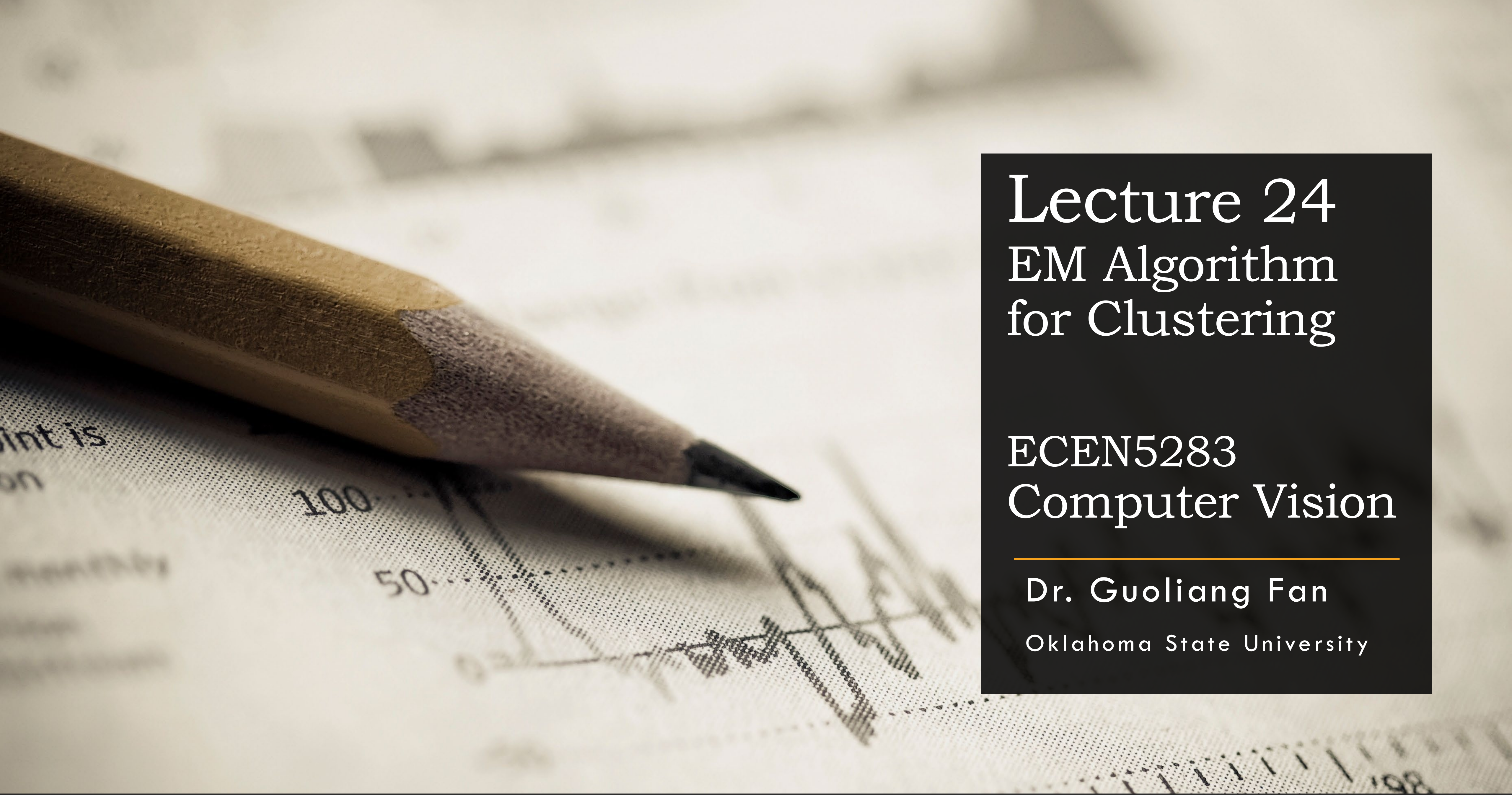


Lecture 24
EM Algorithm for Clustering

ECEN5283 Computer Vision

Dr. Guoliang Fan

Oklahoma State University

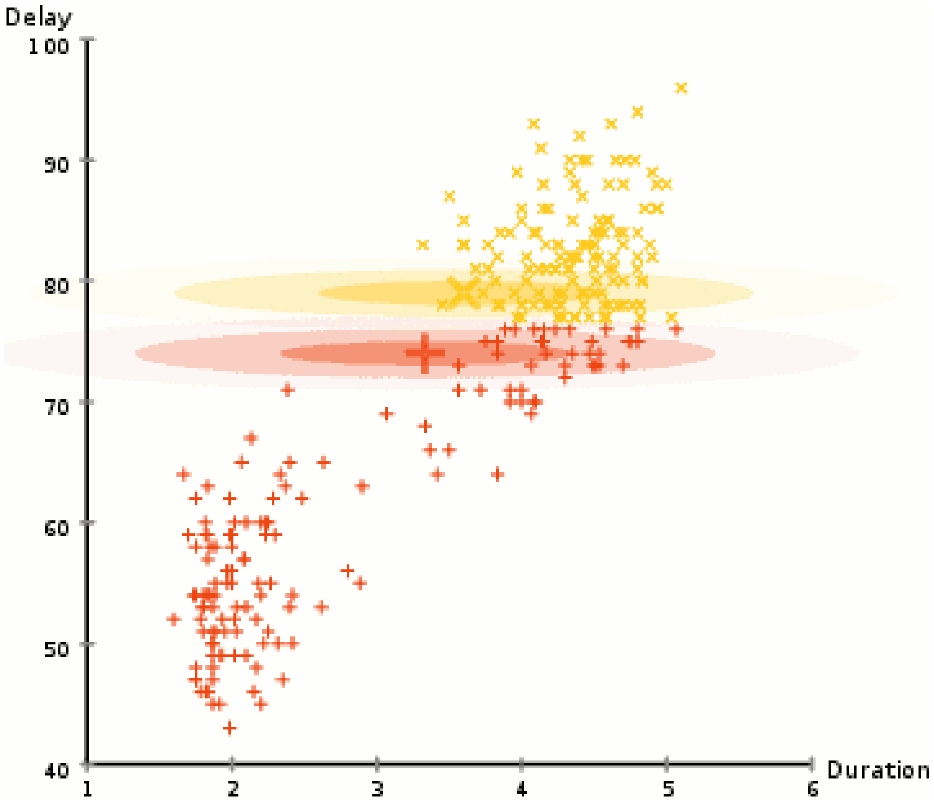
Goals

To revisit the missing data problem and the EM algorithm

To illustrate the EM clustering result via a toy demo

To compare EM with K-means for unsupervised clustering

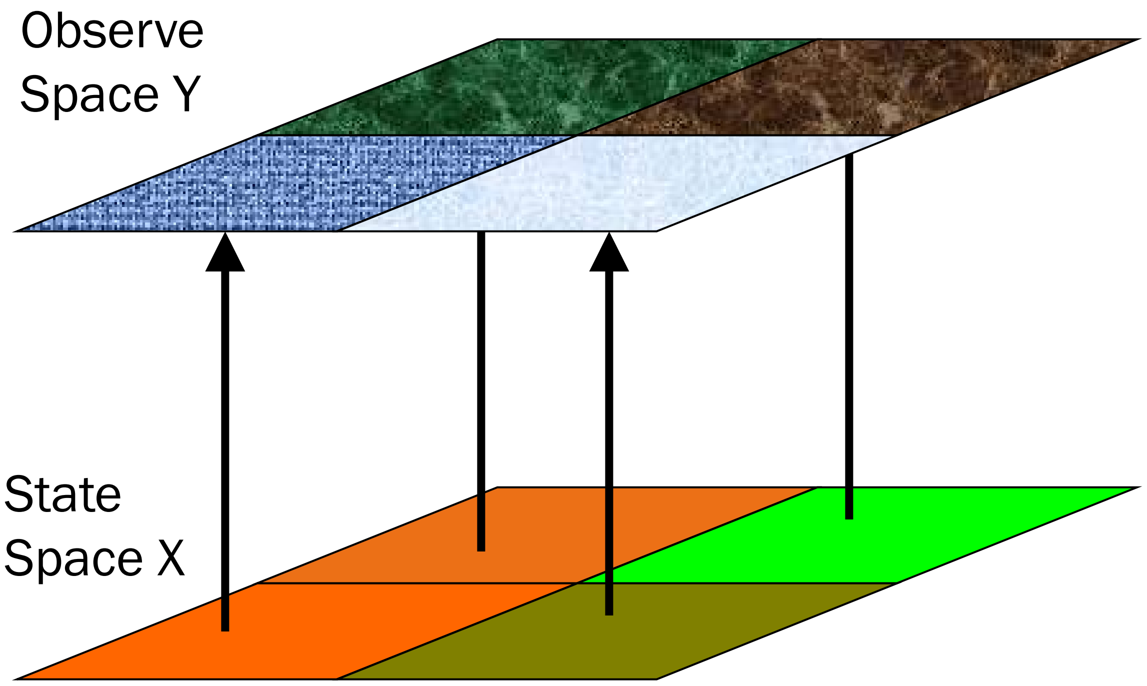
Old Faithful Eruption Data Modeling



EM clustering of Old Faithful eruption data. The random initial model (which, due to the different scales of the axes, appears to be two very flat and wide ellipses) is fit to the observed data. In the first iterations, the model changes substantially, but then converges to the two modes of the geyser.

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Issue (1) Parameter Estimation



$\boldsymbol{y} = (y_1, y_2, \dots, y_N)$
Observation data

Parameter Space
$\Theta = \{\alpha_i, \mu_i, \Sigma_i | i = 1,2,3,4\}$

$p_Y(\boldsymbol{y}|\Theta)$ data likelihood
(the model−data fitness)

$\Theta^* = \arg\max_{\Theta} p_Y(\boldsymbol{y}|\Theta)$
(parameter estimation)

To find the parameter Θ that can best explain the current observation $\boldsymbol{y}$.

Issue (2) Data Classification

To classify data, we need to compute the probability of data sample y belonging to class x, i.e., the posterior probability $p(x|y)$, which is computed during parameter estimation.

