Zoltan Onodi-Szucs
zo2131

Homework 4
Bayesian Machine Learning

Problem Set-up

We are given observations $X = \{x_1, \ldots, x_n\}$ where each $x_i \in \mathbb{R}^d$. We model this as being generated from a Gaussian mixture model of the form

$$x_i \,|\, c_i \sim Normal(\mu_{c_i}, \Lambda_{c_i}^{-1}), \qquad c_i \overset{iid}{\sim} Discrete(\pi)$$

In this homework, you will implement three algorithms for learning this mixture model, one based on maximum likelihood EM, one on variational inference and one on Gibbs sampling. Use the data provided for all experiments.

Problem 1. (30 points)

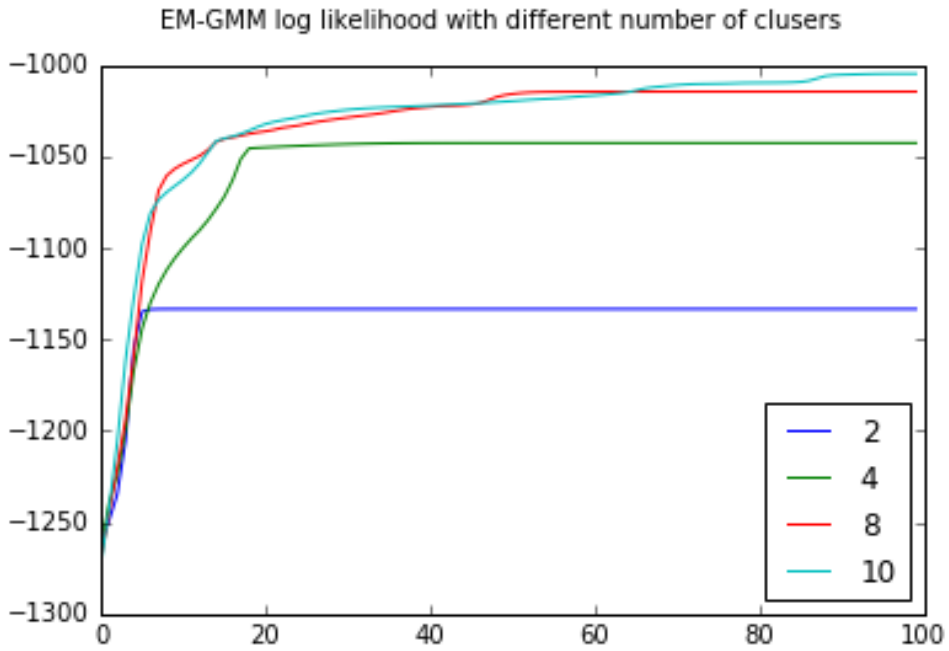
In this problem, you will implement the EM algorithm for learning maximum likelihood values of π and each (μ_j, Λ_j) for $j = 1, \ldots, K$. The algorithm is given in the notes, and also in Section 9.2 of Bishop's book.

a) Implement the EM-GMM algorithm and run it for 100 iterations on the data provided for $K = 2, 4, 8, 10$.

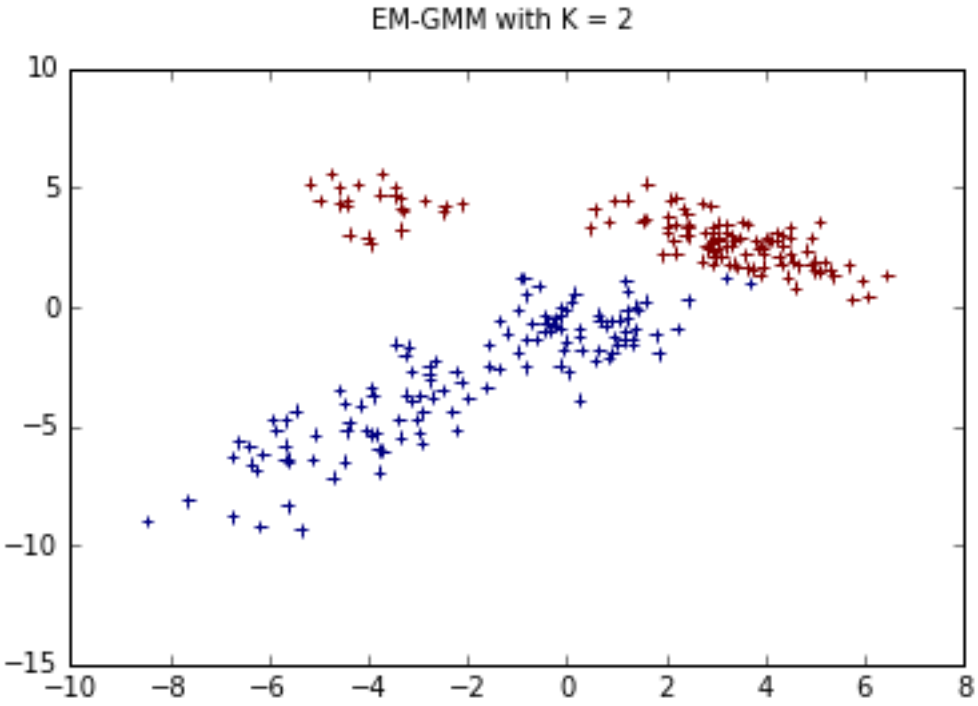
b) For each K, plot the log likelihood over the 100 iterations. What pattern do you observe and why might this not be the best way to do model selection?

