HOMEWORK 4

Name: Yang Bai UNI: yb2356

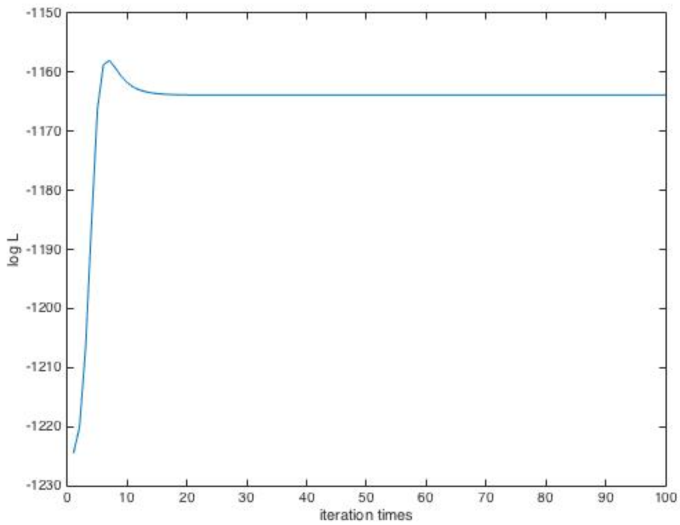
- **Problem1**

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

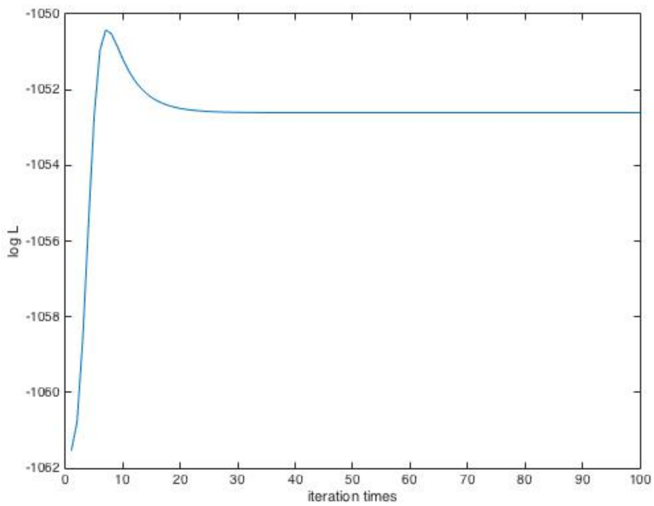
When K =8, 10, I set the iteration number to 200 because 100 seems not enough. If we initialize the parameter randomly, the result usually becomes strange. So I used some initialization algorithm for gaussian mixture model.

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?

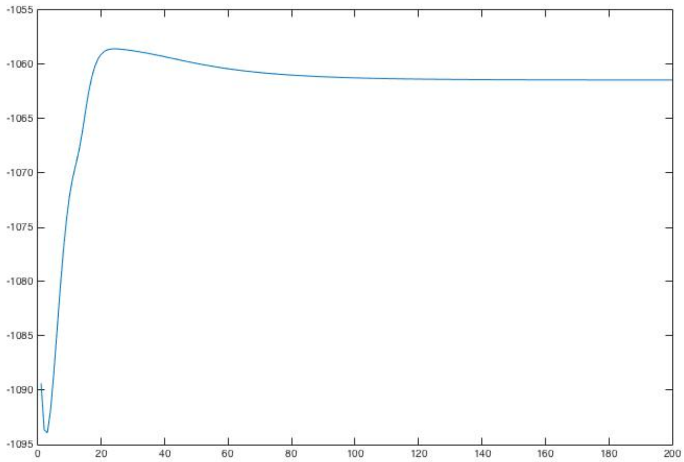
K =2:



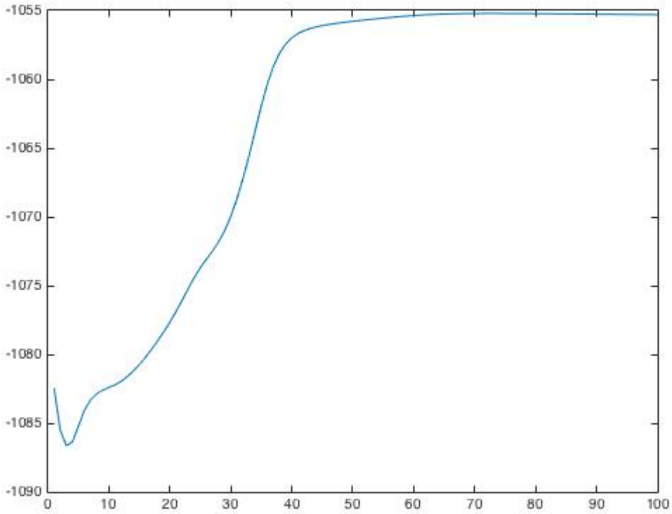
K =4:



K =8:



K =10

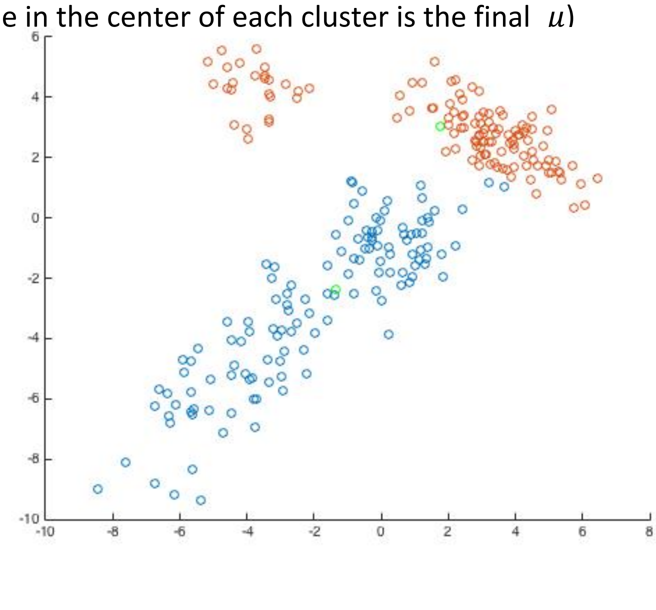


Log likelihoods all converge after several hundred iterations.

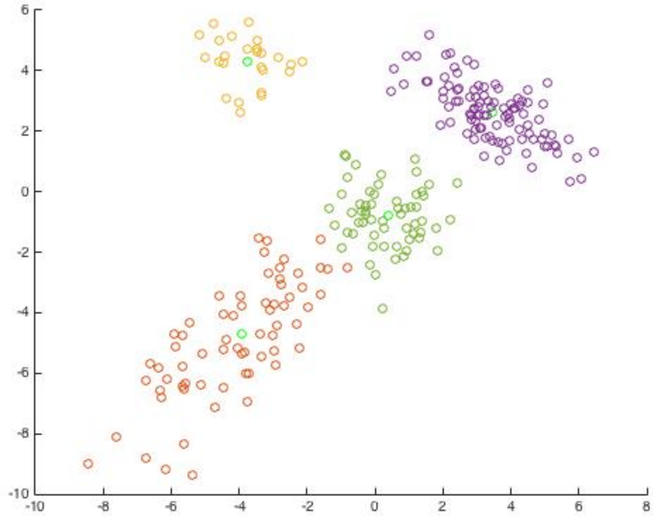
c) For the final iteration of each model, plot the data and indicate the most probable cluster of each observation according to q(ci) by a cluster-specific symbol. What do you notice about these plots as a function of K?

(The green circle in the center of each cluster is the final u)

