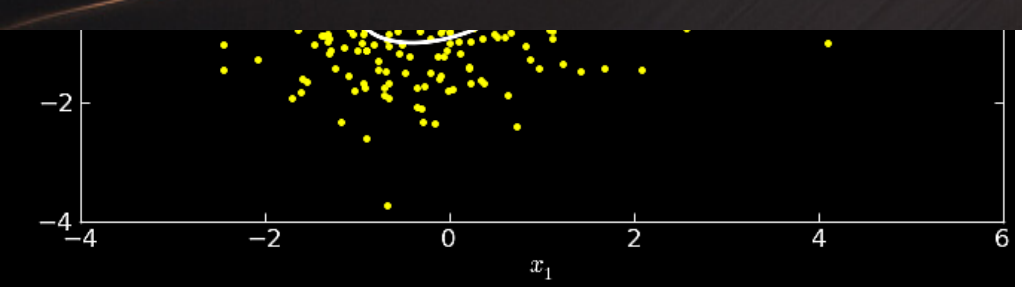
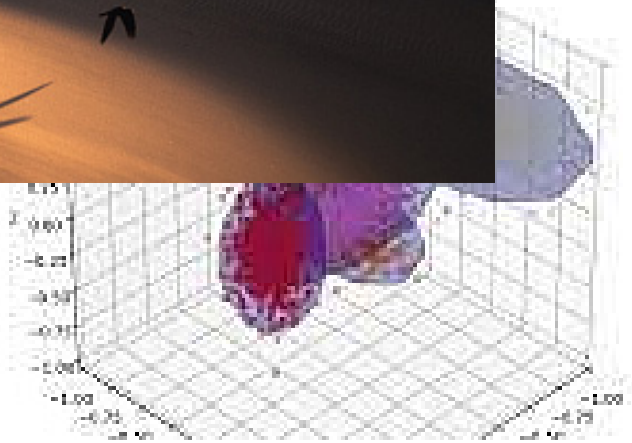
Gaussian Mixture Model (GMM) using Expectation Maximization (EM) Technique

Book : C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

aussian Mixture Model
ctual Data
=10.58, σ=0.35, w=0.53
=9.65, σ=0.36, w=0.27
=11.82, σ=0.31, w=0.08
=8.61, σ=0.52, w=0.11

14





The Gaussian Distribution

- Univariate Gaussian Distribution

$$G(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

mean variance

- Multi-Variate Gaussian Distribution

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

mean covariance

We need to estimate these parameters (Σ, μ) of a distribution:

One method – Maximum Likelihood (ML) Estimation.

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi).$$

Which is MAP, and which one MLE ??

$$= \arg\max_{\theta} \frac{f(x \mid \theta)\, g(\theta)}{\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta}$$

$$= \arg\max_{\theta} f(x \mid \theta)\, g(\theta).$$

ML Method for estimating parameters

- Consider log of Gaussian Distribution

$$\ln p(x \mid \mu, \Sigma) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

- Take the derivative and equate it to zero

$$\frac{\partial \ln p(x \mid \mu, \Sigma)}{\partial \mu} = 0 \qquad \frac{\partial \ln p(x \mid \mu, \Sigma)}{\partial \Sigma} = 0$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T$$

Where, N is the number of samples or data points

Gaussian Mixtures

- Linear super-position of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Number of Gaussians

Mixing coefficient: weightage for each Gaussian dist.

- Normalization and positivity require:

$$\mathbf{0 \le \pi_k \le 1}, \qquad \sum_{\mathbf{k=1}}^{\mathbf{K}} \boldsymbol{\pi}_{\mathbf{k}} = \mathbf{1}$$

- Consider log-likelihood:

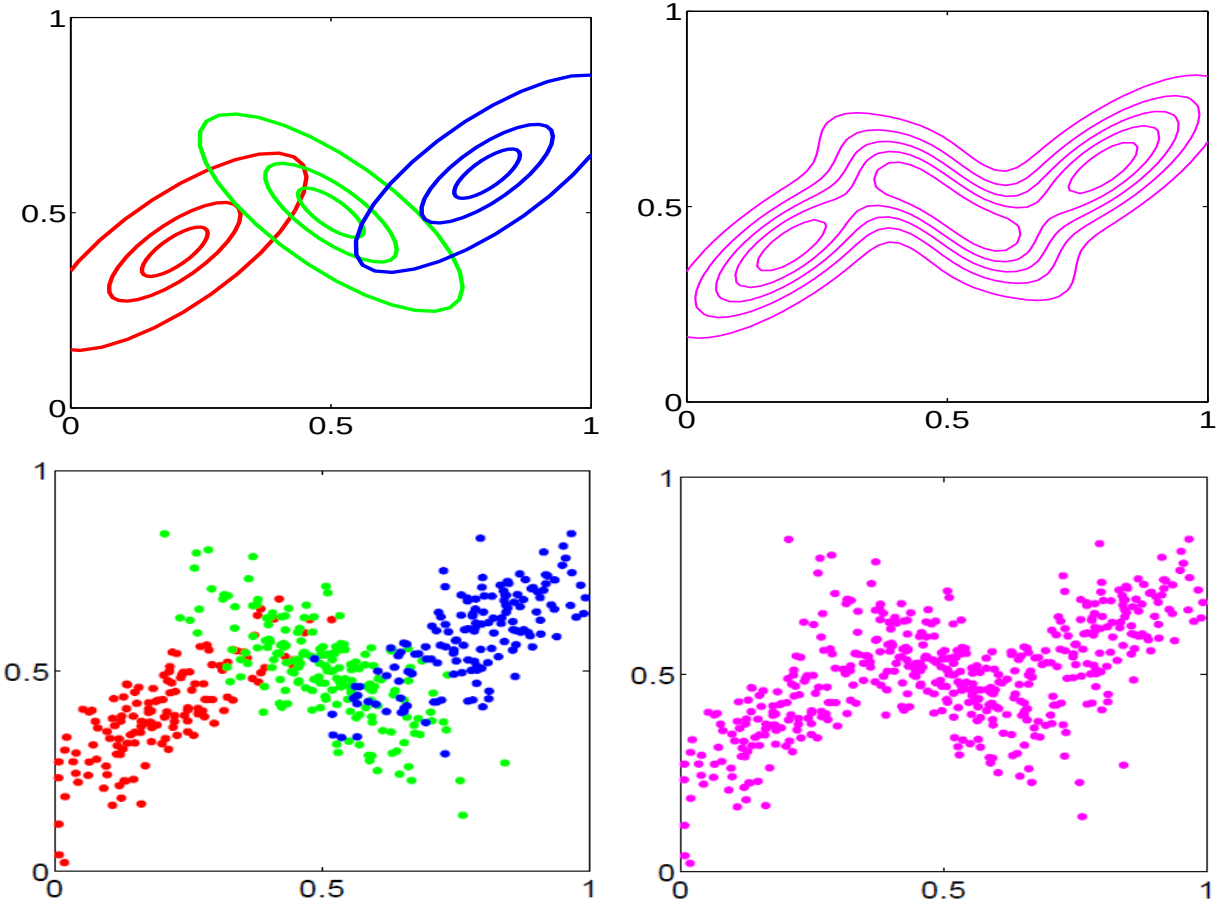
$$\ln p(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln p(x_n) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) \right\}$$

