Mixtures of Gaussians

Sargur Srihari
srihari@cedar.buffalo.edu

9. Mixture Models and EM

Topics in Mixtures of Gaussians

- Goal of Gaussian Mixture Modeling
- Latent Variables
- Maximum Likelihood
- EM for Gaussian Mixtures

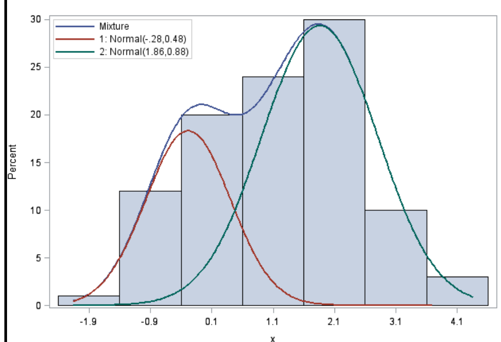
Goal of Gaussian Mixture Modeling

- A linear superposition of Gaussians in the form

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

- Goal of Modeling:
  - Find maximum likelihood parameters π_k, $\boldsymbol{\mu}_k$, Σ_k
  - Examples of data sets and models

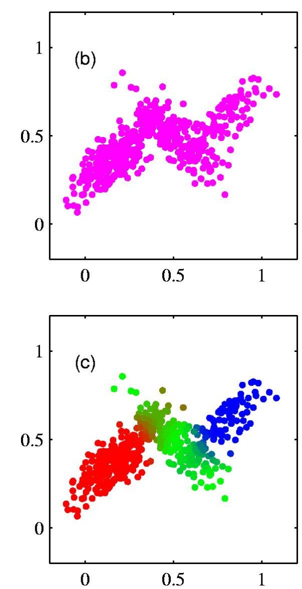
1-D data, K=2 subclasses



k	1	2
π	0.4	0.6
μ	–28	1.86
σ	0.48	0.88

Each data point is associated with a subclass k with probability π_k

2-D data, K=3



GMMs and Latent Variables

- A GMM is a linear superposition of Gaussian components
  - Provides a richer class of density models than the single Gaussian
- We formulate a GMM in terms of discrete latent variables
  - This provides deeper insight into this distribution
  - Serves to motivate the EM algorithm
    - Which gives a maximum likelihood solution to no. of components and their means/covariances

Latent Variable Representation

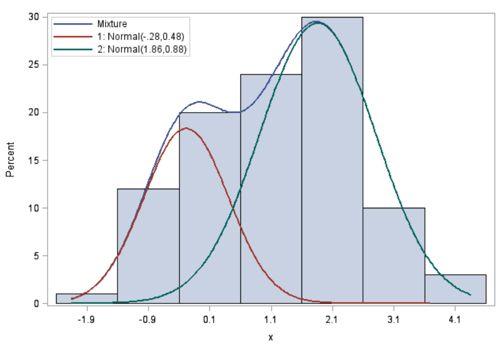
- Linear superposition of K Gaussians:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

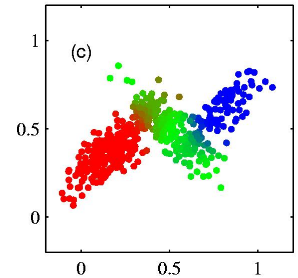
- Introduce a K-dimensional binary variable $\boldsymbol{z}$
  - Use 1-of-K representation (one-hot vector)
    - Let $\boldsymbol{z} = z_1,..,z_K$ whose elements are

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

    - K possible states of $\boldsymbol{z}$ corresponding to K components



k	1	2
$\boldsymbol{z}$	10	01
π_k	0.4	0.6
μ_k	–28	1.86
σ_k	0.48	0.88



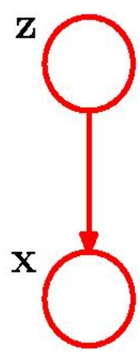
k	1	2	3
$\boldsymbol{z}$	100	010	001

Joint Distribution

- Define joint distribution of latent variable and observed variable
  - $p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{x}|\boldsymbol{z})\bullet\ p(\boldsymbol{z})$
  - $\boldsymbol{x}$ is observed variable
  - $\boldsymbol{z}$ is the hidden or missing variable
  - Marginal distribution $p(\boldsymbol{z})$
  - Conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$

