Statistical Methods for Big Data

Exercise Sheet 5

1. The Women’s Heptathlon is a seven-event Olympic sport for women. It includes:
   * hurdles
   * high jump
   * shot put
   * 200m run
   * long jump
   * javelin
   * 800m run

We wish to examine the results from the heptathlon at the 1988 Olympic games in Seoul. The results of each individual event for all 25 competitors were recorded along with the final position of each competitor. Position 1 corresponds to first place and position 25 corresponds to last place etc.

The aim of this exercise sheet is run a principal components analysis on the data set and to use principal components analysis to find patterns in the heptathlon results.

Download the heptathlon data set from Blackboard.

In the heptathlon data set a ‘good’ performance in the running events corresponded to a low time being recorded. We need to modify the dataset so that large values indicate a ‘good’ performance. We can transform the running events so that a ‘good’ performance corresponds to a large value.

e.g.

heptathlon$run200m<-max(heptathlon$run200m)-heptathlon$run200m

1. Examine the data set using boxplots, histograms and scatterplots.
2. Do you find any outliers in the data set?
3. Assign the data in the Position column to the row names and remove the column containing the position of the competitors.

row.names(heptathlon)<-heptathlon$Position

heptathlon$Position<- NULL

1. Perform principal components analysis on the data set.
2. If we use the Kaiser Criterion to select the number of principal components used for the analysis, how many components will be retained?
3. Produce a scores plot and a loading plot.
4. Comment on the score plot.
5. Use the loadings plot to identify which four events had the highest correlation with PC1 and which event had the lowest correlation with PC1.
6. Which events are correlated with PC2?
7. Looking at PC2, if a competitor performed well in the high jump, which event are they less likely to have performed well in?
8. Which events did the competitor that finished in 1st place perform well in?
9. Produce a scores plot and a loading plot for PC1 only. How well does PC1 correspond with the criteria used to determine a participants position?
10. Produce a scores plot and a loading plot for PC1 vs PC3.
11. Which event is highly correlated with PC3?
12. Calculate the model goodness of fit for each of the original variables for a model with one principal component (R1²X1… R1²X7) and plot this.
13. Calculate the model goodness of fit for each of the original variables for a model with two principal components (R2²X1… R2²X7) and plot this.
14. Calculate the model goodness of fit for each of the original variables for a model with three principal components (R3²X1… R3²X7) and plot this.
15. Using the plots produced for (m), (n) and (o), comment on how well each variable is explained by a PCA model with
    * 1. 1 principal component ()
      2. 2 principal components )
      3. 3 principal components )
16. An important issue in archaeology is the analysis of soil samples to determine their origin and age. It was proposed that the concentrations of trace elements found in soil samples vary significantly depending on the history of the soil. Samples taken from areas of undisturbed soil may exhibit different patterns of trace element composition to samples taken from areas where the soil has been modified by cultivation or dwelling.

A study was carried out on 22 soil samples where the concentrations of 9 trace elements were measured for each sample. The area that the soil sample was taken from was recorded as one of three categories:

Category A – on-site, specific feature e.g. hearth

Category B – on-site, general

Category C – off-site

The categories represent three levels of historical human activity.

A subset of the soil sample data is shown in the table below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sample** | **Fe** | **Cu** | **P** | **Mn** | **V** | **Co** | **Zn** | **Cr** | **Ca** |
| A | 2.21 | 36 | 0.914 | 950 | 50 | 7 | 130 | 34 | 2.26 |
| B | 1.98 | 27 | 0.337 | 786 | 45 | 7 | 110 | 29 | 1.38 |
| C | 2.08 | 64 | 1.436 | 1998 | 45 | 7 | 352 | 29 | 3.33 |

A Principal Components Analysis was carried out on the samples to determine whether the soil samples from the three different categories can be distinguished from one another and which variables are important for determining this. The analysis was carried out using **R**, and the output from the analysis is presented on the following pages.