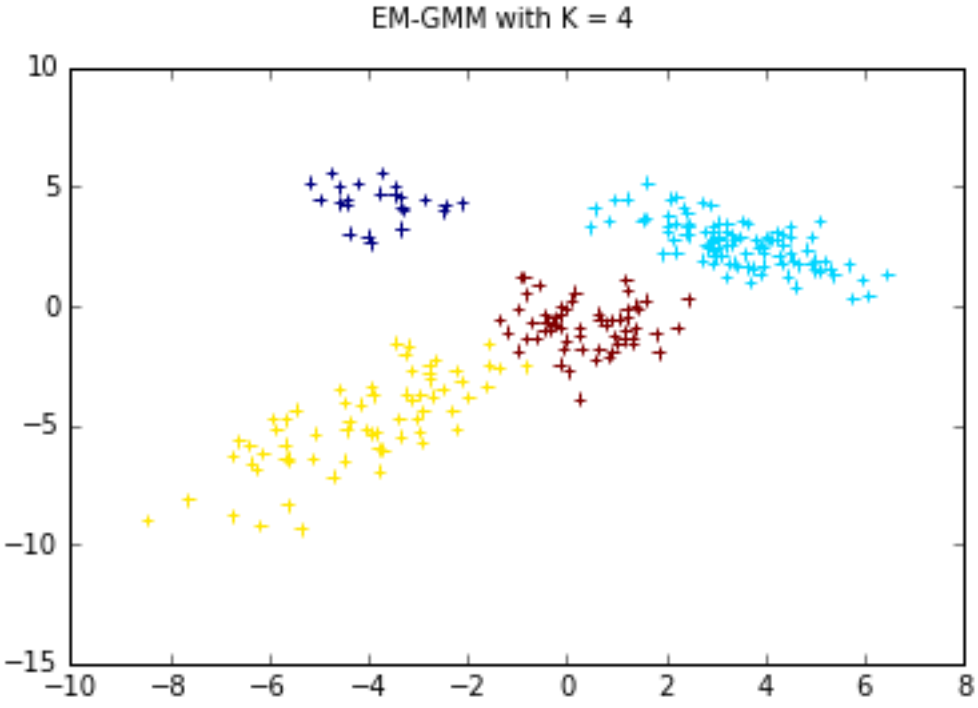
c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to $q(c_i)$ by a cluster-specific symbol. What do you notice about these plots as a function of K?