$\alpha_i = p_X(x = i)$
(prior probability)

$p_{Y|X}(y|x = i) = \mathrm{N}(y|\mu_i, \Sigma_i)$
(likelihood function)

$$\text{(Bayes' law)}\quad p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{i=1}^{k} p_{Y|X}(y|x=i)p_X(x=i)} \quad \text{(posterior probability)}$$

$$x^* = \arg_{x\in X}\max p_{X|Y}(x|y) \quad (\text{maximum } a\ posteriori \text{ or MAP})$$

Log-likelihood Function

For a data sample y_j $(i = 1, \ldots, N)$, the likelihood function is defined

$$p_Y(y_j|\Theta) = \sum_{i=1}^{k} p_Y(x_j = i, y_j|\Theta) = \sum_{i=1}^{k} \underbrace{p_Y\left(y_j \middle| x_j = i, \Theta\right)}_{N(y_i|\mu_i, \Sigma_i)} \underbrace{p_X(x_j = i|\Theta)}_{\alpha_i}$$

Under independent assumption, the likelihood of the whole data set $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ is written as $p_Y(\boldsymbol{y}|\Theta) = \prod_{i=1}^{N} p_Y(y_i|\Theta)$, then we define the log-likelihood function as

$$\log p_Y(\boldsymbol{y}|\Theta) = \sum_{j=1}^{N} \log\left(\sum_{i=1}^{k} p(y_j|x_j = i, \Theta)\alpha_i\right)$$

EM Algorithm: E-step

Initialization: set $s = 0$ and

$$\Theta^0 = (\alpha_1^{(0)}, \alpha_2^{(0)}, \ldots, \alpha_g^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_g^{(0)}).$$

Expectation (E-step):

Posterior probability

$$\mathbf{I}(l, m) = p_{X|Y}(x_l = m | y_l, \Theta^{(s)}) = \frac{\alpha_m^{(s)} p_Y\left(y_l \middle| \theta_m^{(s)}\right)}{\sum_{i=1}^k \alpha_i^{(s)} p_{Y|X}\left(y_l \middle| \theta_i^{(s)}\right)}$$

$$p_{X|Y}(x = i | y) = \frac{p_{Y|X}(y|x) p_X(x)}{\sum_{i=1}^k p_{Y|X}(y | x = i) p_{X(x=i)}}$$

$$p_{Y|X}(y_l | \theta_m^{(s)}) = \frac{\exp\left\{-\frac{1}{2}(y_l - \mu_m)^T \Sigma_m^{-1} (y_l - \mu_m)\right\}}{(2\pi)^{d/2} \det(\Sigma_l)^{1/2}}$$

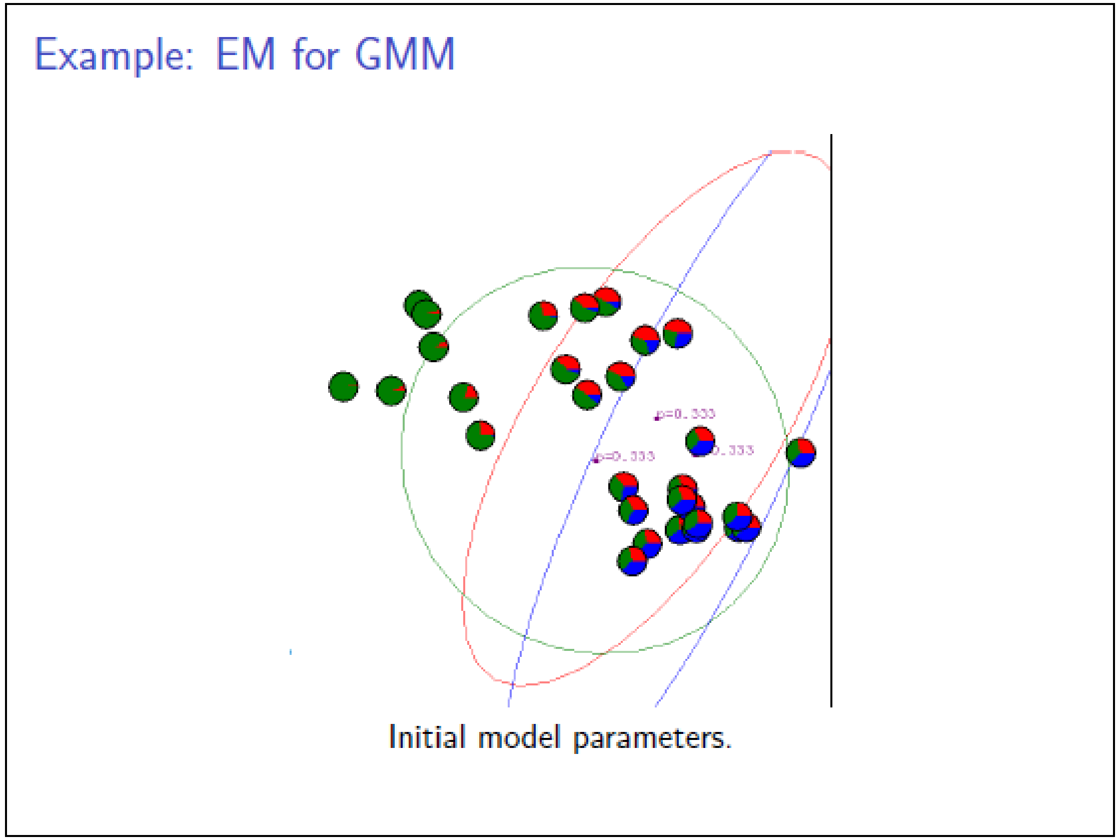
EM Algorithm: M-step

Maximization (M-step): $\Theta^* = \arg\max_{\Theta} \log p(\boldsymbol{y}|\Theta)$

