CSCI 5521 - INTRO TO MACHINE LEARNING

TANVEER SINGH VIRDI

ASSIGNMENT 3

NOVEMBER 4, 2019

PROBLEM 1:

INTRO TO MACHINE LEARNING

ASSIGNMENT 3

PROBLEM 1

SOLUTION: EM algorithm for estimating a mixture of multinomial distributions. Given the probability mass function of a multinomial distribution for class C_i:

$$P(x=(x_1,\dots,x_n)|C_i) = P(x=(x_1,\dots,x_n)|p_{i1},\dots,p_{in})$$
$$= \frac{m!}{x_1!\dots x_n!} p_{i1}^{x_1}\dots p_{in}^{x_n}$$

where $\sum_{j=1}^{n} p_{ij} = 1$ and $\sum_{j=1}^{n} x_j = m$ for all K distributions

The prior probability of the hidden variable is given as -

$$P(z^t) = \prod_{i=1}^{K} \pi_i^{z_i^t} \quad \text{where } z_i^t = 1 \text{ if } x^t \text{ belongs to cluster } i$$

Likelihood of an observation x^t is equal to its probability specified by the component that generated it. This is given by the conditional probability:

$$P(x^t|z^t) = \prod_{i=1}^{K} p_i(x^t)^{z_i^t} = \prod_{i=1}^{K} P(x^t|C_i)^{z_i^t}$$

The joint density $P(x^t,z^t) = P(z^t)\,P(x^t|z^t)$

Complete data log likelihood is:

$$\mathcal{L}_c(\Phi|\mathcal{X},\mathcal{Z}) = \log \prod_{t=1}^{n} P(x^t,z^t) = \log \prod_{t=1}^{n}\prod_{i=1}^{k} \pi_i^{z_i^t} P(x^t|C_i)^{z_i^t}$$
$$= \sum_{t=1}^{n}\sum_{i=1}^{k} \left(\log\left(\pi_i^{z_i^t} P(x^t|C_i)^{z_i^t}\right)\right)$$
$$= \sum_{t=1}^{n}\sum_{i=1}^{k} z_i^t \left(\log \pi_i + \log P(x^t|C_i)\right)$$
$$= \sum_{t=1}^{n}\sum_{i=1}^{k} z_i^t \left(\log \pi_i + \log\left(\frac{m!}{x_1!\dots x_n!} p_{i1}^{x_1}\dots p_{in}^{x_n}\right.\right.$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{K} z_i^t \left(\log \pi_i + \log \left(\frac{m!}{x_1! \cdots x_n!} \right) + x_1 \log p_{i1} + \dots + x_n \log p_{in} \right)$$

$$= \sum_{t=1}^{n} \sum_{i=1}^{K} z_i^t \left(\log \pi_i + \log \left(\frac{m!}{x_1! \cdots x_n!} \right) + \sum_{j=1}^{n} x_j^t \log p_{ij} \right)$$

Thus the complete log likelihood function is given as:

$$\mathcal{L}_c(\phi | \mathcal{X}, Z) = \sum_{t=1}^{n} \sum_{i=1}^{K} z_i^t \left(\log \pi_i + \log \left(\frac{m!}{x_1! \cdots x_n!} \right) + \sum_{j=1}^{n} x_j^t \log p_{ij} \right)$$

In the E step:

Expectation for responsibility $\gamma(z_i^t) = E[z_i^t | x^t] = P(z_i^t = 1 | x^t)$

$$= P(x^t | z_i^t = 1)\, P(z_i^t = 1) / P(x^t)$$

$$= p_i(x^t)\, \pi_i / \Sigma_j\, p_j(x^t)\, \pi_j$$

$$= P(x^t | C_i)\, \pi_i / \Sigma_j\, P(x^t | C_j)\, \pi_j$$

$$= \frac{\frac{m!}{x_1! \cdots x_n!} \left(p_{i1}^{x_1} \cdots p_{in}^{x_n} \right) \pi_i}{\sum_{j=1}^{K} \frac{m!}{x_1! \cdots x_n!} \left(p_{j1}^{x_1} \cdots p_{jn}^{x_n} \right) \pi_j}$$

