MAST90138 Assignment 2

Haonan Zhong 867492

Question 1a

```
wheat <- read.csv("Wheat data.txt", sep = "", header = F)
X <- scale(wheat[, -c(8)], scale = FALSE)
PCX <- prcomp(X, retx = TRUE)
(lambda <- PCX$sdev^2)
```

```
## [1] 1.079333e+01 2.129455e+00 7.363003e-02 1.288749e-02 2.748227e-03
## [6] 1.570450e-03 2.965544e-05
```

```
(gamma <- PCX$rotation)
```

```
##             PC1          PC2         PC3         PC4          PC5          PC6
## V1 -0.884228505  0.100805775  0.26453354  0.19944949 -0.137172970  0.280639558
## V2 -0.395405417  0.056489625 -0.28251995 -0.57881686  0.574756029 -0.301558638
## V3 -0.004311324 -0.002894744  0.05903584  0.05776023 -0.053104536 -0.045229054
## V4 -0.128544478  0.030621731 -0.40014946 -0.43610024 -0.786997760 -0.113437606
## V5 -0.111059139  0.002372293  0.31923869  0.23416358 -0.144802899 -0.896267845
## V6  0.127615624  0.989410476  0.06429754 -0.02514736 -0.001575639  0.003287998
## V7 -0.128966499  0.082233392 -0.76193973  0.61335659  0.087653609 -0.109923643
##             PC7
## V1 -0.025398239
## V2  0.065839904
## V3  0.994125646
## V4  0.001431435
## V5 -0.081549900
## V6  0.001142692
## V7  0.008971926
```

```
# Percentage of the variability of the data explained by each component
(fracvar <- lambda/sum(lambda))
```

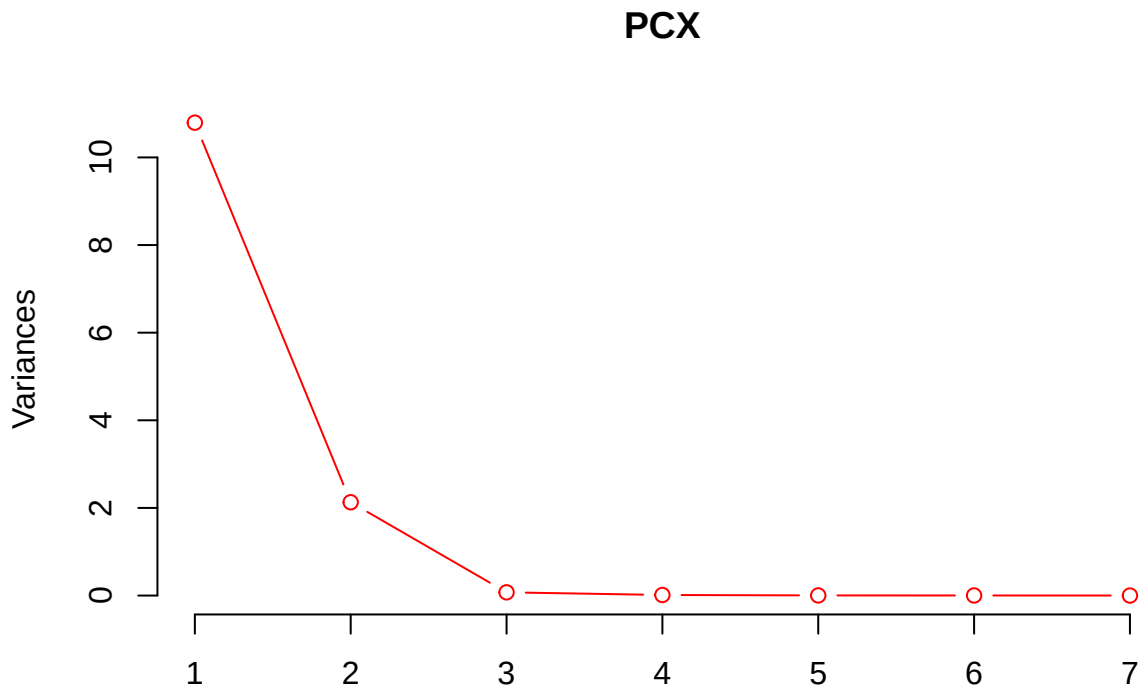
```
## [1] 8.293852e-01 1.636325e-01 5.657909e-03 9.903061e-04 2.111803e-04
## [6] 1.206771e-04 2.278796e-06
```

