MAST90138 Multivariate Statistics for Data Science Assignment2

Yini Lai (Student ID: 1127650)

09/28/2020

```
library(knitr)
library(tidyverse)
library(ggforce)
library(cowplot)
library(broom)
library(latex2exp)
```

Problem 1

```
# Loading data
wheat <- read.table("Wheat data.txt")
```

(a)

```
# Calculating the Principle Components
PCX <- prcomp(wheat[, 1:7],retx=T)

# Eigen vectors
gamma <- PCX$rotation
kable(gamma, caption = "Eigenvectors")
```

Table 1: Eigenvectors

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
V1	-0.8842285	0.1008058	0.2645335	0.1994495	-0.1371730	0.2806396	-0.0253982
V2	-0.3954054	0.0564896	-0.2825200	-0.5788169	0.5747560	-0.3015586	0.0658399
V3	-0.0043113	-0.0028947	0.0590358	0.0577602	-0.0531045	-0.0452291	0.9941256
V4	-0.1285445	0.0306217	-0.4001495	-0.4361002	-0.7869978	-0.1134376	0.0014314
V5	-0.1110591	0.0023723	0.3192387	0.2341636	-0.1448029	-0.8962678	-0.0815499
V6	0.1276156	0.9894105	0.0642975	-0.0251474	-0.0015756	0.0032880	0.0011427
V7	-0.1289665	0.0822334	-0.7619397	0.6133566	0.0876536	-0.1099236	0.0089719

```
# Eigen values
lambda <- PCX$sdev^2
kable(lambda, col.names = NULL, caption = "Eigenvalues")
```

Table 2: Eigenvalues

10.7933269

Table 2: Eigenvalues

2.1294551
0.0736300
0.0128875
0.0027482
0.0015704
0.0000297

```
# Percentage of Variance explained by each PC
(lambda/sum(lambda)) %>% kable(col.names = NULL, caption = "Percentage of Variance")
```

Table 3: Percentage of Variance

0.8293852
0.1636325
0.0056579
0.0009903
0.0002112
0.0001207
0.0000023

```
# Cummulative Percentage explained by the PCs
(cumsum(lambda)/sum(lambda)) %>% kable(col.names = NULL,
                                       caption = "Cummulative Percentage of Variance")
```

Table 4: Cummulative Percentage of Variance

0.8293852
0.9930176
0.9986756
0.9996659
0.9998770
0.9999977
1.0000000

From the table showing above, we can see that

PC1 explains 82.94% of the variability

PC2 explains 16.36% of the variability

PC3 explains 0.57% of the variability

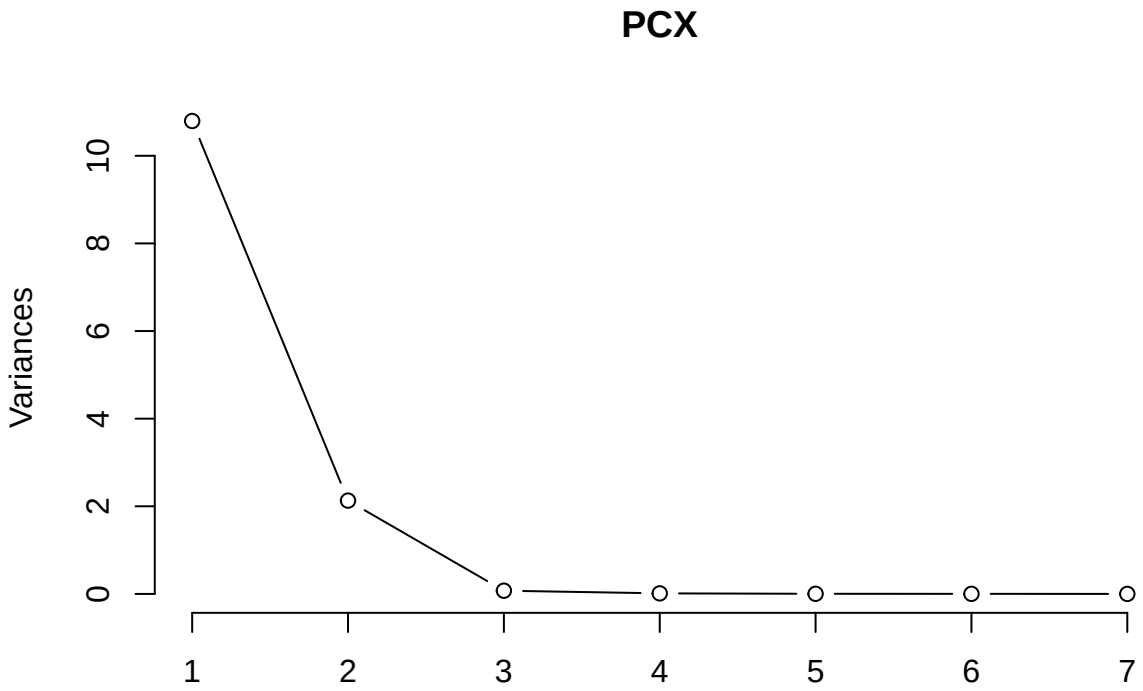
PC4 explains 0.09903% of the variability