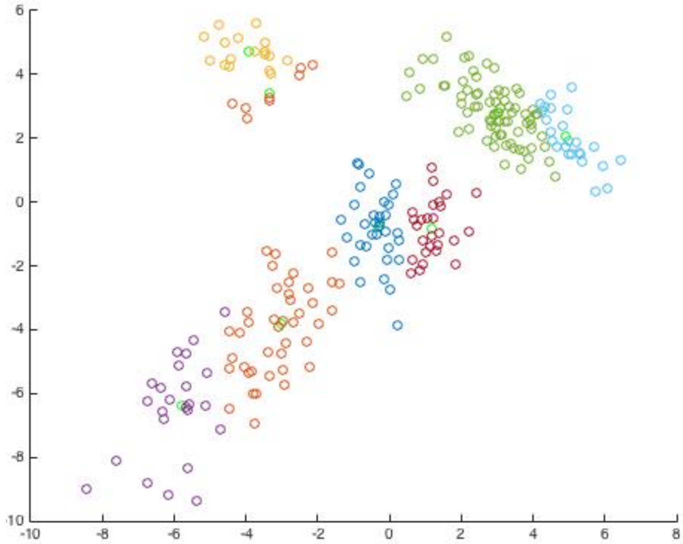
K=2



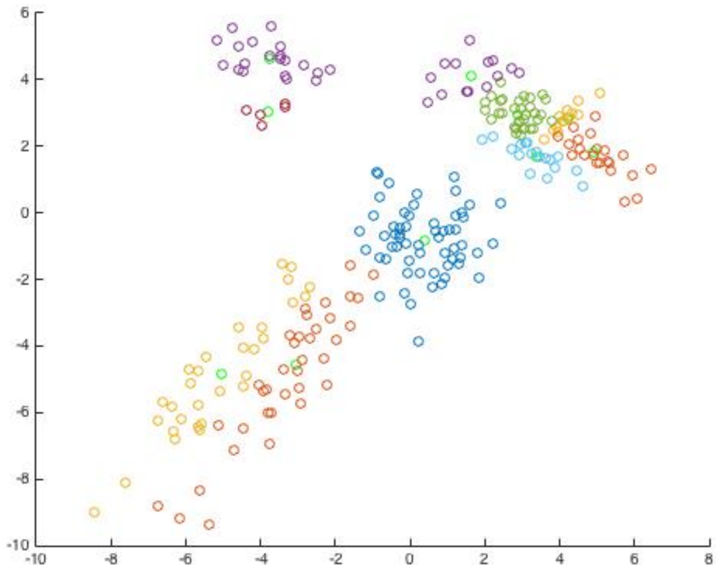
K=4



K =8



K=10



When k = 8 ,10, the clustering result looks mixed up. The probable explanation is that the data does not have so many distinct clusters, so when K is too big, the algorithm can not return a correct result.

- **Problem 2**

variational objective function

$$\mathcal{L} = \mathrm{E}[\ln p(x, c, \pi, \mu, \Lambda)] - \mathrm{E}[\ln q]$$

The first part of L:

$$\mathrm{E}[\ln p(x, c, \pi, \mu, \Lambda)]$$

$$= \sum_{i=1}^{n} \mathrm{E}\left[\ln p\left(x_i \middle| c_i, \mu_{ci}, {\Lambda_{ci}}^{-1}\right)\right] + \sum_{i=1}^{n} \mathrm{E}[\ln p(c_i|\pi)] + \mathrm{E}[\ln p(\pi)]$$

$$+ \sum_{j=1}^{K} \mathrm{E}\left[\ln p(\mu_j)\right] + \sum_{j=1}^{K} \mathrm{E}\left[\ln p(\Lambda_j)\right]$$

$$\mathrm{E}\left[\ln p\left(x_i \middle| c_i, \mu_{ci}, {\Lambda_{ci}}^{-1}\right)\right] = \frac{1}{2}\mathrm{E}[\ln|\Lambda_{ci}|] - \frac{1}{2}\mathrm{E}[(x_i - \mu_{ci})^T \Lambda_{ci} (x_i - \mu_{ci})] + -\frac{d}{2}\ln 2\pi$$

$$= -\frac{d}{2}\ln 2\pi + \frac{1}{2}\sum_{j=1}^{K} \phi_i(j)\mathrm{E}\left[\ln|\Lambda_j|\right]$$

$$-\frac{1}{2}\sum_{j=1}^{K} \phi_i(j)\left[\left(x_i - \mathrm{E}[\mu_j]\right)^T \mathrm{E}[\Lambda_j]\left(x_i - \mathrm{E}[\mu_j]\right) - trace\left(\mathrm{E}[\Lambda_j]\Sigma_j'\right)\right]$$

$$\mathrm{E}[\ln p(c_i|\pi)] = \sum_{j=1}^{K} \phi_i(j)\mathrm{E}[\ln \pi_j] = \sum_{j=1}^{K} \phi_i(j)\left(\psi(\alpha_j') - \psi\left(\sum\nolimits_k \alpha_k'\right)\right)$$

$$\mathrm{E}[\ln p(\pi)] = \mathrm{E}\left[\ln\left(\frac{\Gamma\left(\sum_{j=1}^{K}\alpha_j\right)}{\prod_{j=1}^{K}\Gamma(\alpha_j)}\prod_{j=1}^{K}\pi_j^{\alpha_j - 1}\right)\right]$$

