Hw 7

r12631055

2024-11-11

Problem 1

Use the engineer data. Combine the two groups into a single sample. (a) Using a scree plot, the number of eigenvalues greater than 1, and the percentages, is there a clear choice of m? (b) Extract three factors by the principal component method and carry out a varimax rotation. (c) Extract three factors by the principal axis factor method and carry out a varimax rotation. (d) Compare the results of parts (b) and (c)

First, load the data:

```
pilot <- read.table("pilots.dat", header = FALSE)
str(pilot)
```

```
## 'data.frame':    40 obs. of  7 variables:
##  $ V1: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ V2: int  121 108 122 77 140 108 124 130 149 129 ...
##  $ V3: int  22 30 49 37 35 37 39 34 55 38 ...
##  $ V4: int  74 80 87 66 71 57 52 89 91 72 ...
##  $ V5: int  223 175 266 178 175 241 194 200 198 162 ...
##  $ V6: int  54 40 41 80 38 59 72 85 50 47 ...
##  $ V7: int  254 300 223 209 261 245 242 242 277 268 ...
```

Remove grouping row

```
pilot_data <- pilot[, -1]
str(pilot_data)
```

```
## 'data.frame':    40 obs. of  6 variables:
##  $ V2: int  121 108 122 77 140 108 124 130 149 129 ...
##  $ V3: int  22 30 49 37 35 37 39 34 55 38 ...
##  $ V4: int  74 80 87 66 71 57 52 89 91 72 ...
##  $ V5: int  223 175 266 178 175 241 194 200 198 162 ...
##  $ V6: int  54 40 41 80 38 59 72 85 50 47 ...
##  $ V7: int  254 300 223 209 261 245 242 242 277 268 ...
```

scree plot

```
install.packages("psych")
```

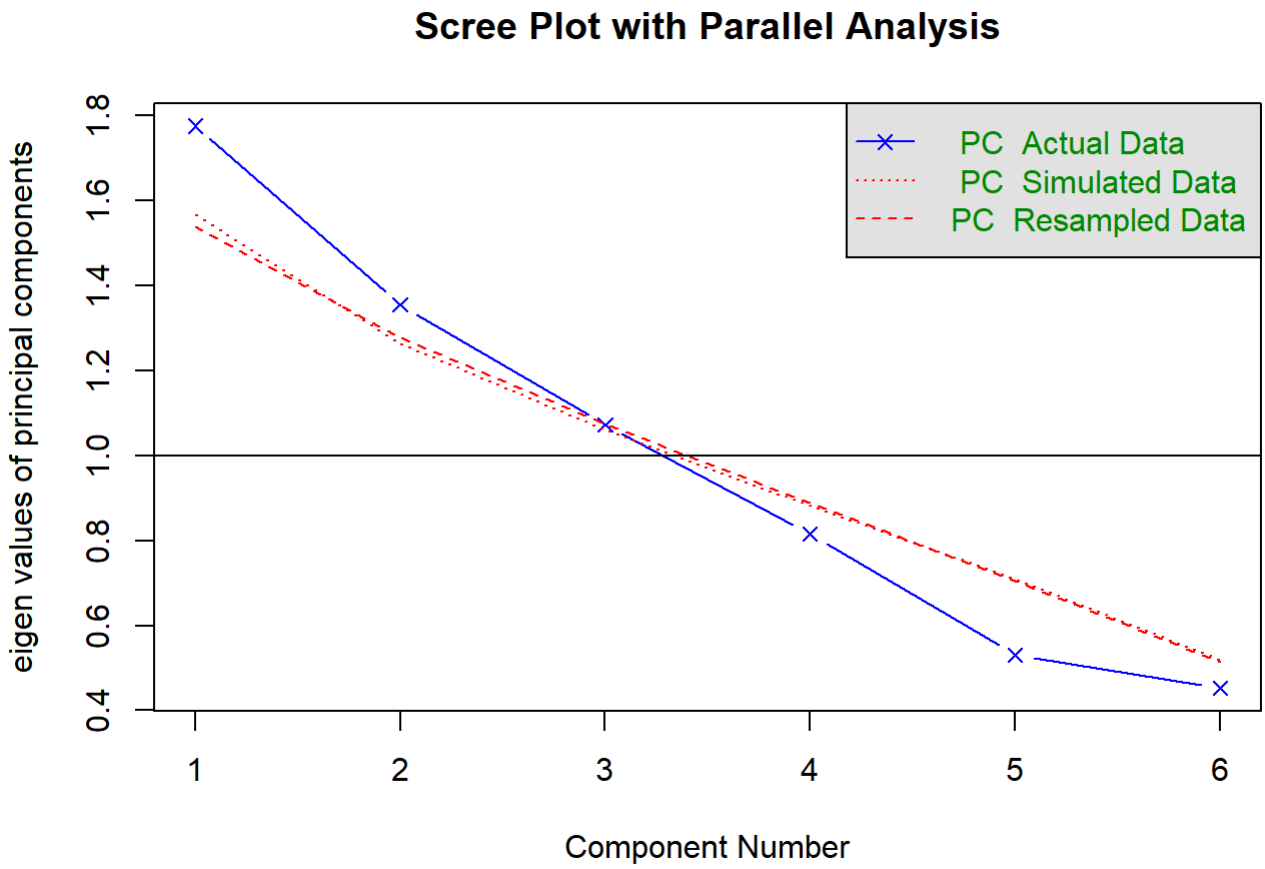
```
## 將程式套件安載入 'C:/Users/Paul/AppData/Local/R/win-library/4.4'
## (因為 'lib' 沒有被指定)
```

```
## 程式套件 'psych' 開啟成功，MD5 和檢查也透過
##
## 下載的二進位程式套件在
##  C:\Users\Paul\AppData\Local\Temp\RtmpuQbGB6\downloaded_packages 裡
```

```
library(psych)
```

```
## Warning: 套件 'psych' 是用 R 版本 4.4.2 來建造的
```

```
fa.parallel(pilot_data, fa = "pc", n.iter = 100, show.legend = TRUE, main = "Scree Plot with Parallel Analysis")
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  0
```

