Chapter 15
Cluster analysis

15.1 INTRODUCTION AND SUMMARY

The objective of cluster analysis is to assign observations to groups ("clusters") so that observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves stand apart from one another. In other words, the objective is to divide the observations into homogeneous and distinct groups.

In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups.

There are a number of clustering methods. One method, for example, begins with as many groups as there are observations, and then systematically merges observations to reduce the number of groups by one, two, . . ., until a single group containing all observations is formed. Another method begins with a given number of groups and an arbitrary assignment of the observations to the groups, and then reassigns the observations one by one so that ultimately each observation belongs to the nearest group.

Cluster analysis is also used to group variables into homogeneous and distinct groups. This approach is used, for example, in revising a questionnaire on the basis of responses received to a draft of the questionnaire. The grouping of the questions by means of cluster analysis helps to identify redundant questions and reduce their number, thus improving the chances of a good response rate to the final version of the questionnaire.

15.2 AN EXAMPLE

Cluster analysis embraces a variety of techniques, the main objective of which is to group observations or variables into homogeneous and distinct clusters. A simple numerical example will help explain these objectives.

Example 15.1 The daily expenditures on food (X_1) and clothing (X_2) of five persons are shown in Table 15.1.

Table 15.1
Illustrative data,
Example 15.1

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

The numbers are fictitious and not at all realistic, but the example will help us explain the essential features of cluster analysis as simply as possible. The data of Table 15.1 are plotted in Figure 15.1.

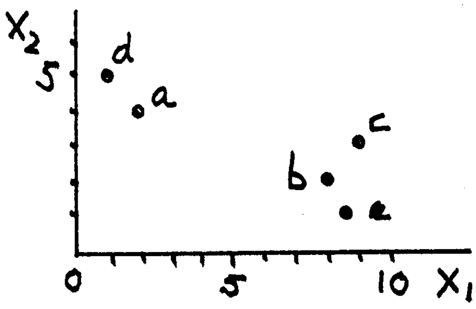


Figure 15.1
Grouping of observations, Example 15.1

Inspection of Figure 15.1 suggests that the five observations form two clusters. The first consists of persons a and d, and the second of b, c and e. It can be noted that the observations in each cluster are similar to one another with respect to expenditures on food and clothing, and that the two clusters are quite distinct from each other.

These conclusions concerning the number of clusters and their membership were reached through a visual inspection of Figure 15.1. This inspection was possible because only two variables were involved in grouping the observations. The question is: Can a procedure be devised for similarly grouping observations when there are more than two variables or attributes?

It may appear that a straightforward procedure is to examine all possible clusters of the available observations, and to summarize each clustering according to the degree of proximity among the cluster elements and of the separation among the clusters. Unfortunately, this is not feasible because in most cases in practice the number of all possible clusters is very large and out of reach of current computers. Cluster analysis offers a number of methods that operate much as a person would in attempting to reach systematically a reasonable grouping of observations or variables.

15.3 MEASURES OF DISTANCE FOR VARIABLES

Clustering methods require a more precise definition of "similarity" ("closeness", "proximity") of observations and clusters.

When the grouping is based on variables, it is natural to employ the familiar concept of distance. Consider Figure 15.2 as a map showing two points, i and j, with coordinates (X_{1i}, X_{2i}) and (X_{1j}, X_{2j}), respectively.

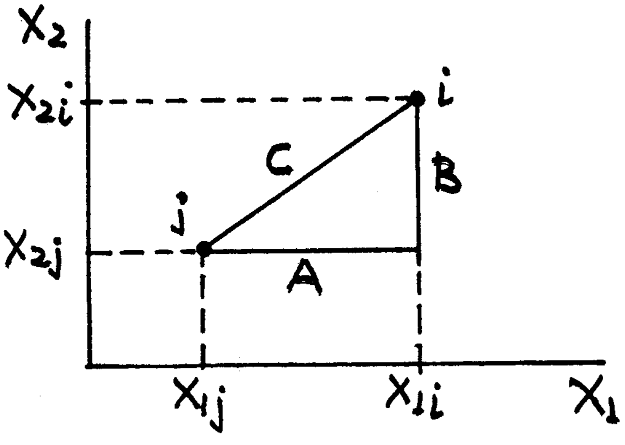


Figure 15.2
Distance measures illustrated

The *Euclidean distance* between the two points is the hypotenuse of the triangle ABC:

$$D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$

An observation i is declared to be closer (more similar) to j than to observation k if $D(i,j) < D(i,k)$.

An alternative measure is the *squared Euclidean distance.* In Figure 15.2, the squared distance between the two points i and j is

$$D_2(i,j) = A^2 + B^2 = (X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2.$$

Yet another measure is the *city block distance*, defined as

$$D_3(i,j) = |A| + |B| = |X_{1i} - X_{1j}| + |X_{2i} - X_{2j}|.$$

As the name suggests, it is the distance one would travel if the points i and j were located at opposite corners of a city block.

The distance measures can be extended to more than two variables. For example, the Euclidean distance between an observation $(X_{1i}, X_{2i}, \ldots, X_{ki})$ and another $(X_{1j}, X_{2j}, \ldots, X_{kj})$ is

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}.$$

All three measures of distance depend on the units in which X_1 and X_2 are measured, and are influenced by whichever variable takes numerically larger values. For this reason, the variables are often standardized so that they have mean 0 and variance 1 before cluster analysis is applied. Alternatively, weights w_1, w_2, ..., w_k reflecting the importance of the variables could be used and a weighted measure of distance calculated. For example,

$$D(i,j) = \sqrt{w_1(X_{1i} - X_{1j})^2 + w_2(X_{2i} - X_{2j})^2 + \cdots + w_k(X_{ki} - X_{kj})^2}.$$

15.4 CLUSTERING METHODS

Given a distance measure, a reasonable procedure for grouping n observations proceeds in the following steps.

