# HW15

106022103

2021/6/5

**import libary**

```
library(openxlsx) # read.xlsx()
library(ggplot2)
library(psych) # principal()
library(factoextra) # fviz_pca_biplot()
```

## Q1 parallel analysis

**Read File**

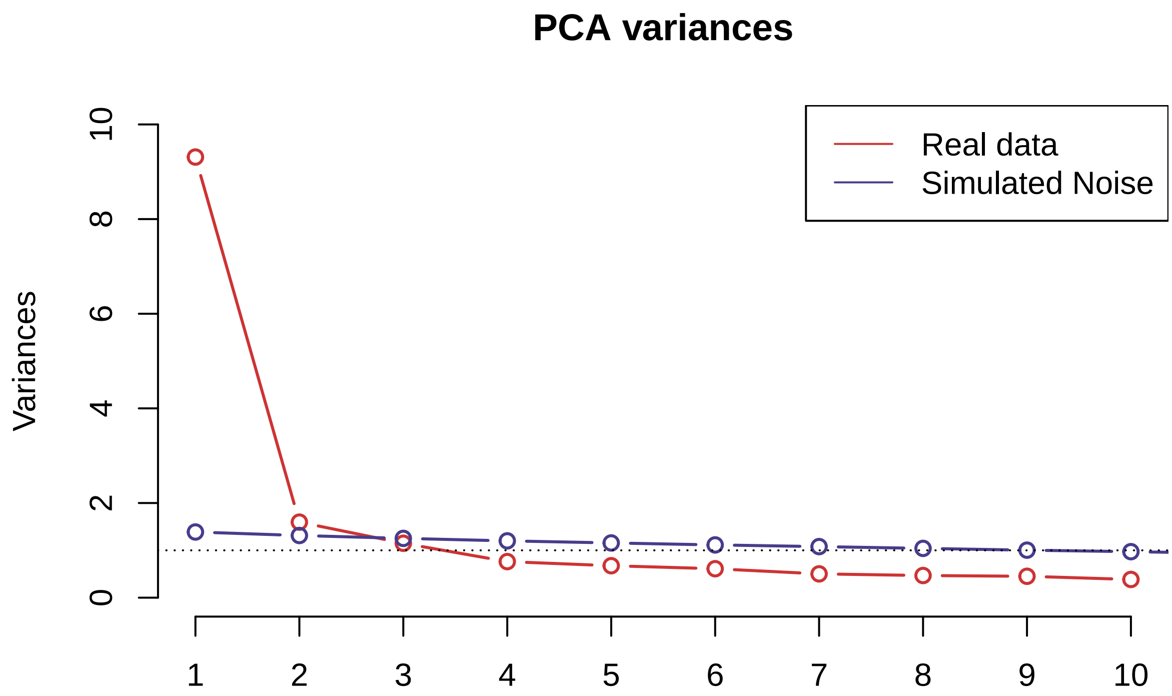
```
data <- read.xlsx(xlsxFile="data/security_questions.xlsx", sheet = 2, colNames = TRUE)
```

**(a) Show a single visualization with scree plot of data, scree plot of simulated noise, and a horizontal line showing the eigenvalue = 1 cutoff.**

```
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(data)[1], dim(data)[2]))
```

```
# draw
evalues_mean <- apply(evalues_noise, 1, mean)
pca <- prcomp(data, scale. = TRUE)
screeplot(pca, type="lines",col="brown3",main = "PCA variances",lwd=1.5,ylim = c(0, 10))
lines(evalues_mean,col="slateblue4" , type="b",lwd=1.5)
abline(h=1, lty="dotted")
legend("topright", c("Real data","Simulated Noise"),lty=c(1,1),col=c("brown3","slateblue4"))
```




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

```
eigenvalues <- eigen(cor(data))$values
sprintf("We should retain %d dimensions ", length(eigenvalues[eigenvalues>1]))
```

```
## [1] "We should retain 3 dimensions "
```

