BACS HW14 - 109006234

Credit: 109006278

May 21st 2023

Loading the data

```
security <- read_excel(
        "G:/My Drive/111_2_BACS/HW14/security_questions.xlsx",
        sheet = "data")
```

Supporting Functions

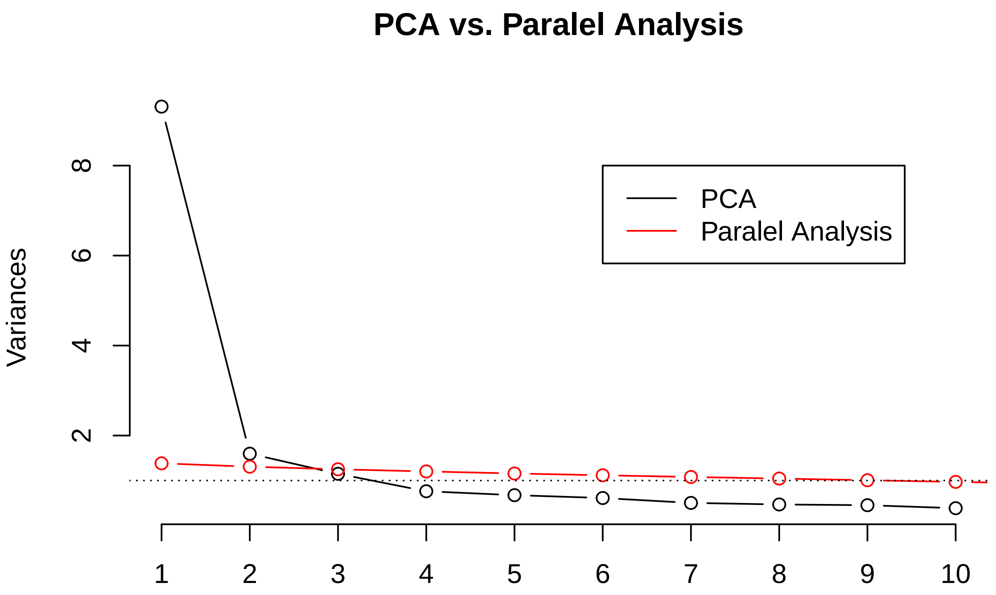
```
sim_noise_ev <- function(n, p){
  noise <- data.frame(replicate(p, rnorm(n)))
  eigen(cor(noise))$values
}

evaluate_loadings <- function(df, range) {
    return(
        (
           abs(df[range, 1] - df[range, 2])<0.1 |
           abs(df[range, 2] - df[range, 3])<0.1 |
           abs(df[range, 1] - df[range, 3])<0.1
        ) &
        (
           df[range, 1] < 0.7 &
           df[range, 2] < 0.7 &
           df[range, 3] < 0.7
        )
    )
}
```

Problem 1

(a) Show a single visualization with scree plot of data, scree plot of simulated noise

```
evalues_noise <- replicate(100, sim_noise_ev(405, 18)) # The same size
evalues_mean <- apply(evalues_noise, 1, mean)
sec_pca <- prcomp(security, scale. = TRUE)
```

```
screeplot(sec_pca, type="lines", main="PCA vs. Paralel Analysis")
lines(evalues_mean, type="b", col='red')
abline(h=1, lty="dotted")
legend(6,8,c("PCA","Paralel Analysis"),
        lty=c(1,1), col=c("black","red"))
```



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

As the red line intersect with the original line, I think I will retain about three dimensions.

Problem 2

(a) Looking at the loadings of the first 3 principal components, to which components does each item seem to best belong?

```
pca_three <- principal(security, nfactor=3, rotate="none", scores=TRUE)
loads <- pca_three$loadings
apply(loads[, 1:3], 1, function(x) which.max(x))
```

```
##  Q1  Q2  Q3  Q4  Q5  Q6  Q7  Q8  Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18
##   1   1   1   2   1   1   1   1   1   1   1   2   1   1   1   1   2   1
```

Q4, Q12, and Q17 seem to belong to PC2. And the others belong to PC1.