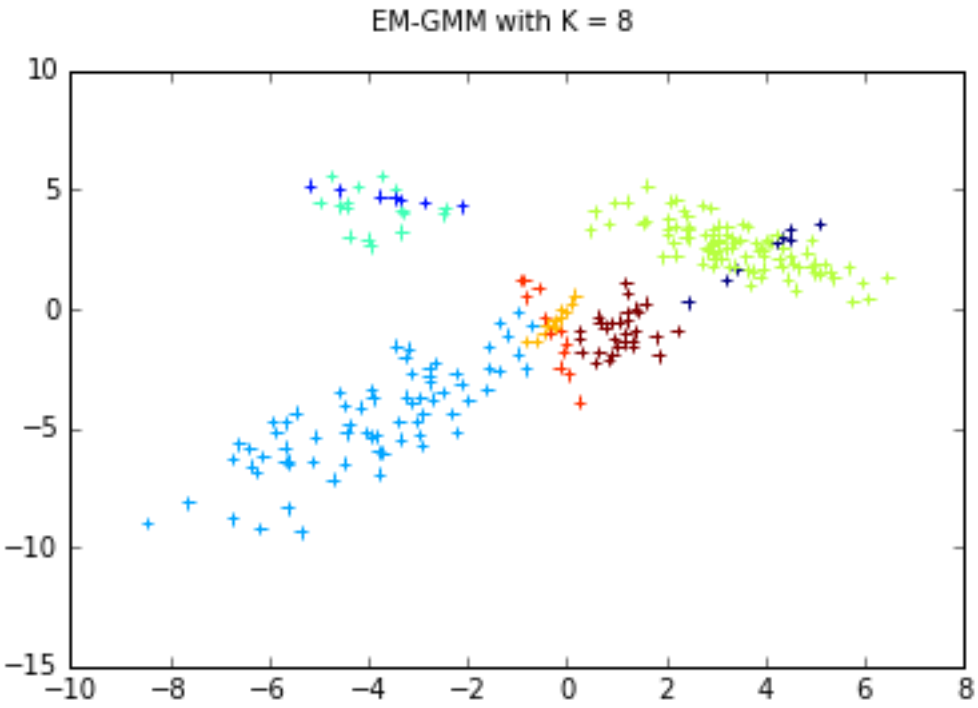
B,

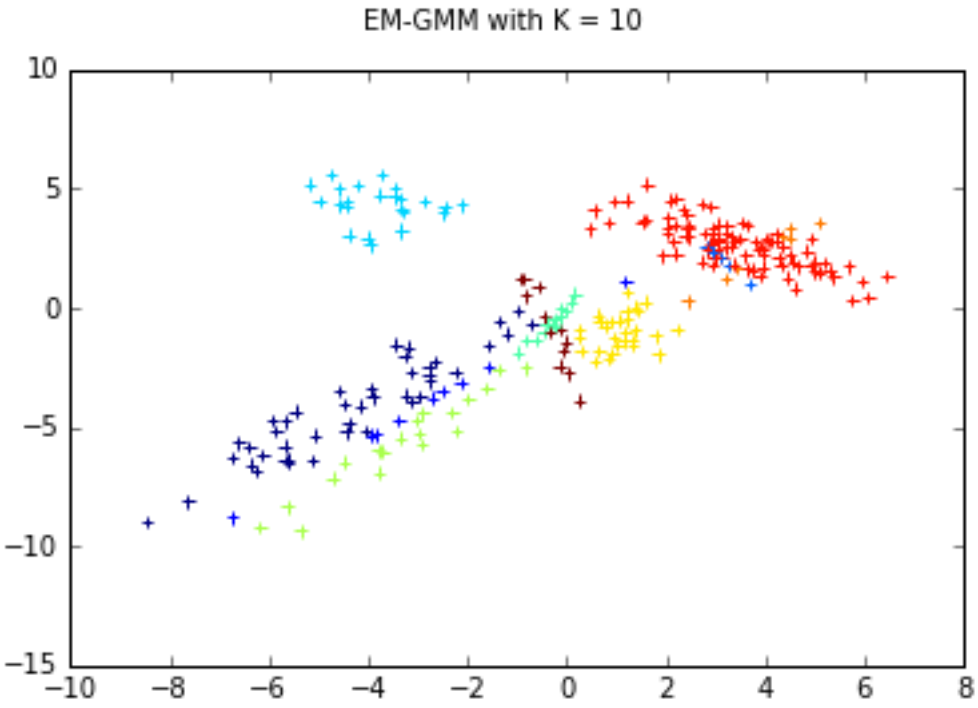


C,


EM-GMM with K = 2
10
5
0
−5
−10
−15
−10
−8
−6
−4
−2
0
2
4
6
8


EM-GMM with K = 4
10
5
0
−5
−10
−15
−10
−8
−6
−4
−2
0
2
4
6
8





As k increases, the number of non-empty clusters increase with it, which is the property of this type of classification. Originally I would have separated them into 4 clusters, but the higher number of K revealed some interesting structure too as it decomposed previous clusters into smaller one (like the brown one in the last plot) which indeed seems like it deserves its own distribution.

Problem 2. (35 points)

In this problem, you will implement a variational inference algorithm for approximating the posterior distribution of the GMM variables. We therefore require prior distributions on these variables. For this problem, we use

$$\pi \sim Dirichlet(\alpha), \qquad \mu_j \sim Normal(0, cI), \qquad \Lambda_j \sim Wishart(a, B)$$

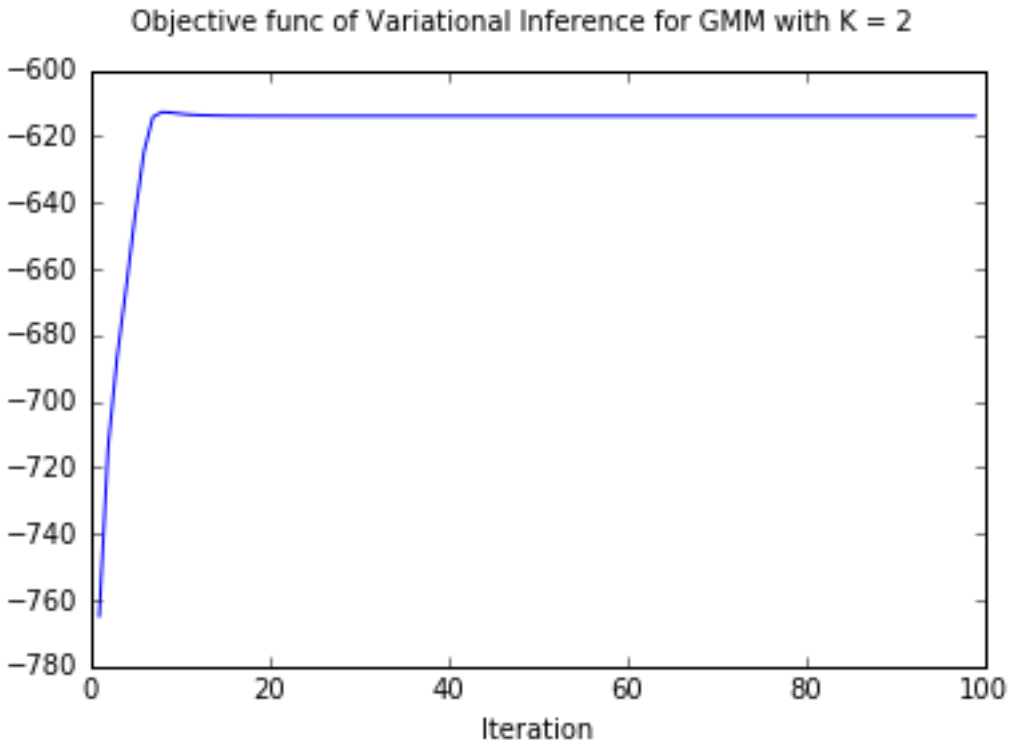
For this problem, set $\alpha = 1$, $c = 10$, $a = d$ and $B = \frac{d}{10}A$ where A is the empirical covariance of the data. Approximate the posterior distribution of these variables with q distributions factorized on π, and each μ_j, Λ_j and c_i as discussed in class.

