Assignment1 MAST90138

Muhan Guan(1407870)

11th 09 2023

Tutor's name:Shangyu Chen

Time of Tutorial class: Thursday 10AM

Problem 1

(a)

By definition, the covariance matrix Σ should be **1)** symmetric and **2)** positive semi-definite matrix.

Therefore, for **1)**

$$\Sigma_{12} = \Sigma_{21}$$

then a=3.

For **2)** we have

$$X^T \Sigma X \geqslant 0 \text{ for all non} - \text{zero } X$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 3 & b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geqslant 0$$

$$\begin{bmatrix} x_1 + 3x_2 & 3x_1 + bx_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geqslant 0$$

$$x_1(x_1 + 3x_2) + x_2(3x_1 + bx_2) \geqslant 0$$

$$x_1^2 + 6x_1x_2 + bx_2^2 \geqslant 0$$

To make this inequality hold, we have $6^2 - 4b \leqslant 0$, hence $b \geqslant 9$

In conclusion, $a = 3$ and $b \geqslant 9$

(b)

We have formula $\Sigma V = \lambda V$, then

$$(\Sigma - I\lambda)V = 0$$

$$\det(\begin{bmatrix} 4-\lambda & -\sqrt{3} \\ -\sqrt{3} & 2-\lambda \end{bmatrix}) = 0$$

$$(4-\lambda)(2-\lambda) - 3 = 0$$

$$5 - 6\lambda + \lambda^2 = 0$$

$$\text{solve it, } \lambda_1 = 1, \lambda_2 = 5$$

$$\begin{bmatrix} 4 & -\sqrt{3} \\ -\sqrt{3} & 2 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}$$

$$\begin{bmatrix} 4v_{11} - \sqrt{3}v_{12} \\ -\sqrt{3}v_{11} + 2v_{12} \end{bmatrix} = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}$$

$$\text{solve it, } v_{11} = \frac{\sqrt{3}}{3} v_{12}$$

$$\text{let } v_{12} = a \text{ then } a^2 + \frac{1}{3}a^2 = 1, \ a = \frac{\sqrt{3}}{2} \text{or} - \frac{\sqrt{3}}{2}$$

$$\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{bmatrix}$$

$$\begin{bmatrix} 4 & -\sqrt{3} \\ -\sqrt{3} & 2 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 5 \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix}$$

$$\begin{bmatrix} 4v_{21} - \sqrt{3}v_{22} \\ -\sqrt{3}_{21} + 2v_{22} \end{bmatrix} = \begin{bmatrix} 5v_{21} \\ 5v_{22} \end{bmatrix}$$

$$\text{solve it, } v_{21} = -\sqrt{3}v_{22}$$

$$\text{let } v_{22} = a \text{ then } a^2 + 3a^2 = 1, \ a = \frac{1}{2} \text{or} - \frac{1}{2}$$

$$\begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{bmatrix}$$

then we have orthogonal matrix $\Gamma = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}$ and diagonal matrix $\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$,which can make $\Sigma = \Gamma\Lambda\Gamma^T$

To check if the equation can hold or not:

```
gamma=matrix(c(1/2,sqrt(3)/2,-sqrt(3)/2,1/2 ),2,2)
lambda=matrix(c(1,0,0,5),2,2)
gamma%*%lambda%*%t(gamma)
```

```
##           [,1]      [,2]
## [1,]  4.000000 -1.732051
## [2,] -1.732051  2.000000
```

```
#the sigma is same with the matrix in the question
```

(c)

```
data=read.table(file='/Users/guanmuhan/Downloads/Wheat (1).txt')#read data
X=as.matrix(data[,1:7]) #Take the first 7 features and convert them into matrix form
cat("dimension of X:", dim(X), "\n")
```

```
## dimension of X: 210 7
```

```
variety=as.vector(data[,8])#Take out the variety features and convert them into vectors
cat("wheat variety :", variety, "\n")
```

```
## wheat variety : 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

```
#length
n=length(variety)
```

(d)

