# Mixture of Gaussians

Sargur Srihari
srihari@cedar.buffalo.edu

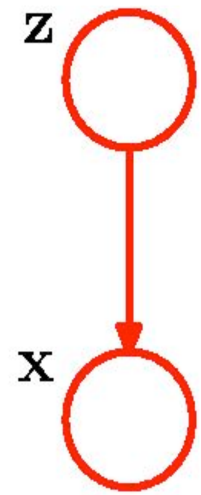
# Gaussian Mixture Model

- 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:

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




- Introduce $1$-of-$K$ representation
  - with $\mathrm{z}$ whose elements are $z_k\ \varepsilon\ \{0,1\}$ and $\Sigma_k z_k = 1$
  - Thus there are $K$ possible states of $z$
- Define joint distribution $p(\mathrm{x},\mathrm{z}) = \underbrace{p(\mathrm{x}|\mathrm{z})}_{\text{conditional}}\underbrace{p(\mathrm{z})}_{\text{marginal}}$

# Properties of marginal distribution

- Denote $p(z_k=1)=\pi_k$

  where parameters $\{\pi_k\}$ satisfy $0 \le \pi_k \le 1$ and $\Sigma_k \pi_k = 1$

- Because $z$ uses *1-of-K* it follows that

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

  – since components are mutually exclusive and hence are independent

# Conditional distribution

- For a particular value of $\mathbf{z}$

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

- Which can be written in the form

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

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

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

- This is the standard form of a Gaussian mixture

# Latent Variable

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

# Another conditional probability (Responsibility)

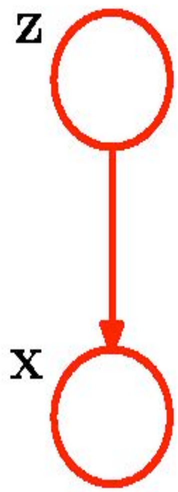
- In EM $p(\mathrm{z}\,|\mathrm{x})$ plays a role
- The probability $p(z_k=1|\mathrm{x})$ is denoted $\gamma(z_k)$
- From Bayes theorem

$$\boxed{\begin{aligned}\gamma(\mathrm{z}_k) \equiv p(\mathrm{z}_k = 1 \mid \mathrm{x}) &= \frac{p(\mathrm{z}_k = 1)\, p(\mathrm{x} \mid \mathrm{z}_k = 1)}{\sum_{\mathrm{j=1}}^{\mathrm{K}} p(\mathrm{z}_j = 1)\, p(\mathrm{x} \mid \mathrm{z}_j = 1)} \\ &= \frac{\pi_{\mathrm{k}} N(\mathrm{x} \mid \mu_k, \Sigma_k)}{\sum_{\mathrm{j=1}}^{\mathrm{K}} \pi_{\mathrm{j}} N(\mathrm{x} \mid \mu_k, \Sigma_j)}\end{aligned}}$$

- We view $\pi_k$ as the prior probability of $z_k=1$ and $\gamma(z_k)$ as the posterior probability after observing $\mathrm{x}$

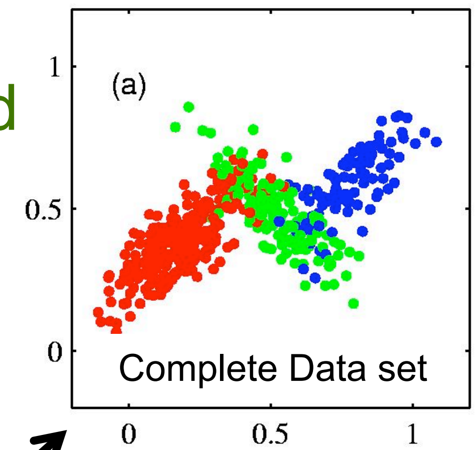
  $\gamma(z_k)$ is also the responsibility that component $k$ takes for explaining the observation $\mathrm{x}$

