HW12

113078506

2025-05-10

- Gen AI Usage: I use Gen AI to refine my grammar, verify my reasoning and adjust to cleaner code.
- Students who helped: 113078505, 113078502, and 113078514, helped with conclusion reasoning.

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:

```
library(readxl)
sec_questions <- data.frame(read_excel("security_questions.xlsx", sheet = "data", col_names = T))
```

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.

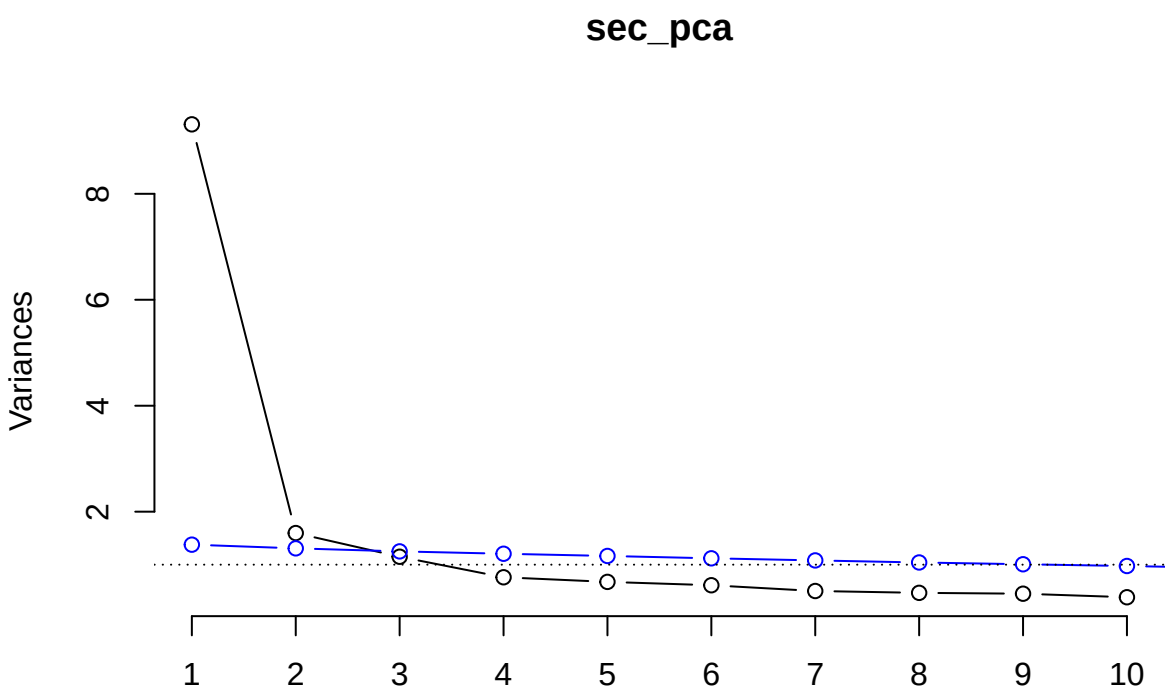
```
# find eigenvalues
sec_pca <- prcomp(sec_questions, scale. = TRUE)
sec_ev <- (sec_pca$sdev)^2

# Simulate noise and run PCA many times
sim_noise_ev <- function(n, p) {
  fake_data <- matrix(rnorm(n*p), nrow=n, ncol=p)
  pca <- prcomp(fake_data, scale. = TRUE)
  return((pca$sdev)^2)
}

# Repeat k times
k <- 100
evalues_noise <- replicate(k, sim_noise_ev(n = nrow(sec_questions),
                                           p = ncol(sec_questions)))

# get means of each row (10 means total)
evalues_mean <- rowMeans(evalues_noise)

screeplot(sec_pca, type="lines")
lines(evalues_mean, type="b", col = "blue")
abline(h=1, lty="dotted")
```



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

Based on the results of the parallel analysis, we observe that only the first two principal components have eigenvalues that exceed the average eigenvalues derived from randomly generated noise datasets. This suggests that these two components capture more systematic variance than what would be expected by chance alone. In contrast, the remaining components fall below or approximately equal to the noise threshold, indicating that they likely reflect random variation rather than meaningful structure in the data. Therefore, following the logic of parallel analysis, which emphasizes retaining only components that outperform noise, we conclude that only the first two dimensions should 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)

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

```
library("knitr")
library(psych)
sec_pca_fact <- principal(sec_questions, nfactors = 3, rotate = "none", scores = TRUE)

# print loadings table
```

```r
loadings_mat <- as.matrix(unclass(sec_pca_fact$loadings))
rownames(loadings_mat) <- rownames(sec_pca_fact$loadings)
kable(loadings_mat, digits = 2, caption = "Factor Loadings")
```

Table 1: Factor Loadings

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

Most items (e.g., Q1, Q2, Q3, Q6, Q8, Q11, Q14, Q18) load most strongly on PC1, indicating that they are primarily explained by the first principal component. A few items such as Q4, Q12, and Q17 show their highest loadings on PC2, while items like Q5 and Q10 have their highest (negative) loadings on PC3, suggesting that these components capture unique aspects not represented in PC1.

