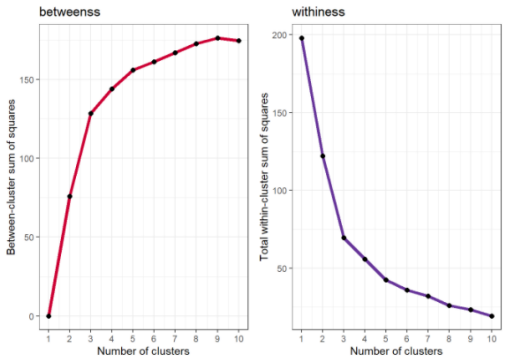
Question 2

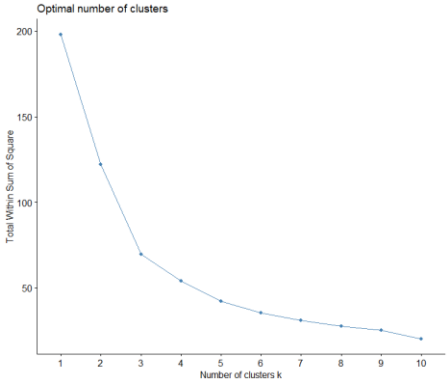
(a)

演算法:

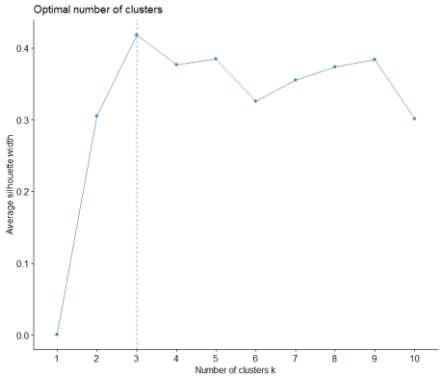
1. 首先生成兩個 bivariate normal distribution.
2. 利用第一題所推導出來的結果，生成 X1, X2, Z1, Z2.
3. Mixing probability 是 0.6，因此利用 runif 的函式生成在(0,1)之間的 100 個值，若取出的值小於 0.6, 則生成 X1, X2，反之，則生成 Z1, Z2.
4. 決定 Kmeans 的分群數:
   A. 利用分群後的 within 值和 between 值，並使得 between 值愈大，within 值愈小。