As we can see from the output, the first and the second principle components was able to explain about 82.9% and 16.4% of the variability of the data, respectively. And the third component explains around 0.57% of variances. Whilst, the rest of the PCs only explains a tiny portion of variances.

```r
# Cumulative sum of explained variance
(cumuprop <- cumsum(lambda)/sum(lambda))
```

```
## [1] 0.8293852 0.9930176 0.9986756 0.9996659 0.9998770 0.9999977 1.0000000
```

```r
screeplot(PCX, type = "line", col = "red")
```



Visually, one can look for an elbow in the screeplot and stop there. In our case, we should keep the first three principle components, with the first three PCs we explain almost 100% of the variability of the data.

Question 1b

According to the eigenvectors given in the question 1a, we have

$$Y_1 = -0.884X_1 - 0.395X_2 - 0.004X_3 - 0.129X_4 - 0.111X_5 + 0.127X_6 - 0.129X_7$$

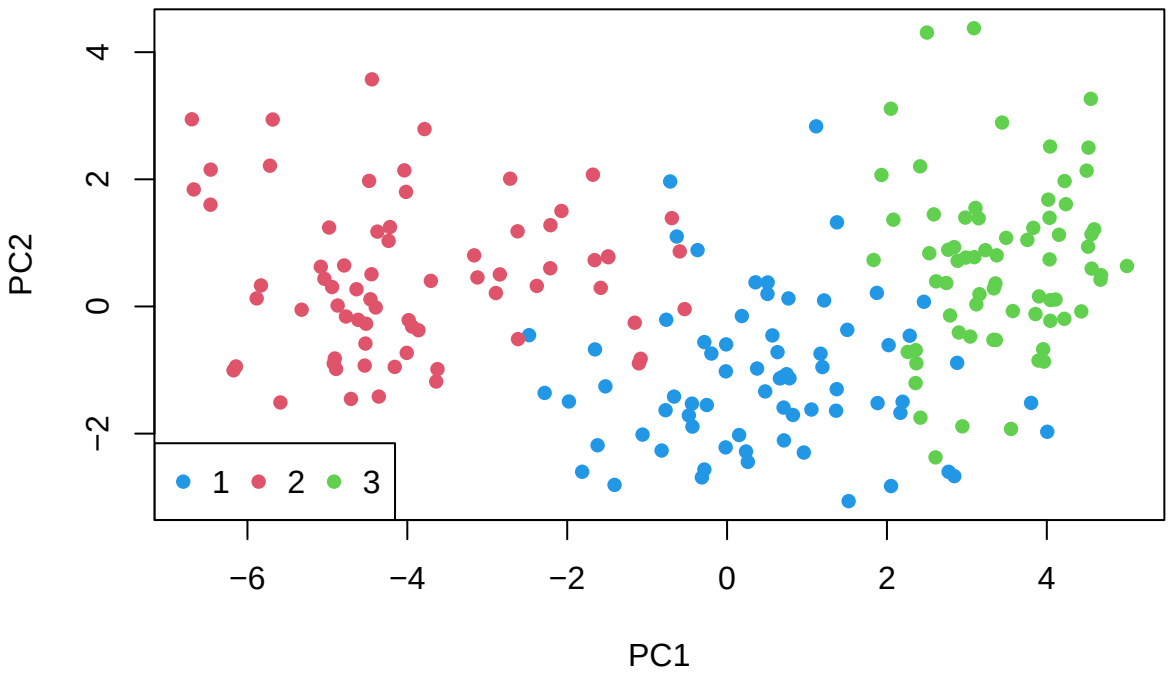
Thus, the first PC puts the most weight on the first variable X_1, which has an negative effect on PC1, followed by X_2 with a positive effect. On the other hand, X_4, X_5, X_6, and X_7 are quite similar in terms of contribution, whilst X_3 plays the smallest role in the construction of the first PC.

