R instructions for the 3rd seminar

In the dataset *Countries.RData*, there are data of employment in particular sectors in 1979. Analyze associations in employment between particular sectors (variables) using PCA. Assess how the particular countries (objects) are different in terms of economical structure. In the beginning, standardize the data. Work with all variables (There are 8 variables. Vector of these variables shall be further signed as **X**.)

R Instructions for the problem 1:

•Get familiar with data (correlation matrix,scatterplot matrix, Chernoff faces...)

```
!library(DescTools )
!library(ellipse)
!library(car)
```

```
x<-cor(Countries[1:8],use="pairwise.complete.obs")
pairs(Countries,panel=panel.smooth)
scatterplotMatrix(Countries[1:8],smooth=F,diagonal="histogram",col=c(2,1,4))
PlotCorr(x)
plotcorr(x)
PlotFaces(Countries)
```

R Instructions for the problem 2:

•Is PCA appropriate method for our data? (=Is correlation matrix significantly different from unit matrix?)

Visually asses graphs provided by correlation plots from previous task. (There exists test, no idea where in R)

R Instructions for the problem 3:

•Create the object in R bearing all essential results of PCA

```
p<-prcomp(x=Countries, center=T,scale.=T)
```

center and scale = T means all calculations are based on correlation (not covariance) matrix, which means that PCA works at standardized data.

R Instructions for the problem 4:

•Find the eigenvalues of a correlation matrix (of standardized variables)

`p$sdev` representr square root of eigenvalues
`p$sdev ˆ 2` eigenvalues
or
`eigen(x)$values`

R Instructions for the problem 5:

•Determine the number of components to extract(Use criterion: scree plot; percentage of explained variability is >75%; number of eigenvalues >1). What portion of total variance of **X** is explained by first two (resp. three) principal components? - this represents importance of particular components.

Scree plot:
`plot(p,type="l")`
`summary(p)` look at the line "Proportion of Variance"

R Instructions for the problem 6:

Find eigenvectors associated with the first 3 eigenvalues and consequently express first three principal components Y_1, Y_2, Y_3. ($D(Y_1) = \lambda_1$; ...). Finaly express first three standardized principal components Y_{1S}, Y_{2S}, Y_{3S}. ($D(Y_{rS}) = 1$).

`p$rotation` In the first column of the output there is the first eigenvector,...

or

`eigen(x)$vectors` In the first column of the output there is the first eigenvector,...

or

```
print(p)
v<-matrix(rep(p$sdev,8),8,8,byrow=T)
p$rotation/v
```

... standardized

R Instructions for the problem 7:

•Calculate the correlation between original variables and principal components:

```
p$rotation*v
```

As we are processing PCA on standardized data, σ_j in the following formula is equal to unit: $R(X_j, Y_r) = \frac{\sqrt{\lambda_r} v_{rj}}{\sigma_j}$

R Instructions for the problem 8:

•Express the cases in a new system of all (resp. first three) principal components:

```
NewCoordinate<-predict(p,newdata=Countries)
NewCoordinate[,1:3]
```

R Instructions for the problem 9:

•plot the graphs 1) where the cases are expressed in a system of first two principal components; 2) Represent all 8 original variables in a new system of first two components.

```
biplot(p,xlim=c(-0.3,0.7))
```

Black points representing countries are axpressed in a new system of first two components; notice which countries have similar economics. Which country is an ”outlier”?
Coordinates of red points(original variables) can be interpreted as correlations of original variable with particular principal components.

R Instructions for the problem 10:

zloka Promnn: Communality (kosinus 2)