$$E[z_i^t | \mathcal{X}, \phi] = \gamma(z_i^t) = \frac{\left(p_{i1}^{x_1} \cdots p_{in}^{x_n} \right) \pi_i}{\sum_{j=1}^{K} \left(p_{j1}^{x_1} \cdots p_{jn}^{x_n} \right) \pi_j}$$

$$Q(\phi | \phi^l) = E[\mathcal{L}_c(\phi | \mathcal{X}, Z)] = \sum_{t=1}^{n} \sum_{i=1}^{K} E[z_i^t | \mathcal{X}, \phi] \left(\log \pi_i + \log \left(\frac{m!}{x_1! \cdots x_n!} \right) + \sum_{j=1}^{n} x_j^t \log p_{ij} \right)$$

$$E[\mathcal{L}_c(\phi | \mathcal{X}, Z)] = \sum_{t=1}^{n} \sum_{i=1}^{K} \gamma(z_i^t) \left(\log \pi_i + \log \left(\frac{m!}{x_1! \cdots x_n!} \right) + \sum_{j=1}^{n} x_j^t \log p_{ij} \right)$$

In the M step:

We know that $\sum_{j=1}^{n} p_{ij} = 1$.

Using the lagrange constraint $\sum_{j=1}^{n} p_{ij} = 1$ and taking partial derivative with respect to p_{ij}, we get:

$$\frac{\partial \left[E[\mathcal{L}_c(\phi \mid x, z)] - \lambda \left[\sum_{j=1}^{n} p_{ij} - 1 \right] \right]}{\partial p_{ij}}$$

$$= \frac{\partial}{\partial p_{ij}} \left(\sum_{t=1}^{n} \sum_{i=1}^{K} \gamma(z_i^t) \left(\log \pi_i + \log \left(\frac{m!}{x_1! \dots x_n!} \right) + \sum_{j=1}^{n} x_j^t \log p_{ij} \right) - \lambda \left(\sum_{j=1}^{n} p_{ij} - 1 \right) \right)$$

$$= \sum_{t=1}^{n} \gamma(z_i^t) \frac{x_j^t}{p_{ij}} - \lambda$$

Setting the derivative to zero and solving for λ we get -

$$\sum_{t=1}^{n} \gamma(z_i^t) x_j^t - \lambda p_{ij} = 0 \Rightarrow \sum_{t=1}^{n} \gamma(z_i^t) \sum_{j=1}^{n} x_j^t - \lambda \sum_{j=1}^{n} p_{ij} = 0$$

We know that $\sum_{j=1}^{n} x_j = m$ and $\sum_{j=1}^{n} p_{ij} = 1$

$$\lambda = m \sum_{t=1}^{n} \gamma(z_i^t)$$

Using this value of λ we get:

$$\boxed{p_{ij} = \frac{\sum_{t=1}^{n} \gamma(z_i^t) x_j^t}{m \sum_{t=1}^{n} \gamma(z_i^t)}}$$

For π_i, we know that $\sum_{i=1}^{K} \pi_i = 1$. Using the lagrange constraint $\sum_{i=1}^{K} \pi_i = 1$ and partial derivative taken with respect to π_i, we get :

$$\frac{\partial}{\partial \pi_i}\left(E[L_c(\Phi|x,z)] - \lambda\left[\sum_{i=1}^{K} \pi_i - 1\right]\right)$$

$$= \frac{\partial}{\partial \pi_i}\left(\sum_{t=1}^{n}\sum_{i=1}^{K} \gamma(z_i^t)\left(\log \pi_i + \log\left(\frac{m!}{x_1!\dots x_n!}\right) + \sum_{j=1}^{n} x_j^t \log p_{ij}\right) - \lambda\left(\sum_{i=1}^{K} \pi_i - 1\right)\right)$$

$$= \frac{\sum_{t=1}^{n} \gamma(z_i^t)}{\pi_i} - \lambda$$

Setting the derivative to zero and solving for λ we get -

$$\sum_{t=1}^{n} \gamma(z_i^t) - \lambda \pi_i = 0 \Rightarrow \sum_{t=1}^{n}\sum_{i=1}^{K} \gamma(z_i^t) - \lambda \sum_{i=1}^{K} \pi_i = 0$$

We know that $\sum_{i=1}^{K} \pi_i = 1$.

