BACS-hw14-107070004

Let's reconsider the security questionnaire from last week, where consumers were asked security related questions about one of the e-commerce websites they had recently used.

```
security_questions <- read.csv("./security_questions.csv")
sq_pca <- prcomp(security_questions)
```

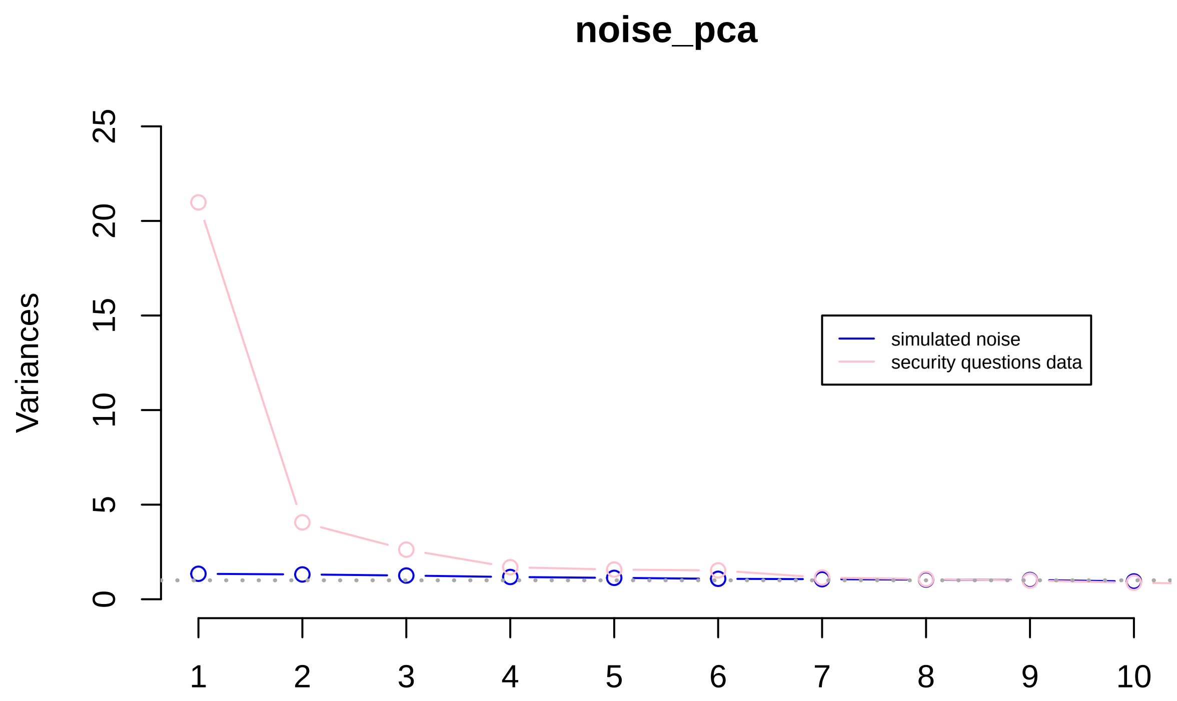
Question 1) Earlier, we examined a dataset from a security survey sent to customers of e-commerce websites. However, we only used the eigenvalue > 1 criteria and the screeplot to find a suitable number of components. Let's perform a parallel analysis as well this week:

(a) 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.

```
var_sq_pca<-sq_pca$sdev^2
noise<-data.frame(replicate(ncol(security_questions),rnorm(nrow(security_questions))))
eigen(cor(noise))$values|>round(2)
```

