Applied Multivariate Data Analysis

Second Edition

Brian S. Everitt
Institute of Psychiatry, King's College London, UK

and

Graham Dunn
School of Epidemiology and Health Sciences, University of Manchester, UK

John Wiley & Sons, Ltd

First published in Great Britain in 2001 by Arnold
This impression printed by Hodder Education,
a part of Hachette Livre UK,
338 Euston Road, London NW1 3BH

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN 978-0-4707-1117-0

8 9 10

Typeset in 10/12pt Times by Academic & Technical Typesetting. Bristol

6

Cluster analysis

6.1 Introduction

An important component of virtually all scientific research is the classification of the phenomena being studied. In the behavioural sciences, for example, these may be individuals or societies or even patterns of behaviour or perception. The investigator is usually interested in finding a classification in which the items of interest are sorted into a small number of homogeneous groups or clusters. Most commonly, the classification sort is one in which the groups are mutually exclusive rather than overlapping, although this is not always appropriate. At the very least the derived classification scheme may represent a convenient method for organizing a large set of multivariate data so that the class labels provide a parsimonious way of describing the patterns of similarities and differences in the data. In market research, for example, it may be useful to group a large number of potential customers according to their needs in a particular product area.

But often a classification may serve more fundamental purposes. In psychiatry, for example, the classification of mental disturbances should help in the search for their causes and lead to improved methods of therapy. And these twin areas of *prediction* (separating diseases that require different treatments) and *aetiology* (searching for the causes of a disease) will be the same in other branches of medicine.

The two aims may not necessarily lead to the same classification, and a variety of alternative classifications for the same set of objects or individuals will always exist. Human beings for example, could be classified with respect to economic status into groups such as *lower class*, *middle class* and *upper class*, or they might be classified by annual consumption of alcohol into *low*, *medium* and *high*. Clearly different classifications may not collect the same set of individuals into groups. Some classifications will, however, be more useful than others, a point clearly made by Needham (1965) in his discussion of the classification of human beings into men and women:

> The usefulness of this classification does not begin and end with all that can, in one sense, be strictly inferred from it – namely a statement about sexual organs. It is a very useful classification because classing a person as man or woman conveys a great deal more information, about probable relative size, strength, certain types of dexterity and so on. When we say that persons in class *man* are more suitable than persons in class *woman* for certain tasks and conversely, we are only incidentally making a remark about sex, our primary concern being with strength, endurance etc. The point is that we have been able to use a classification of persons which conveys information on many properties. On the contrary a classification of persons into those with hair on their forearms between $\frac{3}{16}$ and $\frac{1}{4}$ inch long and those without, though it may serve some particular use, is certainly of no general use, for imputing membership in the former class to a person conveys information on this property alone. Put another way, there are no known properties which divide up a set of people in a similar manner.

In a similar vein, a classification of books based on subject matter into classes such as dictionaries, novels, biographies and so on is likely to be far more useful than one based on say the colour of the book's binding – the former will indicate more of a book's characteristics than the latter.

These examples illustrate that any classification is simply a division of the objects or individuals of interest into groups based on a set of rules – *it is neither true or false* (unlike, say, a theory) and should be judged largely on its usefulness.

Perhaps the simplest approach to identifying groups or clusters in multivariate data is by the examination of scattergrams. These could be based on the raw data, but we might also use the results of a principal components analysis or even a multidimensional scaling. Figure 6.1, for example, shows a scatterplot matrix for the chest, waist and hip measurements of 20 individuals. The data are given in Table 6.1. Several of the scatterplots suggest two groups, which is entirely reasonable given that the sample includes both males and females.

Many of the other graphical techniques discussed in Chapter 2 might be useful in the search for clusters or for providing evidence that one of the many more formal methods of cluster analysis to be described later in this chapter might usefully be applied to the data.

Comprehensive reviews of cluster analysis are provided in Cormack (1971), Gordon (1987; 1996; 1999) and Everitt (1993). In this chapter we shall concentrate on three classes of technique which probably account for the majority of cluster analysis applications reported in the literature. These three classes are:

- agglomerative hierarchical clustering techniques;
- optimization methods;
- mixture models.

It should be noted here that this chapter is concerned only with the problem of classifying previously unclassified material, that is, when at the start of the investigation the number and composition of classes is unknown. An alternative aspect of classification, namely that involved when the groups are known a

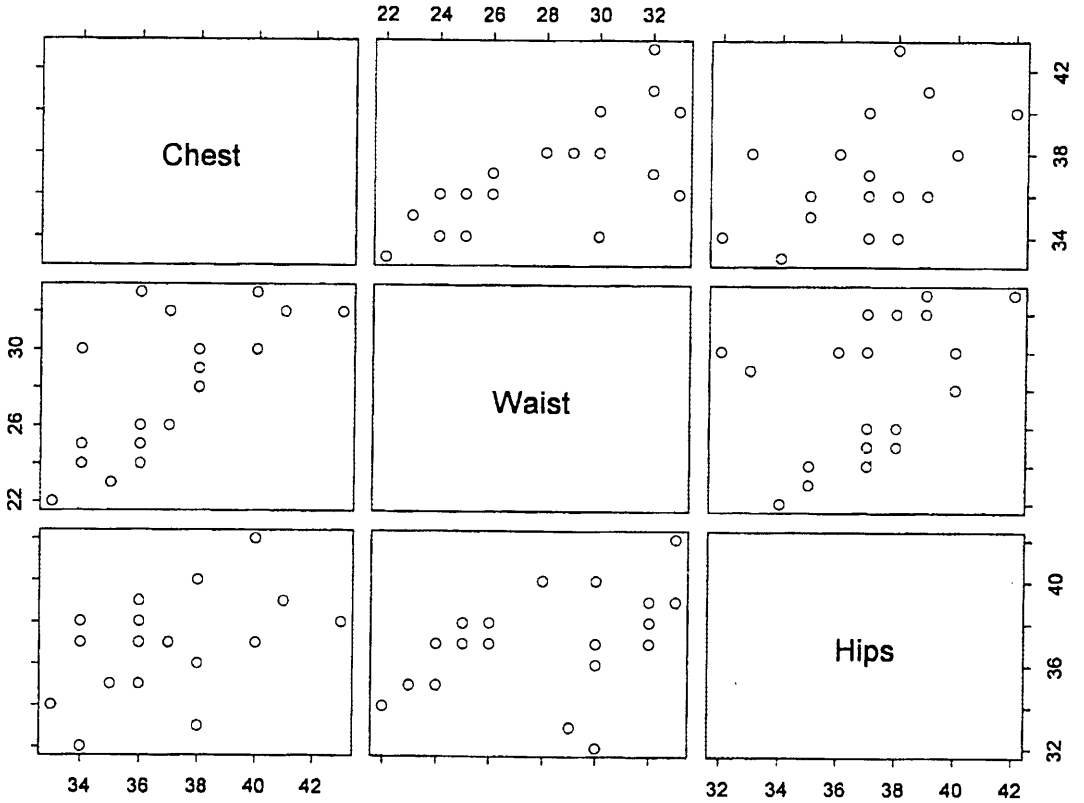

Figure 6.1 Scatterplot matrix of chest, waist and hip measurements on 20 individuals.

Table 6.1 Chest, waist and hip measurements (inches) for 20 individuals

Individual	Chest	Waist	Hips
1	34	30	32
2	37	32	37
3	38	30	36
4	36	33	39
5	38	29	33
6	43	32	38
7	40	33	42
8	38	30	40
9	40	30	37
10	41	32	39
11	36	24	35
12	36	25	37
13	34	24	37
14	33	22	34
15	36	26	38
16	37	26	37
17	34	25	38
18	36	26	37
19	38	28	40
20	35	23	35

priori and the aim is to produce a rule for classifying new individuals, is taken up in Chapter 11.

6.2 Agglomerative hierarchical clustering techniques

In a *hierarchical* classification the data are not partitioned into a particular number of classes or clusters at a single step. Instead, the classification consists of a series of partitions which may run from a single 'cluster' containing all individuals, to n clusters each containing a single individual. Agglomerative hierarchical clustering techniques produce partitions by a series of successive fusions of the n individuals into groups. With such methods fusions, once made, are irreversible, so that when an agglomerative algorithm has placed two individuals in the same group they cannot subsequently appear in different groups. Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, the investigator seeking the solution with the 'best' fitting number of clusters will need to decide which division to choose. This problem of deciding on the correct number of clusters will be taken up in Section 6.2.5.

Hierarchic classifications may be represented by a two-dimensional diagram known as a *dendrogram*, which illustrates the fusions made at each stage of the analysis. An example of such a diagram is given in Figure 6.2. The structure of Figure 6.2 resembles an *evolutionary tree* (see Figure 6.3), and it is in biological

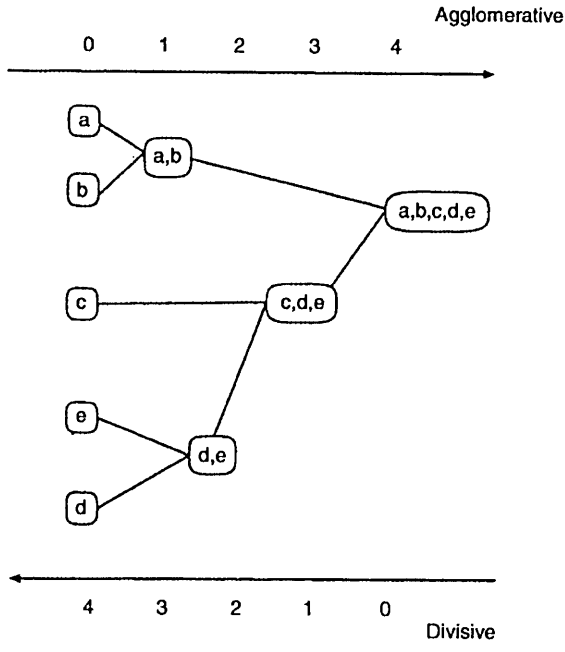

Figure 6.2 An example of a dendrogram. From Kaufman and Rousseeuw (1990), with permission from Wiley.

applications that hierarchical classifications are most relevant and most justified (although, as we shall see later, this type of clustering has now been used in many other areas). According to Rohlf (1970), a biologist, 'all things being equal', aims for a system of nested clusters. Hawkins *et al.* (1982), however, issue the following caveat: 'users should be very wary of using hierarchic methods if they are not clearly necessary'.

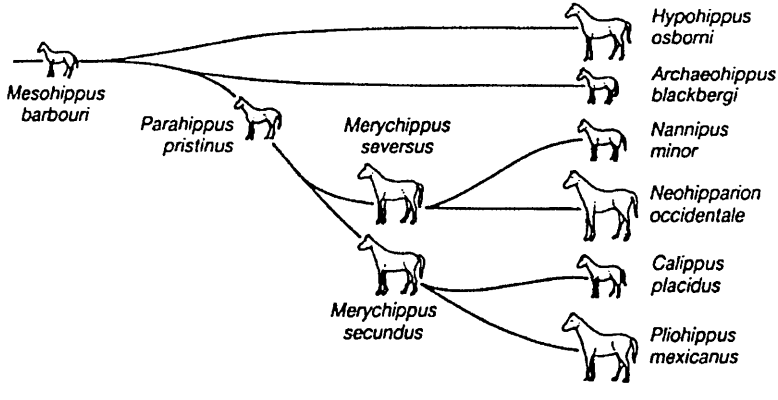

Figure 6.3 An evolutionary tree. From Kaufman and Rousseeuw (1990), with permission from Wiley.

