# 5
# Multidimensional Scaling and Correspondence Analysis

## 5.1 Introduction

In Chapter 3 we noted in passing that one of the most useful ways of using principal component analysis was to obtain a low-dimensional "map" of the data that preserved as far as possible the Euclidean distances between the observations in the space of the original $q$ variables. In this chapter we will make this aspect of principal component analysis more explicit and also introduce some other, more direct methods, which aim to produce similar maps of data that have a different form from the usual multivariate data matrix, $\mathbf{X}$. We will consider two such techniques The first, *multidimensional scaling*, is used, essentially, to represent an observed *proximity matrix* geometrically. Proximity matrices arise either directly from experiments in which subjects are asked to assess the similarity of pairs of stimuli, or indirectly; as a measure of the correlation, covariance, or distance of the pair of stimuli derived from the raw profile data, that is, the variable values in $\mathbf{X}$.

An example of the former is shown in Table 5.1. Here, judgements about various brands of cola made by two subjects using a visual analogue scale with anchor points "same" (having a score of 0) and "different" (having a score of 100). In this example, the resulting rating for a pair of colas is a *dissimilarity*—low values indicate that the two colas are regarded as more alike than high values, and vice versa. A *similarity measure* would have been obtained had the anchor points been reversed, although similarities are often scaled to lie in the interval $[0, 1]$. An example of a proximity matrix arising from the basic data matrix is shown in Table 5.2. Here, the Euclidean distances between a number of pairs of countries have been calculated from the birth and death rates of each country.

The second technique that will be described in this chapter is *correspondence analysis*, which is essentially an approach to displaying the associations among a set of categorical variables in a type of *scatterplot* or *map*, thus allowing a visual examination of any structure or pattern in the data. Table 5.3, for example, shows a cross classification of 538 cancer patients by histological type, and by their response

**Table 5.1** Dissimilarity Data for All Pairs of 10 Colas for Two Subjects

Subject 1:

| | Cola number | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 0 | | | | | | | | | |
| 2 | 16 | 0 | | | | | | | | |
| 3 | 81 | 47 | 0 | | | | | | | |
| 4 | 56 | 32 | 71 | 0 | | | | | | |
| 5 | 87 | 68 | 44 | 71 | 0 | | | | | |
| 6 | 60 | 35 | 21 | 98 | 34 | 0 | | | | |
| 7 | 84 | 94 | 98 | 57 | 99 | 99 | 0 | | | |
| 8 | 50 | 87 | 79 | 73 | 19 | 92 | 45 | 0 | | |
| 9 | 99 | 25 | 53 | 98 | 52 | 17 | 99 | 84 | 0 | |
| 10 | 16 | 92 | 90 | 83 | 79 | 44 | 24 | 18 | 98 | 0 |

Subject 2:

| | Cola number | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 0 | | | | | | | | | |
| 2 | 20 | 0 | | | | | | | | |
| 3 | 75 | 35 | 0 | | | | | | | |
| 4 | 60 | 31 | 80 | 0 | | | | | | |
| 5 | 80 | 70 | 37 | 70 | 0 | | | | | |
| 6 | 55 | 40 | 20 | 89 | 30 | 0 | | | | |
| 7 | 80 | 90 | 90 | 55 | 87 | 88 | 0 | | | |
| 8 | 45 | 80 | 77 | 75 | 25 | 86 | 40 | 0 | | |
| 9 | 87 | 35 | 50 | 88 | 60 | 10 | 98 | 83 | 0 | |
| 10 | 12 | 90 | 96 | 89 | 75 | 40 | 27 | 14 | 90 | 0 |

**Table 5.2** Euclidean Distance Matrix Based on Birth and Death Rates for Five Countries

(1) Raw data

| Country | Birth rate | Death rate |
|---|---|---|
| Algeria | 36.4 | 14.6 |
| France | 18.2 | 11.7 |
| Hungary | 13.1 | 9.9 |
| Poland | 19.0 | 7.5 |
| New Zealand | 25.5 | 8.8 |

(2) Euclidean distance matrix

| | Algeria | France | Hungary | Poland | New Zealand |
|---|---|---|---|---|---|
| Algeria | 0.00 | | | | |
| France | 18.43 | 0.00 | | | |
| Hungary | 23.76 | 5.41 | 0.00 | | |
| Poland | 18.79 | 4.28 | 6.37 | 0.00 | |
| New Zealand | 12.34 | 7.85 | 12.45 | 6.63 | 0.00 |

**Table 5.3** Hodgkin's Disease

| | Response | | | |
|---|---|---|---|---|
| Histological type | Positive | Partial | None | Total |
| LP | 74 | 18 | 12 | 104 |
| NS | 68 | 16 | 12 | 96 |
| MC | 154 | 54 | 58 | 266 |
| LD | 18 | 10 | 44 | 72 |
| Total | 314 | 98 | 126 | 538 |

to treatment three months after it had begun. A correspondence analysis of these data will be described later.

## 5.2 Multidimensional Scaling (MDS)

There are many methods of multidimensional scaling, and most of them are described in detail in Everitt and Rabe-Hesketh (1997). Here we shall concentrate on just one method, *classical multidimensional scaling*. Firstly, like all MDS techniques, classical scaling seeks to represent a proximity matrix by a simple geometrical model or map. Such a model is characterized by a set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, in $q$ dimensions, each point representing one of the stimuli of interest, and a measure of the distance between pairs of points. The objective of MDS is to determine both the dimensionality, $q$, of the model, and the $n$, $q$-dimensional coordinates, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ so that the model gives a "good" fit for the observed proximities. Fit will often be judged by some numerical index that measures how well the proximities and the distances in the geometrical model match. In essence this simply means that the larger an observed dissimilarity between two stimuli (or the smaller their similarity), the further apart should be the points representing them in the final geometrical model.

The question now arises as to how we estimate $q$, and the coordinate values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, from the observed proximity matrix? Classical scaling provides an answer to this question based on the work of Young and Householder (1938). To begin we must note that there is no unique set of coordinate values that give rise to these distances, since they are unchanged by shifting the whole configuration of points from one place to another, or by rotation or reflection of the configuration. In other words, we cannot uniquely determine either the location or the orientation of the configuration. The location problem is usually overcome by placing the mean vector of the configuration at the origin. The orientation problem means that any configuration derived can be subjected to an arbitrary orthogonal transformation. Such transformations can often be used to facilitate the interpretation of solutions as will be seen later.

The essential mathematical details of classical multidimensional scaling are given in Display 5.1.