PC5 explains 0.02112% of the variability

PC6 explains 0.01207% of the variability

PC7 explains 0.00023% of the variability

```r
screeplot(PCX, type = "lines")
```



According to the screeplot, we should keep 2 components.

(b)

```r
gamma[, 1:2] %>% kable()
```

	PC1	PC2
V1	-0.8842285	0.1008058
V2	-0.3954054	0.0564896
V3	-0.0043113	-0.0028947
V4	-0.1285445	0.0306217
V5	-0.1110591	0.0023723
V6	0.1276156	0.9894105
V7	-0.1289665	0.0822334

$$Y_{i1} = -0.8842V1_i - 0.3954V2_i - 0.0043V3_i - 0.1285V4_i - 0.1110V5_i + 0.1276V6_i - 0.1290V7_i$$

PC1 puts weight -0.8842285, -0.3954054, -0.0043113, -0.1285445, -0.1110591, 0.1276156 -0.1289665 on, respectively, V1(area), V2(perimeter), V3(compactness), V4(length of kernel), V5(width of kernel), V6(asymmetry coefficient) and V7(length of kernel groove).

```r
as.data.frame(abs(gamma[, 1])) %>% arrange(desc(abs(gamma[, 1]))) %>%
  kable(caption = "PC1 by order", col.names = NULL)
```

Table 6: PC1 by order

V1	0.8842285
V2	0.3954054
V7	0.1289665
V4	0.1285445
V6	0.1276156
V5	0.1110591
V3	0.0043113

PC1 puts the most weight on **V1(area)** and also some weight on **V2(perimeter)**; **Except V6(asymmetry coefficient)**, all contribute **negatively** to PC1. PC1 is essentially the difference between V1 and V2.

$$Y_{i2} = 0.1008V1_i + 0.0565V2_i - 0.0029V3_i + 0.0306V4_i + 0.0024V5_i + 0.9894V6_i + 0.0822V7_i$$

PC2 puts weight 0.1008058, 0.0564896, -0.0028947, 0.0306217, 0.0023723, 0.9894105 0.0822334 on, respectively, V1(area), V2(perimeter), V3(compactness), V4(length of kernel), V5(width of kernel), V6(asymmetry coefficient) and V7(length of kernel groove).

```r
as.data.frame(abs(gamma[, 2])) %>% arrange(desc(abs(gamma[, 2]))) %>%
  kable(caption = "PC2 by order", col.names = NULL)
```