a. 1.Eigenvalues Greater Than 1: From the scree plot, the first three components have eigenvalues greater than 1. According to the Kaiser criterion, this suggests retaining 3 components.

2.Scree Plot Analysis: The scree plot shows a significant drop after the first component, followed by a more gradual decline for the subsequent components. There is a clear “elbow” after the first component, which suggests that the first component explains the majority of the variance, and the contribution of subsequent components starts to decrease. However, since the eigenvalues of the first three components are greater than 1, it indicates that all three components are important.

3.Percentage of Variance Explained: The first component has the highest eigenvalue and thus explains the largest portion of the total variance. The second and third components also contribute significantly, with each of them having an eigenvalue greater than 1, indicating that they still explain a meaningful portion of the variance.

Conclusion: Considering the Kaiser criterion (eigenvalues > 1) and the scree plot, 3 components seem to be an appropriate choice for m. There is not a single, perfectly clear answer since the scree plot shows the first component as very dominant, but the next two components also contribute significantly.

Therefore, the clear choice for m would be 3 components, as they all have eigenvalues greater than 1 and contribute meaningfully to explaining the variance.

b.

```
pca_result <- principal(pilot_data, nfactors = 3, rotate = "varimax")

print(pca_result)
```

```
## Principal Components Analysis
## Call: principal(r = pilot_data, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC3   RC1   RC2   h2   u2 com
## V2 -0.06  0.83  0.17 0.73 0.27 1.1
## V3 -0.35  0.10  0.82 0.81 0.19 1.4
## V4  0.72 -0.03  0.07 0.53 0.47 1.0
## V5  0.74  0.30 -0.19 0.67 0.33 1.5
## V6 -0.49 -0.01 -0.73 0.77 0.23 1.7
## V7  0.24  0.80 -0.07 0.70 0.30 1.2
##
##                        RC3  RC1  RC2
## SS loadings           1.49 1.43 1.28
## Proportion Var        0.25 0.24 0.21
## Cumulative Var        0.25 0.49 0.70
## Proportion Explained  0.36 0.34 0.30
## Cumulative Proportion 0.36 0.70 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.14
##  with the empirical chi square  22.6  with prob <  NA
##
## Fit based upon off diagonal values = 0.56
```