**Display 5.1**
**Mathematical Details of Classical Multidimensional Scaling**

- To begin our account of the method we shall assume that the proximity matrix we are dealing with is a matrix of Euclidean distances derived from a raw data matrix, **X**.
- In Chapter 1, we saw how to calculate Euclidean distances from **X**. Multidimensional scaling is essentially concerned with the reverse problem: Given the distances (arrayed in the $n \times n$ matrix, **D**) how do we find **X**?
- To begin, define an $n \times n$ matrix **B** as follows

$$\mathbf{B} = \mathbf{X}\mathbf{X}' \tag{a}$$

- The elements of **B** are given by

$$b_{ij} = \sum_{k=1}^{q} x_{ik} x_{jk}. \tag{b}$$

- It is easy to see that the squared Euclidean distances between the rows of **X** can be written in terms of the elements of **B** as

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{c}$$

- If the $b$'s could be found in terms of the $d$'s in the equation above, then the required coordinate value could be derived by factoring **B** as in (a).
- No unique solution exists unless a location constraint is introduced. Usually the center of the points $\bar{\mathbf{x}}$ is set at the origin, so that $\sum_{i=1}^{n} x_{ik} = 0$ for all $k$.
- These constraints and the relationship given in (b) imply that the sum of the terms in any row of **B** must be zero.
- Consequently, summing the relationship given in (c) over $i$, over $j$, and finally over both $i$ and $j$, leads to the following series of equations:

$$\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj},$$

$$\sum_{i=1}^{n} d_{ij}^2 = nb_{ii} + T,$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nT,$$

  where $T = \sum_{i=1}^{n} b_{ii}$ is the trace of the matrix **B**.
- The elements of **B** can now be found in terms of squared Euclidean distances as

$$b_{ij} = -\frac{1}{2}\left[d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2\right],$$

where

$$d_{i\cdot}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2,$$

$$d_{\cdot j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2,$$

$$d_{\cdot\cdot}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2.$$

- Having now derived the elements of $\mathbf{B}$ in terms of Euclidean distances, it remains to factor it to give the coordinate values.
- In terms of its singular value decomposition $\mathbf{B}$ can be written as

$$\mathbf{B} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}',$$

  where $\boldsymbol{\Lambda} = \text{diag}[\lambda_1, \ldots, \lambda_n]$ is the diagonal matrix of eigenvalues of $\mathbf{B}$ and $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_n]$, the corresponding matrix of eigenvectors, normalized so that the sum of squares of their elements is unity, that is, $\mathbf{V}_i'\mathbf{V}_i = 1$. The eigenvalues are assumed labeled such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.
- When $\mathbf{D}$ arises from an $n \times q$ matrix of full rank, then the rank of $\mathbf{B}$ is $q$, so that the last $n - q$ of its eigenvalues will be zero.
- So $\mathbf{B}$ can be written as

$$\mathbf{B} = \mathbf{V}_1\boldsymbol{\Lambda}_1\mathbf{V}_1',$$

  where $\mathbf{V}_1$ contains the first $q$ eigenvectors and $\boldsymbol{\Lambda}_1$ the $q$ nonzero eigenvalues.
- The required coordinate values are thus

$$\mathbf{X} = \mathbf{V}_1\boldsymbol{\Lambda}_1^{1/2}$$

  where $\Lambda_1^{1/2} = \text{diag}[\lambda_1^{1/2}, \ldots, \lambda_p^{1/2}]$.
- The best fitting $k$-dimensional representation is given by the $k$ eigenvectors of $\mathbf{B}$ corresponding to the $k$ largest eigenvalues.
- The adequacy of the $k$-dimensional representation can be judged by the size of the criterion

$$P_k = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n-1}\lambda_i}.$$

- Values of $P_k$ of the order of 0.8 suggest a reasonable fit.
- When the observed dissimilarity matrix is not Euclidean, the matrix $\mathbf{B}$ is not positive-definite.
- In such cases some of the eigenvalues of $\mathbf{B}$ will be negative; correspondingly, some coordinate values will be complex numbers.

- If, however, **B** has only a small number of small negative eigenvalues, a useful representation of the proximity matrix may still be possible using the eigenvectors associated with the $k$ largest positive eigenvalues.
- The adequacy of the resulting solution might be assessed using one of the following two criteria suggested by Mardia et al. (1979)

$$P_k^{(1)} = \frac{\sum_{i=1}^{k} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|}$$

$$P_k^{(2)} = \frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$$

- Alternatively, Sibson (1979) recommends the following:

1. *Trace criterion*: Choose the number of coordinates so that the sum of their positive eigenvalues is approximately equal to the sum of all the eigenvalues.
2. *Magnitude criterion*: Accept as genuinely positive only those eigenvalues whose magnitude substantially exceeds that of the largest negative eigenvalue.

### 5.2.1 *Examples of Classical Multidimensional Scaling*

For our first example we will use the small set of multivariate data shown in Table 5.4, and the associated matrix of Euclidean distances will be our proximity matrix. To apply classical scaling to this matrix in R and S-PLUS® we can use the `dist` function to calculate the Euclidean distances combined with the `cmdscale` function to do the scaling

```
cmdscale(dist(x),k=5)
```

Here the five-dimensional solution (see Table 5.5) achieves complete recovery of the observed distance matrix. We can see this by comparing the original distances with those calculated from the scaling solution coordinates using the following R and S-PLUS code:

```
dist(x)- dist(cmdscale(dist(x), k=5)
```

The result is essentially a matrix of zeros.

The best fit in lower numbers of dimensions uses the coordinate values from the scaling solution in order from one to five. In fact, when the proximity matrix contains Euclidean distances derived from the raw data matrix, **X**, classical scaling can be shown to be equivalent to principal component analysis (see Chapter 3), with the derived coordinate values corresponding to the scores on the principal components derived from the covariance matrix. One result of this duality is the classical MDS

**Table 5.4** Multivariate Data and Associated Euclidean Distances

(1) Data

$$\mathbf{X} = \begin{pmatrix} 3 & 4 & 4 & 6 & 1 \\ 5 & 1 & 1 & 7 & 3 \\ 6 & 2 & 0 & 2 & 6 \\ 1 & 1 & 1 & 0 & 3 \\ 4 & 7 & 3 & 6 & 2 \\ 2 & 2 & 5 & 1 & 0 \\ 0 & 4 & 1 & 1 & 1 \\ 0 & 6 & 4 & 3 & 5 \\ 7 & 6 & 5 & 1 & 4 \\ 2 & 1 & 4 & 3 & 1 \end{pmatrix}$$

(2) Euclidean distances

$$\mathbf{D} = \begin{pmatrix} 0.00 & & & & & & & & & \\ 5.20 & 0.00 & & & & & & & & \\ 8.37 & 6.08 & 0.00 & & & & & & & \\ 7.87 & 8.06 & 6.32 & 0.00 & & & & & & \\ 3.46 & 6.56 & 8.37 & 9.27 & 0.00 & & & & & \\ 5.66 & 8.42 & 8.83 & 5.29 & 7.87 & 0.00 & & & & \\ 6.56 & 8.60 & 8.19 & 3.87 & 7.42 & 5.00 & 0.00 & & & \\ 6.16 & 8.89 & 8.37 & 6.93 & 6.00 & 7.07 & 5.70 & 0.00 & & \\ 7.42 & 9.05 & 6.86 & 8.89 & 6.56 & 7.55 & 8.83 & 7.42 & 0.00 & \\ 4.36 & 6.16 & 7.68 & 4.80 & 7.14 & 2.64 & 5.10 & 6.71 & 8.00 & 0.00 \end{pmatrix}$$

is often referred to as *principal coordinates analysis* (see Gower, 1966). The low-dimensional representation achieved by classical MDS for Euclidean distances (and that produced by principal component analysis) is such that the function $\phi$ given by

$$\phi = \sum_{r,s}^{n} (d_{rs}^2 - \hat{d}_{rs}^2)$$

is minimized. In this expression, $d_{rs}$ is the Euclidean distance between observations $r$ and $s$ in the original $q$-dimensional space, and $\hat{d}_{rs}$ is the corresponding distance in

