

Lecture 09

Gaussian Mixture Model

Mahdi Roozbahani
Georgia Tech

Outline

- Overview 
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm

Recap

Conditional probabilities:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

Bayes rule:

$$p(A|B) = \frac{p(A, B)}{p(B)} = \frac{p(B|A)p(A)}{p(B)}$$

$$p(A = 1) = \sum_{i=1}^K p(A = 1, B_i) = \sum_{i=1}^K p(A|B_i) p(B_i)$$

	Tomorrow=Rainy	Tomorrow=Cold	P(Today)
Today=Rainy	$\frac{4}{9}$	$\frac{2}{9}$	$[\frac{4}{9} + \frac{2}{9}] = \frac{2}{3}$
Today=Cold	$\frac{2}{9}$	$\frac{1}{9}$	$[\frac{2}{9} + \frac{1}{9}] = \frac{1}{3}$
P(Tomorrow)	$[\frac{4}{9} + \frac{2}{9}] = \frac{2}{3}$	$[\frac{2}{9} + \frac{1}{9}] = \frac{1}{3}$	

$P(\text{Tomorrow} = \text{Rainy}) =$

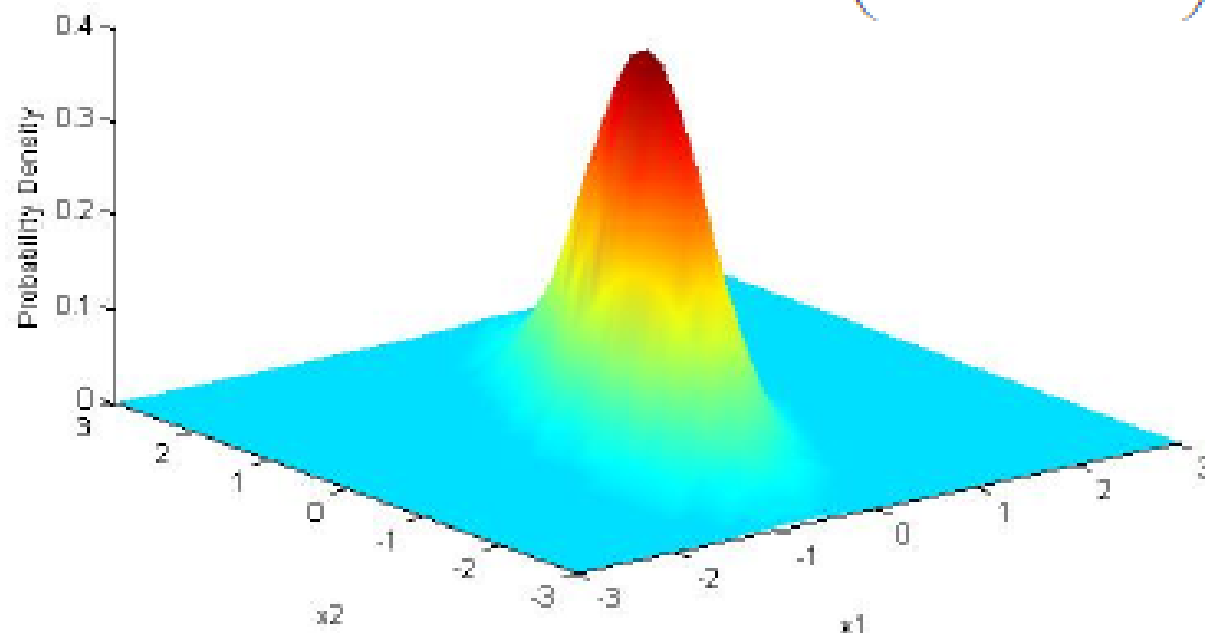
What is a Gaussian?

For d dimensions, the Gaussian distribution of a vector $x = (x^1, x^2, \dots, x^d)^T$ is defined by:

$$N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

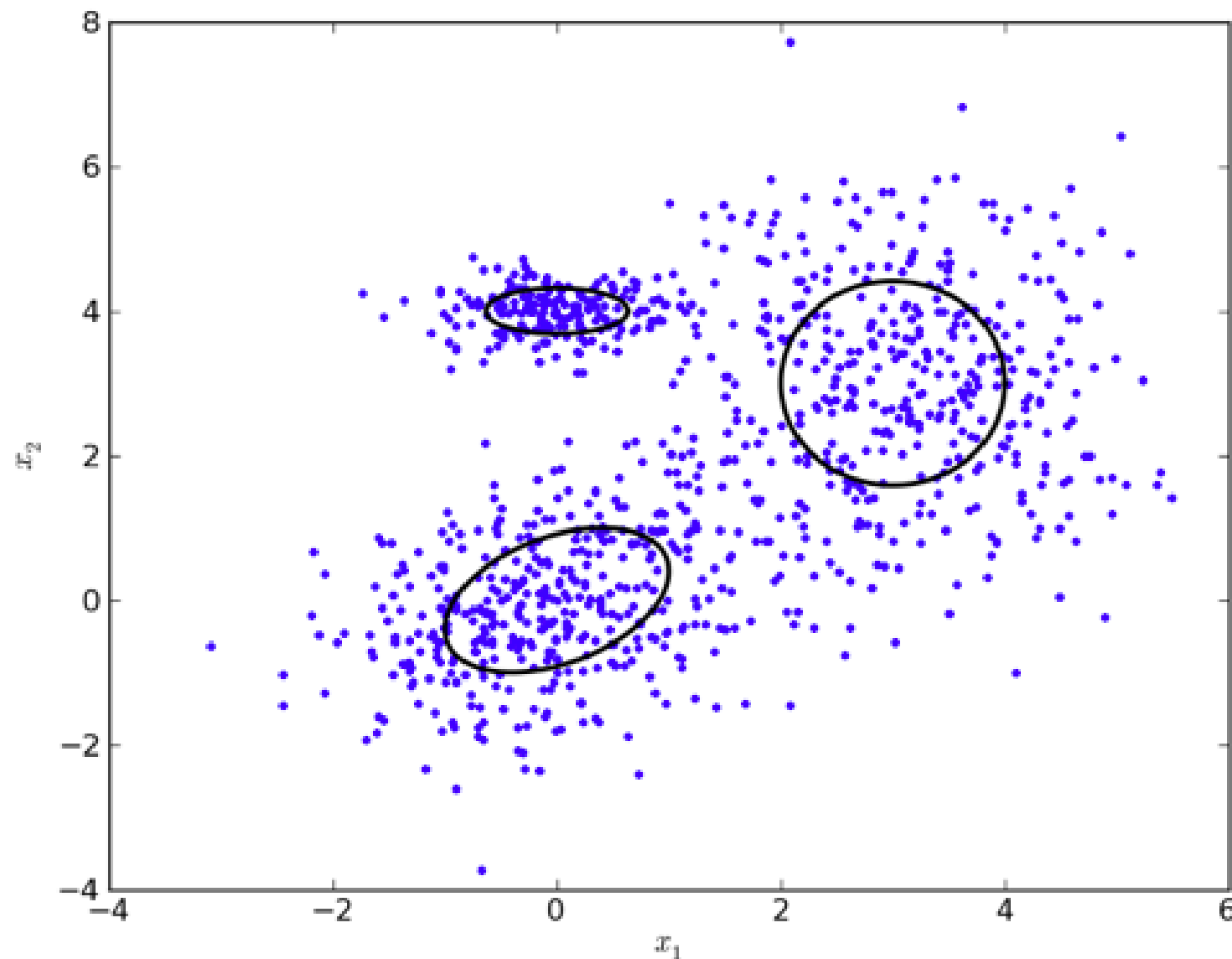
where μ is the mean and Σ is the covariance matrix of the Gaussian.

Example: $\mu = (0,0)^T$ $\Sigma = \begin{pmatrix} 0.25 & 0.30 \\ 0.30 & 1.00 \end{pmatrix}$