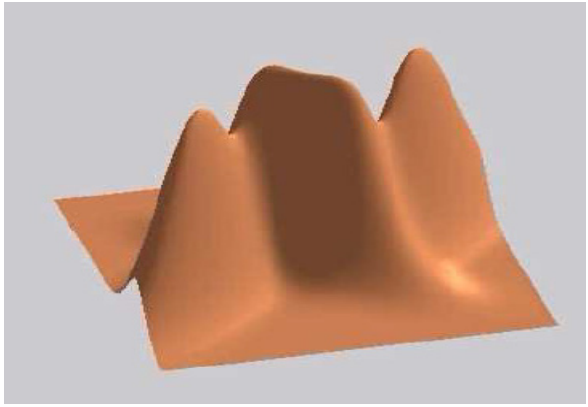
ML does not work here as there is no closed form solution

Parameters can be calculated using -

Expectation Maximization (EM) technique

Example: Mixture of 3 Gaussians





Latent variable: posterior prob.

- We can think of the mixing coefficients as prior probabilities for the components

- For a given value of ‘x’, we can evaluate the corresponding posterior probabilities, called responsibilities

- From Bayes rule

$$\gamma_k(\mathbf{x}) = p(k \mid \mathbf{x}) = \frac{p(k)p(\mathbf{x} \mid k)}{p(\mathbf{x})}$$

Latent Variable

$$= \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma_j)} \qquad \text{where,} \quad \pi_k = \frac{N_k}{N}$$

Interpret N_k as the effective no. of points assigned to cluster k.

Expectation Maximization

- EM algorithm is an iterative optimization technique which is operated locally
  - Estimation step: for given parameter values we can compute the expected values of the latent variable.
  - Maximization step: updates the parameters of our model based on the latent variable calculated using ML method.

EM Algorithm for GMM

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters comprising the means and covariances of the components and the mixing coefficients.

1. Initialize the means μ_j, covariances Σ_j and mixing coefficients π_j, and evaluate the initial value of the log likelihood.

2. E step. Evaluate the responsibilities using the current parameter values

$$\gamma_k(x) = \frac{\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x \mid \mu_j, \Sigma_j)}$$

EM Algorithm for GMM

3. M step. Re-estimate the parameters using the current responsibilities

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n) x_n}{\sum_{n=1}^{N} \gamma_j(x_n)}$$

$$\Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(x_n)}$$

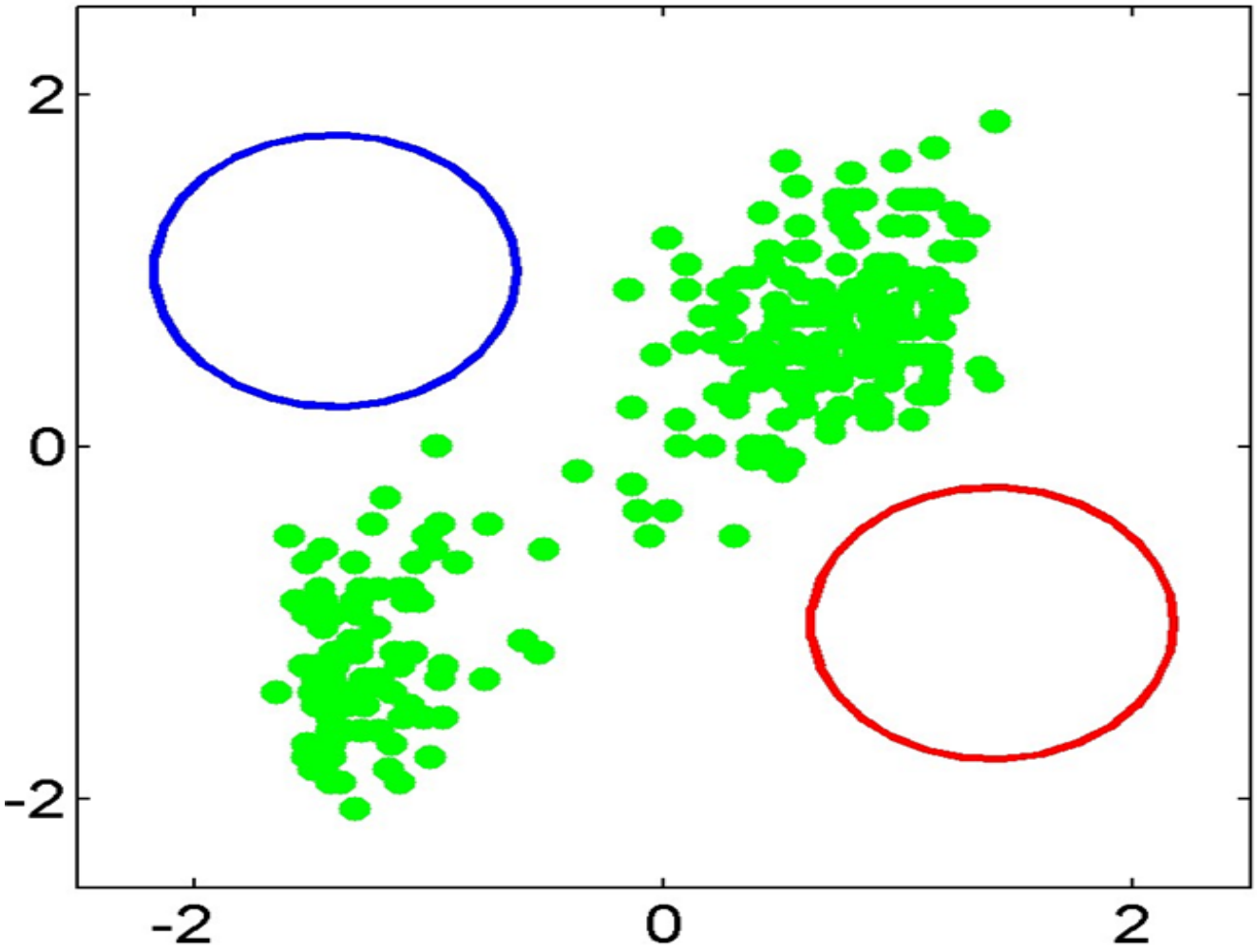
$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n)$$

