# CHAPTER 15

# CLUSTER ANALYSIS

## 15.1 INTRODUCTION

In *cluster analysis* we search for patterns in a data set by grouping the (multivariate) observations into clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar but the clusters are dissimilar to each other. We hope to find the natural groupings in the data, groupings that make sense to the researcher.

Cluster analysis differs fundamentally from classification analysis (Chapter 9). In classification analysis, we allocate the observations to a known number of predefined groups or populations. In cluster analysis, neither the number of groups nor the groups themselves are known in advance.

To group the observations into clusters, many techniques begin with similarities between all pairs of observations. In many cases the similarities are based on some measure of distance. Other cluster methods use a preliminary choice for cluster centers or a comparison of within- and between-cluster variability. It is also possible to cluster the variables, in which case the similarity could be a correlation; see Section 15.7.

*Methods of Multivariate Analysis, Third Edition.* By Alvin C. Rencher and William F. Christensen 

We can search for clusters graphically by plotting the observations. If there are only two variables ($p = 2$), this can be done in a scatterplot (see Section 3.3). For $p > 2$, we can plot the data in two dimensions using principal components (see Section 12.4) or biplots (see Section 16.3). For an example of a principal component plot, see Figure 12.7 in Section 12.4, in which four clear groupings of points can be observed. Another approach to plotting is provided by *projection pursuit*, which seeks two-dimensional projections that reveal clusters [see Friedman and Tukey (1974); Huber (1985); Sibson (1984); Jones and Sibson (1987); Yenyukov (1988); Posse (1990); Nason (1995); Ripley (1996, pp. 296–303)].

Cluster analysis has also been referred to as classification, pattern recognition (specifically, unsupervised learning), and numerical taxonomy. The techniques of cluster analysis have been extensively applied to data in many fields, such as medicine, psychiatry, sociology, criminology, anthropology, archaeology, geology, geography, remote sensing, market research, economics, and engineering.

We shall concentrate largely on quantitative variables [for categorical variables, see Gordon (1999) or Everitt (1993)]. The data matrix [see (3.17)] can be written as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix} = (\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \cdots, \mathbf{y}_{(p)}), \tag{15.1}$$

where $\mathbf{y}_i'$ is a row (observation vector) and $\mathbf{y}_{(j)}$ is a column (corresponding to a variable). We generally wish to group the $n$ $\mathbf{y}_i'$'s (rows) into $g$ clusters. We may also wish to cluster the columns $\mathbf{y}_{(j)}, j = 1, 2, \ldots, p$ (see Section 15.7).

Two common approaches to clustering the observation vectors are hierarchical clustering and partitioning. In *hierarchical* clustering we typically start with $n$ clusters, one for each observation, and end with a single cluster containing all $n$ observations. At each step, an observation or a cluster of observations is absorbed into another cluster. We can also reverse this process, that is, start with a single cluster containing all $n$ observations and end with $n$ clusters of a single item each (see Section 15.3.10). In *partitioning*, we simply divide the observations into $g$ clusters. This can be done by starting with an initial partitioning or with cluster centers and then reallocating the observations according to some optimality criterion. Other clustering methods that we will discuss are based on fitting mixtures of multivariate normal distributions or searching for regions of high density sometimes called modes.

There is an abundant literature on cluster analysis. Useful monographs and reviews have been given by Gordon (1999), Everitt (1993), Khattree and Naik (2000, Chapter 6), Kaufman and Rousseuw (1990), Seber (1984, Chapter 7), Anderberg (1973), and Hartigan (1975a).

## 15.2 MEASURES OF SIMILARITY OR DISSIMILARITY

Since cluster analysis attempts to identify the observation vectors that are similar and group them into clusters, many techniques use an index of *similarity* or *proximity*

between each pair of observations. A convenient measure of proximity is the distance between two observations. Since distance increases as two units become further apart, distance is actually a measure of *dissimilarity*.

A common distance function is the Euclidean distance between two vectors $\mathbf{x} = (x_1, x_2, ..., x_p)'$ and $\mathbf{y} = (y_1, y_2, ..., y_p)'$, defined as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}. \tag{15.2}$$

To adjust for differing variances and covariances among the $p$ variables, we could use the statistical distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y})} \tag{15.3}$$

[see (3.79)], where $\mathbf{S}$ is the sample covariance matrix. After the clusters are formed, $\mathbf{S}$ could be computed as the pooled within-cluster covariance matrix, but we do not know beforehand what the clusters will be. If we compute $\mathbf{S}$ on the unpartitioned sample, there will be distortion of the variances and covariances because of the groups in the data (assuming there really are some natural clusters). We therefore usually use the Euclidean distance given by (15.2). In some clustering procedures, it is not necessary to take the square root in (15.2) or (15.3).

Other distance measures have been suggested, for example, the Minkowski metric

$$d(\mathbf{x}, \mathbf{y}) = \left[ \sum_{j=1}^{p} |x_j - y_j|^r \right]^{1/r}. \tag{15.4}$$

For $r = 2$, $d(\mathbf{x}, \mathbf{y})$ in (15.4) becomes the Euclidean distance given in (15.2). For $p = 2$ and $r = 1$, (15.4) measures the "city block" distance between two observations. There are distance measures for categorical data; see Gordon (1999, Chapter 2).

For the $n$ observation vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, we can compute an $n \times n$ matrix $\mathbf{D} = (d_{ij})$ of distances (or dissimilarities) where $d_{ij} = d(\mathbf{y}_i, \mathbf{y}_j)$ is usually given by (15.2), $d(\mathbf{y}_i, \mathbf{y}_j) = \sqrt{(\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j)}$. We sometimes use $\mathbf{D} = (d_{ij}^2)$, where $d_{ij}^2 = d^2(\mathbf{y}_i, \mathbf{y}_j) = (\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j)$ is the square of (15.2). The matrix $\mathbf{D}$ will typically be symmetric with diagonal elements equal to zero.

The scale of measurement of the variables is an important consideration when using the Euclidean distance measure in (15.2). Changing the scale can affect the relative distances among the items. For example, suppose three items have the following bivariate measurements $(y_1, y_2)$: (2, 5), (4, 2), (7, 9). Using $d_{ij}$ as given by (15.2), the matrix $\mathbf{D} = (d_{ij})$ for these items is

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 3.6 & 6.4 \\ 3.6 & 0 & 7.6 \\ 6.4 & 7.6 & 0 \end{pmatrix}.$$

However, if we multiply $y_1$ by 100 as, for example, in changing from meters to centimeters, the matrix becomes

$$\mathbf{D}_2 = \begin{pmatrix} 0 & 200 & 500 \\ 200 & 0 & 300 \\ 500 & 300 & 0 \end{pmatrix},$$

and the largest distance is now $d_{13}$ instead of $d_{23}$. The distance rankings have been altered by scaling.

To counter this problem, each variable could be standardized in the usual way by subtracting the mean and dividing by the standard deviation of the variable. However, such scaling would ordinarily be based on the entire data set, that is, on all $n$ values in each column of $\mathbf{Y}$ in (15.1). In this case, the variables that best separate clusters might no longer do so after division by standard deviations that include between-cluster variation. If we use standardized variables, the clusters could be less well separated. The question of scaling is therefore not an easy one. However, standardization of this type is recommended by many authors.

By (15.2), the squared Euclidean distance between two observations $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ and $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$ is $d^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p}(x_j - y_j)^2$. This can be expressed as

$$d^2(\mathbf{x}, \mathbf{y}) = (v_x - v_y)^2 + p(\overline{x} - \overline{y})^2 + 2v_x v_y(1 - r_{xy}), \tag{15.5}$$

where $v_x^2 = \sum_{j=1}^{p}(x_j - \overline{x})^2$ and $\overline{x} = \sum_{j=1}^{p} x_j/p$, with similar expressions for $v_y^2$ and $\overline{y}$. The correlation $r_{xy}$ in (15.5) is given by

$$r_{xy} = \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2}}. \tag{15.6}$$

In Figure 15.1, we illustrate the profile (see Sections 5.9 and 6.8) for each of two observation vectors $\mathbf{x}$ and $\mathbf{y}$. The squared Eulcidean distance in (15.5) can be used to compare the profiles of $\mathbf{x}$ and $\mathbf{y}$ in terms of levels, variation, and shape, where $\overline{x}$ and $\overline{y}$ are the *levels* of the two profiles, $v_x$ and $v_y$ are the *variations* of the profiles, and the correlation $r_{xy}$ is a measure of the closeness of the *shapes* of the two profiles. The closer $r_{xy}$ is to 1, the greater is the similarity in shape of the two profiles. Note that $\overline{x}$ and $v_x$ are the mean and variation of the $p$ variables within the observation vector $\mathbf{x}$, not over the $n$ observations in the data set. A similar comment can be made about $\overline{y}$ and $v_y$. Likewise, the correlation $r_{xy}$ is between the two observation vectors $\mathbf{x}$ and $\mathbf{y}$, not between two variables. The use of $r_{xy}$ has been questioned by Jardine and Sibson (1971) and Wishart (1971), but Strauss et al. (1973) found the correlation to be superior to the Euclidean distance for finding the clusters in a particular data set.

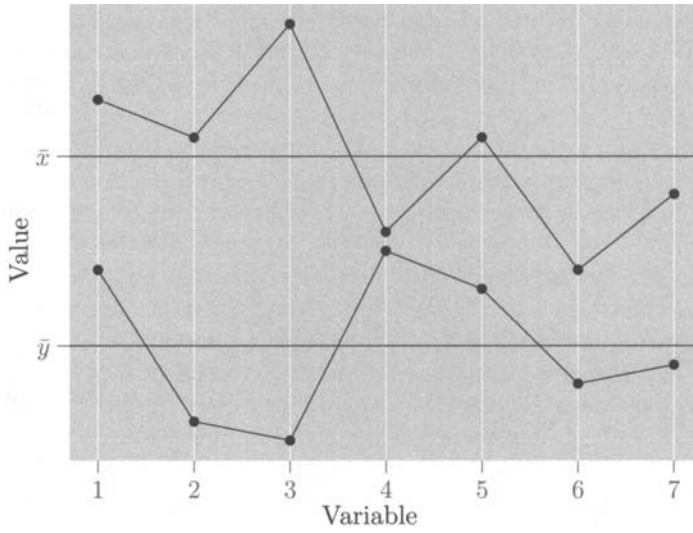


**Figure 15.1** Profiles for two observation vectors $\mathbf{x}$ and $\mathbf{y}$.

## 15.3 HIERARCHICAL CLUSTERING

### 15.3.1 Introduction

Hierarchical methods and other clustering algorithms represent an attempt to find "good" clusters in the data using a computationally efficient technique. It is not generally feasible to examine all possible clustering possibilities for a data set, especially a large one. The number of ways of partitioning a set of $n$ items into $g$ clusters is given by

$$N(n,g) = \frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^n \tag{15.7}$$

[see Duran and Odell (1974, Chapter 4), Jensen (1969), Seber (1984, p. 379)]. This can be approximated by $g^n/g!$, which is large even for moderate values of $n$ and $g$. For example, $N(25,10) \cong 2.8 \times 10^{18}$. The total possible number of clusters for a set of $n$ items is $\sum_{g=1}^{n} N(n,g)$, which, for $n = 25$, is greater than $10^{19}$. Hence, hierarchical methods and other approaches permit us to search for a reasonable solution without having to look at all possible arrangements.

As noted in Section 15.1, hierarchical clustering algorithms involve a sequential process. In each step of the *agglomerative* hierarchical approach, an observation or a cluster of observations is merged into another cluster. In this process, the number of clusters shrinks and the clusters themselves grow larger. We start with $n$ clusters (individual items) and end with one single cluster containing the entire data set. An alternative approach, called the *divisive* method, starts with a single cluster

containing all $n$ items and partitions a cluster into two clusters at each step (see Section 15.3.10). The end result of the divisive approach is $n$ clusters of one item each. Agglomerative methods are more commonly used than divisive methods. In either type of hierarchical clustering, a decision must be made as to the "optimal" number of clusters (see Section 15.5).

At each step of an agglomerative hierarchical approach, the two "closest" clusters are merged into a single new cluster. The process is therefore irreversible in the sense that any two items that are once lumped together in a cluster cannot be separated later in the procedure; any early mistakes cannot be corrected. Similarly, in a divisive hierarchical method, items cannot be moved to other clusters. An optional approach is to carry out a hierarchical procedure followed by a partitioning procedure in which items can be moved from one cluster to another (see Section 15.4.1).

Since an agglomerative hierarchical procedure combines the two "closest" clusters at each step, we must consider the question of measuring the similarity or dissimilarity of two clusters. Different approaches to measuring distance between clusters give rise to different hierarchical methods. Agglomerative techniques are discussed in Sections 15.3.2–15.3.9, and the divisive approach is considered in Section 15.3.10.

### 15.3.2 Single Linkage (Nearest Neighbor)

In the *single linkage* method, the distance between two clusters $A$ and $B$ is defined as the *minimum* distance between a point in $A$ and a point in $B$:

$$D(A, B) = \min \{d(\mathbf{y}_i, \mathbf{y}_j), \text{ for } \mathbf{y}_i \text{ in } A \text{ and } \mathbf{y}_j \text{ in } B\}, \tag{15.8}$$

where $d(\mathbf{y}_i, \mathbf{y}_j)$ is the Euclidean distance in (15.2) or some other distance between the vectors $\mathbf{y}_i$ and $\mathbf{y}_j$. This approach is also called the *nearest neighbor* method.

At each step in the single linkage method, the distance (15.8) is found for every pair of clusters, and we merge the two clusters with smallest distance. The number of clusters is therefore reduced by one. After two clusters are merged, the procedure is repeated for the next step: the distances between all pairs of clusters are calculated again, and the pair with minimum distance is merged into a single cluster.

The results of a hierarchical clustering procedure can be displayed graphically using a *tree diagram*, also known as a *dendrogram*, which shows all the steps in the hierarchical procedure, including the distances at which clusters are merged. Dendrograms are shown in Figures 15.2 and 15.3 in Examples 15.3.2(a) and 15.3.2(b).

■ **EXAMPLE 15.3.2(a)**

Hartigan (1975a, p. 28) compared the crime rates per 100,000 population for various cities. The data are in Table 15.1 (taken from the 1970 US Statistical Abstract). In order to illustrate the use of the distance matrix in single linkage clustering, we use the first six observations in Table 15.1 (Atlanta through Detroit).

**Table 15.1** City Crime Rates per 100,000 Population

| City | Murder | Rape | Robbery | Assault | Burglary | Larceny | AutoTheft |
|---|---|---|---|---|---|---|---|
| Atlanta | 16.5 | 24.8 | 106 | 147 | 1112 | 905 | 494 |
| Boston | 4.2 | 13.3 | 122 | 90 | 982 | 669 | 954 |
| Chicago | 11.6 | 24.7 | 340 | 242 | 808 | 609 | 645 |
| Dallas | 18.1 | 34.2 | 184 | 293 | 1668 | 901 | 602 |
| Denver | 6.9 | 41.5 | 173 | 191 | 1534 | 1368 | 780 |
| Detroit | 13.0 | 35.7 | 477 | 220 | 1566 | 1183 | 788 |
| Hartford | 2.5 | 8.8 | 68 | 103 | 1017 | 724 | 468 |
| Honolulu | 3.6 | 12.7 | 42 | 28 | 1457 | 1102 | 637 |
| Houston | 16.8 | 26.6 | 289 | 186 | 1509 | 787 | 697 |
| Kansas City | 10.8 | 43.2 | 255 | 226 | 1494 | 955 | 765 |
| Los Angeles | 9.7 | 51.8 | 286 | 355 | 1902 | 1386 | 862 |
| New Orleans | 10.3 | 39.7 | 266 | 283 | 1056 | 1036 | 776 |
| New York | 9.4 | 19.4 | 522 | 267 | 1674 | 1392 | 848 |
| Portland | 5.0 | 23.0 | 157 | 144 | 1530 | 1281 | 488 |
| Tucson | 5.1 | 22.9 | 85 | 148 | 1206 | 756 | 483 |
| Washington | 12.5 | 27.6 | 524 | 217 | 1496 | 1003 | 793 |

The distance matrix **D** is given by

| CITY | DISTANCE BETWEEN CITIES | | | | | |
|---|---|---|---|---|---|---|
| Atlanta | 0 | 536.6 | 516.4 | 590.2 | 693.6 | 716.2 |
| Boston | 536.6 | 0 | 447.4 | 833.1 | 915.0 | 881.1 |
| Chicago | 516.4 | 447.4 | 0 | 924.0 | 1073.4 | 971.5 |
| Dallas | 590.2 | 833.1 | 924.0 | 0 | 527.7 | 464.5 |
| Denver | 693.6 | 915.0 | 1073.4 | 527.7 | 0 | 358.7 |
| Detroit | 716.2 | 881.1 | 971.5 | 464.5 | 358.7 | 0 |

The smallest distance is 358.7 between Denver and Detroit, and therefore these two cities are joined at the first step to form $C_1$ = {Denver, Detroit}. In the next step, the distance matrix is calculated for Atlanta, Boston, Chicago, Dallas, and $C_1$:

| | | | | | |
|---|---|---|---|---|---|
| Atlanta | 0 | 536.6 | 516.4 | 590.2 | 693.6 |
| Boston | 536.6 | 0 | 447.4 | 833.1 | 881.1 |
| Chicago | 516.4 | 447.4 | 0 | 924.0 | 971.5 |
| Dallas | 590.2 | 833.1 | 924.0 | 0 | 464.5 |
| $C_1$ | 693.6 | 881.1 | 971.5 | 464.5 | 0 |

Note that all elements of this distance matrix are contained in the original distance matrix. This same pattern will hold in subsequent distance matrices below and is also characteristic of the complete linkage method [see Example 15.3.3(a)]. The smallest distance is 447.4 between Boston and Chicago. Therefore $C_2$ = {Boston, Chicago}. At the next step, the distance matrix is calculated for Atlanta, Dallas, $C_1$, and $C_2$:

| | | | | |
|---|---|---|---|---|
| Atlanta | 0 | 516.4 | 590.2 | 693.6 |
| $C_2$ | 516.4 | 0 | 833.1 | 881.1 |
| Dallas | 590.2 | 833.1 | 0 | 464.5 |
| $C_1$ | 693.6 | 881.1 | 464.5 | 0 |

The smallest distance is 464.5 between Dallas and $C_1$, so that $C_3 = \{\text{Dallas}, C_1\}$. The distance matrix for Atlanta, $C_2$, and $C_3$ is given by

| | | | |
|---|---|---|---|
| Atlanta | 0 | 516.4 | 590.2 |
| $C_2$ | 516.4 | 0 | 833.1 |
| $C_3$ | 590.2 | 833.1 | 0 |

The smallest distance is 516.4, which defines $C_4 = \{\text{Atlanta}, C_2\}$. The distance matrix for $C_3$ and $C_4$ is

| | | |
|---|---|---|
| $C_3$ | 0 | 590.2 |
| $C_4$ | 590.2 | 0 |

The last cluster is given by $C_5 = \{C_3, C_4\}$. The dendrogram for the steps in this example is given in Figure 15.2. The order in which the clusters were formed and the relative distances at which they formed can all be seen. Note that the distance scale runs from right to left. □

■ **EXAMPLE 15.3.2(b)**

To further illustrate the single linkage method of clustering, we use the complete city crime data from Table 15.1. The dendrogram in Figure 15.3 shows the cluster groupings attained by the single linkage method. □

### 15.3.3 Complete Linkage (Farthest Neighbor)

In the *complete linkage* approach, also called the *farthest neighbor* method, the distance between two clusters $A$ and $B$ is defined as the *maximum* distance between a point in $A$ and a point in $B$:

$$D(A, B) = \max \{d(\mathbf{y}_i, \mathbf{y}_j) \text{ for } \mathbf{y}_i \text{ in } A \text{ and } \mathbf{y}_j \text{ in } B\}. \tag{15.9}$$

At each step, the distance (15.9) is found for every pair of clusters, and we merge the two clusters with the smallest distance.

