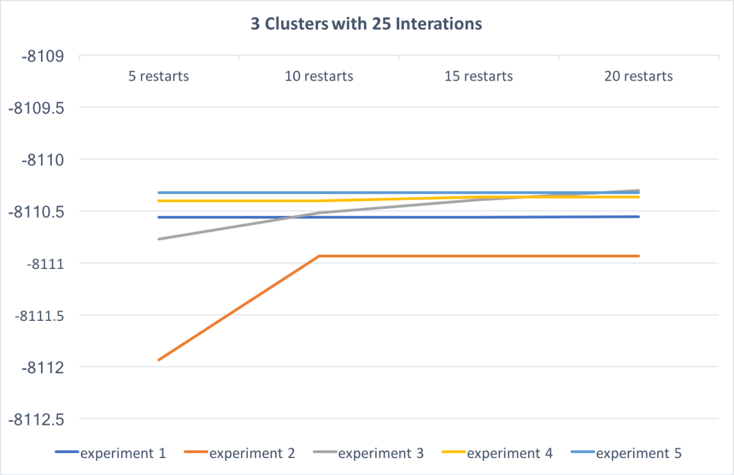
CS 534
Assignment #3: EM
Mingquan Liu, Daniel Wivagg, Mengdi Li, Ying Fang

Write up:

1. The Expectation step of the algorithm begins by computing the likelihood that a point belongs to each cluster for all points in the data set. This probability is computed using the equation for a normal distribution, weighted by the likelihood of the given cluster. The non-normalized probability values for each point are summed, the logarithm is taken, and the resulting values are summed again to compute Log Likelihood. Then, the probabilities for each point are normalized. Maximization now computes the likelihood of the clusters by summing the likelihood of each one and dividing each sum by the total number of data points. To recompute the means, the weighted average of points is taken for each cluster. To recompute covariance, the weighted distance of each point from the new mean is calculated for each cluster. Weighting factors are used because this is a soft clustering algorithm, where each point belongs to each cluster with some probability, even if very small. These values are equal to the probability a point came from the cluster over the total probability of the cluster. After the Maximization step, the algorithm loops.

2. We experimented with a cluster number of 3 and tried different numbers of restart times (5, 10, 15, 20) with 25 iterations of each. For each scenario, we experiment 5 times. Then, recorded the maximum log likelihood of each scenario, as shown in the graph below.



As you can see, the max log likelihood which the value is the most to the convergence can be found in 10 restarts. Thus, we chose 10 as the number of random restarts.

3. For the mean, we picked random points in the dataset to become the mean. For covariance, we took the covariance of the entire data set and divided it by a constant. These steps ensured that the initial parameters were always appropriate to the given data set.

4. When the difference between log_likelihood(t) (the log likelihood value after t iterations) and log_likelihood(t+1) (the log likelihood value after t+1 iterations) less than 0.05, the EM terminates.

5. Notes on random data set:

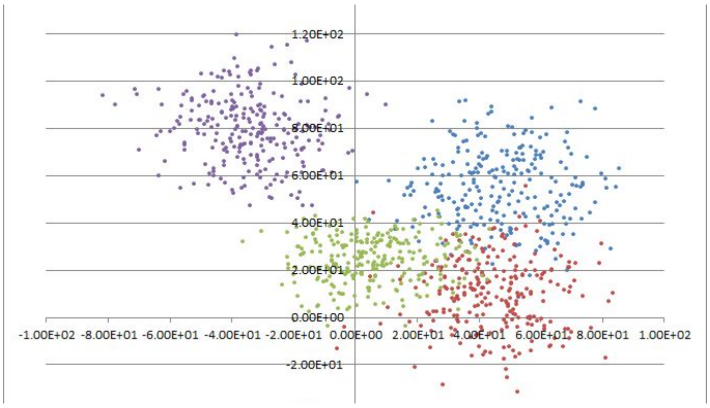
   Cluster 1: Mean (45, 55) Covariance (15, 0; 0, 15) Size: 250

