**Statistical Computing and Simulation(0521)** 統碩一 106354003 林健宏  
＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

**Expectation-maximization algorithm** (EM-Algorithm)

常用於機率模型中，（1）尋找參數使得likelihood function最大或是（2）最大後驗估計，而這個方法仰賴於無法觀測的隱性變量。

EM-Algorithm包含兩個步驟 : E – step與M – step

E- step

在給定當期參數下，計算整筆資料(包含隱性變量)的likelihood function對隱性變量的期望值。

M- step

讓E- step的期望值最大化，並令E- step的最大期望值為下一期E- step的參數。

重複上述兩個動作，直到參數收斂。

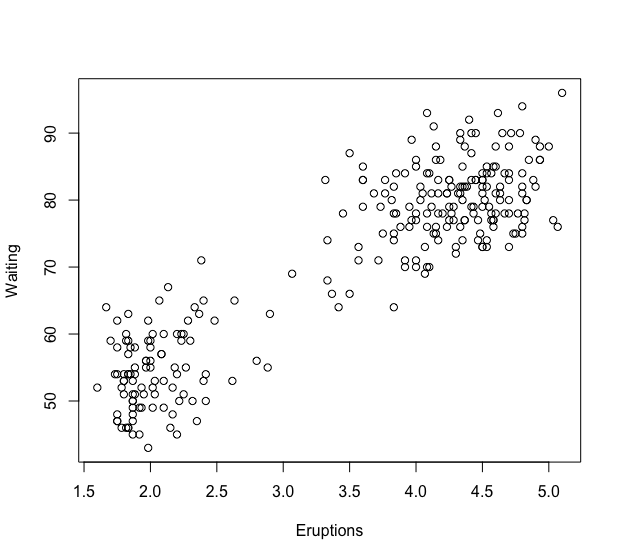
本次使用資料為Mixture normal model的資料 – faithful，黃石公園的噴泉(Old Faithful geyser)資料

其中包含一個成對觀測量：

eruptions: 噴發時間(以分為單位)；

waiting: 前一次噴發結束到下一次噴發開始的間隔時間(以分為單位)。

總共為272筆資料，下表僅呈現前5筆資料。



|  |  |
| --- | --- |
| eruption time(m) | waiting time(m) |
| 3.600 | 79 |
| 1.800 | 54 |
| 3.333 | 74 |
| 2.283 | 62 |
| 4.533 | 85 |

假定資料是由兩個二元常態的線性組合所構成(為eruption time，為waiting time)

其中除了 為觀測已知資料，其餘參數皆為未知，且。

隱性變量為, *j =* 1, 2*, i =* 1~272。

使用EM-algorithm方法去找尋兩組二元常態參數，使用老師在講義推導的最後結果

定義



透過給定起始

經過R的運算後，結果如下

|  |  |  |  |
| --- | --- | --- | --- |
| t |  |  |  |
| 0 | (0.5,0.5) | (4, 60) |  |
| 1 | (0.3558696,0.6441304) | (2.036380, 54.478436) |  |
| 2 | (0.3558721,0.6441279) | (2.036387 ,54.478497) |  |
| 3 | (0.3558727,0.6441273) | (2.036388 ,54.478512) |  |
| 4 | (0.3558728,0.6441272) | (2.036388, 54.478515) |  |
| 5 | (0.3558728,0.6441272) | (2.036388, 54.478516) |  |
| 6 | (0.3558729,0.6441271) | (2.036388, 54.478516) |  |
| 7 | (0.3558729,0.6441271) | (2.036388, 54.478516) |  |
| 8 | (0.3558729,0.6441271) | (2.036388, 54.478516) |  |
| 9 | (0.3558729,0.6441271) | (2.036388, 54.478516) |  |
| 10 | (0.3558729,0.6441271) | (2.036388, 54.478516) |  |

(table 1)

|  |  |  |  |
| --- | --- | --- | --- |
| t |  |  |  |
| 0 | (0.5,0.5) | (3, 70) |  |
| 1 | (0.3558696,0.6441304) | (4.289655, 79.968030) |  |
| 2 | (0.3558721,0.6441279) | (4.289660 ,79.968090) |  |
| 3 | (0.3558727,0.6441273) | (4.289662 ,79.968110) |  |
| 4 | (0.3558728,0.6441272) | (4.289662, 79.968114) |  |
| 5 | (0.3558728,0.6441272) | (4.289662, 79.968115) |  |
| 6 | (0.3558729,0.6441271) | (4.289662, 79.968115) |  |
| 7 | (0.3558729,0.6441271) | (4.289662, 79.968115) |  |
| 8 | (0.3558729,0.6441271) | (4.289662, 79.968115) |  |
| 9 | (0.3558729,0.6441271) | (4.289662, 79.968115) |  |
| 10 | (0.3558729,0.6441271) | (4.289662, 79.968115) |  |

