CS 274A Homework 5

Zachary DeStefano, 15247592

Due Date: Wednesday March 5th

Plots for Dataset1

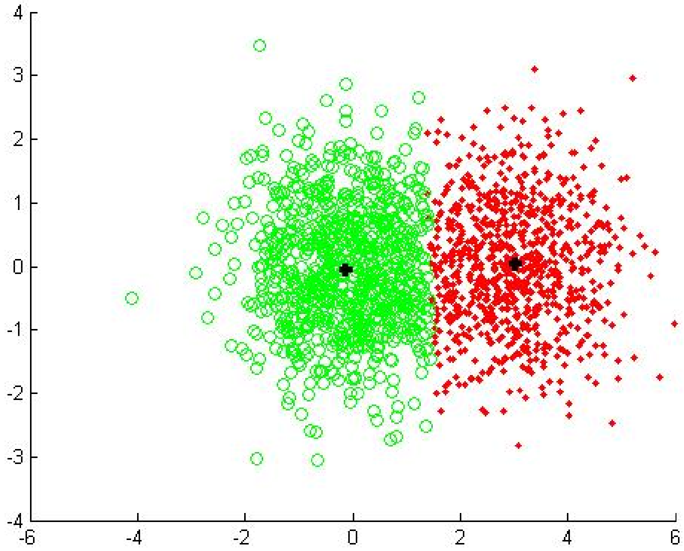

Figure 1: The k-Means plot for dataset1, K=2

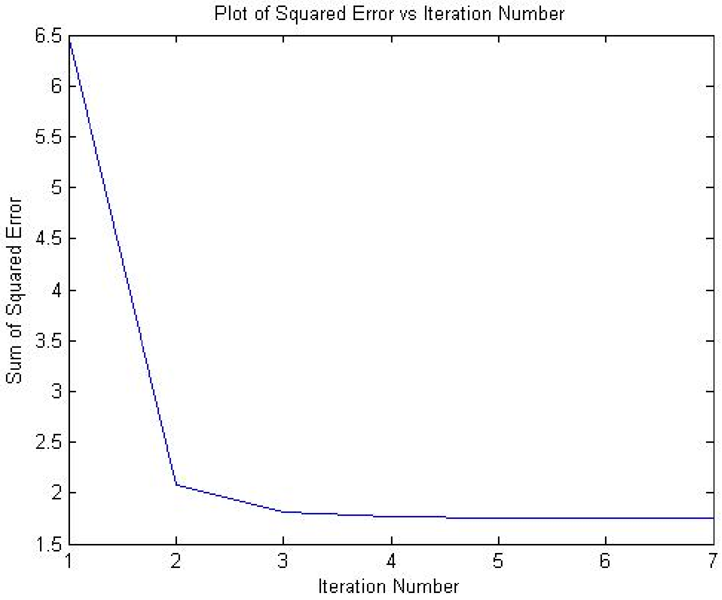

Figure 2: The Sum of Squared Error plot for dataset1, K=2

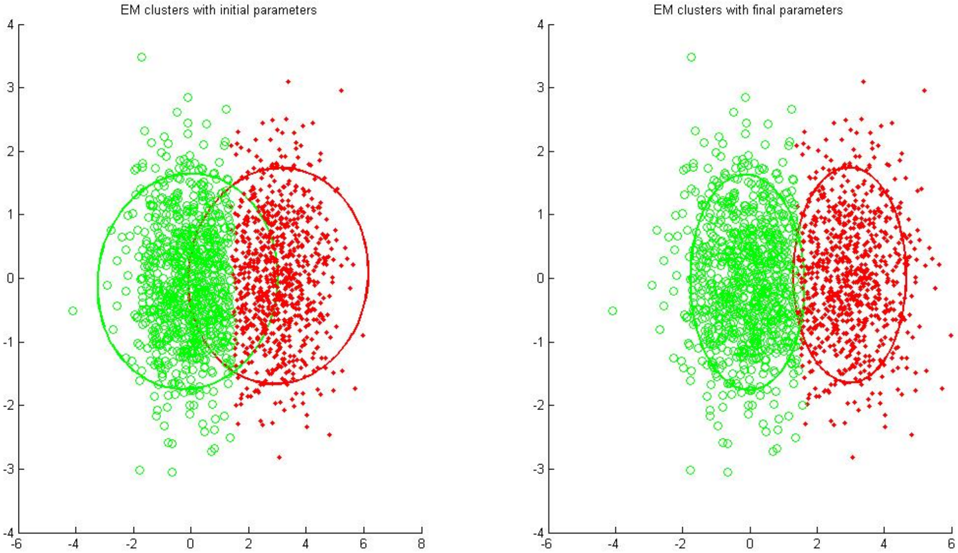

Figure 3: The EM cluster plots for dataset1, K=2, using initialization method 3

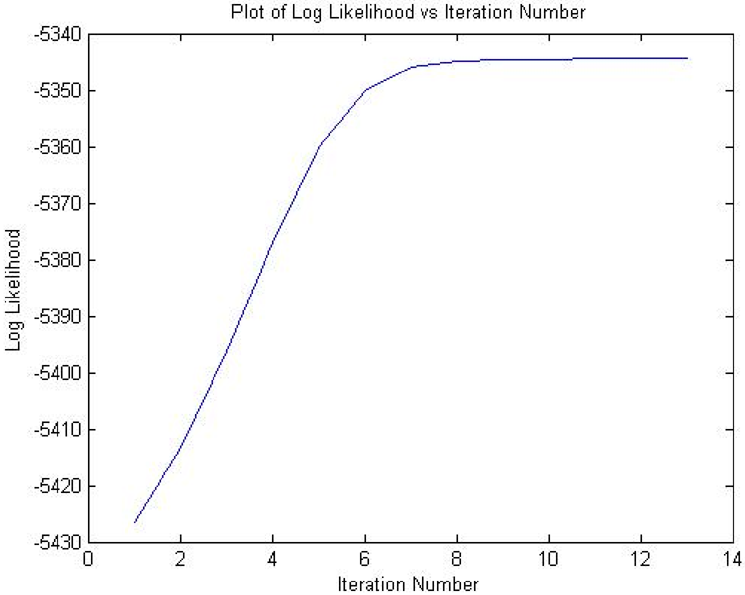

Figure 4: The likelihood plot for dataset1, K=2, using initialization method 3

Plots for Dataset2

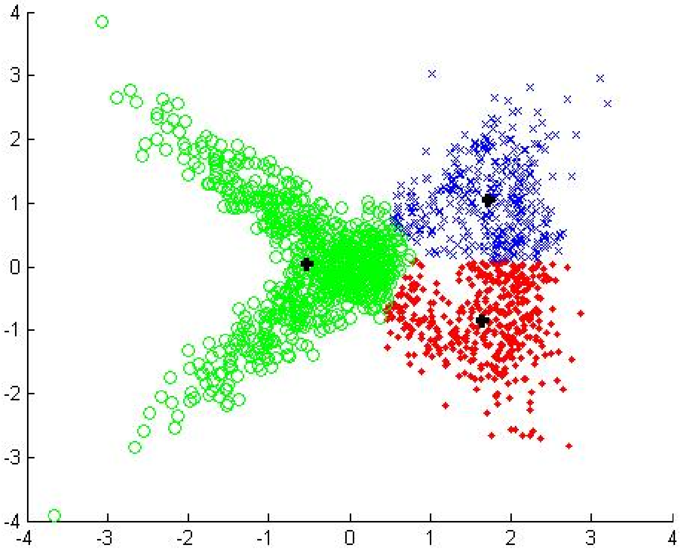

Figure 5: The k-Means plot for dataset2, K=3