**Table 5.5** Five-Dimensional Solution from Classical MDS Applied to the Distance Matrix in Table 5.4

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1.60 | 2.38 | 2.23 | −0.37 | 0.12 |
| 2 | 2.82 | −2.31 | 3.95 | 0.34 | 0.33 |
| 3 | 1.69 | −5.14 | −1.29 | 0.65 | −0.05 |
| 4 | −3.95 | −2.43 | −0.38 | 0.69 | 0.03 |
| 5 | 3.60 | 2.78 | 0.26 | 1.08 | −1.26 |
| 6 | −2.95 | 1.35 | 0.19 | −2.82 | 0.12 |
| 7 | −3.47 | 0.76 | −0.30 | 1.64 | −1.94 |
| 8 | −0.35 | 2.31 | −2.22 | 2.92 | 2.00 |
| 9 | 2.94 | −0.01 | −4.31 | −2.51 | −0.19 |
| 10 | −1.93 | 0.33 | 1.87 | −1.62 | 0.90 |

$k$-dimensional space ($k < q$) chosen for the classical scaling solution (equivalently the first $k$ components).

Now let us look at an example involving distances that are not Euclidean and for this we shall use the data shown in Table 5.6 giving the airline distances between 10 U.S. cities and available as the dataframe `airline.dist`. These distances are not Euclidean since they relate essentially to journeys along the surface of a sphere. To apply classical scaling to these distances and to see the eigenvalues we can use the following R and S-PLUS code:

```
airline.mds<-cmdscale(airline.dist, k=9, eig=T)
airline.mds$eig
```

The eigenvalues are shown in Table 5.7. Some are negative for these non-Euclidean distances (and there are some small differences between R and S-PLUS after the fourth eigenvalue). We will assess how many coordinates we need to adequately represent the observed distance matrix using the criterion, $P_k^{(1)}$ in Display 5.1. The values of the criterion calculated from the eigenvalues in Table 5.7 for the one-dimensional and two-dimensional solutions are

$$P_1^{(1)} = 0.74, \qquad P_1^{(2)} = 0.93,$$
$$P_2^{(1)} = 0.91, \qquad P_2^{(1)} = 0.99.$$

These values suggest that the first two coordinates will give an adequate representation of the observed distances.

The plot of the two-dimensional coordinate values is obtained using

```
#
par(pty="s")
#use same limits for x and y axes
#
plot(airline.mds$points[,1],airline.mds$points[,2],
type="n",xlab="Coordinate 1",ylab="Coordinate 2",
xlim=c(-2000,1500), ylim=c(-2000,1500))
```

**Table 5.6** Airline Distances Between 10 U.S. Cities

| | Atla | Chic | Denv | Hous | LA | Mia | NY | SF | Seat | Wash |
|---|---|---|---|---|---|---|---|---|---|---|
| Atlanta | — | 587 | 1212 | 701 | 1936 | 604 | 748 | 2139 | 218 | 543 |
| Chicago | 587 | — | 920 | 940 | 1745 | 1188 | 713 | 1858 | 1737 | 597 |
| Denver | 1212 | 920 | — | 879 | 831 | 1726 | 1631 | 949 | 1021 | 1494 |
| Houston | 701 | 940 | 879 | — | 1374 | 968 | 1420 | 1645 | 1891 | 1220 |
| Los Angeles | 1936 | 1745 | 831 | 1374 | — | 2338 | 2451 | 347 | 959 | 2300 |
| Miami | 604 | 1188 | 1726 | 968 | 2338 | — | 1092 | 2594 | 2734 | 923 |
| New York | 748 | 713 | 1631 | 1420 | 2451 | 1092 | — | 2571 | 2408 | 205 |
| San Francisco | 2139 | 1858 | 949 | 1645 | 347 | 2594 | 2571 | — | 678 | 2442 |
| Seattle | 218 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | — | 2329 |
| Wash. D.C | 543 | 597 | 1494 | 1220 | 2300 | 923 | 205 | 2442 | 2329 | — |

In dataframe `airline.dist`

**Table 5.7** Eigenvalues and Eigenvectors Arising from Classical Multidimensional Scaling Applied to Distance in Table 5.6

| Eigenvalues | City | 1 | 2 |
|---|---|---|---|
| $9.21 \times 10^6$ | Atlanta | 434.76 | −724.22 |
| $2.20 \times 10^6$ | Chicago | 412.61 | −55.04 |
| $1.08 \times 10^6$ | Denver | −468.20 | 180.66 |
| $3.32 \times 10^3$ | Houston | 175.58 | 515.22 |
| $3.86 \times 10^2$ | Los Angeles | −1206.68 | 465.64 |
| $-3.26 \times 10^{-1}$ | Miami | 1161.69 | 477.98 |
| $-9.30 \times 10$ | New York | 1115.56 | −199.79 |
| $-2.17 \times 10^3$ | San Francisco | −1422.69 | 308.66 |
| $-9.09 \times 10^3$ | Seattle | −1221.54 | −887.20 |
| $-1.72 \times 10^6$ | Wash. D.C | 1018.90 | −81.90 |

```
text(airline.mds$points[,1],airline.mds$points[,2],
labels=row.names(airline.dist))
```

and is shown in Figure 5.1. (The coordinates obtained from R may have different signs in which case some small amendments to the above code will be needed to get the same diagram as Figure 5.1.)

Our last example of the use of classical multidimensional scaling will involve the data shown in Table 5.8. These data show four measurements on male Egyptian skulls from five epochs. The measurements are

MB: Maximum breadth
BH: Basibregmatic height

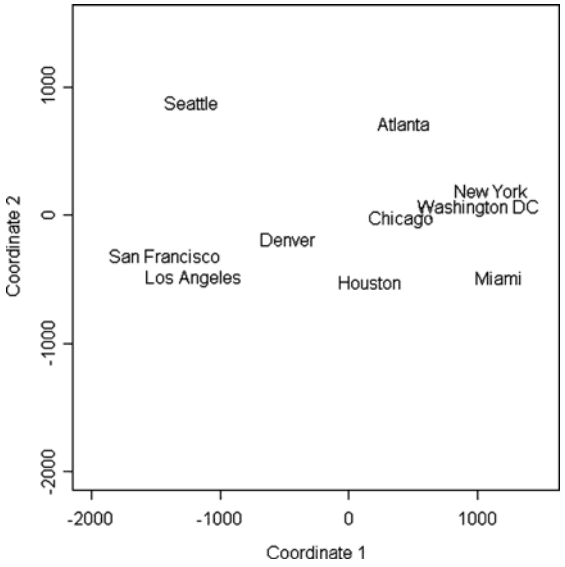


**Figure 5.1** Two-dimensional classical MDS solution for airline distances from S-PLUS.

**Table 5.8** Contents of Skull Dataframe. From *The Ancient Races of the Thebaid*, Arthur Thomson & R. Randall-Maciver (1905). By permission of Oxford University Press

