Dimension reduction via PCA and FA

Kyle Barisone

Principal Components

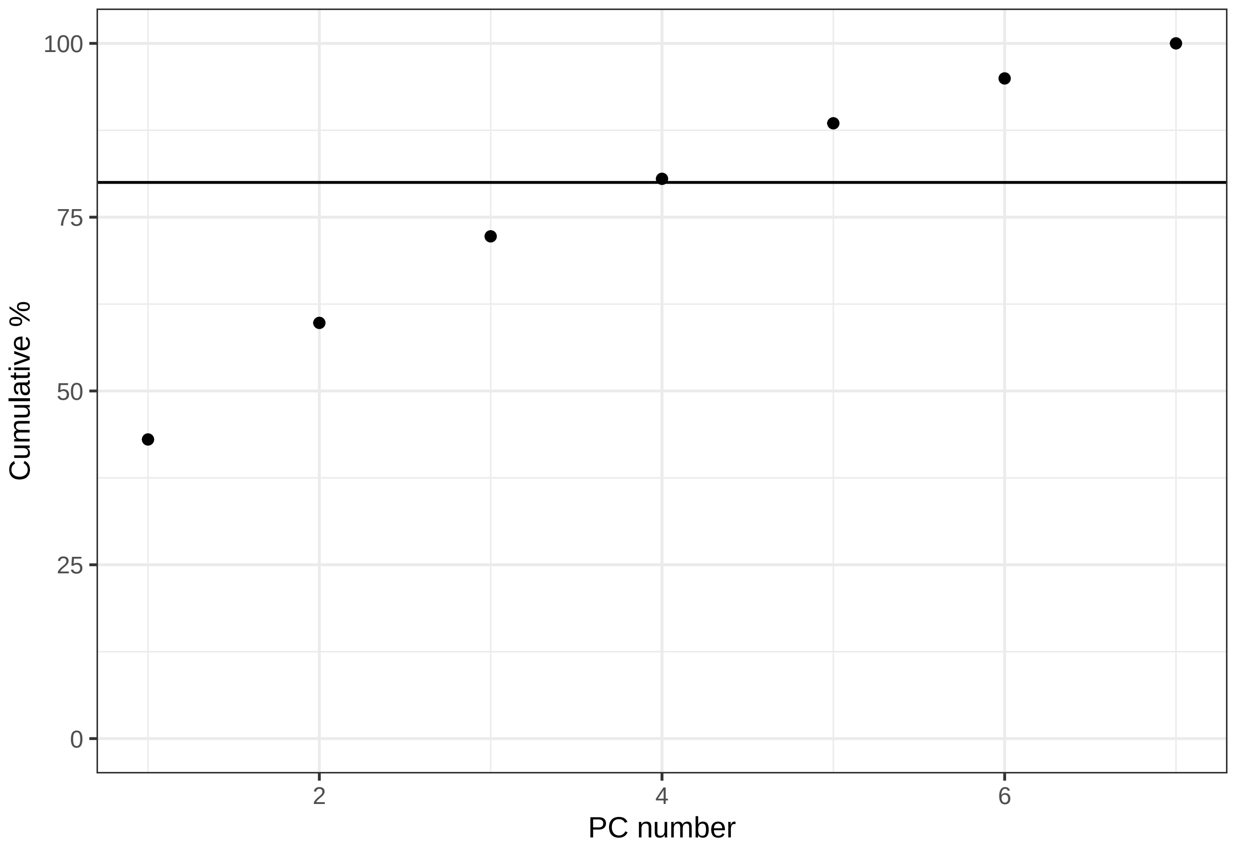
1. PMA6 14.1 (modified): For the depression data set, perform a PCA on the last seven variables DRINK-CHRONILL.

Use the covariance matrix, but *do not center or scale* the data. You should have the codebook for this data set open during this homework.

```
Depress <- readRDS("C:/Users/KBari/OneDrive/Desktop/Math 456/Depression_data.rds")
```

a) Determine the number of PC's to retain that contain 80% of the original variance.

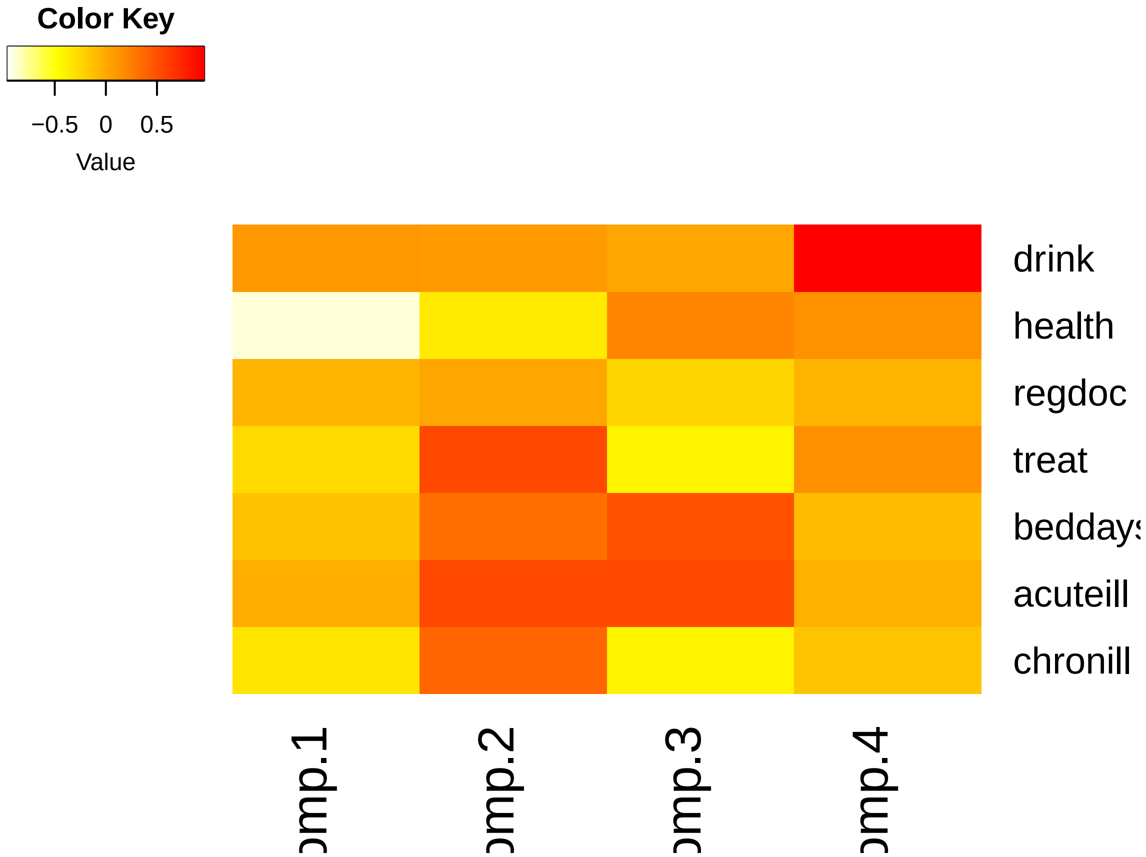
```
pc_dep <- princomp(Depress[,31:37])


var_pc <- (pc_dep$sdev)^2

qplot(x=1:7, y=cumsum(var_pc)/sum(var_pc)*100, geom="point") +
  xlab("PC number") + ylab("Cumulative %") + ylim(c(0,100)) +
  geom_hline(aes(yintercept=80))
```



4 principal components retain about 81% of the original variance. So we need to retain 4 principal components to maintain at least 80% of the original variance.

b) Examine the loadings of the retained PC's using a heatmap. Interpret each PC as it relates to the individual questions.

```r
heatmap.2(pc_dep$loadings[,1:4], scale="none", Rowv=NA, Colv=NA, density.info="none",
          dendrogram="none", trace="none", col=rev(heat.colors(256)))
```



PC1 has no variables with high positive scores and has a very low negative score for health.