   Cluster 2: Mean (45, 10) Covariance (15, 0; 0, 15) Size: 250

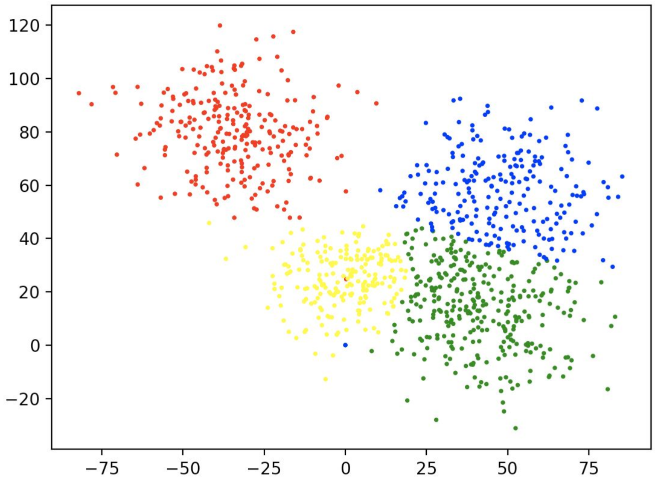
   Cluster 3: Mean (5, 25) Covariance (15, 0; 0, 10) Size: 250

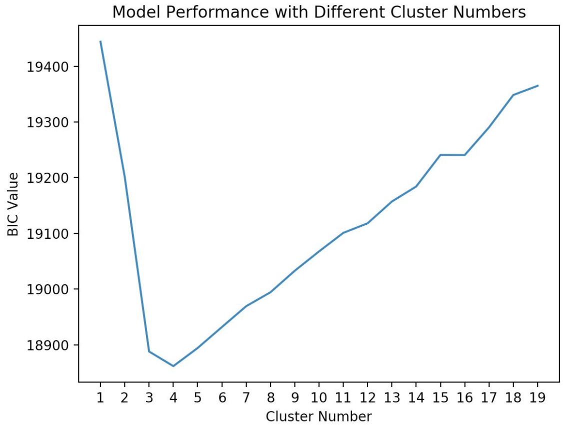
   Cluster 4: Mean(-35, 80) Covariance (15, 0; 0, 15) Size: 250

   The data is shown below. Although there are four clusters, three of them form a larger cluster that is harder to discern than the given data.



   The output of the algorithm is shown below:





Ans: According to the graph above, when the number of clusters is 4, the model has the best performance. So, our EM model with BIC can determine the number of clusters correctly.

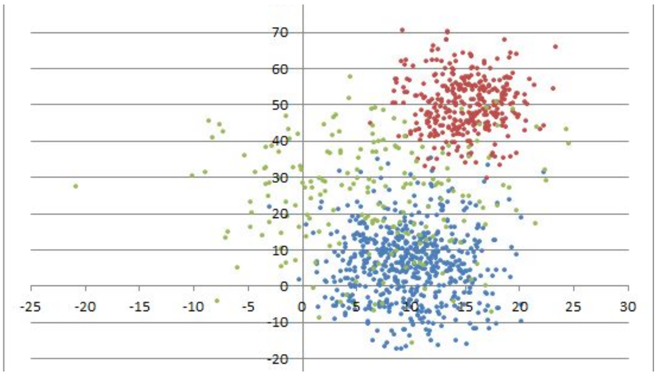
With our clustering:

Mean of cluster 1: [-34.56, 80.03]
Variance of cluster 1: [[219.64, -26.32], [-26.32, 210.40]]
Mean of cluster 2: [25.37   15.90]
Variance of cluster 2: [[526.93, -249.06], [-249.06, 224.43]]
Mean of cluster 3: [-1.76, 23.47]
Variance of cluster 3: [[106.06, 24.18], [24.18, 115.73]]
Mean of cluster 4: [45.02, 38.21]
Variance of cluster 4: [[261.91, 33.97], [33.97, 639.22]]