$$\alpha_m^{(s+1)} = p_X(x) = \frac{1}{N}\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})$$

$$\theta_m^{(s+1)} \begin{cases} \mu_m^{(s+1)} = \dfrac{\sum_{l=1}^{N} y_l \cdot p_{X|Y}(x_l = m|y_l, \Theta^{(s)})}{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})} \\ \Sigma_m^{(s+1)} = \dfrac{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)}) \left\{(y_l - \mu_m^{(s)})(y_l - \mu_m^{(s)})^T\right\}}{\sum_{l=1}^{N} p_{X|Y}(x_l = m|y_l, \Theta^{(s)})} \end{cases}$$

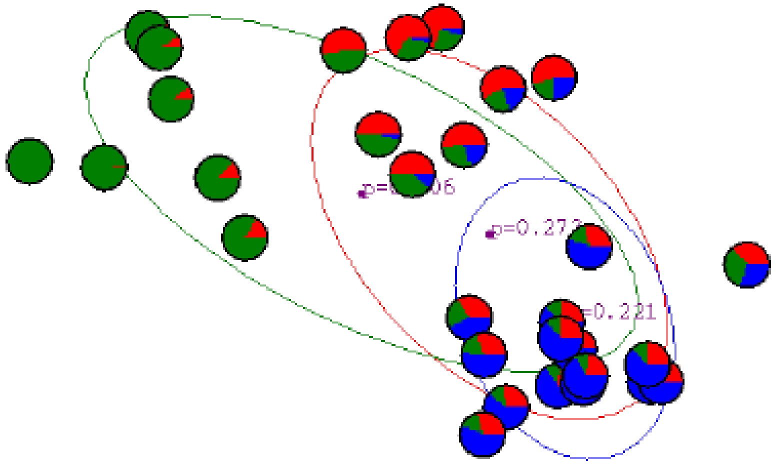
Demo: Initialization



https://www2.cs.duke.edu/courses/fall07/cps271/EM.pdf

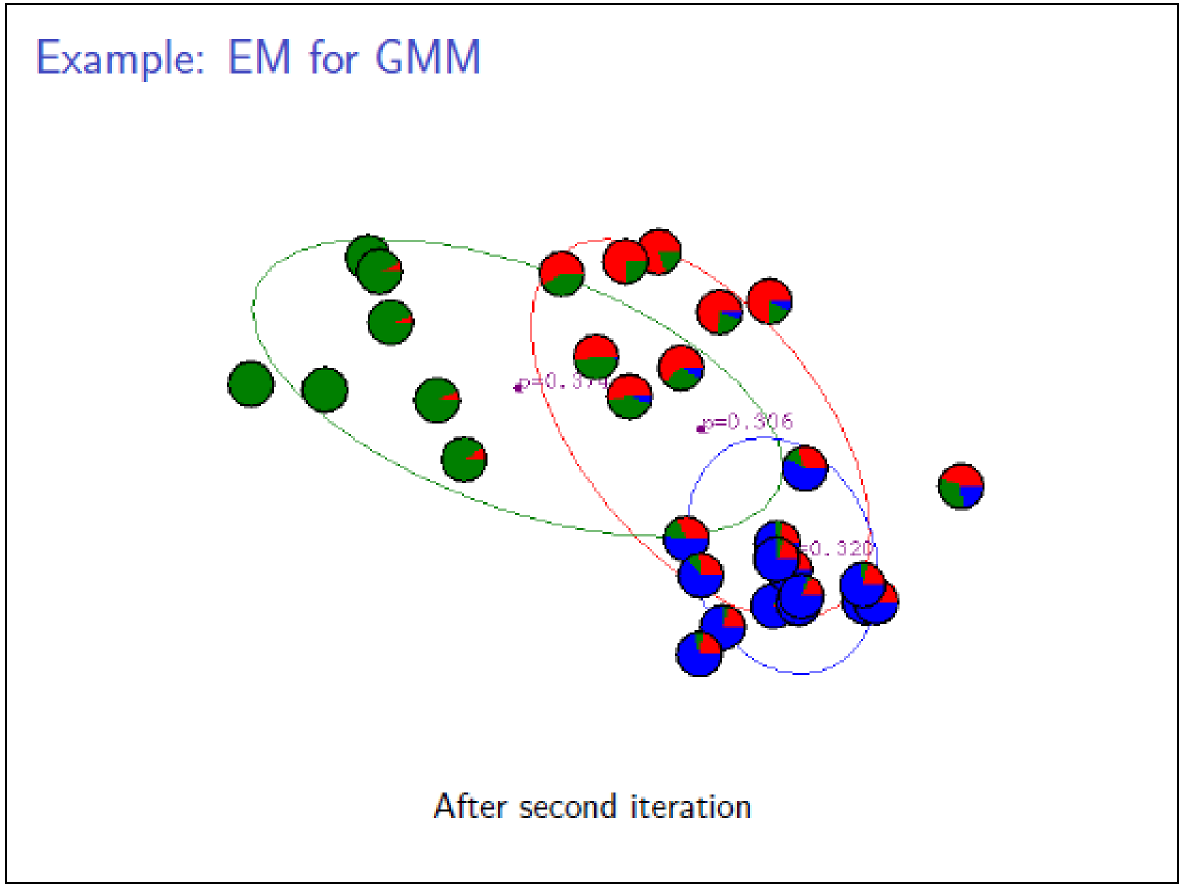
Demo: 1st iteration

