MVA_Assignment_6

Aman

10/22/2020

Assignment 6 - Factor Analysis

This document does a Factor Analysis on the Heart Failure Prediction dataset. We start by understanding if factor analysis is appropriate on our dataset , then perform checks to see how many factors are appropriate and finally interpret the factors.

Let us load libraries and data

```
# clear environment
rm(list = ls())

# defining libraries

library(ggplot2)
library(dplyr)
library(PerformanceAnalytics)
library(data.table)
library(sqldf)
library(nortest)
library(MASS)
library(rpart)
library(class)
library(ISLR)
library(scales)
library(ClustOfVar)
library(GGally)
library(reticulate)
library(ggthemes)
library(RColorBrewer)
library(gridExtra)
library(kableExtra)
library(Hmisc)
library(corrplot)
library(energy)
library(nnet)
library(Hotelling)
library(car)
library(devtools)
library(ggbiplot)
```

```r
library(factoextra)
library(rgl)
library(FactoMineR)
library(psych)
library(nFactors)
library(scatterplot3d)
```

```r
# reading data
data <- read.csv('/Users/mac/Downloads/heart_failure_clinical_records_dataset.csv')
str(data)
```

```
## 'data.frame':    299 obs. of  13 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
##  $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
```

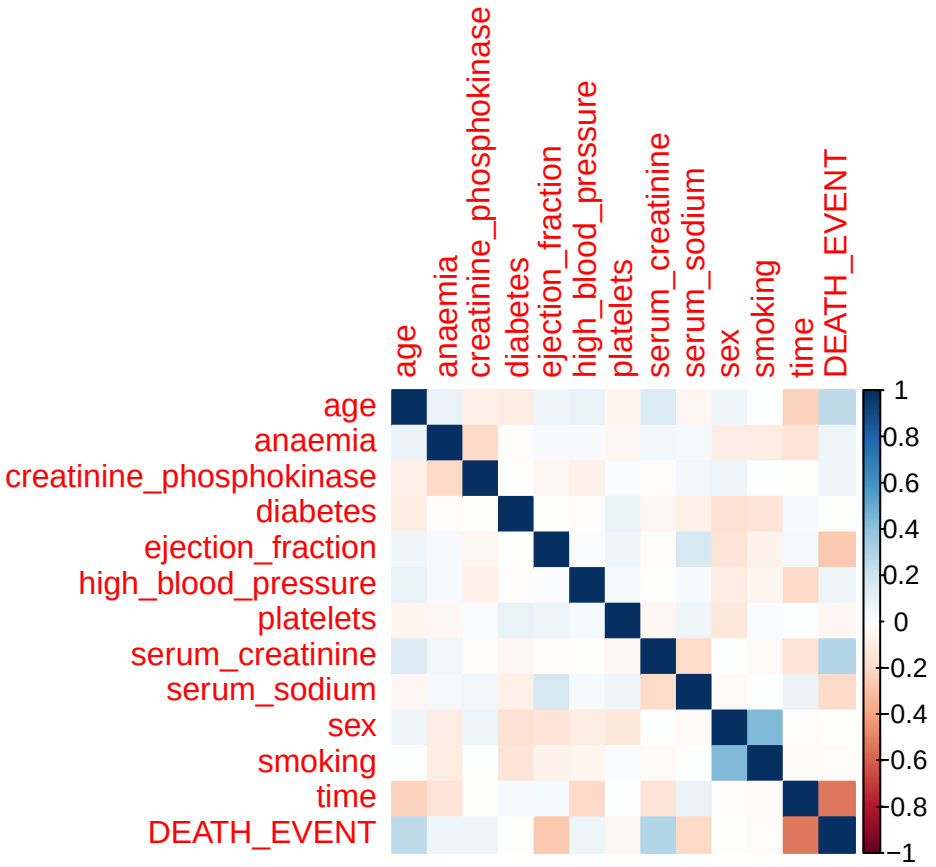
Let's quickly revise our correlation plot and see if factor analysis is appropriate

```r
# Correlation plot
M<-cor(data)
head(round(M,2))
```

```
##                            age anaemia creatinine_phosphokinase diabetes
## age                       1.00    0.09                   -0.08    -0.10
## anaemia                   0.09    1.00                   -0.19    -0.01
## creatinine_phosphokinase -0.08   -0.19                    1.00    -0.01
## diabetes                 -0.10   -0.01                   -0.01     1.00
## ejection_fraction         0.06    0.03                   -0.04     0.00
## high_blood_pressure       0.09    0.04                   -0.07    -0.01
##                          ejection_fraction high_blood_pressure platelets
## age                                   0.06               0.09     -0.05
## anaemia                               0.03               0.04     -0.04
## creatinine_phosphokinase             -0.04              -0.07      0.02
## diabetes                              0.00              -0.01      0.09
## ejection_fraction                     1.00               0.02      0.07
## high_blood_pressure                   0.02               1.00      0.05
##                          serum_creatinine serum_sodium   sex smoking  time
## age                              0.16        -0.05  0.07    0.02 -0.22
## anaemia                          0.05         0.04 -0.09   -0.11 -0.14
## creatinine_phosphokinase        -0.02         0.06  0.08    0.00 -0.01
## diabetes                        -0.05        -0.09 -0.16   -0.15  0.03
```

```
## ejection_fraction                  -0.01         0.18 -0.15   -0.07  0.04
## high_blood_pressure                 0.00         0.04 -0.10   -0.06 -0.20
##                            DEATH_EVENT
## age                               0.25
## anaemia                           0.07
## creatinine_phosphokinase          0.06
## diabetes                          0.00
## ejection_fraction                -0.27
## high_blood_pressure               0.08
```

```
corrplot(M, method="color")
```



Since most of the correlations are low (Pearson's r < 0.25)), we don't particularly see a need for Factor Analysis since we use Factor Analysis to understand the latent factors in the data
However, we can see that given these are patient details, we may try and understand factors such as patient demographics (age, sex), patient lifestyle (smoking, diabetes, high bp), patient physiological makeup (serum sodidum, creatinine_phosphokinase), patient genetics (bp, anaemia). While this is our intuition before we begin, only once we see the factor analysis results will we be able to comment more appropriately.

```
#scale the data
data_fact <- as.data.frame(scale(data[,1:12],center = TRUE, scale = TRUE))
```

Tests to see if factor analysis is appropriate on the data

Kaiser-Meyer-Oklin Test (KMO)

Measured by the Kaiser-Meyer-Olkin (KMO) statistics, sampling adequacy (MSA) predicts if data are likely to factor well, based on correlation and partial correlation.