4. Evaluate log likelihood

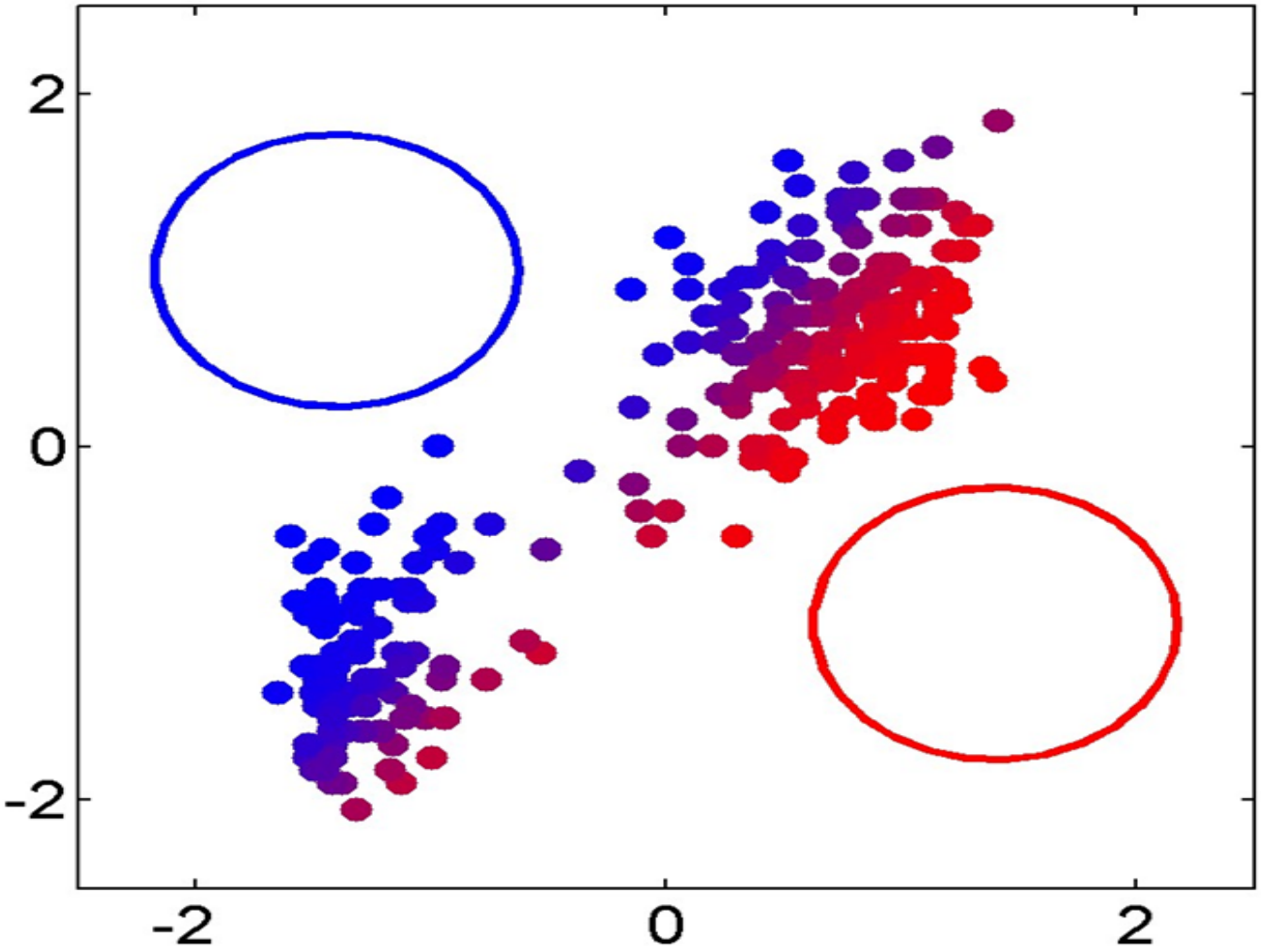
$$\ln p(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}$$

If there is no convergence, return to step 2.

