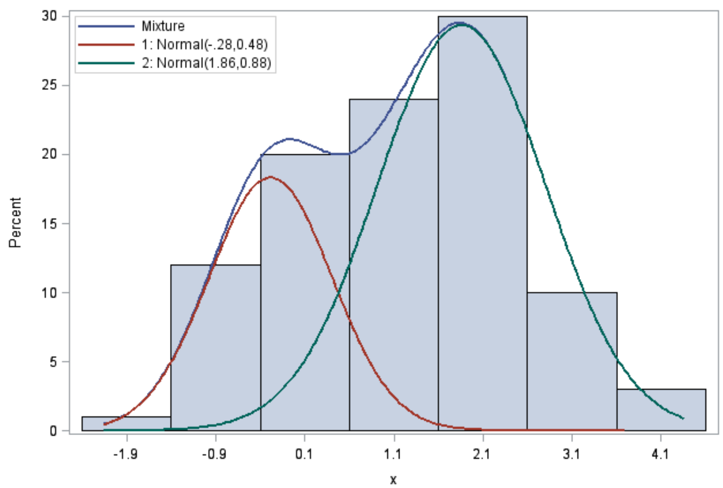
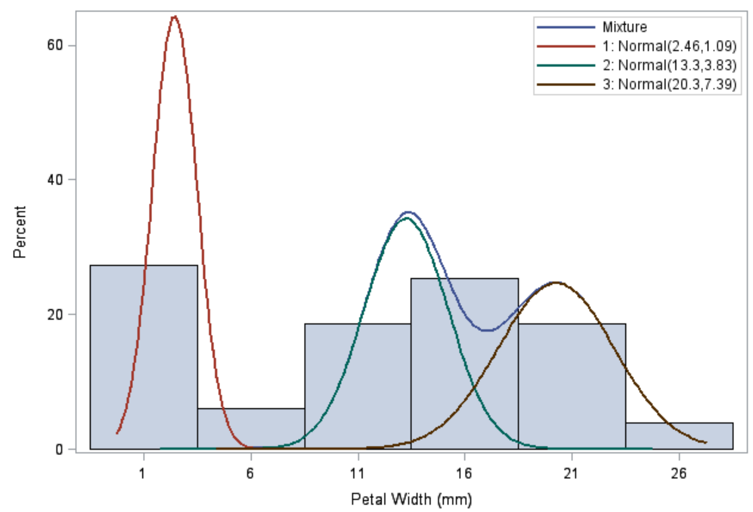
Mixture of Gaussians

Sargur Srihari
srihari@cedar.buffalo.edu

Mixtures of Gaussians

- Also Called as Finite Mixture Models (FMMs)





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

Goal:
Determine the three sets of parameters π, μ, Σ using maximum likelihood

Gaussian Mixture Model

- A linear superposition of Gaussian components
- Provides a richer class of density models than the single Gaussian
- GMM are formulated in terms of discrete latent variables
  - Provides deeper insight
  - Motivates EM algorithm
    - Which gives a maximum likelihood solution to no. of components and their means/covariances

GMM Formulation

- Linear superposition of K Gaussians:

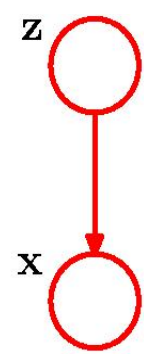
$$p(\mathrm{x}) = \sum_{k=1}^{K} \pi_k N(\mathrm{x} \mid \mu_k, \Sigma_k)$$

- Introduce a latent variable z that has K values
  - Use *1*-of-K representation
    - Let $\mathrm{z} = z_1,..,z_K$ whose elements are

$$z_k \in \{0,1\} \text{ and } \sum_k z_k = 1$$

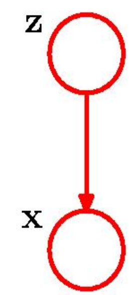
    - There are K possible states of z corresponding to K components

