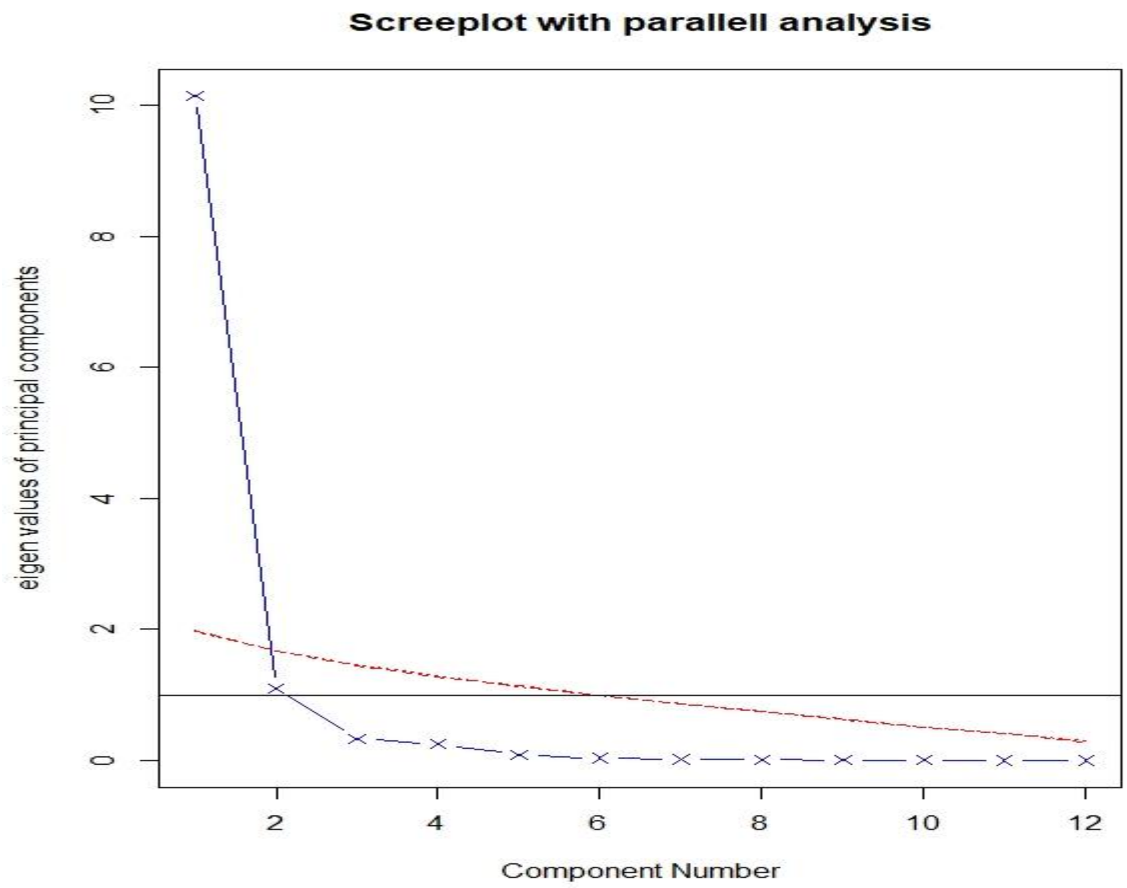
ASSIGNMENT 2

Problem 1:

To Determine the number of components to extract:

library(psych)
fa.parallel(USJudgeRatings,fa="pc",n.iter = 200,show.legend = FALSE, main="Screeplot with parallell analysis")
abline(h=1)



From the graph it can be understood that 2 components are appropriate for extraction from the raw data matrix

To extract the principal components:

pcomp<-principal(USJudgeRatings,nfactors=2,rotate="none")
pcomp

OUTPUT:

```
Principal Components Analysis
Call: principal(r = USJudgeRatings, nfactors = 2, rotate = "none")
```

```
Standardized loadings (pattern matrix) based upon correlation matrix
       PC1   PC2   h2     u2 com
CONT -0.01  0.98 0.96 0.0390 1.0
INTG  0.92 -0.19 0.88 0.1197 1.1
DMNR  0.91 -0.21 0.88 0.1229 1.1
DILG  0.97  0.04 0.94 0.0599 1.0
CFMG  0.96  0.18 0.96 0.0410 1.1
DECI  0.96  0.13 0.94 0.0584 1.0
PREP  0.98  0.03 0.97 0.0287 1.0
FAMI  0.98  0.00 0.95 0.0469 1.0
ORAL  1.00  0.00 0.99 0.0091 1.0
WRIT  0.99 -0.03 0.98 0.0184 1.0
PHYS  0.89  0.09 0.81 0.1927 1.0
RTEN  0.99 -0.04 0.97 0.0258 1.0

                        PC1  PC2
SS loadings           10.13 1.10
Proportion Var         0.84 0.09
Cumulative Var         0.84 0.94
Proportion Explained   0.90 0.10
Cumulative Proportion  0.90 1.00

Mean item complexity =  1
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.03
 with the empirical chi square  4.46  with prob <  1

Fit based upon off diagonal values = 1
```

Rotating Principal Components:

rcomp<-principal(USJudgeRatings,nfactors = 2,rotate="varimax")

rcomp

OUTPUT

```
Principal Components Analysis
Call: principal(r = USJudgeRatings, nfactors = 2, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
      RC1   RC2   h2     u2 com
CONT 0.00  0.98 0.96 0.0390 1.0
INTG 0.92 -0.20 0.88 0.1197 1.1
DMNR 0.91 -0.22 0.88 0.1229 1.1
DILG 0.97  0.03 0.94 0.0599 1.0
CFMG 0.97  0.17 0.96 0.0410 1.1
DECI 0.96  0.12 0.94 0.0584 1.0
PREP 0.99  0.02 0.97 0.0287 1.0
FAMI 0.98 -0.01 0.95 0.0469 1.0
ORAL 1.00 -0.01 0.99 0.0091 1.0
WRIT 0.99 -0.04 0.98 0.0184 1.0
PHYS 0.89  0.08 0.81 0.1927 1.0
RTEN 0.99 -0.05 0.97 0.0258 1.0

                        RC1  RC2
SS loadings           10.13 1.11
Proportion Var         0.84 0.09
Cumulative Var         0.84 0.94
Proportion Explained   0.90 0.10
Cumulative Proportion  0.90 1.00

Mean item complexity =  1
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.03
```

```
 with the empirical chi square  4.46  with prob <  1
Fit based upon off diagonal values = 1
```

Computing Principal components scores:

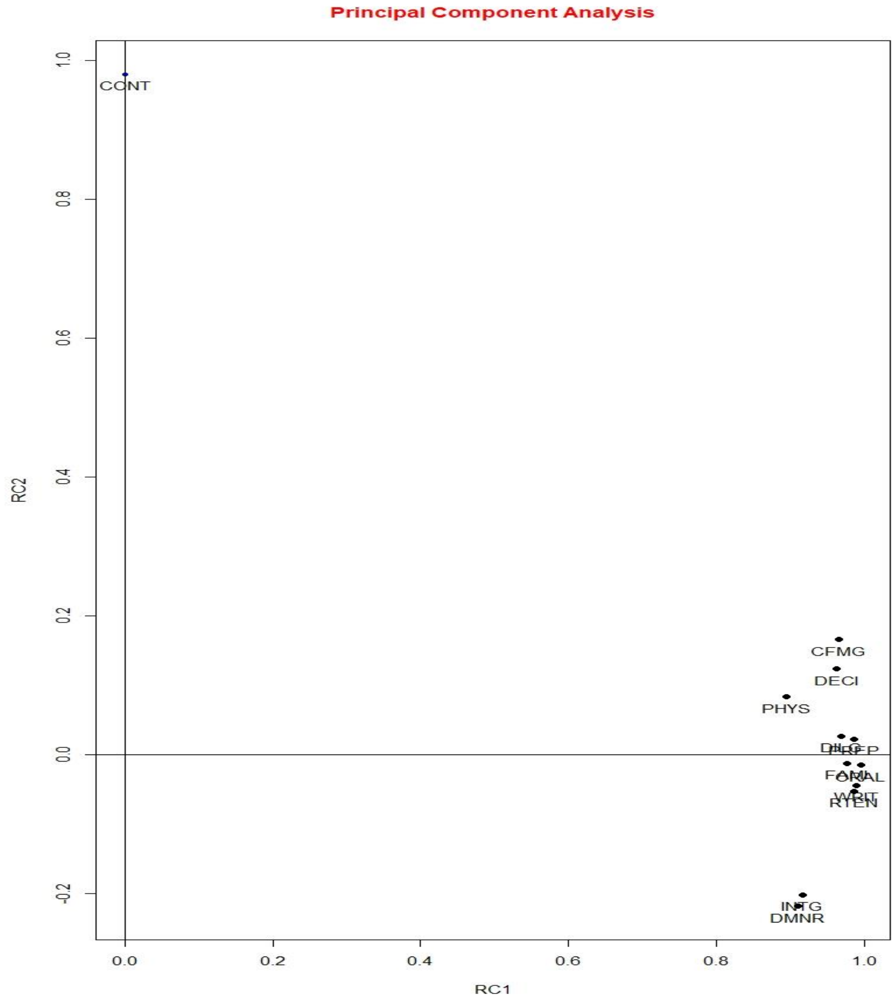
```
###### To compute the principal component scores
pcscore<-principal(USJudgeRatings,nfactors=2,score=TRUE)
pcscore$scores
```

OUTPUT

```
                          RC1         RC2
AARONSON,L.H.     -0.20325626 -1.72331155
ALEXANDER,J.M.     0.73838502 -0.83424278
ARMENTANO,A.J.     0.06754621 -0.28511211
BERDON,R.I.        1.13018700 -0.57353518
BRACKEN,J.J.      -2.15690547  0.15327560
BURNS,E.B.         0.75323535 -1.34633687
CALLAHAN,R.J.      1.25540269  2.94671590
COHEN,S.S.        -2.51529276 -0.22687786
DALY,J.J.          1.14802147 -0.15076333
DANNEHY,J.F.       0.34122462  1.07698120
DEAN,H.H.         -0.11082704 -0.44131655
DEVITA,H.J.       -0.46548561 -1.15436775
DRISCOLL,P.J.     -0.22376048 -1.21826256
GRILLO,A.E.       -1.01926220 -0.52102234
HADDEN,W.L.JR.     0.41607631 -0.80298340
HAMILL,E.C.       -0.14164457 -0.17004189
HEALEY.A.H.       -0.90622223  0.37120061
HULL,T.C.         -0.22065336  0.43830283
LEVINE,I.          0.21093652  0.84834443
LEVISTER,R.L.     -1.34264058  2.31433525
MARTIN,L.F.       -0.57966636 -0.74781509
MCGRATH,J.F.      -0.97853162  0.12282502
MIGNONE,A.F.      -2.01134769 -1.34401901
MISSAL,H.M.       -0.04138858 -1.39473443
MULVEY,H.M.        1.05417343  0.11272083
NARUK,H.J.         1.43392323  0.43525747
O'BRIEN,F.J.       0.46671359 -0.40775927
O'SULLIVAN,T.J.    1.08998995 -0.01105547
PASKEY,L.          0.59726969  0.22487938
RUBINOW,J.E.       1.49640955 -0.43815809
SADEN.G.A.         0.34164192 -0.32753568
SATANIELLO,A.G.    0.18109428  0.95974870
SHEA,D.M.          0.80718871 -0.38478854
SHEA,J.F.JR.       1.12805800 -0.16517354
SIDOR,W.J.        -2.15222233  0.33477799
SPEZIALE,J.A.      0.64881770  1.11509092
SPONZO,M.J.        0.37340904 -0.50449568
STAPLETON,J.F.     0.21653124 -0.86280515
TESTO,R.J.        -0.65165017  0.90048917
TIERNEY,W.L.JR.    0.43619512  1.03020107
WALL,R.A.         -0.79287328  1.82128546
WRIGHT,D.B.        0.45370182 -0.52814656
ZARRILLI,K.J.     -0.27250187  1.35822883
```

Orthogonal Solution graph:

```
factor.plot(rcomp,labels=rownames(rcomp$loadings),col.main="red")
```

