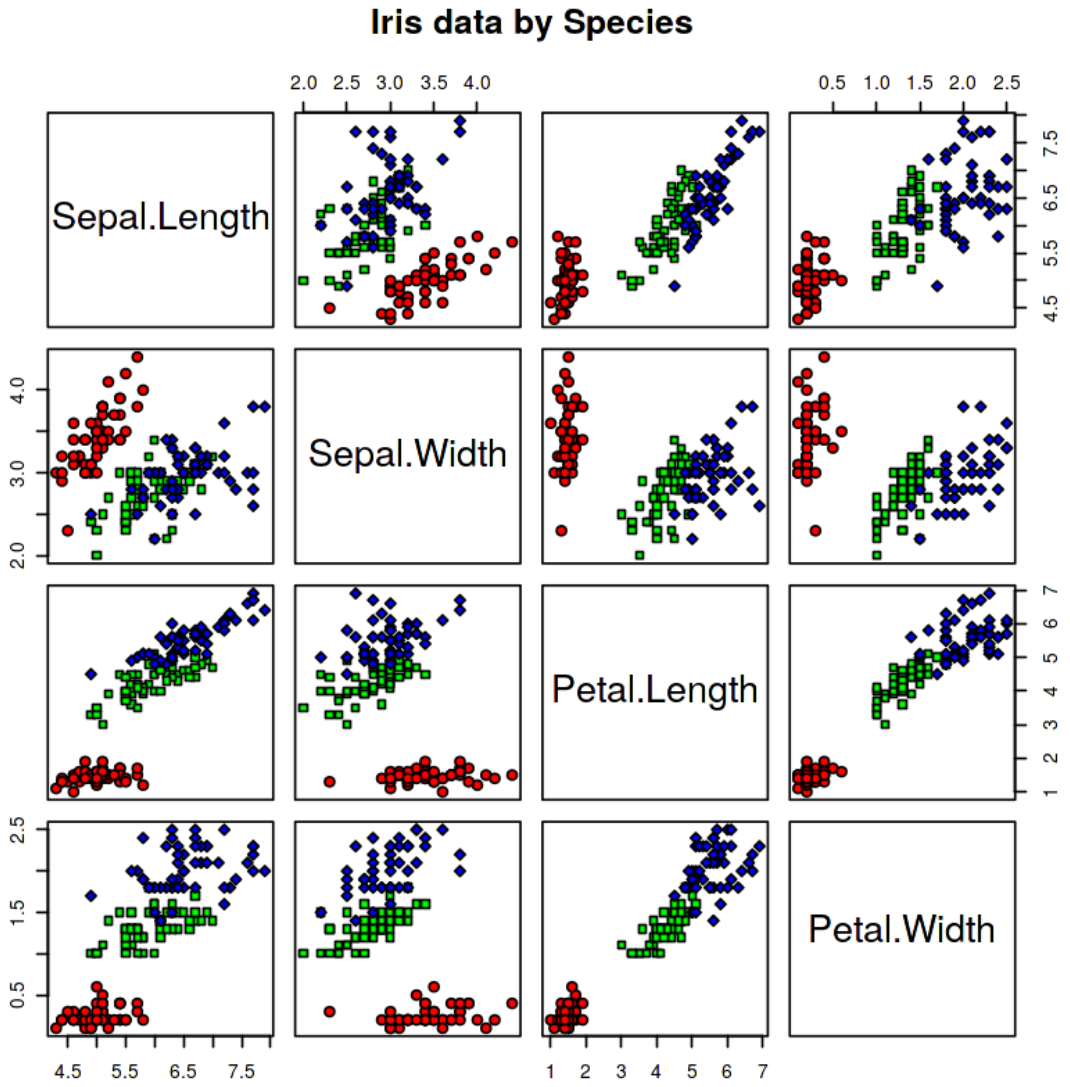
STAT3006 Assignment 2

Ruyun Qi

44506065

Task 1: Exploratory data analysis and basic modelling

Exercise 1:



- Setosa: red circle
- Versicolor: green square
- Virginica: blue diamond

Exercise 2:

	Setosa	Versicolor	Virginica
Sepal L.	5.006	5.936	6.588
Sepal W.	3.428	2.770	2.974
Petal L.	1.462	4.260	5.552
Petal W.	0.246	1.326	2.026

Sample Mean

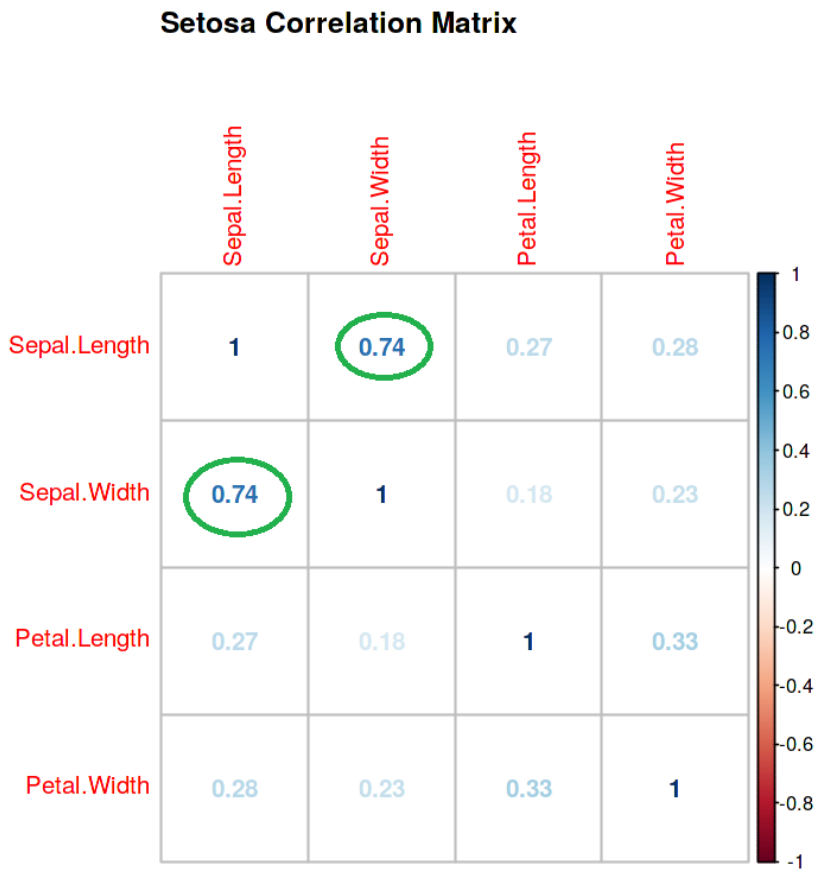
	Setosa	Versicolor	Virginica
Sepal L.	0.3524897	0.5161711	0.6358796
Sepal W.	0.3790644	0.3137983	0.3224966
Petal L.	0.1736640	0.4699110	0.5518947
Petal W.	0.1053856	0.1977527	0.2746501

Standard Deviation

Versicolor and Virginica seem more similar based on their means and standard deviations of each 4 measurements, while Setosa is apparently different from those two.

Exercise 3:

library(corrplot)



Versicolor Correlation Matrix

Virginica Correlation Matrix

By conducting the correlation test of the Sepal Length and Sepal Width for Setosa,

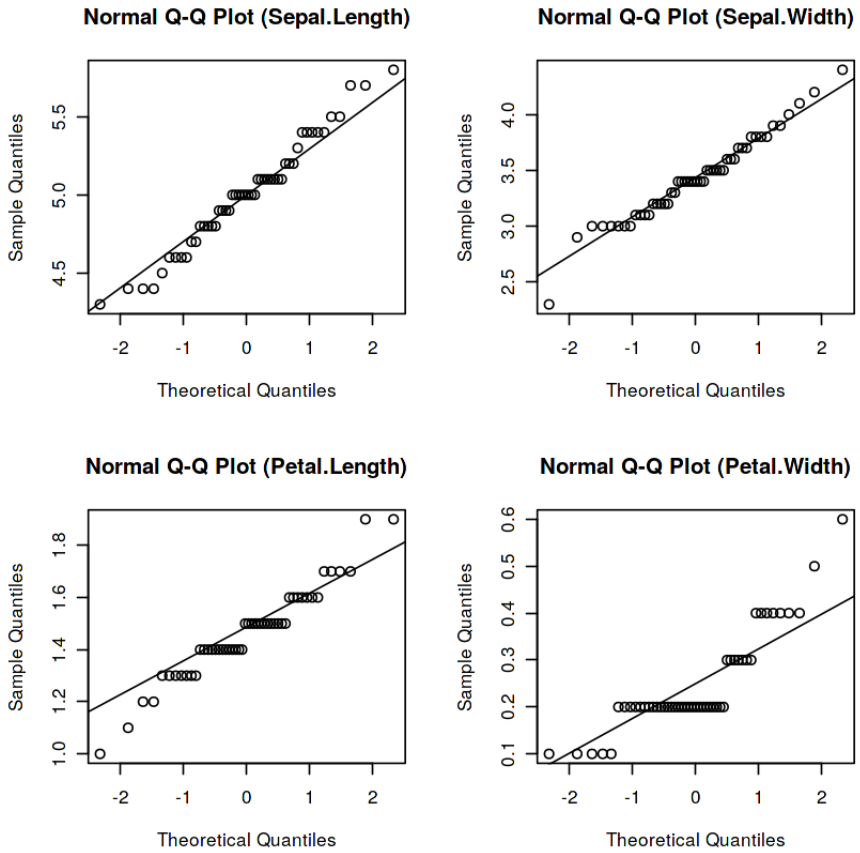
```
95 percent confidence interval:
 0.5851391 0.8460314
sample estimates:
      cor
0.7425467
```

Since $0.5851391 < 0.7425467 < 0.8460314$, the sample correlation confidence interval is not significant at the 0.05 level.

