

# Assignment 2 STAT 315-463: Multivariable Statistical Methods and Applications

Lisa Lu 31088272

**Due date: Friday 24 March 2023**

- Your assignment needs to show the R code you used, and your well discussed answers to the questions.
- Submit your assignments on Learn.

## Background

In the dataset, `USJudgeRatings.csv`, you are presented with ratings of State Judges on the Superior Court on 12 variables provided by 43 Lawyers in 1977.

CONT	Number of contacts of lawyer with judge	INTG	Judicial integrity	DMNR	Demeanour
DILG	Diligence	CFMG	Case flow managing	DECI	Prompt decisions
PREP	Preparation for trial	FAMI	Familiarity with law	ORAL	Sound oral rulings
WRIT	Sound written rulings	PHYS	Physical ability	RTEN	Worthy of retention

```
# Read in the data
dataset <- read.csv("USJudgeRatings.csv")
```

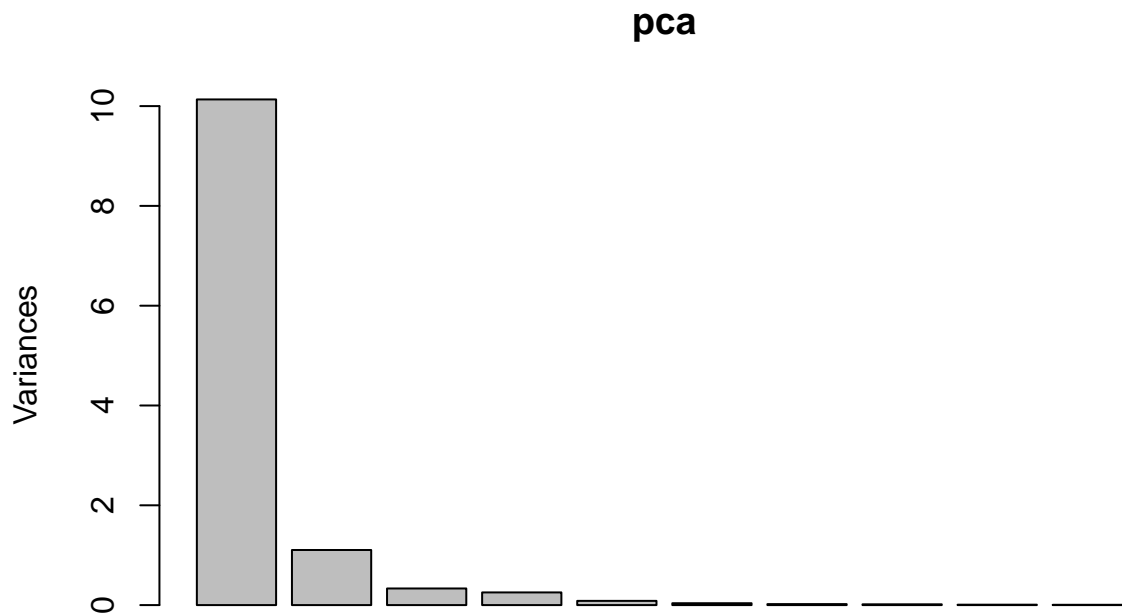
## Principal Component Analysis of the Rating Data.

Perform a PCA on the standardised ratings. Note, you will need to standardise the ratings yourself. Then answer the following questions.

```
# Standardise the ratings
dataset_std <- data.frame(scale(dataset[, -1]))
```

1. How many principal components do you believe should be retained. Justify your answer by looking at the variation in the data explained by each component.

```
# Perform Principal Component Analysis on the standardised dataset
pca <- prcomp(dataset_std)
# Use the scree plot
screeplot(pca)
```



From the scree plot, we can clearly see that the first two principal components can capture most of the variation from the data. Third and fourth components still capture a few variation but not as much. From the fifth onward, the values are very close to 0. Therefore, I think no more than the first four principal components should be retained.

```
# Get the percentage of variance explained by each component
summary(pca)
```

```
## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation  3.1833  1.05078  0.57698  0.50383  0.29061  0.19310  0.14030
## Proportion of Variance 0.8445  0.09201  0.02774  0.02115  0.00704  0.00311  0.00164
## Cumulative Proportion 0.8445  0.93647  0.96421  0.98537  0.99240  0.99551  0.99715
##              PC8      PC9      PC10     PC11     PC12
## Standard deviation  0.12416  0.08851  0.07491  0.05708  0.04539
## Proportion of Variance 0.00128  0.00065  0.00047  0.00027  0.00017
## Cumulative Proportion 0.99844  0.99909  0.99956  0.99983  1.00000
```

It is shown that the first principal component (PC1) explains 84.4% of variation, with PC2 accounts for 9.2%, PC3 for 2.8%, and PC4 for 2.1%. The first 3 PCs explain 96.4% of the variation and the first 4 PCs explain 98.5% of the variation. So from this result, I think even just take the first 3 PCs should be enough to cover a good percentage of variation.

