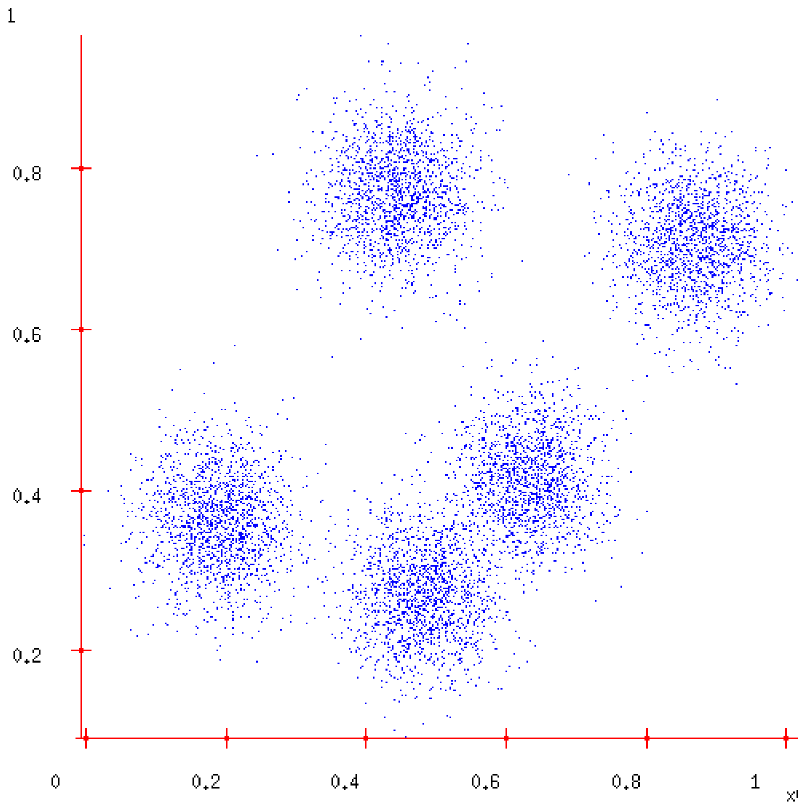
Outline

- K-Means
- Model-based Clustering (GMM and Expectation Maximization)
- Hierarchical Clustering
- Evaluation of Clustering Algorithms

Model-based Clustering: Gaussian Mixture Model

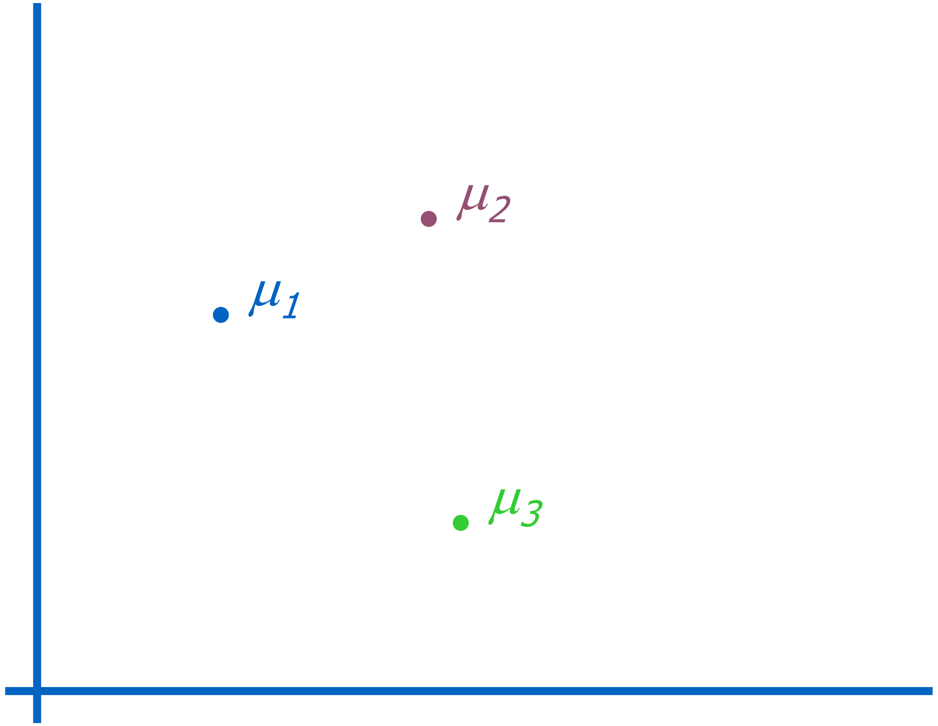
- Density estimation with multimodal/clumpy data



Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

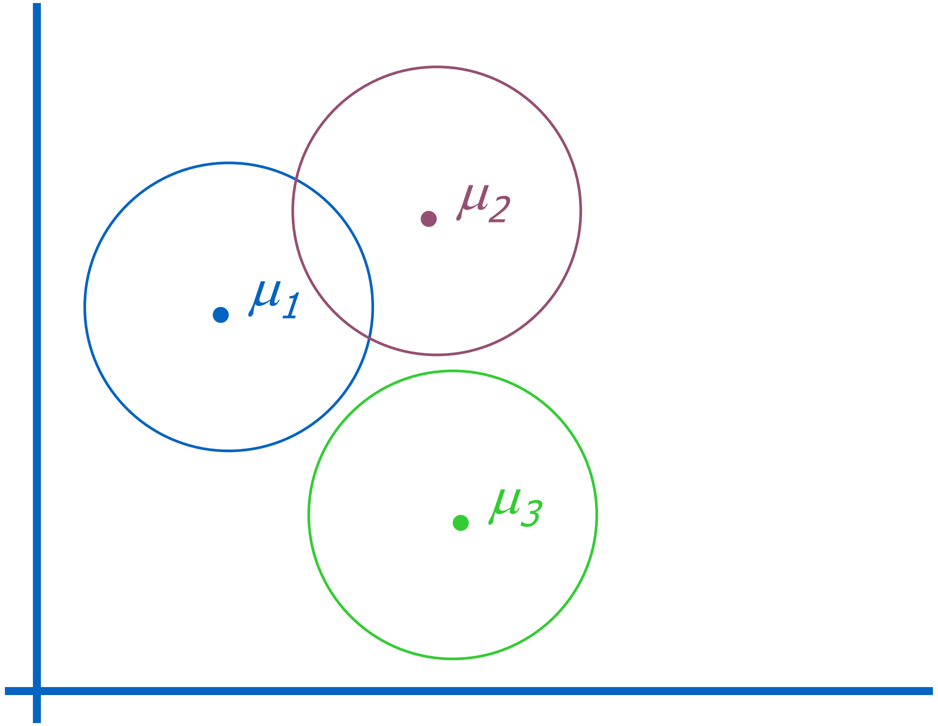
- The GMM assumption
- There are k components. The i[th] component is called ω_i
- Component ω_i has an associated mean vector μ_i



Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

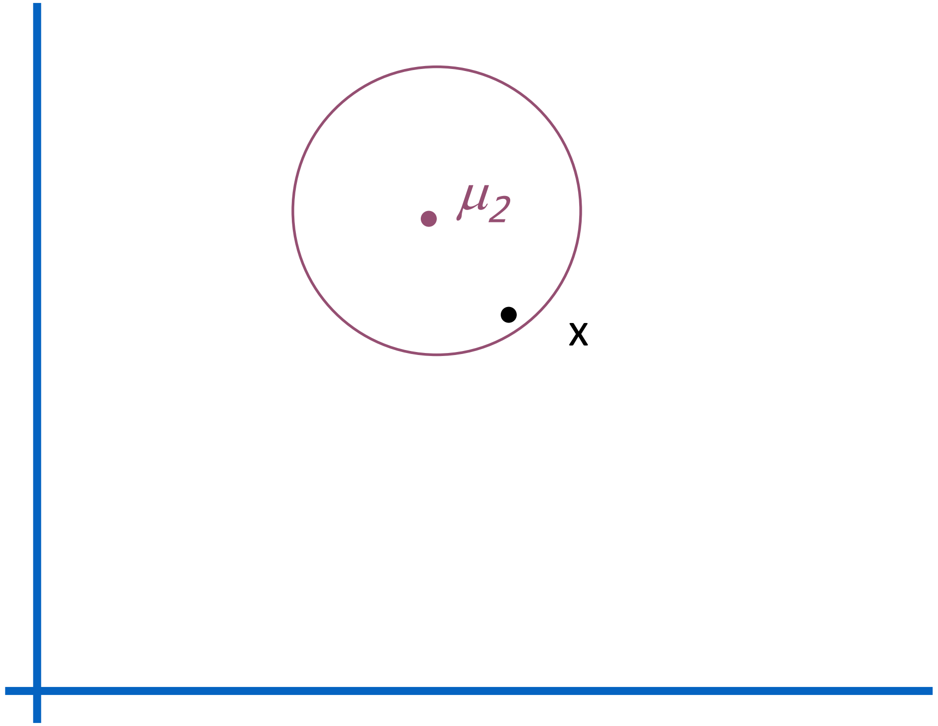
- The GMM assumption
- There are k components. The i^{th} component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 \boldsymbol{I}$



Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

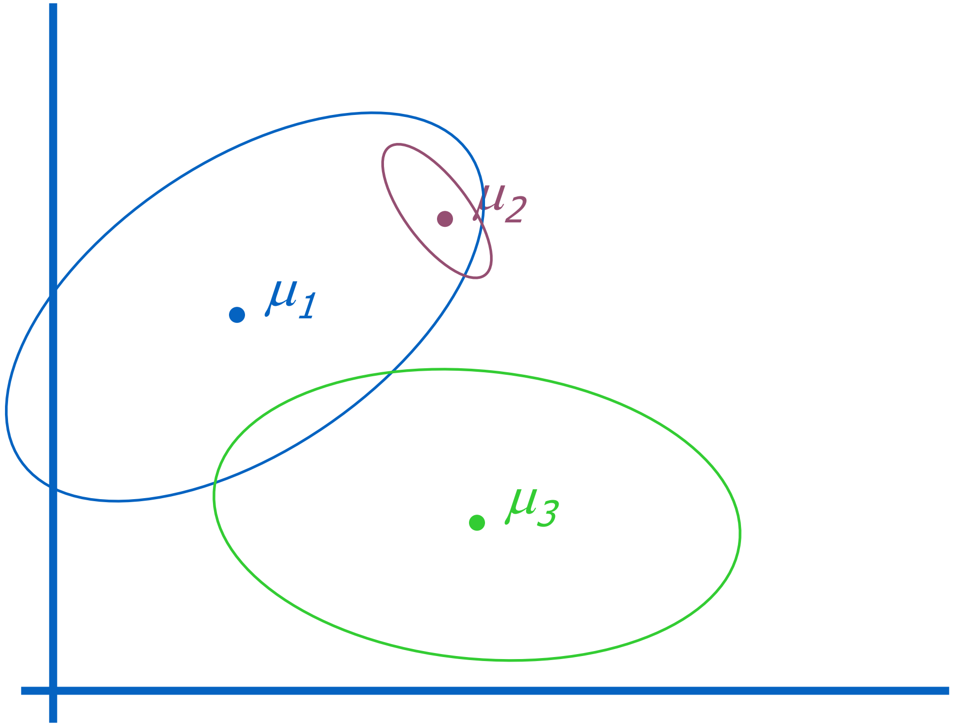
- The GMM assumption
- There are k components. The i[th] component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 \mathbf{I}$
- Assume that each datapoint is generated according to the following recipe:
  - Pick a component at random. Choose component i with probability $P(\omega_i)$.
  - Datapoint ~ N(μ_i, Σ_i)



Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

- The GMM assumption
- There are k components. The i^{th} component is called ω_i
- Component ω_i has an associated mean vector μ_i
- Each component generates data from a Gaussian with mean μ_i and covariance matrix $\sigma^2 \boldsymbol{I}$
- Assume that each datapoint is generated according to the following recipe:
  - Pick a component at random. Choose component i with probability $P(\omega_i)$.
  - Datapoint ~ $N(\mu_i, \Sigma_i)$



Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

- Given the means and σ^2, we can compute P(data | $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2..\boldsymbol{\mu}_k,\ \sigma^2$). How do we find the $\boldsymbol{\mu}_i$ s and σ^2 which give max likelihood?
- The normal max likelihood trick:

  Set $\dfrac{d}{d\mu_i}$ log Prob (….) = 0

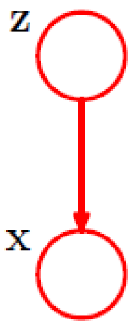
  and solve for μ_i's.
- Use gradient descent

Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

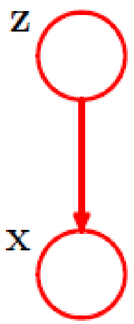
$$p(z_k = 1) = \pi_k \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

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



Gaussian Mixture Model (GMM)

$$p(z_k = 1) = \pi_k \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$



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

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

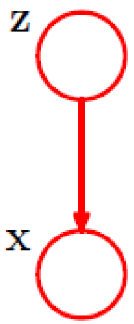
$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

