# Homework 4

Xiao Wang

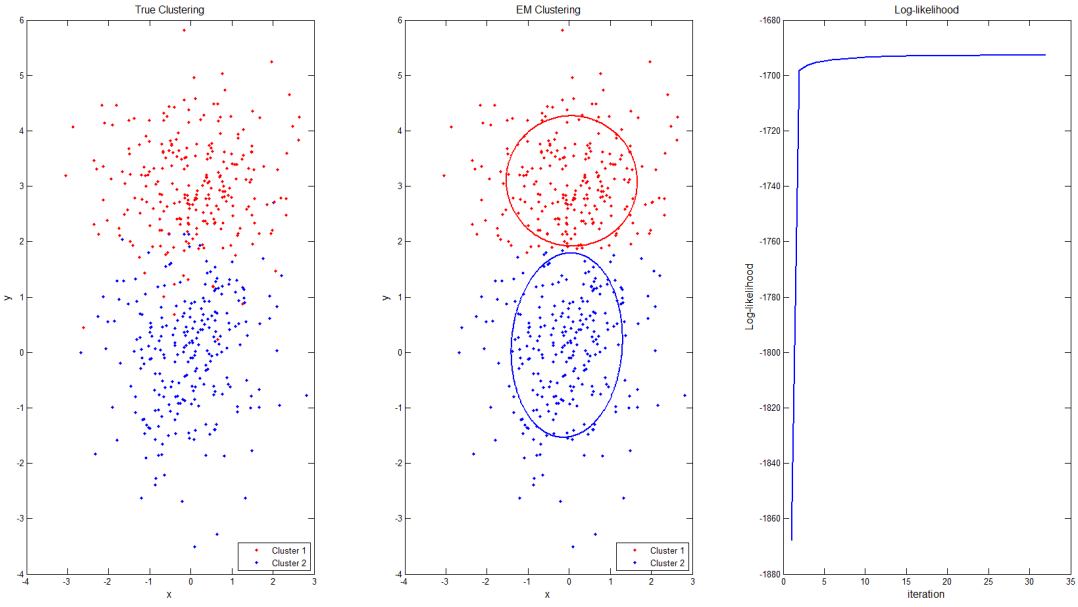
October 26, 2014

## 1

### (a)

The Expectation Maximization (EM) function is in appendix.

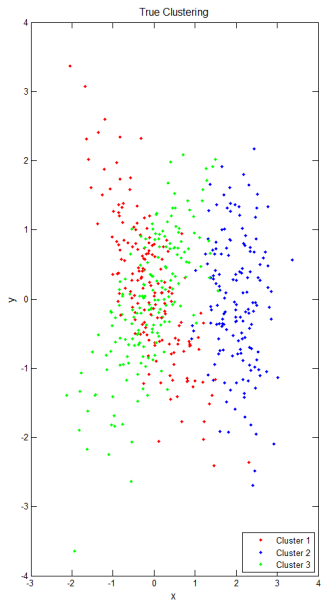
### (b)

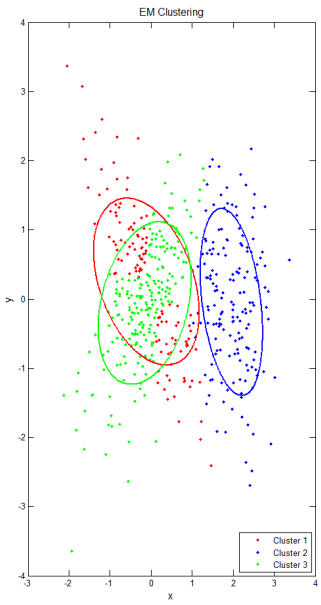


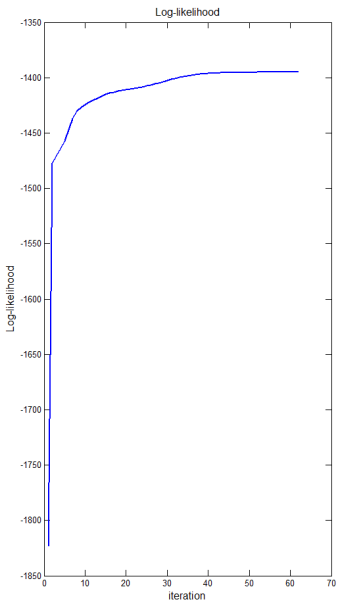

Answer:

For dataset1.m, EM performs as good as K-means. They all did a good solution with low error rate. Apparently, this dataset cannot tell the difference between EM clustering and K-means clustering.