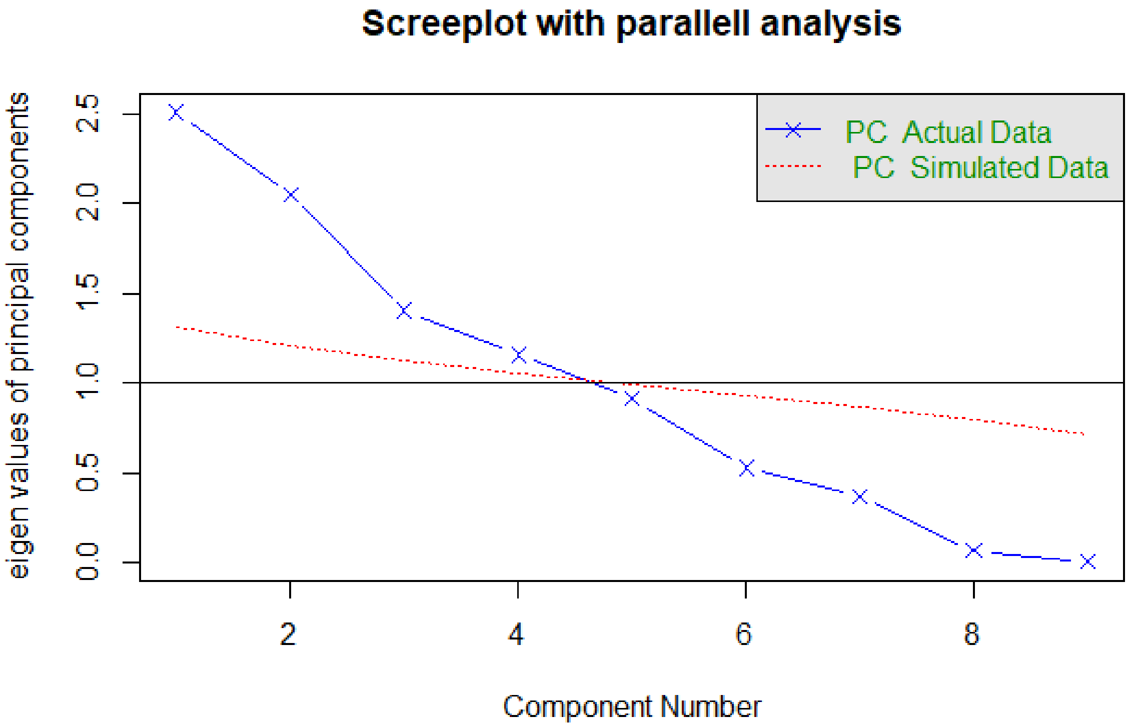


From the Graph above it can be seen that INTG, DMNR, RTEN, ORAL, FAMI, WRIT load on the first component while CONT load on the second component.
The variables PHYS, CFMG, DECI, DILG, PREP loads on both the components.

Problem 2:

To Determine the number of components to extract:

```
fa.parallel(df[c(2:10)],n.obs =214,fa="pc",n.iter = 100,show.legend = TRUE, main="Screeplot with parallell analysis")
abline(h=1)
```



To extract the principal components:

pcomp<-principal(df[c(2:10)],nfactors = 4,rotate ='none')
pcomp

OUTPUT

```
Principal Components Analysis
Call: principal(r = df[c(2:10)], nfactors = 4, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
     PC1   PC2   PC3   PC4   h2    u2  com
RI -0.86  0.41  0.10 -0.16 0.95 0.051 1.5
Na  0.41  0.39 -0.46 -0.53 0.80 0.195 3.8
Mg -0.18 -0.85  0.01 -0.41 0.92 0.081 1.5
Al  0.68  0.42  0.39  0.15 0.81 0.186 2.5
Si  0.36 -0.22 -0.54  0.70 0.97 0.031 2.7
K   0.35 -0.22  0.79  0.04 0.79 0.212 1.6
CA -0.78  0.49  0.00  0.30 0.94 0.058 2.0
Ba  0.40  0.69  0.09 -0.14 0.67 0.333 1.7
Fe -0.29 -0.09  0.34  0.25 0.27 0.730 3.0

                       PC1  PC2  PC3  PC4
SS loadings           2.51 2.05 1.40 1.16
Proportion Var        0.28 0.23 0.16 0.13
Cumulative Var        0.28 0.51 0.66 0.79
Proportion Explained  0.35 0.29 0.20 0.16
Cumulative Proportion 0.35 0.64 0.84 1.00

Mean item complexity =  2.3
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  102.53  with prob <  7.4e-20

Fit based upon off diagonal values = 0.92
```

Rotating Principal Components:

rcomp<-principal(df[c(2:10)],nfactors = 4,rotate = "varimax") rcomp

OUTPUT

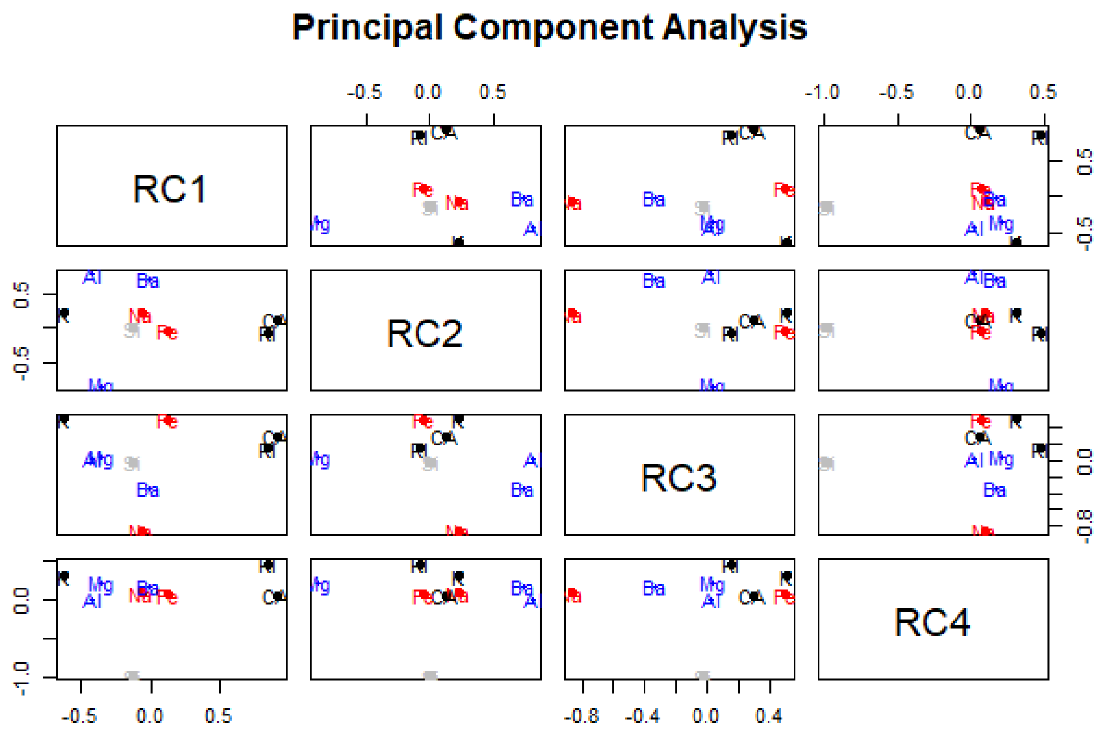
```
Principal Components Analysis
Call: principal(r = df[c(2:10)], nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
     RC1   RC2   RC3   RC4   h2    u2 com
RI  0.84 -0.07  0.15  0.47 0.95 0.051 1.7
Na -0.06  0.22 -0.86  0.09 0.80 0.195 1.2
Mg -0.35 -0.86  0.04  0.21 0.92 0.081 1.5
Al -0.42  0.80  0.03  0.01 0.81 0.186 1.5
Si -0.13  0.00 -0.02 -0.98 0.97 0.031 1.0
K  -0.62  0.22  0.51  0.30 0.79 0.212 2.7
CA  0.91  0.12  0.30  0.06 0.94 0.058 1.3
Ba -0.01  0.72 -0.33  0.17 0.67 0.333 1.5
Fe  0.12 -0.04  0.50  0.07 0.27 0.730 1.2

                       RC1  RC2  RC3  RC4
SS loadings           2.26 2.03 1.48 1.36
Proportion Var        0.25 0.23 0.16 0.15
Cumulative Var        0.25 0.48 0.64 0.79
Proportion Explained  0.32 0.28 0.21 0.19
Cumulative Proportion 0.32 0.60 0.81 1.00

Mean item complexity =  1.5
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  102.53  with prob <  7.4e-20

Fit based upon off diagonal values = 0.92
```