2. In your own words, describe what you believe the first principal component is measuring.

```
pca$rotation[, 1:2]
```

```
##           PC1           PC2
## CONT  0.003075143 -0.932890644
## INTG -0.288550775  0.182040993
## DMNR -0.286884206  0.197565743
## DILG -0.304354091 -0.036304667
## CFMG -0.302572733 -0.168393523
## DECI -0.301891969 -0.127877299
## PREP -0.309406446 -0.032230248
## FAMI -0.306679527  0.001315183
## ORAL -0.312708348  0.003625720
## WRIT -0.311061231  0.031378756
## PHYS -0.280723624 -0.089037698
## RTEN -0.309790218  0.039381306
```

Rotations, or loadings, specify the weight that each variable contribute to the component. In the result above, we can see that all the variables, except CONT, have negative values. Among those, the component has largest negative associations with ORAL, WRIT. It seems like the first PC is measuring the judges' work skills and qualities.

### 3. What do you think the second principal component represents?

The second principal component is the linear combination of variables that covers as much of the remaining variation as possible. From the result above, we can see that it has a large negative association with CONT. Therefore, this component is primarily measuring the judges' interpersonal communication with the lawyers. This component has 0 correlation with the first PC.

### 4. You are told *Judicial Integrity* and *Demeanour* are particularly important traits, and should be given 5 times the weight of the other variables. Re-run the Principal Component Analysis such that Integrity and Demeanour is given 5 times the weight of all other variables. What impact does this have?

```
# Apply weights to INTG and DMNR
dataset_weighed <- dataset_std
dataset_weighed$INTG <- dataset_std$INTG * 5
dataset_weighed$DMNR <- dataset_std$DMNR * 5

# Run the principal component analysis on the weighed dataset
pca1 <- prcomp(dataset_weighed)
summary(pca1)
```

```
## Importance of components:
##           PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation  7.4897 1.36069 0.98113 0.88014 0.41322 0.28404 0.16642
## Proportion of Variance 0.9349 0.03086 0.01604 0.01291 0.00285 0.00134 0.00046
## Cumulative Proportion 0.9349 0.96579 0.98183 0.99474 0.99759 0.99893 0.99940
##           PC8      PC9      PC10     PC11     PC12
## Standard deviation  0.12790 0.09368 0.07540 0.05781 0.04546
## Proportion of Variance 0.00027 0.00015 0.00009 0.00006 0.00003
## Cumulative Proportion 0.99967 0.99982 0.99991 0.99997 1.00000
```

```
pca1$rotation[, 1:3]
```

```
##           PC1           PC2           PC3
## CONT  0.01683514  0.4550909  0.08123501
## INTG -0.66064656 -0.1535575  0.69375788
## DMNR -0.66007098 -0.2704810 -0.63435275
## DILG -0.11860621  0.2787158  0.08527462
## CFMG -0.11383471  0.3580957 -0.06369157
## DECI -0.11263597  0.3564752 -0.07551669
## PREP -0.12040199  0.2809923  0.03303414
## FAMI -0.11888533  0.2746566  0.04884273
## ORAL -0.12545426  0.2256858 -0.03326969
## WRIT -0.12442206  0.2228469  0.01365766
## PHYS -0.10681562  0.2886185 -0.28780330
## RTEN -0.12882861  0.1577163 -0.07117101
```

From the results above, we can see that with the increasing of the weights to variables, it increases the associations between the components and the variables and the associations between the components with other variables decrease.

## Factor Analysis for the Rating Data

Perform Factor Analysis on the standardised Ratings Data.