PC2 has high positive scores for chronic illness, acute illness, days spent in bed, and treatment given by a doctor.

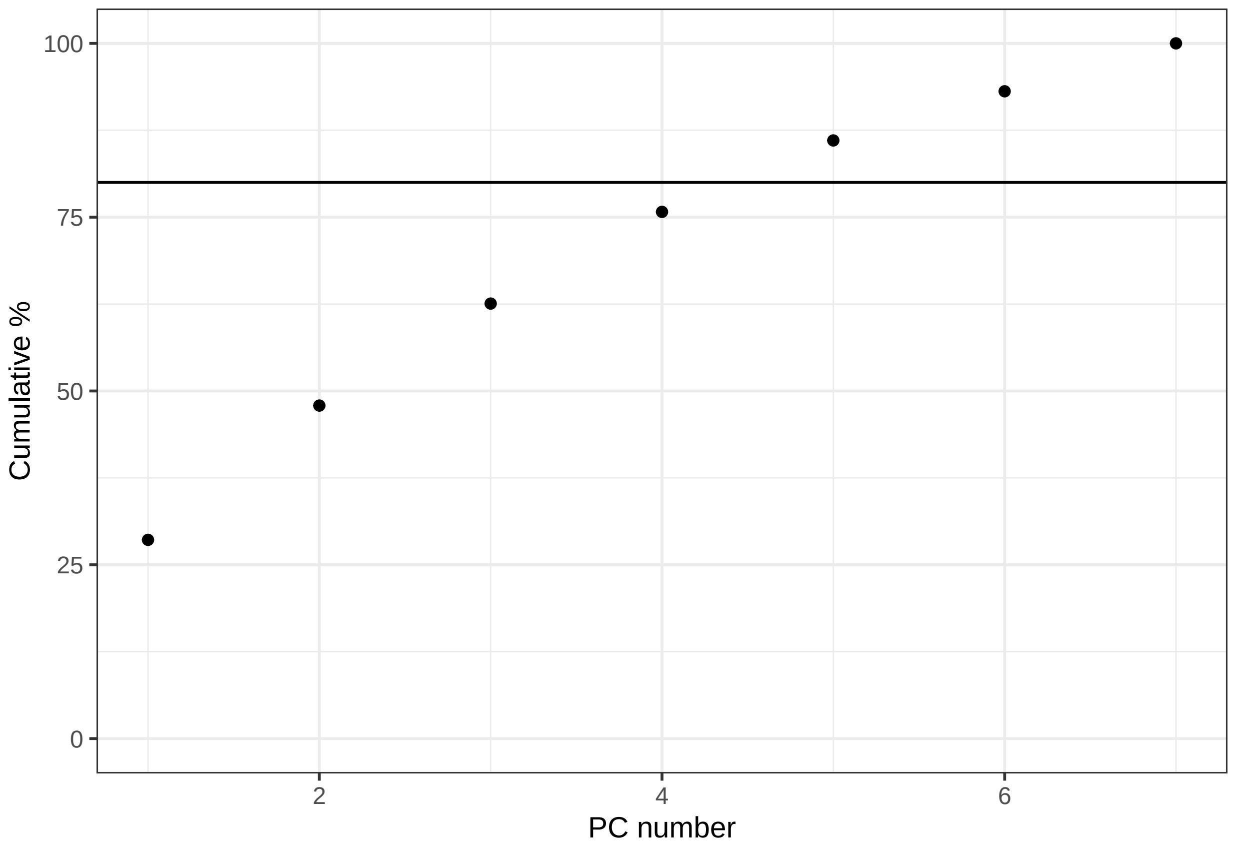
PC3 has high positive scores for acute illness, days spent in bed, and health.

PC4 has high positive scores for overall health, treatment from a doctor, and if the individual is a regular drinker.

2. Repeat question #1 using the correlation matrix instead of the covariance matrix.

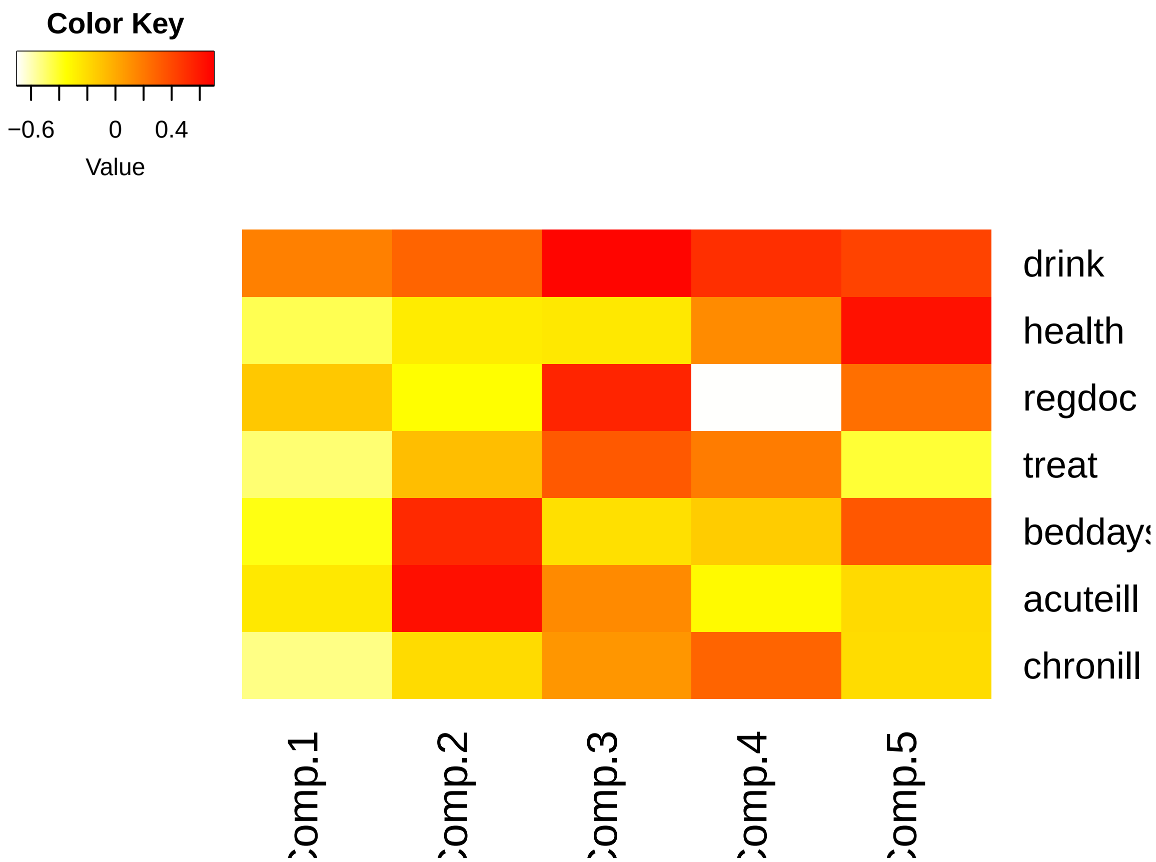
Compare the results and comment. (Are the same number of PC's retained? Are the loadings different?)

```
pc_dep <- princomp(Depress[,31:37], cor = TRUE)

var_pc <- (pc_dep$sdev)^2

qplot(x=1:7, y=cumsum(var_pc)/sum(var_pc)*100, geom="point") +
  xlab("PC number") + ylab("Cumulative %") + ylim(c(0,100)) +
  geom_hline(aes(yintercept=80))
```



When using the correlation matrix, we need to retain 5 principal components to maintain at least 80% of the original variance.

```
heatmap.2(pc_dep$loadings[,1:5], scale="none", Rowv=NA, Colv=NA, density.info="none",
          dendrogram="none", trace="none", col=rev(heat.colors(256)))
```



Whether or not the respondant is a regular drinker seems to have a high positive loading for all components but is highest in component 3.

