EuroGroup

Contents

Overview

The data file 'europegroup.txt' contains data for the percentage of employment by country (n=30 countries). The first variable identifies the region of the country (Group) and the next nine variables represent different employment sectors: AGR=agriculture, forests and fishing; MIN=mining; MAN=manufacturing; PS=power and water supplies; CON=construction; SER=services; FIN=finance; SPS=social and personal services; TC=transport and communication.

Testing For Multivariate Normality (MVN)

Structure of the data

```
#read and preview data
library(RCurl)
```

```
## Warning: package 'RCurl' was built under R version 3.6.3
```

```
online_data <- getURL("https://raw.githubusercontent.com/harjomand/R/main/Data/europegroup.txt")
eg <- read.table(text = online_data, header = TRUE)
str(eg)
```

```
## 'data.frame':    30 obs. of  10 variables:
##  $ Group: Factor w/ 4 levels "Eastern","EFTA",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ AGR  : num  2.6 5.6 5.1 3.2 22.2 13.8 8.4 3.3 4.2 11.5 ...
##  $ MIN  : num  0.2 0.1 0.3 0.7 0.5 0.6 1.1 0.1 0.1 0.5 ...
##  $ MAN  : num  20.8 20.4 20.2 24.8 19.2 19.8 21.9 19.6 19.2 23.6 ...
##  $ PS   : num  0.8 0.7 0.9 1 1 1.2 0 0.7 0.7 0.7 ...
##  $ CON  : num  6.3 6.4 7.1 9.4 6.8 7.1 9.1 9.9 0.6 8.2 ...
##  $ SER  : num  16.9 14.5 16.7 17.2 18.2 17.8 21.6 21.2 18.5 19.8 ...
##  $ FIN  : num  8.7 9.1 10.2 9.6 5.3 8.4 4.6 8.7 11.5 6.3 ...
##  $ SPS  : num  36.9 36.3 33.1 28.4 19.8 25.5 28 29.6 38.3 24.6 ...
##  $ TC   : num  6.8 7 6.4 5.6 6.9 5.8 5.3 6.8 6.8 4.8 ...
```

```
attach(eg)
```

The data consists of measurements of 30 cases (countries) and 10 variables. The Group variable is a factor with 4 levels describing the region of the country. The other nine variables are numerical measuring %employment in 9 different employment sectors. No missing data.

Univariate QQ plots and histograms and univariate Shapiro-Wilks tests

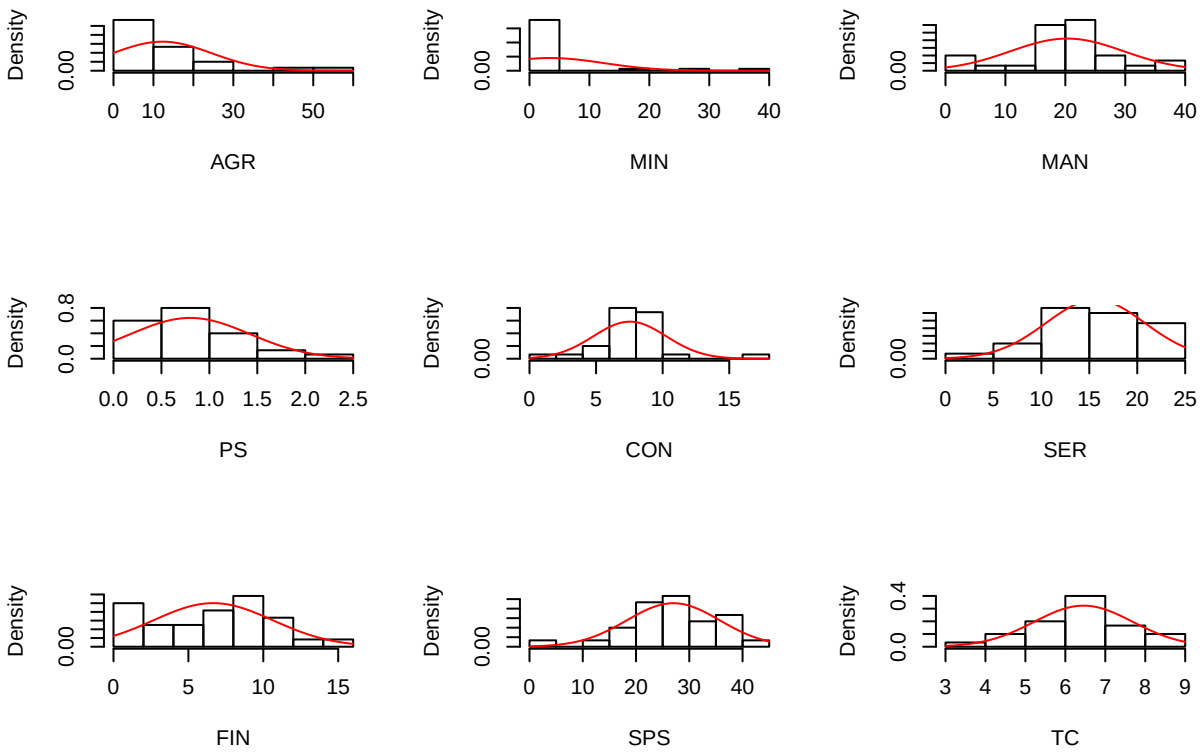
```
library(MVN)
```

```
## Warning: package 'MVN' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## sROC 0.1-2 loaded
```