   B. 利用 Elbow Method 找出一個 n 值使得群內總變異最小

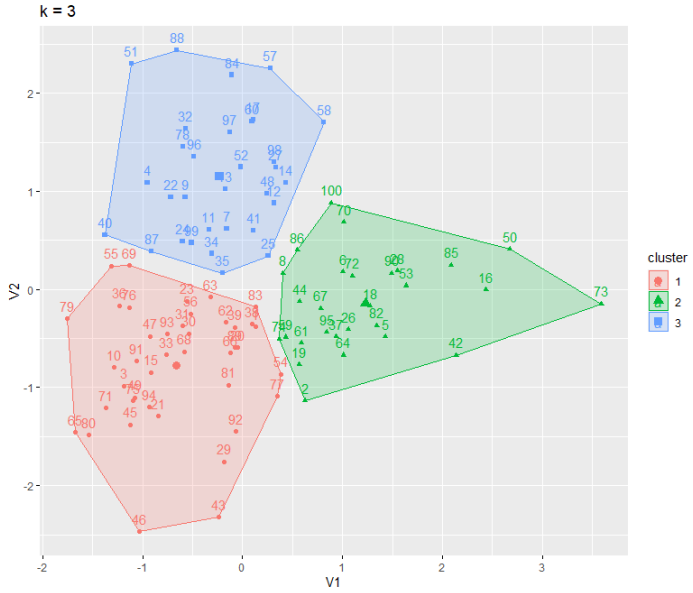


   C. 利用平均側影法(Average silhouette Method)，並算出側影系數(Silhouette Coefficient)，會根據每個資料點(i)的內聚力和分散力，衡量分群的效果(quality)，最終找出最適當的分群數。



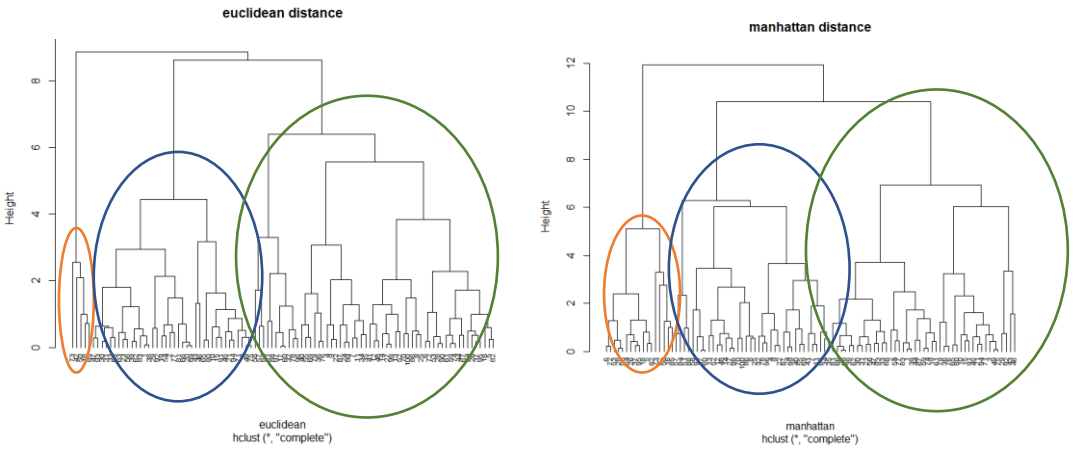
5. 利用以下三個方法，得出 Kmeans 的最佳群數為三。

最後得到 K-means 分群的圖



結論: 三個群之間並沒有交錯的部分，因此設定群數為 3 使得分群的結果十分成功。

(b)



1. 利用歐式距離和曼哈頓距離建立出 Mixed Bivariate Normal distribution 的距離矩陣
2. 並利用套件畫出樹狀圖，根據資料間的距離，來進行階層式分群
3. 左邊的圖為歐式距離，右邊的是曼哈頓距離

結論:

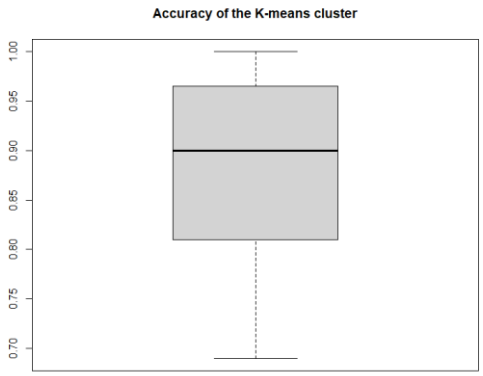
就距離結果來看，不論是歐式距離或曼哈頓距離，都可以很明確地看出分群結果為三群。

Q3

(a)

演算法:

1. 生成 200 個 n 為 100 的 bivariate normal distribution.
2. p1 為 0.6，因此利用 runif 的函式生成在(0,1)之間的 100 個值，若取出的值小於 0.6, 則生成 X1, 反之則生成 X2
3. 利用 for 迴圈，計算 200 個 random sample 各自的 K-means，設置分群數為 2，並計算準確率，算式為正確分群個數/總個數(100)
4. 總共會生成 200 個準確率值
5. 畫出這些值的 boxplot。



結論: 從準確度分布的 boxplot 可以看出，利用 K-means 做分群的準確度為大約 90%。

(b)

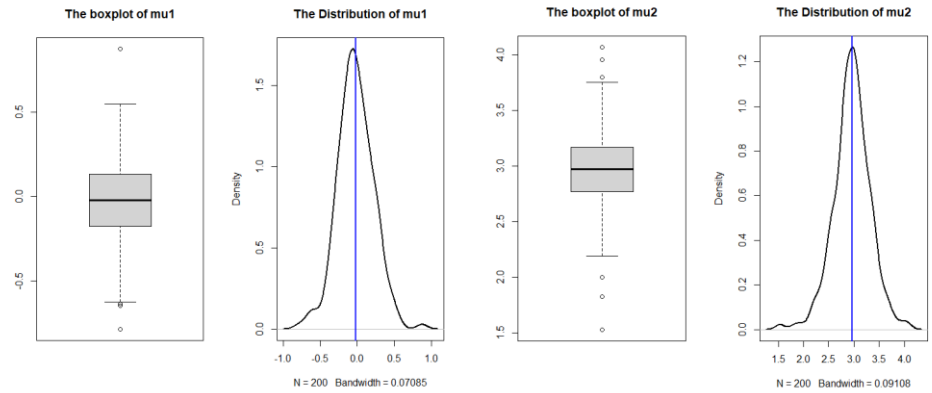
演算法:

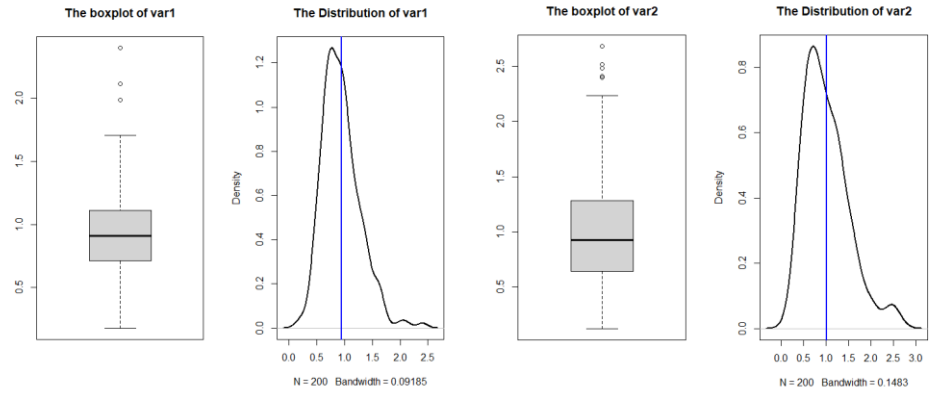
1. 生成 200 個 n 為 100 的 bivariate normal distribution.
2. 首先對每一個 random sample 進行 k-means，以獲得資料的硬標籤。有了這些硬標籤，我們就可以用 MLE 來估計初始化的成分參數
3. 我們現在必須確定資料點（xi）屬於 cluster 的概率是多少，這是 MLE 的期望步驟（E-step），我們要計算每個數據點的軟標籤的期望值。並利用 Bayes' rule 算出後驗概率
4. 得出後驗概率之後，我們可以重新估計我們的成分參數。我們只需要對我們早期指定的 MLE 方程做一點調整。具體來說，在每個方程中，k 個成分的資料點的數量被替換為後驗概率 P（xi∈kj|xi）。
5. Log likelihood 越大=模型參數越適合資料
6. 為了測試收斂性，我們可以在每個 EM 步驟結束時計算對數 Log likelihood，然後測試它是否比上一個 EM 步驟有 "顯著 "變化。如果有，那麼我們就重複另一個 EM 步驟。如果沒有，那麼我們就認為 EM 已經

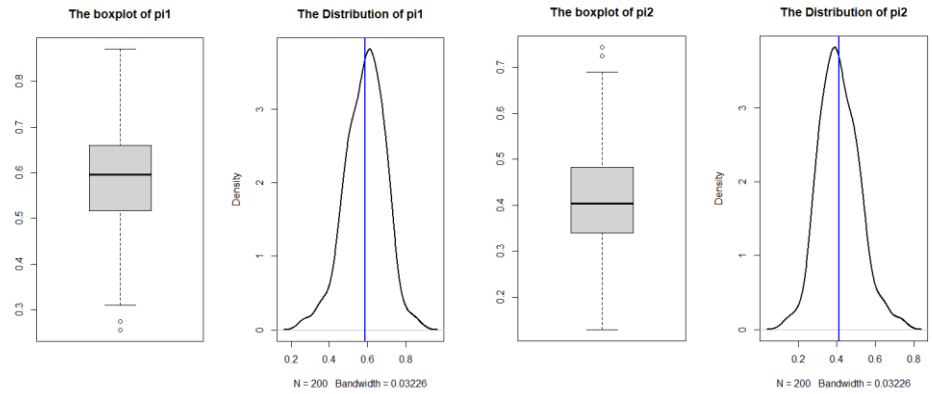
收斂了，那麼就會得到最終參數。

7. 得到 200 個參數組之後，我畫出了各個參數的 boxplot 和 distribution
8. 除此之外，我也計算出各參數的 95%信賴區間，p-value，平均值，和原本的參數值，並將其合併為一個 data frame.

	original_paraneter	parameter	lower_95CI	upper_95CI	p_value
1	0.0	0.003460573	-0.03041167	0.03733281	0.8414940
2	3.0	2.998622479	2.94867000	3.04857496	0.9569494
3	1.0	0.994504965	0.94769119	1.04131874	0.8182766
4	1.0	0.947407623	0.88271482	1.01210042	0.1126591
5	0.6	0.602091915	0.58821577	0.61596806	0.7679336
6	0.4	0.397908085	0.38403194	0.41178423	0.7679336







結論: 對於所有參數的估計，在 200 個 random sample 之中每個參數的平均

值都很接近原本設定的參數，並且各參數的平均值也都落在 95%信心區間之內，並且 p-value 也都大於 0.05，因此可以說明以 GMM 的方法做 clustering 得到的效果也很良好。