MAST90138 Multivariate Statistics for Data Science Assignment1

Yini Lai (Student ID: 1127650)

08/30/2020

Problem 1

(a)

Since Σ is a covariance matrix, it is symmetric ($\Sigma = \Sigma^T$).

$$\Sigma = \begin{pmatrix} 1 & 2 \\ a & b \end{pmatrix} = \Sigma^T = \begin{pmatrix} 1 & a \\ 2 & b \end{pmatrix}$$

Thus, we can know that $\mathbf{a = 2}$ and $\Sigma = \begin{pmatrix} 1 & 2 \\ 2 & b \end{pmatrix}$

Since Σ is a covariance matrix, it should be semi positive definite. In this case, all principal minors of this matrix are non-negative. Then we have

$$\Sigma_1 = |1| \geqslant 0$$

$$\Sigma_2 = \begin{vmatrix} 1 & 2 \\ 2 & b \end{vmatrix} \geqslant 0$$

$$b - 4 \geqslant 0$$

Thus $\mathbf{b \geqslant 4}$

In general, in order to let matrix Σ to be a covariance matrix, it should be symmetry and semi positive definite. Based on the definition of symmetry and extending Sylversters Criterion on semi positive definite, we can know that $\mathbf{a = 2}$ and $\mathbf{b \geqslant 4}$.

(b)

Since $\Sigma = \begin{pmatrix} 13 & -4 \\ -4 & 7 \end{pmatrix} = \Sigma^T$, Σ is a square and symmetric matrix. All eigenvalues of Σ satisfy $|\Sigma - \lambda I_p| = 0$.

$$\begin{aligned} |\Sigma - \lambda E| &= \begin{vmatrix} 13 - \lambda & -4 \\ -4 & 7 - \lambda \end{vmatrix} \\ &= (13 - \lambda)(7 - \lambda) - 16 \\ &= (\lambda - 15)(\lambda - 5) = 0 \end{aligned}$$

Thus $\lambda_1 = 15, \lambda_2 = 5$

When $\lambda = 15$, $(\Sigma - 15E)x = 0$.

$$\begin{bmatrix} 13-15 & -4 \\ -4 & 7-15 \end{bmatrix} \overset{r_2 - 2r_1}{\longleftrightarrow} \begin{bmatrix} -2 & -4 \\ 0 & 0 \end{bmatrix} \overset{-\frac{1}{2}r_1}{\longleftrightarrow} \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

$$x_1 + 2x_2 = 0$$

$$x_1 = -2x_2$$

Therefore, one of the eigenvectors corresponding to egievnvalue $\lambda_1 = 15$ is $\alpha_1 = (-2, 1)^T$

When $\lambda = 5$, $(\Sigma - 5E)x = 0$.

$$\begin{bmatrix} 13-5 & -4 \\ -4 & 7-5 \end{bmatrix} \overset{r_2 + \frac{1}{2}r_1}{\longleftrightarrow} \begin{bmatrix} 8 & -4 \\ 0 & 0 \end{bmatrix} \overset{\frac{1}{4}r_1}{\longleftrightarrow} \begin{bmatrix} 2 & -1 \\ 0 & 0 \end{bmatrix}$$

$$2x_1 = x_2$$

Therefore, one of the eigenvectors corresponding to egievnvalue $\lambda_2 = 5$ is $\alpha_2 = (1, 2)^T$

Since

$$\alpha_1^T \alpha_2 = \begin{bmatrix} -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = -2 + 2 = 0$$

two vectors α_1 and α_2 are orthogonal.

Then scale them so that they have norm 1:

$$v_1 = \frac{1}{\| \alpha_1 \|} \alpha_1^T = \frac{1}{\sqrt{(-2)^2 + 1^2}} \begin{bmatrix} -2 & 1 \end{bmatrix}^T = \begin{bmatrix} -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{bmatrix}$$

$$v_2 = \frac{1}{\| \alpha_2 \|} \alpha_2^T = \frac{1}{\sqrt{1^2 + 2^2}} \begin{bmatrix} 1 & 2 \end{bmatrix}^T = \begin{bmatrix} \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}$$

Thus

$$\Gamma = \begin{bmatrix} -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 15 & 0 \\ 0 & 5 \end{bmatrix}$$

