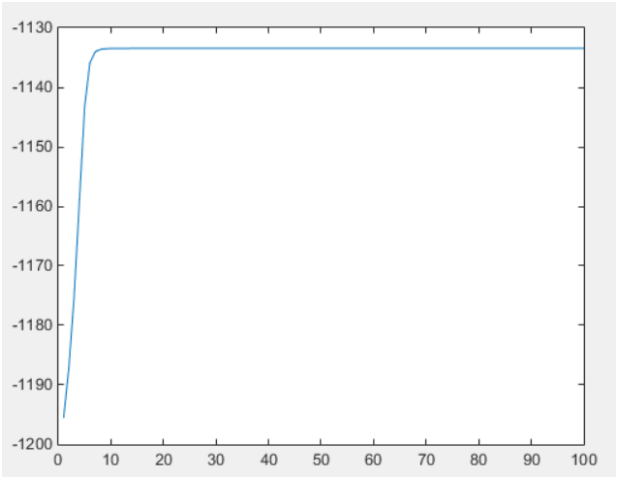
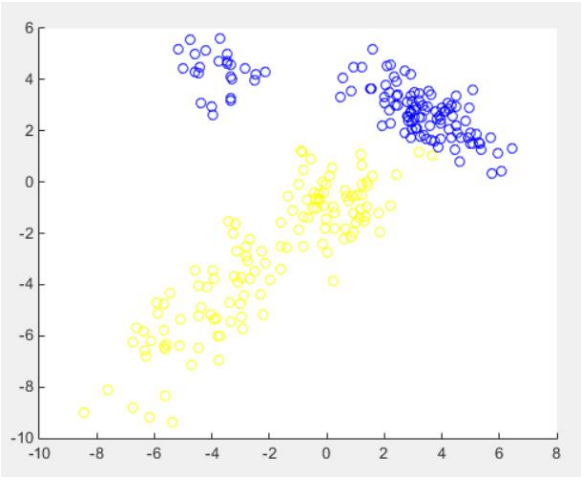
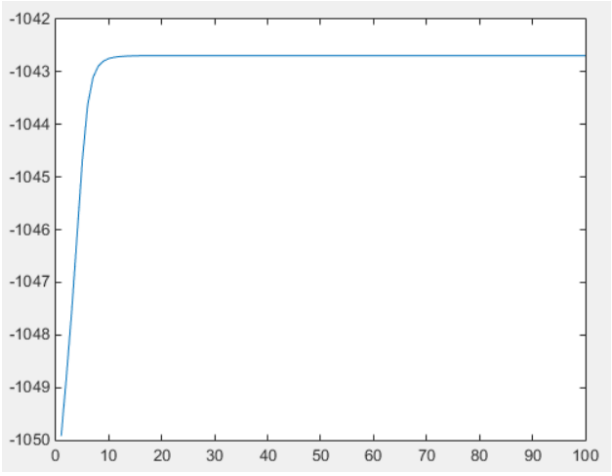
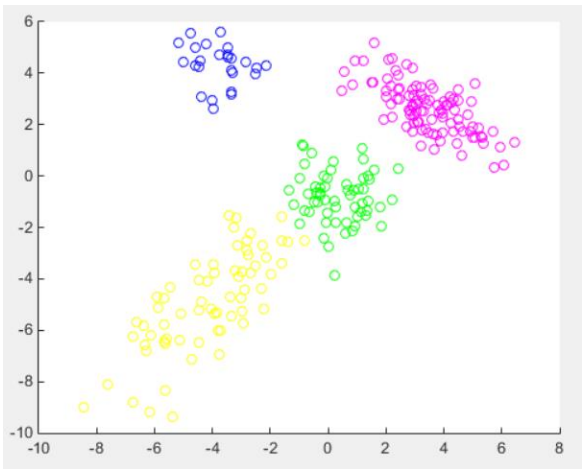
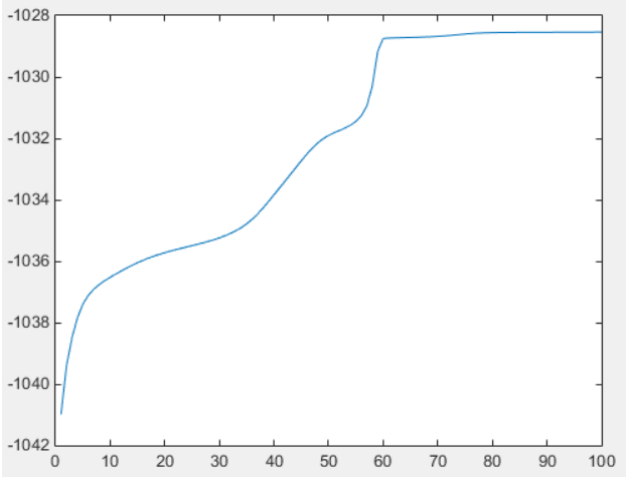