Begin with as many clusters as there are observations, that is, with each observation forming a separate cluster. Merge that pair of observations that are nearest one another, leaving $n-1$ clusters for the next step. Next, merge into one cluster that pair of clusters that are nearest one another, leaving $n-2$ clusters for the next step. Continue in this fashion, reducing the number of clusters by one at each step, until a single cluster is formed consisting of all n observations. At each step, keep track of the distance at which the clusters are formed. In order to determine the number of clusters, consider the step(s) at which the merging distance is relatively large.

A problem with this procedure is how to measure the distance between clusters consisting of two or more observations. Perhaps the simplest method is to treat the distance between the two nearest observations, one from each cluster, as the distance between the two clusters. This is known as the *nearest neighbor* (or *single linkage*) method. Figure 15.3 illustrates.

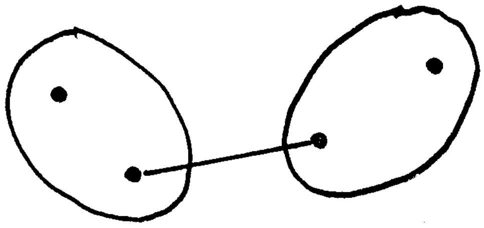
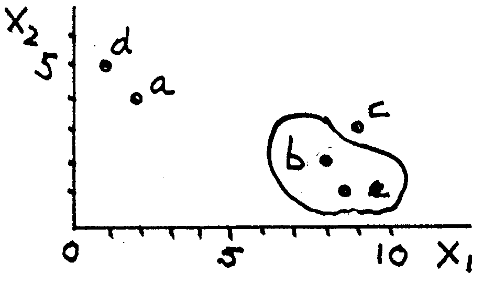


Figure 15.3
Cluster distance, nearest neighbor method

Example 15.1 (Continued) Let us suppose that Euclidean distance is the appropriate measure of proximity. We begin with each of the five observations forming its own cluster. The distance between each pair of observations is shown in Figure 15.4(a).

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0

(a)



(b)

Figure 15.4
Nearest neighbor method, Step 1

For example, the distance between a and b is

$$\sqrt{(2-8)^2+(4-2)^2} = \sqrt{36+4} = 6.325.$$

Observations b and e are nearest (most similar) and, as shown in Figure 15.4(b), are grouped in the same cluster.

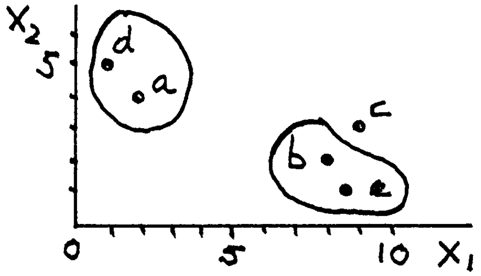
Assuming the nearest neighbor method is used, the distance between the cluster (be) and another observation is the smaller of the distances between that observation, on the one hand, and b and e, on the other. For example,

$$D(be,a) = \min\{D(b,a), D(e,a)\} = \min\{6.325, 7.159\} = 6.325.$$

The four clusters remaining at the end of this step and the distances between these clusters are shown in Figure 15.5(a).

Cluster	(be)	a	c	d
(be)	0	6.325	1.414	7.614
a		0	7.071	1.414
c			0	8.246
d				0

(a)



(b)

Figure 15.5
Nearest neighbor method, Step 2

Two pairs of clusters are closest to one another at distance 1.414; these are (ad) and (bce). We arbitrarily select (a,d) as the new cluster, as shown in Figure 15.5(b).

The distance between (be) and (ad) is

$$D(be, ad) = \min\{D(be,a), D(be,d)\} = \min\{6.325, 7.616\} = 6.325,$$

while that between c and (ad) is

$$D(c,ad) = \min\{D(c,a), D(c,d)\} = \min\{7.071, 8.246\} = 7.071.$$

The three clusters remaining at this step and the distances between these clusters are shown in Figure 15.6(a). We merge (be) with c to form the cluster (bce) shown in Figure 15.6(b).

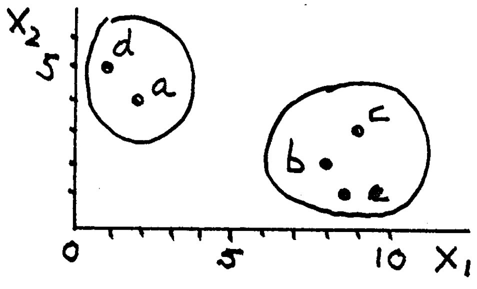
The distance between the two remaining clusters is

$$D(ad, bce) = \min\{D(ad,be), D(ad,c)\} = \min\{6.325, 7.071\} = 6.325.$$

The grouping of these two clusters, it will be noted, occurs at a distance of 6.325, a much greater distance than that at which the earlier groupings took place. Figure 15.7 shows the final grouping.

Cluster	(*be*)	(*ad*)	*c*
(*be*)	0	6.325	1.414
(*ad*)		0	7.071
c			0

(a)

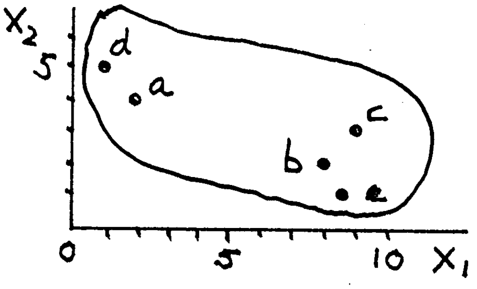


(b)

Figure 15.6
Nearest neighbor method, Step 3

Cluster	(*bce*)	(*ad*)
(*bce*)	0	6.325
(*ad*)		0

(a)



(b)

Figure 15.7
Nearest neighbor method, Step 4

The groupings and the distance at which these took place are also shown in the tree diagram (*dendrogram*) of Figure 15.8.

