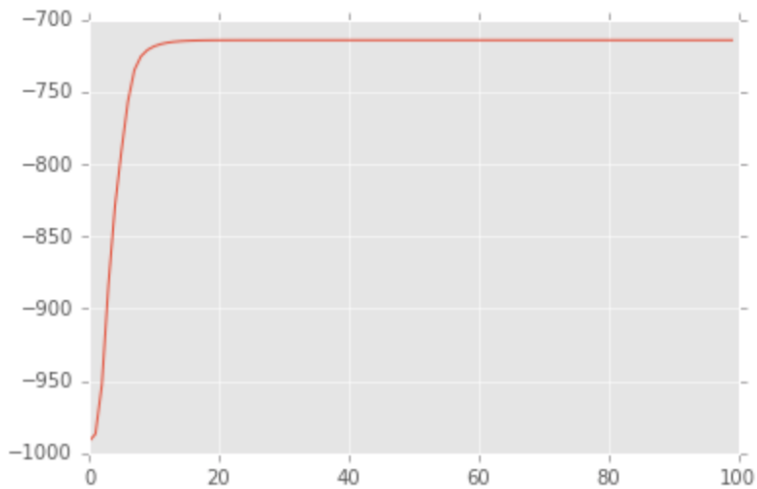
Question 1. EM Algorithm

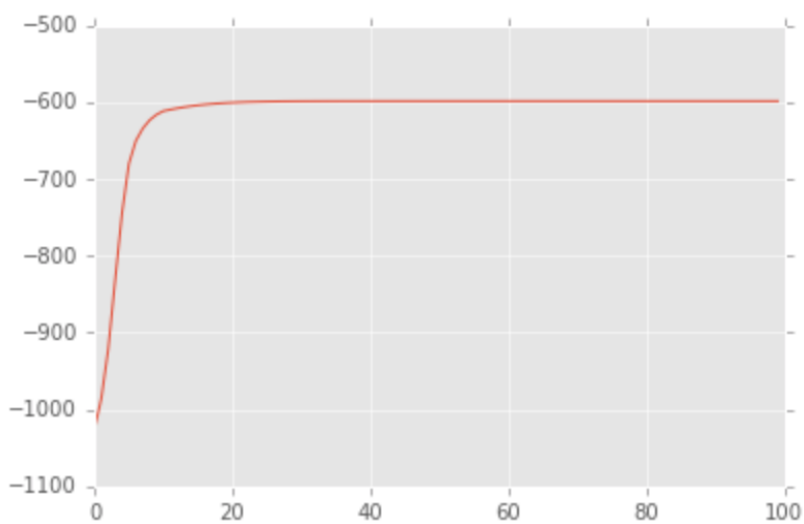
a) **Implement the EM-GMM algorithm and run it for 100 iterations on the data provided for K = 2; 4; 8; 10.**
Answer: Please find the codes attached.

b) **For each K, plot the log likelihood over the 100 iterations. What pattern do you observe and why might this not be the best way to do model selection?**

