esults of nine randomised trials
ıe prevention of caries develop-
outcomes in each trial was the
ısing (due to caries) and filled
ı appropriate measure of effect
ta-analysis of the results. What

alysis on trials comparing
ber of individuals in the
ındard deviation for each

nB	meanB	sdB
13	4.72	4.72
51	5.07	5.38
40	2.51	3.22
79	3.20	2.46
39	5.81	5.14
36	4.76	5.29
09	10.90	7.90
12	3.01	3.32
73	4.37	5.37

your own meta-analysis func-
ect sizes and their associated
he overall effect size and its
l random effect models, and

CHAPTER 13

Principal Component Analysis: The Olympic Heptathlon

13.1 Introduction

The pentathlon for women was first held in Germany in 1928. Initially this consisted of the shot put, long jump, 100m, high jump and javelin events held over two days. In the 1964 Olympic Games the pentathlon became the first combined Olympic event for women, consisting now of the 80m hurdles, shot, high jump, long jump and 200m. In 1977 the 200m was replaced by the 800m and from 1981 the IAAF brought in the seven-event heptathlon in place of the pentathlon, with day one containing the events 100m hurdles, shot, high jump, 200m and day two, the long jump, javelin and 800m. A scoring system is used to assign points to the results from each event and the winner is the woman who accumulates the most points over the two days. The event made its first Olympic appearance in 1984.

In the 1988 Olympics held in Seoul, the heptathlon was won by one of the stars of women's athletics in the USA, Jackie Joyner-Kersee. The results for all 25 competitors in all seven disciplines are given in Table 13.1 (from Hand et al., 1994). We shall analyse these data using *principal component analysis* with a view to exploring the structure of the data and assessing how the derived principal component scores (see later) relate to the scores assigned by the official scoring system.

13.2 Principal Component Analysis

The basic aim of principal component analysis is to describe variation in a set of correlated variables, $x_1, x_2, \ldots, x_q$, in terms of a new set of uncorrelated variables, $y_1, y_2, \ldots, y_q$, each of which is a linear combination of the x variables. The new variables are derived in decreasing order of 'importance' in the sense that y_1 accounts for as much of the variation in the original data amongst all linear combinations of $x_1, x_2, \ldots, x_q$. Then y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1 – and so on, i.e., forming an orthogonal coordinate system. The new variables defined by this process, $y_1, y_2, \ldots, y_q$, are the principal components.

The general hope of principal component analysis is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, x_2, \ldots, x_q$, and can, consequently, be used to provide a conve-

nient lower-dimensional summary of these variables that might prove useful for a variety of reasons.

In some applications, the principal components may be an end in themselves and might be amenable to interpretation in a similar fashion as the factors in an *exploratory factor analysis* (see Everitt and Dunn, 2001). More often they are obtained for use as a means of constructing a low-dimensional informative graphical representation of the data, or as input to some other analysis.

The low-dimensional representation produced by principal component analysis is such that

$$\sum_{r=1}^{n}\sum_{s=1}^{n}\left(d_{rs}^2 - \hat{d}_{rs}^2\right)$$

is minimised with respect to $\hat{d}_{rs}^2$. In this expression, d_{rs} is the Euclidean distance (see Chapter 14) between observations r and s in the original q dimensional space, and $\hat{d}_{rs}$ is the corresponding distance in the space of the first m components.

As stated previously, the first principal component of the observations is that linear combination of the original variables whose sample variance is greatest amongst all possible such linear combinations. The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component. Subsequent components are defined similarly. The question now arises as to how the coefficients specifying the linear combinations of the original variables defining each component are found? The algebra of *sample* principal components is summarised briefly.

The first principal component of the observations, y_1, is the linear combination

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots, a_{1q}x_q$$

whose sample variance is greatest among all such linear combinations. Since the variance of y_1 could be increased without limit simply by increasing the coefficients $\mathbf{a}_1^\top = (a_{11}, a_{12}, \ldots, a_{1q})$ (here written in form of a vector for convenience), a restriction must be placed on these coefficients. As we shall see later, a sensible constraint is to require that the sum of squares of the coefficients, $\mathbf{a}_1^\top\mathbf{a}_1$, should take the value one, although other constraints are possible.

The second principal component $y_2 = \mathbf{a}_2^\top\mathbf{x}$ with $\mathbf{x} = (x_1, \ldots, x_q)$ is the linear combination with greatest variance subject to the two conditions $\mathbf{a}_2^\top\mathbf{a}_2 = 1$ and $\mathbf{a}_2^\top\mathbf{a}_1 = 0$. The second condition ensures that y_1 and y_2 are uncorrelated. Similarly, the jth principal component is that linear combination $y_j = \mathbf{a}_j^\top\mathbf{x}$ which has the greatest variance subject to the conditions $\mathbf{a}_j^\top\mathbf{a}_j = 1$ and $\mathbf{a}_j^\top\mathbf{a}_i = 0$ for $(i < j)$.

To find the coefficients defining the first principal component we need to choose the elements of the vector $\mathbf{a}_1$ so as to maximise the variance of y_1 subject to the constraint $\mathbf{a}_1^\top\mathbf{a}_1 = 1$.

Table 13.1: heptathlon data. Results Olympic heptathlon, Seoul, 1988.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

To maximise a function of several variables subject to one or more constraints, the method of *Lagrange multipliers* is used. In this case this leads to the solution that $\mathbf{a}_1$ is the eigenvector of the sample covariance matrix, $\mathbf{S}$, corresponding to its largest eigenvalue - full details are given in Morrison (2005).

The other components are derived in similar fashion, with $\mathbf{a}_j$ being the eigenvector of $\mathbf{S}$ associated with its jth largest eigenvalue. If the eigenvalues of $\mathbf{S}$ are $\lambda_1, \lambda_2, \ldots, \lambda_q$, then since $\mathbf{a}_j^\top \mathbf{a}_j = 1$, the variance of the jth component is given by λ_j.

The total variance of the q principal components will equal the total variance of the original variables so that

$$\sum_{j=1}^{q} \lambda_j = s_1^2 + s_2^2 + \ldots + s_q^2$$

where s_j^2 is the sample variance of x_j. We can write this more concisely as

$$\sum_{j=1}^{q} \lambda_j = \text{trace}(\mathbf{S}).$$

Consequently, the jth principal component accounts for a proportion P_j of the total variation of the original data, where

$$P_j = \frac{\lambda_j}{\text{trace}(\mathbf{S})}.$$

The first m principal components, where $m < q$, account for a proportion

$$P^{(m)} = \frac{\sum_{j=1}^{m} \lambda_j}{\text{trace}(\mathbf{S})}.$$

13.3 Analysis Using R

To begin it will help to score all the seven events in the same direction, so that 'large' values are 'good'. We will recode the running events to achieve this;

```
R> data("heptathlon", package = "HSAUR")
R> heptathlon$hurdles <- max(heptathlon$hurdles) -
+     heptathlon$hurdles
R> heptathlon$run200m <- max(heptathlon$run200m) -
+     heptathlon$run200m
R> heptathlon$run800m <- max(heptathlon$run800m) -
+     heptathlon$run800m
```

Figure 13.1 shows a scatterplot matrix of the results from the 25 competitors on the seven events. We see that most pairs of events are positively correlated to a greater or lesser degree. The exceptions all involve the javelin event - this is the only really 'technical' event and it is clear that training to become

```
R> score <- which(colnames(heptathlon) == "score")
R> plot(heptathlon[, -score])
```