Graphical Representation of Mixture Model

- The joint distribution $p(\boldsymbol{x},\boldsymbol{z})$ is represented in the form $p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$



Latent variable $\boldsymbol{z}=[z_1,..z_K]$ represents subclass

Observed variable $\boldsymbol{x}$

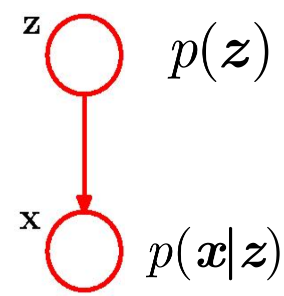
– We now specify marginal $p(\boldsymbol{z})$and conditional $p(\boldsymbol{x}|\boldsymbol{z})$
  - Using them we specify $p(\boldsymbol{x})$ in terms of observed and latent variables

Specifying the marginal $p(\boldsymbol{z})$

- Associate a probability with each component z_k
  - Denote $p(z_k = 1) = \pi_k$ where parameters $\{\pi_k\}$ satisfy

$$0 \le \pi_k \le 1 \text{ and } \sum_k \pi_k = 1$$

- Because $\boldsymbol{z}$ uses 1-of-K it follows that



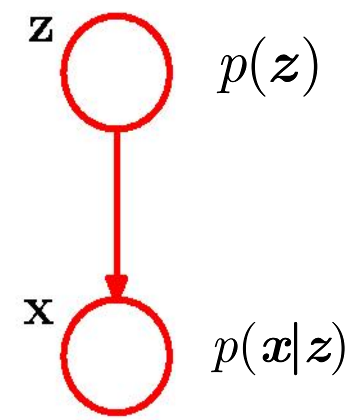
$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

  - since $z_k \in \{0,1\}$ and components of $\boldsymbol{z}$ are mutually exclusive and hence are independent

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

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

Specifying the Conditional $p(\boldsymbol{x}|\boldsymbol{z})$



- For a particular component (value of $\boldsymbol{z}$)

$$p(\boldsymbol{x} \mid z_k = 1) = N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$$

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

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

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

Marginal distribution $p(\boldsymbol{x})$

- The joint distribution $p(\boldsymbol{x},\boldsymbol{z})$ is given by $p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$
- Thus marginal distribution of $\boldsymbol{x}$ is obtained by summing over all possible states of $\boldsymbol{z}$ to give

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

  – Since $z_k \in \{0,1\}$
- This is the standard form of a Gaussian mixture

Value of Introducing Latent Variable

- If we have observations $\boldsymbol{x}_1,..,\boldsymbol{x}_N$
- Because marginal distribution is in the form

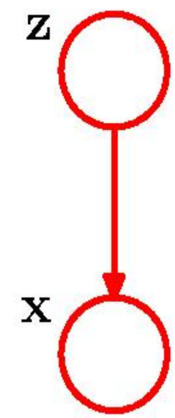
$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z})$$

  - It follows that for every observed data point $\boldsymbol{x}_n$ there is a corresponding latent vector $\boldsymbol{z}_n$, 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 $p(\boldsymbol{x},\boldsymbol{z})$ instead of marginal $p(\boldsymbol{x})$
- Leads to significant simplification through introduction of expectation maximization

Another conditional probability (Responsibility)