■ **EXAMPLE 15.3.3(a)**

As in Example 15.3.2(a) for single linkage clustering, we illustrate the use of the distance matrix in complete linkage clustering with the first six observations of the city crime data in Table 15.1. The initial distance matrix is exactly the same as in Example 15.3.2(a):

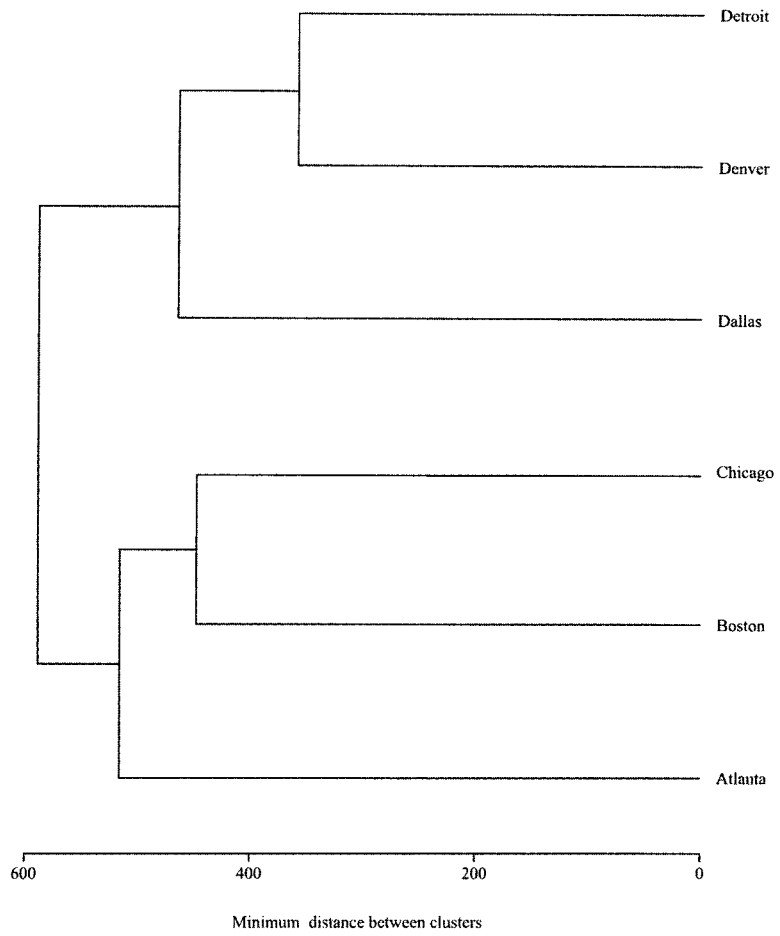


**Figure 15.2** Dendrogram for single linkage of the first six observations in the city crime data in Table 15.1 [See Example 15.3.2(a)].

| CITY | DISTANCE BETWEEN CITIES | | | | | |
|---|---|---|---|---|---|---|
| Atlanta | 0 | 536.6 | 516.4 | 590.2 | 693.6 | 716.2 |
| Boston | 536.6 | 0 | 447.4 | 833.1 | 915.0 | 881.1 |
| Chicago | 516.4 | 447.4 | 0 | 924.0 | 1073.4 | 971.5 |
| Dallas | 590.2 | 833.1 | 924.0 | 0 | 527.7 | 464.5 |
| Denver | 693.6 | 915.0 | 1073.4 | 527.7 | 0 | 358.7 |
| Detroit | 716.2 | 881.1 | 971.5 | 464.5 | 358.7 | 0 |

The smallest distance is 358.7 between Denver and Detroit, and these two therefore form the first cluster, $C_1 = \{\text{Denver, Detroit}\}$. Note that since the first cluster is based on the initial distance matrix, it will be the same regardless of which hierarchical clustering method is used.

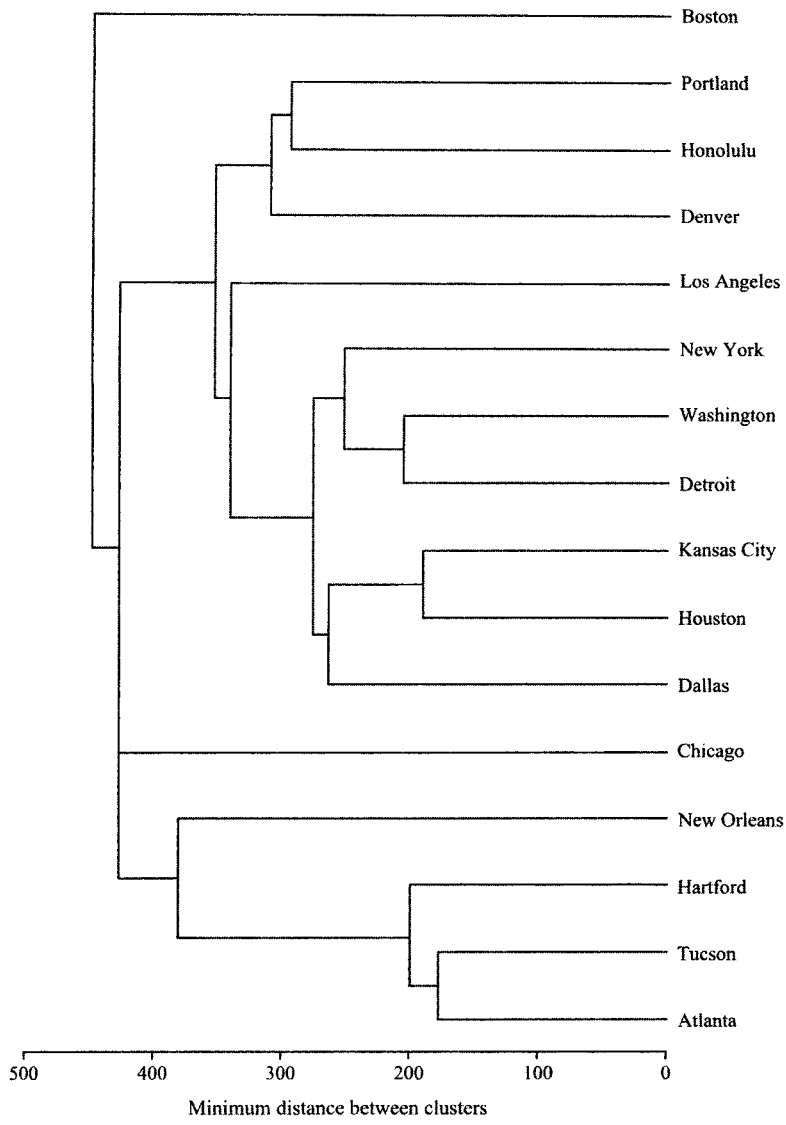


**Figure 15.3** Dendrogram for single linkage of the complete city crime data from Table 15.1 [see Example 15.3.2(b)].

In the next step, the distance matrix is calculated for Atlanta, Boston, Chicago, Dallas, and $C_1$:

| | | | | | |
|---|---|---|---|---|---|
| Atlanta | 0 | 536.6 | 516.4 | 590.2 | 716.2 |
| Boston | 536.6 | 0 | 447.4 | 833.1 | 915.0 |
| Chicago | 516.4 | 447.4 | 0 | 924.0 | 1073.4 |
| Dallas | 590.2 | 833.1 | 924.0 | 0 | 527.7 |
| $C_1$ | 716.2 | 915.0 | 1073.4 | 527.7 | 0 |

Note that this distance matrix differs from its analog for the second step in Example 15.3.2(a) only in the distances between $C_1$ and the other cities. All elements of this matrix and subsequent distance matrices below are contained in the original distance matrix for the six cities. The smallest distance is 447.4 between Boston and Chicago. Therefore, $C_2$ = {Boston, Chicago}. At the next step, distances are calculated for Atlanta, Dallas, $C_1$, and $C_2$:

| | | | | |
|---|---|---|---|---|
| Atlanta | 0 | 536.6 | 590.2 | 716.2 |
| $C_2$ | 536.6 | 0 | 924.0 | 833.1 |
| Dallas | 590.2 | 924.0 | 0 | 527.7 |
| $C_1$ | 693.6 | 881.1 | 527.7 | 0 |

The smallest distance, 527.7, defines $C_3$ = {Dallas, $C_1$}. The distance matrix for Atlanta, $C_2$, and $C_3$ is given by

| | | | |
|---|---|---|---|
| Atlanta | 0 | 536.6 | 716.2 |
| $C_2$ | 536.6 | 0 | 1073.4 |
| $C_3$ | 590.2 | 1073.4 | 0 |

The smallest distance is 536.6 between Atlanta and $C_3$, so that $C_4$ = {Atlanta, $C_2$}. The distance matrix for $C_3$ and $C_4$ is

| | | |
|---|---|---|
| $C_3$ | 0 | 1073.4 |
| $C_4$ | 1073.4 | 0 |

The last cluster is given by $C_5 = \{C_3, C_4\}$. The dendrogram in Figure 15.4 shows the steps in this example. □

■ **EXAMPLE 15.3.3(b)**

To further illustrate the complete linkage method, we use the complete crime data in Table 15.1. The dendrogram in Figure 15.5 shows the clusters found for this data set by the complete linkage approach. There are some differences between these groupings and the groupings from single linkage in Figure 15.3. □

### 15.3.4 Average Linkage

In the *average linkage* approach, the distance between two clusters $A$ and $B$ is defined as the average of the $n_A n_B$ distances between the $n_A$ points in $A$ and the $n_B$

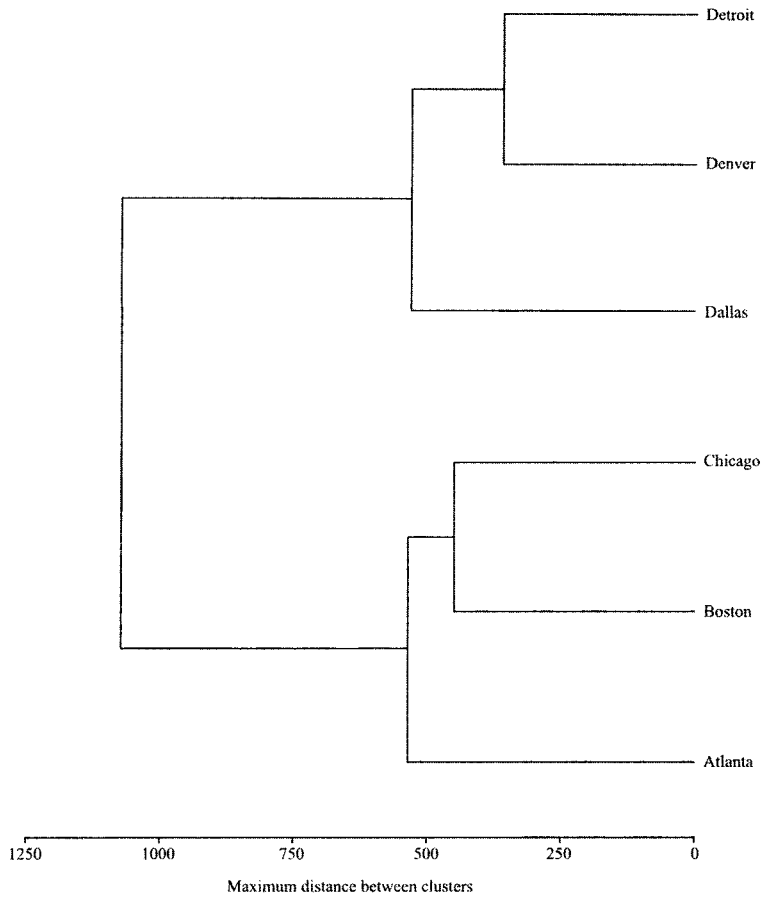


**Figure 15.4** Dendrogram for complete linkage of the first six observations in the city crime data in Table 15.1 [see Example 15.3.3(a)].

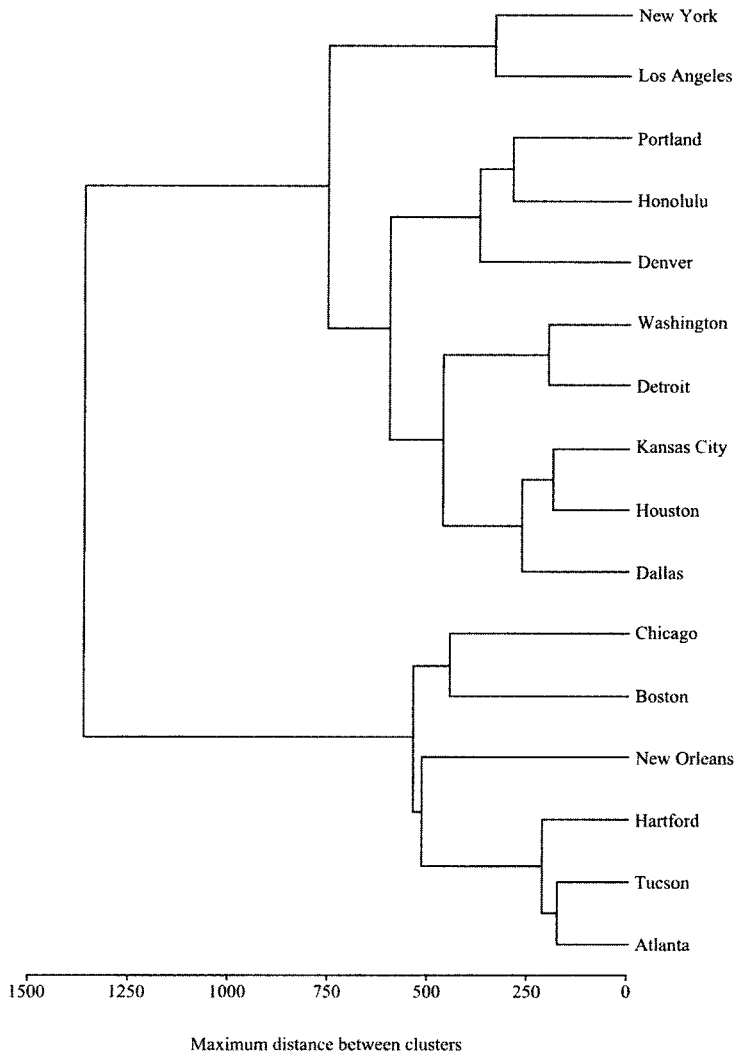


**Figure 15.5** Dendrogram for complete linkage of the complete city crime data of Table 15.1 [see Example 15.3.3(b)].

points in $B$:

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{y}_i, \mathbf{y}_j), \tag{15.10}$$

where the sum is over all $\mathbf{y}_i$ in $A$ and all $\mathbf{y}_j$ in $B$. At each step, we join the two clusters with the smallest distance, as measured by (15.10).

■ **EXAMPLE 15.3.4**

Figure 15.6 shows the dendrogram resulting from the average linkage method applied to the city crime data in Table 15.1. The solution is the same as the complete linkage solution for this data set as given in Example 15.3.3(b) and Figure 15.5. □

### 15.3.5 Centroid

In the *centroid* method, the distance between two clusters $A$ and $B$ is defined as the Euclidean distance between mean vectors (often called centroids) of the two clusters:

$$D(A,B) = d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B), \tag{15.11}$$

where $\bar{\mathbf{y}}_A$ and $\bar{\mathbf{y}}_B$ are the mean vectors for the observation vectors in $A$ and the observation vectors in $B$, respectively, and $d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B)$ is defined in (15.2). We define $\bar{\mathbf{y}}_A$ and $\bar{\mathbf{y}}_B$ in the usual way, that is, $\bar{\mathbf{y}}_A = \sum_{i=1}^{n_A} \mathbf{y}_i / n_A$. The two clusters with the smallest distance between centroids are merged at each step.

After two clusters $A$ and $B$ are joined, the centroid of the new cluster $AB$ is given by the weighted average

$$\bar{\mathbf{y}}_{AB} = \frac{n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B}{n_A + n_B}. \tag{15.12}$$

■ **EXAMPLE 15.3.5**

Figure 15.7 shows the dendrogram resulting from using the centroid clustering method on the complete city crime data in Table 15.1.

