Lab 3 732A97

Raymond Sseguya

2019-12-08

Data

```
library(dplyr)

trackrcs <- read.table("T1-9.dat",
  col.names = c("countries", "x100m", "x200m",
        "x400m", "x800m", "x1500m", "x3000m", "marathon"))
rownames(trackrcs) <- trackrcs[,1]
```

Question 1: Principal components, including interpretation of them

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

```
S <- cov(trackrcs[,-1])
R <- cov2cor(S); R
```

```
##              x100m     x200m     x400m     x800m    x1500m    x3000m
## x100m    1.0000000 0.9410886 0.8707802 0.8091758 0.7815510 0.7278784
## x200m    0.9410886 1.0000000 0.9088096 0.8198258 0.8013282 0.7318546
## x400m    0.8707802 0.9088096 1.0000000 0.8057904 0.7197996 0.6737991
## x800m    0.8091758 0.8198258 0.8057904 1.0000000 0.9050509 0.8665732
## x1500m   0.7815510 0.8013282 0.7197996 0.9050509 1.0000000 0.9733801
## x3000m   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
## x100m    0.6689597
## x200m    0.6799537
## x400m    0.6769384
## x800m    0.8539900
## x1500m   0.7905565
## x3000m   0.7987302
## marathon 1.0000000
```

```
eigen(R)$values
```

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

```r
eigen(R)$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 components, and the cumulative percentage of the total (standardized) sample variance explained by the two components.

the first two principal components for the standardized variables

```r
res=prcomp((trackrcs)[,-1], scale. = FALSE)
# No sacling at this point because we are going to use the correlation matrix later

# Each PC is a linear combination of the original variables
#### res$rotation
res$rotation[,1:2]
```

```
##                  PC1         PC2
## x100m    -0.016123307  0.11485619
## x200m    -0.038657909  0.29039299
## x400m    -0.107793074  0.93844399
## x800m    -0.004504024  0.01340703
## x1500m   -0.013072642  0.03631915
## x3000m   -0.039484872  0.07871002
## marathon -0.992409201 -0.11878027
```

correlation of the standardized variables with the components

This is following the formula in the text book on page 433, "Result 8.3" and "Equation 8-8"

```
eigenvalues=res$sdev^2
CorWithPC <-
  t(res$rotation[,1:2])%*%sqrt(diag(eigenvalues))%*%solve(sqrt(diag(diag(R))))
colnames(CorWithPC) <- colnames(trackrcs[,-1])
t(CorWithPC)
```

```
##                   PC1          PC2
## x100m    -0.267064796  1.902466031
## x200m    -0.077476875  0.581995811
## x400m    -0.055829878  0.486053614
## x800m    -0.001524372  0.004537567
## x1500m   -0.001607685  0.004466563
## x3000m   -0.002011251  0.004009272
## marathon -0.024646395 -0.002949898
```

cummulative percentage of total standardized sample variance explained by the 2 components

```
CorWithPC %>% apply(MARGIN=1,FUN=abs) %>% t() %>%
  apply(MARGIN=1,FUN=function(a) 100*cumsum(a)/sum(a))
```

```
##                PC1       PC2
## x100m     62.08481  63.70265
## x200m     80.09593  83.19034
## x400m     93.07476  99.46548
## x800m     93.42913  99.61742
## x1500m    93.80287  99.76698
## x3000m    94.27043  99.90122
## marathon 100.00000 100.00000
```

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 athletic excellence of a given nation. The second component might measure the relative strength of a nation at various running distances.)

It seems the first principal component is a meausure of how much less time the atheletes of a particular nation take to complete a race relative to their other fellow competitiors from other nations. We can notice that athletes generally concentrate a lot of efforts in running very fast so that they complete their races in the shortest times possible in the 100 metre races and the marathon.

Clearly, the second principal component tells us about the physical strength of an athlete in a given race. We see that for 100 meters, the values are very high because athletes are giving their all while for the marathan, the values are actually negative because the athletes are usually very tired and they try to minimize using a lot of energy as a tactic to run for the very long distance.

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?