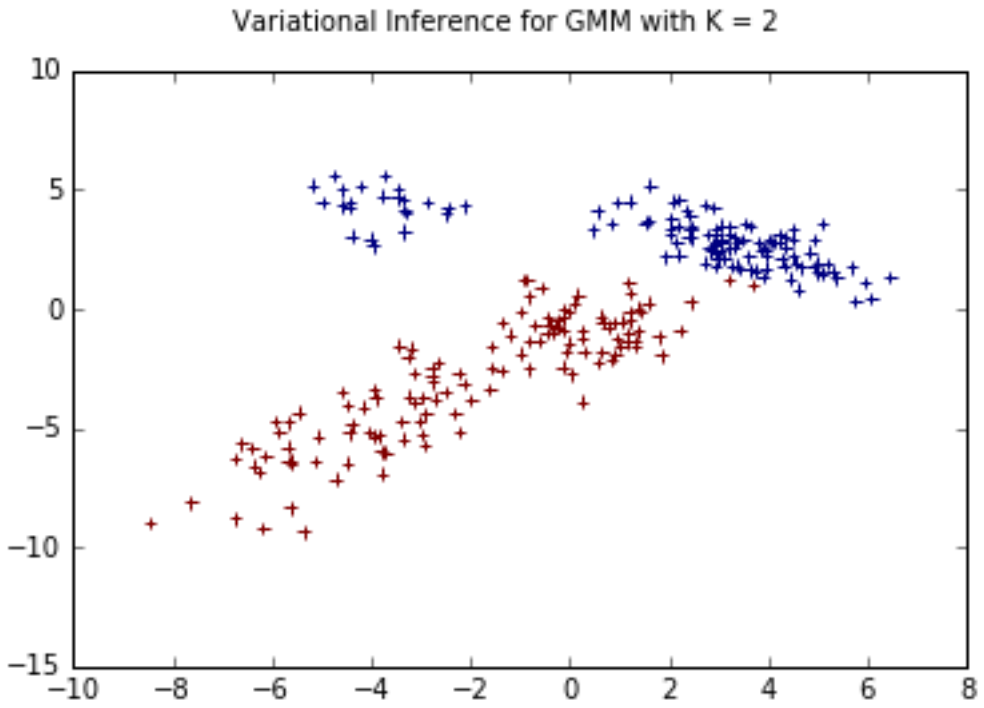
a) Implement the variational inference algorithm discussed in class and in the notes for $K = 2, 4, 10, 25$ and 100 iterations each.

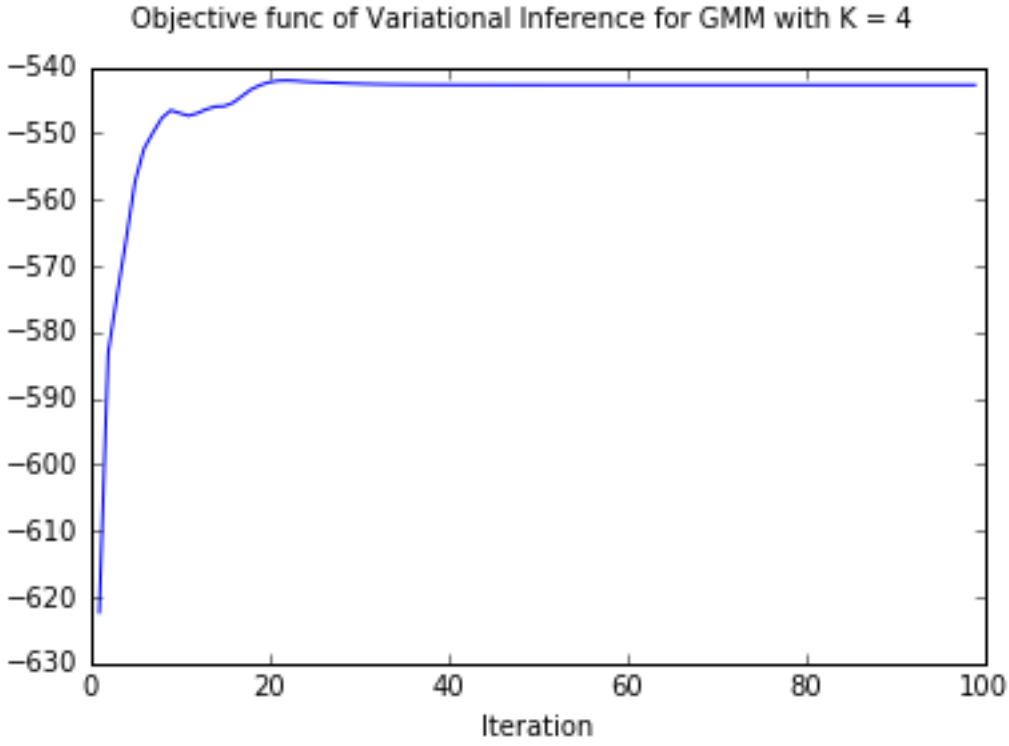
b) For each K, plot the variational objective function over the 100 iterations. What pattern do you observe?

