bacs\_hw12

110071010

2024-05-11

**110034002** walked me through the parameters of the principal(), and i realizes that h2 stands for “commonality”, u2 “uniqueness”, and com “item complexity”

**109048231** notified me that in Question 3e, the factor loadings themselves reflected the correlation-like relationship. And clearly, we got different meanings from the component (from 3 to 2 in Question 3 case) since the greater-than-0.7 loadings shifted significantly.

# Question 1

Earlier, we examined a dataset from a security survey sent to customers of e-commerce websites. However, we only used the “eigenvalue > 1” criteria and the “elbow rule” on the screeplot to find a suitable number of components. Let’s perform a parallel analysis as well this week

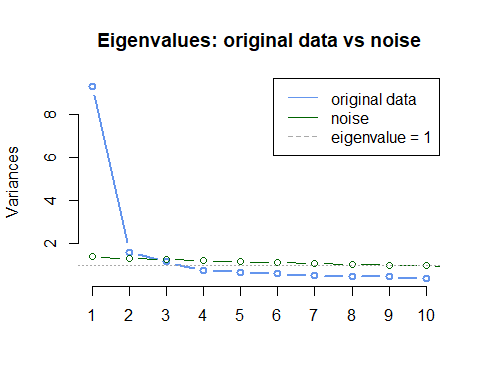
sq <- read.csv('security\_questions.csv')  
pca\_sq <- prcomp(sq, scale. = TRUE)  
summary(pca\_sq)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 3.0514 1.26346 1.07217 0.87291 0.82167 0.78209 0.70921  
## Proportion of Variance 0.5173 0.08869 0.06386 0.04233 0.03751 0.03398 0.02794  
## Cumulative Proportion 0.5173 0.60596 0.66982 0.71216 0.74966 0.78365 0.81159  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.68431 0.67229 0.6206 0.59572 0.54891 0.54063 0.51200  
## Proportion of Variance 0.02602 0.02511 0.0214 0.01972 0.01674 0.01624 0.01456  
## Cumulative Proportion 0.83760 0.86271 0.8841 0.90383 0.92057 0.93681 0.95137  
## PC15 PC16 PC17 PC18  
## Standard deviation 0.48433 0.4801 0.4569 0.4489  
## Proportion of Variance 0.01303 0.0128 0.0116 0.0112  
## Cumulative Proportion 0.96440 0.9772 0.9888 1.0000

## 1a

Show a single visualization with scree plot of data, scree plot of simulated noise (use average eigenvalues of ≥ 100 noise samples), and a horizontal line showing the eigenvalue = 1 cutoff.

set.seed(1111313131)  
  
sim\_noise\_ev <- function(n,p){  
 noise <- data.frame(replicate(p, rnorm(n)))  
 eigen(cor(noise))$values  
}  
evalues\_noise <- replicate(100 ,sim\_noise\_ev(406,18))  
evalues\_mean <- apply(evalues\_noise,1,mean)  
  
  
# Plotting  
screeplot(pca\_sq, col= "cornflowerblue", type = "lines", lwd = 2, main = "Eigenvalues: original data vs noise")  
lines(evalues\_mean, type = "b",col = "darkgreen")  
abline(h = 1,col = "darkgray", lty = "dotted")  
  
legend("topright", legend=c("original data", "noise", "eigenvalue = 1"),  
 col=c("cornflowerblue", "darkgreen", "darkgray"), lty=c(1, 1, 2))



## 1b

**Question**

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

**Answer**

As observed in the screeplot above, we should only **take PC1 and PC2**. Starting from PC3, the variances of the garbage data (noise) start surpassing those of the original data.

# Question 2

Earlier, we treated the underlying dimensions of the security dataset as composites and examined their eigenvectors (weights). Now, let’s treat them as factors and examine factor loadings (use the principal() method from the psych package)

## 2a

**Question**

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

sq <- read.csv('security\_questions.csv')  
library(psych)  
sq\_principal <- principal(sq, nfactor = 3, rotate= "none", scores = TRUE)  
sq\_principal$loadings[,1:3]