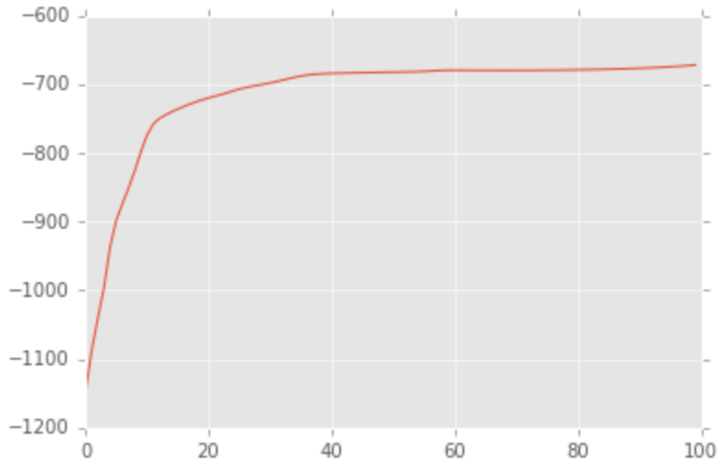
i) K =2



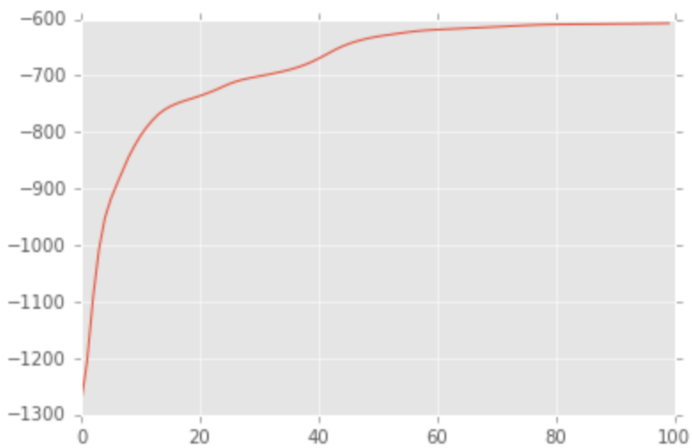
ii) K = 4



iii) K = 8



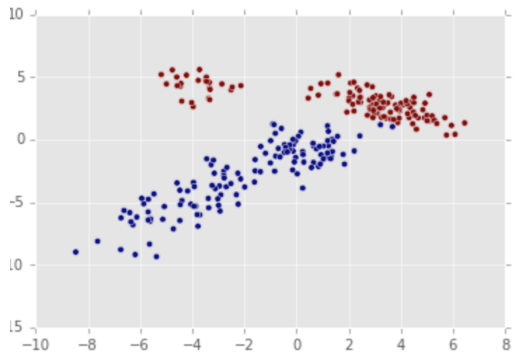
iv) K = 10



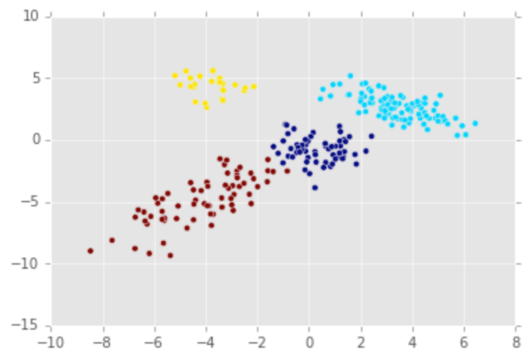
This might not be a good way of model selection because we have to pre-set the number of clusters, and when the number K goes large the clusters are all mixed up together. After many iterations they all converge. It might be hard to tell by eyes what number should be an ideal number of clusters, especially when **d** is large, we can hardly visualize the data, not to mention choosing a reasonable number of clusters.

c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to q(c_i) by a cluster-specific symbol. What do you notice about these plots as a function of K?

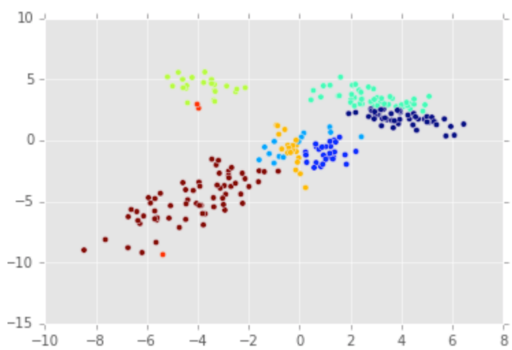
K = 2



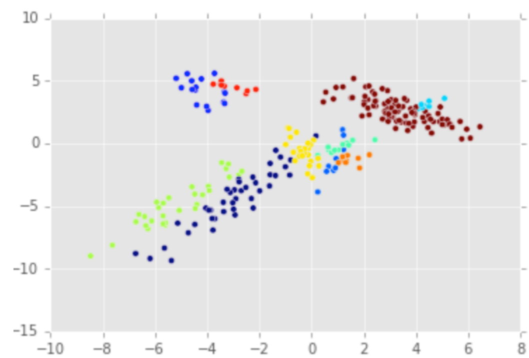
K = 4



K = 8



K = 10



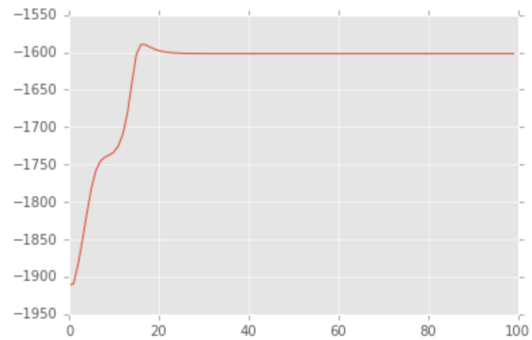
I differentiated the clusters using different colors. From the above four graphs we observe that, for K=2,4 the clusters still have a clear boundary, whereas when K rise to 8 and 10, it looks like we are forcing them into different clusters and all seem to mix up together.
In the end we have number of clusters equal to what we've set previously.