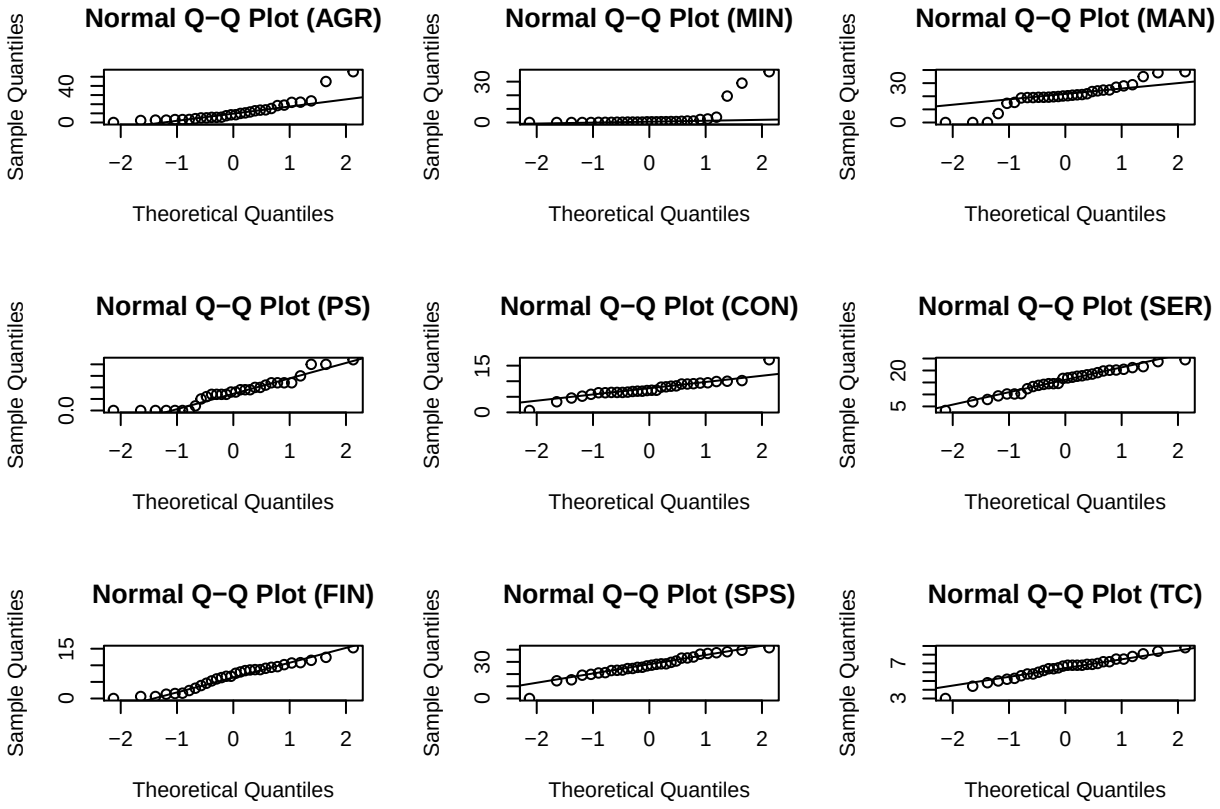
```
mvn(eg[,2:10],univariatePlot="histogram")
```




```
## $multivariateNormality
##               Test        Statistic              p value Result
## 1 Mardia Skewness 297.612934431293 1.14990159186688e-09     NO
## 2 Mardia Kurtosis 3.15252502688008  0.00161864945575441     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    AGR       0.7611  <0.001      NO
## 2 Shapiro-Wilk    MIN       0.4272  <0.001      NO
## 3 Shapiro-Wilk    MAN       0.9004  0.0086      NO
## 4 Shapiro-Wilk    PS        0.9175  0.0232      NO
## 5 Shapiro-Wilk    CON       0.8913  0.0052      NO
## 6 Shapiro-Wilk    SER       0.9735  0.6384      YES
## 7 Shapiro-Wilk    FIN       0.9624  0.3562      YES
## 8 Shapiro-Wilk    SPS       0.9513  0.1829      YES
## 9 Shapiro-Wilk    TC        0.9710  0.5656      YES
##
## $Descriptives
##      n      Mean    Std.Dev Median Min  Max   25th   75th        Skew
## AGR 30 12.186667 12.3069012   8.45 0.0 55.5  4.400 14.925  1.98436586
## MIN 30  3.446667  8.8657315   0.50 0.0 37.3  0.125  1.050  2.80444530
## MAN 30 20.286667  9.4567886  20.30 0.0 38.7 19.000 24.550 -0.44555182
## PS  30  0.800000  0.6209003   0.80 0.0  2.2  0.275  1.175  0.41860269
## CON 30  7.530000  2.7330859   7.05 0.6 16.9  6.400  9.100  0.72113883
## SER 30 15.636667  5.1601579  16.80 3.3 24.5 12.625 19.625 -0.41261682
```

```
## FIN 30  6.650000  3.9866804   7.15 0.0 15.3  3.300  9.325 -0.04249182
## SPS 30 26.993333  8.7320627  27.00 0.0 41.6 22.950 33.175 -0.71956875
## TC  30  6.453333  1.2333675   6.75 3.0  8.8  5.800  7.150 -0.55724363
##       Kurtosis
## AGR  3.9380917
## MIN  6.7127969
## MAN  0.3867478
## PS  -0.4544634
## CON  3.3313202
## SER -0.5827700
## FIN -0.9449859
## SPS  1.0420028
## TC   0.3949991
```

```
mvn(eg[,2:10], univariateTest="SW",univariatePlot="qqplot")
```




```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 297.612934431293 1.14990159186688e-09     NO
## 2 Mardia Kurtosis 3.15252502688008  0.00161864945575441     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    AGR       0.7611  <0.001      NO
```

```
## 2 Shapiro-Wilk    MIN       0.4272  <0.001      NO
## 3 Shapiro-Wilk    MAN       0.9004  0.0086      NO
## 4 Shapiro-Wilk    PS        0.9175  0.0232      NO
## 5 Shapiro-Wilk    CON       0.8913  0.0052      NO
## 6 Shapiro-Wilk    SER       0.9735  0.6384      YES
## 7 Shapiro-Wilk    FIN       0.9624  0.3562      YES
## 8 Shapiro-Wilk    SPS       0.9513  0.1829      YES
## 9 Shapiro-Wilk    TC        0.9710  0.5656      YES
##
## $Descriptives
##      n      Mean    Std.Dev Median Min  Max   25th   75th        Skew
## AGR 30 12.186667 12.3069012   8.45 0.0 55.5  4.400 14.925  1.98436586
## MIN 30  3.446667  8.8657315   0.50 0.0 37.3  0.125  1.050  2.80444530
## MAN 30 20.286667  9.4567886  20.30 0.0 38.7 19.000 24.550 -0.44555182
## PS  30  0.800000  0.6209003   0.80 0.0  2.2  0.275  1.175  0.41860269
## CON 30  7.530000  2.7330859   7.05 0.6 16.9  6.400  9.100  0.72113883
## SER 30 15.636667  5.1601579  16.80 3.3 24.5 12.625 19.625 -0.41261682
## FIN 30  6.650000  3.9866804   7.15 0.0 15.3  3.300  9.325 -0.04249182
## SPS 30 26.993333  8.7320627  27.00 0.0 41.6 22.950 33.175 -0.71956875
## TC  30  6.453333  1.2333675   6.75 3.0  8.8  5.800  7.150 -0.55724363
##       Kurtosis
## AGR  3.9380917
## MIN  6.7127969
## MAN  0.3867478
## PS  -0.4544634
## CON  3.3313202
## SER -0.5827700
## FIN -0.9449859
## SPS  1.0420028
## TC   0.3949991
```