c.

```
factor_result <- fa(pilot_data, nfactors = 3, fm = "pa", rotate = "varimax")
```

```
## maximum iteration exceeded
```

```
print(factor_result)
```

```
## Factor Analysis using method =  pa
## Call: fa(r = pilot_data, nfactors = 3, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PA3  PA1   PA2   h2   u2 com
## V2  0.03 0.83  0.10 0.70 0.30 1.0
## V3 -0.38 0.16  0.64 0.58 0.42 1.8
## V4  0.37 0.06  0.04 0.14 0.86 1.1
## V5  0.72 0.14 -0.05 0.54 0.46 1.1
## V6 -0.33 0.00 -0.58 0.44 0.56 1.6
## V7  0.34 0.44  0.02 0.32 0.68 1.9
##
##                        PA3  PA1  PA2
## SS loadings           1.02 0.94 0.76
## Proportion Var        0.17 0.16 0.13
## Cumulative Var        0.17 0.33 0.45
## Proportion Explained  0.38 0.35 0.28
## Cumulative Proportion 0.38 0.72 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 3 factors are sufficient.
##
## df null model =  15  with the objective function =  0.68 with Chi Square =  24.75
## df of  the model are 0  and the objective function was  0.03
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  NA
##
## The harmonic n.obs is  40 with the empirical chi square  0.85  with prob <  NA
## The total n.obs was  40  with Likelihood Chi Square =  0.92  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  -Inf
## Fit based upon off diagonal values = 0.98
## Measures of factor score adequacy
##                                                  PA3  PA1  PA2
## Correlation of (regression) scores with factors   0.81 0.85 0.78
## Multiple R square of scores with factors          0.66 0.72 0.60
## Minimum correlation of possible factor scores     0.31 0.44 0.21
```

d.

Let’s compare the results of parts (b) and (c), (b)PCA Results: We extracted three factors using the principal component method and applied varimax rotation. The results show the loadings of each variable on the different components. Factor Loadings (RC1, RC2, RC3): V2 has a high loading on RC1 (0.83), indicating that it is strongly explained by RC1. V3 has the highest loading on RC2 (0.82), indicating a strong correlation with RC2. V4 and V5 have high loadings on RC3, with values of 0.72 and 0.74, respectively. V7 has substantial loadings on both RC1 (0.80) and RC3 (0.24). Communality (h2): Represents the amount of variance in each variable explained by the extracted factors. For example, V2 has a communality of 0.729, meaning about 72.9% of its variance is explained by the extracted factors. Uniqueness (u2): Represents the proportion of variance not explained by the factors. For example, V4 has a uniqueness of 0.471, meaning about 47.1% of its variance is not explained.

c. PAF Results:

We extracted three factors using the principal axis factoring method and applied varimax rotation. Factor Loadings (PA1, PA2, PA3): V2 has the highest loading on PA1 (0.83), which is consistent with the PCA result, indicating a similar explanatory power for this variable. V3 has a loading on PA2 (0.64), which is slightly lower

compared to its loading on RC2 in PCA. V4 and V5 have high loadings on PA3, with values of 0.37 and 0.72, respectively, similar to PCA, but with slightly different magnitudes. V6 has a loading of -0.58 on PA2, which is different from the PCA result. Communality (h2): In PAF, V2 has a communality of 0.704, which is slightly lower than in PCA, indicating that less variance is explained by the factors in PAF. V4 has a communality of 0.139 in PAF, showing that the extracted factors explain very little of V4's variance compared to a higher communality in PCA. Uniqueness (u2): In PAF, the uniqueness for some variables is higher than in PCA, such as V4 with a uniqueness of 0.861, meaning most of the variance is not explained.

Comparison and Analysis:

Differences in Factor Loadings: PCA and PAF show some similarities in the factor loading matrix, such as V2 having a high loading on the first factor in both methods, but there are differences in the loadings for other variables. PAF tends to emphasize extracting variance in a different way, resulting in noticeably lower loading values for some variables compared to PCA, especially for V4 and V6.

Communality and Uniqueness: In PCA, communalities are typically higher because the principal components are directly extracted from the total variance. In PAF, since factors are extracted incrementally while reducing error variance, some variables, like V4, have much lower communalities, indicating weaker representation in the factor structure. Uniqueness tends to be higher in PAF for certain variables, such as V4 and V6, indicating that a large portion of their variance is not explained by the extracted factors.

Conclusion: PCA focuses on maximizing the total variance in the data, resulting in a more even distribution of explanatory power across the variables. It is often used to determine which factors explain the most total variance. PAF attempts to explain the common variance among variables while ignoring unique variance, making it more suitable for exploring the underlying factor structure.