Hard Clustering Can Be Difficult

- Hard Clustering: K-Means, Hierarchical Clustering, DBSCAN



Towards Soft Clustering

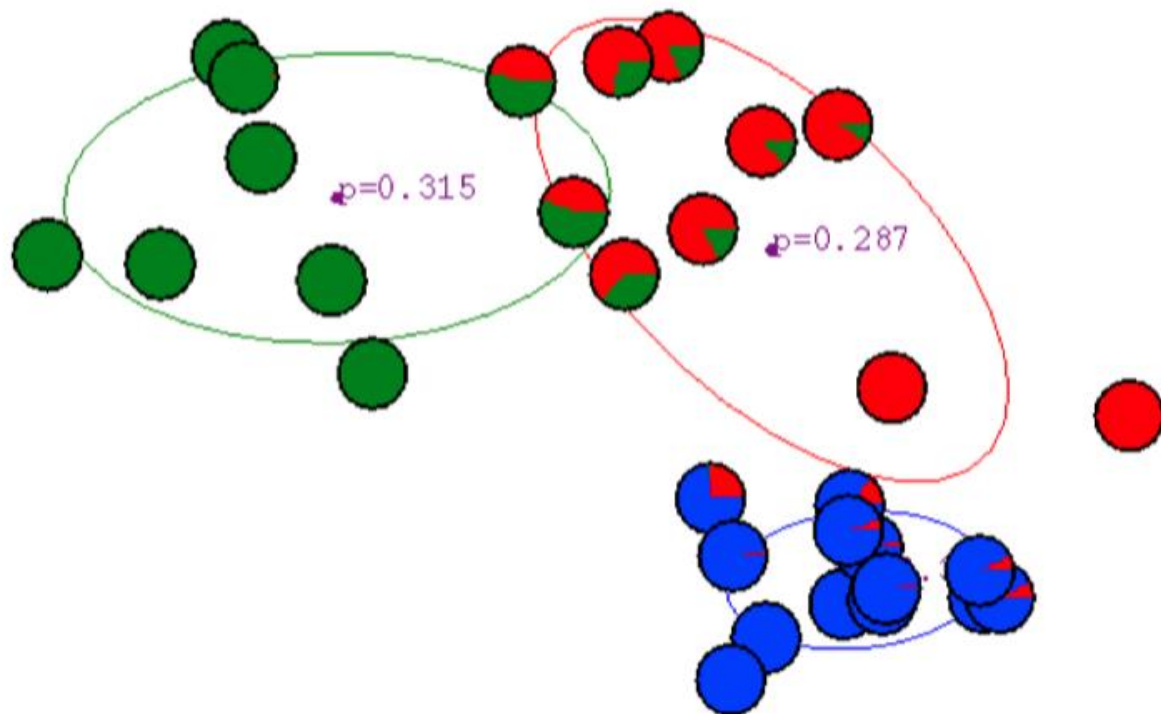
- **K-means**

- hard assignment:** each object belongs to only one cluster


$$\theta_i \in \{\theta_1, \dots, \theta_K\}$$

- **Mixture modeling**

- soft assignment:** probability that an object belongs to a cluster

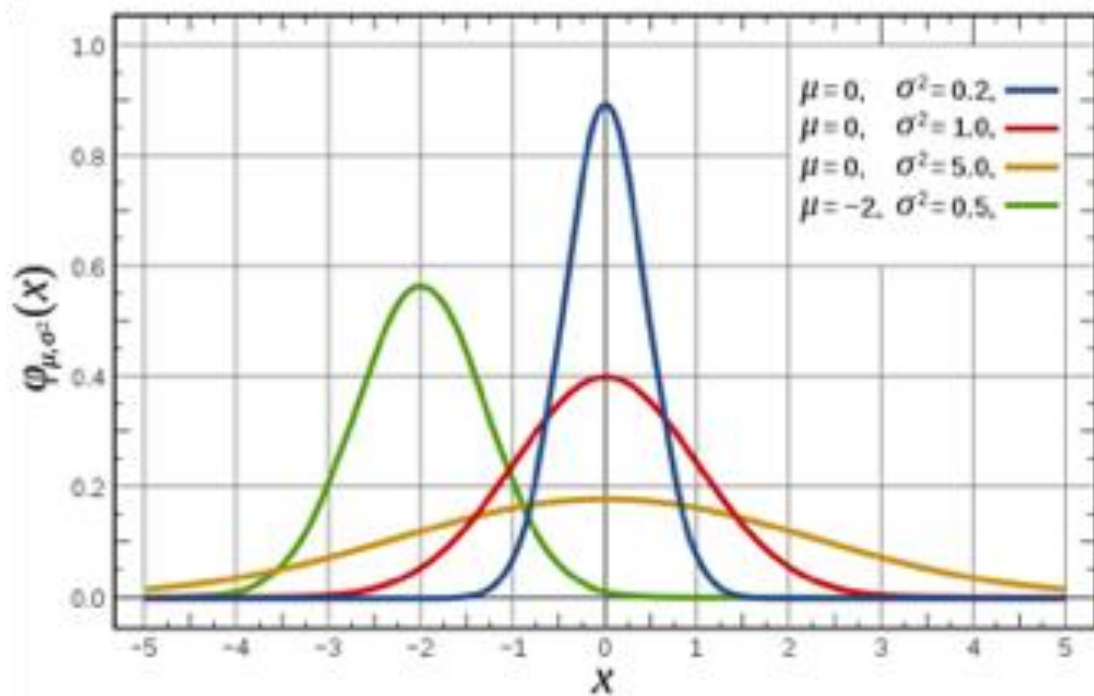


Outline

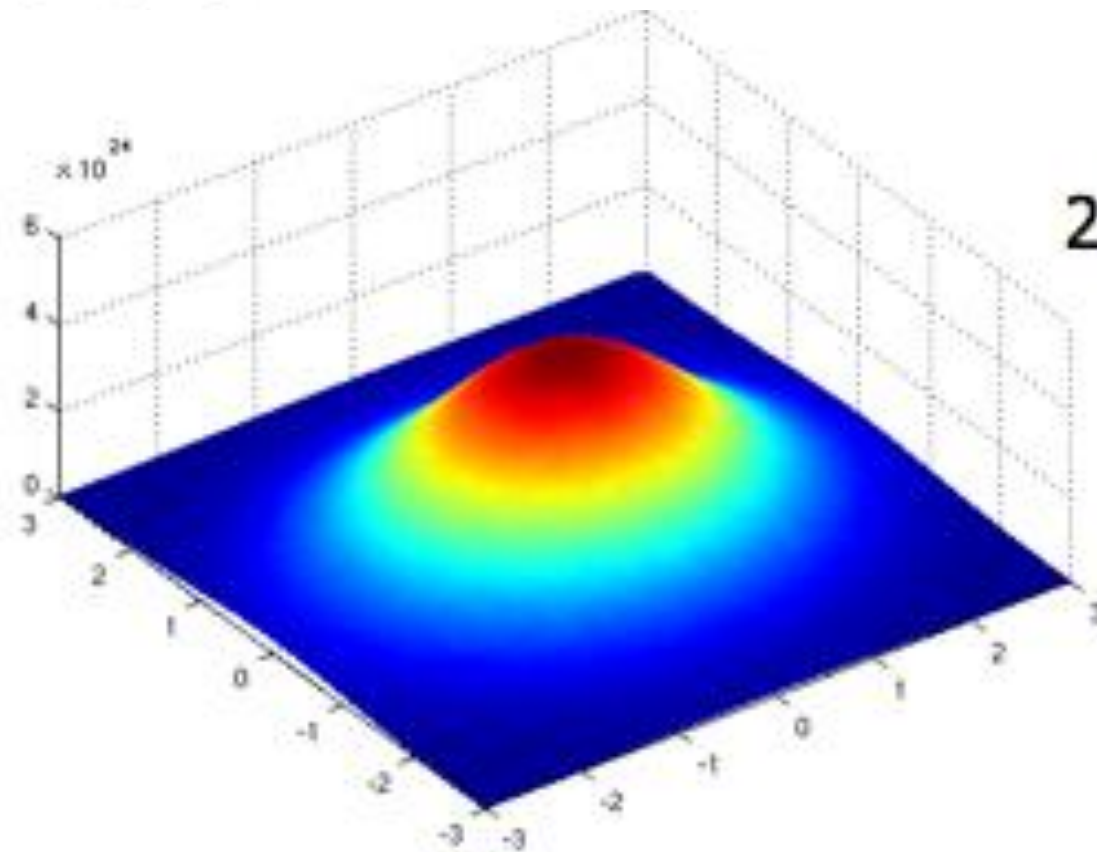
- Overview
- Gaussian Mixture Model 
- The Expectation-Maximization Algorithm

Gaussian Distribution

1-d Gaussian



$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



2-d Gaussian

Mixture Models

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

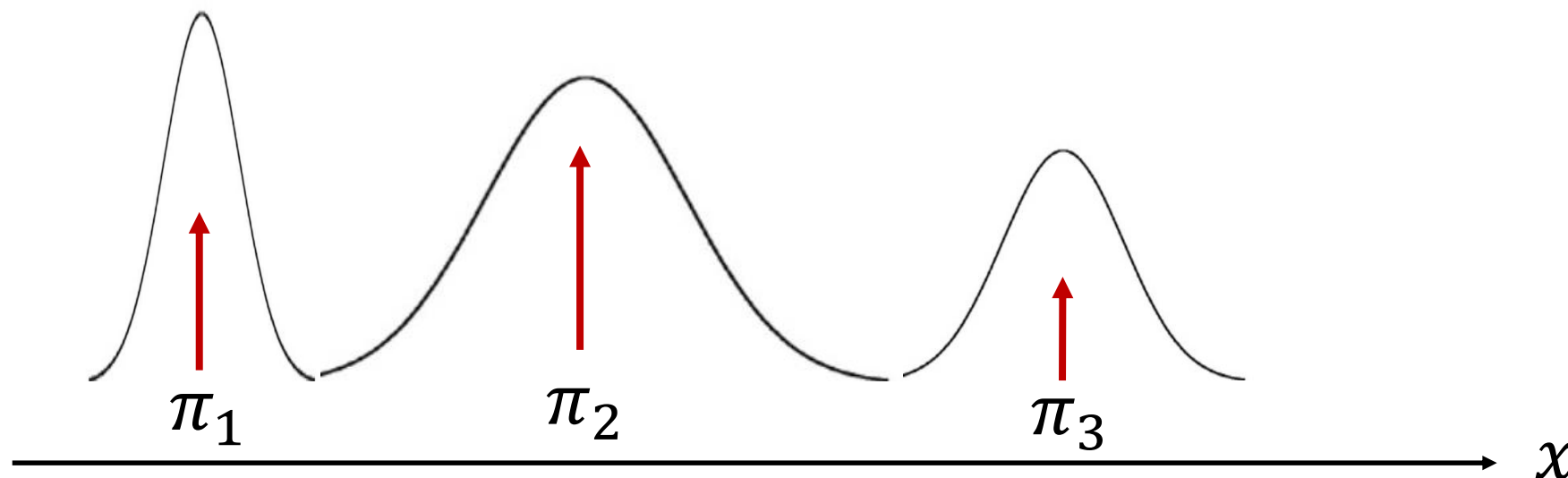
$$p(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x) + \dots + \pi_k f_k(x)$$

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

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



What is **f** in GMM?



Some notes:

Is summation of a bunch of Gaussians a Gaussian itself?

$p(x)$ is a Probability density function or it is also called a marginal distribution function.

$p(x)$ = the density of selecting a data point from the pdf which is created from a mixture model. The previous slide picture shows that $p(x)$ is pdf created by combination of mixing three Gaussians.

Also, we know that the area under a density function is equal to 1.