Gaussian Mixture Model

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

Gaussian Mixture Model (GMM)

$$p(z_k = 1) = \pi_k \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

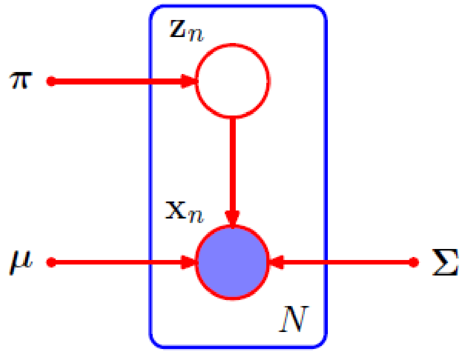


$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

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

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

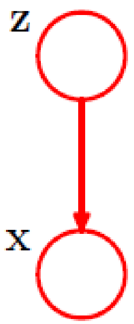
Gaussian Mixture Model

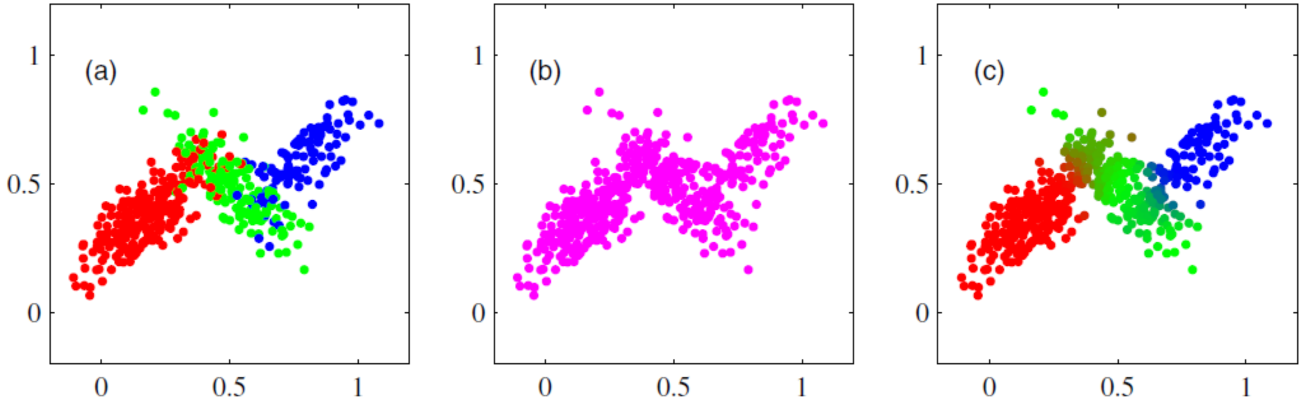


$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

Gaussian Mixture Model (GMM)

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

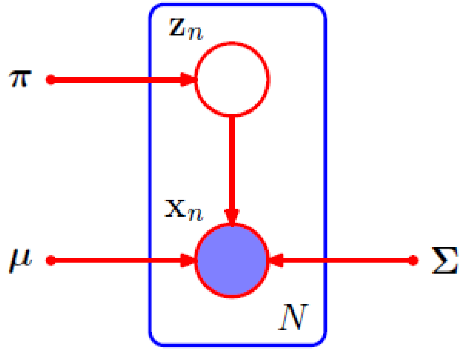


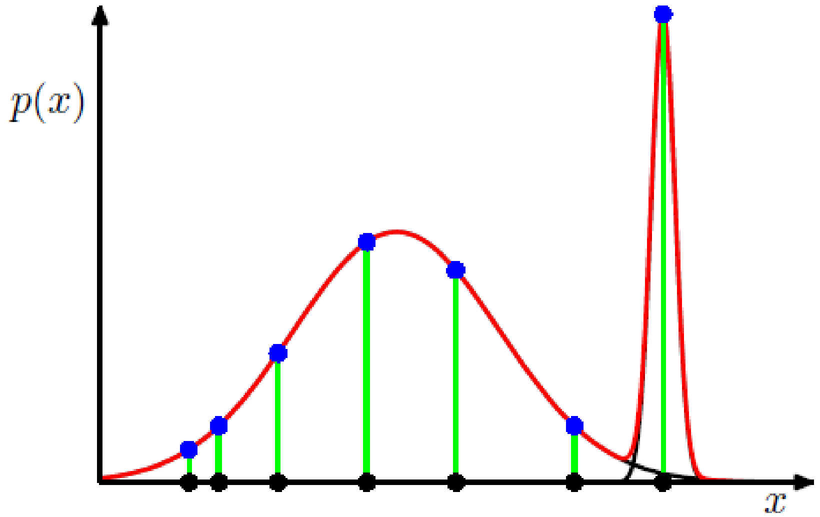


Gaussian Mixture Model (GMM)

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

No closed form solution





Expectation Maximization (EM)

- We'll get back to unsupervised learning/clustering/GMM soon.
- The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.
- They pointed out that the method had been "proposed many times in special circumstances" by earlier authors.
- EM is typically used to compute maximum likelihood estimates given incomplete samples.
  - An excellent way of doing our unsupervised learning problem, as we'll see
  - Many, many other uses, including inference of Hidden Markov Models
- The EM algorithm estimates the parameters of a model iteratively. Starting from some initial guess, each iteration consists of
  - an E step (Expectation step)
  - an M step (Maximization step)

Slide Courtesy: Andrew Moore, CMU

EM: Trivial Example

Let events be "grades in a class"

w_1 = Gets an A $\quad P(A) = \frac{1}{2}$

w_2 = Gets a B $\quad P(B) = \mu$

w_3 = Gets a C $\quad P(C) = 2\mu$

w_4 = Gets a D $\quad P(D) = \frac{1}{2}-3\mu$

(Note $0 \leq \mu \leq 1/6$)

Assume we want to estimate μ from data. In a given class, there were

a A's
b B's
c C's
d D's

What's the maximum likelihood estimate of μ given a,b,c,d ?

Slide Courtesy: Andrew Moore, CMU

EM: Trivial Example

P(A) = ½ P(B) = μ P(C) = 2μ P(D) = ½-3μ

$P(a,b,c,d \mid \mu) = (\tfrac{1}{2})^a (\mu)^b (2\mu)^c (\tfrac{1}{2}-3\mu)^d$

$\log P(a,b,c,d \mid \mu) = a\log \tfrac{1}{2} + b\log \mu + c\log 2\mu + d\log (\tfrac{1}{2}-3\mu)$

$$\text{FOR MAX LIKE } \mu, \text{ SET } \frac{\partial \text{LogP}}{\partial \mu} = 0$$

$$\frac{\partial \text{LogP}}{\partial \mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{1/2 - 3\mu} = 0$$

$$\text{Gives max like } \mu = \frac{b+c}{6(b+c+d)}$$

So if class got

A	B	C	D
14	6	9	10

$$\text{Max likelihood estimate}: \mu = \frac{1}{10}$$

Slide Courtesy: Andrew Moore, CMU

EM: Same Example with Hidden Info

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max likelihood estimate of μ now?

REMEMBER

$P(A) = \frac{1}{2}$

$P(B) = \mu$

$P(C) = 2\mu$

$P(D) = \frac{1}{2}-3\mu$

Slide Courtesy: Andrew Moore, CMU

EM: Same Example with Hidden Info

Someone tells us that

Number of High grades (A's + B's) = h

Number of C's = c

Number of D's = d

What is the max likelihood estimate of μ now?