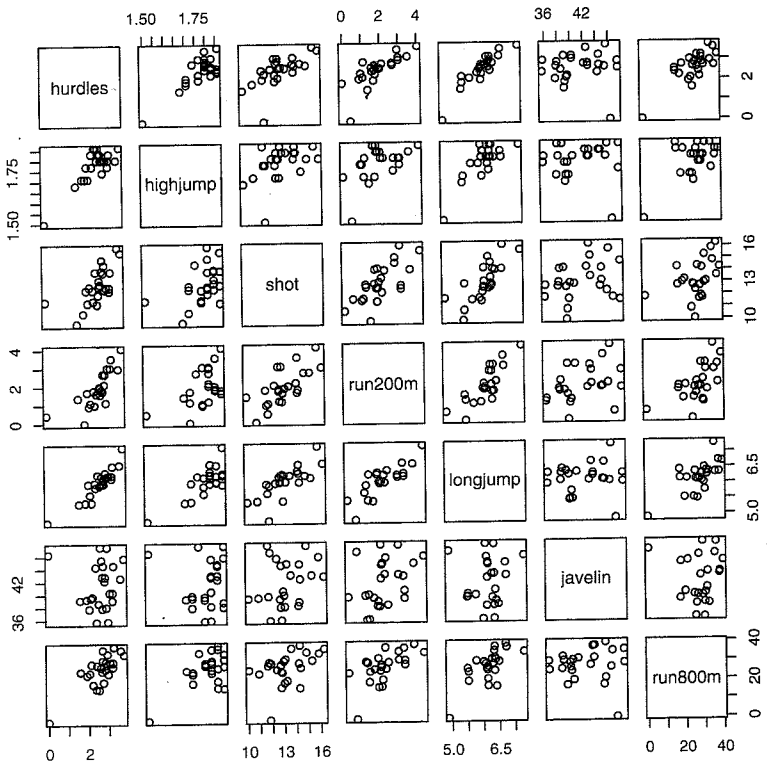

Figure 13.1 Scatterplot matrix for the `heptathlon` data.

successful in the other six 'power'-based events makes this event difficult for the majority of the competitors. We can examine the numerical values of the correlations by applying the `cor` function

```
R> round(cor(heptathlon[, -score]), 2)
```

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	0.81	0.65	0.77	0.91	0.01	0.78
highjump	0.81	1.00	0.44	0.49	0.78	0.00	0.59
shot	0.65	0.44	1.00	0.68	0.74	0.27	0.42
run200m	0.77	0.49	0.68	1.00	0.82	0.33	0.62
longjump	0.91	0.78	0.74	0.82	1.00	0.07	0.70
javelin	0.01	0.00	0.27	0.33	0.07	1.00	-0.02
run800m	0.78	0.59	0.42	0.62	0.70	-0.02	1.00

This correlation matrix demonstrates again the points made earlier.

A principal component analysis of the data can be applied using the `prcomp` function. The result is a list containing the coefficients defining each component (sometimes referred to as *loadings*), the principal component scores, etc. The required code is (omitting the `score` variable)

```
R> heptathlon_pca <- prcomp(heptathlon[, -score], scale = TRUE)
R> print(heptathlon_pca)
```

```
Standard deviations:
[1] 2.1119364 1.0928497 0.7218131 0.6761411 0.4952441 0.2701029
[7] 0.2213617

Rotation:
              PC1         PC2         PC3         PC4
hurdles  -0.4528710  0.15792058 -0.04514996  0.02653873
highjump -0.3771992  0.24807386 -0.36777902  0.67999172
shot     -0.3630725 -0.28940743  0.67618919  0.12431725
run200m  -0.4078950 -0.26038545  0.08359211 -0.36106580
longjump -0.4562318  0.05587394  0.13931653  0.11129249
javelin  -0.0754090 -0.84169212 -0.47156016  0.12079924
run800m  -0.3749594  0.22448984 -0.39585671 -0.60341130
              PC5         PC6         PC7
hurdles  -0.09494792 -0.78334101  0.38024707
highjump  0.01879888  0.09939981 -0.43393114
shot      0.51165201 -0.05085983 -0.21762491
run200m  -0.64983404  0.02495639 -0.45338483
longjump -0.18429810  0.59020972  0.61206388
javelin   0.13510669 -0.02724076  0.17294667
run800m   0.50432116  0.15555520 -0.09830963
```