$$= \ln\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \sum_{j=1}^{K}\ln\Gamma(\alpha_j) + \sum_{j=1}^{K}(\alpha_j - 1)\mathrm{E}[\ln \pi_j]$$

$$= \sum_{j=1}^{K}(\alpha_j - 1)\left(\psi(\alpha_j') - \psi\left(\sum\nolimits_k \alpha_k'\right)\right) + \ln\Gamma\left(\sum_{j=1}^{K}\alpha_j\right) - \sum_{j=1}^{K}\ln\Gamma(\alpha_j)$$

$$\mathrm{E}\left[\ln p(\mu_j)\right] = -\frac{1}{2c}\left(m_j'^T m_j' + trace(\Sigma_j')\right) + const$$

$$\mathrm{E}\left[\ln p(\Lambda_j)\right] = \frac{a-d-1}{2}\mathrm{E}\left[\ln|\Lambda_j|\right] - \frac{1}{2}trace\left(B \cdot \mathrm{E}[\Lambda_j]\right) + \frac{ad}{2}\ln 2 - \ln\Gamma_d\left(\frac{a}{2}\right)$$

$$\mathrm{E}[\ln q\,(c_i|\pi)] = \sum_{j=1}^{K} \phi_i(j)\ln\phi_i(j)$$

$$\mathrm{E}[\ln q(\Lambda_j)] = \frac{a_j' - d - 1}{2}\mathrm{E}[\ln|\Lambda_j|] - \frac{1}{2} trace(B_j' \cdot \mathrm{E}[\Lambda_j]) - \frac{a_j' d}{2}\ln 2 + \frac{a_j'}{2}\ln|B_j'|$$
$$- \ln\Gamma_d\left(\frac{a_j'}{2}\right)$$

$$\mathrm{E}[\ln q(\pi)] = \mathrm{E}\left[\ln\left(\frac{\Gamma(\sum_{j=1}^{K}\alpha_j')}{\prod_{j=1}^{K}\Gamma(\alpha_j')}\prod_{j=1}^{K}\pi_j^{\alpha_j'-1}\right)\right]$$
$$= \sum_{j=1}^{K}(\alpha_j' - 1)\mathrm{E}[\ln \pi_j] + \ln\Gamma\left(\sum_{j=1}^{K}\alpha_j'\right) - \sum_{j=1}^{K}\ln\Gamma(\alpha_j')$$
$$= \sum_{j=1}^{K}(\alpha_j' - 1)\left(\psi(\alpha_j') - \psi\left(\sum\nolimits_k \alpha_k'\right)\right) + \ln\Gamma\left(\sum_{j=1}^{K}\alpha_j'\right) - \sum_{j=1}^{K}\ln\Gamma(\alpha_j')$$
$$\mathrm{E}[\ln q(\mu_j)] = -\frac{1}{2}\ln|\Sigma_j'| + const$$