<u>Computing Principal components scores:</u>

```
pcscore<-principal(scale(df[c(2:10)]),nfactors = 4,rotate = "varimax",scores = TRUE)
pcscore$scores
```

<u>Orthogonal Solution graph</u>

factor.plot(rcomp,labels=rownames(rcomp$loadings),col.main="red")

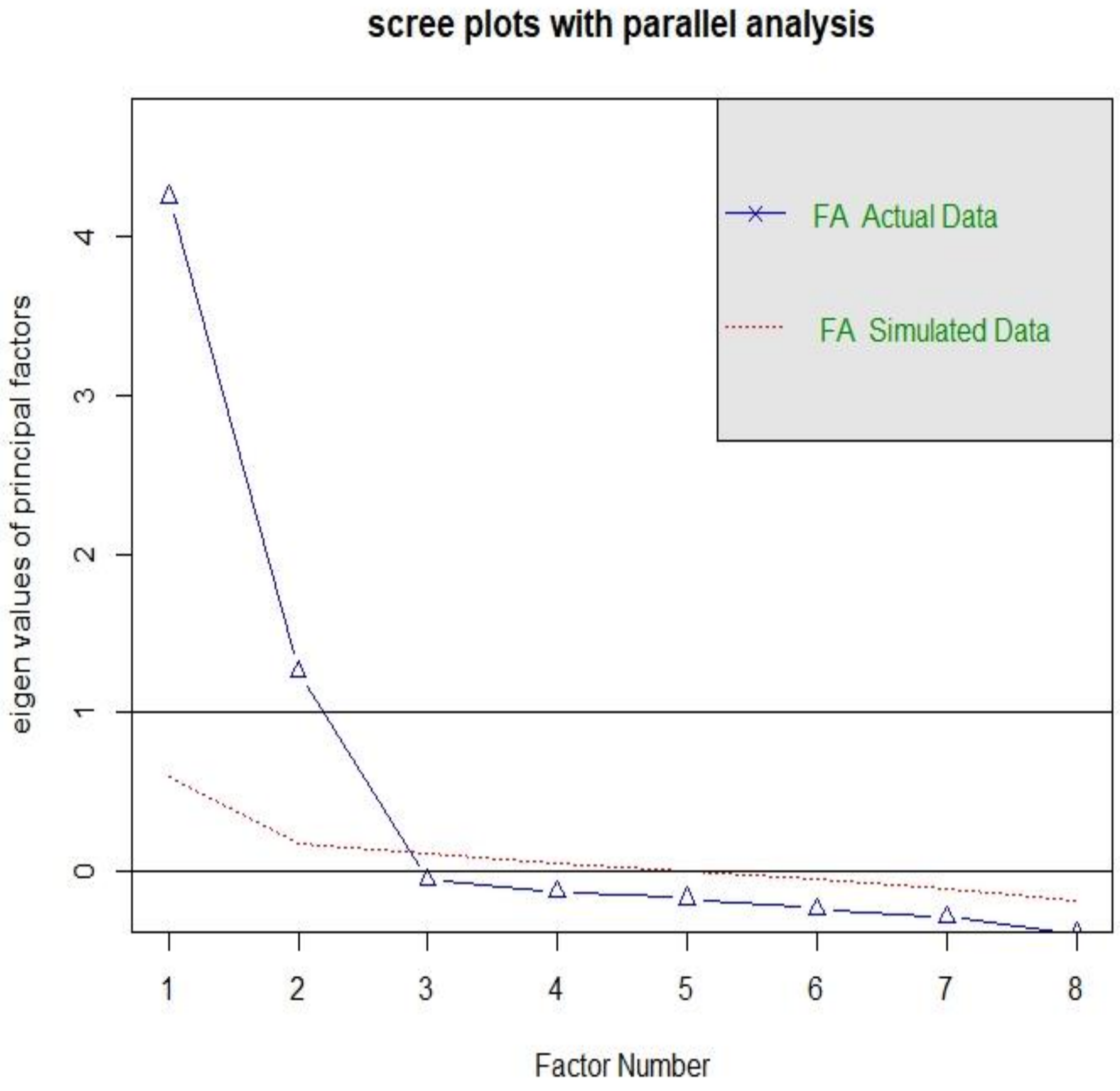


From the above graph we could see that the components are orthogonal,i.e the PCA we performed has got the best fit with maximum variance and minimum errors

Problem 3:

To Determine the number of factors to retain:

fa.parallel(Harman23.cor$cov,n.obs = 305,fa ="fa",n.iter = 200,main="scree plots with parallel analysis")
abline(h=0)



From the graph it can be understood that 2 factors are appropriate for extraction from the correlation matrix

<u>To extract factors:</u>

efa<-fa(Harman23.cor$cov,nfactors = 2,rotate = "none",fm="pa")
efa

OUTPUT:

```
Factor Analysis using method =  pa
Call: fa(r = Harman23.cor$cov, nfactors = 2, rotate = "none", 
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matr
ix
```

```
                 PA1   PA2   h2   u2 com
height          0.86 -0.32 0.84 0.16 1.3
arm.span        0.85 -0.41 0.89 0.11 1.4
forearm         0.81 -0.41 0.82 0.18 1.5
lower.leg       0.83 -0.34 0.81 0.19 1.3
weight          0.75  0.57 0.89 0.11 1.9
bitro.diameter  0.63  0.49 0.64 0.36 1.9
chest.girth     0.57  0.51 0.58 0.42 2.0
chest.width     0.61  0.35 0.49 0.51 1.6

                       PA1  PA2
SS loadings           4.45 1.51
Proportion Var        0.56 0.19
Cumulative Var        0.56 0.74
Proportion Explained  0.75 0.25
Cumulative Proportion 0.75 1.00

Mean item complexity =  1.6
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  28  and the objecti
ve function was  6.94
The degrees of freedom for the model are 13  and the objective fun
ction was  0.26

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                  PA1  PA2
Correlation of (regression) scores with factors   0.98 0.94
Multiple R square of scores with factors          0.96 0.89
Minimum correlation of possible factor scores     0.93 0.77
```

