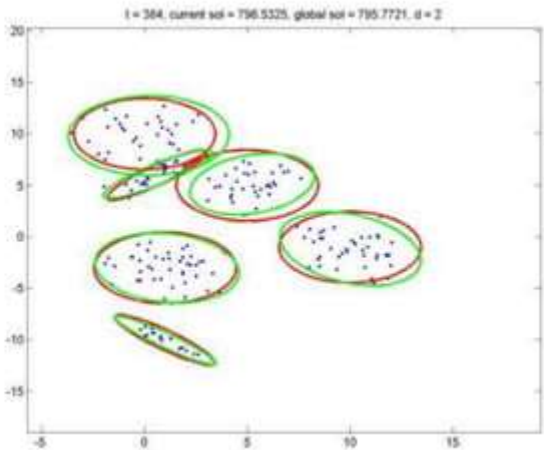
SNS COLLEGE OF TECHNOLOGY
Coimbatore – 35
An Autonomous Institution
Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

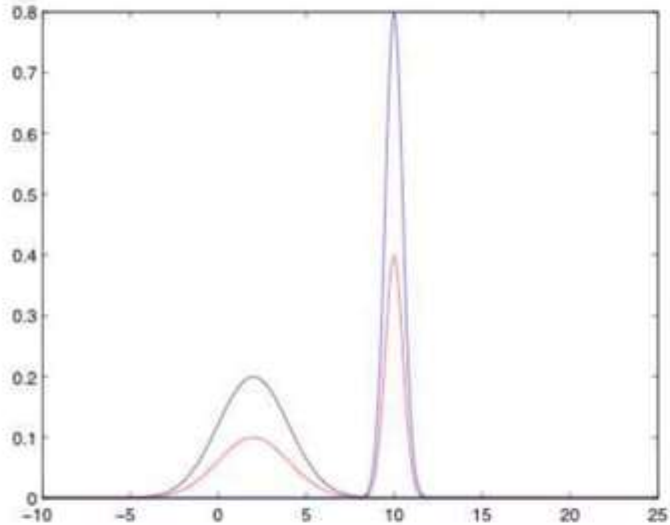
Gaussian Mixture Models and Expectation Maximization

Gaussian Mixture Models

- Rather than identifying clusters by "nearest" centroids
- Fit a Set of *k* Gaussians to the data
- Maximum Likelihood over a mixture model



GMM example



$$f_0(x) = N(x; 2, 2) \qquad f_1(x) = N(x; 10, .5)$$

$$\pi = [.5 \quad .5]^T$$

Mixture Models

- Formally a Mixture Model is the weighted sum of a number of pdfs where the weights are determined by a distribution, π

$$p(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x) + \ldots + \pi_k f_k(x)$$

$$\text{where } \sum_{i=0}^{k} \pi_i = 1$$

$$\boxed{p(x) = \sum_{i=0}^{k} \pi_i f_i(x)}$$

Gaussian Mixture Models

- GMM: the weighted sum of a number of Gaussians where the weights are determined by a distribution, π

$$p(x) = \pi_0 N(x|\mu_0, \Sigma_0) + \pi_1 N(x|\mu_1, \Sigma_1) + \ldots + \pi_k N(x|\mu_k, \Sigma_k)$$

$$\text{where } \sum_{i=0}^{k} \pi_i = 1$$

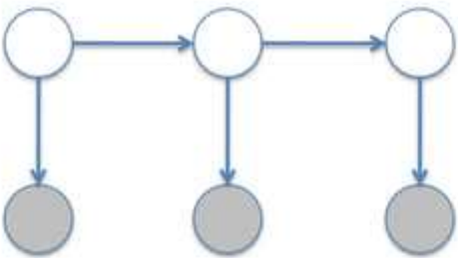
$$\boxed{p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)}$$

Graphical Models with unobserved variables

- What if you have variables in a Graphical model that are **never** observed?
  - Latent Variables
- Training latent variable models is an unsupervised learning application

Latent Variable HMMs

- We can cluster sequences using an HMM with unobserved state variables



- We will train latent variable models using Expectation Maximization

Expectation Maximization

- Both the training of GMMs and Graphical Models with latent variables can be accomplished using Expectation Maximization
  - Step 1: Expectation (E-step)
    - Evaluate the "responsibilities" of each cluster with the current parameters
  - Step 2: Maximization (M-step)
    - Re-estimate parameters using the existing "responsibilities"
- Similar to k-means training.