- Define joint distribution $p(\mathrm{x},\mathrm{z}) = \underbrace{p(\mathrm{x}|\mathrm{z})}_{\text{conditional}}\underbrace{p(\mathrm{z})}_{\text{marginal}}$
  - x is observed variable
  - z is hidden or missing variable



Specifying $p(\mathrm{z})$, $\mathrm{z}=\{z_1,..,z_K\}$

- Associate a probability with each component z_k
  - Denote $p(z_k=1)=\pi_k$ where parameters satisfy
    $$0 \le \pi_k \le 1 \text{ and } \sum_k \pi_k = 1$$
- Because z uses *1-of-K* it follows that



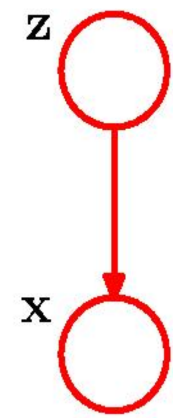
$$p(\mathrm{z})=\prod_{k=1}^{K}\pi_k^{z_k}$$

With one component $p(z_1)=\pi_1^{z_1}$

With two components $p(z_1,z_2)=\pi_1^{z_1}\pi_2^{z_2}$

  - since components are mutually exclusive and hence are independent

Conditional distribution $p(\mathrm{x}|\mathrm{z})$



- For a particular component (value of z)

$$p(\mathrm{x} \mid z_k = 1) = N(\mathrm{x} \mid \mu_k, \Sigma_k)$$

- Thus $p(\mathrm{x}|\mathrm{z})$ can be written in the form

$$p(\mathrm{x} \mid \mathrm{z}) = \prod_{k=1}^{K} N\left(x \mid \mu_k, \Sigma_k\right)^{z_k}$$

Due to the exponent z_k all product terms except for one equal one

- Thus marginal distribution of x is obtained by summing over all possible states of z

$$p(\mathrm{x}) = \sum_{z} p(\mathrm{z}) p(\mathrm{x} \mid \mathrm{z}) = \sum_{z} \prod_{k=1}^{K} \pi_{\mathrm{k}}^{z_k} N\left(\mathrm{x} \mid \mu_k, \Sigma_k\right)^{z_k} = \sum_{k=1}^{K} \pi_{\mathrm{k}} N\left(\mathrm{x} \mid \mu_k, \Sigma_k\right)$$

- This is the standard form of a Gaussian mixture

Latent Variable

- If we have observations $x_1,..,x_N$
- Because marginal distribution is in the form

$$p(\mathrm{x}) = \sum_{z} p(\mathrm{x},\mathrm{z})$$

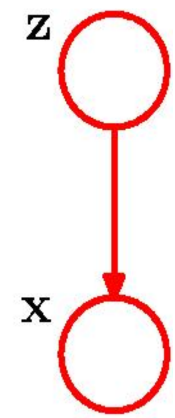
  - It follows that for every observed data point x_n there is a corresponding latent vector z_n , i.e., its sub-class
- Thus we have found a formulation of Gaussian mixture involving an explicit latent variable
  - We are now able to work with joint distribution $p(\mathrm{x},\mathrm{z})$ instead of marginal $p(\mathrm{x})$
- Leads to significant simplification through introduction of expectation maximization

Another conditional probability (Responsibility)

- In EM $p(\mathrm{z}\,|\mathrm{x})$ plays a role
- The probability $p(z_k=1|\mathrm{x})$ is denoted $\gamma(z_k)$
- From Bayes theorem

$$\gamma(\mathrm{z}_k) \equiv p(\mathrm{z}_k=1\,|\,\mathrm{x}) = \frac{p(\mathrm{z}_k=1)\,p(\mathrm{x}\,|\,\mathrm{z}_k=1)}{\sum_{\mathrm{j}=1}^{\mathrm{K}} p(\mathrm{z}_j=1)\,p(\mathrm{x}\,|\,\mathrm{z}_j=1)}$$

$$= \frac{\pi_{\mathrm{k}} N(\mathrm{x}\,|\,\mu_k,\Sigma_k)}{\sum_{\mathrm{j}=1}^{\mathrm{K}} \pi_{\mathrm{j}} N(\mathrm{x}\,|\,\mu_k,\Sigma_j)}$$