Table 7: PC2 by order

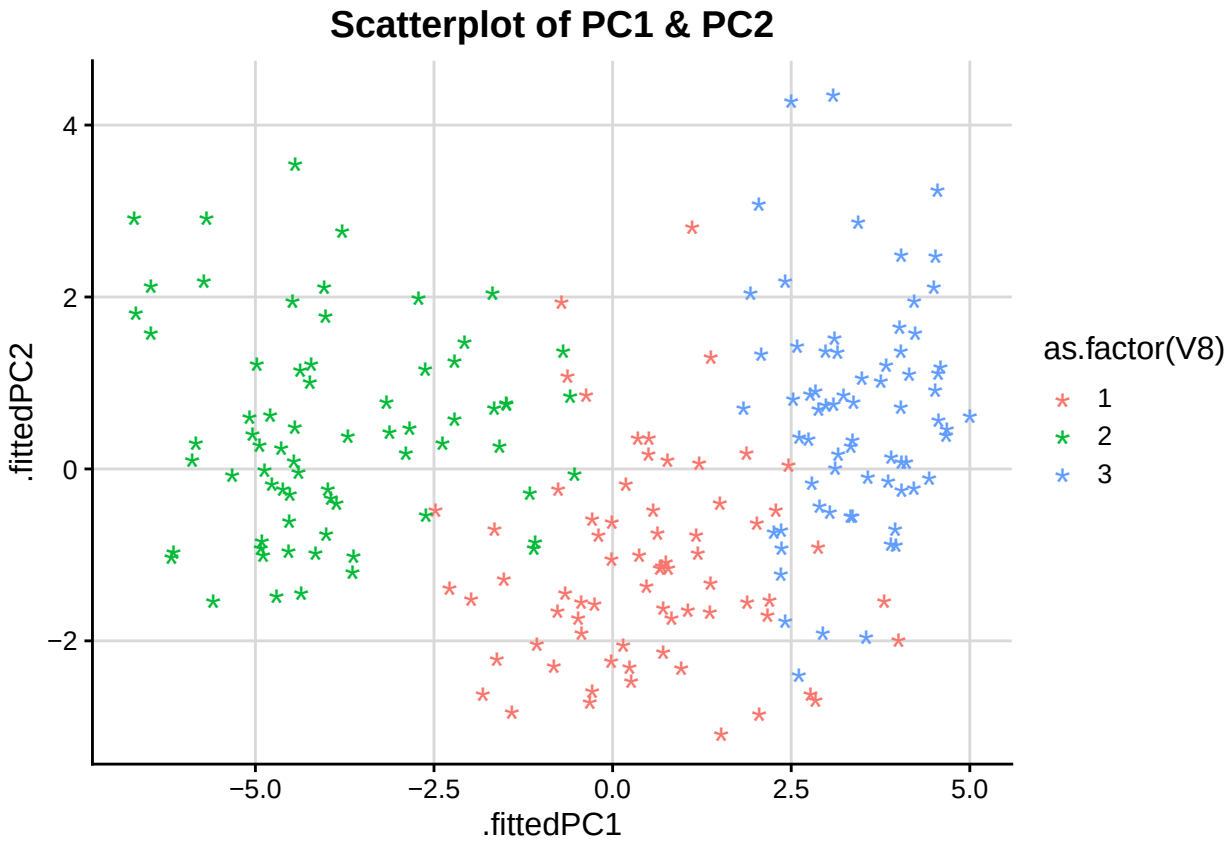
V6	0.9894105
V1	0.1008058
V7	0.0822334
V2	0.0564896
V4	0.0306217
V3	0.0028947
V5	0.0023723

PC2 puts the most weight on **V6(asymmetry coefficient)** and also some weight on **V1(area)**; **Except V3(compactness)**, all contribute **positively** to PC2. PC2 is essentially the difference of V6.