Latent Variable Representation

- We can represent a GMM involving a latent variable

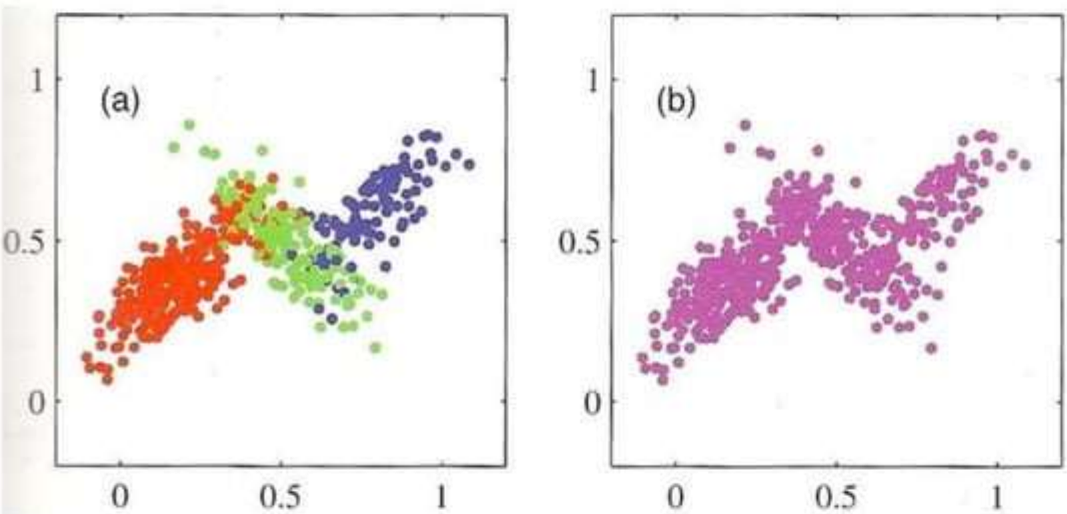
$$p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k) = \sum_{z} p(z)p(x|z)$$

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$$

- What does this give us?

TODO: plate notation

GMM data and Latent variables



One last bit

- We have representations of the joint p(x,z) and the marginal, p(x)…
- The conditional of p(z|x) can be derived using Bayes rule.
  - The **responsibility** that a mixture component takes for explaining an observation x.

$$\tau(z_k) = p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$

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

Maximum Likelihood over a GMM

- As usual: Identify a likelihood function

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

- And set partials to zero…

Maximum Likelihood of a GMM

- Optimization of means.

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

$$\frac{\partial \ln p(x|\pi, \mu, \Sigma)}{\partial \mu_k} = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j} \Sigma_k^{-1}(x_k - \mu_k) = 0$$

$$= \sum_{n=1}^{N} \tau(z_{nk}) \Sigma_k^{-1}(x_k - \mu_k) = 0$$

$$\boxed{\mu_k = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{\sum_{n=1}^{N} \tau(z_{nk})}}$$

Maximum Likelihood of a GMM

- Optimization of covariance

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

$$\Sigma_k = \frac{1}{\sum_{n=1}^{N} \tau(z_{nk})} \sum_{n=1}^{N} \tau(z_{nk})(x_k - \mu_k)(x_k - \mu_k)^T$$

- Note the similarity to the regular MLE without **responsibility terms**.

Maximum Likelihood of a GMM

- Optimization of mixing term

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

$$0 = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j, \Sigma_j)} + \lambda$$

$$\boxed{\pi_k = \frac{\sum_{n=1}^{N} \tau(z_n k)}{N}}$$

MLE of a GMM

$$\mu_k = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{N_k}$$

$$\Sigma_k \quad = \quad \frac{1}{N_k} \sum_{n=1}^{N} \tau(z_{nk})(x_k - \mu_k)(x_k - \mu_k)^T$$

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

$$N_k = \sum_{n=1}^{N} \tau(z_n k)$$

EM for GMMs

- Initialize the parameters
  - Evaluate the log likelihood
- Expectation-step: Evaluate the responsibilities
- Maximization-step: Re-estimate Parameters
  - Evaluate the log likelihood
  - Check for convergence

EM for GMMs

- E-step: Evaluate the Responsibilities

$$\tau(z_{nk}) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j, \Sigma_j)}$$

EM for GMMs

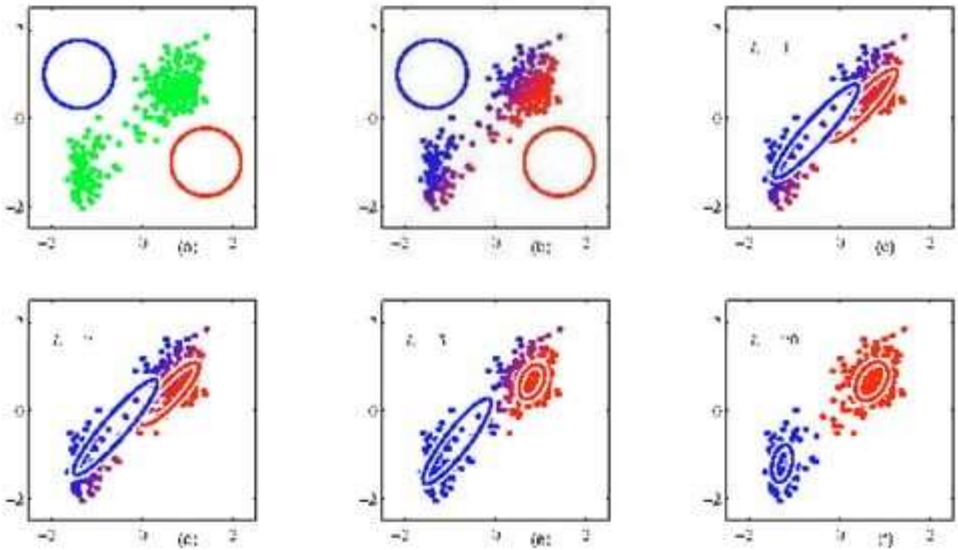
- M-Step: Re-estimate Parameters

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{N_k}$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \tau(z_{nk})(x_k - \mu_k^{new})(x_k - \mu_k^{new})^T$$

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

Visual example of EM

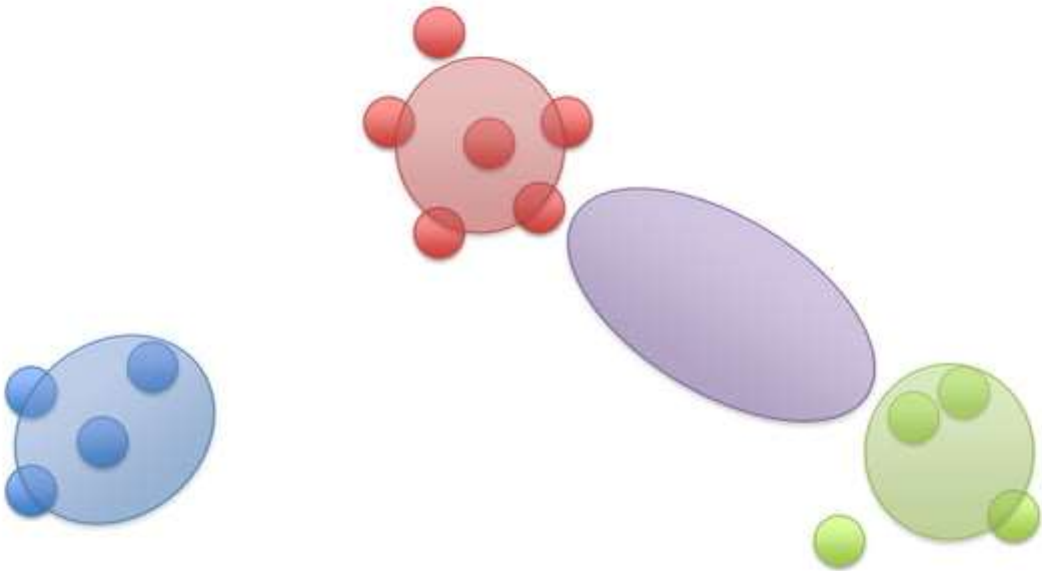


Potential Problems

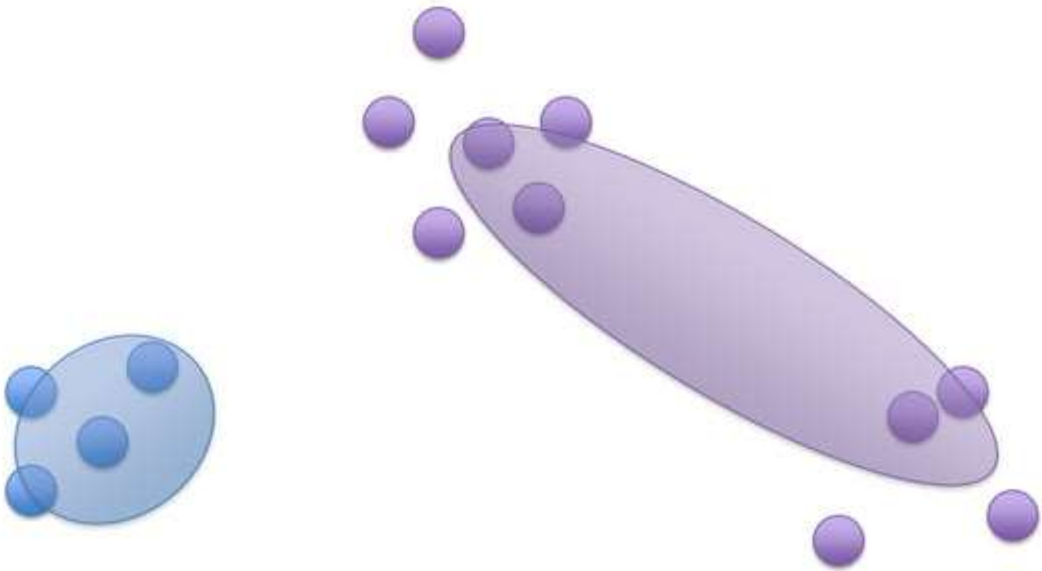
- Incorrect number of Mixture Components

- Singularities

Incorrect Number of Gaussians



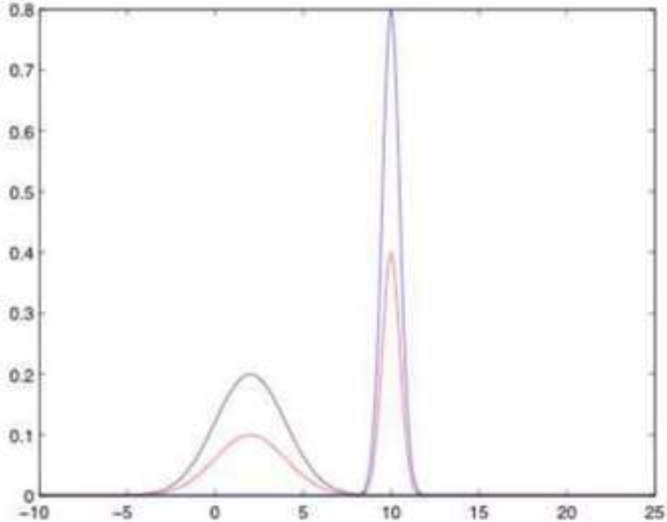
Incorrect Number of Gaussians



Singularities

- A minority of the data can have a disproportionate effect on the model likelihood.
- For example…

GMM example



$$f_0(x) = N(x; 2, 2) \qquad f_1(x) = N(x; 10, .5)$$

$$\pi = [.5 \quad .5]^T$$

Singularities

- When a mixture component collapses on a given point, the mean becomes the point, and the variance goes to zero.
- Consider the likelihood function as the covariance goes to zero.

$$N(x_n|x_n, \sigma^2 I) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_j}$$

- The likelihood approaches infinity.

$$\boxed{p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)}$$

Relationship to K-means

- K-means makes **hard** decisions.
  - Each data point gets assigned to a single cluster.
- GMM/EM makes **soft** decisions.
  - Each data point can yield a posterior $p(z|x)$
- Soft K-means is a special case of EM.

Soft means as GMM/EM

- Assume equal covariance matrices for every mixture component: $\epsilon\mathbf{I}$
- Likelihood:

$$p(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{M/2}} \exp\left\{-\frac{1}{2\epsilon}\|x-\mu_k\|^2\right\}$$

- Responsibilities:

$$\tau(z_{nk}) = \frac{\pi_k \exp\left\{-\|x_n-\mu_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|x_n-\mu_j\|^2/2\epsilon\right\}}$$

- As epsilon approaches zero, the responsibility approaches unity.

Soft K-Means as GMM/EM

- Overall Log likelihood as epsilon approaches zero:

$$\mathbb{E}_z[\ln p(X, Z|\mu, \Sigma, \pi)] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 + const.$$

- The expectation of soft k-means is the intercluster variability
- Note: only the means are reestimated in Soft K-means.
  - The covariance matrices are all tied.

General form of EM

- Given a joint distribution over observed and latent variables: $p(X, Z|\theta)$
- Want to maximize: $p(X|\theta)$

1. Initialize parameters θ^{old}
2. E Step: Evaluate:

$$p(Z|X, \theta^{old})$$

3. M-Step: Re-estimate parameters (based on expectation of complete-data log likelihood)

$$\theta^{new} = \text{argmax}_\theta \sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta)$$

4. Check for convergence of params or likelihood

Next Time

- Homework 4 due…
- Proof of Expectation Maximization in GMMs
- Generalized EM – Hidden Markov Models