

BACS HW15

109062710

6/2/2021

Question 1

a. Show a scree plot of data

```
df <- read.csv("/home/johnbjohn/Documents/git-repos/bacs-hw/hw15/security_questions.csv")
pca <- prcomp(df, scale. = TRUE)
```

Create a noise simulation function

```
noise_simulation <- function(x, y) {
  random_normal_dist <- replicate(y, rnorm(x))
  noise <- data.frame(random_normal_dist)
  eigenvalues <- eigen(cor(noise))$values
  return(eigenvalues)
}

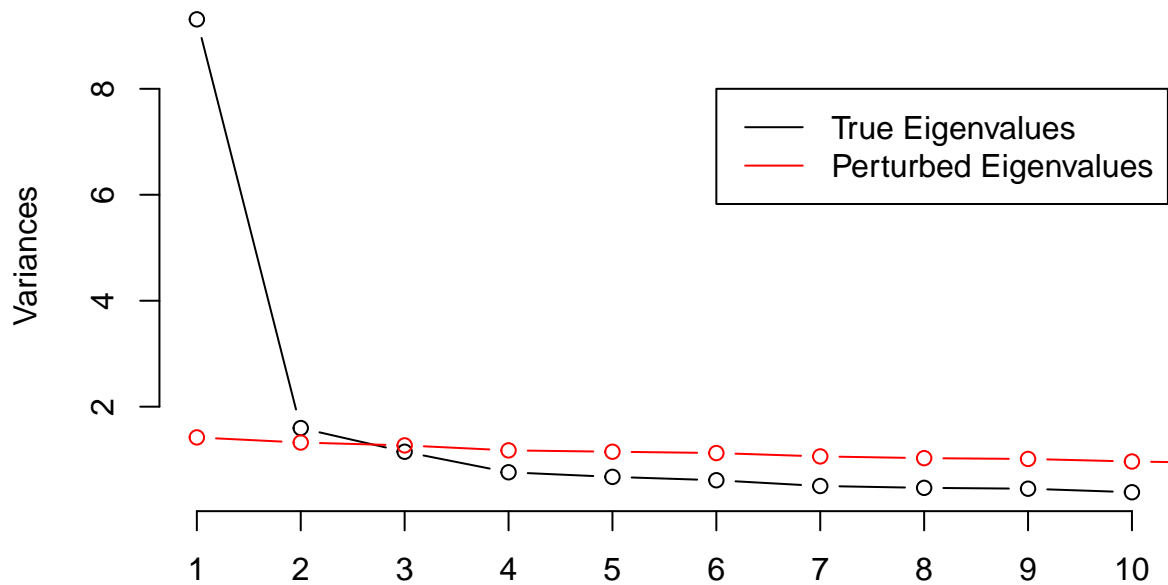
set.seed(2000)
x <- dim(df)[1]
y <- dim(df)[2]
simulate_noise <- replicate(1, noise_simulation(x, y))
evaluated_mean <- apply(simulate_noise, 1, mean)

screeplot(
  pca,
  type = "line",
  col = "black",
  main = "Security Question PCA Scree Plot"
)

lines(evaluated_mean, type = "b", col = "red")

legend(
  6,
  8,
  c("True Eigenvalues", "Perturbed Eigenvalues"),
  lty=c(1,1),
  col = c("black", "red")
)
```

Security Question PCA Scree Plot



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

Question 2

```
pca_result <- principal(df, nfactors = 3, rotate = "none", scores = TRUE)
pca_result
```

```
## Principal Components Analysis
## Call: principal(r = df, 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. To which principal components does each question seems to best belong?

Let's set a threshold of 0.7, we then have Q1, Q3, Q8, Q9, Q11, Q13, Q14, Q16, Q18 that belong to PC1 and the other questions seem to belong to either PC2 or PC3.

b. How much variance captured by PC1, PC2, and PC3.