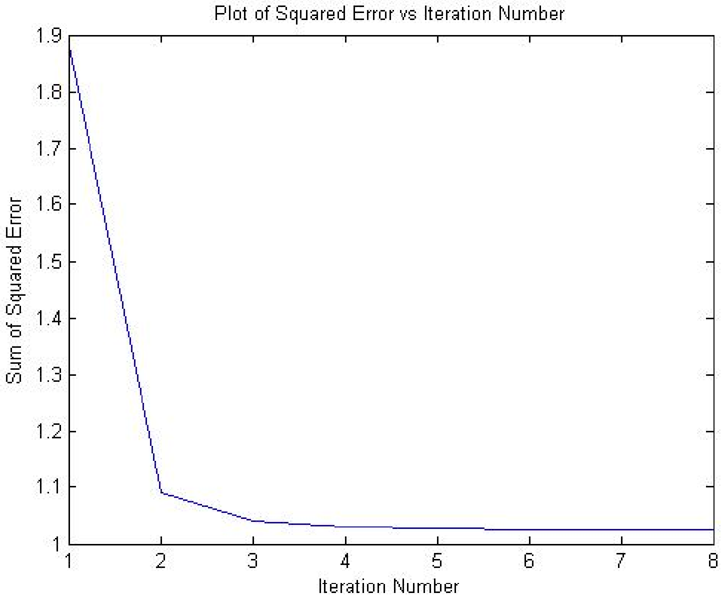

Figure 6: The Sum of Squared Error plot for dataset2, K=3

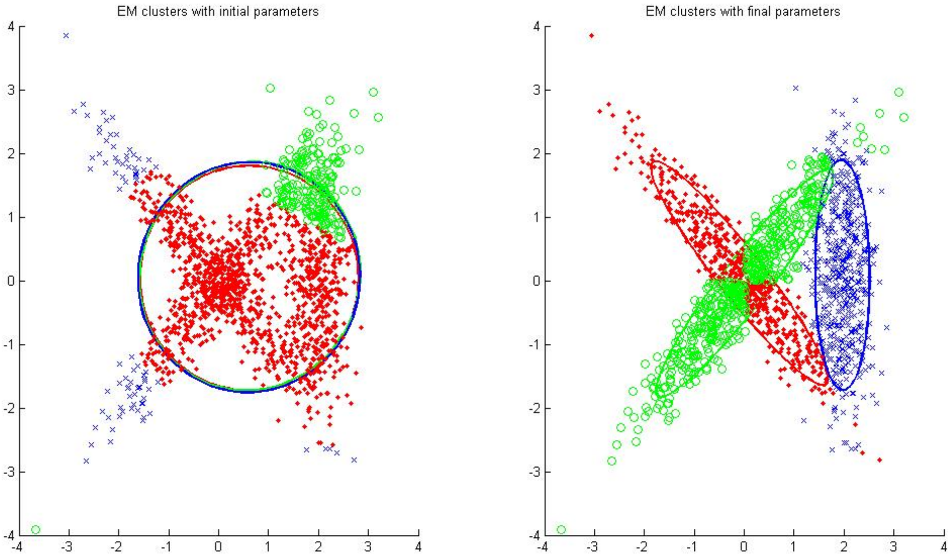

Figure 7: The EM cluster plots for dataset2, K=3, using initialization method 1

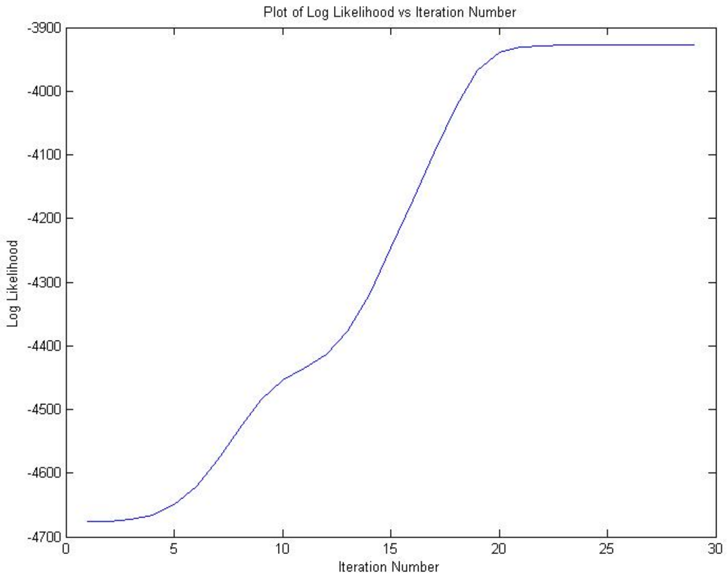

Figure 8: The likelihood plot for dataset2, K=3, using initialization method 1

Plots for Dataset3

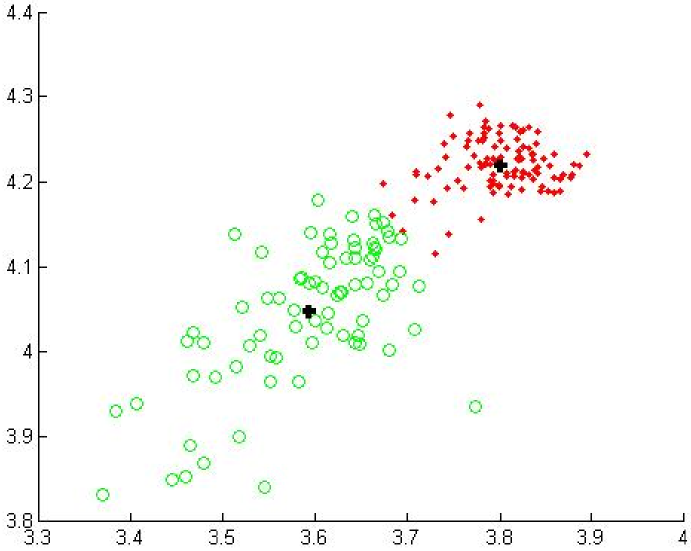

Figure 9: The k-Means plot for dataset3, K=2

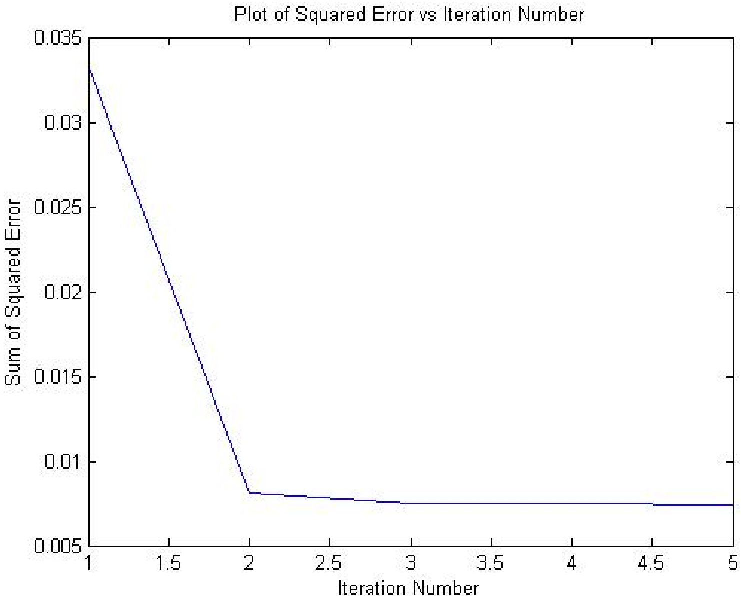

Figure 10: The Sum of Squared Error plot for dataset3, K=2

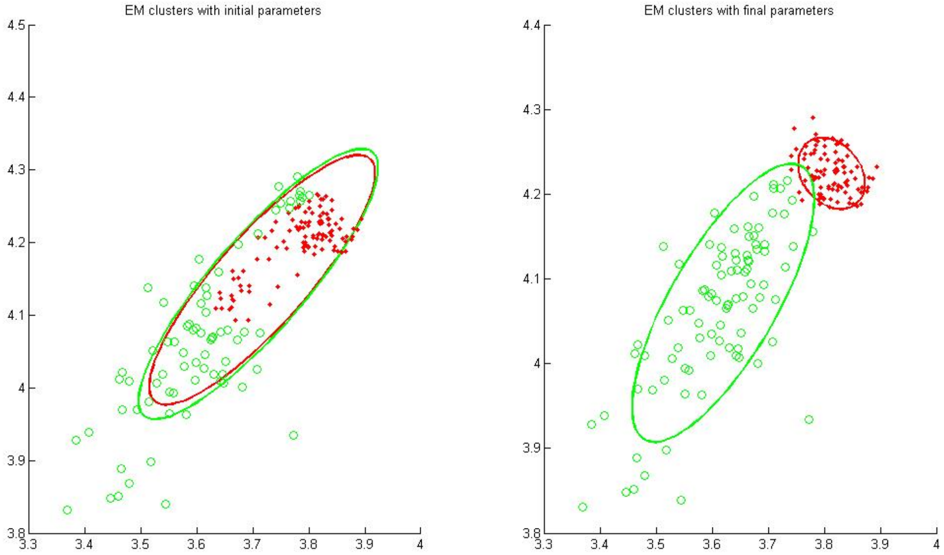