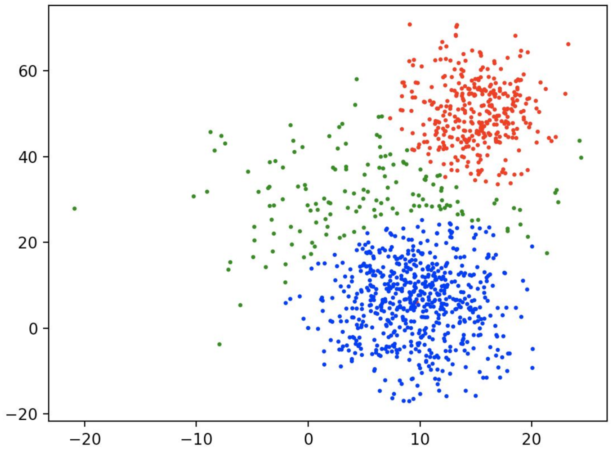
Ans: Our estimated cluster 1 is close to the correct cluster 4, estimated cluster 3 is close to the correct cluster 3. The two other estimated clusters are not very similar with the correct clusters.

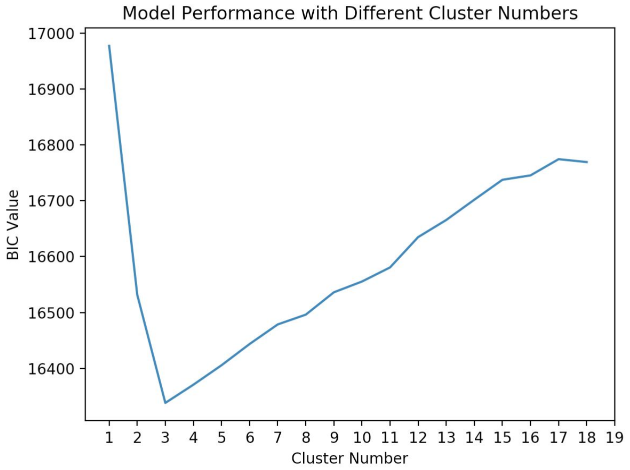
Notes on sample data set:

Parameters of sample data set (These are measured from the given data and known clusters and may not be the true values from which the data was drawn):
Cluster 1: Mean (9.803, 6.371) Covariance (16.1, 0; 0, 97.17) Size = 575
Cluster 2: Mean (14.869, 50.37) Covariance (9, 0; 0, 59)   Size = 336
Cluster 3: Mean (6.341, 25.174) Covariance (56.196, 0; 0, 212.54) Size = 208
The actual data is pictured below.



The output of the algorithm is shown:





Ans: As shown in the graph above, the model has the best performance when the number of clusters is 3. Our EM model with BIC determined it correctly.

With our clustering:

Mean of cluster 1: [9.72, 5.77]

Variance of cluster 1: [[16.05, -0.97], [-0.97, 86.38]]

Mean of cluster 2: [6.26, 29.09]

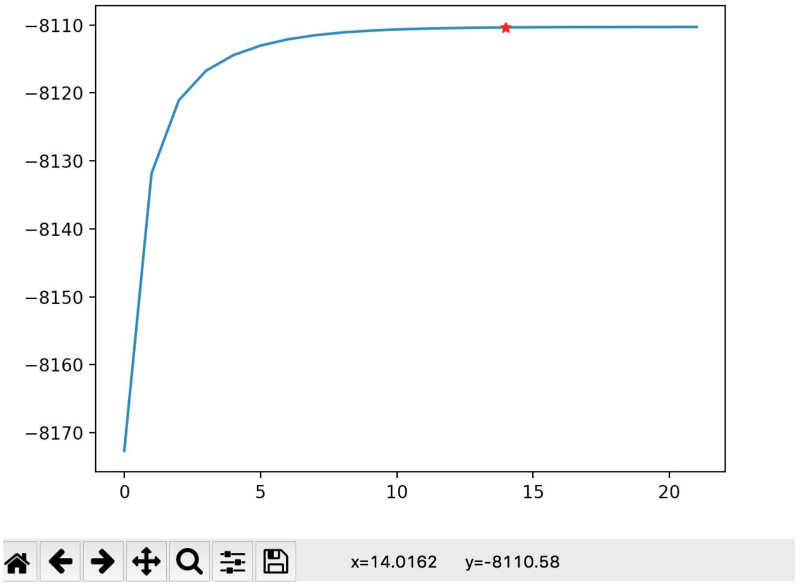
Variance of cluster 2: [[58.06, 9.17], [9.17, 145.68]]