An agglomerative hierarchical clustering procedure produces a series of partitions of the data, $P_n, P_{n-1}, \ldots, P_1$. The first, P_n, consists of n single-member clusters, and the last, P_1, consists of a single group containing all n individuals. The basic operation of all methods is similar:

(START) Clusters $C_1, C_2, \ldots, C_n$ each containing a single individual.
(1) Find the nearest pair of distinct clusters, say C_i and C_j, merge C_i and C_j, delete C_j and decrease the number of clusters by one.
(2) If number of clusters equals one then stop, else return to 1.

At each stage in the process the methods fuse individuals or groups of individuals which are closest (or most similar). Difference between methods arise because of the different ways of defining distance (or similarity) between an individual and a group containing several individuals, or between two groups of individuals.

6.2.1 Measuring inter-cluster dissimilarity

Agglomerative hierarchical clustering techniques differ primarily in how they measure the distances between or similarity of two clusters (where a cluster may, at times, consist of only a single individual). Two simple inter-group measures are

$$d_{AB} = \min_{\substack{i \in A \\ i \in B}}(d_{ij}), \tag{6.1}$$

$$d_{AB} = \max_{\substack{i \in A \\ i \in B}}(d_{ij}), \tag{6.2}$$

where d_{AB} is the distance between two clusters A and B, and d_{ij} is the distance between individuals i and j. (This could be Euclidean distance or one of a variety of other distance measures – see Everitt, 1993, for details.)

The inter-group dissimilarity measure in (6.1) is the basis of *single linkage* clustering, that in (6.2) of *complete linkage* clustering. Both these techniques have the desirable property that they are invariant under monotone transformations of the original inter-individual dissimilarities or distances (cf. non-metric multidimensional scaling in Section 5.4).

A further possibility for measuring inter-cluster distance or dissimilarity is

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}, \tag{6.3}$$

where n_A and n_B are the number of individuals in clusters A and B. This measure is the basis of a commonly used procedure known as *group average* clustering. All three inter-group measures described here are illustrated in Figure 6.4.

6.2.2 Illustrative examples of the application of single linkage, complete linkage and group average clustering

To illustrate the operation of agglomerative hierarchical clustering techniques, we shall apply each of single linkage, complete linkage and group average

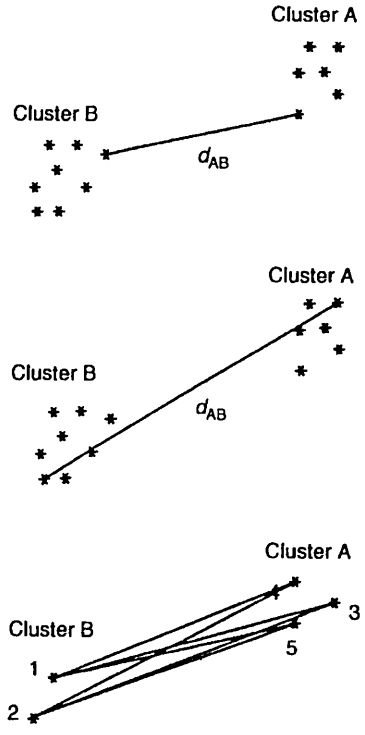

Figure 6.4 Three inter-group distance measures.

clustering to the following dissimilarity matrix for five individuals:

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}$$

Single linkage

At stage one of the procedure, individuals 1 and 2 are merged to form a cluster, since d_{12} is the smallest entry in the matrix **D**. The distances between this group and the three remaining individuals, 3, 4 and 5, are obtained from **D** as follows:

$$d_{(12)3} = \min(d_{13}, d_{23}) = d_{23} = 5.0,$$

$$d_{(12)4} = \min(d_{14}, d_{24}) = d_{24} = 9.0,$$

$$d_{(12)5} = \min(d_{15}, d_{25}) = d_{25} = 8.0.$$

We may now form a new distance matrix $\mathbf{D}_1$ giving inter-individual, and cluster individual values:

$$\mathbf{D}_1 = \begin{array}{c} \\ (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix} \end{array}.$$

The smallest entry in $\mathbf{D}_1$ is d_{45} and so individuals 4 and 5 are now merged to form a second cluster, and distances involving this cluster become:

$$d_{(12)(45)} = \min(d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 8.0,$$

$$d_{(45)3} = \min(d_{34}, d_{35}) = d_{34} = 4.0.$$

The new distances may be arranged to give the matrix

$$\mathbf{D}_2 = \begin{array}{c} \\ (12) \\ 3 \\ (45) \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & (45) \end{array} \\ \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix} \end{array}.$$

The smallest entry is now $d_{(45)3}$, and so individual 3 is merged with the {45} cluster. Finally, fusion of the two remaining groups takes place to form a single group containing all five individuals. The partitions produced at each stage are:

Stage	Groups
P_5	[1], [2], [3], [4], [5]
P_4	[12], [3], [4], [5]
P_3	[12], [3], [45]
P_2	[12], [345]
P_1	[1 2 3 4 5]

The corresponding dendrogram is shown in Figure 6.5.

Complete linkage

As with single linkage, complete linkage begins by merging individuals 1 and 2. The distances between the cluster {12} and individuals 3, 4 and 5 are now obtained as

$$d_{(12)3} = \max(d_{13}, d_{23}) = d_{13} = 6.0,$$

$$d_{(12)4} = \max(d_{14}, d_{24}) = d_{14} = 10.0,$$

$$d_{(12)5} = \max(d_{15}, d_{25}) = d_{15} = 9.0.$$

The final result is the dendrogram shown in Figure 6.6.

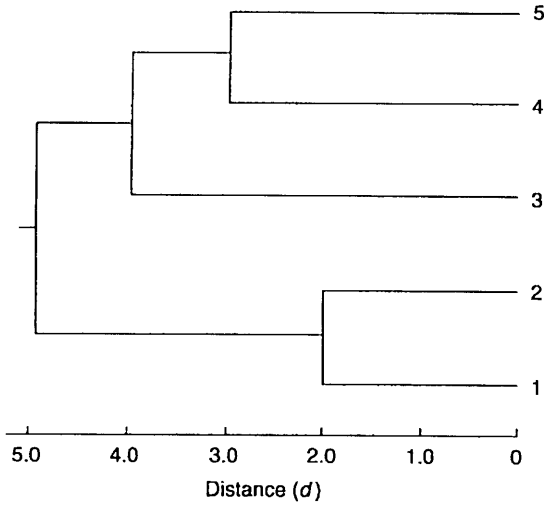

Figure 6.5 Single linkage dendrogram.

Group average

Again the first step is the merger of individuals 1 and 2. The next set of distances is then found as:

$$d_{(12)3} = \tfrac{1}{2}(d_{13} + d_{23}) = 5.5,$$
$$d_{(12)4} = \tfrac{1}{2}(d_{14} + d_{24}) = 9.5,$$
$$d_{(12)5} = \tfrac{1}{2}(d_{15} + d_{25}) = 8.5.$$

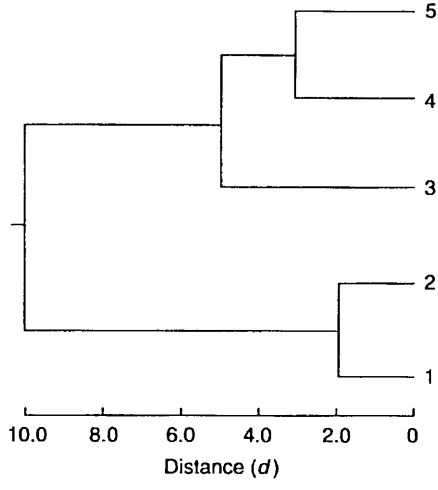

Figure 6.6 Complete linkage dendrogram.

At this stage individuals 4 and 5 merge to form a second cluster. The group average distance between the 2 two-member clusters is given by

$$d_{(12)(45)} = \tfrac{1}{4}(d_{14} + d_{15} + d_{24} + d_{25}) = 9.0. \tag{6.4}$$

6.2.3 Some properties of agglomerative hierarchical clustering techniques

Single linkage can often give unsatisfactory results if 'intermediates' are present between relatively distinct clusters, because of a phenomenon known as *chaining*, which refers to the tendency to incorporate these intermediate points into an existing cluster rather than initiating a new one. The problem is illustrated in Figures 6.7 and 6.8. A result of this problem is that single linkage tends to lead to the formation of long 'straggly' clusters.

Several hierarchical techniques, among them group average and complete linkage, tend to produce solutions in which the clusters are 'spherical' even when the data appear to contain relatively well-separated clusters of other shapes. Consequently, they may *impose* a structure on the data rather than uncover the actual structure present. This problem is illustrated for some two-dimensional data in Figure 6.9.

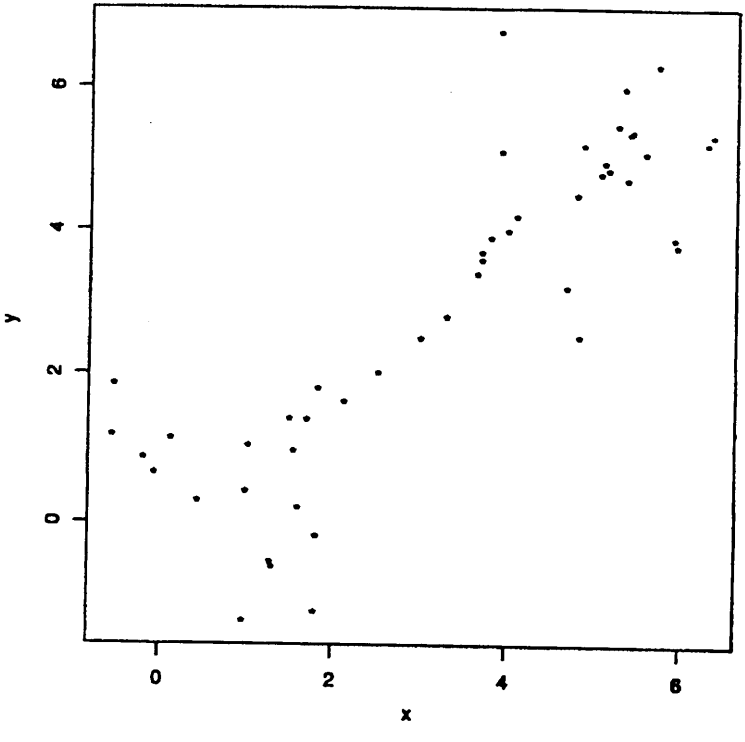

Figure 6.7 Two well-separated clusters with intermediate 'noise' points.