| | EPOCH | MB | BH | BL | NH |
|---|---|---|---|---|---|
| 1 | c4000BC | 131 | 138 | 89 | 49 |
| 2 | c4000BC | 125 | 131 | 92 | 48 |
| 3 | c4000BC | 131 | 132 | 99 | 50 |
| 4 | c4000BC | 119 | 132 | 96 | 44 |
| 5 | c4000BC | 136 | 143 | 100 | 54 |
| 6 | c4000BC | 138 | 137 | 89 | 56 |
| 7 | c4000BC | 139 | 130 | 108 | 48 |
| 8 | c4000BC | 125 | 136 | 93 | 48 |
| 9 | c4000BC | 131 | 134 | 102 | 51 |
| 10 | c4000BC | 134 | 134 | 99 | 51 |
| 11 | c4000BC | 129 | 138 | 95 | 50 |
| 12 | c4000BC | 134 | 121 | 95 | 53 |
| 13 | c4000BC | 126 | 129 | 109 | 51 |
| 14 | c4000BC | 132 | 136 | 100 | 50 |
| 15 | c4000BC | 141 | 140 | 100 | 51 |
| 16 | c4000BC | 131 | 134 | 97 | 54 |
| 17 | c4000BC | 135 | 137 | 103 | 50 |
| 18 | c4000BC | 132 | 133 | 93 | 53 |
| 19 | c4000BC | 139 | 136 | 96 | 50 |
| 20 | c4000BC | 132 | 131 | 101 | 49 |
| 21 | c4000BC | 126 | 133 | 102 | 51 |
| 22 | c4000BC | 135 | 135 | 103 | 47 |
| 23 | c4000BC | 134 | 124 | 93 | 53 |
| 24 | c4000BC | 128 | 134 | 103 | 50 |
| 25 | c4000BC | 130 | 130 | 104 | 49 |
| 26 | c4000BC | 138 | 135 | 100 | 55 |
| 27 | c4000BC | 128 | 132 | 93 | 53 |
| 28 | c4000BC | 127 | 129 | 106 | 48 |
| 29 | c4000BC | 131 | 136 | 114 | 54 |
| 30 | c4000BC | 124 | 138 | 101 | 46 |
| 31 | c3300BC | 124 | 138 | 101 | 48 |
| 32 | c3300BC | 133 | 134 | 97 | 48 |
| 33 | c3300BC | 138 | 134 | 98 | 45 |
| 34 | c3300BC | 148 | 129 | 104 | 51 |
| 35 | c3300BC | 126 | 124 | 95 | 45 |
| 36 | c3300BC | 135 | 136 | 98 | 52 |
| 37 | c3300BC | 132 | 145 | 100 | 54 |
| 38 | c3300BC | 133 | 130 | 102 | 48 |
| 39 | c3300BC | 131 | 134 | 96 | 50 |
| 40 | c3300BC | 133 | 125 | 94 | 46 |
| 41 | c3300BC | 133 | 136 | 103 | 53 |
| 42 | c3300BC | 131 | 139 | 98 | 51 |
| 43 | c3300BC | 131 | 136 | 99 | 56 |
| 44 | c3300BC | 138 | 134 | 98 | 49 |
| 45 | c3300BC | 130 | 136 | 104 | 53 |
| 46 | c3300BC | 131 | 128 | 98 | 45 |
| 47 | c3300BC | 138 | 129 | 107 | 53 |
| 48 | c3300BC | 123 | 131 | 101 | 51 |
| 49 | c3300BC | 130 | 129 | 105 | 47 |
| 50 | c3300BC | 134 | 130 | 93 | 54 |
| 51 | c3300BC | 137 | 136 | 106 | 49 |
| 52 | c3300BC | 126 | 131 | 100 | 48 |
| 53 | c3300BC | 135 | 136 | 97 | 52 |

(*Continued*)

**Table 5.8** (*Continued*)

| | EPOCH | MB | BH | BL | NH |
|---|---|---|---|---|---|
| 54 | c3300BC | 129 | 126 | 91 | 50 |
| 55 | c3300BC | 134 | 139 | 101 | 49 |
| 56 | c3300BC | 131 | 134 | 90 | 53 |
| 57 | c3300BC | 132 | 130 | 104 | 50 |
| 58 | c3300BC | 130 | 132 | 93 | 52 |
| 59 | c3300BC | 135 | 132 | 98 | 54 |
| 60 | c3300BC | 130 | 128 | 101 | 51 |
| 61 | c1850BC | 137 | 141 | 96 | 52 |
| 62 | c1850BC | 129 | 133 | 93 | 47 |
| 63 | c1850BC | 132 | 138 | 87 | 48 |
| 64 | c1850BC | 130 | 134 | 106 | 50 |
| 65 | c1850BC | 134 | 134 | 96 | 45 |
| 66 | c1850BC | 140 | 133 | 98 | 50 |
| 67 | c1850BC | 138 | 138 | 95 | 47 |
| 68 | c1850BC | 136 | 145 | 99 | 55 |
| 69 | c1850BC | 136 | 131 | 92 | 46 |
| 70 | c1850BC | 126 | 136 | 95 | 56 |
| 71 | c1850BC | 137 | 129 | 100 | 53 |
| 72 | c1850BC | 137 | 139 | 97 | 50 |
| 73 | c1850BC | 136 | 126 | 101 | 50 |
| 74 | c1850BC | 137 | 133 | 90 | 49 |
| 75 | c1850BC | 129 | 142 | 104 | 47 |
| 76 | c1850BC | 135 | 138 | 102 | 55 |
| 77 | c1850BC | 129 | 135 | 92 | 50 |
| 78 | c1850BC | 134 | 125 | 90 | 60 |
| 79 | c1850BC | 138 | 134 | 96 | 51 |
| 80 | c1850BC | 136 | 135 | 94 | 53 |
| 81 | c1850BC | 132 | 130 | 91 | 52 |
| 82 | c1850BC | 133 | 131 | 100 | 50 |
| 83 | c1850BC | 138 | 137 | 94 | 51 |
| 84 | c1850BC | 130 | 127 | 99 | 45 |
| 85 | c1850BC | 136 | 133 | 91 | 49 |
| 86 | c1850BC | 134 | 123 | 95 | 52 |
| 87 | c1850BC | 136 | 137 | 101 | 54 |
| 88 | c1850BC | 133 | 131 | 96 | 49 |
| 89 | c1850BC | 138 | 133 | 100 | 55 |
| 90 | c1850BC | 138 | 133 | 91 | 46 |
| 91 | c200BC | 137 | 134 | 107 | 54 |
| 92 | c200BC | 141 | 128 | 95 | 53 |
| 93 | c200BC | 141 | 130 | 87 | 49 |
| 94 | c200BC | 135 | 131 | 99 | 51 |
| 95 | c200BC | 133 | 120 | 91 | 46 |
| 96 | c200BC | 131 | 135 | 90 | 50 |
| 97 | c200BC | 140 | 137 | 94 | 60 |
| 98 | c200BC | 139 | 130 | 90 | 48 |
| 99 | c200BC | 140 | 134 | 90 | 51 |
| 100 | c200BC | 138 | 140 | 100 | 52 |
| 101 | c200BC | 132 | 133 | 90 | 53 |
| 102 | c200BC | 134 | 134 | 97 | 54 |
| 103 | c200BC | 135 | 135 | 99 | 50 |
| 104 | c200BC | 133 | 136 | 95 | 52 |
| 105 | c200BC | 136 | 130 | 99 | 55 |
| 106 | c200BC | 134 | 137 | 93 | 52 |
| 107 | c200BC | 131 | 141 | 99 | 55 |
| 108 | c200BC | 129 | 135 | 95 | 47 |
| 109 | c200BC | 136 | 128 | 93 | 54 |
| 110 | c200BC | 131 | 125 | 88 | 48 |