Rotating Factors:

rfa<-fa(Harman23.cor$cov,nfactors = 2,rotate = "varimax",fm="pa")

rfa

OUTPUT:

```
Factor Analysis using method =  pa
Call: fa(r = Harman23.cor$cov, nfactors = 2, rotate = "varimax",
    fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matr
ix
                 PA1  PA2   h2   u2 com
height          0.87 0.29 0.84 0.16 1.2
arm.span        0.92 0.21 0.89 0.11 1.1
forearm         0.89 0.19 0.82 0.18 1.1
lower.leg       0.86 0.26 0.81 0.19 1.2
weight          0.23 0.91 0.89 0.11 1.1
bitro.diameter  0.18 0.78 0.64 0.36 1.1
chest.girth     0.12 0.75 0.58 0.42 1.1
chest.width     0.25 0.65 0.49 0.51 1.3

                       PA1  PA2
SS loadings           3.29 2.67
Proportion Var        0.41 0.33
Cumulative Var        0.41 0.74
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00

Mean item complexity =  1.1
Test of the hypothesis that 2 factors are sufficient.
```

```
The degrees of freedom for the null model are  28  and the objecti
ve function was  6.94
The degrees of freedom for the model are 13  and the objective fun
ction was  0.26

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.03

Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   PA1  PA2
Correlation of (regression) scores with factors   0.97 0.95
Multiple R square of scores with factors          0.94 0.91
Minimum correlation of possible factor scores     0.88 0.81
```

To compute scores

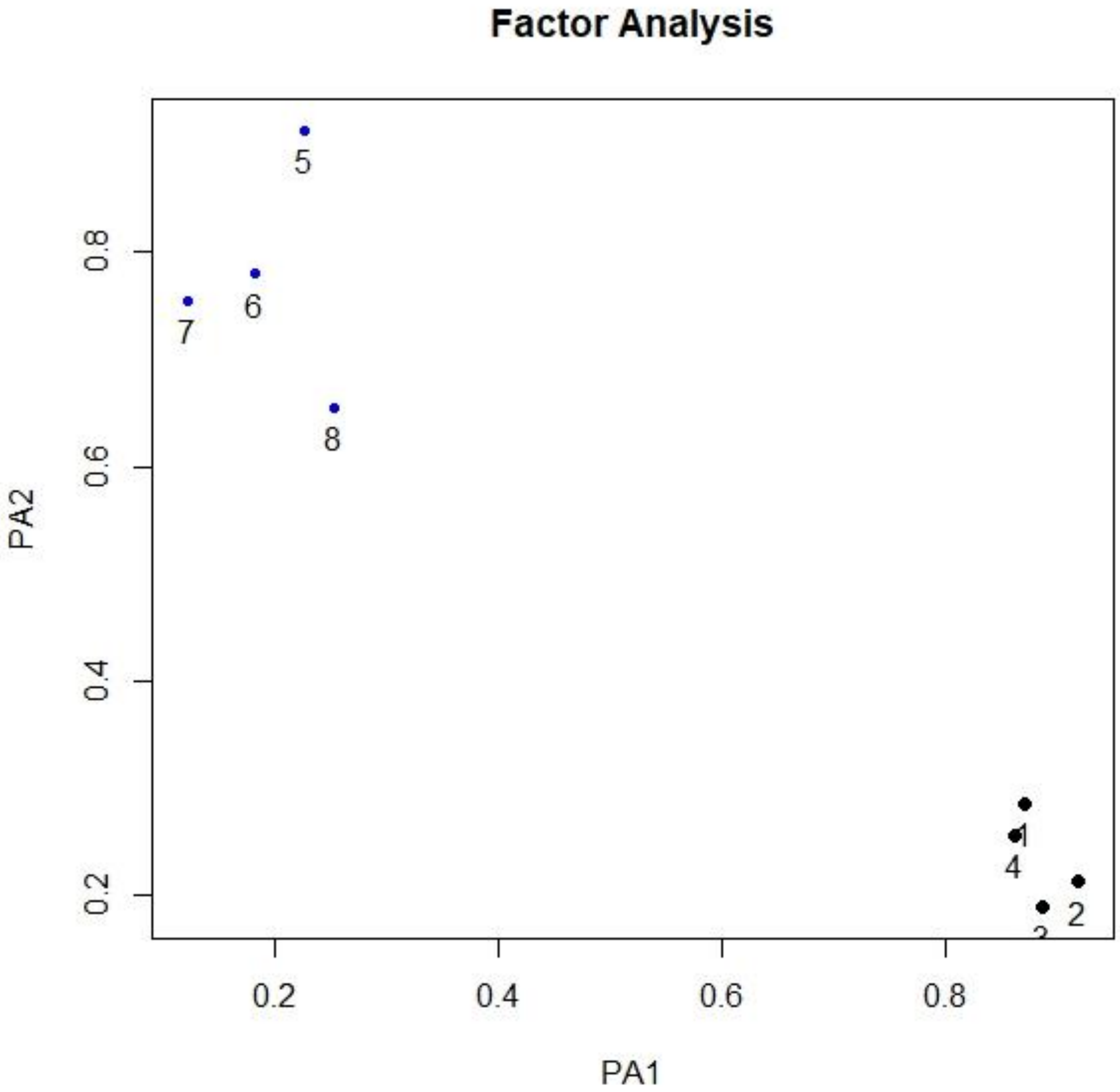
rfa$weights

OUTPUT:

```
                       PA1         PA2
height          0.26986562 -0.08577618
arm.span        0.40058016  0.01206800
forearm         0.23435026 -0.11748521
lower.leg       0.21784507 -0.03943021
weight         -0.15083423  0.70029052
bitro.diameter -0.04313770  0.17643121
chest.girth    -0.04273191  0.13860216
chest.width    -0.03462066  0.12047861
```

To graph the orthogonal solution:

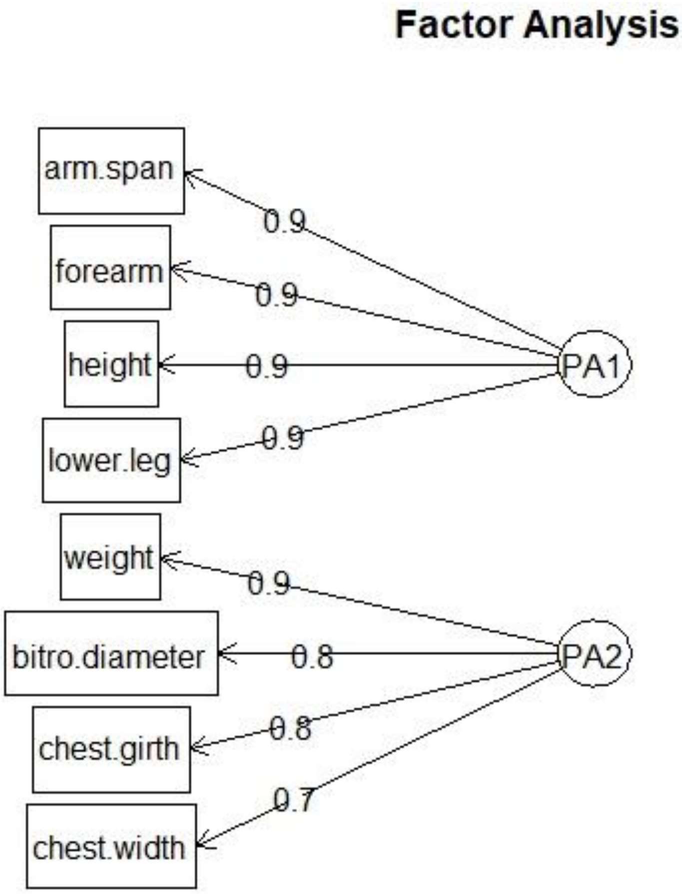
factor.plot(rfa,labels = rownames(rfa$loaadings))



To graph the oblique solution

To graph a oblique solution
fa.diagram(rfa,simple = FALSE)

OUTPUT:

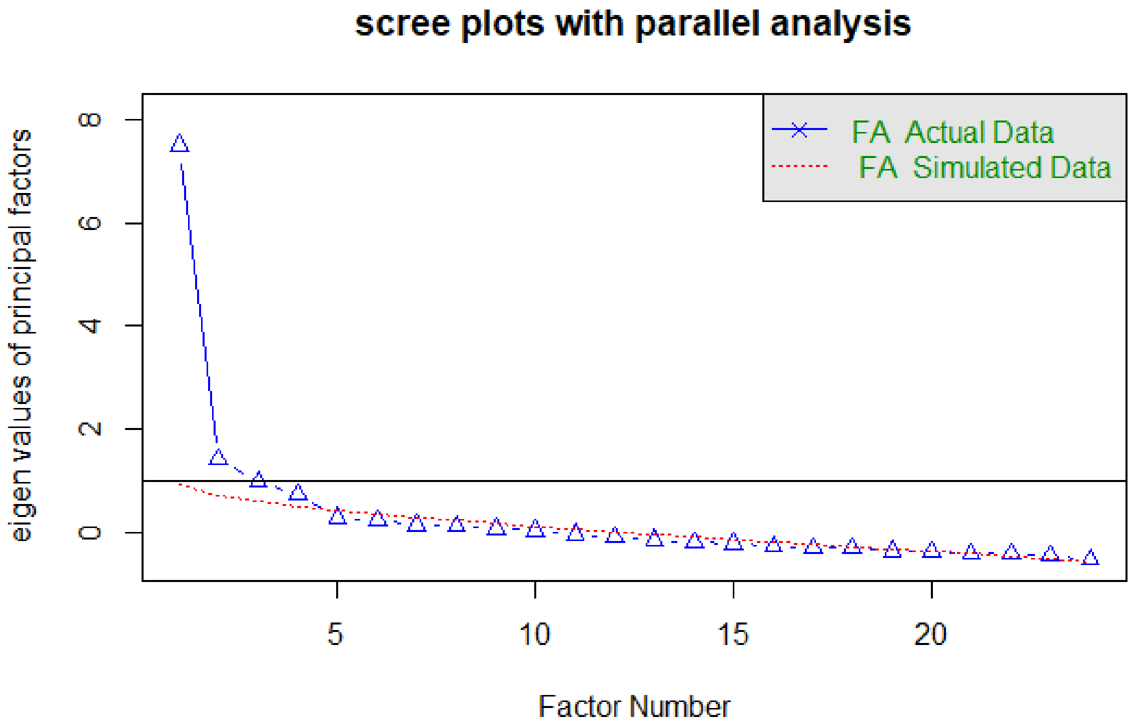


From the above graph it can be understood that the variables height, arm.span, forearm, lower.leg loads on the first factor and weight, bitro.diameter, chest.girth, chest.width loads on the second factor.

<u>Problem4</u>

<u>To Find Number of Components</u>

```
df<-Harman74.cor
fa.parallel(df$cov,n.obs = 145 ,fa='fa',n.iter=100,main="scree plots with parallel analysis")
abline(h=1)
```



To Extract Components

fctr<-fa(df$cov,nfactors = 2,rotate = "none",fm="pa")
fctr

Output

```
Factor Analysis using method =  pa
Call: fa(r = df$cov, nfactors = 2, rotate = "none", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                        PA1   PA2   h2   u2 com
VisualPerception       0.59  0.06 0.35 0.65 1.0
Cubes                  0.37 -0.01 0.14 0.86 1.0
PaperFormBoard         0.41 -0.08 0.18 0.82 1.1
Flags                  0.48 -0.08 0.24 0.76 1.1
GeneralInformation     0.69 -0.31 0.56 0.44 1.4
PargraphComprehension  0.69 -0.41 0.64 0.36 1.6
SentenceCompletion     0.67 -0.42 0.63 0.37 1.7
WordClassification     0.68 -0.20 0.50 0.50 1.2
WordMeaning            0.70 -0.46 0.70 0.30 1.7
Addition               0.45  0.41 0.38 0.62 2.0
Code                   0.55  0.35 0.43 0.57 1.7
CountingDots           0.46  0.47 0.44 0.56 2.0
StraightCurvedCapitals 0.59  0.26 0.42 0.58 1.4
WordRecognition        0.42  0.06 0.18 0.82 1.0
NumberRecognition      0.39  0.10 0.16 0.84 1.1
FigureRecognition      0.50  0.11 0.26 0.74 1.1
ObjectNumber           0.46  0.20 0.25 0.75 1.4
NumberFigure           0.52  0.34 0.38 0.62 1.7
FigureWord             0.44  0.12 0.21 0.79 1.1
Deduction              0.62 -0.12 0.40 0.60 1.1
NumericalPuzzles       0.60  0.23 0.41 0.59 1.3
ProblemReasoning       0.61 -0.09 0.38 0.62 1.0
SeriesCompletion       0.69 -0.05 0.48 0.52 1.0
ArithmeticProblems     0.65  0.16 0.45 0.55 1.1

                       PA1  PA2
SS loadings           7.56 1.58
```

```
Proportion Var        0.32 0.07
Cumulative Var        0.32 0.38
Proportion Explained  0.83 0.17
Cumulative Proportion 0.83 1.00

Mean item complexity =  1.3
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  276  and the objective func
tion was  11.44
The degrees of freedom for the model are 229  and the objective function w
as  3.2

The root mean square of the residuals (RMSR) is  0.07
The df corrected root mean square of the residuals is  0.08

Fit based upon off diagonal values = 0.95
Measures of factor score adequacy
                                                   PA1  PA2
Correlation of (regression) scores with factors   0.97 0.88
Multiple R square of scores with factors          0.93 0.77
Minimum correlation of possible factor scores     0.87 0.53
```

