HW14

110077443

5/17/2022

Please note that all code in this document is presented in a grey box and the output reflected below each box

- The below code allows lengthy lines of comments to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

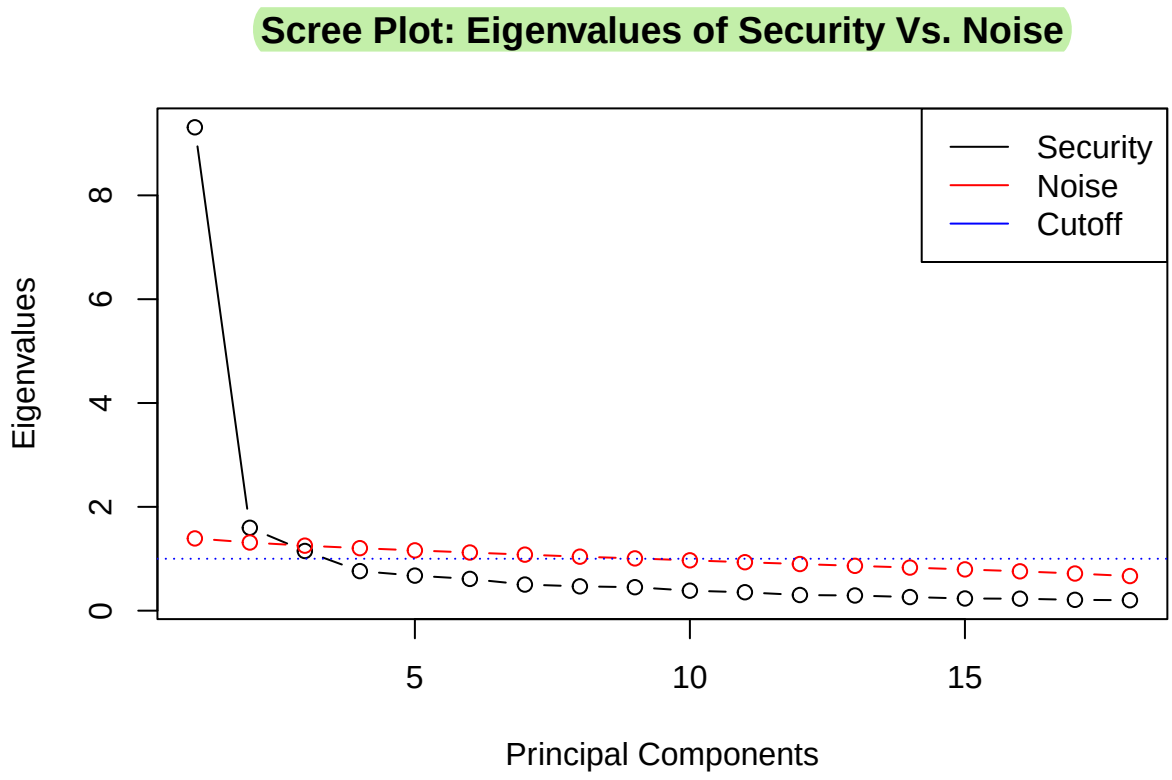
1) Let's perform a parallel analysis from security survey dataset

```
# Importing data
require(readxl)  # Used for read_excel function to import 'xls' and 'xlsx' files
sec <- read_excel("security_questions.xlsx", sheet = "data")
```

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

```
# To extract Eigenvalues from security data frame
eigen_sec <- eigen(cor(sec))

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

# Create noise data frame
set.seed(2022)  # For replication
# Repeat this 100 times
evalues_noise <- replicate(100, sim_noise_ev(405, 18))

# Average each of the noise eigenvalues
evalues_mean <- apply(evalues_noise, 1, mean)

# Creating Scree Plot with an eiganvalue = 1 threshold
plot(eigen_sec$values, type = "b", xlab = "Principal Components",
    ylab = "Eigenvalues", main = "Scree Plot: Eigenvalues of Security Vs. Noise")
```

```
lines(evalues_mean, type = "b", col = "red")
abline(h = 1, col = "blue", lty = "dotted")  # eigenvalue = 1 cut off

# Adding legend
legend("topright", lty = 1, c("Security", "Noise", "Cutoff"),
    col = c("black", "red", "blue"))
```



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

ANSWER ##

- Based on the **Parallel Analysis**, we would retain **two dimensions** because only two PCs have higher eigenvalues than the average "noise".