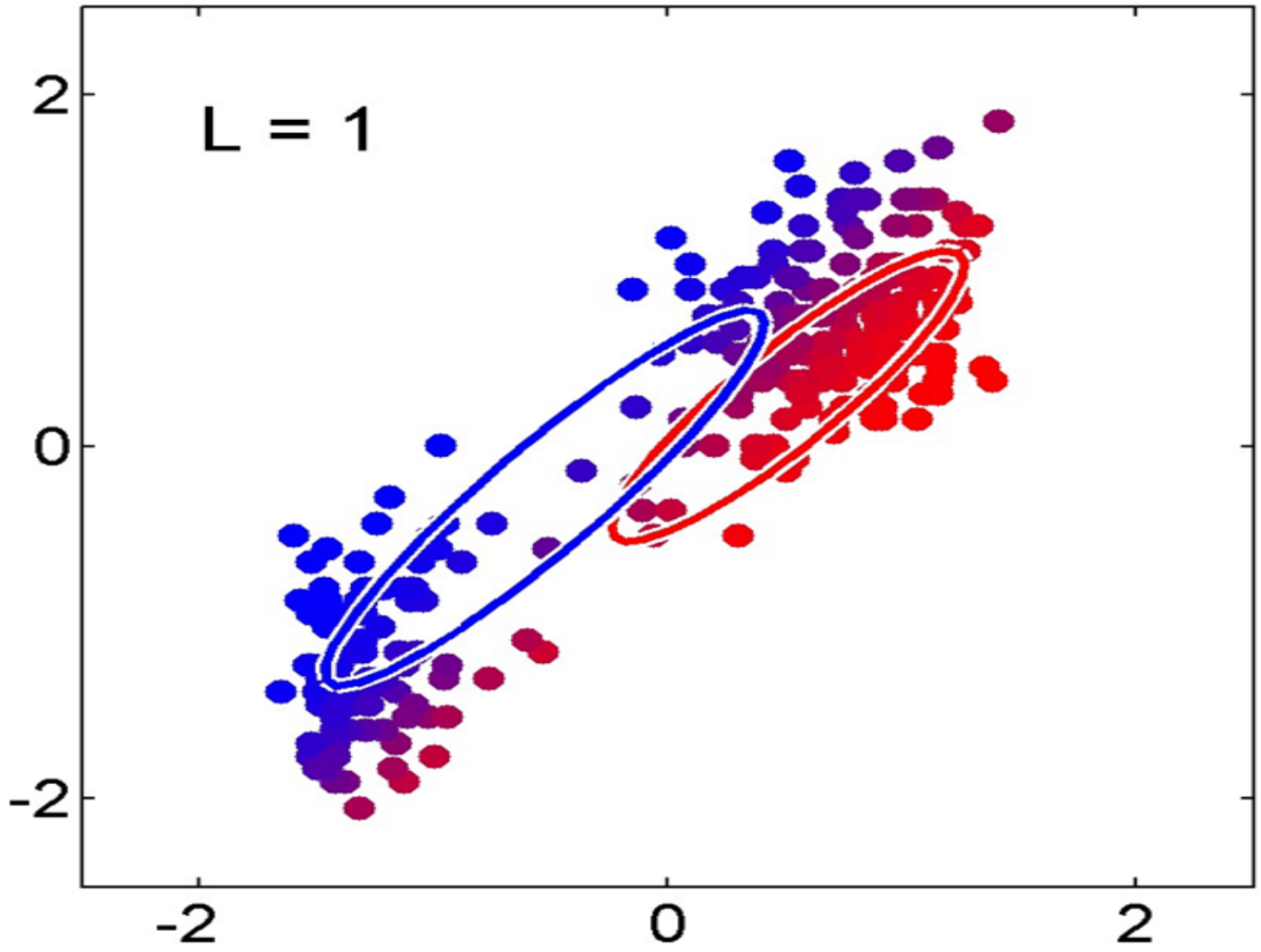
EM Algorithm : Example



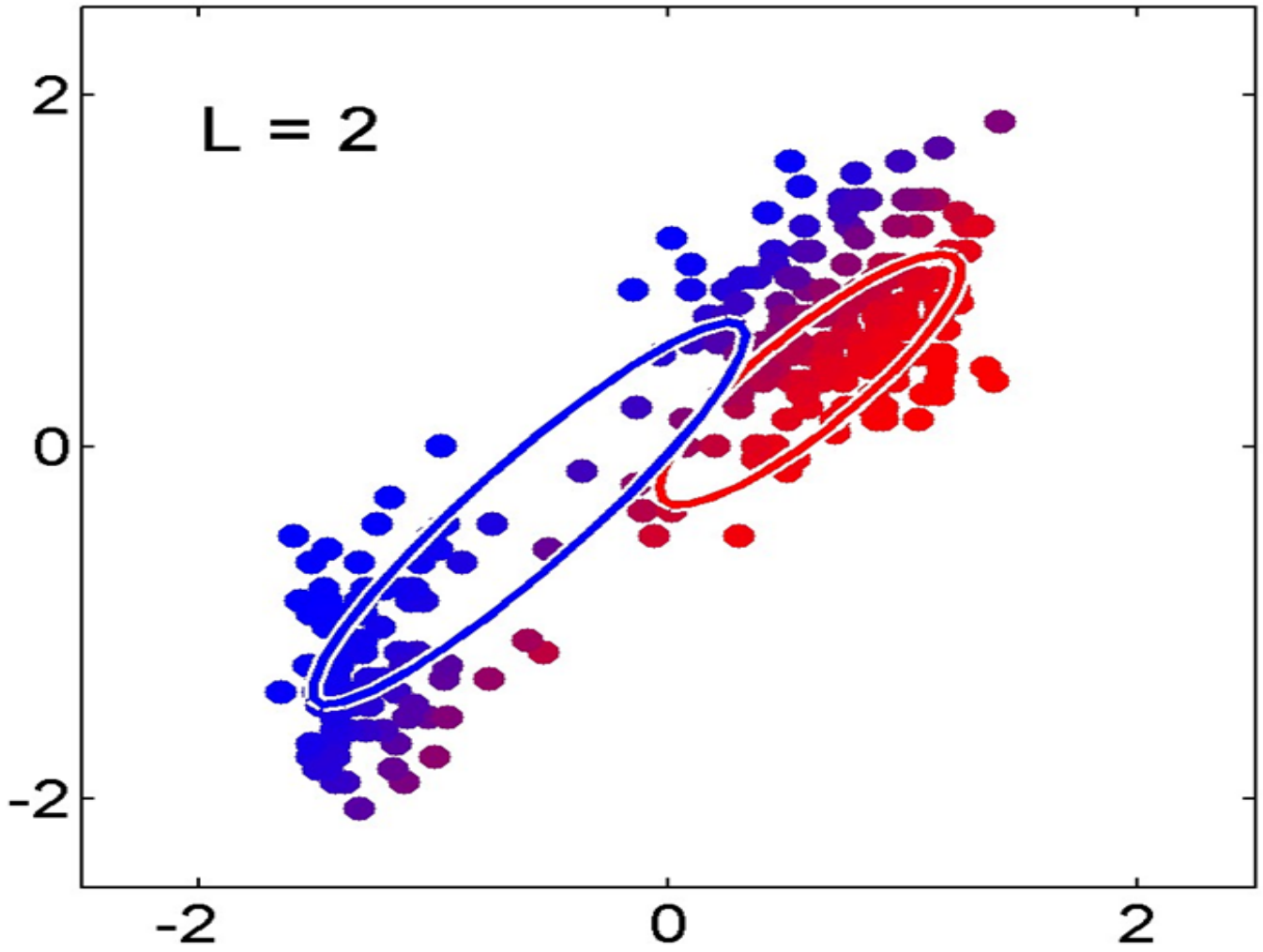
EM Algorithm : Example



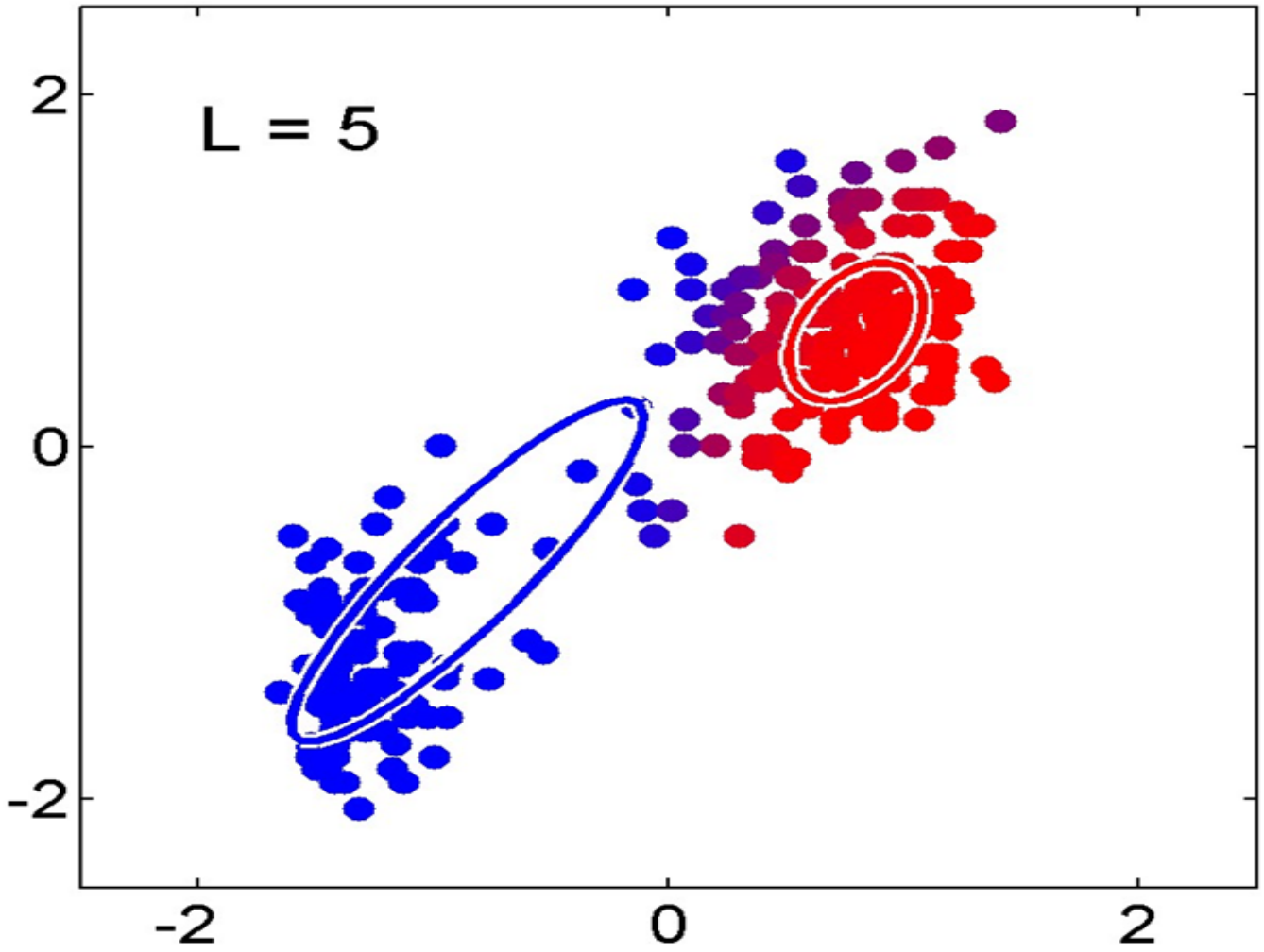
EM Algorithm : Example



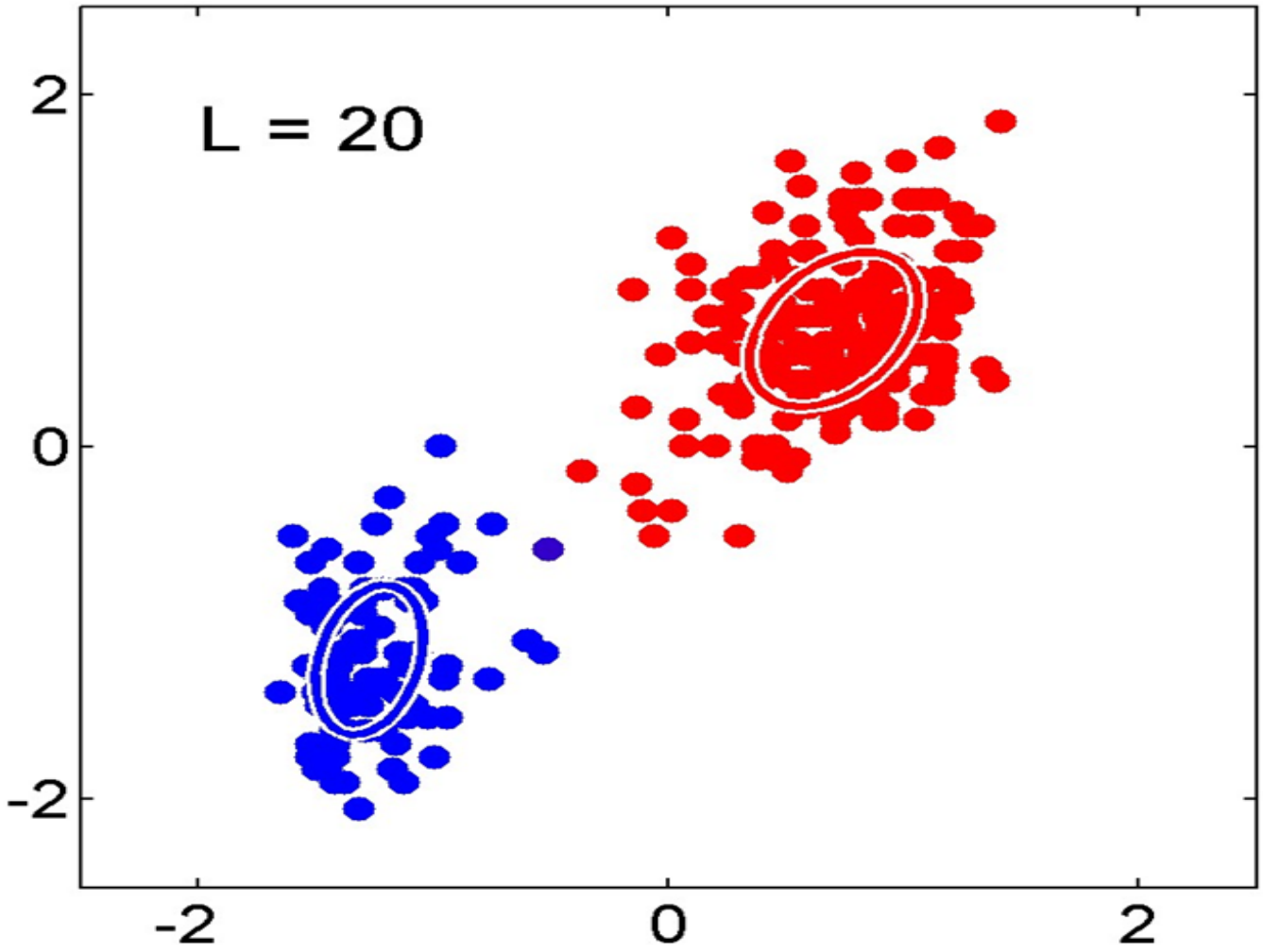