The `summary` method can be used for further inspection of the details:

```
R> summary(heptathlon_pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     2.112 1.093 0.7218 0.6761 0.4952 0.2701
Proportion of Variance 0.637 0.171 0.0744 0.0653 0.0350 0.0104
Cumulative Proportion  0.637 0.808 0.8822 0.9475 0.9826 0.9930
                          PC7
Standard deviation     0.221
Proportion of Variance 0.007
Cumulative Proportion  1.000
```

The linear combination for the first principal component is

```
R> a1 <- heptathlon_pca$rotation[, 1]
R> a1
```

```
   hurdles   highjump       shot    run200m   longjump
-0.4528710 -0.3771992 -0.3630725 -0.4078950 -0.4562318
   javelin    run800m
-0.0754090 -0.3749594
```

We see that the 200m and long jump competitions receive the highest weight but the javelin result is less important. For computing the first principal component, the data need to be rescaled appropriately. The center and the scaling used by `prcomp` internally can be extracted from the `heptathlon_pca` via

```
R> center <- heptathlon_pca$center
R> scale <- heptathlon_pca$scale
```

Now, we can apply the `scale` function to the data and multiply with the loadings matrix in order to compute the first principal component score for each competitor

```
R> hm <- as.matrix(heptathlon[, -score])
R> drop(scale(hm, center = center, scale = scale) %*%
+     heptathlon_pca$rotation[, 1])
```

```
Joyner-Kersee (USA)          John (GDR)         Behmer (GDR)
       -4.121447626        -2.882185935        -2.649633766
 Sablovskaite (URS)    Choubenkova (URS)         Schulz (GDR)
       -1.343351210        -1.359025696        -1.043847471
      Fleming (AUS)       Greiner (USA)     Lajbnerova (CZE)
       -1.100385639        -0.923173639        -0.530250689
      Bouraga (URS)       Wijnsma (HOL)      Dimitrova (BUL)
       -0.759819024        -0.556268302        -1.186453832
     Scheider (SWI)         Braun (FRG)  Ruotsalainen (FIN)
        0.015461226         0.003774223         0.090747709
       Yuping (CHN)         Hagger (GB)         Brown (USA)
       -0.137225440         0.171128651         0.519252646
      Mulliner (GB)    Hautenauve (BEL)         Kytola (FIN)
        1.125481833         1.085697646         1.447055499
     Geremias (BRA)       Hui-Ing (TAI)       Jeong-Mi (KOR)
        2.014029620         2.880298635         2.970118607
       Launa (PNG)
        6.270021972
```

or, more conveniently, by extracting the first from all precomputed principal components

```
R> predict(heptathlon_pca)[, 1]
```

```
Joyner-Kersee (USA)          John (GDR)         Behmer (GDR)
       -4.121447626        -2.882185935        -2.649633766
 Sablovskaite (URS)    Choubenkova (URS)         Schulz (GDR)
       -1.343351210        -1.359025696        -1.043847471
      Fleming (AUS)       Greiner (USA)     Lajbnerova (CZE)
       -1.100385639        -0.923173639        -0.530250689
      Bouraga (URS)       Wijnsma (HOL)      Dimitrova (BUL)
       -0.759819024        -0.556268302        -1.186453832
     Scheider (SWI)         Braun (FRG)  Ruotsalainen (FIN)
        0.015461226         0.003774223         0.090747709
       Yuping (CHN)         Hagger (GB)         Brown (USA)
       -0.137225440         0.171128651         0.519252646
      Mulliner (GB)    Hautenauve (BEL)         Kytola (FIN)
```

```
R> plot(heptathlon_pca)
```