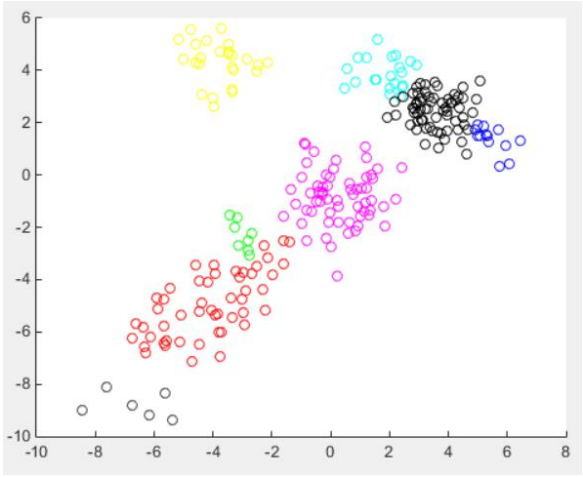
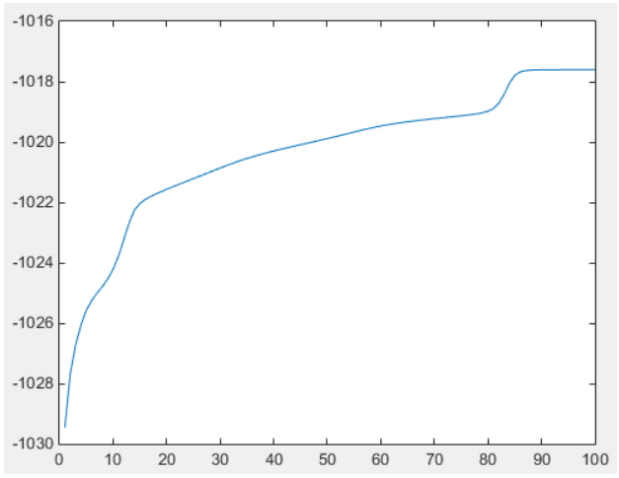
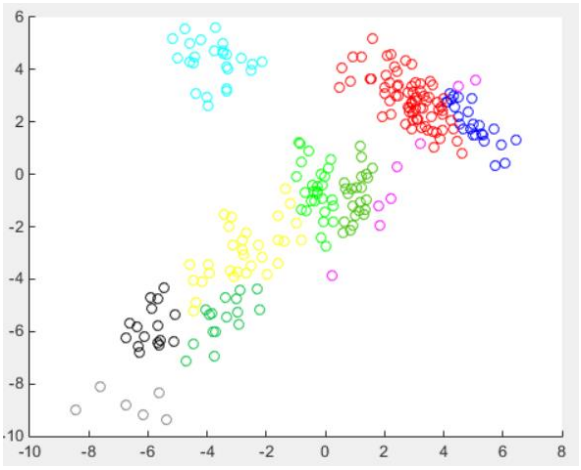
EECS 6892 BMML HW#3 UNI: sl3763 Name: Sung-Yen Liu