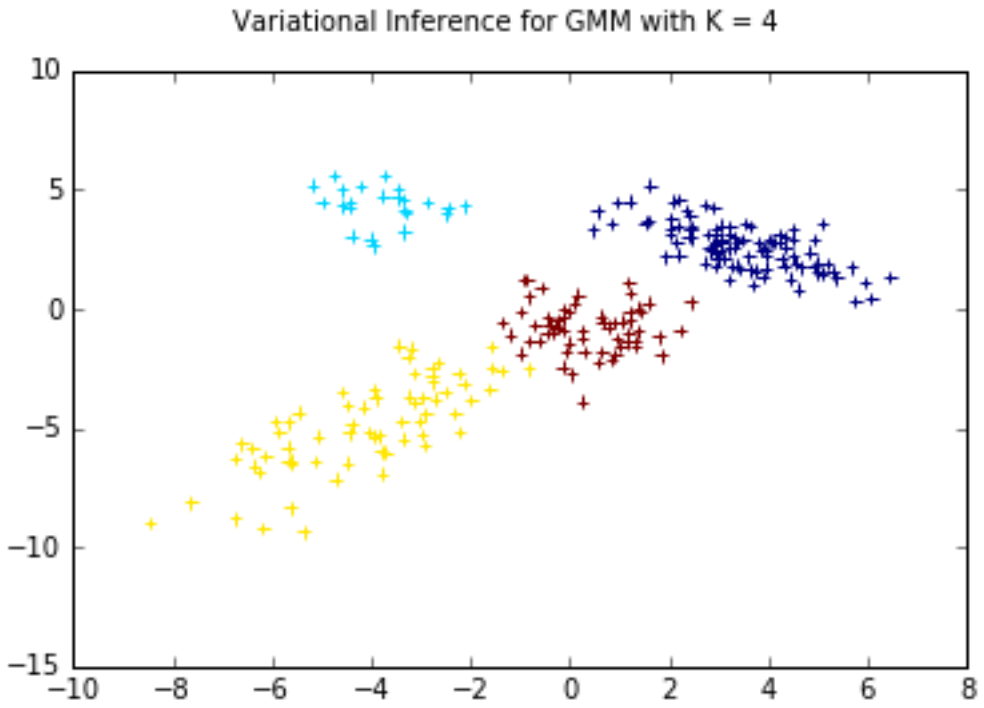
c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to $q(c_i)$ by a cluster-specific symbol. What do you notice about these plots as a function of K?