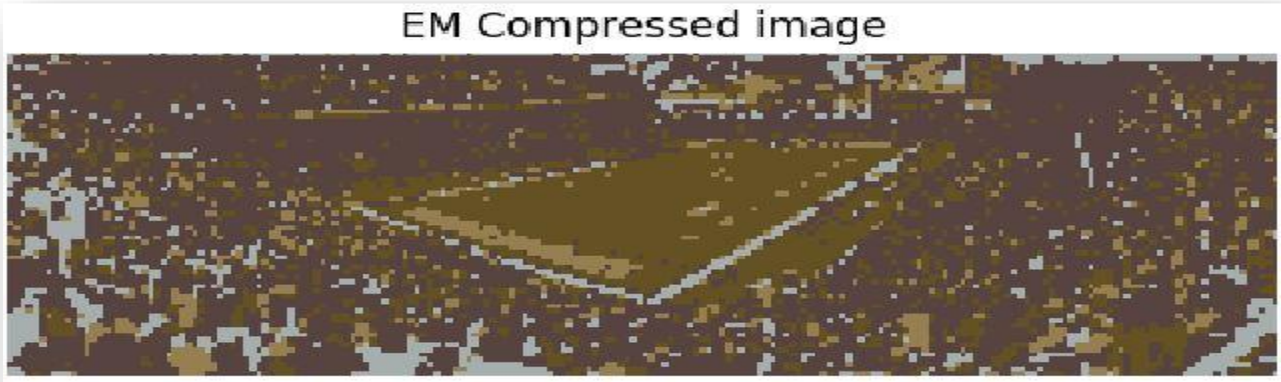
$$\lambda = N$$

Using this value of λ we get :

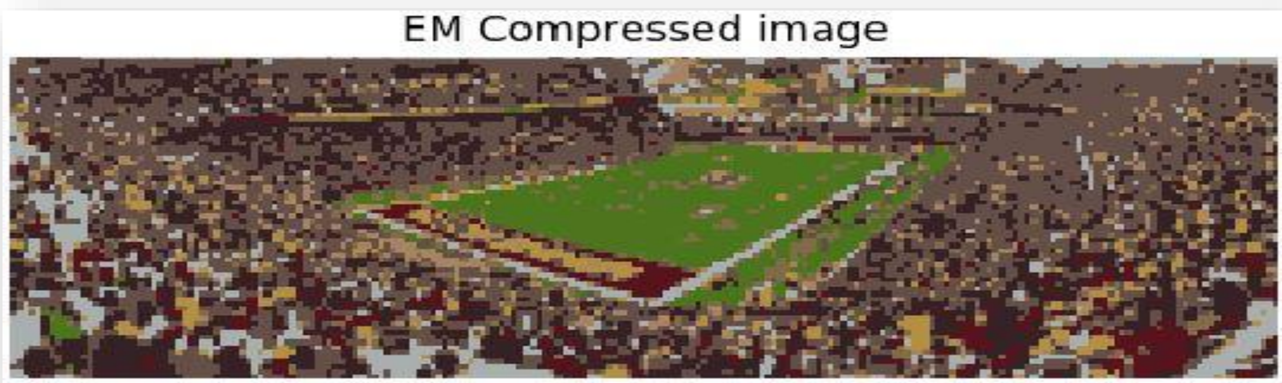
$$\boxed{\pi_i = \frac{\sum_{t=1}^{n} \gamma(z_i^t)}{N}}$$

PROBLEM 2 :