Example: EM for GMM



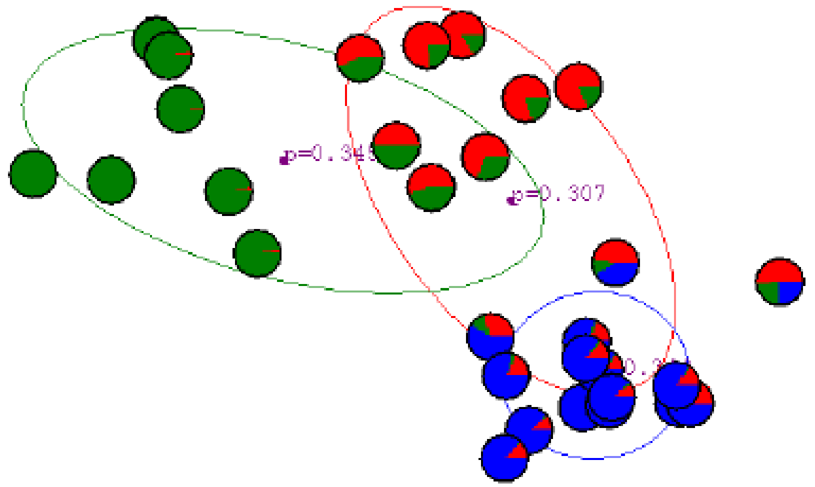
After first iteration

Demo: 2^{nd} iteration



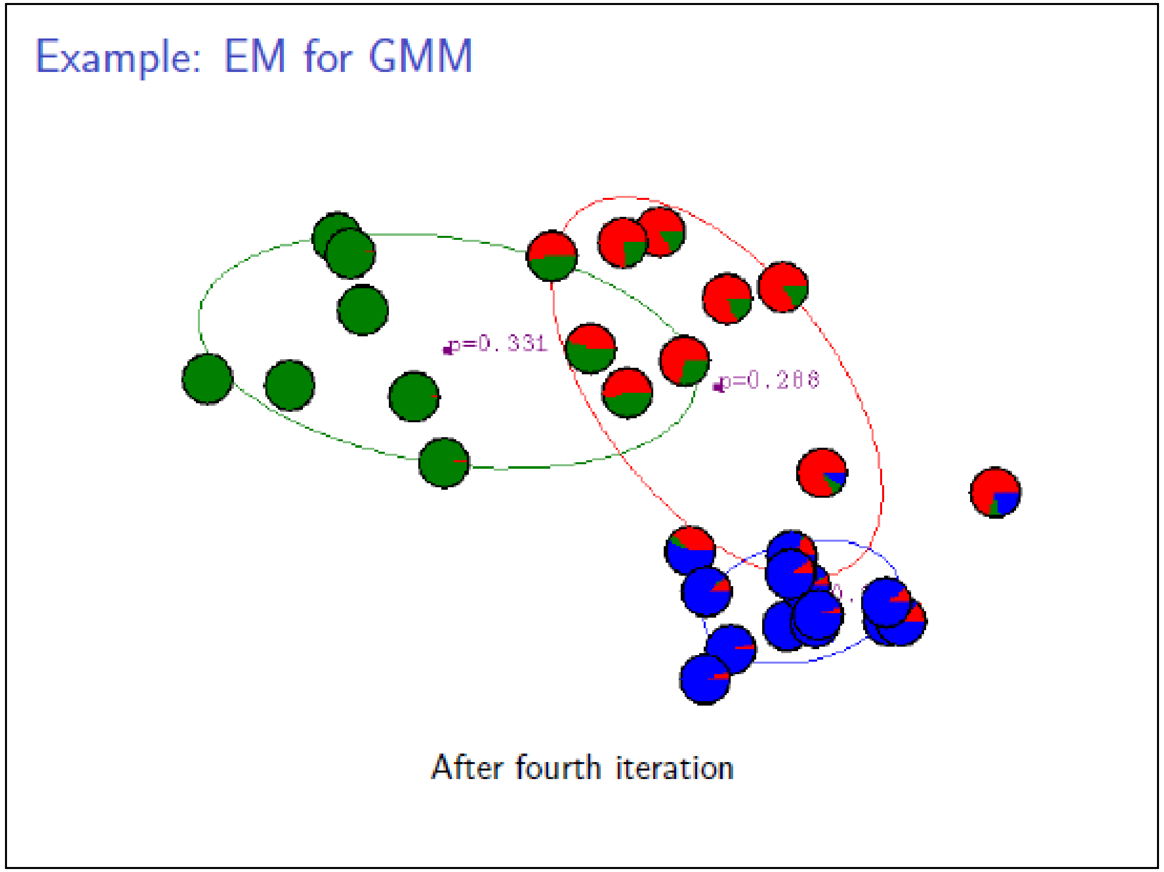
Demo: 3rd iteration

Example: EM for GMM

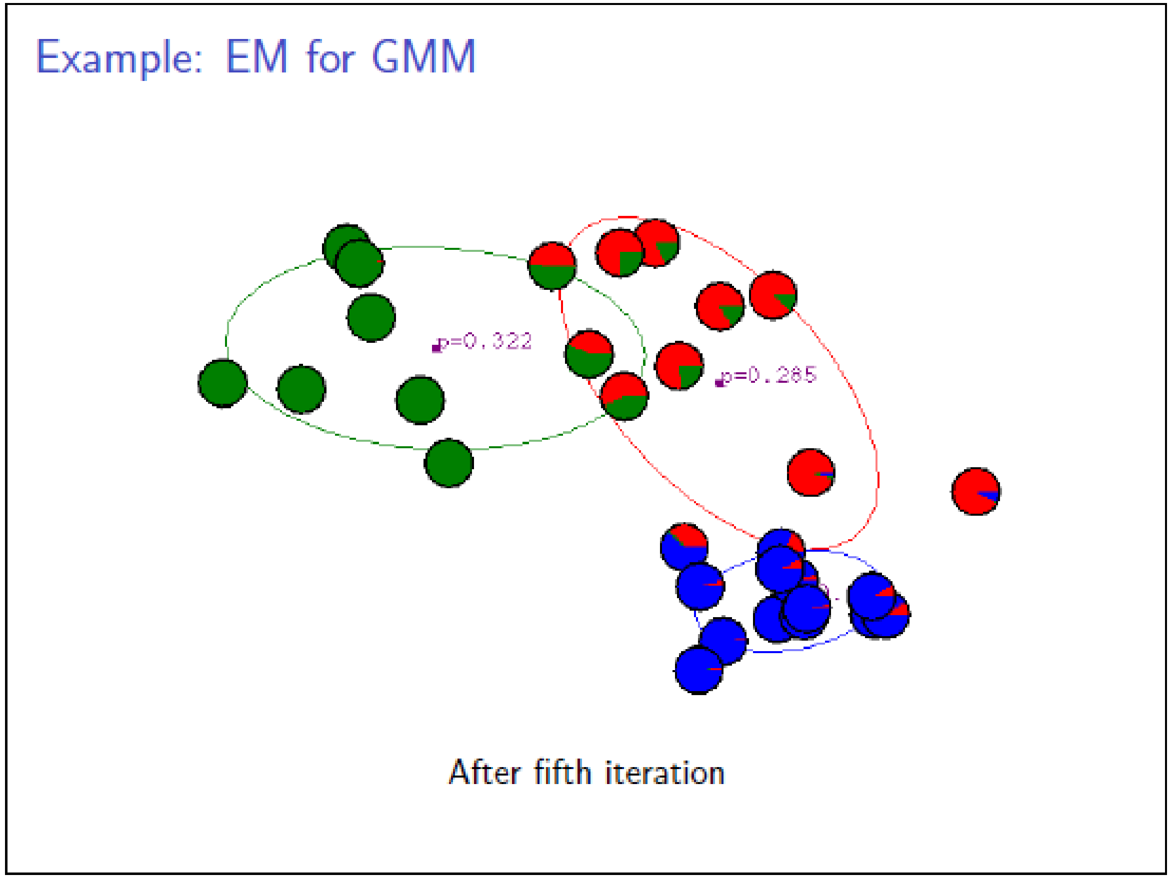


After third iteration

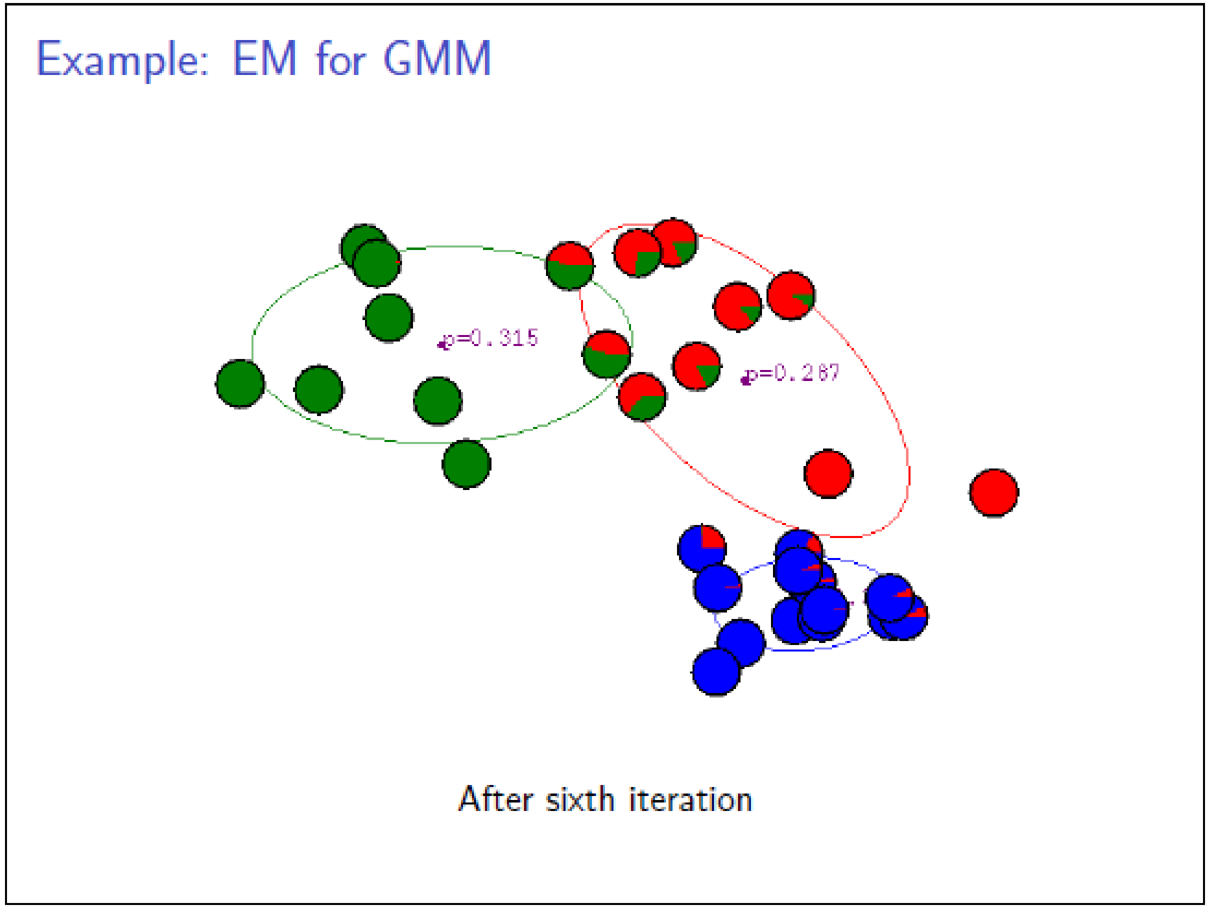
Demo: 4th iteration