(c)

```r
#Y <- PCX$x
#plot(Y[,1], Y[,2], col=wheat$V8, xlab='PC1', ylab='PC2')

PCX %>%
  augment(wheat) %>%  # add original dataset
  ggplot(aes(.fittedPC1, .fittedPC2, color = as.factor(V8))) +
```

```
  geom_point(size = 5, shape = "*") +
  theme_half_open(12) +
  background_grid() +
  ggtitle("Scatterplot of PC1 & PC2") +
  theme(plot.title = element_text(hjust=0.5))
```



The broom package takes the messy output of built-in functions in R, such as lm, nls, or t.test, and turns them into tidy tibbles.

broom::augment() adds information about observations to a dataset

From the graph we can see that first two PCs separate all 3 species pretty well.

The **2** species tends to have low value in PC1 which means that it tends to have low value of the variables that have positive relation with PC1. It also tends to have little bit higher value in PC2 which means that it tends to have a bit higher value (comparing with species 1) of the variables that have positive relationship with PC2.

The **1** species tends to have around zero value in PC1. This may indicates that for those variables which have strong relationship (more likely to be negative relationship based on the results in (a)) with PC1, "species 1" may have small values on them. It also tends to have lower value in PC2. This means that for those variables which have positive relationship with PC2, "species 1" tend to have low value on them.

The **3** species tends to have higher value in PC1. This shows that it is more likely to have lower value of the variables that have negative relationship with PC1 and higher value of the variables that have positive relationship with PC1. It also tends to have higher value in PC2 which means that it is more likely to have high value of the variables that have positive relationship with PC2.

Based on the results in (a), PC1 has strong negative relationship with V1(area). PC2 has strong positive

relationship with V6(asymmetry coefficient).

This may say that **species 2** has large value of V1(area) and a bit higher value of V6(asymmetry coefficient). **species 1** has small value of V1(area). **species 3** has small value of V1(area) and a bit higher value of V6(asymmetry coefficient).

(d)

From the textbook: The covariance between the PC vector Y and the original vector X is calculated with the help of (11.4) as follows:

$$\begin{aligned} Cov(X,Y) &= E(XY^T) - E(X)E(Y) = E(XY^T) \\ &= E(XX^T - \mu\mu^T\Gamma) = Var(X)\Gamma \\ &= \Sigma\Gamma \\ &= \Gamma\Lambda\Gamma^T\Gamma \\ &= \Gamma\Lambda \end{aligned}$$

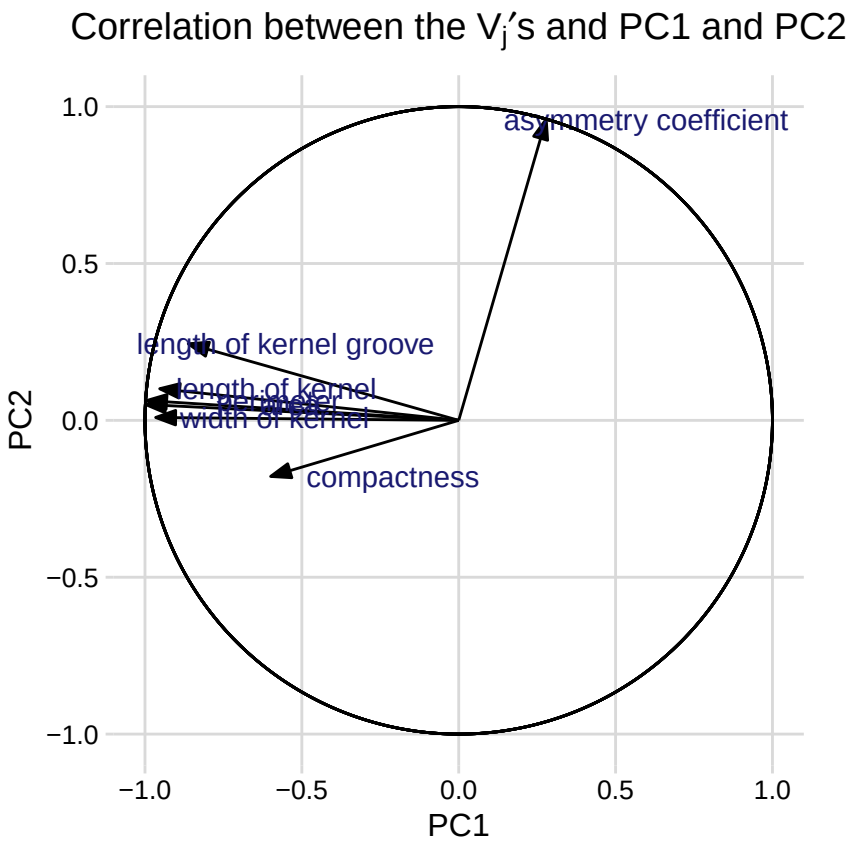
Hence, the correlation between variable X_i and the PC Y_j is