```
summary(pca)$importance[2, c(1:3)]
```

```
##      PC1      PC2      PC3
## 0.51728 0.08869 0.06386
```

All we need is the first three columns of the second row, which is Proportion of Variance. Thus, the variance explained by the PC1 is 51%, PC2 is 8%, and PC3 is 6%. In total, PC1, PC2, and PC3 explain roughly 66% of total variance.

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

It's Q2 since it has 'h2' of 0.4605433.

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

```

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
    )
  )
}

evaluate_loadings(pca_result$loading, 1:18)

```

```

##      Q1      Q2      Q3      Q4      Q5      Q6      Q7      Q8      Q9      Q10     Q11     Q12     Q13
## FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
##      Q14     Q15     Q16     Q17     Q18
## FALSE FALSE FALSE  TRUE FALSE

```

Only Q4, Q12, and Q17 share similar loadings between 2 or more components.

Question 3

a. Individually, does each RC explain the same amount of variance?

```

rc <- principal(df, nfactors = 3, rotate = "varimax", scores = TRUE)$loadings
rc

```

```

##
## Loadings:
##      RC1    RC3    RC2
## Q1  0.660 0.450 0.221
## Q2  0.544 0.286 0.288
## Q3  0.621 0.337 0.311
## Q4  0.218 0.193 0.854
## Q5  0.244 0.828 0.162
## Q6  0.652 0.199 0.234
## Q7  0.790 0.103
## Q8  0.382 0.706 0.305
## Q9  0.738 0.234 0.138
## Q10 0.277 0.823 0.102
## Q11 0.757 0.278 0.118
## Q12 0.233 0.186 0.854
## Q13 0.593 0.315 0.259
## Q14 0.719 0.310 0.283
## Q15 0.342 0.656 0.244

```

```
## Q16 0.740 0.267 0.174
## Q17 0.205 0.187 0.870
## Q18 0.609 0.495 0.227
##
##          RC1    RC3    RC2
## SS loadings  5.613 3.490 2.954
## Proportion Var 0.312 0.194 0.164
## Cumulative Var 0.312 0.506 0.670
```

Looking at the proportion variance, we can see RC1 is 30% and PC1 is 51%. RC2 is 19% and PC3 is 8%. While RC3 is 16% and PC3 is 6%. Clearly, they are different.

b. Together, do these rotated components explain the same amount of variance like the principal components?

Yes, both the rotated components and the principal components have the same total amount of variances explained.

c. do Q4, Q12, and Q17 have more clearly differentiated loadings among rotated components?

```
rc[c(4,12,17), 1:3]
```

```
##          RC1          RC3          RC2
## Q4  0.2182880 0.1933627 0.8536838
## Q12 0.2327616 0.1861745 0.8542346
## Q17 0.2054021 0.1869028 0.8703910
```

Since these rotate components have loading value above 0.7, clearly they are more clearly differentiated.

d. Can you interpret the meaning of the three rotated components?

```
rc[rc[, 1] > 0.7, 1]
```

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

RC1 means data protection.

```
rc[rc[, 2] > 0.7, 2]
```

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

RC2 means transaction processing.

```
rc[rc[, 3] > 0.7, 3]
```

```
##           Q4           Q12           Q17  
## 0.8536838 0.8542346 0.8703910
```

RC3 means providing evidence to protect against its denial.

e. If we reduced the number of rotated components, does the meaning of our rotated components change?

```
reduced_rc <- principal(df, nfactors = 2, rotate = "varimax", scores = TRUE)  
reduced_rc$loadings[,1][reduced_rc$loadings[,1] > 0.7]
```

```
##           Q1           Q7           Q9           Q11           Q14           Q16           Q18  
## 0.7830951 0.7284256 0.7451939 0.7855784 0.7591295 0.7615661 0.7616746
```

Yes. When we decreased `nfactors` by 1, the number of questions belong to 'RC1' would increase.