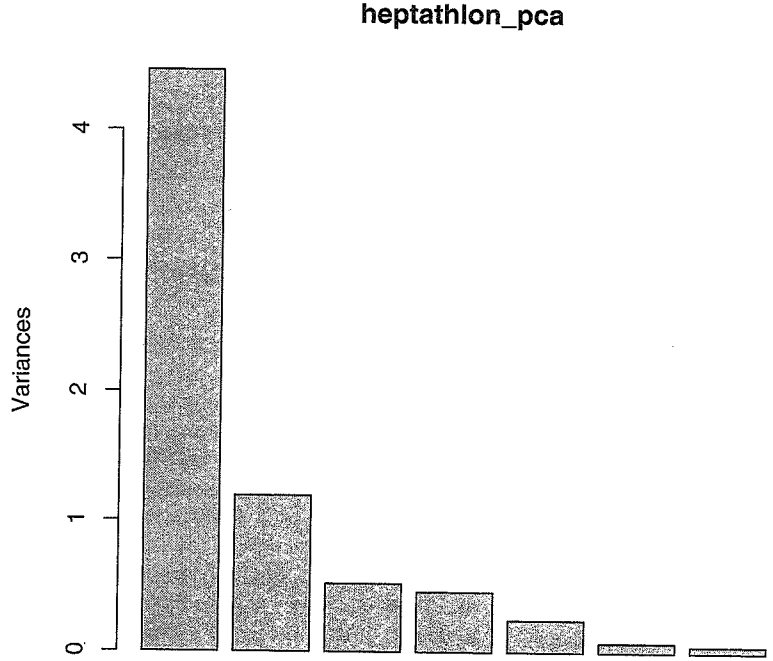

Figure 13.2 Barplot of the variances explained by the principal components.

```
1.125481833          1.085697646          1.447055499
Geremias (BRA)       Hui-Ing (TAI)        Jeong-Mi (KOR)
2.014029620          2.880298635          2.970118607
Launa (PNG)
6.270021972
```

The first two components account for 81% of the variance. A barplot of each component's variance (see Figure 13.2) shows how the first two components dominate. A plot of the data in the space of the first two principal components, with the points labelled by the name of the corresponding competitor can be produced as shown with Figure 13.3. In addition, the first two loadings for the events are given in a second coordinate system, also illustrating the special role of the javelin event. This graphical representation is known as *biplot* (Gabriel, 1971).

```
R> biplot(heptathlon_pca, col = c("gray", "black"))
```

