# HW15

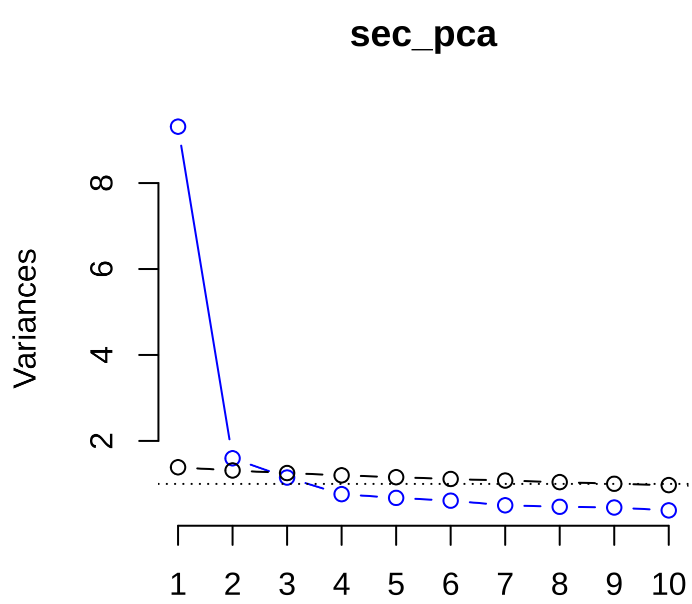
106022113

6/2/2021

## Question 1. Perform Parallel Analysis

### a. Show visualization of screeplot of data, noise and eigenvalue = 1 cutoff

```
sec <- read.csv("security_questions.csv")
sec_pca <- prcomp(sec,scale.=TRUE)
sim_noise_ev <- function(n,p){
  noise <- data.frame(replicate(p,rnorm(n)))
  return(eigen(cor(noise))$values)
}
set.seed(42)
evalues_noise <- replicate(100,sim_noise_ev(dim(sec)[1],dim(sec)[2]))
evalues_mean <- apply(evalues_noise,1,mean)
screeplot(sec_pca,type = "lines",col = "blue")
lines(evalues_mean, type = "b")
abline(h = 1, lty = "dotted")
```



As we can see PC1, 2. 3 are above the eigenvalue = 1 cutoff, which is approximately 67% of the total variance, and the simulated noise is closing near the cutoff line.

**b.How many dimensions would you retain if we used Parallel Analysis?**

**ANSWER:** PCA1, 2, acquire higher value than the random simulated noise. Hence, it is appropriate for us to chose these two.

## Question 2. Examine factor loadings

```
library(psych)
principal(sec, nfactor =3, rotate = "none", scores = TRUE)
```

```
## Principal Components Analysis
## Call: principal(r = sec, nfactors = 3, rotate = "none", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1   PC2   PC3   h2   u2 com
## Q1  0.82 -0.14  0.00 0.69 0.31 1.1
## Q2  0.67 -0.01  0.09 0.46 0.54 1.0
## Q3  0.77 -0.03  0.09 0.60 0.40 1.0
## Q4  0.62  0.64  0.11 0.81 0.19 2.1
## Q5  0.69 -0.03 -0.54 0.77 0.23 1.9
## Q6  0.68 -0.10  0.21 0.52 0.48 1.2
## Q7  0.66 -0.32  0.32 0.64 0.36 2.0
## Q8  0.79  0.04 -0.34 0.74 0.26 1.4
## Q9  0.72 -0.23  0.20 0.62 0.38 1.4
## Q10 0.69 -0.10 -0.53 0.76 0.24 1.9
## Q11 0.75 -0.26  0.17 0.66 0.34 1.4
## Q12 0.63  0.64  0.12 0.82 0.18 2.1
## Q13 0.71 -0.06  0.08 0.52 0.48 1.0
## Q14 0.81 -0.10  0.16 0.69 0.31 1.1
## Q15 0.70  0.01 -0.33 0.61 0.39 1.4
## Q16 0.76 -0.20  0.18 0.65 0.35 1.3
## Q17 0.62  0.66  0.11 0.83 0.17 2.0
## Q18 0.81 -0.11 -0.07 0.67 0.33 1.1
##
##                        PC1  PC2  PC3
## SS loadings           9.31 1.60 1.15
## Proportion Var        0.52 0.09 0.06
## Cumulative Var        0.52 0.61 0.67
## Proportion Explained  0.77 0.13 0.10
## Cumulative Proportion 0.77 0.90 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.05
##  with the empirical chi square  258.65  with prob <  1.4e-15
##
## Fit based upon off diagonal values = 0.99
```

**a.Looking at the first 3 principal components, which components does each item belong?**

**ANSWER:** Setting threshold to 70 %: PCA1: Q1, 3, 8, 9 ,11, 13, 14, 15, 16, 18 belongs here. PCA2: Q4, 12, 17 are close to the threshold, however similar to their score with PCA3: Q5, 10 are significantly higher than the other values, but not close to 0.7

**b.How much of the total variance of the security dataset do the first 3 PCs capture?**

**ANSWER:** According to the summary, the cumulated variance is 67%.

**c.Which items are less than adequately explained by the first 3 principal components?**

**ANSWER:** H2 is the communalities(cummulated variance), and Q2 only accumulated 46% percent, which is less than adequate to be explained.