```
# CorWithPC2 <- cor( t(scale((trackrcs)[,-1])),  res$rotation[,1:2] )
NewScore <- as.matrix(trackrcs[,-1])%*%as.matrix(res$rotation[,1])
NewScore = cbind.data.frame(countries = trackrcs[,1], NewScore = NewScore)
NewScore[,1][order(NewScore[,2], decreasing = TRUE)]
```

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

This ranking indeed corresponds with my inituitive notion of the athletic excellence for the various countries. The principal components are indeed capturing well the ranking of the countries.

Question 2: Factor analysis

```r
factanal(trackrcs[,-1], factors = 3, covmat = S) # varimax is the default
```

```
##
## Call:
## factanal(x = trackrcs[, -1], factors = 3, covmat = S)
##
## Uniquenesses:
##    x100m    x200m    x400m    x800m   x1500m   x3000m marathon
##    0.106    0.005    0.133    0.047    0.005    0.041    0.225
##
## Loadings:
##          Factor1 Factor2 Factor3
## x100m    0.815   0.413   0.245
## x200m    0.886   0.410   0.203
## x400m    0.797   0.311   0.367
## x800m    0.512   0.617   0.556
## x1500m   0.449   0.849   0.270
## x3000m   0.361   0.866   0.280
## marathon 0.380   0.553   0.571
##
##                Factor1 Factor2 Factor3
## SS loadings      2.824   2.593   1.022
## Proportion Var   0.403   0.370   0.146
## Cumulative Var   0.403   0.774   0.920
##
## The degrees of freedom for the model is 3 and the fit was 0.2033
```

```r
factanal(trackrcs[,-1], factors = 3, covmat = R)
```

```
##
## Call:
## factanal(x = trackrcs[, -1], factors = 3, covmat = R)
##
## Uniquenesses:
##    x100m    x200m    x400m    x800m   x1500m   x3000m marathon
##    0.106    0.005    0.133    0.047    0.005    0.041    0.225
##
## Loadings:
##          Factor1 Factor2 Factor3
## x100m    0.815   0.413   0.245
## x200m    0.886   0.410   0.203
## x400m    0.797   0.311   0.367
## x800m    0.512   0.617   0.556
## x1500m   0.449   0.849   0.270
## x3000m   0.361   0.866   0.280
## marathon 0.380   0.553   0.571
##
##                Factor1 Factor2 Factor3
## SS loadings      2.824   2.593   1.022
```

```
## Proportion Var   0.403   0.370   0.146
## Cumulative Var   0.403   0.774   0.920
##
## The degrees of freedom for the model is 3 and the fit was 0.2033
```

```
psych::principal(cov2cor(S), nfactors=3, rotate="varimax")
```

```
## Principal Components Analysis
## Call: psych::principal(r = cov2cor(S), nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           RC2  RC1  RC3   h2    u2 com
## x100m    0.85 0.41 0.23 0.94 0.061 1.6
## x200m    0.86 0.40 0.25 0.96 0.037 1.6
## x400m    0.86 0.26 0.36 0.93 0.065 1.5
## x800m    0.54 0.59 0.54 0.93 0.072 3.0
## x1500m   0.44 0.82 0.34 0.99 0.010 1.9
## x3000m   0.35 0.85 0.37 0.98 0.020 1.7
## marathon 0.33 0.44 0.82 0.98 0.019 1.9
##
##                        RC2  RC1  RC3
## SS loadings           2.92 2.33 1.47
## Proportion Var        0.42 0.33 0.21
## Cumulative Var        0.42 0.75 0.96
## Proportion Explained  0.43 0.35 0.22
## Cumulative Proportion 0.43 0.78 1.00
##
## Mean item complexity =  1.9
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.02
##
## Fit based upon off diagonal values = 1
```

```
psych::fa(cov2cor(S), nfactors=3, rotate="varimax")
```

```
## Factor Analysis using method =  minres
## Call: psych::fa(r = cov2cor(S), nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##           MR2  MR3  MR1   h2     u2 com
## x100m    0.83 0.41 0.23 0.90 0.0993 1.6
## x200m    0.88 0.40 0.21 0.98 0.0160 1.5
## x400m    0.80 0.31 0.35 0.87 0.1338 1.7
## x800m    0.53 0.60 0.54 0.94 0.0622 3.0
## x1500m   0.46 0.85 0.26 1.00 0.0018 1.8
## x3000m   0.38 0.85 0.30 0.95 0.0457 1.7
## marathon 0.37 0.56 0.59 0.80 0.2002 2.7
##
##                        MR2  MR3  MR1
## SS loadings           2.88 2.54 1.03
## Proportion Var        0.41 0.36 0.15
## Cumulative Var        0.41 0.77 0.92
## Proportion Explained  0.45 0.39 0.16
## Cumulative Proportion 0.45 0.84 1.00
```

```
##
## Mean item complexity =  2
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are  21  and the objective function was  11.62
## The degrees of freedom for the model are 3  and the objective function was  0.23
##
## The root mean square of the residuals (RMSR) is  0
## The df corrected root mean square of the residuals is  0.01
##
## Fit based upon off diagonal values = 1
```