The first 4 variables are not UVN while the last 4 are according to the SW tests. From the histograms and QQplots AGR, MIN and PS are all skewed right while MAN appears to have very heavy tails and may be bimodal. Based on the S west both AGR and MIN are equally significantly deviating from normality ($p<0.001$) however, MIN has the lowest test statistic (SW=0.427) and the QQplot shows that it is most strongly right skewed and the histogram shows that nearly all observations are clumped together at the lower end with just a few large outliers

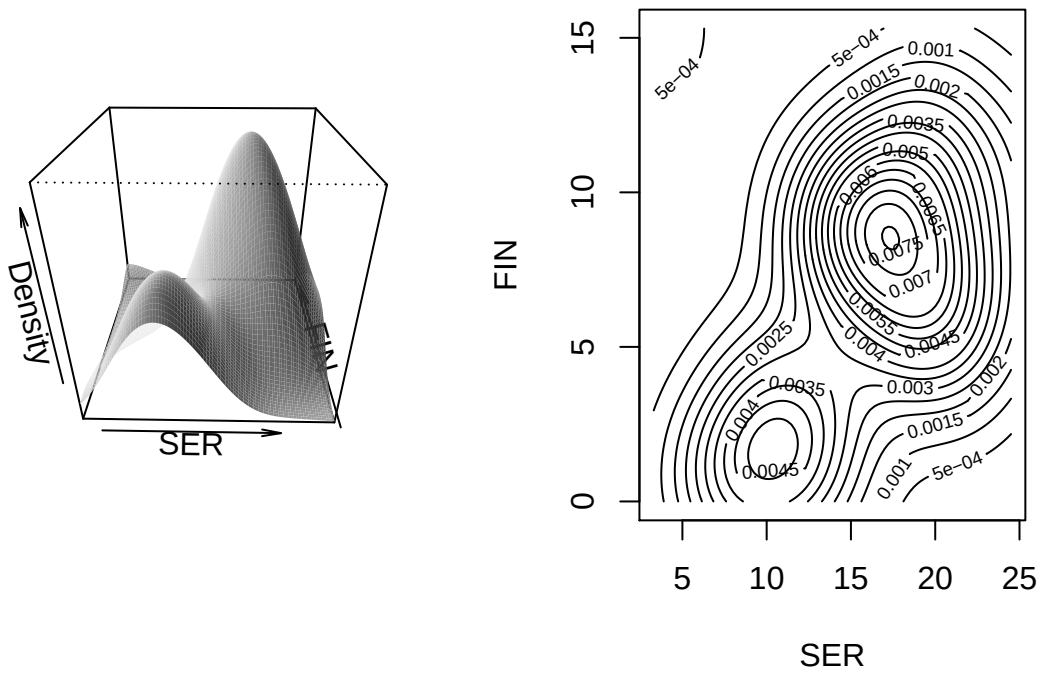
Perspective and contour plots for the SER and FIN variables

```
#perspective and contour plots for FIN and SER
par(mfrow = c(1,2))
mvn(eg[,7:8], multivariatePlot="persp")
```

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 24.3102345291055 6.92140669736533e-05     NO
## 2 Mardia Kurtosis 3.56453295161054 0.000364504659268849     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
```

```
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    SER       0.9735    0.6384    YES
## 2 Shapiro-Wilk    FIN       0.9624    0.3562    YES
##
## $Descriptives
##      n     Mean  Std.Dev Median Min  Max   25th   75th        Skew   Kurtosis
## SER 30 15.63667 5.160158  16.80 3.3 24.5 12.625 19.625 -0.41261682 -0.5827700
## FIN 30  6.65000 3.986680   7.15 0.0 15.3  3.300  9.325 -0.04249182 -0.9449859
```

```
mvn(eg[,7:8], multivariatePlot="contour")
```



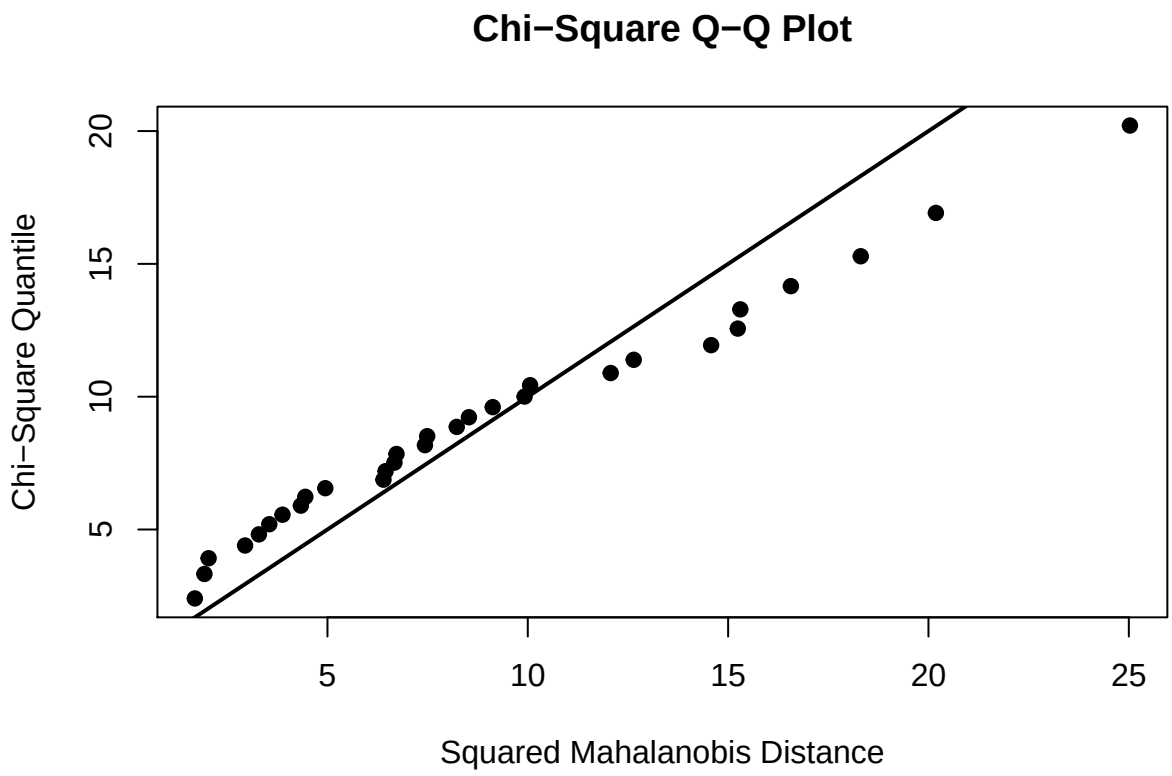

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 24.3102345291055 6.92140669736533e-05     NO
## 2 Mardia Kurtosis 3.56453295161054 0.000364504659268849     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    SER       0.9735    0.6384    YES
## 2 Shapiro-Wilk    FIN       0.9624    0.3562    YES
##
## $Descriptives
##      n     Mean  Std.Dev Median Min  Max   25th   75th        Skew   Kurtosis
## SER 30 15.63667 5.160158  16.80 3.3 24.5 12.625 19.625 -0.41261682 -0.5827700
## FIN 30  6.65000 3.986680   7.15 0.0 15.3  3.300  9.325 -0.04249182 -0.9449859
```

```r
par(mfrow = c(1,1))
```