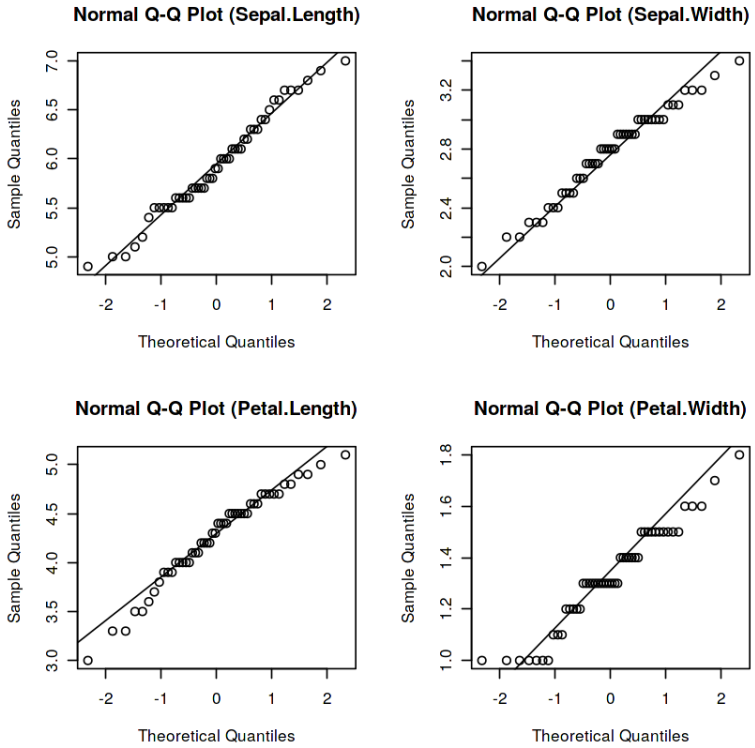
Exercise 4:

library(MVN)

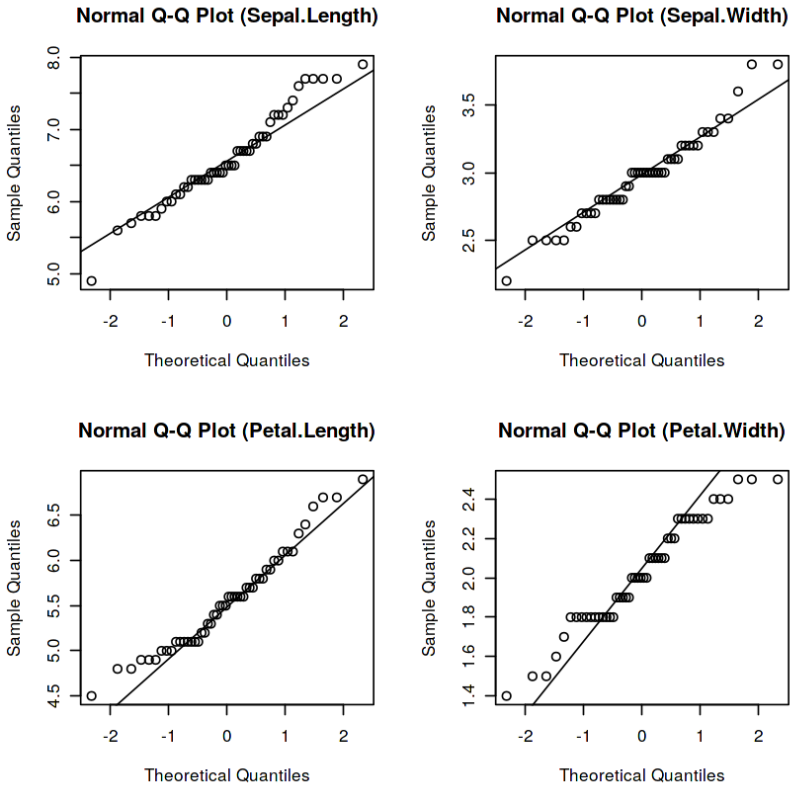
Setosa Q-Q Plot



Versicolor Q-Q Plot



Virginica Q-Q Plot

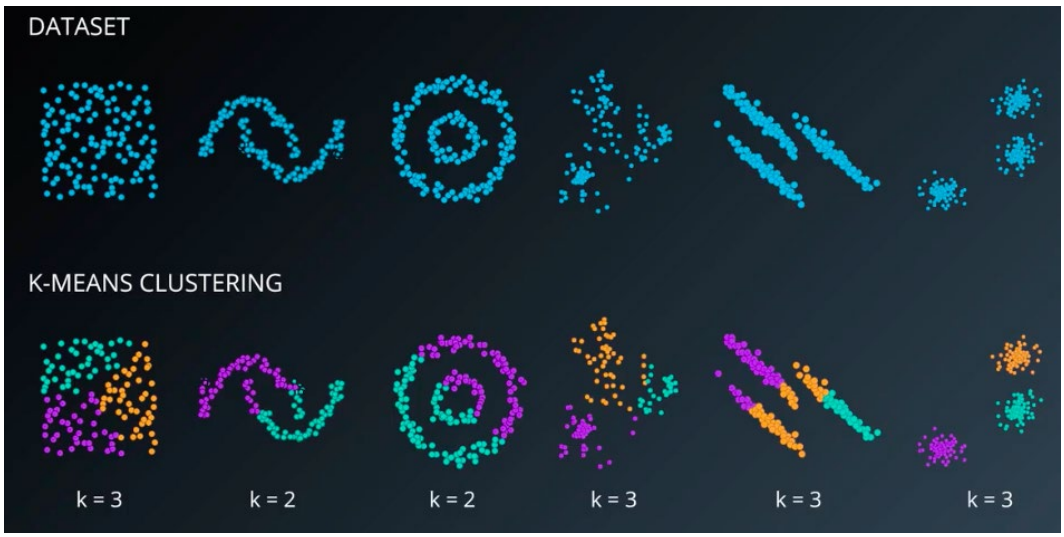


Reference: Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. The R Journal, 6(2), 151–162. https://doi.org/10.32614/RJ-2014-031

- For Setosa, the Petal Length and Width are not normally distributed, especially for the Petal Width.
- For Versicolor, the Petal Width are roughly normal distributed from the Q-Q plot, compared to the other 2 classes.
- For Virginica, the Petal Length and Width are hardly normally distributed, but they look more normal than those for Setosa.

1. In hypothesis testing, he null hypothesis states that the population is normally distributed, against the alternative hypothesis that it is not normally distributed. Therefore, it will be hard to test fairly and accurately if the class is not normally distributed.

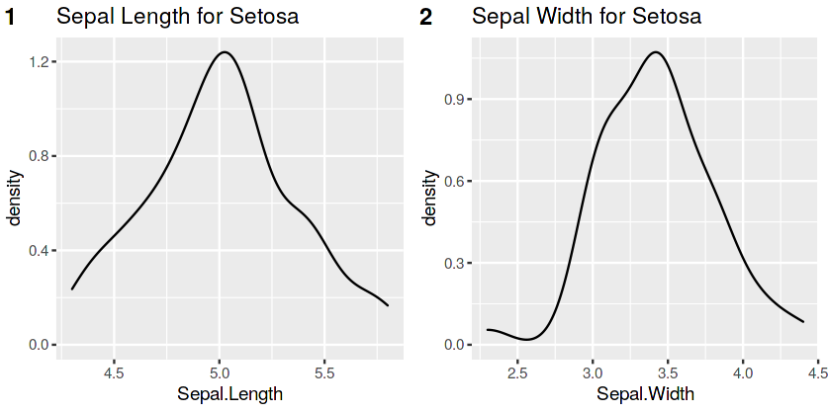
2. In clustering, mostly the data is not required to be normal, but many researchers suggest to standardize the data before clustering. Using K-means clustering as the example, we can see the clustering results of different datasets:

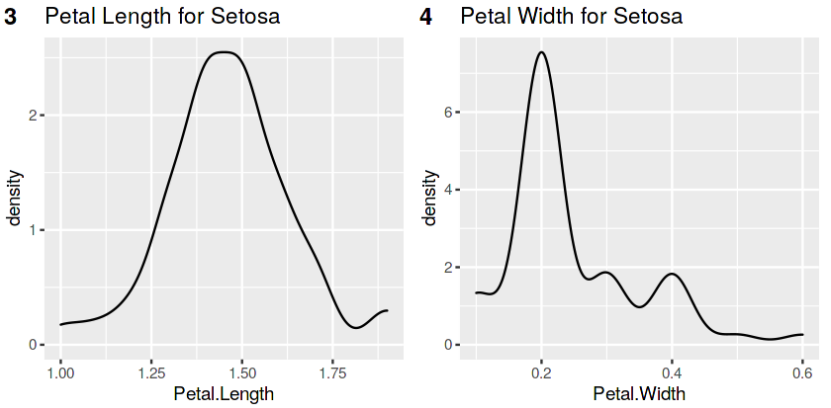


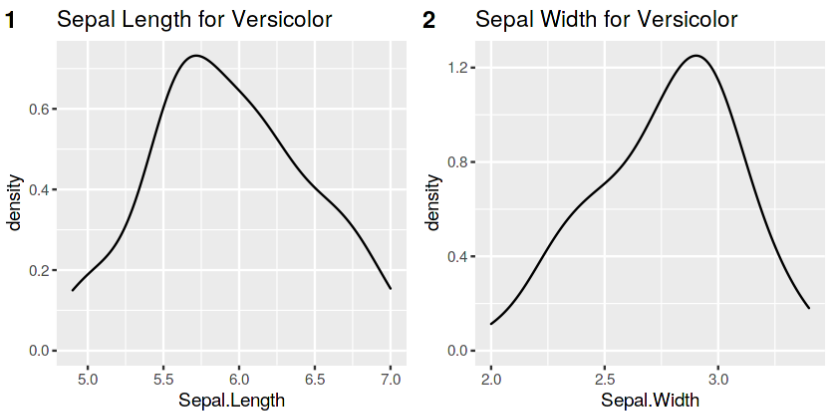
The first dataset is normally distributed, so the clustering result is reasonable. However, in some cases such as the second, third and fifth data sets, the clustering results are affected by their distribution in a bad way. Therefore, non-normality probably causes unstable clustering result.