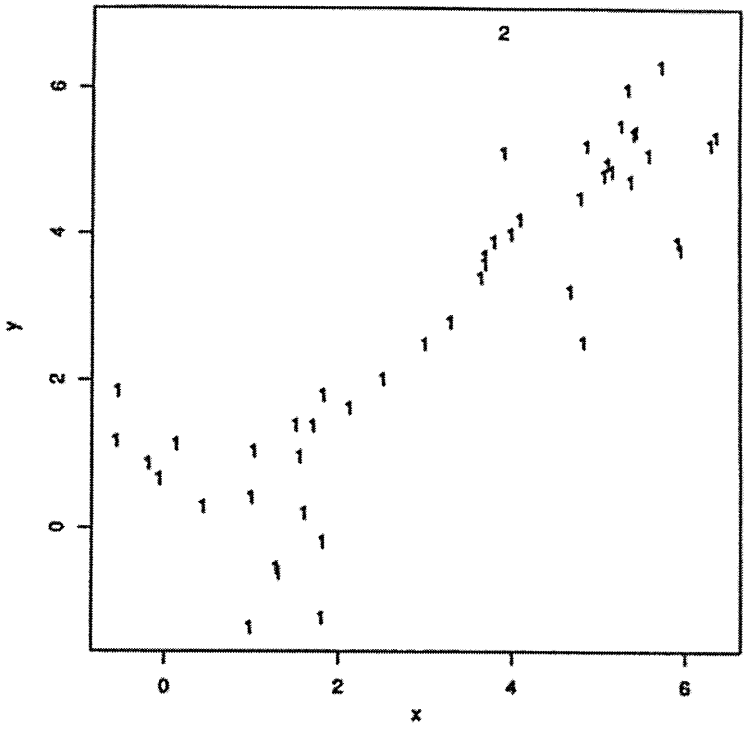

Figure 6.8 Single linkage two-group solution for data in Figure 6.7.

A number of empirical investigations of hierarchical clustering techniques have been performed, the results of which, although not entirely consistent, may be helpful in indicating which of the many methods are likely to be of most use in practice. Baker (1974) and Hubert (1974), for example, both produce evidence that complete linkage clustering is less sensitive to particular types of error than single linkage. Cunningham and Ogilvie (1972) compare seven hierarchical techniques and find that group average clustering performs most satisfactorily overall for the data sets considered. In addition, however, they found a strong interaction in the results between type of input data and the particular clustering method used. Kuiper and Fisher (1975) investigate six hierarchical techniques and find that, for equal numbers of points from different multivariate normal distributions, Ward's (1963) method (see Exercise 6.3) classifies almost as well as Fisher's linear discriminant function (see Chapter 11), *knowing* all the parameters. With unequal sample sizes, however, group average and complete linkage proved more successful. Blashfield (1976) reaches similar conclusions. Applying four hierarchical clustering methods to data generated from multivariate normal mixtures (see Section 6.4), he found the following levels of agreement between cluster solutions and actual structure, agreement being quantified by the *kappa statistic* (see

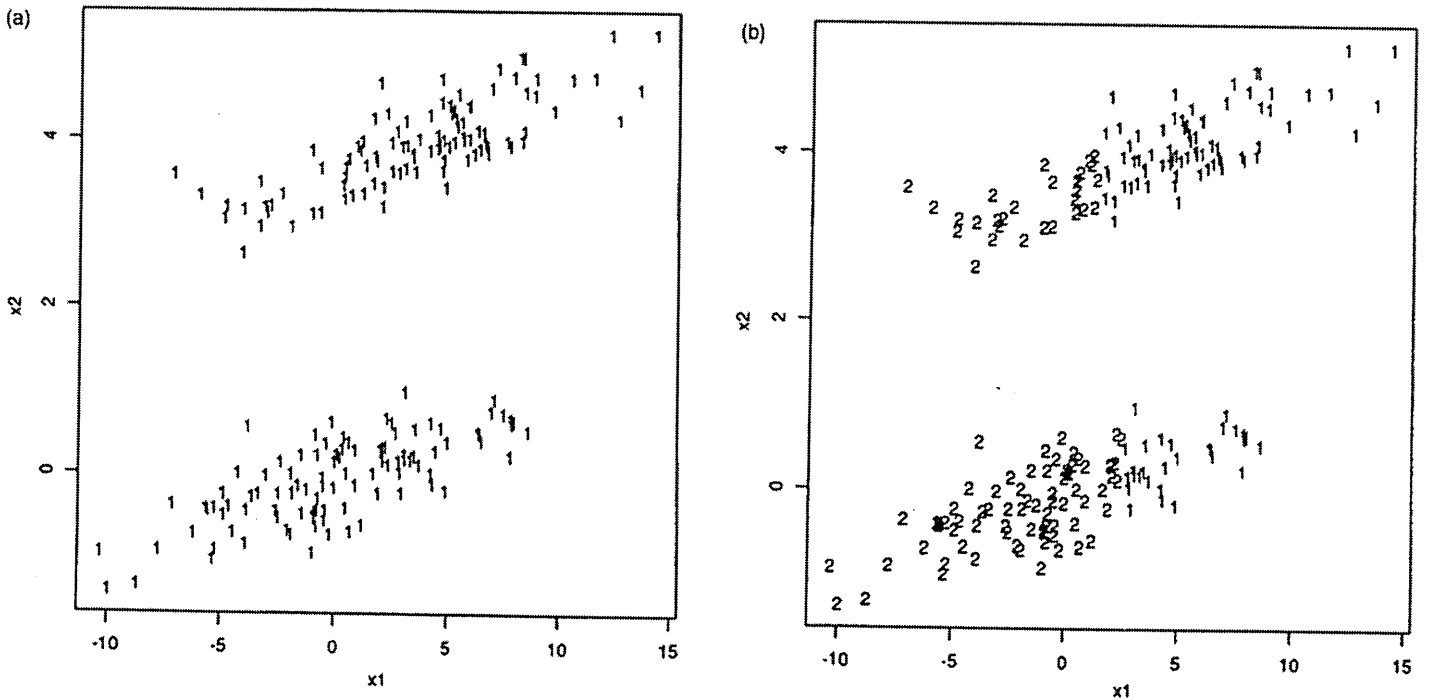

Figure 6.9 (a) Two well-separated elliptical clusters and (b) two-group solution given by hierarchical clustering methods which 'impose' a spherical structure.

Cohen, 1960):

Method	Median kappa	Interquartile range of kappa
Single linkage	0.06	0.034–0.100
Complete linkage	0.42	0.220–0.580
Group average	0.17	0.060–0.460
Ward's method	0.77	0.420–0.940

Clearly, for the data sets considered, single linkage performs very poorly and Ward's method rather well.

A comprehensive study reported by Milligan (1980) demonstrated clearly that *no* single method could be claimed superior for all types of data. The presence of outliers, for example, left the results of single linkage clustering virtually unaffected, but led to very poor performance by group average and Ward's method. In contrast, when the data were such that they contained a true cluster structure masked by the addition of 'noise', single linkage gave poor results, with Ward's method and group average being far superior.

6.2.4 Global fit of a hierarchical clustering solution

Hierarchical clustering techniques impose a hierarchical structure on data and it is usually wise to consider whether this is merited or whether it introduces unacceptable distortions of the original relationships among the individuals as implied by their observed distances. The method most commonly used for assessing the match between the derived dendrogram and the original dissimilarities or distances is the *cophenetic correlation coefficient*. This is simply the product–moment correlation of the $n(n-1)/2$ entries in the lower half of the observed proximity matrix and the corresponding entries in the so-called *cophenetic matrix*, **C**, the elements, c_{ij}, of which are defined to be the first level in the dendrogram at which individuals i and j appear together in the same cluster.

To illustrate the use of the cophenetic correlation coefficient we shall use the results from the application of single linkage clustering described in Section 6.2.1. The elements of **D** and **C** to be correlated are as follows:

d_{ij}:	2.0	5.0	10.0	9.0	4.0	9.0	8.0	5.0	3.0
c_{ik}:	2.0	5.0	5.0	5.0	4.0	5.0	5.0	4.0	3.0

The value of the cophenetic correlation is 0.82.

Rohlf and Fisher (1968) studied the distribution of this type of correlation under the hypothesis that the individuals are randomly chosen from a single multivariate normal distribution. They found that the average value of the coefficient tends to decrease with n and to be almost independent of the number of variables. They also suggested that values of the cophenetic correlation above 0.8 were usually sufficient to reject the null hypothesis. In a later paper by Rohlf (1970), however, a warning is given that 'even a cophenetic

correlation near 0.9 does not guarantee that the dendrogram serves as a sufficiently good summary of the phenetic relationships'.

6.2.5 Partitions from a hierarchy: the number-of-groups problem

It is often the case, when hierarchical clustering techniques are used in practice, that the investigator is not interested in the complete hierarchy but only in one or two partitions obtained from it. In hierarchical clustering, partitions are obtained by 'cutting' a dendrogram or selecting one of the solutions in the nested sequence of clusterings that make up the complete hierarchical classification. Trying to determine the appropriate number of groups, that is, the appropriate partition, is not straightforward. One informal approach often used is to examine the size of the difference between fusion levels in the resulting dendrogram. Large changes might be taken to indicate a particular number of clusters. The dendrogram in Figure 6.10, for example, would be strongly suggestive of a two-cluster solution.

Other more formal methods for determining number of groups from a hierarchical classification are described in Everitt (1993).

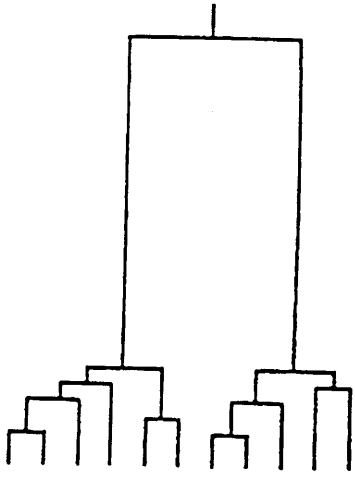

Figure 6.10 Dendrogram suggestive of two clusters.

6.2.6 Examples of the application of agglomerative hierarchical clustering techniques

As a first example of the application of agglomerative hierarchical clustering techniques, single linkage, complete linkage and group average clustering were applied to the data on chest, waist and hip measurements of 20 individuals shown in Table 6.1. The three dendrograms are shown in Figure 6.11. The group average and complete linkage dendrograms both suggest a two-group solution (if we ignore the first observation), and both solutions contain almost the same individuals in each group. This structure is less clear in the single linkage dendrogram.