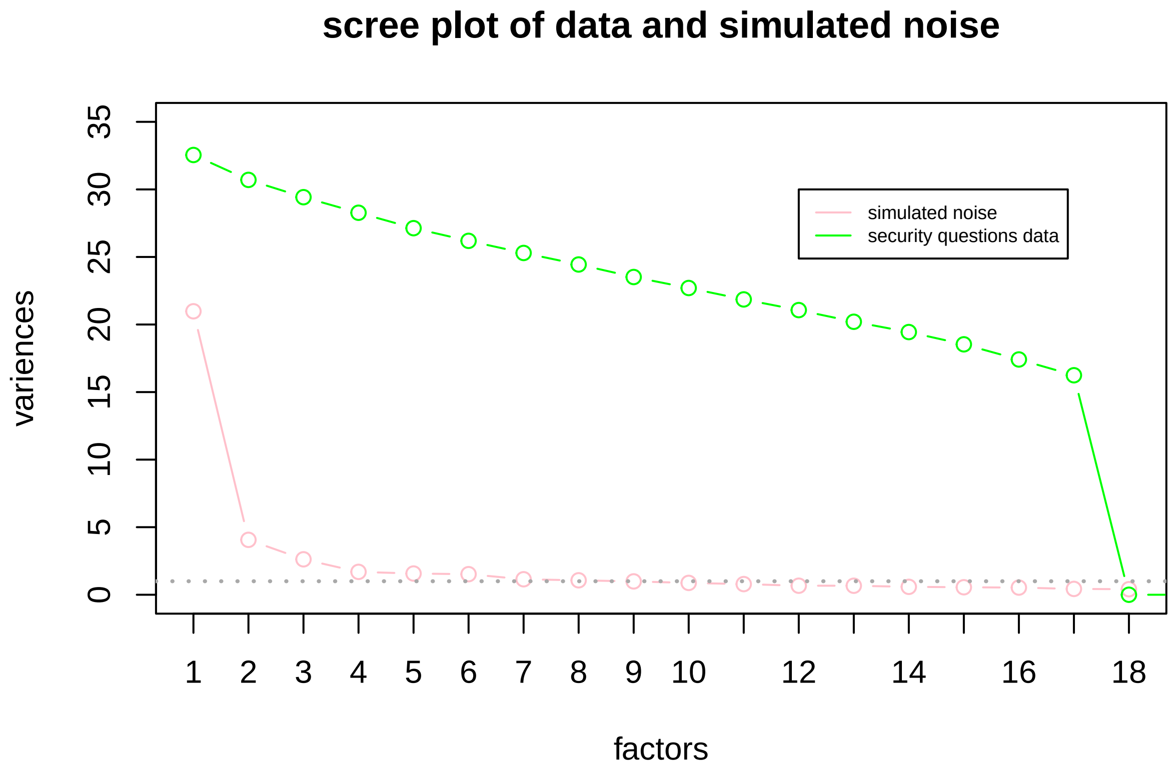
```
##  [1] 1.35 1.31 1.25 1.18 1.13 1.08 1.06 1.04 1.02 0.94 0.93 0.91 0.88 0.87 0.82
## [16] 0.76 0.73 0.73
```

```
noise_pca<-prcomp(noise,scale. = TRUE)
screeplot(noise_pca, type="lines", col="blue" , ylim=c(0,25))
lines(var_sq_pca, type="b", col="pink")
abline(h=1, lty="dotted", col="darkgray", lwd=2)
legend(7, 15, legend = c("simulated noise", "security questions data"),
       , lty = 1, col = c("blue", "pink"), cex = 0.6)
```



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

```
sim_noise_ev<-function(n, p) {
  noise<-data.frame(replicate(p, rnorm(n)))
  eigen(cor(noise))$values|>round(2)
}
evalues_noise<-replicate(100, sim_noise_ev(ncol(security_questions), nrow(security_questions)))
evalues_mean<-apply(evalues_noise, 1, mean)
plot(var_sq_pca, type="b", col="pink", xaxt = 'n'
    , ylim=c(0,35), xlim=c(1,18), xlab="factors", ylab="variences"
    , main="scree plot of data and simulated noise")
axis(1, at = 1:18)
lines(evalues_mean, type="b", col="green")
abline(h=1, lty="dotted", col="darkgray", lwd=2)
legend(12, 30, legend = c("simulated noise", "security questions data")
       , lty = 1, col = c("pink", "green"), cex = 0.6)
```



ans: 0 dimensions will be retained.

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)

```
sq_principal<-principal(security_questions, nfactor=3, rotate="none", scores=TRUE)
sq_principal
```

```
## Principal Components Analysis
## Call: principal(r = security_questions, nfactors = 3, rotate = "none",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##         PC1   PC2   PC3   h2   u2 com
## X...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 best belong?

- PC1 : Q1, Q3, Q8, Q9, Q11, Q13, Q14, Q15, Q16, Q18
- PC2 : none
- PC3 : none

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