Note the two crossovers in the dendrogram in Figure 15.7. Boston and Chicago join at a distance of 447.4. Then that cluster joins with {Atlanta, Tucson, Hartford} at a distance of 441.1. Finally, all five join with New Orleans at a distance of 393.8. Crossovers are discussed in Section 15.3.9a. □

### 15.3.6 Median

If two clusters $A$ and $B$ are combined using the centroid method, and if $A$ contains a larger number of items than $B$, then the new centroid $\bar{\mathbf{y}}_{AB} = (n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B)/(n_A + n_B)$ may be much closer to $\bar{\mathbf{y}}_A$ than to $\bar{\mathbf{y}}_B$. To avoid weighting the mean vectors according to cluster size, we can use the median (midpoint) of the line

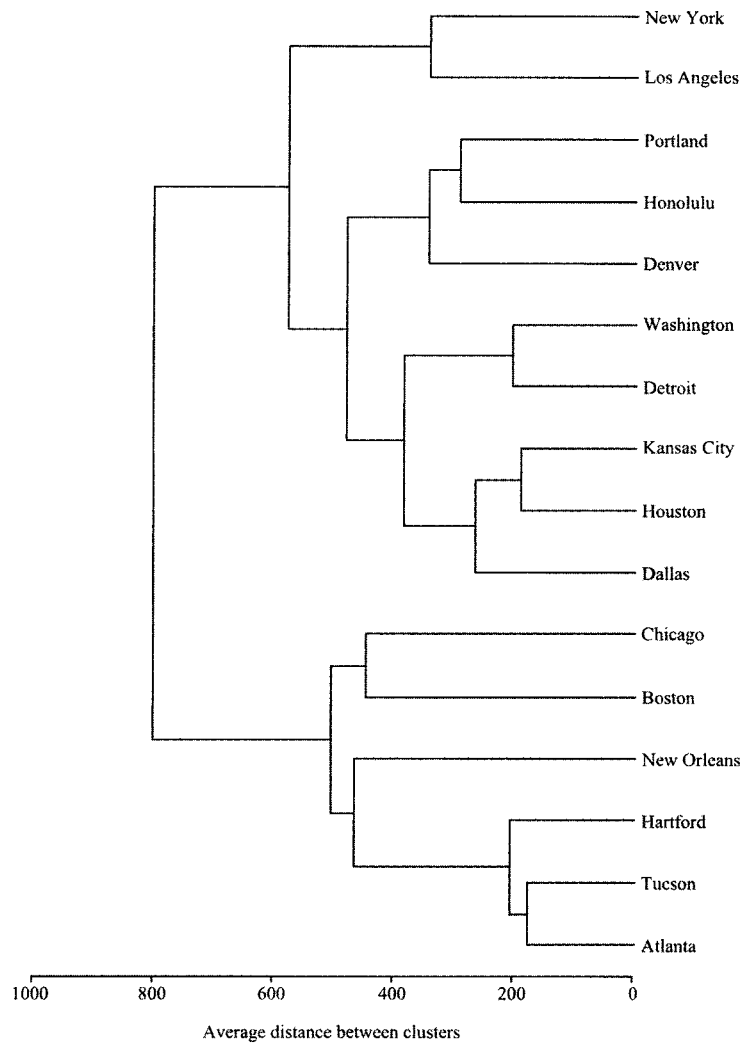


**Figure 15.6** Dendrogram for average linkage clustering of the data in Table 15.1 (see Example 15.3.4).

joining $A$ and $B$ as the point for computing new distances to other clusters:

$$\mathbf{m}_{AB} = \tfrac{1}{2}(\bar{\mathbf{y}}_A + \bar{\mathbf{y}}_B). \tag{15.13}$$

The two clusters with the smallest distance between medians are merged at each step.

Note that the "median" in (15.13) is not the ordinary median in the statistical sense. The terminology arises from a median of a triangle, namely, the line from a vertex to the midpoint of the opposite side.

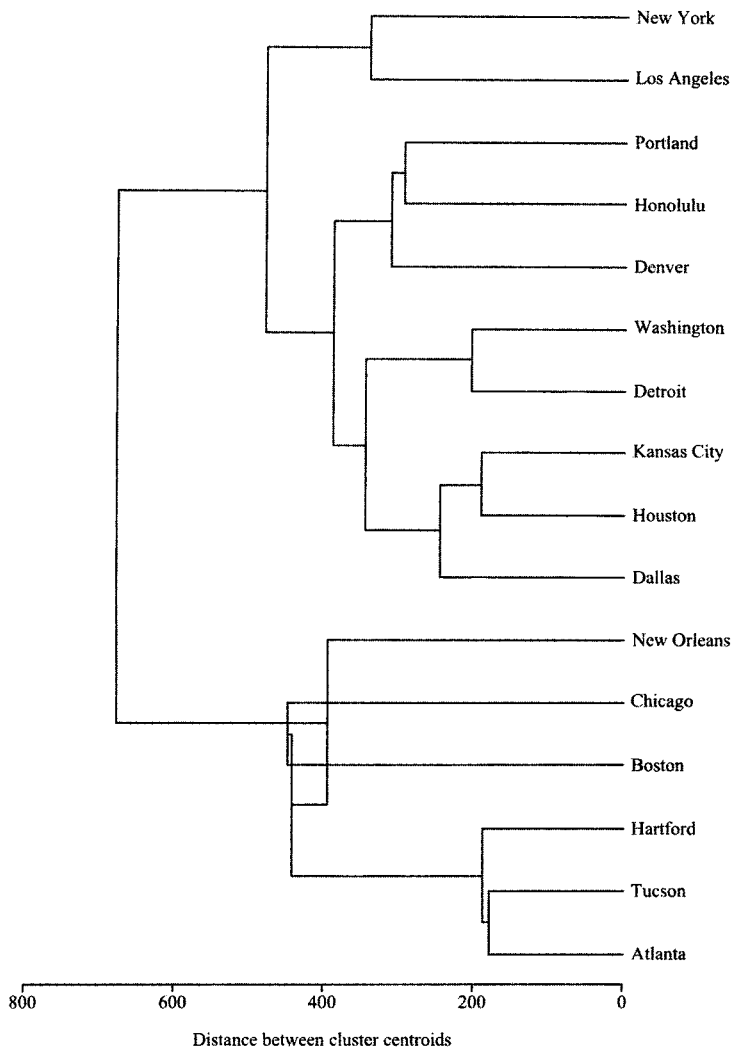


**Figure 15.7** Dendrogram for the centroid clustering of the complete city crime data in Table 15.1 (see Example 15.3.5).

## ■ EXAMPLE 15.3.6

Figure 15.8 shows the dendrogram resulting from using the median distance clustering method on the complete city crime data in Table 15.1. In Figure 15.8, we see the same two crossovers as in Figure 15.7. □

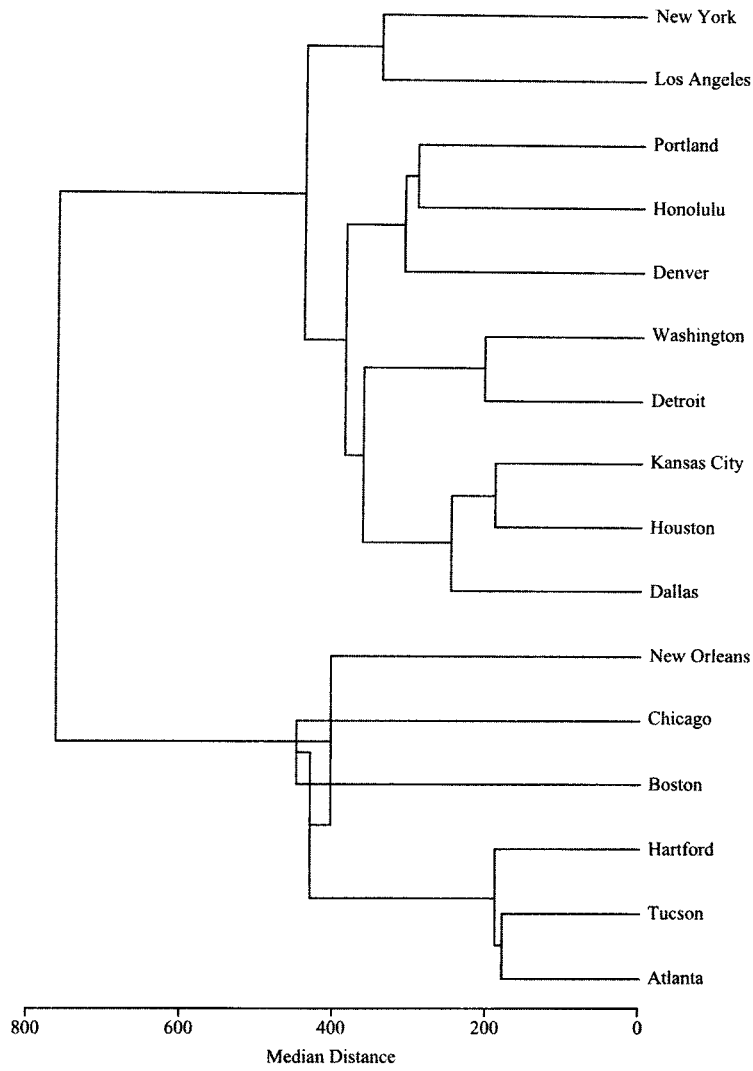


**Figure 15.8** Dendrogram for the median clustering method applied to the complete city crime data in Table 15.1 (see Example 15.3.6).

### 15.3.7 Ward's Method

*Ward's method*, also called the *incremental sum of squares method*, uses the within-cluster (squared) distances and the between-cluster (squared) distances (Ward 1963, Wishart 1969a). If $AB$ is the cluster obtained by combining clusters $A$ and $B$, then

the sums of within-cluster distances (of the items from the cluster mean vectors) are

$$\text{SSE}_A = \sum_{i=1}^{n_A} (\mathbf{y}_i - \bar{\mathbf{y}}_A)'(\mathbf{y}_i - \bar{\mathbf{y}}_A), \tag{15.14}$$

$$\text{SSE}_B = \sum_{i=1}^{n_B} (\mathbf{y}_i - \bar{\mathbf{y}}_B)'(\mathbf{y}_i - \bar{\mathbf{y}}_B), \tag{15.15}$$

$$\text{SSE}_{AB} = \sum_{i=1}^{n_{AB}} (\mathbf{y}_i - \bar{\mathbf{y}}_{AB})'(\mathbf{y}_i - \bar{\mathbf{y}}_{AB}), \tag{15.16}$$

where $\bar{\mathbf{y}}_{AB} = (n_A\bar{\mathbf{y}}_A + n_B\bar{\mathbf{y}}_B)/(n_A + n_B)$ as in (15.12) and $n_A, n_B$, and $n_{AB} = n_A + n_B$ are the numbers of points in $A$, $B$, and $AB$ respectively. Since these sums of distances are equivalent to within-cluster sums of squares, they are denoted by $\text{SSE}_A$, $\text{SSE}_B$, and $\text{SSE}_{AB}$.

Ward's method joins the two clusters $A$ and $B$ that minimize the increase in SSE, defined as

$$I_{AB} = \text{SSE}_{AB} - (\text{SSE}_A + \text{SSE}_B). \tag{15.17}$$

It can be shown that the increase $I_{AB}$ in (15.17) has the following two equivalent forms:

$$\begin{aligned} I_{AB} &= n_A(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB}) \\ &\quad + n_B(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB}) \qquad (15.18) \\ &= \frac{n_A n_B}{n_A + n_B}(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B). \qquad (15.19) \end{aligned}$$

Thus by (15.19), minimizing the increase in SSE is equivalent to minimizing the *between cluster* distances. If $A$ consists only of $\mathbf{y}_i$ and $B$ consists only of $\mathbf{y}_j$, then $\text{SSE}_A$ and $\text{SSE}_B$ are zero, and (15.17) and (15.19) reduce to

$$I_{ij} = \text{SSE}_{AB} = \tfrac{1}{2}(\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j) = \tfrac{1}{2}d^2(\mathbf{y}_i, \mathbf{y}_j).$$

Ward's method is related to the centroid method in Section 15.3.5. If the distance $d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B)$ in (15.11) is squared and compared to (15.19), the only difference is the coefficient $n_A n_B/(n_A + n_B)$ for Ward's method. Thus the cluster sizes have an impact on Ward's method but not on the centroid method. Writing $n_A n_B/(n_A + n_B)$ in (15.19) as

$$\frac{n_A n_B}{n_A + n_B} = \frac{1}{1/n_A + 1/n_B},$$

we see that as $n_A$ and $n_B$ increase, $n_A n_B/(n_A + n_B)$ increases. Writing the coefficient as

$$\frac{n_A n_B}{n_A + n_B} = \frac{n_A}{1 + n_A/n_B},$$

we see that as $n_B$ increases with $n_A$ fixed, $n_A n_B/(n_A + n_B)$ increases. Therefore, compared to the centroid method, Ward's method is more likely to join smaller clusters or clusters of equal size.

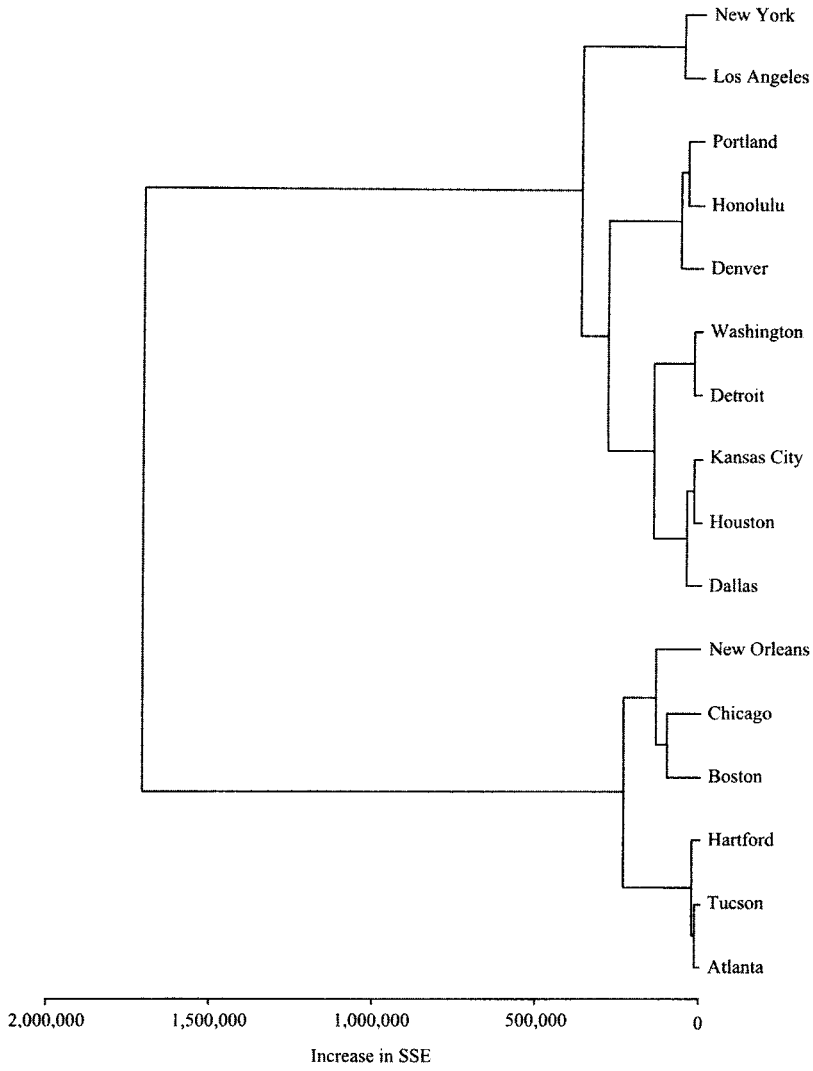


**Figure 15.9** Dendrogram for Ward's method applied to the complete city crime data in Table 15.1 (see Example 15.3.7).

## ■ EXAMPLE 15.3.7

Figure 15.9 shows the dendrogram resulting from using Ward's clustering method on the complete city crime data in Table 15.1. The vertical axis is $I_{AB}/\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})'(\mathbf{y}_i - \bar{\mathbf{y}})$, where $\bar{\mathbf{y}}$ is the overall mean vector for the data. □

### 15.3.8 Flexible Beta Method

Suppose clusters $A$ and $B$ have just been merged to form cluster $AB$. A general formula for the distance between $AB$ and any other cluster $C$ was given by Lance and Williams (1967):

$$D(C, AB) = \alpha_A D(C, A) + \alpha_B D(C, B) + \beta D(A, B) + \gamma |D(C, A) - D(C, B)|. \quad (15.20)$$

The distances $D(C, A)$, $D(C, B)$, and $D(A, B)$ are from the distance matrix before joining $A$ and $B$. The distances from $AB$ to other clusters as given by (15.20) would be used (along with distances between other pairs of clusters) to form the next distance matrix for choosing the pair of clusters with smallest distance. This pair would then be joined at the next step.

To simplify (15.20), Lance and Williams (1967) suggested the following constraints on the parameter values:

$$\begin{gathered} \alpha_A + \alpha_B + \beta = 1 \\ \alpha_A = \alpha_B \\ \gamma = 0 \\ \beta < 1. \end{gathered}$$

With $\alpha_A = \alpha_B$ and $\gamma = 0$, we have $2\alpha_A = 1 - \beta$ or $\alpha_A = \alpha_B = (1 - \beta)/2$, and we need only choose a value of $\beta$. The resulting hierarchical clustering procedure is called the *flexible beta* method.

The choice of $\beta$ determines the characteristics of the flexible beta clustering procedure. Lance and Williams (1967) suggested the use of a small negative value of $\beta$, such as $\beta = -.25$. If there are (or might be) outliers in the data, the use of a smaller value of $\beta$, such as $\beta = -.5$, may be more likely to isolate these outliers into simple clusters.

The distances defined for the agglomerative hierarchical methods in Sections 15.3.2–15.3.7 can all be expressed as special cases of (15.20). The requisite parameter values are given in Table 15.2. For the centroid, median, and Ward's methods, the distances in (15.20) must be squared distances (assuming Euclidean distances). For the other methods in Table 15.2, the distances may be either squared or unsquared.

We illustrate the choice of parameter values in Table 15.2 for the single linkage method. Using $\alpha_A = \alpha_B = \frac{1}{2}, \beta = 0$, and $\gamma = -\frac{1}{2}$ as in the first row of Table 15.2, (15.20) becomes

$$D(C, AB) = \tfrac{1}{2} D(C, A) + \tfrac{1}{2} D(C, B) - \tfrac{1}{2} |D(C, A) - D(C, B)|. \quad (15.21)$$

If $D(C, A) > D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, A) - D(C, B)$, and (15.21) reduces to

$$D(C, AB) = D(C, B). \quad (15.22)$$

On the other hand, if $D(C, A) < D(C, B)$, then $|D(C, A) - D(C, B)| = D(C, B) - D(C, A)$, and (15.21) reduces to

$$D(C, AB) = D(C, A). \quad (15.23)$$

**Table 15.2** Parameter Values for (15.20)

| Cluster Method | $\alpha_A$ | $\alpha_B$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Single linkage | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $-\frac{1}{2}$ |
| Complete linkage | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| Average linkage | $\frac{n_A}{n_A+n_B}$ | $\frac{n_B}{n_A+n_B}$ | 0 | 0 |
| Centroid | $\frac{n_A}{n_A+n_B}$ | $\frac{n_B}{n_A+n_B}$ | $\frac{-n_An_B}{(n_A+n_B)^2}$ | 0 |
| Median | $\frac{1}{2}$ | $\frac{1}{2}$ | $-\frac{1}{4}$ | 0 |
| Ward's method | $\frac{n_A+n_C}{n_A+n_B+n_C}$ | $\frac{n_B+n_C}{n_A+n_B+n_C}$ | $\frac{-n_C}{n_A+n_B+n_C}$ | 0 |
| Flexible beta | $(1-\beta)/2$ | $(1-\beta)/2$ | $\beta(<1)$ | 0 |

Thus, (15.21) can be written as

$$D(C, AB) = \min[D(C, A), D(C, B)], \tag{15.24}$$

which is equivalent to (15.8), the definition of distance for the single linkage method.