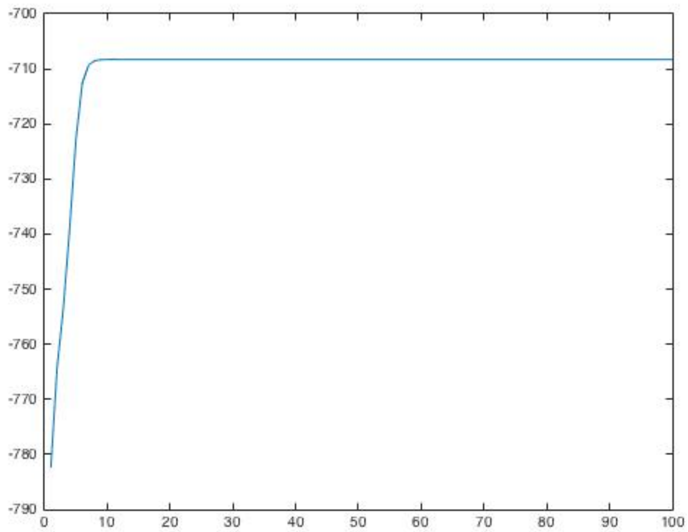
And we can calculate that

$$\mathrm{E}[\Lambda_j] = a_j' B_j'^{-1}$$

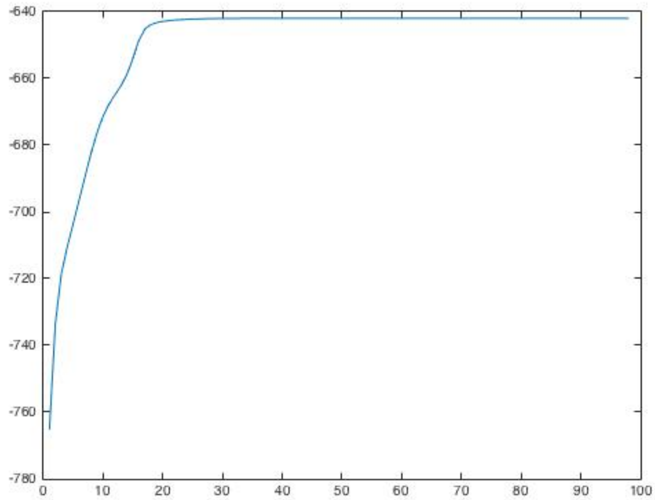
$$\mathrm{E}[\ln|\Lambda_j|] = \frac{d(d-1)}{4}\ln\pi + d\ln 2 - \ln|B_j'| + \sum_{k=1}^{d}\psi(\frac{a_j'}{2} + (1-k)/2)$$

$$\mathrm{E}[\mu_j] = m_j'$$

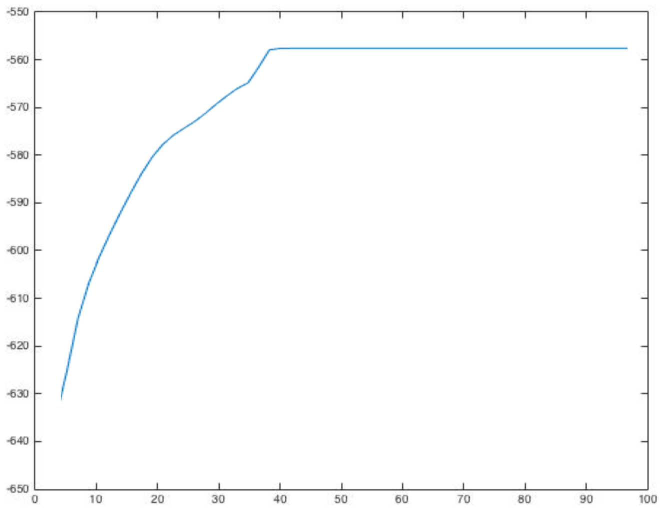
a) Implement the variational inference algorithm discussed in class and in the notes for K = 2, 4, 10, 25 and 100 iterations each.
b) For each K, plot the variational objective function over the 100 iterations. What pattern do you observe?
When calculating L in the code, I did not add any constant in the above formula.
K =2:



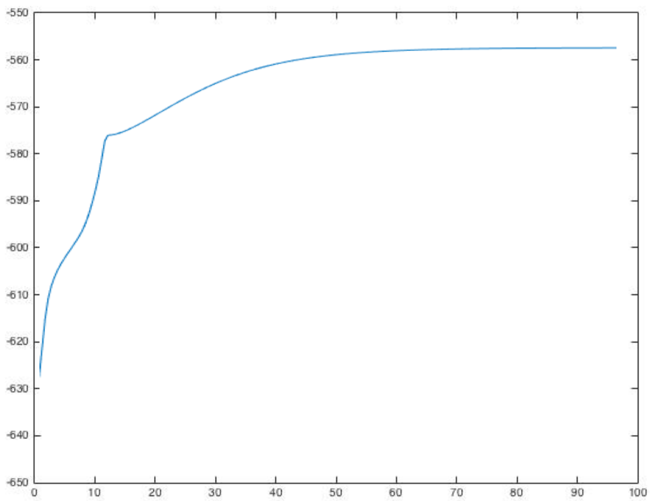
K =4


-640
-660
-680
-700
-720
-740
-760
-780
0
10
20
30
40
50
60
70
80
90
100

K =10


-550
-560
-570
-580
-590
-600
-610
-620
-630
-640
-650
0
10
20
30
40
50
60
70
80
90
100