Problem 1

K=2, Log-Likelihood	Most probable cluster
	

K=4, Log-Likelihood	Most probable cluster
	

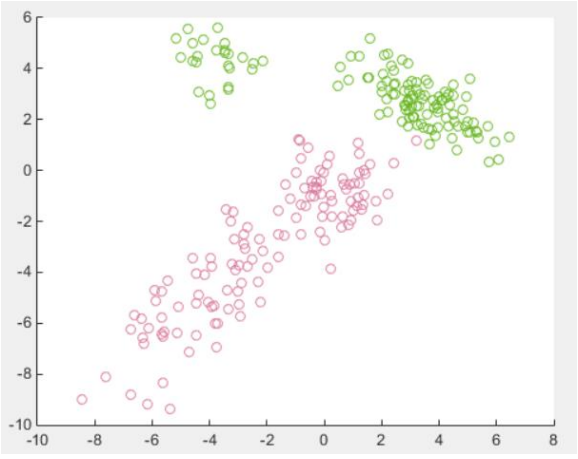
K=8, Log-Likelihood	Most probable cluster
	

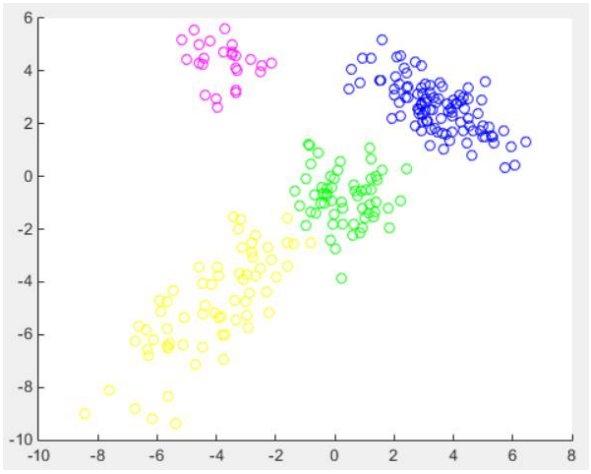
K=10, Log-Likelihood	Most probable cluster
	

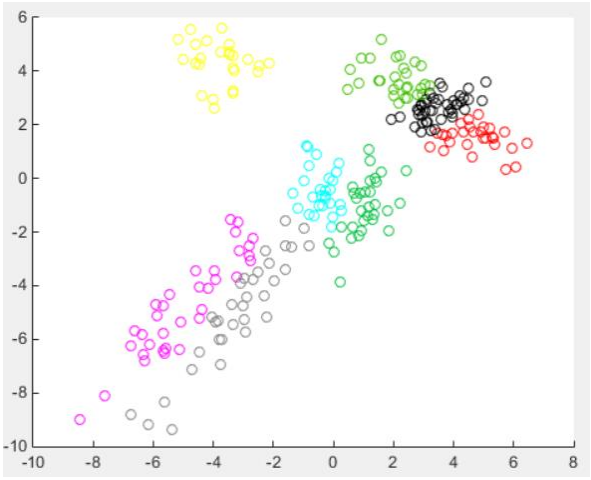
Observation & Discussion:
By using EM algorithm for GMM, as the number of clusters increases (K=8, 10), we can observe that the results may not be the same. GMM-EM is pretty sensitive to initialization, therefore I use K-means to initialize the centers. This leads to global optimum while K is small (K=2, 4), as the plots show convergence in a few iterations. Nevertheless, in the case of larger K, they reach convergence in more iterations; since the results are not always the same, we know that they only reach local maximum.