We can answer this circularly as below

REMEMBER

P(A) = ½

P(B) = μ

P(C) = 2μ

P(D) = ½-3μ

EXPECTATION

If we know the value of μ we could compute the expected value of *a* and *b*

Since the ratio a:b should be the same as the ratio ½ : μ

$$a = \frac{1/2}{1/2 + \mu} h \qquad b = \frac{\mu}{1/2 + \mu} h$$

MAXIMIZATION

If we know the expected values of *a* and *b* we could compute the maximum likelihood value of μ

$$\mu = \frac{b + c}{6(b + c + d)}$$

Slide Courtesy: Andrew Moore, CMU

EM: Solution for Trivial Example

We begin with a guess for μ

We iterate between EXPECTATION and MAXIMIZATION to improve our estimates of μ and *a* and *b*.

Define μ(t) the estimate of μ on the t^th iteration

b(t) the estimate of *b* on t^th iteration

E-step

$$\mu(0) = \text{initial guess}$$

$$b(t) = \frac{\mu(t)h}{\frac{1}{2} + \mu(t)} = \mathrm{E}\left[b \mid \mu(t)\right]$$

M-step

$$\mu(t+1) = \frac{b(t)+c}{6\left(b(t)+c+d\right)}$$

$$= \text{max like est of } \mu \text{ given } b(t)$$

Continue iterating until converged.

Good news: Converging to local optimum is assured.

Bad news: "local" optimum.

Slide Courtesy: Andrew Moore, CMU

EM: Convergence

In our example, suppose we had

$h = 20$
$c = 10$
$d = 10$
$\mu(0) = 0$

t	μ(t)	b(t)
0	0	0
1	0.0833	2.857
2	0.0937	3.158
3	0.0947	3.185
4	0.0948	3.187
5	0.0948	3.187
6	0.0948	3.187

Slide Courtesy: Andrew Moore, CMU

Back to GMM

Given a training data set: X={x(1),x(2),…,x(n)}
Z={z(1),z(2),…,z(n)}
z(i) is the calss/group label of sample x(i).
As we are in Clustering setting,
X is Given and $\boxed{\text{Z is unknown}}$

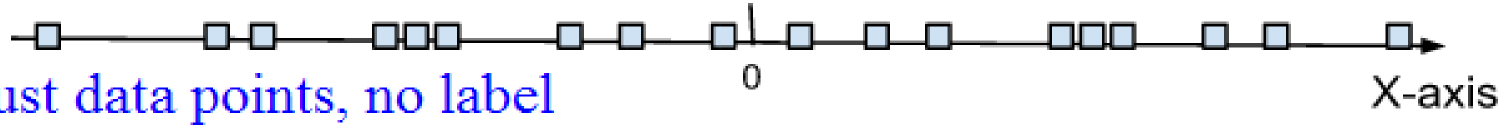
Now, we model the data by specifying a joint distribution p(x(i), z(i))=p(x(i)|z(i))p(z(i))

$$\boxed{\begin{aligned} &z_{(i)} \sim \text{Multinomial}(\phi) \\ &\qquad \phi_j \geq 0,\ \textstyle\sum_{j=1}^{k} \phi_j = 1 \\ &\qquad k = \#\ \text{of}\ z_{(i)}\text{'s values} \\ &\qquad \phi_j = p(z_{(i)} = j) \\ &x_{(i)} | z_{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j) \end{aligned}}$$

$\Rightarrow$

each $x_{(i)}$ was generated by randomly choosing $z_{(i)}$ from $\{1, \ldots, k\}$, and then $x_{(i)}$ was drawn from one of k Gaussians .

The parameters of our model are thus ϕ, μ and Σ.



Just data points, no label

Slide Courtesy: Andrew Moore, CMU

EM for GMM

Incomplete Data

X={x(1),x(2),…,x(n)} Given
Z={z(1),z(2),…,z(n)} unknown

What is the value of z(i)?

unknown

The parameters of our model ϕ, μ, Σ

We can answer this question circularly:

EXPECTATION

If we know the values of ϕ, μ, Σ we could compute the expected values of Z

MAXIMIZATION

If we know the expected values of Z we could compute the maximum likelihood value of ϕ, μ, Σ

We begin with a guess for ϕ, μ, Σ, and then iterate between EXPECTATION and MAXIMALIZATION to improve our estimates of ϕ, μ, Σ and Z

Continue iterating until converged.

Slide Courtesy: Andrew Moore, CMU

EM for GMM

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)} | z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi)$$

Maximizing this with respect to ϕ, μ and Σ gives the parameters:

$$\begin{aligned}
\phi_j &= \frac{1}{m} \sum_{i=1}^{m} 1\{z^{(i)} = j\}, \\
\mu_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}, \\
\Sigma_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}.
\end{aligned}$$

Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)

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

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

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

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

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

Gaussian Mixture Model (GMM)

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

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

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

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

EM for GMM

Repeat until convergence: {

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M-step) Update the parameters:

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)},$$

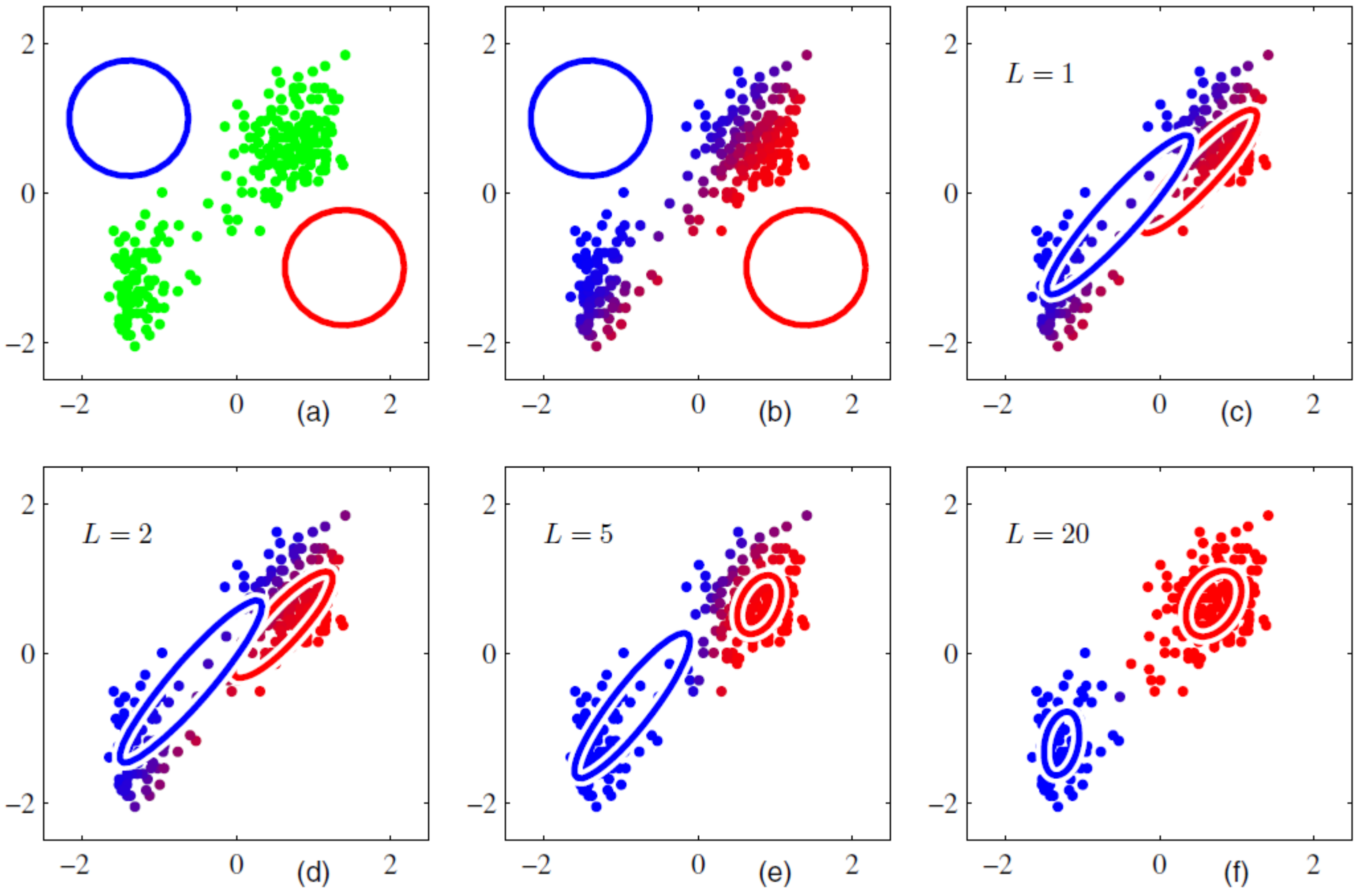
$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

}

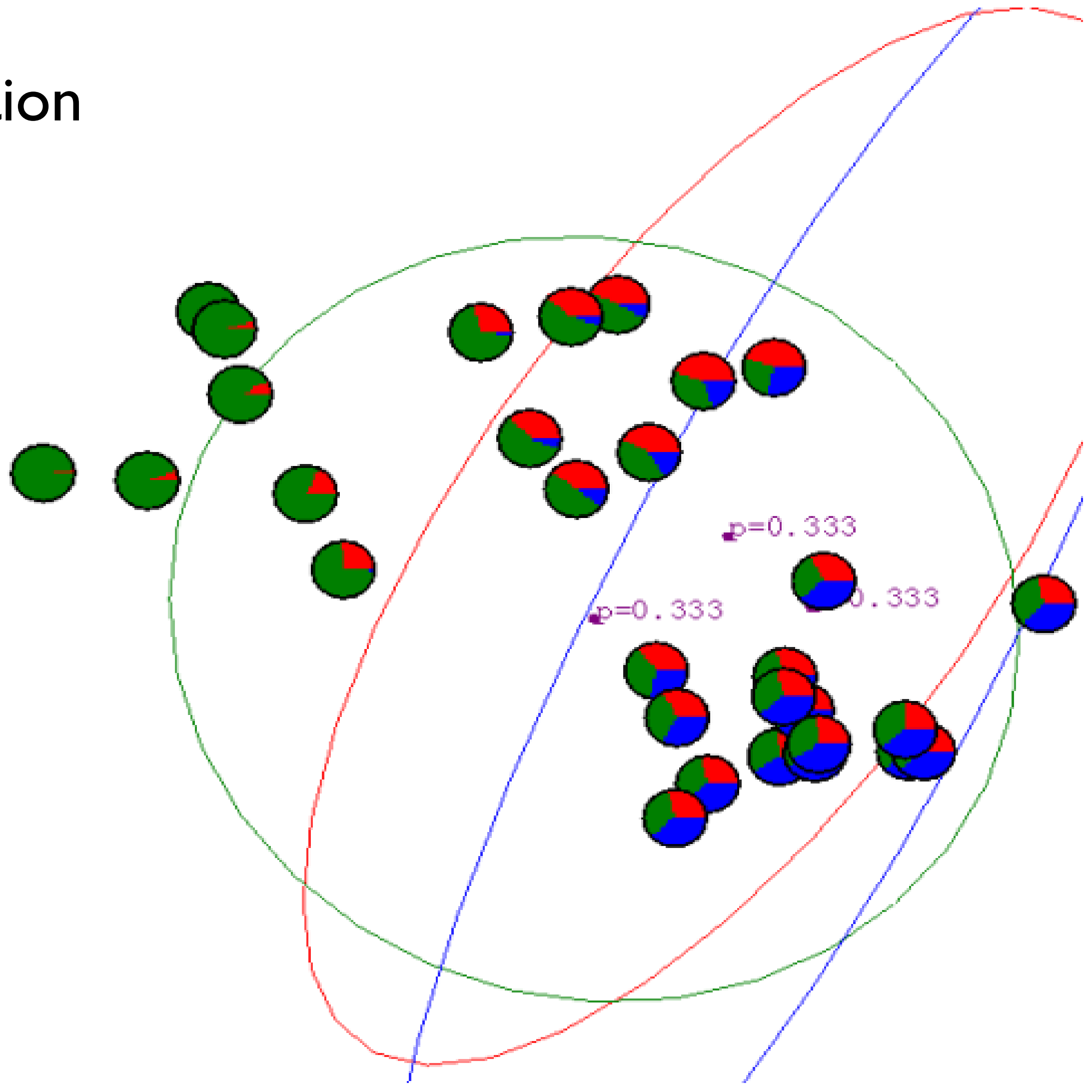
Slide Courtesy: Andrew Moore, CMU

Gaussian Mixture Model (GMM)