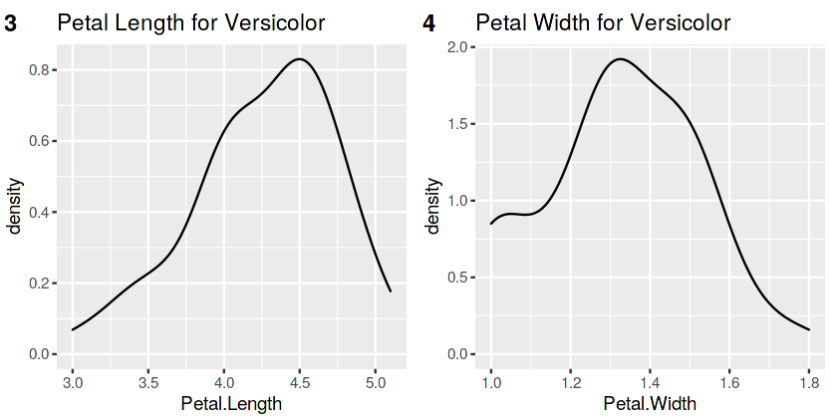
Exercise 5:

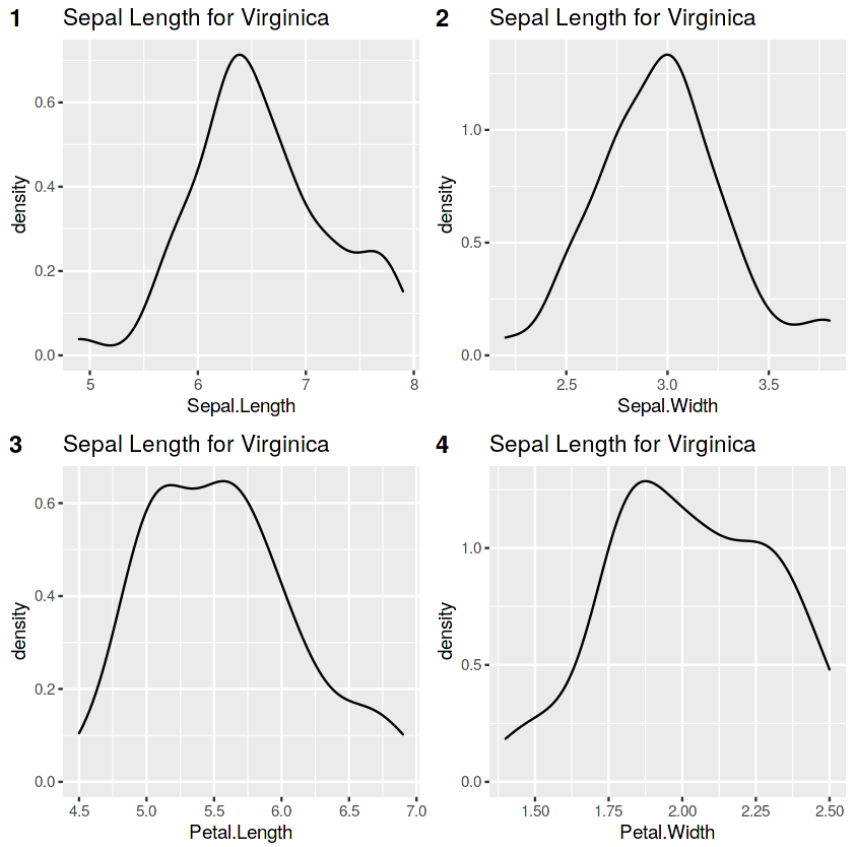
Plots of marginal distributions for each dimension for each class:












```
library(ggplot2)

library(ggpubr)
```

After I have the fitted distributions for 3 classes from MLEs, according to (1.24) in lecture notes, I denote X as the fitted data of Setosa and X with number from 1 to 4 as the data of Sepal Length/Width and Petal Length/Width.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{bmatrix}$$

eg. means of setosa:

Sepal.Length: -2.30926389122033e-16 **Sepal.Width:** 6.21724893790088e-17

Petal.Length: 2.22044604925031e-17 **Petal.Width:** 1.16573417585641e-17

and covariances of setosa:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.12424898	0.099216327	0.016355102	0.010330612
Sepal.Width	0.09921633	0.143689796	0.011697959	0.009297959
Petal.Length	0.01635510	0.011697959	0.030159184	0.006069388
Petal.Width	0.01033061	0.009297959	0.006069388	0.011106122

The marginal distributions are:

$$X_n = N(\mu_n, \Sigma_{nn})$$

- Sentosa

	Mean	Variance
	<dbl>	<dbl>
Sepal.Length	-2.309264e-16	0.12424898
Sepal.Width	6.217249e-17	0.14368980
Petal.Length	2.220446e-17	0.03015918
Petal.Width	1.165734e-17	0.01110612

- Versicolor

	Mean	Variance
	<dbl>	<dbl>
Sepal.Length	3.552714e-17	0.26643265
Sepal.Width	-2.664535e-17	0.09846939
Petal.Length	2.131628e-16	0.22081633
Petal.Width	-6.217249e-17	0.03910612

- Virginica

	Mean	Variance
	<dbl>	<dbl>
Sepal.Length	-1.776357e-17	0.40434286
Sepal.Width	-2.042810e-16	0.10400408
Petal.Length	3.375078e-16	0.30458776
Petal.Width	1.865175e-16	0.07543265

Reference: STAT3006 Tutorial 1 Q4 solutions

Exercise 6:

library(mvtnorm)

First, I denote X_A as the data of Sepal Length and Width, and X_B as for Petal Length and Width. According to (1.30) in lecture notes, I have:

$$\mu_{A|B}(x_B) = \mu_A + \Sigma_{AB}\Sigma^{-1}{}_{BB}(x_B - \mu_B)\,;$$

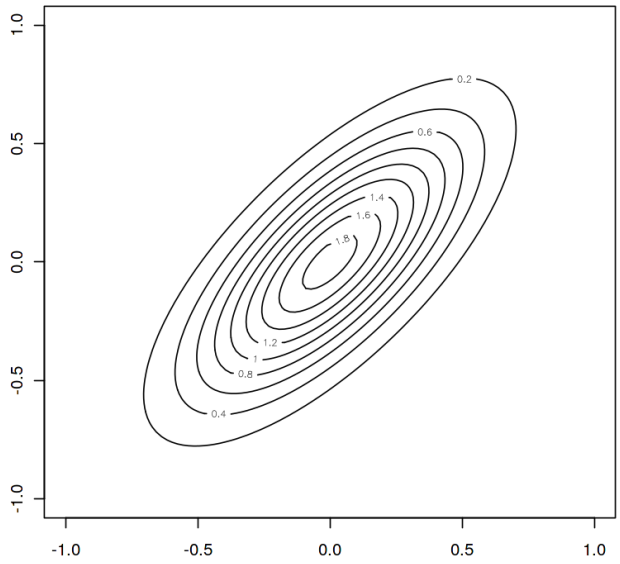
$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma^{-1}{}_{BB}\Sigma_{BA}\,;$$

$$X_A|(X_B = x_B) \sim N(\mu_{A|B}(x_B), \Sigma_{A|B})$$

$$where\ x_B = mean(X_{B)}\,.$$

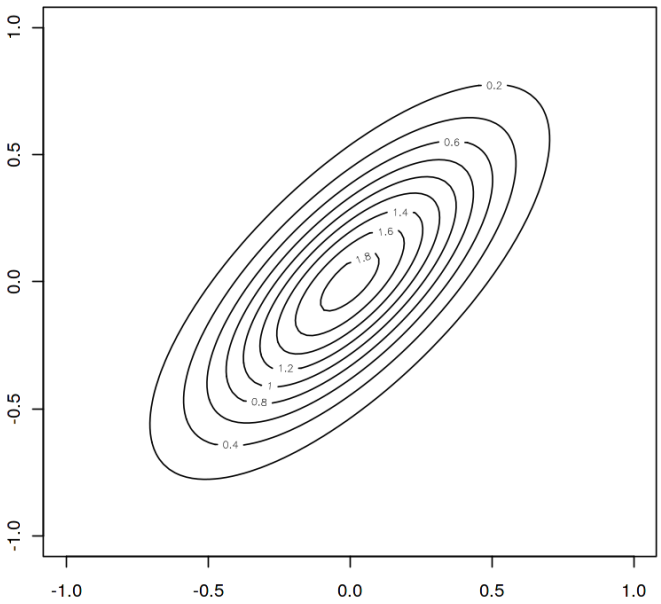
Then I have:

$$\mu_{A|B} = \begin{pmatrix} -2.309264e-16 \\ 6.217249e-17 \end{pmatrix}, \qquad \Sigma_{A|B} = \begin{pmatrix} 0.11036685 & 0.08792769 \\ 0.08792769 & 0.13427458 \end{pmatrix}$$