2) Let's treat the underlying dimensions of the security dataset as factors and examine factor loadings using the principal() method from the psych package

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

```
require(psych)  # Used for principal function
sec_fac <- principal(r = sec, nfactors = 3, rotate = "none",
    scores = TRUE)
sec_fac  # Print
```

```
> 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
```

ANSWER ##

- **Q4, Q12, Q17** as highlighted seem to best belong to **PC2**. The remaining questions seem to best belong to PC1 with **Q1, Q14, and Q18** being above the 0.8 threshold.

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

```
# Using results from sec_fac to check
require(knitr)  # For creating tables with kable function
kable(sec_fac$Vaccounted |>
    round(2), caption = "Variance Accounted", align = "c")  # Print table of variance
```

Table 1: Variance Accounted

	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

```
# Calculating actual value
pca_sec <- prcomp(sec, scale. = TRUE)
var_explained = pca_sec$sdev^2/sum(pca_sec$sdev^2)
sum(var_explained[1:3])  # Print
```

```
ANSWER > [1] 0.6698246
```

ANSWER ##

- The first 3 PCs capture **67%** of the total variance of the security dataset.

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

```
sec_fac  # Print
```

```
> 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
```

```
sec_fac$communality[2]  # Print
```

```
ANSWER >        Q2
ANSWER > 0.4605433
```

ANSWER ##

- Variance of **Q2** is least explained with an h^2 (commonality) of **0.46** and an unexplained variance (uniqueness) of **0.54**.

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

```
loadings <- sec_fac$loadings[, 1:3] |>
    round(2)
kable(loadings, caption = "Loadings", align = "c")  # Print table of loadings
```

Table 2: Loadings

	PC1	PC2	PC3
Q1	0.82	-0.14	0.00
Q2	0.67	-0.01	0.09

	PC1	PC2	PC3
Q3	0.77	-0.03	0.09
Q4	0.62	0.64	0.11
Q5	0.69	-0.03	-0.54
Q6	0.68	-0.10	0.21
Q7	0.66	-0.32	0.32
Q8	0.79	0.04	-0.34
Q9	0.72	-0.23	0.20
Q10	0.69	-0.10	-0.53
Q11	0.75	-0.26	0.17
Q12	0.63	0.64	0.12
Q13	0.71	-0.06	0.08
Q14	0.81	-0.10	0.16
Q15	0.70	0.01	-0.33
Q16	0.76	-0.20	0.18
Q17	0.62	0.66	0.11
Q18	0.81	-0.11	-0.07

ANSWER ##

- **Three** measurement item as appear to share similar loadings.
- **Q4, Q12, Q17** share similar loadings between **PC1 and PC2** as highlighted.

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

```
pc1_loadings <- sec_fac$loadings[, 1] |>
    round(2)
kable(pc1_loadings, caption = "PC1 Loadings", align = "c", col.names = "PC1")  # Print table of loading
```

Table 3: PC1 Loadings

	PC1
Q1	0.82
Q2	0.67
Q3	0.77
Q4	0.62
Q5	0.69
Q6	0.68
Q7	0.66
Q8	0.79
Q9	0.72
Q10	0.69
Q11	0.75
Q12	0.63
Q13	0.71
Q14	0.81
Q15	0.70
Q16	0.76

	PC1
Q17	0.62
Q18	0.81

ANSWER ##

- Since **Q1, Q14, and Q18** are above a 0.8 threshold for PC1, we could interpret that consumers value the site's **confidentiality of transactions**.