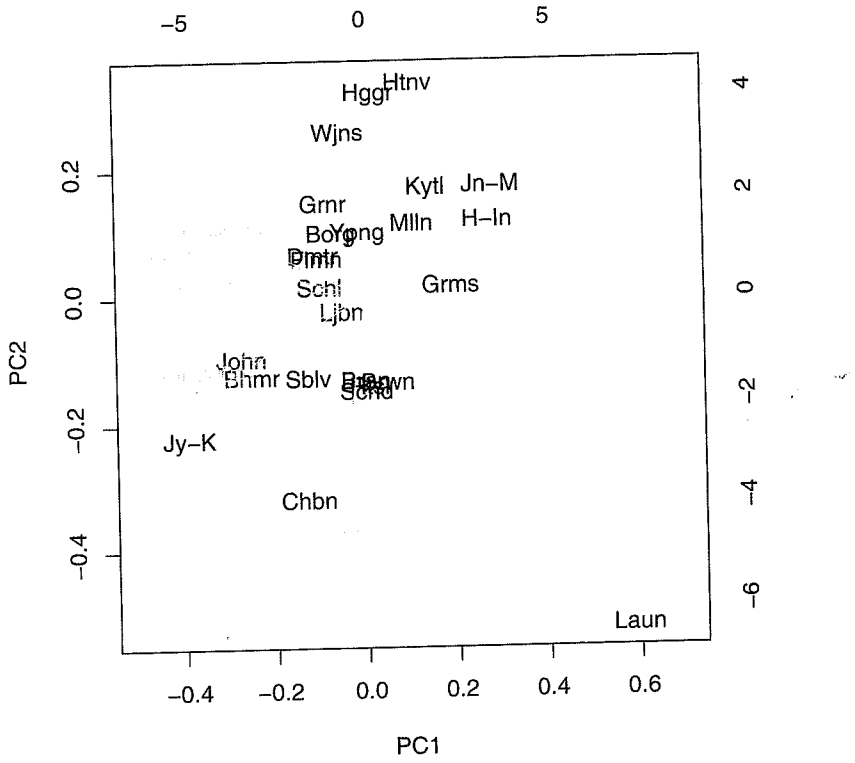

Figure 13.3 Biplot of the (scaled) first two principal components.

The correlation between the score given to each athlete by the standard scoring system used for the heptathlon and the first principal component score can be found from

```
R> cor(heptathlon$score, heptathlon_pca$x[, 1])
```

```
[1] -0.9910978
```

This implies that the first principal component is in good agreement with the score assigned to the athletes by official Olympic rules; a scatterplot of the official score and the first principal component is given in Figure 13.4.

13.4 Summary

Principal components look for a few linear combinations of the original variables that can be used to summarise a data set, losing in the process as little

```
R> plot(heptathlon$score, heptathlon_pca$x[, 1])
```

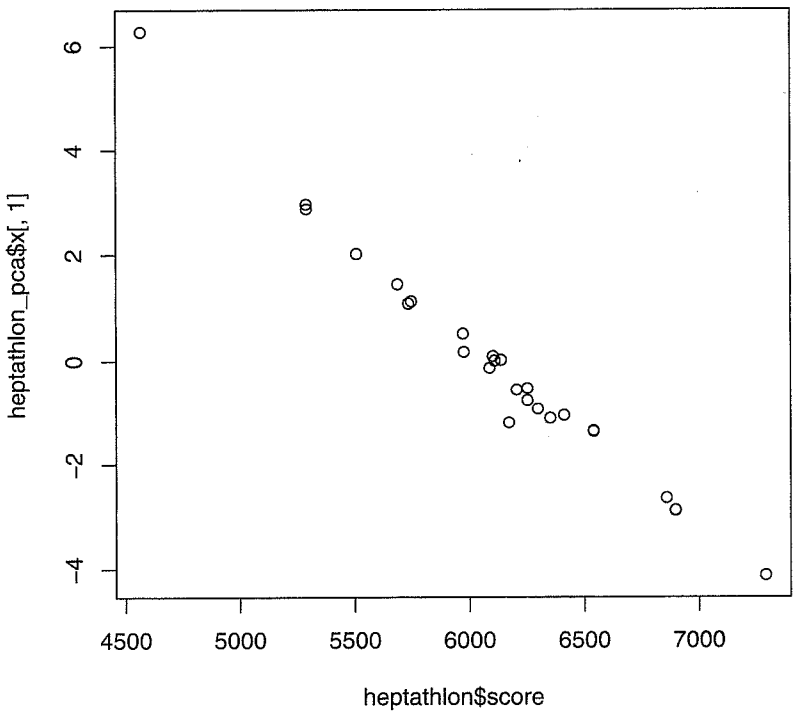

Figure 13.4 Scatterplot of the score assigned to each athlete in 1988 and the first principal component.

information as possible. The derived variables might be used in a variety of ways, in particular for simplifying later analyses and providing informative plots of the data. The method consists of transforming a set of correlated variables to a new set of variables which are uncorrelated. Consequently it should be noted that if the original variables are themselves almost uncorrelated there is little point in carrying out a principal components analysis, since it will merely find components which are close to the original variables but arranged in decreasing order of variance.

Exercises

Ex. 13.1 Apply principal components analysis to the covariance matrix of the heptathlon data (excluding the score variable) and compare your results with those given in the text, derived from the correlation matrix of the data. Which results do you think are more appropriate for these data?