**(c)**



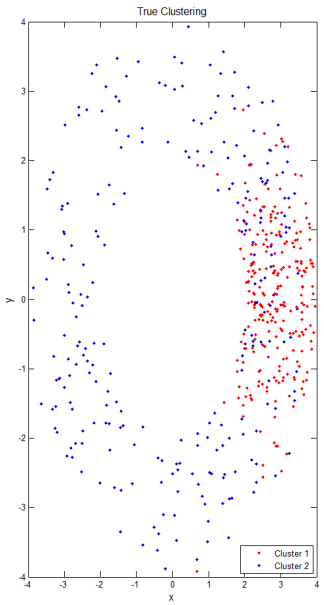



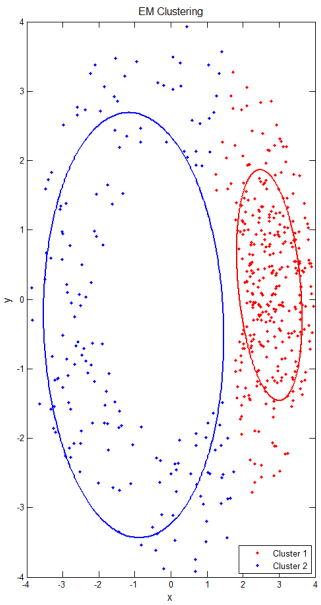



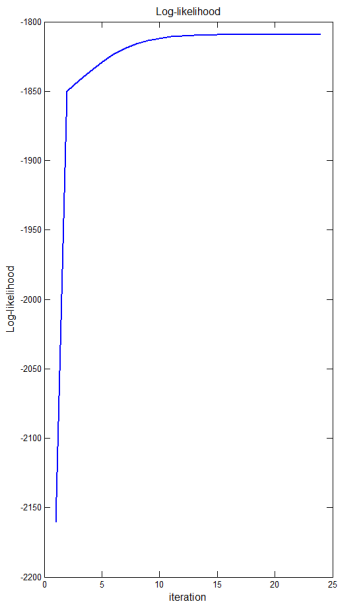

Answer:

For dataset2.m, when compared to K-means, EM is obviously better. It sucsuccessfully separates red group and green group. But it is still incorrect in the middle of the mixture part.

**(d)**



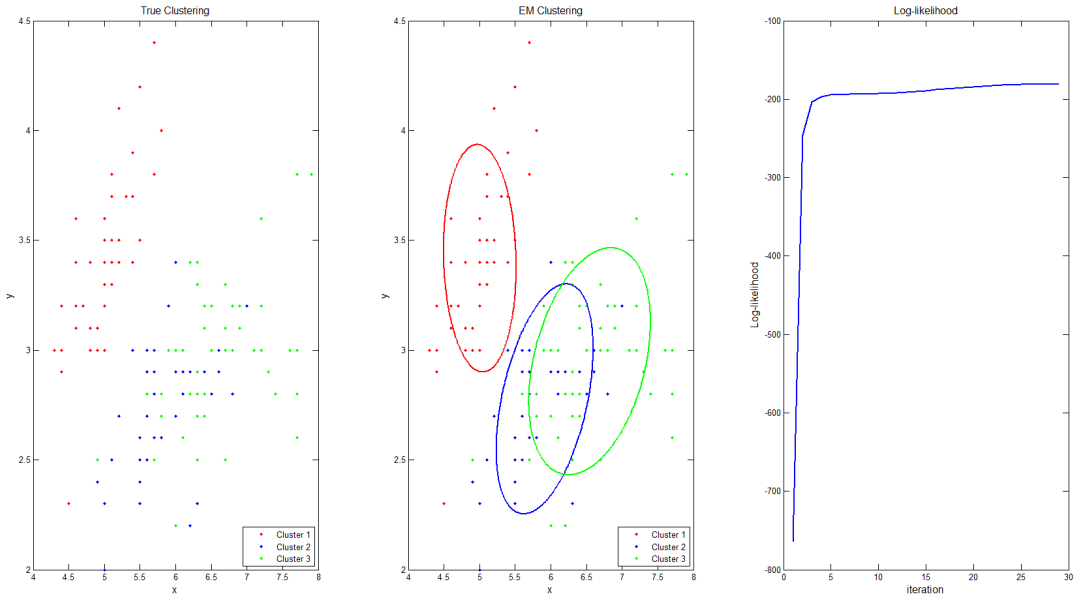



Answer:

For dataset3.mat, when compared to K-means, EM is a little better. However,it also failed to find the true clusters, especially in the overlap area of these two clusters. Therefore, EM algorithm has its limitation.