$$\rho_{X_iY_j} = \frac{\gamma_{ij}\lambda_j}{(\sigma_{X_iX_i}\lambda_j)^{1/2}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_iX_i}}\right)^{1/2}$$

Replacing each population quantity by its empirical estimator and calculating in R as follow:

```
# Unbiased sample covariance matrix S of X at (c)
S <- cov(wheat[, 1:7])

# Correlation matrix based on the formula and a bit tidy
correlation <- as.data.frame(diag(diag(S)^(-1/2)) %*% gamma %*% diag(lambda^(1/2)))

correlation <- correlation %>% rename(c("PC1"="V1", "PC2"="V2","PC3"="V3", "PC4"="V4",
                                       "PC5"="V5", "PC6"="V6", "PC7"="V7"))
correlation$column <- c("area", "perimeter", "compactness", "length of kernel",
                        "width of kernel", "asymmetry coefficient",
                        "length of kernel groove")

# Correlation Graph

# Define the type of arrow
arrow_style <- arrow(angle = 20, ends = "first", type = "closed",
                     length = grid::unit(8, "pt"))

# Draw the correlation graph
correlation %>% ggplot(aes(PC1, PC2)) +
  geom_segment(xend = 0, yend = 0, arrow = arrow_style) +
  geom_text(aes(label = column), hjust = 0.7, nudge_x = 0.5, color = "midnightblue") +
  geom_ellipse(aes(x0 = 0, y0 = 0, a = 1, b = 1, angle = 0)) +
  coord_fixed() +
  theme_minimal_grid(12) +
```

```
ggtitle(TeX("Correlation between the $V_j's$ and PC1 and PC2$")) +
theme(plot.title = element_text(hjust=0.5, face="bold"))
```



geom_segment() draws a straight line between points (x, y) and (xend, yend): Here it draw a straight line from point (0, 0) to point(PC1, PC2)

The correlation between V_i and PC1 or PC2 may be seen as the proportion of variance of V_i explained by PC1 or PC2. The plot above is showing which original variables are most strongly correlated with PC1 and PC2.

From it, we can see that most of the variables except for V3(compactness) are close to the periphery of the circle. This means that these variables are well explained by first two PCs.

This plot also shows that V1(area) and V2(perimeter) have largest correlation with PC1 which is the as same the conclusion shown in (b) (PC1 is essentially the difference between V1 and V2). Also, PC1 is strongly negatively correlated with V4(length of kernel), V5(width of kernel), V7(length of kernel groove), V1(area), and V2(perimeter). It is only positively correlated with V6(asymmetry coefficient).

According to the result from (c), **species 2** tends to have **low** value in **PC1**. This means that **species 2** may have **large** value in **V1(area)** and **V2(perimeter)**. **Species 3** tends to have **large** value in **PC1** which means that **species 3** may have **small** value of V1(area) and V2(perimeter).

Most of these variables are positively correlated with PC2. Among them, V6(asymmetry coefficient) has the largest correlation with PC2 and this is consistent with the conclusion in (b) which says that PC2 is essentially the difference of V6(asymmetry coefficient). V4(length of kernel), V7(length of kernel groove), V1(area), and V2(perimeter) are positively correlated with PC2 as well. PC2 is only negatively correlated with V3(compactness).