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

```r
var1 <- sum(pca_three$loadings[,"PC1"]^2)
var2 <- sum(pca_three$loadings[,"PC2"]^2)
var3 <- sum(pca_three$loadings[,"PC3"]^2)
total_var <- var1 + var2 + var3
```

```
## [1] "Total Variance:  12.0568434463861"
```

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

```r
which(pca_three$communality < 0.7)
```

```
##  Q1  Q2  Q3  Q6  Q7  Q9 Q11 Q13 Q14 Q15 Q16 Q18 
##   1   2   3   6   7   9  11  13  14  15  16  18
```

Q1, Q2, Q3, Q6, Q7, Q9, Q11, Q13, Q14, Q15, Q16, and Q18 is less than adequately explained by the first 3 principal components.

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

```r
shared_loads <- apply(pca_three$loadings, 1,
        function(x) sum(abs(x) > 0.5))
with_shared_loads <- sum(shared_loads >= 2)
```

(e) Can you interpret a 'meaning' behind the first principal component from the items that load best upon it?

```r
first_pc_loads <- abs(loads[, 1])
sort(first_pc_loads[first_pc_loads > 0.7], decreasing = TRUE)
```

```
##        Q1       Q14       Q18        Q8        Q3       Q16       Q11        Q9 
## 0.8169846 0.8114677 0.8067284 0.7861054 0.7655215 0.7575616 0.7529735 0.7230295 
##       Q13       Q15 
## 0.7119085 0.7040428
```

The meaning behind the first principal component might be related to the personal information security.

Problem 3

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

```
pca_three_rotate <- principal(security, nfactor=3, rotate="varimax", scores=TRUE)
loads_rot <- pca_three_rotate$loadings
```

From this result compared to the Question 2a, all are different.

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

```
var1_rot <- sum(pca_three_rotate$loadings[,"RC1"]^2)
var2_rot <- sum(pca_three_rotate$loadings[,"RC2"]^2)
var3_rot <- sum(pca_three_rotate$loadings[,"RC3"]^2)
total_var_rot <- var1_rot + var2_rot + var3_rot
```

```
## [1] "Total Variance Rotation:  12.0568434463861"
```

Three rotated components explain the same cumulative variance as the three principal components combined.

(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?

```
loads[c(4, 5, 10, 12,17), 1:3]
```

```
##           PC1         PC2        PC3
## Q4  0.6233733  0.64307826  0.1080319
## Q5  0.6900841 -0.03126466 -0.5423546
## Q10 0.6861529 -0.09868038 -0.5326787
## Q12 0.6303505  0.63753124  0.1215228
## Q17 0.6175336  0.66426051  0.1100612
```

```
loads_rot[c(4, 5, 10, 12,17), 1:3]
```

```
##           RC1       RC3       RC2
## Q4  0.2182880 0.1933627 0.8536838
## Q5  0.2441735 0.8279850 0.1617475
## Q10 0.2768895 0.8229206 0.1020988
## Q12 0.2327616 0.1861745 0.8542346
## Q17 0.2054021 0.1869028 0.8703910
```

After the rotation, the items are slightly different to the previous loading.

(d) Can you now more easily interpret the "meaning" of the 3 rotated components from the items that load best upon each of them?

```
loads_rot[loads_rot[, 1] > 0.7, 1]
```

```
##        Q7        Q9       Q11       Q14       Q16
## 0.7895344 0.7378148 0.7573493 0.7187578 0.7396241
```

From this, we can conclude that RC1 might represent data protection

```
loads_rot[loads_rot[, 2] > 0.7, 2]
```

```
##        Q5        Q8       Q10
## 0.8279850 0.7062018 0.8229206
```

From this, we can conclude that RC2 might represent transaction processing.

```
loads_rot[loads_rot[, 3] > 0.7, 1]
```

```
##        Q4       Q12       Q17
## 0.2182880 0.2327616 0.2054021
```

From this, we can conclude that RC3 might represent the evidences provided to protect against denial. ## (e) If we reduced the number of extracted and rotated components to 2, does the meaning of our rotated components change?

```
security_pca_rot2 <- principal(security, nfactor=2, rotate="varimax", scores=TRUE)
security_pca_rot2
```

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

Reducing the number of both components to only two components might change the meaning of the rotated components.

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.

We should extract and analyze more than one component to be able to accurately analyze and interpret the data.