Mixtures of Gaussians

What is the probability of picking a mixture component (Gaussian model) = $p(z) = \pi_i$

AND

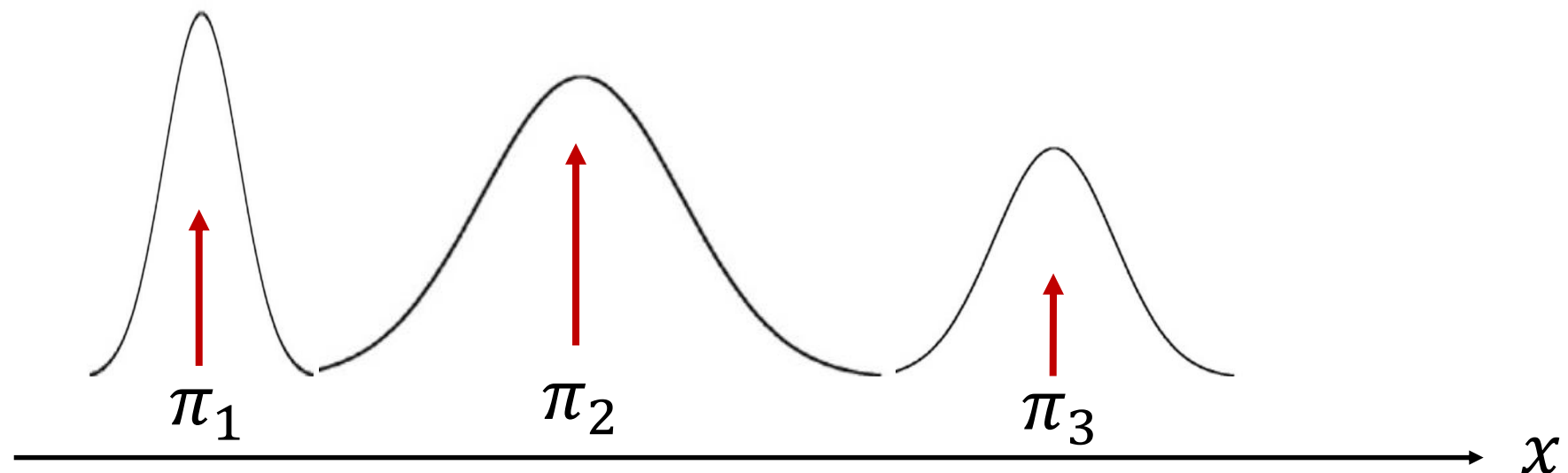
Picking data from that specific mixture component = $p(x|z)$

z is latent, we observe x , but z is hidden



$p(x, z) = p(x|z)p(z) \rightarrow$ Generative model, Joint distribution

$$p(x, z) = N(x|\mu_k, \sigma_k)\pi_k$$



Start with parameters describing each cluster:

Mean μ_k

Variance σ_k

Size π_k

Marginal probability distribution

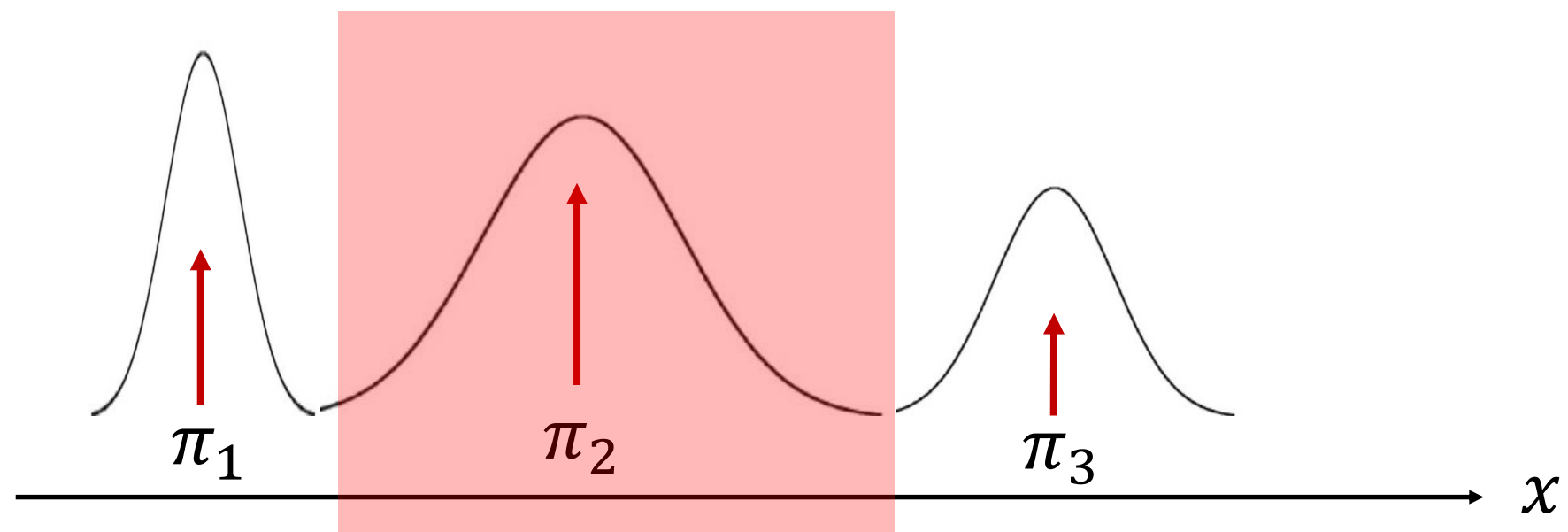
$$p(x|\theta) = \sum_k p(x, z_{nk}|\theta) = \sum_k \underbrace{p(x|z_{nk}, \theta)}_{f_k(x)} \underbrace{p(z_{nk}|\theta)}_{\pi_k} = \sum_k N(x|\mu_k, \sigma_k) \pi_k$$

$$p(z_{nk}|\theta) = \pi_k$$

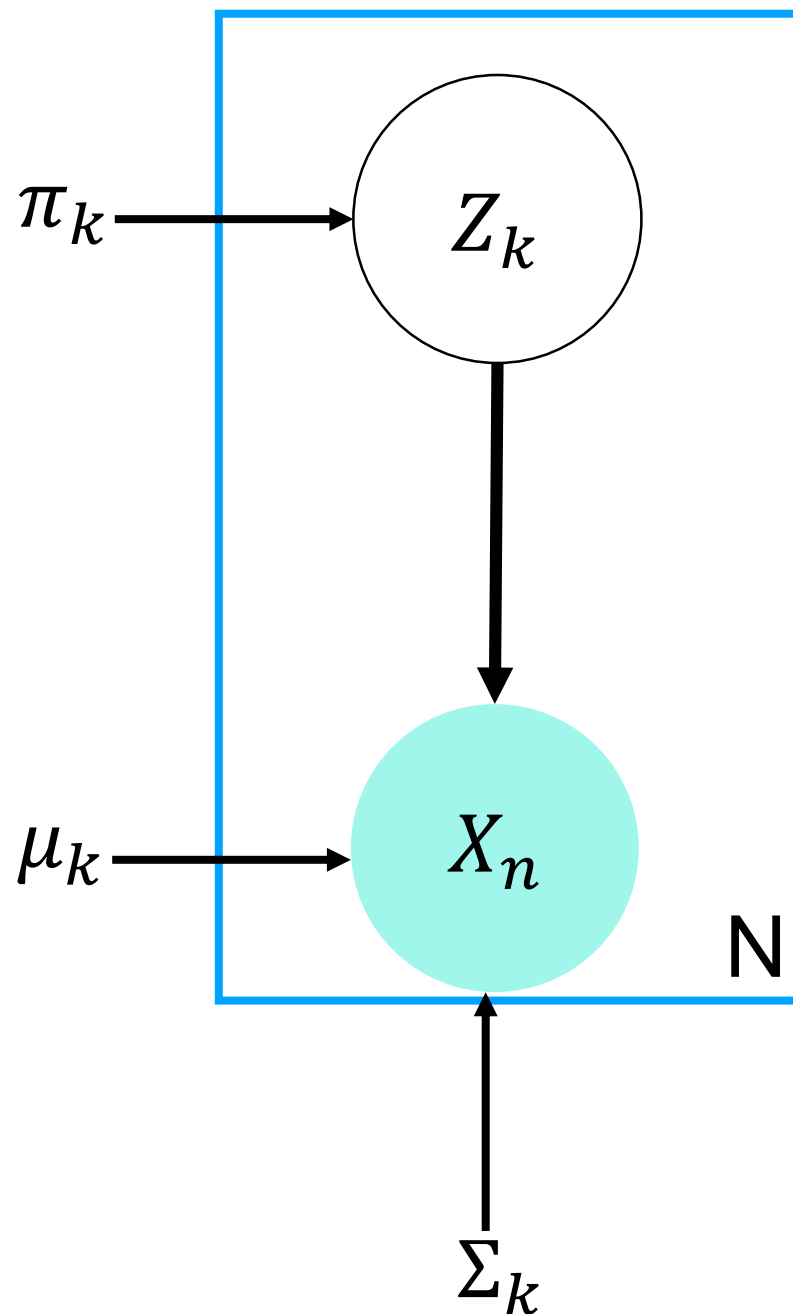
Select a mixture component with probability π

$$p(x|z_{nk}, \theta) = N(x|\mu_k, \sigma_k)$$

Sample from that component's Gaussian



GMM with graphical model concept

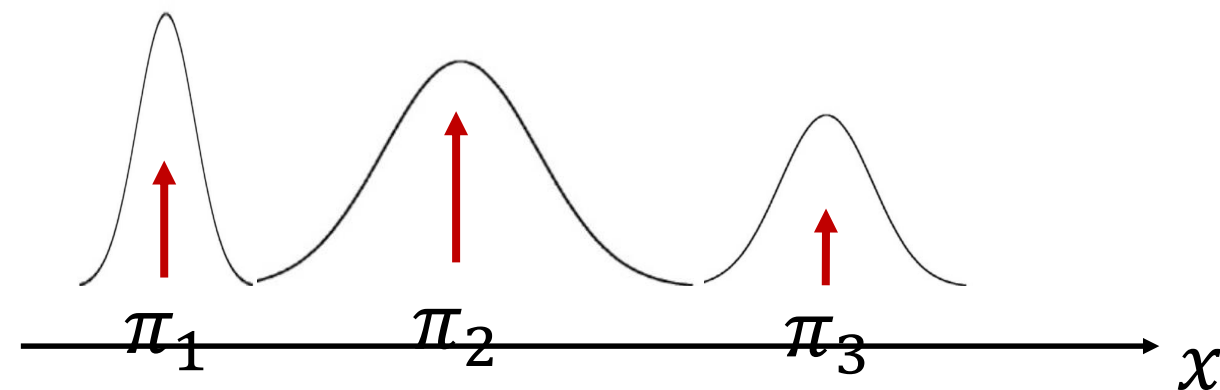


$$p(z_{nk} | \pi_k) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

Z_k is the latent variable
1-of-K representation

$$p(x | z_{nk}, \pi, \mu, \Sigma) = \prod_{k=1}^K \left(N(x | \mu_k, \Sigma_k) \right)^{z_{nk}}$$

Given z, π, μ , and Σ , what is the probability of x in component k



Why having “Latent variable”

- A variable can be unobserved (latent) because:
 - it is an imaginary quantity meant to provide some simplified and abstractive view of the data generation process.
 - e.g., speech recognition models, mixture models (soft clustering)...
 - it is a real-world object and/or phenomena, but difficult or impossible to measure
 - e.g., the temperature of a star, causes of a disease, evolutionary ancestors ...
 - it is a real-world object and/or phenomena, but sometimes wasn't measured, because of faulty sensors, etc.
- Discrete latent variables can be used to partition/cluster data into sub-groups.
- Continuous latent variables (factors) can be used for dimensionality reduction (factor analysis, etc).