According to the result from (c), **species 2** and **species 3** tend to have a bit **large** value in **PC2**. This

means that **species 2** and **species 3** may have **large** value in **V6(asymmetry coefficient)**. **Species 1** tends to have **small** value in **PC1** which means that **species 1** may have **small** value of **V6(asymmetry coefficient)**.

Problem 2

(a)

Based on **12.4** and **12.7**, we know that the factor analysis model can be written as

$$X = QF + U + \mu$$

where Q is the *loadings of common factors* F, and U is a matrix of the (random) *specific factor*.

The variance of X can be written as

$$\Sigma = QQ^T + \psi$$

where $\psi = var(U) = diag\begin{pmatrix} \psi_{11} & \psi_{22} & \psi_{33} \end{pmatrix}$

From the question, we have k = 1, p = 3 and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix} = \begin{pmatrix} q_1^2 + \psi_{11} & q_1 q_2 & q_1 q_3 \\ q_1 q_2 & q_2^2 + \psi_{22} & q_2 q_3 \\ q_1 q_3 & q_2 q_3 & q_3^2 + \psi_{33} \end{pmatrix}$$

with Q and ψ as follow

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

$$\Psi = \begin{pmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{pmatrix}$$

In this case, according to Conditions **12.11**, the degrees of freedom of a model with 1 factor is

$$d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$$

$$= \frac{1}{2}(3-1)^2 - \frac{1}{2}(3+1) = 0$$

This indicates that we may find an unique factor loadings and specific variance.

Accoring to the formula given in **Example 12.1**, we can get

$$q_1^2 = \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} = \frac{0.9 \times 0.7}{0.4} = 1.575$$

$$q_2^2 = \frac{\sigma_{12}\sigma_{23}}{\sigma_{13}} = \frac{0.9 \times 0.4}{0.7} = 0.5142857$$

$$q_3^2 = \frac{\sigma_{13}\sigma_{23}}{\sigma_{12}} = \frac{0.7 \times 0.4}{0.9} = 0.311111$$

Therefore

$$q_1 = 1.25499 || q_1 = -1.25499$$
$$q_2 = 0.7171372 || q_2 = -0.7171372$$
$$q_3 = 0.5577734 || q_3 = -0.5577734$$

and

$$\psi_{11} = \sigma_{11} - q_1^2 = 1 - 1.575 = -0.575$$
$$\psi_{22} = \sigma_{22} - q_2^2 = 1 - 0.5142857 = 0.4857143$$
$$\psi_{33} = \sigma_{33} - q_3^2 = 1 - 0.3111111 = 0.6888889$$

Thus

$$Q = \begin{pmatrix} 1.25499 \\ 0.7171372 \\ 0.5577734 \end{pmatrix}$$

or

$$Q = \begin{pmatrix} -1.25499 \\ -0.7171372 \\ -0.5577734 \end{pmatrix}$$

$$\Psi = \begin{pmatrix} -0.575 & 0 & 0 \\ 0 & 0.4857143 & 0 \\ 0 & 0 & 0.6888889 \end{pmatrix}$$

However, this solution does not make sense. This is because ψ is the variance of specific factors, it should be positive semi-definite which means that every entry in this matrix should be equal to or greater than 0. From the computational results, we can see that $\psi_{11} = -0.575$ which does not satisfy the condition and indicates the solution may be wrong.

(b)

```
Harman23 <- data.frame(Harman23.cor)[,1:8]
```

(i)

Since there is no unique solution for the loadings Q and in order to make it easier to interpret, we need to impose some constraints on Σ. According to **12.11** and **12.12** which indicates that $Q^T\psi^{-1}Q$ or $Q^T D^{-1}Q$ is diagonal, we need to impose $\frac{1}{2}\{k(k-1)\}$ constraints where k is the number of factors.

