Computer Lab 1

Thomas Zhang

2015-11-10

We have a data set of Womens national running records from all nations for seven different distances. Some descriptive statistics for the records for each running distance is given by

```
## [1] "Means"
```

```
##    100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
##     11.358     23.119     51.989      2.022      4.189      9.081
##  Mara(min)
##    153.619
```

```
## [1] "standard deviations:"
```

```
##    100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
##      0.394      0.929      2.597      0.087      0.272      0.815
##  Mara(min)
##     16.440
```

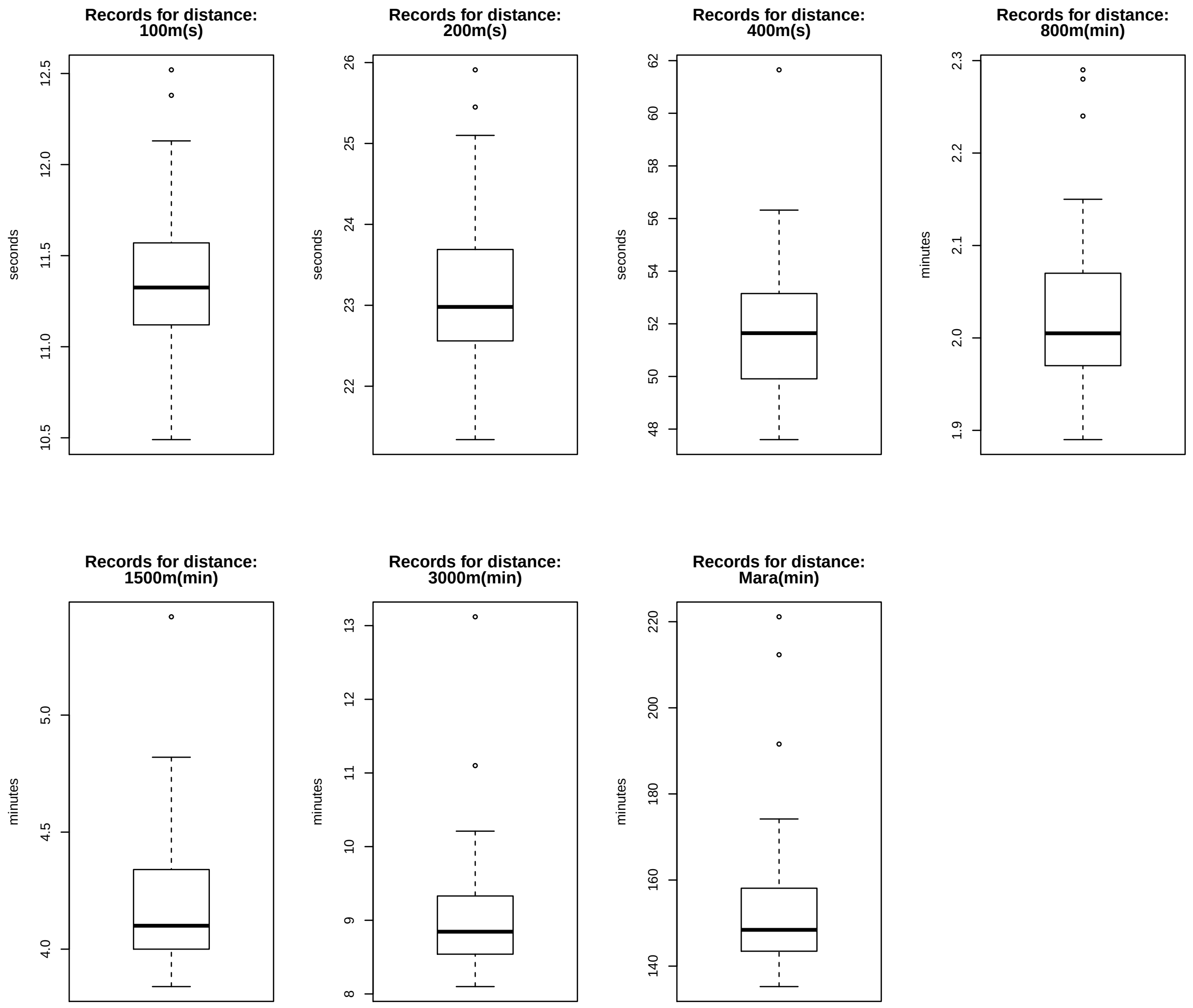
```
## [1] "Min values:"
```

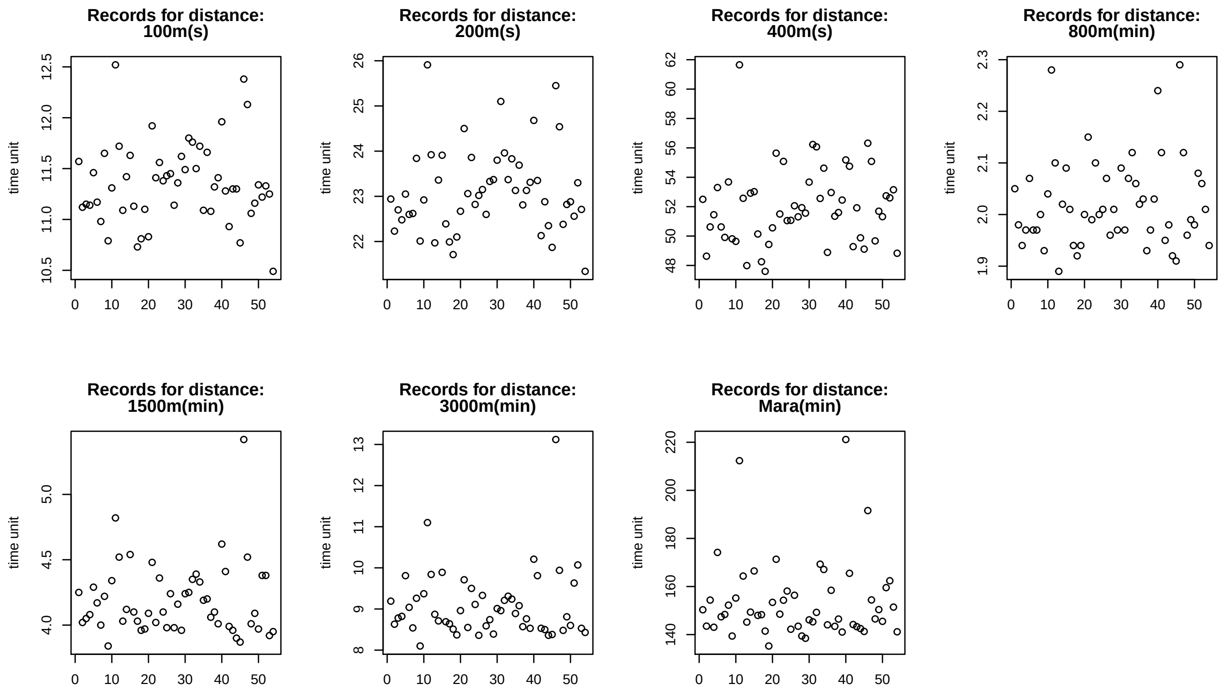
```
##    100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
##      10.49      21.34      47.60       1.89       3.84       8.10
##  Mara(min)
##     135.25
```

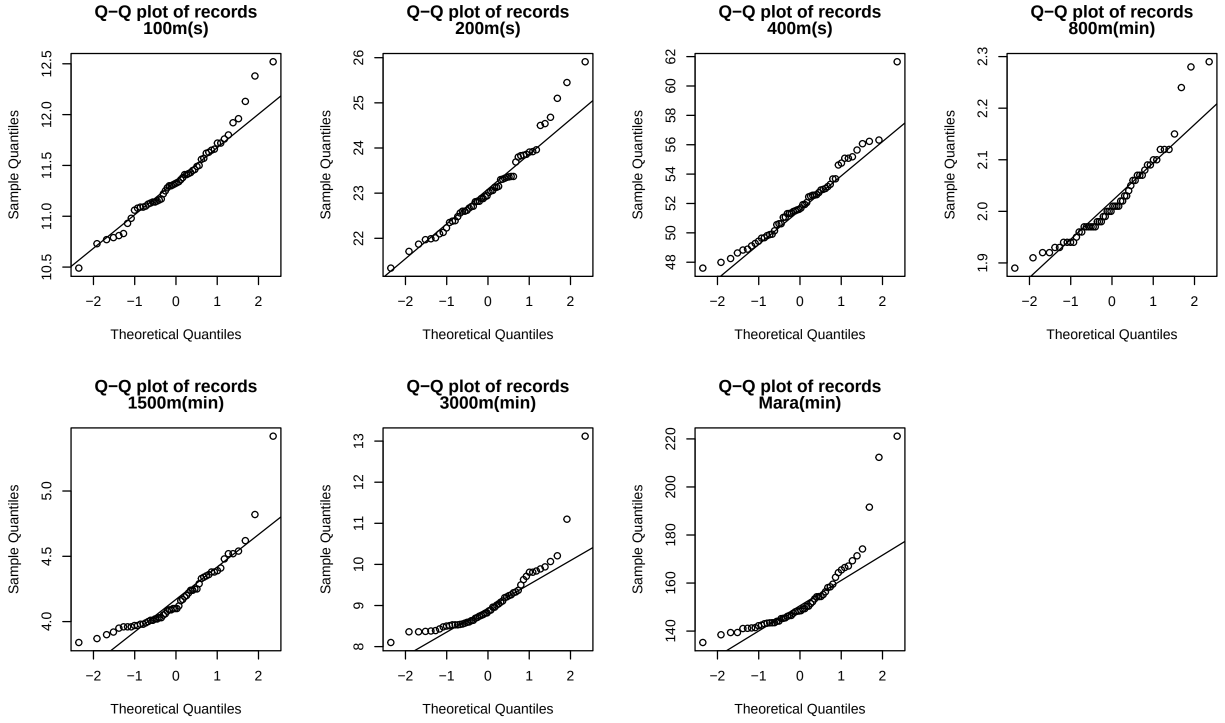
```
## [1] "Max values:"
```

```
##    100m(s)    200m(s)    400m(s)  800m(min) 1500m(min) 3000m(min)
##      12.52      25.91      61.65       2.29       5.42      13.12
##  Mara(min)
##     221.14
```

We plot the distributions of the records of each distance as a box plot and a scatter plot and we study the quantile-quantile plot (Q-Q plot) of each distance.


Records for distance:
100m(s)
seconds
10.5
11.0
11.5
12.0
12.5
Records for distance:
200m(s)
seconds
22
23
24
25
26
Records for distance:
400m(s)
seconds
48
50
52
54
56
58
60
62
Records for distance:
800m(min)
minutes
1.9
2.0
2.1
2.2
2.3
Records for distance:
1500m(min)
minutes
4.0
4.5
5.0
Records for distance:
3000m(min)
minutes
8
9
10
11
12
13
Records for distance:
Mara(min)
minutes
140
160
180
200
220





The box and scatter plots of records for differenct distances indicate that there are some outliers (nations for which the record time is bad). The Q-Q plots show that for shorter distances, the record distributions are approximately normally distributed, but with increasing distance a normal distribution of record times describes the data less and less well.