## Q2 Examine factor loadings

```
dec_pca3_orig <- principal(data,nfactors = 3,rotate="none",scores = TRUE)
dec_pca3_orig
```

```
## Principal Components Analysis
## Call: principal(r = data, 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 loadings of the first 3 principal components, to which components does each item seem to best belong?**

```
dec_pca3_orig[["loadings"]]
```

```
##
## Loadings:
##     PC1    PC2    PC3
## Q1   0.817 -0.139
## Q2   0.673
## Q3   0.766
## Q4   0.623  0.643  0.108
## Q5   0.690        -0.542
## Q6   0.683 -0.105  0.207
## Q7   0.657 -0.318  0.324
## Q8   0.786        -0.343
## Q9   0.723 -0.232  0.204
## Q10  0.686        -0.533
## Q11  0.753 -0.261  0.173
## Q12  0.630  0.638  0.122
## Q13  0.712
## Q14  0.811         0.157
## Q15  0.704        -0.333
## Q16  0.758 -0.203  0.183
## Q17  0.618  0.664  0.110
## Q18  0.807 -0.114
##
##                  PC1   PC2   PC3
## SS loadings    9.311 1.596 1.150
## Proportion Var 0.517 0.089 0.064
```

```
## Cumulative Var 0.517 0.606 0.670
```

- It seems all components belongs to PC1.

- Take the threshold of loading to 0.5, `Q4`,`Q12`,`Q17`belongs to PC2.
- Take the threshold of loading to 0.5, `Q5`,`Q10`belongs to PC3.

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

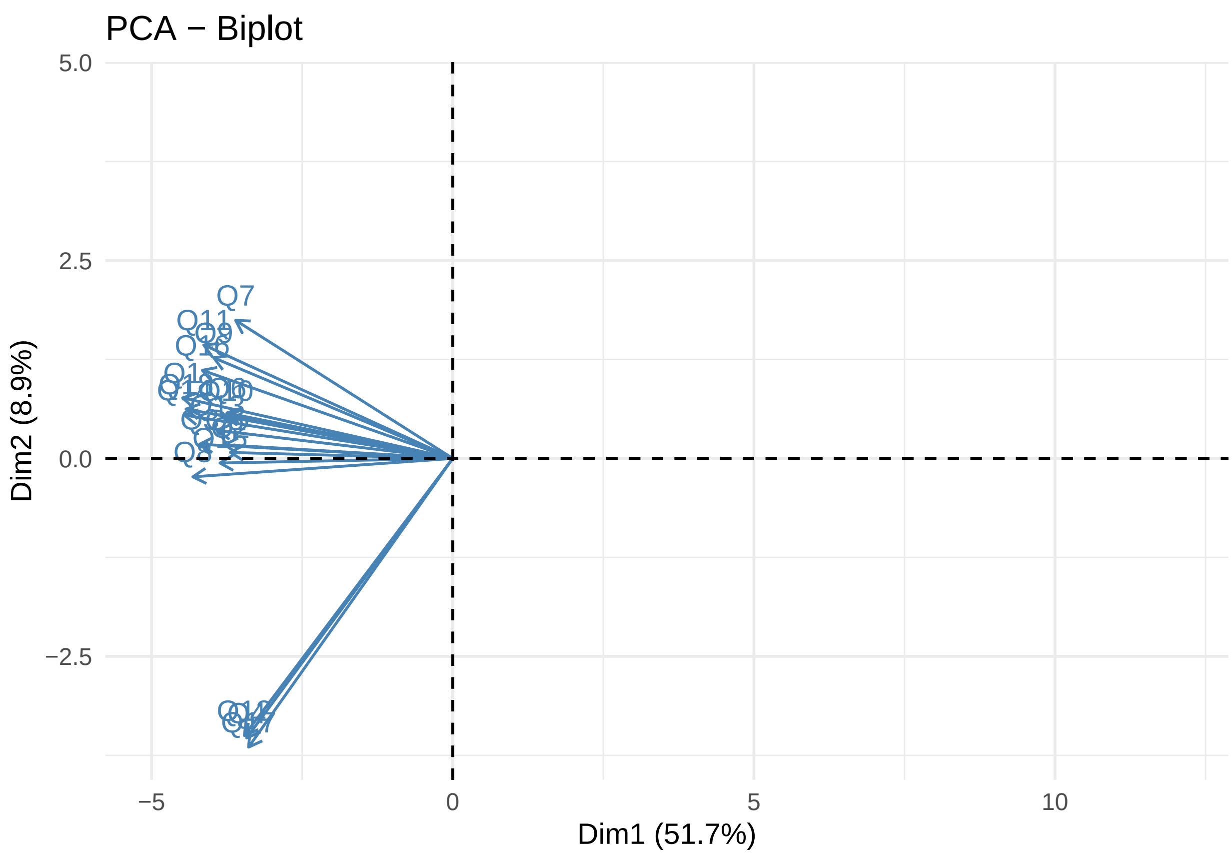
**ANSWER:** 67 of variance captured from the first 3 PCs.

**(c) Looking at commonality and uniqueness, which items are less than adequately explained by the first 3 principal components?**

**ANSWER:** According to the table, `Q2` is the least adequately explained component. The commonality of `Q2` is 1.035995 and uniqueness is 0.5394567

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