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

$$\boxed{\begin{aligned}\gamma(z_k) \equiv p(z_k=1\,|\,\boldsymbol{x}) &= \frac{p(z_k=1)p(\boldsymbol{x}\,|\,z_k=1)}{\sum_{j=1}^{K} p(z_j=1)p(\boldsymbol{x}\,|\,z_j=1)} \\ &= \frac{\pi_k N(\boldsymbol{x}\,|\,\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}\,|\,\mu_k,\Sigma_j)}\end{aligned}}$$



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

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

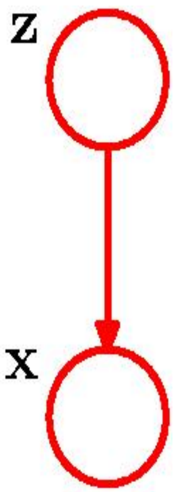
  $\gamma(z_k)=p(z_k=1\,|\,\boldsymbol{x})$ as the posterior probability

  it is also the responsibility that component k takes for explaining the observation $\boldsymbol{x}$

Plan of Discussion

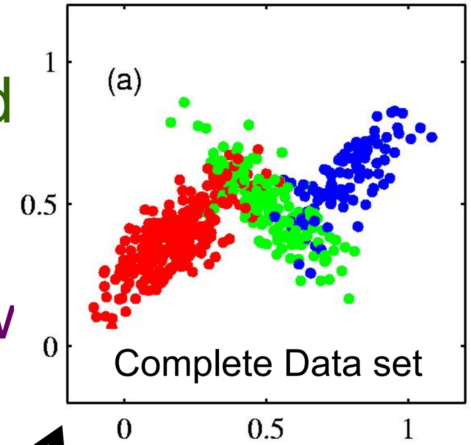
- Next we look at
  1. How to get data from a mixture model synthetically and then
  2. Given a data set $\{\boldsymbol{x}_1,..\boldsymbol{x}_N\}$ how to model the data using a mixture of Gaussians

Synthesizing data from mixture



- Use ancestral sampling
  - Start with lowest numbered node and draw a sample,
    - Generate sample of z, called $\hat{z}$
    - move to successor node and draw a sample given the parent value, etc.
    - Then generate a value for x from conditional $p(x|\hat{z})$
- 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



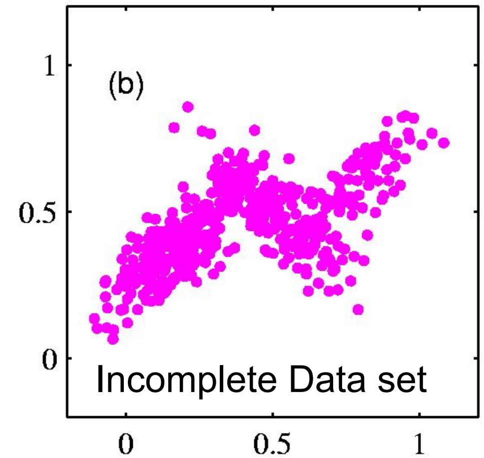
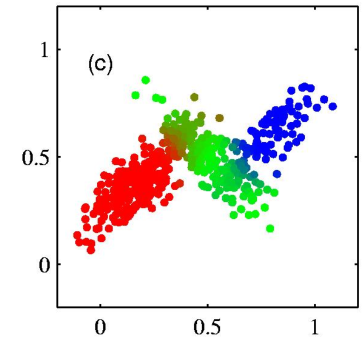


Illustration of responsibilities



- Evaluate for every data point
  - Posterior probability of each component
- Responsibility $\gamma(z_{nk})$ is associated with data point $\boldsymbol{x}_n$
- 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 $\{\boldsymbol{x}_1,..\boldsymbol{x}_N\}$ using a mixture of Gaussians (N items each of dimension D)
- Represent by N x D matrix X
  - n^{th} row is given by $\boldsymbol{x}_n{}^T$

$$X = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \\ \boldsymbol{x}_N \end{bmatrix}$$