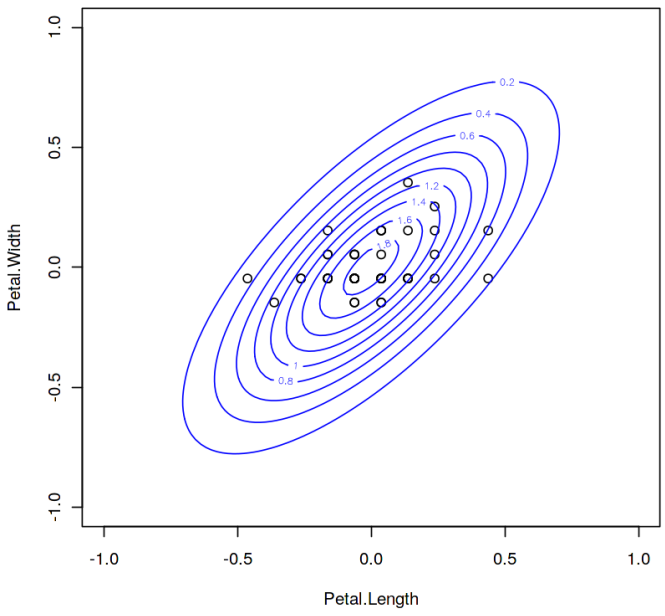
Reversely, I denote X_A as the data of Petal Length and Width, and X_B as for Sepal Length and Width, then I have:

$$\mu_{A|B} = \begin{pmatrix} 2.220446e-17 \\ 1.165734e-17 \end{pmatrix}, \Sigma_{A|B} = \begin{pmatrix} 0.11036685 & 0.08792769 \\ 0.08792769 & 0.13427458 \end{pmatrix}$$



Exercise 7:

The previous contour plot includes the conditional distribution of fitted Sepal Length and Width which are conditioned on the fitted Petal Length and Width. It's a 2D plot which only displays the range of the 2 Sepal measurements. To observe the data with 4 measurements, we can add the points of the 2 Petal measurements in the sample plot.

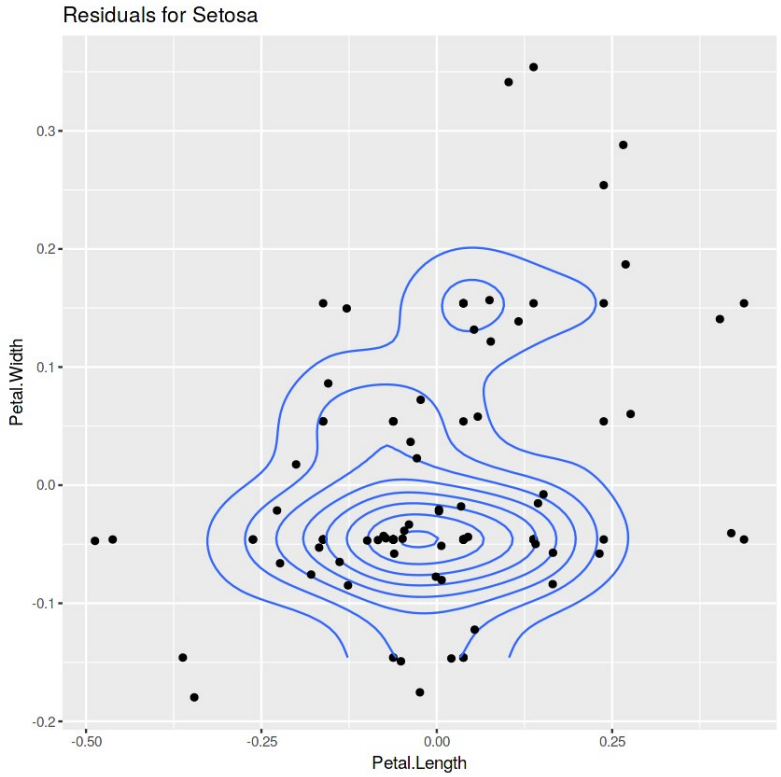


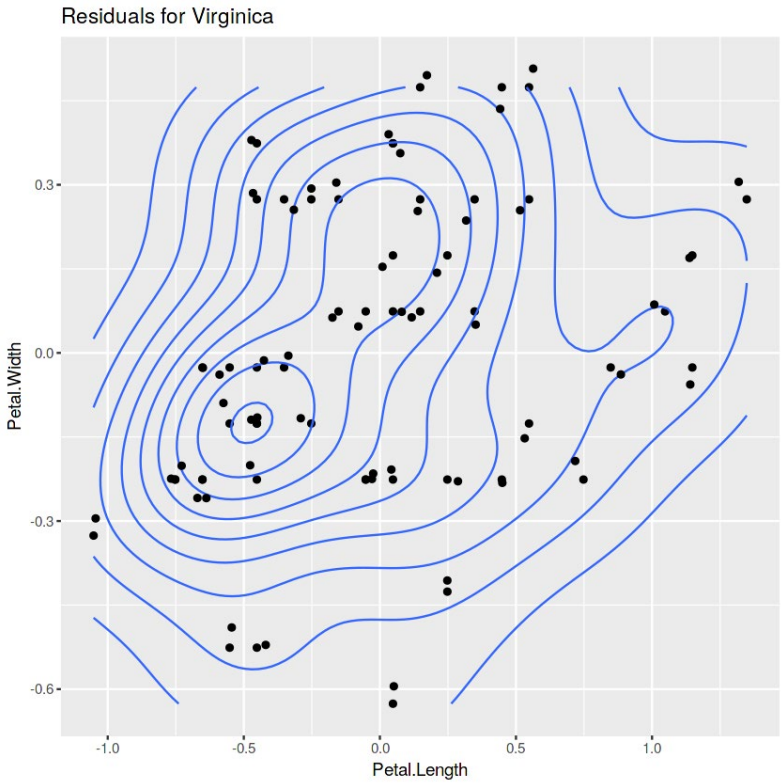
It’s easier to check whether our conditional distribution is rationally formed by looking at the points and the contour. Since the points are inside the contour not separated away from it, the conditional distribution should be formed correctly. However, we can hardly explain the relationship between the conditional distribution and its dependences from this plot overlapping two 2D-plots together.

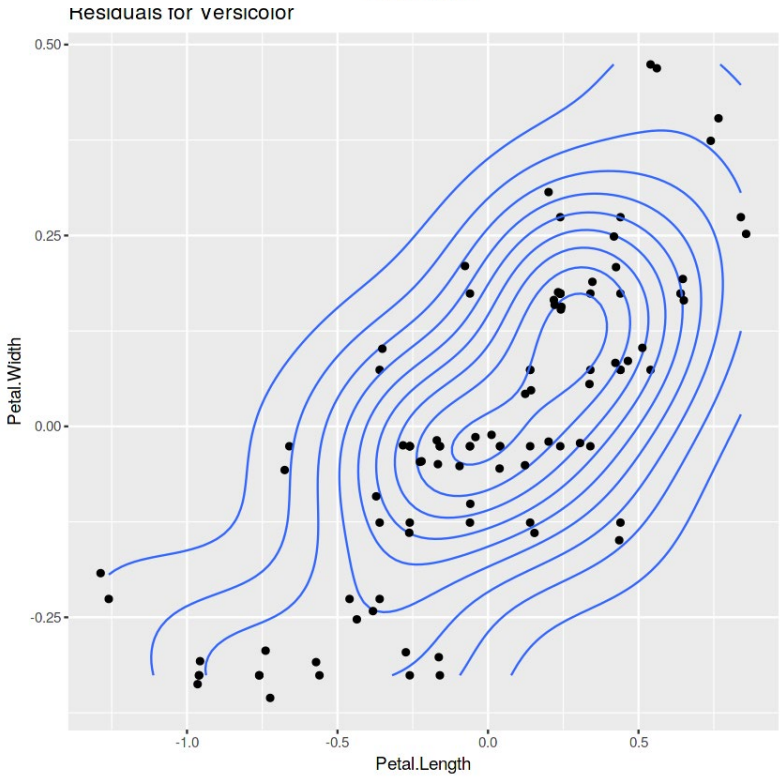
Exercise 8:

library(ggplot2)

The residual is calculated from the differences between the observed values of the fitted distributions and the predictive values which is the mean vectors of the conditional distributions.







We can see that the residuals for conditioned Petal Length and Width for Setosa and Virginica are not normally distributed, but those for Virginica look more normal than those for Setosa. While those for Versicolor seems roughly normally distributed. It is similar with the results of my Q-Q plots in question (iv).