```
sq_eigen <- eigen(cor(security_questions))
sq_eigen$values[1:3]/sum(sq_eigen$values)
```

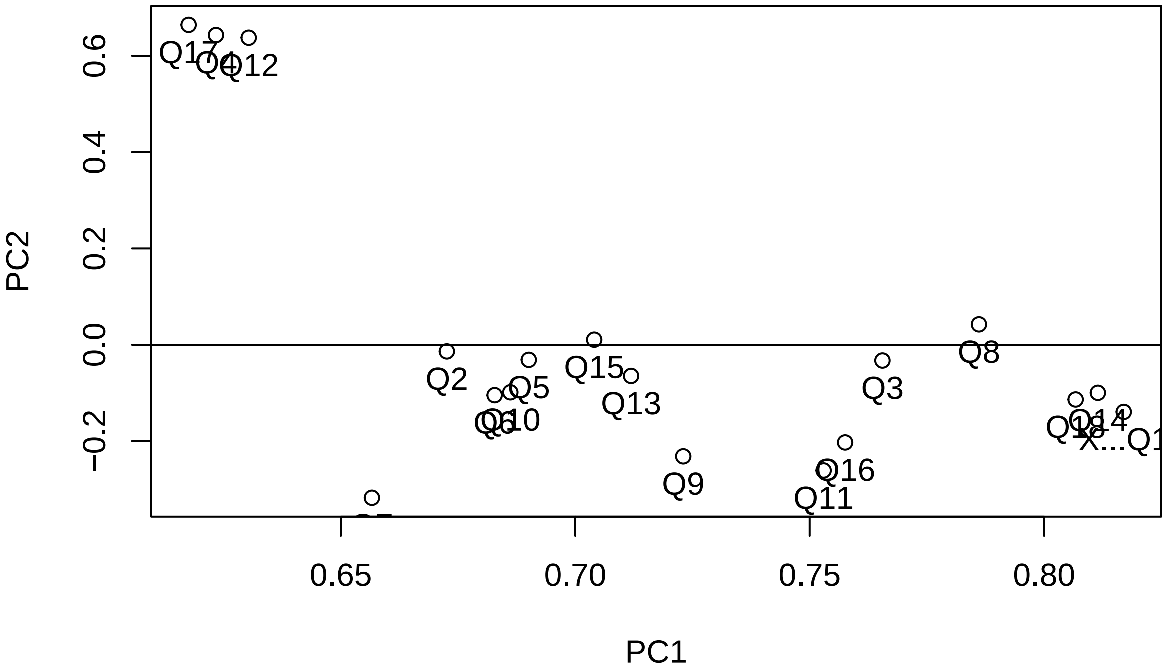
```
## [1] 0.51727518 0.08868511 0.06386435
```

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

ans: Q2 is less than adequately explained by the first 3 principal components, because its u2 is 0.54 (more than 0.5).

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

```
plot(sq_principal$loadings)
text(sq_principal$loadings, pos=1, labels=rownames(sq_principal$loadings))
abline(h=0, v=0)
```



- Q1 & Q14 & Q18
- Q6 & Q10
- Q4 & Q12 & Q17

ans: 8

(e) 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)

ans: The items with best loadings with first principal component means that correlation of PC and items are high. Due to relatively high correlations among items, this would be a good candidate for factor analysis. Recall that the goal of factor analysis is to model the interrelationships between items with fewer (latent) variables. These interrelationships can be broken up into multiple components.

Question 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)

```
sq__pca_rot<-principal(security_questions, nfactor=3, rotate="varimax", scores=TRUE)
sq__pca_rot
```

```
## Principal Components Analysis
## Call: principal(r = security_questions, nfactors = 3, rotate = "varimax",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##         RC1  RC3  RC2   h2   u2 com
## X...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 (RC) explain the same, or different, amount of variance than the corresponding principal components (PCs)?

ans: No, there are totally different.

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

ans:

- RC1 < PC1
- RC2 > PC2
- RC3 > PC3

(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?

ans: The items that shared similar loadings with multiple principal components, share similar loadings with rotated components.

(d) 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)

ans: After rotating the principal components, the original dimensions are closer to axes. Therfore, Rotated components rotate principal components to maximize interpretation of loadings (increase interpretabilty).

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

```r
sq__pca_rot<-principal(security_questions, nfactor=2, rotate="varimax", scores=TRUE)
sq__pca_rot
```

```
## Principal Components Analysis
## Call: principal(r = security_questions, nfactors = 2, rotate = "varimax",
##     scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##         RC1  RC2   h2   u2 com
## X...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
```

ans: Yes, as you can see the RC1 and RC2's results are different from the 3 rotated components.