GMM: Example

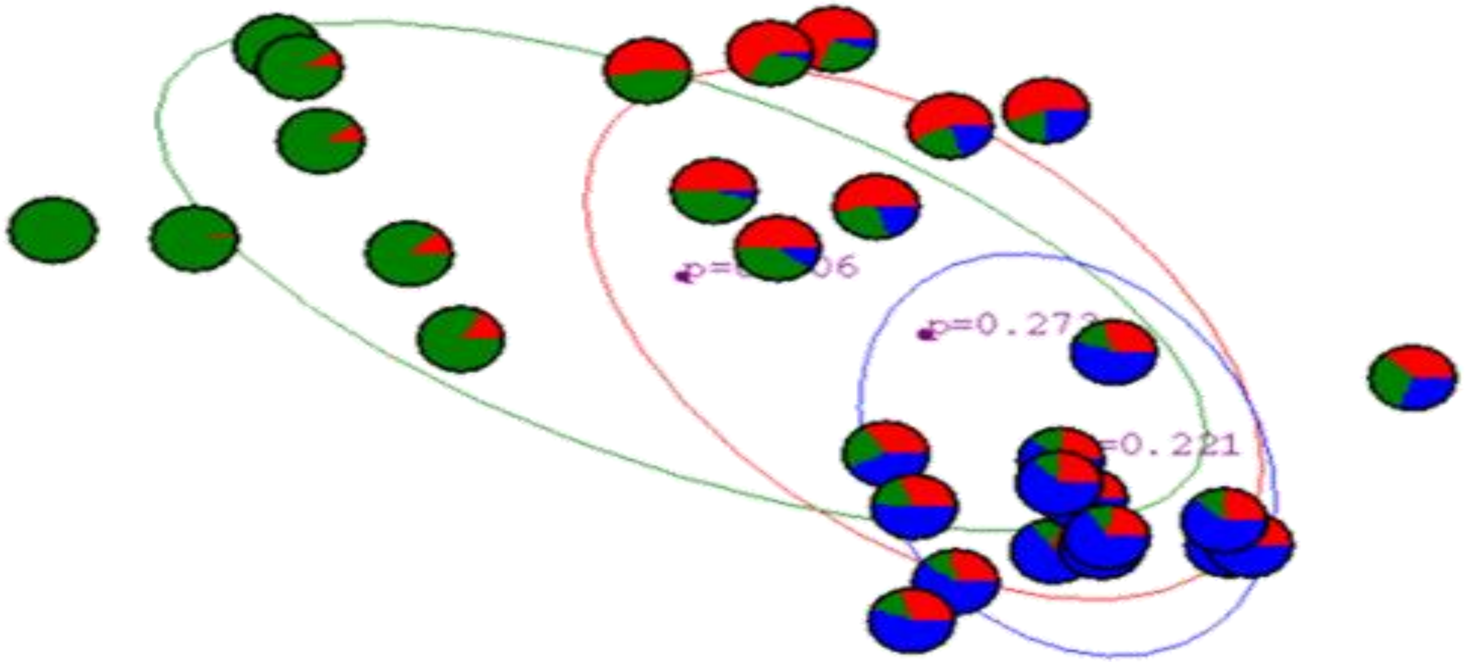
Start: 0th iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

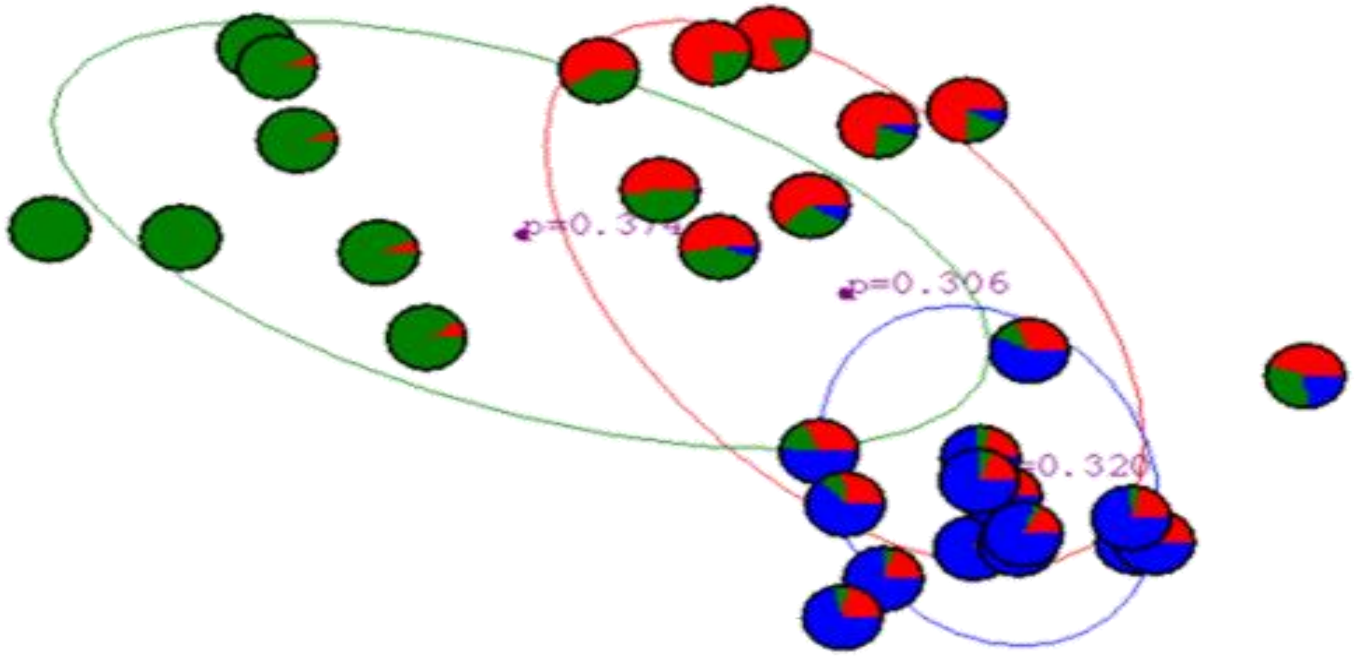
After 1st iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

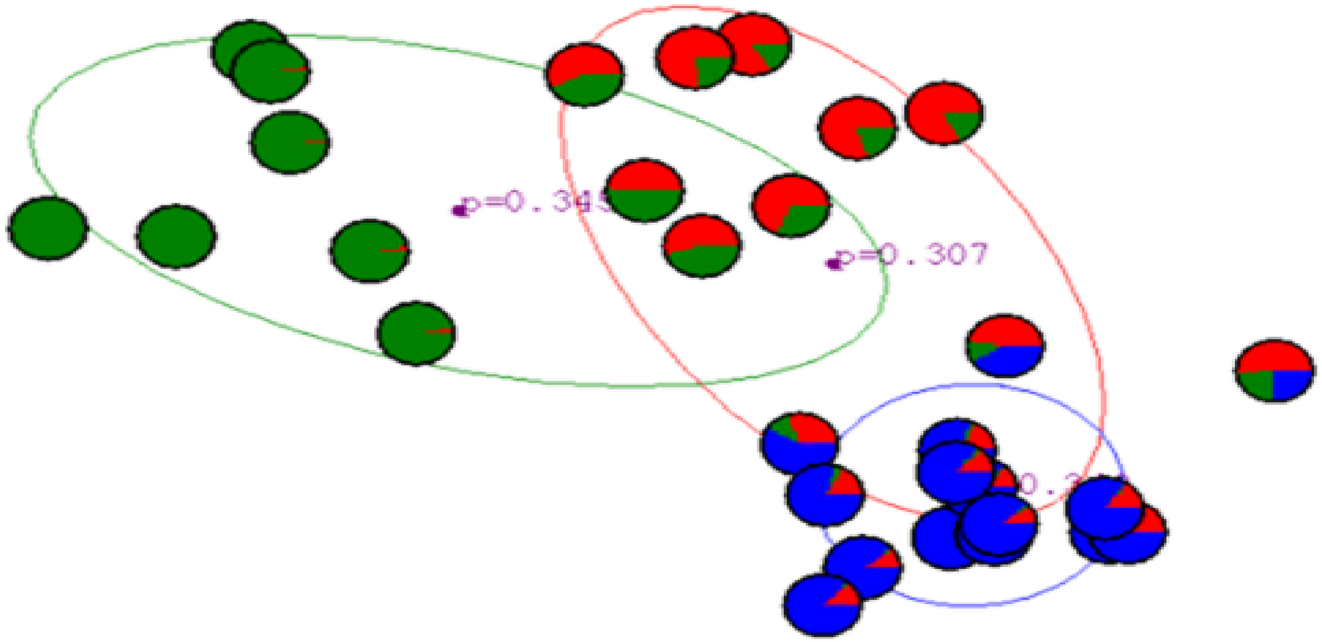
After 2nd iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