$p(\mathrm{x},\mathrm{z})=p(\mathrm{x}|\mathrm{z})p(\mathrm{z})$

- View $p(z_k=1)=\pi_k$ as prior probability of component k

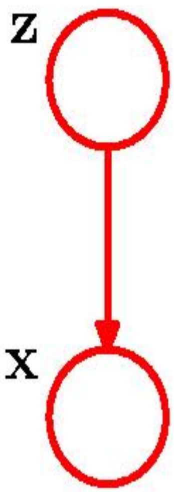
  $\gamma(z_k)=p(z_k=1\,|\,\mathrm{x})$ as the posterior probability

it is also the responsibility that component k takes for explaining the observation x

Plan of Discussion

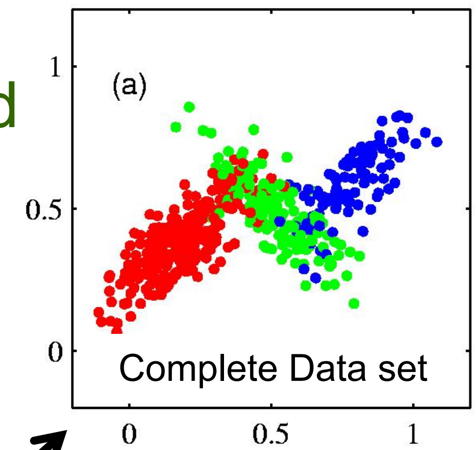
- Next we look at
  1. How to get data from a mixture model synthetically and then
  2. How to estimate the parameters from the data

Synthesizing data from mixture

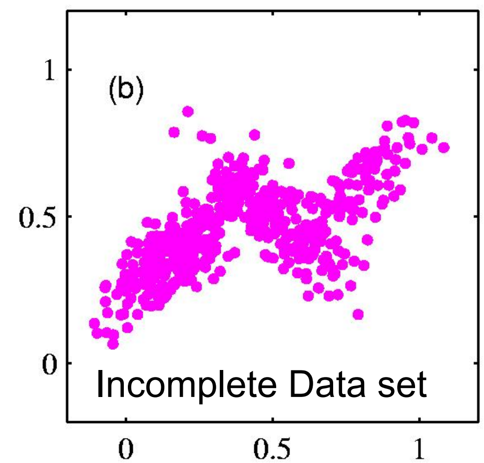


- Use ancestral sampling
  - Start with lowest numbered node and draw a sample,
    - Generate sample of z, called $z^{\wedge}$
  - move to successor node and draw a sample given the parent value, etc.
    - Then generate a value for x from conditional $p(x|z^{\wedge})$
- Samples from $p(x,z)$ are plotted according to value of x and colored with value of z
- Samples from marginal $p(x)$ obtained by ignoring values of z

500 points from three Gaussians

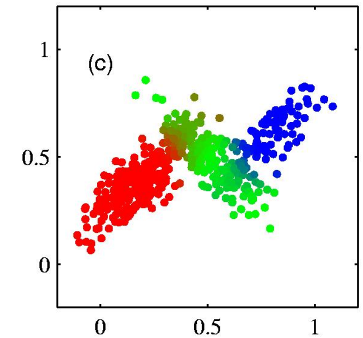


Complete Data set



Incomplete Data set

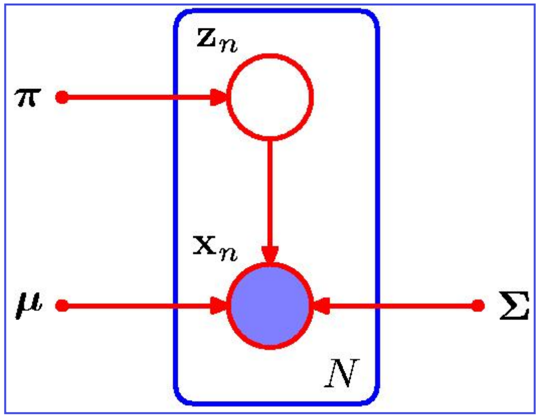
Illustration of responsibilities



- Evaluate for every data point
  - Posterior probability of each component
- Responsibility $\gamma(z_{nk})$ is associated with data point $\mathbf{x}_n$
- Color using proportion of red, blue and green ink
  - If for a data point $\gamma(z_{n1})=1$ it is colored red
  - If for another point $\gamma(z_{n2})=\gamma(z_{n3})=0.5$ it has equal blue and green and will appear as cyan