Both variables were UVN however they are not jointly MVN. Their joint distribution looks like it has one main peak at the centre of the distribution for FIN, but slightly off-centre for SER. There is also a second peak 'in front' the first and both the perspective and the contour plot show that the 'valley' between the two peals is relatively deep. The joint distribution is quite 'peaked' or leptopkurtic rather than platykurtic (flat). The inherent problem with these plots is that they only consider bivariate normality between 2 selected variables at a time from the dataset, not MVN of the whole dataset.

Applying Mardia, Henze-Zirkler and Royston tests of MVN based on all nine employment variables

```r
mvn(eg[,2:10], mvnTest="mardia", desc=FALSE, multivariatePlot="qq")
```




```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 297.612934431293 1.14990159186688e-09     NO
## 2 Mardia Kurtosis 3.15252502688008  0.00161864945575441     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
```

```
## 1 Shapiro-Wilk    AGR       0.7611  <0.001      NO
## 2 Shapiro-Wilk    MIN       0.4272  <0.001      NO
## 3 Shapiro-Wilk    MAN       0.9004  0.0086      NO
## 4 Shapiro-Wilk    PS        0.9175  0.0232      NO
## 5 Shapiro-Wilk    CON       0.8913  0.0052      NO
## 6 Shapiro-Wilk    SER       0.9735  0.6384      YES
## 7 Shapiro-Wilk    FIN       0.9624  0.3562      YES
## 8 Shapiro-Wilk    SPS       0.9513  0.1829      YES
## 9 Shapiro-Wilk    TC        0.9710  0.5656      YES
```

The Chi-Square QQ plot : The joint distribution of all 4 variables does not appear to be MVN from the Chi-square QQ plot – countries with larger mahalanobis distances deviate most from expected.

Mardia's MVN for skew is highly significant ($p<0.0001$) indicating significant deviation from expected under MVN. The kurtosis statistic is also significant ($p>0.01$), suggesting that kurtosis (peakedness) or lack of, is also a problem in the MV distribution. Note: kurtosis p-value is < 0.05 means the distribution is significantly different from normal –i.e. there is excessive kurtosis. Platykurtic distributions have negative excessive kurtosis (tests statistic is negative) because they are too flat. Leptokurtic distributions have positive excessive kurtosis because they are too peaked (test statistic is positive) Overall the Mardia's test suggests that there is significant deviation from MVN.

```
mvn(eg[,2:10], mvnTest="hz", desc=FALSE)
```

```
## $multivariateNormality
##            Test       HZ     p value MVN
## 1 Henze-Zirkler 1.122414 8.01692e-13  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    AGR       0.7611  <0.001      NO
## 2 Shapiro-Wilk    MIN       0.4272  <0.001      NO
## 3 Shapiro-Wilk    MAN       0.9004  0.0086      NO
## 4 Shapiro-Wilk    PS        0.9175  0.0232      NO
## 5 Shapiro-Wilk    CON       0.8913  0.0052      NO
## 6 Shapiro-Wilk    SER       0.9735  0.6384      YES
## 7 Shapiro-Wilk    FIN       0.9624  0.3562      YES
## 8 Shapiro-Wilk    SPS       0.9513  0.1829      YES
## 9 Shapiro-Wilk    TC        0.9710  0.5656      YES
```

```
mvn(eg[,2:10], mvnTest="royston", desc=FALSE)
```

```
## $multivariateNormality
##      Test        H      p value MVN
## 1 Royston 63.91799 1.972026e-11  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    AGR       0.7611  <0.001      NO
## 2 Shapiro-Wilk    MIN       0.4272  <0.001      NO
## 3 Shapiro-Wilk    MAN       0.9004  0.0086      NO
## 4 Shapiro-Wilk    PS        0.9175  0.0232      NO
## 5 Shapiro-Wilk    CON       0.8913  0.0052      NO
## 6 Shapiro-Wilk    SER       0.9735  0.6384      YES
```

```
## 7 Shapiro-Wilk    FIN       0.9624  0.3562      YES
## 8 Shapiro-Wilk    SPS       0.9513  0.1829      YES
## 9 Shapiro-Wilk    TC        0.9710  0.5656      YES
```

Both the HZ and Royston tests also conclude that there is significant deviation from MVN. Key limitation of these tests : To meet the MVN assumption we 'want' to accept the null hypothesis and we can not control Type II error rate.

