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。

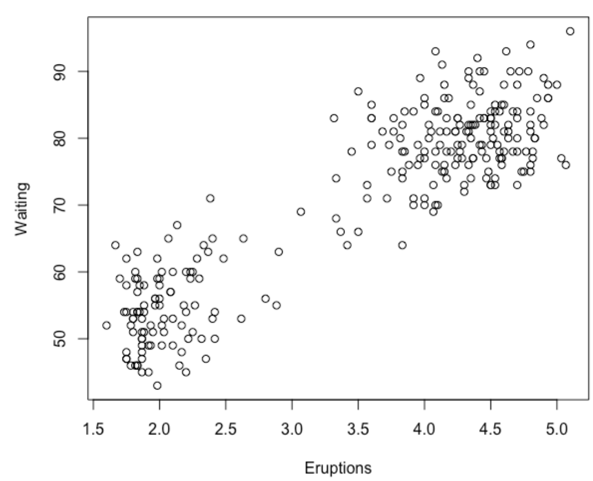
M- step

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(x_1為eruption time，x_2為waiting time)

$$\binom{x_1}{x_2} \sim \alpha_1 BN\left(\binom{\mu_{x_1 1}}{\mu_{x_2 1}}, \begin{pmatrix} \sigma_{x_1 1}^2 & \rho_1\ \sigma_{x_1 1}\sigma_{x_2 1} \\ \rho_1\ \sigma_{x_1 1}\sigma_{x_2 1} & \sigma_{x_2 1}^2 \end{pmatrix}\right) + \alpha_2 BN\left(\binom{\mu_{x_1 2}}{\mu_{x_2 2}}, \begin{pmatrix} \sigma_{x_1 2}^2 & \rho_1\ \sigma_{x_1 2}\sigma_{x_2 2} \\ \rho_1\ \sigma_{x_1 2}\sigma_{x_2 2} & \sigma_{x_2 2}^2 \end{pmatrix}\right)$$

其中除了 $\binom{x_1}{x_2}$ 為觀測已知資料，其餘參數皆為未知，且$\alpha_1 + \alpha_2 = 1$。

隱性變量為$y_i = j$, 若$\binom{x_1}{x_2}$來自 $BN\left(\binom{\mu_{x_1 j}}{\mu_{x_2 j}}, \begin{pmatrix} \sigma_{x_1 j}^2 & \rho_j\ \sigma_{x_1 j}\sigma_{x_2 j} \\ \rho_j\ \sigma_{x_1 j}\sigma_{x_2 j} & \sigma_{x_2 j}^2 \end{pmatrix}\right)$, $j = 1, 2$, $i = 1 \sim 272$。

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θ為所有未知參數集合

At iteration t, $\theta^{(t)} = \{(\alpha_j^{(t)}, \mu_j^{(t)}, \Sigma_j^{(t)}), j = 1, \ldots, k\}$,

$$p(y_i = j | x_i, \theta^{(t)}) = \frac{\alpha_j^{(t)} \phi(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j=1}^{k} \alpha_j^{(t)} \phi(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}$$

$$\alpha_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p(y_i = j | x_i, \theta^{(t)})$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} x_i p(y_i = j | x_i, \theta^{(t)})}{\sum_{i=1}^{n} p(y_i = j | x_i, \theta^{(t)})}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n} (x_i - \mu_j^{(t)})(x_i - \mu_j^{(t)})^T p(y_i = j | x_i, \theta^{(t)})}{\sum_{i=1}^{n} p(y_i = j | x_i, \theta^{(t)})}$$

透過給定起始θ^0值去計算第 t 期參數，並重複運算直到參數收斂。

經過R的運算後，結果如下

t	(α_1, α_2)	(μ_{x_11}, μ_{x_21})	$\begin{pmatrix} \sigma^2_{x_11} & \rho_1\ \sigma_{x_11}\sigma_{x_21} \\ \rho_1\ \sigma_{x_11}\sigma_{x_21} & \sigma^2_{x_21} \end{pmatrix}$
0	(0.5,0.5)	(4, 60)	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
1	(0.3558696,0.6441304)	(2.036380, 54.478436)	$\begin{pmatrix} 0.06916132 & 0.43510139 \\ 0.43510139 & 33.69683083 \end{pmatrix}$
2	(0.3558721,0.6441279)	(2.036387 ,54.478497)	$\begin{pmatrix} 0.06916614 & 0.43515167 \\ 0.43515167 & 33.69717330 \end{pmatrix}$
3	(0.3558727,0.6441273)	(2.036388 ,54.478512)	$\begin{pmatrix} 0.06916730 & 0.43516380 \\ 0.43516380 & 33.69725590 \end{pmatrix}$
4	(0.3558728,0.6441272)	(2.036388, 54.478515)	$\begin{pmatrix} 0.06916758 & 0.43516670 \\ 0.43516670 & 33.69727576 \end{pmatrix}$
5	(0.3558728,0.6441272)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916765 & 0.43516740 \\ 0.43516740 & 33.69728055 \end{pmatrix}$
6	(0.3558729,0.6441271)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916767 & 0.43516757 \\ 0.43516757 & 33.69728171 \end{pmatrix}$
7	(0.3558729,0.6441271)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916767 & 0.43516761 \\ 0.43516761 & 33.69728198 \end{pmatrix}$
8	(0.3558729,0.6441271)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916767 & 0.43516762 \\ 0.43516762 & 33.69728205 \end{pmatrix}$
9	(0.3558729,0.6441271)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916767 & 0.43516762 \\ 0.43516762 & 33.69728207 \end{pmatrix}$
10	(0.3558729,0.6441271)	(2.036388, 54.478516)	$\begin{pmatrix} 0.06916767 & 0.43516762 \\ 0.43516762 & 33.69728207 \end{pmatrix}$