We also know that, before imposing this constriant, the dimension(or degrees of freedom) for Σ is $\frac{1}{2}p(p+1)$.

After introducing the constraints, the effective dimension of a factor model with respect to the covariance structure is $pk + p - \frac{1}{2}k(k-1)$.

In this case, the degrees of freedom of a model with k factors is

$$d = \frac{1}{2}p(p+1) - (pk + p - \frac{1}{2}k(k-1))$$

If d < 0, it indicates that the dimension for unconstrained covariance matrix is less than the dimension for the restricted one. This means that the number of parameters of the factorial model is larger than the number of parameters of the original model. Thus it may result in "overparameterization".

```
dim(Harman23) %>% kable(caption = "Dimensions", col.names = NULL, align = c('c'))
```

Table 8: Dimensions

8
8

In this dataset, we can see that p $= 8$, in order to avoid "overparameterization", we should make sure the degrees of freedom d is equal to or greater than 0.

$$\begin{aligned} d &= \frac{1}{2} \times 8 \times (8+1) - (8k + 8 - \frac{1}{2}k(k-1)) \geq 0 \\ &= 36 - 8k - 8 + \frac{1}{2}k^2 - \frac{1}{2}k \\ &= \frac{1}{2}k^2 - \frac{17}{2}k + 28 \geq 0 \end{aligned}$$

Therefore

$$k = 12.53113 \approx 13$$

or

$$k = 4.468871 \approx 4$$

Thus, in this dataset, maximum number of factors we can fit is **4**.

(ii)

```
(Harman23.FA <- factanal(factors = 1, covmat = Harman23.cor))
```

```
##
## Call:
## factanal(factors = 1, 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
```

```
for(factors in 2:4)
  print(update(Harman23.FA, factors = factors))
```

```
##
## Call:
## factanal(factors = factors, 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 = factors, 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 = factors, 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
```

Null Hypothesis 1: 1 Factor is good fit

```
qchisq(.95, df=20) %>% kable(caption = "Chi-sq w df = 20",
                             col.names = NULL, align = c('c'))
```

Table 9: Chi-sq w df = 20

31.41043

From the test result, we can see that in this case, the chi square statistic is 611.44 on 20 degrees of freedom, which is much greater than 31.41043. Therefore, we should reject the null hypothesis and conclude that 1 factor is not a good fit under 5% significant level.

Null Hypothesis 2: 2 Factor is good fit

```
qchisq(.95, df=13) %>% kable(caption = "Chi-sq w df = 13",
                             col.names = NULL, align = c('c'))
```

Table 10: Chi-sq w df = 13

22.36203

From the test result, we can see that in this case, the chi square statistic is 75.74 on 13 degrees of freedom, which is greater than 22.36203. Therefore, we should reject the null hypothesis and conclude that 2-factor is not a good fit under 5% significant level.

Null Hypothesis 3: 3 Factor is good fit

```
qchisq(.95, df=7) %>% kable(caption = "Chi-sq w df = 7",
                            col.names = NULL, align = c('c'))
```

Table 11: Chi-sq w df = 7

14.06714

From the test result, we can see that in this case, the chi square statistic is 22.81 on 7 degrees of freedom, which is greater than 14.06714. Therefore, we should reject the null hypothesis and conclude that 3-factor is not a good fit under 5% significant level.

Null Hypothesis 4: 4 Factors is good fit

```
qchisq(.95, df = 2) %>% kable(caption = "Chi-sq w df = 2",
                              col.names = NULL, align = c('c'))
```

Table 12: Chi-sq w df = 2

5.991465

From the test result, we can see that in this case, the chi square statistic is 4.63 on 2 degrees of freedom, which is less than 5.991465. Therefore, we do not reject the null hypothesis and conclude that 4-factor is a

good fit under 5% significant level.

Thus, based on these LR statistics, 4-factor model has the best fit.