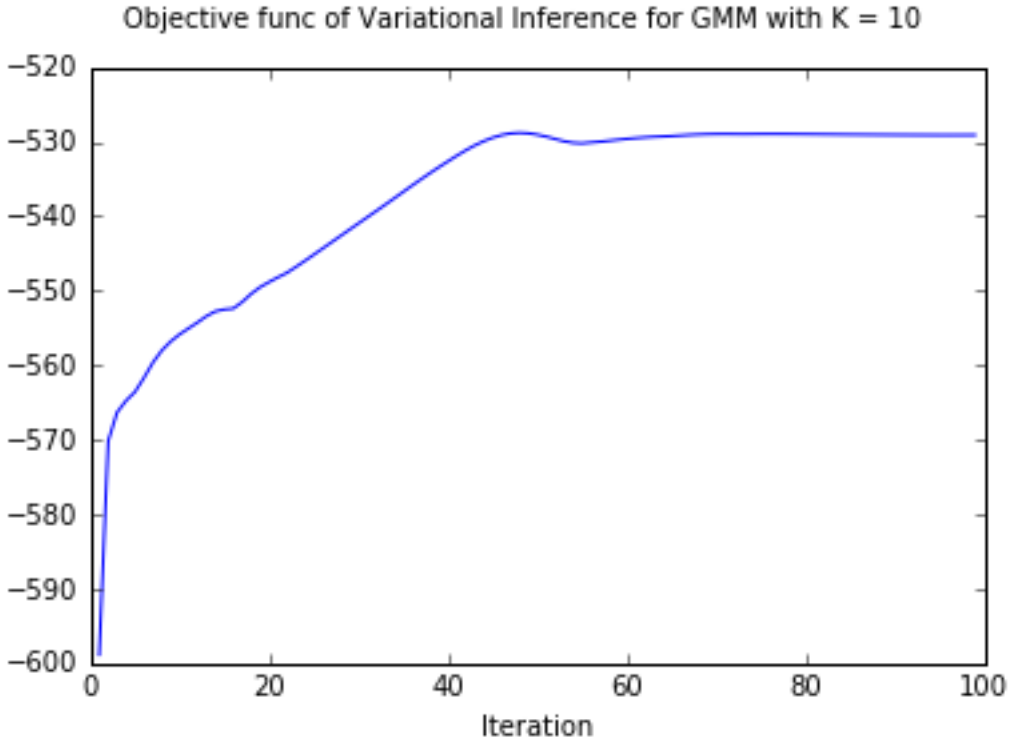
B & C,

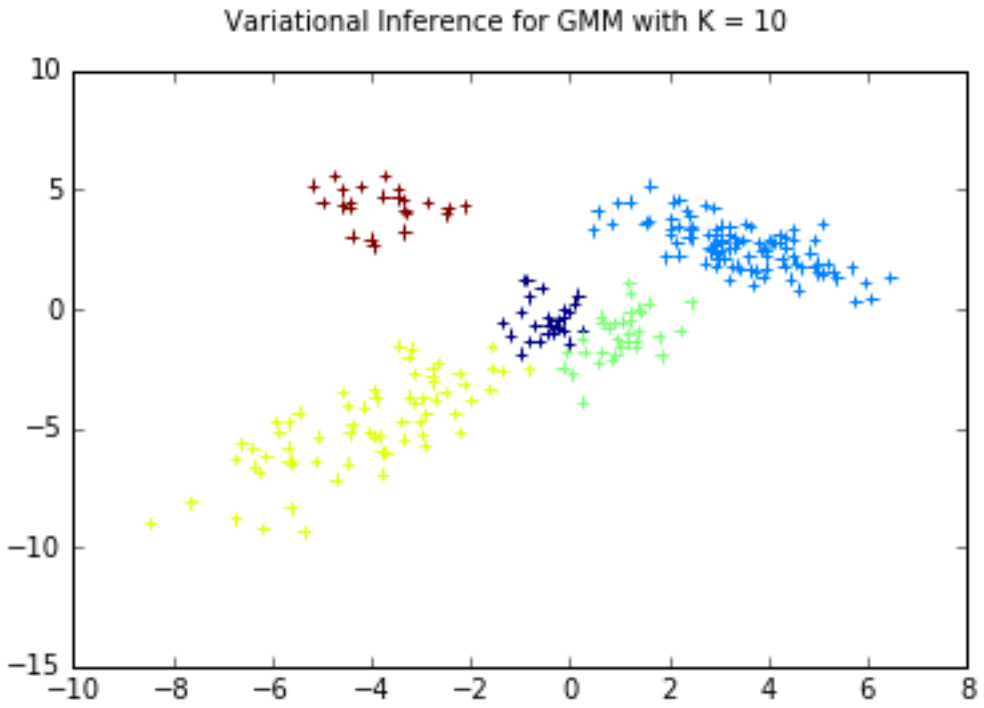


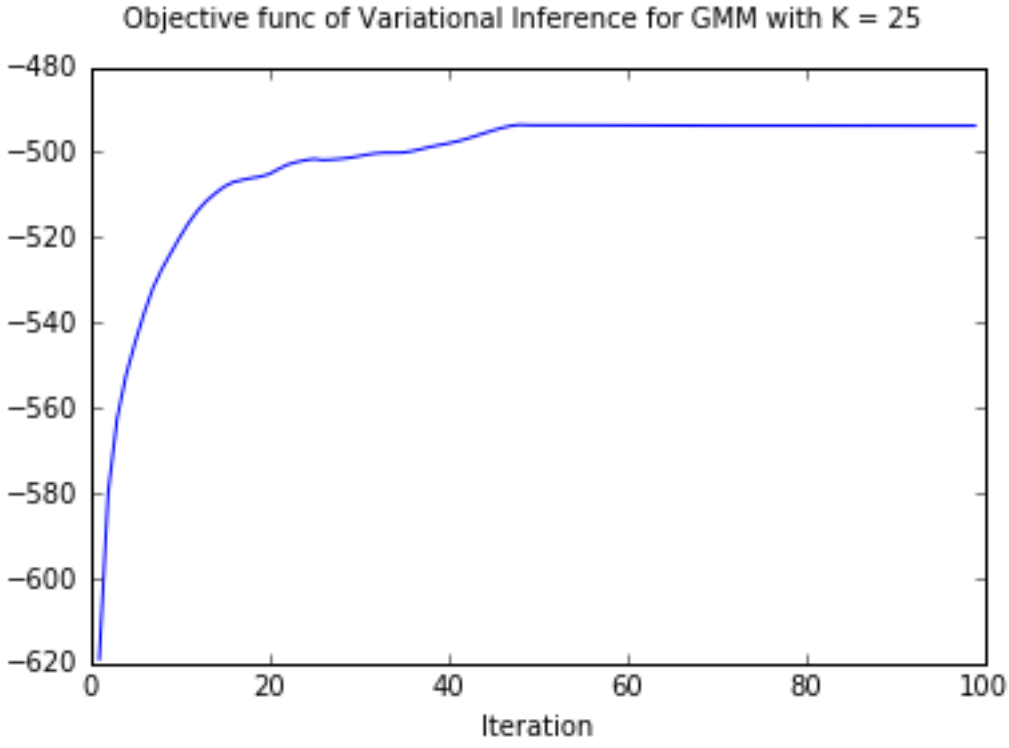

Variational Inference for GMM with K = 2
10
5
0
−5
−10
−15
−10
−8
−6
−4
−2
0
2
4
6
8