$$Y_2 = 0.101X_1 + 0.056X_2 - 0.003X_3 + 0.031X_4 + 0.002X_5 + 0.989X_6 + 0.082X_7$$

Thus, X_6 plays the major role in the construction of the second PC with a positive effect, followed by X_1.

Question 1c

```
plot(PCX$x[,1], PCX$x[,2], col =c(4, 2, 3)[wheat$V8], pch=16, xlab = 'PC1', ylab = 'PC2')
legend(x = "bottomleft", horiz = TRUE, legend = unique(wheat$V8), col=c(4, 2, 3), pch=16)
```



Based on the scatter plot between the first two principle components, we can see that PC1 captured most of the variation driven by the different varieties of wheat, as it divides the data points into three distinct clusters, where blue points indicates data point from group 1, red points indicates data point from group 2, and green points represents data point from group 3. On the other hand, no clear group separation can be seen from PC2, other than group 1 seems to have a low value of PC2 compare to the other two groups.

Question 1d

```
corr_matrix <- matrix(0, 7, 7)

for (j in 1:7) {
  for (k in 1:7) {
    corr_matrix[j, k] <- (gamma[j, k] * lambda[k]) / sqrt(cov(X)[j, j] * lambda[k])
  }
}
corr_matrix[, c(1, 2)]
```

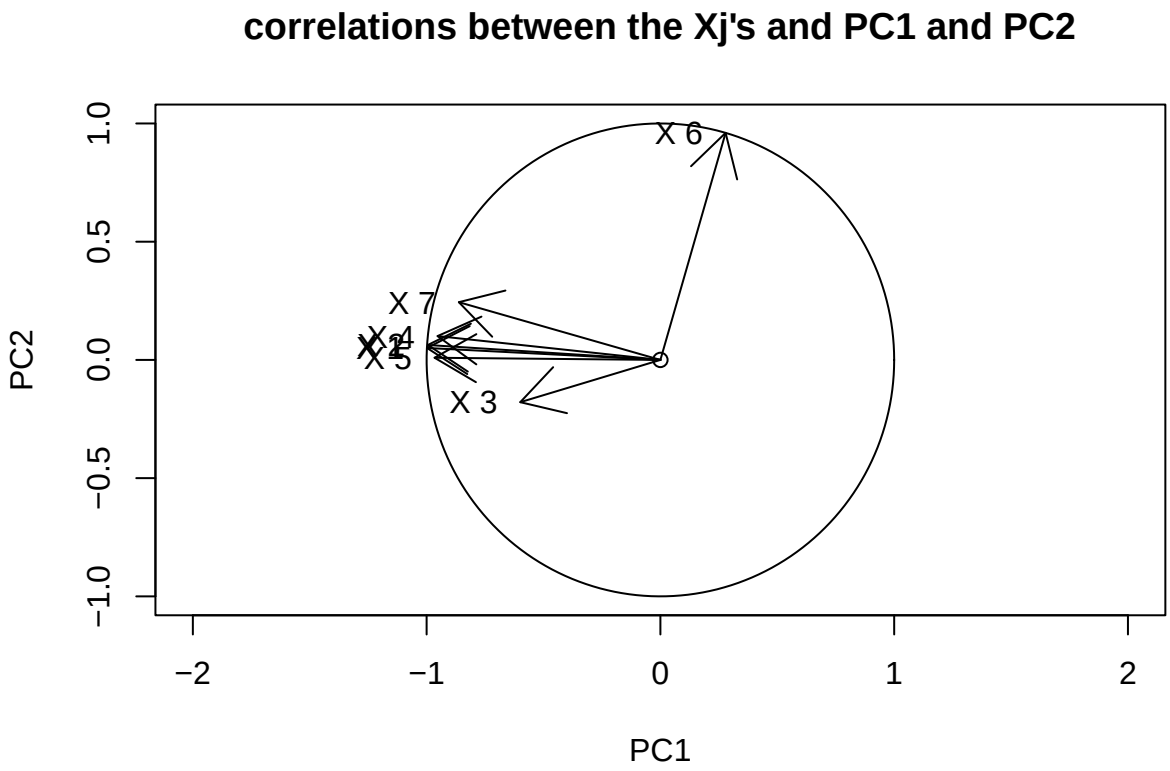
```
##            [,1]         [,2]
```

```
## [1,] -0.9983758  0.050555863
## [2,] -0.9946970  0.063120944
## [3,] -0.5994257 -0.178768653
## [4,] -0.9531585  0.100855139
## [5,] -0.9659805  0.009165136
## [6,]  0.2788442  0.960264370
## [7,] -0.8620814  0.244160925
```

```
corr_matrix[,1]^2 + corr_matrix[,2]^2
```

```
## [1] 0.9993101 0.9934065 0.3912694 0.9186829 0.9332024 0.9998617 0.8027989
```