■ **EXAMPLE 15.3.8**

Figures 15.10 and 15.11 show dendrograms produced when using the flexible beta clustering method on the complete city crime data in Table 15.1, with $\beta = -.25$ and $\beta = -.75$. The two results are similar. □

### 15.3.9 Properties of Hierarchical Methods

#### *15.3.9a Monotonicity*

If an item or a cluster joins another cluster at a distance that is less than the distance for the previous merger of two clusters, we say that an *inversion* or a *reversal* has occurred. The reversal is represented by a *crossover* in the dendrogram. Examples of crossovers can be found in Figures 15.7 and 15.8.

A hierarchical method in which reversals cannot occur is said to be *monotonic*, because the distance at each step is greater than the distance at the previous step. A distance measure or clustering method that is monotonic is also called *ultrametric*.

We now show that the single linkage and complete linkage methods are monotonic. Let $d_k$ be the distance at which two clusters are joined at the $k$th step. We can describe steps $k$ and $k+1$ in terms of four clusters $A, B, C$, and $D$. Suppose $D(A, B)$ is less than the distance between any other pair among these four clusters,

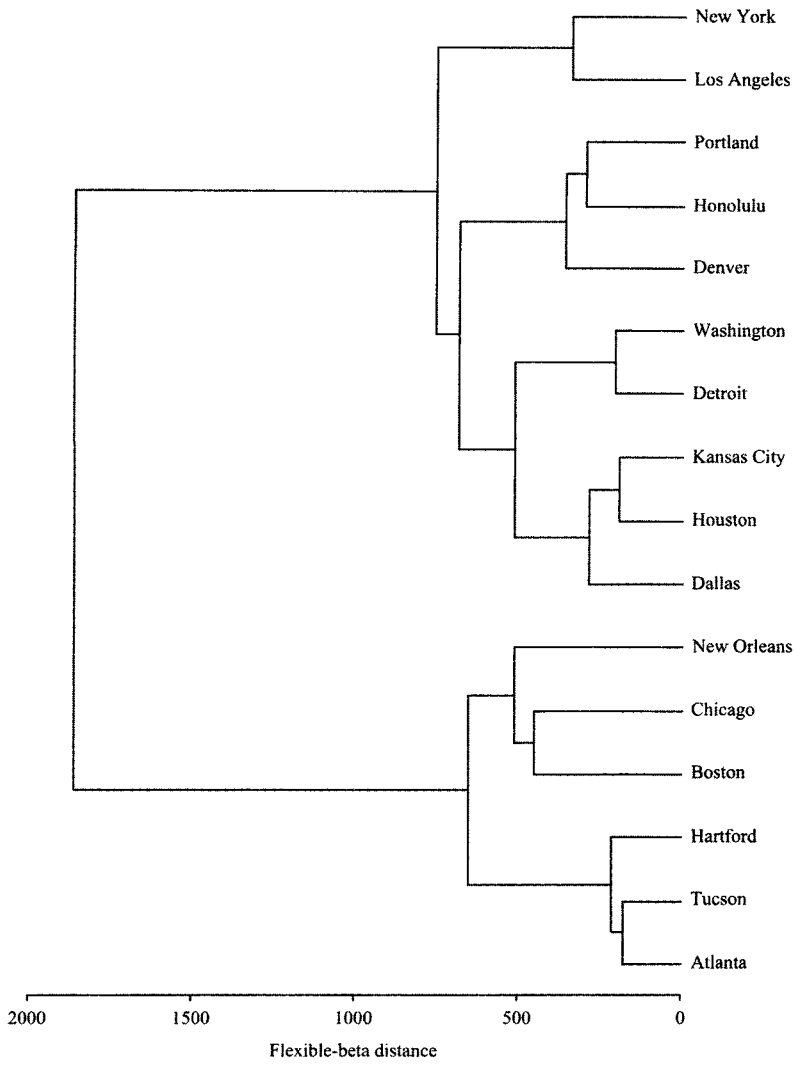


**Figure 15.10** Dendrogram for the flexible beta method with $\beta = -.25$ applied to the complete city crime data in Table 15.1(see Example 15.3.8).

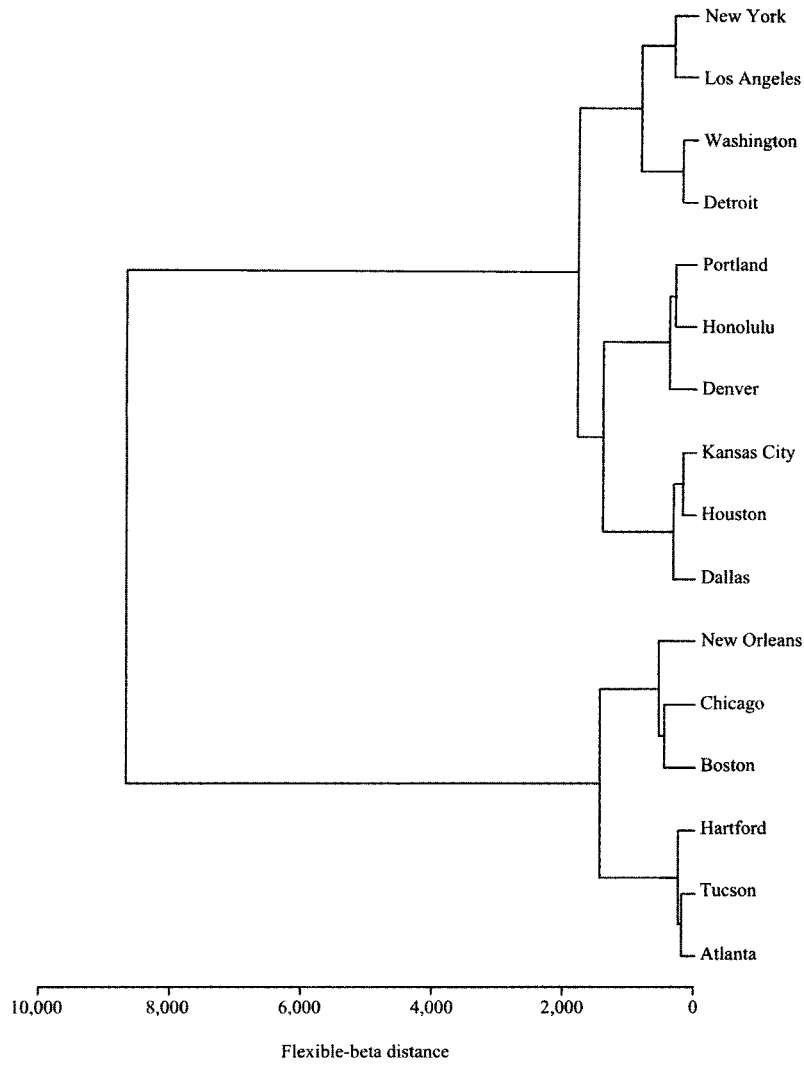


**Figure 15.11** Dendrogram for the flexible beta method with $\beta = -.75$ applied to the complete city crime data in Table 15.1 (see Example 15.3.8).

so that $A$ and $B$ are joined at step $k$ to form $AB$. Then

$$d_k = D(A,B) < \min\{D(A,C), D(B,C), D(C,D)\}. \tag{15.25}$$

[If $D(A,B)$ is less than these three distances, it is less than the other two possible distances, $D(A,D)$ and $D(B,D)$.] Suppose at step $k+1$ we join $AB$ and $C$ or we join $C$ and $D$. If we merge $C$ and $D$, then by (15.25), $d_k = D(A,B) < D(C,D) = d_{k+1}$. If we join $AB$ and $C$, then for single linkage (15.24) gives

$$d_{k+1} = D(C,AB) = \min\{D(A,C), D(B,C)\} > d_k = D(A,B).$$

By (15.25), both of $D(A,C)$ and $D(B,C)$ exceed $D(A,B)$, and this also holds for complete linkage. Thus, the single linkage and complete linkage methods are monotonic.

For the methods in Table 15.2 other than single linkage and complete linkage, we have $\gamma = 0$ and by (15.20) and (15.25),

$$D(C,AB) > (\alpha_A + \alpha_B + \beta)D(A,B). \tag{15.26}$$

Thus we need $\alpha_A + \alpha_B + \beta \geq 1$ for monotonicity. Using this criterion, we see that all methods in Table 15.1 (beyond the first two) are monotonic except the centroid and median methods. (These two methods showed crossovers in the dendrograms in Figures 15.7 and 15.8.) Because of lack of monotonicity, some authors do not recommend the centroid and median methods.

#### *15.3.9b Contraction or Dilation*

We now consider the characteristics of the distances or proximities between the original points. As clusters form, the properties of this space of distances may be altered somewhat. A clustering method that does not alter the spatial properties is referred to by Lance and Williams (1967) as *space-conserving*. A method that is not space-conserving may either *contract* or *dilate* the space.

A method is *space-contracting* if newly formed clusters appear to move closer to individual observations, so that an individual item tends to join an existing cluster rather than join with another individual item to form a new cluster. This tendency is also called *chaining*.

A method is *space-dilating* if newly formed clusters appear to move away from individual observations, so that individual items tend to form new clusters rather than join existing clusters. In this case, clusters appear to be more distinct than they are.

Dubien and Warde (1979) described the spatial properties as follows. Suppose that the distances among three clusters satisfy

$$D(A,B) < D(A,C) < D(B,C).$$

Then a cluster method is space-conserving if

$$D(A,C) < D(AB,C) < D(B,C). \tag{15.27}$$

A method is space-contracting if the first inequality in (15.27) does not hold and space-dilating if the second inequality does not hold.

The single linkage method is very space-contracting, with marked chaining tendencies. For this reason, single linkage is not recommended by some authors. Complete linkage on the other hand, is very space-dilating, with a tendency to artificially impose cluster boundaries.

Other hierarchical methods fall in between the extremes represented by single linkage and complete linkage. The centroid and average linkage methods are largely space-conserving, while Ward's method is space-contracting. Whenever a method produces reversals for a particular data set, it can be considered to be space-contracting. Thus the centroid method is space-conserving unless it has reversals, whereupon it becomes space-contracting.

The flexible beta method is space-contracting for $\beta > 0$, space-conserving for $\beta = 0$, and space-dilating for $\beta < 0$. A small degree of dilation may help define cluster boundaries, but too much dilation may lead to too many clusters in the early stages. Thus the recommended value of $\beta = -.25$ may represent a good compromise.

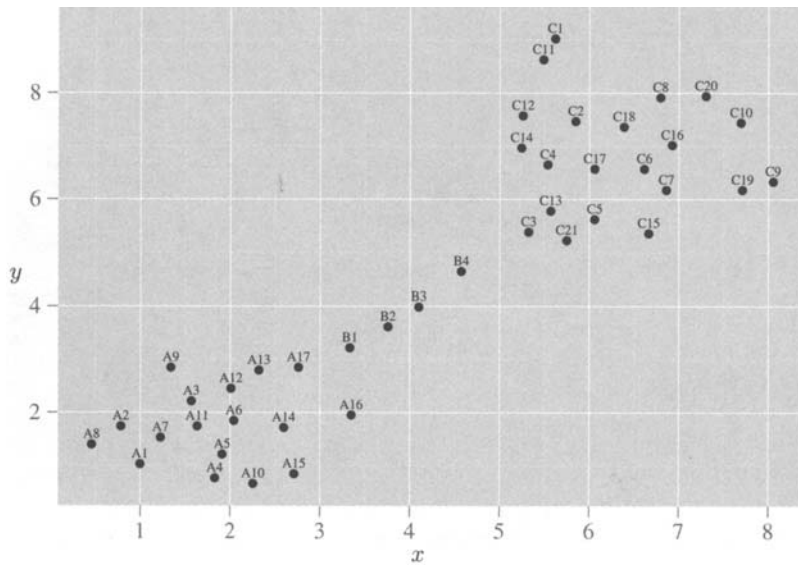


**Figure 15.12** Two distinct clusters with intervening individuals.

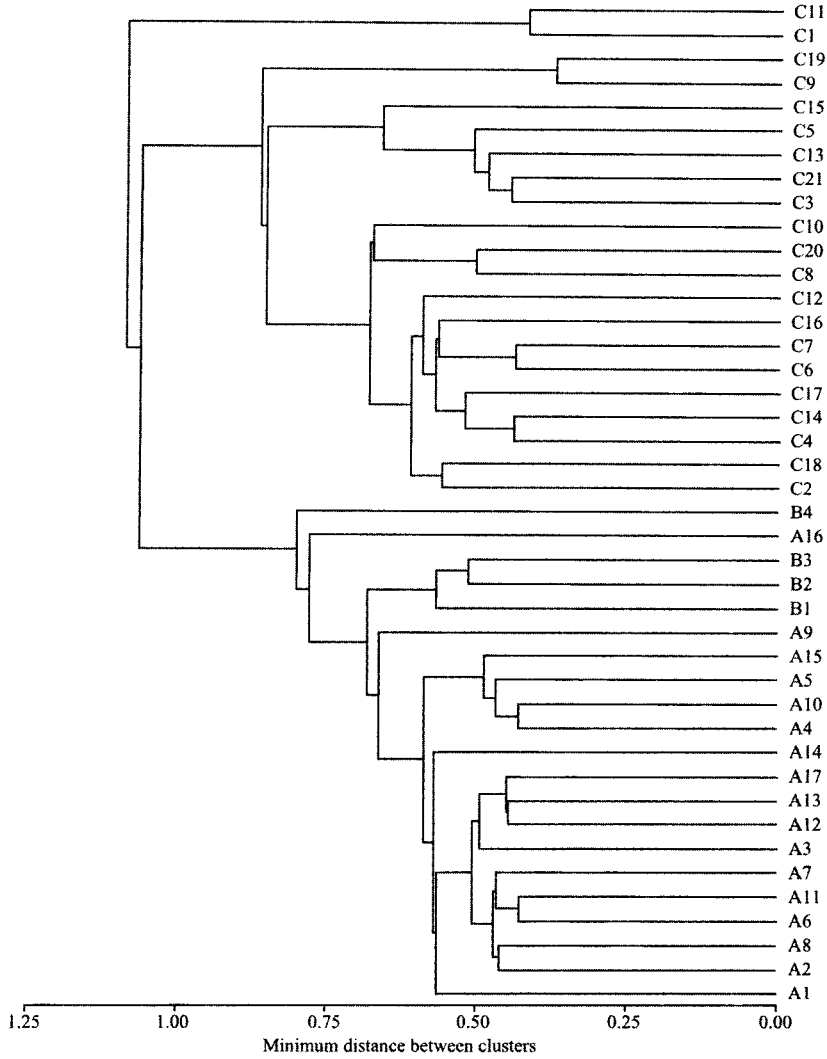


**Figure 15.13** Single linkage clustering of the data in Figure 15.12.

## ■ EXAMPLE 15.3.9b

To illustrate chaining in the single linkage method, consider the data plotted in Figure 15.12 (see Everitt 1993, p. 68). Two distinct clusters, $A$ and $C$, have points between them labeled $B$ that do not belong to $A$ or $C$.

In Figure 15.13, the two-cluster solution for single linkage clustering places $C_1$ and $C_{11}$ into one cluster and all other points into another cluster. The three-cluster solution has two clusters with $C$'s and a cluster with $A$'s and $B$'s.

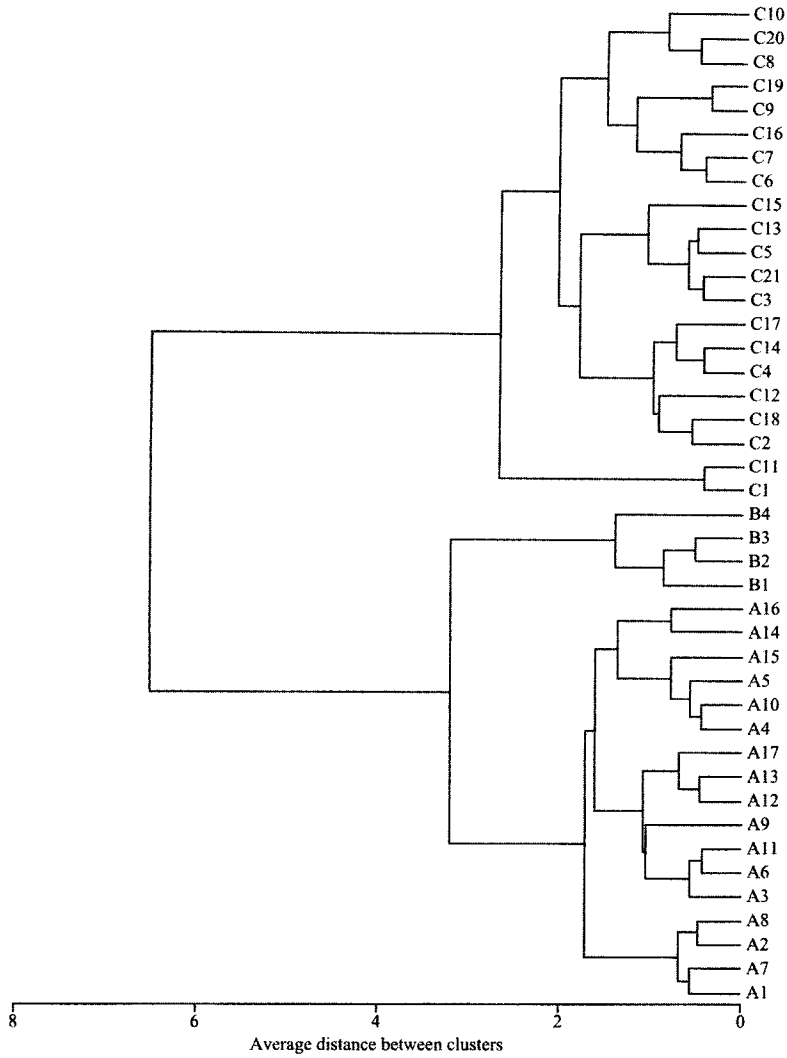


**Figure 15.14** Average linkage clustering of the data in Figure 15.12.

A dendrogram for average linkage clustering of the data in Figure 15.12 is given in Figure 15.14. For this data set, the average linkage method is more robust to chaining. The two-cluster solution separates the $C$'s from the $A$'s and $B$'s. The three-cluster solution completely separates the three groups, $A, B$, and $C$. □

#### *15.3.9c Other Properties*

The single linkage method has been criticized by many authors because of its chaining tendencies and because it is sensitive to errors in distances between observations.