Mean of cluster 3: [14.91, 50.29]

Variance of cluster 3: [[9.59, 0.36], [0.36, 59.83]]

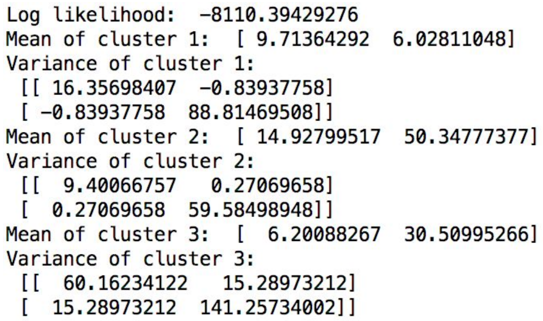
Ans: Our estimated cluster 1 is close to the correct cluster 1, estimated cluster 2 is close to the correct cluster 3, estimated cluster 3 is close to the correct cluster 2.

6. The program would normally stop at the point marked as red star. When it keep going, the value of log likelihood only changed slightly.



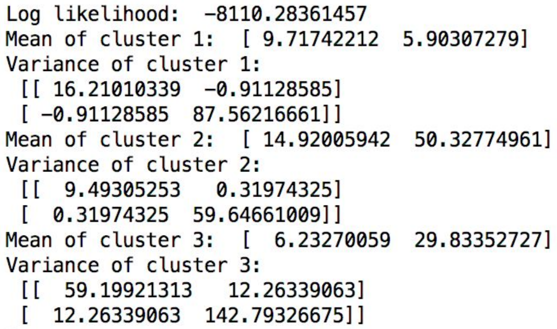
x=14.0162 y=-8110.58

The parameters of clusters of 14th iteration:



```
Log likelihood:  -8110.39429276
Mean of cluster 1:  [ 9.71364292  6.02811048]
Variance of cluster 1:
 [[ 16.35698407  -0.83937758]
 [ -0.83937758  88.81469508]]
Mean of cluster 2:  [ 14.92799517  50.34777377]
Variance of cluster 2:
 [[  9.40066757   0.27069658]
 [  0.27069658  59.58498948]]
Mean of cluster 3:  [  6.20088267  30.50995266]
Variance of cluster 3:
 [[  60.16234122   15.28973212]
 [  15.28973212  141.25734002]]
```

The parameters of clusters of the last iteration:



```
Log likelihood:  -8110.28361457
Mean of cluster 1:  [ 9.71742212  5.90307279]
Variance of cluster 1:
 [[ 16.21010339  -0.91128585]
 [ -0.91128585  87.56216661]]
Mean of cluster 2:  [ 14.92005942  50.32774961]
Variance of cluster 2:
 [[  9.49305253   0.31974325]
 [  0.31974325  59.64661009]]
Mean of cluster 3:  [  6.23270059  29.83352727]
Variance of cluster 3:
 [[  59.19921313   12.26339063]
 [  12.26339063  142.79326675]]
```

As we compare the value in 14th iteration and last iteration, we notice that as the log likelihood converges, the model parameters also converge. Thus, we can draw the conclusion that the convergence of log-likelihood correspond to convergence of model parameter estimates.

7. We used the equation below to calculate BIC value.

$$\mathrm{BIC} = \ln(N) * \mathrm{k} - 2\ln(\hat{L})$$

where:

N is the sample size i.e. 1119 for the sample data set, 1000 for the new data set.

k is the number of parameters estimated by the model. In our case:

$$\mathrm{k} = (\mathrm{M} - 1) + \mathrm{M}(\mathrm{D} + \frac{1}{2}\mathrm{D}(\mathrm{D} + 1))$$

where M is the number of clusters, D is the number of dimensions i.e. 2.

L(hat) is the maximized value of the likelihood function of the model.

For each cluster number, we recorded the BIC value and compared with them to find the lowest one. When the BIC value does not become lower, the searching will be terminated.