EM Algorithm : Example

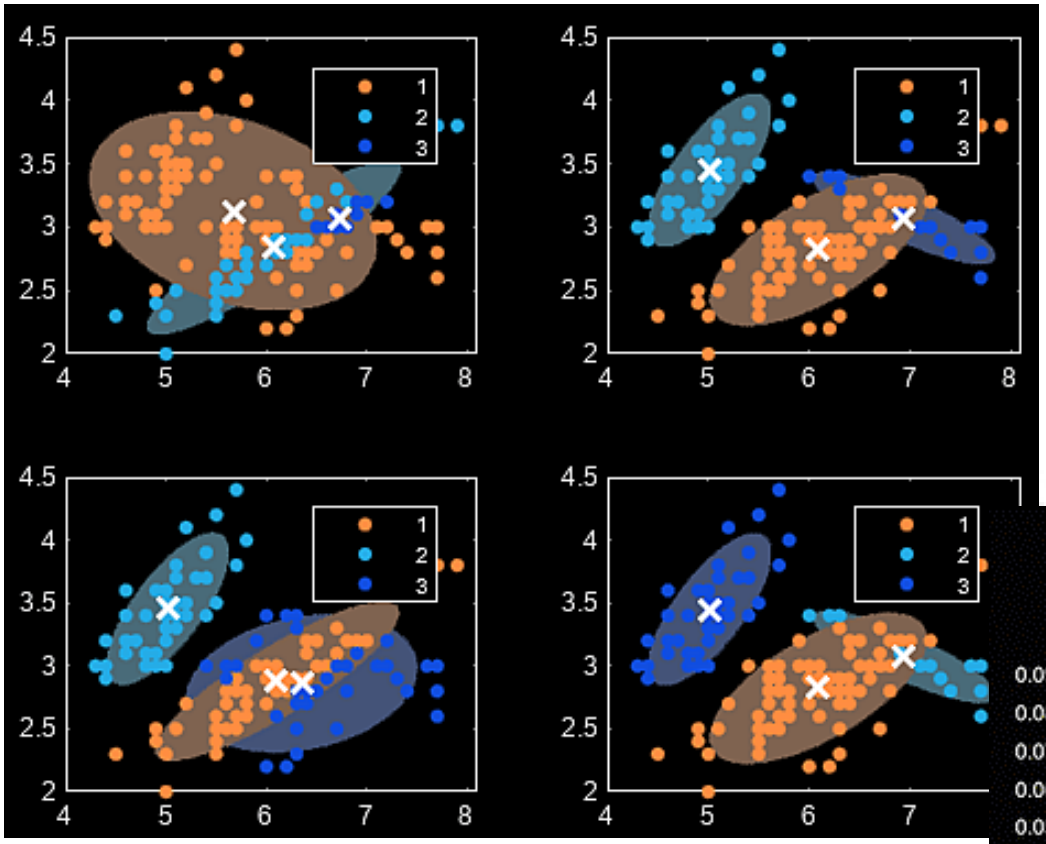


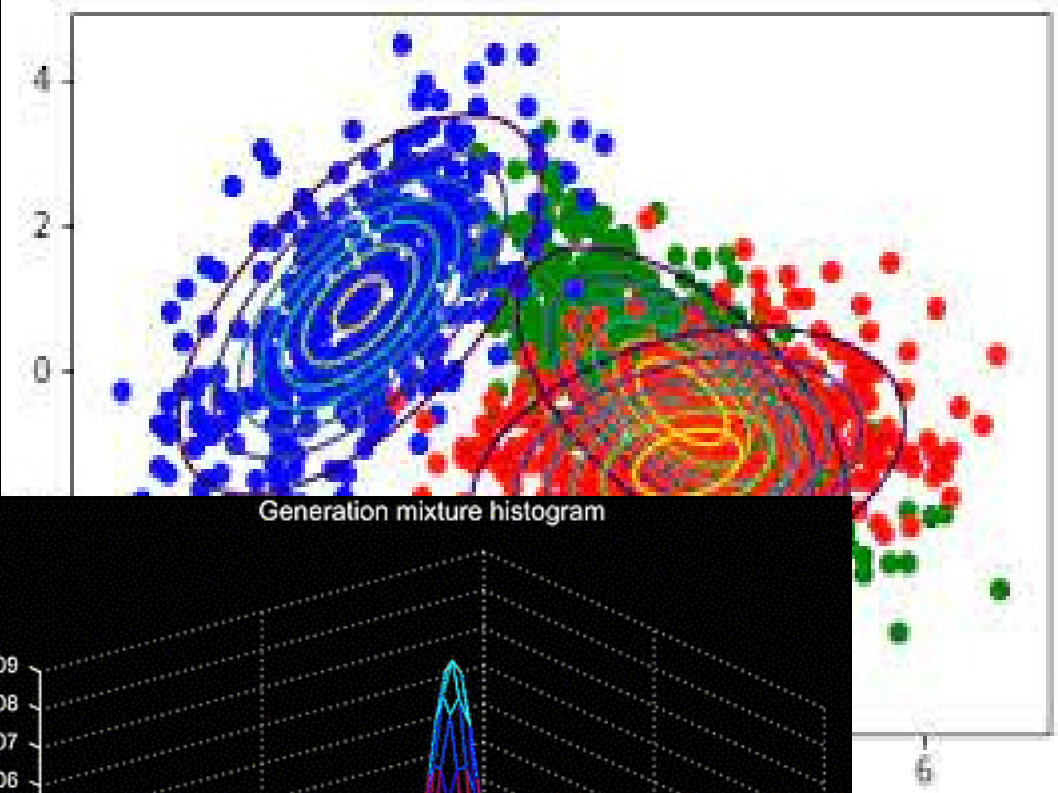
EM Algorithm : Example

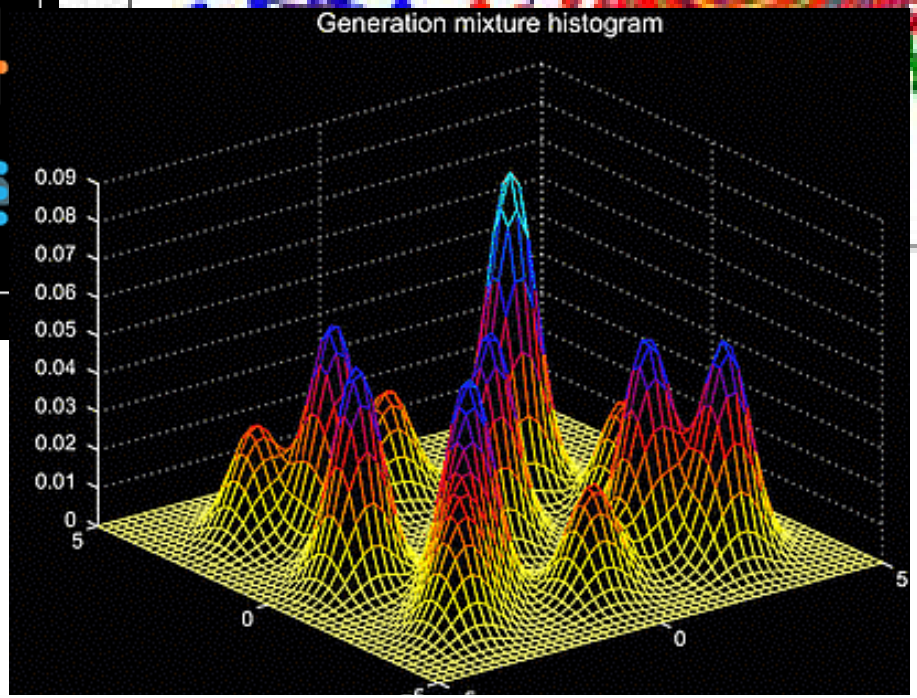


EM Algorithm : Example




1
