calculating the Euclidean distance between the points. This measurement is:

$$\begin{aligned} d_{jh}^2 &= \sum_{i=1}^{n}(x_{ij}-x_{ih})^2 && (8.8)\\ &= \sum x_{ij}^2+\sum x_{ih}^2-2\sum x_{ij}x_{ih}. && (8.9)\end{aligned}$$

So that the distance does not depend on the units, the variables must be standardized. If they are not, the distance between the two variables could arbitrarily change as a result of linear transformations of the variables (for example, in measuring height in meters instead of centimeters). We suppose then that if we are working with standardized variables with a zero mean and a variance of one, we have that (8.8) is reduced to:

$$d_{jh}^2 = 2n(1-r_{jh}).$$

We can see that:

(a) if $r_{jh}=1$, the distance is zero, indicating that the two variables are identical.

(b) if $r_{jh}=0$, the two variables are uncorrelated and the distance is $d_{jh}=\sqrt{2n}$.

(c) if $r_{jh}<0$, the two variables are negatively correlated, and the distance will take on its maximum value, $\sqrt{4n}$, when the two variables have a correlation of -1.

This distance measurement can be standardized so that its values are between zero and one, disregarding the constant n and taking $d_{jh}=\sqrt{(1-r_{jh})/2}$.

For qualitative binary variables, we can construct a similarity measurement in much the same way as we did when constructing an association table between items. In order to do this, we take the number of items where both characteristics are present (a), where only one is present (b) and (c), and where neither of the two are present (d). These tables show that if n is the number of individuals, then $n=a+b+c+d$, and we can build similarity coefficients as we did with the items. Alternatively, this association table between variables is a contingency table (see chapter 7) and a measure of distance is the value of chi-squared (see Appendix 8.1)

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(a+c)(c+d)(b+d)}.$$

It is more common to define the distance by the contingency coefficient

$$d_{ij} = 1 - \sqrt{\frac{\chi^2}{n}}.$$

8.5 ADDITIONAL READINGS

Extensive literature is available on clustering techniques: Anderberg (1973), Everitt (1993), Gordon (1981), Hartigan (1975), Mirkin (1996), Spath and Bull (1980) and Spath (1985). The majority of general multivariate textbooks also dedicate a chapter to these methods.

Example: Figure 8.21 shows the dendogram of the variables of the EUROSEC data using the Ward criterion. It is observed that the grouping of the variables coincides as expected: first is the merging of mining and energy, services and industrial services, and industry and construction. At the second level we have services (which encompass the three variables - service, industrial services and finance), agriculture, which is alone, and industry, which picks up the rest of the industrial variables.

Figura 8.21: Cluster of the EUROSEC data by variables.

Example: The dendogram in Figure 8.22 shows the cluster of the variables for body measurements, MEDIFIS. The closest correlation is found between foot length and height, and the cranial diameter variable shows little relation to the rest as we have seen previously. If we wished to make groups at a first level we have three groups of variables: length, with four variables, width, with two, and cranial diameter. At a higher level we find all the variables on one side and cranial diameter on the other.

Figura 8.22: Dendrogram of body measurements using the Ward criterion.

Example: Figure 8.23 presents the results for the INVES variables. At a lower level we have four groups of variables: chemistry, engineering, agriculture and biology, and the rest, which include 4 variables. At a higher level the two groups join and the greatest distance is found between the chemistry databank and the rest.

Figura 8.23: Dendrogram of the INVES variables.

EXERCISES

Exercise: Apply the k-means algorithm to the household budget data. How many groups are there in the data?

Exercise: Apply a hierarchical cluster to the household budget data. Compare the results with different clustering methods. Compare with the k-means results.

Exercise: Prove that Hartigan's criterion for the k-means algorithm is equivalent to continuing to add groups until $tr(\mathbf{W}_G) < tr(\mathbf{W}_{G+1})(n-G+9)/(n-G-1)$ (Suggestion - make $tr(\mathbf{W}) = SSW$, and impose the condition that the value of F be greater than 10).

Exercise: Prove that if we define $\mathbf{T} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}})(\mathbf{x}_{ig} - \overline{\mathbf{x}})'$ for the total sum of squares we can write $\mathbf{T} = \mathbf{B} + \mathbf{W}$, where W was defined in section 8.2 and $\mathbf{B}$ is the matrix of the sum of squares between groups.

Exercise: Prove that the distances between groups with simple, complete and average linkage can be calculated with $\alpha d_{CA} + \alpha d_{CB} + \beta \left| d_{CA} - d_{CB} \right|$ and obtain the values of α and β that produce these distances.

Exercise: Prove that in hierarchical clustering we can calculate the squared Euclidean distances between a group C to the union of the groups A, with n_a items and B n_b using $d^2(C;AB) = \frac{n_a}{n_a+n_b} d^2_{CA} + \frac{n_b}{n_a+n_b} d^2_{CB} - \frac{n_a n_b}{(n_a+n_b)^2} d^2_{AB}$.

(Suggestion - the average of the union of the groups A and B will have the coordinates $\overline{x}_{AB} = \frac{n_a}{n_a+n_b}\overline{x}_A + \frac{n_b}{n_a+n_b}\overline{x}_B$, use this equation in the distance of C to this point $(\overline{x}_C - \overline{x}_{AB})'(\overline{x}_C - \overline{x}_{AB})$ and expand.

APPENDIX 8.1. CALCULATION OF CHI-SQUARED STATISTICS IN 2×2 TABLES

In the contingency table $\{a,b,c,d\}$ the expected frequencies are $\frac{1}{n}\{(a+c)(a+b),(a+b)(b+d),(b+d)(c+d)\}$ and the value of the χ^2 defined in section 7.3 is:

$$\chi^2 = \left(\frac{ad-bc}{n}\right)^2 \left[\frac{n}{(a+c)(a+b)} + \frac{n}{(a+b)(b+d)} + \frac{n}{(a+c)(c+d)} + \frac{n}{(b+d)(c+d)}\right]$$

Indeed, as the table has a degree of freedom, the discrepancies between observed and expected frequencies must be equal, for example for the first box

$$\left(a - \frac{(a+c)\,(a+b)}{n}\right)^2 = \left(\frac{na - a(a+b+c) - bc}{n}\right)^2 = \left(\frac{ad-bc}{n}\right)^2$$

and the same is obtained in the remaining. Since:

$$(b+d)(d+c) + (a+c)(c+d) + (a+b)(b+d) + (a+c)(a+b) =$$

$$(b+d)n + (a+c)n = (a+b+c+d)n = n^2$$

the final result is:

$$\chi^2 = \frac{(ab-bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}.$$