Meeting the MVN assumption

One way to try and meet the MVN assumption could be to remove some of the variables from the multivariate analysis. Let's check if the data is MVN if only those variables that are univariate normal (UVN) are used:

```
mvn(eg[,7:10], mvnTest="mardia", desc=FALSE)
```

```
## $multivariateNormality
##              Test        Statistic              p value Result
## 1 Mardia Skewness 80.5121769695863 3.21311530240524e-09     NO
## 2 Mardia Kurtosis 3.52317642864713 0.000426407408346208     NO
## 3             MVN             <NA>                  <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    SER       0.9735    0.6384    YES
## 2 Shapiro-Wilk    FIN       0.9624    0.3562    YES
## 3 Shapiro-Wilk    SPS       0.9513    0.1829    YES
## 4 Shapiro-Wilk    TC        0.9710    0.5656    YES
```

The data is still not MVN. Just because variables are UVN there is no expectation that they will also be jointly MVN, so this outcome is reasonable.

MANOVA

Fr rest of the project, I do not use all nine employment variables. I use only those 4 identified aboved as UVN.

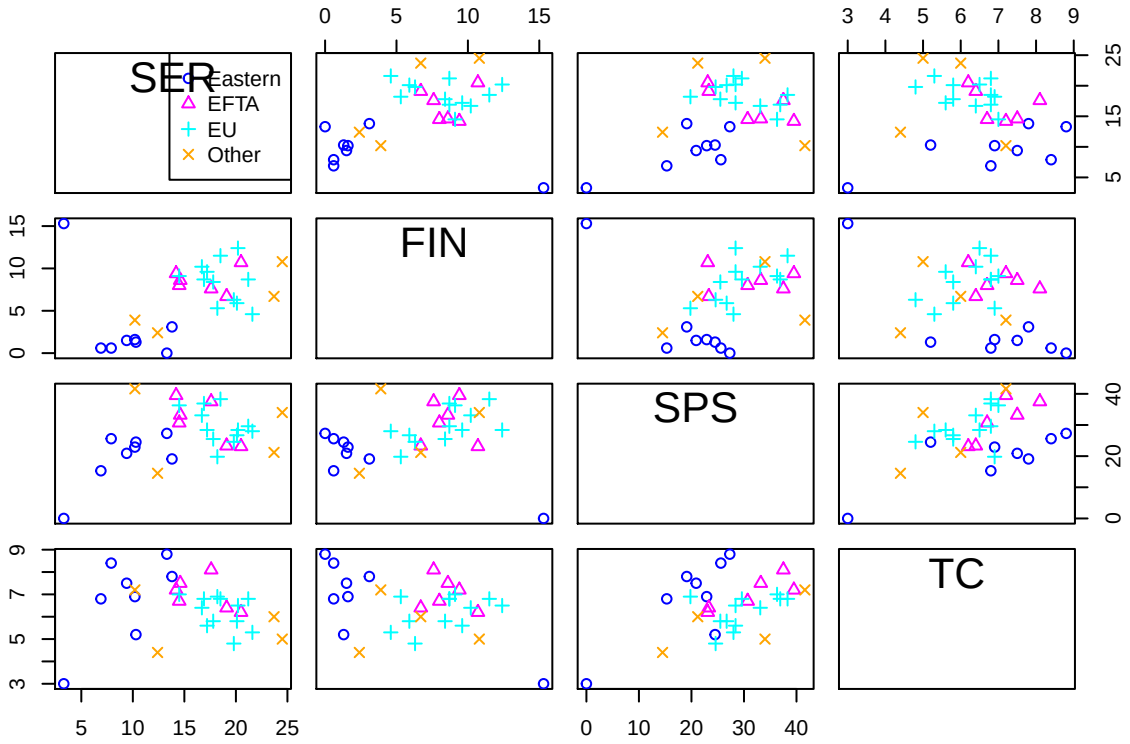
Daftsman display

I plot the observations grouped by regional group. I also include smoothing, regression lines, or distribution curves in the diagonal panels of the plot

```
# Plot draftsman
library(car)
```

```
## Loading required package: carData
```

```
scatterplotMatrix(eg[7:10], groups=eg$Group, smooth=FALSE, regLine=FALSE, diagonal=FALSE)
```



There is one Eastern (dark blue) country that has a much higher FIN value and lower values for SER, SPS and TC. Looking back at the original data this is country 19. While there is a fair bit of overlap between regions (Groups) it also looks like there is some separation between regions (colours) in the bivariate relationships, particularly between Eastern and EU. Eastern countries tend to have lower SER and FIN (except country 19) and higher TC.

Applying MANOVA:

The 4 dependent variables are SER, FIN, SPS AND TC which measure the percentage employment within each of these 4 employment sectors for 30 countries. The independent variable is Group which is a factor that represents the country regions. The MANOVA tests whether there is a difference between the vector of dependent variable means among the country groups based on the ratio of between and within group variances among the dependent variables.

Checking for for differences in 'percentage of employment' between the four country regions

```
eg.manova1<-manova(cbind(SER,FIN,SPS,TC) ~ factor(Group), data=eg)
summary(eg.manova1) #default test is Pillai's
```

```
##               Df Pillai approx F num Df den Df   Pr(>F)
## factor(Group)  3 1.0329   3.2819     12     75 0.000763 ***
## Residuals     26
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(eg.manova1, test="Wilks")
```

```
##               Df   Wilks approx F num Df den Df    Pr(>F)
## factor(Group)  3 0.17678   4.7134     12 61.144 2.184e-05 ***
## Residuals     26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(eg.manova1, test="Roy")
```

```
##               Df    Roy approx F num Df den Df    Pr(>F)
## factor(Group)  3 3.1076   19.423      4     25 2.236e-07 ***
## Residuals     26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(eg.manova1, test="Hotelling-Lawley")
```

```
##               Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## factor(Group)  3           3.4813   6.2856     12     65 3.206e-07 ***
## Residuals     26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assuming MVN distribution and equal covariance matrices, Pillai's, Wiks', Roys' and Hotelling's statistics are all significant ($p<0.001$) indicating that at least two of the three country regions (Groups) differ significantly in composition of their percent employment across the four employment sectors

Comparing each of the regions (Group) with each other

I use Hotelling's T2 test and a significance level of 0.05 to produce the following table:

Comparison	Hotelling's p-value	Significant (Y/N)	Significant after correction (Y/N)
EU-EFTA	0.249	No	No
EU-Eastern	2.253e-07	Yes	Yes
EU-Other	0.5728	No	No
EFTA-Eastern	0.0001096	Yes	Yes
EFTA-Other	0.02368	Yes	No
Eastern-Other	0.07146	No	No

These sample sizes are quite unbalanced with the biggest difference between EU (n=12) and Other (n=4) countries. MANOVA and Hotelling's T2 t-test are fairly robust to deviations from MVN and equal covariance when sample sizes are roughly equal. The sample size differences in this analysis could have caused issues with the significance tests. However, the comparison between EU and Other is non-significant. Two of the significant comparisons are highly significant ($p<0.0001$), which would indicate that there is strong evidence

for the distinction between these regions. The third significant comparison between EFTA v Other becomes non-significant once the correction for multiple testing is applied. Therefore, the lack of MVN is unlikely to be of concern when interpreting the significant results (i.e. Type I error not a concern) but the lack of MVN may have led to erroneous non-significant results (Type II error)

PCA analysis

Checking the correlation and covariance matrices