Let us take a look at the covariance and correlation matrices for the records of the seven running distances.

```
## [1] "Covariance matrix:"
```

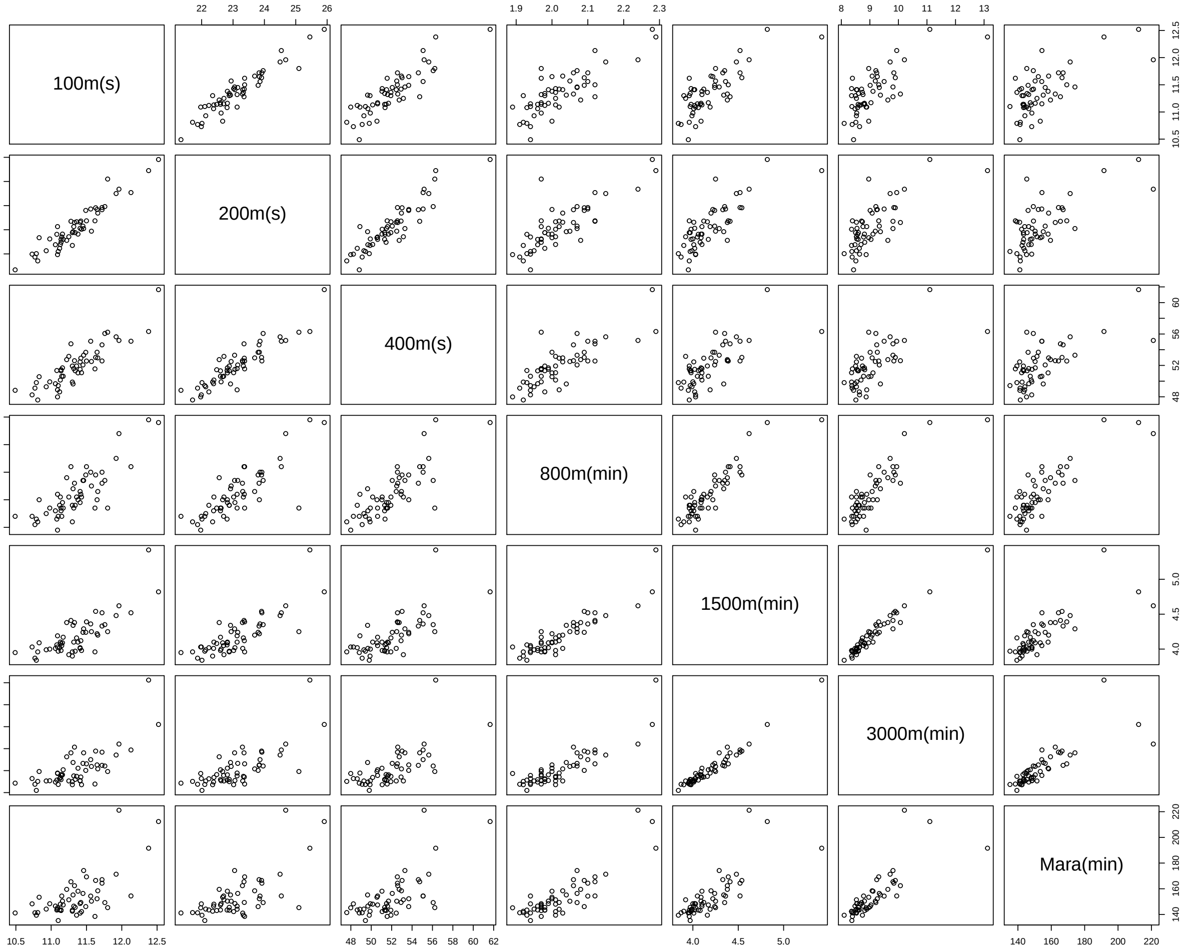
```
##            100m(s) 200m(s) 400m(s) 800m(min) 1500m(min) 3000m(min)
## 100m(s)     0.1550  0.3450   0.891   0.02770     0.0839     0.2340
## 200m(s)     0.3450  0.8630   2.190   0.06620     0.2030     0.5540
## 400m(s)     0.8910  2.1900   6.750   0.18200     0.5090     1.4300
## 800m(min)   0.0277  0.0662   0.182   0.00755     0.0214     0.0614
## 1500m(min)  0.0839  0.2030   0.509   0.02140     0.0742     0.2160
## 3000m(min)  0.2340  0.5540   1.430   0.06140     0.2160     0.6650
## Mara(min)   4.3300 10.4000  28.900   1.22000     3.5400    10.7000
##            Mara(min)
## 100m(s)         4.33
## 200m(s)        10.40
## 400m(s)        28.90
## 800m(min)       1.22
## 1500m(min)      3.54
## 3000m(min)     10.70
## Mara(min)     270.00
```

```
## [1] "Correlation matrix:"
```

```
##            100m(s) 200m(s) 400m(s) 800m(min) 1500m(min) 3000m(min)
## 100m(s)      1.000   0.941   0.871     0.809      0.782      0.728
## 200m(s)      0.941   1.000   0.909     0.820      0.801      0.732
## 400m(s)      0.871   0.909   1.000     0.806      0.720      0.674
## 800m(min)    0.809   0.820   0.806     1.000      0.905      0.867
## 1500m(min)   0.782   0.801   0.720     0.905      1.000      0.973
## 3000m(min)   0.728   0.732   0.674     0.867      0.973      1.000
## Mara(min)    0.669   0.680   0.677     0.854      0.791      0.799
##            Mara(min)
## 100m(s)        0.669
## 200m(s)        0.680
## 400m(s)        0.677
## 800m(min)      0.854
## 1500m(min)     0.791
## 3000m(min)     0.799
## Mara(min)      1.000
```