## PC1 PC2 PC3  
## Q1 0.8169846 -0.13941235 -0.002115927  
## Q2 0.6726084 -0.01375526 0.089174403  
## Q3 0.7655215 -0.03269651 0.089686106  
## Q4 0.6233733 0.64307826 0.108031860  
## Q5 0.6900841 -0.03126466 -0.542354570  
## Q6 0.6828029 -0.10462094 0.207232000  
## Q7 0.6566249 -0.31763196 0.324176779  
## Q8 0.7861054 0.04235983 -0.343212951  
## Q9 0.7230295 -0.23164618 0.203556038  
## Q10 0.6861529 -0.09868038 -0.532678749  
## Q11 0.7529735 -0.26100673 0.172516196  
## Q12 0.6303505 0.63753124 0.121522834  
## Q13 0.7119085 -0.06463837 0.084335919  
## Q14 0.8114677 -0.09970016 0.156787046  
## Q15 0.7040428 0.01057936 -0.332546876  
## Q16 0.7575616 -0.20281591 0.183170175  
## Q17 0.6175336 0.66426051 0.110061160  
## Q18 0.8067284 -0.11360432 -0.065189145

**Answer**

The first principal component

## 2b

**Question**

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

sq\_principal$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

**Answer**

PC1: 0.517 PC2: 0.089 PC3: 0.064 (individual)

0.670 (cumulative)

## 2c

**Question**

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

sq\_principal$communality

## Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8   
## 0.6869041 0.4605433 0.5951359 0.8138147 0.7713420 0.5201104 0.6371369 0.7375512   
## Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16   
## 0.6178667 0.7642903 0.6648554 0.8185557 0.5181043 0.6930021 0.6063756 0.6485852   
## Q17 Q18   
## 0.8347032 0.6679663

**Answer**

Commonality (h2) + uniqueness(u2) =1 , and high uniqueness(u2) means data being less explained. It is known that the first three PCs capture 0.67 of total variance, and items whose commonality is less than 0.67 is exactly what we are looking for: Q2, 3, 6, 7, 9, 11, 13, 15, 16, 18.

## 2d

**Question**

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

sq\_principal$complexity

## Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8   
## 1.058202 1.035995 1.031144 2.055762 1.899001 1.233397 1.959957 1.374540   
## Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16   
## 1.373796 1.932541 1.351862 2.072501 1.044775 1.105810 1.425577 1.266376   
## Q17 Q18   
## 2.047594 1.052956

**Answer**

It can be seen that Q4, 12, 17 have their item complexity value greater than 2.

## 2e

**Question**

Can you interpret 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**

The patterns looks no clear for interpretation, but my guess is that the first principal component vaguely captures “confidentiality”

# Question 3

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

## 3a

**Question**

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

**Answer**

Different.

PC1 = 0.517 , RC1 = 0.312

PC2 = 0.089 , RC2 = 0.164

PC3 = 0.064 , RC3 = 0.194

sq\_pca\_rot <- principal(sq, nfactor = 3, rotate = "varimax", scores = TRUE)  
sq\_pca\_rot$loadings

##   
## 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

## 3b

**Question**

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

**Answer**

The same. Proven by the fact that cumulative variance = 0.67.

## 3c

**Question**

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**

Yes.

sq\_pca\_rot$loadings

##   
## 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

## 3d

**Question**

Can you now more easily 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**

RC1: Q7, Q9, Q11, Q14, Q16 –> “personal information”

RC2: Q4,Q12,Q17 –> “evidence showing the transaction is not denied”

RC3: Q5,Q8,Q10 –> “authenticity & user-to-website security”

## 3e

**Question**

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

**Answer**

Yes. It can be observed that both RC1 and RC2 (column) have different questions (row) yield correlations greater than 0.7, so the underlying meanings might have changed.

sq\_pca\_rot <- principal(sq, nfactor = 2, rotate = "varimax", scores = TRUE)  
sq\_pca\_rot$loadings

##   
## Loadings:  
## RC1 RC2   
## Q1 0.783 0.271  
## Q2 0.596 0.312  
## Q3 0.687 0.340  
## Q4 0.236 0.864  
## Q5 0.620 0.305  
## Q6 0.649 0.237  
## Q7 0.728   
## Q8 0.668 0.416  
## Q9 0.745 0.145  
## Q10 0.649 0.244  
## Q11 0.786 0.134  
## Q12 0.245 0.862  
## Q13 0.655 0.286  
## Q14 0.759 0.304  
## Q15 0.612 0.348  
## Q16 0.762 0.187  
## Q17 0.221 0.880  
## Q18 0.762 0.289  
##   
## RC1 RC2  
## SS loadings 7.521 3.387  
## Proportion Var 0.418 0.188  
## Cumulative Var 0.418 0.606