Problem Set 2: PCA

Stat 154, Spring 2018

Due date: Fri Feb-16 (before midnight)

The purpose of this assignment is to perform an exhaustive Principal Component Analysis (PCA) from scratch in R. This means that you will have to carry out the calculations WITHOUT using any of the exisiting functions—or packages—for PCA (e.g. can't use `prcomp()` or `princomp()`). In other words, you must perform all of the steps using either matrix operations or auxiliary functions such as `apply()`, `sweep()`, `scale()`, `crossprod()`, `eigen()` and/or `svd()`.

Use an R markdown (`.Rmd`) file to write your code and answers. You can *knit* the `Rmd` file as html or pdf. Please submit both your `Rmd` and knitted file to bCourses. Make sure to include your name, and your lab section. No late assignments will be accepted.

Data Set

In this problem, you will investigate the results of decathlon events using the data file `decathlon.csv`, available in the `data/` folder of the course github repository.

Data file

You can download the file to your working directory (or any other location) with `download.file()`:

```r
# download data to your working directory
# (do NOT include this code in your Rmd file)
repo <- 'https://raw.githubusercontent.com/ucb-stat154/'
file <- 'stat154-spring-2018/master/data/decathlon.csv'

download.file(
  url = paste0(repo, file),
  destfile = 'decathlon.csv'
)
```

Please get your own copy of the data file. This operation should not be part of the commands in your Rmd file. Otherwise, everytime you knit the Rmd file, you will (unnecessarily) be downloading the data file over and over.

Description

The Decathlon dataset contains the results of decathlon events during two athletic meetings which took place one month apart in 2004: 1) the Olympic Games in Athens which took place on 23 and 24 August, and 2) the Decastar 2004 which took place on 25 and 26 September.

For both competitions, the following information is available for each athlete: performance for each of the 10 events, total number of points (for each event, an athlete earns points based on performance; here the sum of points scored) and final ranking.

The events took place in the following order: 100 meters, long jump, shot put, high jump, 400 meters (first day) and 110 meter hurdles, discus, pole vault, javelin, 1500 meters (second day).

There are 12 quantitative variables (the results for the 10 events, the ranking of the athlete, and the total number of points earned) and one categorical variable (the competition in which the athlete took part).

The overall goal is to obtain a *typology* of the performance profiles based on the performances for each of the 10 events, such that two performance profiles might be as close as they are similar.

In addition, we want to obtain a review of the relationships between the results for the different events by studying the correlation coefficients between the variables taken pairwise.

Exploratory Phase (not graded)

- Before computing PCA outputs, start with an Exploratory Data Analysis (EDA). Get descriptive statistics for each variable; produce visualizations for each variable, maybe some scatterplots.
- You don't need to report all the results, summaries, and plots that you obtain in this phase. However, you should carry out some exploration to "get to know" the data better.
- Report two or three comprehensive graphs (e.g. maybe a `stars()` plot for the athletes, a correlogram of the pairwise correlations, or a scatterplot matrix with `pairs()`).
- Report descriptions (e.g. summary statistics) that show an *interesting* pattern (something unique) of the variables that catch your attention.
- Tell the reader what things are eye-catching, why it is important to know the specific details that you found in the EDA.

1) Calculation of primary PCA outputs (30 pts)

As we saw in lecture, the primary outputs of a PCA can be obtained via various approaches. Perhaps the two most common approaches consist of using either a Singular Value Decomposition (SVD) of the data matrix, or an Eigen-Value Decomposition (EVD) of one of the *sum of squares and cross products* (SSCP) matrices.

Regardless of the approach you decide to use, please keep in mind the following specs:

- Work with standardized data: mean = 0, sample variance = 1.
- Active individuals: the first 28 rows are the active individuals (those that competed in the Olympic Games);
- Active variables: the 10 events.
- Supplementary individuals: the rows 29 to 41 are supplementary individuals (those who competed in Decastar)
- Supplementary variables: `rank`, `points`, `competition`.
- Show your computations (do NOT use `results = 'hide'` or `echo = FALSE` or `eval = FALSE` in any of the code chunks in your Rmd file).

a. Obtain the loadings and store them in a matrix, include row and column names. Display the first four loadings (10 pts).

b. Obtain the principal components and store them in a matrix, include row and column names. Display the first four PCs (10 pts).

c. Obtain the eigenvalues and store them in a vector. Display the entire vector, and compute their sum. (10 pts)

2) Choosing the number of dimensions to retain/examine (30 pts)

a. Make a summary table of the eigenvalues: eigenvalue in the first column (each eigenvalue represents the variance captured by each component); percentage of variance in the second column; and cumulative percentage in the third column. Comment on the table. (10 pts)

b. Create a scree-plot (with axis labels) of the eigenvalues. What do you see? How do you read/interpret this chart? (10 pts)

c. If you had to choose a number of dimensions (i.e. a number of PCs), how many would you choose and why? (10 pts)

3) Studying the cloud of individuals (30 pts)

a. Create a scatter plot of the athletes on the 1st and 2nd PCs (10 pts).
    - In this plot, you should also project the supplementary individuals.
    - Make sure to add a visual cue (e.g. size, font, shape) to differentiate between active and supplementary individuals.
    - Color the individuals according to the variable `competition`.
    - Comment on general patterns, as well as on particular patterns.

b. Compute the quality of individuals representation, that is, the squared cosines given by:

$$cos^2(i,k) = \frac{z_{ik}^2}{d^2(\mathbf{x_i}, \mathbf{g})}$$

where:

- z_{ik} is the square value of the i-th individual on PC k
- $\mathbf{x_i}$ represents the row-vector of the i-th individual
- $\mathbf{g}$ is the centroid (i.e. average individual)

Store the squared cosines in a matrix or data frame, include row and column names. Display the first four columns. What athletes are best represented on the first two PCs? What athletes have the worst representation on the first two PCs? (10 pts).

c. Compute the contributions of the individuals to each extracted PC.

$$ctr(i,k) = \frac{m_i z_{ik}^2}{\lambda_k} \times 100$$

where:

- m_i is the mass or weight of individual i, in our case: $(\frac{1}{n-1})$
- z_{ik} is the value of k-th PC for individual i
- λ_k is the eigenvalue associated to k-th PC

Store the individuals contributions in a matrix or data frame, including row and column names. Display the first four columns. Based on the contributions, are there any influential athletes on the first two PCs? (10 pts).

4) Studying the cloud of variables (30 pts)

a. Calculate the correlation of all quantitative variables (active and supplementary) with the principal components. Store the correlations in a matrix or data frame, include row and column names. Display the first four columns. (10 pts)

b. Make a Circle of Correlations plot between the PCs and all the quantitative variables (10 pts).

   - For visualization purposes, include the circumference of a circle of radius one.
   - Represent each variable in the plot as an arrow.
   - Use color to distinguish between active and supplementary variables.
   - Also include names of variables.

c. Based on the above parts (a) and (b), how are the active and supplementary variables related to the components? (10 pts)

5) Conclusions (10 pts)

Write summarizing conclusions for the performed PCA.