Rotate Components

rfctr<-fa(df$cov,nfactors = 2,rotate = "varimax",fm="pa")
rfctr

output:

```
Factor Analysis using method =  pa
Call: fa(r = df$cov, nfactors = 2, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
                        PA1  PA2   h2   u2 com
VisualPerception       0.40 0.43 0.35 0.65 2.0
Cubes                  0.28 0.24 0.14 0.86 1.9
PaperFormBoard         0.36 0.21 0.18 0.82 1.6
Flags                  0.42 0.25 0.24 0.76 1.6
GeneralInformation     0.72 0.22 0.56 0.44 1.2
PargraphComprehension  0.79 0.14 0.64 0.36 1.1
SentenceCompletion     0.78 0.13 0.63 0.37 1.1
WordClassification     0.64 0.29 0.50 0.50 1.4
WordMeaning            0.83 0.11 0.70 0.30 1.0
Addition               0.07 0.61 0.38 0.62 1.0
Code                   0.19 0.63 0.43 0.57 1.2
CountingDots           0.04 0.66 0.44 0.56 1.0
StraightCurvedCapitals 0.28 0.59 0.42 0.58 1.4
WordRecognition        0.28 0.32 0.18 0.82 2.0
NumberRecognition      0.23 0.33 0.16 0.84 1.8
FigureRecognition      0.31 0.41 0.26 0.74 1.9
ObjectNumber           0.22 0.45 0.25 0.75 1.4
NumberFigure           0.17 0.59 0.38 0.62 1.2
FigureWord             0.26 0.38 0.21 0.79 1.8
Deduction              0.55 0.31 0.40 0.60 1.6
NumericalPuzzles       0.30 0.56 0.41 0.59 1.5
ProblemReasoning       0.52 0.33 0.38 0.62 1.7
SeriesCompletion       0.55 0.42 0.48 0.52 1.9
ArithmeticProblems     0.39 0.55 0.45 0.55 1.8

                       PA1  PA2
SS loadings           5.00 4.15
Proportion Var        0.21 0.17
Cumulative Var        0.21 0.38
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00
```

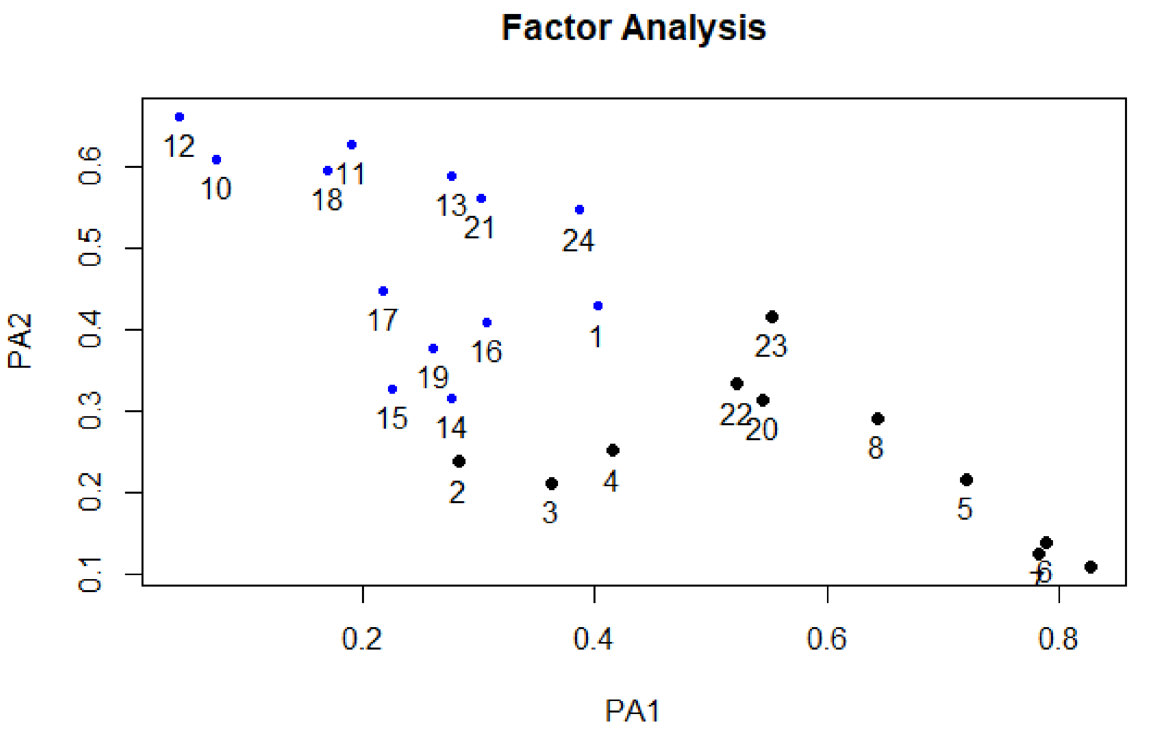
```
Mean item complexity =  1.5
Test of the hypothesis that 2 factors are sufficient.

The degrees of freedom for the null model are  276  and the objective func
tion was  11.44
The degrees of freedom for the model are 229  and the objective function w
as  3.2

The root mean square of the residuals (RMSR) is  0.07
The df corrected root mean square of the residuals is  0.08

Fit based upon off diagonal values = 0.95
Measures of factor score adequacy
                                                  PA1  PA2
Correlation of (regression) scores with factors   0.94 0.90
Multiple R square of scores with factors          0.88 0.82
Minimum correlation of possible factor scores     0.76 0.64
```