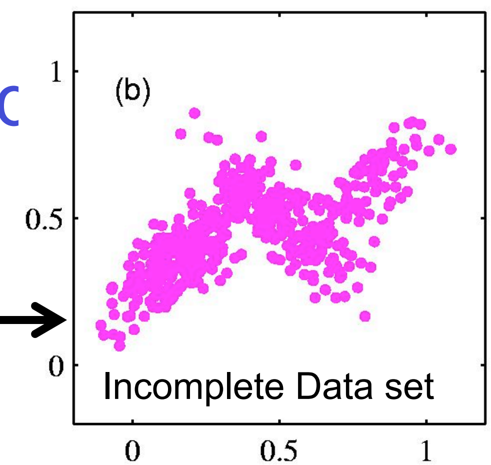
# Synthesizing data from mixture




- 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 $z$

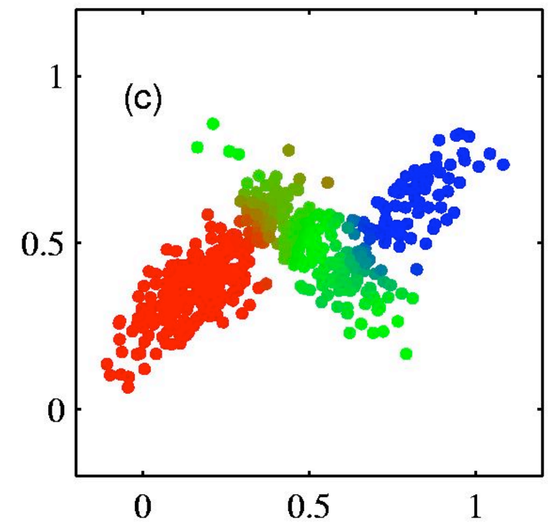
500 points from three Gaussians



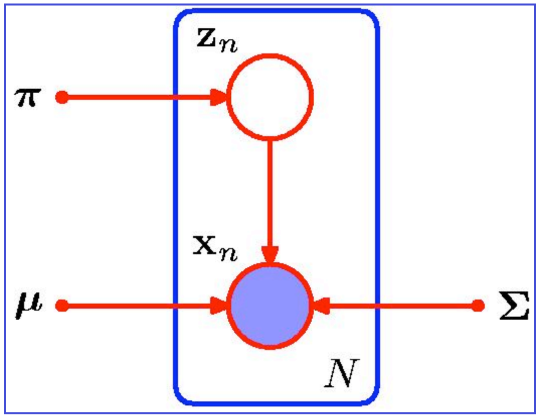
# Illustration of responsibilities

- 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 $\{x_1,..x_N\}$ using a mixture of Gaussians
- Represent by $N$ x $D$ matrix $X$
  - $N^{th}$ row is given by $x_n^T$
- Represent latent variables with $N$ x $K$ matrix $Z$ with rows $z_n^T$
- Graphical representation is →



# Likelihood Function for GMM

Mixture density function is

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

Therefore Likelihood function is

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

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

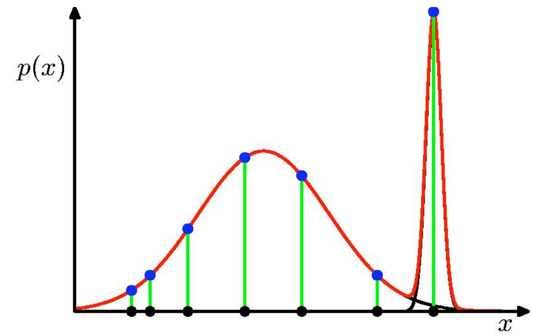
A more difficult problem than for a single Gaussian

# Singularities with Gaussian mixtures

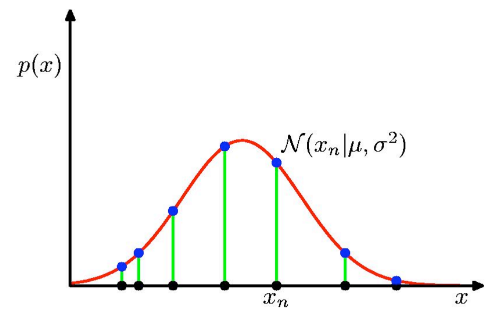
- Consider Gaussian mixture
  - components with covariance matrices $\Sigma_k=\sigma_k^2 I$
- Data point that falls on a mean $x_n=\mu_j$ will contribute to the likelihood function

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

- 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



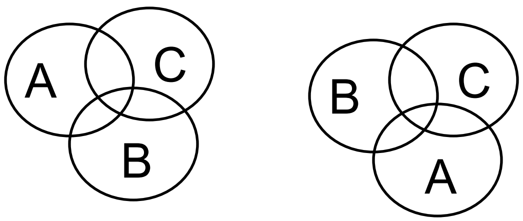

One component assigns finite values and other to large value




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

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

- We wish to find $\pi, \mu, \Sigma$ that maximize this quantity
- Take derivatives in turn w.r.t
  - means $\mu_k$ and set to zero
  - covariance matrices $\Sigma_k$ and set to zero
  - mixing coefficients $\pi_k$ and set to zero

# EM for GMM: Derivative wrt $\mu_k$

- Begin with log-likelihood function

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

- Take derivative w.r.t the means $\mu_k$ 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

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

Inverse of covariance matrix

# M.L.E. solution for Means

- Multiplying by $\Sigma_k$ (assuming non-singularity)

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

Mean of $k^{\text{th}}$ Gaussian component is the weighted mean of <u>all</u> the points in the data set:

where data point $\mathrm{x_n}$ is weighted by the posterior probability that component $\mathrm{k}$ was responsible for generating $\mathrm{x_n}$

- 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 $\Sigma_k$ 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 $\ln p(\mathbf{X}|\pi,\mu,\Sigma)$ w.r.t. $\pi_k$
  - 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 $\pi_k$ to zero and solving gives

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

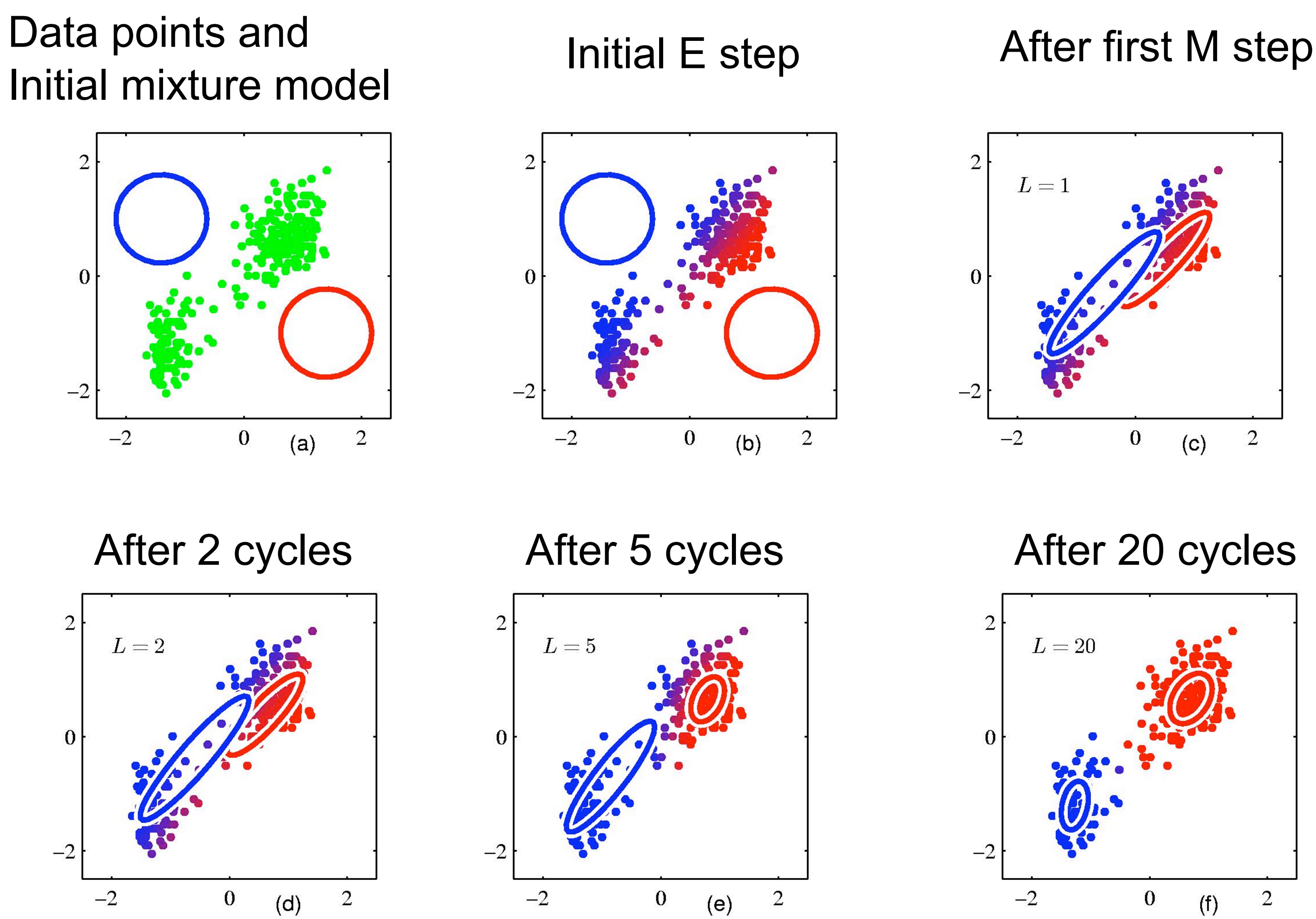
# EM Formulation

- The results for $\mu_k$, $\Sigma_k$ and $\pi_k$ 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 to re-estimate means, covariances and mixing coefficients

# EM using Old Faithful



# 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)

Step1: Initialize the means $\mu_k$, covariances $\Sigma_k$ and mixing coefficients $\pi_k$ and evaluate initial value of log-likelihood

# EM continued

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

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

- 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