```
fviz_pca_biplot(pca, invisible ="ind")+
 theme_minimal()
```




**ANSWER:** `Q1`,`Q4`,`Q12` share similar loadings between 2 or more components.

**(e) Can you distinguish a 'meaning' behind the first principal component from the items that load best upon it? (see the wording of the questions of those items)**

**ANSWER:** Since the highest component of `PC1` is `Q1`,`Q14`, `Q18`, let's take a look at these question:

- `Q1`:I am convinced that this site respects the confidentiality of the transactions received from me.
- `Q14`:This site devotes time and effort to verify the accuracy of the information in transit.
- `Q18`:This site uses some security controls for the confidentiality of the transactions received from me.

I would give a conclusion about the users care about how website protect the security.

### Q3 rotate the our principal component axes

```
dec_pca3_rot <- principal(data,nfactors = 3,rotate="varimax",scores = TRUE)
dec_pca3_rot
```

```
## Principal Components Analysis
## Call: principal(r = data, 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:** The variance of RCs are **different** to original PCs.

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

**ANSWER:** The cumulative variance 3 RCs is same to 3 PCs, which is 67%.

**(c) Looking back at the items that shared similar loadings with multiple principal components (#2d), do those items have more clearly differentiated loadings among rotated components?**

**ANSWER:** According to the components of `RC1`, those items have more clearly differentiated loadings now.

**(d) Can you now interpret the "meaning" of the 3 rotated components from the items that load best upon each of them? (see the wording of the questions of those items)**

**ANSWER:** Since the highest component of `RC1` is `Q7`,`Q9`,`Q11`, `Q14`, `Q16`, let's take a look at these question:

- `Q7`:This site never sells my personal information in their computer databases to other companies
- `Q9`:I can remove my personal information from this site when I want to.
- `Q11`:This site devotes time and effort to preventing unauthorized access to my personal information.
- `Q14`:This site devotes time and effort to verify the accuracy of the information in transit.
- `Q16`:Databases that contain my personal information are protected from unauthorized access

I would give a conclusion about the users care about the **personal information** should be protected well.

**(e) If we reduced the number of extracted and rotated components to 2, does the meaning of our rotated components change?**

```r
dec_pca2_rot <- principal(data,nfactors = 2,rotate="varimax",scores = TRUE)
dec_pca2_rot
```

```
## Principal Components Analysis
## Call: principal(r = data, 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
```

**ANSWER:** Yes, the components in RC1 is acturally changed.

## (ungraded) Looking back at all our results and analyses of this dataset (from this week and previous), how many components (1-3) do you believe we should extract and analyze to understand the security dataset? Feel free to suggest different answers for different purposes.

**ANSWER:** According to the cumulative variance explained, I think 3 components is better to understand the security dataset.

## Reference Link

- Colors code in R
- screeplot: Draw a SCREE plot, showing the distribution of explained…
- fviz_pca: Quick Principal Component Analysis data visualization - R software and data mining