1. What happens when you try to fit a 3 and a 4 factor solution with no rotation. Hint: For the three factor solution, you may need to add `control=list(nstart=100)` as an additional argument in the `factanal` function.

```
factor3 <- factanal(dataset_std, factors = 3, scores = "regression",
                    rotation = "none", control=list(nstart=100))
```

```
factor3
```

```
##
## Call:
## factanal(x = dataset_std, factors = 3, scores = "regression",      rotation = "none", control = list(
##
## Uniquenesses:
##  CONT  INTG  DMNR  DILG  CFMG  DECI  PREP  FAMI  ORAL  WRIT  PHYS  RTEN
## 0.709 0.052 0.020 0.050 0.009 0.027 0.011 0.005 0.006 0.005 0.189 0.016
##
## Loadings:
##      Factor1 Factor2 Factor3
## CONT      0.418  0.342
## INTG  0.905 -0.359
## DMNR  0.894 -0.409  0.119
## DILG  0.969
## CFMG  0.961  0.171  0.195
## DECI  0.961  0.175  0.139
## PREP  0.991
## FAMI  0.988      -0.126
## ORAL  0.996
```

```
## WRIT 0.996
## PHYS 0.877      0.203
## RTEN 0.975 -0.155
##
##          Factor1 Factor2 Factor3
## SS loadings    10.067    0.574    0.262
## Proportion Var   0.839    0.048    0.022
## Cumulative Var   0.839    0.887    0.909
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 112.37 on 33 degrees of freedom.
## The p-value is 1.38e-10
```

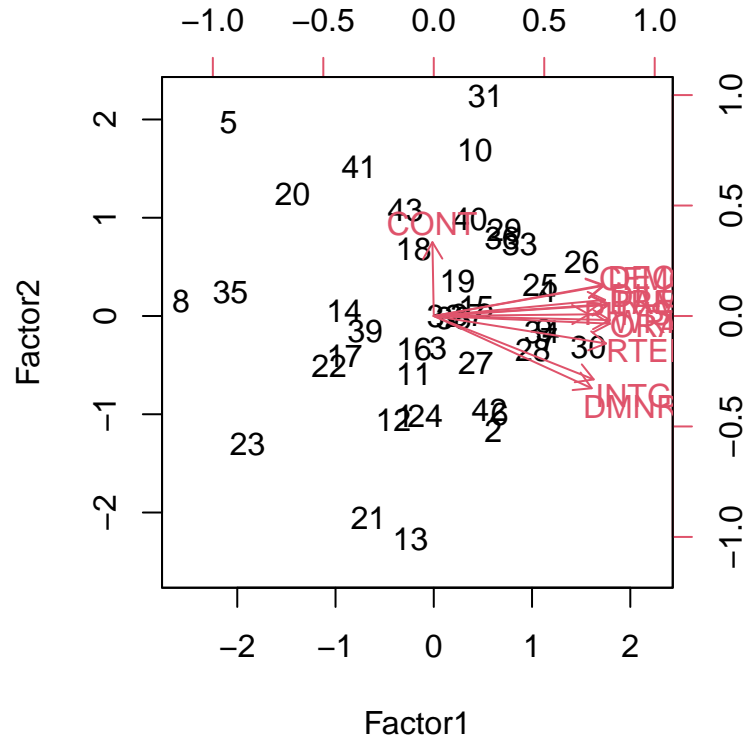
```
factor4 <- factanal(dataset_std, factors = 4, scores = "regression",
                    rotation = "none")
factor4
```

```
##
## Call:
## factanal(x = dataset_std, factors = 4, scores = "regression",      rotation = "none")
##
## Uniquenesses:
##  CONT  INTG  DMNR  DILG  CFMG  DECI  PREP  FAMI  ORAL  WRIT  PHYS  RTEN
## 0.749 0.013 0.032 0.023 0.014 0.025 0.008 0.005 0.005 0.005 0.047 0.006
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4
## CONT          0.316   0.343   0.183
## INTG  0.916  -0.356  -0.101
## DMNR  0.904  -0.385
## DILG  0.968          0.177
## CFMG  0.956   0.160   0.177   0.123
## DECI  0.956   0.179   0.148
## PREP  0.989   0.102
## FAMI  0.985   0.112  -0.101
## ORAL  0.997
## WRIT  0.995
## PHYS  0.879          0.361  -0.221
## RTEN  0.981  -0.144   0.102
##
##          Factor1 Factor2 Factor3 Factor4
## SS loadings    10.091    0.487    0.338    0.154
## Proportion Var   0.841    0.041    0.028    0.013
## Cumulative Var   0.841    0.881    0.910    0.922
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 49.78 on 24 degrees of freedom.
## The p-value is 0.00151
```

From the results of loadings above, it seems like there is not much difference between a 3 and a 4 factor without a rotation.