Question 2. Variational Inference

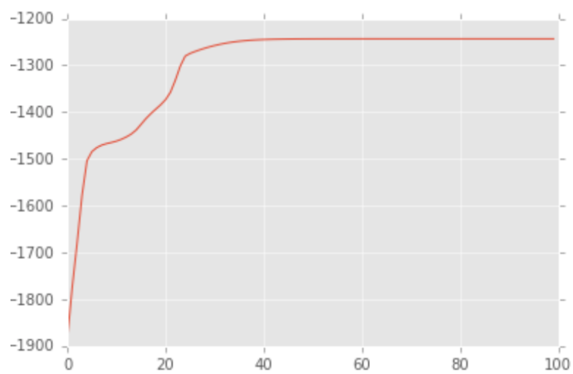
a) **Implement the variational inference algorithm discussed in class and in the notes for K = 2; 4; 10; 25 and 100 iterations each.**
Answer: Please kindly find the code ☺

b) **For each K, plot the variational objective function over the 100 iterations. What pattern do you observe?**

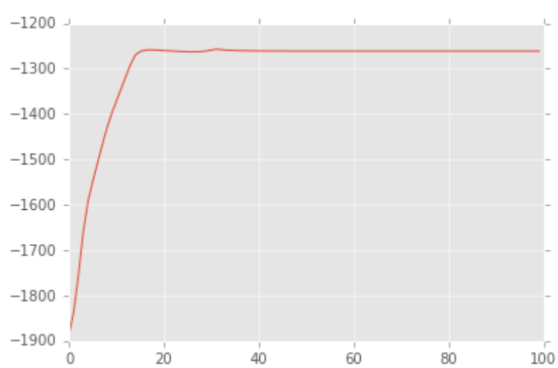
i) K =2



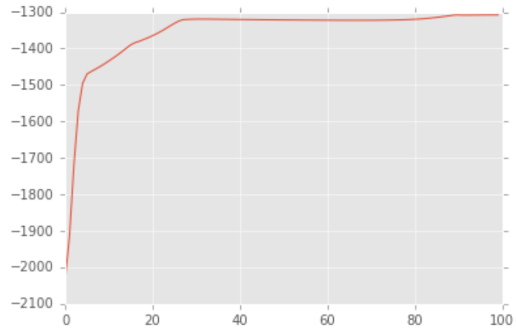
ii) K = 4



iii) K =10



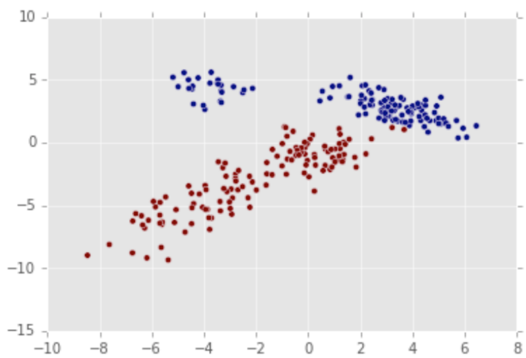
iv) K = 25



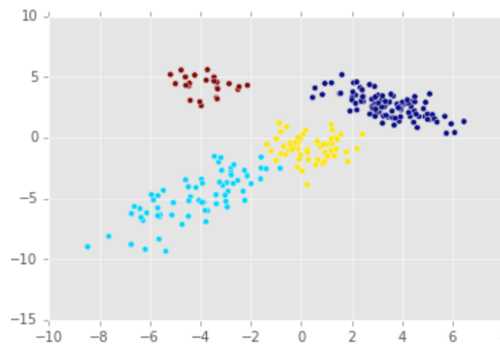
Still, they all seem converge after around 40 iterations.

c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to q(ci) by a cluster-specific symbol. What do you notice about these plots as a function of K?