(*Continued*)

**Table 5.8** (*Continued*)

| | EPOCH | MB | BH | BL | NH |
|---|---|---|---|---|---|
| 111 | c200BC | 139 | 130 | 94 | 53 |
| 112 | c200BC | 144 | 124 | 86 | 50 |
| 113 | c200BC | 141 | 131 | 97 | 53 |
| 114 | c200BC | 130 | 131 | 98 | 53 |
| 115 | c200BC | 133 | 128 | 92 | 51 |
| 116 | c200BC | 138 | 126 | 97 | 54 |
| 117 | c200BC | 131 | 142 | 95 | 53 |
| 118 | c200BC | 136 | 138 | 94 | 55 |
| 119 | c200BC | 132 | 136 | 92 | 52 |
| 120 | c200BC | 135 | 130 | 100 | 51 |
| 121 | cAD150 | 137 | 123 | 91 | 50 |
| 122 | cAD150 | 136 | 131 | 95 | 49 |
| 123 | cAD150 | 128 | 126 | 91 | 57 |
| 124 | cAD150 | 130 | 134 | 92 | 52 |
| 125 | cAD150 | 138 | 127 | 86 | 47 |
| 126 | cAD150 | 126 | 138 | 101 | 52 |
| 127 | cAD150 | 136 | 138 | 97 | 58 |
| 128 | cAD150 | 126 | 126 | 92 | 45 |
| 129 | cAD150 | 132 | 132 | 99 | 55 |
| 130 | cAD150 | 139 | 135 | 92 | 54 |
| 131 | cAD150 | 143 | 120 | 95 | 51 |
| 132 | cAD150 | 141 | 136 | 101 | 54 |
| 133 | cAD150 | 135 | 135 | 95 | 56 |
| 134 | cAD150 | 137 | 134 | 93 | 53 |
| 135 | cAD150 | 142 | 135 | 96 | 52 |
| 136 | cAD150 | 139 | 134 | 95 | 47 |
| 137 | cAD150 | 138 | 125 | 99 | 51 |
| 138 | cAD150 | 137 | 135 | 96 | 54 |
| 139 | cAD150 | 133 | 125 | 92 | 50 |
| 140 | cAD150 | 145 | 129 | 89 | 47 |
| 141 | cAD150 | 138 | 136 | 92 | 46 |
| 142 | cAD150 | 131 | 129 | 97 | 44 |
| 143 | cAD150 | 143 | 126 | 88 | 54 |
| 144 | cAD150 | 134 | 124 | 91 | 55 |
| 145 | cAD150 | 132 | 127 | 97 | 52 |
| 146 | cAD150 | 137 | 125 | 85 | 57 |
| 147 | cAD150 | 129 | 128 | 81 | 52 |
| 148 | cAD150 | 140 | 135 | 103 | 48 |
| 149 | cAD150 | 147 | 129 | 87 | 48 |
| 150 | cAD150 | 136 | 133 | 97 | 51 |

BL: Basialiveolar length
NH: Nasal height

We shall calculate Mahalanobis generalized distances (see Chapter 1) between each pair of epochs using the `mahalanobis` function, and apply classical scaling to the resulting distance matrix. In this calculation we shall use the following estimate of the assumed common covariance matrix $\mathbf{S}$,

$$\mathbf{S} = \frac{29\mathbf{S}_1 + 29\mathbf{S}_2 + 29\mathbf{S}_3 + 29\mathbf{S}_4 + 29\mathbf{S}_5}{149},$$

where $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_5$ are the covariance matrices of the data in each epoch. We shall then use the first two coordinate values to provide a map of the data showing the relationships between epochs. The necessary R and S-PLUS code is

```
labs<-rep(1:5,rep(30,5))
centers<-matrix(0,nrow=5,ncol=4)
S<-matrix(0,nrow=4,ncol=4)
#
for(i in 1:5) {
               centers[i,]<-apply(skulls[labs==i,-1],2,mean)
              S<-S+29*var(skulls[,-1])
}
#
S<-S/145
#
mahal<-matrix(0,5,5)
#
for(i in 1:5) {
    mahal[i,]<-mahalanobis(centers,centers[i,],S)
}
#
win.graph()
par(pty="s")
coords<-cmdscale(mahal)
#set up plotting area
xlim<-c(-1.5,1.5)
plot(coords,xlab="C1",ylab="C2",type="n",xlim=xlim,
   ylim=xlim,lwd=2)
text(coords,labels=c("c4000BC","c3300BC","c1850BC","c200BC",
   "cAD150"),lwd=3)
```

The resulting plot is shown in Figure 5.2.

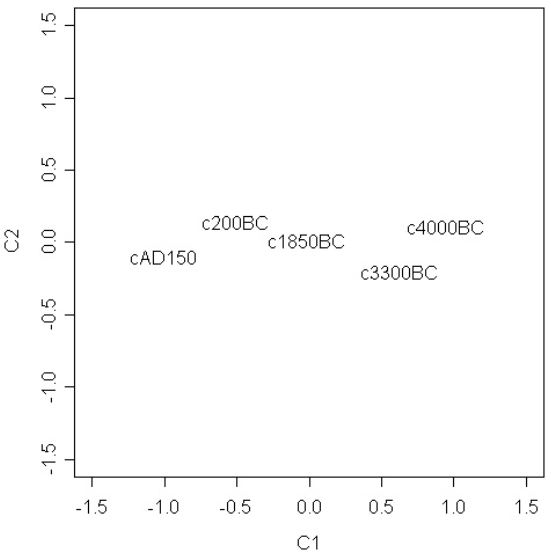


**Figure 5.2** Two-dimensional solution from classical multidimensional scaling applied to the Mahalanobis distances between epochs for the skull data.

The scaling solution for the skulls data is essentially unidimensional, with this single-dimension time ordering the five epochs. There appears to be a change in the "shape" of the skulls over time with maximum breadth increasing and basialiveolar length decreasing.

## 5.3 Correspondence Analysis

Correspondence analysis has a relatively long history (see de Leeuw, 1985), but for a long period was only routinely used in France, largely due to the almost evangelical efforts of Benzécri (1992). But nowadays the method is used more widely and is often applied to supplement say a standard chi-squared test of independence for two categorical variables forming a contingency table.

Mathematically, correspondence analysis can be regarded as either:

- A method for decomposing the chi-squared statistic used to test for independence in a contingency table into components corresponding to different dimensions of the heterogeneity between its columns; or
- A method for simultaneously assigning a scale to rows and a separate scale to columns so as to maximize the correlation between the two scales.

Quintessentially however, correspondence analysis is a technique for displaying multivariate (most often bivariate) categorical data graphically, by deriving coordinates to represent the categories of both the row and column variables, which may then be plotted so as to display the pattern of association between the variables graphically.

In essence, correspondence analysis is nothing more than the application of classical multidimensional scaling to a specific type of distance suitable for categorical data, namely what is known as the *chi-squared distance*. Such distances are defined in Display 5.2. (A detailed account of correspondence analysis in which its similarity to principal components analysis is stressed is given in Greenacre, 1992.)