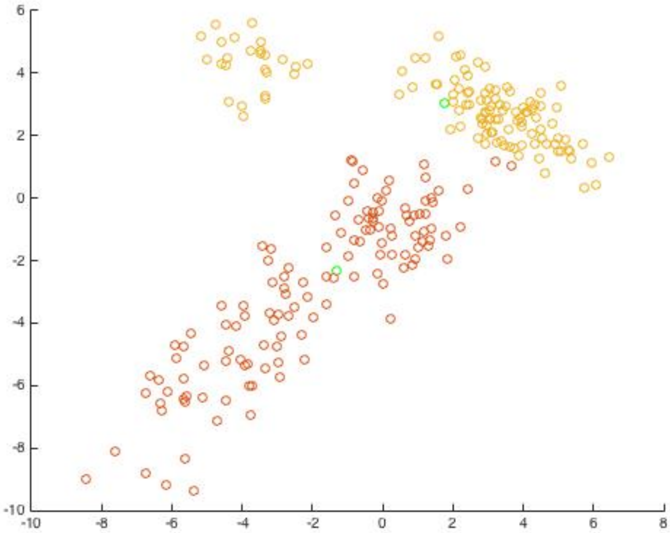
K =25


-550
-560
-570
-580
-590
-600
-610
-620
-630
-640
-650
0
10
20
30
40
50
60
70
80
90
100

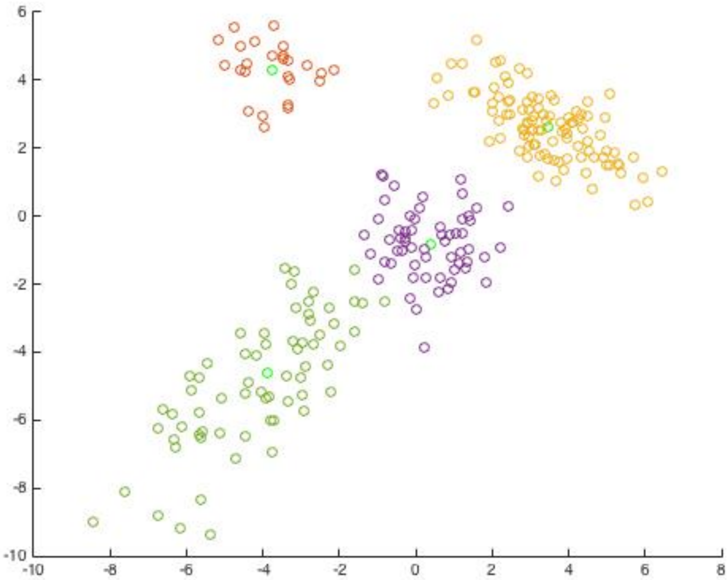
c) For the final iteration of each model, plot the data. What do you notice about these plots as a function of K?

(The green circle in the center of each cluster is the final μ)