For PC1 signs of loading are all fairly low negatively except for if they are a regular drinker or not and if they regularly see a doctor.

People with high values for PC2 tend to have more acute illness and more days spent in bed.

Respondants with high values for PC3 tend to regularly see a doctor and get treatment or medicine prescribed by a doctor.

People with high values for PC4 tend to score high on health, treatment, and chronic illness but do not regularly see a doctor.

People with high positive values for PC5 tend to regularly see a doctor and spend a lot of days in bed but have good overall health.

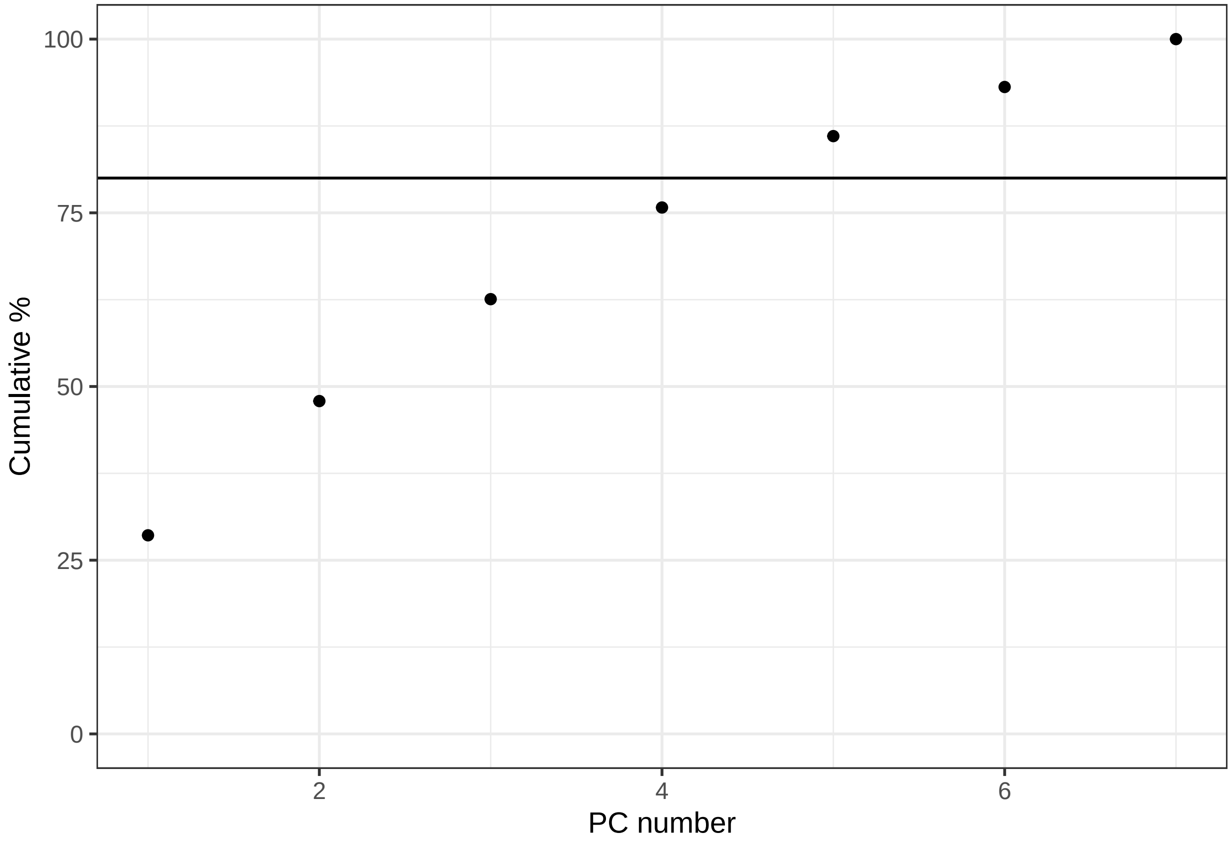
People who regularly see a doctor score high on pc3 and pc5 but low on pc4.

The score on the general health of the respondant increases with each component.

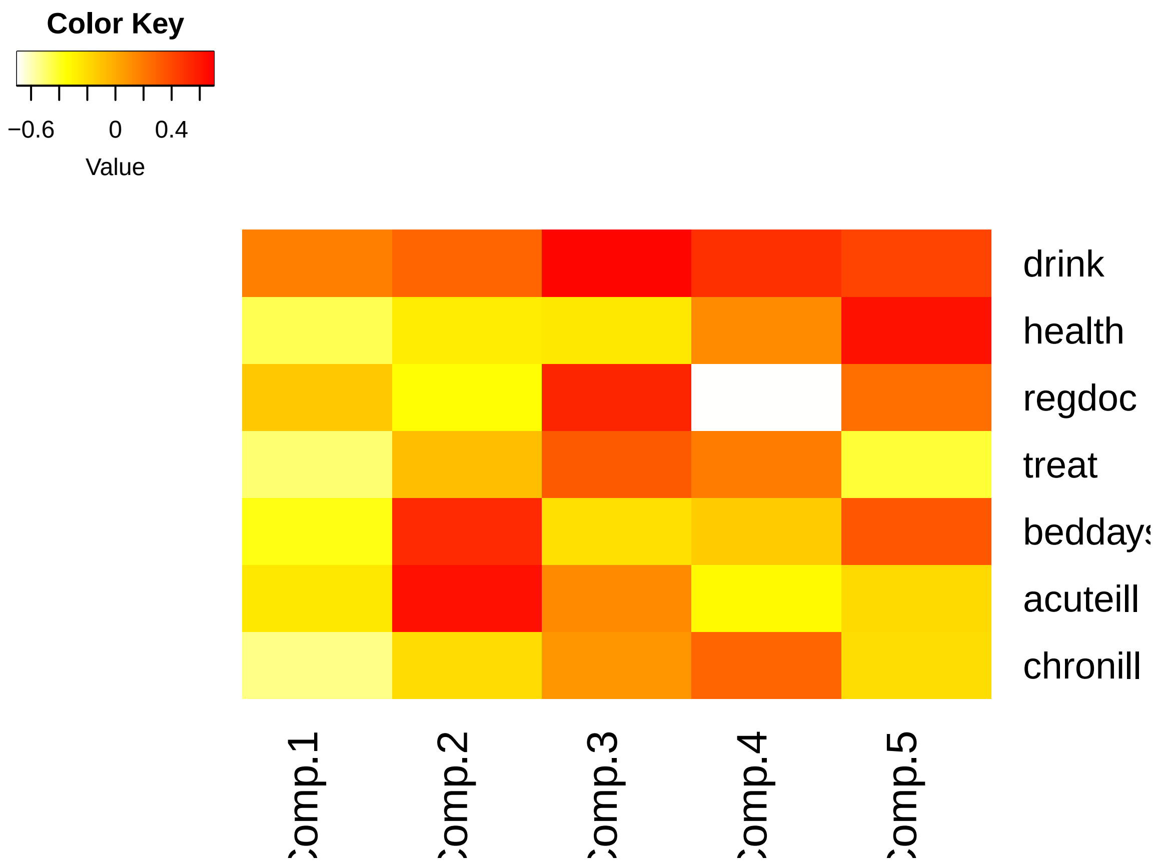
The loadings and the number of PC's retained are different from the covariant matrix.