Exercise 9:

library("HDMD")

Mahalanobis distances among the 3 species:

	Setosa	**Versicolor**	**Virginica**
Setosa	0.0000	91.65640	178.01916
Versicolor	91.6564	0.00000	14.52879
Virginica	178.0192	14.52879	0.00000

Three clusters are created for the 3 classes. Each cluster maintains a mean vector and an inverse Covariance matrix. By calculating the Mahalanobis distance, we can see the differences among the 3 species. The pair of Versicolor and Virginica has a small distance compared to the other 2 pairs. It matches the observations of the bivariate plots in part(i). Petal Length and Width discriminate best among the 3 species from all previous observations.

Task 2: Hypothesis testing

Exercise 1:

After applying manova procedure and Hotelling T^2 test, we have:

```
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
Species    1           3.6273   86.148      4     95 < 2.2e-16 ***
Residuals 98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

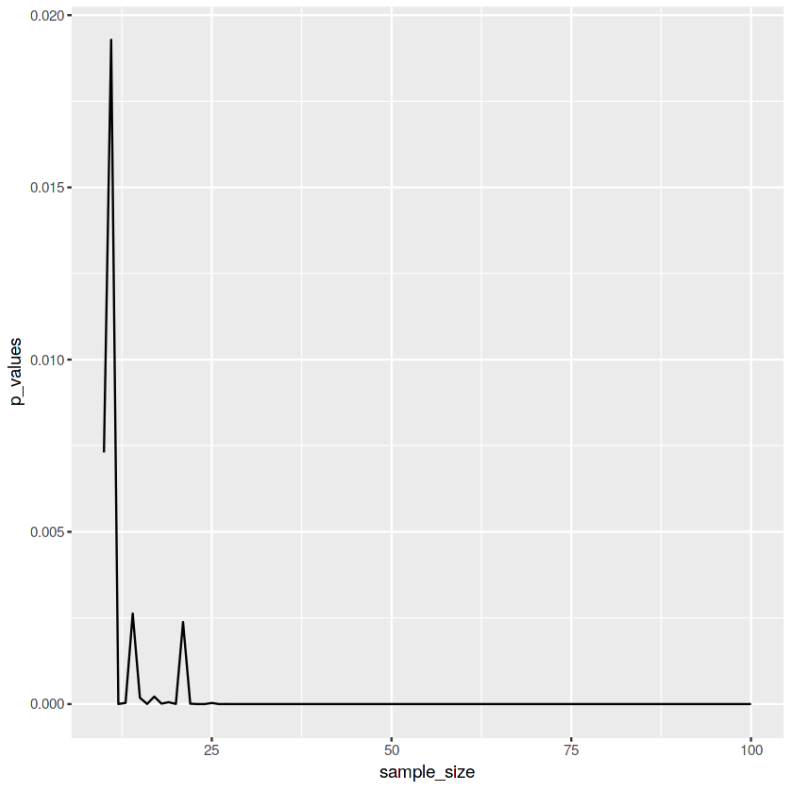
When alpha = 0.05, the critical value from F-statistic table is F(4, 95) = 2.46749362.

The probability of the test F-statistic greater than the critical value is less than 2.2e-16 which is extremely small. Therefore, there is a significant difference.

Exercise 2:

library(ggplot2)

By randomly choose sample sizes from 10 to 100, their respective p-value are calculated.



Obviously, when the sample size is roughly fewer than 25, the p-values are unreliable.

Task 3: Clustering

Exercise 1:

Assume the normal mixture pdf:

$$\overset{\circ}{f}(x;\theta) = \sum_{h=1}^{H} \pi_h \varphi\left(\frac{x-\mu_h}{\Sigma_h}\right)$$

where φ is the pdf of $N\sim(0,1)$, $\theta = (\mu, \Sigma, \pi)$,

with $\mu = (\mu_1, \dots, \mu_H)$, $\Sigma = (\Sigma_1, \dots, \Sigma_H)$,

$\pi = (\pi_1, \dots, \pi_H)$

$$L(\theta;x) = \prod_{i=1}^{n} \overset{\circ}{f}(x_i;\theta) = \prod_{i=1}^{n} \sum_{h=1}^{H} \pi_h \varphi\left(\frac{x-\mu_h}{\Sigma_h}\right)$$

Introduce $Z = (Z_1, \dots, Z_n)$, Z_i in $\{1, \dots, H\}$

$$(X_i \mid Z_i = z_i) \sim N(\mu_{z_i}, \sigma^2_{z_i})$$

$$L(\theta;x,z) = \prod_{i=1}^{n} \pi_{z_i} \varphi\left(\frac{x_i-\mu_{z_i}}{\Sigma_{z_i}}\right)$$

Suppose θ^k is the current guess for θ.

In E-step, derive the discrete pdf of Z,

$$T(z) = f_{z|x}(z|x;\theta_k) \propto f(x,z;\theta_k)$$

$$T_h(x_i; z^{(k)}) = \frac{\pi_h^{(k)} f_h(x_i; \theta_h^{(k)})}{\sum_{m=1}^{H} \pi_m^{(k)} f_m(x_i; \theta_h^{(k)})}$$

