Computer Lab 4

Thomas Zhang

2015-11-26

Assignment 1

This is the sample correlation matrix for the womens national track records data provided, its eigenvalues and its eigenvectors.

```
##              100m(s)   200m(s)   400m(s) 800m(min) 1500m(min) 3000m(min)
## 100m(s)    1.0000000 0.9410886 0.8707802 0.8091758  0.7815510  0.7278784
## 200m(s)    0.9410886 1.0000000 0.9088096 0.8198258  0.8013282  0.7318546
## 400m(s)    0.8707802 0.9088096 1.0000000 0.8057904  0.7197996  0.6737991
## 800m(min)  0.8091758 0.8198258 0.8057904 1.0000000  0.9050509  0.8665732
## 1500m(min) 0.7815510 0.8013282 0.7197996 0.9050509  1.0000000  0.9733801
## 3000m(min) 0.7278784 0.7318546 0.6737991 0.8665732  0.9733801  1.0000000
## Mara(min)  0.6689597 0.6799537 0.6769384 0.8539900  0.7905565  0.7987302
##            Mara(min)
## 100m(s)    0.6689597
## 200m(s)    0.6799537
## 400m(s)    0.6769384
## 800m(min)  0.8539900
## 1500m(min) 0.7905565
## 3000m(min) 0.7987302
## Mara(min)  1.0000000
```

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

We perform a principal component analysis (PCA) on this data. Particularly, we are interested in the first two principal components (PCs). The first two PCs and their correlations with the seven standardized data variables (running distances) and the percentage of total sample variance explained by these two PCs, each and cumulatively, are shown below.

```
##                    PC1        PC2
## 100m(s)    -0.3777657  0.4071756
## 200m(s)    -0.3832103  0.4136291
## 400m(s)    -0.3680361  0.4593531
## 800m(min)  -0.3947810 -0.1612459
## 1500m(min) -0.3892610 -0.3090877
## 3000m(min) -0.3760945 -0.4231899
## Mara(min)  -0.3552031 -0.3892153
```

```
## [1] "Correlations between the two first PCs and the standardized data variables:"
```

```
##        100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
## PC1 -0.9103780 -0.9234990 -0.8869307 -0.9513832 -0.9380805 -0.9063506
## PC2  0.3228503  0.3279673  0.3642220 -0.1278522 -0.2450762 -0.3355481
##      Mara(min)
## PC1 -0.8560043
## PC2 -0.3086096
```

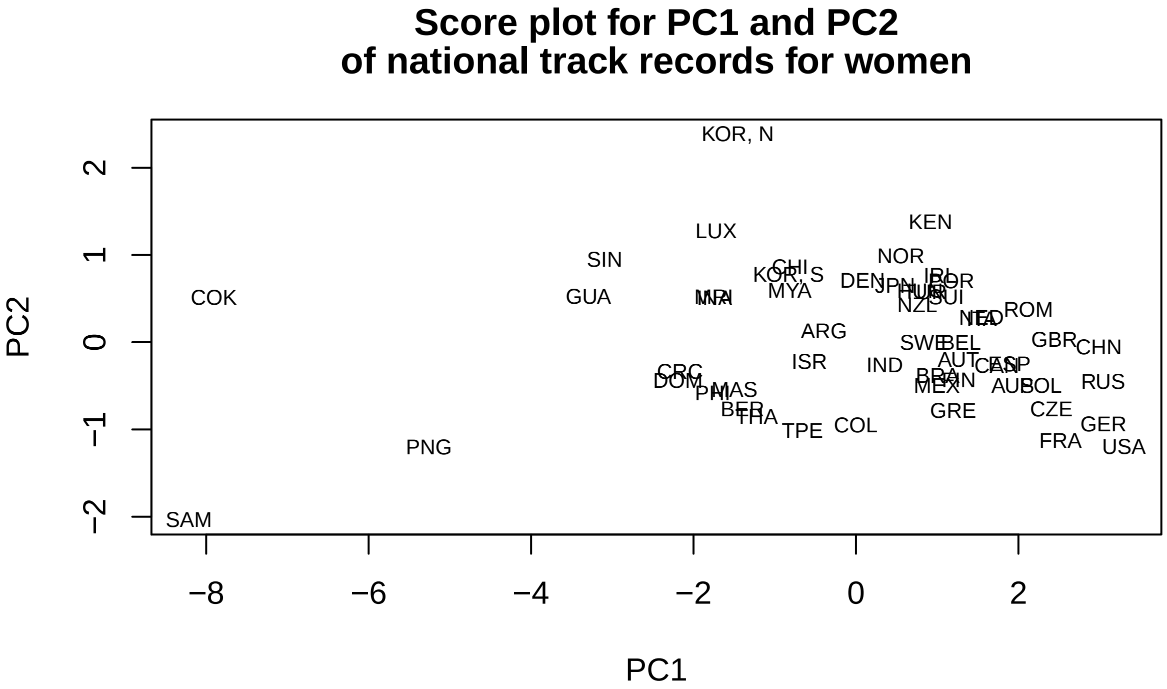
```
## [1] "Percent of total sample variance explained by first PC: 82.966"
```

```
## [1] "Percent of total sample variance explained by second PC: 8.981"
```

```
## [1] "Cumulative Percent of total sample variance explained by first two PCs: 91.947"
```