Maximum Likelihood for GMM

- We wish to model data set $\{x_1,..x_N\}$ using a mixture of Gaussians (N items each of dimension D)
- Represent by N x D matrix X $X = \begin{bmatrix} x_1 \\ x_2 \\ x_n \\ \end{bmatrix}$
  - N^{th} row is given by x_n^T
- Represent N latent variables with N x K matrix Z with rows z_n^T $Z = \begin{bmatrix} z_1 \\ z_2 \\ z_n \end{bmatrix}$
- Bayesian Network for $p(x,z)$ is →



We are given N i.i.d. samples $\{x_n\}$ with *unknown* latent points $\{z_n\}$ with parameters π_n.

The samples have parameters: mean μ and covariance Σ

Goal is to estimate the three sets of parameters

Likelihood Function for GMM

Mixture density function is

$$p(\mathrm{x}) = \sum_{\mathrm{z}} p(\mathrm{z})p(\mathrm{x} \mid \mathrm{z}) = \sum_{k=1}^{K} \pi_{\mathrm{k}} N\left(\mathrm{x} \mid \mu_k, \Sigma_k\right)$$

Summation is for Marginalizing the joint

Therefore Likelihood function is

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

Product is over the N i.i.d. samples

Therefore log-likelihood function is

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

Which we wish to maximize

A more difficult problem than for a single Gaussian

Maximization of Log-Likelihood

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

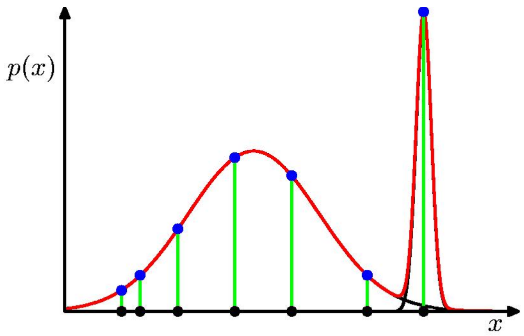
- Goal is to estimate the three sets of parameters π, μ, Σ
- Before proceeding with the m.l.e. briefly mention two technical issues:
  - Problem of singularities
  - Identifiability of mixtures

Problem of Singularities with Gaussian mixtures

- Consider Gaussian mixture
  - components with covariance matrices $\Sigma_k = \sigma_k{}^2 I$
- Data point that falls on a mean $\mu_j = \mathrm{x}_n$ will contribute to the likelihood function

$$\boxed{N(\mathrm{x}_n \mid \mathrm{x}_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}}$$

since $\exp(\mathrm{x}_n-\mu_j)^2=1$



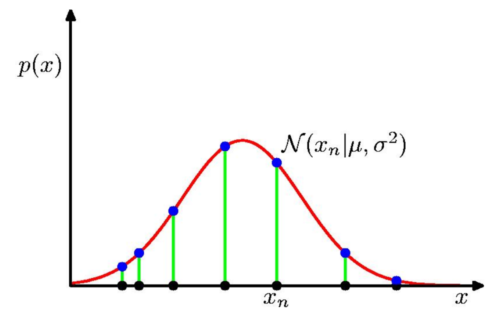
One component assigns finite values and other to large value

- As $\sigma_j \to 0$ term goes to infinity
- Therefore maximization of log-likelihood

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

is not well-posed

  - Does not happen with a single Gaussian
    - Multiplicative factors go to zero
  - Does not happen in the Bayesian approach
- Problem is avoided using heuristics
  - Resetting mean or covariance



