A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 19

Principal Component Analysis: The Olympic Heptathlon

19.1 Introduction

19.2 Principal Component Analysis

19.3 Analysis Using R

To begin it will help to score all 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 = "HSAUR3")
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 19.1 shows a scatterplot matrix of the results from all 25 competitors for the seven events. Most of the scatterplots in the diagram suggest that there is a positive relationship between the results for each pairs of events. The exception are the plots involving the javelin event which give little evidence of any relationship between the result for this event and the results from the other six events; we will suggest possible reasons for this below, but first we will examine the numerical values of the between pairs events 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
```

Examination of these numerical values confirms that most pairs of events are positively correlated, some moderately (for example, high jump and shot) and others relatively highly (for example, high jump and hurdles). And we see that the correlations involving the javelin event are all close to zero. One possible explanation for the latter finding is perhaps that training for the other six events does not help much in the javelin because it is essentially a 'technical'

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

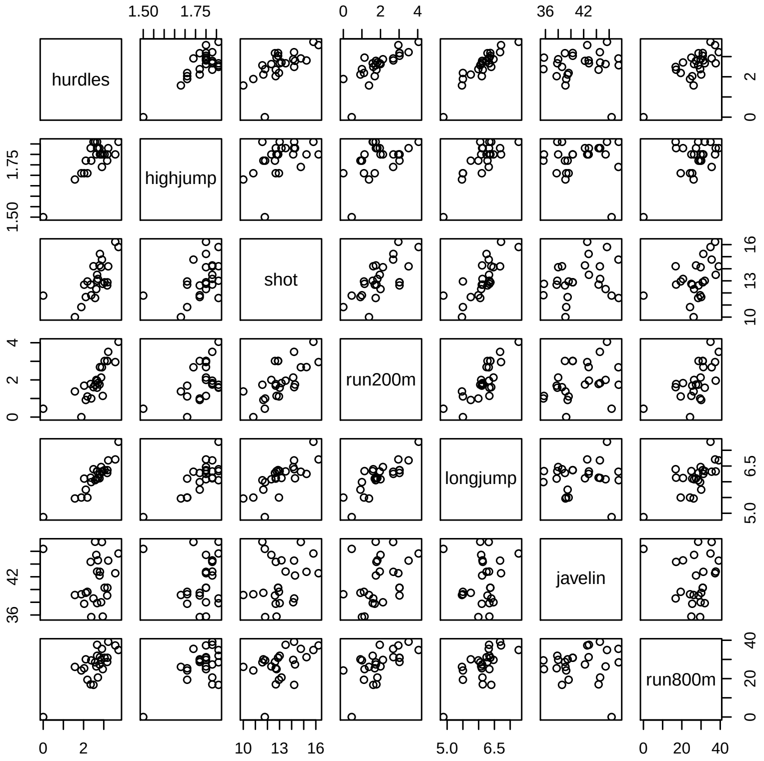

Figure 19.1 Scatterplot matrix for the `heptathlon` data (all countries).

event. An alternative explanation is found if we examine the scatterplot matrix in Figure 19.1 a little more closely. It is very clear in this diagram that for all events except the javelin there is an outlier, the competitor from Papua New Guinea (PNG), who is much poorer than the other athletes at these six events and who finished last in the competition in terms of points scored. But surprisingly in the scatterplots involving the javelin it is this competitor who again stands out but because she has the third highest value for the event. It might be sensible to look again at both the correlation matrix and the scatterplot matrix after removing the competitor from PNG; the relevant R code is