We see that more than 90% of total sample variance can be explained by the first two PCs. We also see that the first PC correlates with all standardized data variables equally, and can be seen as a measure of speediness (or fastidity) of a nations runners. In other words, the athletic excellence of a nation is measured by the first PC. The second PC has correlations with data which favors good long distance running records and penalizes good short distance running records. This can be interpreted as a measure of how much better a country is at long distance running than at short distance running.

Let us plot the PC1-PC2 score plot of the nations and find out which nations are the most excellent in terms of womens track records (highest PC1 value).



It looks as if big countries, such as USA,Germany,Russia,China and France produce the most excellent female runners. This conclusion coincides with intuitive explanations that factors such as the size of the talent pool, the athletics programmes available and resources available leads to good results for these large and developed countries.

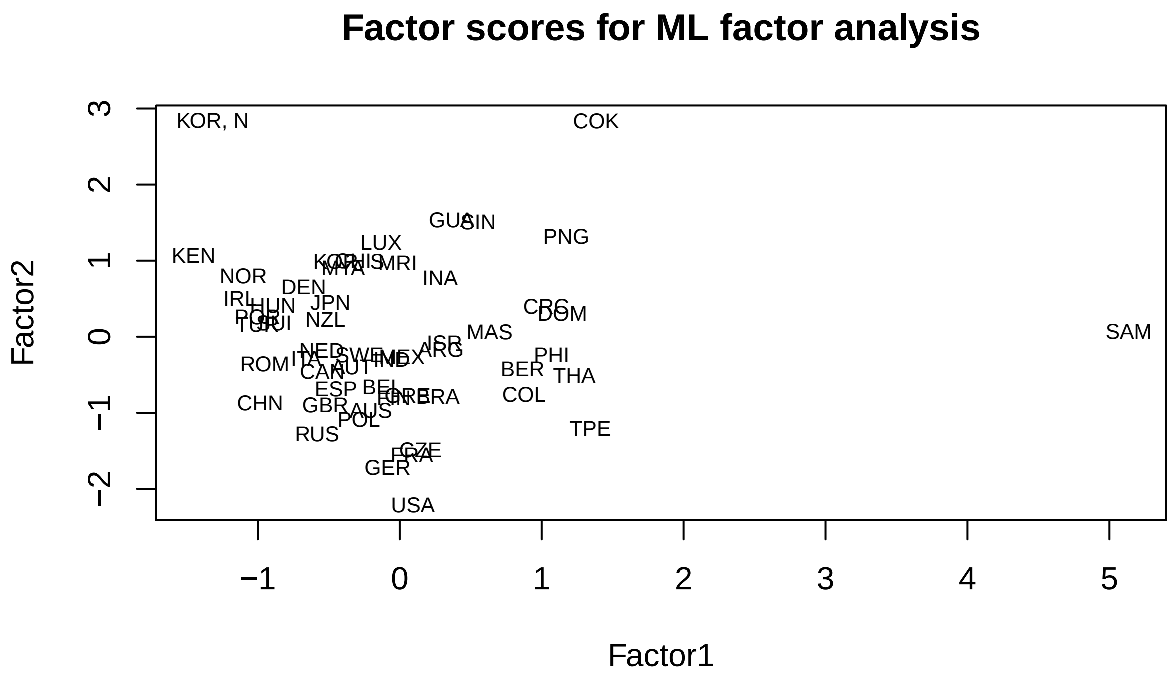
Assignment 2

Maximum Liklihood method

We want to create a factor model for the very same data set. Let us start out with factor analysis of the data using the Maximum Likelihood method used in the built-in R function `factanal`. Like in PCA, we want two factors which can help us explain the data sample variance.