**Display 5.2**
**Chi-Squared Distance**

- The general contingency table in which there are $r$ rows and $c$ columns can be written as

| | | Columns | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | $c$ | |
| | 1 | $n_{11}$ | $n_{12}$ | $n_{1c}$ | $n_{1.}$ |
| Rows | 2 | $n_{21}$ | | | |
| | ⋮ | | | | |
| | $r$ | $n_{r1}$ | | $n_{rc}$ | $n_{r.}$ |
| | | $n_{.1}$ | | $n_{.c}$ | $N$ |

  using an obvious dot notation.

- From this we can construct tables of column proportions and row proportions given by

  (a) Column proportions

$$\begin{array}{cccc} & 1 & & c \\ 1 & p_{11} = n_{11}/n_{1} & \cdots & p_{1c} = n_{1c}/n_{1.} \\ 2 & & & \\ \vdots & & & \\ r & p_{r1} = n_{r1}/n_{r.} & & p_{rc} = n_{rc}/n_{r.} \end{array}$$

  (b) Row proportions

$$\begin{array}{cccc} & 1 & & c \\ 1 & p_{11} = n_{11}/n_{.1} & \cdots & p_{1c} = n_{1c}/n_{.1} \\ 2 & & & \\ \vdots & & & \\ r & p_{r1} = n_{r1}/n_{.1} & & p_{rc} = n_{rc}/n_{.1} \end{array}$$

- The chi-squared distance between columns $i$ and $j$ is now defined as

$$d_{ij}^{(\text{cols})} = \sum_{k=1}^{r} \frac{1}{p_{k.}} (p_{ki} - p_{kj})^2$$

  where

$$p_{k.} = \frac{n_{k.}}{N}$$

  The chi-square distance is seen to be a weighted Euclidean distance based on column proportions. It will be zero if the two columns have the same values for these proportions. It can also be seen from the weighting factors $1/p_{k.}$ that rare categories of the column variable have a greater influence on the distance than common ones.
- A similar distance measure can be defined for rows $i$ and $j$ as

$$d_{ij}^{(\text{rows})} = \sum_{k=1}^{c} \frac{1}{p_{.k}} (p_{ik} - p_{jk})^2$$

  where

$$p_{.k} = \frac{n_{.k}}{N}$$

- A correspondence analysis results from applying classical MDS to each distance matrix in turn and plotting say the first two coordinates for column categories and those for row categories on the same diagram, suitably labelled to differentiate the points representing row categories from those representing column categories.

- An explanation of how to interpret the derived coordinates is simplified by considering only a one-dimensional solution.
- When the coordinates for both row and columns category are large and positive (or both large and negative), it indicates a *positive* association between row $i$ and column $j$; $n_{ij}$ is greater than expected under the assumption of independence.
- Similarly, when the coordinates are both large in absolute values but have different signs, the corresponding row and column have a negative association; $n_{ij}$ is less than expected under independence.
- Finally, when the product of the coordinates is near zero, the association between the row and column the column is low; $n_{ij}$ is close to the value expected under independence.

As a simple introductory example, consider the data shown in Table 5.9 concerned with the influence of a girl's age on her relationship with her boyfriend. In this table each of 139 girls has been classified into one of three groups:

- No boyfriend;
- Boyfriend/no sexual intercourse;
- Boyfriend/sexual intercourse.

In addition, the age of each girl was recorded and used to divide the girls into five age groups. The calculation of the chi-squared distance measure can be illustrated using the proportions of girls in age groups 1 and 2 for each relationship type from Table 5.9:

$$\text{Chi-squared distance} = \sqrt{\frac{(0.68-0.64)^2}{0.55} + \frac{(0.26-0.27)^2}{0.24} + \frac{(0.06-0.09)^2}{0.21}}$$
$$= 0.09.$$

**Table 5.9** The Influence of Age on Relationships with Boyfriends

| | Age group | | | | |
|---|---|---|---|---|---|
| | 1(AG1) | 2(AG2) | 3(AG3) | 4(AG4) | 5(AG5) |
| No boyfriend (nbf) | 21 | 21 | 14 | 13 | 8 |
| (row percentage) | (68) | (64) | (58) | (42) | (40) |
| Boyfriend/no sexual intercourse (bfns) | 8 | 9 | 6 | 8 | 2 |
| (row percentage) | (26) | (27) | (25) | (26) | (10) |
| Boyfriend/sexual intercourse (bfs) | 2 | 3 | 4 | 10 | 10 |
| (row percentage) | (6) | (9) | (17) | (32) | (50) |
| Totals | 31 | 33 | 24 | 31 | 20 |

NOTE: Age groups: (1) less than 16, (2) 16–17, (3) 17–18, (4) 18–19, (5) 19–20.

This is similar to ordinary Euclidean distance but differs in the division of each term by the corresponding average proportion. In this way the chi-squared distance measure compensates for the different levels of occurrence of the categories. (More formally, the choice of the chi-squared distance for measuring interprofile similarity can be justified as a way of standardizing variables under a multinomial or Poisson distributional assumption; see Greenacre, 1992.)

The complete set of chi-squared distances for all pairs of age groups can be arranged into the following matrix:

$$dcols = \begin{array}{c} \\ \text{age group1} \\ \text{age group2} \\ \text{age group3} \\ \text{age group4} \\ \text{age group5} \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.00 & 0.09 & 0.26 & 0.66 & 1.07 \\ 0.09 & 0.00 & 0.19 & 0.59 & 1.01 \\ 0.26 & 0.19 & 0.00 & 0.41 & 0.83 \\ 0.66 & 0.59 & 0.41 & 0.00 & 0.51 \\ 1.07 & 1.01 & 0.83 & 0.51 & 0.00 \end{pmatrix} \end{array}$$

The corresponding matrix for rows is

$$drows = \begin{array}{c} \\ \text{No boyfriend} \\ \text{Boyfriend/no sex} \\ \text{Boyfriend/sex} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.00 & 0.21 & 0.93 \\ 0.21 & 0.00 & 0.93 \\ 0.93 & 0.93 & 0.00 \end{pmatrix} \end{array}$$

Applying classical MDS to each of these distance matrices gives the two-dimensional coordinates shown in Table 5.10. Plotting those with suitable labels and with the axes suitably scaled to reflect the greater variation on dimension one than on dimension two is achieved using the R and S-PLUS code:

```
r1<-cmdscale(dcols,eig=T)
c1<-cmdscale(drows,eig=T)
par(pty="s")
plot(r1$points,xlim=range(r1$points[,1],c1$points[,1]),
   ylim=range(r1$points[,1],c1$points[,1]),type="n",
xlab="Coordinate 1",ylab="Coordinate 2",lwd=2)
```

**Table 5.10** Derived Correspondence Analysis Coordinates for Table 5.9

| | *x* | *y* |
|---|---|---|
| No boyfriend | −0.304 | −0.102 |
| Boyfriend/no sexual intercourse | −0.312 | 0.101 |
| Boyfriend/sexual intercourse | 0.617 | 0.000 |
| Age group 1 | −0.402 | 0.062 |
| Age group 2 | −0.340 | 0.004 |
| Age group 3 | −0.153 | −0.003 |
| Age group 4 | 0.225 | −0.152 |
| Age group 5 | 0.671 | 0.089 |

```
text(r1$points,labels=c("AG1","AG2","AG3","AG4","AG5"),lwd=2)
text(c1$points,labels=c("nobf","bfns","bfs"),lwd=4)
abline(h=0,lty=2)
abline(v=0,lty=2)
```

to give Figure 5.3.