3. Repeat question #1 after normalizing the data (centering and scaling). Use the `scale()` function here to help. Compare the results and comment. (Are the same number of PC's retained? Are the loadings different?)

```
Depress_scaled <- scale(Depress[,31:37], center = TRUE, scale = TRUE)

pc_dep <- princomp(Depress_scaled)

var_pc <- (pc_dep$sdev)^2

qplot(x=1:7, y=cumsum(var_pc)/sum(var_pc)*100, geom="point") +
  xlab("PC number") + ylab("Cumulative %") + ylim(c(0,100)) +
  geom_hline(aes(yintercept=80))
```



```
heatmap.2(pc_dep$loadings[,1:5], scale="none", Rowv=NA, Colv=NA, density.info="none",
          dendrogram="none", trace="none", col=rev(heat.colors(256)))
```



After Centering and scaling, the number of pc's we retain are 5 which is the same amount as when we used the correlation matrix from problem 2. The loadings are also the same from the correlation matrix. It seems that scaling and centering the data makes the covariant matrix more similar to the correlation matrix.

4 (PMA6 14.2 modified): Perform a regression analysis of `CASES` on your retained PC's *from question 3* Interpret the results. Recall that PC's are on the scale of a standard deviation of the PC scale. So you can say "for every one standard deviation on the PC1 scale an individual is. . . .". But don't just call it "PC1". Use your interpretation from the loadings visualized in question 3.

```
dim(pc_dep$scores); kable(pc_dep$scores[1:7, 1:7])
```

```
## [1] 294   7
```

Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
-1.1833169	-1.7622066	-0.6956641	-0.5417424	-1.2224590	-0.1296571	-0.1985339
-0.2143059	-0.7265186	1.2768213	0.5525415	-0.9242934	0.0441769	-0.7822709
0.2979064	-0.6553579	0.8038988	0.1326571	0.2290807	-1.3584922	0.4203031
0.3756555	-1.2837367	-1.0513930	-1.0721879	-1.1181198	1.2009052	-0.9483610
-0.6289723	2.3379282	0.8539207	-1.0622428	-0.1034075	-1.2009070	-0.4411602
-0.7728325	0.6672917	1.5641769	-0.1547904	-1.3423968	0.6829014	0.3851814
-2.2892317	-0.6919495	-0.7094596	-1.0957841	-0.8959783	0.4510110	1.5587501

```
Depress$pc1 <- pc_dep$scores[,1]
Depress$pc2 <- pc_dep$scores[,2]
Depress$pc3 <- pc_dep$scores[,3]
Depress$pc4 <- pc_dep$scores[,4]
Depress$pc5 <- pc_dep$scores[,5]

PC.model <- glm(cases~pc1+pc2+pc3+pc4+pc5, data=Depress, family='binomial')

summary(PC.model)
```

```
##
## Call:
## glm(formula = cases ~ pc1 + pc2 + pc3 + pc4 + pc5, family = "binomial",
##     data = Depress)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4727  -0.6188  -0.5007  -0.4124   2.2391
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7256     0.1729  -9.979  < 2e-16 ***
## pc1          -0.3233     0.1110  -2.914  0.00357 **
## pc2           0.2497     0.1275   1.959  0.05014 .
## pc3          -0.1210     0.1576  -0.768  0.44259
## pc4           0.2592     0.1656   1.565  0.11758
## pc5           0.3170     0.1902   1.667  0.09555 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.12  on 293  degrees of freedom
## Residual deviance: 248.42  on 288  degrees of freedom
## AIC: 260.42
##
## Number of Fisher Scoring iterations: 4
```

```
confint(PC.model)
```

```
##                     2.5 %     97.5 %
## (Intercept) -2.0814415870 -1.4010013
## pc1         -0.5462259038 -0.1091326
## pc2         -0.0006852559  0.5016635
## pc3         -0.4279539034  0.1947544
## pc4         -0.0651229781  0.5879940
## pc5         -0.0516730106  0.6975304
```

Controlling for all other PCs, for every 1 standard deviation on the PC1 scale (or individuals who are regular drinkers who see a doctor regularly and are fairly healthy) an individual is 0.323 (0.109, 0.546) less likely to be depressed (p=0.004).

Controlling for all other PCs, for every 1 standard deviation on the PC2 scale (or people who spend a lot of days in bed and have acute illness) an individual is .250 (0, .502) more likely to be depressed (p=0.050).

Controlling for all other PCs, for every 1 standard deviation on the PC3 scale (or people who regularly see a doctor and receive treatment) scale an individual is .121 (-.428, .195) times less likely to be depressed (p=0.443). However, this is a highly insignificant p-value.

Controlling for all other PCs, for every 1 standard deviation on the PC4 scale (or people with chronic illness who do not see a doctor regularly) an individual is .259 (-.065, .588) more likely to be depressed (p=0.118). However, this is not a significant p-value.

Controlling for all other PCs, for every 1 standard deviation on the PC5 scale (or people in good health who regularly see a doctor but do not recieve much treatment) an individual is .317 (-.052, .698) more likely to be depressed (p=0.096).

5 (PMA6 14.11 modified): This question uses the Parental HIV data set.

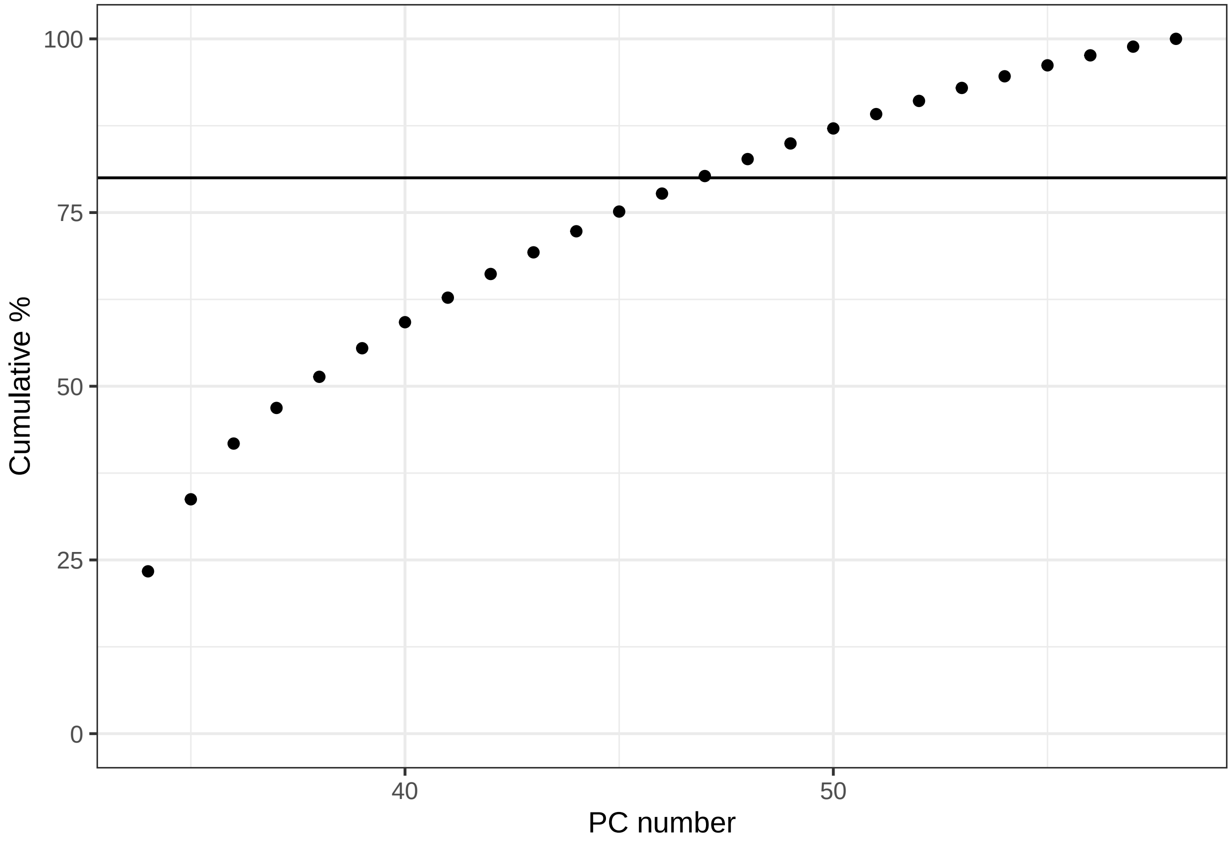
a) If you were to conduct a PCA on the items of the Parental Bonding scale, *a priori* how many PC's would you expect to retain for this scale?

```
HIV <- read.csv("C:/Users/KBari/OneDrive/Desktop/Math 456/Parhiv.csv")
View(HIV)
```