Multiplicative values Take it to zero

Problem of Identifiability

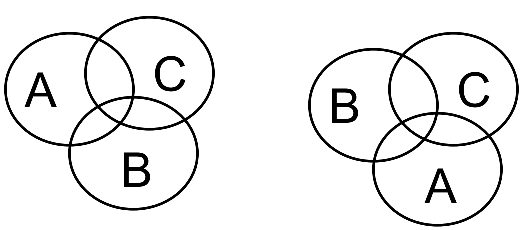
A density $p(\mathrm{x}\,|\,\theta)$ is identifiable if $\theta \neq \theta'$ then there is an x for which $p(\mathrm{x}\,|\,\theta) \neq p(\mathrm{x}\,|\,\theta')$

A K-component mixture will have a total of $K!$ equivalent solutions

- Corresponding to $K!$ ways of assigning K sets of parameters to K components
  - E.g., for $K=3$ $K!=6$: 123, 132, 213, 231, 312, 321
- For any given point in the space of parameter values there will be a further $K!$-1 additional points all giving exactly same distribution

- However any of the equivalent solutions is as good as the other

Two ways of labeling three Gaussian subclasses



EM for Gaussian Mixtures

- EM is a method for finding maximum likelihood solutions for models with latent variables
- Begin with log-likelihood function

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

- We wish to find π, μ, Σ that maximize this quantity
- Take derivatives in turn w.r.t
  - Means μ_k and set to zero
  - covariance matrices Σ_k and set to zero
  - mixing coefficients π_k and set to zero

EM for GMM: Derivative wrt μ_k

- Begin with log-likelihood function

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

- Take derivative w.r.t the means μ_k and set to zero
  - Making use of exponential form of Gaussian
  - Use formulas: $\frac{d}{dx}\ln u = \frac{u'}{u}$ and $\frac{d}{dx}e^u = e^u u'$
  - We get

$$0 = \sum_{n=1}^{N} \underbrace{\boxed{\frac{\pi_k N(\mathrm{x}_n \mid \mu_k,\Sigma_k)}{\sum_j \pi_j N(\mathrm{x}_n \mid \mu_j,\Sigma_j)}}}_{\gamma(z_{nk})} \boxed{\Sigma_k^{-1}} (\mathrm{x}_n - \mu_k)$$

Inverse of covariance matrix

$\gamma(z_{nk})$ the posterior probabilities

M.L.E. solution for Means

- Multiplying by Σ_k (assuming non-singularity)

$$\boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n}$$

Mean of k^{th} Gaussian component is the weighted mean of <u>all</u> the points in the data set:
where data point x_n is weighted by the posterior probability that component k was responsible for generating x_n

- Where we have defined

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

  – Which is the effective number of points assigned to cluster k

M.L.E. solution for Covariance

- Set derivative wrt Σ_k to zero
  - Making use of mle solution for covariance matrix of single Gaussian

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathrm{x}_n - \mu_k)(\mathrm{x}_n - \mu_k)^T$$

  - Similar to result for a single Gaussian for the data set but each data point weighted by the corresponding posterior probability
  - Denominator is effective no of points in component

M.L.E. solution for Mixing Coefficients

- Maximize $\ln p(X \mid \pi, \mu, \Sigma)$ w.r.t. π_k
  - Must take into account that mixing coefficients sum to one
  - Achieved using Lagrange multiplier and maximizing

$$\ln p(X \mid \pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

  - Setting derivative wrt π_k to zero and solving gives

$$\boxed{\pi_k = \frac{N_k}{N}}$$

Summary of m.l.e. expressions

- GMM maximum likelihood estimates

$$\boxed{\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathrm{x}_n}$$

Parameters (means)

$$\boxed{\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathrm{x}_n - \mu_k)(\mathrm{x}_n - \mu_k)^T}$$

Parameters(covariance matrices)

$$\boxed{\pi_k = \frac{N_k}{N}} \qquad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

Parameters
(Mixing Coefficients)