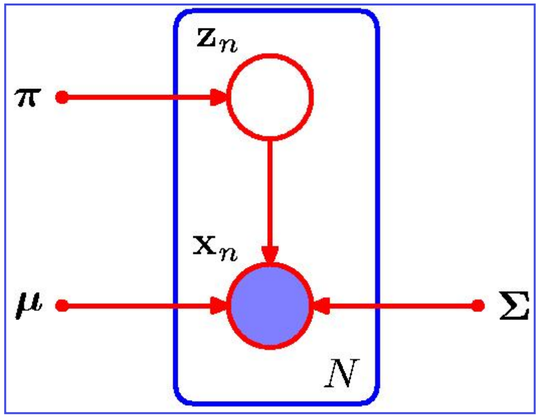
- Represent N latent variables with N x K matrix Z
  - n^{th} row is given by $\boldsymbol{z}_n{}^T$

$$Z = \begin{bmatrix} \mathrm{z}_1 \\ \mathrm{z}_2 \\ \\ \mathrm{z}_N \end{bmatrix}$$

- Goal is to state the likelihood function
  - so as to estimate the three sets of parameters
  - by maximizing the likelihood

Graphical representation of GMM

- For a set of i.i.d. data points $\{\boldsymbol{x}_n\}$ with corresponding latent points $\{\boldsymbol{z}_n\}$ where $n=1,..,N$
- Bayesian Network for $p(\boldsymbol{X},\boldsymbol{Z})$ using plate notation
  - N x D matrix $\boldsymbol{X}$
  - N x K matrix $\boldsymbol{Z}$



Likelihood Function for GMM

Mixture density function is

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

Since z has values $\{z_k\}$ with probabilities $\{\pi_k\}$

Therefore Likelihood function is

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

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

Therefore log-likelihood function is

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

Which we wish to maximize

A more difficult problem than for a single Gaussian

Maximization of Log-Likelihood

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

- Goal is to estimate the three sets of parameters

$$\pi_k, \mu_k, \Sigma_k$$

  - By taking derivatives in turn w.r.t each while keeping others constant
  - But there are no closed-form solutions
    - Task is not straightforward since summation appears in Gaussian and logarithm does not operate on Gaussian
- While a gradient-based optimization is possible, we consider the iterative EM algorithm

Some issues with GMM m.l.e.

- Before proceeding with the m.l.e. briefly mention two technical issues:
  1. Problem of singularities with Gaussian mixtures
  2. Problem of 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 $\mu_j = \boldsymbol{x}_n$ will contribute to the likelihood function

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

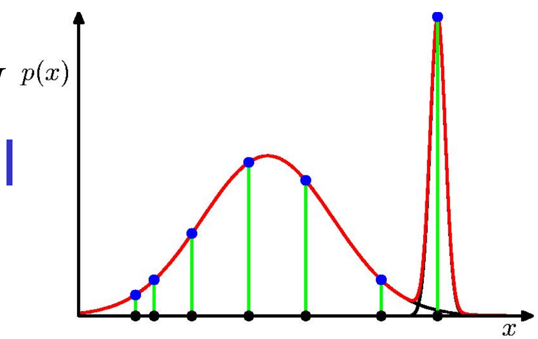
since $\exp(\boldsymbol{x}_n\text{-}\mu_j)^2=1$

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

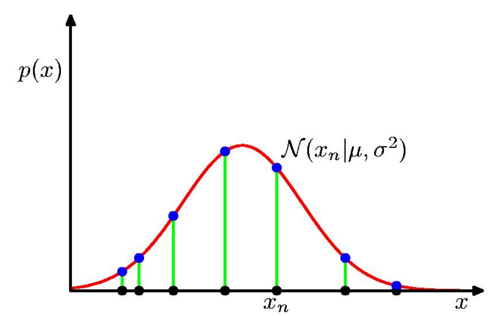
$$\ln p(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n \mid \mu_k, \Sigma_k) \right\}$$

