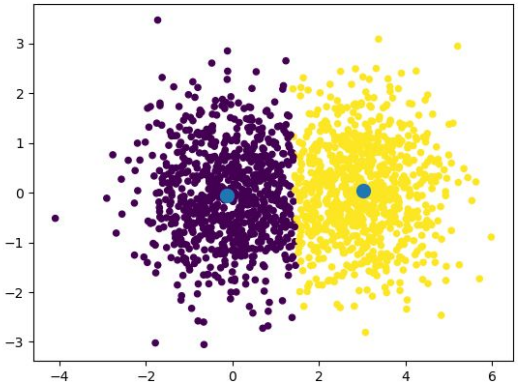
CS 274A HW6

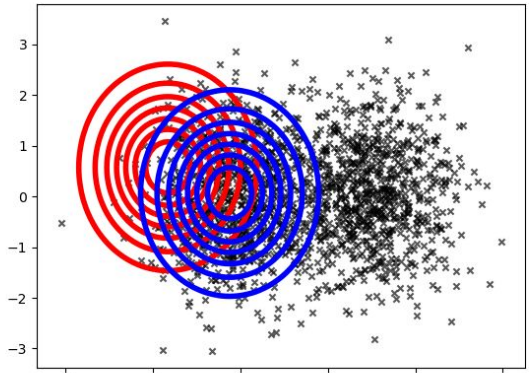
Caleb Nelson

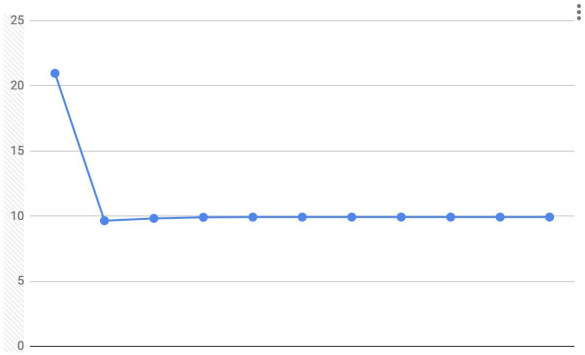
Each page corresponds to a separate dataset, ordered from 1 to 3, with the figures and plots laid out as such

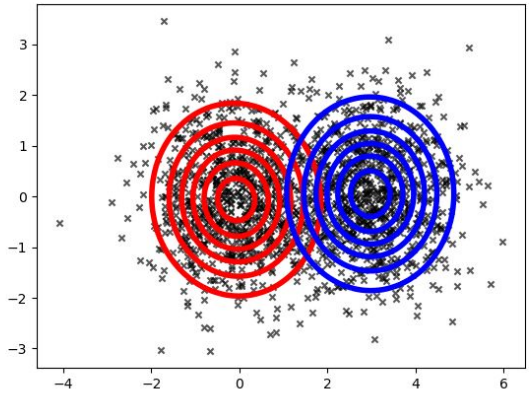
K-means clustering with each data colored to match its assigned cluster	The initial parameters for the EM/Gaussian mixtures code for the highest-likelihood solution.
A plot of the sum-squared-error (divided by n) as a function of iteration number in the K-means algorithm.	The final parameters for the EM/Gaussian mixtures code for the highest-likelihood solution.
Table of BIC values for various K's, with the minimum highlighted in green	A plot of the log-likelihood as a function of iteration number during EM.