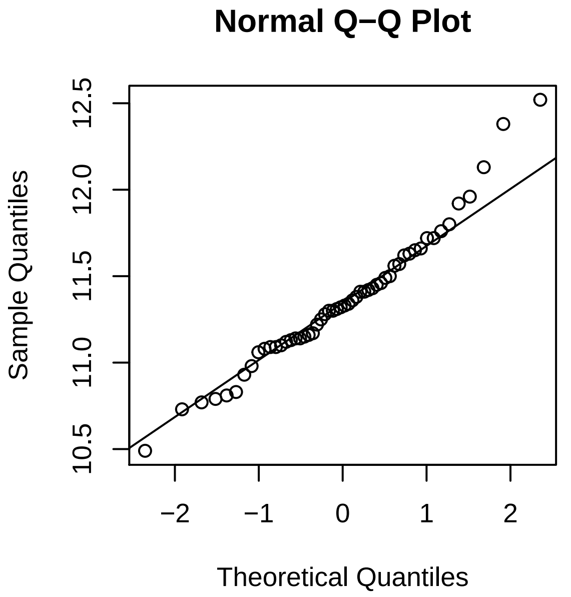
```
##
## Call:
## factanal(x = data, factors = 2, scores = "Bartlett")
##
## Uniquenesses:
##    100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
##      0.094      0.024      0.152      0.144      0.016      0.028
##  Mara(min)
##      0.338
##
## Loadings:
##            Factor1 Factor2
## 100m(s)    0.461   0.833
## 200m(s)    0.455   0.877
## 400m(s)    0.401   0.829
## 800m(min)  0.732   0.566
## 1500m(min) 0.882   0.454
## 3000m(min) 0.918   0.361
## Mara(min)  0.693   0.427
##
##                Factor1 Factor2
## SS loadings      3.216   2.987
## Proportion Var   0.459   0.427
## Cumulative Var   0.459   0.886
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 31.43 on 8 degrees of freedom.
## The p-value is 0.000118
```

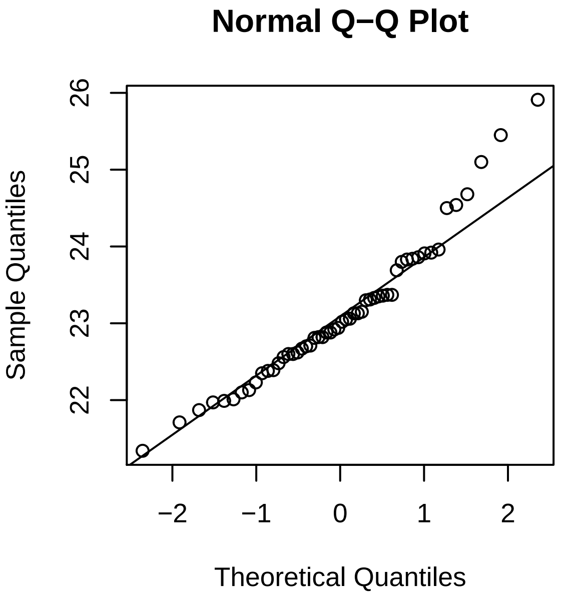
The loadings tell us that Factor 1 focuses on performance records of longer distances while Factor 2 focuses on performance records of shorter distances. This long-short distance performance focus appears almost symmetric, and this could be why we notice that the two factors explain about the same fraction of the sample variance. Together they explain about 90% of sample variance. We plot the factor scores (obtained by the weighted least squares method) and see if there are any outliers.

