HW14

Answer are in black text.

Question 1) We saw that identifying the number of principal components to keep can be challenging

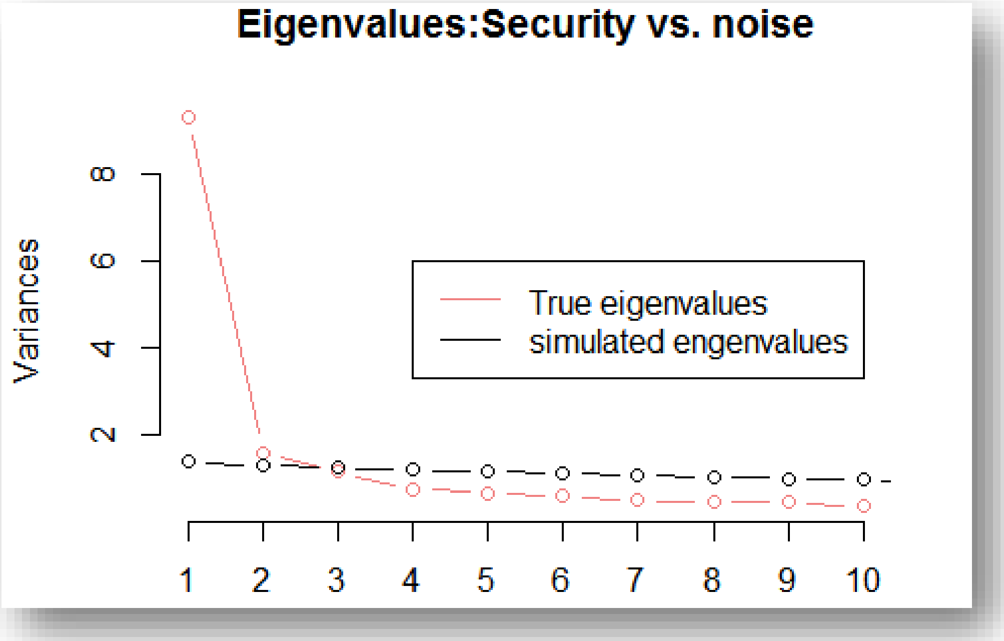
a). Report your earlier findings from applying the “eigenvalue > 1” and screeplot criteria to the security dataset.

Previosly, we use the criteria mentioned above and selected the first three principle components. They could explained about 67% of the variance.

b). Perform a parallel analysis to find out how many principal components have higher eigenvalues than their counterparts in random datasets of the same dimensions as the security dataset.

```
sec_q <- read.csv("security_questions.csv")
sec_pca <- prcomp(sec_q,scale. = TRUE)

#create noise data frame
sim_noise <- function(n,p){
  noise <- data.frame(replicate(p,rnorm(n)))
  return(eigen(cor(noise))$values)
}
set.seed(38)
evalues_noise <- replicate(500, sim_noise(dim(sec_q)[1],dim(sec_q)[2]))
evalues_mean <- apply(evalues_noise,1,mean)
screeplot(sec_pca,type = "line",col="lightcoral",main = "Eigenvalues:Se
curity vs. noise")
lines(evalues_mean,type = "b")
legend(4,6,c("True eigenvalues","simulated engenvalues"),lty=c(1,1),col
=c("lightcoral","black"))
```



From the above screeplot, we can find out that the first **two principal** components have higher eigenvalues than random ones.

Question 2) Earlier, we examined the eigenvectors of the security dataset. This time, let's examine loadings of our principal components (use the principal() method from the psych package)

```
library(psych)

## Warning: package 'psych' was built under R version 3.3.2

principal(sec_q,nfactors = 3,rotate="none",scores = TRUE)

## Principal Components Analysis
## Call: principal(r = sec_q, nfactors = 3, rotate = "none", scores = T
RUE)
## 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 belong?

By setting the threshold of loading to 0.7, I have the following discovers:

- Q1, Q3, Q8, Q9, Q11, Q13, Q14, Q16 and Q18 seem to belong to PC1.
- No loading above 0.7 in PC2, but Q4, Q12, Q17 might belong to it.
- Also, in PC3 no loading above 0.7, Q5 and Q10 are explained by it more, relatively.

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

67 percent of variance captured by the first 3 PCs.

c). Looking at commonality and uniqueness, which item's variance is least explained by the first 3 principal components?

Q2 is least explained by the first 3 principal components with h2 value of 0.46 .

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

About four, Q4, Q7, Q12 and Q17.

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)

Whether users satisfied about the control of the confidentiality of the transactions.

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

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

## Principal Components Analysis
## Call: principal(r = sec_q, 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 explain the same, or different, amount of variance than the three principal components?

Variance explained are **different**.

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

Cumulative variance is the same, 67% percent.

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

According to RC1 it's obviously that those items have more clearly differentiated.

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)

RC1: Q7 Q9 Q11 Q14 Q16 => Personal information protection RC3: Q5 Q8 Q10 => Process of transaction RC2: Q4 Q12 Q17 => Evidence to protect against its denial

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

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

## Principal Components Analysis
## Call: principal(r = sec_q, 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. We can find out that RC1 have more items belong to than the previous model.

(ungraded) Looking back at all our results and analyses, 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 implement PCA for dealing with the curse of dimensionality, but here we could see that the dimension is not that big. Besides, from the above analysis the cumulative variance explained is not exceed 80%, so I think 3 components should be better.