Figure 11: The EM cluster plots for dataset3, K=2, using initialization method 1

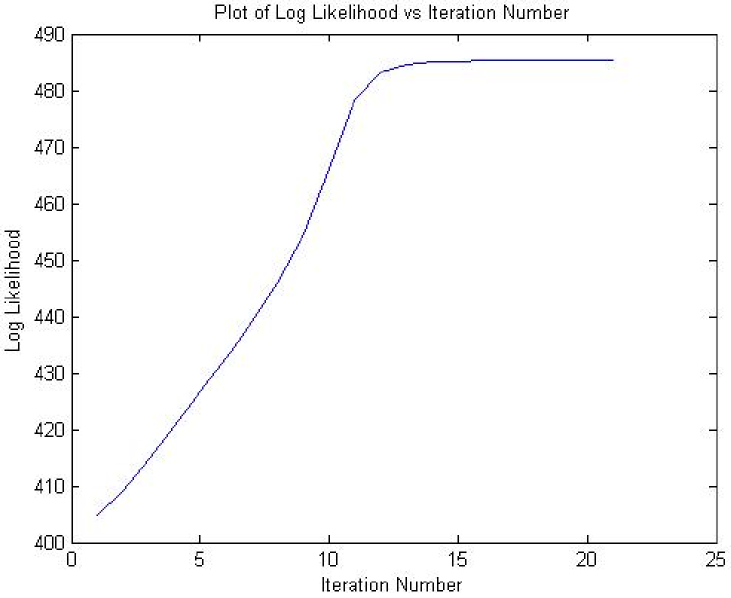

Figure 12: The likelihood plot for dataset3, K=2, using initialization method 1

Comments on Results

Make new plots after comparing with Bailey

For dataset1, there was not much of a difference between the K-means result and the EM algorithm result. This was likely due to the fact that the two Gaussians did not overlap so K-means put the points in the correct Gaussian cluster more easily. With dataset2, there was a substantial difference. The EM algorithm seemed to capture the Guassians much better than K-means. This is likely due