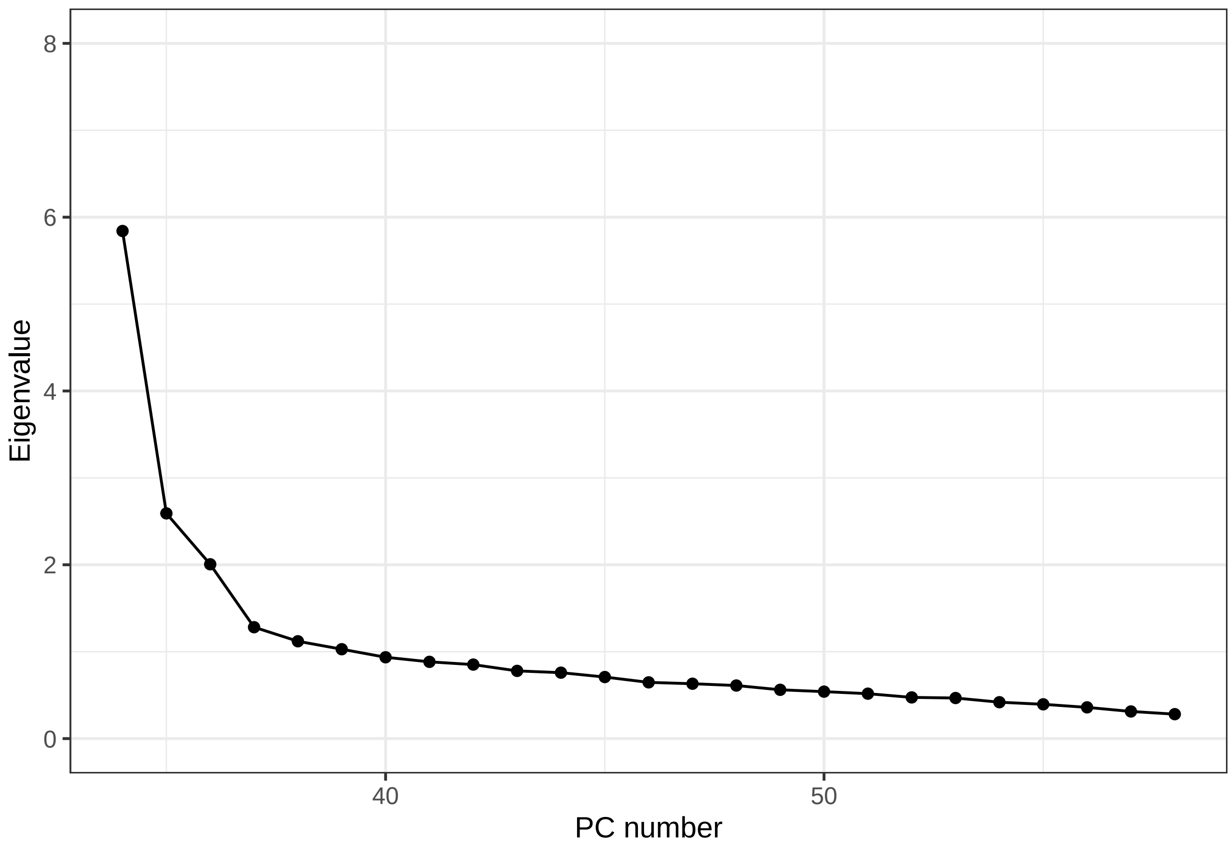
I would expect to retain around 10 variables because looking at the codebook, a lot of these questions seem like they are redundant and might have similar answers.

b) Perform this PCA. How many components should be retained based on the rules of thumb mentioned in Section 14.5 (% variance retained, Eigenvalues > 1, scree plot)?

```
pc_HIV <- princomp(na.omit(HIV[,34:58]), cor = TRUE)

var_pc <- (pc_HIV$sdev)^2

qplot(x=34:58, y=cumsum(var_pc)/sum(var_pc)*100, geom="point") +
  xlab("PC number") + ylab("Cumulative %") + ylim(c(0,100)) +
  geom_hline(aes(yintercept=80))
```



```r
qplot(x=34:58, y=var_pc, geom=c("point", "line")) +
  xlab("PC number") + ylab("Eigenvalue") + ylim(c(0,8))
```



At least 4 components should be retained looking at the scree plot. Using the elbow rule, it starts to flatten out around the 4th component. 14 components lets us retain about 80% of the original variance if we are using the first plot. Finally 7 components have eigenvalues greater than 1.

By these graphs, there is not clear amount of principal components to retain, so based off the total variance graph and the eigenvalues, we should retain at least 7 principal components for analysis.