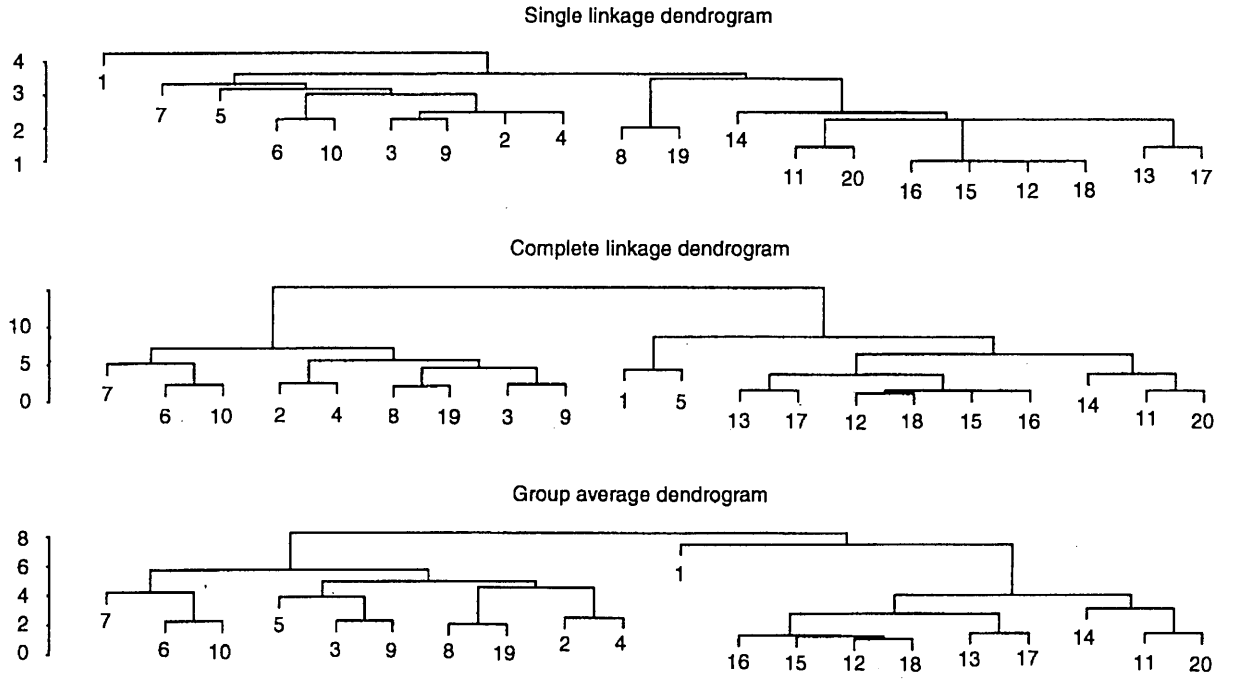

Figure 6.11 Single linkage, complete linkage and group average dendrograms for the data in Table 6.1.

As a second example, data collected by Colonel L.A. Waddell on 32 skulls found in south-western and eastern districts of Tibet will be used. According to Morant (1923), the data can be divided into two groups. The first comprises skulls 1 to 17 found in graves in Sikkim and neighbouring areas of Tibet. The remaining 15 skulls were picked up on a battlefield in the Lhasa district and were believed to be those of native soldiers from the eastern province of Khams. These skulls were of particular interest since it was thought at the time that Tibetans from Khams might be survivors of a particular fundamental human type, unrelated to the Mongolian and Indian types which surrounded them. On each skull the following five measurements (all in millimetres) were obtained: greatest length of skull (x_1), greatest horizontal breadth of skull (x_2), height of skull (x_3), upper face height (x_4), and face breadth, between outermost points of cheek bones (x_5). The data are given in Table 6.2. Here we shall ignore the a priori grouping of the skulls and investigate how a cluster analysis solution reproduces this particular partition. The dendrograms from complete linkage and group average clustering are shown in Figure 6.12. The complete

Table 6.2 Tibetan skull data

Obs.	x_1	x_2	x_3	x_4	x_5
1	190.5	152.5	145.0	73.5	136.5
2	172.5	132.0	125.5	63.0	121.0
3	167.0	130.0	125.5	69.5	119.5
4	169.5	150.5	133.5	64.5	128.0
5	175.0	138.5	126.0	77.5	135.5
6	177.5	142.5	142.5	71.5	131.0
7	179.5	142.5	127.5	70.5	134.5
8	179.5	138.0	133.5	73.5	132.5
9	173.5	135.5	130.5	70.0	133.5
10	162.5	139.0	131.0	62.0	126.0
11	178.5	135.0	136.0	71.0	124.0
12	171.5	148.5	132.5	65.0	146.5
13	180.5	139.0	132.0	74.5	134.5
14	183.0	149.0	121.5	76.5	142.0
15	169.5	130.0	131.0	68.0	119.0
16	172.0	140.0	136.0	70.5	133.5
17	170.0	126.5	134.5	66.0	118.5
18	182.5	136.0	138.5	76.0	134.0
19	179.5	135.0	128.5	74.0	132.0
20	191.0	140.5	140.5	72.5	131.5
21	184.5	141.5	134.5	76.5	141.5
22	181.0	142.0	132.5	79.0	136.5
23	173.5	136.5	126.0	71.5	136.5
24	188.5	130.0	143.0	79.5	136.0
25	175.0	153.0	130.0	76.5	142.0
26	196.0	142.5	123.5	76.0	134.0
27	200.0	139.5	143.5	82.5	146.0
28	185.0	134.5	140.0	81.5	137.0
29	174.5	143.5	132.5	74.0	136.5
30	195.5	144.0	138.5	78.5	144.0
31	197.0	131.5	135.0	80.5	139.0
32	182.5	131.0	135.0	68.5	136.0

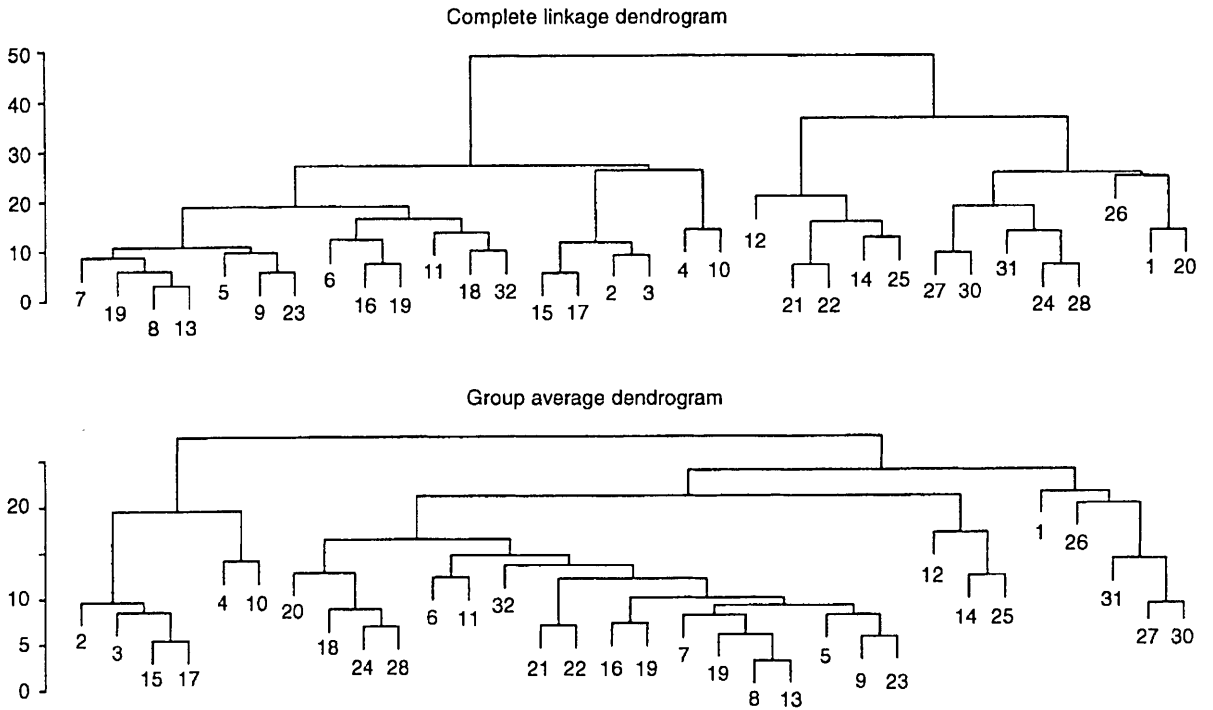

Figure 6.12 Complete linkage and group average dendrograms for Tibetan skull data.

linkage dendrogram suggests a main division into two groups, with a possible further subdivision into a number of smaller groups. The average linkage dendrogram appears to imply perhaps three or four groups, some of which are similar to those given by complete linkage, although there are a number of differences in the two solutions. Neither clustering method produces a solution that matches the original two-group partition of skulls 1 to 17 and skulls 18 to 32.

Other more extensive applications of agglomerative hierarchical clustering techniques are given in Everitt (1993).

6.3 Optimization methods

In this section, a class of clustering techniques is considered which produces a partition of the individuals for a particular number of groups, by optimizing some numerical criterion, low (or, for some criteria, high) values of which indicate a 'good' clustering. With a single variable, for example, we might choose the partition into the chosen number of groups which minimizes the within-group sum of squares of the variable. In the multivariate situation many clustering criteria have been suggested, but the most commonly used arise from considering the following three matrices which can be calculated for each particular partition of the data into g groups:

$$\mathbf{T} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})', \tag{6.5}$$

$$\mathbf{W} = \frac{1}{n-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)', \tag{6.6}$$

$$\mathbf{B} = \sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'. \tag{6.7}$$

where $\mathbf{x}_{ij}$ is the vector of variable values for the jth observation in the ith group, $\bar{\mathbf{x}}$ is the mean vector of all n observations, $\bar{\mathbf{x}}_i$ is the mean vector of the observations in group i, and n_i is the number of observations in group i.

These three $p \times p$ matrices represent respectively *total dispersion*, *within-group dispersion* and *between-group dispersion*; they satisfy the equation

$$\mathbf{T} = \mathbf{W} + \mathbf{B}. \tag{6.8}$$

For $p = 1$, this equation represents a relationship between scalars; it is then simply the separation of the total sum of squares for a variable into the within- and between-groups sum of squares, familiar from a one-way analysis of variance. In this case a natural criterion for grouping is, as suggested above, to choose the partition corresponding to the minimum value of the within-group sum of squares, or, equivalently, the maximum value of the between-group sum of squares.

For $p > 1$, a number of criteria based on (6.8) have been suggested; we shall look at two of these.

1. *Minimization of trace*(**W**). An obvious extension of the minimization of the within-group sum-of-squares criterion applicable for $p = 1$, is minimization of the *sum* of the within-group sum of squares for each variable – that is, the minimization of trace(**W**). This criterion is commonly used, although it suffers from the problem of not being scale-invariant, and of imposing a 'spherical' structure on the data. See Everitt (1993) for details.
2. *Minimization of det*(**W**). In multivariate analysis of variance (see Chapter 8), one of the test statistics for assessing differences in group mean vectors is the ratio of the determinants of the within and total dispersion matrices. Large values of det(**T**)/det(**W**) indicate that the group mean vectors do indeed differ. Such considerations led Friedman and Rubin (1967) to suggest as a clustering criterion the maximization of this ratio. Since for all partitions of the n individuals into g groups, T remains the same, maximization of det(**T**)/det(**W**) is equivalent to minimization of det(**W**). This particular criterion has been studied in detail by Marriott (1971; 1982). Minimization of det(**W**) has the advantage of being scale-invariant. But although it does not impose a spherical structure on the data, it does assume that all clusters have the same shape – again, see Everitt (1993) for details.