2. Which variables are grouped by the first two factors? (e.g. threshold  $|\text{loading}| \geq 0.25$ )

```
biplot(factor3$scores, factor3$loadings)
```



From the plot, we can see that INTG and DMNR are grouped together, CONT is in its own group, and all the rest of the variables seems to be in one group.

```
factor3$correlation
```

```
##          CONT          INTG          DMNR          DILG          CFMG          DECI
## CONT  1.00000000 -0.1331909 -0.1536885  0.0123920  0.1369123  0.08653823
## INTG -0.13319089  1.0000000  0.9646153  0.8715111  0.8140858  0.80284636
## DMNR -0.15368853  0.9646153  1.0000000  0.8368510  0.8133582  0.80411683
## DILG  0.01239200  0.8715111  0.8368510  1.0000000  0.9587988  0.95616608
## CFMG  0.13691230  0.8140858  0.8133582  0.9587988  1.0000000  0.98113590
## DECI  0.08653823  0.8028464  0.8041168  0.9561661  0.9811359  1.00000000
## PREP  0.01146921  0.8777965  0.8558175  0.9785684  0.9579140  0.95708831
## FAMI -0.02563656  0.8688580  0.8412415  0.9573634  0.9354684  0.94280452
## ORAL -0.01199681  0.9113992  0.9067729  0.9544758  0.9505657  0.94825640
## WRIT -0.04381025  0.9088347  0.8930611  0.9592503  0.9422470  0.94610093
## PHYS  0.05424827  0.7419360  0.7886804  0.8129211  0.8794874  0.87176277
## RTEN -0.03364343  0.9372632  0.9437002  0.9299652  0.9270827  0.92499241
##          PREP          FAMI          ORAL          WRIT          PHYS          RTEN
## CONT  0.01146921 -0.02563656 -0.01199681 -0.04381025  0.05424827 -0.03364343
## INTG  0.87779650  0.86885798  0.91139915  0.90883469  0.74193597  0.93726315
## DMNR  0.85581749  0.84124150  0.90677295  0.89306109  0.78868038  0.94370017
## DILG  0.97856839  0.95736345  0.95447583  0.95925032  0.81292115  0.92996523
## CFMG  0.95791402  0.93546838  0.95056567  0.94224697  0.87948744  0.92708271
```

```
## DECI 0.95708831 0.94280452 0.94825640 0.94610093 0.87176277 0.92499241
## PREP 1.00000000 0.98986345 0.98310045 0.98679918 0.84867350 0.95029259
## FAMI 0.98986345 1.00000000 0.98133905 0.99069557 0.84374436 0.94164495
## ORAL 0.98310045 0.98133905 1.00000000 0.99342943 0.89116392 0.98213227
## WRIT 0.98679918 0.99069557 0.99342943 1.00000000 0.85594002 0.96755639
## PHYS 0.84867350 0.84374436 0.89116392 0.85594002 1.00000000 0.90654782
## RTEN 0.95029259 0.94164495 0.98213227 0.96755639 0.90654782 1.00000000
```

From the correlation matrix, we can also see that CONT does not have much correlation with other variables. INTG has high correlation with DMNR (0.96). All the other variables have very high correlations with one another.

