Multivariate Statistical Methods

Assignment 3

Allan Gholmi, Emma Wallentinsson, Rasmus Holm

2017-12-14

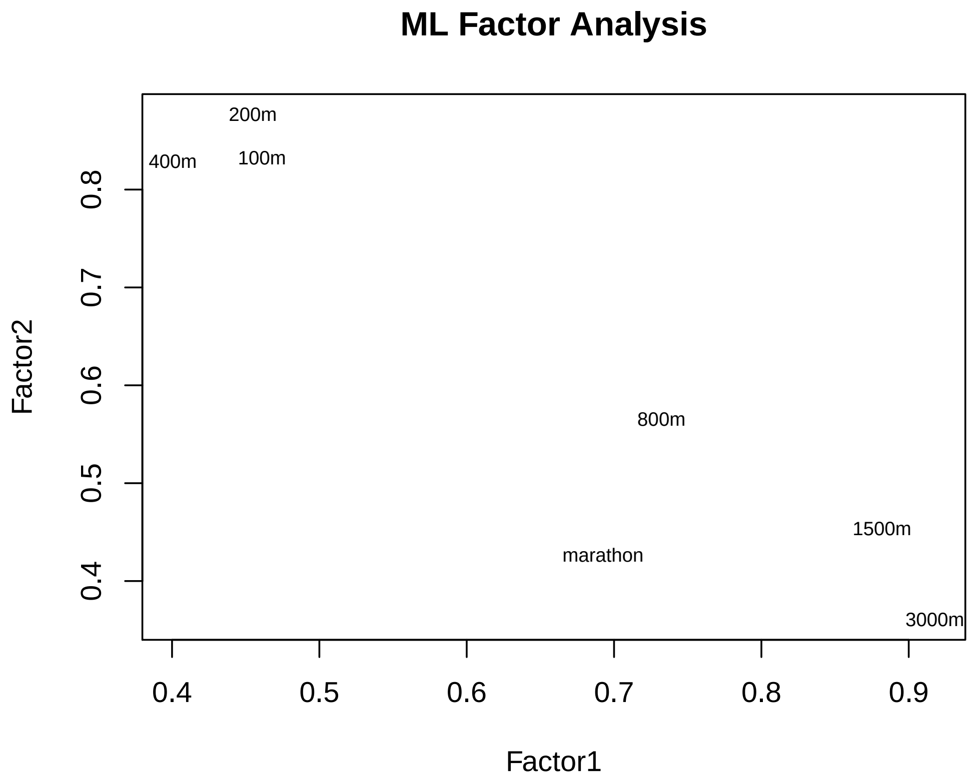
Question 2

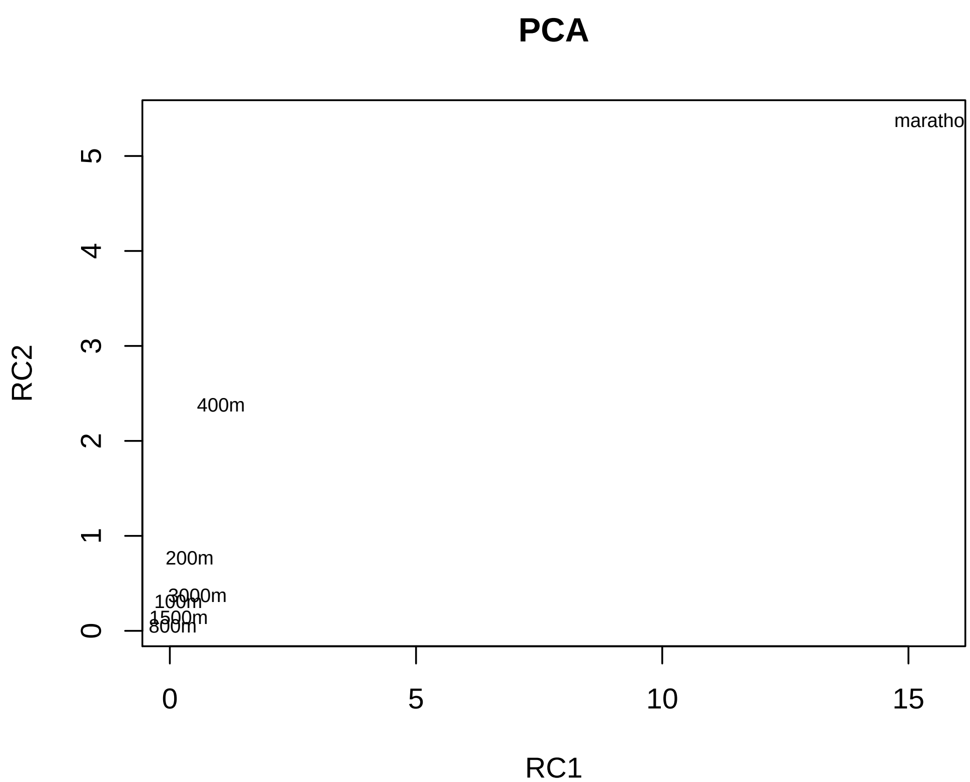
```
library(psych)

data <- read.table("../data/T1-9.dat")
names(data) <- c("country", "100m", "200m", "400m", "800m", "1500m", "3000m", "marathon")
numeric_data <- data[, -1]
countries <- as.character(data$country)

S <- cov(numeric_data)
R <- cor(numeric_data)
factors <- 2
```