$$Q(\theta; \theta^{(k)}) = E_{\tau_h}\, l(\theta; x, z)$$

$$= \sum_{i=1}^{n} \sum_{h=1}^{H} \tau_h(x_i; \theta^{(k)}) \left[\ln \pi_h + \ln \varphi\left(\frac{x-\mu_h}{\Sigma_h}\right) \right]$$

In M-step, Since $l(\theta; x, z) = Q(\theta; \theta^{(k)}) - E_{\tau_h}$

we should find $\theta_k = \underset{\theta}{\operatorname{argmax}}\, Q(\theta; \theta^{(k)})$

Maximize Q_k with respect to π under constraint

$\sum_h \vartheta_h = 1$, $\vartheta_h \geq 0$, solution to $\nabla Q_k(\theta) = 0$ is

$$\pi_h^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_h(x_i; \theta^{(k)}),$$

$$\mu_h^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_h(x_i; \theta^{(k)})\, x_i}{\sum_{i=1}^{n} \tau_h(x_i; \theta^{(k)})},$$

$$\Sigma_h^{(k+1)} = \frac{\sum_{i=1}^{n} \tau_h(x_i; \theta^{(k)}) (x_i - \mu_h^{(k+1)})(x_i - \mu_h^{(k+1)})^T}{\sum_{i=1}^{n} \tau_h(x_i; \theta^{(k)})}.$$

Exercise 2:

In the condition of spherical covariance matrices such that ① $\Sigma_h = a_h I_p$, GMM detect spherical clusters like k-means. However, when we look at the EM-Algorithm, there are differences between GMM and k-means.

Changes needed to achieve k-means :

② $\pi_h = \frac{1}{H}$, ③ $\tau_h(x_i; z^{(k)}) = 1$

Take this 3 conditions into Q function in last question.

$$Q(\theta; z^{(k)}) = \sum_{i=1}^{n} \sum_{h=1}^{H} \tau_h(x_i; z^{(k)}) \left[-\frac{1}{2} (x_i - \mu_h)^T (x_i - \mu_h) \right]$$

$$= \sum_{h=1}^{H} \sum_{x \in \Theta_h} \left[-\frac{1}{2} (x_i - \mu_h)^T (x_i - \mu_h) \right]$$

In M-step, to maximize Q function, we have

$$\min_{\mu} \sum_{h=1}^{H} \sum_{x \in \Theta_h} \| x_i - \mu_h \|^2$$

which is the objective of k-means.

Exercise 3:

For k-means, use Gap statistic method.

W^k is the value of the within-cluster scatter cost function.

Take B samples of size n, $b = 1, \ldots, B$, we have W_b^k

$\bar{W}^k = \frac{1}{B} \sum_{b=1}^{B} W_b^k$, $Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \log W_b^k - \log W^k$

$s'(k+1) = \sqrt{1 + \frac{1}{B}}$ s.d. $(\log W_1^k, \ldots \log W_B^k)$

When $Gap(k) \geq Gap(k+1) - s'(k+1)$ first appears, this k is the optimal value.

For mixture models, use BIC,

$k^* = \arg\min_k BIC = \arg\min_k -2\mathcal{L}(\hat{\varphi}_k, X_{all}, M_k) + d_k \log n$

Hence select the k with lowest BIC value.

try K from 1 to k

Choose k for K-means