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

```r
sec_pca_fact$Vaccounted
```

```
##                             PC1        PC2        PC3
## SS loadings           9.3109533 1.59633195 1.14955822
## Proportion Var        0.5172752 0.08868511 0.06386435
## Cumulative Var        0.5172752 0.60596029 0.66982464
## Proportion Explained  0.7722546 0.13240049 0.09534487
## Cumulative Proportion 0.7722546 0.90465513 1.00000000
```

The first three principal components together account for approximately 64% of the total variance in the security dataset. Specifically, PC1 explains around 52%, PC2 explains 9%, and PC3 explains 6%. This suggests that the majority of the variance is captured by the first component, with the second and third components adding modest but potentially meaningful contributions.

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

```
round(sec_pca_fact$communality, 2)
```

```
##   Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9  Q10  Q11  Q12  Q13  Q14  Q15  Q16
## 0.69 0.46 0.60 0.81 0.77 0.52 0.64 0.74 0.62 0.76 0.66 0.82 0.52 0.69 0.61 0.65
##  Q17  Q18
## 0.83 0.67
```

```
round(sec_pca_fact$uniquenesses, 2)
```

```
##   Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9  Q10  Q11  Q12  Q13  Q14  Q15  Q16
## 0.31 0.54 0.40 0.19 0.23 0.48 0.36 0.26 0.38 0.24 0.34 0.18 0.48 0.31 0.39 0.35
##  Q17  Q18
## 0.17 0.33
```

Looking at the commonalities and uniqueness values, items with communalities below 0.40 are considered to be less adequately explained by the first three principal components. In this case, Q2 (0.46), Q6 (0.52), Q7 (0.64), Q8 (0.74), Q9 (0.62), Q13 (0.52), Q15 (0.61), and Q18 (0.67) fall within a moderate range and are reasonably explained. However, Q2 (commonality = 0.46, uniqueness = 0.54), Q6 (0.52 / 0.48), Q7 (0.64 / 0.36), Q8 (0.74 / 0.26), and Q13 (0.52 / 0.48) are on the lower end. More critically, Q15 (commonality = 0.61) and Q18 (0.67) are just below the common threshold for concern, but Q2, Q6, Q13, and Q7 show relatively higher uniqueness (0.48), suggesting that a substantial portion of their variance remains unexplained. These items may reflect noise, measurement error, or constructs not captured by the current factor structure.

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

```
kable(loadings_mat, digits = 2, caption = "Factor Loadings")
```

Table 2: Factor Loadings

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

	PC1	PC2	PC3
Q16	0.76	-0.20	0.18
Q17	0.62	0.66	0.11
Q18	0.81	-0.11	-0.07

Based on the commonly used threshold of 0.30 for interpreting meaningful loadings, six items (Q4, Q5, Q7, Q8, Q10, Q12, and Q17) display cross-loadings across two or more principal components. These items load significantly on more than one component, suggesting that they may reflect overlapping or multidimensional constructs rather than being clearly associated with a single latent factor.

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)

The first principal component appears to represent a general sense of perceived data security and transactional trustworthiness in the e-commerce platform. Items that load most strongly on this component (such as Q1, Q3, Q8, Q11, Q14, and Q18) focus on the confidentiality of personal information (Q1, Q18), the verification and accuracy of communicated data (Q3, Q14), identity authentication (Q8), and the site's efforts to protect against unauthorized access (Q11). Together, these items reflect a user's holistic confidence that the platform handles sensitive transactions responsibly and securely. Hence, PC1 may be interpreted as capturing the construct of "overall trust in data protection and transaction integrity."

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)

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

```
sec_pca_rot <- principal(sec_questions, nfactors = 3, rotate = "varimax", scores = TRUE)
sec_pca_rot$Vaccounted
```

```
##                            RC1       RC3       RC2
## SS loadings           5.6131484 3.4901395 2.9535556
## Proportion Var        0.3118416 0.1938966 0.1640864
## Cumulative Var        0.3118416 0.5057382 0.6698246
## Proportion Explained  0.4655570 0.2894737 0.2449692
## Cumulative Proportion 0.4655570 0.7550308 1.0000000
```

```
sec_pca_fact$Vaccounted
```

```
##                            PC1        PC2        PC3
## SS loadings           9.3109533 1.59633195 1.14955822
## Proportion Var        0.5172752 0.08868511 0.06386435
## Cumulative Var        0.5172752 0.60596029 0.66982464
## Proportion Explained  0.7722546 0.13240049 0.09534487
## Cumulative Proportion 0.7722546 0.90465513 1.00000000
```

Individually, each rotated component (RC) explains a different amount of variance than its corresponding unrotated principal component (PC). For example, PC1 explains 51.7% of the variance, while RC1 explains only 31.2%. Conversely, PC2 and PC3 explain very little variance individually (8.9% and 6.4%, respectively), whereas RC2 and RC3 explain 16.4% and 19.4%, respectively.

