# EM for old faithful dataset

EM algorithm:

Observed data : X

Likelihood :

Goal : maximize with respect to

Idea:

- augment y (missing data or latent variable)

- complete data Likelihood

- & updata

E-step (update y):

Find the expectation of the complete-data loglikelihood with respect to the missing data y given the oberserved data x and the current parameter estimates

So we find

M-step (update):

Maximize the expectation obtained in E-step.

So we find

Repeat these two steps until convergence.

重複步驟直到參數收斂{

(E – step) For each i , set

(M – step) Set

}

其中 (E-step) 為

目的:找到使Q函數的最大值參數

(M-step)

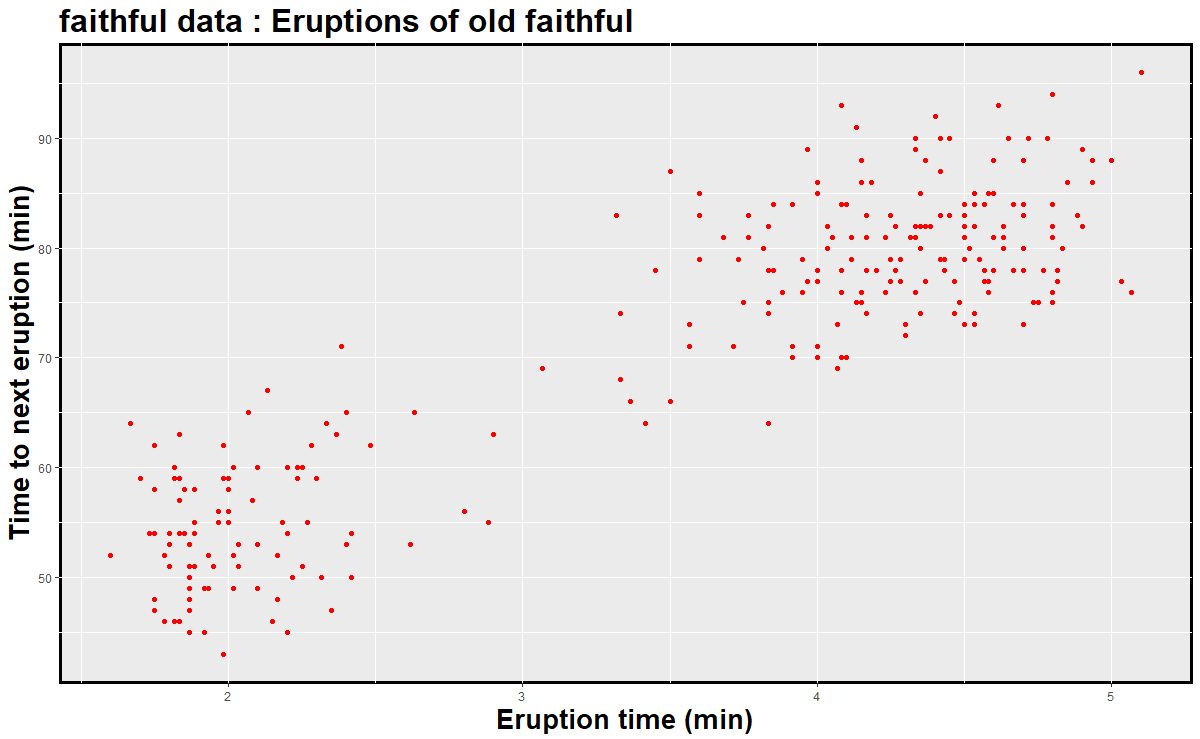
我們從(E-step)得到概似函數，並將其最大概似函數的MLE計算出來，得到的結果如下

結論:

我們發現兩步驟相同的函數為 ，也就是說我們在撰寫EM演算法的時候，我們只要根據每次迭代得到新的參數，在給定原本的樣本下，把每一次隱藏變數的條件分配計算出來，並帶入我們(M-step)的結果算出新的參數，經過數次迭代後，就能找到收斂的參數。

**Old Faithful dataset :**

資料為美國黃石國家公園週期性噴發時間及等待時間資料(成對共272筆資料)



由上圖我們可以發生資料呈現為兩群的分佈，所以此題我們使用的分配為兩個混合的二元常態分配，並使用EM演算法來進行參數估計

我們先計算樣本中的平均數與變異數

|  |  |  |
| --- | --- | --- |
|  | Eruption(min) | Waiting(min) |
| Means | 3.487783 | 70.89706 |
| Covariance matrix |  | |

