A Handbook of Statistical Analyses Using R

Brian S. Everitt and Torsten Hothorn

CHAPTER 13

Principal Component Analysis: The Olympic Heptathlon

13.1 Introduction

13.2 Principal Component Analysis

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

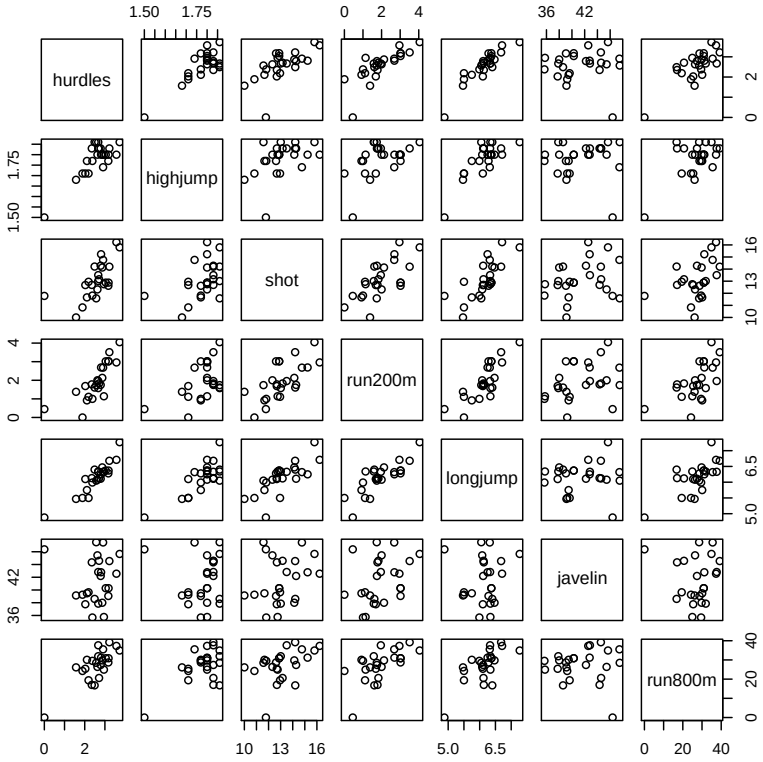

Figure 13.1 Scatterplot matrix for the `heptathlon` data.

```
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)
       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* (**?**).

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])
```