One usually searches the dendrogram for large jumps in the grouping distance as guidance in arriving at the number of groups. In this illustration, it is clear that the elements in each of the clusters (*ad*) and (*bce*) are close (they were merged at a small distance), but the clusters are distant (the distance at which they merge is large).

The nearest neighbor is not the only method for measuring the distance between clusters. Under the *furthest neighbor* (or *complete linkage*) method,

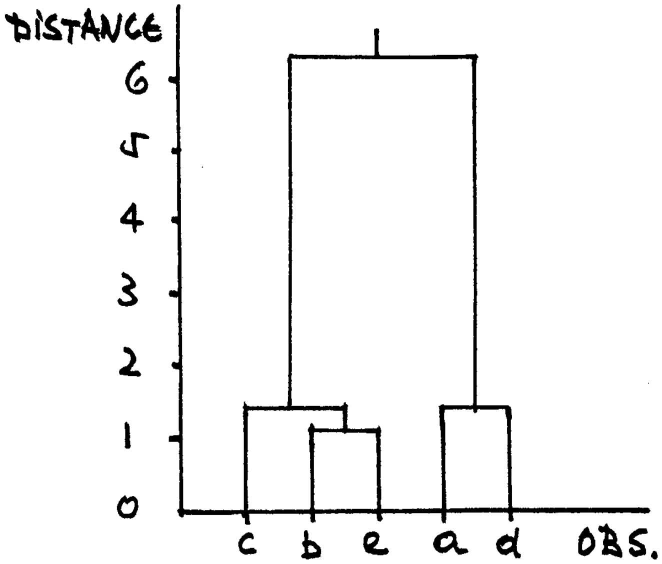


Figure 15.8
Nearest neighbor method, dendrogram

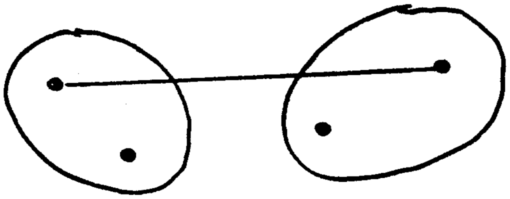


Figure 15.9
Cluster distance, furthest neighbor method

the distance between two clusters is the distance between their two most distant members. Figure 15.9 illustrates.

Example 15.1 (Continued) The distances between all pairs of observations shown in Figure 15.4 are the same as with the nearest neighbor method. Therefore, the furthest neighbor method also calls for grouping *b* and e at Step 1. However, the distances between (be), on the one hand, and the clusters (a), (c), and (d), on the other, are different:

$$
\begin{aligned}
D(be,a) &= \max\{D(b,a), D(e,a)\} = \max\{6.325, 7.159\} = 7.159 \\
D(be,c) &= \max\{D(b,c), D(e,c)\} = \max\{1.414, 2.062\} = 2.062 \\
D(be,d) &= \max\{D(b,d), D(e,d)\} = \max\{7.616, 8.500\} = 8.500
\end{aligned}
$$

The four clusters remaining at Step 2 and the distances between these clusters are shown in Figure 15.10(a).

Cluster	(be)	a	c	d
(be)	0	7.159	2.062	8.500
a		0	7.071	1.414
c			0	8.246
d				0

(a)

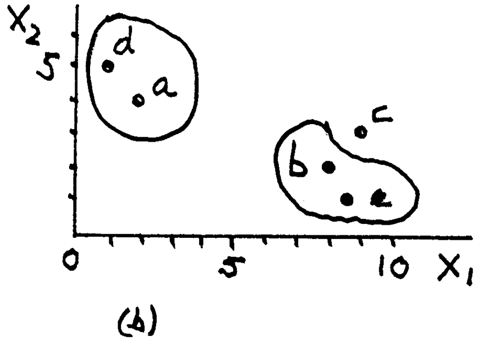


Figure 15.10
Furthest neighbor method, Step 2

The nearest clusters are (a) and (d), which are now grouped into the cluster (ad). The remaining steps are similarly executed.

The reader is asked to confirm in Problem 15.1 that the nearest and furthest neighbor methods produce the same results in this illustration. In other cases, however, the two methods may not agree.

Consider Figure 15.11(a) as an example. The nearest neighbor method will probably not form the two groups percived by the naked eye. This is so because at some intermediate step the method will probably merge the two "nose" points joined in Figure 15.11(a) into the same cluster, and proceed to string along the remaining points in chain-link fashion. The furthest neighbor method, will probably identify the two clusters because it tends to resist merging clusters the elements of which vary substantially in distance from those of the other cluster. On the other hand, the nearest neighbor method will probably succeed in forming the two groups marked in Figure 15.11(b), but the furthest neighbor method will probably not.

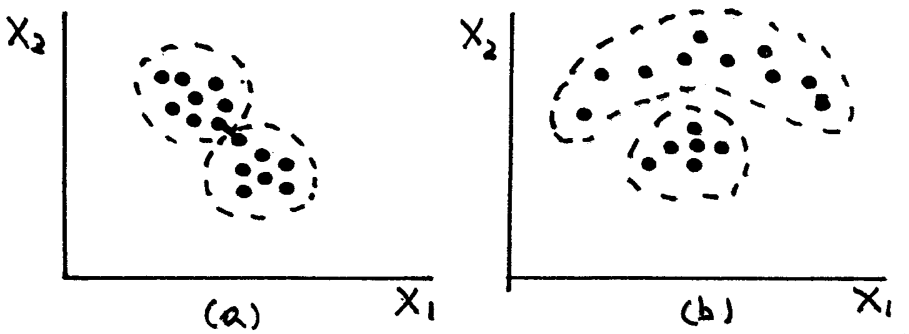


Figure 15.11
Two cluster patterns

A compromise method is *average linkage*, under which the distance between two clusters is the average of the distances of all pairs of observations, one observation in the pair taken from the first cluster and the other from the second cluster as shown in Figure 15.12.