To Compute Weights

rfctr$weights

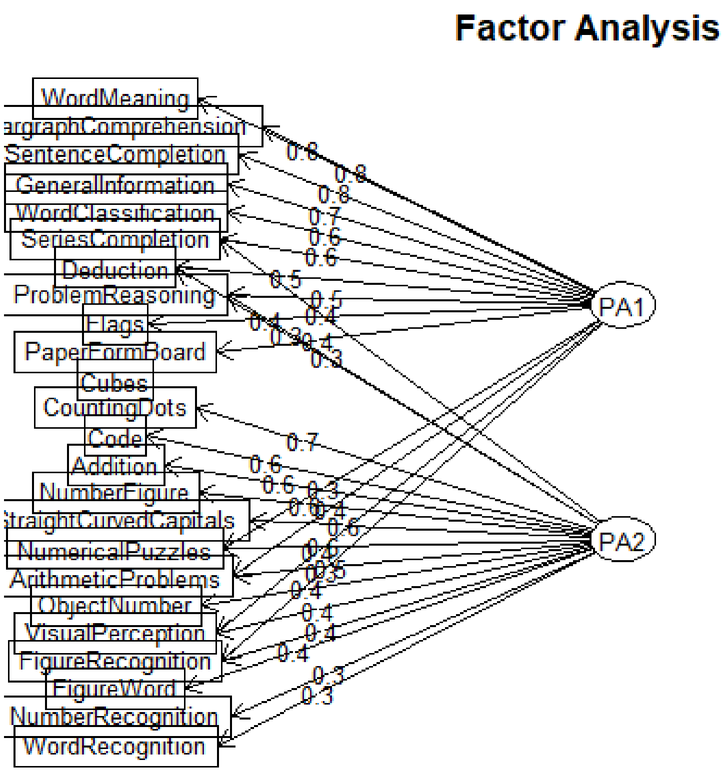
To Graph Orthogonal Solutions

factor.plot(rfctr,labels = rownames(rfctr$loaadings))



To Graph Oblique Solution

fa.diagram(rfctr,simple = FALSE)

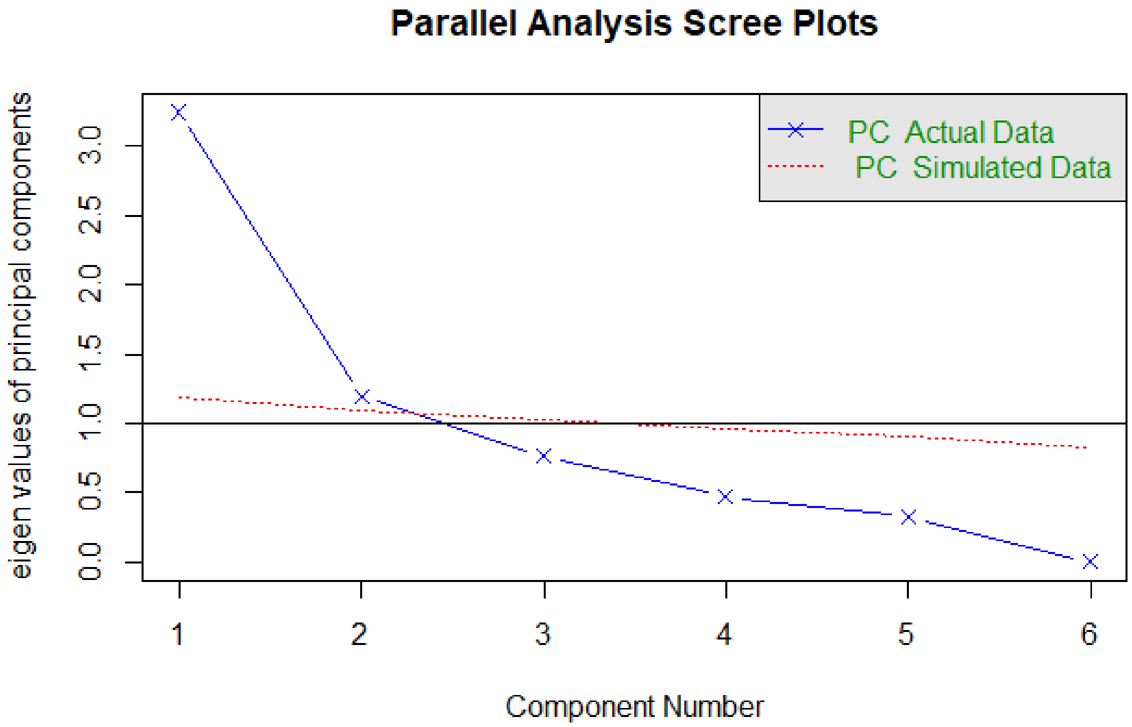


After performing factor analysis the attributes are split into two clusters with no correlation .The oblique graph explains how much information is contributed by each attributed to the factor components 1 and 2

<u>Problem 5</u>

<u>Extract Components</u>

df<-read.csv("C:\\Users\\sundarakishore\\Desktop\\neu classes\\2nd sem\\Data Mining\\assignment-2\\Vertebral Column Data.csv")
fa.parallel(df[c(1:6)],n.obs=310,fa='pc',n.iter=100)
abline(h=1)

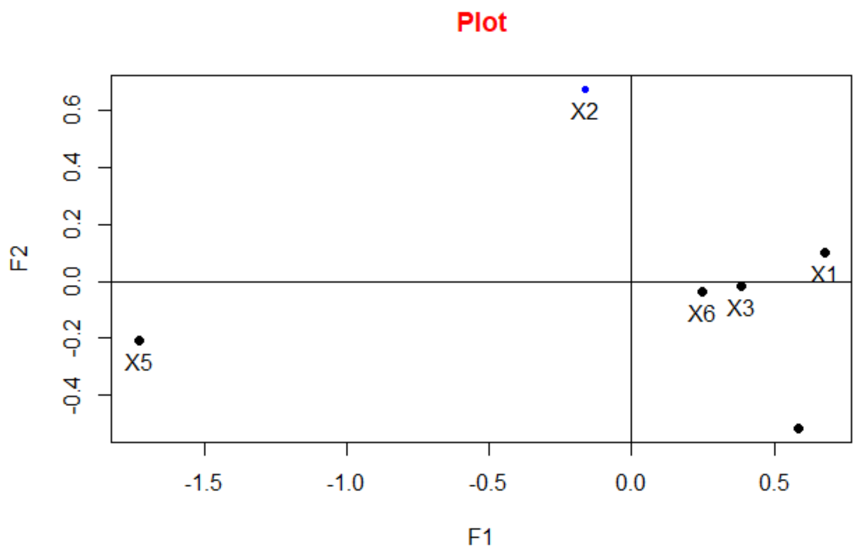


<u>Perform Multi-dimensional scaling</u>:

pc=cor(df[c(1:6)],method = "p")
dmatrix<-dist(pc)
d1<-as.matrix(dmatrix)
c<-cmdscale(d1,k=2)

<u>Graph Orthogonal Solution</u>

factor.plot(c,labels = rownames(c),col.main='red')



From the above graph we could see that we have plotted 6 attributes into two dimensional space with points of no correlation and maximum variance