Artificial Intelligence Project 3

Nishant Shah, Rishi Khajuriwala, Ganesh Prasanna Balakrishnan, Mohamad El-Rifai

1. **Write a paragraph explaining the expectation and maximization steps in your algorithm.**

ANSWER:

Expectation:

The data is assumed to be produced by k, n-dimensional Gaussians. We choose k data points randomly as the cluster centers and assign large covariance (identity matrix x variance of the data). We get the probability of the data points belonging to each individual cluster and multiply it with the corresponding probability of the clusters. We normalize it to get the probability of cluster given data point.

Maximization:

We recompute the probability of clusters by taking the average of the probabilities of cluster given data. We recompute the mean and using the recomputed means we recompute the covariances using $(\text{data} - \text{mean})^T$ (data – mean). The data in our case is a 2d matrix, the probabilities of each data point belonging to a cluster is k x m matrix, where m is the number of data points. The means of all the clusters are stored in a 2d matrix of size n x k. Covariance is a list of k 2d matrices each of size n x n.

2. **How many random restarts did you choose to do? What was your approach for determining a good number?**

ANSWER:

We choose the means from the data itself. The number of restarts is based on the number of data points we have. The more the number of data points the more the number of restarts we might need because there is a lot of data to choose the means from and choosing a good mean to start with is less probable.

3. **How did you initialize the cluster centers?**

ANSWER:

We randomly chose points from the data as the cluster centers. This ensured that the initial assumption of means was fairly sensible because space in infinite.

4. **How did you decide on the terminating criteria for EM?**

ANSWER:

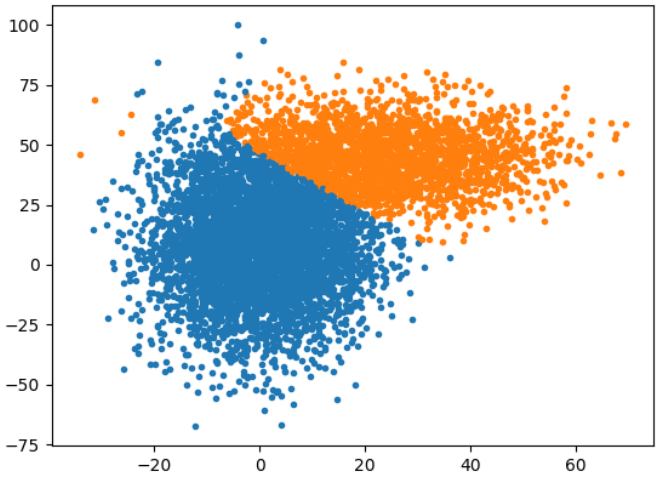
We keep a track of the log likelihood and once the log likelihood converges we terminate the EM algorithm. In our code we have a small epsilon (around 10^{-4}) which we use to check for convergence. If $\text{Log-likelihood}_{n+1}$ - Log-likelihood_n < epsilon we terminate EM. We do this because once Log-likelihood converges it is highly unlikely for the log likelihood to increase further at a later point of time.

5. **Create a data file whose clusters are less-easily separable than the provided example (but more easily separable than the first file I provided :-). How does EM perform? Specifically, how accurate is it in determining the correct number of clusters? If given the correct number of clusters, does it find the correct means and variances of the clusters? Does it assign points to the correct cluster? Answer the same questions for the provided sample data file.**

ANSWER:

Hard to separate data:

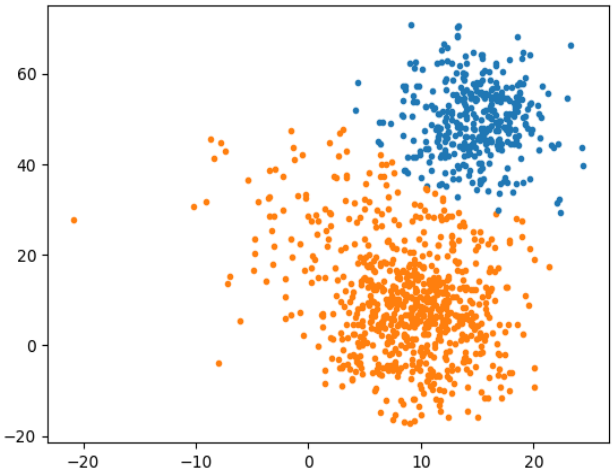
EM performs decently for data that is hard to separate. When the algorithm with BIC was run ten times it was able to determine the correct number of clusters (2) 6 times out of ten and thrice it guessed the number of clusters as 1 and once as 3.