Ex. 13.2 The data in Table 13.2 below give measurements on five meteorological variables over an 11-year period (taken from Everitt and Dunn, 2001). The variables are;

`year`: the corresponding year,
`rainNovDec`: rainfall in November and December (mm),
`temp`: average July temperature,
`rainJuly`: rainfall in July (mm),
`radiation`: radiation in July (curies), and
`yield`: average harvest yield (quintals per hectare).

Carry out a principal components analysis of both the covariance matrix and the correlation matrix of the data and compare the results. Which set of components leads to the most meaningful interpretation?

Table 13.2: `meteo` data. Meteorological measurements in an 11-year period.

year	rainNovDec	temp	rainJuly	radiation	yield
1920-21	87.9	19.6	1.0	1661	28.37
1921-22	89.9	15.2	90.1	968	23.77
1922-23	153.0	19.7	56.6	1353	26.04
1923-24	132.1	17.0	91.0	1293	25.74
1924-25	88.8	18.3	93.7	1153	26.68
1925-26	220.9	17.8	106.9	1286	24.29
1926-27	117.7	17.8	65.5	1104	28.00
1927-28	109.0	18.3	41.8	1574	28.37
1928-29	156.1	17.8	57.4	1222	24.96
1929-30	181.5	16.8	140.6	902	21.66
1930-31	181.4	17.0	74.3	1150	24.37

Source: From Everitt, B. S. and Dunn, G., *Applied Multivariate Data Analysis, 2nd Edition*, Arnold, London, 2001. With permission.

Ex. 13.3 The correlations below are for the calculus measurements for the six anterior mandibular teeth. Find all six principal components of the data and use a screeplot to suggest how many components are needed to adequately account for the observed correlations. Can you interpret the components?

Table 13.3. Correlations for calculus measurements for the six anterior mandibular teeth.

1.00					
0.54	1.00				
0.34	0.65	1.00			
0.37	0.65	0.84	1.00		
0.36	0.59	0.67	0.80	1.00	
0.62	0.49	0.43	0.42	0.55	1.00

CHAPTER 14

Multidimensional Scaling: British Water Voles and Voting in US Congress

14.1 Introduction

Corbet et al. (1970) report a study of water voles (genus *Arvicola*) in which the aim was to compare British populations of these animals with those in Europe, to investigate whether more than one species might be present in Britain. The original data consisted of observations of the presence or absence of 13 characteristics in about 300 water vole skulls arising from six British populations and eight populations from the rest of Europe. Table 14.1 gives a distance matrix derived from the data as described in Corbet et al. (1970).

Romesburg (1984) gives a set of data that shows the number of times 15 congressmen from New Jersey voted differently in the House of Representatives on 19 environmental bills. Abstentions are not recorded, but two congressmen abstained more frequently than the others, these being Sandman (nine abstentions) and Thompson (six abstentions). The data are available in Table 14.2 and one question of interest is can party affiliations be detected?

14.2 Multidimensional Scaling

The data in Tables 14.1 and 14.2 are both examples of *proximity matrices*. The elements of such matrices attempt to quantify how similar are stimuli, objects, individuals, etc. In Table 14.1 the values measure the 'distance' between populations of water voles; in Table 14.2 it is the similarity of the voting behaviour of the congressmen that is measured. Models are fitted to proximities in order to clarify, display and possibly explain any structure or pattern not readily apparent in the collection of numerical values. In some areas, particularly psychology, the ultimate goal in the analysis of a set of proximities is more specifically theories for explaining similarity judgements, or in other words, finding an answer to the question 'what makes things seem alike or seem different?' Here though we will concentrate on how proximity data can be best displayed to aid in uncovering any interesting structure.

The class of techniques we shall consider here, generally collected under the label *multidimensional scaling* (MDS), has the unifying feature that they seek to represent an observed proximity matrix by a simple geometrical model or map. Such a model consists of a series of say q-dimensional coordinate values,