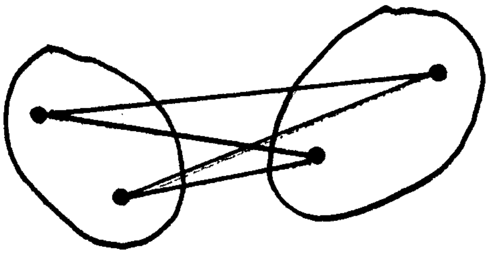
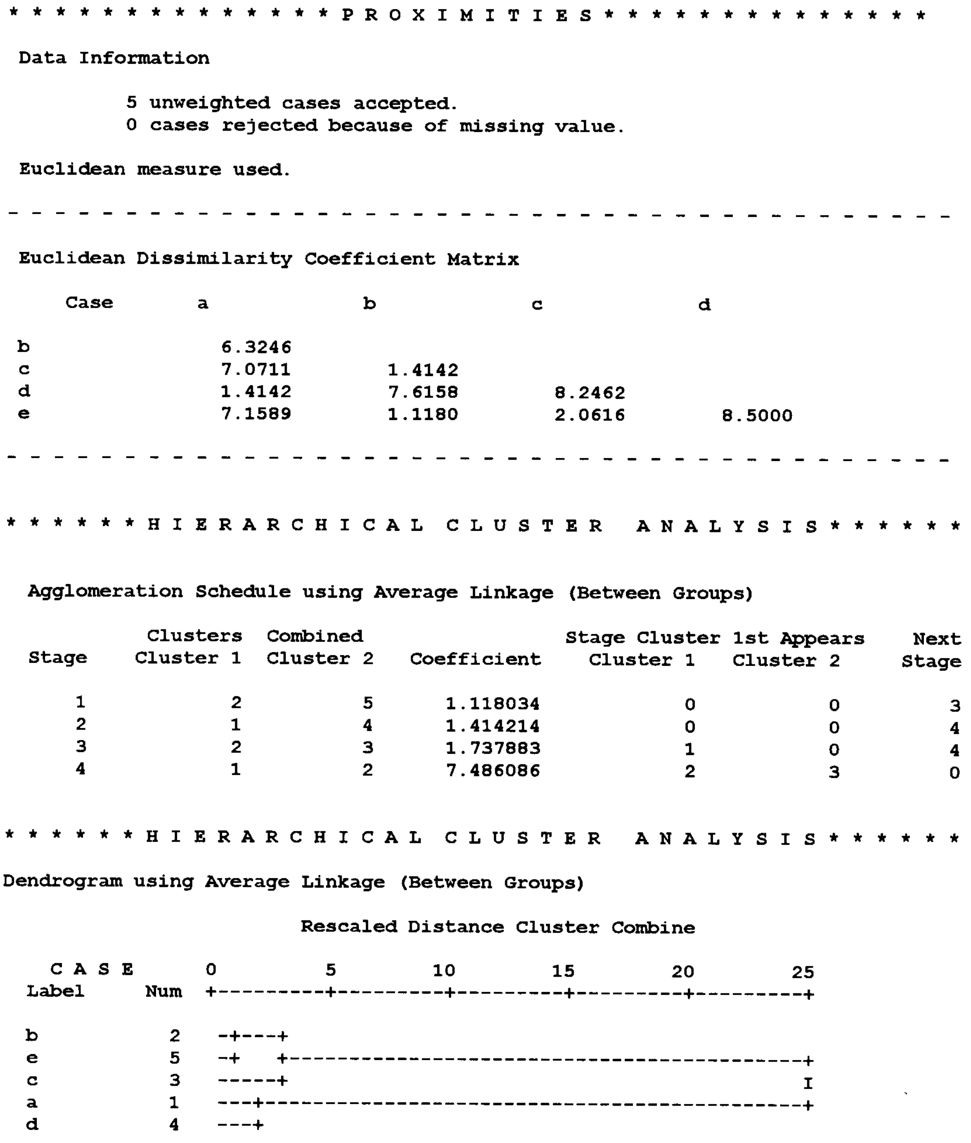


Figure 15.12
Cluster distance, average linkage method

Figure 15.13 shows the slightly edited output of program SPSS, instructed to apply the average linkage method to the data of Table 15.1. In Problem 15.2, we let the reader confirm these results and compare them to those of earlier methods.

The three methods examined so far are examples of *hierarchical agglomerative* clustering methods. "Hierarchical" because all clusters formed by these methods consist of mergers of previously formed clusters. "Agglomerative" because the methods begin with as many clusters as there are observations and end with a single cluster containing all observations.



```
* * * * * * * * * * * * * * P R O X I M I T I E S * * * * * * * * * * * * * *

 Data Information

          5 unweighted cases accepted.
          0 cases rejected because of missing value.

 Euclidean measure used.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 Euclidean Dissimilarity Coefficient Matrix

     Case        a             b             c             d

 b               6.3246
 c               7.0711        1.4142
 d               1.4142        7.6158        8.2462
 e               7.1589        1.1180        2.0616        8.5000

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *


  Agglomeration Schedule using Average Linkage (Between Groups)

            Clusters  Combined              Stage Cluster 1st Appears     Next
  Stage    Cluster 1  Cluster 2  Coefficient   Cluster 1  Cluster 2      Stage

      1            2          5     1.118034           0          0          3
      2            1          4     1.414214           0          0          4
      3            2          3     1.737883           1          0          4
      4            1          2     7.486086           2          3          0

* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

                         Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  b           2   -+---+
  e           5   -+   +-------------------------------------------+
  c           3   -----+                                           I
  a           1   ---+---------------------------------------------+
  d           4   ---+
```

Figure 15.13
SPSS output, average linkage method

There are many other clustering methods. For example, a *hierarchical divisive* method follows the reverse procedure in that it begins with a single cluster consisting of all observations, forms next 2, 3, etc. clusters, and ends with as many clusters as there are observations. It is not our intention to examine all clustering methods.* We do want to describe, however, an example of *non-hierarchical* clustering method, the so-called *k-means method.* In its simplest form, the k-means method follows the following steps.

