MAST90138_A2

QINGTAN SHEN (1130945)

06/09/2022

Question 1

(a)

Firsly, we load the wheat dataset:

```
wheat_data_2 = read.table(file='/Users/apple/Desktop/MAST90138/wheat_data_2.txt', sep = "", header=F)
```

Then we will get gamma and lambda which are defined in the question, and get the summary of the principle components analysis:

```
PCX_wheat = prcomp(wheat_data_2[,1:7])
gamma = PCX_wheat$rotation
lambda =PCX_wheat$sdev^2
summary(PCX_wheat)
```

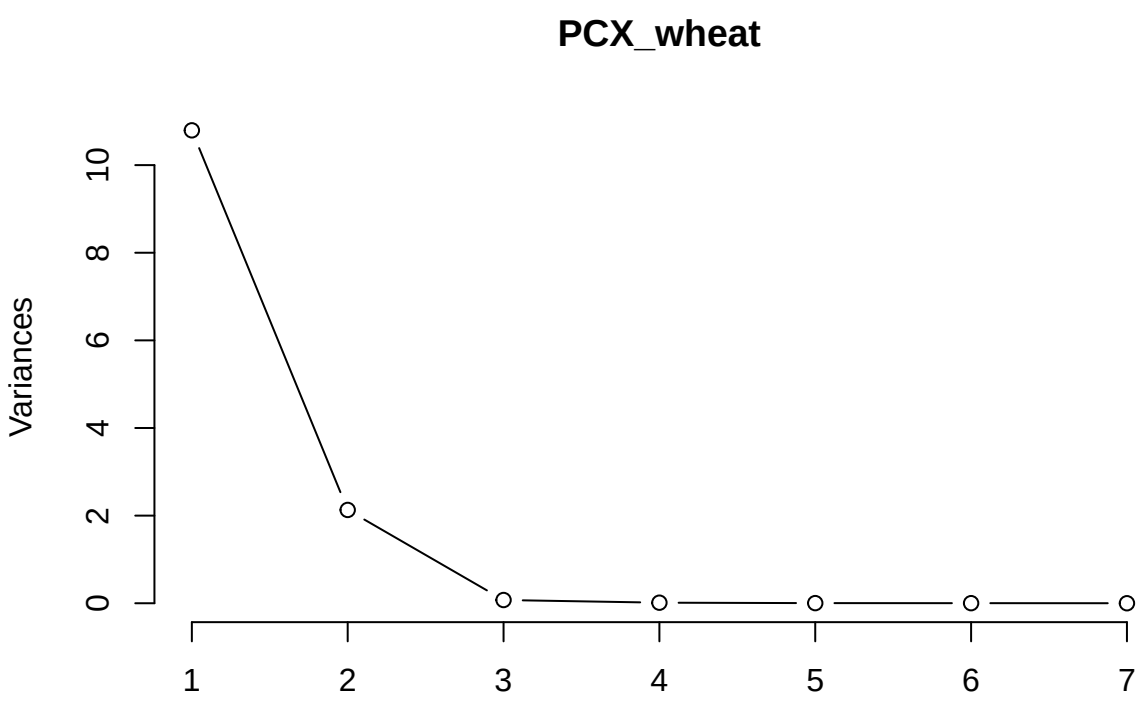
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6      PC7
## Standard deviation     3.2853 1.4593 0.27135 0.11352 0.05242 0.03963 0.005446
## Proportion of Variance 0.8294 0.1636 0.00566 0.00099 0.00021 0.00012 0.000000
## Cumulative Proportion  0.8294 0.9930 0.99868 0.99967 0.99988 1.00000 1.000000
```

For the percentage each principal component explain, we can see that PC1 explains 82.94% variability of the data, PC2 explains 16.36% variability, PC3 explains 0.566%, PC4 explains 0.099%, PC5 explains 0.021%, PC6 explains 0.012% and PC7 explains 0.000% of the variability.

For the cumulative percentage of the principle components, we can see from the third line of the table, for PC1 to PC7, the cumulative percentage we get is 82.94%, 99.3%, 99.868%, 99.967%, 99.988%, 100%, 100% respectively.

Then I will draw the screeplot:

```
screeplot(PCX_wheat, type = 'l')
```



From this plot, we can see that the line decreases dramatically from {lambda}_1 to {lambda}_3, and the line is more smoothly after {lambda_3}, which means the first two principle components contain most information of this dataset, also we can see the cumulative proportion for the second component is 99.3%. This percentage is large enough, so we should keep the first two principle components.

(b)

Here is the eigenvectors:

```
gamma
```

```
##             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
```

For PC1, we get the formula for different samples:

$$Y1_i = -0.8842 * V1_i - 0.3954 * V2_i - 0.0043 * V3_i - 0.1285 * V4_i - 0.1111 * V5_i + 0.1276 * V6_i - 0.1290 * V7_i$$

From the formula above, V1 has the largest percentage in PC1, which has the negative impact on PC1.

For PC2, we get the formula for different samples:

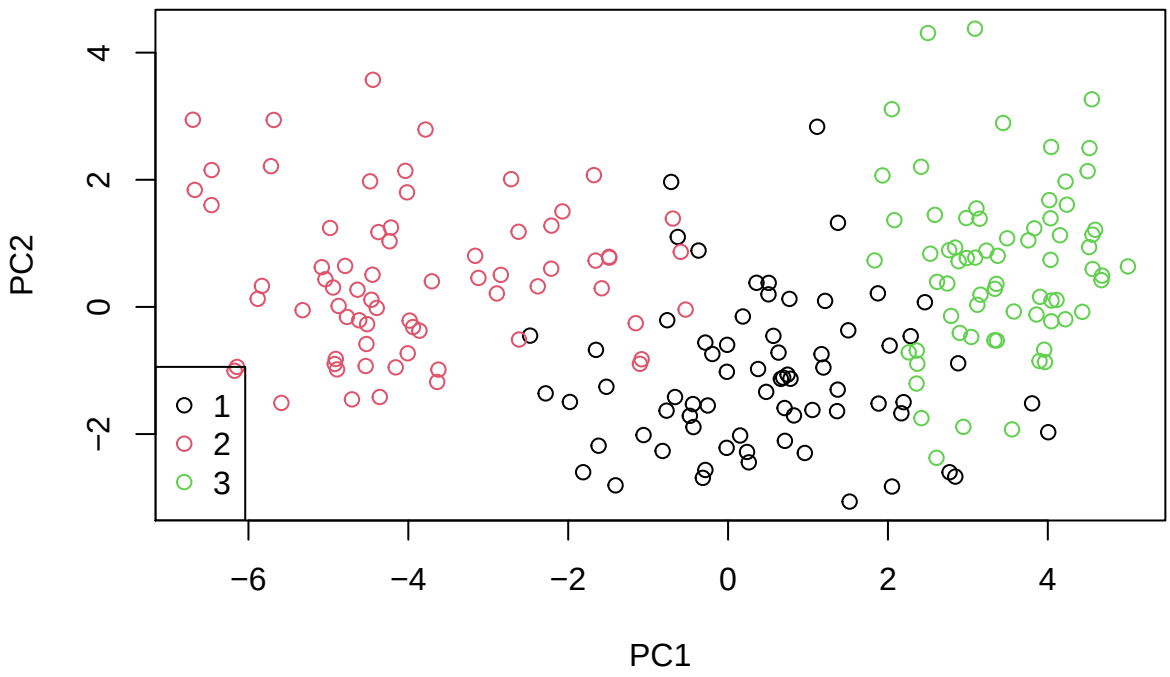
$$Y2_i = 0.1008 * V1_i + 0.0565 * V2_i - 0.0029 * V3_i + 0.0306 * V4_i + 0.0024 * V5_i + 0.9894 * V6_i + 0.0822 * V7_i$$

From the formula above, V6 has the largest percentage in PC1, which has the positive impact on PC1.

(c)

We will draw a scatter plot with different colors:

```
VA <- PCX_wheat$x
plot(VA[,1], VA[,2], col=c(1,2,3)[wheat_data_2$V8], xlab='PC1', ylab='PC2')
legend('bottomleft', legend=c(1,2,3), col=c(1,2,3), pch=c(1,1,1))
```



From this scatter plot, we can see that the black points in the middle are group1, the red points in the left is group2, the green points in the right is group3. So all the three groups are visible. We can also know group3 has the largest value in PC1, group2 has the smallest value in PC1. Also group1 has a little smaller value in PC2.

For the PC1 and PC2 comparison, it is much easier for us to identify these three groups when using PC1. While for PC2, it is had to identify these three groups.

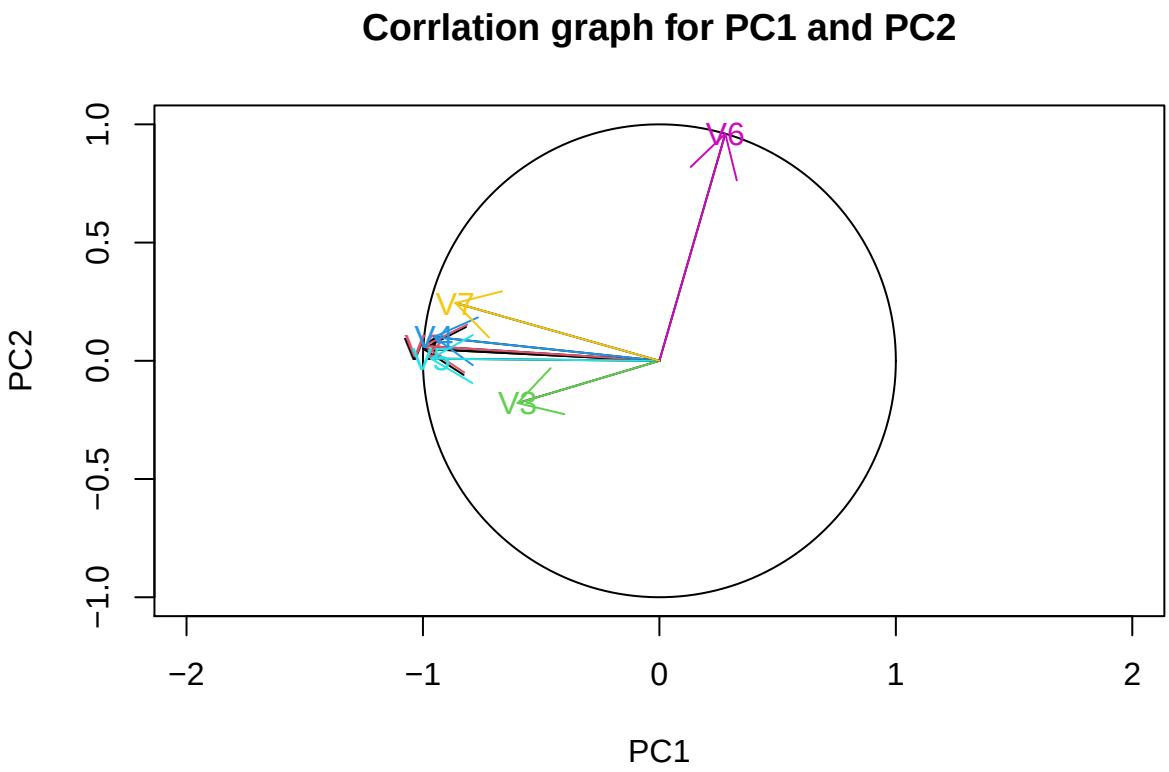
(d)

First, we need to get the correlation table:

```
cor_data <- wheat_data_2[,1:7]
correlation = cor(cor_data, VA)
correlation
```

```
##            PC1          PC2         PC3          PC4            PC5            PC6
## V1 -0.9983758  0.050555863  0.02466949  0.007781603 -2.471421e-03  3.822195e-03
## V2 -0.9946970  0.063120944 -0.05870123 -0.050314826  2.307174e-02 -9.150701e-03
## V3 -0.5994257 -0.178768653  0.67793837  0.277498186 -1.178162e-01 -7.585366e-02
## V4 -0.9531585  0.100855139 -0.24506641 -0.111738960 -9.311804e-02 -1.014619e-02
## V5 -0.9659805  0.009165136  0.22933979  0.070378501 -2.009740e-02 -9.403429e-02
## V6  0.2788442  0.960264370  0.01160385 -0.001898702 -5.493678e-05  8.666098e-05
## V7 -0.8620814  0.244160925 -0.42067027  0.141674274  9.349530e-03 -8.863329e-03
##              PC7
## V1 -4.753436e-05
## V2  2.745440e-04
## V3  2.291081e-01
## V4  1.759373e-05
## V5 -1.175742e-03
## V6  4.138678e-06
## V7  9.941035e-05
```

```
V <- c('V1','V2','V3','V4','V5','V6','V7')
radius <- 1
theta <- seq(0, 2*pi, length = 200)
plot(0, 0, type='l', xlim=c(-1,1), ylim=c(-1,1), xlab='PC1', ylab='PC2', asp=1,  main='Corrlation graph
lines(x = radius * cos(theta), y = radius * sin(theta))
for (i in 1:7){
  lines(c(0, correlation[i,1]), c(0, correlation[i,2]))
  arrows(0, 0, correlation[i,1], correlation[i,2], col=i)
  text(correlation[i,1], correlation[i,2], V[i], col=i)
}
```



From this graph, we can see that most variables are close to the round circle except V3. Also we can see the arrow directions of variables V1,V2,V4,V5 and V7 are all to the left, and they are similar to each other, which means they are negative correlated with PC1. However, V6 has a little positive correlated with PC1, and V6 also has the largest positive correlated with PC2.

When the variables V1, V2, V4, V5, V7 are all small, we can get a large PC1 value. When the variable V6 is large, we can get a large PC2 value.

From the group information we got in the scatter plot, we can get that group3 has smallest values in V1, V2, V4, V5, V7, because PC1 for group3 is large. And group1 has largest values in V1, V2, V4, V5, V7, because PC1 for group1 is small.

Also we can get that group1 has the smallest value of V6, because PC2 for group1 is smaller than other two groups.

Question 2

(a)

We define matrix Q and matrix ϕ as follow:

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}$$

$$\phi = \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix}$$

From the lecture, we know that $Q^TQ + \phi = \sum$, so:

$$\begin{bmatrix} q_1^2 + m_1 & q_1q_2 & q_1q_3 \\ q_1q_2 & q_2^2 + m_2 & q_2q_3 \\ q_1q_3 & q_2q_3 & q_3^2 + m_3 \end{bmatrix} = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

In this matrix equation we have six formulas, which are $q_1^2 + m_1 = 1$, $q_2^2 + m_2 = 1$, $q_3^2 + m_3 = 1$, $q_1q_2 = 0.9$, $q_1q_3 = 0.7$, $q_2q_3 = 0.4$, and we also have six variables:$m_1, m_2, m_3, q_1, q_2, q_3$, so we can get the result of Q and ϕ.

The single factor we get is $Q = \begin{bmatrix} 1.255 \\ 0.717 \\ 0.558 \end{bmatrix}$, the specific variance we get is $\phi = \begin{bmatrix} -0.575 & 0 & 0 \\ 0 & 0.486 & 0 \\ 0 & 0 & 0.689 \end{bmatrix}$

We see that $m_1 = -0.575$, but the variance value should be greater than zero, so the solution is wrong and we cannot use it as our factor analysis model.

(b) i.

Firstly, we load data Harman:

```
data('Harman23.cor')
Harman23.cor
```

```
## $cov
##                height arm.span forearm lower.leg weight bitro.diameter
## height          1.000    0.846   0.805     0.859  0.473          0.398
## arm.span        0.846    1.000   0.881     0.826  0.376          0.326
## forearm         0.805    0.881   1.000     0.801  0.380          0.319
## lower.leg       0.859    0.826   0.801     1.000  0.436          0.329
## weight          0.473    0.376   0.380     0.436  1.000          0.762
## bitro.diameter  0.398    0.326   0.319     0.329  0.762          1.000
## chest.girth     0.301    0.277   0.237     0.327  0.730          0.583
## chest.width     0.382    0.415   0.345     0.365  0.629          0.577
##                chest.girth chest.width
## height               0.301       0.382
## arm.span             0.277       0.415
## forearm              0.237       0.345
## lower.leg            0.327       0.365
## weight               0.730       0.629
## bitro.diameter       0.583       0.577
## chest.girth          1.000       0.539
## chest.width          0.539       1.000
##
## $center
## [1] 0 0 0 0 0 0 0 0
##
## $n.obs
## [1] 305
```

From this table, we can see that it has 8 parameters. So the number of parameters after operation must be smaller than 8. We define the number of parameters after operation is q, the original variable number $p = 8$. So we have this formula:

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

, we take $p = 8$ into this formula, then we get:

$$8 * (8 + 1)/2 > 8 * q + 8 - q * (q - 1)/2$$

, then the result is:

$$q \geq 12.53 or q \leq 4.47$$

, because q must smaller than 8, so we only consider $q \leq 4.47$. So the maximum number is 4.

(b) ii.

Firstly, we will get four results by using factanal method in four conditions (variable number from 1 to 4), and the four results are printed. Also the p-value results are stored in a matrix. Here is the result for factanal method:

```
num <- Harman23.cor$n.obs
coviance <- Harman23.cor$cov
inform_matrix <- matrix(nrow=2, ncol=5)
inform_matrix[1,1] = "factor_num"
inform_matrix[2,1] = "p-value"
for (i in 1:4){
  facta_result = factanal(covmat=coviance,factors=i,n.obs=num)
  print(facta_result)
  inform_matrix[1,i+1] <- i
  inform_matrix[2,i+1] = facta_result$PVAL
}
```

```
##
## Call:
## factanal(factors = i, covmat = coviance, n.obs = num)
##
## 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 = coviance, n.obs = num)
##
## 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 = coviance, n.obs = num)
##
## 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 = coviance, n.obs = num)
##
## 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
```

Here is the result for the matrix with p-value

```
inform_matrix
```

```
##      [,1]         [,2]                    [,3]
## [1,] "factor_num" "1"                     "2"
## [2,] "p-value"    "1.11769132260669e-116" "6.93677512221004e-11"
##      [,4]                 [,5]
## [1,] "3"                  "4"
## [2,] "0.0018401369746635" "0.0987789639529568"
```

In this matrix, it has two lines, the first line is the variable number, the second line is p-value for these four variable numbers. We will consider these four conditions from left to the right. When we have one variable, or two variables, or three variables, the p-value of these three are $1.1177e^{-116}$, $6.9368e^{-11}$, 0.0018 respectively. They are all smaller than 0.05, so we should reject them. For the last one, if we have four parameters, the p-value is 0.0988, which is greater than 0.05, so we keep this one. And q=4 is correct.