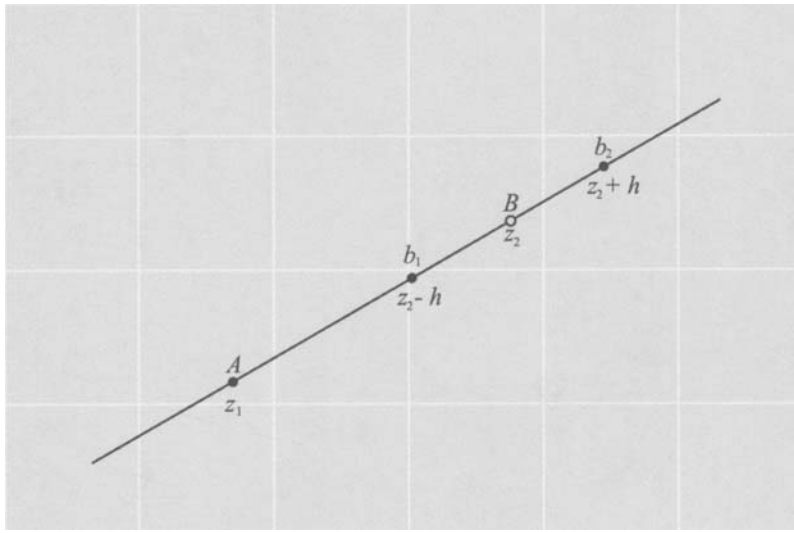


**Figure 15.15** Clusters in a single dimension.

On the other hand, the single linkage approach is better than the other methods at identifying clusters that have curvy shapes instead of spherical or elliptical shapes, and it is somewhat robust to outliers in the data.

Ward's method and the average linkage method are also relatively insensitive to outliers. For example, in the average linkage method, outliers tend to remain isolated in the early stages and to join with other outliers rather than to join with large clusters or with less compact clusters. This is due to two properties of the average linkage method: (1) the average distance between two groups (squared Euclidean distance) increases as the points in the groups are more spread out, and (2) the average distance increases as the size of the groups increases.

These two properties of the average linkage method are illustrated in one dimension in Figure 15.15 (see Jobson 1992, pp. 524–525), where cluster $A$ has one point at $z_1$ and cluster $B$ has two points, $b_1$ and $b_2$, located at $z_2 - h$ and $z_2 + h$. The average squared distance between $A$ and $B$ is

$$\begin{aligned}\overline{d^2} &= \frac{1}{2}[(z_1 - z_2 + h)^2 + (z_1 - z_2 - h)^2] \\ &= \frac{1}{2}[(z_1 - z_2)^2 + h^2 + 2h(z_1 - z_2) + (z_1 - z_2)^2 + h^2 - 2h(z_1 - z_2)] \\ &= (z_1 - z_2)^2 + h^2.\end{aligned}$$

Thus the average distance between $A$ and $B$ increases as the spread of $b_1$ and $b_2$ increases (that is, as $h$ increases).

To illustrate the second property of the average linkage method, suppose cluster $B$ in Figure 15.15 consists of a single point located at $z_2$. Then, the distance between $A$ and $B$ is $(z_1 - z_2)^2$, and $A$ is closer to $B$ than it is if $B$ consists of two points.

The centroid method is fairly robust to outliers. Complete linkage is somewhat sensitive to outliers and tends to produce clusters of the same size and shape. Ward's method tends to yield spherical clusters of the same size.

Many studies conclude that the best overall performers are Ward's method and the average linkage method. However, there seems to be an "interaction" between methods and data sets; that is, some methods work better for certain data sets, and other methods work better for other data sets.

A good strategy is to try several methods. If the results agree to some extent, the researcher may have found some natural clusters in the data.

### 15.3.10 Divisive Methods

In the agglomerative hierarchical methods covered in Sections 15.3.2–15.3.9, we begin with $n$ items and end with a single cluster containing all $n$ items. As noted in the second paragraph of Section 15.3.1, a divisive hierarchical method starts with a single cluster of $n$ items and divides it into two groups. At each step thereafter, one of the groups is divided into two subgroups. The ultimate result of a divisive algorithm is $n$ clusters of one item each. The results can be shown in a dendrogram.

Divisive methods suffer from the same potential drawback as the agglomerative methods, namely, that once a partition is made, an item cannot be moved into another group it does not belong to at the time of the partitioning. However, if larger clusters are of interest, then the divisive approach may sometimes be preferred over the agglomerative approach, in which the larger clusters are reached only after a large number of joinings of smaller groups.

Divisive algorithms are generally of two classes: monothetic and polythetic. In a *monothetic* approach, the division of a group into two subgroups is based on a single variable, whereas the *polythetic* approach uses all $p$ variables to make the split.

If the variables are binary (quantitative variables can be converted to binary variables), the monothetic approach can easily be applied. Division into two groups is based on presence or absence of an attribute. The variable (attribute) is chosen which maximizes a chi-square statistic or an information statistic; see Everitt (1993, pp. 87–88) or Gordon (1999, pp. 130–134).

For a monothetic approach using a quantitative variable $y$, we seek to maximize the between-group sum of squares,

$$\text{SSB} = n_1(\overline{y}_1 - \overline{y})^2 + n_2(\overline{y}_2 - \overline{y})^2,$$

where $n_1$ and $n_2$ are the two group sizes (with $n_1 + n_2 = n$), $\overline{y}_1$ and $\overline{y}_2$ are the group means, and $\overline{y}$ is the overall mean based on all $n$ observations. The sum of squares SSB would be calculated for all possible splits into two groups of sizes $n_1$ and $n_2$ and for each of the $p$ variables. The final division would be based on the variable that maximizes $\text{SSB}/\sum_{i=1}^{n}(y_i - \overline{y})^2$.

**Table 15.3** Athletic Records for Eight Events in Eight Countries

| Country | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Australia | 10.31 | 20.06 | 44.84 | 1.74 | 3.57 | 13.28 | 27.66 | 128.30 |
| Belgium | 10.34 | 20.68 | 45.04 | 1.73 | 3.60 | 13.22 | 27.45 | 129.95 |
| Canada | 10.17 | 20.22 | 45.68 | 1.76 | 3.63 | 13.55 | 28.09 | 130.15 |
| GDR | 10.12 | 20.33 | 44.87 | 1.73 | 3.56 | 13.17 | 27.42 | 129.92 |
| GB | 10.11 | 20.21 | 44.93 | 1.70 | 3.51 | 13.01 | 27.51 | 129.13 |
| Kenya | 10.46 | 20.66 | 44.92 | 1.73 | 3.55 | 13.10 | 27.80 | 129.75 |
| USA | 9.93 | 19.75 | 43.86 | 1.73 | 3.53 | 13.20 | 27.43 | 128.22 |
| USSR | 10.07 | 20.00 | 44.60 | 1.75 | 3.59 | 13.20 | 27.53 | 130.55 |

Event: (1) 100 m (s), (2) 200 m (s), (3) 400 m (s), (4) 800 m (min), (5) 1500 m (min), (6) 5000 m (min), (7) 10,000 m (min), (8) marathon (min).

**Table 15.4** Average Distance from Each Country to the Other Seven

| Country | Average Distance | Country | Average Distance |
|---|---|---|---|
| USA | 2.068 | USSR | 1.513 |
| Aust | 1.643 | Canada | 1.594 |
| GB | 1.164 | Kenya | 1.156 |
| GDR | 1.083 | Belgium | 1.16 |

For a polythetic approach, we consider a technique proposed by MacNaughton-Smith et al. (1964). To divide a group, we work with a splinter group and the remainder. We seek the item in the remainder whose average distance (dissimilarity) from other items in the remainder, minus its average distance from items in the splinter group, is largest. If the largest difference is positive, the item is shifted to the splinter group. If the largest difference is negative, the procedure stops, and the division is complete. We can start the splinter group with the item that has the largest average distance from the other items in the group.

■ **EXAMPLE 15.3.10**

In Table 15.3 we have the track records of eight countries (Dawkins 1989). Based on the distance matrix for these eight observations, the average distance from each observation to the other seven observations is given in Table 15.4. Since USA has the greatest average distance to the other countries, USA becomes the first observation in the splinter group. Now, the average distance between each observation in the remainder to the other six observations in the remainder is calculated. Then the (average) distance between USA and each

**Table 15.5** Average Distances to Remainder and Splinter Groups for Seven Countries

| Country | Average Distance to Remainder (1) | Average Distance to Splinter Group (2) | Difference (1) − (2) |
|---|---|---|---|
| Australia | 1.729 | 1.126 | .603 |
| GB | 1.108 | 1.504 | −.396 |
| GDR | .918 | 2.070 | −1.151 |
| USSR | 1.355 | 2.464 | −1.111 |
| Canada | 1.392 | 2.808 | −1.416 |
| Kenya | .986 | 2.173 | −1.186 |
| Belgium | .975 | 2.329 | −1.353 |

**Table 15.6** Average Distances to Remainder and Splinter Groups for Six Countries

| Country | Average Distance to Remainder (1) | Average Distance to Splinter Group (2) | Difference (1) − (2) |
|---|---|---|---|
| GB | 1.144 | 1.216 | −.072 |
| GDR | .767 | 1.872 | −1.105 |
| USSR | 1.169 | 2.373 | −1.203 |
| Canada | 1.249 | 2.457 | −1.208 |
| Kenya | .865 | 1.884 | −1.019 |
| Belgium | .813 | 2.058 | −1.245 |

item in the remainder is calculated. (This may be found using the distance matrix since there is only one observation in the splinter group.) Finally, we calculate the difference between the average distance to the remainder and the average distance to the splinter group. The results are in Table 15.5. Because Australia has a positive difference in Table 15.5, it is added to the splinter group with USA. This process is repeated for the six countries in the remainder, and the results are given in Table 15.6. Since no difference in Table 15.6 is positive, the process stops, and we have the following clusters: $C_1 = \{\text{USA, Australia}\}$, $C_2 = \{\text{GB, GDR, USSR, Canada, Kenya, Belgium}\}$. We could continue and divide $C_2$ into two groups in the same way. □

## 15.4 NONHIERARCHICAL METHODS

In this section, we discuss three nonhierarchical techniques: partitioning, mixtures of distributions, and density estimation. Among these three methods, partitioning is the most commonly used.

### 15.4.1 Partitioning

In the partitioning approach, the observations are separated into $g$ clusters without using a hierarchical approach based on a matrix of distances or similarities between all pairs of points. The methods described in this section are sometimes called *optimization methods* rather than *partitioning*.

An attractive strategy would be to examine all possible ways to partition $n$ items into $g$ clusters and find the optimal clustering according to some criterion. However, the number of possible partitions as given by (15.7) is prohibitively large for even moderate values of $n$ and $g$. Thus we seek simpler techniques.

#### *15.4.1a k-Means*

We now consider an approach to partitioning that is usually called the *k-means method*. (We will continue to use the notation $g$ rather than $k$ for the number of clusters.) The method allows the items to be moved from one cluster to another, a reallocation that is not available in the hierarchical methods.

We first select $g$ items to serve as "seeds." These are later replaced by the centroids (mean vectors) of the clusters. There are various ways we can choose the seeds: select $g$ items at random (perhaps separated by some minimum distance), choose the first $g$ points in the data set (again subject to a minimum distance apart), select the $g$ points that are mutually farthest apart, find the $g$ points of maximum density, or specify $g$ regularly spaced points in a grid-like pattern (these would not be actual data points).

For these methods of selecting seeds, the number of clusters, $g$, must be specified. Alternatively, a minimum distance between seeds may be specified, and then all items that satisfy this criterion are chosen as seeds.

After the seeds are chosen, each remaining point in the data set is assigned to the cluster with the nearest seed (based on Euclidean distance). As soon as a cluster has more than one member, the cluster seed is replaced by the centroid.

After all items are assigned to clusters, each item is examined to see if it is closer to the centroid of another cluster than to the centroid of its own cluster. If so, the item is moved to the new cluster and the two cluster centroids are updated. This process is continued until no further improvement is possible.

The $k$-means procedure is somewhat sensitive to the initial choice of seeds. It might be advisable to try the procedure again with a new initial choice of seeds. If different initial choices of seeds produce widely different final clusters, or if convergence is extremely slow, there may be no natural clusters in the data.

The $k$-means partitioning method can also be used as a possible improvement on hierarchical techniques. We first cluster the items using a hierarchical method and then use the centroids of these clusters as seeds for a $k$-means approach, which will allow points to be reallocated from one cluster to another.

■ **EXAMPLE 15.4.1a**

Protein consumption in twenty-five European countries for nine food groups is given in Table 15.7 (Hand et al. 1994, p. 298). In order to illustrate the

**Table 15.7** Protein Data

| Country | Red Meat | White Meat | Eggs | Milk | Fish | Cereals | Starchy Foods | Nuts | Fruit/Veg |
|---|---|---|---|---|---|---|---|---|---|
| Albania | 10.1 | 1.4 | .5 | 8.9 | .2 | 42.3 | .6 | 5.5 | 1.7 |
| Austria | 8.9 | 14.0 | 4.3 | 19.9 | 2.1 | 28.0 | 3.6 | 1.3 | 4.3 |
| Belgium | 13.5 | 9.3 | 4.1 | 17.5 | 4.5 | 26.6 | 5.7 | 2.1 | 4.0 |
| Bulgaria | 7.8 | 6.0 | 1.6 | 8.3 | 1.2 | 56.7 | 1.1 | 3.7 | 4.2 |
| Czech. | 9.7 | 11.4 | 2.8 | 12.5 | 2.0 | 34.3 | 5.0 | 1.1 | 4.0 |
| Denmark | 10.6 | 10.8 | 3.7 | 25.0 | 9.9 | 21.9 | 4.8 | .7 | 2.4 |
| E. Germany | 8.4 | 11.6 | 3.7 | 11.1 | 5.4 | 24.6 | 6.5 | .8 | 3.6 |
| Finland | 9.5 | 4.9 | 2.7 | 33.7 | 5.8 | 26.3 | 5.1 | 1.0 | 1.4 |
| France | 18.0 | 9.9 | 3.3 | 19.5 | 5.7 | 28.1 | 4.8 | 2.4 | 6.5 |
| Greece | 10.2 | 3.0 | 2.8 | 17.6 | 5.9 | 41.7 | 2.2 | 7.8 | 6.5 |
| Hungary | 5.3 | 12.4 | 2.9 | 9.7 | .3 | 40.1 | 4.0 | 5.4 | 4.2 |
| Ireland | 13.9 | 10.0 | 4.7 | 25.8 | 2.2 | 24.0 | 6.2 | 1.6 | 2.9 |
| Italy | 9.0 | 5.1 | 2.9 | 13.7 | 3.4 | 36.8 | 2.1 | 4.3 | 6.7 |
| Netherlands | 9.5 | 13.6 | 3.6 | 23.4 | 2.5 | 22.4 | 4.2 | 1.8 | 3.7 |
| Norway | 9.4 | 4.7 | 2.7 | 23.3 | 9.7 | 23.0 | 4.6 | 1.6 | 2.7 |
| Poland | 6.9 | 10.2 | 2.7 | 19.3 | 3.0 | 36.1 | 5.9 | 2.0 | 6.6 |
| Portugal | 6.2 | 3.7 | 1.1 | 4.9 | 14.2 | 27.0 | 5.9 | 4.7 | 7.9 |
| Romania | 6.2 | 6.3 | 1.5 | 11.1 | 1.0 | 49.6 | 3.1 | 5.3 | 2.8 |
| Spain | 7.1 | 3.4 | 3.1 | 8.6 | 7.0 | 29.2 | 5.7 | 5.9 | 7.2 |
| Sweden | 9.9 | 7.8 | 3.5 | 24.7 | 7.5 | 19.5 | 3.7 | 1.4 | 2.0 |
| Switzerland | 13.1 | 10.1 | 3.1 | 23.8 | 2.3 | 25.6 | 2.8 | 2.4 | 4.9 |
| UK | 17.4 | 5.7 | 4.7 | 20.6 | 4.3 | 24.3 | 4.7 | 3.4 | 3.3 |
| USSR | 9.3 | 4.6 | 2.1 | 16.6 | 3.0 | 43.6 | 6.4 | 3.4 | 2.9 |
| W. Germany | 11.4 | 12.5 | 4.1 | 18.8 | 3.4 | 18.6 | 5.2 | 1.5 | 3.8 |
| Yugoslavia | 4.4 | 5.0 | 1.2 | 9.5 | .6 | 55.9 | 3.0 | 5.7 | 3.2 |

sensitivity of the $k$-means clustering method to the initial choice of seeds we use the following four methods of choosing seeds:

1. Select at random $g$ observations that are at least a distance $r$ apart.
2. Select the first $g$ observations that are at least a distance $r$ apart.
3. Select the $g$ observations that are mutually farthest apart.
4. Use the $g$ centroids from the $g$-cluster solution from the average linkage (hierarchical) clustering method.

To help choose $g$, the number of clusters, we plot the first two principal components in Figure 15.16. It appears that there may be at least five clusters. For the first method, we select five observations at random that are at least a distance $r = 1$ from each other. The five chosen seeds are West Germany, Austria, Ireland, France, and Czechoslovakia. Using these seeds, the $k$-means

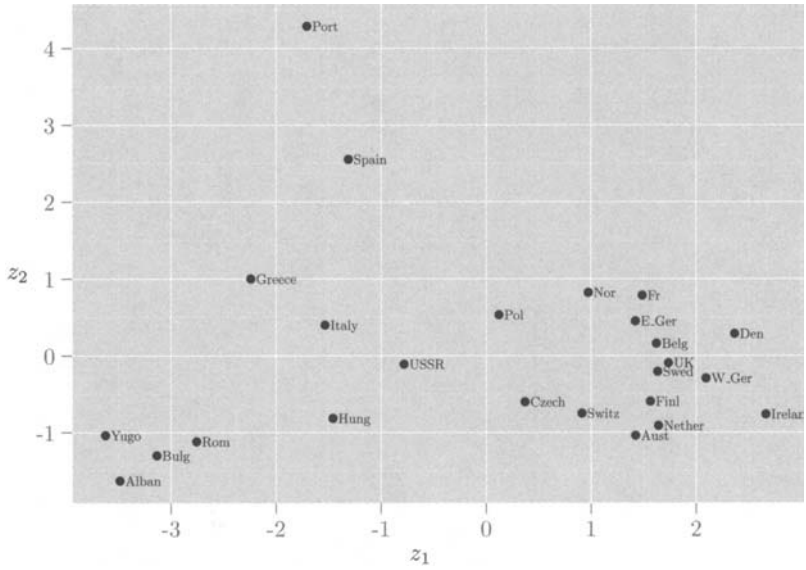


**Figure 15.16** First two principal components $z_1$ and $z_2$ for the protein data in Table 15.7.

method produced the clusters identified in Table 15.8 along with the distance of each observation from its cluster centroid.