Latent variable representation

$$p(\mathbf{x}|\theta) = \sum_k p(x, z_{nk}|\theta) = \sum_k p(z_{nk}|\theta)p(x|z_{nk}, \theta) = \sum_{k=0}^K \pi_k N(x|\mu_k, \Sigma_k)$$

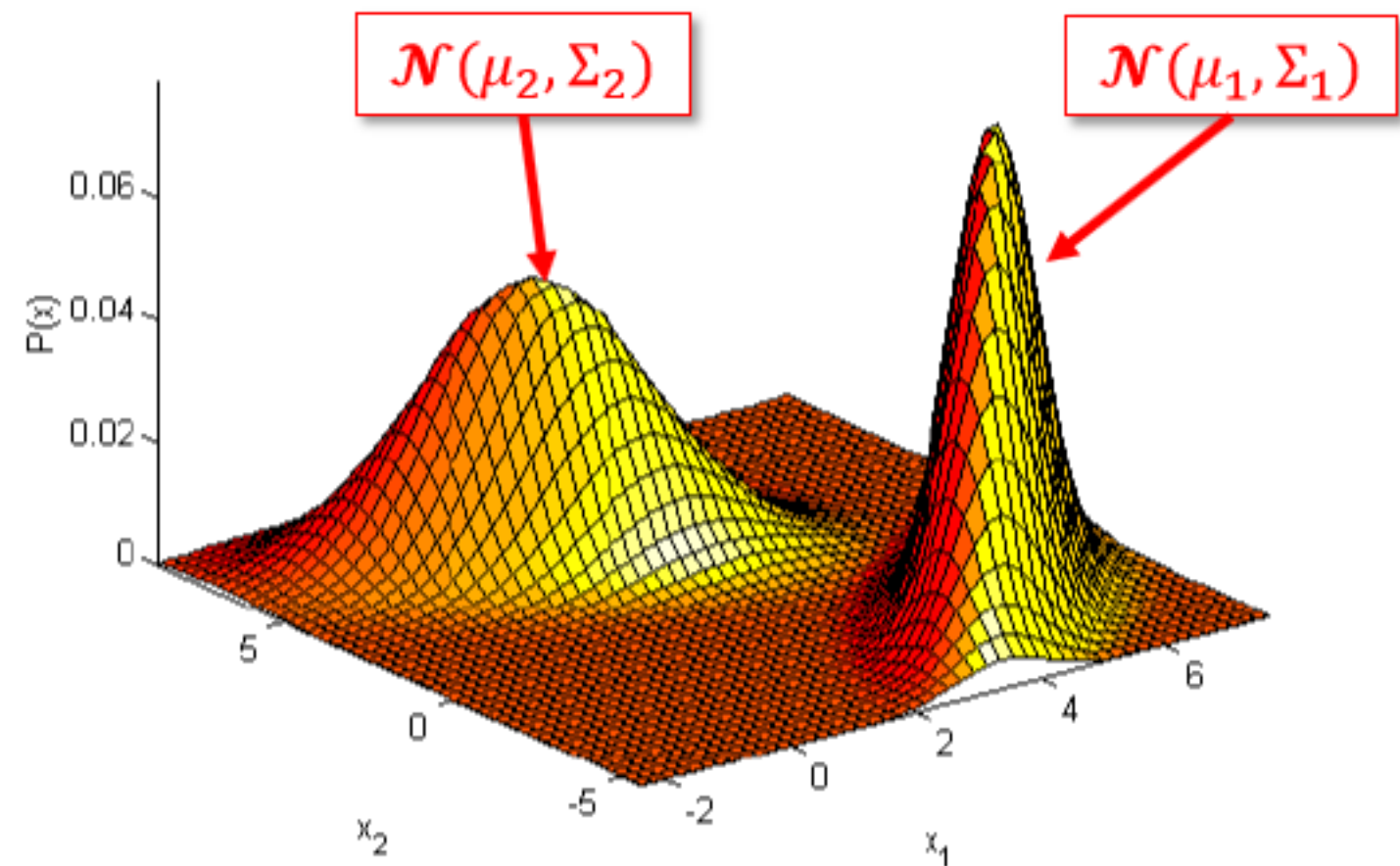
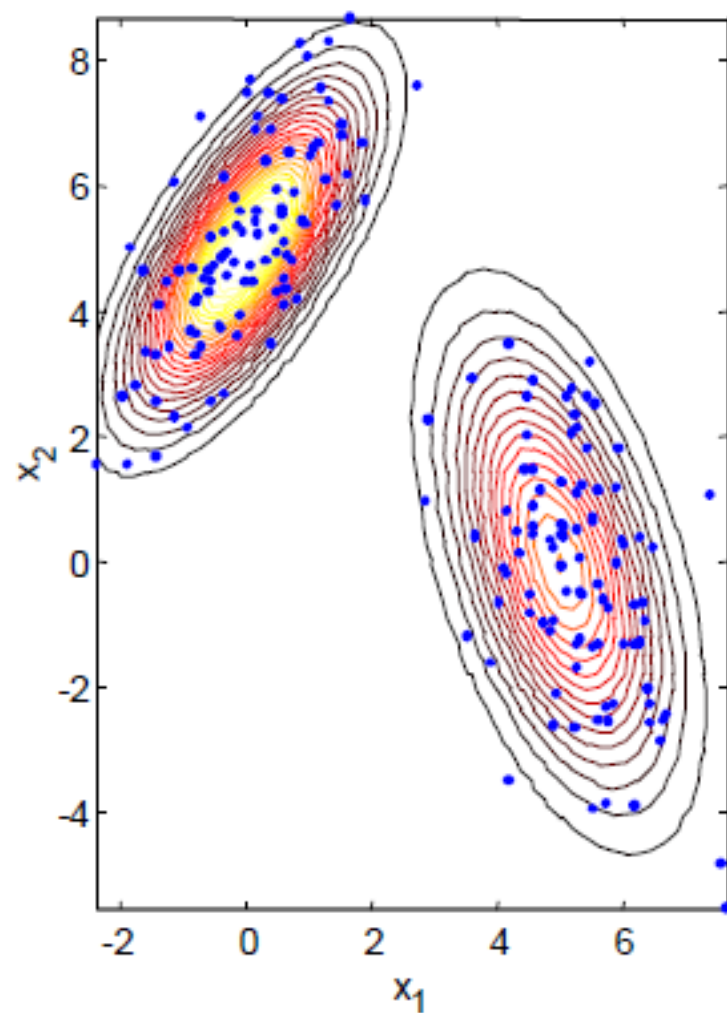
$$p(z_{nk}|\theta) = \prod_{k=1}^K \pi_k^{z_{nk}} \quad p(x|z_{nk}, \theta) = \prod_{k=1}^K \left(N(x|\mu_k, \Sigma_k) \right)^{z_{nk}}$$

Why having the latent variable?

The distribution that we can model using a mixture of Gaussian components is much more expressive than what we could have modeled using a single component.

Multimodal distribution

- What if we know the data consists of a few Gaussians
- What if we want to fit parametric models



Gaussian Mixture Model

- A density model $p(X)$ may be multi-modal: model it as a mixture of uni-modal distributions (e.g. Gaussians)