2
3



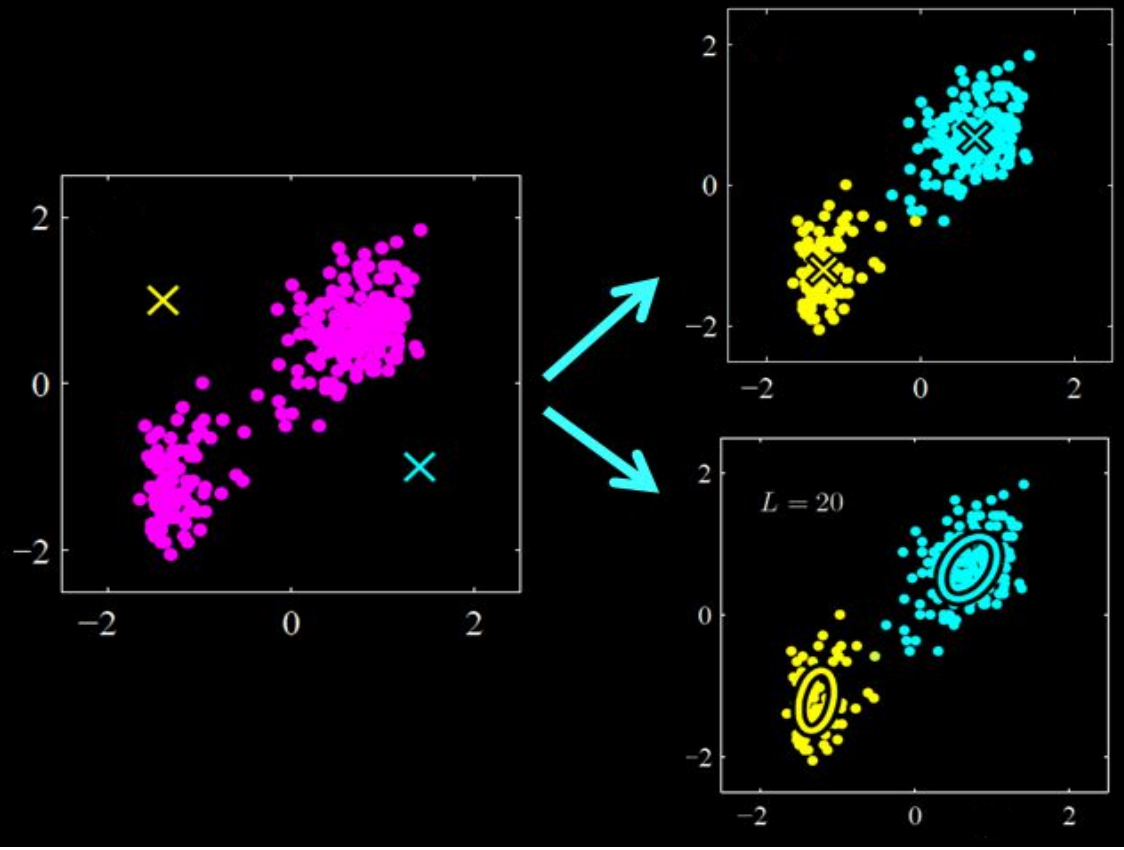

Generation mixture histogram

Imperial College London

EE462 MLCV

K-means vs GMM

Two representative techniques are k-means and Gaussian Mixture Model (GMM). K-means assigns