**(e)**

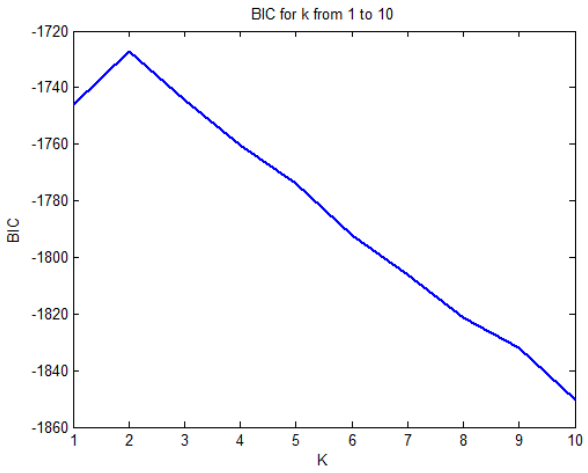



Answer:

For dataset4.m, EM and K-means both perform well. Maybe EM is still a little better. It is difficult to estimate the true figure, but I think the low quantity and the high dimension of data help to increase the precision of clustering.

## 2

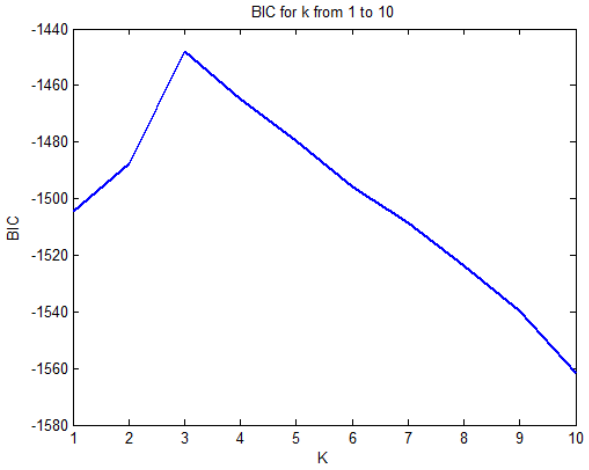
**(a)**




Answer:

For dataset1.m, the maximum matches the true number of clusters (K = 2).

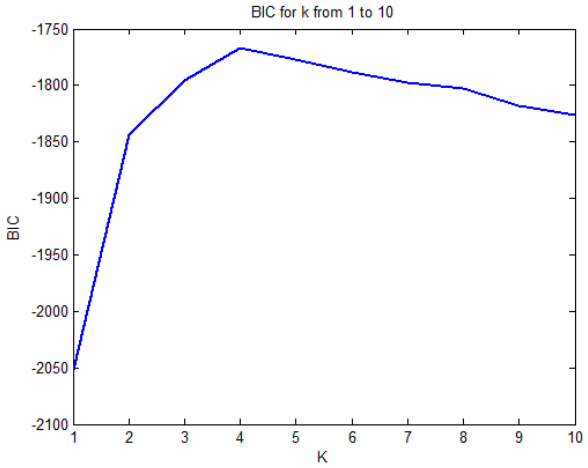
**(b)**




Answer:

For dataset2.m, the maximum matches the true number of clusters (K = 3).

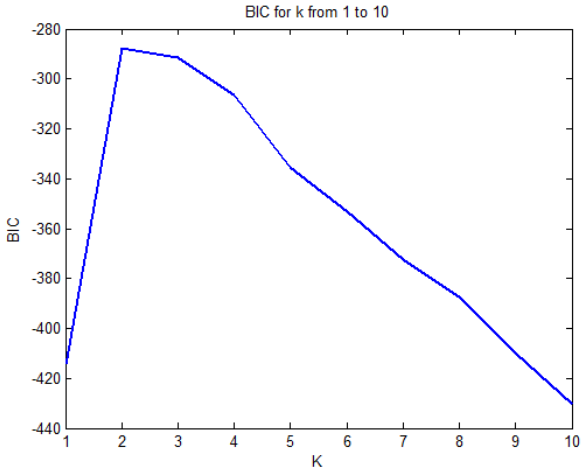
**(c)**




Answer:

For dataset3.mat, the K value that maximizes the BIC criterion is K = 4, which does not match the true number of clusters (K = 2). Therefore, BIC criterion is not a perfect method to find K. When data model is bad, both EM and BIC will fail.