(table 1)

t	(α_1, α_2)	(μ_{x_12}, μ_{x_22})	$\begin{pmatrix} \sigma^2_{x_12} & \rho_1\ \sigma_{x_12}\sigma_{x_22} \\ \rho_1\ \sigma_{x_12}\sigma_{x_22} & \sigma^2_{x_22} \end{pmatrix}$
0	(0.5,0.5)	(3, 70)	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
1	(0.3558696,0.6441304)	(4.289655, 79.968030)	$\begin{pmatrix} 0.1699774 & 0.9407236 \\ 0.9407236 & 36.0474986 \end{pmatrix}$
2	(0.3558721,0.6441279)	(4.289660 ,79.968090)	$\begin{pmatrix} 0.1699706 & 0.9406368 \\ 0.9406368 & 36.0465212 \end{pmatrix}$
3	(0.3558727,0.6441273)	(4.289662 ,79.968110)	$\begin{pmatrix} 0.1699690 & 0.9406159 \\ 0.9406159 & 36.0462859 \end{pmatrix}$
4	(0.3558728,0.6441272)	(4.289662, 79.968114)	$\begin{pmatrix} 0.1699686 & 0.9406109 \\ 0.9406109 & 36.0462293 \end{pmatrix}$

5	(0.3558728,0.6441272)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699685 & 0.9406097 \\ 0.9406097 & 36.0462156 \end{pmatrix}$
6	(0.3558729,0.6441271)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699684 & 0.9406094 \\ 0.9406094 & 36.0462124 \end{pmatrix}$
7	(0.3558729,0.6441271)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699684 & 0.9406093 \\ 0.43516761 & 36.0462116 \end{pmatrix}$
8	(0.3558729,0.6441271)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699684 & 0.9406093 \\ 0.9406093 & 36.0462114 \end{pmatrix}$
9	(0.3558729,0.6441271)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699684 & 0.9406093 \\ 0.9406093 & 36.0462113 \end{pmatrix}$
10	(0.3558729,0.6441271)	(4.289662, 79.968115)	$\begin{pmatrix} 0.1699684 & 0.9406093 \\ 0.9406093 & 36.0462113 \end{pmatrix}$

(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
$\boldsymbol{\alpha_1}$	0.3558729	0.3558854
$(\boldsymbol{\mu_{x_1 1}}, \boldsymbol{\mu_{x_2 1}})$	(2.036388, 54.478516)	(2.036421, 54.478880)
$\begin{pmatrix} \sigma_{x_1 1}^2 & \rho_1\, \sigma_{x_1 1}\sigma_{x_2 1} \\ \rho_1\, \sigma_{x_1 1}\sigma_{x_2 1} & \sigma_{x_2 1}^2 \end{pmatrix}$	$\begin{pmatrix} 0.06916767 & 0.43516762 \\ 0.43516762 & 33.69728207 \end{pmatrix}$	$\begin{pmatrix} 0.06991677 & 0.4400355 \\ 0.4400355 & 34.0510784 \end{pmatrix}$

(table 3)

第二組常態資料比較

	自己撰寫結果	網路運算結果
$\boldsymbol{\alpha_2}$	0.6441271	0.6441146
$(\boldsymbol{\mu_{x_1 2}}, \boldsymbol{\mu_{x_2 2}})$	(4.289662, 79.968115)	(4.289688, 79.968413)
$\begin{pmatrix} \sigma_{x_1 1}^2 & \rho_1\, \sigma_{x_1 1}\sigma_{x_2 1} \\ \rho_1\, \sigma_{x_1 1}\sigma_{x_2 1} & \sigma_{x_2 1}^2 \end{pmatrix}$	$\begin{pmatrix} 0.1699684 & 0.9406093 \\ 0.9406093 & 36.0462113 \end{pmatrix}$	$\begin{pmatrix} 0.1709121 & 0.9456219 \\ 0.9456219 & 36.2489543 \end{pmatrix}$

(table 4)