data points to the nearest clusters, while GMM assigns data to the Gaussian densities that best represent the data.

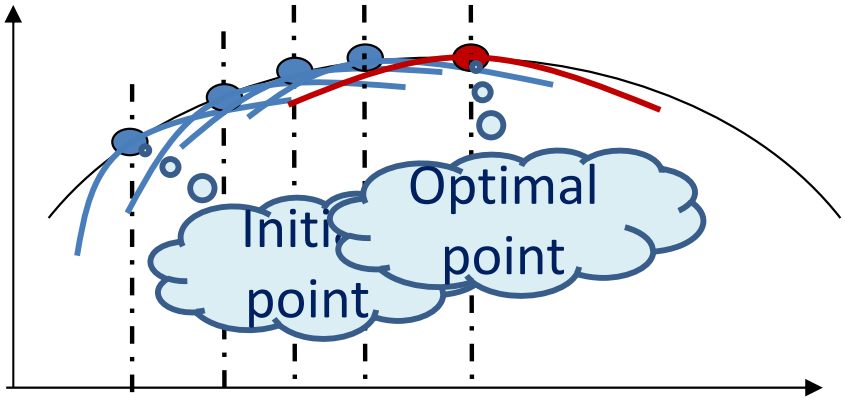


Hard clustering: a data point is assigned only one cluster.