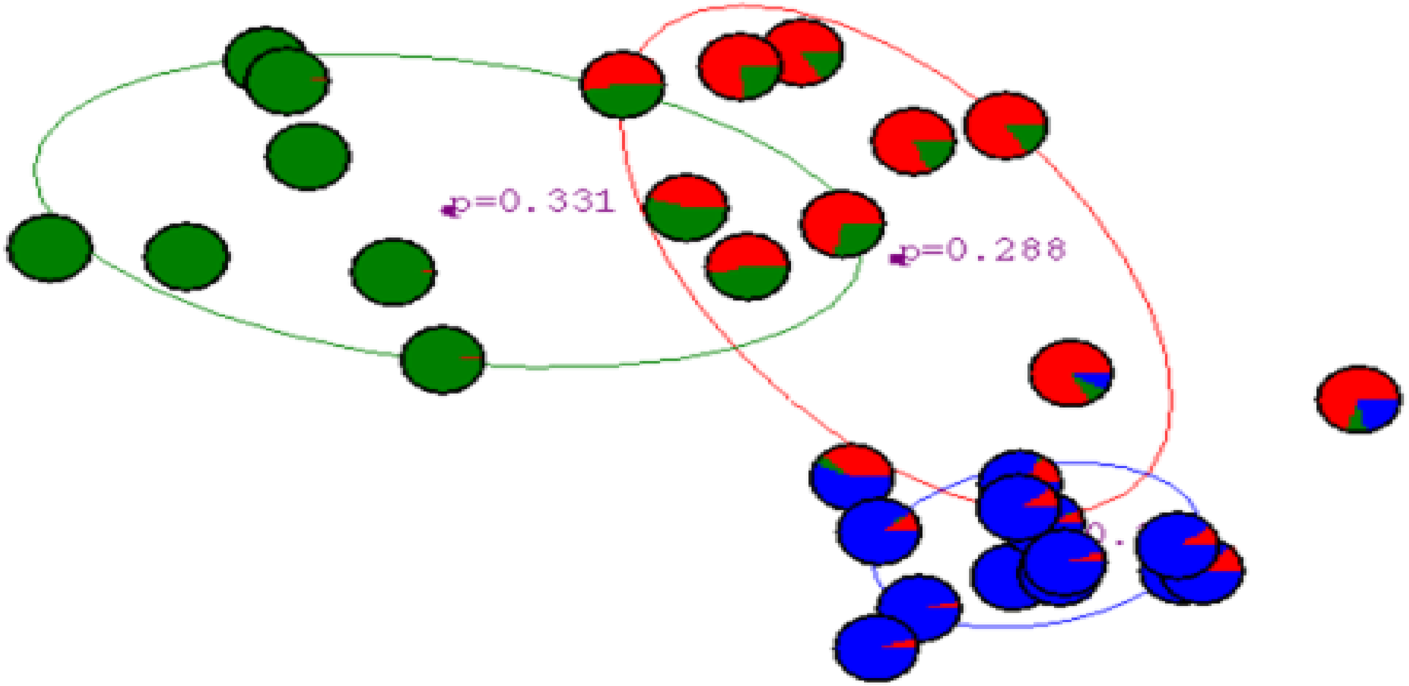
After 3rd iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

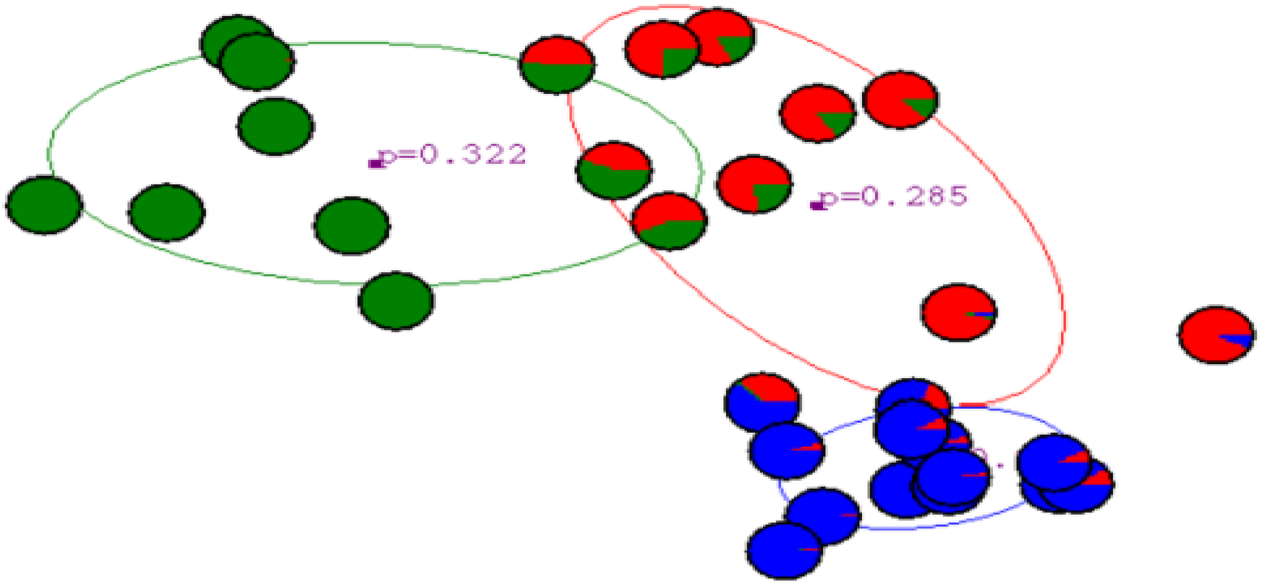
After 4th iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

