Assignment 3: Principle component and factor analysis

Group 12

Dávid Hrabovszki (davhr856), Laura Julia Melis (lauju103), Spyridon Dimitriadis (spydi472), Vasileia Kampouraki (vaska979)

08/12/2019

Question 1: Principal components, including interpretation of them.

1.a.

The sample correlation matrix (R) is:

```
##             100m    200m    400m    800m   1500m   3000m marathon
## 100m     1.00000 0.94109 0.87078 0.80918 0.78155 0.72788  0.66896
## 200m     0.94109 1.00000 0.90881 0.81983 0.80133 0.73185  0.67995
## 400m     0.87078 0.90881 1.00000 0.80579 0.71980 0.67380  0.67694
## 800m     0.80918 0.81983 0.80579 1.00000 0.90505 0.86657  0.85399
## 1500m    0.78155 0.80133 0.71980 0.90505 1.00000 0.97338  0.79056
## 3000m    0.72788 0.73185 0.67380 0.86657 0.97338 1.00000  0.79873
## marathon 0.66896 0.67995 0.67694 0.85399 0.79056 0.79873  1.00000
```

The eigenvalues of R are:

```
##               100m    200m    400m    800m   1500m   3000m marathon
## Eigenvalue 5.80762 0.62869 0.27933 0.12455 0.09097 0.05452   0.0143
```

And its eigenvectors:

```
##         e1       e2       e3       e4       e5       e6       e7
## 1 -0.37777 -0.40718 -0.14058  0.58706 -0.16707  0.53970  0.08894
## 2 -0.38321 -0.41363 -0.10078  0.19408  0.09350 -0.74493 -0.26566
## 3 -0.36804 -0.45935  0.23703 -0.64543  0.32727  0.24009  0.12660
## 4 -0.39478  0.16125  0.14754 -0.29521 -0.81905 -0.01651 -0.19521
## 5 -0.38926  0.30909 -0.42199 -0.06669  0.02613 -0.18899  0.73077
## 6 -0.37609  0.42319 -0.40606 -0.08016  0.35170  0.24050 -0.57151
## 7 -0.35520  0.38922  0.74106  0.32108  0.24701 -0.04827  0.08208
```

1.b.

First two principal components

In general, the i-th principal component is given by

$$\hat{y}_i = \hat{e}_i' z = \hat{e}_{i1}' z_1 + \hat{e}_{i2}' z_2 + \cdots + \hat{e}_{ip}' z_p \quad i = 1, 2, \ldots, p.$$

where z_i are the standardized variables.

So given the eigenvectors e_1 and e_2 obtained in part (a), the first two principal components for the standardized variables are:

$$\hat{y}_1 = \hat{e}_1' z = -0.37777 z_1 - 0.38321 z_2 - 0.36804 z_3 - 0.39478 z_4 - 0.38926 z_5 - 0.37609 z_6 - 0.35520 z_7.$$
$$\hat{y}_2 = \hat{e}_2' z = -0.40718 z_1 - 0.41363 z_2 - 0.45935 z_3 + 0.16125 z_4 + 0.30909 z_5 + 0.42319 z_6 + 0.38922 z_7.$$

Correlations of the standardized variables with the components

The correlation coefficients between the components Y_i and the standarized variables z_i are given by

$$r_{\hat{y}_i, z_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

In these data we have the following coefficients:

```
##          100m     200m     400m     800m    1500m    3000m marathon
## r_Y1 -2.31001 -1.05622 -0.13044 16.28542 -1.47824  1.59521  0.01304
## r_Y2 -0.77099 -0.35302 -0.03077  1.77135  0.27220 -0.72444 -0.01281
```

Cumulative percentage of the total (standardized) sample variance.

The percentage of the total sample variance due to the k-th principal component is given by

$$\left(\frac{\hat{\lambda}_k}{\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p} \right) \cdot 100\%, \quad k = 1, 2, \ldots, p$$

So, the total sample variance in the first two components are:

```
##                   PC1       PC2
## percentage 82.96606  8.981335
## cumulative 82.96606 91.947398
```

The percentage of the total standarized sample variance explained by the first two principal components is 91.95%. Then, as the majority of the total sample variance is attributed to these first two components, and in our case, if we remove the other variables, we won't lose much information.

1.c.

The projections in PC1 are more or less the same (around 0.3) which means that all the variables contribute almost the same. So, PC1 might measure the athletic excellence of a given nation because all the variables contribute the same.

Regarding the projections in PC2, the first three elements (100m, 200m and 400m) of the PC2 have the smallest values, so they contribute more to the second principal component. Because of this, we can interpret it as it captures how a nation performes in short distance runtypes.

1.d.