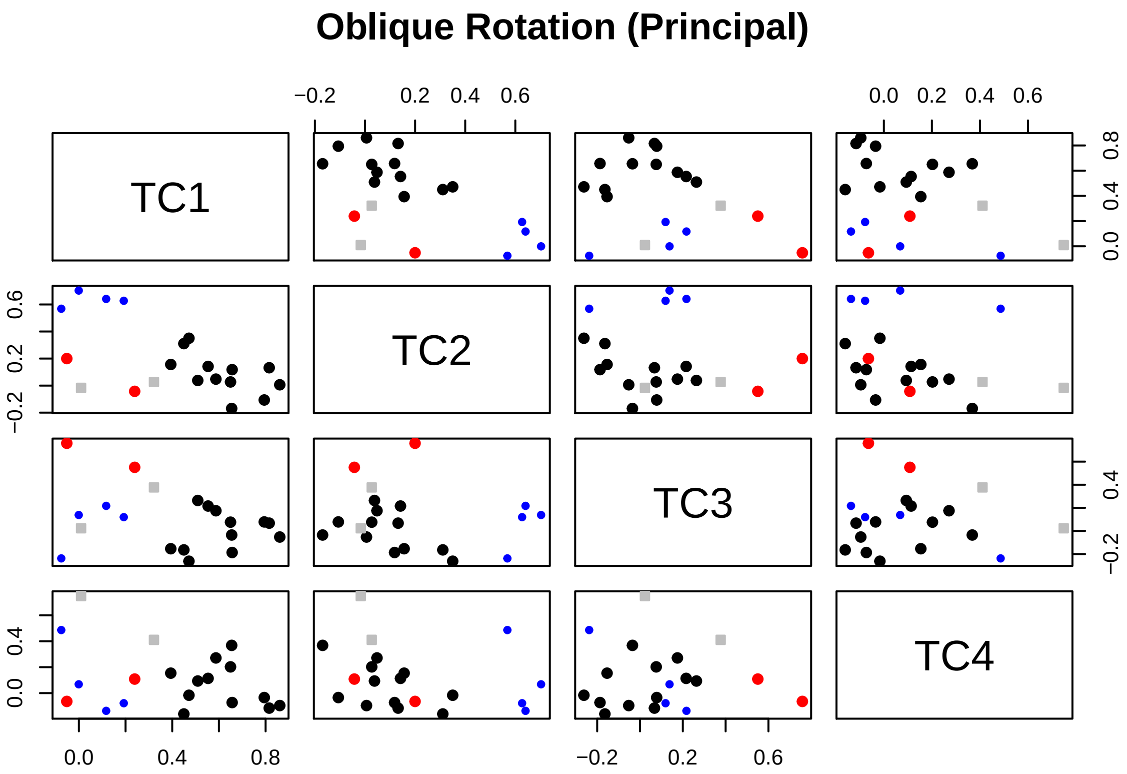
Factor Analysis

1. (PMA6 15.1). The CESD scale items (`C1-C20`) from the depression data set were used to obtain the factor loadings listed in Table 15.7. The initial factor solution was obtained from the principal components method, and a varimax rotation was performed. Analyze this same data set by using an oblique rotation such as the direct quartimin procedure. Compare the results.

```
Depress <- readRDS("C:/Users/KBari/OneDrive/Desktop/Math 456/Depression_data.rds")

cesd.scale <- scale(Depress[c(9:28)])

pc.extract.quartimin <- principal(cesd.scale, nfactors=4, rotate="quartimin")

print(pc.extract.quartimin)
```

```
## Principal Components Analysis
## Call: principal(r = cesd.scale, nfactors = 4, rotate = "quartimin")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       TC1   TC2   TC3   TC4   h2   u2 com
## c1   0.65  0.03  0.08  0.20 0.58 0.42 1.2
## c2   0.82  0.13  0.07 -0.12 0.76 0.24 1.1
## c3   0.79 -0.11  0.08 -0.03 0.61 0.39 1.1
## c4   0.65 -0.17 -0.04  0.37 0.61 0.39 1.7
## c5   0.86  0.01 -0.05 -0.10 0.69 0.31 1.0
## c6   0.66  0.12 -0.19 -0.07 0.44 0.56 1.3
## c7   0.59  0.05  0.17  0.27 0.62 0.38 1.6
## c8   0.01 -0.02  0.02  0.75 0.57 0.43 1.0
## c9   0.24 -0.04  0.55  0.11 0.45 0.55 1.5
## c10  0.55  0.14  0.22  0.11 0.55 0.45 1.5
## c11  0.51  0.04  0.26  0.09 0.46 0.54 1.6
## c12  0.45  0.31 -0.16 -0.16 0.36 0.64 2.4
## c13 -0.08  0.57 -0.24  0.49 0.58 0.42 2.4
## c14  0.00  0.70  0.14  0.07 0.55 0.45 1.1
## c15  0.47  0.35 -0.26 -0.02 0.44 0.56 2.5
## c16  0.12  0.64  0.22 -0.14 0.56 0.44 1.4
## c17  0.19  0.63  0.12 -0.08 0.55 0.45 1.3
## c18  0.39  0.16 -0.15  0.15 0.26 0.74 2.0
## c19 -0.05  0.20  0.76 -0.06 0.62 0.38 1.2
## c20  0.32  0.03  0.38  0.41 0.58 0.42 2.9
##
##                        TC1  TC2  TC3  TC4
## SS loadings           5.48 2.26 1.61 1.50
## Proportion Var        0.27 0.11 0.08 0.07
## Cumulative Var        0.27 0.39 0.47 0.54
## Proportion Explained  0.51 0.21 0.15 0.14
## Cumulative Proportion 0.51 0.71 0.86 1.00
##
##  With component correlations of
##      TC1  TC2  TC3  TC4
## TC1 1.00 0.36 0.28 0.25
## TC2 0.36 1.00 0.11 0.07
## TC3 0.28 0.11 1.00 0.06
## TC4 0.25 0.07 0.06 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  526.92  with prob <  1.1e-53
##
## Fit based upon off diagonal values = 0.96
```

```
plot(pc.extract.quartimin, title="Oblique Rotation (Principal)")
```



Using the oblique quartimin rotation, we see similar results from the table in the textbook. We see a 3% higher proportion of variance in TC1 and a slightly lower proportion of variance for TC2,TC3, and TC4. All values are fairly similar though.

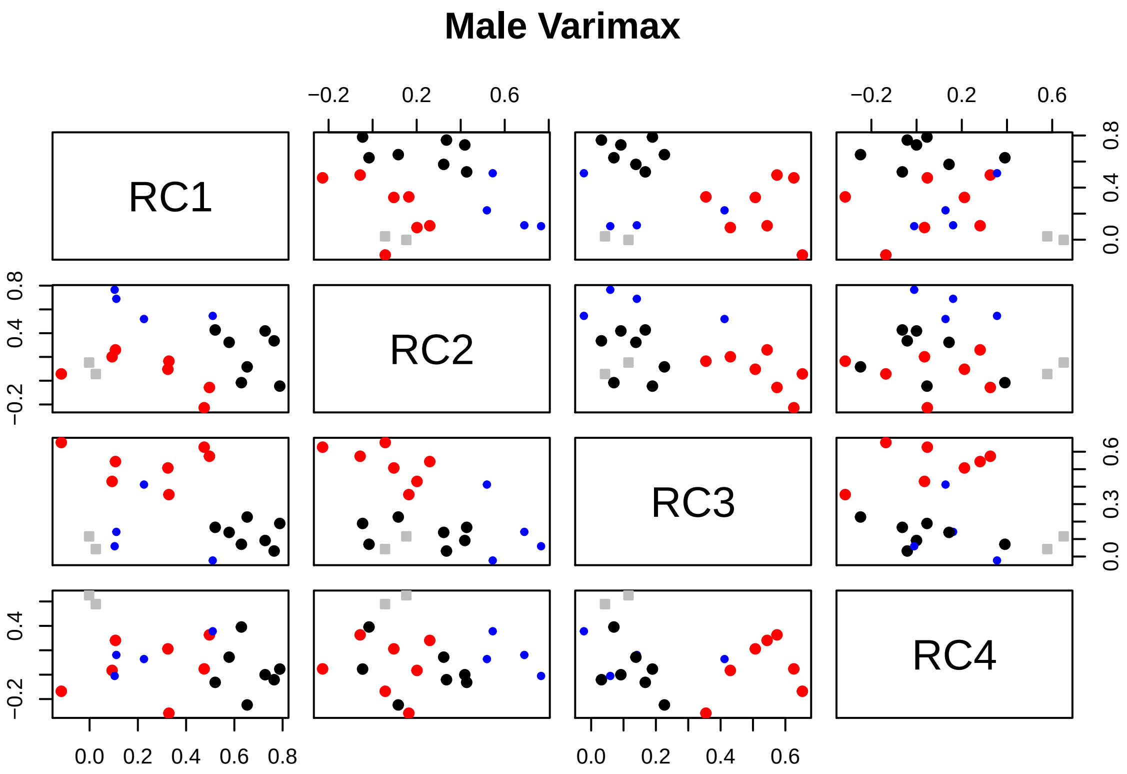
2. (PMA6 15.6) Separate the depression data set into two subgroups, men and women. Using four factors, repeat the factor analysis in Table 15.7. Compare the results of your two factor analyses to each other, and do the results in Table 15.7.

```
Depress$sex <- ifelse(Depress$sex == 0, "male", "female")

#separates males and females
dep.male<- filter(Depress, sex=="male")
dep.female<- filter(Depress, sex=="female")

male.scale<- scale(dep.male[c(9:28)])
male.factor <- principal(male.scale, nfactors=4, rotate="varimax")

female.scale<- scale(dep.female[c(9:28)])
female.factor <- principal(female.scale, nfactors=4, rotate="varimax")
```

```r
#plot for males
print(male.factor)
```

```
## Principal Components Analysis
## Call: principal(r = male.scale, nfactors = 4, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       RC1   RC2   RC3   RC4   h2   u2 com
## c1   0.33  0.16  0.35 -0.32 0.36 0.64 3.4
## c2   0.73  0.42  0.09  0.00 0.71 0.29 1.6
## c3   0.79 -0.05  0.19  0.05 0.66 0.34 1.1
## c4   0.47 -0.23  0.63  0.05 0.67 0.33 2.2
## c5   0.76  0.34  0.03 -0.04 0.70 0.30 1.4
## c6   0.65  0.12  0.23 -0.25 0.55 0.45 1.6
## c7   0.50 -0.06  0.57  0.33 0.69 0.31 2.6
## c8  -0.12  0.06  0.65 -0.14 0.46 0.54 1.2
## c9   0.00  0.15  0.12  0.65 0.46 0.54 1.2
## c10  0.23  0.52  0.41  0.13 0.51 0.49 2.5
## c11  0.11  0.26  0.54  0.28 0.45 0.55 2.1
## c12  0.11  0.69  0.14  0.16 0.53 0.47 1.3
## c13  0.09  0.20  0.43  0.04 0.24 0.76 1.5
## c14  0.52  0.43  0.17 -0.06 0.49 0.51 2.2
## c15  0.10  0.77  0.06 -0.01 0.60 0.40 1.0
## c16  0.51  0.55 -0.02  0.36 0.69 0.31 2.7
## c17  0.58  0.32  0.14  0.14 0.48 0.52 1.9
## c18  0.03  0.06  0.04  0.58 0.34 0.66 1.0
## c19  0.63 -0.02  0.07  0.39 0.55 0.45 1.7
## c20  0.32  0.10  0.51  0.21 0.42 0.58 2.2
##
##                        RC1  RC2  RC3  RC4
## SS loadings           4.22 2.45 2.37 1.52
## Proportion Var        0.21 0.12 0.12 0.08
## Cumulative Var        0.21 0.33 0.45 0.53
## Proportion Explained  0.40 0.23 0.22 0.14
## Cumulative Proportion 0.40 0.63 0.86 1.00
##
## Mean item complexity =  1.8
## 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  251.98  with prob <  4.5e-12
##
## Fit based upon off diagonal values = 0.93
```

```r
plot(male.factor, title="Male Varimax")
```