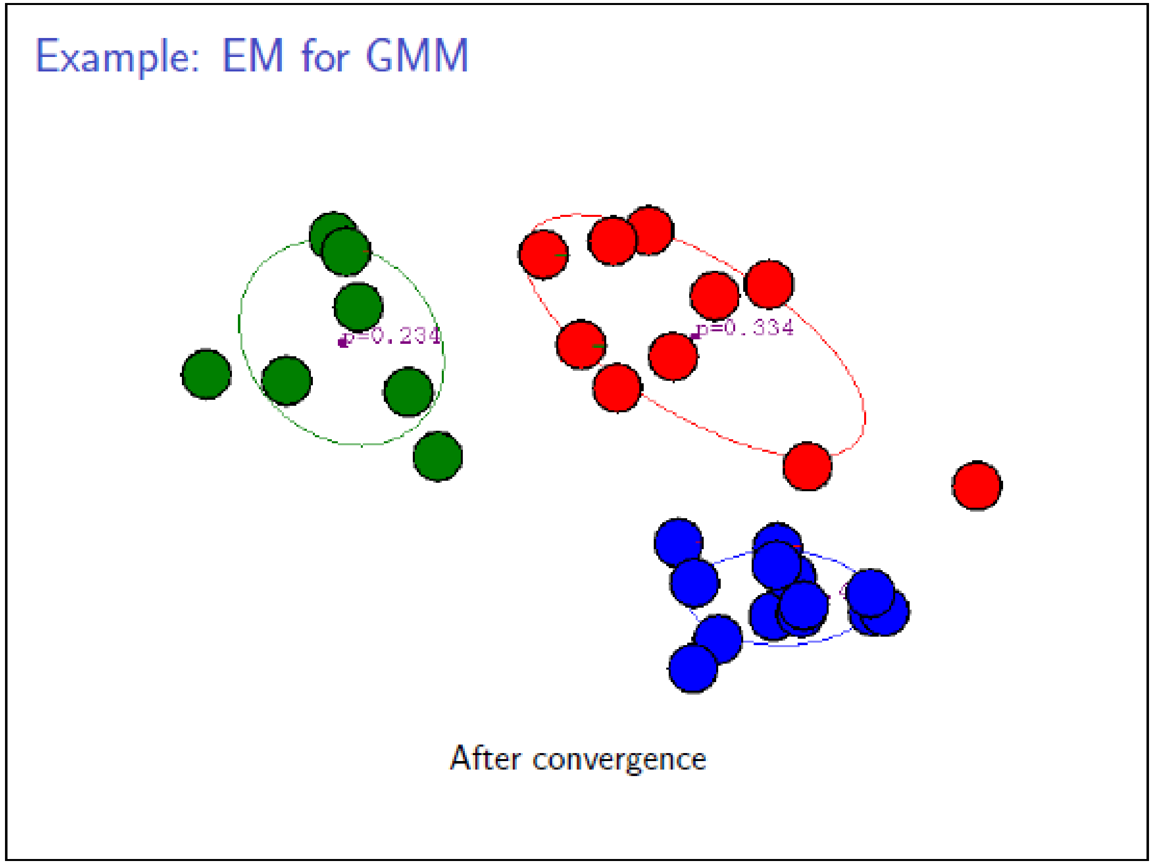
Demo: 5th iteration



Demo: 6^{th} iteration



Demo: Convergence



EM Algorithm: Stop Condition

Iteration still the stop criteria is satisfied, e.g., no much change of the data log-likelihood

$$\log p_Y\left(\boldsymbol{y}|\Theta^{(s)}\right) - \log p_Y\left(\boldsymbol{y}|\Theta^{(s-1)}\right) < \Delta'$$