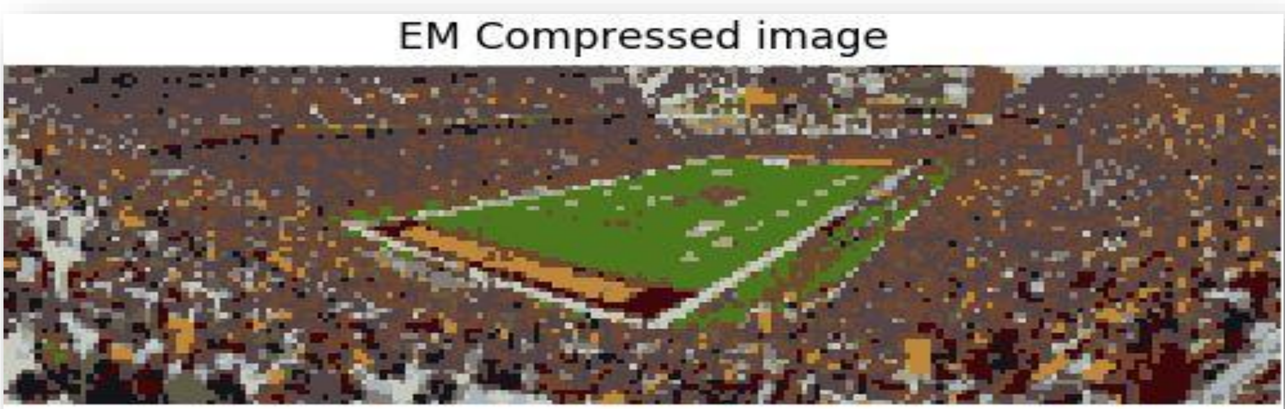
a. K = 4



K = 8

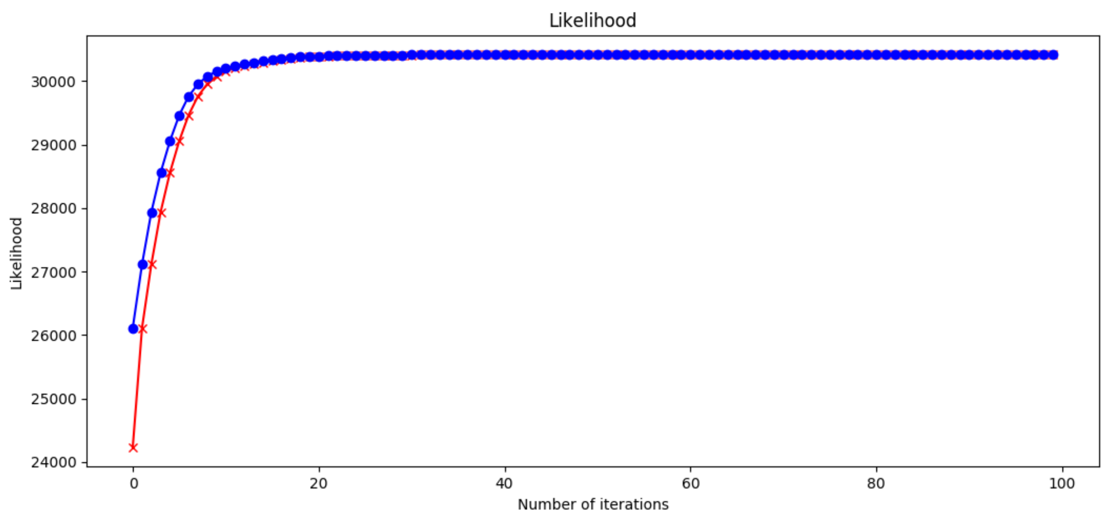


K = 12

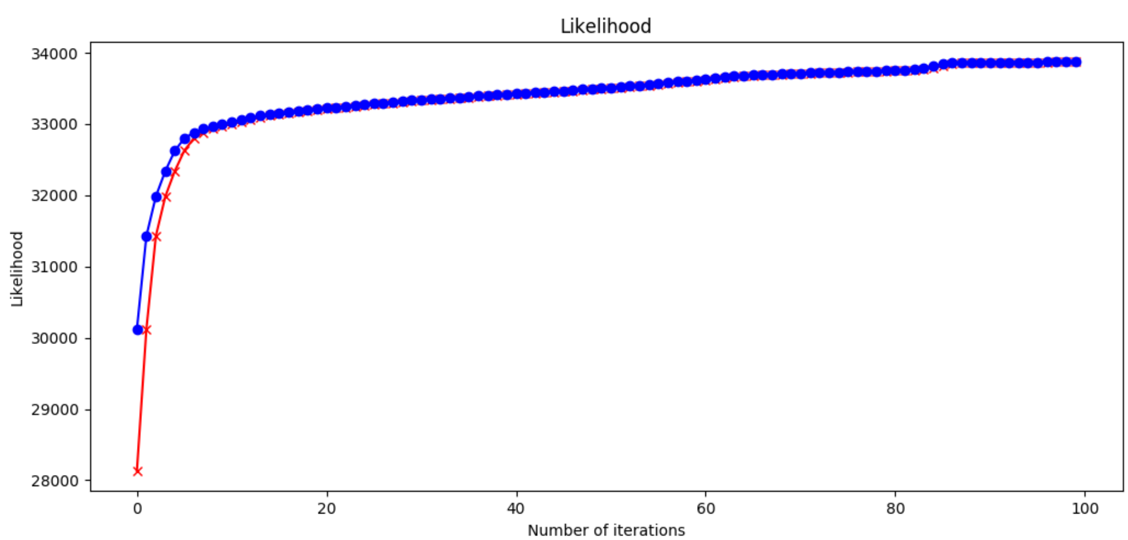


b. Plots below show the complete log likelihood after each E step and M step vs number of iterations. The red colored crosses denote the log likelihood after every E step and the blue dots denote the log likelihood after every M step. After the M-step, the means, covariances and the prior probabilities are recalculated, hence the log likelihood changes. We observe that the values are different initially but they converge as the number of iterations increase.