KMO varies from 0 to 1.0 and KMO overall should be .60 or higher to proceed with factor analysis.

```
KMO(data_fact)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data_fact)
## Overall MSA =  0.55
## MSA for each item =
##                       age                   anaemia creatinine_phosphokinase
##                      0.61                      0.56                      0.47
##                  diabetes         ejection_fraction      high_blood_pressure
##                      0.63                      0.56                      0.55
##                 platelets          serum_creatinine              serum_sodium
##                      0.49                      0.59                      0.50
##                       sex                   smoking                      time
##                      0.55                      0.53                      0.56
```

We see that other than age and diabetes, there is no variable with MSA > 0.6 but there are many variables who are close to the cut-off.

Bartlett's test

We also perform the Bartlett's test which allows us to compare the variance of two or more samples to determine whether they are drawn from populations with equal variance.

```
bartlett.test(data_fact)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data_fact
## Bartlett's K-squared = 3.3045e-14, df = 11, p-value = 1
```

From the output we can see that the p-value of 1 is not less than the significance level of 0.05. This means we cannot reject the null hypothesis that the variance is the same for all patients.

Let us now perform Factor Anaysis on our dataset

```
# perform factor analysis
data.fa <- factanal(data_fact, factors = 2)
data.fa
```

```
##
## Call:
## factanal(x = data_fact, factors = 2)
##
## Uniquenesses:
##                       age                   anaemia creatinine_phosphokinase
##                     0.794                     0.924                     0.972
##                  diabetes         ejection_fraction      high_blood_pressure
##                     0.953                     0.972                     0.927
##                 platelets          serum_creatinine              serum_sodium
##                     0.976                     0.904                     0.975
##                       sex                   smoking                      time
##                     0.233                     0.742                     0.737
##
## Loadings:
##                           Factor1 Factor2
## age                        0.139   0.432
## anaemia                            0.264
## creatinine_phosphokinase          -0.154
## diabetes                  -0.204
## ejection_fraction         -0.167
## high_blood_pressure                0.258
## platelets                 -0.141
## serum_creatinine                   0.304
## serum_sodium                      -0.147
## sex                        0.866  -0.132
## smoking                    0.498  -0.103
## time                              -0.503
##
##                 Factor1 Factor2
## SS loadings       1.139   0.752
## Proportion Var    0.095   0.063
## Cumulative Var    0.095   0.158
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 57.21 on 43 degrees of freedom.
## The p-value is 0.0721
```

Factor analysis creates linear combinations of factors to abstract the variable's underlying communality. To the extent that the variables have an underlying communality, fewer factors capture most of the variance in the data set.

Here, we see high uniqueness (>0.7) for most variables indicating that factors don't account well for the variance. But we do note that sex variable has the least uniqueness (0.233).

We also note that cumulative variance explained is only 15.8% which isn't great and we may have to use more than 2 factors

```r
#squaring the loadings to assess communality
apply(data.fa$loadings^2,1,sum)
```

```
##                      age                  anaemia creatinine_phosphokinase
##               0.20588167               0.07630296               0.02819032
##                 diabetes        ejection_fraction      high_blood_pressure
##               0.04721557               0.02808125               0.07296775
##                 platelets         serum_creatinine             serum_sodium
##               0.02402435               0.09550733               0.02508073
##                      sex                  smoking                     time
##               0.76665418               0.25821777               0.26304715
```

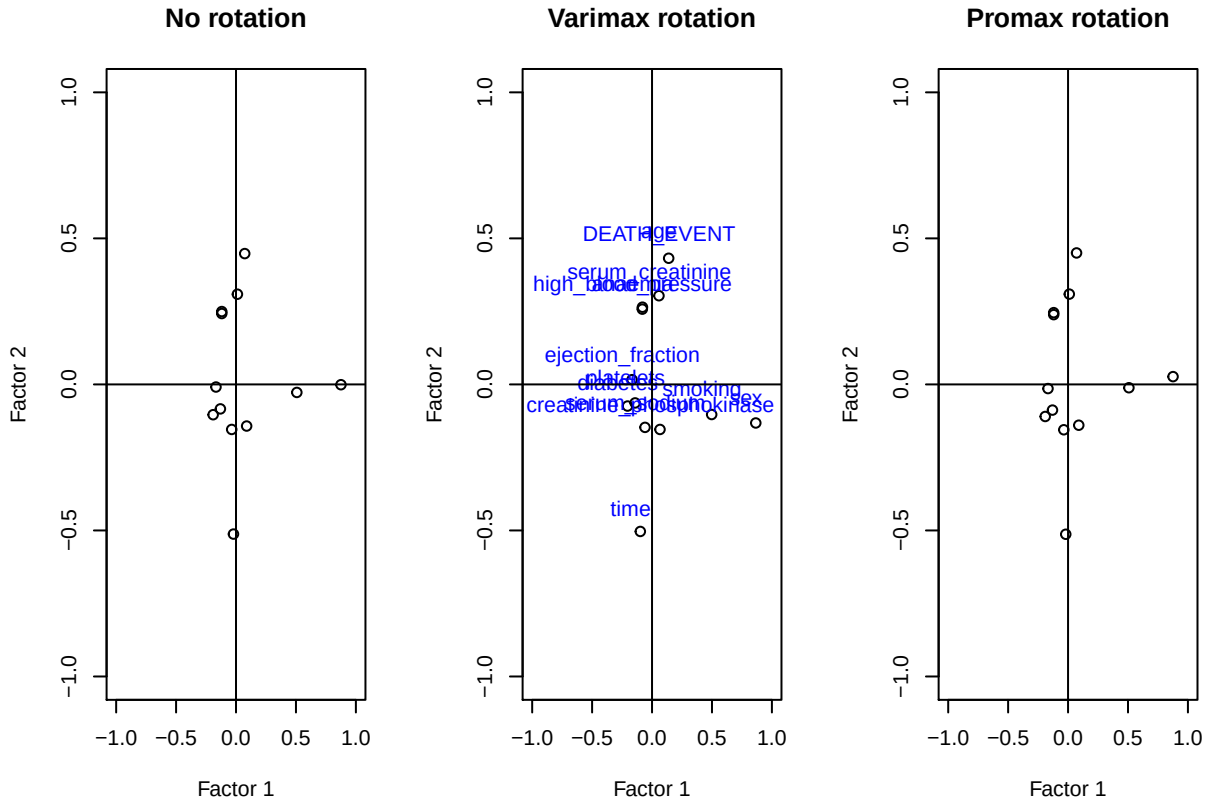
We see mostly low values of loadings other than sex here indicating the model for factor analysis isn't appropriate. A good model would indicate high values of communality and low values of uniqueness.