Having chosen a suitable clustering criterion, the optimization process would appear to be relatively straightforward: consider every partition of the n individuals into g groups and select the one with optimal value. Unfortunately the number of partitions, N, of n objects into g groups quickly becomes too large to deal with. For example:

n	g	N
15	3	2 375 101
20	4	45 232 115 900
25	8	690 223 721 118 365 580
100	5	10^{68}

The impracticability of examining every possible partition has led to the development of algorithms designed to search for the optimum value of a clustering criterion by rearranging existing partitions and keeping the new one only if it provides an improvement. The essential steps in such a *hill climbing algorithm* are:

- Find some initial partition of the individuals into the required number of groups.
- Calculate the change in the clustering criterion produced by moving each individual from its own to another cluster.
- Make the change which leads to the greatest improvement in the value of the clustering criterion.
- Repeat the previous two steps until no move of a single individual causes the clustering criterion to improve.

In most applications of optimization methods of cluster analysis, the investigator will have to 'estimate' the number of clusters in the data. A variety of methods have been suggested which may be helpful in particular situations. Most are relatively informal and involve, essentially, plotting the value of the clustering criterion against number of groups. Large changes in level in the plot are taken as suggestive of a particular number of groups. Like the analogous procedure for judging dendrograms discussed in the previous section, this approach may, of course, be very subjective. More formal techniques have been suggested by a number of authors – for example, Beale (1969), Calinski and Harabasz (1974) and Marriott (1971) – none of which is totally successful, as indicated by the results of the simulation study reported in Milligan and Cooper (1985).

The use of the optimization approach to clustering will be illustrated using again the Tibetan skull data given in Table 6.2. The two-group solutions found by minimization of trace(**W**) and by minimization of det(**W**) are displayed graphically in Figure 6.13 by placing them in the space of the first two principal components of the correlation matrix of the data. (These two components are the only ones with eigenvalues greater than 1, and together they account for 60% of the variation in the data.) The two solutions are quite different. For comparison, the two groups as defined by Morant (1923) are shown in Figure 6.14, also placed in the space of the first two principal components. The two-group solution produced by minimizing det(**W**) is perhaps a little closer to the a priori grouping.

Scott and Symons (1971) demonstrate how the two clustering criteria described above, and others that have been suggested, arise from consideration of a formal probability model for clustering. The model assumes that the

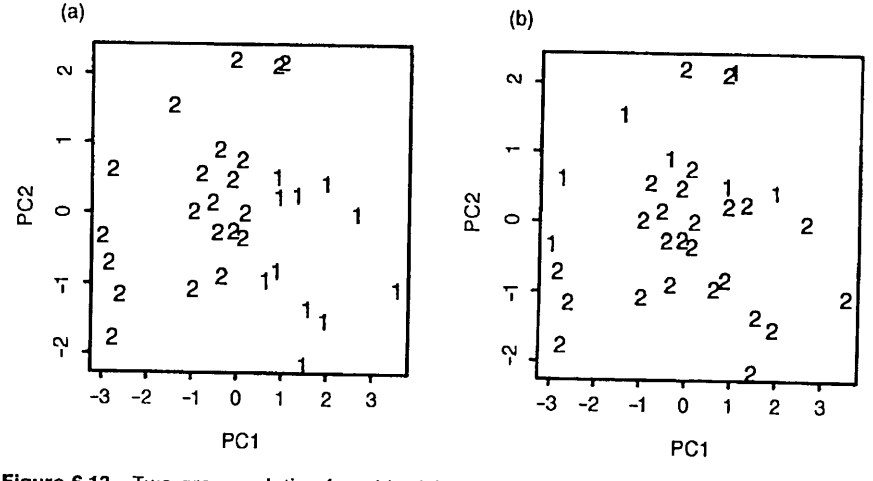

Figure 6.13 Two-group solution found by (a) minimization of trace(**W**) and (b) minimization of det(**W**) clustering applied to the Tibetan skull data, displayed in the space of the first two principal components of the correlation matrix of the data.

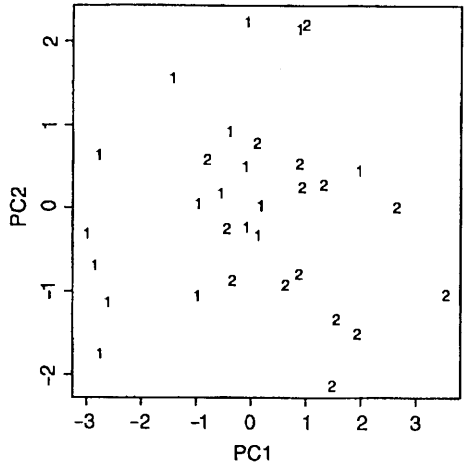

Figure 6.14 Original groups for Tibetan skull data displayed in the space of the first two principal components of the correlation matrix of the data.

population of interest consists of g different subpopulations and that the density of a p-dimensional observation, $\mathbf{x}$, from the kth subpopulation is $f_k(\mathbf{x};\boldsymbol{\theta}_k)$ for some unknown vector of parameters $\boldsymbol{\theta}_k$. Given observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and letting $\boldsymbol{\gamma}' = [\gamma_1,\ldots,\gamma_n]$ denote the identifying labels, where $\gamma_i = k$ if $\mathbf{x}_i$ comes from the kth subpopulation, then $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_k$ and $\boldsymbol{\gamma}$ are chosen so as to maximize the likelihood

$$L(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\gamma}) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i;\boldsymbol{\theta}_{\gamma_i}). \tag{6.9}$$

Scott and Symons show that when $f_k(\mathbf{x};\boldsymbol{\theta}_k)$ is multivariate normal with $\boldsymbol{\theta}_k$ now being the mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, then the trace(W) and det(W) clustering criteria described previously are equivalent to this likelihood approach under the following conditions.

- trace(W): equivalent to maximizing L in (6.9) under the assumption $\boldsymbol{\Sigma}_k = \sigma^2\mathbf{I}$.
- det(W): equivalent to maximizing L under the assumption $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, $k = 1,\ldots,g$.

Banfield and Raftery (1993) extend Scott and Symons's approach, showing that when the covariance matrices $\boldsymbol{\Sigma}_k$ are not constrained, maximizing L is equivalent to the minimization of $\sum_{k=1}^{g} n_k \log|\mathbf{W}_k/n_k|$. They then develop a number of new clustering criteria which are more general than those of Friedman and Rubin such as minimization of det(W), but more parsimonious than when completely unconstrained covariance matrices are allowed. The key to their approach is a reparameterization of the covariance matrix $\boldsymbol{\Sigma}_k$ in terms of its eigenvalue decomposition

$$\boldsymbol{\Sigma}_k = \mathbf{D}_k\boldsymbol{\Lambda}_k\mathbf{D}_k', \tag{6.10}$$

where $\mathbf{D}_k$ is the matrix of eigenvectors and $\mathbf{\Lambda}_k$ is a diagonal matrix with the eigenvalues of $\mathbf{\Sigma}_k$ on the diagonal.

To illustrate the Banfield and Raftery techniques one of their criteria which allows both size and orientation to vary between clusters will be used to cluster the air pollution data used and described previously in Chapter 2. We shall use the ecology and climate variables to cluster the states. The sulphur dioxide concentration will then be used in an attempt to partially validate the derived cluster solution. Since both Chicago (city 11) and Philadelphia (city 29) were identified as outliers in Chapter 3 (Exercise 3.7), they will be left out of the cluster analysis. The dendrogram resulting from applying Banfield and Raftery's S^* criterion is shown in Figure 6.15. This is labelled with the *original*

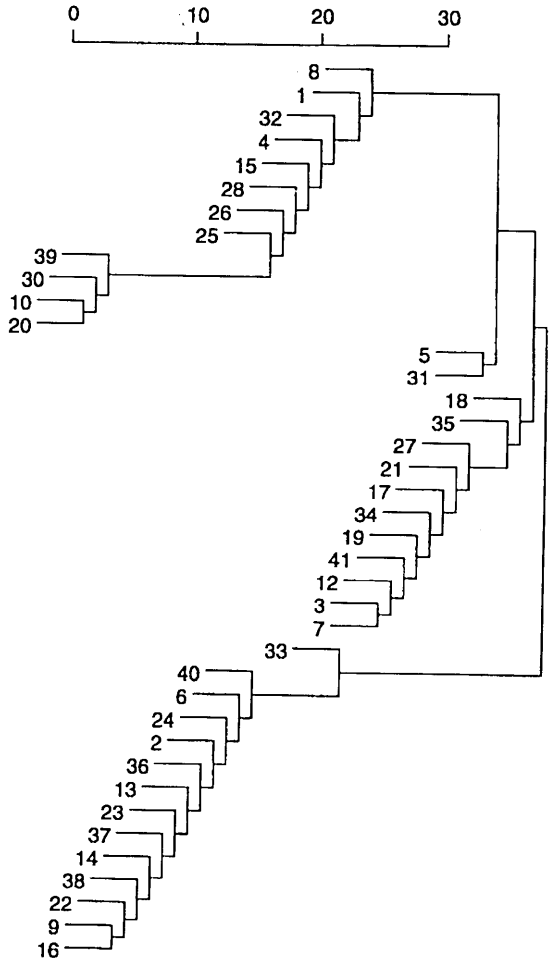

Figure 6.15 Dendrogram produced by clustering method of Banfield and Raftery applied to the air pollution data in US cities.

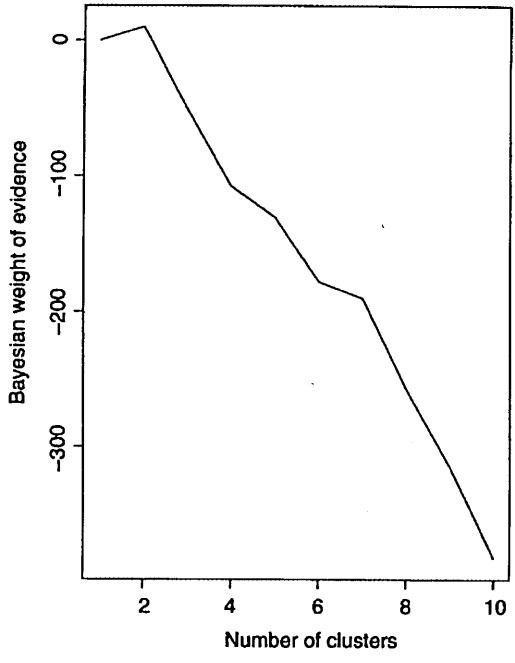

Figure 6.16 Bayesian weight of evidence indicator for number of groups.

number labelling each city. The dendrogram appears to suggest that the data contain two groups. A further indicator of number of groups is provided by an index described in Banfield and Raftery (1993) and called *Bayesian weight of evidence*. Plotting this index against number of groups gives Figure 6.16. The largest value of the index indicates the number of groups to choose. Here this is the value 2. The cities in each group are as follows:

Group 1: Phoenix, San Francisco, Denver, Hartford, Washington, Jacksonville, Atlanta, Indianapolis, Louisville, Baltimore, Detroit, Minneapolis-St. Paul, Kansas City, Albany, Buffalo, Cincinnati, Cleveland, Columbus, Pittsburgh, Providence, Memphis, Dallas, Houston, Seattle, Milwaukee.

Group 2: Little Rock, Wilmington, Miami, Des Moines, Wichita, New Orleans, St. Louis, Omaha, Albuquerque, Nashville, Salt Lake City, Norfolk, Richmond, Charleston.