```
(egcov<-round(cov(eg[,7:10]), digits=2))
```

```
##       SER   FIN   SPS    TC
## SER 26.63  7.80 17.48 -0.54
## FIN  7.80 15.89  5.78 -1.92
## SPS 17.48  5.78 76.25  5.11
## TC  -0.54 -1.92  5.11  1.52
```

```
(egcor<-round(cor(eg[,7:10]), digits=2))
```

```
##       SER   FIN  SPS    TC
## SER  1.00  0.38 0.39 -0.08
## FIN  0.38  1.00 0.17 -0.39
## SPS  0.39  0.17 1.00  0.47
## TC  -0.08 -0.39 0.47  1.00
```

The covariance matrix is calculated based on data where the mean of each variable is subtracted from each observation in that variable. This centres each variable around a mean of 0. The correlation matrix is calculated based on data where the centred values are divided by the relevant variable standard deviation. This gives a unit variance (variance=1) within each variable. There can be a benefit in centring and standardising variables if the MV dataset is made up of variables measured on very different scales (units) or measured on similar scales but with very different magnitudes of variances – disparity in scale and variance can lead to one (or a few) variables dominating analysis). In the correlation matrix values have been centred and standardised, but only centred in the covariance matrix. The europegroup dataset contains variables measured in the same units (% employed) however the range of values varies quite a bit among countries, so the covariance matrix would not be appropriate for PCA analysis. The much larger variance of the variable 'SPS' in the covariance matrix (76.25) compared to TC (1.52) is an example of the scale of one variable potentially overwhelming the effect of another variable. Overall the correlations between variables are moderate to low (less than 0.47) indicating that PCA may not return very informative results. There two negative correlations between TC and SER and FIN

PCA analysis

I use the prcomp function fo PCA analysis.

```
(eg.prcomp <- prcomp(eg[,7:10], center=TRUE, scale=TRUE ))
```

```
## Standard deviations (1, .., p=4):
## [1] 1.2775894 1.2411057 0.7408298 0.5278194
```

```
##
## Rotation (n x k) = (4 x 4):
##             PC1         PC2        PC3        PC4
## SER  0.65367075 -0.01871173 -0.7263543  0.2115983
## FIN  0.55609574 -0.39632844  0.6221862  0.3828389
## SPS  0.51103815  0.52598829  0.2638248 -0.6265563
## TC  -0.04808409  0.75226987  0.1252609  0.6450486
```

```
names(eg.prcomp)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
#eigen values
(eigen<-(eg.prcomp$sdev)^2)
```

```
## [1] 1.6322347 1.5403433 0.5488287 0.2785933
```

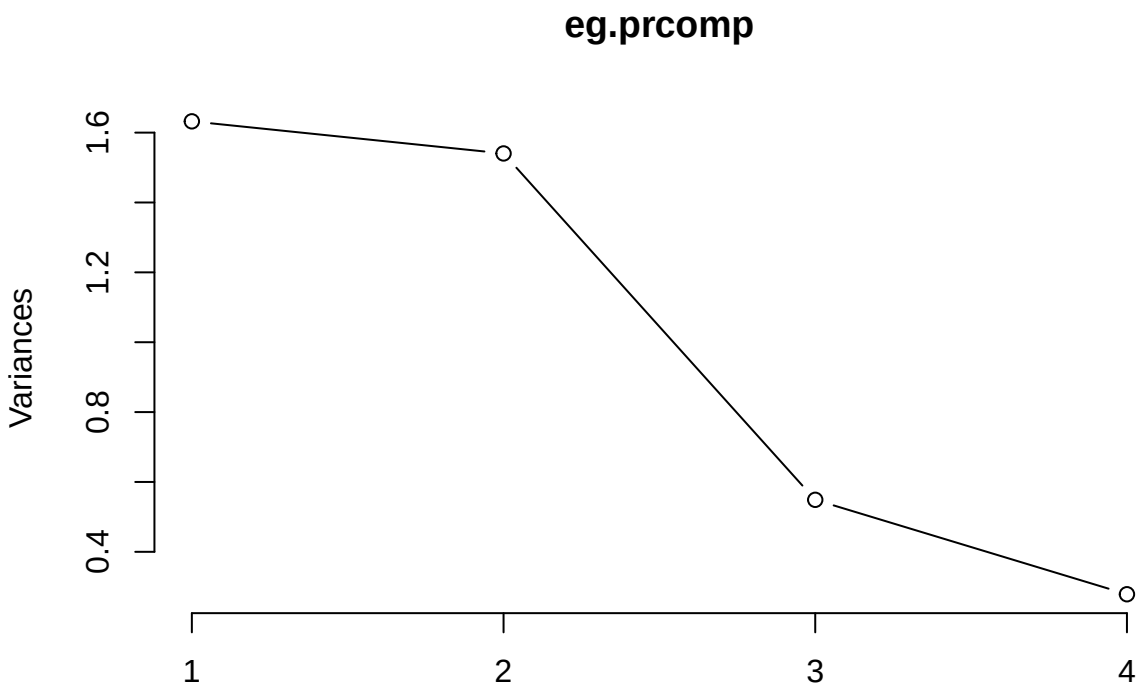
```
# %variance
(pervar<-((eg.prcomp$sdev^2/sum(eg.prcomp$sdev^2))*100))
```

