Multivariate Statistical Methods - Lab 03

Lakshidaa Saigiridharan (laksa656)

12/8/2019

```r
# Loading required packages
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

```r
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.5.2
```

Question 1: Principal components, including interpretation of them

Solve Exercise 8.18 of Johnson, Wichern. The data on the national track records for women, which you have studied earlier, can be found in the file T1-9.dat.

```r
# Loading the required data
track_data <- read.table("T1-9.dat")
colnames(track_data) <- c("Country", "100", "200", "400", "800",
                          "1500", "3000", "Marathon")
head(track_data)
```

```
##   Country   100   200   400  800 1500 3000 Marathon
## 1     ARG 11.57 22.94 52.50 2.05 4.25 9.19   150.32
## 2     AUS 11.12 22.23 48.63 1.98 4.02 8.63   143.51
## 3     AUT 11.15 22.70 50.62 1.94 4.05 8.78   154.35
## 4     BEL 11.14 22.48 51.45 1.97 4.08 8.82   143.05
## 5     BER 11.46 23.05 53.30 2.07 4.29 9.81   174.18
## 6     BRA 11.17 22.60 50.62 1.97 4.17 9.04   147.41
```

(a) Obtain the sample correlation matrix R for these data, and determine its eigenvalues and eigenvectors.

```r
# Obtaining sample correlation matrix R
R <- cor(track_data[,2:8])

cat("Sample Correlation Matrix R : \n")
```

```
## Sample Correlation Matrix R :
```

```r
R
```

```
##                100       200       400       800      1500      3000
## 100      1.0000000 0.9410886 0.8707802 0.8091758 0.7815510 0.7278784
## 200      0.9410886 1.0000000 0.9088096 0.8198258 0.8013282 0.7318546
## 400      0.8707802 0.9088096 1.0000000 0.8057904 0.7197996 0.6737991
## 800      0.8091758 0.8198258 0.8057904 1.0000000 0.9050509 0.8665732
## 1500     0.7815510 0.8013282 0.7197996 0.9050509 1.0000000 0.9733801
```

```
## 3000     0.7278784 0.7318546 0.6737991 0.8665732 0.9733801 1.0000000
## Marathon 0.6689597 0.6799537 0.6769384 0.8539900 0.7905565 0.7987302
##           Marathon
## 100      0.6689597
## 200      0.6799537
## 400      0.6769384
## 800      0.8539900
## 1500     0.7905565
## 3000     0.7987302
## Marathon 1.0000000
```

```
# Obtaining eigenvalues and eigenvectors of R
eigen_values <- eigen(R)$values
eigen_vectors <- eigen(R)$vectors

cat("\nEigen values of R : \n")
```

```
##
## Eigen values of R :
```

```
eigen_values
```

```
## [1] 5.80762446 0.62869342 0.27933457 0.12455472 0.09097174 0.05451882
## [7] 0.01430226
```

```
cat("\nEigen vectors of R : \n")
```

```
##
## Eigen vectors of R :
```

```
eigen_vectors
```

```
##            [,1]       [,2]       [,3]        [,4]        [,5]        [,6]
## [1,] -0.3777657 -0.4071756 -0.1405803  0.58706293 -0.16706891  0.53969730
## [2,] -0.3832103 -0.4136291 -0.1007833  0.19407501  0.09350016 -0.74493139
## [3,] -0.3680361 -0.4593531  0.2370255 -0.64543118  0.32727328  0.24009405
## [4,] -0.3947810  0.1612459  0.1475424 -0.29520804 -0.81905467 -0.01650651
## [5,] -0.3892610  0.3090877 -0.4219855 -0.06669044  0.02613100 -0.18898771
## [6,] -0.3760945  0.4231899 -0.4060627 -0.08015699  0.35169796  0.24049968
## [7,] -0.3552031  0.3892153  0.7410610  0.32107640  0.24700821 -0.04826992
##             [,7]
## [1,]  0.08893934
## [2,] -0.26565662
## [3,]  0.12660435
## [4,] -0.19521315
## [5,]  0.73076817
## [6,] -0.57150644
## [7,]  0.08208401
```

(b) Determine the first two principal components for the standardized variables. Prepare a table showing the correlations of the standardized variables with the compo- nents, and the cumulative percentage of the total (standardized) sample variance explained by the two components.

