Multivariate Statistical Methods

Assignment 3

Allan Gholmi, Emma Wallentinsson, Rasmus Holm

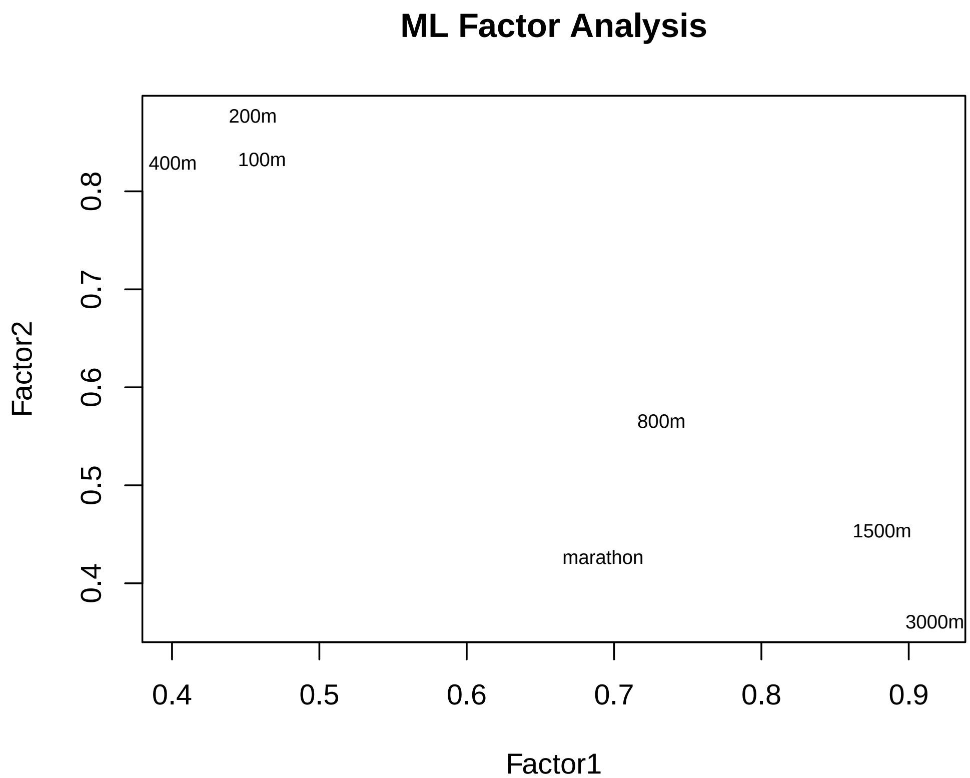
2017-12-14

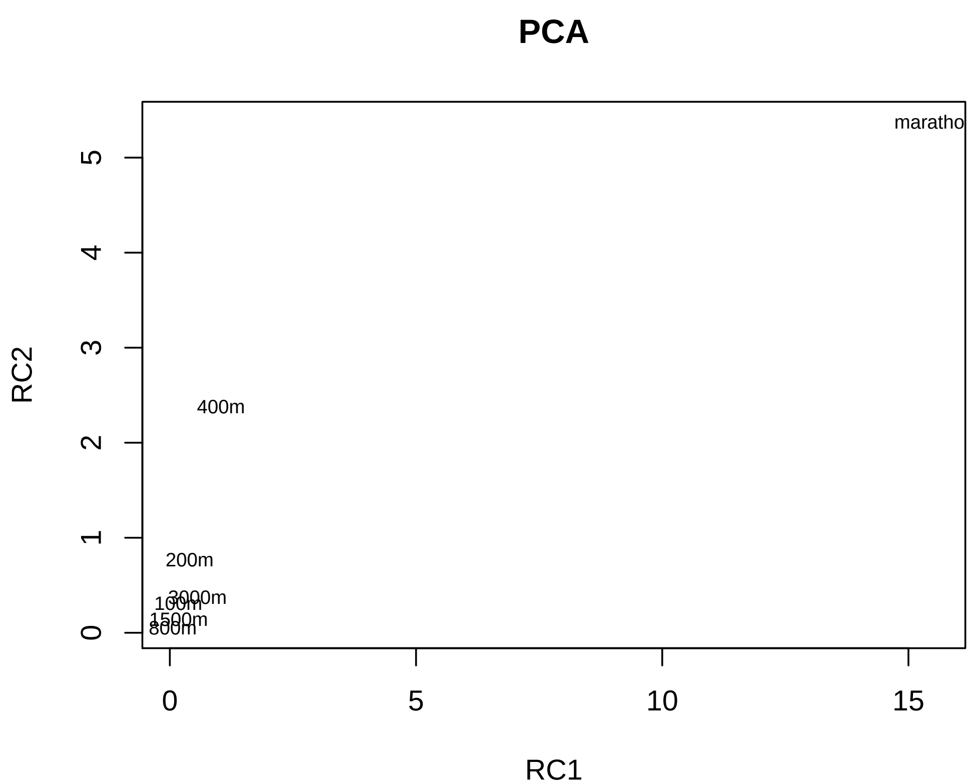
Question 2

```
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

print(S)
#>                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
```