Problem 2

Compute the orthogonal factor model for:

Define the covariance matrix

```
Sigma <- matrix(c(1, 0.9, 0.7,
                  0.9, 1, 0.4,
                  0.7, 0.4, 1), nrow = 3, byrow = TRUE)

eigen_result <- eigen(Sigma)
```

Extract eigenvalues and eigenvectors

```
eigenvalues <- eigen_result$values
eigenvectors <- eigen_result$vectors

print("Eigenvalues:")
```

```
## [1] "Eigenvalues:"
```

```
print(eigenvalues)
```

```
## [1] 2.35363603 0.61601660 0.03034736
```

```
print("Eigenvectors:")
```

```
## [1] "Eigenvectors:"
```

```
print(eigenvectors)
```

```
##            [,1]       [,2]       [,3]
## [1,] -0.6436239 -0.1110798  0.7572381
## [2,] -0.5766348 -0.5801800 -0.5752248
## [3,] -0.5032303  0.8068782 -0.3093652
```

Compute the factor loadings matrix

```
loadings <- eigenvectors %*% diag(sqrt(eigenvalues))

print("Factor Loadings Matrix:")
```

```
## [1] "Factor Loadings Matrix:"
```

```
print(loadings)
```

```
##            [,1]       [,2]        [,3]
## [1,] -0.9874197 -0.0871829  0.13191463
## [2,] -0.8846479 -0.4553643 -0.10020701
## [3,] -0.7720339  0.6332923 -0.05389295
```

Problem 3

A 12-year-old girl made five ratings on a nine-point semantic differential scale for each of seven of her acquaintances. The ratings were based on the five adjectives “kind,” “intelligent,” “happy,” “likeable,” and “just.” Her ratings are given in the lecture slides. The correlation matrix for the five variables (adjectives) is as follows,

$$\mathbf{R} = \begin{bmatrix} 1.000 & .296 & .881 & .995 & .545 \\ .296 & 1.000 & -.022 & .326 & .837 \\ .881 & -.022 & 1.000 & .867 & .130 \\ .995 & .326 & .867 & 1.000 & .544 \\ .545 & .837 & .130 & .544 & 1.000 \end{bmatrix}$$

(a) Compute factor loadings and communalities by the principal component method. (b) Use the hypothesis test to choose m factors for the data. (Hint: Please solve (a) and (b) step by step.)

a. using the given correlation matrix and calculate the factor loadings and communalities.

```
R <- matrix(c(1.000, 0.296, 0.881, 0.995, 0.545,
              0.296, 1.000, -0.022, 0.326, 0.837,
              0.881, -0.022, 1.000, 0.867, 0.130,
              0.995, 0.326, 0.867, 1.000, 0.544,
              0.545, 0.837, 0.130, 0.544, 1.000), nrow = 5, byrow = TRUE)

r_pca_result <- principal(R, nfactors = 2, rotate = "none")
```

```
print("Factor Loadings:")
```

```
## [1] "Factor Loadings:"
```

```
print(r_pca_result$loadings)
```

```
##
## Loadings:
##      PC1    PC2
## [1,]  0.970 -0.231
## [2,]  0.519  0.807
## [3,]  0.785 -0.588
## [4,]  0.971 -0.210
## [5,]  0.704  0.667
##
##                  PC1   PC2
## SS loadings    3.263 1.538
## Proportion Var 0.653 0.308
## Cumulative Var 0.653 0.960
```

```
print("Communalities (h2):")
```

```
## [1] "Communalities (h2):"
```

```
print(r_pca_result$communality)
```

```
## [1] 0.9933223 0.9207769 0.9608635 0.9865106 0.9402854
```

b. To determine the appropriate number of factors, we can use Bartlett's test of sphericity or parallel analysis. These methods help us decide the number of factors to retain:

Parallel Analysis: Compares the eigenvalues from the correlation matrix to those from randomly generated matrices of the same size. Factors with eigenvalues greater than those from the random data are retained.