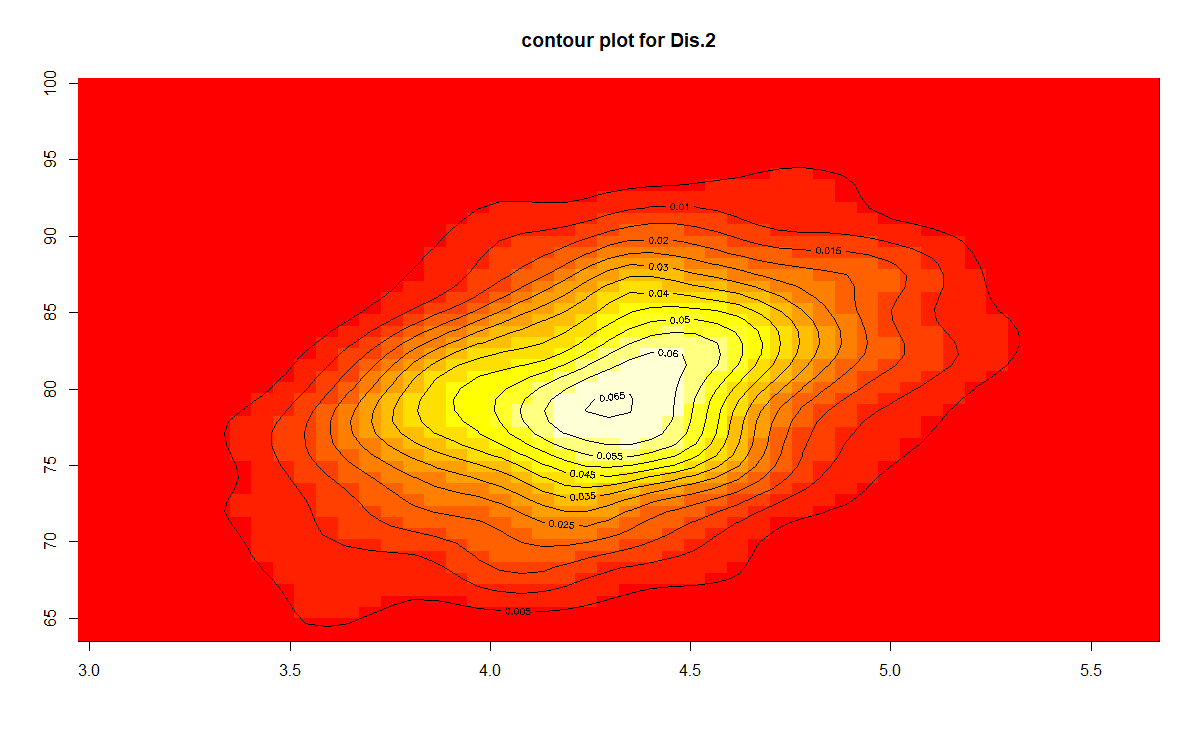
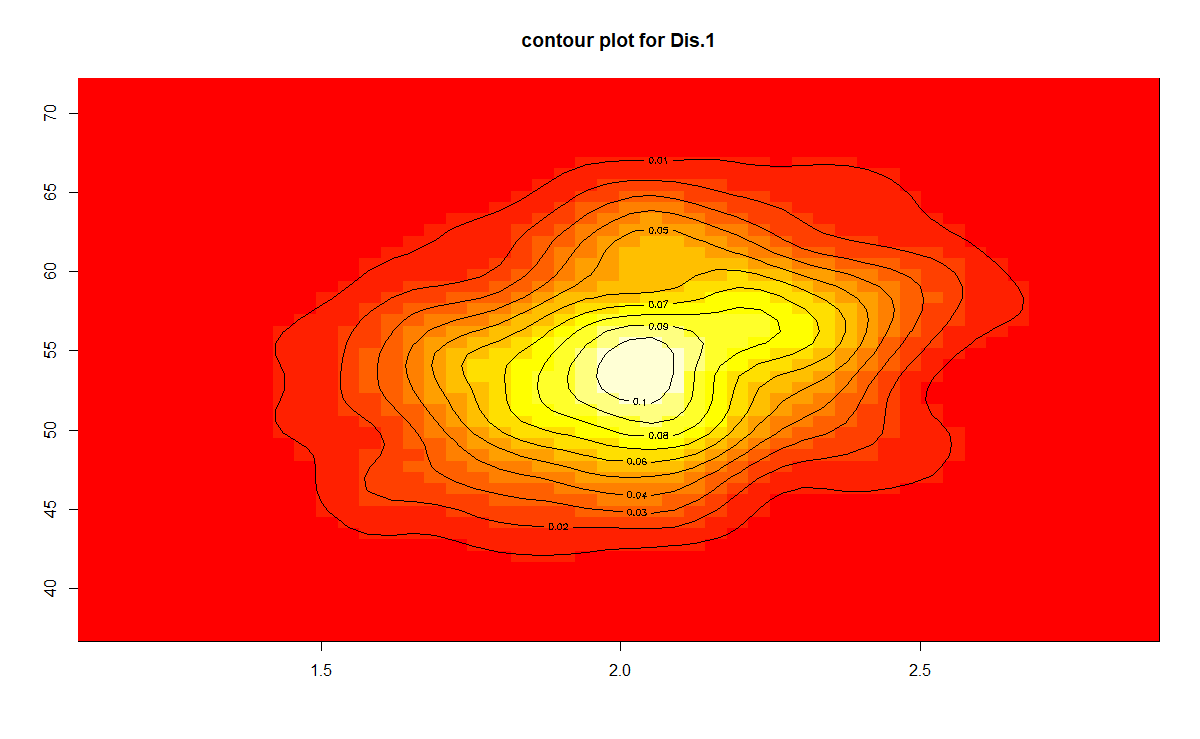
根據上圖，我們選取下列數值作為起始值

|  |  |  |
| --- | --- | --- |
|  | Dis.1 | Dis.2 |
| Weights |  |  |
| Means |  |  |
| Covariance matrix |  |  |