In order to rank the countries based on the first principal component ($\hat{y}_1$) first we need the standarize our observations:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}} \quad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, p$$

And then, we only have to compute the following formula, replacing the z_j values with the standarized observations:

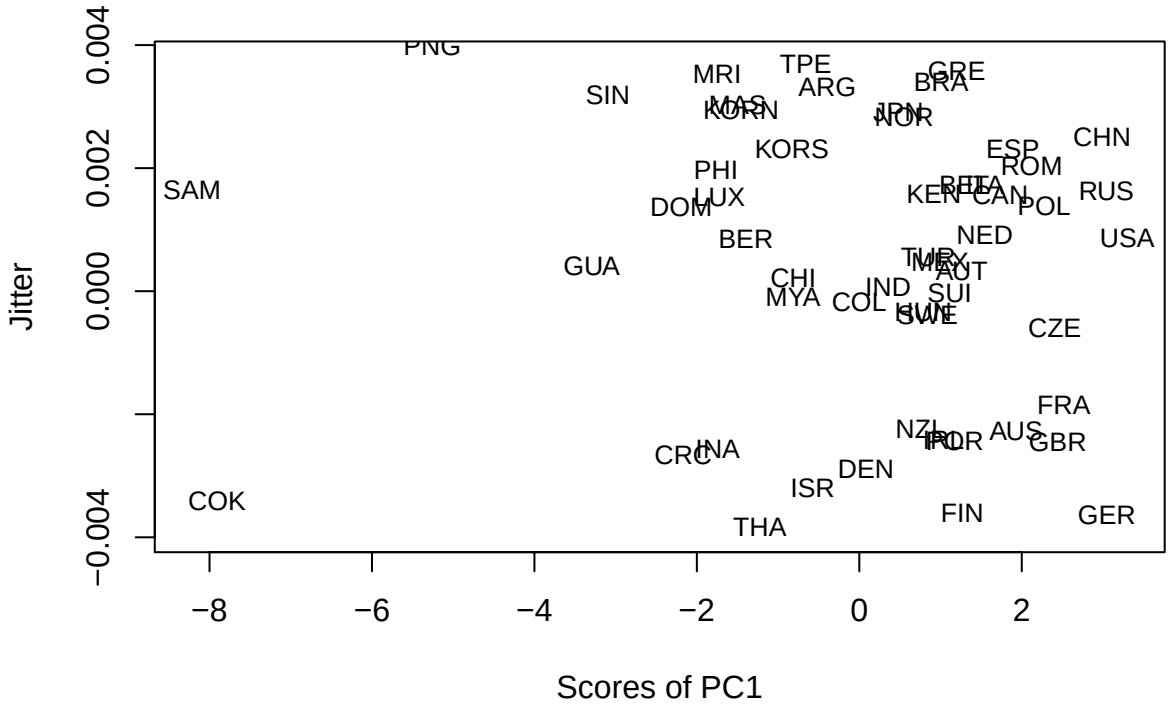
$$\hat{y}_1 = \hat{e}_1' z = -0.37777 z_1 - 0.38321 z_2 - 0.36804 z_3 - 0.39478 z_4 - 0.38926 z_5 - 0.37609 z_6 - 0.35520 z_7.$$

.

As a result, the 5 contries with highest scores are:

```
##      USA      GER      RUS      CHN      FRA
## 3.299149 3.047517 3.042948 2.989467 2.518346
```

This ranking corresponds with the group of countries that perform better in all runtypes.



NOTE: In the above plot the x-axis represent the scores of the PC1 and the y-axis is just the jitter of this points.

Question 2: Factor analysis.

Solve Exercise 9.28 of Johnson, Wichern, the same data as above. Try both PC and ML as estimation methods. Notice that R's factanal() only does ML estimation. For the PC method you can use the principal() function of the psych package. What does it mean that the parameter rotation of factanal() is set to "varimax" by default (equivalently rotate of principal())? Do not forget to check the adequacy of your model

Tip: Read section "A Large Sample Test for the Number of Common Factors".

EXERCISE: Perform a factor analysis of the national track records for women given in Table 1.9. Use the sample covariance matrix S and interpret the factors. Compute factor scores, and check for outliers in the data. Repeat the analysis with the sample correlation matrix R. Does it make a difference if R, rather than S, is factored? Explain.

- Covariance matrix:

```
##                100m       200m       400m        800m      1500m
## 100m     0.15531572  0.3445608  0.8912960 0.027703564 0.08389119
## 200m     0.34456080  0.8630883  2.1928363 0.066165898 0.20276331
## 400m     0.89129602  2.1928363  6.7454576 0.181807932 0.50917683
## 800m     0.02770356  0.0661659  0.1818079 0.007546925 0.02141457
## 1500m    0.08389119  0.2027633  0.5091768 0.021414570 0.07418270
## 3000m    0.23388281  0.5543502  1.4268158 0.061379315 0.21615514
## marathon 4.33417757 10.3849876 28.9037314 1.219654647 3.53983732
##                3000m   marathon
## 100m      0.23388281   4.334178
## 200m      0.55435017  10.384988
## 400m      1.42681579  28.903731
## 800m      0.06137932   1.219655
## 1500m     0.21615514   3.539837
## 3000m     0.66475793  10.706091
## marathon 10.70609113 270.270150
```

Appendix