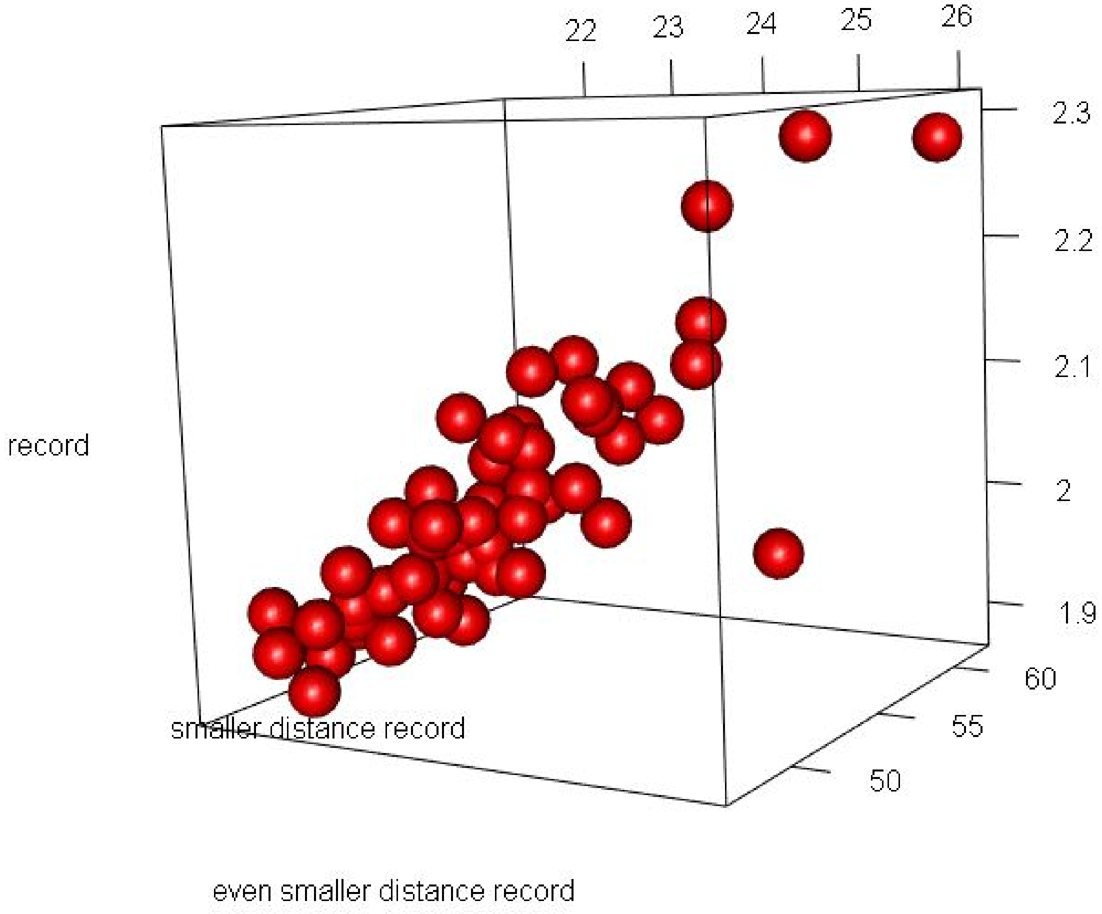
We see that the records for almost every distance is heavily correlated with the records of almost every other distance. Covariance varies in magnitude greatly while correlation is between zero and one.

We now scatter plot the records of different distances against each other.



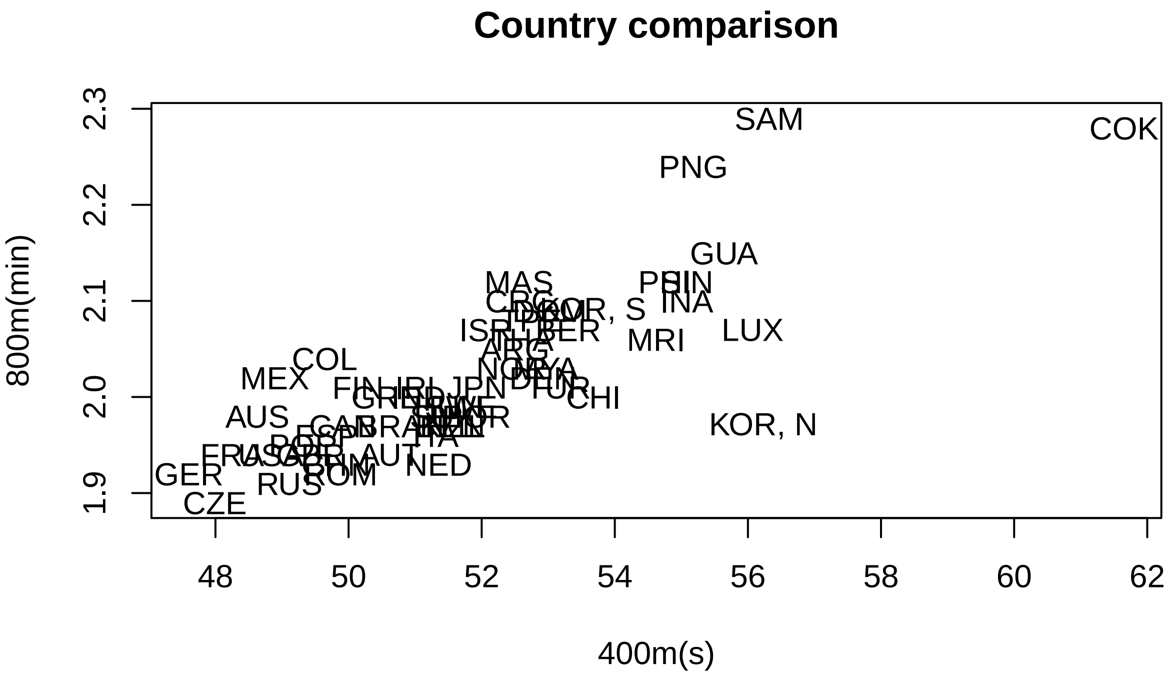
We see that there is significant positive correlation between many of the records, which indicate that countries which perform well in one distance of running has a good chance of performing well in another distance as well. There are also a few data points which show particularly poor records, as indicated previously, and these could be considered outliers.