Black dots have high positive correlation in RC1 and lower correlation in RC3.

Blue dots tend to have high positive correlation in RC2.

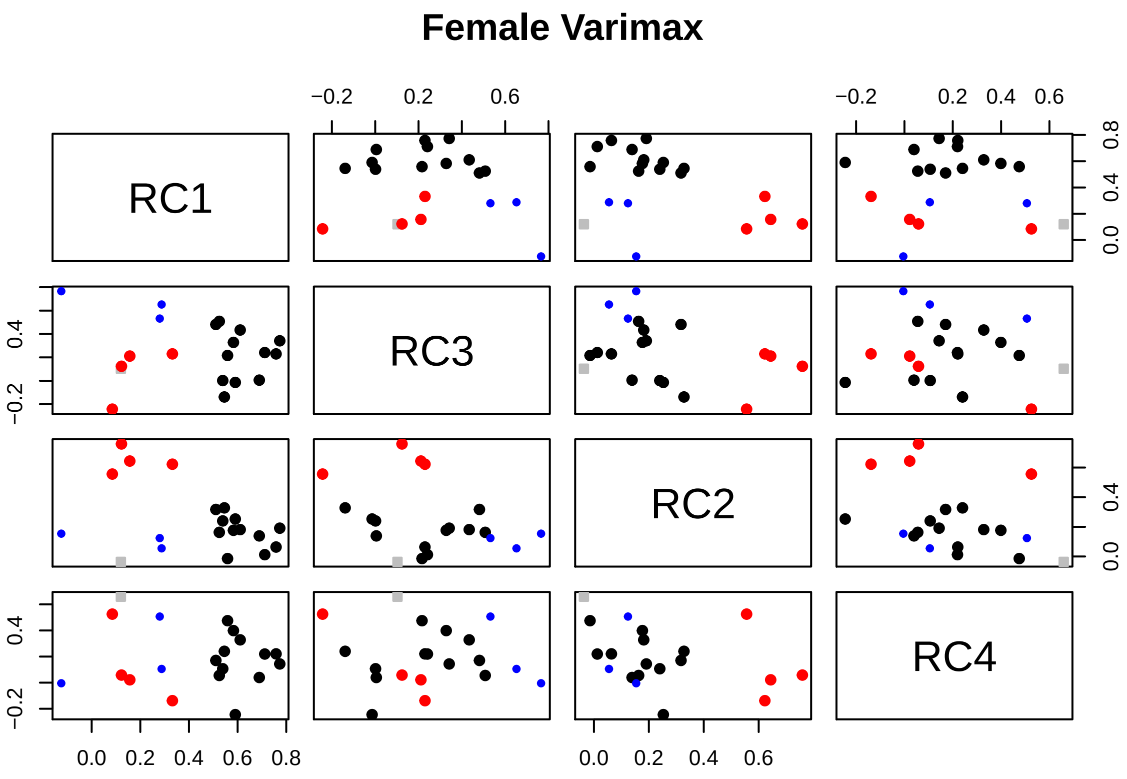
Red dots tend to have high positive correlation in RC3.

Gray squares have a high correlation in RC4.

```
#plot for females
print(female.factor)
```

```
## Principal Components Analysis
## Call: principal(r = female.scale, nfactors = 4, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##       RC1   RC3   RC2   RC4   h2   u2 com
## c1   0.61  0.43  0.18  0.33 0.70 0.30 2.6
## c2   0.77  0.34  0.19  0.14 0.77 0.23 1.6
## c3   0.71  0.24  0.01  0.22 0.61 0.39 1.4
## c4   0.56  0.22 -0.01  0.48 0.58 0.42 2.3
## c5   0.76  0.23  0.06  0.22 0.68 0.32 1.4
## c6   0.69  0.01  0.14  0.04 0.50 0.50 1.1
## c7   0.58  0.33  0.18  0.40 0.64 0.36 2.7
## c8   0.12  0.10 -0.04  0.66 0.46 0.54 1.1
## c9   0.29  0.65  0.05  0.11 0.52 0.48 1.5
## c10  0.51  0.48  0.32  0.17 0.62 0.38 2.9
## c11  0.52  0.51  0.16  0.06 0.56 0.44 2.2
## c12  0.59 -0.01  0.25 -0.25 0.47 0.53 1.7
## c13  0.08 -0.24  0.56  0.53 0.65 0.35 2.4
## c14  0.12  0.12  0.76  0.06 0.61 0.39 1.1
## c15  0.55 -0.14  0.33  0.24 0.48 0.52 2.2
## c16  0.16  0.21  0.64  0.02 0.48 0.52 1.3
## c17  0.33  0.23  0.62 -0.14 0.57 0.43 2.0
## c18  0.54  0.00  0.24  0.11 0.36 0.64 1.5
## c19 -0.12  0.77  0.15  0.00 0.63 0.37 1.1
## c20  0.28  0.53  0.12  0.51 0.63 0.37 2.6
##
##                        RC1  RC3  RC2  RC4
## SS loadings           5.00 2.56 2.21 1.77
## Proportion Var        0.25 0.13 0.11 0.09
## Cumulative Var        0.25 0.38 0.49 0.58
## Proportion Explained  0.43 0.22 0.19 0.15
## Cumulative Proportion 0.43 0.65 0.85 1.00
##
## Mean item complexity =  1.8
## Test of the hypothesis that 4 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  341.54  with prob <  4.5e-24
##
## Fit based upon off diagonal values = 0.96
```

```
plot(female.factor, title="Female Varimax")
```



Black dots on the graph tend to have a high positive correlation within RC1 and a negative correlation within RC2.