結論

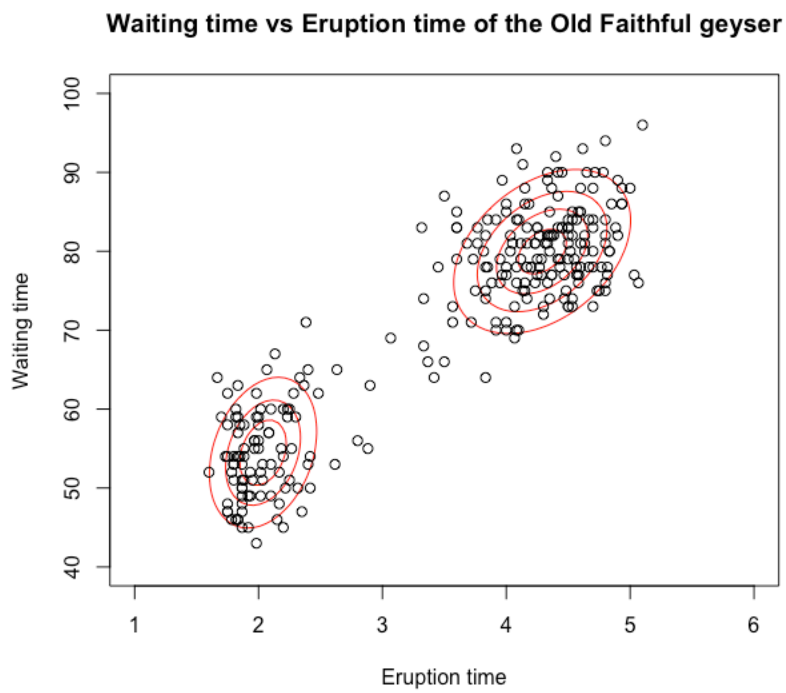
從table 1以及table 2中看出，當t = 3之後各個參數都呈現收斂的狀態。我亦比較網路上他人所撰寫的程式結果(table 3)，得到的結果相差無幾。因此我推估這兩個二元常態為

第一組

$$BN\left(\begin{pmatrix}2.036388\\54.478516\end{pmatrix},\begin{pmatrix}0.06916767 & 0.43516762\\0.43516762 & 33.69728207\end{pmatrix}\right),$$ 其線性係數為 0.3558729

第二組

$$BN\left(\begin{pmatrix}4.289662\\79.968115\end{pmatrix},\begin{pmatrix}0.1699684 & 0.9406093\\0.9406093 & 36.0462113\end{pmatrix}\right),$$ 其線性係數為0.6441271

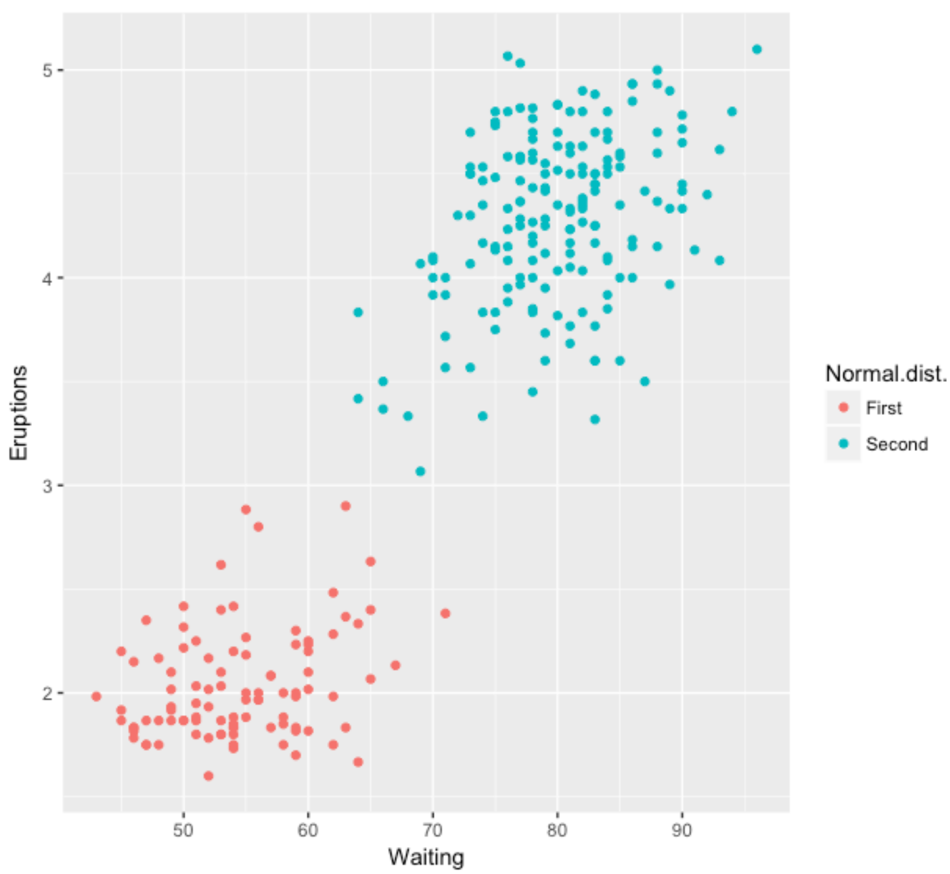


(左下為第一組二元常態，右上為第二組二元常態)

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