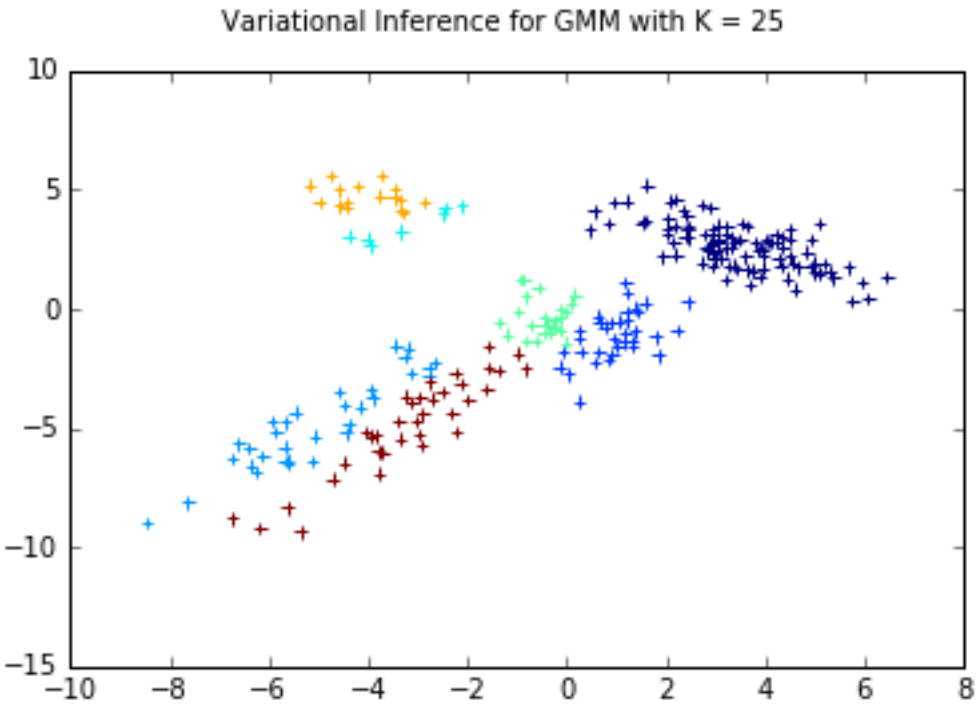

Objective func of Variational Inference for GMM with K = 4
−540
−550
−560
−570
−580
−590
−600
−610
−620
−630
0
20
40
60
80
100
Iteration


Variational Inference for GMM with K = 4
10
5
0
−5
−10
−15
−10
−8
−6
−4
−2
0
2
4
6
8


Objective func of Variational Inference for GMM with K = 10
−520
−530
−540
−550
−560
−570
−580
−590
−600
0
20
40
60
80
100
Iteration


Variational Inference for GMM with K = 10
10
5
0
−5
−10
−15
−10
−8
−6
−4
−2
0
2
4
6
8


Objective func of Variational Inference for GMM with K = 25
−480
−500
−520
−540
−560
−580
−600
−620
0
20
40
60
80
100
Iteration



For the objective functions, we can see that as the number of k increases, the objective function increases, as well as it takes more time to converge (in K=2 it converges pretty soon, and takes a lot more time for higher K).

As for the number of cluster used, we see that from 4 on, we don't actually end up having the max amount of clusters. For K= 10 its 6, for K = 25 its 7. That is because our dirichlet prior.

Problem 3. (35 points)

In this problem, you will implement a Bayesian nonparametric sampler for a marginalized version of the GMM. In contrast to Problem 2, in this problem we will use a joint prior on (μ_j, Λ_j). This is done for computational convenience in calculating the marginal distribution of the data. Specifically, we use the prior distribution

$$\mu_j \,|\, \Lambda_j \sim Normal(m, (c\Lambda)^{-1}), \qquad \Lambda_j \sim Wishart(a, B)$$