```
R> heptathlon <- heptathlon[-grep("PNG", rownames(heptathlon)),]
```

Now, we again look at the scatterplot and correlation matrix;

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

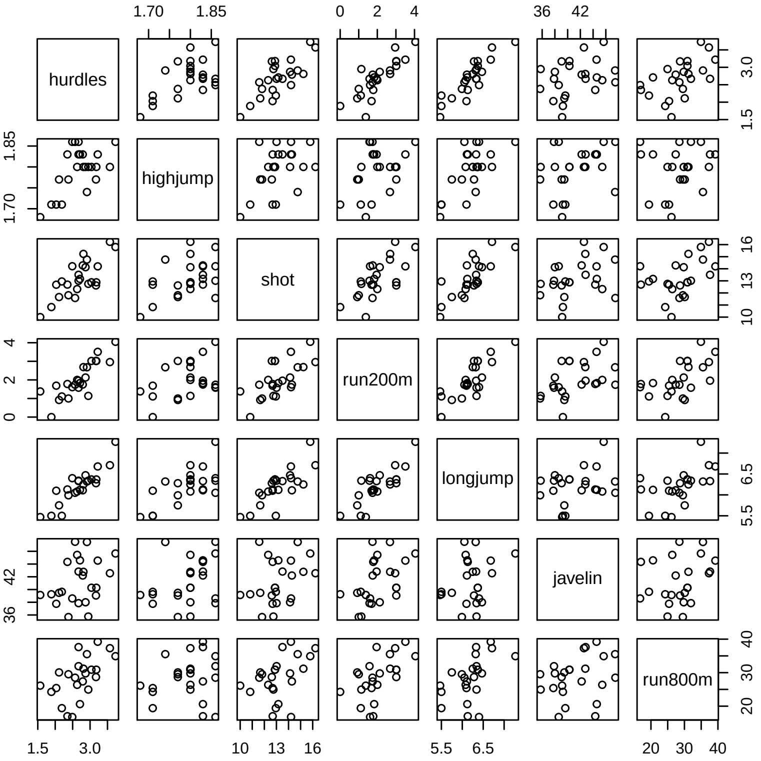

Figure 19.2 Scatterplot matrix for the `heptathlon` data after removing observations of the PNG competitor.

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

```
         hurdles highjump shot run200m longjump javelin run800m
hurdles     1.00     0.58 0.77    0.83     0.89    0.33    0.56
highjump    0.58     1.00 0.46    0.39     0.66    0.35    0.15
shot        0.77     0.46 1.00    0.67     0.78    0.34    0.41
run200m     0.83     0.39 0.67    1.00     0.81    0.47    0.57
longjump    0.89     0.66 0.78    0.81     1.00    0.29    0.52
javelin     0.33     0.35 0.34    0.47     0.29    1.00    0.26
run800m     0.56     0.15 0.41    0.57     0.52    0.26    1.00
```

The correlations change quite substantially and the new scatterplot matrix in Figure 19.2 does not point us to any further extreme observations. In the remainder of this chapter we analyze the `heptathlon` data with the observations of the competitor from Papua New Guinea removed.

Because the results for the seven heptathlon events are on different scales we shall extract the principal components from the correlation matrix. A principal component analysis of the data can be applied using the `prcomp` function with the `scale` argument set to `TRUE` to ensure the analysis is carried out on the correlation matrix. 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] 2.0793 0.9482 0.9109 0.6832 0.5462 0.3375 0.2620

Rotation:
             PC1      PC2     PC3      PC4      PC5      PC6
hurdles  -0.4504  0.05772 -0.1739  0.04841 -0.19889  0.84665
highjump -0.3145 -0.65133 -0.2088 -0.55695  0.07076 -0.09008
shot     -0.4025 -0.02202 -0.1535  0.54827  0.67166 -0.09886
run200m  -0.4271  0.18503  0.1301  0.23096 -0.61782 -0.33279
longjump -0.4510 -0.02492 -0.2698 -0.01468 -0.12152 -0.38294
javelin  -0.2423 -0.32572  0.8807  0.06025  0.07874  0.07193
run800m  -0.3029  0.65651  0.1930 -0.57418  0.31880 -0.05218
              PC7
hurdles  -0.06962
highjump  0.33156
shot      0.22904
run200m   0.46972
longjump -0.74941
javelin  -0.21108
run800m   0.07719
```

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     2.079 0.948 0.911 0.6832 0.5462 0.3375
Proportion of Variance 0.618 0.128 0.119 0.0667 0.0426 0.0163
Cumulative Proportion  0.618 0.746 0.865 0.9313 0.9739 0.9902
                           PC7
Standard deviation     0.26204
Proportion of Variance 0.00981
Cumulative Proportion  1.00000
```

The linear combination for the first principal component is

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