In this way, we have

$$\Sigma = \Gamma \Lambda \Gamma^T$$

$$\begin{bmatrix} 13 & -4 \\ -4 & 7 \end{bmatrix} = \begin{bmatrix} -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} 15 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}$$

(c)

```
# Read the wheat data in R
wheat <- read.table("Wheat data.txt")

# Create a data matrix X
X <- as.matrix(wheat[, 1:7])
# Example(first 6 rows) of the matrix X
knitr::kable(head(X), col.names = NULL, caption = "Matrix X")
```

Table 1: Matrix X

15.26	14.84	0.8710	5.763	3.312	2.221	5.220
14.88	14.57	0.8811	5.554	3.333	1.018	4.956
14.29	14.09	0.9050	5.291	3.337	2.699	4.825
13.84	13.94	0.8955	5.324	3.379	2.259	4.805
16.14	14.99	0.9034	5.658	3.562	1.355	5.175
14.38	14.21	0.8951	5.386	3.312	2.462	4.956

```r
# Create a vector of wheat variety
variety <- as.matrix(wheat[, 8])
# Example(first 6 rows) of the vector variety
knitr::kable(head(variety), caption = "Vector 'variety'", align = c('c'))
```

Table 2: Vector 'variety'

1
1
1
1
1
1

(d)

```r
# Unbiased sample covariance matrix S of X at (c)
S <- cov(X)
knitr::kable(S, caption = "Sample Covariance")
```

Table 3: Sample Covariance

	V1	V2	V3	V4	V5	V6	V7
V1	8.4663508	3.7784432	0.0418226	1.2247037	1.0669114	-1.0043558	1.2351329
V2	3.7784432	1.7055282	0.0163320	0.5626656	0.4660649	-0.4267660	0.5717525
V3	0.0418226	0.0163320	0.0005583	0.0038518	0.0067977	-0.0117766	0.0026342
V4	1.2247037	0.5626656	0.0038518	0.1963052	0.1439917	-0.1142900	0.2031251
V5	1.0669114	0.4660649	0.0067977	0.1439917	0.1426682	-0.1465429	0.1390682
V6	-1.0043558	-0.4267660	-0.0117766	-0.1142900	-0.1465429	2.2606840	-0.0081871
V7	1.2351329	0.5717525	0.0026342	0.2031251	0.1390682	-0.0081871	0.2415531

```r
EE <- eigen(S)
# Eigen vectors
Evect <- EE$vectors
knitr::kable(Evect, caption = "Eigenvectors")
```

Table 4: Eigenvectors

0.8842285	-0.1008058	0.2645335	-0.1994495	0.1371730	-0.2806396	-0.0253982
0.3954054	-0.0564896	-0.2825200	0.5788169	-0.5747560	0.3015586	0.0658399
0.0043113	0.0028947	0.0590358	-0.0577602	0.0531045	0.0452291	0.9941256
0.1285445	-0.0306217	-0.4001495	0.4361002	0.7869978	0.1134376	0.0014314
0.1110591	-0.0023723	0.3192387	-0.2341636	0.1448029	0.8962678	-0.0815499
-0.1276156	-0.9894105	0.0642975	0.0251474	0.0015756	-0.0032880	0.0011427
0.1289665	-0.0822334	-0.7619397	-0.6133566	-0.0876536	0.1099236	0.0089719

```r
# Eigen values
Eval <- EE$values
knitr::kable(Eval, col.names = NULL, caption = "Eigenvalues")
```

Table 5: Eigenvalues

10.7933269
2.1294551
0.0736300
0.0128875
0.0027482
0.0015704
0.0000297

```
# Spectral decomposition of S
SampleCov <- Evect%*%diag(Eval)%*%t(Evect)
```

Problem 2

(a)

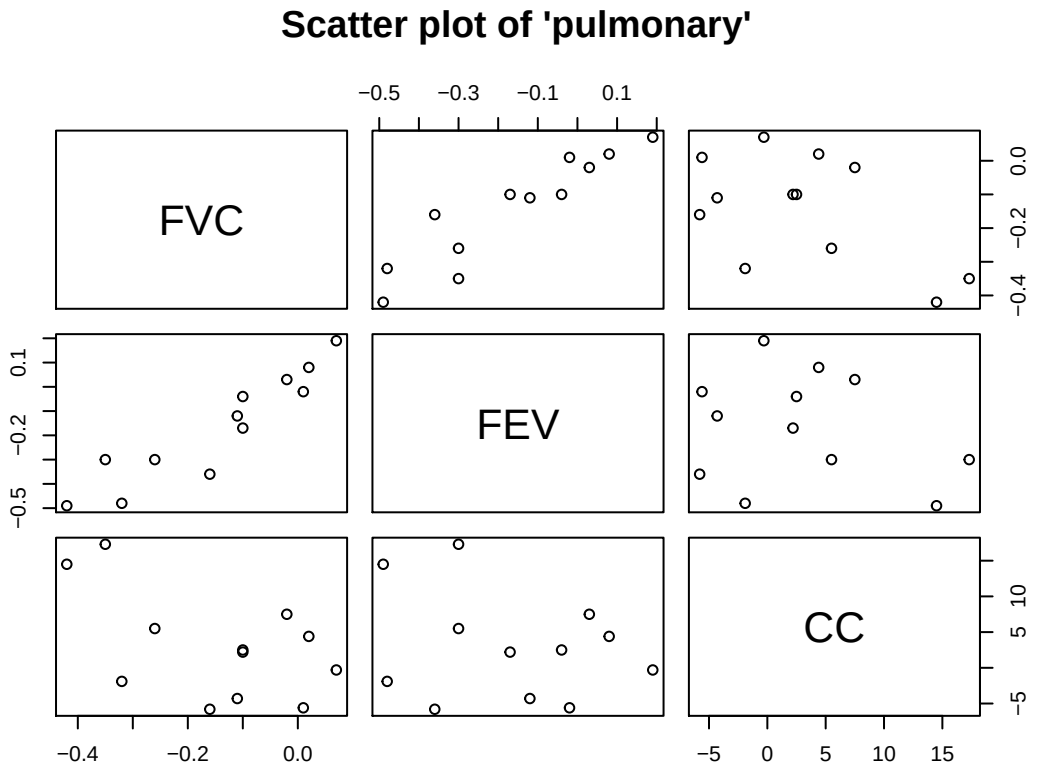
```
library(ICSNP)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: ICS
```

```
library(mvnormtest)
data(pulmonary)

# Make a scatter plot of this dataset
plot(pulmonary, main = "Scatter plot of 'pulmonary'")
```



(b)

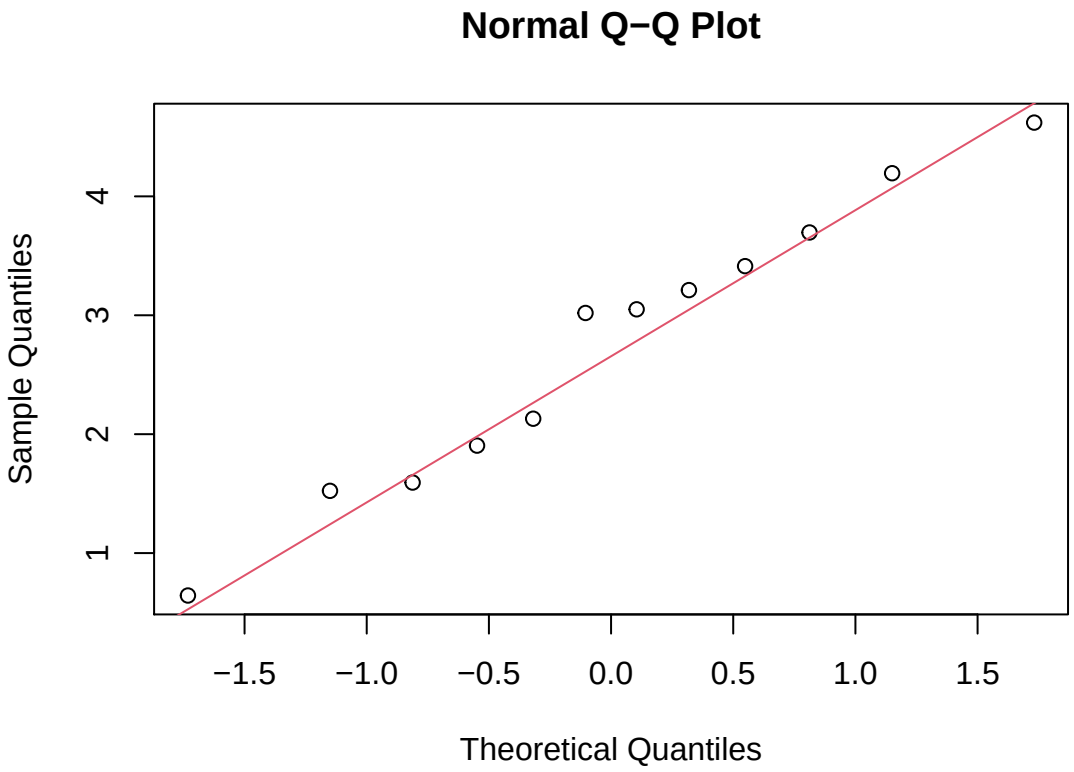
According to the statement of this question, we expect the squared Mahalanobis distances to roughly follow a chi-squared distribution with 3 degree of freedom. In this case, the assumption that the data come from a multivariate normal distribution is valid.

Firstly, we can check whether the squared Magalanovis distances roughly follow a chi-squared distribution or not by looking at the Q-Q Plot

```
# Compute the Unbiased sample covariance matrix of the data
Sx <- cov(pulmonary)

# Get the mahalanobis distances
D2 <- mahalanobis(pulmonary, colMeans(pulmonary), Sx)

# Plot the mahalanobis distances
qqnorm(D2, distribution = "chisq", df = 3)
qqline(D2, col = 2)
```



From the Q-Q Plot, we can see that the data points stay around the 45 degree line(diagonal line). There is no outlier even at the two tails. In this case, we may say that the mahalanobis distances roughly follow a chi-squared distribution with 3 degree of freedom and the multivariate normal assumption is valid.

We can also do the Shapiro-Wilk Test to check the normality.

```
C <- t(pulmonary[,])
mshapiro.test(C)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  Z
## W = 0.94032, p-value = 0.5022
```

From the test results, we can see that the p-value is 0.5022 which is much greater than 0.05. Therefore, we do not have strong evidence to reject the null hypothesis and conclude that the data come from multivariate normal distribution.

(c)

```
# Do the Hotelling's T2 test in R
HotellingsT2(pulmonary, mu = c(0, 0, 0), test = "f")
```

```
##
##  Hotelling's one sample T2-test
##
## data:  pulmonary
## T.2 = 3.8231, df1 = 3, df2 = 9, p-value = 0.05123
## alternative hypothesis: true location is not equal to c(0,0,0)
```

Null Hypothesis: The means of the three variables are all zero.

Alternative Hypothesis: At least one of the variables have non-zero mean.

From the test result, we can see that the p-value is 0.05123 which is slightly greater than 0.05. Therefore, we do not have strong evidence to reject the null hypothesis and may say that the means of the three variables are all closed to zero.

Compute the p-value "manually"

```
samplemean <- matrix(colMeans(pulmonary),1, 3)
knitr::kable(samplemean, caption = "Sample Mean")
```

Table 6: Sample Mean

-0.145	-0.165	3

```
expectedmean <- matrix(c(0, 0, 0), 1, 3)
knitr::kable(expectedmean, caption = "Expected Mean")
```

Table 7: Expected Mean

0	0	0

```
# unbiased sample covariance matrix of the Yi's: Sx
T2 <- 12*((samplemean - expectedmean)%*%solve(Sx)%*%(t(samplemean - expectedmean)))

# Test Statistic
Fvalue <- ((12 - 1 - 3 + 1)/(3*(12 - 1)))*T2
knitr::kable(Fvalue, caption = "Test Statistic")
```

Table 8: Test Statistic

3.823146

```r
# Corresponding quantile of test statistic
knitr::kable(pf(Fvalue, df1=3, df2=(12 - 1 - 3 + 1), lower.tail = FALSE), caption = "Corresponding Quant
```

Table 9: Corresponding Quantile

0.0512288

```r
#crit <- qf(.95, df1=3, df2=(12 - 1 - 3 + 1))
#knitr::kable(crit, caption = "Critical value")
```

Note that the degree of freedom is 3 and (12 - 1 - 3 + 1) respectively. This is because we are calculating these values based on sample data

From the result, we can see that the p-value is 0.0512, which is slightly greater than 0.05 and not fall in the reject region. Therefore, we do not have strong evidence to reject the null hypothesis and conclude that the means of three variables are all closed to zero at 5% significant level.

Note that we assume the significant level is 5% in this case