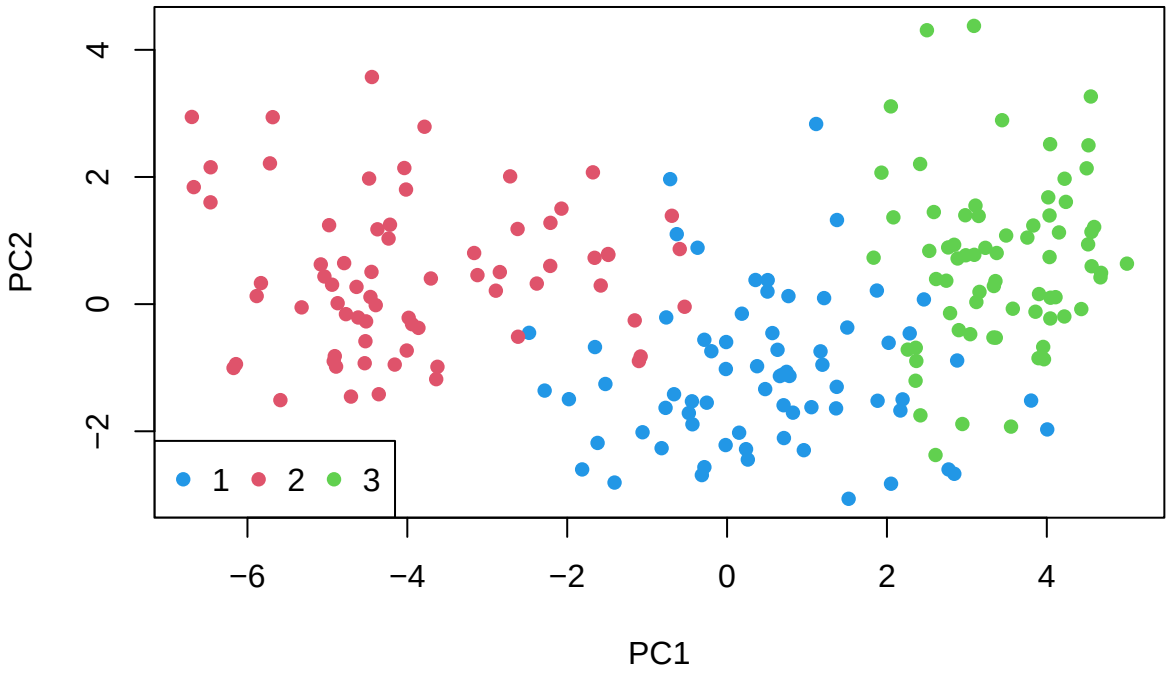
```
plot(x = 0, y = 0, xlim = c(-2, 2), ylim = c(-1, 1), xlab = 'PC1', ylab = 'PC2',
     main = "correlations between the Xj's and PC1 and PC2")
for (i in 1:7) {
  arrows(0, 0, corr_matrix[i, 1], corr_matrix[i, 2])
  text(corr_matrix[i, 1] - 0.2,
       corr_matrix[i, 2],
       paste("X", as.character(i)))
}

radius <- 1
theta <- seq(0, 2 * pi, length = 200)
lines(x = radius * cos(theta), y = radius * sin(theta))
```