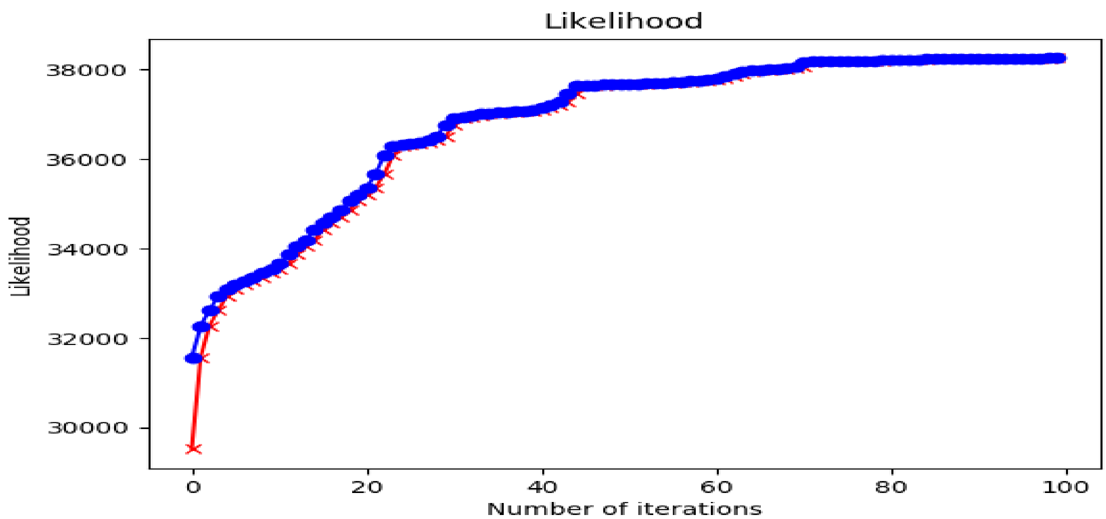
K = 4



K = 8

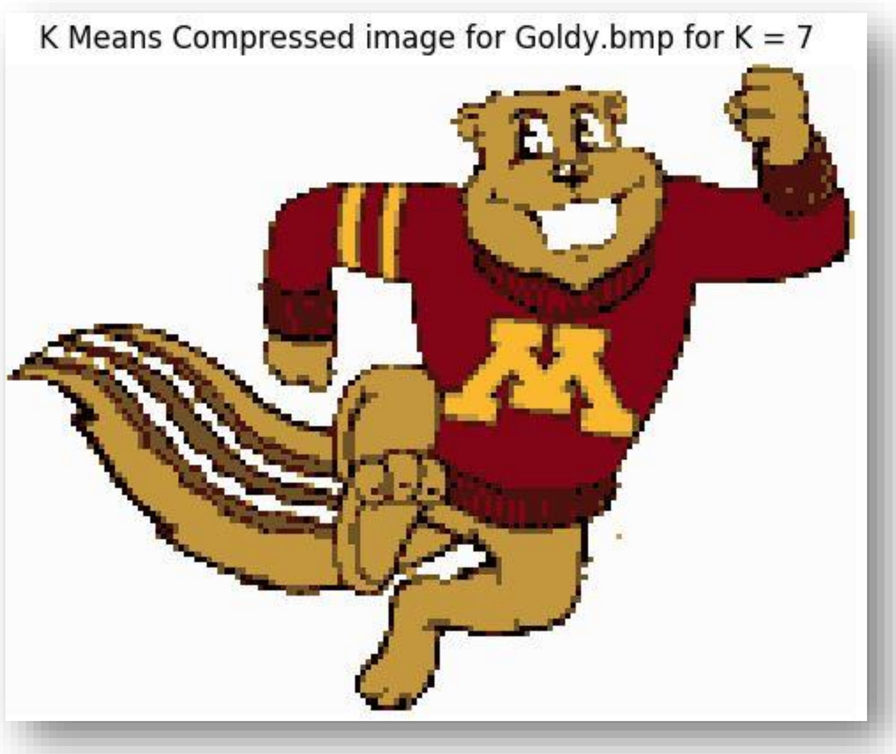


K = 12



c. The EM algorithm implementation does not run for Goldy.bmp. This is because the covariance matrix is singular for this image and is not positive definite. The covariance matrix is singular when the row and column values are interrelated. In this image, several pixels have the same color or are similar to one another. Thus the covariance matrix becomes singular for this image. In K means we are not concerned with covariance and thus we don't see any such error.

K Means compressed image for Goldy.bmp for K = 7 :



d.

PROBLEM 2(d)

SOLUTION: Improved version of EM algorithm to handle singular covariance matrix. Complete data log likelihood for a mixture of gaussians is given as –

$$\mathcal{L}_c(\phi|\mathcal{X},\mathcal{Z}) = \log \prod_{t=1}^{n} p(x^t, z^t) = \log \prod_{t=1}^{n} \prod_{i=1}^{K} P(G_i)\, P(x^t|G_i)$$

$$\mathcal{L}_c(\pi_i, \mu_i, \Sigma_i|\mathcal{X},\mathcal{Z}) = \sum_{t=1}^{n} \log \sum_{i=1}^{K} P(G_i)P(x^t|G_i)$$

$$= \sum_{t=1}^{n} \log \sum_{i=1}^{K} \pi_i \mathcal{N}(x^t|\mu_i,\Sigma_i)$$

where $\mathcal{N}(x^t|\mu_i,\Sigma_i) = \dfrac{1}{(2\pi)^{D/2}\,|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$

Adding the regularization term in the question, we get –

$$\mathcal{L}_c(\pi_i, \mu_i, \Sigma_i|\mathcal{X},\mathcal{Z}) = \sum_{t=1}^{n} \log \sum_{i=1}^{K} \pi_i \mathcal{N}(x^t|\mu_i,\Sigma_i) - \frac{\lambda}{2}\sum_{i=1}^{K}\sum_{j=1}^{d} (\Sigma_i^{-1})_{jj}$$