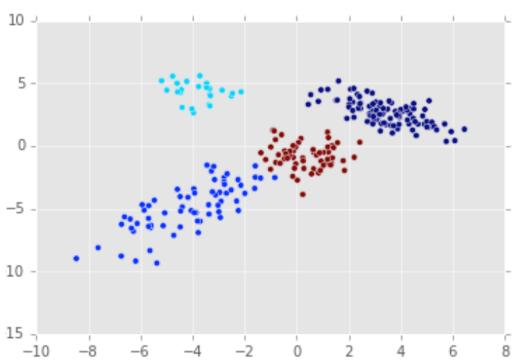
K = 2



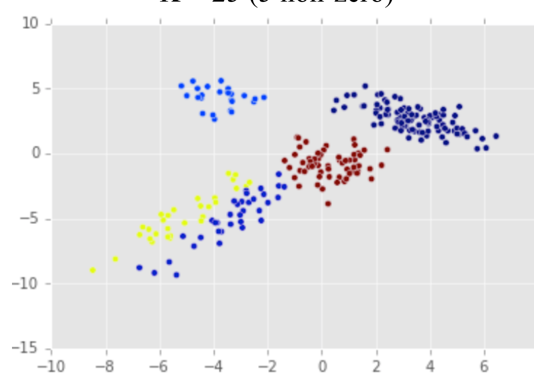
K = 4



K = 8 (4 non-zero)



K = 25 (5 non-zero)



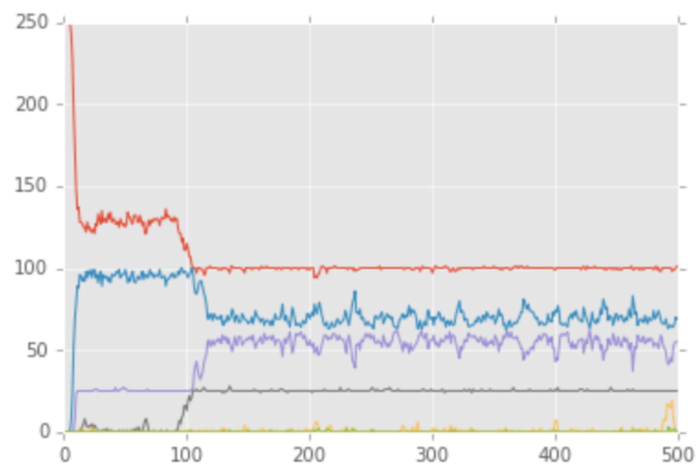
I noticed that as K increases, the number of non-zero clusters does not nessarily go up. Instead, as K changes from 4 to 8, the number of clusters that actually have data there stays unchanged, whereas when K goes up to 25 there are only 5 non-empty clusters.

Question 3: Gibbs Sampling

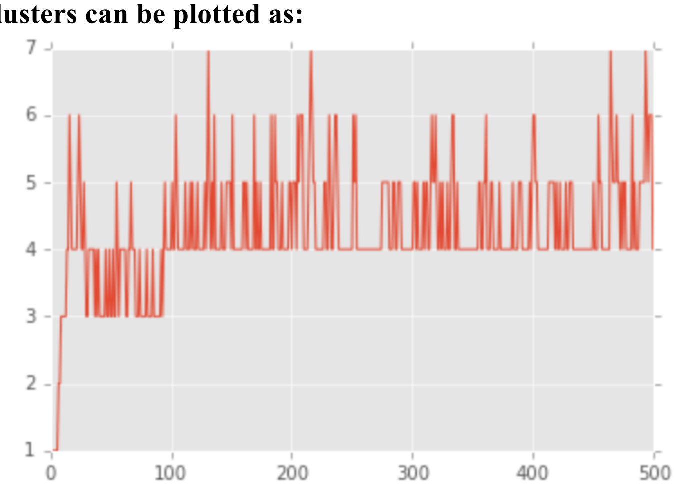
a) **Implement the above-mentioned Gibbs sampling algorithm discussed in class and described in the notes. Run your algorithm on the data provided for 500 iterations.**

Answer: Still, please find the code ☺

b) **Plot the number of observations per cluster as a function of iteration for the six most probable clusters. These should be shown as lines that never cross; for example, the ith value of the \second" line will be the number of observations in the second largest cluster after completing the ith iteration. If there are fewer than six clusters then set the remaining values to zero.**



c) **Plot of the total number of clusters that contain data as a function of iteration. Total number of clusters can be plotted as:**



And since at the last iteration the number of clusters is 4, we pretty much arrived at a familiar figure: