Applied Multivariate analysis-HW10

ID : 111024517 Name : 鄭家豪

Problem 1

$X_1, X_2, \cdots, X_{60} \overset{iid}{\sim} N_4(\mu, \Sigma)$

(a) $\bar{X} = \frac{1}{60}\sum_{i=1}^{60} X_i \sim N_4(\mu, \frac{1}{60}\Sigma)$ #

$(\because Var(X_i) = \Sigma \quad \therefore Var(\bar{X}) = (\frac{1}{60})^2 \cdot 60 \cdot \Sigma = \frac{1}{60}\Sigma)$

(b) By Eigen-decomposition, $\Sigma = \sum_{i=1}^{4} \lambda_i e_i e_i^T$

$(X_i - \mu)^T \Sigma^{-1} (X_i - \mu) = \sum_{i=1}^{4} \lambda_i^{-1} (X_i - \mu)^T e_i e_i^T (X_i - \mu)$

$= \sum_{i=1}^{4} \lambda_i^{-1} Y_i^2$, where $Y_i \sim N(0, \lambda_i)$

$= \sum_{i=1}^{4} \left(\frac{Y_i}{\sqrt{\lambda_i}}\right)^2 \sim \chi^2_4$ #

(c) $60(\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu) = (\sqrt{60})^2 (\bar{X} - \mu)^T \Sigma^{-1} (\bar{X} - \mu)$

$= [\sqrt{60}(\bar{X} - \mu)]^T \Sigma^{-1} [\sqrt{60}(\bar{X} - \mu)] \sim \chi^2_4$ #

$(\because \sqrt{60}(\bar{X} - \mu) \sim N_4(0, \Sigma)$ and by part (b))

(d) $59S = (60-1)S = \sum_{i=1}^{60} (X_i - \bar{X})(X_i - \bar{X})^T$

$\sim Wishart_{60}(\cdot \mid \Sigma)$ #

(e) $e_1^T X_i \sim N_1(e_1^T \mu, \lambda_1)$, $e_2^T X_i \sim N_2(e_2^T \mu, \lambda_2)$.

$\Rightarrow Y_i \sim N_2\left(\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} \mu, \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\right)$

$\text{Cov}(e_1^T X_i, e_2^T X_i) = e_1^T \Sigma e_2 = e_1^T \lambda_2 e_2 = 0$.

And by (a), $\bar{Y} \sim N_2\left(\begin{pmatrix} e_1^T \\ e_2^T \end{pmatrix} \mu, \frac{1}{60}\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\right)$. #

(f) Let $a^T = (e_1\ e_2)^T$, $a^T X_i = Y_i$.

$$(60-1) S_Y = \sum_{i=1}^{60} (Y_i - \bar{Y})(Y_i - \bar{Y})^T = \sum_{i=1}^{60} (a^T X_i - a^T \bar{X})(a^T X_i - a^T \bar{X})^T$$

$$= \sum_{i=1}^{60} a^T (X_i - \bar{X})(X_i - \bar{X})^T a$$

$$= a^T \left[\sum_{i=1}^{60} (X_i - \bar{X})(X_i - \bar{X})^T\right] a$$

$$= a^T (60-1) S\, a. \quad \text{So, } S_Y = a^T S a.$$

By part (d) and slide P. 33,

$$59 S_Y \sim \text{Wishart}_{60}(\cdot \mid a^T \Sigma a)$$

$$= \text{Wishart}_{60}\left(\cdot \mid \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}\right)$$ #

Problem 2

Pre-processing

先觀察這筆資料的 dimension，以及將 Species 的名稱改為比較簡要的名字:

```
The dimension of dataset: 344 17
```

```
The simple name for Species: Adelie Gentoo Chinstrap
```

這裡我們只考慮每個 Species 所對應的 variables:“Culmen Length”, “Culmen Depth”。因此，只檢查這兩個變數的 NA 數量狀況:

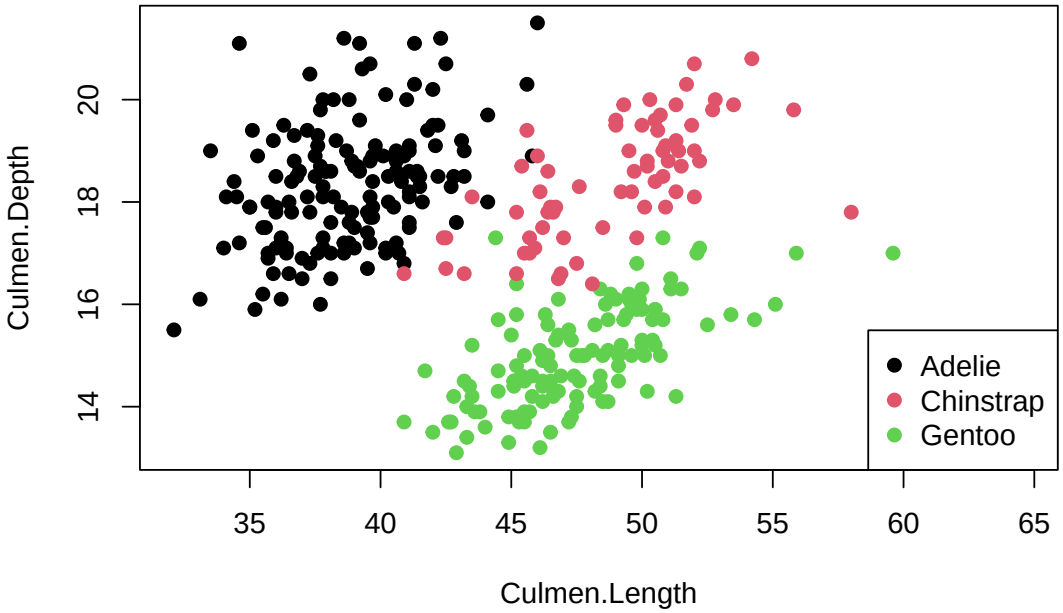
```
The count of NA for Species: 0
```

```
The observations with Length==NA or Depth==NA are 4 272
```

由以上的結果，移除第 4 和第 272 筆觀察樣本。

(a)

移除第 4 和第 272 筆觀察樣本之後，畫出 scatter plot:

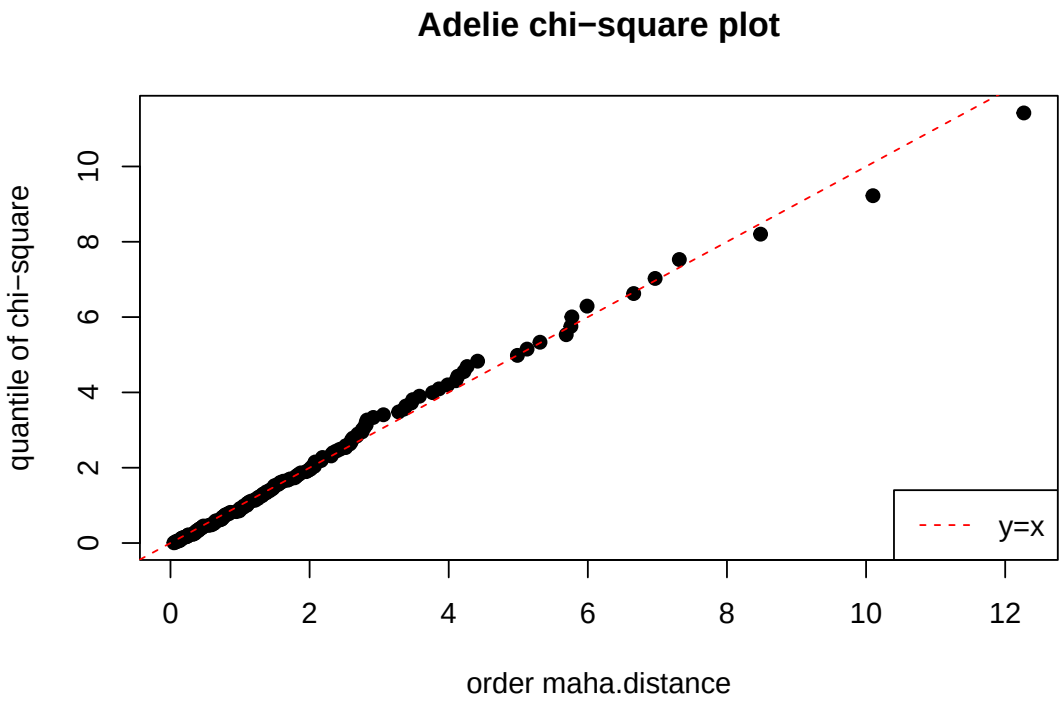


(b)

這裡使用 Slide P.43~P.44 的介紹，劃出每個 Species 的 chi-square plot:

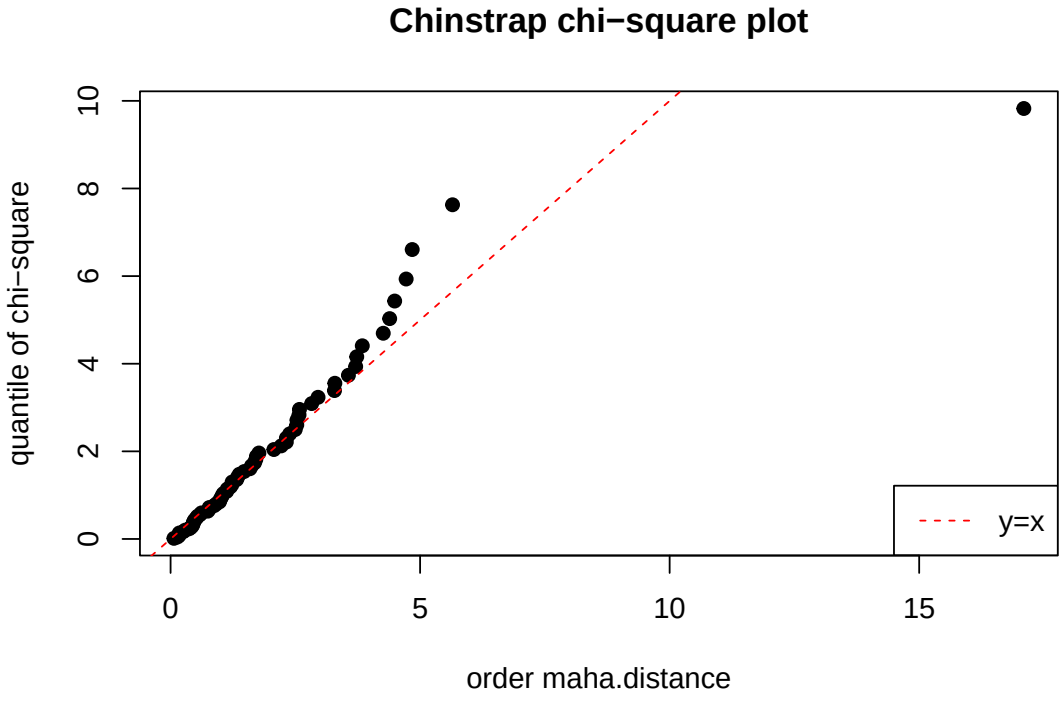
For Adelie,

```
The number of Adelie is 151
```



For Chinstrap,

```
The number of Chinstrap is 68
```



For Gentoo,

```
The number of Gentoo is 123
```

由以上三張 chi-square plot，Species:Adelie 最適合用 Multivariate normal distribution 建模。對於 Species:Gentoo，大部分的點都很貼近 y=x 的紅色虛線 (除了有兩個很明顯偏離的點)，用 Multivariate normal distribution 建模應該也不會不合適。但是，對於 Species:Chinstrap，會發現後面約有 5 個點明顯偏離 y=x 的紅色虛線，即有 $5/68 \approx 7.353\%$ 比例的點是偏離 Multivariate normal distribution ，因此 Species:Chinstrap 可能不太適合用此 distribution 建模。