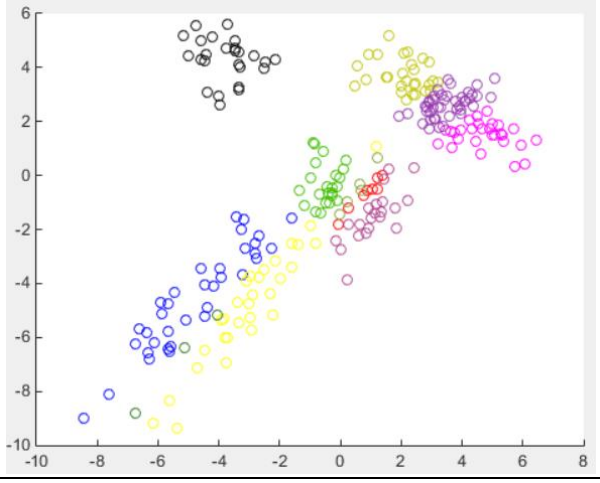
Since EM algorithm only take likelihood into consideration, it may not be a good choice when K is much larger than the number of original clusters. Without any prior distributions, EM is unable to capture the uncertainty (point-wise estimation only), and still classifies all the data into K clusters to optimize the objective function, which may lead to unstable result, as the plots show in this case (by visual classification, the best number of clusters should between 2~4).

Problem 2

K=2, Log-Likelihood	Most probable cluster
	

K=4, Log-Likelihood	Most probable cluster
	

K=10, Log-Likelihood	Most probable cluster, total 8 clusters
	

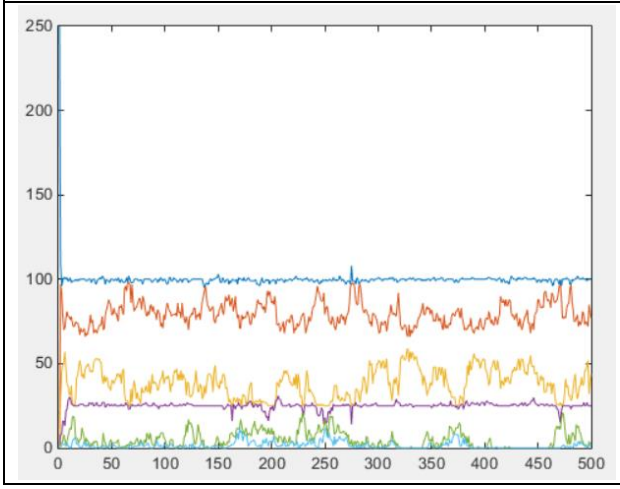
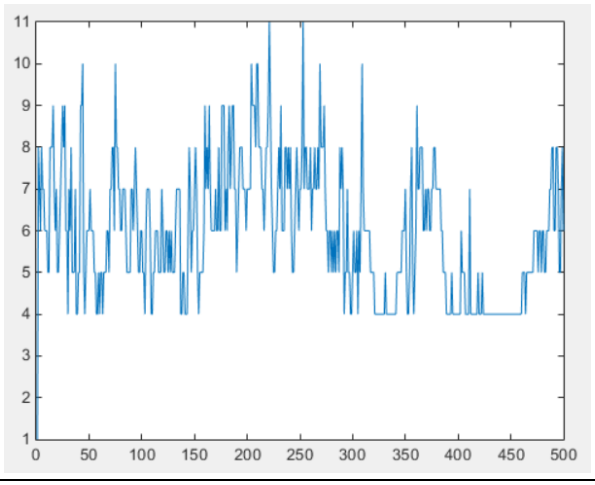
K=25, Log-Likelihood	Most probable cluster, total 11 clusters
	

(The formula of objective function are at the last page.)

Observation & Discussion:

By using Bayesian priors and VI algorithm, the results of “K is large” has been improved, which no longer classify all the data into assigned K clusters. Although there has no plot of objective function, by the cases of K=10, 25, we observe that VI algorithm find a balance between optimizing both the number of the most probable clusters and the objective function. This captures the uncertainty of data, and also avoid the hazard of overfitting while K is large (or the number of data is small). Without enough information of number of clusters, this would be a better way to cluster the data.

Problem 3

Number of observations per cluster	Total number of clusters
	

As the result shows, since the data can be obviously separated into two clusters, the numbers of the largest cluster (blue) and the fourth one (purple) are the relatively more stable. Furthermore, there are still many data which are on the bound of different clusters result to the oscillation both in the observation of other clusters and the number of total clusters.