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

Despite the differences in individual variance explained, the total cumulative variance remains the same (66.98%) in both rotated and unrotated solutions.

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_mat_rot <- as.matrix(sec_pca_rot$loadings)[, 1:3]
kable(loadings_mat_rot, digits = 2, caption = "Rotated Factor Loadings")
```

Table 3: Rotated Factor 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

Yes, the items that previously showed cross-loadings among multiple unrotated principal components now exhibit more clearly differentiated loadings after varimax rotation. For example:

- Q4 and Q12, which originally loaded on both PC1 and PC2, now load strongly on RC2 with minimal loading elsewhere.
- Q5 and Q10, which had strong associations with both PC1 and PC3 (including negative loadings), now load clearly on RC3.
- Q7, which previously had moderate loadings on PC1 and PC3, now loads almost exclusively on RC1.

- Q17, originally ambiguous between PC1 and PC2, now loads cleanly on RC2.

This demonstrates that rotation helped clarify the factor structure by reducing ambiguity and allowing each item to align more distinctly with a single component.

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)

Yes, after rotation, the meanings of the three components become more clearly interpretable based on the items that load strongly on each:

- RC1 (Q1, Q3, Q6, Q7, Q9, Q11, Q14, Q16, Q18) captures the trust in data privacy and control, reflecting their trust in the platform's ability to safeguard personal information, prevent unauthorized access, and allow users to manage their own data.
- RC2 (Q4, Q12, Q17) reflects proof of transaction and accountability, emphasizing the platform's ability to provide evidence of transactions and prevent denial of service.
- RC3 (Q5, Q8, Q10, Q15) pertains to identity verification and message authenticity, covering users' confidence that the site verifies sender/receiver identity and ensures message accuracy and legitimacy.

This rotation thus enhances the semantic clarity of the factor structure and makes each component's psychological construct more distinct and actionable.

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

```
sec_pca_rot_2 <- principal(sec_questions, nfactors = 2, rotate = "varimax", scores = TRUE)
loadings_mat_rot_2 <- as.matrix(sec_pca_rot_2$loadings)[, 1:2]
kable(loadings_mat_rot_2, digits = 2, caption = "2 Rotated Factor Loadings")
```

Table 4: 2 Rotated Factor Loadings

	RC1	RC2
Q1	0.78	0.27
Q2	0.60	0.31
Q3	0.69	0.34
Q4	0.24	0.86
Q5	0.62	0.31
Q6	0.65	0.24
Q7	0.73	0.04
Q8	0.67	0.42
Q9	0.75	0.15
Q10	0.65	0.24
Q11	0.79	0.13
Q12	0.25	0.86
Q13	0.65	0.29
Q14	0.76	0.30
Q15	0.61	0.35
Q16	0.76	0.19

	RC1	RC2
Q17	0.22	0.88
Q18	0.76	0.29

Yes, reducing the number of components from three to two changes the meaning of the rotated components. In the 3-component solution, distinct themes were separated: RC1 focused on data privacy, RC2 on proof of transaction and accountability, and RC3 on identity verification. However, in the 2-component solution, items from the original RC3 (Q5, Q8, Q10, Q15) are now absorbed into RC1. This broadens RC1 to cover multiple aspects of security, including privacy, identity, and message authenticity. While RC2 still reflects transaction and accountability (Q4, Q12, Q17), the interpretability of RC1 becomes less focused and more ambiguous. This demonstrates that while reducing dimensions simplifies the model, it may also weaken conceptual clarity.

(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.

Based on all the analyses conducted, I believe that extracting three components provides the most meaningful and interpretable structure for understanding the security dataset.

From a statistical perspective, parallel analysis suggested retaining two components, as only the first two eigenvalues exceeded the simulated noise threshold.

However, from an interpretability and conceptual clarity perspective, the three-component solution clearly distinguished between (1) trust in data privacy, (2) transaction accountability, and (3) identity verification. These are all distinct and meaningful aspects of perceived online security.

If the goal is to simplify the model for practical application (e.g., building a shorter survey or index), a two-component solution may be sufficient, although it combines multiple themes into broader categories. But for theoretical understanding or detailed behavioral insights, the three-component solution offers better clarity and separation of constructs.