To view the clusters, we plot the first two discriminant functions (see Section 8.4.1) in Figure 15.17. The first two discriminant functions show good separation among clusters 3, 4 and 5. However, clusters 1 and 2 appear to overlap with the other clusters.

We now select the first five observations as cluster seeds. With these seeds, the $k$-means clustering method produced the clusters in Table 15.9. The first two discriminant functions are plotted in Figure 15.18. Good separation of clusters is seen except for clusters 2 and 3.

We next choose as cluster seeds the five observations that are mutually farthest apart. These seeds gave rise to the clusters in Table 15.10. The first two discriminant functions are plotted in Figure 15.19. Clusters 1, 3, and 4 seem very well separated, but clusters 2 and 5 show considerable overlap.

Finally, we obtain a five-cluster solution from average linkage and use the centroids of these clusters as the new seeds. The clusters in Table 15.11 result. The first two discriminant functions are plotted in Figure 15.20. All five clusters are well separated in the first two discriminant functions. These clusters show some resemblance to those in the principal components plot given in Figure 15.16. □

**Table 15.8** $k$-Means Cluster Solution for Seeds Chosen at Random

| Country | Cluster | Distance from Centroid | Country | Cluster | Distance from Centroid |
|---|---|---|---|---|---|
| Denmark | 1 | 1.227 | Finland | 3 | 2.261 |
| Sweden | 1 | 1.247 | France | 4 | 2.273 |
| Norway | 1 | 1.629 | Greece | 4 | 2.273 |
| Belgium | 1 | 1.669 | Romania | 5 | 1.514 |
| E. Germany | 1 | 1.987 | Italy | 5 | 1.981 |
| Netherlands | 2 | 0.991 | Yugoslavia | 5 | 2.040 |
| Austria | 2 | 1.160 | Bulgaria | 5 | 2.225 |
| W. Germany | 2 | 1.303 | USSR | 5 | 2.393 |
| Czech. | 2 | 1.433 | Hungary | 5 | 2.435 |
| Switzerland | 2 | 1.679 | Spain | 5 | 2.871 |
| Poland | 2 | 2.052 | Albania | 5 | 3.180 |
| Ireland | 3 | 1.334 | Portugal | 5 | 4.343 |
| UK | 3 | 1.821 | | | |

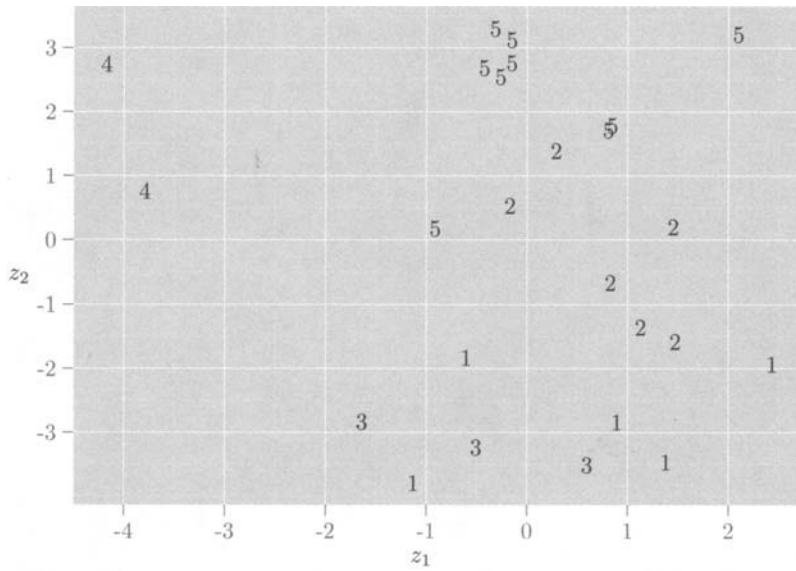


**Figure 15.17** First two discriminant functions $z_1$ and $z_2$ for the clusters in Table 15.8.

**Table 15.9** $k$-Means Cluster Solution Using the First Five Observations as Seeds

| Country | Cluster | Distance from Centroid | Country | Cluster | Distance from Centroid |
|---|---|---|---|---|---|
| Albania | 1 | .000 | Romania | 4 | 1.415 |
| Netherlands | 2 | .648 | Bulgaria | 4 | 1.587 |
| Austria | 2 | 1.000 | Yugoslavia | 4 | 1.784 |
| W. Germany | 2 | 1.087 | Italy | 4 | 1.898 |
| Switzerland | 2 | 1.489 | Greece | 4 | 2.450 |
| Belgium | 3 | 1.368 | Poland | 5 | 1.709 |
| Sweden | 3 | 1.462 | Czech. | 5 | 1.956 |
| Denmark | 3 | 1.666 | USSR | 5 | 2.218 |
| Ireland | 3 | 1.832 | E. Germany | 5 | 2.285 |
| Norway | 3 | 1.927 | Spain | 5 | 2.344 |
| UK | 3 | 2.076 | Hungary | 5 | 2.558 |
| Finland | 3 | 2.341 | Portugal | 5 | 3.859 |
| France | 3 | 2.629 | | | |

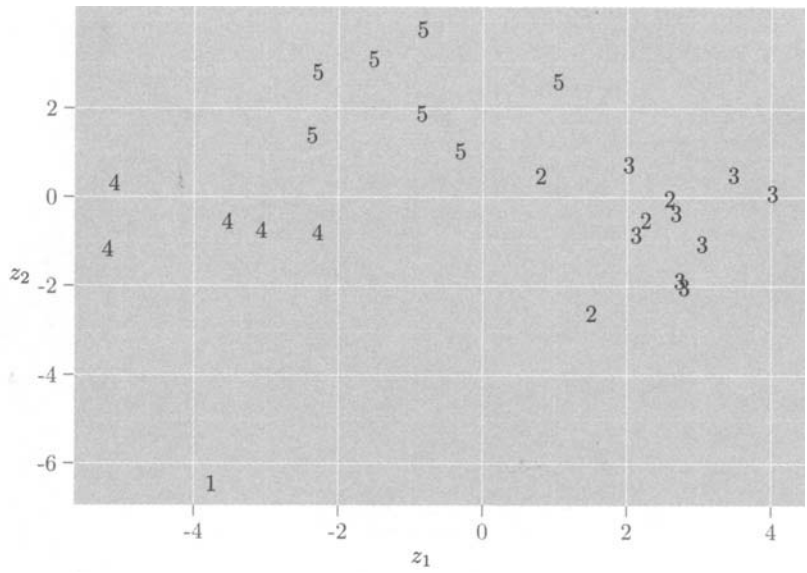


**Figure 15.18** First two discriminant functions $z_1$ and $z_2$ for the clusters in Table 15.9.

**Table 15.10** $k$-Means Cluster Solution Using as Seeds the Five Observations Furthest Apart

| Country | Cluster | Distance from Centroid | Country | Cluster | Distance from Centroid |
|---|---|---|---|---|---|
| Romania | 1 | .601 | France | 2 | 2.358 |
| Yugoslavia | 1 | 1.159 | Poland | 2 | 2.405 |
| Bulgaria | 1 | 1.435 | UK | 2 | 2.537 |
| Albania | 1 | 2.421 | Greece | 3 | 1.075 |
| Hungary | 1 | 2.540 | Italy | 3 | 1.075 |
| Belgium | 2 | .956 | Portugal | 4 | 1.466 |
| W. Germany | 2 | 1.012 | Spain | 4 | 1.466 |
| Netherlands | 2 | 1.416 | Norway | 5 | 1.054 |
| Austria | 2 | 1.663 | Sweden | 5 | 1.191 |
| Czech. | 2 | 1.706 | Finland | 5 | 1.545 |
| Switzerland | 2 | 1.713 | Denmark | 5 | 1.708 |
| Ireland | 2 | 1.839 | USSR | 5 | 2.780 |
| E. Germany | 2 | 2.042 | | | |

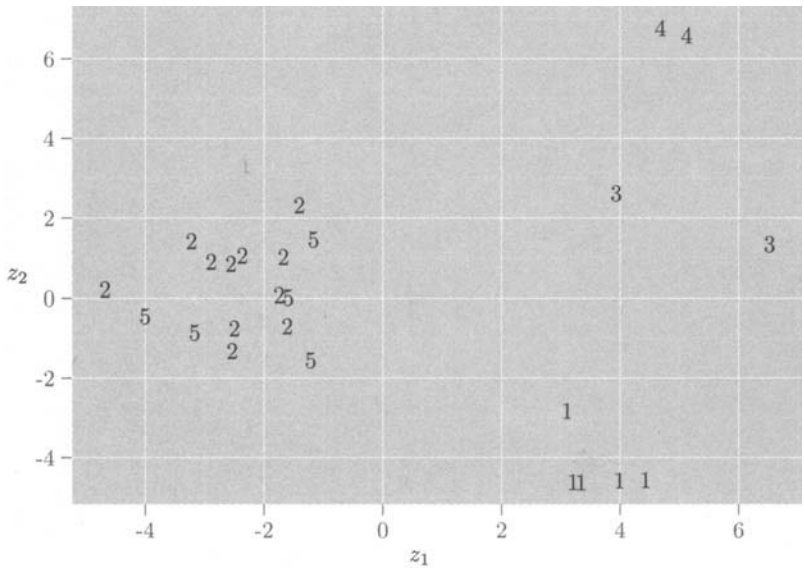


**Figure 15.19** First two discriminant functions $z_1$ and $z_2$ for the clusters in Table 15.10.

**Table 15.11** $k$-Means Cluster Solution Using Seeds from Average Linkage

| Country | Cluster | Distance from Centroid | Country | Cluster | Distance from Centroid |
|---|---|---|---|---|---|
| Romania | 1 | .970 | Norway | 2 | 2.287 |
| Yugoslavia | 1 | 1.182 | UK | 2 | 2.354 |
| Bulgaria | 1 | 1.339 | France | 2 | 2.600 |
| Albania | 1 | 1.970 | Finland | 2 | 2.683 |
| Belgium | 2 | 1.152 | Greece | 3 | 1.075 |
| W. Germany | 2 | 1.245 | Italy | 3 | 1.075 |
| Netherlands | 2 | 1.547 | Portugal | 4 | 1.466 |
| Sweden | 2 | 1.604 | Spain | 4 | 1.466 |
| Ireland | 2 | 1.744 | Czech. | 5 | 1.337 |
| Denmark | 2 | 1.766 | Poland | 5 | 1.579 |
| Switzerland | 2 | 1.831 | USSR | 5 | 1.964 |
| Austria | 2 | 2.037 | Hungary | 5 | 2.023 |
| E. Germany | 2 | 2.251 | | | |

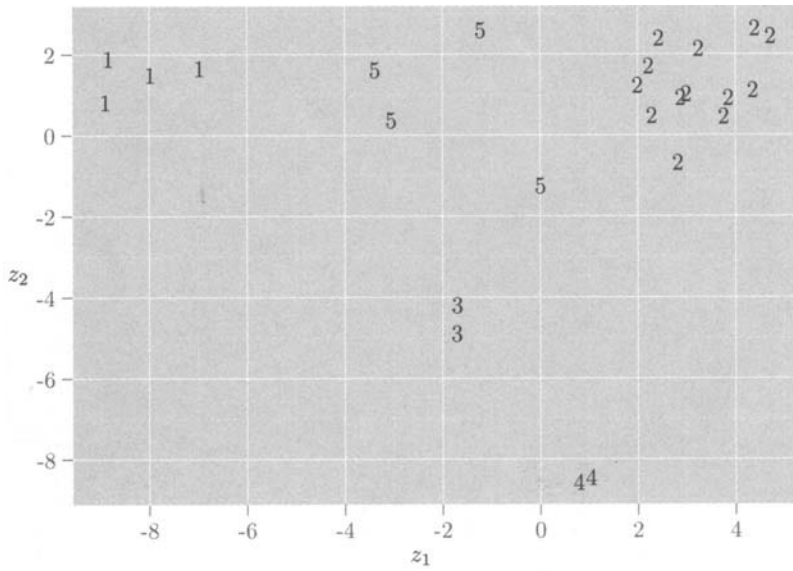


**Figure 15.20** First two discriminant functions $z_1$ and $z_2$ for the clusters in Table 15.11.

### 15.4.1b Other Partitioning Criteria

We now consider three partitioning methods that are not based directly on the distance from a point to the centroid of a cluster. These methods are based on the between-cluster and within-cluster sum of squares and products matrices $\mathbf{H}$ and $\mathbf{E}$ defined in (6.9) and (6.10) for one-way MANOVA. For well-defined clusters, we would like $\mathbf{E}$ to be "small" and $\mathbf{H}$ to be "large."

The three criteria are as follows:

1. Minimize $\text{tr}(\mathbf{E})$
2. Minimize $|\mathbf{E}|$
3. Maximize $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$.

Using criterion 1, for example, we would move an item with observation vector $\mathbf{y}$ to the cluster for which $\text{tr}(\mathbf{E})$ is minimized after the move.

We can express the first criterion in two alternative forms. By (6.10), we have

$$\text{tr}(\mathbf{E}) = \text{tr}\left[\sum_{i=1}^{g}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})'\right] \tag{15.28}$$

$$= \sum_i \text{tr}\left[\sum_j (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})'\right] \quad [\text{by (2.96)}]$$

$$= \sum_i \text{tr}(\mathbf{E}_i), \tag{15.29}$$

where $\mathbf{E}_i = \sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})'$ is the sum of squares and products matrix of deviations of observations from the mean vector for the $i$th cluster. In (15.28) we are using the notation of Section 6.1.2 for a balanced design, in which $n$ is the number of observations in each cluster.

We can write $\text{tr}(\mathbf{E}_i)$ in (15.29) in the form

$$\text{tr}(\mathbf{E}_i) = \text{tr}\sum_j (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})'$$

$$= \sum_j \text{tr}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})' \quad [\text{by (2.96)}]$$

$$= \sum_j (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})'(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) \quad [\text{by (2.97)}]. \tag{15.30}$$

Thus $\text{tr}(\mathbf{E}_i)$ is the sum of the (squared) Euclidean distances from the individual points to the centroid of the $i$th cluster.

A second form of (15.28) was given by Seber (1984, p. 277) as

$$\text{tr}(\mathbf{E}) = \frac{1}{n}\sum_i \sum_{k<m} (\mathbf{y}_{ik} - \mathbf{y}_{im})'(\mathbf{y}_{ik} - \mathbf{y}_{im}). \tag{15.31}$$

Hence minimizing $\text{tr}(\mathbf{E})$ is equivalent to minimizing the sum of squared Euclidean distances between all pairs of points in a cluster.

The second criterion, minimizing $|\mathbf{E}|$, is related to $\Lambda = |\mathbf{E}|/|\mathbf{E} + \mathbf{H}|$ in (6.13). Minimizing $|\mathbf{E}|$ is equivalent to minimizing Wilks' $\Lambda$.

Another way to look at minimizing $|\mathbf{E}|$ is to consider the effect of adding a point $\mathbf{y}$ to a cluster with centroid $\bar{\mathbf{y}}$. Let $\mathbf{u} = \mathbf{y} - \bar{\mathbf{y}}$. By (15.28), $\mathbf{E}$ is a sum of terms of the form $\mathbf{uu}' = (\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})'$. Thus (ignoring the change in centroid with the added observation $\mathbf{y}$), the increase in $|\mathbf{E}|$ is

$$\begin{aligned} |\mathbf{E} + \mathbf{uu}'| - |\mathbf{E}| &= |\mathbf{E}|(1 + \mathbf{u}'\mathbf{E}^{-1}\mathbf{u}) - |\mathbf{E}| \quad \text{[by (2.95)]} \\ &= |\mathbf{E}|\mathbf{u}'\mathbf{E}^{-1}\mathbf{u}. \end{aligned}$$

Hence, the minimum increase in $|\mathbf{E}|$ is obtained by adding $\mathbf{y}$ to the cluster for which the standardized distance $\mathbf{u}'\mathbf{E}^{-1}\mathbf{u}$ of $\mathbf{y}$ from $\bar{\mathbf{y}}$ is the smallest. By comparison, the $\text{tr}(\mathbf{E})$ criterion would add $\mathbf{y}$ to the cluster for which $\mathbf{u}'\mathbf{u}$ is minimum [see (15.30)].

The third criterion, maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$, is related to the Lawley–Hotelling statistic $U^{(s)} = \text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{s} \lambda_i$ in (6.27), where $\lambda_1, \lambda_2, ..., \lambda_s$ are the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ and $s = \min(p, g - 1)$. Associated with each $\lambda_i$ is the eigenvector $\mathbf{a}_i$ and the discriminant function $z_i = \mathbf{a}_i'\mathbf{y}$ (see Section 8.4). The largest eigenvalue, $\lambda_1$, and the accompanying first discriminant function, $z_1 = \mathbf{a}_1'\mathbf{y}$, have the greatest influence on $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$. Maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$ has the inclination to produce elliptical clusters of the same size. These would tend to follow a straight-line trend, especially if the first eigenvalue dominates the others. If the initial clusters or seeds are lined up in a different direction than the "true clusters," maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$ may not correct the error in subsequent iterations.

Since $\text{tr}(\mathbf{E})$ involves only the diagonal elements, the first criterion ignores the correlations and tends to yield spherical clusters. The second criterion, minimizing $|\mathbf{E}|$, takes correlations into account and tends to produce elliptical clusters. These clusters have a tendency to be of the same shape because $\mathbf{E}/\nu_E$ is a pooled estimator of the covariance matrix. A modification that may be useful is $\prod_{i=1}^{g} |\mathbf{E}_i|$, where $\mathbf{E}_i$ is the error matrix for the $i$th cluster [see (15.29)].

Finally, we compare the three criteria in terms of invariance to nonsingular linear transformations $\mathbf{v}_{ij} = \mathbf{A}\mathbf{y}_{ij} + \mathbf{b}$, where $\mathbf{A}$ is a constant nonsingular matrix and $\mathbf{b}$ is a vector of constants. The first criterion, minimizing $\text{tr}(\mathbf{E})$, is not invariant to such linear transformations, while the other two criteria are invariant to these transformations. Therefore, minimizing $\text{tr}(\mathbf{E})$ will likely give different partitions for the raw data and standardized data.

### 15.4.2 Other Methods

We discuss mixtures of distributions in Section 15.4.2a and density estimation in Section 15.4.2b.

#### *15.4.2a Mixtures of Distributions*

In this method, we assume the existence of $g$ distributions (usually multivariate normal), and we wish to assign each of the $n$ items in the sample to the distribution

it most likely belongs to. Such an approach is related to classification analysis in Chapter 9. Along with partitioning in Section 15.4.1, this method has the property that points can be transferred from one cluster to another, but it requires more assumptions than partitioning.