As the plot above depicted, most of the variables, except for $X3$, are relatively close to the periphery of

the circle, which indicates that they are strongly correlated with the first two PCs. Furthermore, PC1 is strongly negatively correlated with $X1$, $X2$, $X4$, $X5$ and $X7$. Whilst PC2 is highly positively correlated with $X6$.

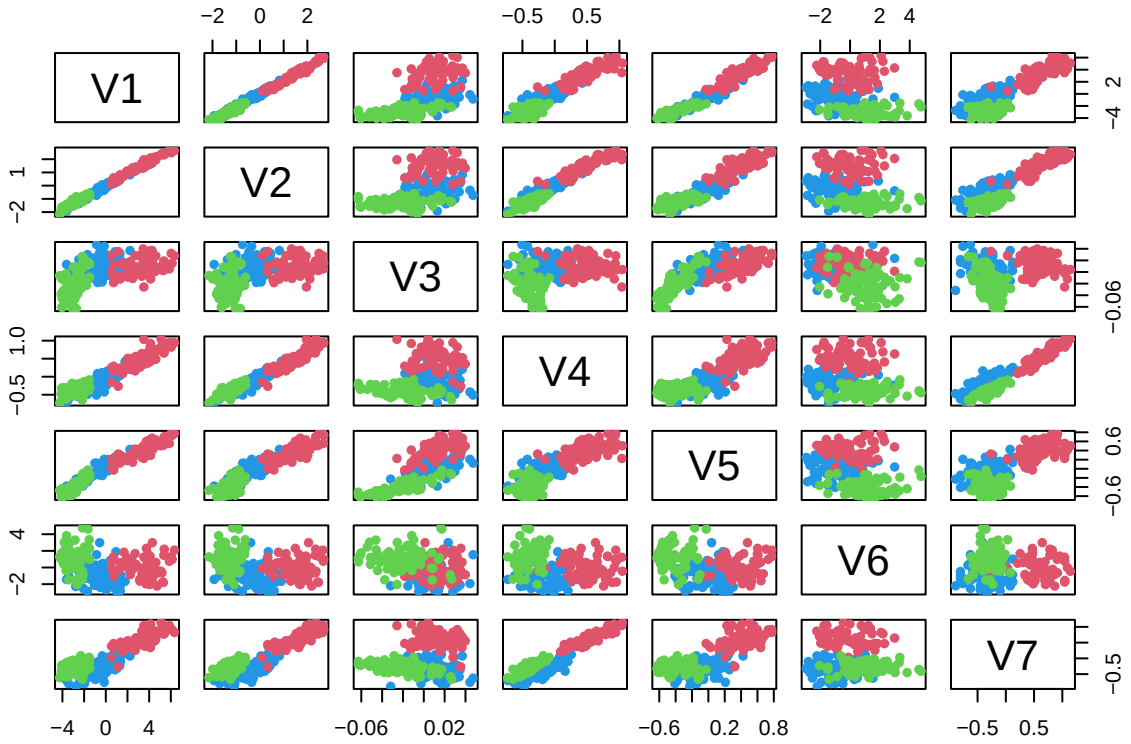
```
plot(PCX$x[,1], PCX$x[,2], col =c(4, 2, 3)[wheat$V8], pch=16, xlab = 'PC1', ylab = 'PC2')
legend(x = "bottomleft", horiz = TRUE, legend = unique(wheat$V8), col=c(4, 2, 3), pch=16)
```