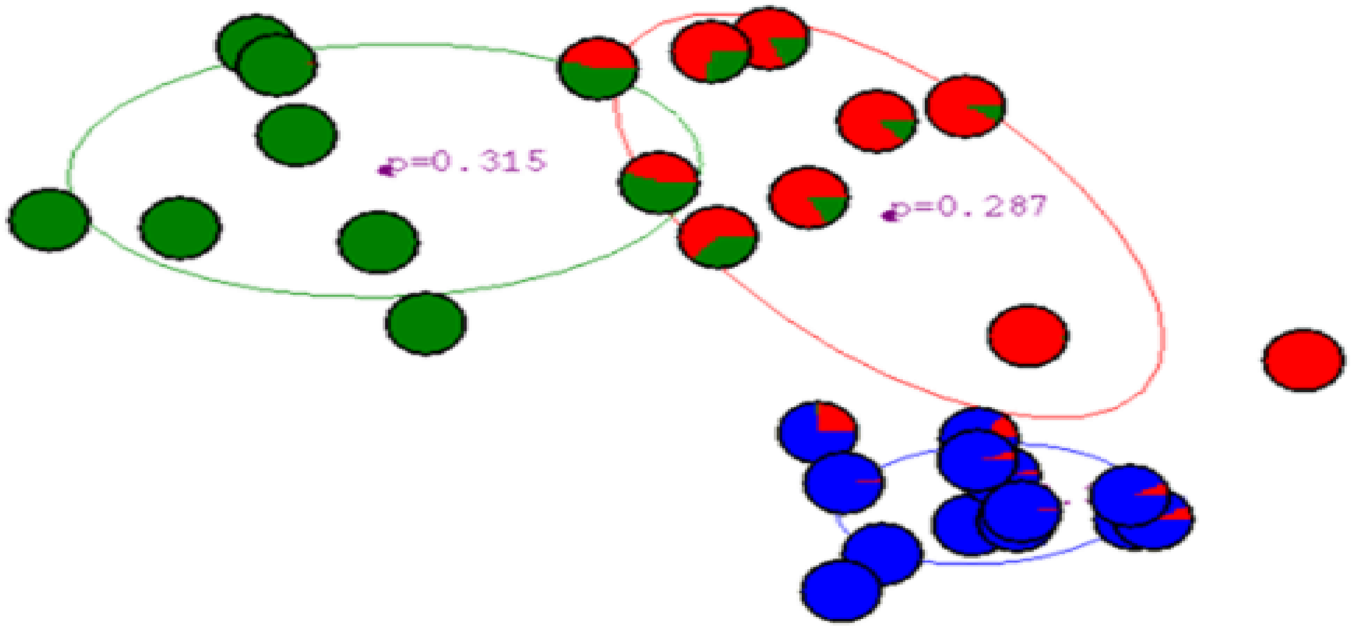
After 5th iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

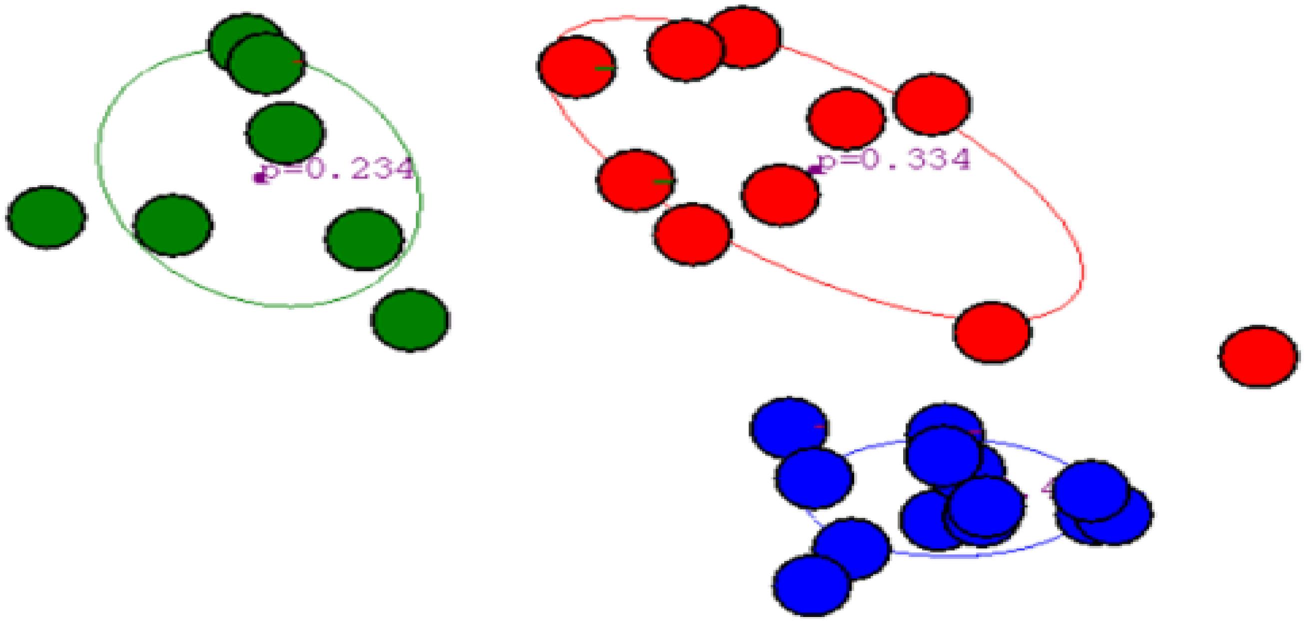
After 6th iteration



Slide Courtesy: Andrew Moore, CMU

GMM: Example

After 20th iteration



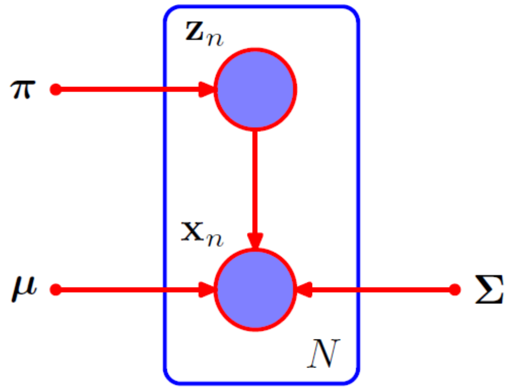
Slide Courtesy: Andrew Moore, CMU

More on EM Algorithm

- What are the EM algorithm initialization methods?
  - Random guess.
  - Initialized by k-means. After a few iterations of k-means, using the parameters to initialize EM
- What are the main advantages of parametric methods?
  - You can easily change the model to adapt to different distribution of data sets.
  - Knowledge representation is very compact. Once the model is selected, the model is represented by a specific number of parameters.
  - The number of parameters does not increase with the increasing of training data .

General EM Algortihm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.



1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter
If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

and return to step 2.

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}.$$

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_{nk}}.$$

Does it maximize the log likelihood ?

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln\left\{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right\}.$$