Blue dots tend to have a high positive correlation with RC3 and a high negative correlation in RC2.

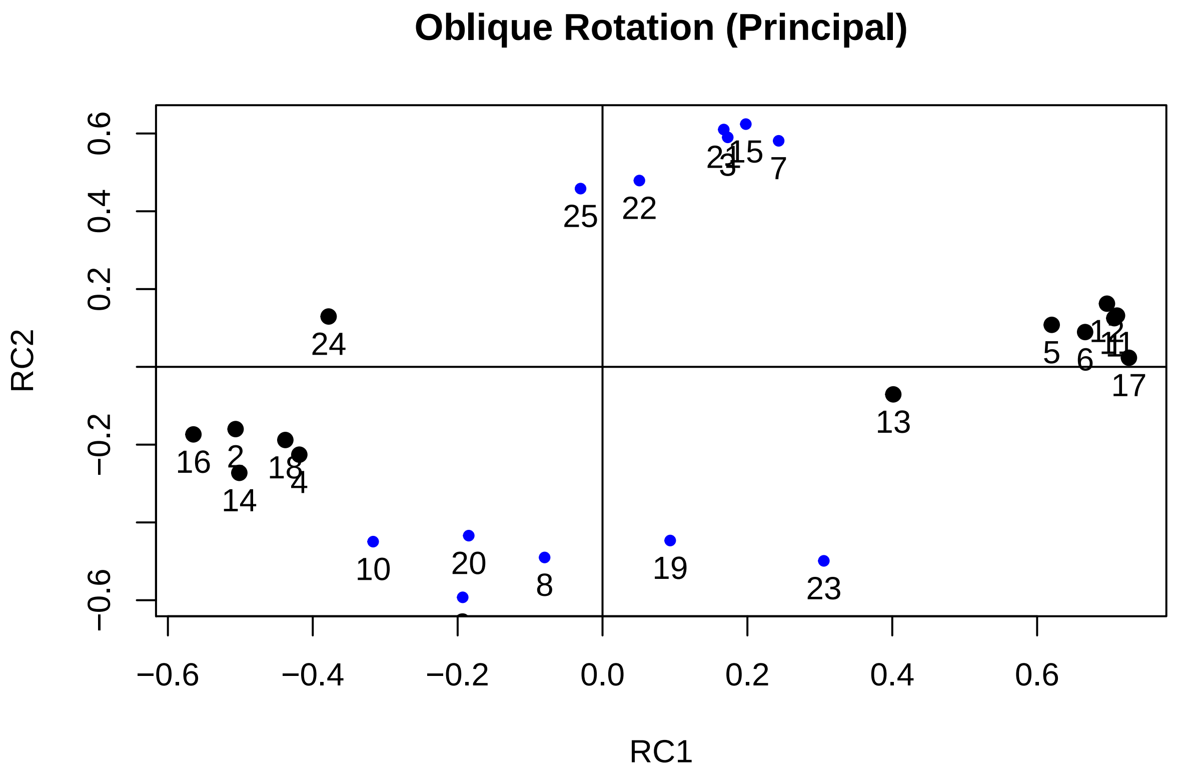
Red dots seem to only be correlated with RC2 and the correlation is high positive.

Gray squares tend to have high correlation within RC4.

There are less Gray points in the graph for females. In addition it seems that blue and red dots in RC2 and RC3 have switched places for females compared to males.

3. (PMA6 15.8) Perform a factor analysis on all of the items of the Parental Bonding scale for the Parental HIV data set. Retain two factors. Rotate the factors using an orthogonal rotation. Do the items with the highest loadings for each of the factors correspond to the items of the overprotection and care scale? Interpret the findings.

```r
parent.cols <- scale(HIV[c(34:58)])
parent.factor <- principal(parent.cols, nfactors = 2, rotate = "varimax")
plot(parent.factor, title="Oblique Rotation (Principal)")
```

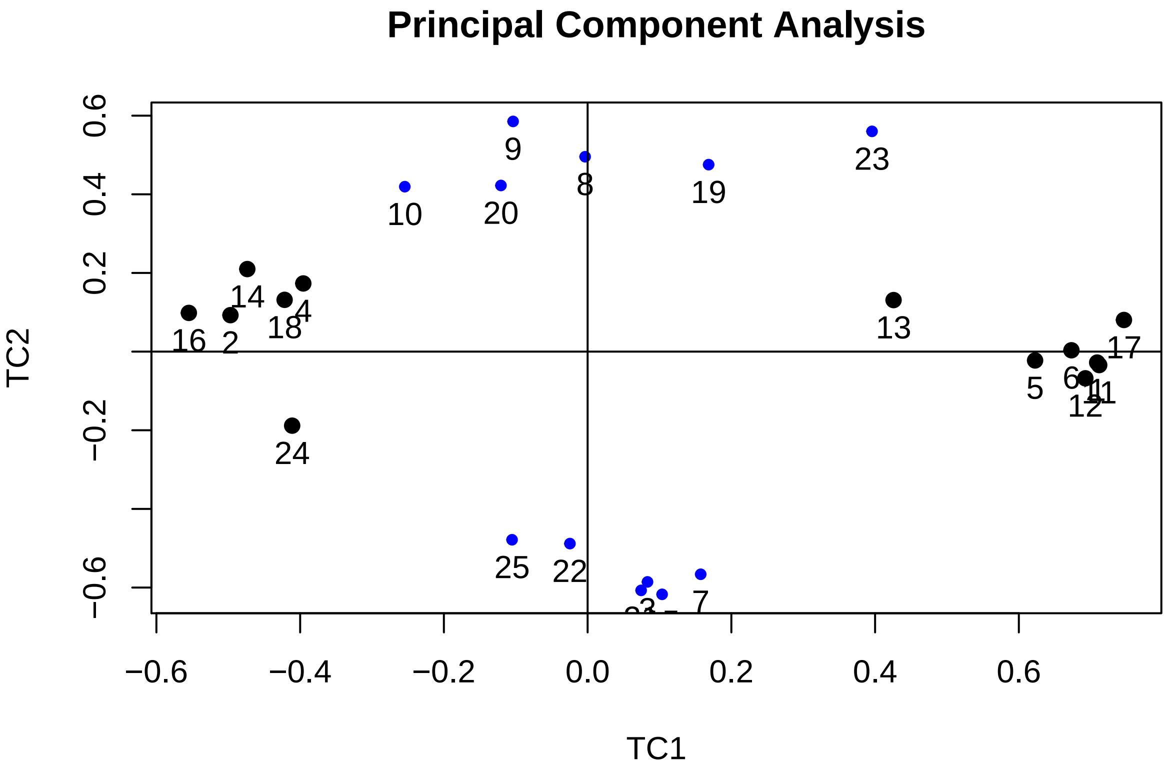


The black points on the graph tend to be correlated positively or negatively with RC1. The black points tend to correspond with the care subscale of the parental bonding variables.

The blue points on the graph tend to be correlated positively or negatively with RC2. The blue points tend to correspond with the overprotection subscale of the parental bonding variables.

4. (PMA6 15.9) Repeat problem #3 (15.8) using an oblique rotation. DO the substantive conclusions change?

```r
parent.factor.ob <- principal(parent.cols, nfactors = 2, rotate = "quartimin")
plot(parent.factor.ob)
```



The values seem pretty similar when observing both the rotations together.The conclusions do not change when using an oblique rotation.