Step 1. Specify the number of clusters and, arbitrarily or deliberately, the members of each cluster.

Step 2. Calculate each cluster's "centroid" (explained below), and the distances between each observation and centroid. If an observation is nearer the centroid of a cluster other than the one to which it currently belongs, re-assign it to the nearer cluster.

Step 3. Repeat Step 2 until all observations are nearest the centroid of the cluster to which they belong.

Step 4. If the number of clusters cannot be specified with confidence in advance, repeat Steps 1 to 3 with a different number of clusters and evaluate the results.

Example 15.1 (Continued) Suppose two clusters are to be formed for the observations listed in Table 15.1. We begin by arbitrarily assigning a, b and d to Cluster 1, and c and e to Cluster 2. The cluster centroids are calculated as shown in Figure 15.14(a).

The cluster centroid is the point with coordinates equal to the average values of the variables for the observations in that cluster. Thus, the centroid of Cluster 1 is the point ($X_1 = 3.67$, $X_2 = 3.67$), and that of Cluster 2 the point (8.75, 2). The two centroids are marked by C1 and C2 in Figure 15.14(a). The cluster's centroid, therefore, can be considered the center of the observations in the cluster, as shown in Figure 15.14(b).

We now calculate the distance between a and the two centroids:

$$\begin{aligned} D(a,abd) &= \sqrt{(2-3.67)^2+(4-3.67)^2} = 1.702, \\ D(a,ce) &= \sqrt{(2-8.75)^2+(4-2)^2} = 7.040. \end{aligned}$$

Observe that a is closer to the centroid of Cluster 1, to which it is currently assigned. Therefore, a is not reassigned.

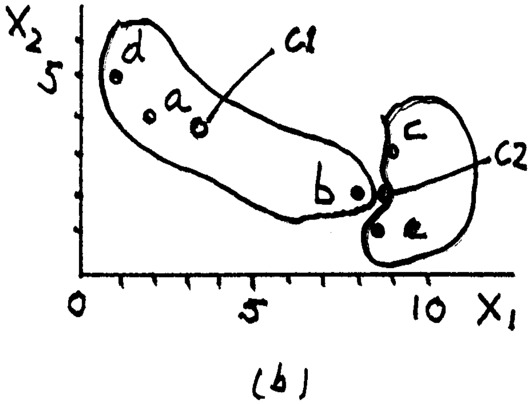
Next, we calculate the distance between b and the two cluster centroids:

$$\begin{aligned} D(b,abd) &= \sqrt{(8-3.67)^2+(2-3.67)^2} = 4.641, \\ D(b,ce) &= \sqrt{(8-8.75)^2+(2-2)^2} = 0.750. \end{aligned}$$

* For additional information, see, for example, Everitt (1993), Kaufman and Rousseeuw (1990).

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	*c*	9	3
b	8	2	*e*	8.5	1
d	1	5			
Ave.	3.67	3.67	Ave.	8.75	2

(a)

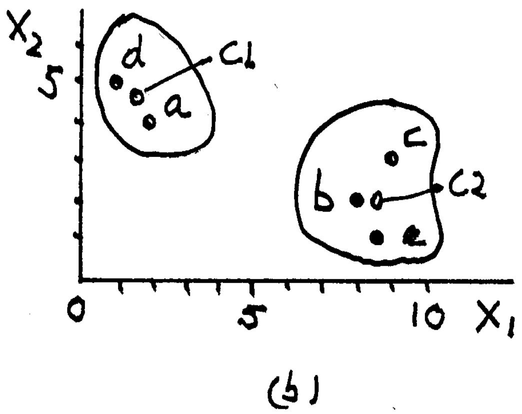


(b)

Figure 15.14
k-means method, Step 1

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	*c*	9	3
d	1	5	*e*	8.5	1
			b	8	2
Ave.	1.5	4.5	Ave.	8.5	2

(a)



(b)

Figure 15.15
k-means method, Step 2

Since *b* is closer to Cluster 2's centroid than to that of Cluster 1, it is reassigned to Cluster 2. The new cluster centroids are calculated as shown in Figure 15.15(a).

The new centroids are plotted in Figure 15.15(b). The distances of the observations from the new cluster centroids are as follows (an asterisk indicates the nearest centroid):

	Distance from	
Obs.	Cluster 1	Cluster 2
a	0.707*	6.801
b	6.964	0.500*
c	7.649	1.118*
d	0.707*	8.078
e	7.826	1.000*

Every observation belongs to the cluster to the centroid of which it is nearest, and the k-means method stops. The elements of the two clusters are shown in Figure 15.15(b).

Other variants of the k-means method require that the first cluster centroids (the “seeds”, as they are sometimes called) be specified. These seeds could be observations. Observations within a specified distance from a centroid are then included in the cluster. In some variants, the first observation found to be nearer another cluster centroid is immediately reassigned and the new centroids recalculated, in others reassignment and recalculation await until all observations are examined and one observation is selected on the basis of certain criteria. The “quick” or “fast” clustering procedures used by computer programs such as SAS or SPSS make use of variants of the k-means method.