Attempting to partially validate derived cluster analysis solutions by examining differences between clusters on variables *not* used in constructing them can often be useful. Here we can examine the group difference on sulphur dioxide concentration. A boxplot of this variable for each of the two derived clusters is shown in Figure 6.17. A t-test of the difference gives $t = 0.91$ with 37 d.f. The associated p-value is 0.37. The clusters found from using climate and ecology variables do not appear to be associated with differences in sulphur dioxide concentration.

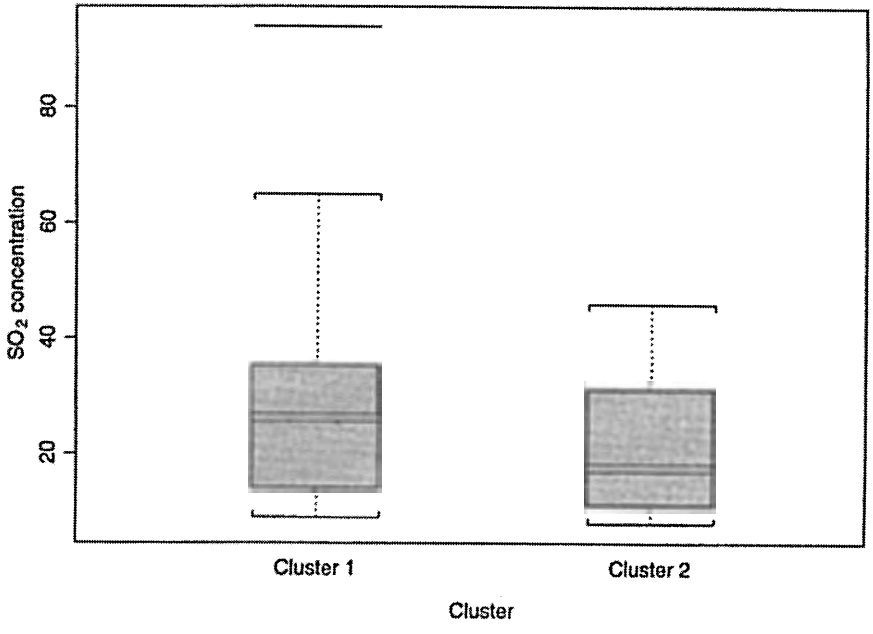

Figure 6.17 Boxplots of SO_2 concentration in each cluster of the two-cluster solution from Banfield and Raftery clustering of air pollution data.

6.4 Finite mixture models for cluster analysis

A further probability model approach to clustering is one based on finite mixture models. To introduce this approach, consider taking a random sample of people living in London and recording for each sample member their height. What might be a sensible model for the distribution of this variable in the population? First, we have to allow for our sample containing both males and females, since it is well known that, on average, males are taller than females. Within each sex it might be reasonable to assume that height is normally distributed with a particular mean and variance. Such considerations lead naturally to the following probability density function for height:

$$f(\text{height}) = pN(\mu_f, \sigma_f) + (1-p)N(\mu_m, \sigma_m), \tag{6.11}$$

where p is the proportion of females in the population.

A density function of the form given in (6.11) is an example of a finite mixture – see Everitt and Hand (1981) and Titterington *et al.* (1985) for details. In our particular example the main concern would be to use the sample of recorded heights to estimate the five parameters of the density function. Of course, if we had been sensible enough to record the sex of each member of the sample, estimation of these quantities would have been straightforward; here this could have been done very easily. In other areas, however, sexing of species is

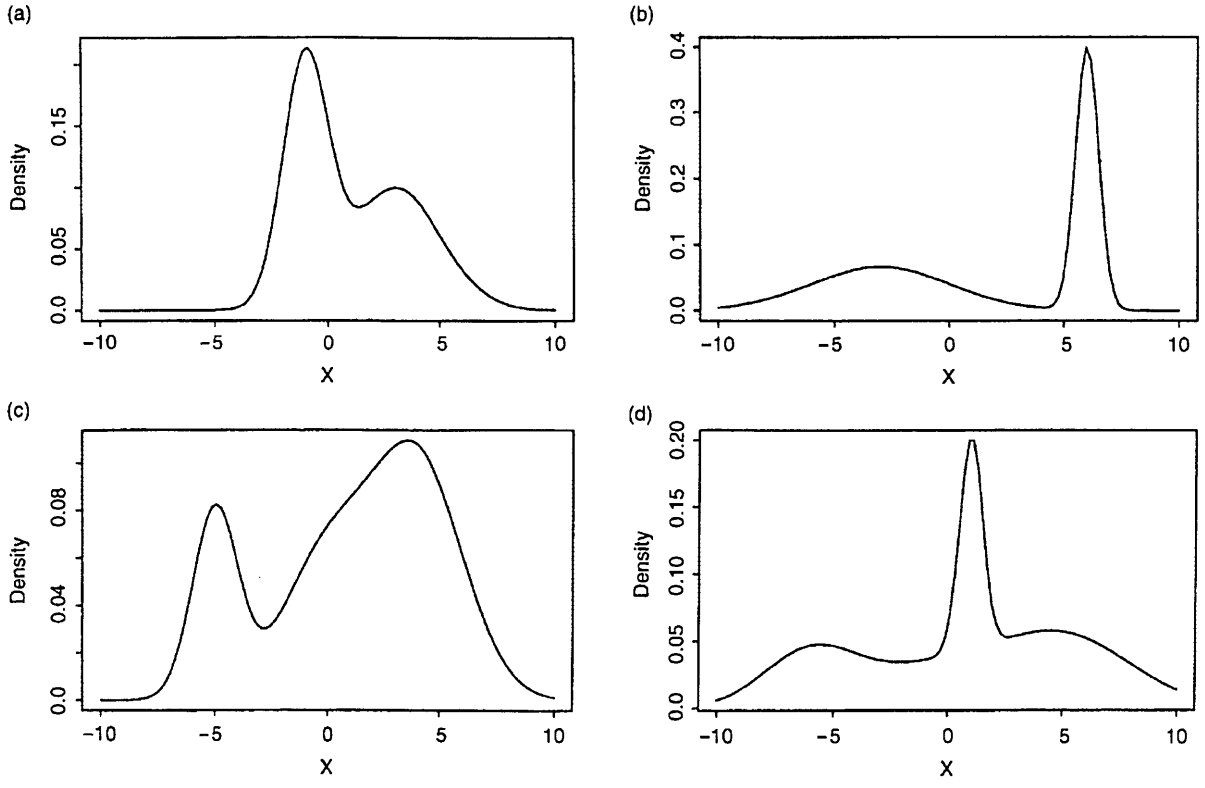

Figure 6.18 Examples of finite mixtures of univariate normal densities (a) $g = 2$, $p_1 = 0.5$, $p_2 = 0.5$, $\mu_1 = -1$, $\sigma_1 = 1$, $\mu_2 = 3$, $\sigma_2 = 2$; (b) $g = 2$, $p_1 = 0.5$, $p_2 = 0.5$, $\mu_1 = 6$, $\sigma_1 = 0.5$, $\mu_2 = -3$, $\sigma_2 = 3$; (c) $g = 3$, $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.5$, $\mu_1 = -5$, $\sigma_1 = 1$, $\mu_2 = 0$, $\sigma_2 = 2$, $\mu_3 = 4$, $\sigma_3 = 2$; (d) $g = 4$, $p_1 = 0.2$, $p_2 = 0.2$, $p_3 = 0.2$, $p_4 = 0.4$, $\mu_1 = -6$, $\sigma_1 = 2$, $\mu_2 = -1$, $\sigma_2 = 3$, $\mu_3 = 1$, $\sigma_3 = 0.5$, $\mu_4 = 5$, $\sigma_4 = 3$.

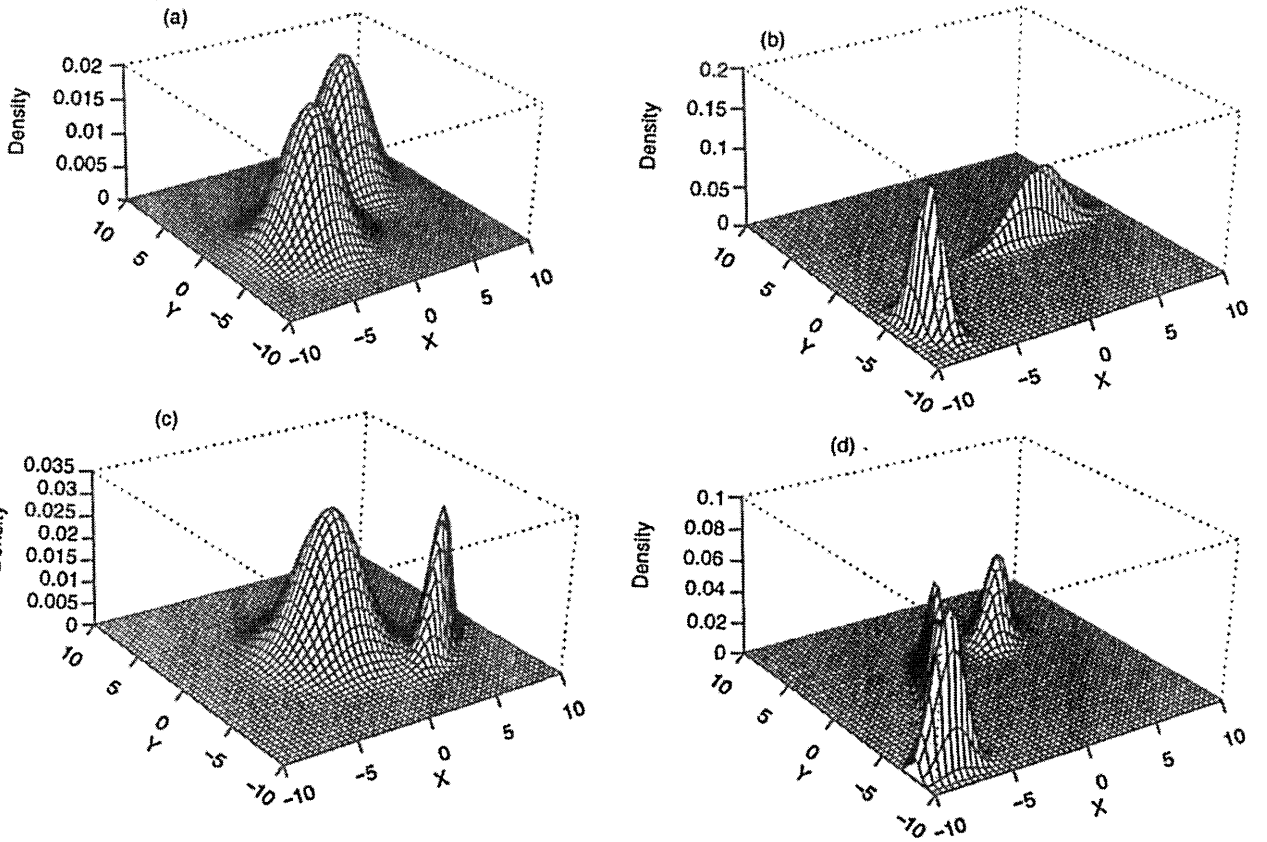
(a)
Density
0.02
0.015
0.01
0.005
0
10
5
0
Y
-5
-10
-10
-5
0
5
10
X
(b)
Density
0.2
0.15
0.1
0.05
0
10
5
0
Y
-5
-10
-10
-5
0
5
10
X
(c)
Density
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
10
5
0
Y
-5
-10
-10
-5
0
5
10
X
(d)
Density
0.1
0.08
0.06
0.04
0.02
0
10
5
0
Y
-5
-10
-10
-5
0
5
10
X