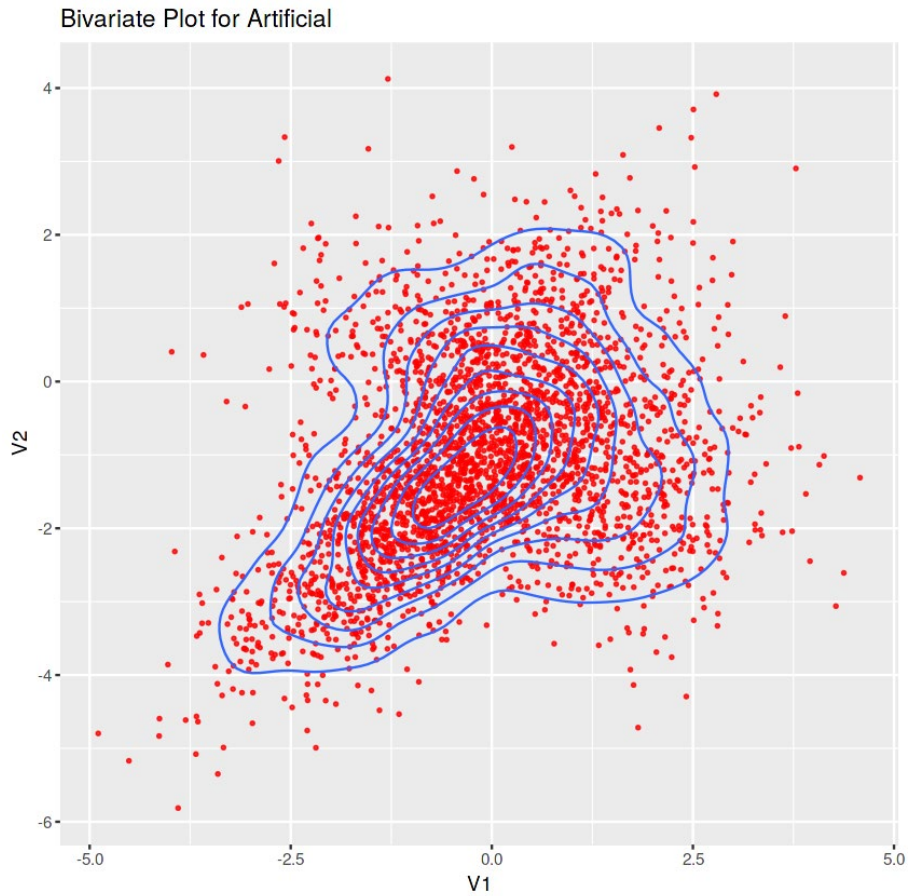
gap_stat <- clusGap(data, FUN=kmeans, nstart = n, K.max = k, B = b)

Choose k for mixture model

BIC <- mclsterBIC (data, G=swq(1,k), modelNames="VVV")

Exercise 4:

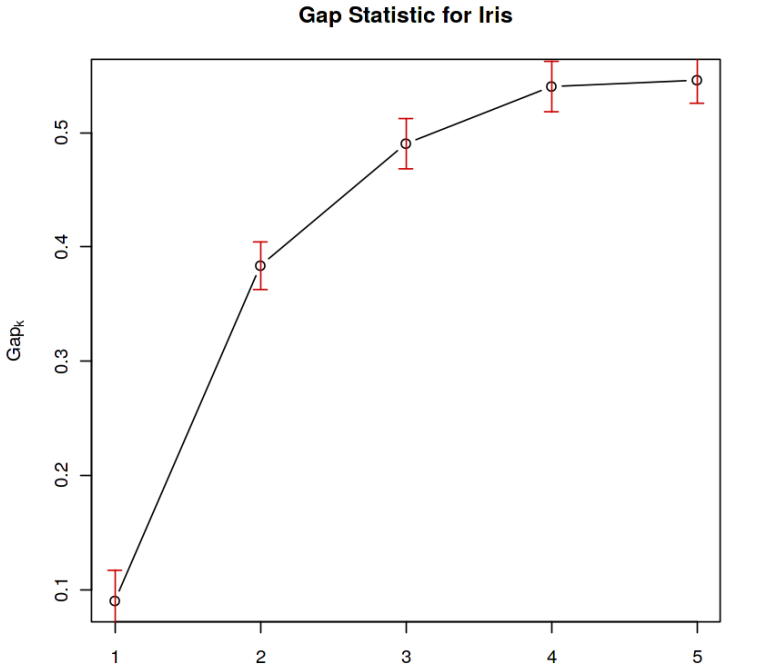
library(ggplot2)

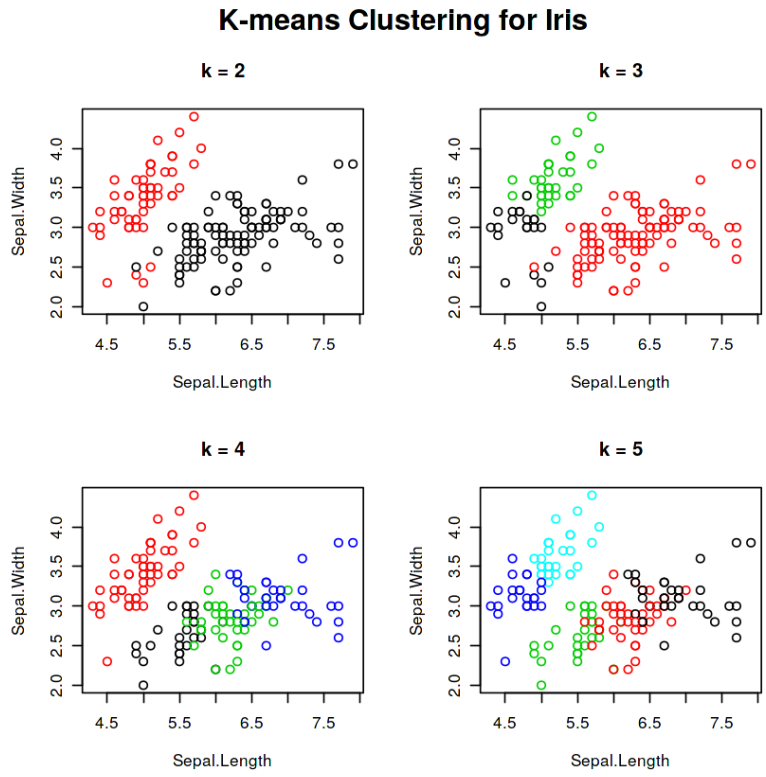


It is difficult to judge how many clusters would be appropriate from this bivariate data, since the data is not separated by several groups like that of Iris in the bivariate plots. I would guess 3 or 4 clusters from the shape of the data.

Exercise 5:

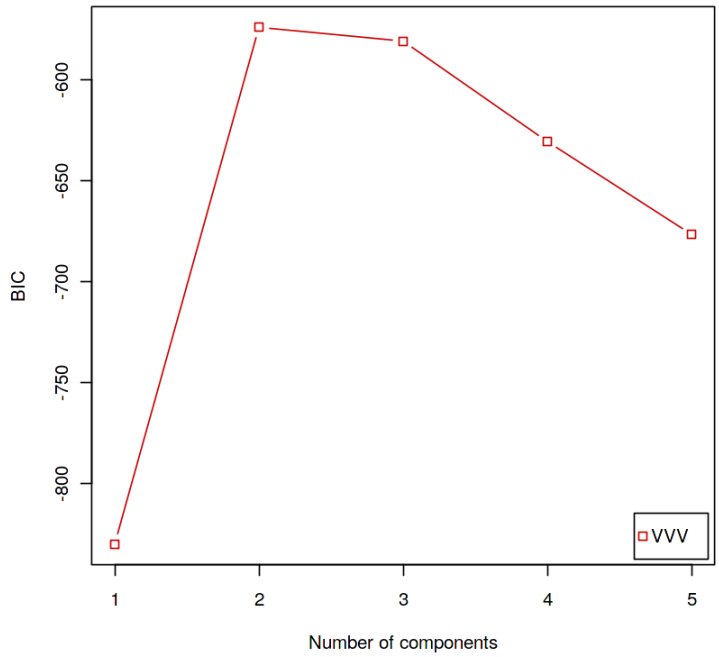
1. K-means Clustering for Iris



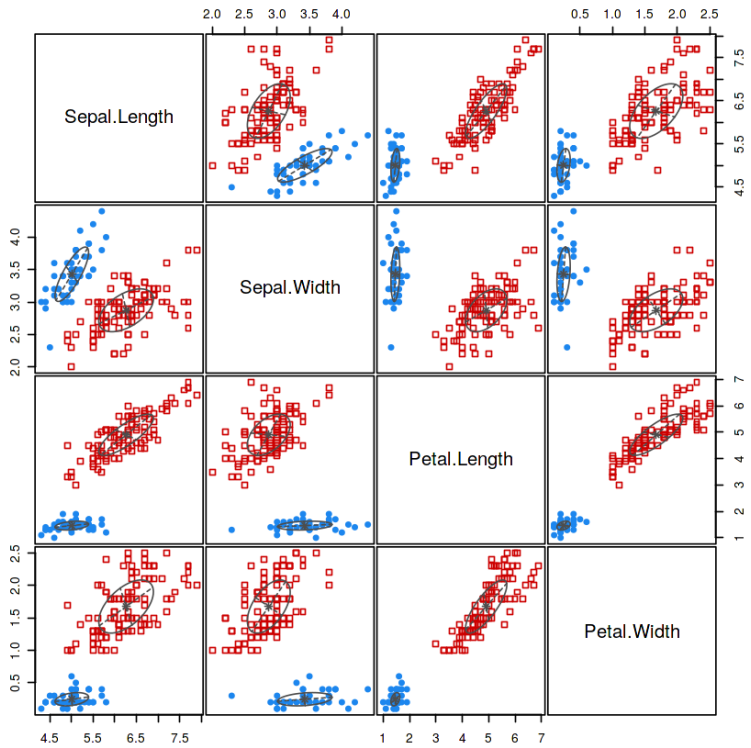


The Gap statistic picks the optimal k value of 4 or 5, bute we can see the K-means clustering does not work well on Iris data from the plots with k = 4 or 5.

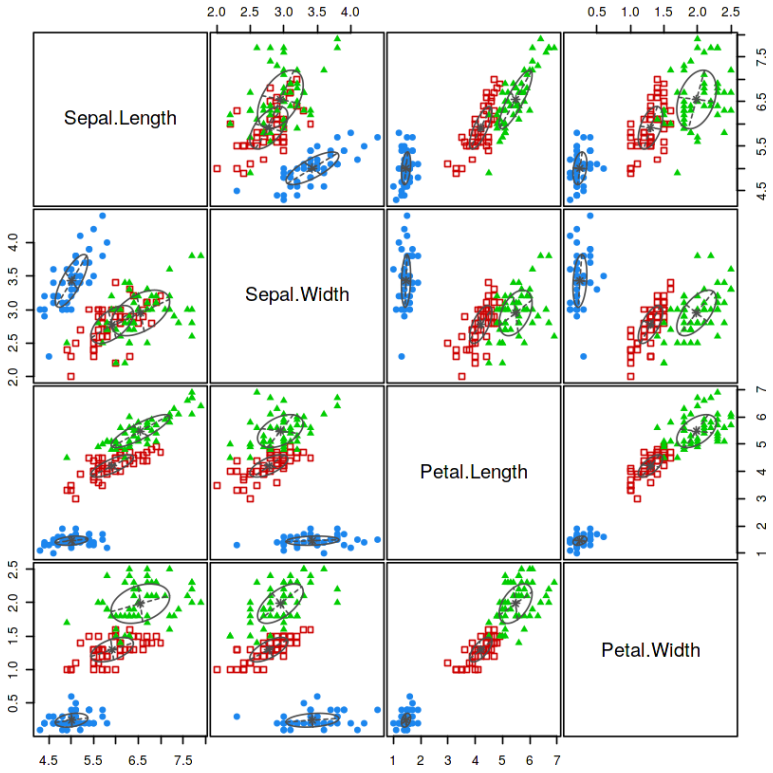
2. GMM for Iris



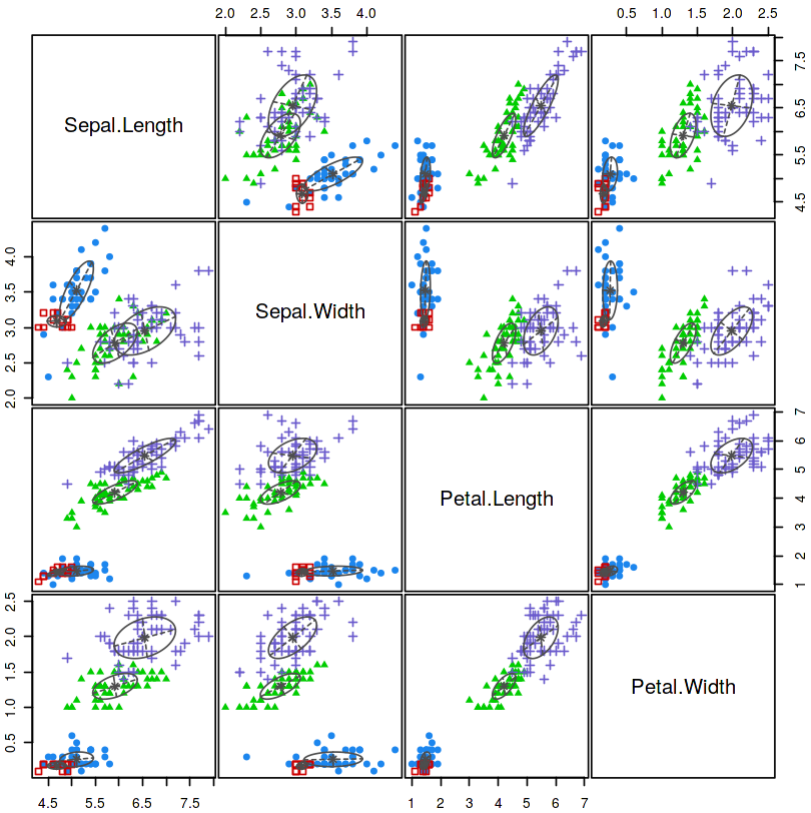
By applying BIC, it seems that k=2 is the optimal value but k=3 is similar with it.



k = 2


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

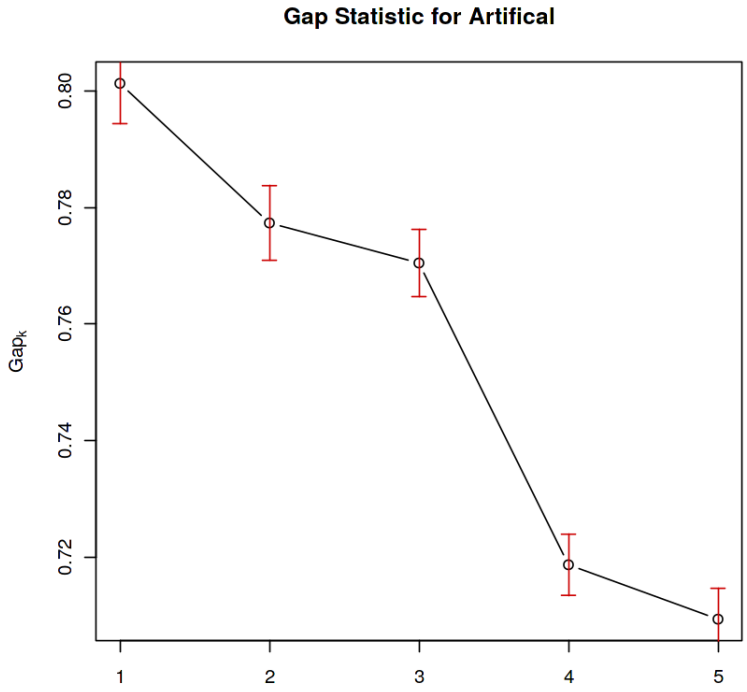
k = 3

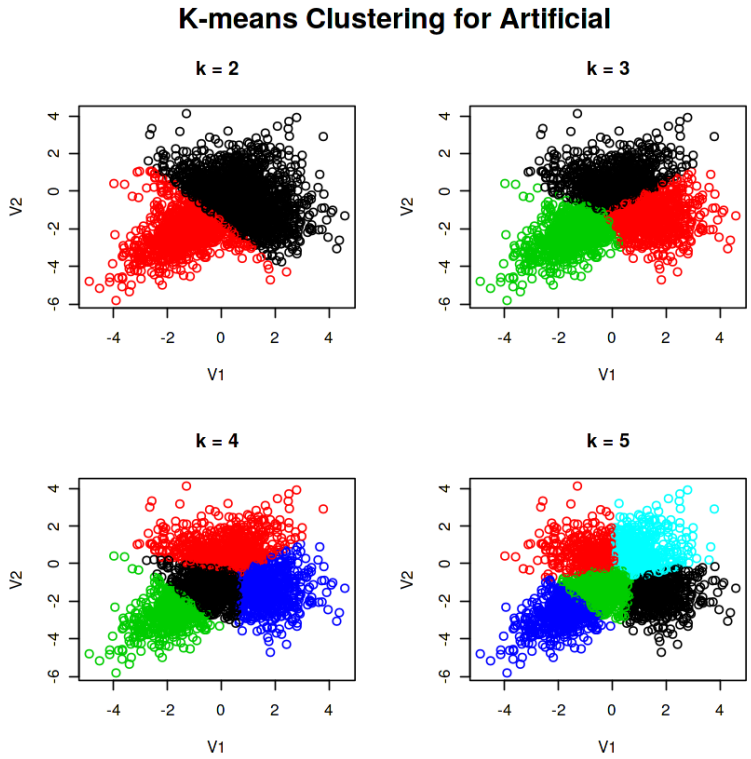

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

k = 4

For k = 2, GMM works well because Versicolor and Virginica are similar. The optimal k value should be 3 and it looks good in GMM clustering. The 3 classes are distinguished and separated mostly.

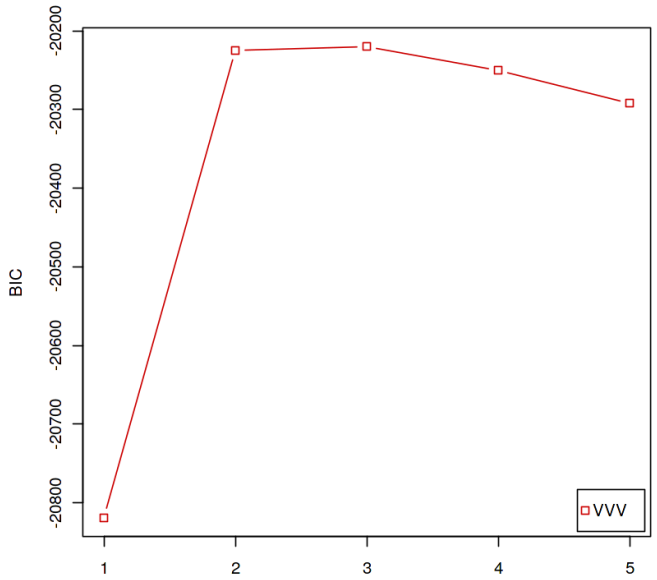
3. K-means clustering for Artificial



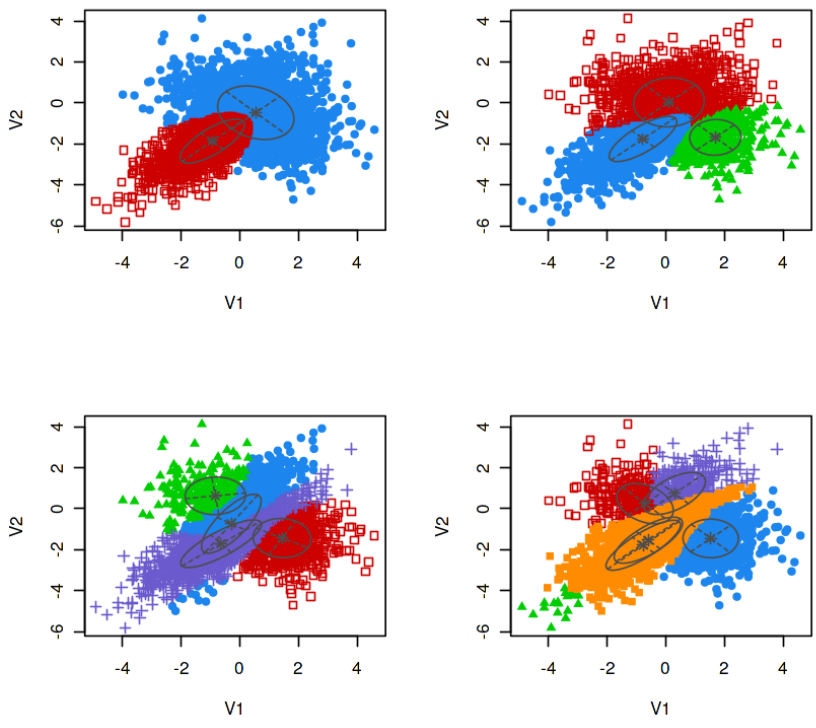


The Gap statistic picks k = 1 as the optimal value and it's not appropriate. It seems not suitable for Artificial which data is crowded without clear segment line. Even though the optimal value is hard to figure out, the K-means clustering works well in this case. The plots with k = 3, 4, 5 all seems appropriate.

4. GMM clustering for Artificial



GMM Clustering for Artificial



By applying BIC, k = 2 or 3 is the optimal value for Artificial. This is proved in the plots of GMM clustering. When k = 4 or 5, the shapes of clusters look weirder rather than those with k = 2 or 3.

In all, the performance of K-means clustering for Iris is not ideal. However, it works well for Artificial which the data is compact. GMM clustering is very suitable for Iris, and it works well for Artificial if the optimal value is chosen appropriately.