Followed by brief comments on each algorithm with each dataset



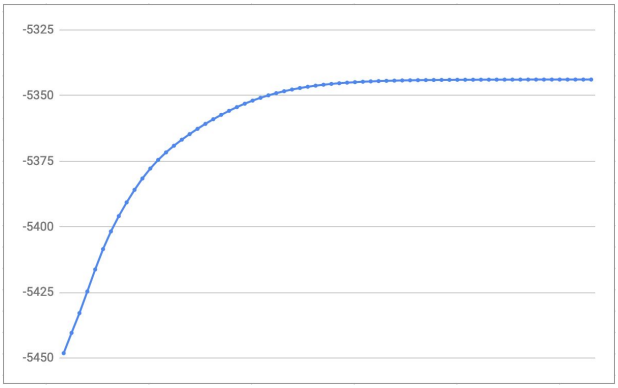




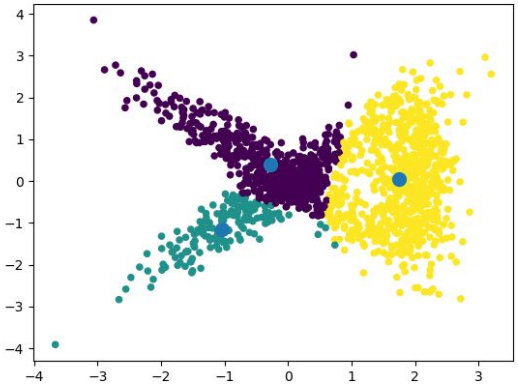


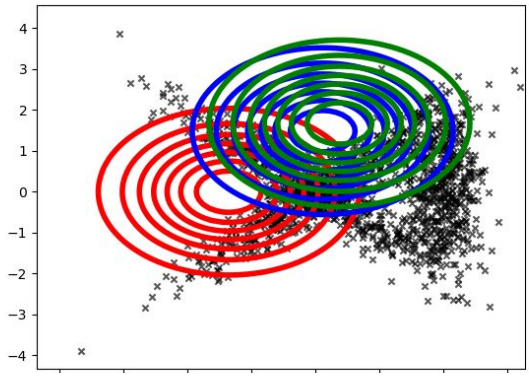
BIC Table

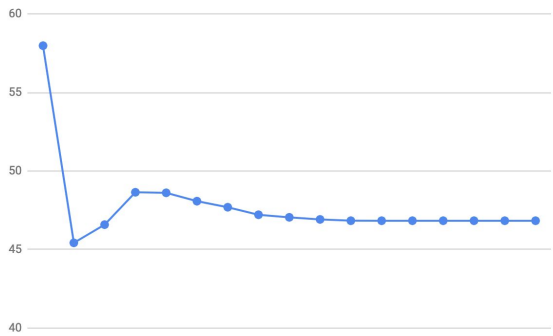
K=	Log Likelihood	BIC
1	-5459.398099570324	-5477.842496840894
2	-5344.462849956579	-5385.040523951832
3	-5339.679047920735	-5402.389998640671
4	-5341.11522998659	-5425.9594574312105
5	-5335.675977467001	-5442.653481636305

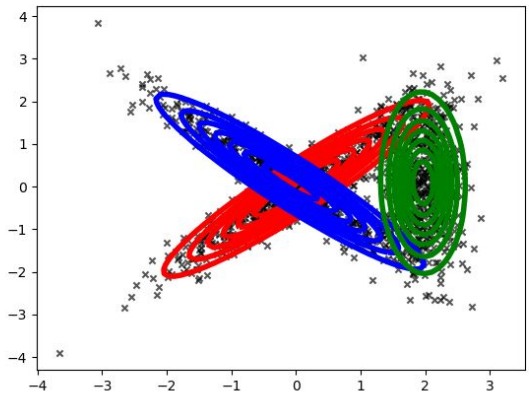


In this dataset, K-means does an okay job separating the data, but it doesn't reflect the overlap between the two gaussians that generate the data. EM, however, does show this by showing the two distributions overlapping in the middle. As for finding the best value of K using BIC scores, as we should expect a K value of 2 produces the highest BIC, which makes sense given the fact that we know the data comes from two separate distributions