**d. How many measurement items share similar loadings between 2 or more components?**

**ANSWER:** Three measurements. Q4, 7, 12 acquire high complexity of the component loadings for the variable.

**e. Can you distinguish a 'meaning' behind the first principal component from the items that load best upon it?**

**ANSWER:** It seems that these questions are more associated with the **security** problems of the site. Referring to the *protection*, *security*, *identity* keywords.

## Question 3. Let's rotate the our principal component axes to get rotated components

```r
principal(sec, nfactors = 3, rotate = "varimax", scores = TRUE)
```

```
## Principal Components Analysis
## Call: principal(r = sec, nfactors = 3, rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1  RC3  RC2   h2   u2 com
## Q1  0.66 0.45 0.22 0.69 0.31 2.0
## Q2  0.54 0.29 0.29 0.46 0.54 2.1
## Q3  0.62 0.34 0.31 0.60 0.40 2.1
## Q4  0.22 0.19 0.85 0.81 0.19 1.2
## Q5  0.24 0.83 0.16 0.77 0.23 1.3
## Q6  0.65 0.20 0.23 0.52 0.48 1.5
## Q7  0.79 0.10 0.06 0.64 0.36 1.0
## Q8  0.38 0.71 0.30 0.74 0.26 2.0
## Q9  0.74 0.23 0.14 0.62 0.38 1.3
## Q10 0.28 0.82 0.10 0.76 0.24 1.3
## Q11 0.76 0.28 0.12 0.66 0.34 1.3
## Q12 0.23 0.19 0.85 0.82 0.18 1.2
## Q13 0.59 0.32 0.26 0.52 0.48 1.9
## Q14 0.72 0.31 0.28 0.69 0.31 1.7
```

```
## Q15 0.34 0.66 0.24 0.61 0.39 1.8
## Q16 0.74 0.27 0.17 0.65 0.35 1.4
## Q17 0.21 0.19 0.87 0.83 0.17 1.2
## Q18 0.61 0.50 0.23 0.67 0.33 2.2
##
##                        RC1  RC3  RC2
## SS loadings           5.61 3.49 2.95
## Proportion Var        0.31 0.19 0.16
## Cumulative Var        0.31 0.51 0.67
## Proportion Explained  0.47 0.29 0.24
## Cumulative Proportion 0.47 0.76 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.05
##  with the empirical chi square  258.65  with prob <  1.4e-15
##
## Fit based upon off diagonal values = 0.99
```

**a. Individually, does each rotated component (RC) explain the same, or different, amount of variance than the corresponding principal components (PCs)?**

**ANSWER:** Individually, they explain the *different* amount of variances.

**b. Together, do the three rotated components explain the same, more, or less cumulative variance as the three principal components combined?**

**ANSWER:** Together, they explain the *same* cumulative variances.

**c. Do those items have more clearly differentiated loadings among rotated components?**

**ANSWER:** Yes, according to the summary, they are more differentiated.

**d. Can you now interpret the "meaning" of the 3 rotated components from the items that load best upon each of them?**

**ANSWER:** RC1: Q7, 9, 11, 14, 16 –> *Unauthorized* seems to be the topic of these questions. RC2: Q4, 12, 17 –> *Denial* and *Deleted* appeared in these questions, indicating some kind of protection RC3: Q5, 8, 10, 15 –> *Transaction* process is mentioned

**e. Change the component to 2**

```r
principal(sec, nfactors = 2, rotate = "varimax", scores = TRUE)
```

```
## Principal Components Analysis
## Call: principal(r = sec, nfactors = 2, rotate = "varimax", scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1  RC2   h2   u2 com
```

```
## Q1  0.78 0.27 0.69 0.31 1.2
## Q2  0.60 0.31 0.45 0.55 1.5
## Q3  0.69 0.34 0.59 0.41 1.5
## Q4  0.24 0.86 0.80 0.20 1.1
## Q5  0.62 0.31 0.48 0.52 1.5
## Q6  0.65 0.24 0.48 0.52 1.3
## Q7  0.73 0.04 0.53 0.47 1.0
## Q8  0.67 0.42 0.62 0.38 1.7
## Q9  0.75 0.15 0.58 0.42 1.1
## Q10 0.65 0.24 0.48 0.52 1.3
## Q11 0.79 0.13 0.64 0.36 1.1
## Q12 0.25 0.86 0.80 0.20 1.2
## Q13 0.65 0.29 0.51 0.49 1.4
## Q14 0.76 0.30 0.67 0.33 1.3
## Q15 0.61 0.35 0.50 0.50 1.6
## Q16 0.76 0.19 0.62 0.38 1.1
## Q17 0.22 0.88 0.82 0.18 1.1
## Q18 0.76 0.29 0.66 0.34 1.3
##
##                        RC1  RC2
## SS loadings           7.52 3.39
## Proportion Var        0.42 0.19
## Cumulative Var        0.42 0.61
## Proportion Explained  0.69 0.31
## Cumulative Proportion 0.69 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.06
##  with the empirical chi square  439.68  with prob <  1.3e-38
##
## Fit based upon off diagonal values = 0.99
```

Yes, RC1 will acquire more items upon it and become more significant.

**NOT GRADED : How many compoents should we extract to understand the dataset?**

I believe we should understand three components. If we extract more, the meanings for each PC will decrease and will be harder for us interpret the meanings, while two PCs are a bit few since the questions are diverse and not able to be forced into 2 classes.