to the fact that two of the Gaussians had an overlapping mean. Since K-means just clusters them into groups, it would not detect the differing Gaussians, whereas since EM is specifically trying to fit to Gaussians, it would detect it. For dataset3, the algorithm performed slightly better. **INSERT COMMENT ABOUT RESULT WHEN COMPARING WITH LABELSET3**

EM algorithm gives accuracy of 45.6% for dataset 3. For k-means it was 42.3%.

Mention the initialization methods used and why those could have been the best ones

In the end, because this data was generated from Gaussian densities and the EM algorithm fits the data to Gaussians, the EM algorithm was ideally suited to find the clusters in our case. However, if our need was to put data into two groups, as with dataset3, without much knowledge of the underlying model, then k-means is the ideal solution. **INSERT COMMENT AFTER RUNNING COMPARISON WITH DATASET3**

Dataset 1			Dataset 2			Dataset 3		
K	log-L	BIC	K	log-L	BIC	K	log-L	BIC
1	-5459.40	-5481.53	1	-4676.77	-4698.72	1	402.79	387.17
2	-5344.47	-5388.73	2	-4449.92	-4493.80	2	485.35	454.13
3	-5343.42	-5409.82	3	-3993.20	-3993.20	3	494.04	447.20
4	-5341.18	-5429.72	4	-3922.58	-4010.34	4	503.97	441.52
5	-5337.96	-5448.63	5	-3921.62	-4031.31	5	507.75	429.69