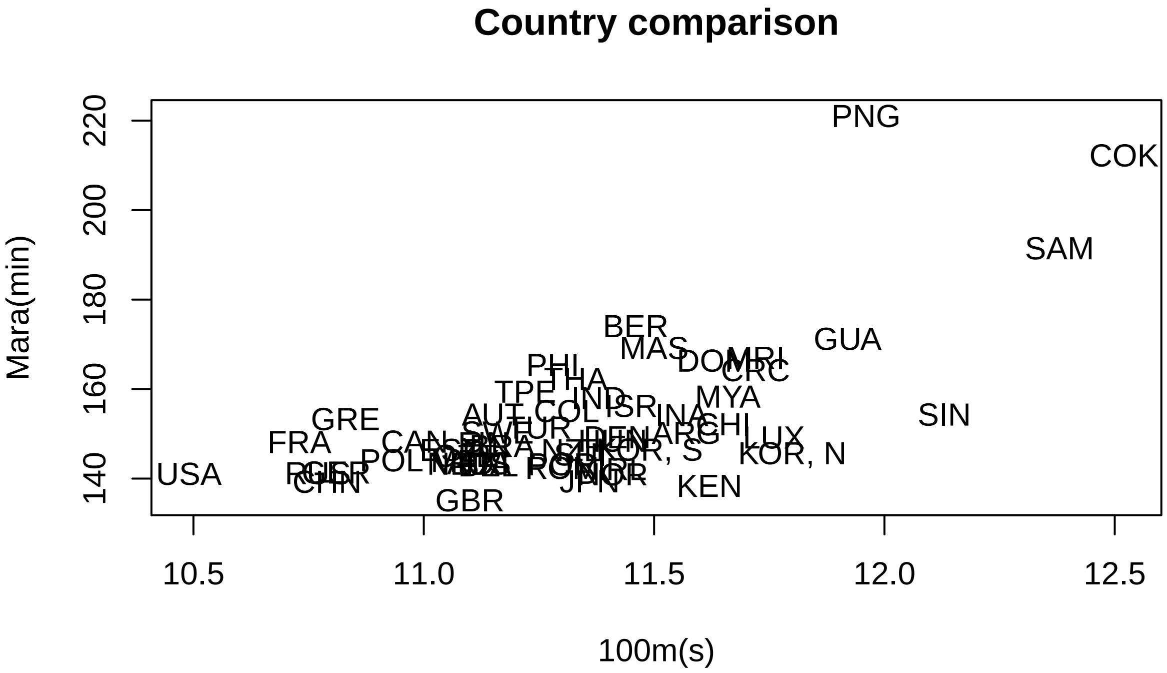
Let us illustrate that further with a 3D scatter plot, where we plot the 800 m records against the 200 m and 400 m records.