```
S=cov(X)#covariance matrix
S
```

```
##             V1          V2            V3           V4           V5           V6
## V1  8.46635078  3.77844320  0.0418225658  1.224703671  1.066911361 -1.004355845
## V2  3.77844320  1.70552820  0.0163319511  0.562665550  0.466064932 -0.426765980
## V3  0.04182257  0.01633195  0.0005583493  0.003851826  0.006797719 -0.011776556
## V4  1.22470367  0.56266555  0.0038518256  0.196305245  0.143991709 -0.114289956
## V5  1.06691136  0.46606493  0.0067977190  0.143991709  0.142668202 -0.146542890
## V6 -1.00435584 -0.42676598 -0.0117765562 -0.114289956 -0.146542890  2.260684046
## V7  1.23513290  0.57175254  0.0026342068  0.203125110  0.139068229 -0.008187052
##              V7
## V1  1.235132905
## V2  0.571752539
## V3  0.002634207
## V4  0.203125110
## V5  0.139068229
## V6 -0.008187052
## V7  0.241553081
```

```
eigen_set=eigen(S)
e_val=eigen_set$values#eigenvalue
e_val_matrix=diag(e_val)
e_vec=eigen_set$vectors#eigenvector

S1=e_vec%*%e_val_matrix%*%t(e_vec)


library(MASS)
#Output the result of the difference between the two matrices in the form of a fraction.
fractions(S1)-fractions(S)
```

```
##    V1 V2 V3 V4 V5 V6 V7
## V1 0  0  0  0  0  0  0
## V2 0  0  0  0  0  0  0
## V3 0  0  0  0  0  0  0
## V4 0  0  0  0  0  0  0
## V5 0  0  0  0  0  0  0
## V6 0  0  0  0  0  0  0
## V7 0  0  0  0  0  0  0
```

we can see the result is 0, so the covariance matrix S of $X = \Gamma\Lambda\Gamma^T$.

Problem2

(a)

1.X8 can not be included because PCA can only be used for numerical variables and cannot be applied to categorical variables, X8 is categorical variable.

2.

$$\psi_1 = 0.718743$$
$$\psi_2 = 0.8898249$$
$$\psi_3 = 0.9866825$$
$$\psi_4 = 0.9964488$$
$$\psi_5 = 0.9991222$$
$$\psi_6 = 0.9998839$$
$$\psi_7 = 1$$

3.Through the screen plot, we recommend keeping the first three principal components, because they explain most of the variation in the data set, explaining a roughly total of 98.7% of the data, which is already reasonable.

```
diag(diag(cov(X)))
```

```
##          [,1]     [,2]         [,3]      [,4]      [,5]     [,6]      [,7]
## [1,] 8.466351 0.000000 0.0000000000 0.0000000 0.0000000 0.000000 0.0000000
## [2,] 0.000000 1.705528 0.0000000000 0.0000000 0.0000000 0.000000 0.0000000
## [3,] 0.000000 0.000000 0.0005583493 0.0000000 0.0000000 0.000000 0.0000000
## [4,] 0.000000 0.000000 0.0000000000 0.1963052 0.0000000 0.000000 0.0000000
## [5,] 0.000000 0.000000 0.0000000000 0.0000000 0.1426682 0.000000 0.0000000
## [6,] 0.000000 0.000000 0.0000000000 0.0000000 0.0000000 2.260684 0.0000000
## [7,] 0.000000 0.000000 0.0000000000 0.0000000 0.0000000 0.000000 0.2415531
```

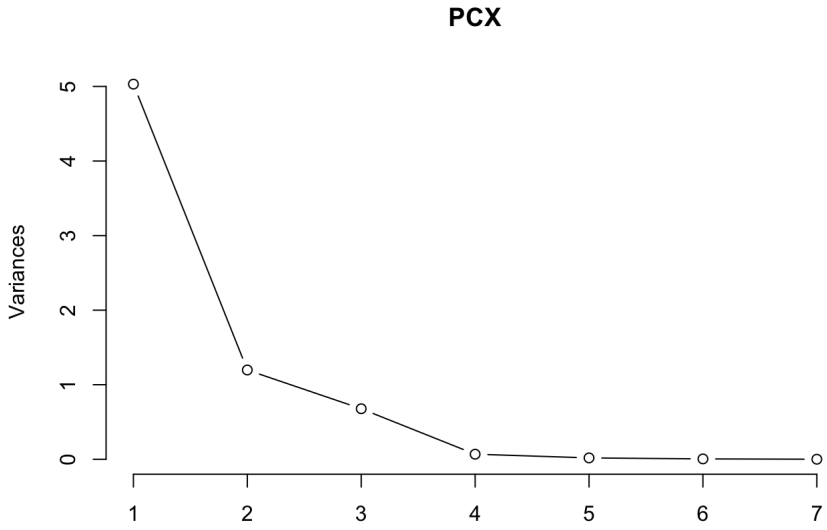
```
#we can observe that the variance of the area is much larger than that of the other variables,so we need to do PC
A of scaled variables,hence we set the parameter 'scale' as True.
#PCA
PCX=prcomp(X,retx = T,scale. = T)
#eigenvalue
Lambda=PCX$sdev^2
#eigenvector
Gamma=PCX$rotation
PCX_summary = summary(PCX)
# The percentage of variability explained by each principal component by checing the information in  "Proportion
of Variance"
cat("The precentage of the variability explained by each PC:" , "\n")
```