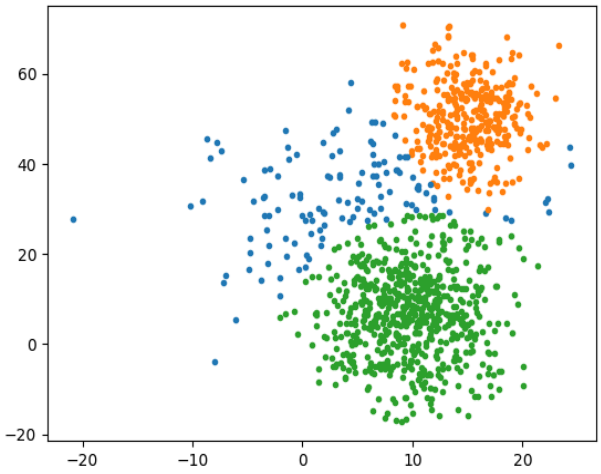
If given the number of clusters given the number of restarts is sufficient enough and the algorithm is allowed to run for a few iterations even after the log-likelihood converges it almost always gives the right means and covariances of the clusters. The algorithm clusters approximately 90 percentage of the points correctly.

Provided Sample File:

EM performs really well for this data because the data is easily separable. When the algorithm with BIC was run ten times it was able to determine the correct number of clusters (3) 9 times out of ten and once it guessed the number of clusters as 2.



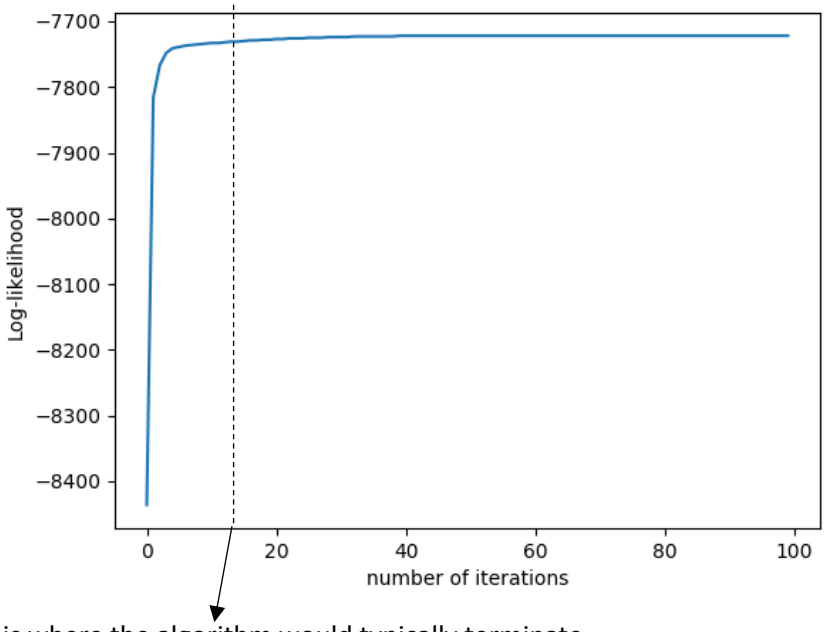
2 clusters



3 clusters

If given the number of clusters given the number of restarts is sufficient enough and the algorithm is allowed to run for a few iterations even after the log-likelihood converges it almost always gives the right means and covariances of the clusters. The algorithm clusters approximately 90 percentage of the points correctly.

6. **Sketch log-likelihood vs. # of iterations for a data set. Run it for longer than your typical termination criteria, and mark where the algorithm would normally stop on the graph. Provide the parameter estimates for where your program would normally stop, and what it would find if it kept going. Does convergence of log-likelihood correspond to convergence of model parameter estimates?**
   ANSWER:

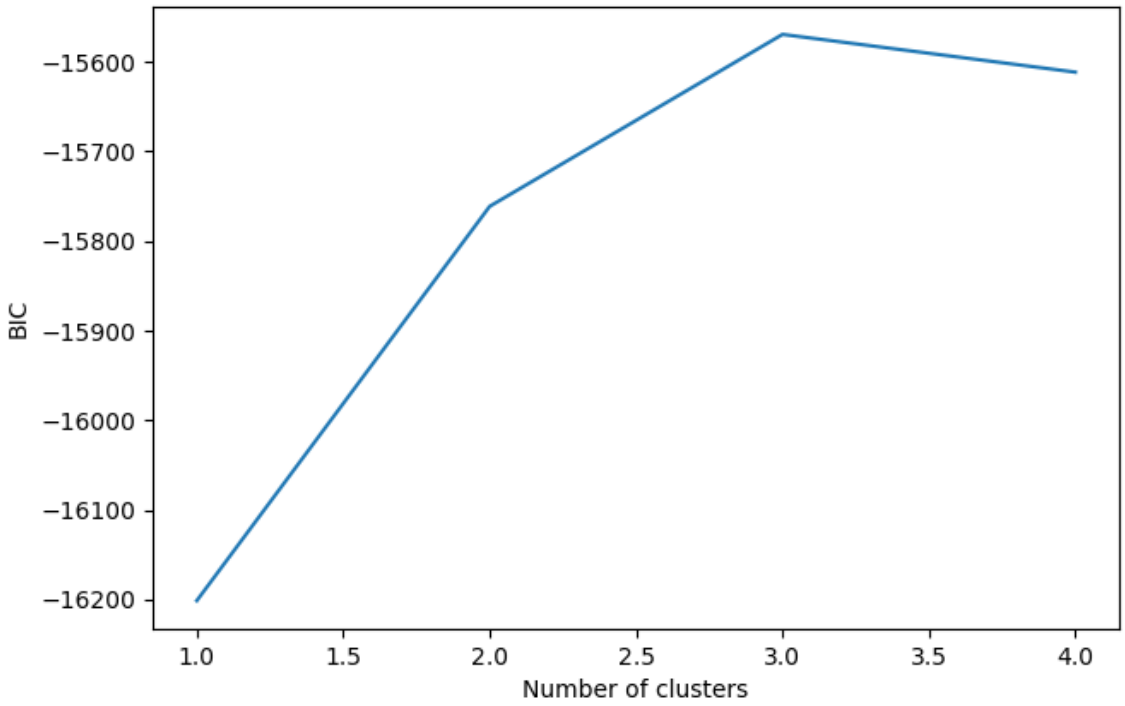


The parameters EM would find if it were terminate based on convergence of log-likelihood would be somewhat close to the true mean and covariance whereas if it kept running for a longer time it would converge to some value almost the same as the true mean and covariance. You can see that in the picture below.

```
Parameters when the logLikelihood has converged:
Cluster 1:
Mean: [9.7543129  5.68720843]
Covariance: [[15.77378414 -0.8262316 ]
 [-0.8262316  86.47360031]]
Cluster 2:
Mean: [ 5.81056621 26.9017275 ]
Covariance: [[ 54.70875242  -0.89951268]
 [ -0.89951268 153.88403827]]
Cluster 3:
Mean: [14.8573178  49.89911871]
Covariance: [[10.18094067  0.65007301]
 [ 0.65007301 63.80311566]]
After plenty of iterations after convergence of logLikelihood:
Cluster 1:
Mean: [9.71894554 5.86763255]
Covariance: [[16.16659824 -0.92834258]
 [-0.92834258 87.22461444]]
Cluster 2:
Mean: [ 6.24249832 29.63779953]
Covariance: [[ 58.90520807  11.43692779]
 [ 11.43692779 143.47047007]]
Cluster 3:
Mean: [14.9182272  50.32059169]
Covariance: [[ 9.51729053  0.33082101]
 [ 0.33082101 59.67878481]]
```

From the picture above we can see that the convergence of log-likelihood does not mean that the model parameters have converged.

7. **Explain how you used BIC (see www.wikipedia.org/wiki/Bayesian_information_criterion) as a modeling fitting criteria for part 2 and how you used it to terminate your search.**
ANSWER:



BIC increases up to a point and then starts decreasing. It is not necessary that BIC would keep decreasing after 3 clusters. There might be other number of clusters for which the algorithm might give similar or better performance than 3 clusters but that would only lead to overfitting. According to Occam's razor a better model between two would be the simpler one. This is why we stop when it reaches the first local maxima.
The number of parameters for the BIC is k-1 probabilities (considering they all add to one), the covariance matrix is symmetric about the diagonal and the number of parameters would be k x n(n+1)/2 and the number of means is k x n, where n is the dimension of data. Total number of parameters is k-1 + k x (n + n(n+1)/2).

References:
Parameter Selection for EM Clustering Using Information Criterion and PDDP
-Ujjwal Das Gupta,Vinay Menon and Uday Babbar