Let's try and interpret the factors

We perform three factor models - one with no rotation, one with varimax rotation, and finally one with promax rotation and see the results

```r
data.fa.none <- factanal(data_fact, factors = 2, rotation = "none")
data.fa.varimax <- factanal(data_fact, factors = 2, rotation = "varimax")
data.fa.promax <- factanal(data_fact, factors = 2, rotation = "promax")

par(mfrow = c(1,3))
plot(data.fa.none$loadings[,1],
     data.fa.none$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "No rotation")
abline(h = 0, v = 0)

plot(data.fa.varimax$loadings[,1],
     data.fa.varimax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation")

text(data.fa.varimax$loadings[,1]-0.08,
     data.fa.varimax$loadings[,2]+0.08,
      colnames(data),
      col="blue")
abline(h = 0, v = 0)

plot(data.fa.promax$loadings[,1],
     data.fa.promax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
```

```
     xlim = c(-1,1),
     main = "Promax rotation")
abline(h = 0, v = 0)
```

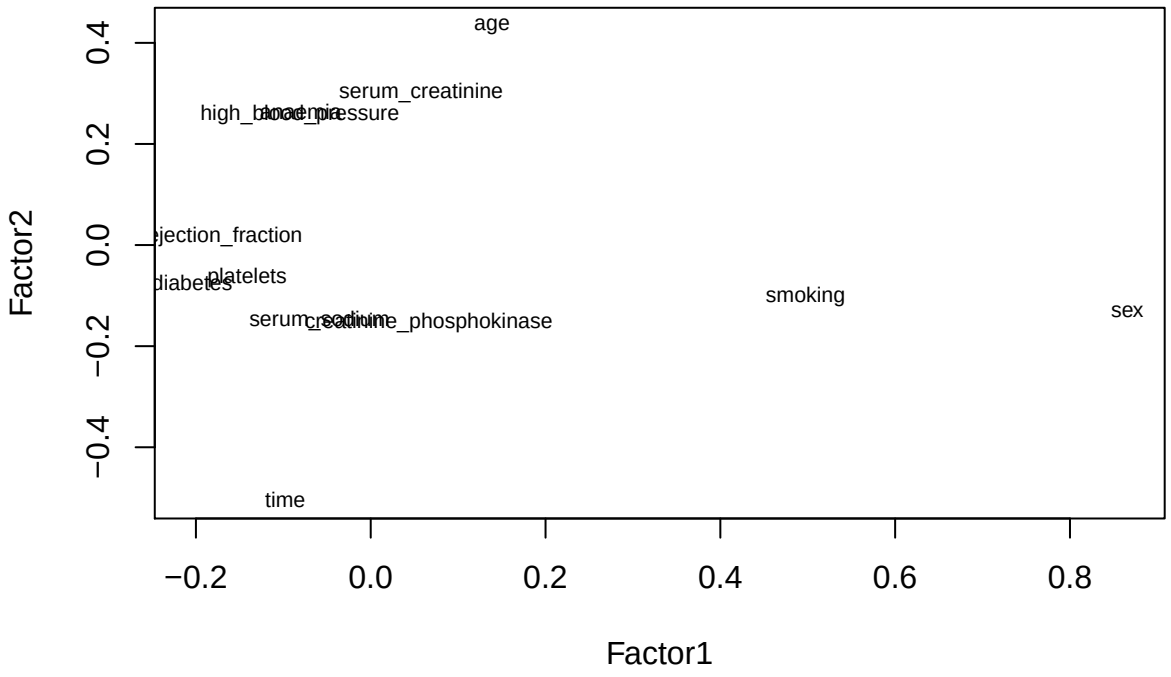


We can see that factor 1 corresponds to smoking, sex, platelets, , ejection_fraction and diabetes whereas factor 2 corresponds to age, anaemia, high bp, serum_creatinine and time among others. We cannot clearly name the factors at this point in line with our intuition.

Let's plot the results

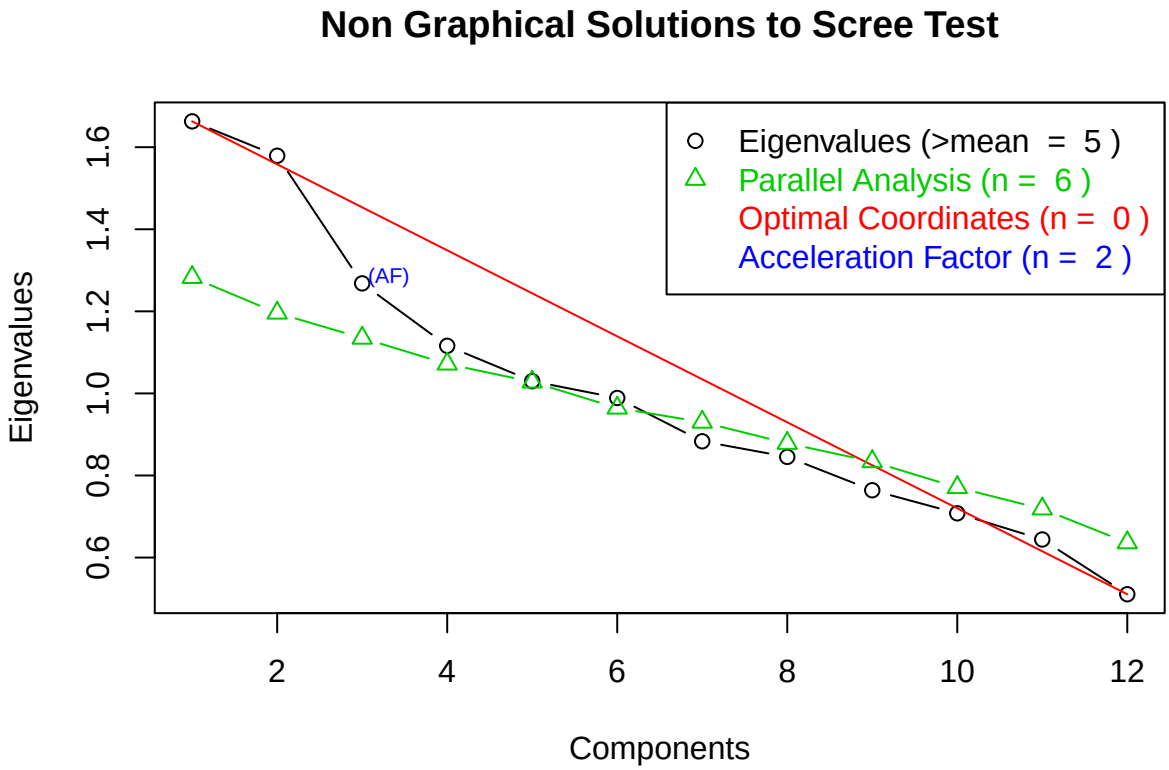
Maximum Likelihood Factor Analysis with 2 factors

```
# Maximum Likelihood Factor Analysis
# entering raw data and extracting 2 factors,
# with varimax rotation
fit <- factanal(data_fact, 2, rotation="varimax")
# plot factor 1 by factor 2
load <- fit$loadings[,1:2]
plot(load,type="n") # set up plot
text(load,labels=names(data_fact),cex=.7) # add variable names```
```