where $(\Sigma_i^{-1})_{jj}$ is the (j,j)th entry of matrix Σ_i^{-1} and $\lambda > 0$

Taking the partial derivative with respect to Σ^{-1} & setting it to zero:

$$\frac{\partial \mathcal{L}}{\partial \Sigma^{-1}} = \sum_{t=1}^{n} \frac{\mathcal{N}(x^t|\mu_i,\Sigma_i)\pi_i}{\sum_{i=1}^{K}\pi_i \mathcal{N}(x^t|\mu_i,\Sigma_i)} \left(\Sigma^{-1} - (x^t-\mu_i)(x^t-\mu_i)^T\right) - \frac{\lambda I}{2} = 0$$

$$\Rightarrow \sum_{t=1}^{n} \gamma(z_i^t)\left(\Sigma^{-1} - (x^t-\mu_i)(x^t-\mu_i)^T\right) - \frac{\lambda I}{2} = 0$$

Solving this equation we get –

$$\Sigma_i = \frac{\sum_{t=1}^{n} \gamma(z_i^t)(x^t-\mu_i)(x^t-\mu_i)^T}{\sum_{t=1}^{n}\gamma(z_i^t)} + \lambda I$$

e. The improved version of EM algorithm ran successfully on Goldy.bmp using lambda λ= 0.000001. Following is the image obtained after compression using this improved version of EM algorithm :

We observe that the image compressed using the improved version of EM algorithm does not have sharp distinctions between the boundary of two colors as observed in K means. This is because EM uses probability of a pixel belonging to a particular cluster and so two similar pixels might belong to two different clusters when the EM algorithm is used. . Plot below show the complete log likelihood after each E step and M step vs number of iterations. The red colored crosses denote the log likelihood after every E step and the blue dots denote the log likelihood after every M step.