(or an easy way to fix the iteration number)

Decide the class label for data point

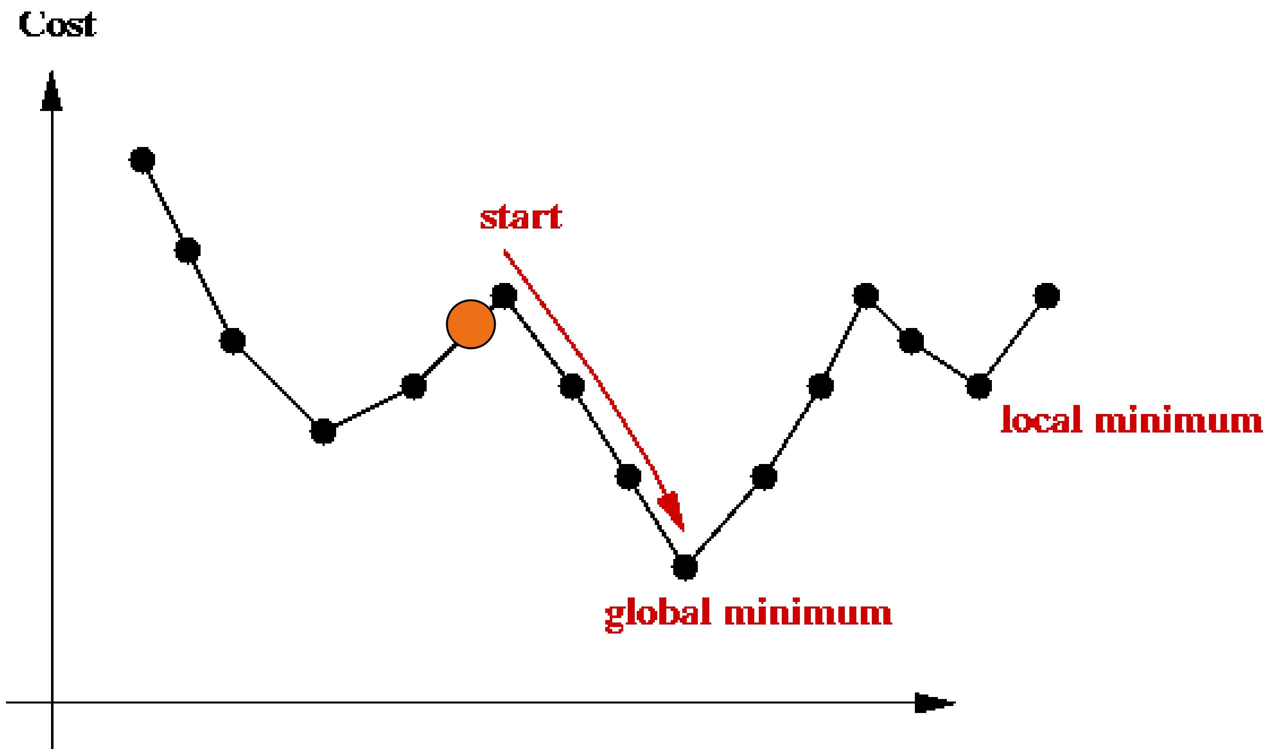
$$\mathbf{I}(l,m) = p_{X|Y}\left(x_l = m|y_l, \Theta^{(s)}\right)$$

$$x_l^* = \arg_{m\in\{1,\ldots,k\}} \max \mathbf{I}\,(l,m)$$

EM Algorithm: Local Optimality

Both K-means and EM can only converge to the local optimum of the objective function.

In other words, both methods are sensitive to the initialization, especially EM.



EM Algorithm: K-means Initialization

Random initialize k centers for the k-means algorithm.

$$C^0 = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$$

Run k-means multiple times and use the results with the minimum divergence.

$$\Phi(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{k} \left\{ \sum_{x_j=i} \left| y_j - \mathbf{c}_i \right|^2 \right\}$$

According to the class labels $\{x_1, x_2, \ldots, x_N\}$ of all data samples $\{y_1, y_2, \ldots, y_N\}$, initialize the multivariate Gaussian models for the EM.

$$\alpha_i^{(0)} = \frac{\#(x_j = i | j = 1, \ldots, N)}{N} \qquad \Sigma_i^{(0)} = \frac{\sum_{x_j=i} (y_j - \mu_i)(y_j - \mu_i)^T}{\#(x_j = i | j = 1, \ldots, N) - 1}$$

$$\mu_i^{(0)} = \frac{\sum_{x_j=i} y_j}{\#(x_j = i | j = 1, \ldots, N)}$$

counting the number

K-means vs. EM

	K-means	EM
Initialization	Initialize k means (cluster centers) $$C^0 = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$$	Initialize k Gaussian models that have equal weights. $$\Theta^0 = \{\alpha_i, \mu_i, \Sigma_i \mid i = 1, \ldots, k\}$$
Step 1.	Assume the cluster centers are known, and classify each data sample to the closest cluster center.	Given the model parameters, estimate the missing data in terms of the posterior probability of each data sample.
Step 2.	Assume the allocation (the class label of each sample) is known, and choose a new set of cluster centers C^{s+1}. Each center is the mean of all points in that cluster.	From the estimated missing data, to obtain the maximum likelihood estimate of the model parameters Θ^{s+1} .