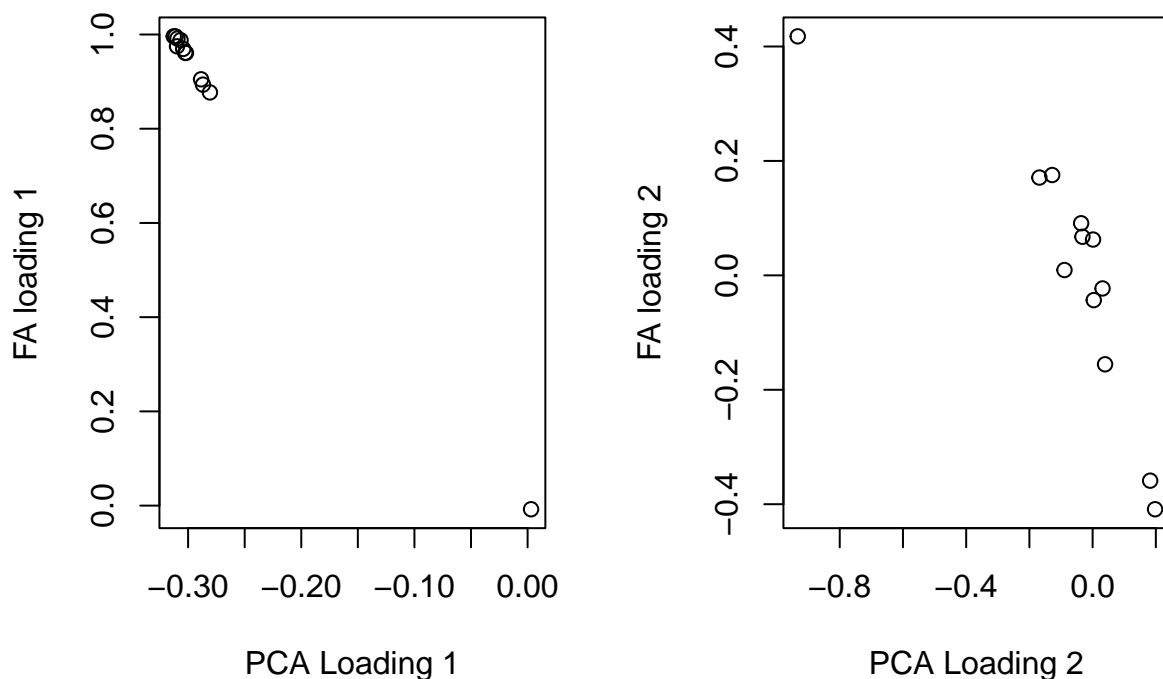
3. Compare the factor loadings for the first two factors to the first two principal components of the standardised data found in the previous section. Comment on any similarities and/or differences.

```
# Comparison of factor loadings from FA and PCA
cbind(factor3$loadings[,1:2], pca$rotation[, 1:2])
```

```
##      Factor1      Factor2      PC1      PC2
## CONT -0.0077011 0.417744860 0.003075143 -0.932890644
## INTG 0.9049222 -0.358851069 -0.288550775 0.182040993
## DMNR 0.8936034 -0.408789153 -0.286884206 0.197565743
## DILG 0.9693851 0.091095816 -0.304354091 -0.036304667
## CFMG 0.9613289 0.170944911 -0.302572733 -0.168393523
## DECI 0.9608196 0.175383924 -0.301891969 -0.127877299
## PREP 0.9914602 0.067491817 -0.309406446 -0.032230248
## FAMI 0.9880282 0.062578708 -0.306679527 0.001315183
## ORAL 0.9960809 -0.043275311 -0.312708348 0.003625720
## WRIT 0.9958562 -0.022884890 -0.311061231 0.031378756
## PHYS 0.8770771 0.009314406 -0.280723624 -0.089037698
## RTEN 0.9749411 -0.155453160 -0.309790218 0.039381306
```

CONT always have the opposite association in the first factor and the first PC and the second factor and second PC have more association with this variable.

```
#Plots of FA vs PCA Loadings
par(mfrow=c(1,2))
plot(factor3$loadings[,1]~pca$rotation[,1],xlab='PCA Loading 1',ylab='FA loading 1')
plot(factor3$loadings[,2]~pca$rotation[,2],xlab='PCA Loading 2',ylab='FA loading 2')
```



4. Comment on the observed variable specific variances (the uniquenesses). Do you believe all observed variables are explained by the factors discovered.

```
factor3$uniquenesses
```

```
##          CONT          INTG          DMNR          DILG          CFMG          DECI
## 0.708710765 0.052024940 0.020204850 0.050204514 0.008681842 0.026765112
##          PREP          FAMI          ORAL          WRIT          PHYS          RTEN
## 0.010896197 0.005000000 0.005934581 0.005000000 0.189451999 0.016146962
```

CONT has a really high uniqueness, which means this variable does not fit neatly into the factors, whereas CFMG, FAMI, ORAL, WRIT have very low uniqueness values.