```
 hurdles highjump     shot  run200m longjump  javelin  run800m
  -0.450   -0.315   -0.402   -0.427   -0.451   -0.242   -0.303
```

We see that the hurdles 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)
           -4.75753            -3.14794            -2.92618
 Sablovskaite (URS)   Choubenkova (URS)        Schulz (GDR)
           -1.28814            -1.50345            -0.95847
      Fleming (AUS)       Greiner (USA)    Lajbnerova (CZE)
           -0.95345            -0.63324            -0.38157
      Bouraga (URS)       Wijnsma (HOL)     Dimitrova (BUL)
           -0.52232            -0.21770            -1.07598
     Scheider (SWI)         Braun (FRG) Ruotsalainen (FIN)
            0.00301             0.10918             0.20887
       Yuping (CHN)         Hagger (GB)         Brown (USA)
            0.23251             0.65952             0.75685
      Mulliner (GB)    Hautenauve (BEL)        Kytola (FIN)
            1.88093             1.82817             2.11820
     Geremias (BRA)       Hui-Ing (TAI)      Jeong-Mi (KOR)
            2.77071             3.90117             3.89685
```

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

```
R> predict(heptathlon_pca)[,1]
Joyner-Kersee (USA)          John (GDR)        Behmer (GDR)
           -4.75753            -3.14794            -2.92618
 Sablovskaite (URS)   Choubenkova (URS)        Schulz (GDR)
           -1.28814            -1.50345            -0.95847
      Fleming (AUS)       Greiner (USA)    Lajbnerova (CZE)
           -0.95345            -0.63324            -0.38157
      Bouraga (URS)       Wijnsma (HOL)     Dimitrova (BUL)
           -0.52232            -0.21770            -1.07598
     Scheider (SWI)         Braun (FRG) Ruotsalainen (FIN)
            0.00301             0.10918             0.20887
       Yuping (CHN)         Hagger (GB)         Brown (USA)
            0.23251             0.65952             0.75685
      Mulliner (GB)    Hautenauve (BEL)        Kytola (FIN)
            1.88093             1.82817             2.11820
     Geremias (BRA)       Hui-Ing (TAI)      Jeong-Mi (KOR)
            2.77071             3.90117             3.89685
```

The first two components account for 75% of the variance. A barplot of each component's variance (see Figure 19.3) shows how the first two components dominate. A plot of the data in the space of the first two principal components, with the points labeled by the name of the corresponding competitor, can be produced as shown with Figure 19.4. 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* (**?**). A biplot is a graphical representation of the information in an $n \times p$ data matrix. The 'bi' is a reflection that the technique produces a diagram that gives variance and covariance information about the variables and information about generalized distances between individuals. The coordinates used to produce the biplot can all be obtained directly from the principal components analysis of the covariance matrix of the data and so the plots can be

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

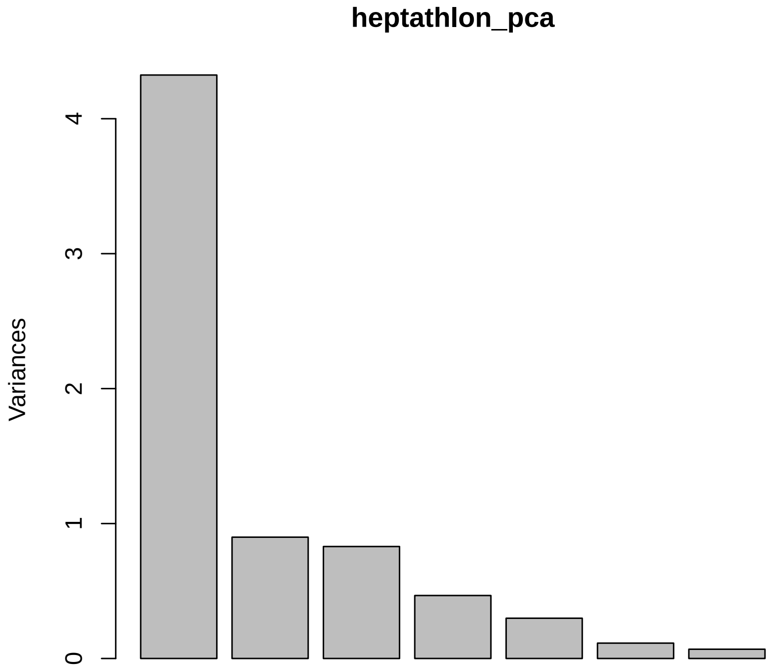

Figure 19.3 Barplot of the variances explained by the principal components (with observations for PNG removed).

viewed as an alternative representation of the results of such an analysis. Full details of the technical details of the biplot are given in **?** and in **?**. Here we simply construct the biplot for the heptathlon data (without PNG); the result is shown in Figure 19.4. The plot clearly shows that the winner of the gold medal, Jackie Joyner-Kersee, accumulates the majority of her points from the three events long jump, hurdles, and 200m.