```
## The precentage of the variability explained by each PC:
```

```
PCX_summary$importance["Proportion of Variance", ]
```

```
##     PC1     PC2     PC3     PC4     PC5     PC6     PC7
## 0.71874 0.17108 0.09686 0.00977 0.00267 0.00076 0.00012
```

```
screeplot(PCX,type='lines')
```




```
#cumulative percentage
cat("Cumulative percentage of variance :" ,cumsum(Lambda)/sum(Lambda), "\n")
```

```
## Cumulative percentage of variance : 0.718743 0.8898249 0.9866825 0.9964488 0.9991222 0.9998839 1
```

(b)

PC1 puts weights on-0.4444735, -0.4415715, -0.2770174, -0.4235633, -0.4328187, 0.1186925 , -0.3871608 on, respectively,the area, the perimeter, the compactness, the length of kernel, the width of kernel, the asymmetry coefficient and the length of kernel groove.PC1 puts the most weight on the area ,the perimeter ,the length of kernel and the width of kernel and also some weight on the length of kernel groove,all of which contribute negatively to PC1.

$$PC1 = -0.4444735 * area - 0.4415715 * perimeter - 0.2770174 * compactness - 0.4235633 * length\ of\ kernel - 0.4328187 * wi$$

PC2 puts weights on 0.02656355, 0.08400282, -0.52915125, 0.20597518, -0.11668963, 0.71688203, 0.37719327 on, respectively,the area, the perimeter, the compactness, the length of kernel, the width of kernel, the asymmetry coefficient and the length of kernel groove.PC2 puts the most weight on the asymmetry coefficient and also put some weight on the compactness and the length of kernel groove.

$$PC2 = 0.02656355 * area + 0.08400282 * perimeter - 0.52915125 * compactness + 0.20597518 * length\ of\ kernel - 0.11668963$$

Do the coefficients of these linear combinations depend on the wheat kernel?

No, the coefficients of these linear combinations are not dependent on individual wheat kernels. These coefficients are derived from the entirety of the dataset and define the direction of the principal components in the original feature space. They remain consistent and are not influenced by individual observations or specific indices of the wheat kernels.

Do the principal components depend on the wheat kernel?

Yes, the principal components are indeed dependent on individual wheat kernels. The principal components represent the projection of each wheat kernel onto the directions delineated by the coefficients of these linear combinations. As such, each wheat kernel will have its unique set of scores for the principal components, indicating its position in the transformed coordinate system defined by these components.