```
## [1] 40.805867 38.508583 13.720718  6.964832
```

```
(cumsum<-cumsum(pervar))
```

```
## [1]  40.80587  79.31445  93.03517 100.00000
```

```
#Screeplot
screeplot(eg.prcomp, type="lines")
```



The first 2 PC's have eigen values>1 and explain 79.3% of the total variance. The scree plot shows an elbow at 3 PC's which would also indicate interpreting only the first 2 PCs Reducing dimensions (the overall purpose of PCA) from 4 to 2 is quite a good result while maintaining a substantial %var explained.

Interpreting the first PC

Z1 = 0.6537(SER) + 0.5561(FIN) + 0.5110(SPS) - 0.0481(TC)

On PC1 SER, FIN and SPS are all moderately positively correlated, while TC is weakly negatively correlated with the other 3 variables. This suggests that the services (SER) sector, the financial sector (FIN) and the social and personal services sector (SPS) all have relatively similar values across countries, while employment percentages in the transport and communications sector is not strongly related to these sectors and does not contribute to the overall % variation explained by this PC. It is important to note that this PC explains only 40.8% of the total variation so caution should be applied to the above interpretation.

Checking the correlation between the first and second PCs

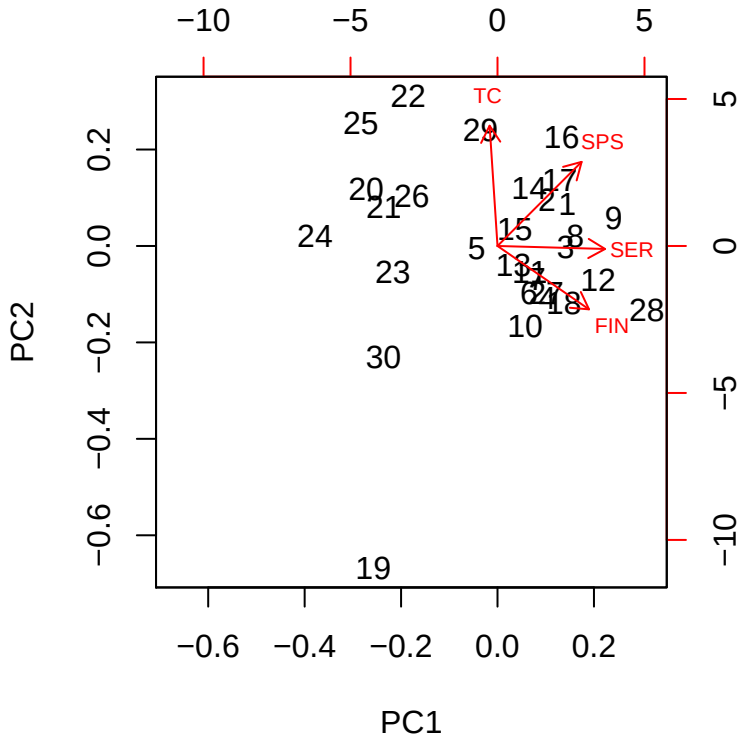
```
cor(eg.prcomp$x[,1],eg.prcomp$x[,2])
```

```
## [1] -1.068474e-16
```

The correlation coefficient is practically zero confirming that the two PCs are orthogonal.

Biplot based on the first 2 PCs

```
biplot(eg.prcomp,cex=c(1,0.7))
```



The biplot shows each of the countries by number in the main plot space. The red vectors indicate the influence of the original variables on the position of countries within the 2D space. From the biplot we can see that TC is associated strongly with PC2 – countries close to the top of the plot are associate with high values of TC. SER is most influential along PC1. Country 19 is associated with the lowest values of TC, SER and SPS and the highest value of FIN. Country 9, in contrast, has moderate to high values for all 4 variables.

Factor Analysis

I use the the factanal function to perfrom Factor Analysis

```
#factanal function using 1 factor and no rotation
(eg.fa3 <- factanal(eg[,7:10], factors=1, rotation="none" ))
```

```
##
## Call:
## factanal(x = eg[, 7:10], factors = 1, rotation = "none")
##
## Uniquenesses:
```

```
##   SER   FIN   SPS    TC
## 0.849 0.973 0.005 0.774
##
## Loadings:
##     Factor1
## SER 0.388
## FIN 0.166
## SPS 0.997
## TC  0.475
##
##                Factor1
## SS loadings      1.399
## Proportion Var   0.350
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 13.35 on 2 degrees of freedom.
## The p-value is 0.00126
```

There are 6 combinations or 6 df available at start.

For F1 we need 1 df for the eigen value and 3 df for the eigen vector (4-1) leaving 6-4=2 df. For F2 we need 1 df for the eigen value and 2df for the eigen vector (3-1) leaving 2-3= -1 df. Therefore, the maximum number of factors we can fit and still rn the chi-sq test is 1 factor.

```
print(eg.fa3$loadings,cutoff=0.5)
```

```
##
## Loadings:
##     Factor1
## SER
## FIN
## SPS 0.997
## TC
##
##                Factor1
## SS loadings      1.399
## Proportion Var   0.350
```

The 1 factor explains 35% of the total variance which is not very good. The chi-square goodness of fit test also indicates that the model deviates significantly from the data ($p<0.00126$) which suggests that the unexplained variance is important, and interpretation of the 1 factor model may not accurately reflect the original data. Only the SPS variable has a loading >0.5 on this factor and at 0.997 this factor is explaining almost all the variance associated with SPS. In comparison the loadings for TC is only moderate at 0.475. The uniqueness is the proportion of variance for a variable not explained by the factor model. For SPS has very little left unexplained by the factor (0.005) while FIN is nerly completely unexplained by the factor (0.973).

Chi-square test: The chi-square goodness of fit test determines how well the factor model reproduces the correlation matrix. The chi-square test indicates that the hypothesis of perfect model fit is rejected. The chi-squared test result is 0.00126 so we should reject the null hypothesis as the p-value is less than 0.05 . Therefore the model with 1 factor cannot sufficiently represents the data. We are losing too much information, its not really describing all the important relationships among our variables.

Variance explained: From the Proportion Variance for Factor 1 we can see that this model only accounts for 35% of the total variance which is low. If we are satisfied with the model explaining this % variation we can proceed with some caution.

Variable loadings: We need to consider how each variable is loaded on our factor. There is no loadings less than 0.1 . We see that SPS is highly loaded on our factor. While SER and TC have a medium load. FIN has the least loading on the factor.

Difference in uniqueness values for the variables FIN and SPS: SPS has uniqueness value of 0.005 and FIN is 0.973 . Uniqueness values are specificity values .Therefore uniqueness is the variance in each variable not explained by the common factors. Large specificity would indicate that the factors are not specific for that Xi. So its not specific for FIN. Also FIN has the least loading on the factor. So SPS is variation explained by the common factor but FIN was not explained.

How about a Rotation ?

How would our results change if we applied a rotation?