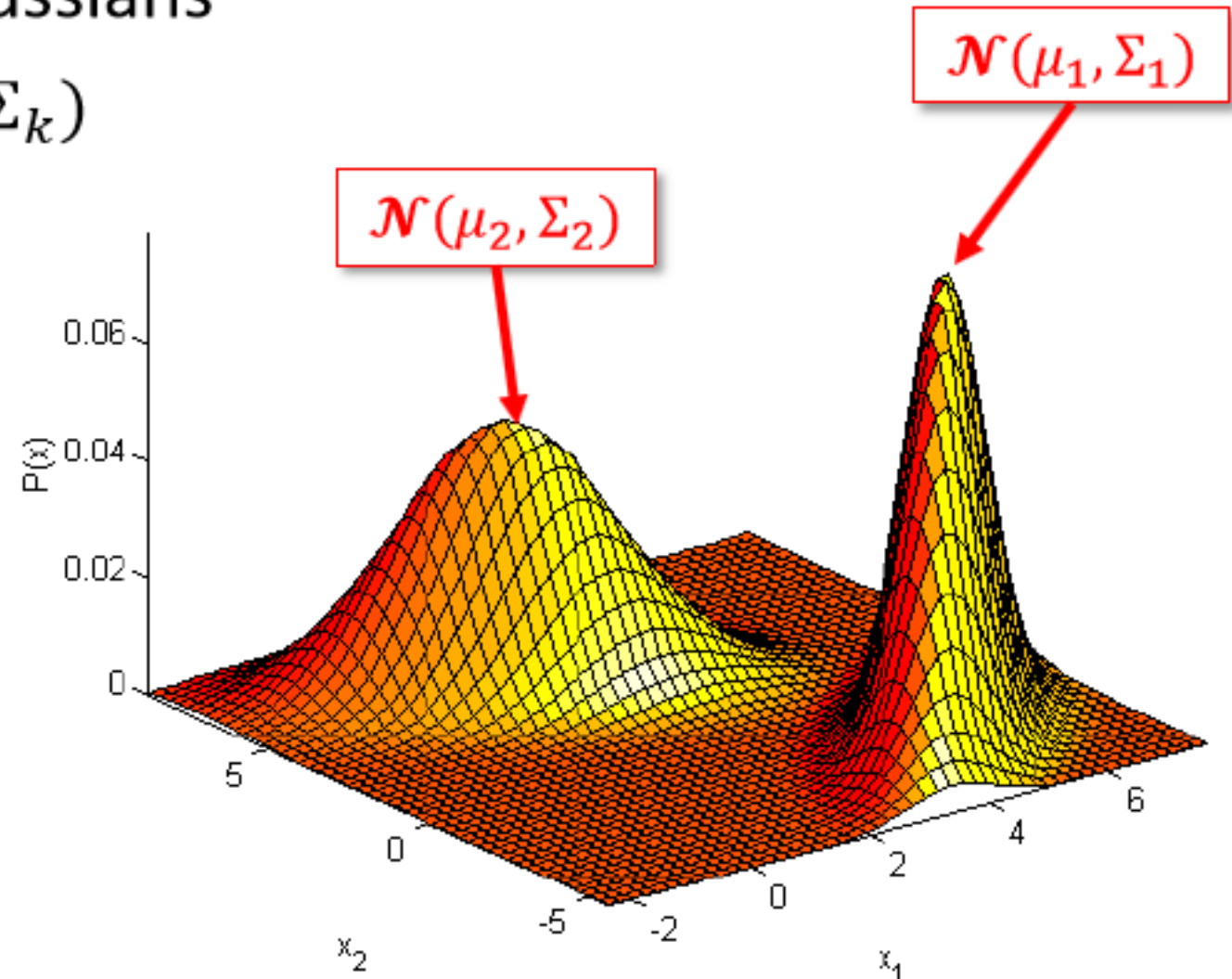
- Consider a mixture of K Gaussians

- $p(X) = \sum_{k=1}^K \pi_k \mathcal{N}(X|\mu_k, \Sigma_k)$

mixing
proportion

mixture
Component

- Learn $\pi_k \in (0,1), \mu_k, \Sigma_k$;



Inferring Cluster Membership

- We have representations of the joint $p(x, z_{nk}|\theta)$ and the marginal, $p(x|\theta)$
- The conditional of $p(z_{nk}|x, \theta)$ can be derived using Bayes rule.
 - The **responsibility** that **a** mixture component takes for explaining an observation x .

$$\begin{aligned}\tau(z_{nk}) = p(z_{nk}|x) &= \frac{p(z_{nk})p(x|z_{nk})}{\sum_{j=1}^K p(z_{nj})p(x|z_{nj})} \\ &= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x|\mu_j, \Sigma_j)}\end{aligned}$$

Well, we don't know π_k, μ_k, Σ_k
What should we do?

We use a method called “Maximum Likelihood Estimation” (MLE) to solve the problem.

$$p(x|\theta) = \sum_k p(x, z_{nk}|\theta) = \sum_k p(z_{nk}|\theta)p(x|z_{nk}, \theta) = \sum_{k=0}^K \pi_k N(x|\mu_k, \Sigma_k)$$

Let's identify a likelihood function :

Now, let's find the missing parameters that maximizes the likelihood:

$$\arg \max p(x|\theta) = p(x|\pi, \mu, \Sigma) = \prod_{n=1}^N p(x_n|\theta) = \prod_{n=1}^N \sum_{k=0}^K \pi_k N(x_n|\mu_k, \Sigma_k)$$

$$\arg \max p(x) = p(x|\pi, \mu, \Sigma) = \prod_{n=1}^N p(x_n|\theta) = \prod_{n=1}^N \sum_{k=0}^K \pi_k N(x_n|\mu_k, \Sigma_k)$$

$$\ln[p(x)] = \ln[p(x|\pi, \mu, \Sigma)]$$

- 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.

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

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

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

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

Maximum Likelihood of a GMM

- Optimization of covariance

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

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

Maximum Likelihood of a GMM

- Optimization of mixing term

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

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

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

MLE of a GMM


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

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

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

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

Outline

- Overview
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm 

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

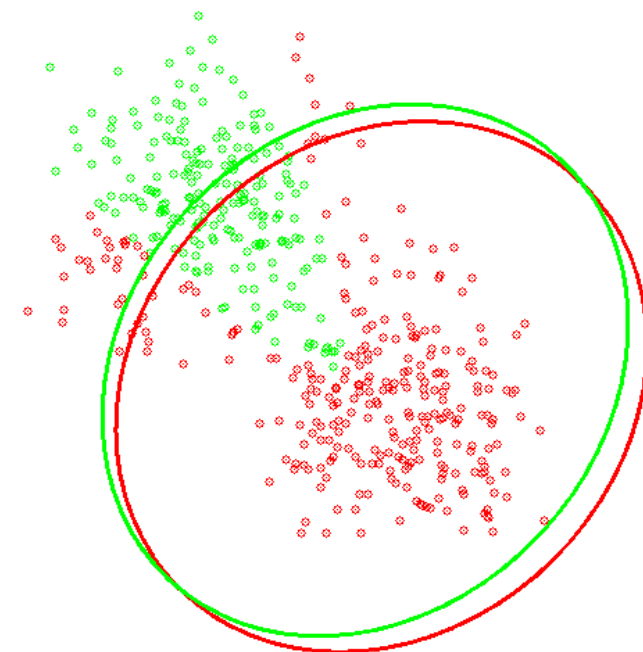
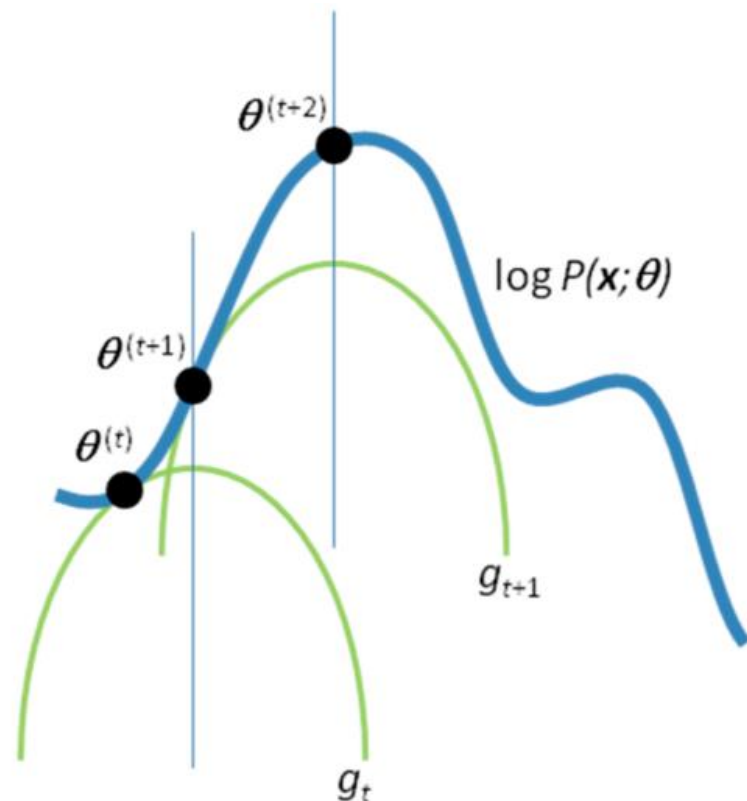
$$\mu_k^{new} = \frac{\sum_{n=1}^N \tau(z_{nk}) x_n}{N_k}$$

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

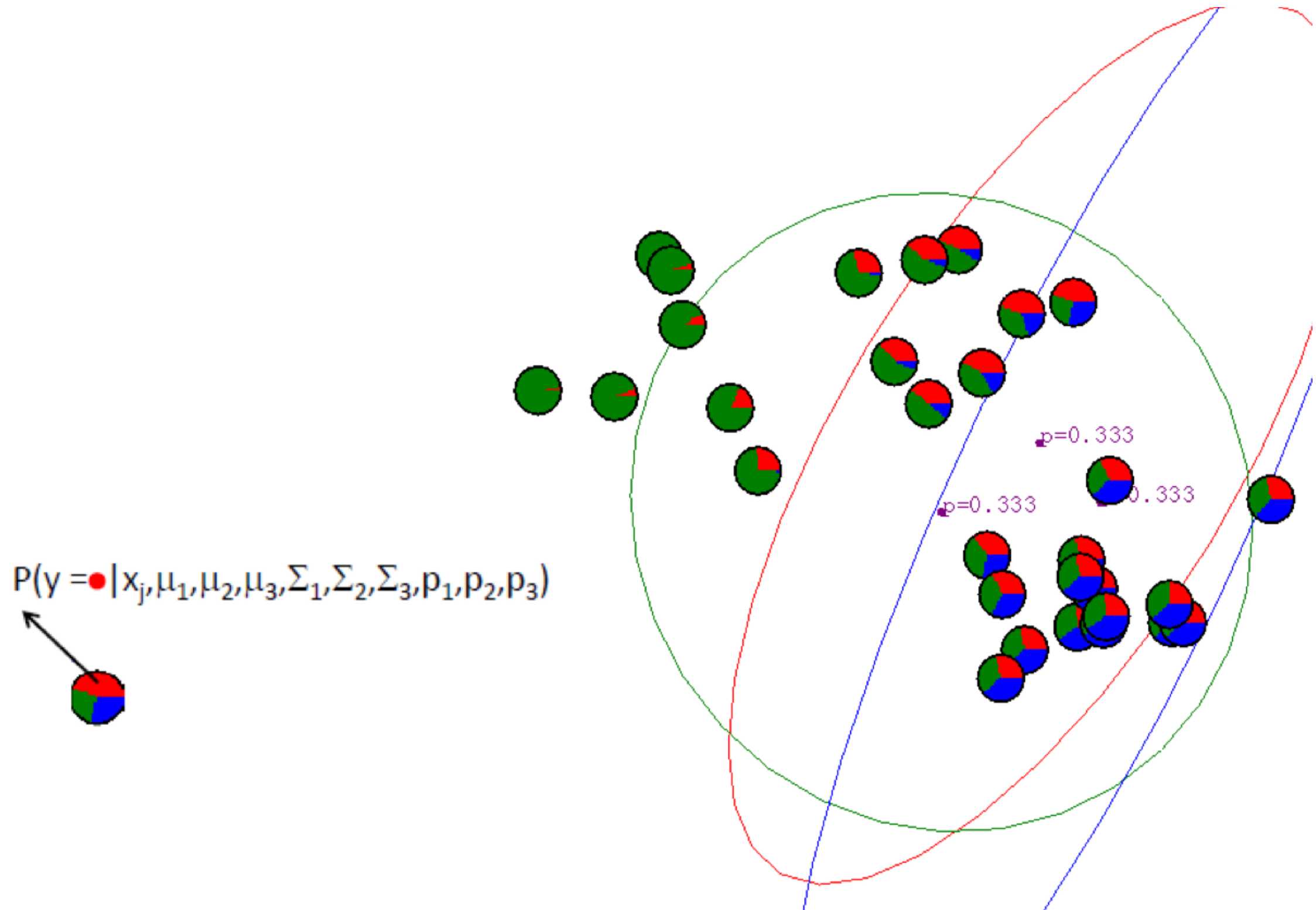
$$\pi_k^{new} = \frac{N_k}{N}$$

Expectation Maximization

- Expectation Maximization (EM) is a general algorithm to deal with hidden variables.
- Two steps:
 - E-Step: Fill-in hidden values using inference
 - M-Step: Apply standard MLE method to estimate parameters
- EM always converges to a local minimum of the likelihood.

