HW 1　Report　　0756604 曹育榕

Q1

利用 python 的 scipy.stats 隨機產生 100 個 N(0,1)的點及 100 個 N(1,1)的點。
(圖 A)

Q2

同樣利用 scipy.stats 將隨機產生的 200 個點 fit 到 N(0.49,1.17)的新的 normal distribution(藍線)。
(圖 B)

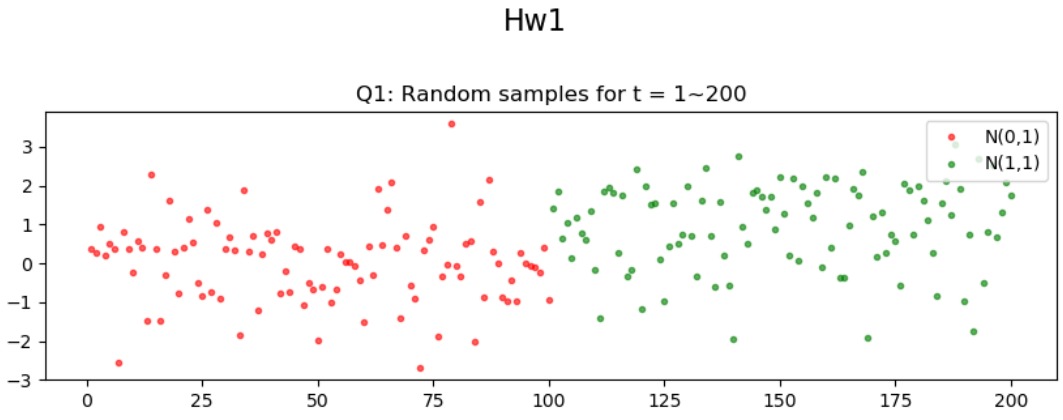

圖 A

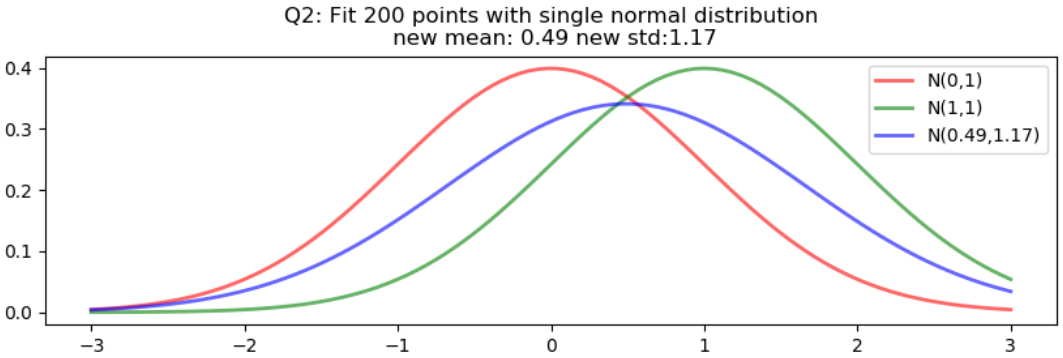

圖 B

Q3

已知群數為 2 群的狀況下，利用 k-means 將 200 個點分為 2 群，下圖 C 中為自動分群的結果。

將這兩群的點各自 fit 到新的 normal distribution，如下圖 D，符合已知的資料分布狀況。

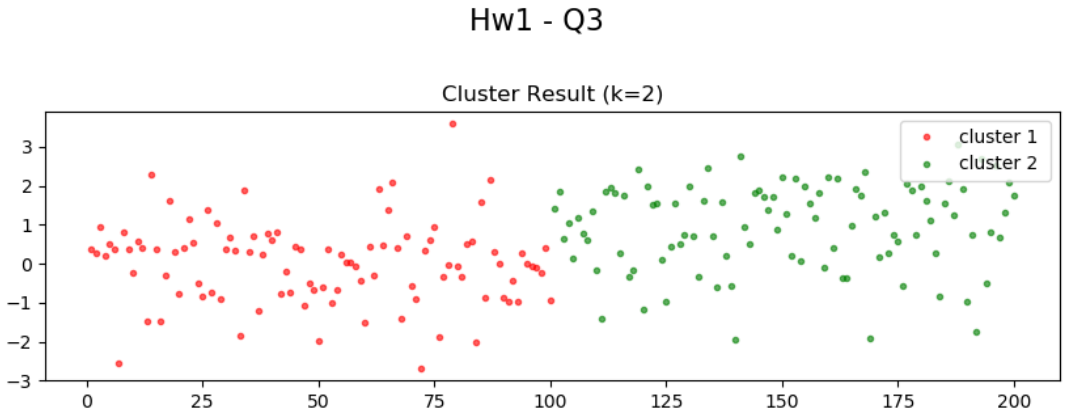

圖 C

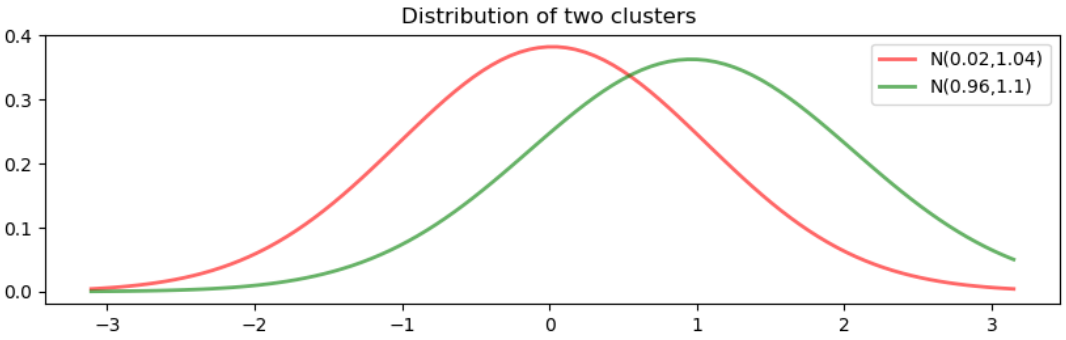

圖 D

Q4

因組數(k)未知，故將 k 分別設為 1~9 組，分別計算 k=1~10 的 k-means 分群結果，利用 calinski_harabaz_score 來衡量分群結果的好壞，決定最佳分群組數。

重複做了 20 次實驗後發現 k=9 時分群的分數皆最高，但明顯不符合實際結果，可能原因為此 200 個資料點的原始分布並無明顯的分群跡象，利用 k-means 方法在分群時 k 越大 score 也會越高，造成 overfitting 的問題。

```
1. best k = 9
2. best k = 9
3. best k = 9
4. best k = 9
5. best k = 9
6. best k = 9
7. best k = 9
8. best k = 9
9. best k = 9
10. best k = 9
11. best k = 9
12. best k = 9
13. best k = 9
14. best k = 9
15. best k = 9
16. best k = 9
17. best k = 9
18. best k = 9
19. best k = 9
20. best k = 9
```

圖 E，Q4 實驗結果

Q5

當已知只有 1 個 cutting point 時將每個時間點(T=1~200)都視為 cutting point，分別針對每個 cutting point 將資料切成兩群，再分別對兩群的資料做 normal distribution test，若兩群資料皆符合 p-value > 0.05，表示兩群皆為 normal distribution，造成這樣分群結果的 T 值就是可能的 cutting point。對於可能的 cutting points 造成的分群結果再取 MLE 得到最佳的 cutting point。

重複做 20 次實驗結果如下，約有一半的實驗能得到 cutting point 在 100 左右的結果，其餘則不太穩定。

```
1. best cut point = 148
2. best cut point = 99
3. best cut point = 12
4. best cut point = 17
5. best cut point = 107
6. best cut point = 26