15.5 DISTANCE MEASURES FOR ATTRIBUTES

The distance measures presented in Section 15.3 and used in earlier examples must be modified if the clustering of observations is based on attributes.

Consider, for example, the following description of four persons according to marital status (single, married, divorced, other) and gender (male, female):

Obs.	Marital status	Gender
a	Single	Female
b	Married	Male
c	Other	Male
d	Single	Female

A reasonable measure of the similarity of two observations is the ratio of the number of matches (identical categories) to the number of attributes. For example, since a and d are both single and female, the similarity measure is 2/2 or 1; b and c do not have the same marital status but are both male, so the similarity measure is 1/2. To be consistent with earlier measures, however, we use instead

$$D_a(i,j) = 1 - \frac{\text{Number of matches}}{\text{Number of attributes}}$$

as the measure of "distance" (dissimilarity) of two observations i and j. We declare two observations to be closer, the smaller this distance. The distances between all pairs of observations in our example are as follows:

Obs.	a	b	c	d
a	0	1	1	0
b		0	0.5	1
c			0	1
d				0

Any of the clustering methods described earlier can be applied to the above distances. For example, in the first step of the nearest neighbor, furthest neighbor, or complete linkage methods a and d would be grouped to form the first cluster. The remaining steps would be carried out in the usual fashion.

When the grouping is to be based on variables *and* attributes, perhaps the simplest approach is to convert the variables to attributes and then apply the measure $D_a(i, j)$ to the distance between any pair of observations. For example, suppose that the four observations will be grouped according to marital status, gender, and age:

	Marital		Age	Age
Obs.	status	Gender	(years)	category
a	Single	Female	15	Y
b	Married	Male	30	M
c	Other	Male	60	O
d	Single	Female	32	M

We could make age an attribute with, say, three categories: Y (under 25 years old), M (25 to 50), and O (more than 50 years old). The "distance" between b and c, for example, is

$$D_a(b, c) = 1 - \frac{1}{3} = \frac{2}{3}.$$

The distances between all pairs of observations are as follows:

Obs.	a	b	c	d
a	0	1	1	1/3
b		0	2/3	2/3
c			0	1
d				0

Any clustering method can now be applied to this table of distances.

Example 15.2 A study* was made to identify clusters of warehouse items that tended to be ordered together. Items in the same cluster could be stored near one another in the warehouse, so as to minimize the effort needed to select those required for particular orders. The study involved a distributor of telecommunications products who stored approximately 1,000 items and was filling approximately 75,000 orders per month on average.

Available was a history of K orders and the items that each order required. To measure the "distance" between two items, a variable V_i for each item i was introduced such that $V_{ik} = 1$ if item i was required by a given order k, otherwise $V_{ik} = 0$. The distance between any pair of items i and j was defined as

$$D(i,j) = \sum_{k=1}^{K} |V_{ik} - V_{jk}|.$$

The following table illustrates the calculation of the distance for two items and a fictitious history of four orders:

Order no., k	Item 1, V_{1k}	Item 2, V_{2k}	$\|V_{1k} - V_{2k}\|$
1	1	1	0
2	0	1	1
3	1	0	1
4	0	0	$\underline{0}$
			2

It is clear that smaller values of the distance measure, $D(1,2) = 2$ in this illustration, indicate that the two items are frequently ordered together.

15.6 GROUPING VARIABLES

Occasionally, clustering methods are applied to group variables rather than observations. One situation where such a grouping is desirable is the design of questionnaires. The first draft of a questionnaire often contains more questions than is prudent to ensure a good response rate. When the draft questionnaire is tested on a small number of respondents it may be observed that the responses to certain groups of questions are highly correlated. Clustering analysis may be applied to identify groups of questions that are similar

* M. B. Rosenwein, "An Application of Cluster Analysis to the Problem of Locating Items Within a Warehouse", *IIE Transactions*, v. 26, no. 1, Jan. 1994, pp. 101-3.

to one another, in the sense that the answers to these questions are correlated. Then, in the final form of the questionnaire only one of the questions in each cluster of similar questions may be used as representative of all the questions in the cluster.

For example, consider the following responses to three questions by five respondents to the first draft of a questionnaire:

Respondent	Q1	Q2	Q3
a	10	5.0	3.00
b	30	7.5	3.10
c	20	6.0	2.90
d	40	8.0	2.95

The correlation coefficient, r, of Q1 and Q2 can be shown to be 0.984, that of Q1 and Q3 0.076, and that of Q2 and Q3 0.230. A measure of the "distance" (dissimilarity) between two questions is $1 - r$, and the starting table of distances between all pairs of questions is

Variable	Q1	Q2	Q3
Q1	0	0.016	0.924
Q2		0	0.770
Q3			0

Any clustering method can now be applied to this table in the usual manner.

15.7 TO SUM UP

• Cluster analysis embraces a variety of methods, the main objective of which is to group observations or variables into homogeneous and distinct clusters.

• For groupings based on variables, frequently used measures of the similarity of observations are the *Euclidean, squared*, or *city block distance*, applied to the original, standardized, or weighted variables. For groupings based on attributes, a measure of the similarity of two observations is the ratio of the number of matches (identical categories) to the number of attributes. Other measures are possible.

• The *nearest neighbor (single linkage), furthest neighbor (complete linkage)* and *average linkage* methods are examples of hierarchical agglomerative clustering methods. These methods begin with as many clusters as there are observations and end with a single cluster containing all observations; all clusters formed by these methods are mergers of previously formed clusters. Other types of clustering methods are the hierarchical divisive (beginning with a single cluster and ending with as many clusters as there are observations) and the non-hierarchical methods (a notable example of which is the *k-means method* often employed for "quick clustering" by some statistical programs).