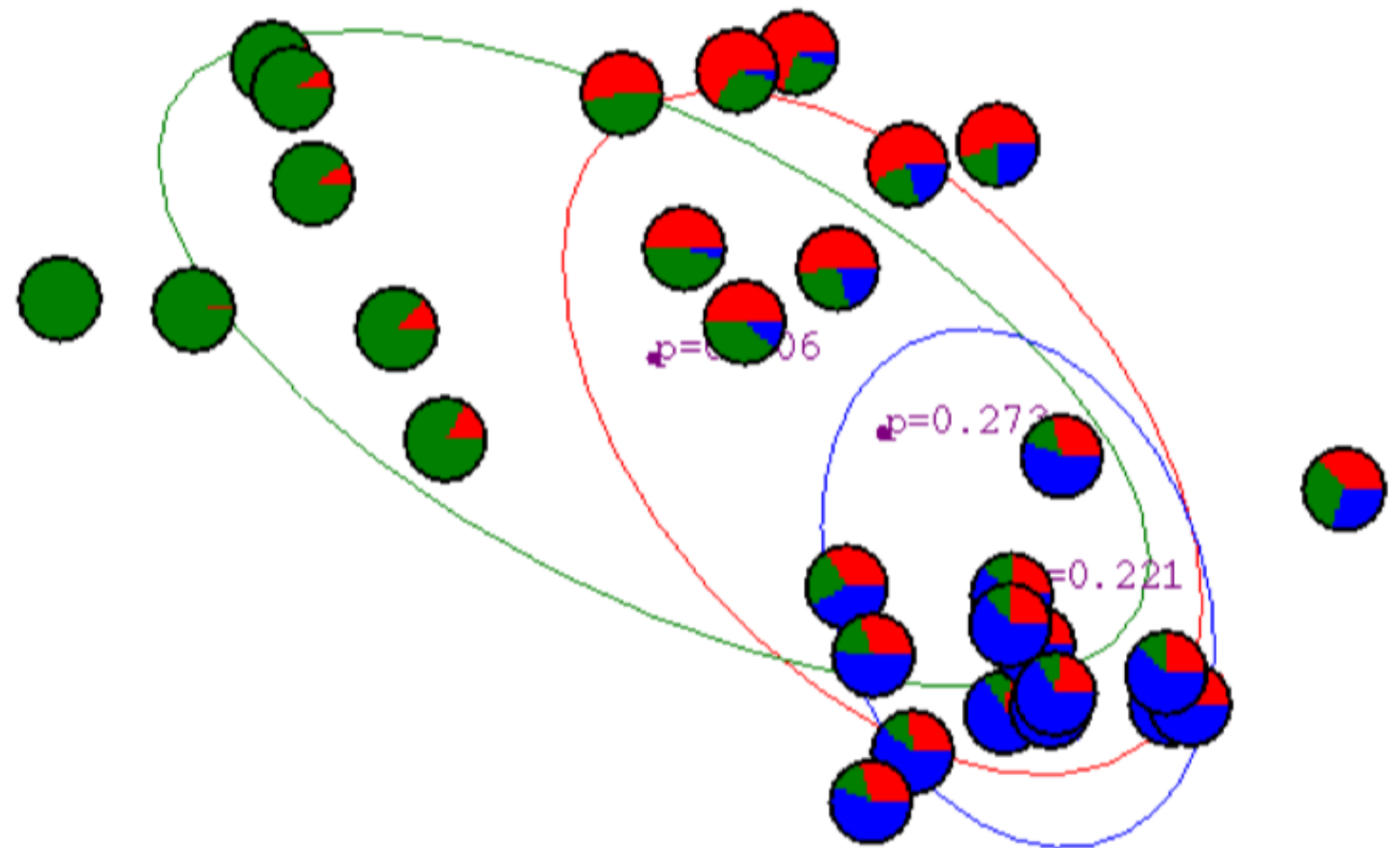


EM for Gaussian Mixture Model: [Example](#)