We define the density of a mixture of $g$ distributions as the weighted average

$$h(\mathbf{y}) = \sum_{i=1}^{g} \alpha_i f(\mathbf{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{15.32}$$

where $0 \leq \alpha_i \leq 1, \sum_{i=1}^{g} \alpha_i = 1$, and $f(\mathbf{y}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the multivariate normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ given in (4.2).

Clusters could be formed in two ways. The first approach is to assign an item with observation vector $\mathbf{y}$ to the cluster $C_i$ with largest value of the estimated posterior probability

$$\hat{P}(C_i|\mathbf{y}) = \frac{\hat{\alpha}_i f(\mathbf{y}, \hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)}{h(\mathbf{y})} \tag{15.33}$$

[see Rencher (1998, Sections 6.2.4 and 6.3.1)], where $\hat{\alpha}_i, \hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\Sigma}}_i$ are maximum likelihood estimates and $h(\mathbf{y})$ is given by (15.32) with estimates inserted for parameters. The posterior probability (15.33) is an estimate of the probability that an item with observation vector $\mathbf{y}$ belongs to the $i$th cluster, $C_i$.

The second approach is to assign an item with observation vector $\mathbf{y}$ to the cluster with largest value of

$$\ln\hat{\alpha}_i - \tfrac{1}{2}\ln|\hat{\boldsymbol{\Sigma}}_i| - \tfrac{1}{2}(\mathbf{y} - \hat{\boldsymbol{\mu}}_i)'\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}_i) \tag{15.34}$$

[see (9.15)]. For either of these approaches [based on (15.33) or (15.34)], we need the estimates $\hat{\alpha}_i, \hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\Sigma}}_i$. These estimates are obtained by maximizing the likelihood function $L = \prod_{j=1}^{n} h(\mathbf{y}_j)$, where $h(\mathbf{y}_j)$ is given by (15.32). The results are

$$\hat{\alpha}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{P}(C_i|\mathbf{y}_j), \qquad i = 1, 2, ..., g-1$$

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n\hat{\alpha}_i}\sum_{j=1}^{n} \mathbf{y}_j \hat{P}(C_i|\mathbf{y}_j), \qquad i = 1, 2, ..., g$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n\hat{\alpha}_i}\sum_{j=1}^{n} (\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_i)'\hat{P}(C_i|\mathbf{y}_j), \qquad i = 1, 2, ..., g$$

(Everitt 1993, p. 111), where $\hat{P}(C_i|\mathbf{y}_j)$ is given by (15.33). These three equations must be solved iteratively. For a given value of $g$, we can begin with initial estimates or guesses for the parameters and adjust them by iteration (this approach is related to the EM algorithm mentioned in Section 3.11). If $g$ is not known, we can begin with $g = 1$, then successively try $g = 2, g = 3$, and so on, until the results are satisfactory.

The total number of parameters to be estimated is large. There are $p$ parameters in each $\boldsymbol{\mu}_i$, $\frac{1}{2}p(p+1)$ unique parameters in each $\boldsymbol{\Sigma}_i$, and $g - 1$ values of $\alpha_i$ (the

remaining $\hat{\alpha}_i$ is found by $\sum_{i=1}^{g} \hat{\alpha}_i = 1$), for a total of

$$\tfrac{1}{2}g(p+1)(p+2) - 1 \tag{15.35}$$

parameters. If the sample size $n$ is not sufficiently large to estimate all of these parameters, we could assume a common covariance matrix $\boldsymbol{\Sigma}$, which reduces the number of parameters by $\frac{1}{2}(g-1)p(p+1)$.

The method of mixtures is invariant to full-rank linear transformations and is somewhat robust to the assumption of normality. The technique works better if the $g$ densities are well separated or the sample sizes are large.

■ **EXAMPLE 15.4.2a**

To illustrate the clustering method based on mixtures of distributions, we use the protein consumption data of Table 15.7. Because of the small number of countries in the data set, there are not enough degrees of freedom to estimate a different covariance matrix for each cluster. Hence we assume equal covariance matrices and estimate a pooled covariance matrix $\hat{\boldsymbol{\Sigma}}$. For illustration purposes, we choose $g = 5$ as in Example 15.4.1a.

We use the five clusters found by Ward's method to obtain initial estimates of $\alpha_i, \boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}$. Then the maximum likelihood equations are solved iteratively to find the following estimates.

$$\hat{\alpha}_1 = 0.2801, \hat{\alpha}_2 = 0.3200, \hat{\alpha}_3 = 0.1199, \hat{\alpha}_4 = 0.1600, \hat{\alpha}_5 = 0.1200$$

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 8.64 \\ 6.87 \\ 2.39 \\ 14.04 \\ 2.54 \\ 39.27 \\ 3.74 \\ 4.21 \\ 4.66 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 13.21 \\ 10.64 \\ 3.99 \\ 21.16 \\ 3.38 \\ 24.70 \\ 4.65 \\ 2.06 \\ 4.18 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix} 6.13 \\ 5.77 \\ 1.43 \\ 9.63 \\ .93 \\ 54.07 \\ 2.40 \\ 4.90 \\ 3.40 \end{pmatrix},$$

$$\hat{\boldsymbol{\mu}}_4 = \begin{pmatrix} 9.85 \\ 7.05 \\ 3.15 \\ 26.68 \\ 8.22 \\ 22.68 \\ 4.55 \\ 1.18 \\ 2.12 \end{pmatrix}, \quad \hat{\boldsymbol{\mu}}_5 = \begin{pmatrix} 7.23 \\ 6.23 \\ 2.63 \\ 8.20 \\ 8.87 \\ 26.93 \\ 6.03 \\ 3.80 \\ 6.23 \end{pmatrix}$$

$$\hat{\Sigma} = \begin{pmatrix} 4.250 & -2.952 & .021 & -.047 & 1.001 & .929 & -.157 & .287 & .035 \\ -2.952 & 9.411 & .963 & .265 & -1.934 & -4.250 & 1.245 & -2.903 & -.319 \\ .021 & .963 & .471 & .552 & -.296 & -.699 & .301 & -.256 & -.008 \\ -.047 & .265 & .552 & 9.706 & -1.254 & .011 & 1.313 & -.801 & .032 \\ 1.001 & -1.934 & -.296 & -1.254 & 3.648 & .167 & .111 & .839 & 1.653 \\ .929 & -4.250 & -.699 & -.011 & .167 & 8.412 & -.777 & 1.708 & .137 \\ -.157 & 1.245 & .301 & 1.313 & .111 & -.777 & 1.634 & -.845 & -.208 \\ .287 & -2.903 & -.256 & -.801 & .839 & 1.708 & -.845 & 2.053 & .503 \\ .035 & -.319 & -.008 & .032 & 1.653 & .137 & -.208 & .503 & 1.808 \end{pmatrix}$$

Then assigning each country to the cluster for which it has the highest posterior probability of membership as in (15.33) yields the following clusters:

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Albania, Czech., Greece, Hungary, Italy, Poland, USSR | Austria, Belgium, France, Ireland, Netherlands, Switzerland, UK, W. Germany | Bulgaria, Romania, Yugoslavia | Denmark, Finland Norway, Sweden | E. Germany, Portugal Spain |

□

### *15.4.2b Density Estimation*

In the method of *density estimation* or *density searching*, we seek regions of high density sometimes called *modes*. No assumption is made about the form of the density, as was done in Section 15.4.2a. We could estimate the density using a kernel function as in Section 9.7.2. Alternatively, we simply attempt to separate regions with a high concentration of points from regions with a low density.

To find regions of high density, we first choose a radius $r$ and a value of $k$, the number of points in a $k$-nearest neighbor scheme. For each of the $n$ points in the data, the number of points within a sphere of radius $r$ is found. A point is called a *dense point* if at least $k$ other points are contained in its sphere.

If a dense point is more than a distance $r$ from all other dense points, it becomes the nucleus of a new cluster. If a dense point is within a distance $r$ from at least one dense point that belongs to a cluster, it is added to the cluster. If the dense point is within a distance $r$ of two or more clusters, these clusters are combined. Two clusters are also combined if the smallest distance between their dense points is less than the average of the $2k$ smallest distances between the original $n$ points. The value of $r$ can be gradually increased so that more points become dense. Another option is to begin with the specified value of $r$ for each point and then gradually increase $r$ until $k$ observations are contained in its sphere.

■ **EXAMPLE 15.4.2b**

To illustrate the density estimation method, we use the protein data. For each pair of values of $k$ and $r$, the value of $r$ was allowed to increase if needed, as described above. For the following values of $k$ and $r$, the number of clusters obtained are given.

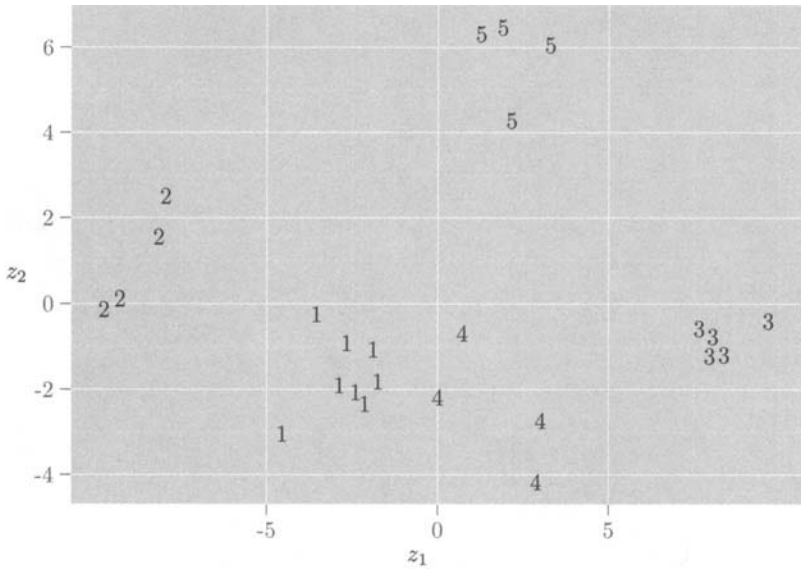


**Figure 15.21** First two discriminant functions for the clusters found in Example 15.4.2b.

| $k/r$ | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 |
| 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 |

The five-cluster solution found by setting $r = 1.8$ and $k = 2$ is

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Austria, Belgium | Denmark | Albania, | Czech., | Greece, |
| France, Ireland, | Finland, | Bulgaria, | E. Germany, | Italy, |
| Netherlands, | Norway, | Hungary, | Poland, | Portugal, |
| Switzerland, | Sweden | Romania, | USSR | Spain |
| UK, W. Germany | | Yugoslavia | | |

This partitioning into five clusters is perhaps more reasonable than any of those found in Example 15.4.2a. The first two discriminant functions for these five clusters are plotted in Figure 15.21. □

## 15.5 CHOOSING THE NUMBER OF CLUSTERS

In hierarchical clustering, we can select $g$ clusters from the dendrogram by cutting across the branches at a given level of the distance measure used by one of the axes.

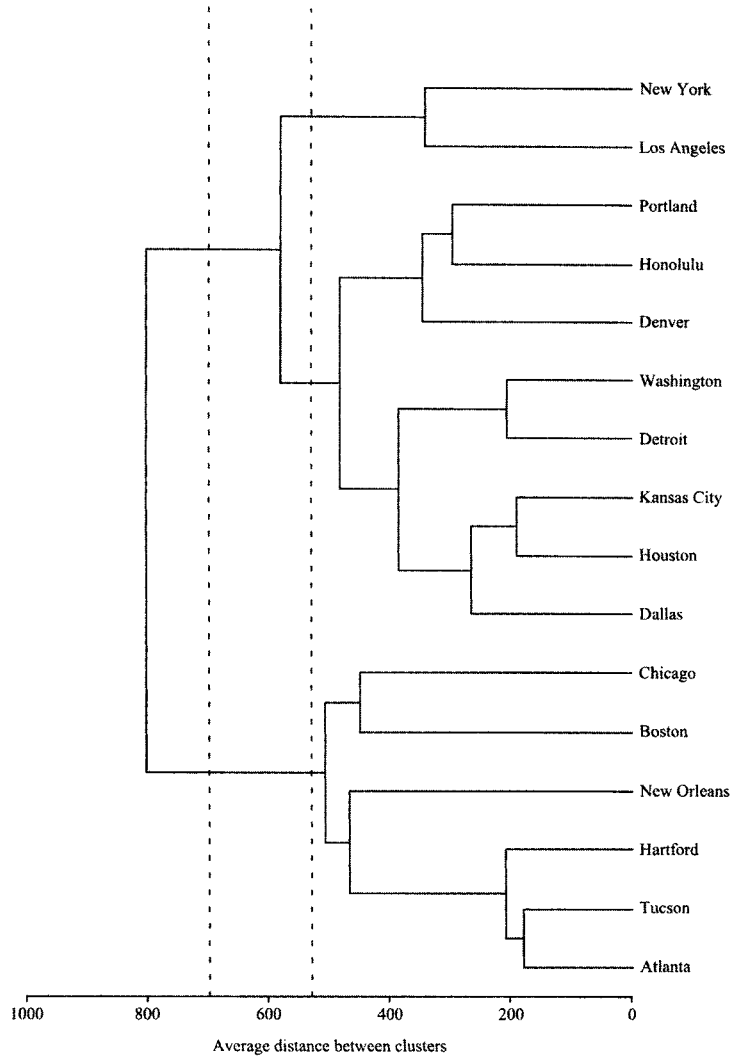


**Figure 15.22** Cutting the dendrogram to choose the number of clusters.

This is illustrated in Figure 15.22, which is the dendrogram for the average linkage method (Section 15.3.4) applied to the city crime data in Table 15.1. Cutting the dendrogram at a level of 700 yields two clusters. Cutting it at 535 gives three clusters.

We wish to determine the value of $g$ that provides the best fit to the data. One approach is to look for large changes in distances at which clusters are formed. For example, in Figure 15.22, the largest change in levels occurs in going from two clusters to a single cluster. The change in distance between the two-cluster solution and the three-cluster solution is 82 units-squared. The difference between the three-

cluster solution and the four-cluster solution is 73 units-squared, and the change between the four- and five-cluster solutions is only 26 units-squared. In this case we would choose two clusters.

A formalization of this procedure was proposed by Mojena (1977): Choose the number of groups given by the first stage in the dendrogram at which

$$\alpha_j > \overline{\alpha} + ks_\alpha, \qquad j = 1, 2, ..., n, \tag{15.36}$$

where $\alpha_1, \alpha_2, ..., \alpha_n$ are the distance values for stages with $n, n-1, ..., 1$ clusters, $\overline{\alpha}$ and $s_\alpha$ are the mean and standard deviation of the $\alpha$'s, and $k$ is a constant. Mojena (1977) suggested using a value of $k$ in the range 2.75 to 3.5, but Milligan and Cooper (1985) recommended $k = 1.25$, based on a simulation study.

An index that can be used with either hierarchical or partitioning methods is

$$c = \frac{\text{tr}(\mathbf{H})/(g-1)}{\text{tr}(\mathbf{E})/(n-g)}. \tag{15.37}$$

The value of $g$ that maximizes $c$ is chosen. A related approach is to choose the value of $g$ that minimizes

$$d = g^2|\mathbf{E}|. \tag{15.38}$$

To compare two cluster solutions with $g_1$ and $g_2$ clusters where $g_2 > g_1$, we can use the test statistic

$$F = \frac{\text{tr}(\mathbf{E}_1) - \text{tr}(\mathbf{E}_2)}{\text{tr}(\mathbf{E}_2)\left[\left(\frac{n-g_1}{n-g_2}\right)\left(\frac{g_2}{g_1}\right)^{2/p} - 1\right]}, \tag{15.39}$$

which has an approximate $F$-distribution with $p(g_2 - g_1)$ and $p(n - g_2)$ degrees of freedom [Beale (1969)]. The matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ are within-cluster sums of squares and products matrices corresponding to $g_1$ and $g_2$. The hypothesis is that the cluster solutions with $g_1$ and $g_2$ clusters are equally valid, and rejection implies that the cluster solution with $g_2$ clusters is better than the solution with $g_1$ clusters ($g_2 > g_1$). The $F$-approximation in (15.39) may not be sufficiently accurate to justify the use of $p$-values.

## 15.6 CLUSTER VALIDITY

To check the validity of a cluster solution, it may be possible to test the hypothesis that there are no clusters or groups in the population from which the sample at hand was taken. For example, the hypothesis could be that the population represents a single unimodal distribution such as the multivariate normal, or that the observations arose from a uniform distribution. Formal tests of hypotheses of this type concerning cluster validity are reviewed by Gordon (1999, Section 7.2).

A cross-validation approach can also be used to check the validity or stability of a clustering result. The data are randomly divided into two subsets, say $A$ and $B$, and

the cluster analysis is carried out separately on each of $A$ and $B$. The results should be similar if the clusters are valid. An alternative approach is the following (Gordon 1999, Section 7.1; Milligan 1996):

1. Use some clustering method to partition subset $A$ into $g$ clusters.
2. Partition subset $B$ into $g$ clusters in two ways:
   (a) Assign each item in $B$ to the cluster in $A$ that it is closest to by using, for example, the distance to cluster centroids.
   (b) Use the same clustering method on $B$ that was used on $A$.
3. Compare the results of (a) and (b) in step 2.

## 15.7 CLUSTERING VARIABLES

In some cases, it may be of interest to cluster the $p$ variables rather than the $n$ observations. For a similarity measure between each pair of variables, we would usually use the correlation. Since most clustering methods use dissimilarities (such as distances), we need to convert the correlation matrix $\mathbf{R} = (r_{ij})$ to a dissimilarity matrix. This can conveniently be done by replacing each $r_{ij}$ by $1 - |r_{ij}|$ or $1 - r_{ij}^2$. Using the resulting dissimilarity matrix, we can apply a clustering method such as a hierarchical technique to cluster the variables.

Clustering of variables can sometimes be done successfully with factor analysis, which groups the variables corresponding to each factor; see Sections 13.1 and 13.5.

### ■ EXAMPLE 15.7

We illustrate clustering of variables using the city crime data in Table 15.1. We first calculate the correlation matrix $\mathbf{R} = (r_{ij})$ and then transform $\mathbf{R}$ to a dissimilarity matrix $\mathbf{D} = (1 - r_{ij}^2)$. The variables are then clustered using both average linkage and Ward's clustering methods, and the dendrograms are given in Figures 15.23 and 15.24, respectively. Both clustering methods yield the same solution.