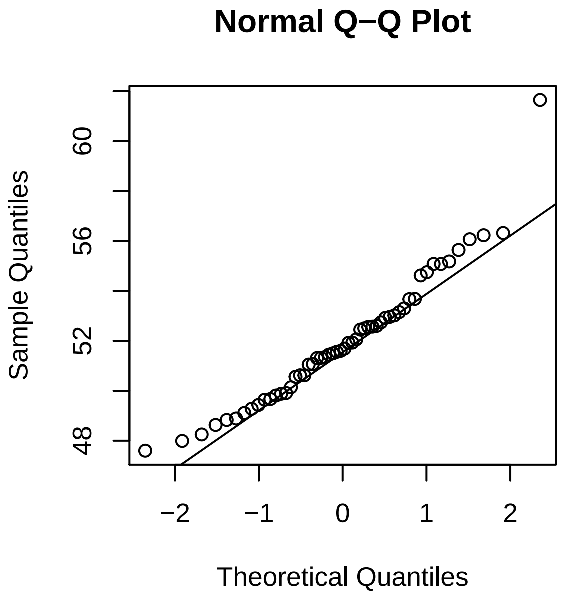


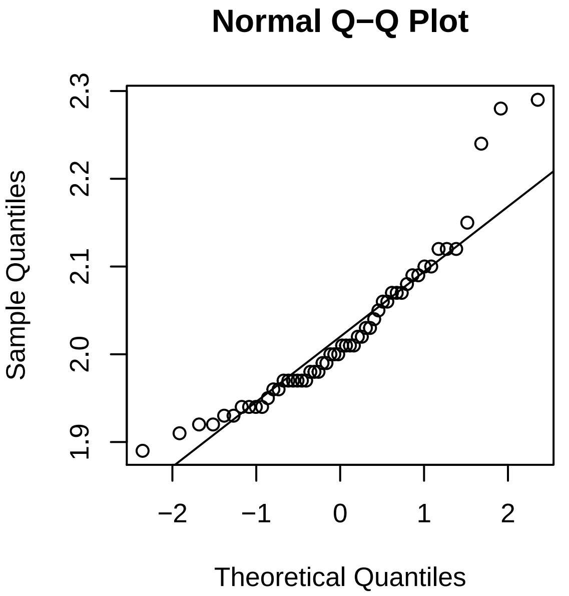
Yes, it appears Samoa is a big outlier in factor 1, which means that Samoa has bad performance in longer distance records. Analogously, North Korea and Cook islands are outliers in factor 2, and thus are bad at shorter distance records.

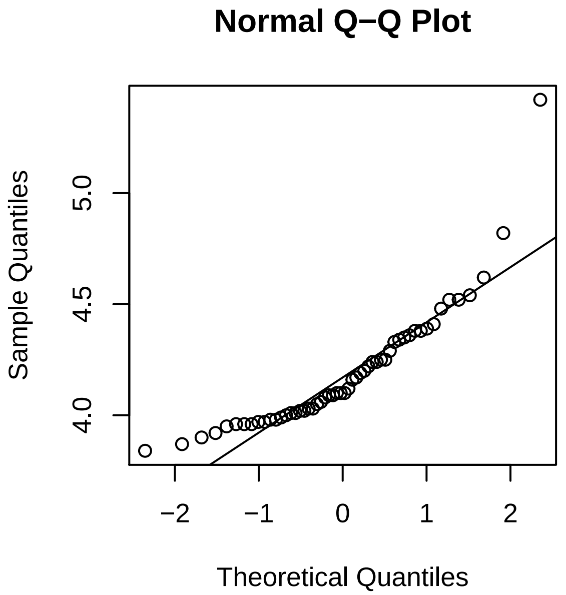
Maximum Liklihood estimation of loadings is adequate whenever the data is approximately multivariate normally distributed. A few quick Q-Q plots indicate that that is the case with womens national track records data, especially for the shorter distances.(corresponding to the first Q-Q-plots)

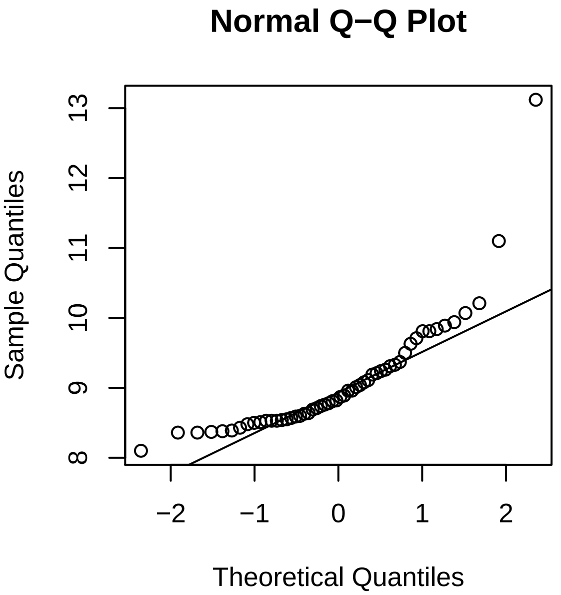

Normal Q−Q Plot
Sample Quantiles
10.5
11.0
11.5
12.0
12.5
Theoretical Quantiles
−2
−1
0
1
2

