

Mixture of Gaussians

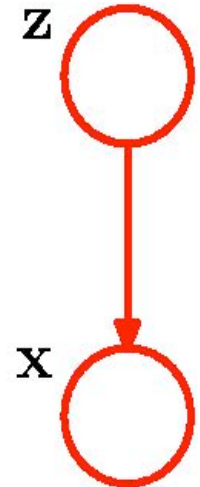
Gaussian Mixture Model

- A simple 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

GMM Formulation

- Linear superposition of K Gaussians:

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



- Introduce 1 -of- K representation

- with z whose elements are $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$
- Thus there are K possible states of z

- Define joint distribution $p(\mathbf{x}, \mathbf{z}) = \underbrace{p(\mathbf{x} | \mathbf{z})}_{\text{conditional}} \underbrace{p(\mathbf{z})}_{\text{marginal}}$

Properties of marginal distribution

- Denote $p(z_k=1)=\pi_k$
where parameters $\{\pi_k\}$ satisfy $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$
- Because z uses 1-of- K it follows that

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

- since components are mutually exclusive and hence are independent

Conditional distribution

- For a particular value of \mathbf{z}

$$p(\mathbf{x}|z_k=1) = N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Which can be written in the form

$$p(\mathbf{x} | \mathbf{z}) = \prod_{k=1}^K N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

- Thus marginal distribution of \mathbf{x} is obtained by summing over all possible states of \mathbf{z}

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x} | \mathbf{z}) = \sum_{k=1}^K \pi_k N(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- This is the standard form of a Gaussian mixture

Latent Variable

- If we have observations $\mathbf{x}_1, \dots, \mathbf{x}_N$
- Because marginal distribution is in the form $p(\mathbf{x}) = \sum_k p(\mathbf{x}, z_k)$
 - It follows that for every observed data point \mathbf{x}_n there is a corresponding latent vector z_n
- Thus we have found a formulation of Gaussian mixture involving an explicit latent variable
 - We are now able to work with joint distribution $p(\mathbf{x}, z)$ instead of marginal $p(\mathbf{x})$
- Leads to significant simplification through introduction of expectation maximization

Another conditional probability (Responsibility)

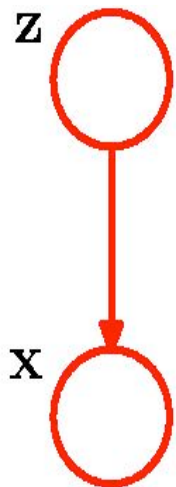
- In EM $p(z | \mathbf{x})$ plays a role
- The probability $p(z_k = 1 | \mathbf{x})$ is denoted $\gamma(z_k)$
- From Bayes theorem

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

- We view π_k as the prior probability of $z_k = 1$ and $\gamma(z_k)$ as the posterior probability after observing \mathbf{x}
 $\gamma(z_k)$ is also the responsibility that component k takes for explaining the observation \mathbf{x}

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

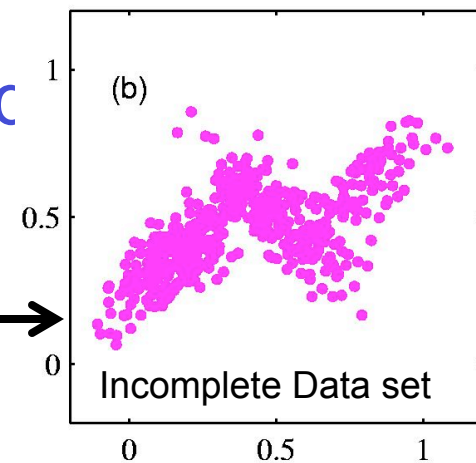
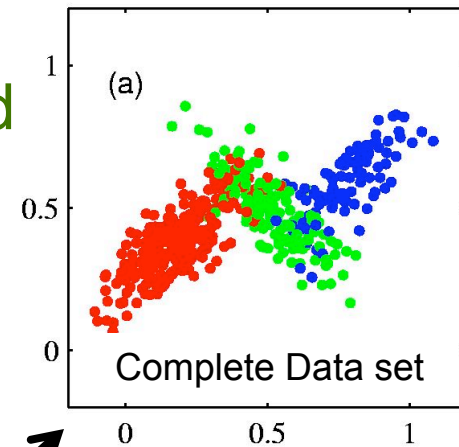
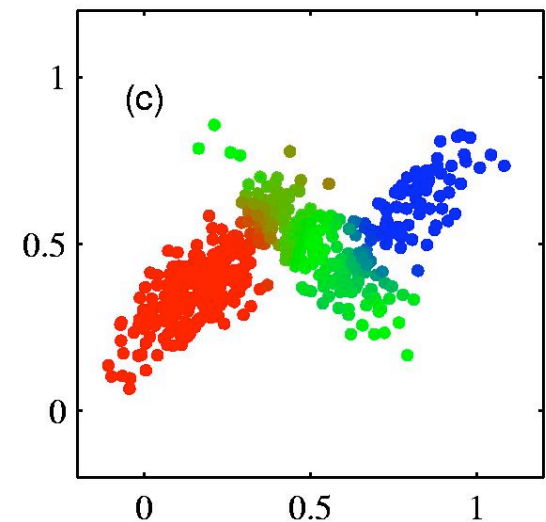


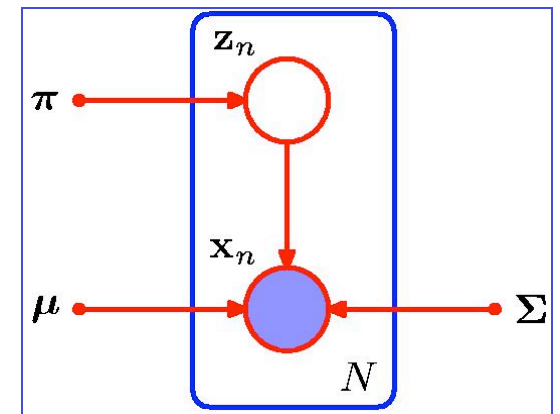
Illustration of responsibilities

- 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 $\gamma(z_{n1})=1$ it is colored red
 - If $\gamma(z_{n2})=\gamma(z_{n3})=0.5$ it is cyan



Maximum Likelihood for GMM

- We wish to model data set $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ using a mixture of Gaussians
- Represent by $N \times D$ matrix X
 - N^{th} row is given by \mathbf{x}_n^T
- Represent latent variables with $N \times K$ matrix Z with rows \mathbf{z}_n^T
- Graphical representation is \longrightarrow



Likelihood Function for GMM

Mixture density function is

$$p(\mathbf{x}) = \sum_z p(z)p(\mathbf{x} | z) = \sum_{k=1}^K \pi_k N(\mathbf{x} | \mu_k, \Sigma_k)$$

Therefore Likelihood function is

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

Therefore log-likelihood function is

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

Which we wish to maximize

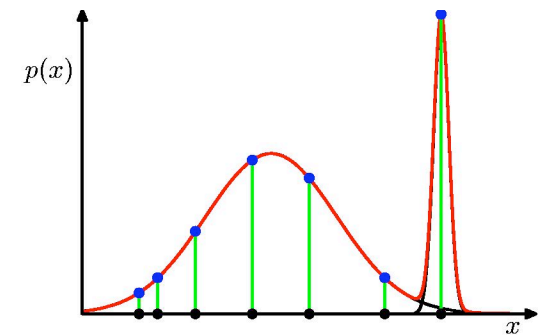
A more difficult problem than for a single Gaussian

Singularities with Gaussian mixtures

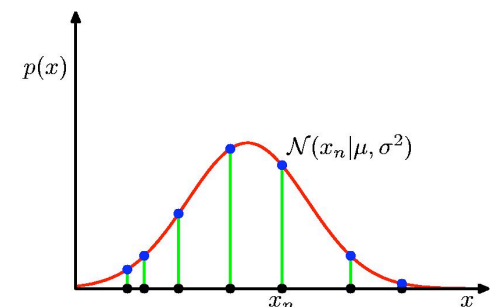
- Consider Gaussian mixture
 - components with covariance matrices $\Sigma_k = \sigma_k^2 I$
- Data point that falls on a mean $\mathbf{x}_n = \mu_j$ will contribute to the likelihood function

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

- As $\sigma_j \rightarrow 0$ term goes to infinity
- Therefore maximization of log-likelihood is not well-posed
 - Interestingly, this does not happen in the case of a single Gaussian
 - Multiplicative factors go to zero
 - Does not happen in the Bayesian approach
- Problem is avoided using heuristics
 - Resetting mean or covariance



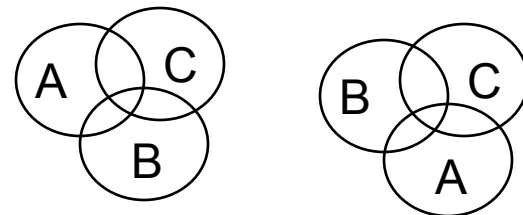
One component assigns finite values and other to large value



Multiplicative values Take it to zero

Identifiability

- For any given m.l.e. solution
- 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
- 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



EM for Gaussian Mixtures

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

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

- 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

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

- 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

$$0 = \sum_{n=1}^N \frac{\pi_k N(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_j \pi_j N(\mathbf{x}_n | \mu_j, \Sigma_j)} \sum_k^{-1} (\mathbf{x}_n - \mu_k)$$

Inverse of covariance matrix

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

M.L.E. solution for Means

- Multiplying by Σ_k (assuming non-singularity)

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

Mean of k^{th} Gaussian component is the weighted mean of all the points in the data set:

where data point \mathbf{x}_n is weighted by the posterior probability that component k was responsible for generating \mathbf{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

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

- 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 $\ln p(X|\pi, \mu, \Sigma)$ w.r.t. π_k
 - Must take into account that mixing coefficients sum to one
 - Achieved using Lagrange multiplier and maximizing
$$\ln p(X|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$
 - Setting derivative wrt π_k to zero and solving gives

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

EM Formulation

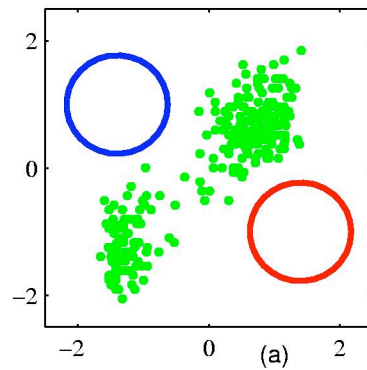
- The results for μ_k , Σ_k and π_k are not closed form solutions for the parameters
 - Since $\gamma(z_{nk})$ depend on those parameters in a complex way
- Results suggest an iterative solution
- An instance of EM algorithm for the particular case of GMM

Informal EM for GMM

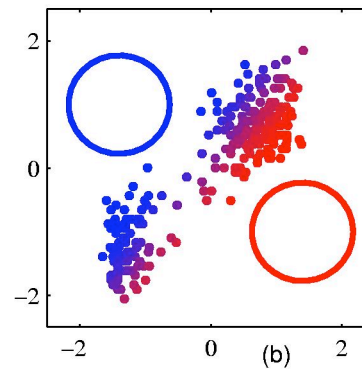
- First choose initial values for means, covariances and mixing coefficients
- Alternate between following two updates
 - Called E step and M step
- In E step use current value of parameters to evaluate posterior probabilities, or responsibilities
- In the M step use these probabilities to re-estimate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

EM using Old Faithful

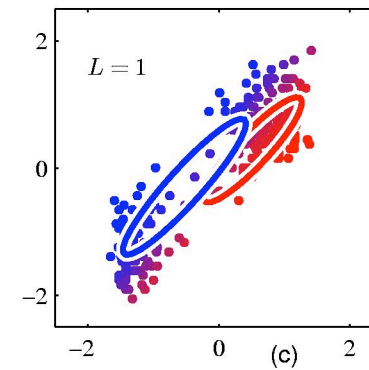
Data points and Initial mixture model



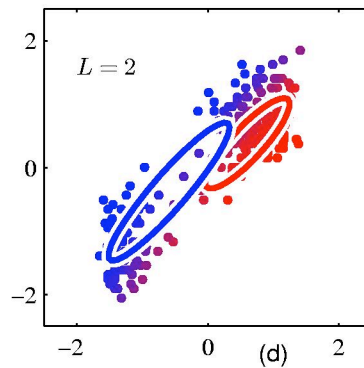
Initial E step



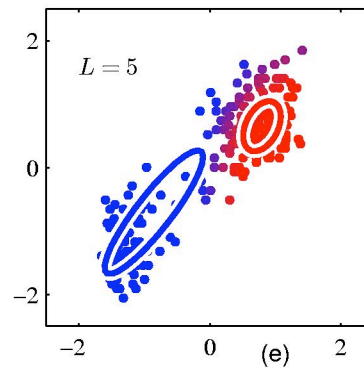
After first M step



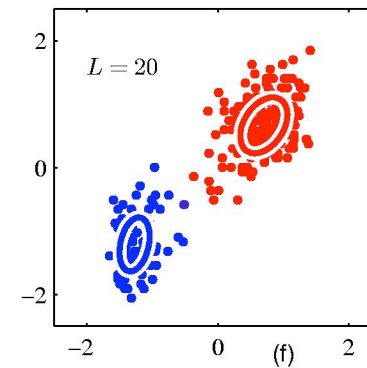
After 2 cycles



After 5 cycles



After 20 cycles



Practical Issues with EM

- Takes many more iterations than K-means
- Each cycles 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)

Step 1: 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

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

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

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

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

$$\pi_k^{\text{new}} = \frac{N_k}{N} \quad \text{where} \quad N_k = \sum_{n=1}^N \gamma(z_{nk})$$

EM Continued

- Step 4: Evaluate the log likelihood

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

- And check for convergence of either parameters or log likelihood
- If convergence not satisfied return to Step 2