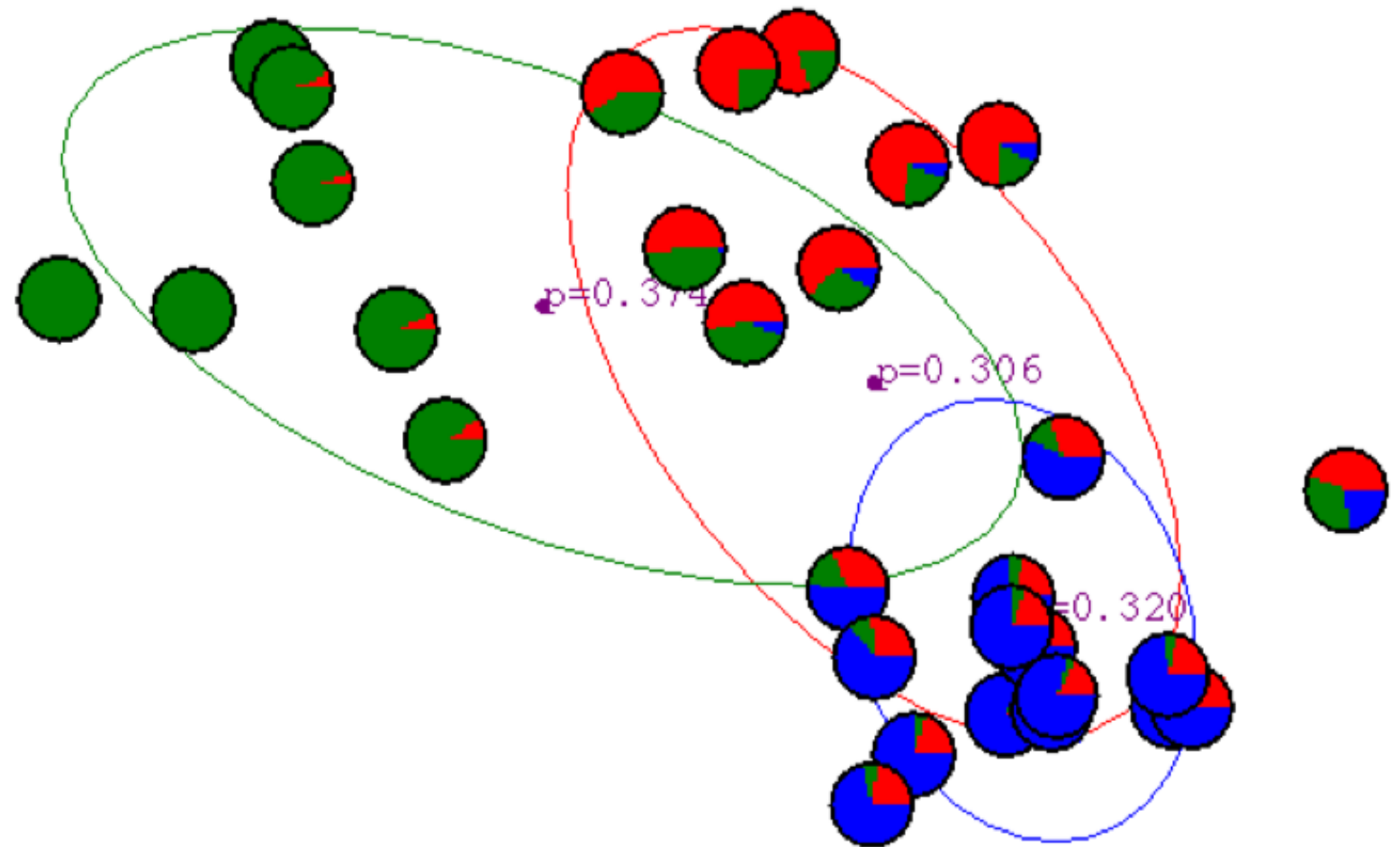
EM for Gaussian Mixture Model: Example

After 1st iteration



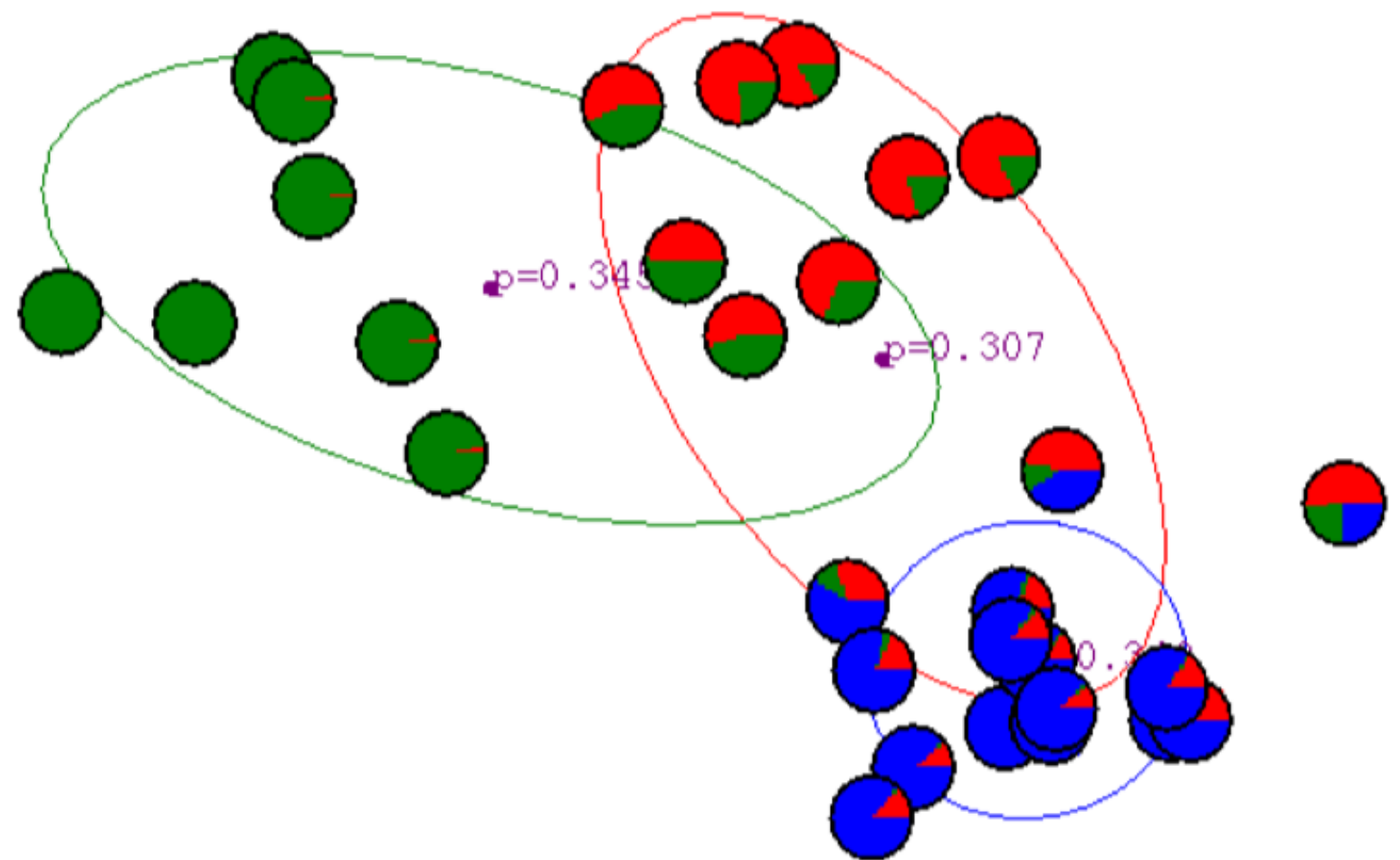
EM for Gaussian Mixture Model: Example

After 2nd iteration



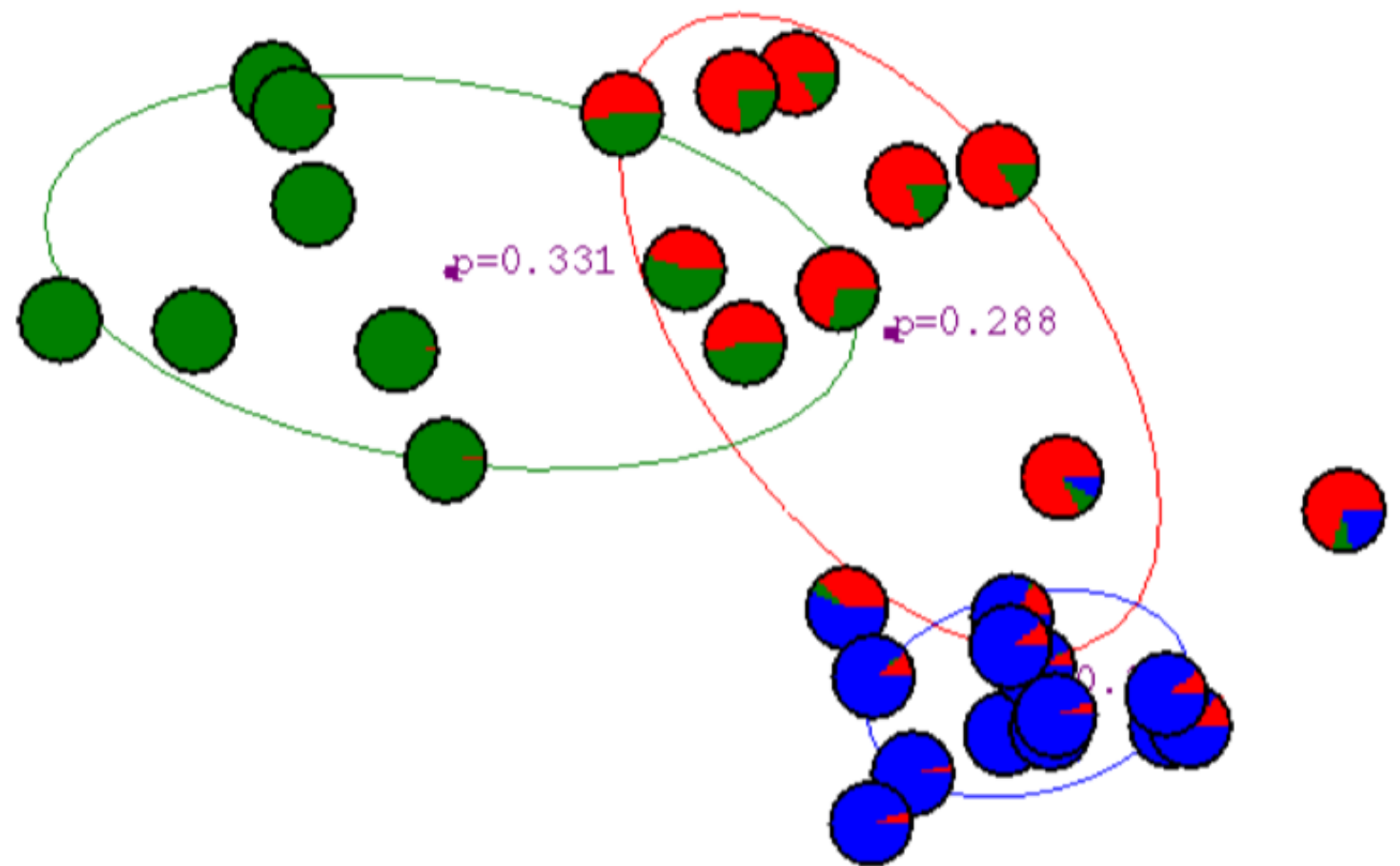
EM for Gaussian Mixture Model: Example

After 3rd iteration



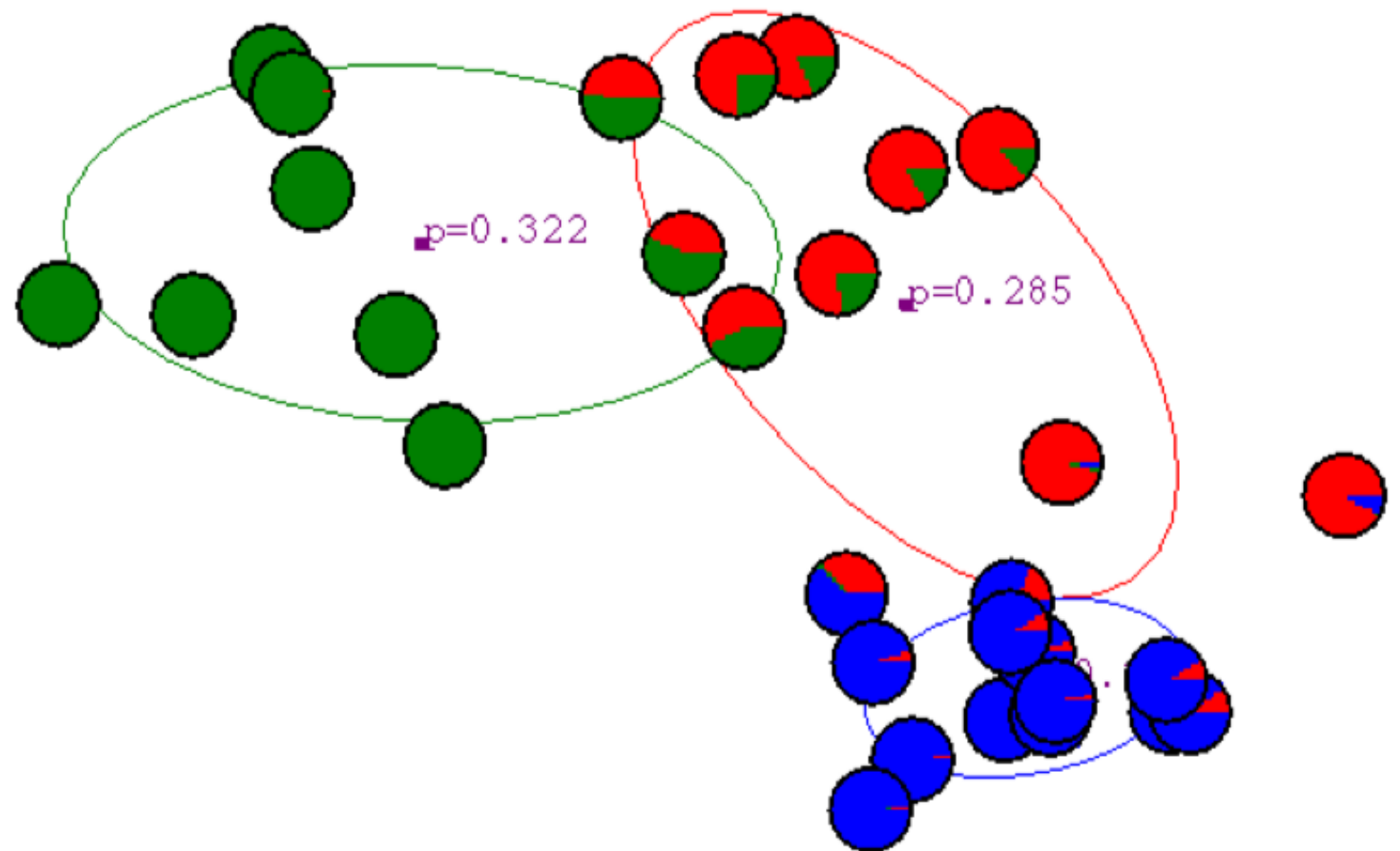
EM for Gaussian Mixture Model: Example

After 4th iteration



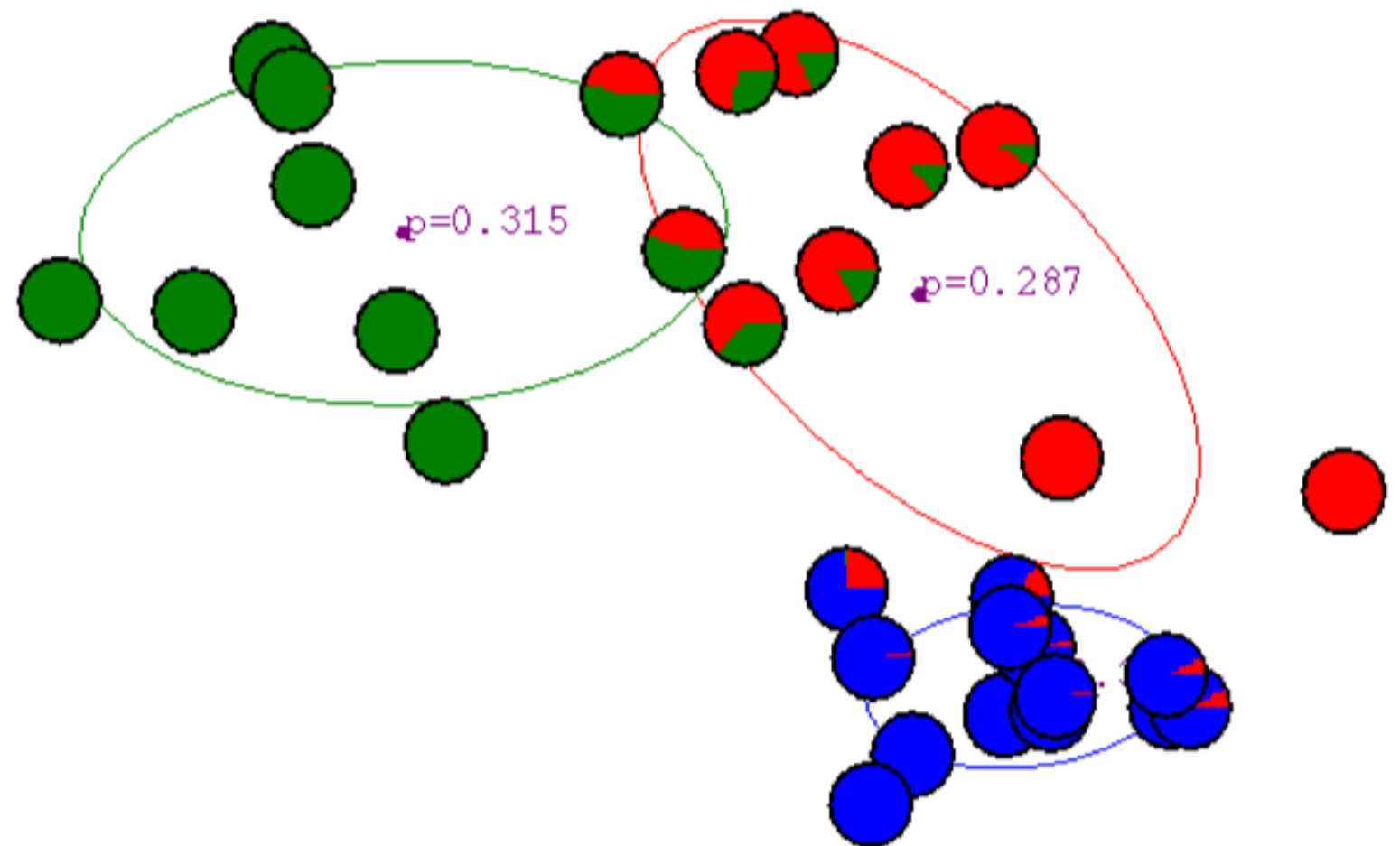
EM for Gaussian Mixture Model: Example

After 5th iteration



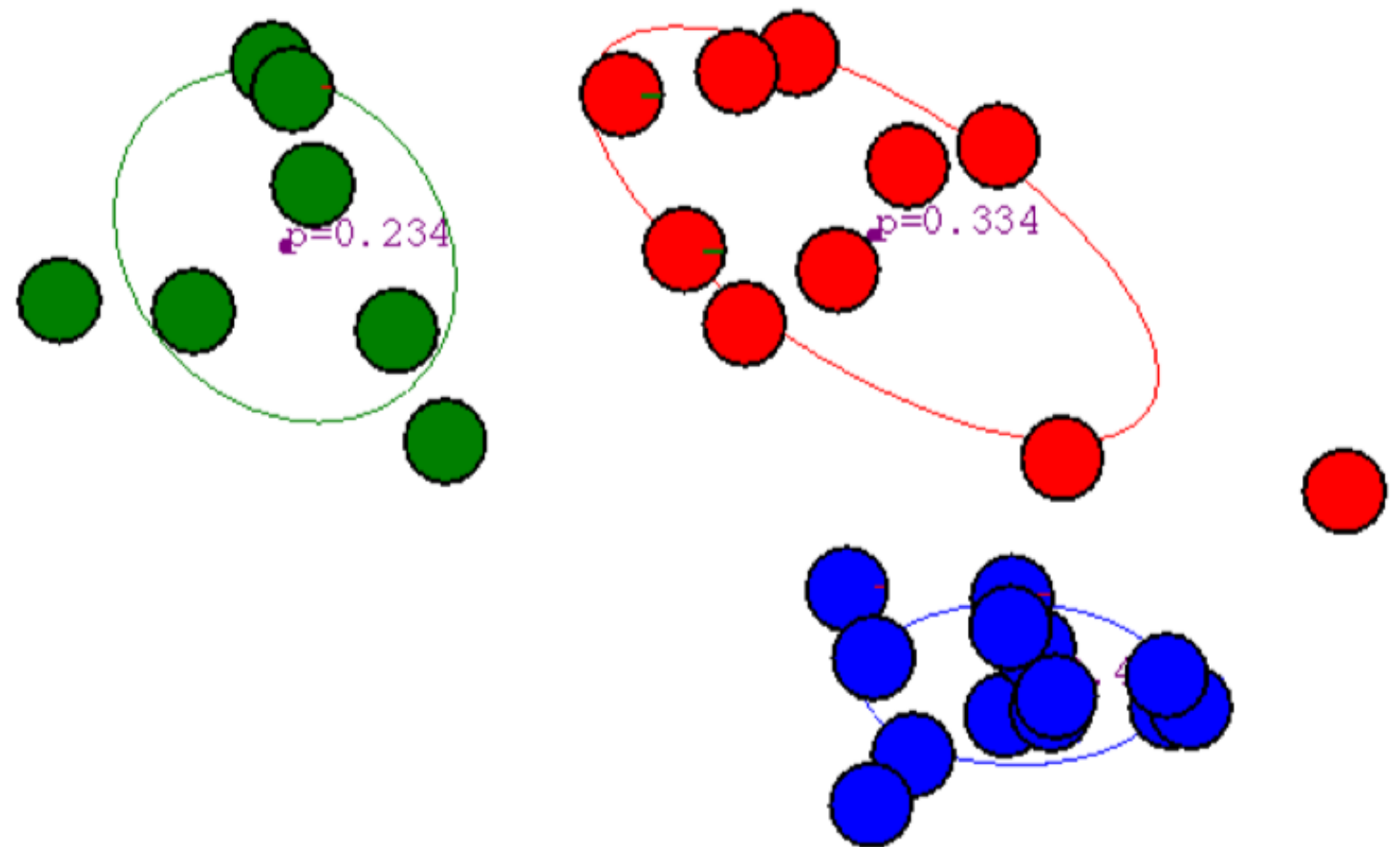
EM for Gaussian Mixture Model: Example

After 6th iteration



EM for Gaussian Mixture Model: Example

After 20th iteration



EM Algorithm for GMM

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means μ_j , covariances Σ_j and mixing coefficients π_j , and evaluate the initial value of the log likelihood.
2. **E step.** Evaluate the responsibilities using the current parameter values

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

EM Algorithm for GMM

3. **M step.** Re-estimate the parameters using the current responsibilities

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

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

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

4. Evaluate log likelihood

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

If there is no convergence, return to step 2.

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 $p(z|x)$
- K-means is a special case of EM.

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)