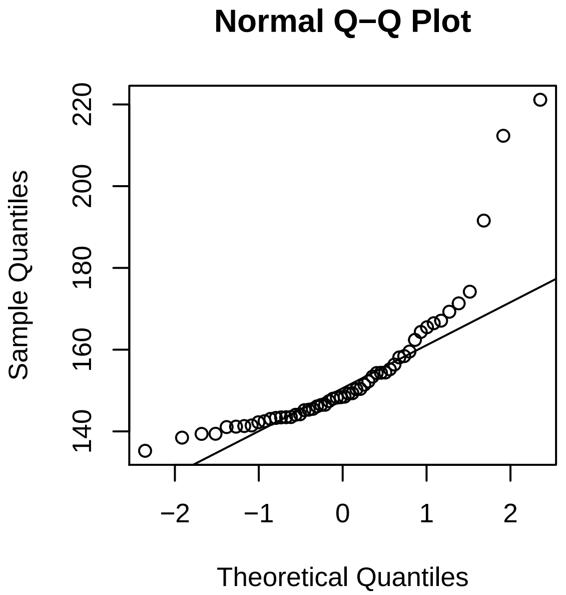

Normal Q−Q Plot
Sample Quantiles
22
23
24
25
26
Theoretical Quantiles
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
48
52
56
60
Theoretical Quantiles
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
1.9
2.0
2.1
2.2
2.3
Theoretical Quantiles
−2
−1
0
1
2







Principal Components method - with sample Covariance matrix S

Now we decide to do the factor analyisis by working with the sample covariance matrix **S** and estimate using the principal components method.

```
## [1] "percentage of sample variance explained by first common factor: 0.984"
```

We observe that the first eigenvalue of **S** explains almost all of the sample variance, so our factor model is only going to have one common factor.

```
## [1] "loadings:"
```

```
##              [,1]
## [1,]  -0.26706480
## [2,]  -0.64032562
## [3,]  -1.78547336
## [4,]  -0.07460419
## [5,]  -0.21653389
## [6,]  -0.65402335
## [7,] -16.43816362
```

```
## [1] "communalities:"
```