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] -0.993
```

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

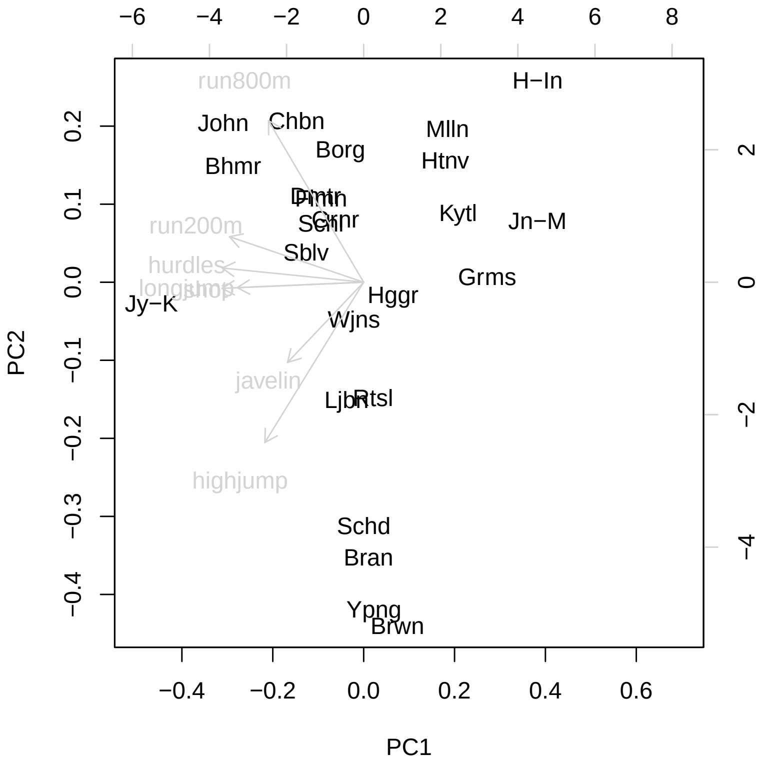

Figure 19.4 Biplot of the (scaled) first two principal components (with observations for PNG removed).

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 19.5.

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

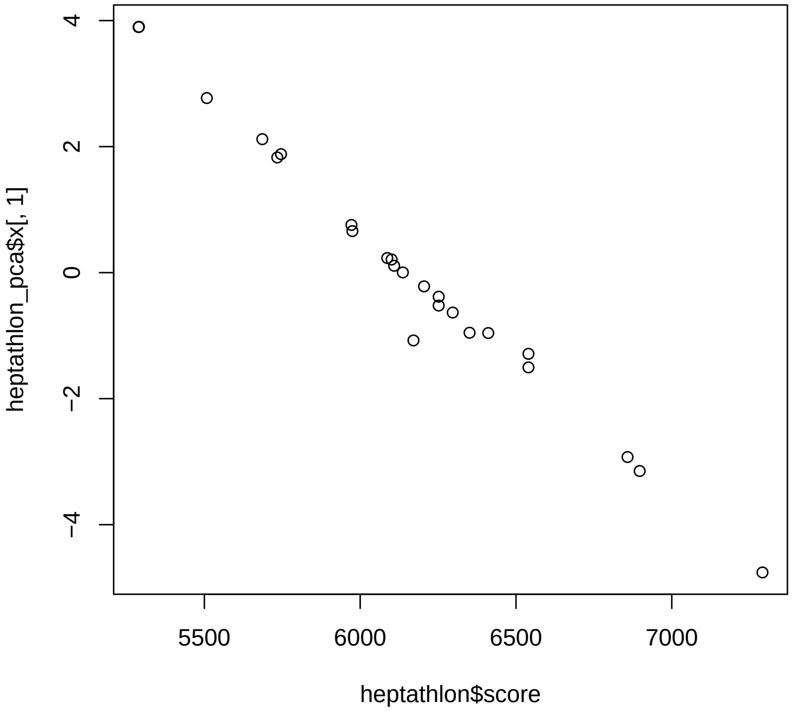

Figure 19.5 Scatterplot of the score assigned to each athlete in 1988 and the first principal component.