```
Gamma
```

```
##           PC1         PC2         PC3         PC4         PC5         PC6
## V1 -0.4444735  0.02656355 -0.02587094  0.19363997 -0.20441167  0.42643686
## V2 -0.4415715  0.08400282  0.05983912  0.29545659 -0.17427591  0.47623853
## V3 -0.2770174 -0.52915125 -0.62969178 -0.33281640  0.33265481  0.14162884
## V4 -0.4235633  0.20597518  0.21187966  0.26340659  0.76609839 -0.27357647
## V5 -0.4328187 -0.11668963 -0.21648338  0.19963039 -0.46536555 -0.70301171
## V6  0.1186925  0.71688203 -0.67950584  0.09246481  0.03625822  0.01964186
## V7 -0.3871608  0.37719327  0.21389720 -0.80414995 -0.11134657 -0.04282974
##            PC7
## V1  0.734805689
## V2 -0.670751532
## V3 -0.072552703
## V4  0.046276051
## V5 -0.039289079
## V6 -0.003723456
## V7 -0.034498098
```

(c)

1.What you can extract from this graph?

The three types of wheat have generally obvious distribution differences in the two-dimensional space of PC1 and PC2, which means that PCA successfully distinguished the three types of wheat on these two principal components.The first type of wheat has a higher density in the negative area of PC1, but is more widely distributed on PC2.The second type of wheat is widely distributed on PC1, but mainly concentrated in the negative value area on PC2.The third type of wheat has a higher density in the positive area of PC1, but is more widely distributed on PC2.

2.Which groups are visible on the graph? What do they correspond to?

The varieties 1 and 3 is more clearly visible, while there is overlap between variety 2 and the other two varieties on the scatter plot.

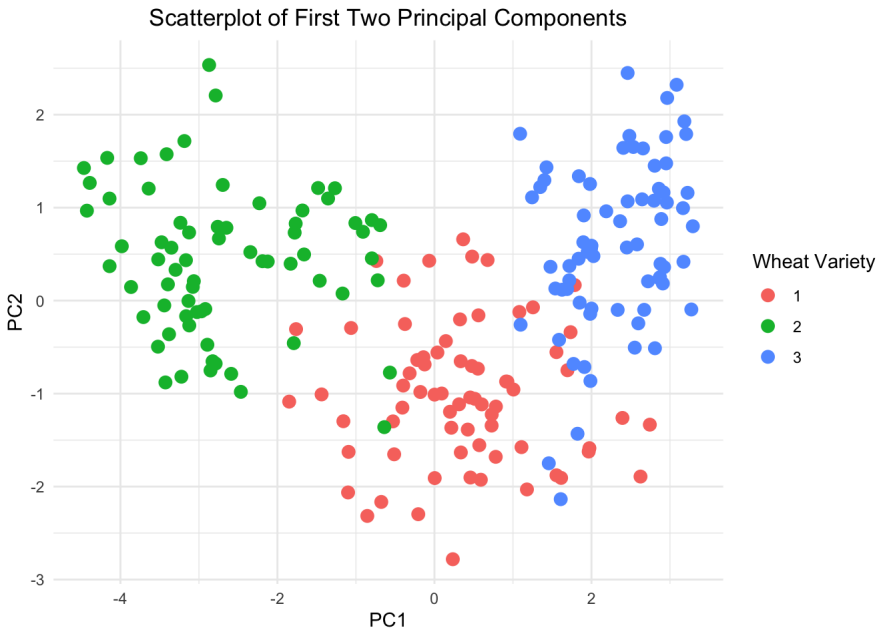
3.How do the original variables contribute to those groups?

Variety 3 exhibits higher values on PC1, suggesting that, in comparison to Varieties 1 and 2, it tends to possess a smaller area, perimeter, length of kernel, and width of kernel. Conversely, Variety 1, with its generally lower PC1 values, indicates that it has larger measurements for these four variables relative to the other two varieties. Furthermore, compared with Variety 1, both Varieties 3 and 2 display larger values on PC2, pointing to a higher asymmetry coefficient and a lower compactness for these two varieties.