All three are in terms of responsibliities

EM Formulation

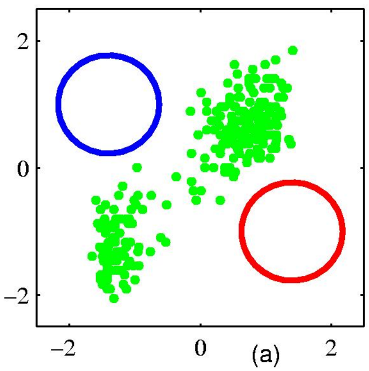
- The results for μ_k, Σ_k, π_k are not closed form solutions for the parameters
  - Since $\gamma(z_{nk})$ the responsibilities depend on those parameters in a complex way
- Results suggest an iterative solution
- An instance of EM algorithm for the particular case of GMM

Informal EM for GMM

- First choose initial values for means, covariances and mixing coefficients
- Alternate between following two updates
  - Called E step and M step
- In E step use current value of parameters to evaluate posterior probabilities, or responsibilities
- In the M step use these posterior probabilities to to re-estimate means, covariances and mixing coefficients

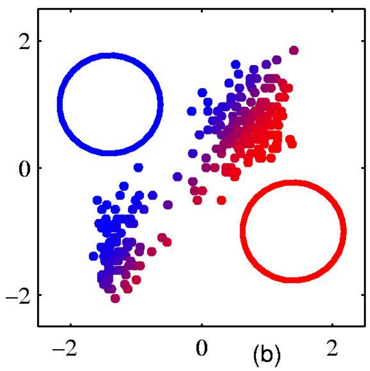
EM using Old Faithful

Data points and
Initial mixture model



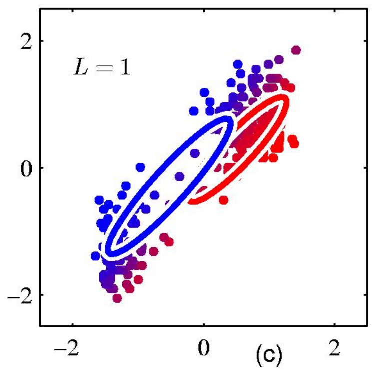
(a)

Initial E step
Determine
responsibilities



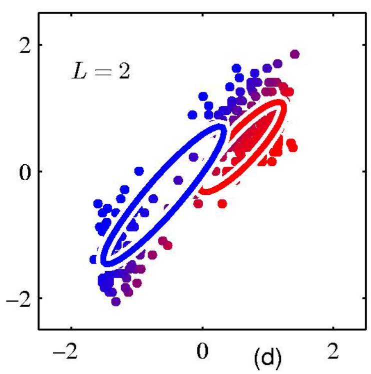
(b)

After first M step
Re-evaluate Parameters



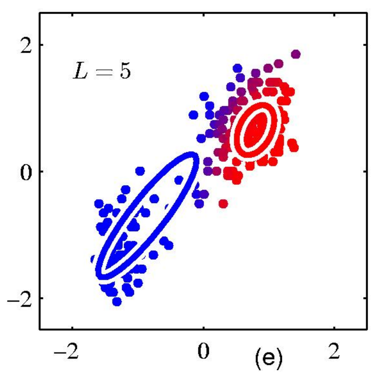
(c)

After 2 cycles



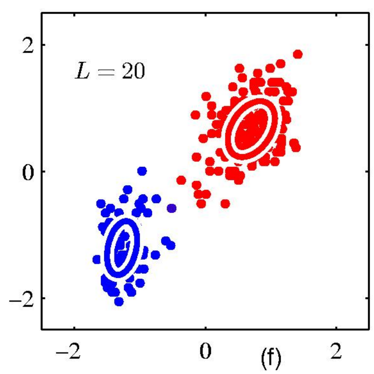
(d)

After 5 cycles



(e)