```r
# Standardizing the data
track <- track_data[,2:8]
track_data_sd <- sapply(track, sd)
track_data_mean <- colMeans(track)
stand_track_data <- matrix(nrow=54, ncol=7) #standardized variables

for(i in 1:7){
  stand_track_data[,i] <- (track[,i] - track_data_mean[i])/track_data_sd[i]
}

colnames(stand_track_data)<-colnames(track)

# Sample correlation matrix of standardized data
R_stand <- cor(stand_track_data)
R_stand_eigen <- eigen(R_stand)

# Principal components
PC1 <- t(R_stand_eigen$vectors[,1] %*% t(stand_track_data))
PC2 <- t(R_stand_eigen$vectors[,2] %*% t(stand_track_data))

pc_table <- data.frame("PrincipalComponent1"=PC1,
                       "PrincipalComponent2"=PC2)

# Proportion of total variance for ith PC
proportion_var12 <- (sum(R_stand_eigen$values[1:2]) / 7)*100

# Table showing the correlations of the standardized variables
kable(R_stand) %>%
  kable_styling(bootstrap_options = c("striped", "hover")) %>%
  add_header_above(c(" ", "Correlation of Standardized Variable" = 7))
```

	Correlation of Standardized Variable						
	100	200	400	800	1500	3000	Marathon
100	1.0000000	0.9410886	0.8707802	0.8091758	0.7815510	0.7278784	0.6689597
200	0.9410886	1.0000000	0.9088096	0.8198258	0.8013282	0.7318546	0.6799537
400	0.8707802	0.9088096	1.0000000	0.8057904	0.7197996	0.6737991	0.6769384
800	0.8091758	0.8198258	0.8057904	1.0000000	0.9050509	0.8665732	0.8539900
1500	0.7815510	0.8013282	0.7197996	0.9050509	1.0000000	0.9733801	0.7905565
3000	0.7278784	0.7318546	0.6737991	0.8665732	0.9733801	1.0000000	0.7987302
Marathon	0.6689597	0.6799537	0.6769384	0.8539900	0.7905565	0.7987302	1.0000000

```r
# Table showing first two principal components of the standardized variables
kable(pc_table) %>%
  kable_styling(bootstrap_options = c("striped", "hover", position = "left"))
```

PrincipalComponent1	PrincipalComponent2
-0.3932402	-0.1316107
1.9316429	0.4910673
1.2625204	0.1931484
1.2917303	-0.0024053
-1.3961086	0.7607806
1.0067789	0.3795169
1.7343406	0.2625383
-0.8118382	-0.8689690
2.9894669	0.0515565
-0.0019277	0.9440511
-7.9062272	-0.5205487
-2.1668115	0.3329829
2.4060303	0.7596584
0.0824955	-0.7134670
-2.1924098	0.4313474
1.2667313	0.4263465
2.5183457	1.1230568
3.0475166	0.9345293
2.4427063	-0.0333740
1.1978004	0.7754294
-3.2941238	-0.5291973
0.7882511	-0.5905189
-1.7419421	-0.5146703
0.3542566	0.2542125
1.0359072	-0.7726532
-0.5741617	0.2181300
1.5474528	-0.2725522
0.4816576	-0.6557135
0.9177354	-1.3818382
-0.8307946	-0.7687521
-1.4553473	-2.3771213
-1.7214677	-1.2782741
-1.4952101	0.5386191
-1.7497278	-0.5254636
0.9957663	0.4905095
-0.8159815	-0.5990664
1.5447606	-0.2873591
0.7552355	-0.4320195
0.5530035	-0.9934747
-5.2574497	1.1953938
-1.7635337	0.5797417
2.2737658	0.4911614
1.1752500	-0.7069616
2.1230057	-0.3810120
3.0429482	0.4460682
-8.2134151	2.0282582
-3.0939195	-0.9564211
1.8894623	0.2470325
0.8391496	0.0001607
1.1135452	-0.5263586
-0.6590931	1.0063775
-1.2238050	0.8469873
0.8501278	-0.5785810
3.2991488	1.1897213

```r
cat(paste("Cumulative percentage of the total(standardized)
          sample variance explained by the two components = ",
          proportion_var12, "%"))
```

```
## Cumulative percentage of the total(standardized)
##           sample variance explained by the two components =  91.9473983845876 %
```

(c) Interpret the two principal components obtained in Part b. (Note that the first component is essentially a normalized unit vector and might measure the athlet- ic excellence of a given nation. The second component might measure the relative strength of a nation at the various running distances.)

From the results obtained in Part (b), it can be seen that all events contribute equally to the first component which measures the track index of the data. The second component exhibits a clear difference between the shorter distance track events (i.e., 100m, 200m, 400m) and the longer distance track events (i.e., 800m, 1500m, 3000m, Marathon).

(d) Rank the nations based on their score on the first principal component. Does this ranking correspond with your inituitive notion of athletic excellence for the various countries?

```r
rank <- order(PC1, decreasing = TRUE)
ranked_countries <- track_data$Country[rank]
ranked_countries
```

```
##  [1] USA  GER  RUS  CHN  FRA  GBR  CZE  POL  ROM  AUS  ESP  CAN  ITA  NED
## [15] BEL  FIN  AUT  GRE  POR  SUI  IRL  BRA  MEX  KEN  TUR  SWE  HUN  NZL
## [29] NOR  JPN  IND  DEN  COL  ARG  ISR  TPE  CHI  MYA  KORS THA  BER  KORN
## [43] MAS  LUX  INA  MRI  PHI  CRC  DOM  SIN  GUA  PNG  COK  SAM
## 54 Levels: ARG AUS AUT BEL BER BRA CAN CHI CHN COK COL CRC CZE DEN ... USA
```