However, there is a better method to first determine number of Factors to Extract

```
ev <- eigen(cor(data_fact)) # get eigenvalues
ap <- parallel(subject=nrow(data_fact),var=ncol(data_fact),
  rep=100,cent=.05)
nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)
```



We plot the components against the eigenvalues and this tells us that there are 5 factors with eigenvalue of above 1. Hence we know our maximum number of factors is clearly 5 but also using the elbow logic of the scree test, we see that there are probably 3 or 4 factors (the levelling off after 3 is more significant).

This is interesting now since our interpretation might be more relevant with 3 factors.

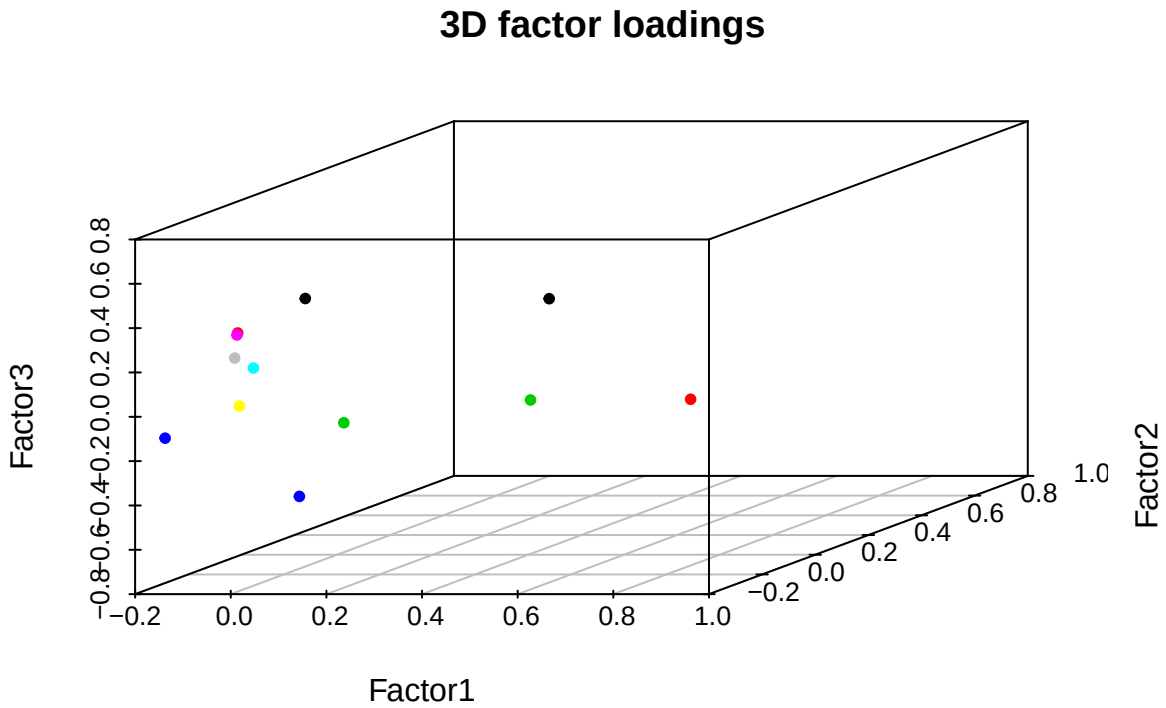
Factor Analysis (n=3 factors)

```
data.fa.none <- factanal(data_fact, factors = 3, rotation = "none")
data.fa.none
```