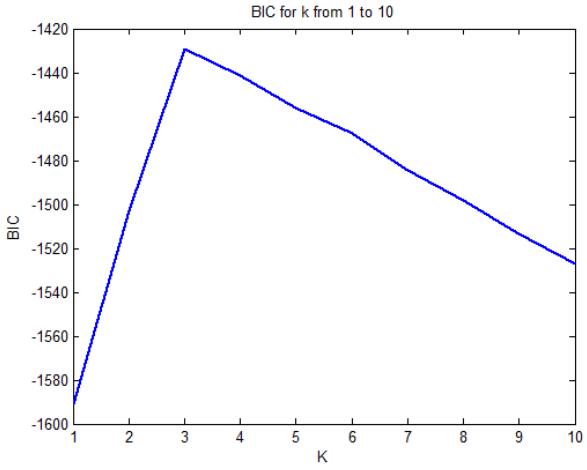
**(d)**

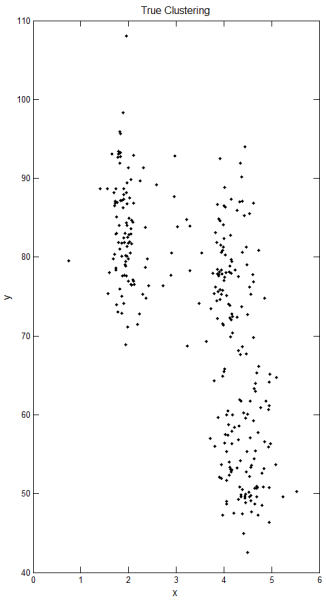


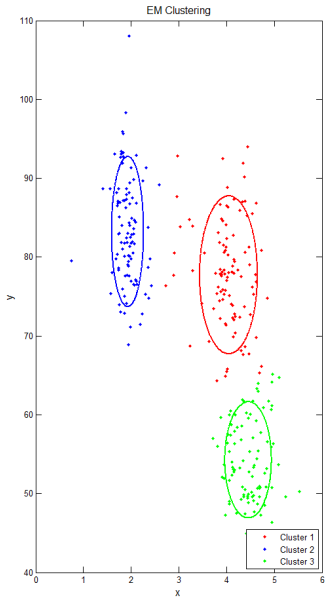

Answer:

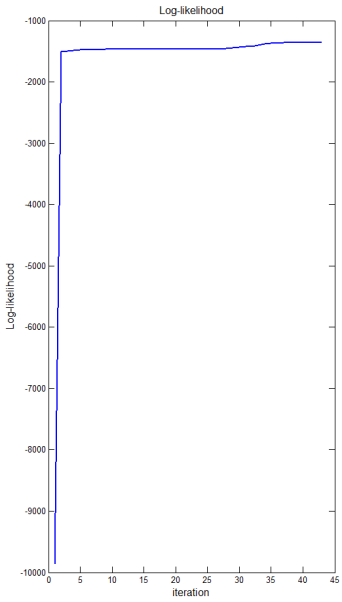
For dataset4.m, the K value that maximizes the BIC criterion is K = 2, which does not match the true number of clusters (K = 3). However, the BIC values are almost the same when K = 2 and K = 3. Therefore, BIC criterion is just a method for reference. The most appropriate K should be chosen through many considers.

(e)


BIC for k from 1 to 10
BIC
K
-1420
-1440
-1460
-1480
-1500
-1520
-1540
-1560
-1580
-1600
1
2
3
4
5
6
7
8
9
10


True Clustering
x
y


EM Clustering
x
y
Cluster 1
Cluster 2
Cluster 3


Log-likelihood
iteration
Log-likelihood

Answer:

For dataset5.mat, the maximum matches the number of clusters found with K-means (K = 3). And EM algorithm performs very well as I observed and much better than K-means.

NOTE: The figures above are the best results after I run the program many times. Sometimes the results are not very satisfing (just like K-means result). And sometimes BIC suddenly drops at K = 3 or K = 4. As I estimate, K = 3 or K = 4 should be the highest BIC value, which is unnormal.
I find that the BIC curve varies to the initiation. When data quantity is low for some clusters, EM algorithm may get a bad result and cause a bad estimation of K if we just use BIC criterion. The problem includes local maxima of the likelihood function and singular solutions.

In conclusion, EM performed pretty good overall. Although, it still has some problems with bad initiation and tricky dataset. Generally speaking, EM is better than K-means, but may take longer time.