Figure 6.19 Examples of finite mixtures of bivariate normal densities:

(a) $g = 2$, $p_1 = 0.5$, $p_2 = 0.5$, $\boldsymbol{\mu}_1' = [4, 4]$, $\boldsymbol{\mu}_2' = [-2, -3]$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 4.00 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 4.00 \end{pmatrix};$$

(b) $g = 2$, $p_1 = 0.5$, $p_2 = 0.5$, $\boldsymbol{\mu}_1' = [0, 4]$, $\boldsymbol{\mu}_2' = [-6, -8]$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.25 & 0.50 \\ 0.50 & 4.00 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.00 & -0.20 \\ -0.20 & 0.25 \end{pmatrix};$$

(c) $g = 2$, $p_1 = 0.8$, $p_2 = 0.2$, $\boldsymbol{\mu}_1' = [0, 0]$, $\boldsymbol{\mu}_2' = [-3, 6]$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 4.00 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.40 & 1.30 \\ 1.30 & 2.00 \end{pmatrix};$$

(d) $g = 3$, $p_1 = 0.3$, $p_2 = 0.3$, $p_3 = 0.4$, $\boldsymbol{\mu}_1' = [4, 4]$, $\boldsymbol{\mu}_2' = [-8, -8]$, $\boldsymbol{\mu}_3' = [-1, -4]$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.25 \end{pmatrix}, \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 4.00 & 1.80 \\ 1.80 & 1.00 \end{pmatrix}.$$

more difficult and estimation of the parameters from the unlabelled sample becomes a practical necessity.

The mixture density given by (6.11) involves two univariate normal components. It is useful as a model for *two* groups in *univariate* data where the single variable is continuous. The extension to more than two groups is relatively simple, involving the mixture given by

$$f(x) = \sum_{i=1}^{g} p_i N(\mu_i, \sigma_i), \tag{6.12}$$

where g is the number of groups assumed. A total of $3g - 1$ parameters now need estimating. Some examples of $f(x)$ in (6.12) are shown in Figure 6.18.

Extending the model to deal with *multivariate data* is also simple in principle. In (6.12) the univariate normal components of the mixture are replaced by the corresponding multivariate densities with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$

$$f(\mathbf{x}) = \sum_{i=1}^{g} p_i MVN(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{6.13}$$

Now there are $g - 1$ mixing proportions, gd means and $gd(d + 1)/2$ variances and covariances to estimate. Some examples of mixtures of bivariate normal densities are shown in Figure 6.19.

Clearly estimation of the large number of parameters in multivariate normal mixtures is going to be a formidable computational problem, but it can be handled in most cases by maximum likelihood methods, the essentials of which are given in Box 6.1. Further details are given in Everitt and Hand (1981). A numerical example of this approach to clustering is given later.

If the variables being recorded are binary rather than continuous then a mixture density model based upon normal components would obviously not be realistic. The same general approach can, however, still be used but with a different choice of component density. One possibility is to assume that, within each cluster, responses to individual binary items are independent, with probabilities which are constant within clusters but different between clusters. If we make such an assumption, what is the probability density function of variables within a particular group? To answer this question, let us begin with an example in which there are three binary variables, x_1, x_2, x_3; within a particular group j, the probabilities of a positive response for each of the variables are θ_{j1}, θ_{j2}, θ_{j3}. Assuming that the three variables are independent of one another within this group, we can now find the probability of observing any value of the vector $\mathbf{x}' = [x_1, x_2, x_3]$. For example,

$$\Pr[\mathbf{x}' = (0, 1, 1)] = (1 - \theta_{j1})\theta_{j2}\theta_{j3} \tag{6.14}$$

or

$$\Pr[\mathbf{x}' = (1, 0, 0)] = \theta_{j1}(1 - \theta_{j2})(1 - \theta_{j3}). \tag{6.15}$$

Box 6.1 Maximum likelihood estimation of parameters in multivariate normal mixture distributions

- Suppose we have n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The log-likelihood function for a multivariate normal mixture of the form given in (6.13) is

$$L = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{g} p_k MVN(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

- The maximum likelihood equations are obtained by equating the first partial derivatives of L with respect to the p_k, the elements of each matrix $\boldsymbol{\Sigma}_k$ and those of the vectors $\boldsymbol{\mu}_k$, to zero.
- Everitt and Hand (1981) show that the resulting equations can be written in the following form:

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{\Pr}(k|\mathbf{x}_i), \qquad k = 1, \ldots, g-1,$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n\hat{p}_k} \sum_{i=1}^{n} \widehat{\Pr}(k|\mathbf{x}_i)\mathbf{x}_i, \qquad k = 1, \ldots, g,$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{np_k} \sum_{i=1}^{n} \widehat{\Pr}(k|\mathbf{x}_i)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)', \qquad k = 1, \ldots, g$$

- In these equations $\widehat{\Pr}$ denotes the estimated posterior probability of an observation $\mathbf{x}_i$ belonging to component k, that is,

$$\widehat{\Pr}(k|\mathbf{x}_i) = \frac{\hat{p}_k MVN(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{j=1}^{g} p_j MVN(\mathbf{x}_{ij}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}.$$

- Written in this form, we can see that the maximum likelihood equations for the parameters in a mixture of multivariate normals are closely analogous to those for a single normal distribution, except that each sample point is now weighted by the estimated posterior probability of belonging to a particular component.
- The equations are solved by an iterative scheme in which initial estimates of the posterior probabilities are used to find initial estimates of the parameters, and these are then used to calculate improved estimates of the posterior probabilities, the procedure alternating between these two steps until convergence. This as an example of the application of the *EM algorithm* (see Dempster *et al.*, 1977).

Both (6.14) and (6.15) may be rewritten in the form

$$\Pr[\mathbf{x}] = \theta_{j1}^{x_1}(1-\theta_{j1})^{1-x_1}\theta_{j2}^{x_2}(1-\theta_{j2})^{1-x_2}\theta_{j3}^{x_3}(1-\theta_{j3})^{1-x_3}$$

$$= \prod_{l=1}^{3} \theta_{jl}^{x_l}(1-\theta_{jl})^{1-x_l}. \tag{6.16}$$

This is known as a *multivariate Bernoulli density*, and it can be extended to the situation with p binary variables in an obvious fashion:

$$\Pr[\mathbf{x}] = \Pr[(x_1, x_2, \ldots, x_p)] = \prod_{l=1}^{p} \theta_{jl}^{x_l}(1-\theta_{jl})^{1-x_l}. \tag{6.17}$$

Such density functions now take the place of the normal components of (6.16). Parameter estimation is again by maximum likelihood; for details, see Goodman (1974) and Everitt and Hand (1981). (Finite mixture densities based upon components with the form given in (6.16) are essentially equivalent to the *latent class model* proposed by Lazarsfeld and Henry, 1968.) A numerical example of this approach to the clustering of binary data is given below.

An obvious candidate for assessing number of groups when fitting mixture distributions is a *likelihood ratio test* of, say, g_1 against g_2 components in the mixture. Under the null hypothesis of g_1 groups, this statistic is generally assumed to be asymptotically distributed as chi-squared with degrees of freedom equal to the difference in the number of parameters in the two mixtures being compared. Unfortunately, such a test suffers from a number of problems which are discussed in McLachlan and Basford (1988). Attempts

Table 6.3 City crime data (per 100 000 population)

City	Murder	Rape
1. Atlanta	16.5	24.8
2. Boston	4.2	13.3
3. Chicago	11.6	24.7
4. Dallas	18.9	34.2
5. Denver	6.9	41.5
6. Detroit	13.0	35.7
7. Hartford	2.5	8.8
8. Honolulu	3.6	12.7
9. Houston	16.8	26.6
10. Kansas City	10.8	43.2
11. Los Angeles	9.7	51.8
12. New Orleans	10.3	39.7
13. New York	9.4	19.4
14. Portland	5.9	23.0
15. Tucson	5.1	22.9
16. Washington	12.5	27.6

Source: Hartigan (1975).

to improve the test using a Bayesian approach are described in Richardson and Green (1997).

In the behavioural sciences, data often contain *both* continuous and categorical variables. A finite mixture model appropriate for such data is described by Everitt (1988) and further elaborated in Everitt and Merette (1990).

Finite mixture models have, over the last decade, become a topic of great research interest and they have been used in a variety of applications. Some particularly interesting examples are given in Everitt and Bullmore (1999), McLaren (1996) and Schork *et al.* (1996).

6.4.1 Some numerical examples of the application of mixture distributions

Our first example concerns the application of the multivariate normal mixture model in (6.13) to the data in Table 6.3, which gives murder/manslaughter and rape rates for 16 cities in the USA. Since the data involve only two variables, they may be plotted as shown in Figure 6.20. Such a diagram will

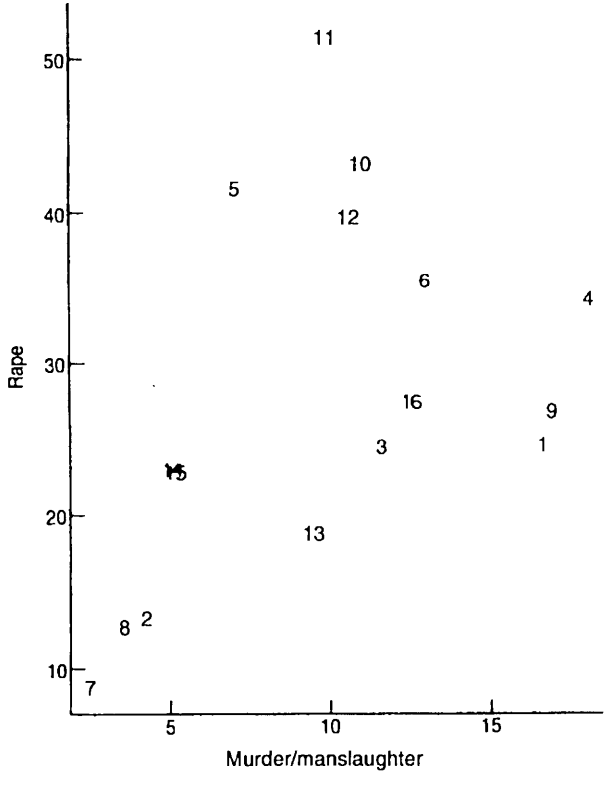

Figure 6.20 Murder/manslaughter and rape rates for 16 cities in the USA.

Table 6.4 Results of fitting two-component bivariate normal mixture to the city crime data in Table 6.3

Starting values	Final values
$\hat{p} = 0.438$	$\hat{p} = 0.385$
$\hat{\boldsymbol{\mu}}_1' = [13.99, 27.57]$	$\hat{\boldsymbol{\mu}}_1' = [14.40, 27.97]$
$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 8.64 & 26.08 \\ 26.08 & 136.19 \end{pmatrix}$	$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 9.00 & 24.53 \\ 24.53 & 136.42 \end{pmatrix}$

allow the results from the mixture analysis to be compared with those obtained by 'visual' analysis.