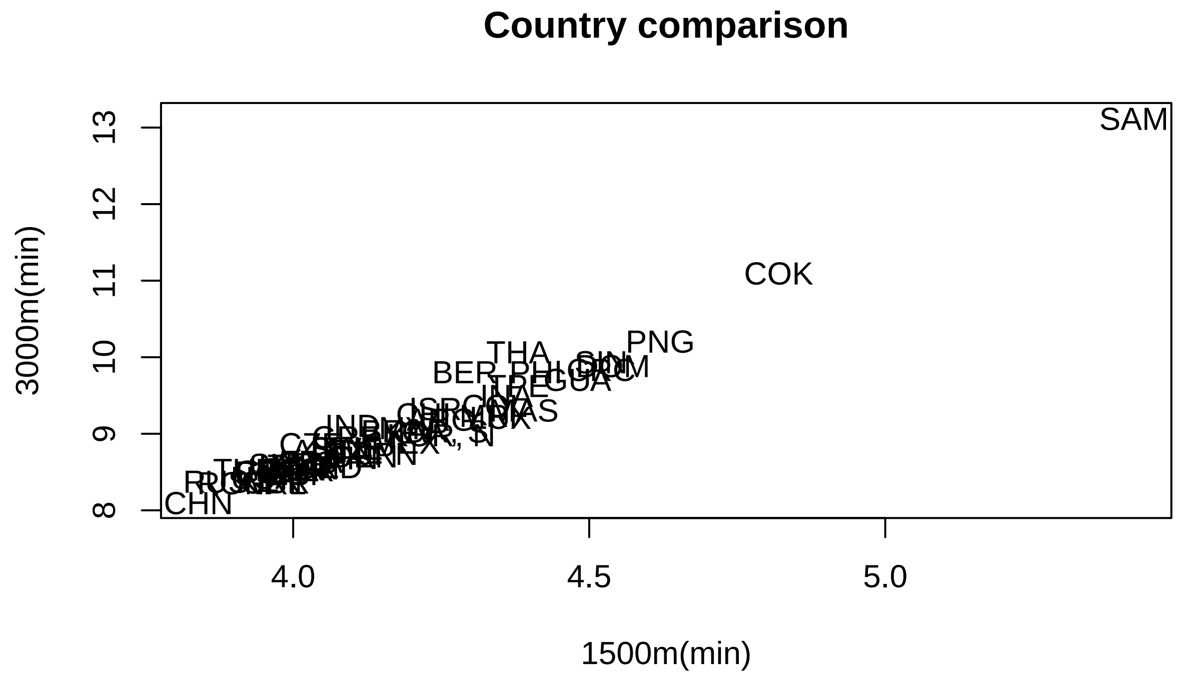


It can be seen that the majority of data points (countries) form a cluster pointing diagonally through the cube, which shows positive correlation between the records. The outliers are clearly identfied here as well.

Let us try and see which countries are performing extremely badly in terms of womens running records.

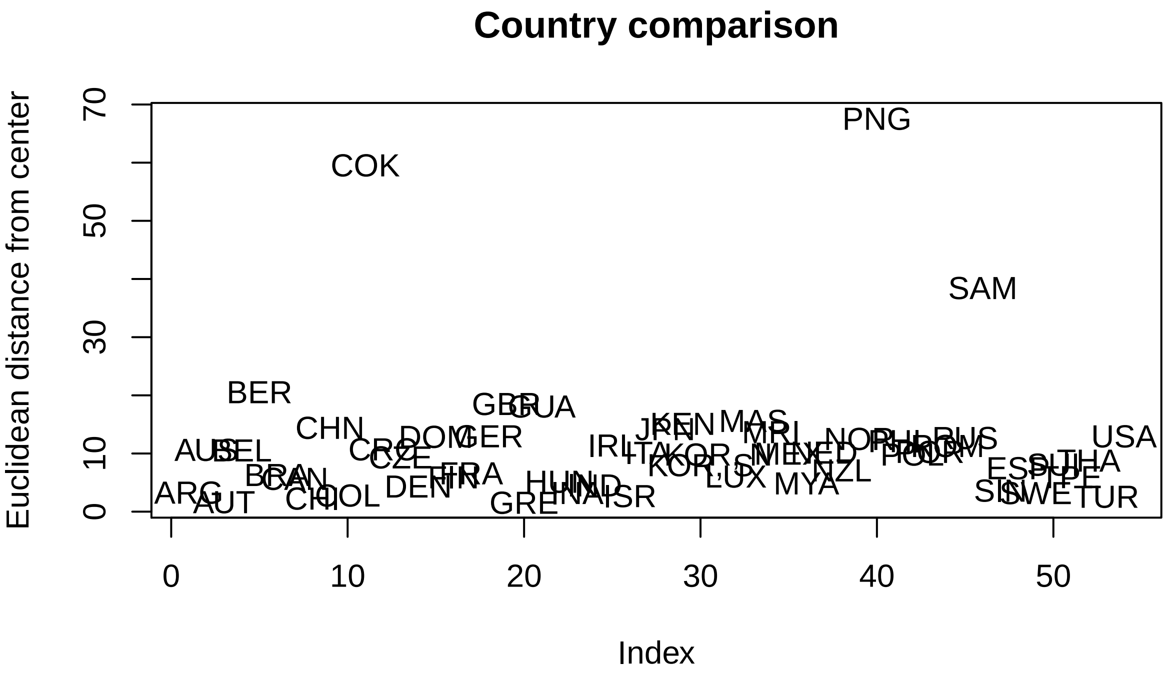

Country comparison
800m(min)
400m(s)
1.9
2.0
2.1
2.2
2.3
48
50
52
54
56
58
60
62
SAM
COK
PNG
GUA
PHI
SIN
INA
MAS
CRC
KOR, S
ISR
BER
MRI
LUX
ARG
COL
MEX
NOR
MYA
DEN
TUR
CHI
FIN
IRL
JPN
GRE
AUS
KOR, N
CAN
BRA
ESP
POR
FRA
USA
AUT
GBR
NED
GER
RUS
ROM
CZE