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

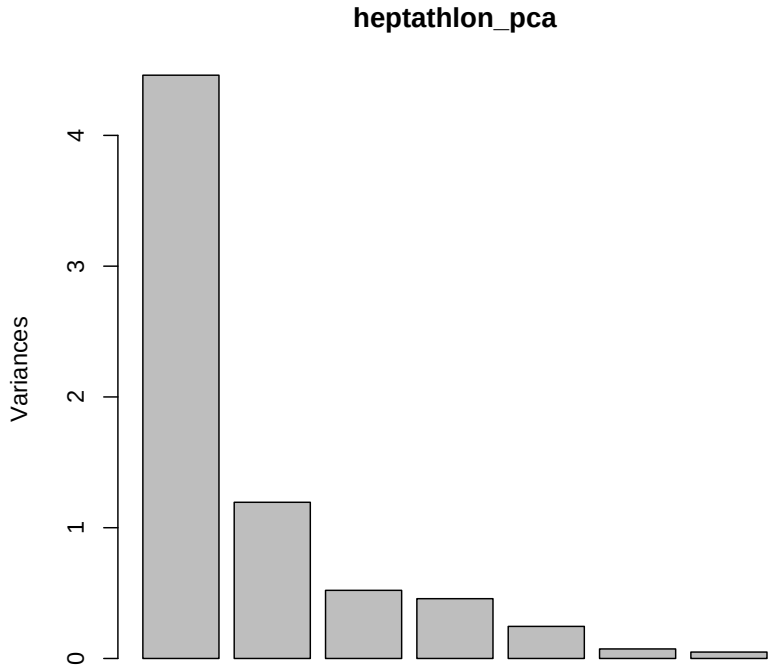

Figure 13.2 Barplot of the variances explained by the principal components.

```
[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.

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

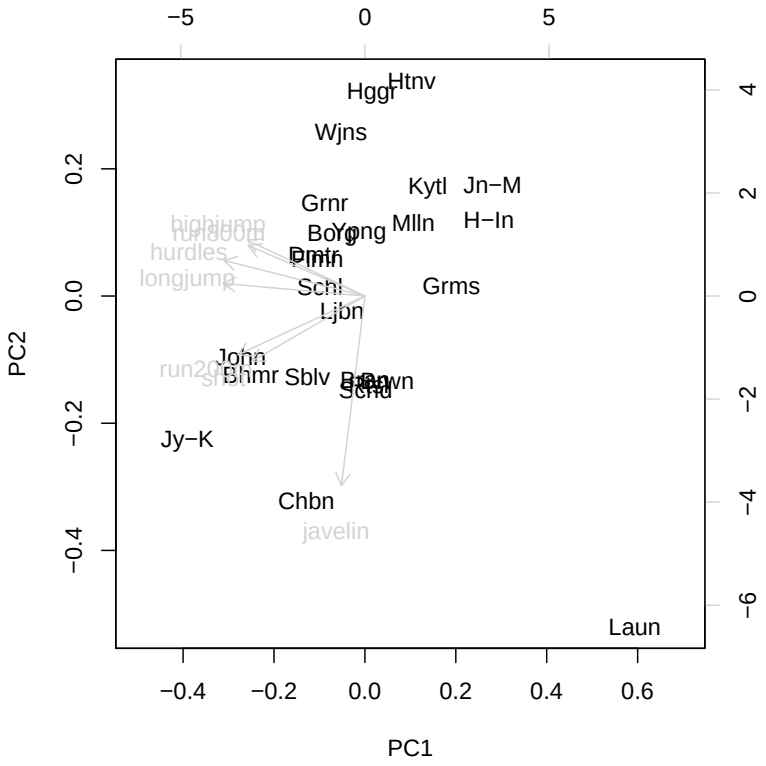

Figure 13.3 Biplot of the (scaled) first two principal components.

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

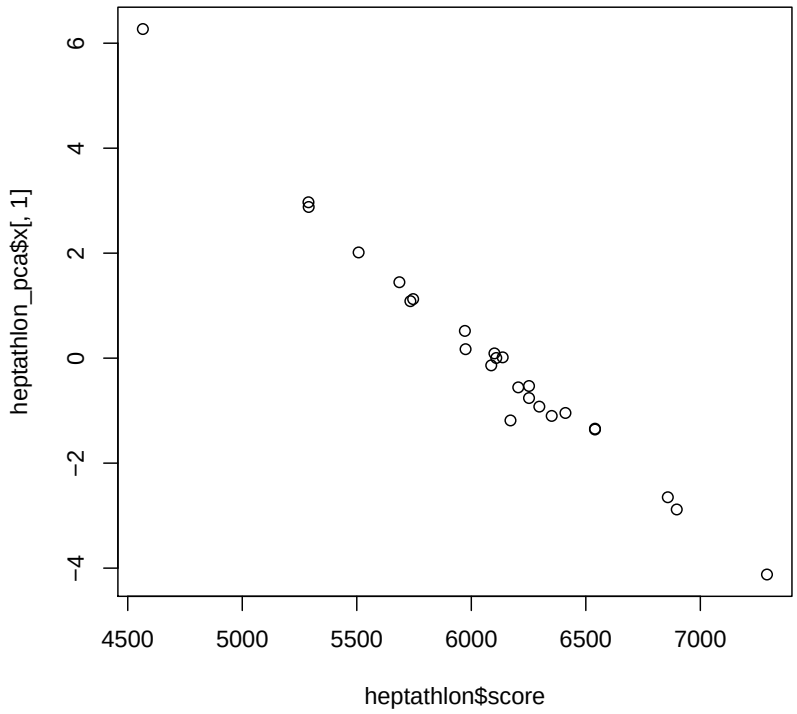

Figure 13.4 Scatterplot of the score assigned to each athlete in 1988 and the first principal component.