(table 2)

**比較網路上現成的程式所得結果**

(資料比較來源: https://commons.wikimedia.org/wiki/File:Em\_old\_faithful.gif)

皆為迭代10次，起始參數皆設為一致

第一組常態資料比較

|  |  |  |
| --- | --- | --- |
|  | 自己撰寫結果 | 網路運算結果 |
|  | 0.3558729 | 0.3558854 |
|  | (2.036388, 54.478516) | (2.036421, 54.478880) |
|  |  |  |

(table 3)

第二組常態資料比較

|  |  |  |
| --- | --- | --- |
|  | 自己撰寫結果 | 網路運算結果 |
|  | 0.6441271 | 0.6441146 |
|  | (4.289662, 79.968115) | (4.289688, 79.968413) |
|  |  |  |

(table 4)

**結論**

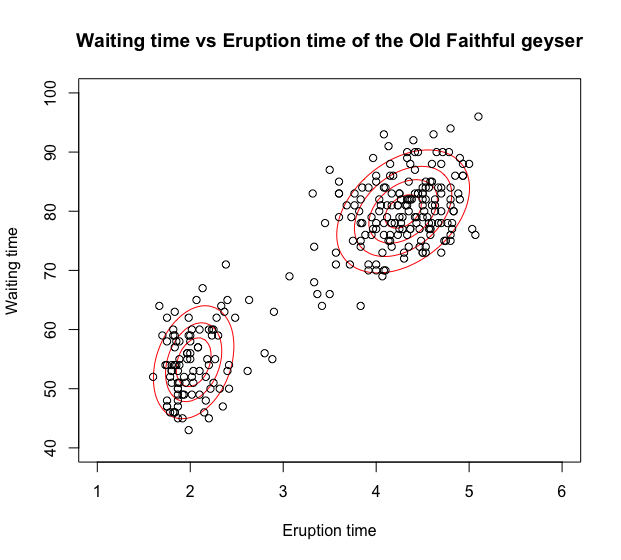
從table 1以及table 2中看出，當t = 3之後各個參數都呈現收斂的狀態。我亦比較網路上他人所撰寫的程式結果(table 3)，得到的結果相差無幾。因此我推估這兩個二元常態為

第一組

, 其線性係數為0.3558729

第二組

, 其線性係數為0.6441271

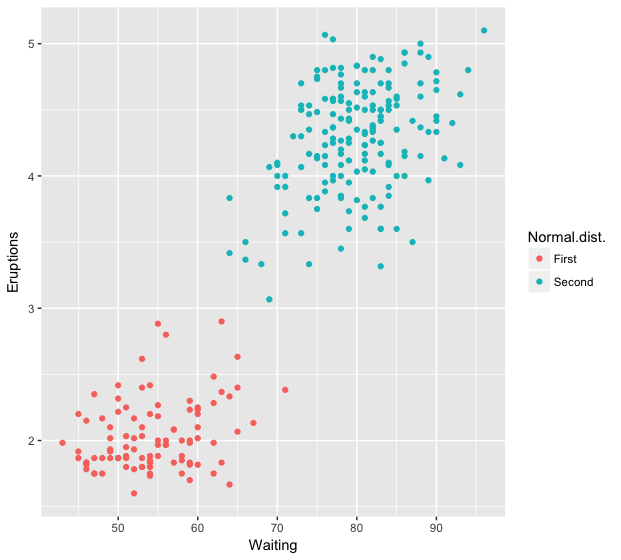


(左下為第一組二元常態，右上為第二組二元常態)

**資料分類分群**

而EM-algorithm也可以用來將資料分類分群，因此我將資料分成兩群，屬於第一組二元常態或是第二組二元常態。分類依據依照將成對的資料分別帶入兩個二元常態，去比較對應之機率大小，並將該筆成對資料判定為機率較大的那群（第一組或地二組二元常態）。

以圖表呈現資料分群



**Appendix**

Code: <https://github.com/kevinpiger/Statistical-Computing-and-Simulation-Nine/tree/master/Hw2>