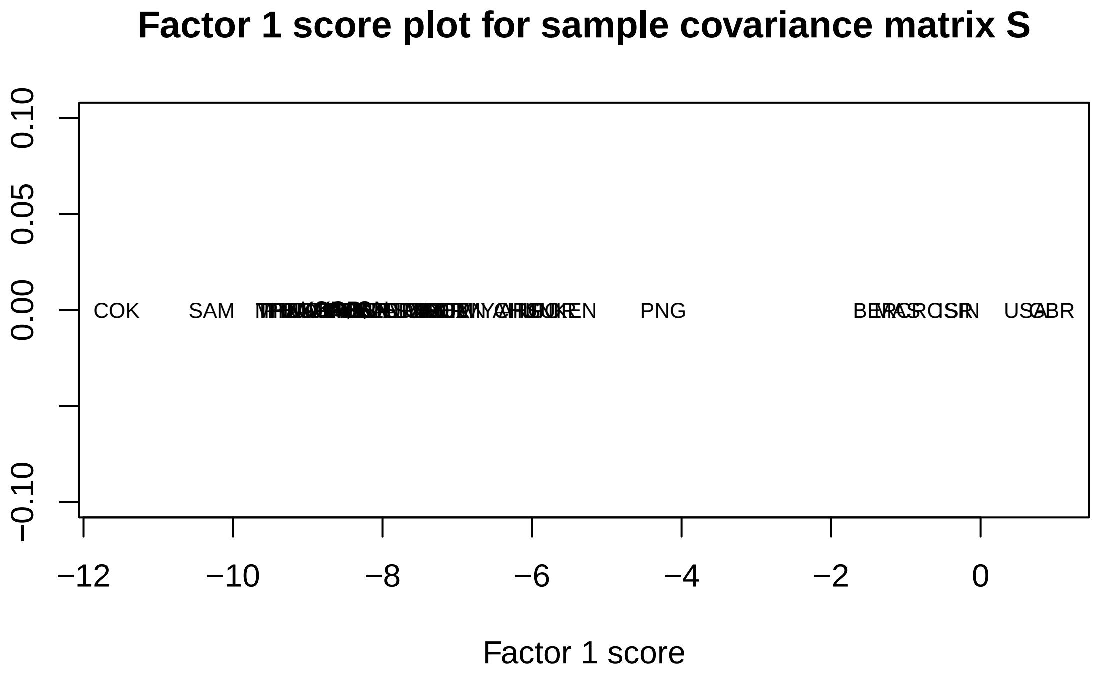
```
## [1] 7.132361e-02 4.100169e-01 3.187915e+00 5.565785e-03 4.688692e-02
## [6] 4.277465e-01 2.702132e+02
```

```
## [1] "Uniquenesses:"
```

```
## [1] 0.08399212 0.45307143 3.55754251 0.00198114 0.02729578 0.23701139
## [7] 0.05692705
```

```
## [1] "Residual Matrix:"
```

```
##                 100m(s)     200m(s)     400m(s)    800m(min)   1500m(min)
## 100m(s)     0.000000000  0.17355236  0.41445894  0.007779411  0.026062617
## 200m(s)     0.173552365  0.00000000  1.04955200  0.018394923  0.064111117
## 400m(s)     0.414458939  1.04955200  0.00000000  0.048604135  0.122561350
## 800m(min)   0.007779411  0.01839492  0.04860414  0.000000000  0.005260235
## 1500m(min)  0.026062617  0.06411112  0.12256135  0.005260235  0.000000000
## 3000m(min)  0.059216197  0.13556227  0.25907453  0.012586432  0.074536920
## Mara(min)  -0.055877248 -0.14078979 -0.44617182 -0.006701258 -0.019582127
##             3000m(min)    Mara(min)
## 100m(s)     0.05921620 -0.055877248
## 200m(s)     0.13556227 -0.140789787
## 400m(s)     0.25907453 -0.446171818
## 800m(min)   0.01258643 -0.006701258
## 1500m(min)  0.07453692 -0.019582127
## 3000m(min)  0.00000000 -0.044851664
## Mara(min)  -0.04485166  0.000000000
```



The plot is not ideal, but One can discern at least that Great Britain and USA score highest and Cook Island and Samoa score lowest. An inspection of the loadings show that the score is based almost solely on the result of the marathon distance, the records of which exhibit the largest sample variance. One notices that the uniquenesses are highest for the distance records which have the highest absolute values (third and seventh distances), regardless of the time unit used. This is a weakness of using the sample covariance matrix. The residual Matrix looks ok specific factors are inferred to be fairly negligable and this model appears to be adequate.

Principal Components method - with sample Correlation matrix R

We repeat the factor analysis just done, only now we replace sample covariance matrix **S** with sample correlation matrix **R**.

```
## [1] "percentage of sample correlation explained by first common factor: 83"
```

```
## [1] "percentage of sample correlation explained by second common factor: 8.98"
```

We see that the first two common factors are sufficient to explain most of the sample correlation. We thus proceed with a factor analysis using two common factors.

```
## [1] "loadings:"
```