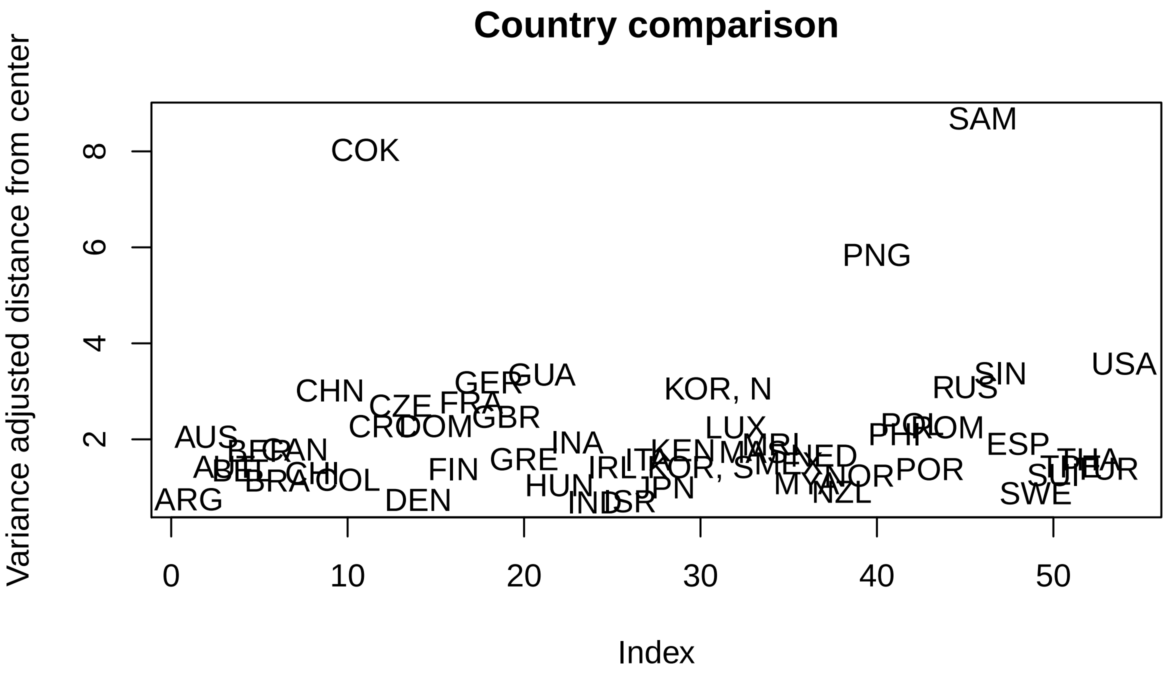

Country comparison
Mara(min)
100m(s)
140
160
180
200
220
10.5
11.0
11.5
12.0
12.5
PNG
COK
SAM
BER
GUA
MAS
DOM
MRI
CRC
PHI
THA
TPE
IND
MYA
ISR
INA
COL
AUT
SIN
TUR
CHI
ARG
DEN
LUX
KOR, S
KOR, N
GRE
FRA
CAN
POL
USA
RUS
GER
CHN
POR
JPN
KEN
GBR

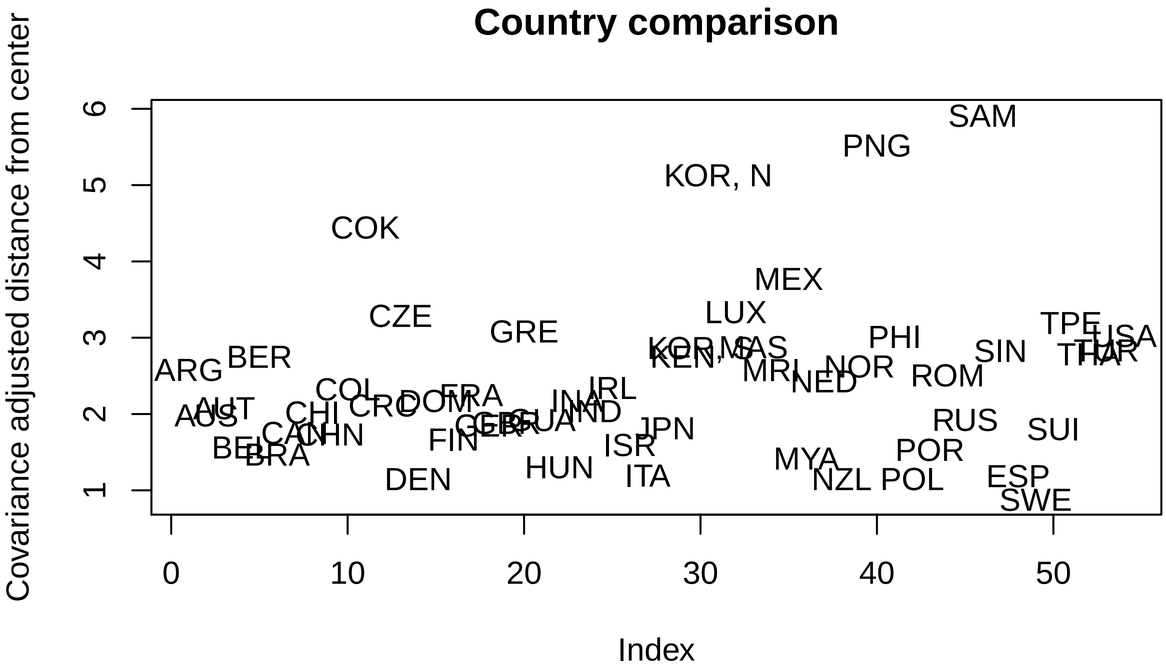


We see a pattern where the pacific island nations of Papua New Guinea, Samoa and Cook Islands are by far the worst performers in most distance categories. I suggest they are so extreme(ly bad) because they are small island nations with limited human talent pools and a culture which do not emphasize fast running. Also their professional runner training organizations are probably lacking in all respects compared to those of rich and/or large nations.

Let us now compute the euclidean distance, the standardized euclidean distance and the Mahalanobis distance from the center of the observations of the records and find out which are the five most extreme nations (fartest from the center of observations) in each case.