Since the data is measured in different units it is more appropriate to use the correlation matrix.

Analysis on Covariance Matrix

```
S_principal <- principal(S, factors, rotate="varimax", covar=TRUE)
S_factanalysis <- factanal(numeric_data, factors=factors, covmat=S, rotation="varimax")

S_factoranalysis_loadings <- S_factanalysis$loadings[, 1:2]
S_principal_loadings <- S_principal$loadings[, 1:2]

old <- par(mfrow=c(2, 1))
plot(S_factoranalysis_loadings, type="n", main="ML Factor Analysis")
text(S_factoranalysis_loadings, labels=names(numeric_data), cex=.7)

plot(S_principal_loadings, type="n", main="PCA")
text(S_principal_loadings, labels=names(numeric_data), cex=.7)
```


ML Factor Analysis
Factor2
0.4
0.5
0.6
0.7
0.8
0.9
Factor1
200m
400m
100m
800m
1500m
marathon
3000m


PCA
RC2
0
1
2
3
4
5
10
15
RC1
maratho
400m
200m
3000m
100m
1500m
800m

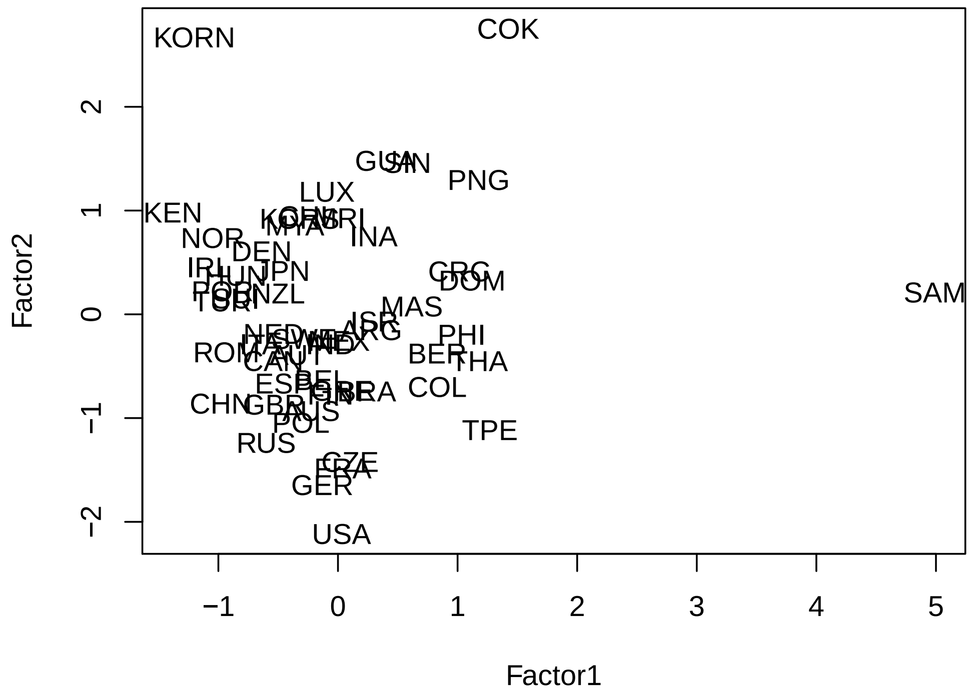
```r
par(old)
```

```r
print("PCA")
#> [1] "PCA"
S_principal$Vaccounted
#>                                RC1        RC2
#> SS loadings           243.0046956 35.3746648
#> Proportion Var          0.8716703  0.1268907
#> Cumulative Var          0.8716703  0.9985611
#> Proportion Explained    0.8729264  0.1270736
#> Cumulative Proportion   0.8729264  1.0000000

print("FA")
#> [1] "FA"
S_factanalysis$loadings
#>
#> Loadings:
#>          Factor1 Factor2
#> 100m     0.461   0.833
#> 200m     0.455   0.877
#> 400m     0.401   0.829
#> 800m     0.732   0.566
#> 1500m    0.882   0.454
#> 3000m    0.918   0.361
#> marathon 0.693   0.427
#>
#>                Factor1 Factor2
#> SS loadings      3.216   2.987
#> Proportion Var   0.459   0.427
#> Cumulative Var   0.459   0.886
```

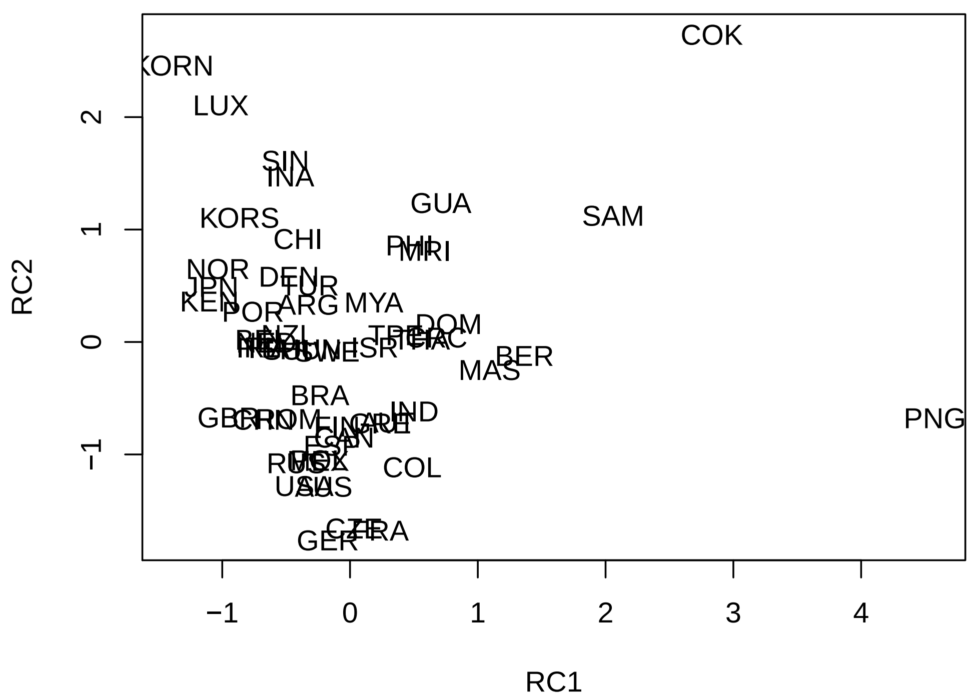
We can see that the first principal component explains about 87% of the variance and the largest loading is associated with the marathon which is clear from the plot.

```r
factor_scores <- factanal(numeric_data, factors=factors,
                          rotation="varimax", scores="regression")$scores
principal_scores <- principal(numeric_data, factors, scores=TRUE, covar=TRUE)$scores

old <- par(mfrow=c(2, 1))
plot(factor_scores, type="n", main="ML Factor Analysis")
text(factor_scores, labels=countries)

plot(principal_scores, type="n", main="PCA")
text(principal_scores, labels=countries)
```

ML Factor Analysis



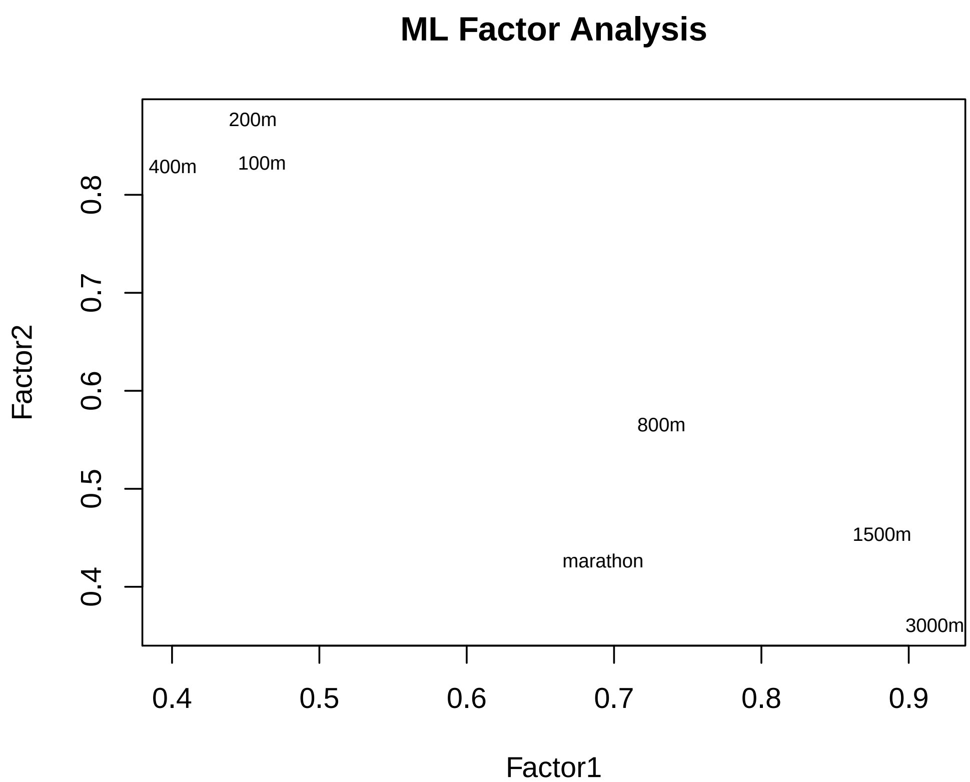
PCA

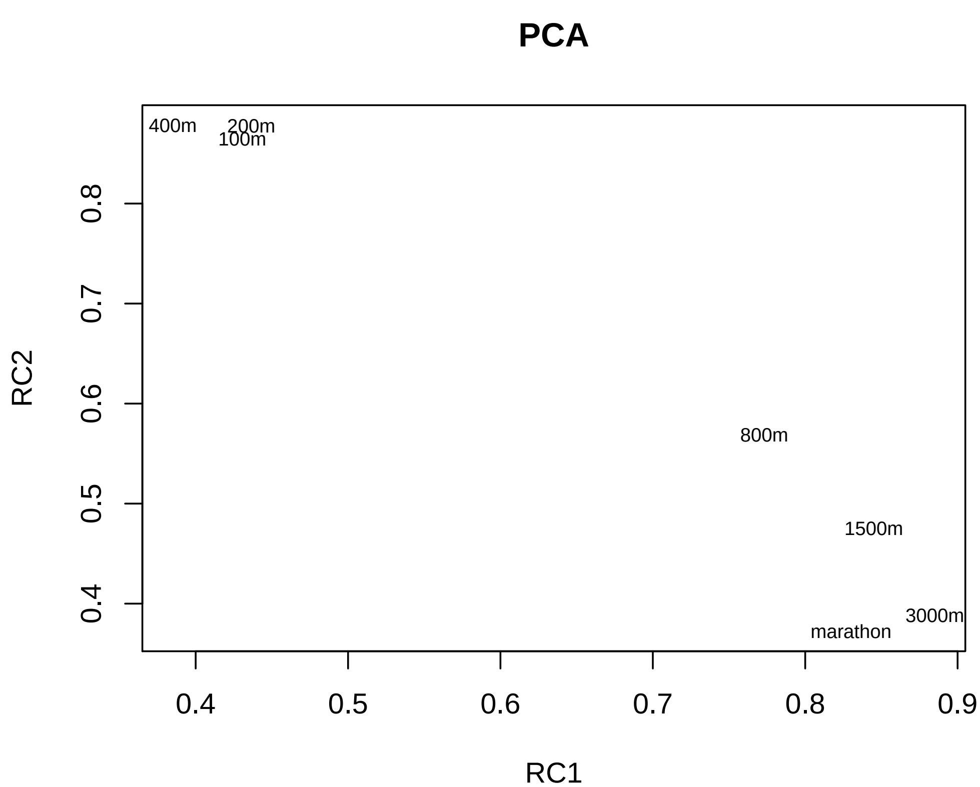


```r
par(old)
```

Analysis on Correlation Matrix

```r
R_principal <- principal(R, factors, rotate="varimax", covar=FALSE)
R_factanalysis <- factanal(numeric_data, factors=factors, covmat=R, rotation="varimax")

R_factoranalysis_loadings <- R_factanalysis$loadings[, 1:2]
R_principal_loadings <- R_principal$loadings[, 1:2]

old <- par(mfrow=c(2, 1))
plot(R_factoranalysis_loadings, type="n", main="ML Factor Analysis")
text(R_factoranalysis_loadings, labels=names(numeric_data), cex=.7)

plot(R_principal_loadings, type="n", main="PCA")
text(R_principal_loadings, labels=names(numeric_data), cex=.7)
```


ML Factor Analysis
200m
400m
100m
800m
1500m
marathon
3000m
Factor2
0.4
0.5
0.6
0.7
0.8
0.9
Factor1


PCA
400m
200m
100m
800m
1500m
3000m
marathon
RC2
0.4
0.5
0.6
0.7
0.8
0.9
RC1

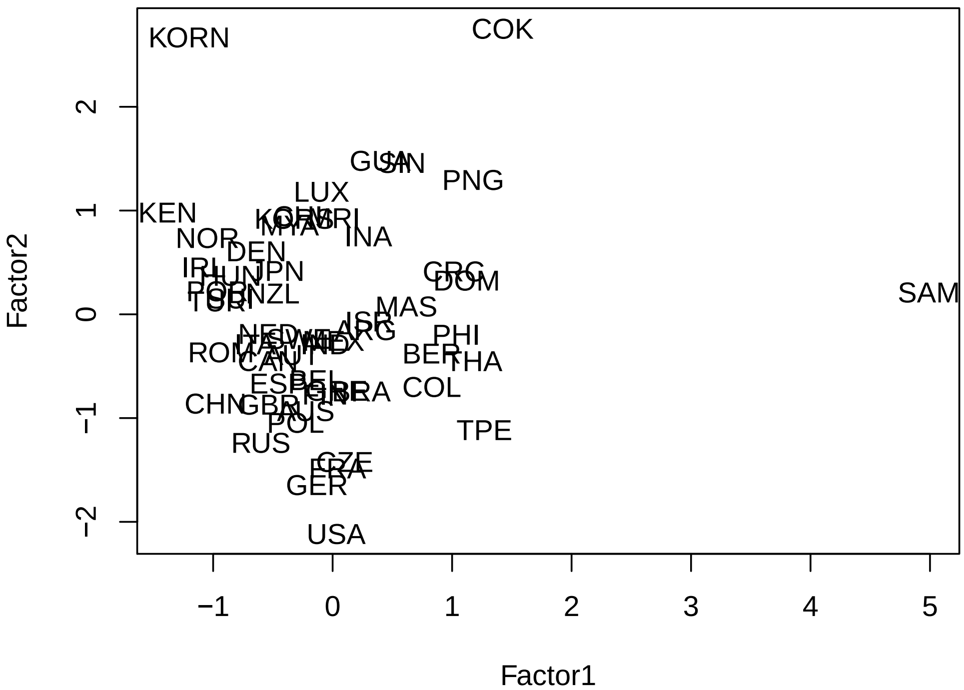
```r
par(old)
```

```r
print("PCA")
#> [1] "PCA"
R_principal$Vaccounted
#>                             RC1       RC2
#> SS loadings           3.3087124 3.1276055
#> Proportion Var        0.4726732 0.4468008
#> Cumulative Var        0.4726732 0.9194740
#> Proportion Explained  0.5140691 0.4859309
#> Cumulative Proportion 0.5140691 1.0000000

print("FA")
#> [1] "FA"
R_factanalysis$loadings
#>
#> Loadings:
#>          Factor1 Factor2
#> 100m     0.461   0.833
#> 200m     0.455   0.877
#> 400m     0.401   0.829
#> 800m     0.732   0.566
#> 1500m    0.882   0.454
#> 3000m    0.918   0.361
#> marathon 0.693   0.427
#>
#>                Factor1 Factor2
#> SS loadings      3.216   2.987
#> Proportion Var   0.459   0.427
#> Cumulative Var   0.459   0.886
```

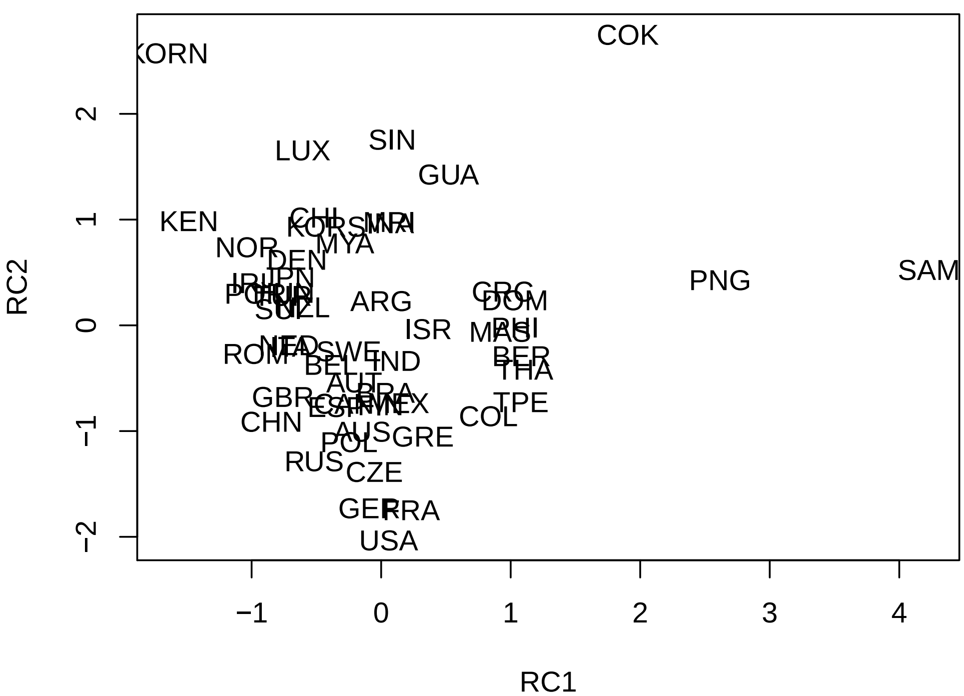
Now the first two principal components explains about the same amount of variance and in total almost 92% so its a pretty good fit. Similar values are true for the factors.

```r
factor_scores <- factanal(numeric_data, factors=factors,
                          rotation="varimax", scores="regression")$scores
principal_scores <- principal(numeric_data, factors, scores=TRUE, covar=FALSE)$scores

old <- par(mfrow=c(2, 1))
plot(factor_scores, type="n", main="ML Factor Analysis")
text(factor_scores, labels=countries)

plot(principal_scores, type="n", main="PCA")
text(principal_scores, labels=countries)
```

ML Factor Analysis



PCA



```r
par(old)
```

Setting rotation to varimax means that the algorithm rotates the loadings such as to maximize their variances. As a result of this rotation, each variable loads more heavily on a single factor making the factors easier to interpret.

Appendix

Code