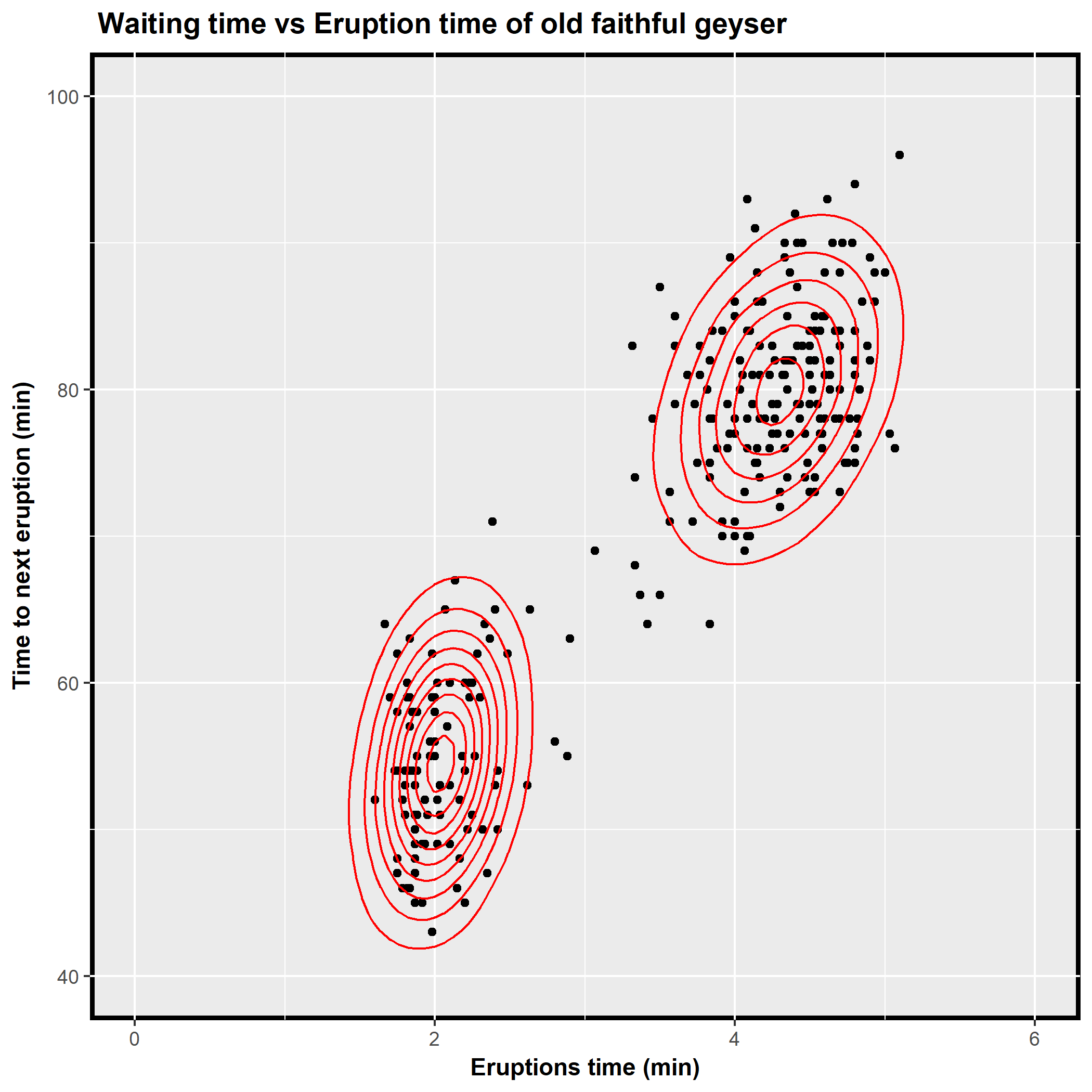
經由EM演算法迭代30次後，兩分配的估計參數如下:

|  |  |  |
| --- | --- | --- |
|  | Dis.1 | Dis.2 |
| Weights |  |  |
| Means |  |  |
| Covariance matrix |  |  |

接著我們根據迭代後得到的參數，將其二元常態機率密度投影在二維平面上



我們將混和常態分配一同繪製在原始的圖表上



以上為從EM演算法估計出參數所繪製的結果

30次迭代過程的動態gif檔 : [網址](https://github.com/CaoCharles/SCSH/blob/master/HW2/Faithful-EM.gif)