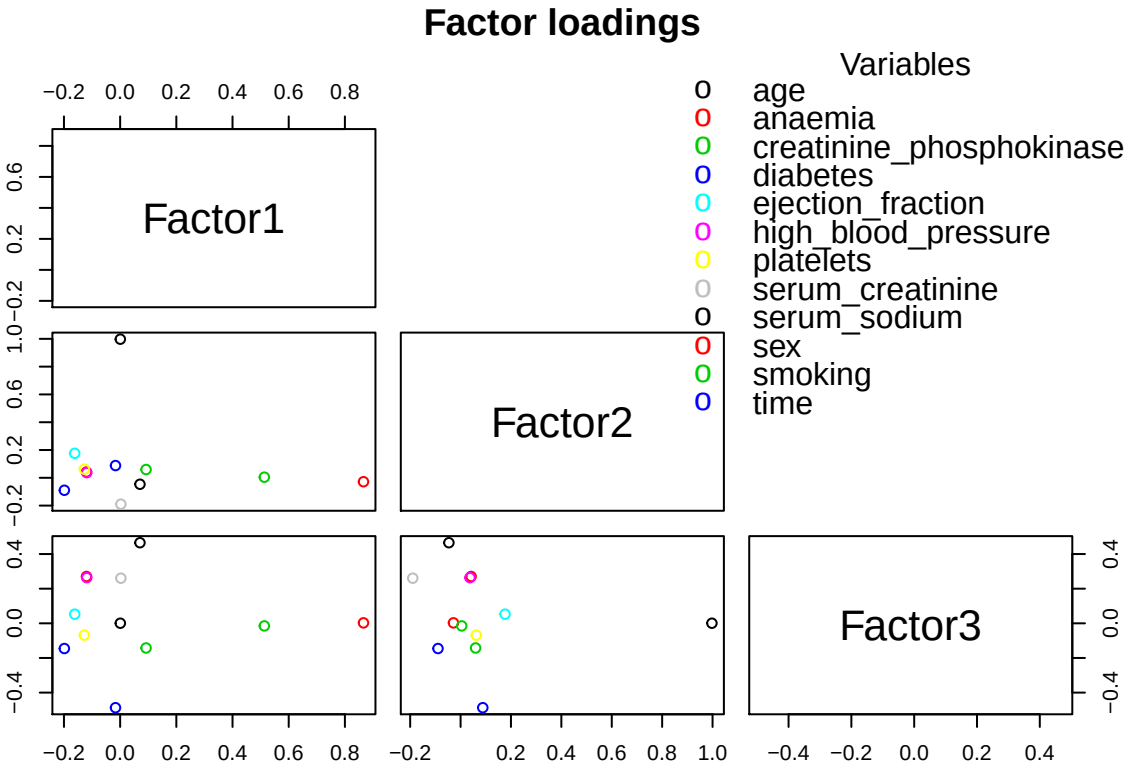
```
##
## Call:
## factanal(x = data_fact, factors = 3, rotation = "none")
##
## Uniquenesses:
##                      age                  anaemia creatinine_phosphokinase
##                    0.776                    0.911                    0.968
##                 diabetes        ejection_fraction      high_blood_pressure
##                    0.931                    0.940                    0.915
##                platelets         serum_creatinine             serum_sodium
##                    0.975                    0.896                    0.005
##                      sex                  smoking                     time
##                    0.249                    0.736                    0.755
##
## Loadings:
##                          Factor1 Factor2 Factor3
```

```
## age                                          0.465
## anaemia                  -0.120              0.270
## creatinine_phosphokinase                    -0.142
## diabetes                 -0.199             -0.146
## ejection_fraction        -0.162   0.176
## high_blood_pressure      -0.119              0.264
## platelets                -0.127
## serum_creatinine                  -0.190     0.261
## serum_sodium                       0.997
## sex                       0.866
## smoking                   0.513
## time                                        -0.487
##
##                Factor1 Factor2 Factor3
## SS loadings      1.138   1.091   0.714
## Proportion Var   0.095   0.091   0.059
## Cumulative Var   0.095   0.186   0.245
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 31.85 on 33 degrees of freedom.
## The p-value is 0.524
```

```
scatterplot3d(as.data.frame(unclass(data.fa.none$loadings)),
  main="3D factor loadings", color=1:ncol(data_fact), pch=20)
```




```
pairs(data.fa.none$loadings, col=1:ncol(data_fact),
  upper.panel=NULL, main="Factor loadings")
par(xpd=TRUE)
legend('topright', bty='n', pch='o', col=1:ncol(data_fact),
attr(data.fa.none$loadings, 'dimnames')[[1]], y.intersp=0.5,
title="Variables")
```



This is a lot more interesting since now if we try and interpret the 3 factors we see that Factor 1 is sex, smoking dominant while factor 2 is ejection_fraction and serum component dominant while factor 3 is age, anaemia, high bp dominant. While not exactly the same as intuition, we do note that Factor 1 can be interpreted as patient demographics/ lifestyle feature as males tend to smoke more, while factor 2 is the physiological makeup we discussed about earlier and factor 3 is the again patient demographics but also genetics as variables with blood pressure and anaemia show up along with age.

Conclusion -

Factor 1 - Patient Demographics / Lifestyle

Factor 2 - Patient Physiological Makeup

Factor 3 - Patient Demographics / Genetics

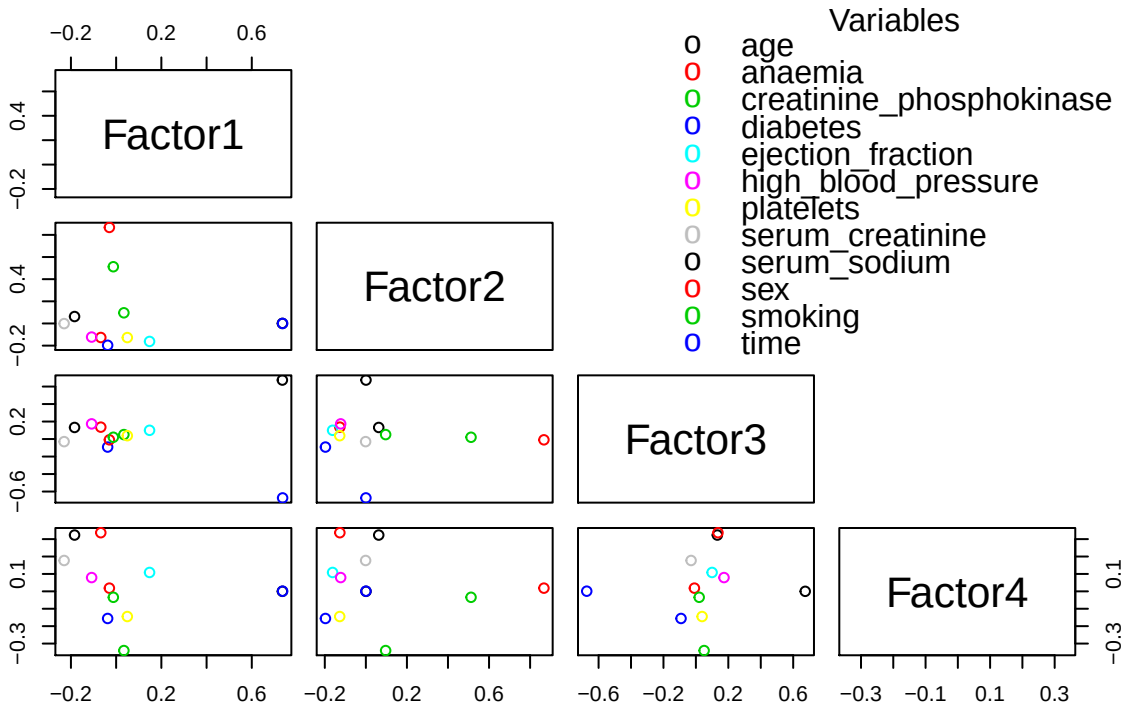
Factor Analysis (n=4 factors)

```
data.fa.none <- factanal(data_fact, factors = 4, rotation = "none")
data.fa.none
```

```
##
## Call:
## factanal(x = data_fact, factors = 4, rotation = "none")
##
## Uniquenesses:
##                      age                  anaemia creatinine_phosphokinase
##                    0.840                    0.847                    0.871
##                 diabetes        ejection_fraction      high_blood_pressure
```

```
##                     0.927                     0.930                     0.937
##                 platelets          serum_creatinine              serum_sodium
##                     0.959                     0.915                     0.005
##                       sex                   smoking                      time
##                     0.246                     0.736                     0.005
##
## Loadings:
##                           Factor1 Factor2 Factor3 Factor4
## age                       -0.184           0.132   0.323
## anaemia                           -0.127   0.136   0.337
## creatinine_phosphokinase                           -0.340
## diabetes                          -0.197           -0.156
## ejection_fraction          0.148  -0.162            0.108
## high_blood_pressure       -0.108  -0.123   0.173
## platelets                         -0.127           -0.145
## serum_creatinine          -0.230                    0.177
## serum_sodium               0.735           0.674
## sex                                0.867
## smoking                            0.512
## time                       0.736          -0.673
##
##                 Factor1 Factor2 Factor3 Factor4
## SS loadings       1.214   1.140   0.997   0.430
## Proportion Var    0.101   0.095   0.083   0.036
## Cumulative Var    0.101   0.196   0.279   0.315
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 20.17 on 24 degrees of freedom.
## The p-value is 0.687
```

```
pairs(data.fa.none$loadings, col=1:ncol(data_fact),
  upper.panel=NULL, main="Factor loadings")
par(xpd=TRUE)
legend('topright', bty='n', pch='o', col=1:ncol(data_fact),
attr(data.fa.none$loadings, 'dimnames')[[1]],  y.intersp=0.5,
title="Variables")
```