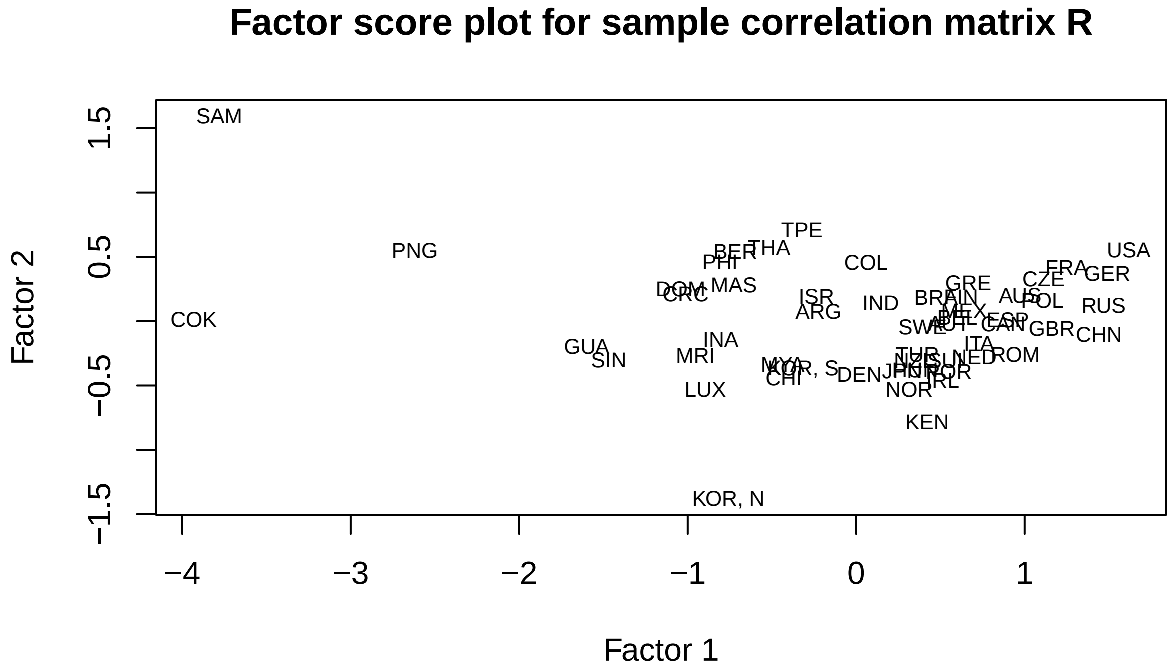
```
##            [,1]       [,2]
## [1,] -0.9103780 -0.3228503
## [2,] -0.3038482 -0.9968052
## [3,] -0.8869307 -0.3642220
## [4,] -0.3130226  0.3885866
## [5,] -0.9380805  0.2450762
## [6,] -0.2982060  1.0198460
## [7,] -0.8560043  0.3086096
```

```
## [1] "Uniquenesses:"
```

```
## [1]  0.06697954 -0.08594441  0.08069628  0.75101729  0.05994263 -0.12901260
## [7]  0.17201676
```

```
## [1] "Residual Matrix:"
```

```
##                 100m(s)   200m(s)     400m(s) 800m(min)   1500m(min)
## 100m(s)     0.000000000 0.3426530 -0.05425120 0.6496623  0.006666115
## 200m(s)     0.342653042 0.0000000  0.27625891 1.1120596  0.760587427
## 400m(s)    -0.054251195 0.2762589  0.00000000 0.6696928 -0.022950646
## 800m(min)   0.649662269 1.1120596  0.66969285 0.0000000  0.516177169
## 1500m(min)  0.006666115 0.7605874 -0.02295065 0.5161772  0.000000000
## 3000m(min)  0.785655778 1.6578330  0.78076134 0.3769295  0.443698844
## Mara(min)  -0.010693128 0.7274819  0.03012429 0.4661198 -0.088077285
##            3000m(min)   Mara(min)
## 100m(s)     0.7856558 -0.01069313
## 200m(s)     1.6578330  0.72748193
## 400m(s)     0.7807613  0.03012429
## 800m(min)   0.3769295  0.46611981
## 1500m(min)  0.4436988 -0.08807728
## 3000m(min)  0.0000000  0.22873040
## Mara(min)   0.2287304  0.00000000
```



I think something is wrong, since we have negative uniquenesses. However, the plot looks very similar to the principal components analyisis score plot, and the two common factors scores have similar interpretations as the PC scores. (although one axis is reversed). The model is adequate for the same reasons as in the previous case.