The points representing the age groups in Figure 5.4 give a two-dimensional representation of this distance, with the Euclidean distance between two points representing the chi-squared distance between the corresponding age groups. (This is similar for the points representing each type of relationship.) For a contingency table with $I$ rows and $J$ columns, it can be shown that the chi-squared distances can be represented *exactly* in $\min\{I-1, J-1\}$ dimensions; here since $I = 3$ and $J = 5$, this means that the Euclidean distances in Figure 5.4 will equal the corresponding chi-squared distances. For example, the correspondence analysis coordinates for age groups 1 and 2 taken from Table 5.10 are

| Age group | $x$ | $y$ |
|---|---|---|
| 1 | −0.403 | 0.062 |
| 2 | −0.339 | 0.004 |

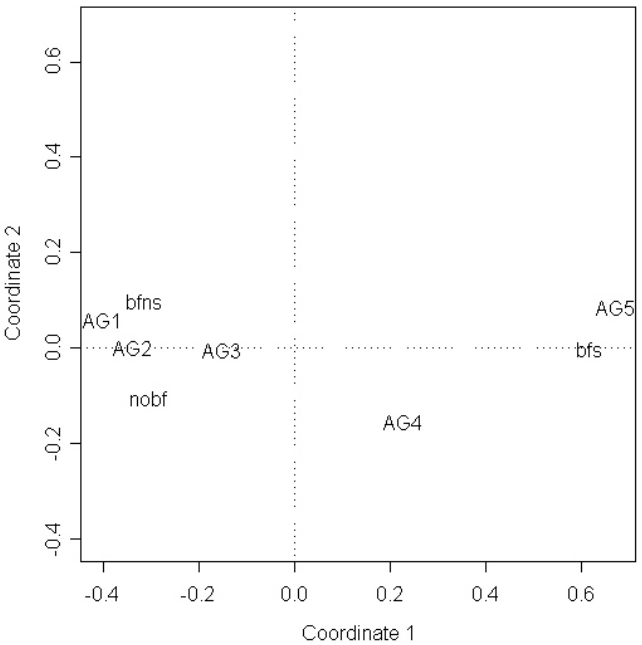


**Figure 5.3** Classical multidimensional scaling of data in Table 5.9.

The corresponding Euclidean distance is calculated as

$$\sqrt{(-0.403 - 0.339)^2 + (0.062 - 0.004)^2} = 0.09$$

which agrees with the chi-squared distance between the two age groups calculated earlier.

When both $I$ and $J$ are greater than 3, an exact two-dimensional representation of the chi-squared distances is not possible. In such cases the derived two-dimensional coordinates will give only an approximate representation, and so the question of the adequacy of the fit will need to be addressed. In some of these cases more than two dimensions may be required to given an acceptable fit.

A correspondence analysis is interpreted by examining the positions of the row categories and the column categories as reflected by their respective coordinate values. The values of the coordinates reflect associations between the categories of the row variable and those of the column variable. If we assume that a two-dimensional solution provides an adequate fit, then row points that are close together indicate row categories that have similar profiles (conditional distributions) across the columns. Column points that are close together indicate columns with similar profiles (conditional distributions) down the rows. Finally, row points that are close to column points represent combinations that occur more frequently than would be expected under an independence model, that is, one in which the categories of the row variable are unrelated to the categories of the column variable.

Let's now look at two further examples of the application of correspondence analysis.

### 5.3.1 *Smoking and Motherhood*

Table 5.11 shows a set of frequency data first reported by Wermuth (1976). The data show the distribution of birth outcomes by age of mother, length of gestation, and whether or not the mother smoked during the prenatal period. We shall consider the data as a two-dimensional contingency table with four row categories and four column categories.

**Table 5.11** Smoking and Motherhood

| | Premature | | Full term | |
|---|---|---|---|---|
| | Died in 1st year (pd) | Alive at year 1 (pa) | Died in 1st year (ftd) | Alive at year 1 (fta) |
| **Young mothers** | | | | |
| Nonsmokers (YN) | 50 | 315 | 24 | 4012 |
| Smokers (YS) | 9 | 40 | 6 | 459 |
| **Older mothers** | | | | |
| Nonsmokers (ONS) | 41 | 147 | 14 | 1594 |
| Smokers (YS) | 4 | 11 | 1 | 124 |

The obvious question of interest for the data in Table 5.11 is whether or not a mother's smoking puts a newborn baby at risk. However, several other questions might also be of interest. Are smokers more likely to have premature babies? Are older mothers more likely to have premature babies? And how does smoking affect premature babies?

The chi-squared statistic for testing the independence of the two variables forming Table 5.11 takes the value 19.11 with 9 degrees of freedom; the associated $p$-value is 0.024. So it appears that "type" of mother is related to what happens to the newborn baby. We shall now examine how the results from a correspondence analysis can shed a little more light on this rather general finding. The relevant chi-squared distance matrices for these data are:

$$dcols = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.00 & 0.30 & 0.27 & 0.37 \\ 0.30 & 0.00 & 0.23 & 0.07 \\ 0.27 & 0.23 & 0.00 & 0.28 \\ 0.37 & 0.07 & 0.28 & 0.00 \end{pmatrix} \end{array}$$

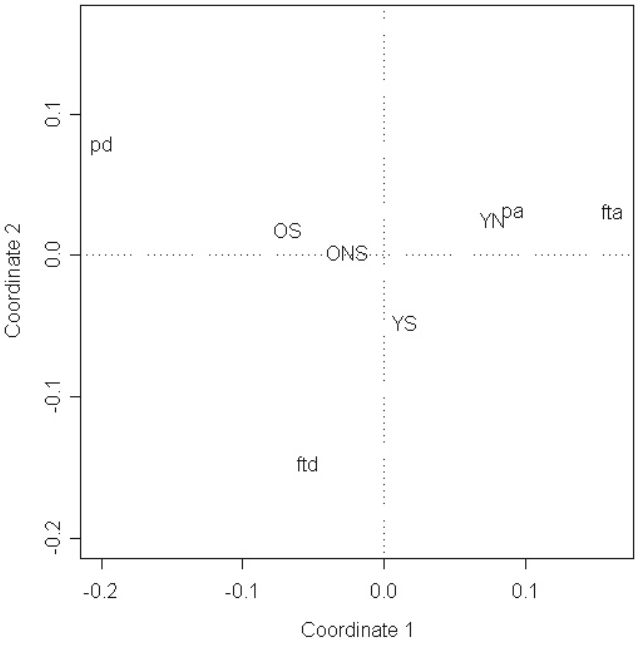


**Figure 5.4** Two-dimensional solution for classical MDS applied to the motherhood and smoking data in Table 5.11.

$$drows = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.00 & 0.10 & 0.11 & 0.15 \\ 0.10 & 0.00 & 0.07 & 0.11 \\ 0.11 & 0.07 & 0.00 & 0.05 \\ 0.15 & 0.11 & 0.05 & 0.00 \end{pmatrix} \end{array}$$

Applying classical MDS and plotting the two-dimensional solution as above gives Figure 5.4. This diagram suggests that young mothers who smoke tend to have more full-term babies who then die in their first year, and older mothers who smoke have rather more than expected premature babies who die in the first year. It does appear that smoking is a risk factor for death in the first year of the baby's life, and that age is associated with length of gestation, with older mothers delivering more premature babies.