7. best cut point = 145
8. best cut point = 101
9. best cut point = 150
10. best cut point = 21
11. best cut point = 190
12. best cut point = 96
13. best cut point = 91
14. best cut point = 190
15. best cut point = 86
16. best cut point = 180
17. best cut point = 102
18. best cut point = 180
19. best cut point = 168
20. best cut point = 101
```

Q6

當 cutting point 的數量未知時，用 ruptures library 提供的 change point detection 方法，設定 model = 'normal'，penalty level = 10 來偵測 cutting point 的位置。

重複 20 次實驗結果如下，大部分實驗皆能正確找到 cut point 在 T = 100 處。

```
1. number of cut points = 2 cut point = [65, 95]
2. number of cut points = 1 cut point = [95]
3. number of cut points = 1 cut point = [100]
4. number of cut points = 2 cut point = [100, 110]
5. number of cut points = 1 cut point = [100]
6. number of cut points = 1 cut point = [100]
7. number of cut points = 4 cut point = [90, 100, 110, 115]
8. number of cut points = 1 cut point = [100]
9. number of cut points = 2 cut point = [20, 100]
10. number of cut points = 1 cut point = [100]
11. number of cut points = 1 cut point = [100]
12. number of cut points = 2 cut point = [100, 190]
13. number of cut points = 2 cut point = [80, 100]
14. number of cut points = 1 cut point = [95]
15. number of cut points = 1 cut point = [100]
16. number of cut points = 1 cut point = [100]
17. number of cut points = 1 cut point = [100]
18. number of cut points = 2 cut point = [20, 100]
19. number of cut points = 1 cut point = [100]
20. number of cut points = 1 cut point = [100]
```