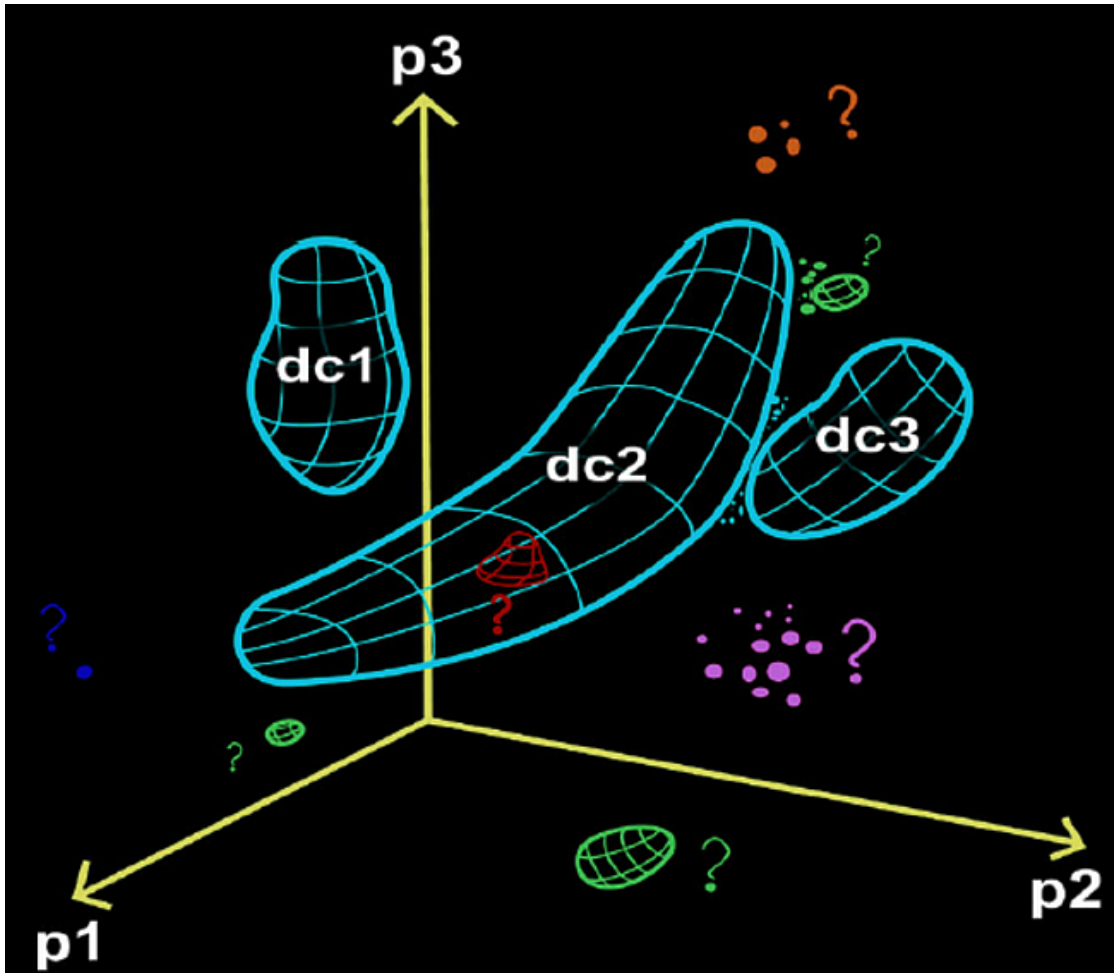
Soft clustering: a data point is assigned multiple Gaussians probabilistically.

Expectation Maximization

- EM algorithm is an iterative optimization technique which is operated locally



- Estimation step: for given parameter values we can compute the expected values of the latent variable.
- Maximization step: updates the parameters of our model based on the latent variable calculated using ML method.



Other Applications of Latent Variable:

- HMM, PGM, LDA (latent Dirichlet Allocation), any mixture models (e.g. multi-variate Bernoulli); Bayesian Learning with mixed graph models (DAG, G-DMG etc.)

Maximum Likelihood Estimation (**MLE**), Expectation-Maximization (**EM**), and Maximum A Posteriori (**MAP**) estimation are techniques to estimate model parameters.

MLE finds parameters maximizing data likelihood.
MAP adds prior knowledge to MLE (Bayes is an example).
EM is an iterative algorithm to find MLE/MAP when data has hidden (latent) variables.

Brief highlights follow:

1. Maximum Likelihood Estimation (**MLE**)

Goal: Find parameters that maximize the likelihood function **P(Data | θ)**. It finds the most likely parameters that generated the observed data. Take derivative of LOG, equate to ZERO and get parameter assignment (or, as done in *Regression* too).

If its log-likelihood, then use:

else:

$$\theta_{MLE} = argmax_{\theta}\ P(x|\theta)$$

2. Expectation-Maximization (**EM**) Algorithm

Goal: Iteratively find the Maximum Likelihood Estimate (MLE) or Maximum A Posteriori (MAP) when the model depends on unobserved latent variables (for GMM, Clustering).

Steps:

E-Step (Expectation): Computes the expected value of the log-likelihood using current parameters to estimate latent variables.

M-Step (Maximization): Updates parameters to maximize the expected log-likelihood from the E-step.

3. Maximum A Posteriori (**MAP**) Estimation

Goal: Maximizes the posterior distribution **P(θ | Data)** . It combines the likelihood with a prior distribution **P(θ)** , aiming for

P(θ | Data) = P(Data | θ) x P(θ) .

Best Used When: Prior knowledge is available, or data is sparse (small datasets)

$$\theta_{MAP} = argmax_\theta \, P(x|\theta)g(\theta)$$

Feature	MLE	MAP	EM
Objective	Maximize $P(D\|\theta)$	Maximize $P(D\|\theta)P(\theta)$	Maximize $P(D\|\theta)$ (iteratively)
Prior	Not used (or uniform)	Uses prior knowledge	Generally uses uniform prior
Data Type	Fully observed	Fully observed	Latent/Missing Data
Method	Closed-form or Optimizer	Closed-form or Optimizer	Two-step iteration (E & M)
Goal	Best parameter point	Best parameter (with prior)	Best parameter with hidden variables

Key Differences

MLE vs. MAP: MAP includes prior belief, reducing over-fitting. MLE is a special case of MAP with a uniform prior.

MLE vs. EM: MLE is the goal; EM is the method to achieve that goal when data is hidden.

Algorithm Structure:

MLE: Directly maximizes the likelihood function to estimate the parameters. It typically involves solving an optimization problem using numerical methods.

EM: Iteratively alternates between two steps:

E-step (Expectation): Computes the expected value of the log-likelihood function with respect to the current estimate of the parameters and the unobserved data.

M-step (Maximization): Maximizes this expected log-likelihood to update the parameter estimates.

EM vs. MAP: EM handles missing data. It can be adapted to perform MAP estimation (often called MAP-EM).

Which is MAP, and which one MLE ??

$$= \arg\max_{\theta} \frac{f(x \mid \theta)\, g(\theta)}{\int_{\Theta} f(x \mid \vartheta)\, g(\vartheta)\, d\vartheta}$$

$$= \arg\max_{\theta} f(x \mid \theta)\, g(\theta).$$