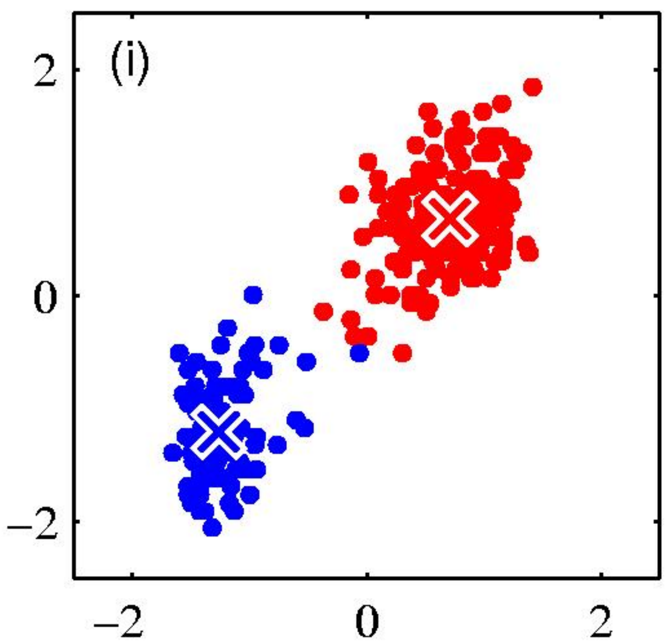
After 20 cycles



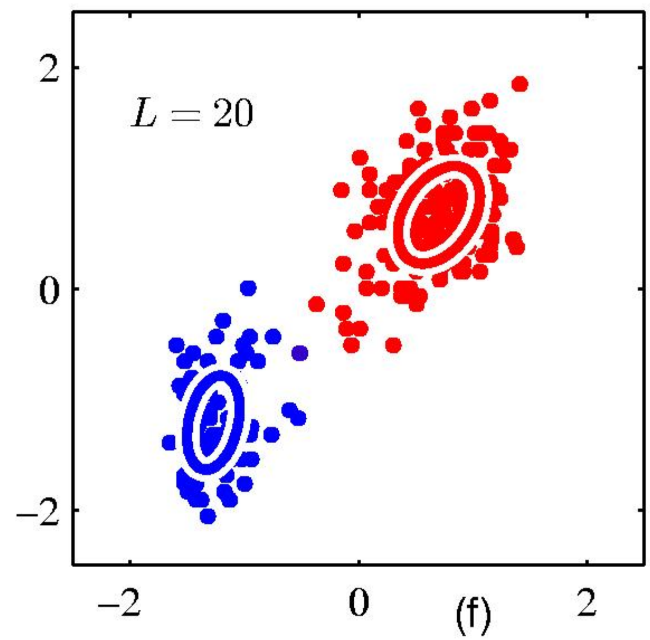
(f)

Comparison with *K*-Means

K-means result



E-M result



Animation of EM for Old Faithful Data

- http://en.wikipedia.org/wiki/File:Em_old_faithful.gif

Code in R

```
#initial parameter estimates (chosen to be deliberately bad)
theta <- list( tau=c(0.5,0.5),
mu1=c(2.8,75),
mu2=c(3.6,58),
sigma1=matrix(c(0.8,7,7,70),ncol=2),
sigma2=matrix(c(0.8,7,7,70),ncol=2) )
```

Practical Issues with EM

- Takes many more iterations than K-means
- Each cycle requires significantly more comparison
- Common to run K-means first in order to find suitable initialization
- Covariance matrices can be initialized to covariances of clusters found by K-means
- EM is not guaranteed to find global maximum of log likelihood function

Summary of EM for GMM

- Given a Gaussian mixture model
- Goal is to maximize the likelihood function w.r.t. the parameters (means, covariances and mixing coefficients)

Step1: Initialize the means μ_k, covariances Σ_k and mixing coefficients π_k and evaluate initial value of log-likelihood

EM continued

- Step 2: E step: Evaluate responsibilities using current parameter values

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

- Step 3: M Step: Re-estimate parameters using current responsibilities

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathrm{x}_n - \mu_k^{\text{new}})(\mathrm{x}_n - \mu_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

EM Continued

- Step 4: Evaluate the log likelihood

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

  – And check for convergence of either parameters or log likelihood

- If convergence not satisfied return to Step 2