```
(eg.fa3 <- factanal(eg[,7:10], factors=1, rotation="varimax" ))
```

```
##
## Call:
## factanal(x = eg[, 7:10], factors = 1, rotation = "varimax")
##
## Uniquenesses:
##   SER   FIN   SPS    TC
## 0.849 0.973 0.005 0.774
##
## Loadings:
##     Factor1
## SER 0.388
## FIN 0.166
## SPS 0.997
## TC  0.475
##
##                Factor1
## SS loadings      1.399
## Proportion Var   0.350
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 13.35 on 2 degrees of freedom.
## The p-value is 0.00126
```

When we apply a rotation , it tries to maximise the differences between the loadings . It will take the variables with high loadings on factor 1 and variables with low loadings and tries to exacerbate the differences between them.

Applying the varimax rotation does not change the chi-square statistic or significance or the Uniquenesses and the subsequent communalities or final cumulative percentage variance (individual percentage contributions do change). Varimax only enhances large loadings and minimises small loadings.

In our case, there was no change in our loadings.

Parallel analysis

An alternative way to determine the number of factors to interpret is to compare the solution to normally distributed random data with the same properties as the real data set - that is the same number of original

variables and the same number of cases (sample size). We can produce the expected eigen values from an 'ideal' sample and if we do this multiple times, say 500 simulations, we can get a distribution of possible eigen values for each component/factor; each with a mean, SD and 95 percentile value. If our 'real data' eigen value is greater than the 95 percentile value we can be confident our factor is a well supported factor.

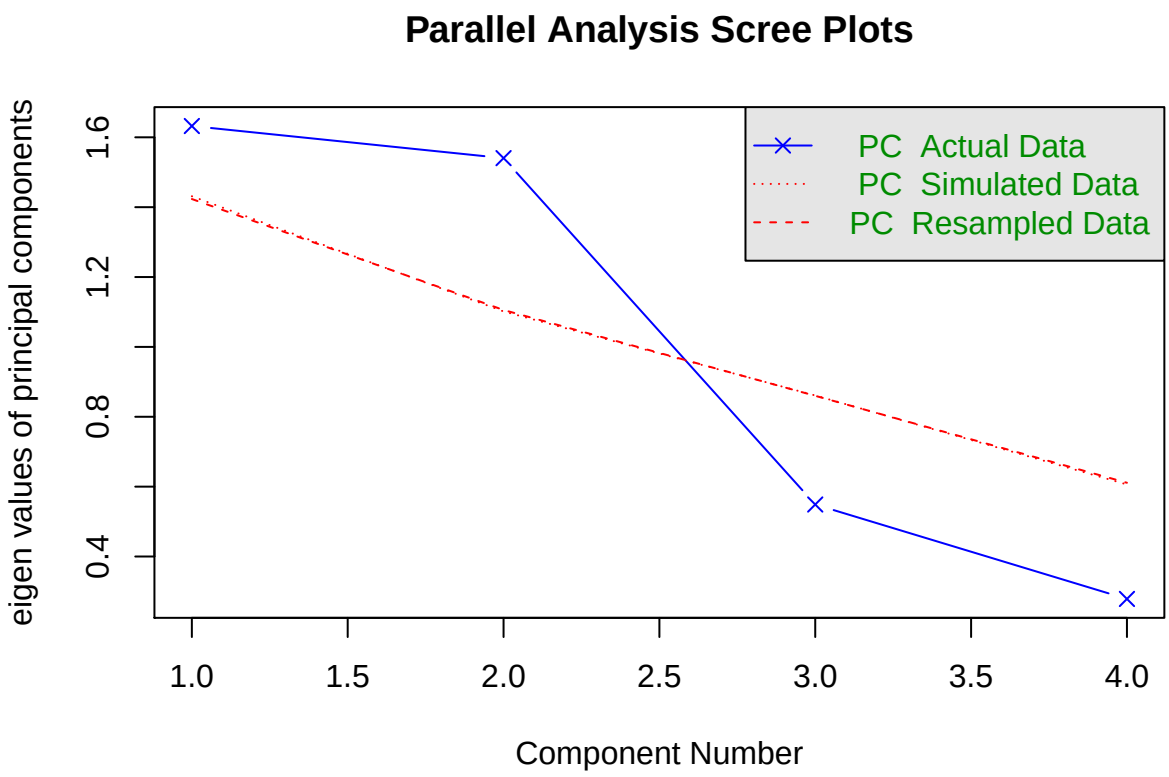
```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.3
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
set.seed(245)
pa<-fa.parallel(eg[,7:10], fm="ml", fa="pc", n.iter=500)
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  0
```

```
pa$pc.values
```

```
## [1] 1.6322347 1.5403433 0.5488287 0.2785933
```

```
pa$pc.sim
```

```
## [1] 1.4312552 1.1015395 0.8609970 0.6062083
```

```
pa.out<-pa$values
quants <- c(0.95)
(pa_95quant<-apply( pa.out[,5:8], 2 , quantile , probs = quants ))
```

```
##      Sim5      Sim6      Sim7      Sim8
## 1.6919020 1.2574337 0.9859400 0.7794194
```

Parallel analysis suggests that the number of factors = NA and the number of components = 0 .

The analysis recommends that no factors be used - FA is not appropriate for this data. The scree plot shows that the observed eigen values for the first 2 components (1.632 and 1.540) are both greater than the mean of the simulated data sets (1.421 and 1.097). However, there is quite a bit of variation in the simulated data sets, creating large standard deviations around the means. The 95th percentile value for component 1 (1.665) is greater than the observed eigen value. Interestingly the 95th percentile value for component 2 (1.256) is below the observed eigen value, however since the first component is not feasible we cannot consider the second component.