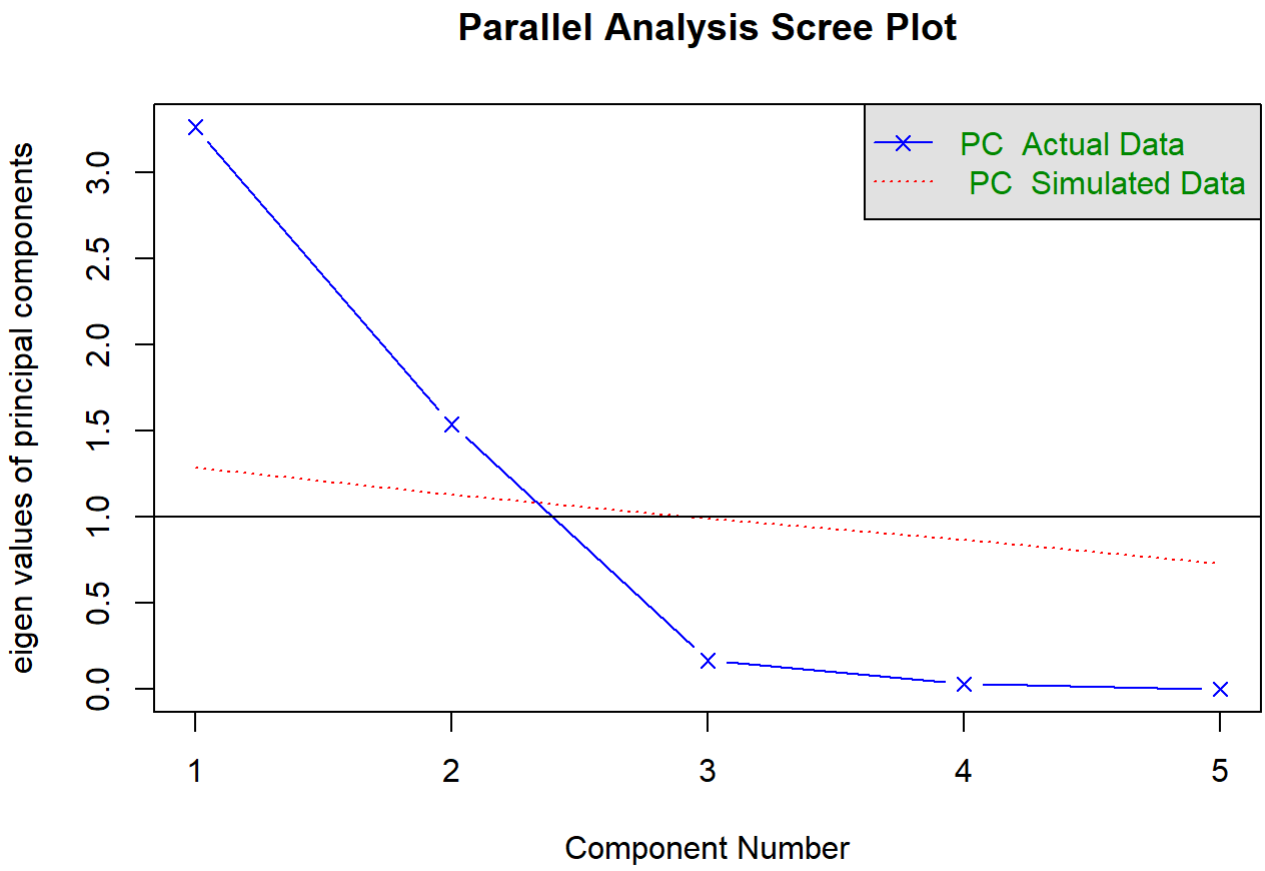
```
fa.parallel(R, fa = "pc", n.iter = 100, show.legend = TRUE, main = "Parallel Analysis Scree P
lot")
```

```
## Warning in fa.parallel(R, fa = "pc", n.iter = 100, show.legend = TRUE, main =
## "Parallel Analysis Scree Plot"): It seems as if you are using a correlation
## matrix, but have not specified the number of cases. The number of subjects is
## arbitrarily set to be 100
```

```
## Warning in cor.smooth(model): Matrix was not positive definite, smoothing was
## done
```

```
## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, :
## The estimated weights for the factor scores are probably incorrect.  Try a
## different factor score estimation method.
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate = rotate, : An
## ultra-Heywood case was detected.  Examine the results carefully
```




```
## Parallel analysis suggests that the number of factors =  NA  and the number of components
=  2
```

Based on the parallel analysis, we should choose m=2 factors.

Bartlett's Test of Sphericity: Tests whether the correlation matrix is significantly different from the identity matrix. A significant result suggests that factor analysis is appropriate.

```
bartlett_result <- cortest.bartlett(R, n = 7)  # since # of her acquaintances are 7
print(bartlett_result)
```

```
## $chisq
## [1] 42.00663
##
## $p.value
## [1] 7.478341e-06
##
## $df
## [1] 10
```

Since the p-value is significantly less than 0.05, we reject the null hypothesis that the correlation matrix is an identity matrix. This result indicates that there is significant shared variance among the variables, and factor analysis is appropriate for the data.

However, Bartlett’s Test of Sphericity is used to assess whether factor analysis is appropriate, but it does not directly determine the number of factors (m) to retain. But from the Parallel Analysis Scree Plot that you provided earlier, we observed that 2 components have eigenvalues greater than the simulated data.

Based on the Parallel Analysis, the optimal number of factors to retain is m = 2. Bartlett’s Test of Sphericity indicates that factor analysis is suitable, and parallel analysis suggests that 2 factors are appropriate.

Therefore, the final answer is m = 2.