Appendix - R-Code

```
www = "http://www.ida.liu.se/~732A37/T1-9.dat"
data <- read.delim(www, header = FALSE, sep="\t")
colnames(data) <- c("NAT","100m(s)","200m(s)","400m(s)",
                    "800m(min)","1500m(min)","3000m(min)","Mara(min)")
#head(data)
nations <- as.character(data[,1])
data <- data[,-1]
col_mu <- colMeans(data)
col_sigmasquared<-apply(data,2,var)

corrmatr <- cor(data)
corrmatr
eigenstuff <-eigen(corrmatr)
eigenstuff
data2 <- scale(data)
std_lambdas <- eigen(cor(data2))$values #Same as eigenstuff$values
pcastuff <- prcomp(data2,center = FALSE,scale.=FALSE)

pcastuff$rotation[,1:2]

pccorrstdvars <- matrix(0,nrow=2,ncol=7)

for(i in 1:2){
  for(j in 1:7){
    pccorrstdvars[i,j] <- pcastuff$rotation[j,i] * sqrt(std_lambdas[i])
  }
}
paste("Correlations between the two first PCs and the standardized data variables:")
rownames(pccorrstdvars) <- c("PC1","PC2")
colnames(pccorrstdvars) <- c("100m(s)","200m(s)","400m(s)",
                       "800m(min)","1500m(min)","3000m(min)","Mara(min)")
pccorrstdvars
paste("Percent of total sample variance explained by first PC:",
      sprintf("%2.3f",eigenstuff$values[1]/sum(eigenstuff$values)*100))
paste("Percent of total sample variance explained by second PC:",
      sprintf("%2.3f",eigenstuff$values[2]/sum(eigenstuff$values)*100))
paste("Cumulative Percent of total sample variance explained by first two PCs:",
      sprintf("%2.3f",(eigenstuff$values[2]+eigenstuff$values[1])
              /sum(eigenstuff$values)*100))
pcscores <- data.frame(PC1=pcastuff$x[,1], PC2=pcastuff$x[,2])
plot(pcscores,pch="",main=c("Score plot for PC1 and PC2",
     "of national track records for women"))
text(pcscores[,1],pcscores[,2],labels=nations,cex=0.7)
MLfactanal <- factanal(data,factors = 2,scores = "Bartlett")
MLfactanal
```

```r
plot(MLfactanal$scores, pch="",main="Factor scores for ML factor analysis")
text(MLfactanal$scores, labels=nations,cex=0.7)
par(mfrow=c(2,2))
for(i in 1:7){
  qqnorm(data[,i])
  qqline(data[,i])
}
par(mfrow=c(1,1))
eigenthings <- eigen(cov(data))
estsqrteigenvalS <- sqrt(eigenthings$values)
paste("percentage of sample variance explained by first common factor:",
      signif(eigenthings$values[1] / sum(diag(cov(data))),3))
L <- estsqrteigenvalS[1] * eigenthings$vectors[,1]
L <- as.matrix(L,drop=FALSE)
paste("loadings:")
L
loadings <- L %*% t(L)
communalities <- diag(loadings)
paste("communalities:")
communalities
specificfactors <- diag(x=diag(cov(data) - loadings))
paste("Uniquenesses:")
diag(specificfactors)
residualmatr <- cov(data) - loadings - specificfactors
paste("Residual Matrix:")
residualmatr

centereddata <- t(as.matrix(data - col_mu))
scores <- as.vector(solve(t(L) %*% L) %*% t(L) %*% centereddata)
scores <- as.data.frame(cbind(scores,rep(0,length(scores))))
plot(scores,pch="",ylim=c(-0.1,0.1),main="Factor 1 score plot for sample covariance matrix S",
     ylab="",xlab="Factor 1 score")
text(scores,labels=nations,cex=0.7)
estsqrteigenvalS2 <- sqrt(eigenstuff$values)
paste("percentage of sample correlation explained by first common factor:",
      signif(eigenstuff$values[1] / 7 * 100,3))
paste("percentage of sample correlation explained by second common factor:",
      signif(eigenstuff$values[2] / 7 * 100,3))
L2 <- estsqrteigenvalS2[1:2] * eigenstuff$vectors[,1:2]
L2 <- as.matrix(L2,drop=FALSE)
paste("loadings:")
L2
loadings2 <- L2 %*% t(L2)
specificfactors2 <- diag(x=diag(corrmatr - loadings2))
paste("Uniquenesses:")
diag(specificfactors2)
residualmatr2 <- corrmatr - loadings2 - specificfactors2
paste("Residual Matrix:")
residualmatr2
scores2 <- (solve(t(L2) %*% L2) %*% t(L2) %*% t(data2))
plot(scores2[1,],scores2[2,], pch = "",main="Factor score plot for sample correlation matrix R",
     xlab="Factor 1",ylab="Factor 2")
text(scores2[1,],scores2[2,],labels=nations,cex=0.7)
```

```
## NA
```