- Clustering methods can also be employed to group variables rather than observations, as in the case of questionnaire design. These groupings are frequently based on the correlation coefficients of the variables.

PROBLEMS

15.1 Continue the application of the furthest neighbor (complete linkage) method past Step 2 shown in Figure 15.10. Compare each step's results with those of the nearest neighbor method shown in Figures 15.4 to 15.7.

15.2 Apply the average linkage method to the data in Table 15.1. Compare the results of this method with those of the nearest and furthest neighbor methods.

15.3 Use the data of Table 15.1 and a program for cluster analysis to confirm as many as possible of the results concerning the nearest neighbor, furthest neighbor, average linkage, and k-means methods given in the text and in Problems 15.1 and 15.2.

15.4 Six observations on two variables are available, as shown in the following table:

Obs.	X_1	X_2
a	3	2
b	4	1
c	2	5
d	5	2
e	1	6
f	4	2

(a) Plot the observations in a scatter diagram. How many groups would you say there are, and what are their members?

(b) Apply the nearest neighbor method and the squared Euclidean distance as a measure of dissimilarity. Use a dendrogram to arrive at the number of groups and their membership.

(c) Same as (b), except apply the furthest neighbor method.

(d) Same as (b), except apply the average linkage method.

(e) Apply the k-means method, assuming that the observations belong to two groups and that one of these groups consists of a and e.

15.5 Six observations on two variables are available, as shown in the following table:

Obs.	X_1	X_2
a	-1	-2
b	0	0
c	2	2
d	-2	-2
e	1	-1
f	1	2

(a) Plot the observations in a scatter diagram. How many groups would you say there are, and what are their members?

(b) Apply the nearest neighbor method and the Euclidean distance as a measure of dissimilarity. Draw a dendrogram to arrive at the number of groups and their membership.

(c) Same as (b), except apply the furthest neighbor method.

(d) Same as (b), except apply the average linkage method.

(e) Apply the k-means method, assuming that the observations belong to two groups and that one of these groups consists of a and e.

15.6 A magazine for audiophiles tested 19 brands of mid-sized loudspeakers. The test results and the list prices of these speakers are shown in Table 15.2.

Table 15.2
Data for Problem 15.6

Brand	Price	Accuracy	Bass	Power
A	600	91	5	38
B	598	92	4	18
C	550	90	4	36
D	500	90	4	29
E	630	90	4	15
F	580	87	5	5
G	460	87	5	15
H	600	88	4	29
I	590	88	3	15
J	599	89	3	23
K	598	85	2	23
L	618	84	2	12
M	600	88	3	46
N	600	82	3	29
O	600	85	2	36
P	500	83	2	45
Q	539	80	1	23
R	569	86	1	21
S	680	79	2	36

File `ldspkr.dat`

'Price' is the manufacturer's suggested list price in dollars. 'Accuracy' measures on a scale from 0 to 100 the ability of the loudspeaker to reproduce every frequency in the musical spectrum. 'Bass' measures on a scale from 1 to 5 how well the loudspeaker handles very loud bass notes. 'Power' measures in watts per channel the minimum amplifier power the loudspeaker needs to reproduce moderately loud music.

The magazine would like to group these brands into homogeneous and distinct groups. How would you advise the magazine?

15.7 A consumer organization carries out a survey of its members every year. Among the questions in the last survey were a number requesting the members' appraisal of 42 national hotel chains with respect to such characteristics as cleanliness, bed comfort, etc. The file `hotels.dat` contains the summary of thousands of responses and is partially listed in Table 15.3.

Table 15.3
Data for Problem 15.7

Chain id. no.	Price ($)	Cleanliness	Room size	Bed comfort	Climate control	Noise	Amenities	Service
1	36	3	3	3	3	3	3	3
2	36	1	2	1	1	1	1	1
3	37	2	2	2	1	1	2	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
42	129	4	3	4	4	4	4	4

File `hotels.dat`

'Price' is the average of the prices paid by members, rounded to the nearest dollar. The numbers under the other columns are averages of the members' ratings for each feature, which ranged from 1 (poor) to 5 (excellent), rounded to the nearest integer.

Group the 42 hotel chains into categories of quality (for example: Poor, Acceptable, Good, Very Good, and Excellent). Is there any relationship between quality and price?

15.8 Fall visitors to Michigan's Upper Peninsula were segmented into six clusters on the basis of their responses to a subset of 22 questions concerning participation in recreational activities.* A hierarchical clustering method with squared Euclidean distance was used. The six clusters, the participation rates in the 22 activities, and the number of respondents assigned to each cluster are shown in Table 15.4.

Table 15.4 shows, for example, that 2% of all visitors and 21% of those of Cluster 6 intended to hunt bear. Alltogether, 1,112 visitors were interviewed; 259 of these were assigned to cluster 1.

The six clusters were labeled as follows:

Cluster	Label
1	Inactives
2	Active recreationists/nonhunters
3	Campers
4	Passive recreationists
5	Strictly fall color viewers
6	Active recreationists/hunters

In your opinion, what was the form of the original data to which cluster analysis was applied? Was standardization advisable? Do you agree with the labels attached to the clusters? Would you say that a visitor to Michigan's Upper Peninsula can be treated as one of the six mutually exclusive and collectively exhaustive types described above?

* D. M. Spotts and E. M. Mahoney, "Understanding the Fall Tourism Market", *Journal of Travel Research*, Fall 1993, pp. 3-15.