We also know that together the first two PCs explains a large portion of the variability of the data, therefore we could also use the direction of the arrow in conjunction with the scatter plot of the first two PCs to learn the effect of these variables on individuals. In particular, group 2 tends to have negative values on PC1, while group 1 tend to have a value centering around 0 and group 3 tend to have large positive values on PC1. Contrarily, group 3 tend to have slightly larger values on PC2, followed by group 2, then group 1.

As we mentioned above, PC1 is strongly negatively correlated with $X1$, $X2$, $X4$, $X5$ and $X7$.Therefore, when a group has a low value of PC1, we expect it tend to go together with large values of $X1$, $X2$, $X4$, $X5$ and $X7$, and vice versa. In addition, given that PC2 is highly positively correlated with $X6$, we expect groups with large value on PC2 will also have a large value on $X6$. We can see that these are indeed the case in the scatter plot below.

```
pairs(X, col = c(4,2,3)[wheat$V8], pch=16)
```



Question 2a

Given that we are computing the orthogonal factor model with single factor $k = 1$ for Σ. Therefore, we have $\Sigma = QQ^T + \Psi$, where

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \; and \; \Psi = \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix} = \begin{bmatrix} q_1^2 + \psi_{11} & q_1 q_2 & q_1 q_3 \\ q_1 q_2 & q_2^2 + \psi_{22} & q_2 q_3 \\ q_1 q_3 & q_2 q_3 & q_3^2 + \psi_{33} \end{bmatrix}$$

Thus, we will have

$$q_1 = \sqrt{\frac{0.9 \times 0.7}{0.4}} = 1.25 \;\; q_2 = \sqrt{\frac{0.9 \times 0.4}{0.7}} = 0.72 \;\; q_3 = \sqrt{\frac{0.7 \times 0.4}{0.9}} = 0.56$$

$$\psi_{11} = 1 - 1.25^2 = -0.57 \;\; \psi_{22} = 1 - 0.72^2 = 0.49 \;\; \psi_{33} = 1 - 0.56^2 = 0.69$$

Therefore, we have the following solution with,

$$Q = \begin{bmatrix} 1.25 \\ 0.72 \\ 0.56 \end{bmatrix} \; and \; \Psi = \begin{bmatrix} -0.57 & 0 & 0 \\ 0 & 0.49 & 0 \\ 0 & 0 & 0.69 \end{bmatrix}$$

The problem we have is that the ψ_{11} is negative, and since Ψ is the variance of the specific factors U and cannot be negative, thus, this solution cannot be interpret as a factor analysis model.

Question 2bi

```
data("Harman23.cor")
Harman23.cor$n.obs
```

```
## [1] 305
```

```
dim(Harman23.cor$cov)
```

```
## [1] 8 8
```

Given that factor modeling is meant for dimension reduction. Therefore, to avoid over-parametrization, we generally requires

$$p(p+1)/2 \geqslant pq + p - q(q-1)/2$$

Therefore, the maximum number of factors we can fit is $p(p+1)/2$. In the case here, we have $p = 8$, hence

$$8(8+1)/2 \geqslant 8q + 8 - q(q-1)/2$$

$$28 \geqslant 8q - q(q-1)/2$$

$$56 \geqslant 17q - q^2$$

$$56 - 17q + q^2 \geqslant 0$$

Solving the above equation we have $q \leqslant 4.46887$ *or* $q \geqslant 12.53112$, thus, the maximum number of factor is 4.