## 5.3.2 *Hodgkin's Disease*

The data shown in Table 5.3 were recorded during a study of Hodgkin's disease, a cancer of the lymph nodes; the study is described in Hancock et al. (1979). Each of 538 patients with the disease was classified by histological type, and by their response to treatment three months after it had begun. The histological classification is:

- lymphocyte predominance (LP),
- nodular sclerosis (NS),
- mixed cellularity (MC),
- lymphocyte depletion (LD).

The key question is, "What, if any, is the relationship between histological type and response to treatment?"

Here the chi-squared statistic takes the value 75.89 with 6 degrees of freedom. The associated $p$-value is very small. Clearly histological classification and response to treatment are related, but can correspondence analysis help in uncovering more about this association?

In this example the two-dimensional solution from applying classical MDS to the chi-squared distances gives a perfect fit. The resulting scatterplot is shown in Figure 5.5. The positions of the points representing histological classification and response to treatment in this diagram imply the following:

- Lymphocyte depletion tends to result in no response to treatment.
- Nodular sclerosis and lymphocyte predominance are associated with a positive response to treatment.
- Mixed cellularity tends to result in a partial response to treatment.

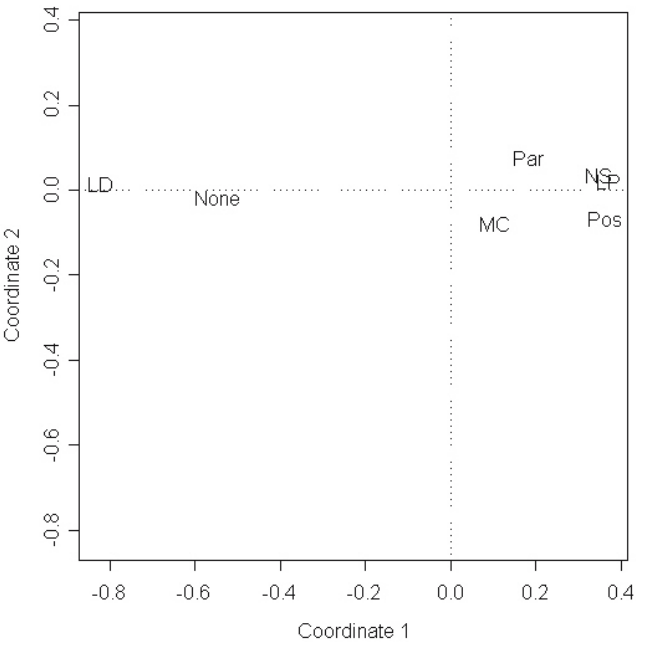


**Figure 5.5** Classical MDS two-dimensional solution for Hodgkin's disease data.

## 5.4 Summary

Multidimensional scaling and correspondence analysis both aim to help in the understanding of particular types of data by displaying the data graphically. Multidimensional scaling applied to proximity matrices is often useful in uncovering the dimensions on which similarity judgments are made, and correspondence analysis often allows more insight into the pattern of relationships in a contingency table then a simple chi-squared test.

## *Exercises*

5.1 What is mean by the *horseshoe effect* in multidimensional scaling solutions? (See Everitt and Rabe-Hesketh, 1997.) Create a similarity matrix as follows:

$$
\begin{aligned}
s_{ij} &= 9 \text{ if } i = j, \\
&= 8 \text{ if } 1 \leq |i - j| \leq 3, \\
&\vdots \\
&= 1 \text{ if } 2 \leq |i - j| \leq 2, \\
&= 0 \text{ if } |i - j| > 25.
\end{aligned}
$$

**Table 5.12** Dissimilarity Matrix for a Set of Eight Legal Offenses

| Offense | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | | | | | | | |
| 2 | 21.1 | 0 | | | | | | |
| 3 | 71.2 | 54.1 | 0 | | | | | |
| 4 | 36.4 | 36.4 | 36.4 | 0 | | | | |
| 5 | 52.1 | 54.1 | 52.1 | 0.7 | 0 | | | |
| 6 | 89.9 | 75.2 | 36.4 | 54.1 | 53.0 | 0 | | |
| 7 | 53.0 | 73.0 | 75.2 | 52.1 | 36.4 | 88.3 | 0 | |
| 8 | 90.1 | 93.2 | 71.2 | 63.4 | 52.1 | 36.4 | 73.0 | 0 |

Offenses: (1) assault and battery, (2) rape, (3) embezzlement, (4) perjury, (5) libel, (6) burglary, (7) prostitution, (8) receiving stolen goods.

Convert the resulting similarities into dissimilarities using $\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$ and find the two-dimensional configuration given by classical multidimensional scaling. The configuration should clearly show the horseshoe effect.

5.2 Show that classical multidimensional scaling applied to Euclidean distances calculated from a multivariate data matrix $\mathbf{X}$ is equivalent to principal components analysis, with the derived coordinate values corresponding to the scores on the principal components found from the covariance matrix of $\mathbf{X}$.

5.3 Write an S-PLUS (or R) function to calculate the chi-squared distance matrices for both rows and columns in a two-dimensional contingency table.

5.4 Table 5.12 summarizes data collected during a survey in which subjects were asked to compare a set of eight legal offenses, and to say for each one how unlike it was, in terms of seriousness, from the others. Each entry in the table shows the percentage of respondents who judged that the two offenses are very dissimilar. Find a two-dimensional scaling solution and try to interpret the dimensions underlying the subjects' judgements.

5.5 The data shown in Table 5.13 given the hair and eye color of a large number of people. Find the two-dimensional correspondence analysis solution for the data and plot the results.

**Table 5.13** Hair Color and Eye Color of a Sample of Individuals

| | Hair color | | | | |
|---|---|---|---|---|---|
| Eye color | Fair | Red | Medium | Dark | Black |
| Light | 688 | 116 | 584 | 188 | 4 |
| Blue | 326 | 38 | 241 | 110 | 3 |
| Medium | 343 | 84 | 909 | 412 | 26 |
| Dark | 98 | 48 | 403 | 681 | 81 |

**Table 5.14** Suicides by Method, Sex, and Age

| | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1970 | 1971 | 1972 | 1973 | 1974 | 1975 | 1976 | 1977 |
| Shooting | 15 | 15 | 31 | 17 | 42 | 49 | 38 | 27 |
| Stabbing | 95 | 113 | 94 | 125 | 124 | 126 | 148 | 127 |
| Blunt instrument | 23 | 16 | 34 | 34 | 35 | 33 | 41 | 41 |
| Poison | 9 | 4 | 8 | 3 | 5 | 3 | 1 | 4 |
| Manual violence | 47 | 60 | 54 | 70 | 69 | 66 | 70 | 60 |
| Strangulation | 43 | 45 | 43 | 53 | 51 | 63 | 47 | 51 |
| Smothering/drowning | 26 | 16 | 20 | 24 | 15 | 15 | 15 | 15 |

5.6 The data in Table 5.14 shows the methods by which victims of persons convicted for murder were killed between 1970 and 1977. How many dimensions would be needed for an *exact* correspondence analysis solution for these data? Use the first three correspondence analysis coordinates to plot a $3 \times 3$ scatterplot matrix (see Chapter 1). Interpret the results.