Again an interesting result since if we try and interpret the 4 factors we see that Factor 1 is serum_sodium dominant (Physiological makeup), while Factor 2 is sex and smoking dominant (Patient Lifestyle) and Factor 3 is serum_sodium and high bp dominant (Physiological makeup & lifestyle) and Factor 4 is age, anaemia dominant (Patient Demographics & genetics). We notice some overlaps here so perhaps, 3 factors would be the ideal choice, however do note that p-values aren't significant in either results.

Conclusion -

Factor 1 - Physiological makeup

Factor 2 - Patient Lifestyle

Factor 3 - Physiological makeup & lifestyle

Factor 4 - Patient demographics & Genetics

Another method - we can try the psych package as well for n=3 factors

```
fit.pc <- principal(data_fact, nfactors=3, rotate="varimax")
fit.pc
```

```
## Principal Components Analysis
## Call: principal(r = data_fact, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                             RC1   RC2   RC3   h2   u2 com
## age                        0.24  0.59 -0.14 0.43 0.57 1.5
## anaemia                   -0.12  0.51  0.05 0.28 0.72 1.1
## creatinine_phosphokinase  0.10 -0.36 -0.01 0.14 0.86 1.2
```

```
## diabetes                -0.55 -0.20 -0.14 0.36 0.64 1.4
## ejection_fraction       -0.08  0.24  0.55 0.37 0.63 1.4
## high_blood_pressure     -0.05  0.47  0.14 0.24 0.76 1.2
## platelets               -0.16 -0.05  0.28 0.11 0.89 1.6
## serum_creatinine        -0.02  0.32 -0.50 0.35 0.65 1.7
## serum_sodium             0.17  0.03  0.76 0.60 0.40 1.1
## sex                      0.76 -0.19 -0.21 0.66 0.34 1.3
## smoking                  0.74 -0.16 -0.04 0.57 0.43 1.1
## time                    -0.12 -0.58  0.26 0.41 0.59 1.5
##
##                        RC1  RC2  RC3
## SS loadings           1.58 1.55 1.38
## Proportion Var        0.13 0.13 0.12
## Cumulative Var        0.13 0.26 0.38
## Proportion Explained  0.35 0.34 0.31
## Cumulative Proportion 0.35 0.69 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.11
##  with the empirical chi square  458.49  with prob <  2.2e-76
##
## Fit based upon off diagonal values = -0.09
```

```
round(fit.pc$values, 3)
```

```
##  [1] 1.663 1.579 1.268 1.116 1.030 0.989 0.883 0.845 0.764 0.708 0.644
## [12] 0.511
```

```
fit.pc$loadings
```

```
##
## Loadings:
##                           RC1    RC2    RC3
## age                        0.240  0.591 -0.142
## anaemia                   -0.122  0.510
## creatinine_phosphokinase        -0.357
## diabetes                  -0.550 -0.201 -0.143
## ejection_fraction                0.240  0.550
## high_blood_pressure              0.465  0.140
## platelets                 -0.158         0.282
## serum_creatinine                 0.318 -0.500
## serum_sodium               0.169         0.757
## sex                        0.759 -0.192 -0.210
## smoking                    0.737 -0.159
## time                      -0.118 -0.576  0.262
##
##                  RC1   RC2   RC3
## SS loadings    1.580 1.548 1.382
## Proportion Var 0.132 0.129 0.115
## Cumulative Var 0.132 0.261 0.376
```