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



One component assigns finite values and other to large value



Multiplicative values Take it to zero

Problem of Identifiability

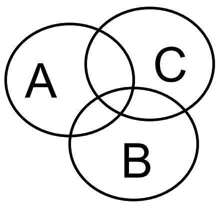
A density $p(\mathrm{x} \mid \theta)$ is identifiable if $\theta \neq \theta'$ then there is an x for which $p(\mathrm{x} \mid \theta) \neq p(\mathrm{x} \mid \theta')$

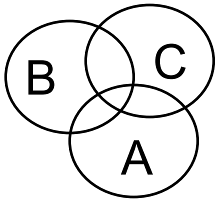
A K-component mixture will have a total of $K!$ equivalent solutions

- 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

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

  - We wish to find π, μ, Σ that maximize this quantity
  - Task is not straightforward since summation appears in Gaussian and logarithm does not operate on Gaussian
- 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

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

- 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

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

Inverse of covariance matrix

M.L.E. solution for Means

- Multiplying by Σ_k (assuming non-singularity)

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

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

- 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

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

  - 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 parameter estimates

Means

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

Covariance matrices

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

Mixing Coefficients

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

- All three are in terms of responsibilities
- and so we have not completely solved the problem

EM Formulation

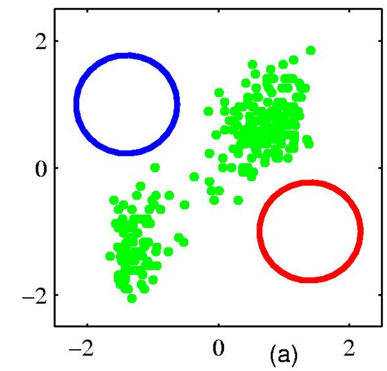
- The results for μ_k, Σ_k, π_k are not closed form solutions for the parameters
  - Since $\gamma(z_{nk})$ the responsibilities depend on those parameters in a complex way
- Results suggest an iterative solution
- An instance of EM algorithm for the particular case of GMM

Informal EM for GMM

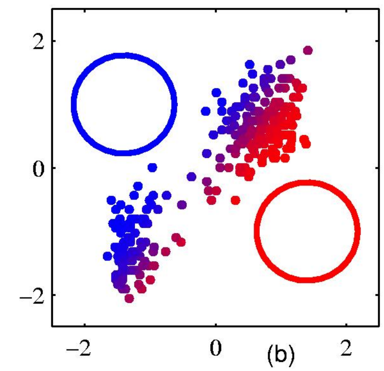
- First choose initial values for means, covariances and mixing coefficients
- Alternate between following two updates
  - Called E step and M step
- In E step use current value of parameters to evaluate posterior probabilities, or responsibilities
- In the M step use these posterior probabilities to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

EM using Old Faithful

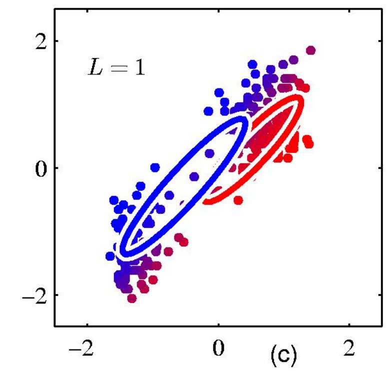
Data points and
Initial mixture model



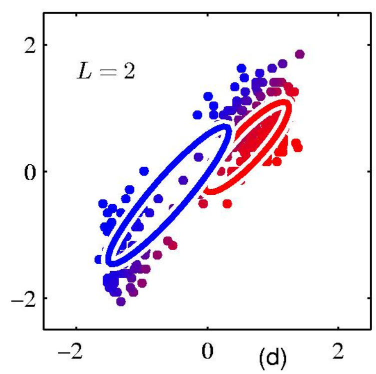
Initial E step
Determine
responsibilities