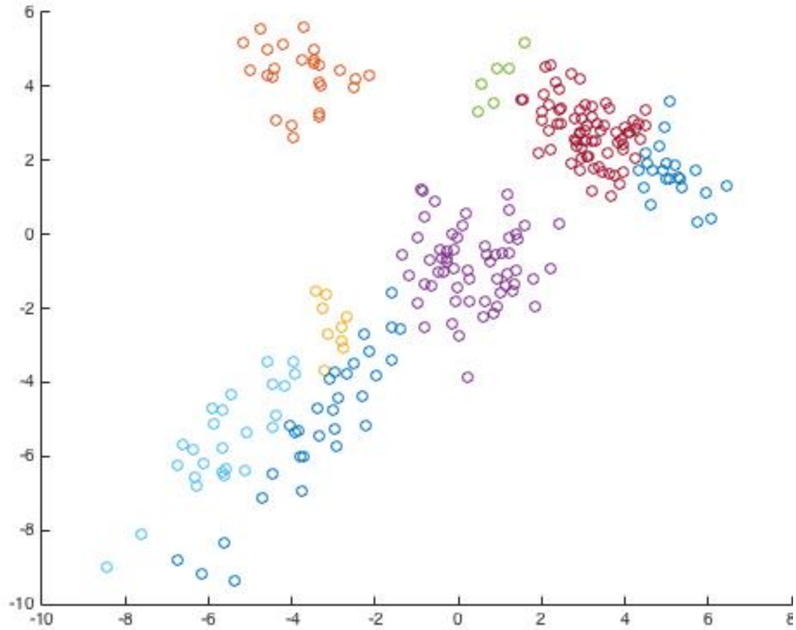
K =2



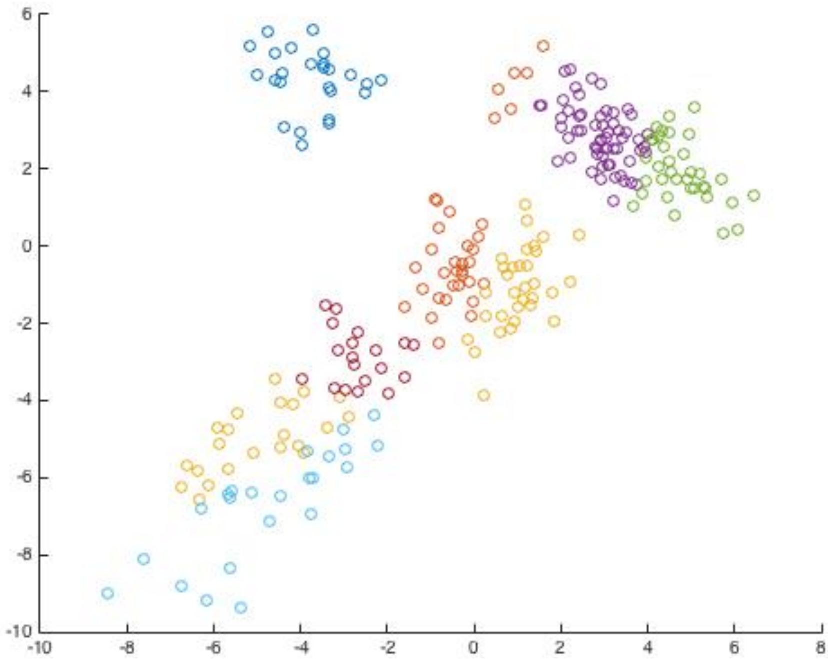
K =4



K =10



K=25

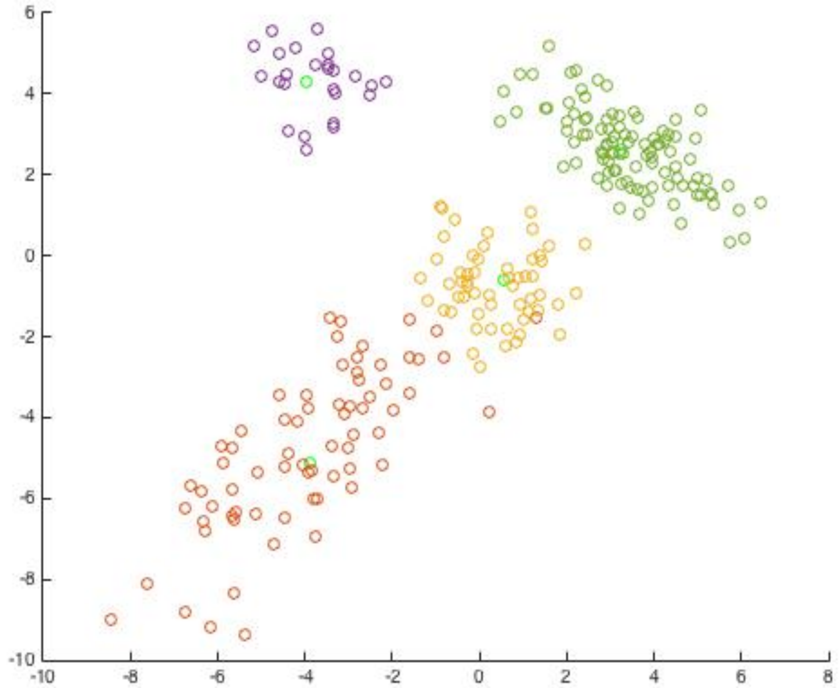


The clustering result of K =10, 25 varies. Generally, the algorithm can not cluster the data into 10 or 25 clusters. The number of clusters are mostly smaller than 10.