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

1. Check for convergence of params or likelihood

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

Jensen's inequality

$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x} | \theta)$$

$$= \log \sum_z p(\mathbf{x}, \mathbf{z} | \theta)$$

$$= \log \sum_z q(\mathbf{z} | \mathbf{x}) \frac{p(\mathbf{x}, \mathbf{z} | \theta)}{q(\mathbf{z} | \mathbf{x})}$$

Will lead to maximize this

$$\geq \sum_z q(\mathbf{z} | \mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{z} | \theta)}{q(\mathbf{z} | \mathbf{x})}$$

Maximizing this

$$\begin{aligned}
F(q, \theta) &= \sum_z q(\mathbf{z} | \mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{z} | \theta)}{q(\mathbf{z} | \mathbf{x})} \\
&= \sum_z q(\mathbf{z} | \mathbf{x}) \log p(\mathbf{x}, \mathbf{z} | \theta) - \sum_z q(\mathbf{z} | \mathbf{x}) \log q(\mathbf{z} | \mathbf{x}) \\
&= \langle \ell_c(\theta; \mathbf{x}, \mathbf{z}) \rangle_q + H_q
\end{aligned}$$

The first term is the expected complete log likelihood and the second term, which does not depend on θ , is the entropy.

Thus, in the M-step, maximizing with respect to θ for fixed q we only need to consider the first term:

$$\theta^{t+1} = \arg \max_{\theta} \langle \ell_c(\theta; \mathbf{x}, \mathbf{z}) \rangle_{q^{t+1}} = \arg \max_{\theta} \sum_z q(\mathbf{z} | \mathbf{x}) \log p(\mathbf{x}, \mathbf{z} | \theta)$$

Take-Home Messages

- The generative process of Gaussian Mixture Model
- Inferring cluster membership based on a learned GMM
- The general idea of Expectation-Maximization
- Expectation-Maximization for GMM