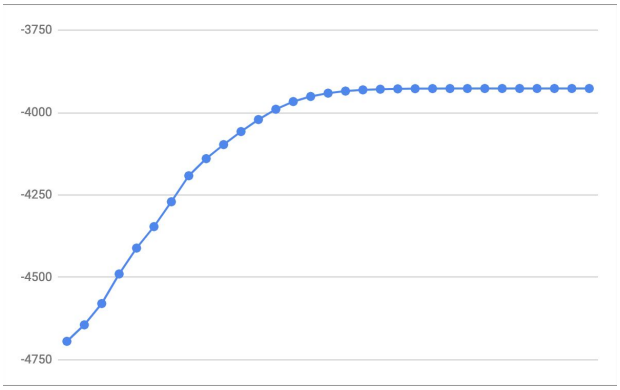




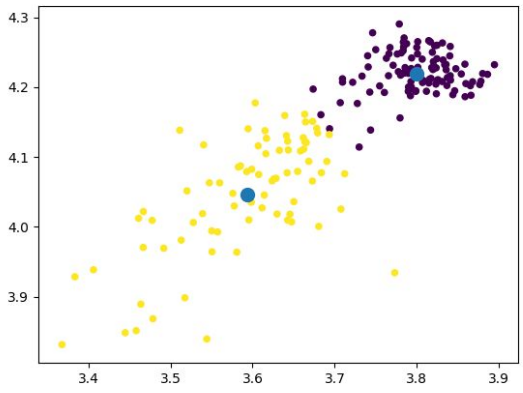


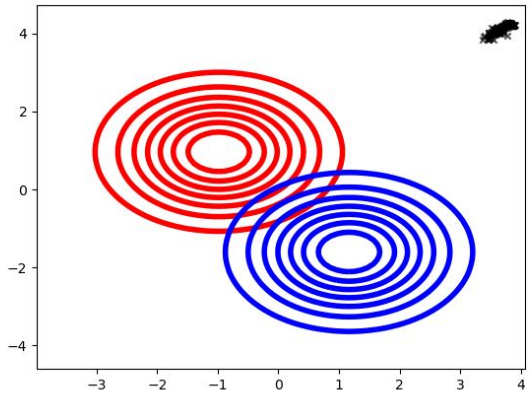
BIC Table

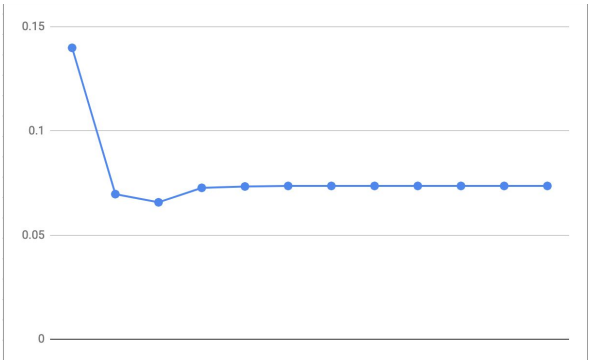
K=	Log Likelihood	BIC
1	-4676.771903597165	-4695.054954564891
2	-4424.361273040321	-4464.583985169318
3	-3927.365095411104	-3989.5274687013716
4	-3923.349177591916	-4007.451212043455
5	-3915.408118084927	-4021.4498136977368

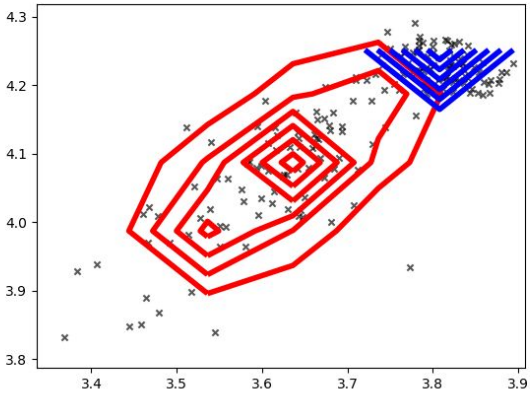


For the second dataset, two of the distributions overlap to such a large degree that K-means completely fails at separating them. Instead, K-means will lump a lot of the data into the third distribution, and split the leftovers between the two remaining clusters. EM, however, does encapsulate the overlapping distributions that make up the data, and it is clear that it presents a much more accurate grouping of the data. Again, for finding the best value of K using BIC scores, a K value of 3 produces the highest BIC, which points to the data coming from 3 separate distributions.



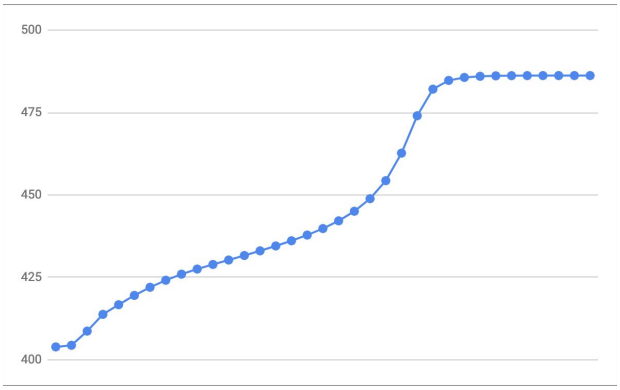






BIC Table

K=	Log Likelihood	BIC
1	402.7884986965029	389.77848197881093
2	485.35392931446	456.7318925355376
3	425.59018297562307	381.3561261354703
4	514.0421753292372	454.1960984278541
5	529.5380271613083	454.07993019869474



Finally, for the last dataset, we actually see K-means performing much better than EM. This is due to a number of reasons, the primary one being that this data is not generated from gaussian distributions at all - it is taken from real life medical data. In addition, the data is so sparse that EM has a hard time defining distributions from such a small sample size. This shows that K-means is a better algorithm for finding some initial exploratory findings in sparse data where the original distributions are unknown. For finding the optimal value of K, a K value of 2 results in the highest BIC, leading 4 and 5 slightly. Seeing as the data is pulled from two separate populations, this is what we would want to see.