- **Problem 3.**

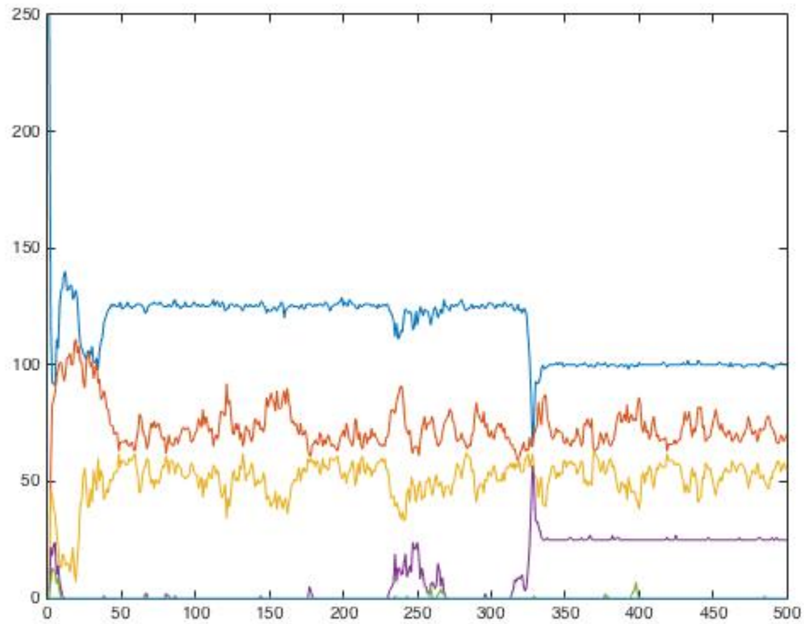
a) Implement the above-mentioned Gibbs sampling algorithm discussed in class and described in the notes. Run your algorithm on the data provided for 500 iterations.

I tested the program for many times and after 500 iterations, the number of clusters is 4.

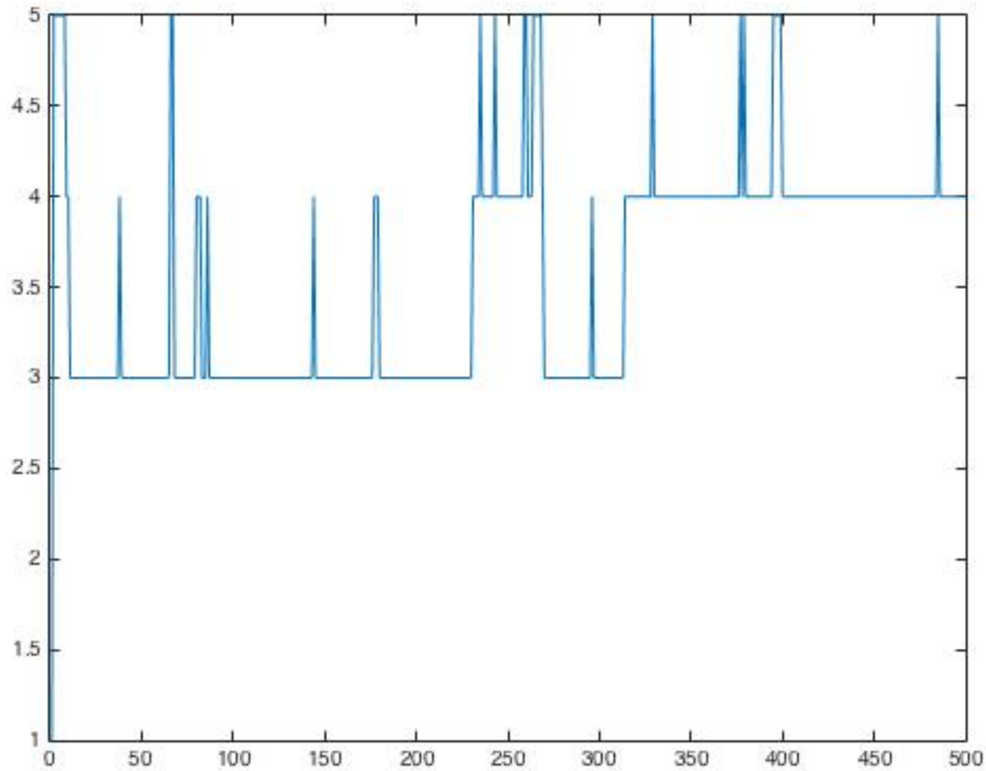


b) Plot the number of observations per cluster as a function of iteration for the six most probable clusters. These should be shown as lines that never cross; for example the ith value of the "second" line will be the number of observations in the second largest cluster after completing the ith iteration.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

The number of data points in each cluster always changes a little although the iteration time is set to 500.



c) Plot of the total number of clusters that contain data as a function of iteration.



There are still some peaks with value 5 when the result becomes stable. This may be caused by the sampling process of ci, μ, Λ in every iteration.