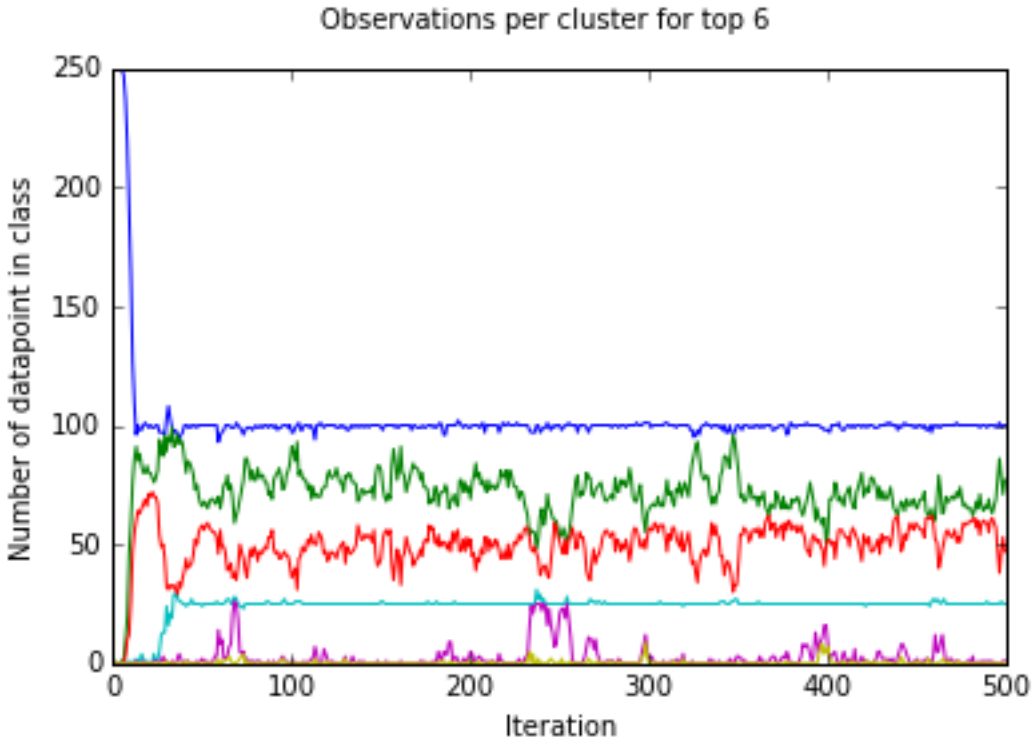
as well as the limit of the prior $\pi \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$ as $K \to \infty$.

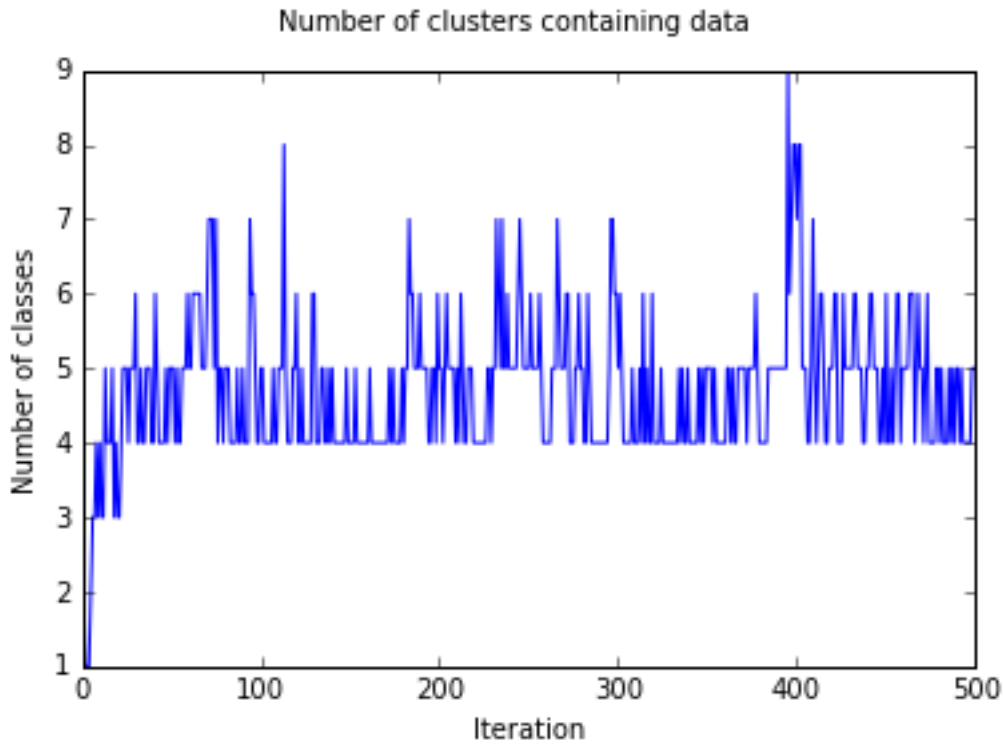
In this problem you will implement the marginal sampler where π is integrated out. For this problem, set m to be the empirical mean of the data, $c = 1/10$, $a = d$ and $B = c \cdot d \cdot A$ where A is the empirical covariance of the data. For the "cluster innovation parameter" set $\alpha = 1$.

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.

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.





It can be seen that the number of clusters stabilizes around 4, occasionally a couple new is created (these are single or very few point clusters) which die soon after. Two of the clusters seems pretty stable, while the other two is constantly trading data points.