Since the data is measured in different units it is more appropriate to use the correlation matrix. We can see that the covariances of marathon is huge compared to the other variables which will pose a problem.

Analysis on Covariance Matrix

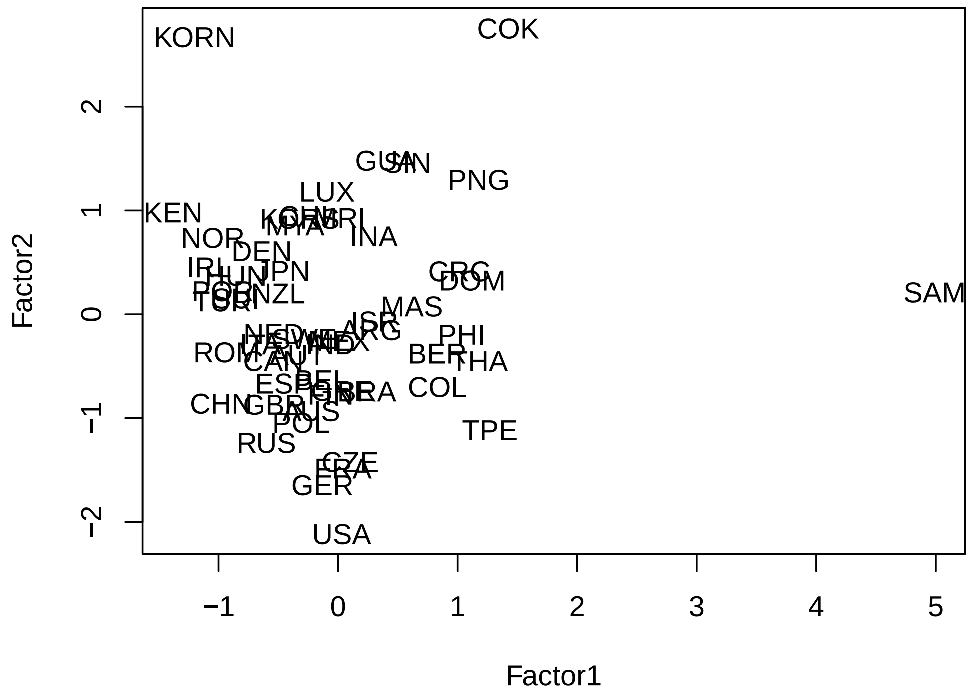




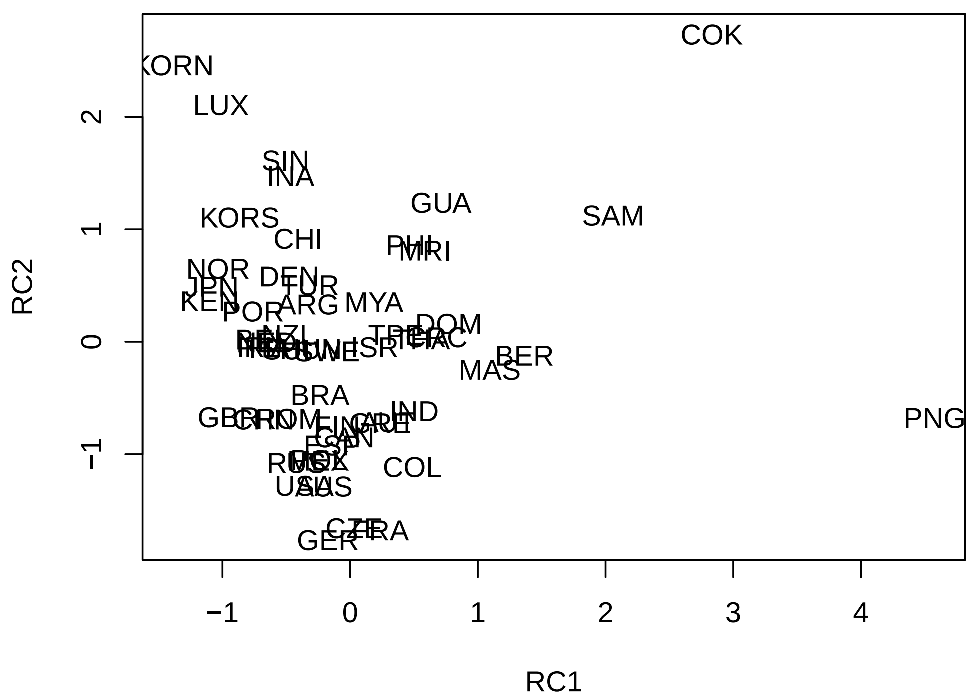
```
#> [1] "PCA"
#>
#> Loadings:
#>          RC1    RC2
#> 100m      0.173  0.307
#> 200m      0.404  0.765
#> 400m      1.038  2.376
#> 800m
#> 1500m     0.179  0.142
#> 3000m     0.561  0.371
#> marathon 15.537  5.375
#>
#>                    RC1    RC2
#> SS loadings    243.005 35.375
#> Proportion Var  34.715  5.054
#> Cumulative Var  34.715 39.768
#> [1] "FA"
#>