Country comparison
Euclidean distance from center
0
10
30
50
70
0
10
20
30
40
50
Index
PNG
COK
SAM
BER
CHN
GBR
GUA
DOM
GER
CRO
CZE
AUS
BEL
BRA
CAN
ARG
AUT
CHI
COL
DEN
FIN
FRA
GRE
HUN
INA
IND
ISR
IRL
ITA
JPN
KEN
KOR
KOR, S
MAS
MRI
MEX
NED
NOR
NZL
MYA
LUX
PHI
POL
POR
ROM
RUS
ESP
SIN
SUI
SWE
TPE
THA
TUR
USA


Country comparison
Variance adjusted distance from center
2
4
6
8
0
10
20
30
40
50
Index
SAM
COK
PNG
USA
SIN
RUS
KOR, N
CHN
GER
GUA
CZE
FRA
GBR
CRO
DOM
AUS
BER
CAN
AUT
BEL
CHI
BRA
COL
ARG
DEN
FIN
GRE
INA
HUN
IND
ISR
IRL
ITA
JPN
KEN
KOR, S
MAS
MRI
LUX
MEX
NED
MYA
NOR
NZL
PHI
POL
ROM
POR
ESP
SUI
SWE
TPE
THA
TUR



For all three distance measures we note that the three pacific island nations PNG, SAM and COK are among the most extreme nations. In the case of euclidean distance measure a distant fourth most extreme we find Bermuda (BER) and fifth we find Great Britain (GBR). Keep in mind tough, we have not compensated for the different magnitudes of records yet by compensating for the variance experienced by the records for each distance.

In the standardized euclidean distance measure case, where variance of records in each distance is compensated for, we find three pacific islanders, USA and Singapore (SIN) to be the most extreme.

In the Mahalanobis distance measure case, where we compensate for covariance of records, we find beside three pacific islanders, North Korea (KOR, N) to be very extreme record nation. A distant fifth we find Mexico (MEX). Given the nature of North Korea, one can not rule out the case that they have fabricated running records.

Appendix - R-Code

```r
library(rgl)
library(scatterplot3d)
library(ggplot2)
www = "http://www.ida.liu.se/~732A37/T1-9.dat"
data <- read.delim(www, header = FALSE, sep="\t")
nations <- as.character(data[,1])
data <- data[,-1]
colnamez <- c("100m(s)","200m(s)","400m(s)",
                    "800m(min)","1500m(min)","3000m(min)","Mara(min)")
colnames(data) <- colnamez
col_mu <- round(colMeans(data),3)

col_sigmasquared<- apply(data,2,var)

mins <- apply(data,2,min)
maxs <- apply(data,2,max)
print('Means')
col_mu
paste("standard deviations:" )
round(sqrt(col_sigmasquared),3)
paste("Min values:" )
mins
paste("Max values:" )
maxs
par(mfrow = c(2,4))
for(i in 1:3){
boxplot(data[,i],ylab="seconds",main=c("Records for distance:",
                                            colnames(data)[i]))
}
for(j in 4:7){
boxplot(data[,j],ylab="minutes",main=c("Records for distance:",
                                            colnames(data)[j]))
}
plot.new()

par(mfrow = c(2,4))
for(i in 1:length(data)){
  plot(seq_along(data[,i]),data[,i],main=c("Records for distance:",
                                            colnames(data)[i]),xlab="",ylab="time unit")
}

plot.new()
for(i in 1:length(data)){
  qqnorm(data[,i],main=c("Q-Q plot of records",
                                            colnames(data)[i]))
  qqline(data[,i])
}
par(mfrow = c(1,1))
covmatr <- var(data)
corrmatr <- cor(data)
```

```
print("Covariance matrix:")
signif(covmatr,3)
print("Correlation matrix:")
signif(corrmatr,3)
pairs(data)
## for(i in 1:(length(data)-2)){
##   plot3d(data[,i],data[,i+1],data[,i+2],
##                  xlab="even smaller distance record",ylab="smaller distance record",
##          zlab="record",col="red",size=3,type="s")
## }
##
par(mfrow = c(1,1))
plot(data[,3],data[,4],pch="", main="Country comparison",xlab=colnames(data)[3],ylab=colnames(data)[4])
text(data[,3],data[,4],labels=nations)
plot(data[,1],data[,7],pch="", main="Country comparison",xlab=colnames(data)[1],ylab=colnames(data)[7])
text(data[,1],data[,7],labels=nations)
plot(data[,5],data[,6],pch="", main="Country comparison",xlab=colnames(data)[5],ylab=colnames(data)[6])
text(data[,5],data[,6],labels=nations)
centerdata <- data - matrix(rep(col_mu, dim(data)[1]),ncol=7,byrow=TRUE)
centerdata <- as.matrix(centerdata)
euclideandist <- c()
varianceadjusteddists <- c()
V <- diag(col_sigmasquared)
covarianceadjusteddists <-c()
for(i in 1:dim(data)[1]){
  euclideandist <- c( euclideandist, sqrt(sum(centerdata[i,] * centerdata[i,])))
  varianceadjusteddists <- c( varianceadjusteddists,
            sqrt(centerdata[i,] %*% solve(V) %*% centerdata[i,]))
  covarianceadjusteddists <- c( covarianceadjusteddists,
            sqrt(centerdata[i,] %*% solve(covmatr) %*% centerdata[i,]))
}

plot(euclideandist,pch="", main="Country comparison",ylab="Euclidean distance from center")
text(euclideandist,labels=nations)
plot(varianceadjusteddists,pch="",
     main="Country comparison",ylab="Variance adjusted distance from center")
text(varianceadjusteddists,labels=nations)
plot(covarianceadjusteddists,pch="",
     main="Country comparison",ylab="Covariance adjusted distance from center")
text(covarianceadjusteddists,labels=nations)
##
##
```