Question 2bii

```
for (i in 1:4) {
  print(factanal(factors = i, covmat = Harman23.cor))
}
```

```
##
## Call:
## factanal(factors = i, covmat = Harman23.cor)
##
## Uniquenesses:
##         height       arm.span        forearm      lower.leg         weight
##          0.158          0.135          0.190          0.187          0.760
## bitro.diameter    chest.girth    chest.width
##          0.829          0.877          0.801
##
## Loadings:
##                Factor1
## height         0.918
## arm.span       0.930
## forearm        0.900
## lower.leg      0.902
## weight         0.490
## bitro.diameter 0.413
## chest.girth    0.351
```

```
## chest.width    0.446
##
##                Factor1
## SS loadings      4.064
## Proportion Var   0.508
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 611.44 on 20 degrees of freedom.
## The p-value is 1.12e-116
##
## Call:
## factanal(factors = i, covmat = Harman23.cor)
##
## Uniquenesses:
##         height       arm.span        forearm      lower.leg         weight
##          0.170          0.107          0.166          0.199          0.089
## bitro.diameter    chest.girth    chest.width
##          0.364          0.416          0.537
##
## Loadings:
##                Factor1 Factor2
## height         0.865   0.287
## arm.span       0.927   0.181
## forearm        0.895   0.179
## lower.leg      0.859   0.252
## weight         0.233   0.925
## bitro.diameter 0.194   0.774
## chest.girth    0.134   0.752
## chest.width    0.278   0.621
##
##                Factor1 Factor2
## SS loadings      3.335   2.617
## Proportion Var   0.417   0.327
## Cumulative Var   0.417   0.744
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 75.74 on 13 degrees of freedom.
## The p-value is 6.94e-11
##
## Call:
## factanal(factors = i, covmat = Harman23.cor)
##
## Uniquenesses:
##         height       arm.span        forearm      lower.leg         weight
##          0.127          0.005          0.193          0.157          0.090
## bitro.diameter    chest.girth    chest.width
##          0.359          0.411          0.490
##
## Loadings:
##                Factor1 Factor2 Factor3
## height          0.886   0.267  -0.130
## arm.span        0.937   0.195   0.280
## forearm         0.874   0.188
## lower.leg       0.877   0.230  -0.145
```

```
## weight          0.242   0.916  -0.106
## bitro.diameter  0.193   0.777
## chest.girth     0.137   0.755
## chest.width     0.261   0.646   0.159
##
##                Factor1 Factor2 Factor3
## SS loadings      3.379   2.628   0.162
## Proportion Var   0.422   0.329   0.020
## Cumulative Var   0.422   0.751   0.771
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 22.81 on 7 degrees of freedom.
## The p-value is 0.00184
##
## Call:
## factanal(factors = i, covmat = Harman23.cor)
##
## Uniquenesses:
##         height       arm.span        forearm      lower.leg         weight
##          0.137          0.005          0.191          0.116          0.138
## bitro.diameter    chest.girth    chest.width
##          0.283          0.178          0.488
##
## Loadings:
##                Factor1 Factor2 Factor3 Factor4
## height          0.879   0.277          -0.115
## arm.span        0.937   0.194           0.277
## forearm         0.875   0.191
## lower.leg       0.887   0.209   0.135  -0.188
## weight          0.246   0.882   0.111  -0.109
## bitro.diameter  0.187   0.822
## chest.girth     0.117   0.729   0.526
## chest.width     0.263   0.644           0.141
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      3.382   2.595   0.323   0.165
## Proportion Var   0.423   0.324   0.040   0.021
## Cumulative Var   0.423   0.747   0.787   0.808
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 4.63 on 2 degrees of freedom.
## The p-value is 0.0988
```

By testing the null hypothesis: q is the number of factors. For each allowed q, we can see that the p-value returned for $q = 1$, $q = 2$, $q = 3$, and $q = 4$ is 1.12e-116, 6.94e-11, 0.00184 and 0.0988, respectively. Among them, only $q = 4$ has a p-value greater than the significance level of 0.05, hence we do not reject the null hypothesis and the factor model with $q = 4$ has the best fit.