We next carry out a factor analysis of the data and compare the resulting groups of variables with the clusters obtained with the average linkage and Ward's methods. The factor loadings are estimated using the principal factor method (Section 13.3.2) with squared multiple correlations as initial communality estimates, and the loadings are then rotated with a varimax rotation (Section 13.5.2b). The rotated factor pattern is given in Table 15.12. The highest loading in each row is bolded. The first factor deals with crimes associated with the home. The second factor involves crimes that are violent in nature. The third factor consists of crimes of theft outside the home. Note that the three-cluster solutions found by both average linkage and Ward's methods are identical to the grouping of variables in the factor analysis solution, which is

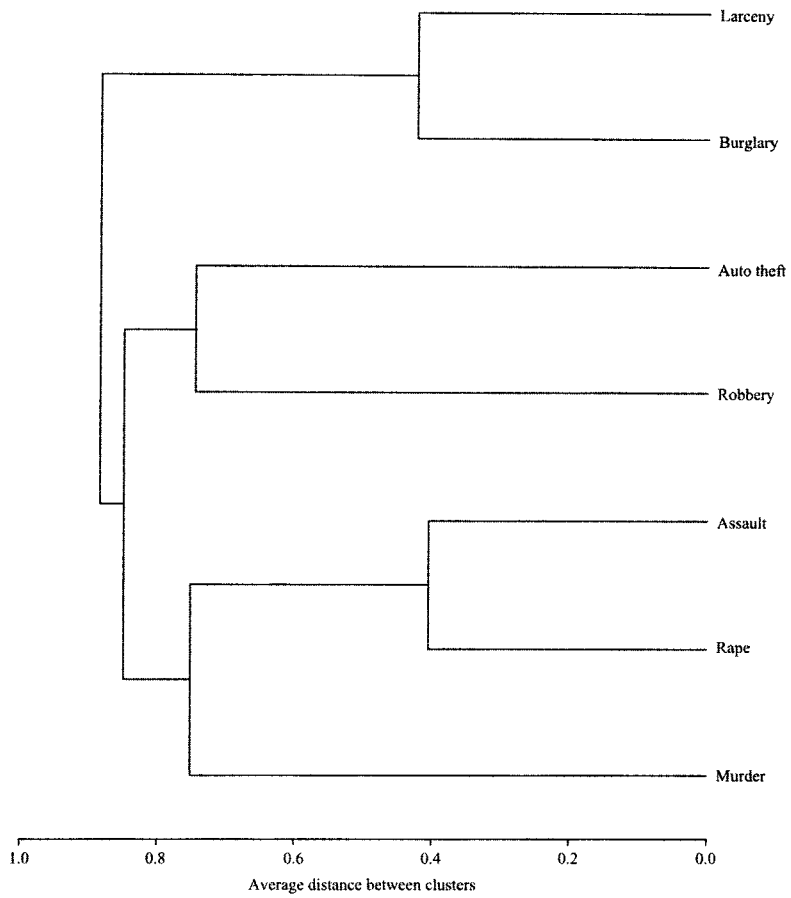


**Figure 15.23** Dendrogram for clustering the variables of Table 15.1 using average linkage (see Example 15.7).

(1) murder, rape, and assault, (2) robbery and auto theft, and (3) burglary and larceny. Since all three methods agree, we have some confidence in the validity of the solution. □

## PROBLEMS

**15.1** Show that $d^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p}(x_j - y_j)^2$ from (15.2) is equal to (15.5), $d^2(\mathbf{x}, \mathbf{y}) = (v_x - v_y)^2 + p(\overline{x} - \overline{y})^2 + 2v_x v_y(1 - r_{xy})$, where $v_x^2 = \sum_{j=1}^{p}(x_j - \overline{x})^2$, $\overline{x} = \sum_{j=1}^{p} x_j/p$, and $r_{yx}$ is defined in (15.6).

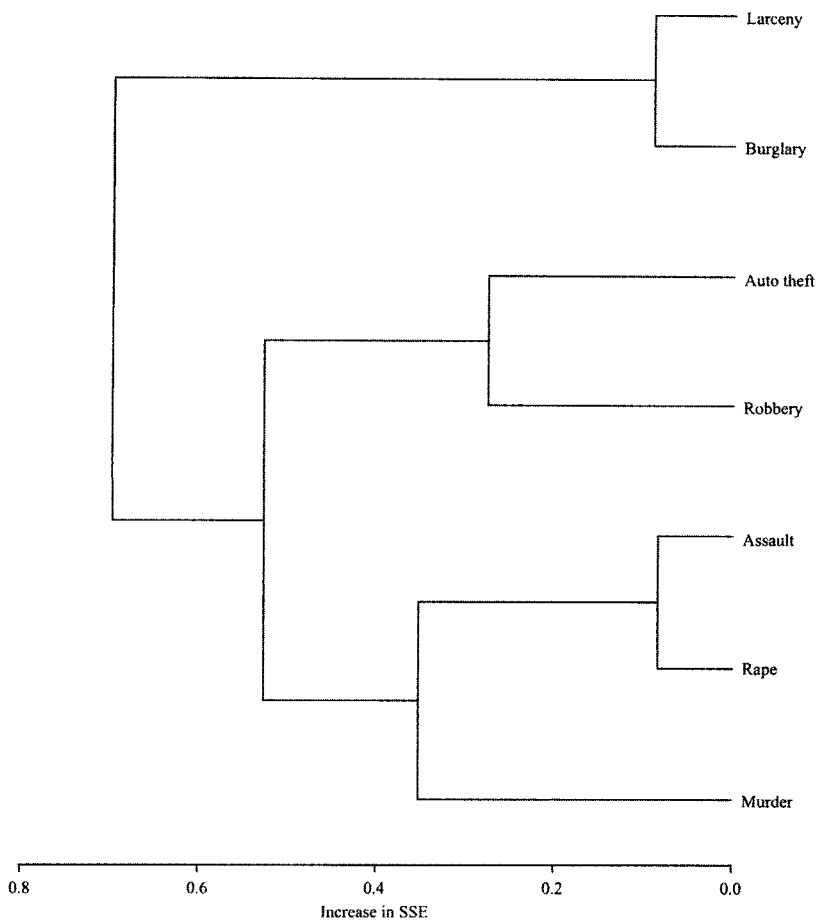


**Figure 15.24** Dendrogram for clustering the variables of Table 15.1 using Ward's method (see Example 15.7).

**15.2** **(a)** Show that $I_{AB} = n_A(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB}) + n_B(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB})$ as in (15.18).

**(b)** Show that (15.18) is equal to (15.19); that is,

$$n_A(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_{AB}) + n_B(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_B - \bar{\mathbf{y}}_{AB}) = \frac{n_A n_B}{n_A + n_B}(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B).$$

**15.3** Using the hints provided below, show that the parameter values for (15.20) in Table 15.2 produce appropriate distances for the following cluster methods.

**Table 15.12** Rotated Factor Loadings for City Crime Data

| Variables | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Murder | −.063 | **.734** | .142 |
| Rape | .504 | **.659** | .160 |
| Robbery | .133 | .355 | **.726** |
| Burglary | **.764** | .221 | .181 |
| Larceny | **.847** | −.014 | .244 |
| Auto theft | .240 | .097 | **.584** |

**(a)** Complete linkage. Use an approach analogous to that in Section 15.3.8 for the single linkage method.

**(b)** Average linkage. Write (15.20) in terms of parameter values for average linkage in Table 15.2. Then use (15.9).

**(c)** Centroid method. Show that

$$(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_{AB})'(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_{AB}) = \frac{n_A}{n_A + n_B}(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_A)'(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_A) + \frac{n_B}{n_A + n_B}(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_B)'(\bar{\mathbf{y}}_C - \bar{\mathbf{y}}_B) - \frac{n_A n_B}{(n_A + n_B)^2}(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)'(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B), \tag{15.40}$$

where $\bar{\mathbf{y}}_{AB} = (n_A\bar{\mathbf{y}}_A + n_B\bar{\mathbf{y}}_B)/(n_A + n_B)$.

**(d)** Median method. Use $n_A = n_B$ in (15.12) and (15.40) [see part (c)].

**(e)** Ward's method. Show that

$$I_{C(AB)} = \frac{n_A + n_C}{n_A + n_B + n_C} I_{AC} + \frac{n_B + n_C}{n_A + n_B + n_C} I_{BC} - \frac{n_C}{n_A + n_B + n_C} I_{AB},$$

where $I_{AB}$ is defined in (15.17).

**15.4** Show that for all methods in Table 15.2 for which $\gamma = 0$, we have $D(C, AB) > (\alpha_A + \alpha_B + \beta)D(A, B)$ as in (15.26).

**15.5** Verify the statement in the last paragraph of Section 15.4.1b, namely, that the first criterion in Section 15.4.1b is not invariant to nonsingular linear transformations $\mathbf{v}_{ij} = \mathbf{A}\mathbf{y}_{ij} + \mathbf{b}$, where $\mathbf{A}$ is a $p \times p$ nonsingular matrix, and that the other two criteria are invariant to such transformations. Use the following approach:

**(a)** Show that $\mathbf{H}_v = \mathbf{A}\mathbf{H}_y\mathbf{A}'$ and $\mathbf{E}_v = \mathbf{A}\mathbf{E}_y\mathbf{A}'$.

**(b)** Show that minimizing $\text{tr}(\mathbf{E})$ is not invariant.

**(c)** Show that minimizing $|\mathbf{E}|$ is invariant.

**(d)** Show that maximizing $\text{tr}(\mathbf{E}^{-1}\mathbf{H})$ is invariant.

**15.6** Verify the statement in Section 15.4.2a that in $\boldsymbol{\mu}_i, i = 1, 2, ..., g$; $\boldsymbol{\Sigma}_i$, $i = 1, 2, ..., g$; and $\alpha_i, i = 1, 2, ..., g - 1$; the total number of parameters is given by $\frac{1}{2}g(p+1)(p+2) - 1$ as in (15.35).

**15.7** Use the ramus bone date of Table 3.7. Carry out the following cluster methods and compare to the principal component plot in Figure 12.5.

**(a)** Find a two-cluster solution using the single linkage method.

**(b)** Find a two-cluster solution using the average linkage method and compare to the result in (a). Which seems better?

**(c)** Carry out a cluster analysis using the Ward's, complete linkage, centroid, and median methods.

**(d)** Use the flexible beta method with $\beta = -0.25, \beta = -0.5$, and $\beta = -0.75$.

**15.8** Use the hematology data of Table 4.2.

**(a)** Carry out a cluster analysis using the centroid method and find the distance between the centroids of the two-cluster solution.

**(b)** Carry out a cluster analysis using the average linkage method. How many clusters are indicated in the dendrogram?

**(c)** Using the two-cluster solution from part (b), label observations from one cluster as group 1 and the observations from the other cluster as group 2. Calculate and plot the discriminant function, as in Example 8.2. Do the two clusters overlap?

**15.9** Use all of the variables of the Seishu data of Table 7.1.

**(a)** Find the three-cluster solution using the single linkage, complete linkage, average linkage, centroid, median, and Ward's methods. Which observation appears to be an outlier? Which cluster is the same in all six solutions?

**(b)** Using the cluster found in part (a) to be common to all solutions as group 1 and the rest of the observations as group 2, calculate and plot the discriminant function, as in Problem 15.8(c). Do the two clusters overlap?

**15.10** Use the first 20 observations of the temperature data of Table 7.2. Standardize the variables (columns) before doing the following:

**(a)** Carry out a $k$-means cluster analysis using as initial seeds the five observations that are mutually farthest apart. Plot the first two discriminant functions using the five clusters as groups.

**(b)** Repeat part (a) using the first five observations as initial seeds.

**(c)** Repeat part (a) using as initial seeds the centroids of the five-cluster solution found using Ward's method. Plot the dendrogram resulting from Ward's method.

**(d)** Repeat part (c) using average linkage instead of Ward's method. Compare the results with those in part (c).

**(e)** Plot the first and second principal components and the second and third components. Which cluster solutions found in parts (a)–(d) seem to agree most with the principal component plots?

**(f)** Repeat parts (a) and (b) using three initial seeds instead of five. How do the cluster solutions compare?

**(g)** Repeat part (c) using three initial seeds instead of five. How does the cluster solution compare to your answer in (f)?

**15.11** Table 15.13 contains air pollution data from 41 US cities (Sokal and Rohlf 1981, p. 619). The variables are as follows:

$y_1 = SO_2$ content of air in micrograms per cubic meter

$y_2 =$ Average annual temperature in °F

$y_3 =$ Number of manufacturing enterprises employing 20 or more workers

$y_4 =$ Population size (1970 census) in thousands

$y_5 =$ Average annual wind speed in miles per hour

$y_6 =$ Average annual precipitation in inches

$y_7 =$ Average number of days with precipitation per year

Standardize each variable to mean 0 and standard deviation 1. Carry out a cluster analysis using the density estimation method with $k$ equal to 2, 3, 4, 5, and values of $r$ ranging from 0.2 to 2 by increments of 0.2 for each value of $k$. What is the maximum value of $k$ that produces a two-cluster solution?

**15.12** Table 15.14 gives the yields of winter wheat in each of the years 1970–1973 at twelve different sites in England (Hand et al. 1994, p. 31).

**(a)** Carry out a cluster analysis using the density estimation method with $k = 2, 3, 4$, and $r = .2, .4, \ldots, 2.0$.

**(b)** Plot the first two discriminant functions from the three-cluster solution obtained with $k = 2$ and $r = 1$.

**(c)** Plot the first two principal components and compare with the plot in part (b).

**(d)** Repeat part (b) using a two-cluster solution obtained with $k = 3$ and $r = 1$. Which two clusters of the three-cluster solution found in part (b) merged into one cluster?

**Table 15.13** Air Pollution Levels in US Cities

| Cities | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ |
|---|---|---|---|---|---|---|---|
| Phoenix | 10 | 70.3 | 213 | 582 | 6.0 | 7.05 | 36 |
| Little Rock | 13 | 61.0 | 91 | 132 | 8.2 | 48.52 | 100 |
| San Francisco | 12 | 56.7 | 453 | 716 | 8.7 | 20.66 | 67 |
| Denver | 17 | 51.9 | 454 | 515 | 9.0 | 12.95 | 86 |
| Hartford | 56 | 49.1 | 412 | 158 | 9.0 | 43.37 | 127 |
| Wilmington | 36 | 54.0 | 80 | 80 | 9.0 | 40.25 | 114 |
| Washington | 29 | 57.3 | 434 | 757 | 9.3 | 38.89 | 111 |
| Jacksonville | 14 | 68.4 | 136 | 529 | 8.8 | 54.47 | 116 |
| Miami | 10 | 75.5 | 207 | 335 | 9.0 | 59.80 | 128 |
| Atlanta | 24 | 61.5 | 368 | 497 | 9.1 | 48.34 | 115 |
| Chicago | 110 | 50.6 | 3344 | 3369 | 10.4 | 34.44 | 122 |
| Indianapolis | 28 | 52.3 | 361 | 746 | 9.7 | 38.74 | 121 |
| Des Moines | 17 | 49.0 | 104 | 201 | 11.2 | 30.85 | 103 |
| Wichita | 8 | 56.6 | 125 | 277 | 12.7 | 30.58 | 82 |
| Louisville | 30 | 55.6 | 291 | 593 | 8.3 | 43.11 | 123 |
| New Orleans | 9 | 68.3 | 204 | 361 | 8.4 | 56.77 | 113 |
| Baltimore | 47 | 55.0 | 625 | 905 | 9.6 | 41.31 | 111 |
| Detroit | 35 | 49.9 | 1064 | 1513 | 10.1 | 30.96 | 129 |
| Minneapolis-St. Paul | 29 | 43.5 | 699 | 744 | 10.6 | 25.94 | 137 |
| Kansas City | 14 | 54.5 | 381 | 507 | 10.0 | 37.00 | 99 |
| St. Louis | 56 | 55.9 | 775 | 622 | 9.5 | 35.89 | 105 |
| Omaha | 14 | 51.5 | 181 | 347 | 10.9 | 30.18 | 98 |
| Albuquerque | 11 | 56.8 | 46 | 244 | 8.9 | 7.77 | 58 |
| Albany | 46 | 47.6 | 44 | 116 | 8.8 | 33.36 | 135 |
| Buffalo | 11 | 47.1 | 391 | 463 | 12.4 | 36.11 | 166 |
| Cincinnati | 23 | 54.0 | 462 | 453 | 7.1 | 39.04 | 132 |
| Cleveland | 65 | 49.7 | 1007 | 751 | 10.9 | 34.99 | 155 |
| Columbus | 26 | 51.5 | 266 | 540 | 8.6 | 37.01 | 134 |
| Philadelphia | 69 | 54.6 | 1692 | 1950 | 9.6 | 39.93 | 115 |
| Pittsburgh | 61 | 50.4 | 347 | 520 | 9.4 | 36.22 | 147 |
| Providence | 94 | 50.0 | 343 | 179 | 10.6 | 42.75 | 125 |
| Memphis | 10 | 61.6 | 337 | 624 | 9.2 | 49.10 | 105 |
| Nashville | 18 | 59.4 | 275 | 448 | 7.9 | 46.00 | 119 |
| Dallas | 9 | 66.2 | 641 | 844 | 10.9 | 35.94 | 78 |
| Houston | 10 | 68.9 | 721 | 1233 | 10.8 | 48.19 | 103 |
| Salt Lake City | 28 | 51.0 | 137 | 176 | 8.7 | 15.17 | 89 |
| Norfolk | 31 | 59.3 | 96 | 308 | 10.6 | 44.68 | 116 |
| Richmond | 26 | 57.8 | 197 | 299 | 7.6 | 42.59 | 115 |
| Seattle | 29 | 51.1 | 379 | 531 | 9.4 | 38.79 | 164 |
| Charleston | 31 | 55.2 | 35 | 71 | 6.5 | 40.75 | 148 |
| Milwaukee | 16 | 45.7 | 569 | 717 | 11.8 | 29.07 | 123 |

**Table 15.14** Yields of Winter Wheat (kg per unit area)

| | Year | | | |
|---|---|---|---|---|
| Site | 1970 | 1971 | 1972 | 1973 |
| Cambridge | 46.81 | 39.40 | 55.64 | 32.61 |
| Cockle Park | 46.49 | 34.07 | 45.06 | 41.02 |
| Harpers Adams | 44.03 | 42.03 | 40.32 | 50.23 |
| Headley Hall | 52.24 | 36.19 | 47.03 | 34.56 |
| Morley | 36.55 | 43.06 | 38.07 | 43.17 |
| Myerscough | 34.88 | 49.72 | 40.86 | 50.08 |
| Rosemaund | 56.14 | 47.67 | 43.48 | 38.99 |
| Seale-Hayne | 45.67 | 27.30 | 45.48 | 50.32 |
| Sparsholt | 42.97 | 46.87 | 38.78 | 47.49 |
| Sutton Bonington | 54.44 | 49.34 | 24.48 | 46.94 |
| Terrington | 54.95 | 52.05 | 50.91 | 39.13 |
| Wye | 48.94 | 48.63 | 31.69 | 59.72 |