**Table 1**: Summary of model

Importance of components:

PC1 PC2 PC3 PC4 PC5 Standard deviation 2.1115 2.0150 0.46992 0.39963 0.20699

Proportion of Variance 0.4954 0.4511 0.02454 0.01775 0.00476

Cumulative Proportion 0.4954 0.9465 0.97105 0.98879 0.99355

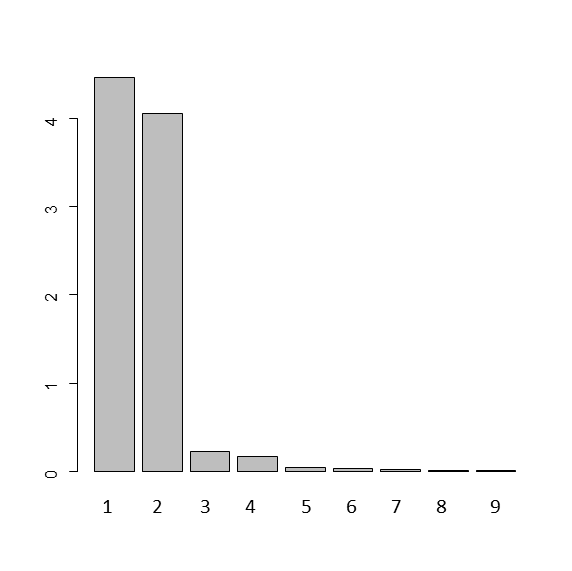
PC6 PC7 PC8 PC9

Standard deviation 0.16886 0.12487 0.10097 0.06114

Proportion of Variance 0.00317 0.00173 0.001130.00042

Cumulative Proportion 0.99672 0.99845 0.99958 1.00000

**Figure 1.** Scree plot from the PCA of the soil samples data set.



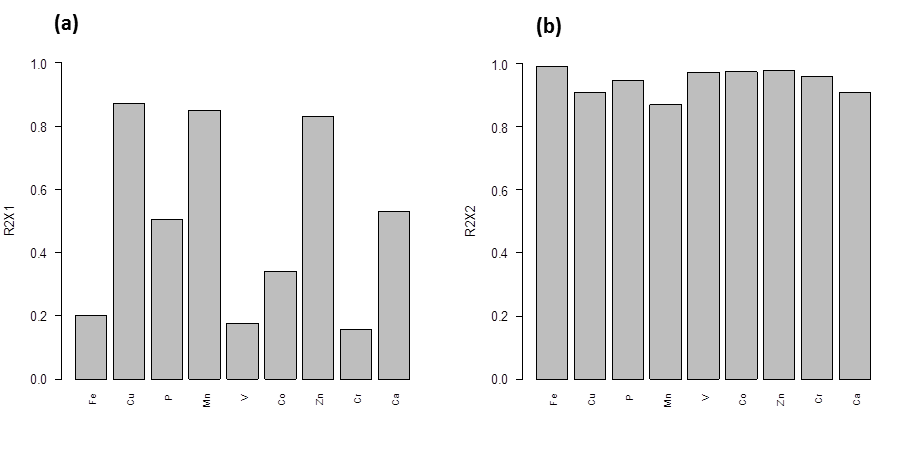
**Figure 2.** Scores plot for the soil samples



**Figure 3.** Loading plot for the soil samples



**Figures 4 (a) and (b).** Barcharts showing the fraction of the explained variation of the variables of the soil samples data set for a PCA model with 4(a) 1 principal component and 4(b) 2 principal components.



1. How many principal components are needed if you wish to account for 95% of variation in the data (please round the output data to 2 decimal places)?
2. Based on the output of the PCA, is it possible to distinguish between the soil samples? Please explain your answer, referencing the output you used to reach your decision.
3. Comment on the scores plot shown in Figure 2.
4. Which category of soil samples has comparatively high levels of the trace elements P, Ca, Zn, Cu and Mn?
5. Comment on the composition of trace elements found in soil samples that were subject to a high level of human activity in the past.
6. Which variables (trace elements) are well explained by the second principal component?