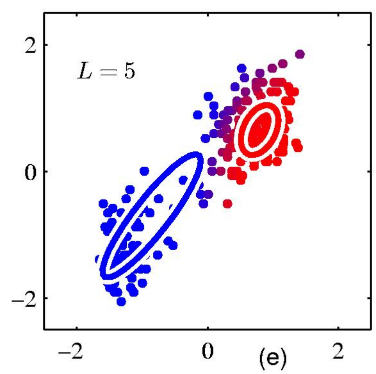
After first M step
Re-evaluate Parameters



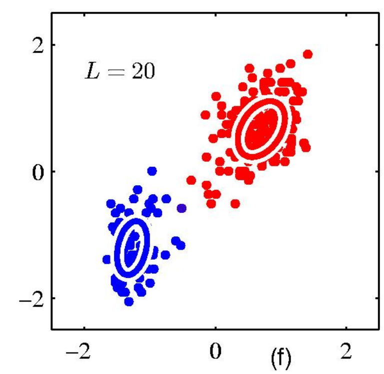
After 2 cycles



After 5 cycles

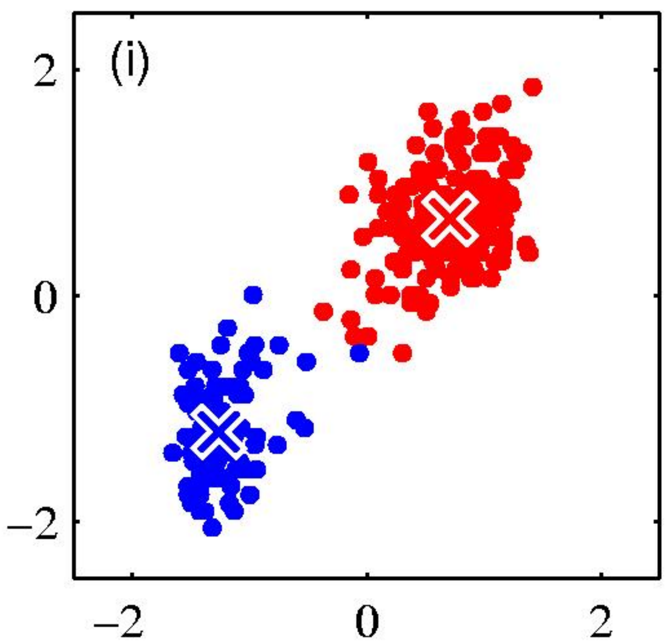


After 20 cycles

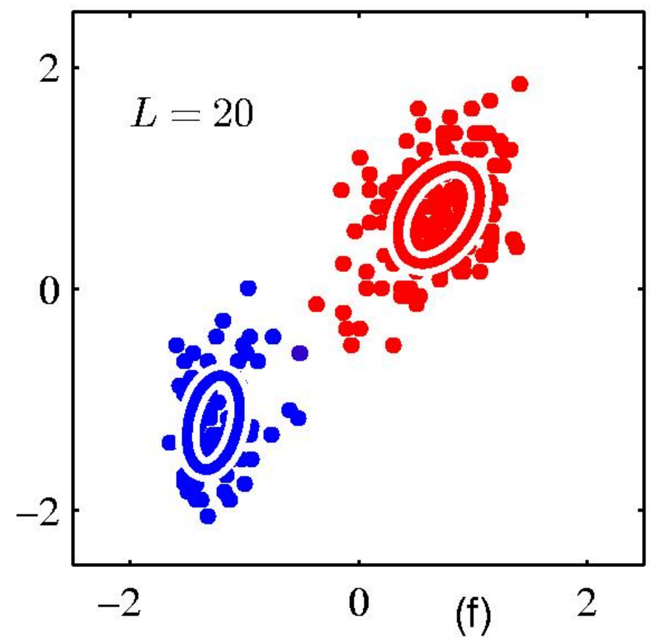


Comparison with K-Means

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

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

- 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