3) To improve interpretability of loadings, let's rotate the our principal component axes using the varimax technique to get rotated components (extract and rotate only three principal components)

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

```
sec_rot <- principal(r = sec, nfactors = 3, rotate = "varimax",
    scores = TRUE)
sec_rot  # Print
```

```
> 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
```

ANSWER ##

- Individually, the amount of variance explained through rotation appears to be **different** from the corresponding principal components.

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

```
# Using results from sec_fac to check
kable(sec_rot$Vaccounted |>
    round(2), caption = "Variance Accounted (RC)", align = "c")  # Print table of variance
```

Table 4: Variance Accounted (RC)

	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

ANSWER ##

- Together, the three rotated components explain the **same** **cumulative variance** as the three principal components combined.
- They both total **0.67**

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?

```
loadings_rot <- sec_rot$loadings[, 1:3] |>
    round(2)
kable(loadings_rot, caption = "Rotated Loadings", align = "c")  # Print table of loadings
```

Table 5: Rotated Loadings

	RC1	RC3	RC2
Q1	0.66	0.45	0.22
Q2	0.54	0.29	0.29
Q3	0.62	0.34	0.31
Q4	0.22	0.19	0.85
Q5	0.24	0.83	0.16
Q6	0.65	0.20	0.23
Q7	0.79	0.10	0.06
Q8	0.38	0.71	0.30
Q9	0.74	0.23	0.14
Q10	0.28	0.82	0.10
Q11	0.76	0.28	0.12
Q12	0.23	0.19	0.85
Q13	0.59	0.32	0.26
Q14	0.72	0.31	0.28
Q15	0.34	0.66	0.24
Q16	0.74	0.27	0.17
Q17	0.21	0.19	0.87
Q18	0.61	0.50	0.23

ANSWER ##

- **Q4, Q12, and Q17** in particular are more clearly differentiated as highlighted.

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

```
loadings_rot <- sec_rot$loadings[, 1:3] |>
    round(2)
kable(loadings_rot, caption = "Rotated Loadings", align = "c")  # Print table of loadings
```

Table 6: Rotated Loadings

	RC1	RC3	RC2
Q1	0.66	0.45	0.22
Q2	0.54	0.29	0.29
Q3	0.62	0.34	0.31
Q4	0.22	0.19	0.85
Q5	0.24	0.83	0.16
Q6	0.65	0.20	0.23
Q7	0.79	0.10	0.06
Q8	0.38	0.71	0.30
Q9	0.74	0.23	0.14
Q10	0.28	0.82	0.10
Q11	0.76	0.28	0.12
Q12	0.23	0.19	0.85
Q13	0.59	0.32	0.26

	RC1	RC3	RC2
Q14	0.72	0.31	0.28
Q15	0.34	0.66	0.24
Q16	0.74	0.27	0.17
Q17	0.21	0.19	0.87
Q18	0.61	0.50	0.23

ANSWER ##

- We can now **more easily** interpret the meaning
- *RC1*: **Q7, Q9, Q11, Q14, and Q16** are above a 0.7 threshold and we could interpret that consumers value the site's **protection of personal information**.
- *RC3*: **Q5, Q8, and Q10** are above a 0.7 threshold and we could interpret that consumers value that the site **verifies the identity**.
- *RC2*: **Q4, Q12, and Q17** are above a 0.8 threshold and we could interpret that consumers value that the site **provides proof to protect against denial of transactions**.

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

```
sec_rot2 <- principal(r = sec, nfactors = 2, rotate = "varimax",
    scores = TRUE)
sec_rot2
```

```
## 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
```

ANSWER ##

- Yes, the meaning changes as more items belong to the first rotated component now.
- In addition, we see a loss in total variance explained from **0.67 to 0.61**

(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 ##

- By looking at the screeplot and performing a parallel analysis, three factors seem to be better than two for explanation and interpretation purposes, although two factors seem to capture enough variance. Ultimately, it depends on the number of dimensions we're dealing with in our data. Since our dimensions have not been that large thus far, three components seem be sufficent.