We shall begin by fitting a two-component bivariate normal mixture using maximum likelihood methods as in Box 6.1. Starting values for the parameters were obtained from the minimization of trace(W) clustering procedure and the final parameter values were obtained after 12 iterations of the estimation algorithm. (In this example we have assumed that the correlation between the two variables, murder rate and rape rate, is the same in each group.) The results are shown in Table 6.4. The final parameter estimates shown in this table may now be used to find estimates of the posterior probabilities of each city belonging to each of the component densities in the mixture. These are given in Table 6.5. The maximum posterior probabilities can be used to partition the cities into two groups. This partition is shown in Figure 6.21. Here the

Table 6.5 Estimated posterior probabilities for city crime data

City	$\Pr(1\|\mathbf{x}_i)$	$\Pr(2\|\mathbf{x}_i)$
1. Atlanta	1.00	0.00
2. Boston	0.00	1.00
3. Chicago	0.91	0.09
4. Dallas	1.00	0.00
5. Denver	0.00	1.00
6. Detroit	0.78	0.22
7. Hartford	0.00	1.00
8. Honolulu	0.00	1.00
9. Houston	1.00	0.00
10. Kansas City	0.00	1.00
11. Los Angeles	0.00	1.00
12. New Orleans	0.01	0.99
13. New York	0.54	0.45
14. Portland	0.00	1.00
15. Tucson	0.00	1.00
16. Washington	0.95	0.05

On the basis of these probabilities group 1 consists of cities 1, 3, 4, 6, 9, 13, 16; group 2 consists of cities 2, 5, 7, 8, 10, 11, 12, 14, 15.

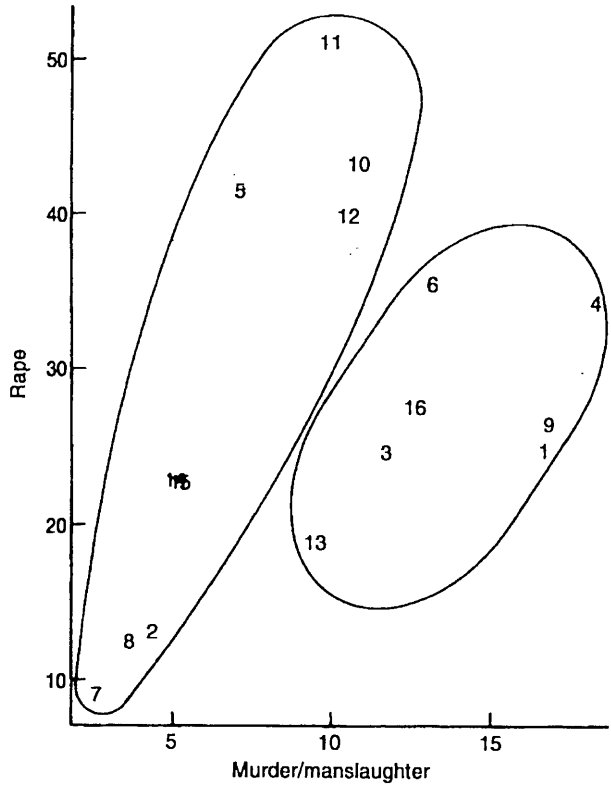

Figure 6.21 Two groups of cities given by estimated posterior probabilities found after fitting two-component bivariate normal mixture to city crime data.

two groups differ predominantly in the murder/manslaughter rate, with those in the first group having very high values.

The next example relates to fitting the mixture density based on components as given in (6.16) to sets of binary data collected on psychiatric patients during an investigation of social networks. Each patient was asked to give the names of all their acquaintances and, for each name supplied, to say whether they regarded the person as a friend or not, somebody they could confide in or not, somebody they would miss or not, and finally whether the acquaintance was an active one or not. Table 6.6 gives the counts of the number of names falling into each of the 16 possible categories. Latent class analysis was applied and two-class and three-class models fitted. The results are shown in Table 6.7. In the former the division is clearly into 'close friends' and perhaps 'just names'. In the three-group solution this division is again clear, with the additional group being 'friends in whom one would not confide'. This description of the data proved extremely useful in other investigations, and more details are given in Dunn and Everitt (1988).

Table 6.6 Data collected from long-stay psychiatric patients

Active	Confides	Friend	Missed	Frequency
1	1	1	1	529
1	1	1	2	424
1	1	2	1	51
1	1	2	2	193
1	2	1	1	185
1	2	1	2	274
1	2	2	1	46
1	2	2	2	311
2	1	1	1	81
2	1	1	2	279
2	1	2	1	13
2	1	2	2	228
2	1	1	1	25
2	2	1	2	256
2	2	2	1	18
2	2	2	2	1893

Table 6.7 Latent class results for psychiatric data in Table 6.5

	p	Pr(Active)	Pr(Confides)	Pr(Friend)	Pr(Missed)
(a) Two-class solution					
Class 1	0.56	0.13	0.00	0.11	0.10
Class 2	0.44	0.78	0.44	0.71	0.84
(b) Three-class solution					
Class 1	0.41	0.08	0.01	0.02	0.01
Class 2	0.25	0.89	0.73	0.73	0.90
Class 3	0.34	0.48	0.03	0.53	0.58

6.5 Summary

Cluster analysis techniques are potentially very useful for the exploration of complex multivariate data, although their use in practice requires care if misleading solutions are to be avoided. The wider use of methods based on relatively sound and sensible models such as mixture distributions or the criteria introduced by Banfield and Raftery (1993) is an encouraging sign and will hopefully lead to more convincing applications of clustering than has often been the case in the past.

Exercises

6.1 Show that the entries of the cophenetic matrix satisfy the *ultrametric inequality*, $d(x, y) \leq \max[d(x, z), d(y, z)]$.

6.2 Show that the inter-cluster distances used by single linkage, complete linkage and group average clustering satisfy the following formula:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \gamma |d_{ki} - d_{kj}|,$$

where

$$\alpha_i = \alpha_j,\ \gamma = -\tfrac{1}{2} \qquad \text{(single linkage)},$$

$$\alpha_i = \alpha_j,\ \gamma = \tfrac{1}{2} \qquad \text{(complete linkage)},$$

$$\alpha_i = \frac{n_i}{n_i + n_j},\ \alpha_j = \frac{n_j}{n_i + n_j},\ \gamma = 0 \qquad \text{(group average)}.$$

($d_{k(ij)}$ is the distance between a group k and a group (ij) formed by the fusion of groups i and j, and d_{ij} is the distance between groups i and j; n_i and n_j are the number of observations in groups i and j.)

6.3 Ward (1963) proposed an agglomerative hierarchical clustering procedure in which, at each step, the union of every possible pair of clusters is considered and the two clusters whose fusion results in the minimum increase in an error sum-of-squares criterion, *ESS*, are combined. For a single variable, *ESS* for a group with n individuals is simply $ESS = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

(a) If ten individuals with variable values $\{2, 6, 5, 6, 2, 2, 2, 0, 0, 0\}$ are considered as a single group, calculate *ESS*. If the individuals are grouped into two groups with individuals 1, 5, 6, 7, 8, 9, 10 in one group and individuals 2, 3, 4 in the other, what does *ESS* become?

(b) Can you fit Ward's method into the general equation given in Exercise 6.2?

6.4 The data in Table 6.8 give life expectancy by country and age for males in 15 countries. Find the estimates of the parameters in a two-component multivariate normal mixture model for the data.

Table 6.8 Male life expectancy, by country and age

	Age			
Country	0	25	50	75
1. Algeria	53	51	30	13
2. Costa Rica	65	48	26	9
3. El Salvador	56	44	25	10
4. Greenland	60	44	22	6
5. Grenada	61	45	22	8
6. Honduras	59	42	22	6
7. Mexico	59	44	24	8
8. Nicaragua	65	48	28	14
9. Panama	65	48	26	9
10. Trinidad	64	43	21	6
11. Chile	59	47	23	10
12. Ecuador	57	46	25	9
13. Argentina	65	46	24	9
14. Tunisia	56	46	24	11
15. Dominican Republic	64	50	28	11

Table 6.9 Acceptability of suicide

Response pattern				
1	2	3	4	Frequency
1	1	1	1	105
2	1	1	1	0
1	2	1	1	1
2	2	1	1	0
1	1	2	1	4
2	1	2	1	0
1	2	2	1	4
2	2	2	1	3
1	1	1	2	10
2	1	1	2	1
1	2	1	2	3
2	2	1	2	3
1	1	2	2	62
2	1	2	2	16
1	2	2	2	444
2	2	2	2	724

Situations: (1) person with an incurable disease, (2) person is tired of living, (3) person has been dishonoured, (4) person has gone bankrupt.

6.5 Respondents in a survey were asked whether or not suicide was acceptable in four different situations. Responses were coded 1 for yes and 2 for no. The result are summarized in Table 6.9. Fit a latent class model with two classes.

Chapter 6

6.1 Let x, y and z be any three objects and suppose that: at fusion level α_j, x and y are in the same cluster; and at fusion level α_k, y and z are in the same cluster. Since the clusters are hierarchical, one of these includes the other. This will be the cluster corresponding to the larger of j and k. Let this be the integer e so that at α_e, x, y and z are all in the same cluster.

Then

$$d(x,z) \leq \alpha_e,$$

but since $e = \max\{j,k\}$,

$$\alpha_e = \max\{\alpha_j, \alpha_k\}$$

so that

$$d(x,z) \leq \max\{\alpha_j, \alpha_k\},$$

that is,

$$d(x,z) \leq \max\{d(x,y), d(y,z)\}.$$

6.2 The single linkage distance between cluster k and the cluster formed by the fusion of clusters i and j is defined to be $\min\{d_{ki}, d_{kj}\}$. According to the formula given in Exercise 6.2, this distance is

$$d_{k(ij)} = \tfrac{1}{2}d_{ki} + \tfrac{1}{2}d_{kj} - \tfrac{1}{2}|d_{ki} - d_{kj}|.$$

If $d_{ki} > d_{kj}$, then

$$|d_{ki} - d_{kj}| = d_{ki} - d_{kj};$$

therefore

$$d_{k(ij)} = d_{kj}.$$

If $d_{ki} < d_{kj}$, then

$$|d_{ki} - d_{kj}| = d_{kj} - d_{ki};$$

therefore

$$d_{k(ij)} = d_{ki}.$$

So the formula gives that $d_{k(ij)} = \min\{d_{ki}, d_{ki}\}$, as required.

6.3 (a) *ESS* for the ten observations is simply

$$ESS = (2 - 2.5)^2 + (6 - 2.5)^2 + \ldots + (0 - 2.5)^2 = 50.5.$$

(b) Here the *ESS* is found as the sum of the values for each group of individuals.

Chapter 7

7.2 For factors A, B and C; Minimal model: $\ln(n) = A + B + C$. Saturated model: $\ln(n) = A + B + C + A.B + A.C + B.C + A.B.C$.

7.3 The reciprocal link.