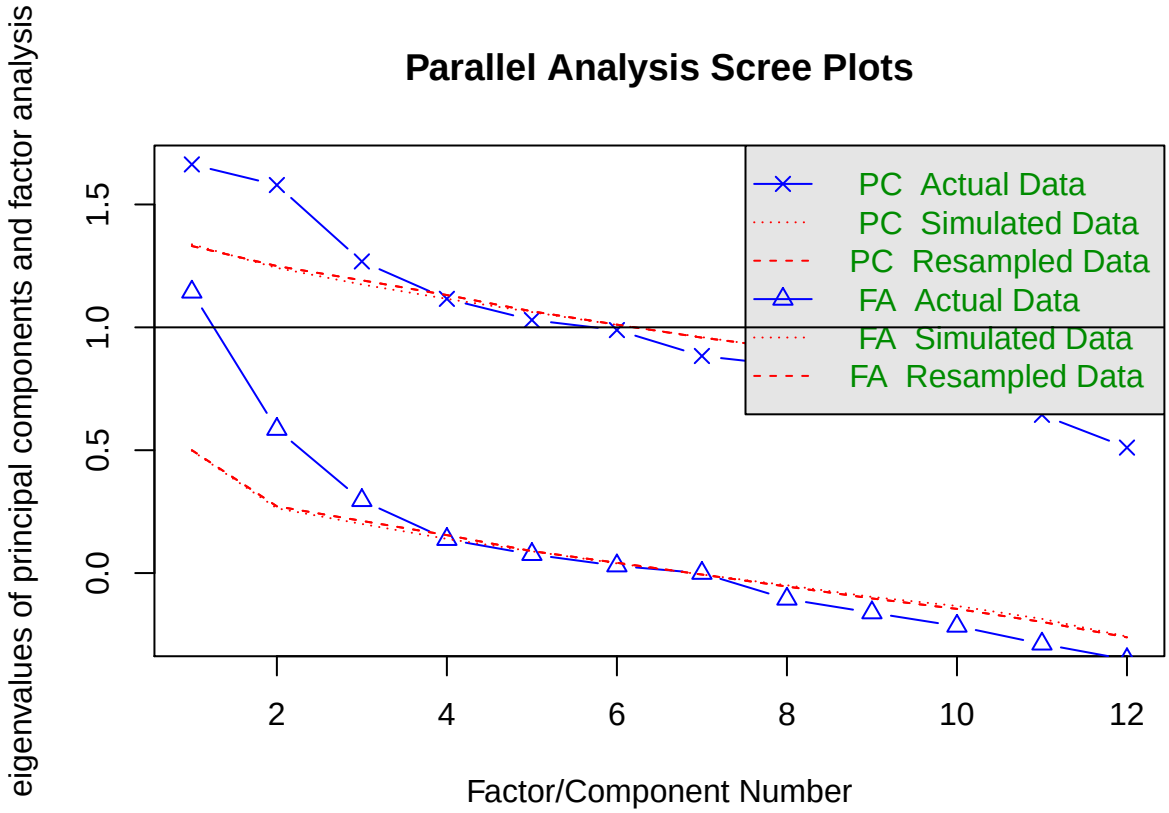
```
# Loadings with more digits
for (i in c(1,2,3)) { print(fit.pc$loadings[[1,i]])}
```

```
## [1] 0.2400349
## [1] 0.5907011
## [1] -0.1417927
```

```
# Communalities
fit.pc$communality
```

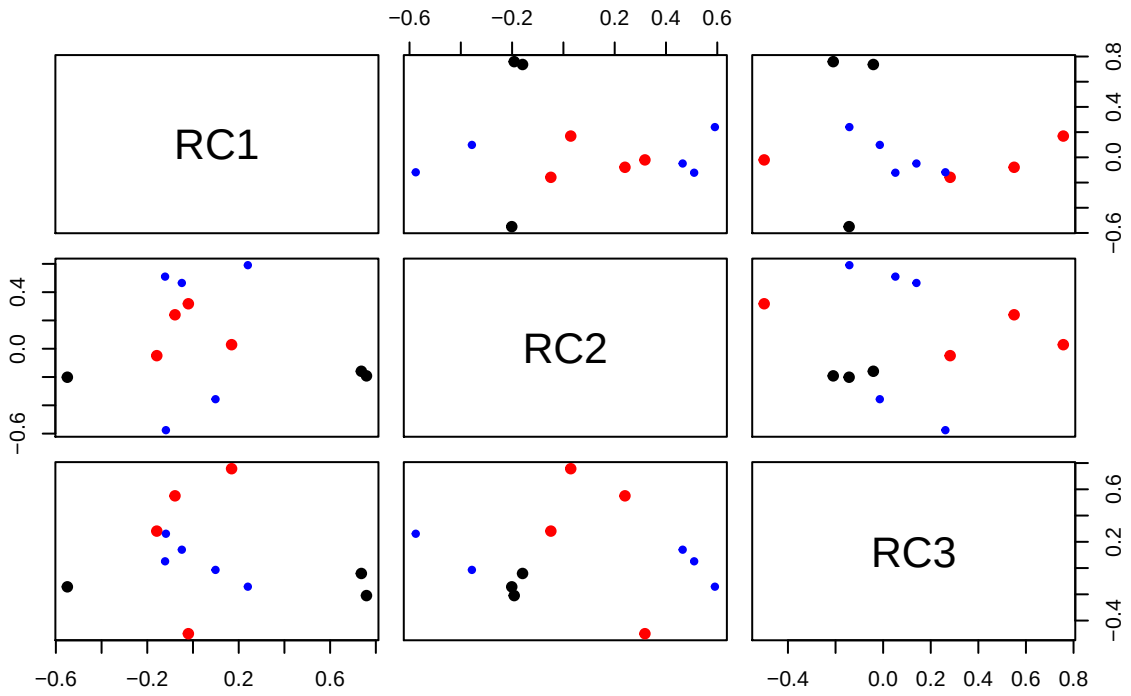
```
##                      age                  anaemia creatinine_phosphokinase
##                0.4266497                0.2774545                0.1372058
##                 diabetes         ejection_fraction      high_blood_pressure
##                0.3632012                0.3666522                0.2381992
##                platelets          serum_creatinine             serum_sodium
##                0.1067861                0.3512908                0.6023730
##                      sex                  smoking                     time
##                0.6569339                0.5695994                0.4138580
```

```
# Play with FA utilities
fa.parallel(data_fact) # See factor recommendation
```

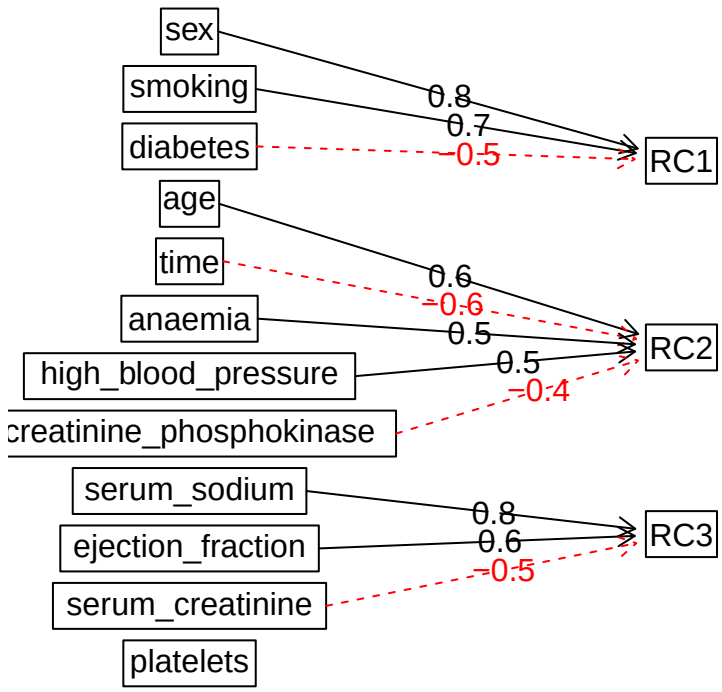



```
## Parallel analysis suggests that the number of factors =  3  and the number of components =  3
```