```
library(cluster)
library(mclust)
```

Reference: STAT3006 Tutorial 2 Q2(h)

Exercise 6:

Parameter Estimate for Iris:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	4.738095	2.904762	1.790476	0.3523810
2	6.314583	2.895833	4.973958	1.7031250
3	5.175758	3.624242	1.472727	0.2727273

K-means - cluster centers

```
Mixing probabilities:
        1         2         3
0.3333333 0.3005423 0.3661243

Means:
              [,1]     [,2]     [,3]
Sepal.Length 5.006 5.915044 6.546807
Sepal.Width  3.428 2.777451 2.949613
Petal.Length 1.462 4.204002 5.482252
Petal.Width  0.246 1.298935 1.985523

Variances:
[,,1]
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.13320850  0.10938369  0.019191764 0.011585649
Sepal.Width    0.10938369  0.15495369  0.012096999 0.010010130
Petal.Length   0.01919176  0.01209700  0.028275400 0.005818274
Petal.Width    0.01158565  0.01001013  0.005818274 0.010695632
[,,2]
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.22572159  0.07613348   0.14689934  0.04335826
Sepal.Width    0.07613348  0.08024338   0.07372331  0.03435893
Petal.Length   0.14689934  0.07372331   0.16613979  0.04953078
Petal.Width    0.04335826  0.03435893   0.04953078  0.03338619
[,,3]
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.42943106  0.10784274   0.33452389  0.06538369
Sepal.Width    0.10784274  0.11596343   0.08905176  0.06134034
Petal.Length   0.33452389  0.08905176   0.36422115  0.08706895
Petal.Width    0.06538369  0.06134034   0.08706895  0.08663823
```

GMM clustering – probabilities, means, variances

Parameter Estimate for Artificial:

	V1	V2
1	1.32284838	-1.3713515
2	0.02635759	0.3968141
3	-1.34194362	-2.2037897

K-means - cluster centers

```
Mixing probabilities:
        1         2         3
0.4274436 0.4372506 0.1353058

Means:
         [,1]       [,2]      [,3]
V1  0.7059662 -0.9442225 0.1450731
V2 -0.8750902 -1.8657054 0.8646786

Variances:
[,,1]
           V1         V2
V1  1.4902706 -0.4616759
V2 -0.4616759  1.2272822
[,,2]
          V1        V2
V1 1.1719097 0.8578488
V2 0.8578488 1.1580333
[,,3]
          V1        V2
V1 1.8313923 0.3176476
V2 0.3176476 0.9564894
```

GMM clustering – probabilities, means, variances

Using Bootstrap Re-smapling with 95% CI,

```
Mixing probabilities:
              1         2         3
2.5%  0.4423961 0.3301965 0.1505798
97.5% 0.4969207 0.3803167 0.1975180

Means:
[,,1]
              V1        V2
2.5%  -0.9487867 -1.860843
97.5% -0.8030537 -1.694465
[,,2]
             V1         V2
2.5%  0.8000518 -1.1462875
97.5% 1.0202658 -0.9633496
[,,3]
              V1        V2
2.5%  -0.1341267 0.6244378
97.5%  0.1417393 0.8602966

Variances:
[,,1]
            V1       V2
2.5%  1.090605 1.092943
97.5% 1.329952 1.383649
[,,2]
            V1        V2
2.5%  1.171167 0.9656294
97.5% 1.528885 1.2411115
[,,3]
            V1        V2
2.5%  1.483822 0.7688167
97.5% 2.028666 1.1944937
```

Before resampling:

```
Mixing probabilities:
        1         2         3
0.4621395 0.3507423 0.1871182
```

After resampling:

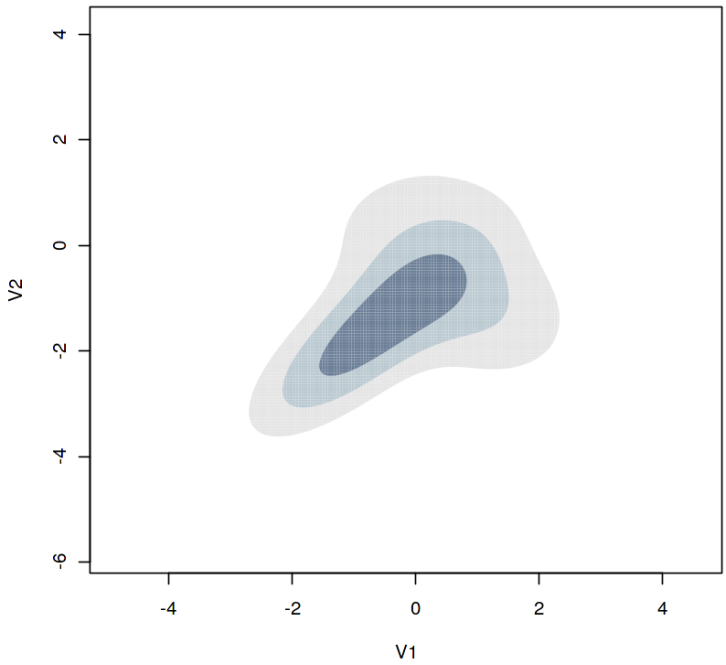
```
Mixing probabilities:
              1         2         3
2.5%  0.4423961 0.3301965 0.1505798
97.5% 0.4969207 0.3803167 0.1975180
```

Some data from Label 1 and 2 were switched to Label 3, since the probability of Label 3 increased. This switch happens when component sizes are not separated enough applying EM algorithm.

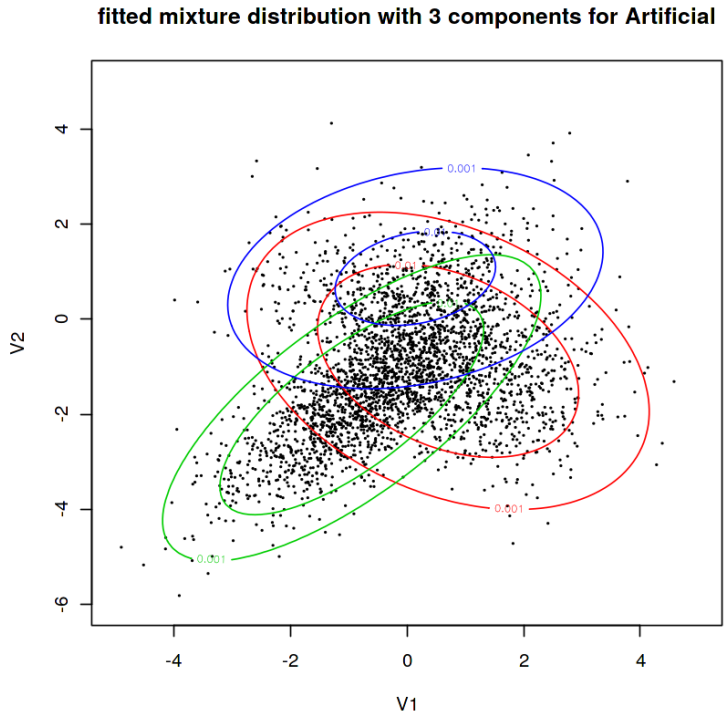
Reference: https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html

Exercise 7:

Contour plot of the overall fitted mixture distribution for Artificial:



Contour plots of the 3 components of the fitted mixture distribution for Artificial:



Reference: STAT3006 Tutorial 2 Q2(e)