#> 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. The other component explains about 13% of the variance and is thus not very informative. These two components do not help us very much in understanding the nature of the data because we did not normalize the data.

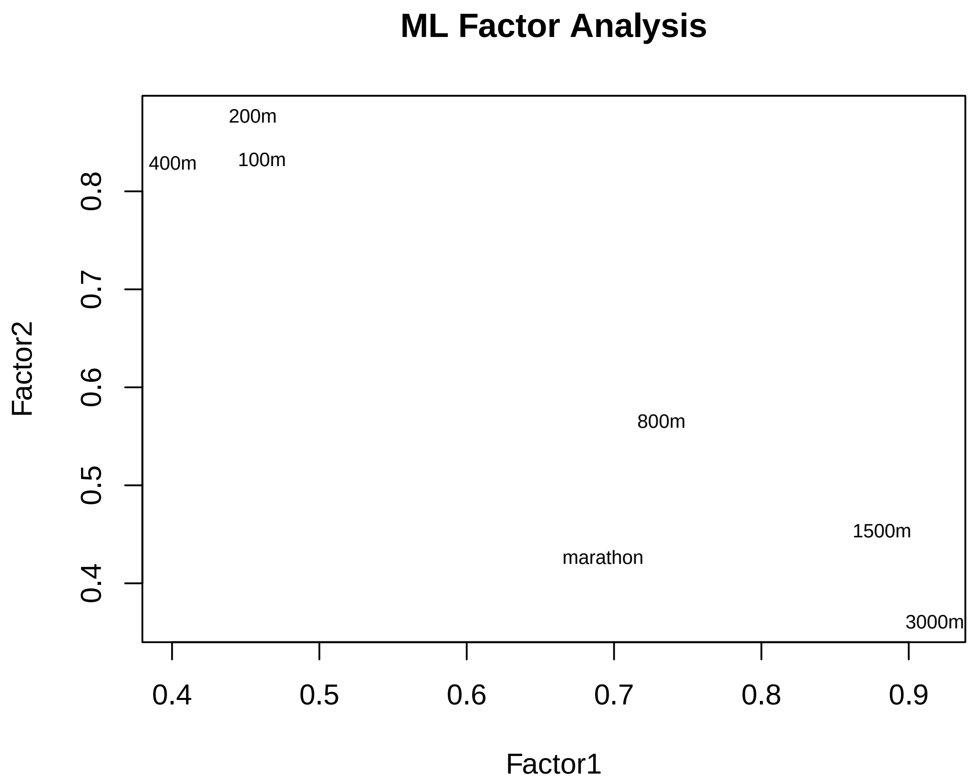
ML Factor Analysis

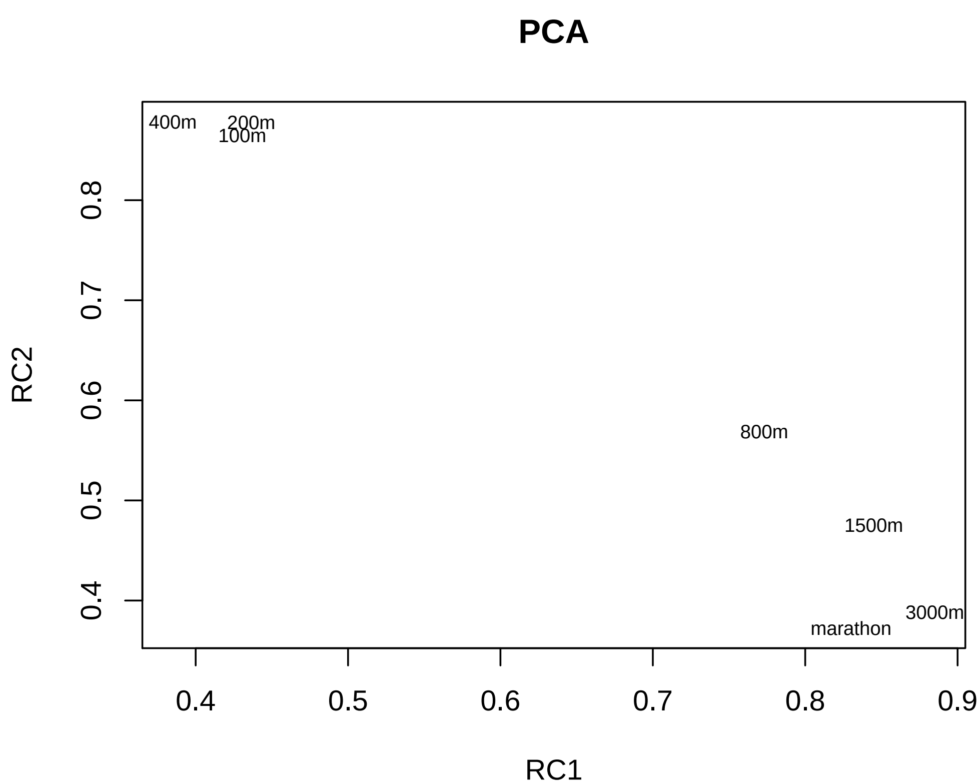


PCA



Analysis on Correlation Matrix

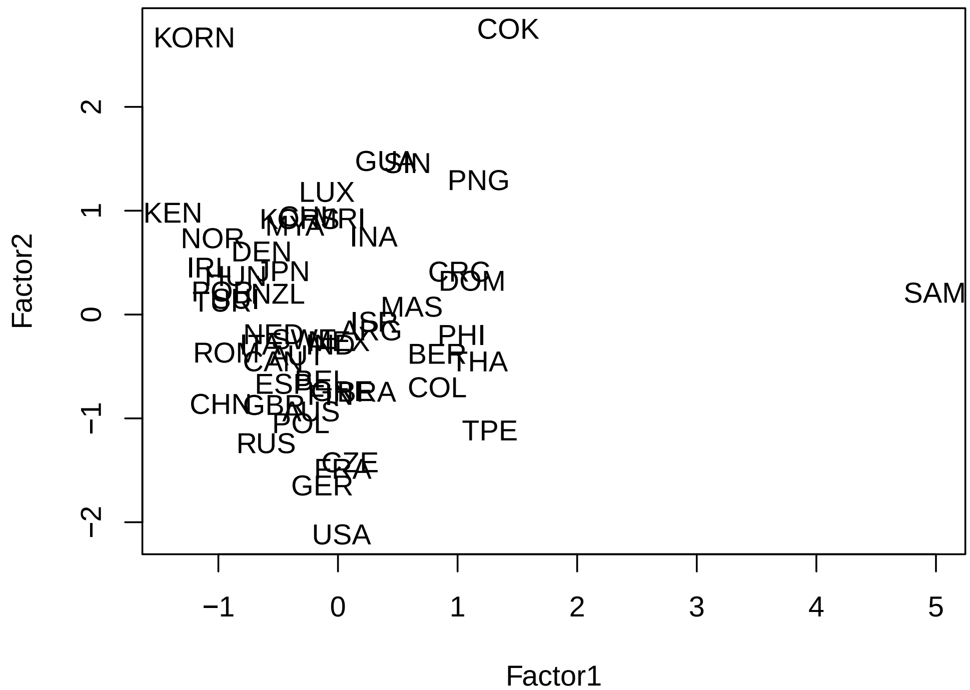




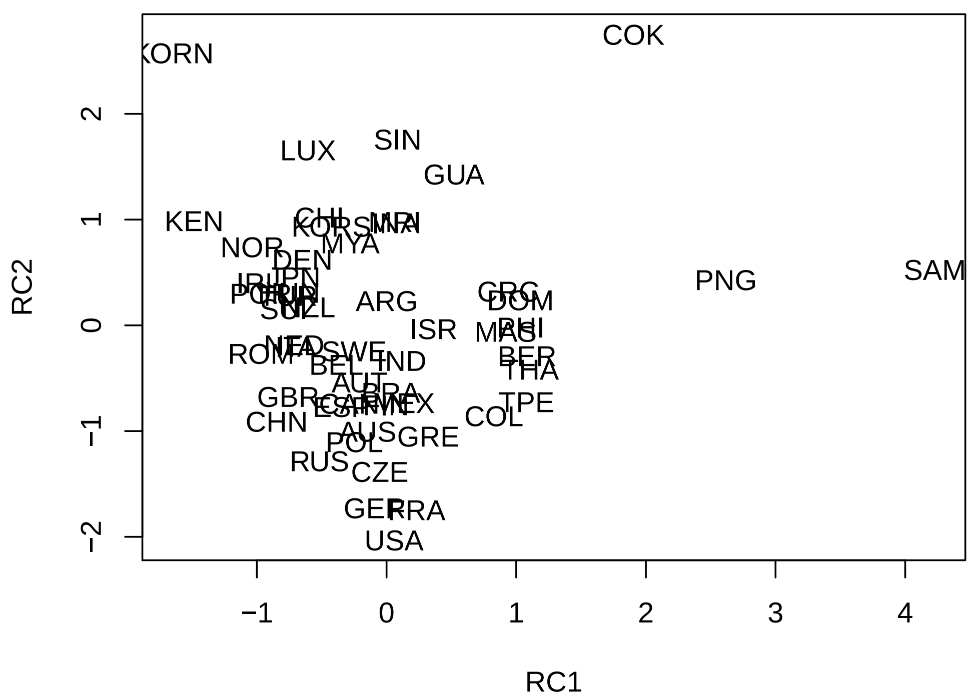
```
#> [1] "PCA"
#>
#> Loadings:
#>          RC1   RC2
#> 100m     0.431 0.865
#> 200m     0.437 0.877
#> 400m     0.385 0.878
#> 800m     0.773 0.569
#> 1500m    0.845 0.475
#> 3000m    0.885 0.388
#> marathon 0.830 0.373
#>
#>                  RC1   RC2
#> SS loadings    3.309 3.128
#> Proportion Var 0.473 0.447
#> Cumulative Var 0.473 0.919
#> [1] "FA"
#>
#> 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 decent fit. Similar values are true for the factors and so the two solutions give similar results. The first factor/principal component seem to represent shorter races since those load highly on it and the other represent longer races, but the opposite loadings are still rather high. So these factors could be interpreted as representing speed versus endurance.

ML Factor Analysis



PCA



We can see from the plots that the factor and principal component scores indicate that North Korea, Cook Islands, Samoa, and Papua New Guinea are outliers.

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

```
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

print(S)

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)
par(old)
print("PCA")
S_principal$loadings

print("FA")
S_factanalysis$loadings
factor_scores <- factanal(numeric_data, factors=factors,
                          rotation="varimax", scores="regression")$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)
par(old)

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]
```

```r
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)
par(old)
print("PCA")
R_principal$loadings

print("FA")
R_factanalysis$loadings
factor_scores <- factanal(numeric_data, factors=factors,
                          rotation="varimax", scores="regression")$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)
par(old)
```