5. Re-fit the 3 factor solution with a varimax rotation. How does this change the interpretation of the factors? In this case, do you find the rotated or non-rotated solution easier to interpret. Explain why or why not?

```
factor3_rotate <- factanal(dataset_std, factors = 3, rotation = "varimax",
                           scores = "regression", control=list(nstart=100))
factor3_rotate
```

```
##
## Call:
```

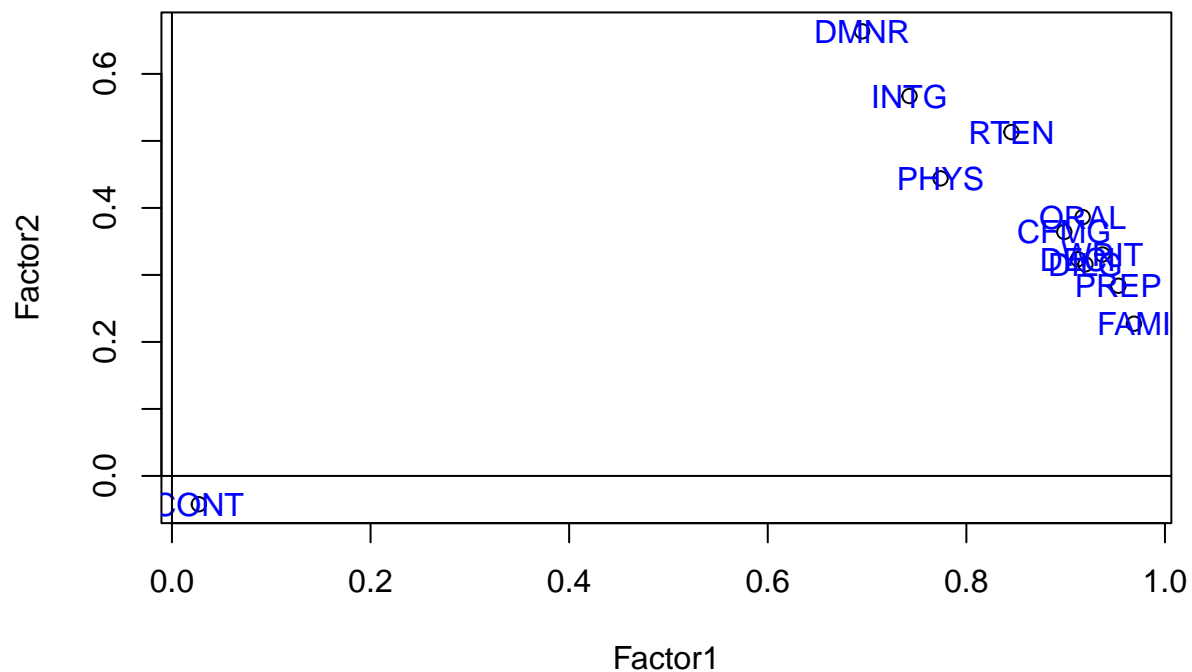


```

## factanal(x = dataset_std, factors = 3, scores = "regression", rotation = "varimax", control = li
##
## Uniquenesses:
##  CONT  INTG  DMNR  DILG  CFMG  DECI  PREP  FAMI  ORAL  WRIT  PHYS  RTEN
## 0.709 0.052 0.020 0.050 0.009 0.027 0.011 0.005 0.006 0.005 0.189 0.016
##
## Loadings:
##      Factor1 Factor2 Factor3
## CONT          0.537
## INTG  0.743   0.567 -0.273
## DMNR  0.695   0.664 -0.238
## DILG  0.920   0.316
## CFMG  0.899   0.364  0.226
## DECI  0.912   0.323  0.190
## PREP  0.953   0.284
## FAMI  0.969   0.227
## ORAL  0.917   0.386
## WRIT  0.937   0.330
## PHYS  0.774   0.444  0.120
## RTEN  0.845   0.513
##
##      Factor1 Factor2 Factor3
## SS loadings      8.402   1.952   0.550
## Proportion Var   0.700   0.163   0.046
## Cumulative Var   0.700   0.863   0.909
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 112.37 on 33 degrees of freedom.
## The p-value is 1.38e-10

plot(loadings(factor3_rotate))
text(factor3_rotate$loadings[,1],
      factor3_rotate$loadings[,2],
      colnames(dataset_std),
      col="blue")
abline(h = 0, v = 0)

```



From the plot, we can clearly see different clusters of the variables, i.e., CONT has its own cluster. ORAL, CFMG, PREP, FAMI, WRIT, DILG, and DECI should be grouped together, and the rest 4 variables can be grouped together. Therefore, in this case, with a varimax rotation, it is easier to interpret.