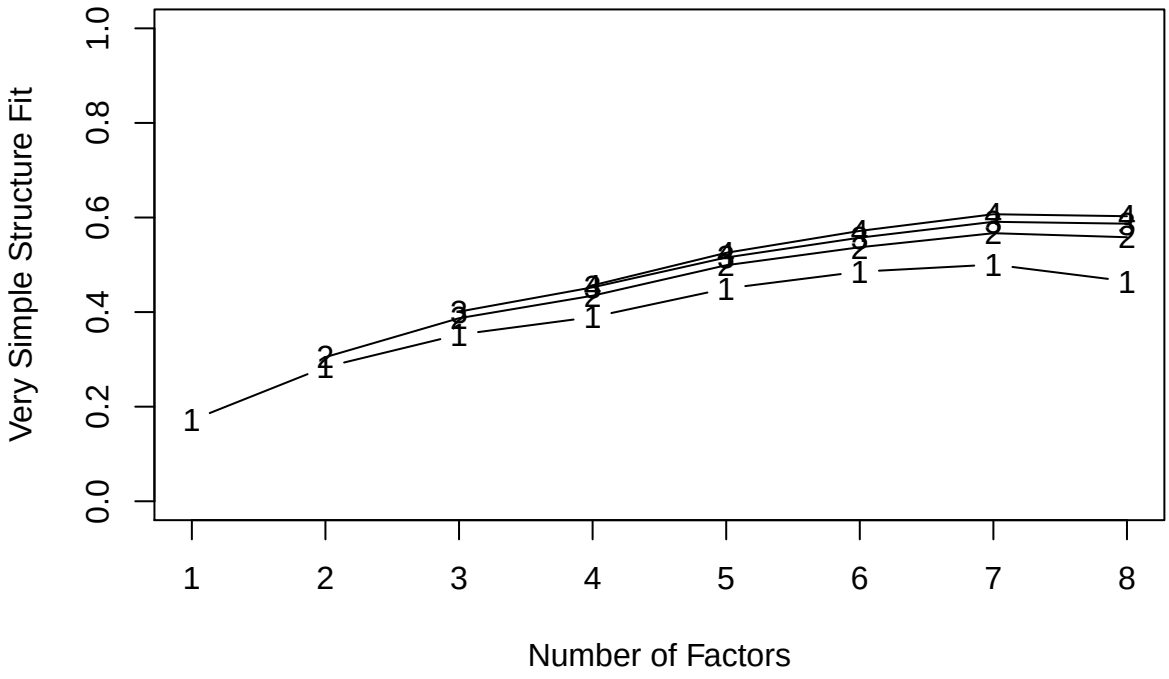
```
fa.plot(fit.pc) # See Correlations within Factors
```




```
fa.diagram(fit.pc) # Visualize the relationship
```




```
vss(data_fact) # See Factor recommendations for a simple structure
```




```
##
## Very Simple Structure
## Call: vss(x = data_fact)
## VSS complexity 1 achieves a maximimum of 0.5  with  7  factors
## VSS complexity 2 achieves a maximimum of 0.57  with  7  factors
##
## The Velicer MAP achieves a minimum of 0.02  with  1  factors
## BIC achieves a minimum of  NA  with  1  factors
## Sample Size adjusted BIC achieves a minimum of  NA  with  2  factors
##
## Statistics by number of factors
##   vss1 vss2   map dof   chisq    prob sqresid  fit  RMSEA  BIC SABIC
## 1 0.17 0.00 0.016  54 1.1e+02 1.6e-05    11.1 0.17 0.0593 -199 -28.1
## 2 0.28 0.30 0.023  43 5.8e+01 6.8e-02     9.3 0.30 0.0351 -188 -51.2
## 3 0.35 0.39 0.035  33 3.2e+01 4.9e-01     8.0 0.40 0.0049 -156 -51.0
## 4 0.39 0.43 0.049  24 2.1e+01 6.3e-01     7.3 0.46 0.0000 -116 -39.5
## 5 0.45 0.50 0.069  16 9.6e+00 8.9e-01     6.3 0.53 0.0000  -82 -30.8
## 6 0.49 0.54 0.096   9 2.6e+00 9.8e-01     5.6 0.58 0.0000  -49 -20.1
## 7 0.50 0.57 0.141   3 1.1e+00 7.9e-01     5.1 0.62 0.0000  -16  -6.5
## 8 0.47 0.56 0.197  -2 8.1e-07      NA     5.1 0.62     NA   NA    NA
##   complex  eChisq    SRMR eCRMS eBIC
## 1     1.0 2.2e+02 7.5e-02 0.083  -85
## 2     1.2 9.6e+01 4.9e-02 0.061 -149
## 3     1.4 4.8e+01 3.5e-02 0.049 -140
## 4     1.6 2.9e+01 2.7e-02 0.045 -108
## 5     1.7 1.3e+01 1.8e-02 0.037  -78
## 6     1.9 3.2e+00 9.0e-03 0.024  -48
## 7     1.8 1.0e+00 5.1e-03 0.024  -16
## 8     1.7 8.6e-07 4.7e-06    NA   NA
```

Again, we note that the analysis recommends 3 factors and we see the interpretation as Factor 1 - Sex, Smoking (Patient Demographics/ Lifestyle) and Factor 2 - Age, anaemia, high_bp (Patient Demographics/ Genetics) and Factor 3 - Serum_sodium, ejection_fraction (Physiological make-up) which confirms our earlier interpretation as well

Note: While we see our data isn't perhaps ideal for Factor Analysis, we can gauge some interesting results and given this dataset is part of a study of only 299 patients, the latent factors may be more prominent in the population distribution.

This concludes our approach to Factor Analysis in our dataset