Table 15.4
Clusters and participation rates, Problem 15.8

Recreational		Cluster					
activity	All	1	2	3	4	5	6
Bear hunting	0.02	0.02	0.01	0.04	0.01	0.00	0.21
Deer hunting	0.05	0.08	0.03	0.04	0.02	0.00	0.58
Small game hunting	0.05	0.06	0.02	0.04	0.02	0.00	0.85
Upland gamebird hunting	0.03	0.00	0.01	0.01	0.01	0.00	0.97
Waterfowl fishing	0.02	0.01	0.00	0.01	0.00	0.00	0.48
Fall color viewing	0.68	0.05	0.78	0.65	0.94	0.96	0.82
Fishing	0.13	0.14	0.55	0.06	0.01	0.01	0.54
Canoeing	0.05	0.00	0.28	0.04	0.01	0.00	0.30
Attending a festival or special event	0.08	0.04	0.40	0.01	0.02	0.01	0.24
Sailing	0.02	0.01	0.04	0.05	0.01	0.00	0.12
Power boating or water skiing	0.03	0.02	0.07	0.04	0.01	0.03	0.06
Tennis	0.01	0.00	0.03	0.00	0.02	0.00	0.12
Off-road vehicle riding	0.06	0.02	0.11	0.02	0.10	0.00	0.30
Swimming	0.07	0.03	0.09	0.13	0.11	0.00	0.15
Day-hiking for at least two hours	0.30	0.13	0.48	0.46	0.46	0.01	0.45
Overnight hiking (backpaking)	0.04	0.00	0.06	0.20	0.00	0.00	0.09
Camping (not backpacking)	0.18	0.02	0.34	0.83	0.02	0.00	0.48
Visiting a place solely to observe birds	0.06	0.01	0.09	0.04	0.13	0.00	0.21
Scuba diving	0.01	0.02	0.00	0.03	0.00	0.00	0.00
Visiting an historic site	0.43	0.22	0.66	0.44	0.78	0.05	0.33
Golfing	0.04	0.01	0.03	0.00	0.03	0.12	0.15
Horseback riding	0.03	0.00	0.09	0.01	0.04	0.01	0.00
Number of visitors	1,112	259	148	138	315	219	33
Percent of visitors	100	23	13	12	28	20	3

15.9 From a national panel of cooperating households, 1,150 wine drinkers were selected and requested to rate each of the 33 motives for drinking wine listed in Table 15.5.[*]

The italized words will be used as abbreviations for each listed motive.

The k-means cluster method was used to form five clusters. The clusters, their labels, and their composition are as follows:

1. The Wine Itself: Taste, food, mild, aroma/bouquet, hearty, refreshing.

2. Introspective: Relax, sleep, lonely, feel good, depressed.

3. Semi-temperate: Light, natural, healthy, low calorie, low alcohol, less filling, watch weight.

4. Social; Familiar, sociable, acceptable, celebrate, friendly.

5. Image Conscious: Stylish, choosing, distinctive.

Describe in sufficient detail the approach you would have used to confirm this clustering if you had access to the original data. State any assumptions you are obliged to make. Comment on the results presented here.

[*] Joel S. Dubow, "Occasion-Based vs. User-Based Benefit Segmentation: A Case Study", *Journal of Advertising Research*, March-April 1992, pp. 11-18.

Table 15.5
Motives for drinking wine, Problem 15.9

I like the *taste*
To *relax*
I want a *refreshing* drink
As a *treat* for myself
To enhance the taste of *food*
I enjoy *choosing* the wine I want
I want a *mild* tasting drink
I want a *familiar* drink
To enjoy the *aroma/bouquet*
I am in *no hurry*
To *feel good*
I want something *light*
Something special to *share*
To be *sociable*
To satisfy a *thirst*
To have *fun*
To be *friendly*

I want something *easy* to serve
To *celebrate* something
It is socially *acceptable*
I want a *low alcohol* drink
I want something *less filling*
I want a *hearty* drink
I want a *natural* drink
I want a *healthy* drink
I want something *low in calories*
To be *romantic*
To be *distinctive*
To help me *sleep*
To be *stylish*
To *watch* my *weight*
I feel *depressed*
I feel *lonely*

15.10 About 20,000 respondents to a survey in the United States were segmented into seven clusters with respect to prime-time TV viewing, radio listening, use of cable TV, movie attendance, video cassette rental, number of books purchased, and number of videogames purchased.*

The clusters were labeled (presumably according to the dominant medium in the cluster) as Prime-Time TV Fans (22% of the respondents), Radio Listeners (22%), Newspaper Readers (20%), Moviegoers (12%), Book Buyers (10%), Videophiles (8%), and CD Buyers (6%).

The profiles of the respondents in each cluster were determined with respect to such personal characteristics as age, gender, household income, whether or not there were children in the household, whether or not the respondent was employed, etc. and the respondent's use or ownership of 32 products and activities.

It was found, for example, that TV Fans were "lackluster consumers" because as a group they had the lowest use, ownership or participation rate in 27 of the 32 products and activities. Newspaper Readers were described as "Mr. and Mrs. Average". Book Buyers were "least likely to drink regular cola and most likely to drink diet cola". Videophiles "could be called the champions of a disposable, fast-paced, consumption lifestyle".

These seven clusters, the study concluded, "... are seven distinct groups. Businesses can use these distinctions to reach their best customers. For example, makers of foreign automobiles or wine might develop co-promotions with music stores to reach the affluent CD buyers. Fast-food companies could profit from developing alliances with video-game manufacturers, video-rental stores, and cable TV companies, because all of these industries depend on families with young children (p. 55)".

Comment on the possible advantages and drawbacks of this type of study.

* Robert Maxwell, "Videophiles and Other Americans", *American Demographics*, July 1992, pp. 48-55.

State any assumptions that you are forced to make. Describe briefly any additional statistical information you may require before you would make any commitments on the basis of this type of study.