```
# loadingnecessary package
library(ggplot2)

# Combine principal component scores and wheat kernel categories into a new data frame
pca_1st2nd_data=data.frame(PC1 = PCX$x[,1], PC2 =PCX$x[,2], Variety = data$V8)


#use ggplot2 draw a scatterplot
ggplot(pca_1st2nd_data, aes(x = PC1, y = PC2, color = as.factor(Variety))) +
  geom_point(size = 3) +
  scale_color_discrete(name = "Wheat Variety") +
  labs(title = "Scatterplot of First Two Principal Components",
       x = "PC1",
       y = "PC2") +
  theme_minimal() +
  theme(legend.position = "right",
            plot.title = element_text(hjust = 0.5))
```



(d)

We have $\rho_{x_{ij},y_{ik}} = \gamma_{kj}\frac{\lambda_k^{\frac{1}{2}}}{\sigma_{jj}^{\frac{1}{2}}}$

Conclusion:

1. Interpretability: The arrows for the variables 'area' and 'perimeter' are located on the periphery, indicating that these two variables are fully accounted for by PC1 and PC2. The variables 'length of kernel', 'width of kernel', and 'length of kernel groove' are also proximate to the periphery, suggesting that the first two principal components capture most of their variation. However, the arrows for 'compactness' and 'asymmetry coefficient' are notably distant from the periphery, implying that the initial two principal components do not entirely encapsulate the variance of these two variables, necessitating further information.

2. Correlation: The variables 'area', 'perimeter', 'length of kernel', 'width of kernel', and 'length of kernel groove' have a minimal angle with PC1 and are oriented in the opposite direction, indicating a strong negative correlation with PC1. Their angle with PC2 is substantial, nearing 90 degrees, suggesting a minor positive correlation with PC2.The 'asymmetry coefficient' has a pronounced angle with PC1, indicating a weaker correlation between them. Conversely, its angle with PC2 is smaller, hinting at a positive correlation with PC2. However, the strength of this correlation cannot be conclusively determined since its arrow is not on the periphery.The variable 'compactness' exhibits discernible angles with both PC1 and PC2. While the direction suggests a negative correlation with both principal components, it doesn't necessarily indicate a strong correlation.Additionally, by observing the direction of the variable arrows, we can infer that the variables 'area', 'perimeter', 'length of kernel', 'width of kernel', and 'length of kernel groove' exhibit positive correlations with each other. Among these, the correlation between 'area' and 'perimeter' appears to be more pronounced, suggesting a higher correlation coefficient.

```
#calculte by formula
corr=(Gamma%*%sqrt(diag(Lambda)))%*%solve(sqrt(diag(diag(cov(scale(X))))))

colnames(corr) =paste0("PC", 1:ncol(corr))
rownames(corr) =gsub("V", "X", rownames(corr))
#rename the names of columns and rows respectively
print(corr)
```

```
##           PC1         PC2         PC3         PC4         PC5          PC6
## X1 -0.9969692  0.02906947 -0.02130238  0.05063027 -0.02796304  0.031138785
## X2 -0.9904598  0.09192737  0.04927210  0.07725186 -0.02384054  0.034775346
## X3 -0.6213594 -0.57906965 -0.51849428 -0.08702018  0.04550641  0.010341859
## X4 -0.9500669  0.22540620  0.17446376  0.06887187  0.10480049 -0.019976789
## X5 -0.9708269 -0.12769775 -0.17825451  0.05219656 -0.06366093 -0.051334518
## X6  0.2662313  0.78451032 -0.55951166  0.02417641  0.00496004  0.001434266
## X7 -0.8684149  0.41277645  0.17612501 -0.21025788 -0.01523195 -0.003127464
##              PC7
## X1  0.0209438545
## X2 -0.0191181460
## X3 -0.0020679389
## X4  0.0013189866
## X5 -0.0011198399
## X6 -0.0001061281
## X7 -0.0009832846
```

```
plot(0, 0, xlim = c(-1, 1), ylim = c(-1, 1), type = "n", xlab = "PC1", ylab = "PC2", main = "Correlation Graph")
radius = 1
theta = seq(0, 2 * pi, length = 200)
lines(x = radius * cos(theta), y = radius * sin(theta))
result=apply(corr, 1, function(row) {arrows(0, 0, row[1], row[2], length = 0.1)})
text(corr[,1], corr[,2], labels = rownames(corr), pos = 4)
```