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

$$= \mathbb{E}[\ln p(x|c, \mu, \Lambda)] + \mathbb{E}[\ln p(c|\pi)] + \mathbb{E}[\ln p(\pi)] + \mathbb{E}[\ln p(\mu)] + \mathbb{E}[\ln p(\Lambda)]$$

$$- \mathbb{E}[\ln q(c)] - \mathbb{E}[\ln q(\pi)] - \mathbb{E}[\ln q(\mu)] - \mathbb{E}[\ln q(\Lambda)]$$

① $$\mathbb{E}[\ln p(x|c, \mu, \Lambda)] = \mathbb{E}\left[\ln \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Lambda_k)^{c_{nk}}\right] = \sum_{k=1}^{K} \mathbb{1}(c_i=k)\, \mathbb{E}[\ln \mathcal{N}(x_i|\mu_k, \Lambda_k)]$$

$$\ln \mathcal{N}_d(x_i|\mu_k, \Lambda_k) = -\frac{1}{2}\ln|\Lambda_k| - \frac{d}{2}\ln(2\pi) - \frac{1}{2}(x_i-\mu_k)^T \Lambda_k^{-1}(x_i-\mu_k)$$

② $$\mathbb{E}[\ln p(c|\pi)] = \mathbb{E}\left[\ln \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{c_{nk}}\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[\ln \pi_k^{c_{nk}}] = \sum_{k=1}^{K} \mathbb{1}(c_i=k)\, \mathbb{E}[\ln \pi_k]$$

③ $$\mathbb{E}[\ln p(\pi)] = \mathbb{E}\left[\ln\left(C(\alpha)\prod_{k=1}^{K} \pi_k^{\alpha_0-1}\right)\right] = \mathbb{E}[\ln C(\alpha)] + \sum_{k=1}^{K}\mathbb{E}[\ln \pi_k^{\alpha_0-1}]$$

④ $$\mathbb{E}[\ln p(\mu)] = \sum_{j=1}^{K} \mathbb{E}[\ln p(\mu_j)] = \sum_{j=1}^{K}\left(-\frac{d}{2}\ln\frac{c}{2\pi} - \frac{1}{2c}\mathbb{E}[\mu_j^T\mu_j]\right)$$

⑤ $$\mathbb{E}[\ln p(\Lambda)] = \sum_{j=1}^{K} \mathbb{E}[\ln p(\Lambda_j)] = \sum_{j=1}^{K}\left(\frac{a-d-1}{2}\mathbb{E}[\ln|\Lambda_j|] - \frac{1}{2}\mathrm{trace}(B\,\mathbb{E}[\Lambda_j]) - \underbrace{\frac{ad}{2}\ln 2 - \frac{a}{2}\ln|B| - \ln \mathcal{I}_d\left(\frac{a}{2}\right)}_{\text{Const.}}\right)$$

<Entropy of each function, from PRML>

⑥ $$-\mathbb{E}[\ln q(c)] = -\sum_{i=1}^{N}\mathbb{E}[\ln q(c_i)] = \sum_{i=1}^{N} H[c_i] = -\sum_{i=1}^{N}\sum_{j=1}^{K}\mu_j \ln \mu_j$$

$$C(\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}$$

⑦ $$-\mathbb{E}[\ln q(\pi)] = H[\pi] = -\sum_{k=1}^{K}(\alpha_k-1)\{\psi(\alpha_k) - \psi(\textstyle\sum_i \alpha_i)\} - \ln C(\alpha)$$

⑧ $$-\mathbb{E}[\ln q(\mu)] = \sum_{j=1}^{K} H[\mu_j] = \sum_{j=1}^{K}\left(\frac{1}{2}\ln|\Sigma| + \frac{D}{2}(1+\ln(2\pi))\right)$$

⑨ $$-\mathbb{E}[\ln q(\Lambda)] = \sum_{j=1}^{K} H[\Lambda_j] = \sum_{j=1}^{K}\left(-\ln B(W, \nu) - \frac{(\nu-D-1)}{2}\mathbb{E}[\ln|\Lambda_j|] + \frac{\nu D}{2}\right)$$