Gaussian Mixture Model

Nakul Gopalan
Georgia Tech

Some of the slides are based on slides from Jiawei Han Chao Zhang, Mahdi Roozbahani and Barnabás Póczos.

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:

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

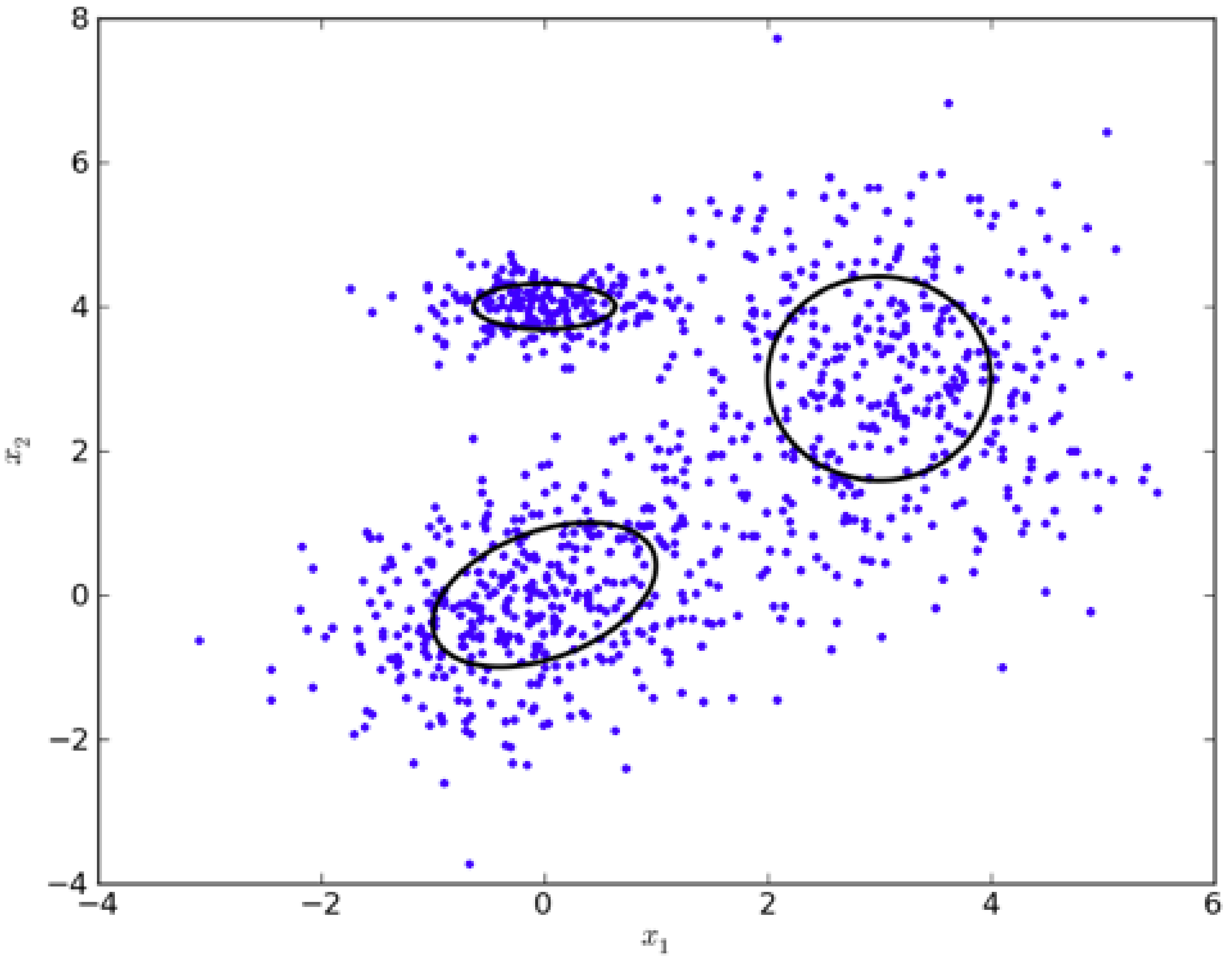
$$p(A=1) = \sum_{i=1}^{K} p(A=1, B_i) = \sum_{i=1}^{K} p(A|B_i)\, p(B_i)$$

	Tomorrow=Rainy	Tomorrow=Cold	P(Today)
Today=Rainy	4/9	2/9	[4/9 + 2/9] = 2/3
Today=Cold	2/9	1/9	[2/9 + 1/9] = 1/3
P(Tomorrow)	[4/9 + 2/9] = 2/3	[2/9 + 1/9] = 1/3	

P(Tomorrow = Rainy) =

Hard Clustering Can Be Difficult

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

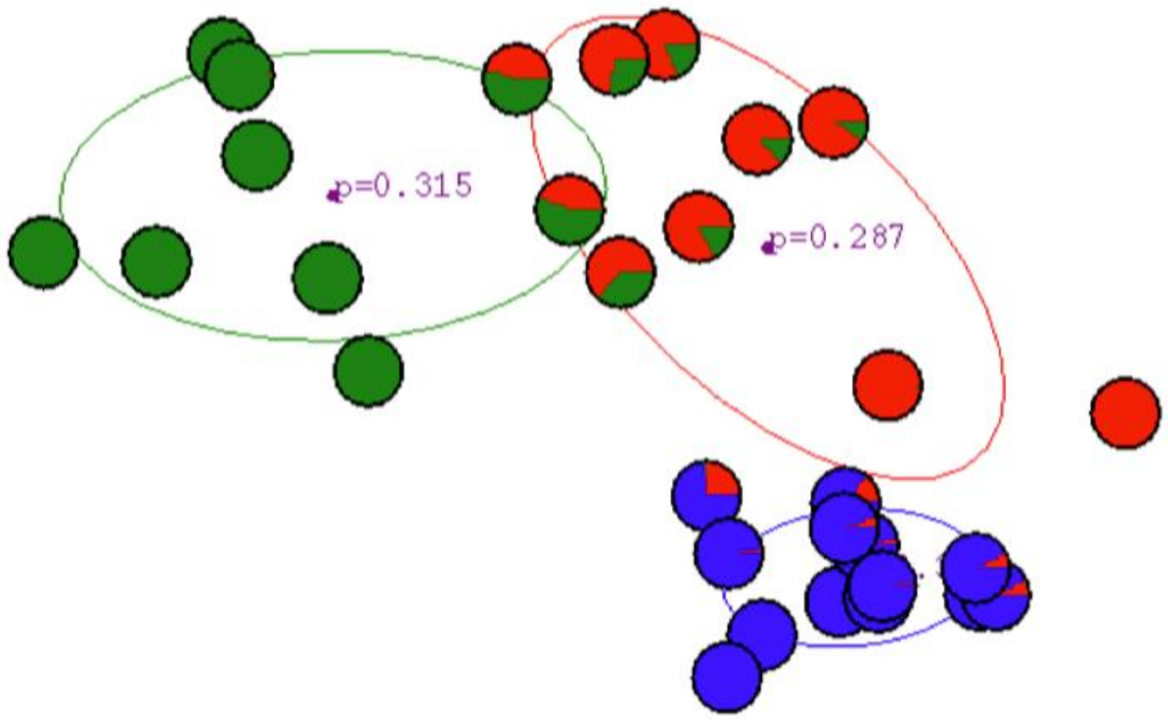


Towards Soft Clustering

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

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

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

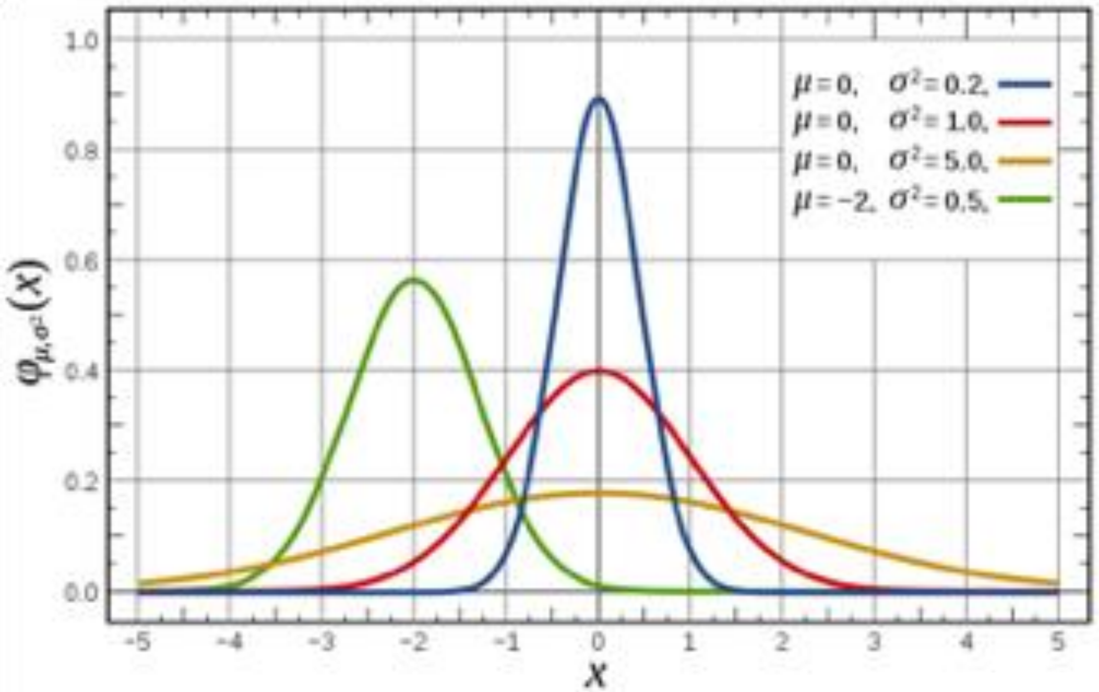


Outline

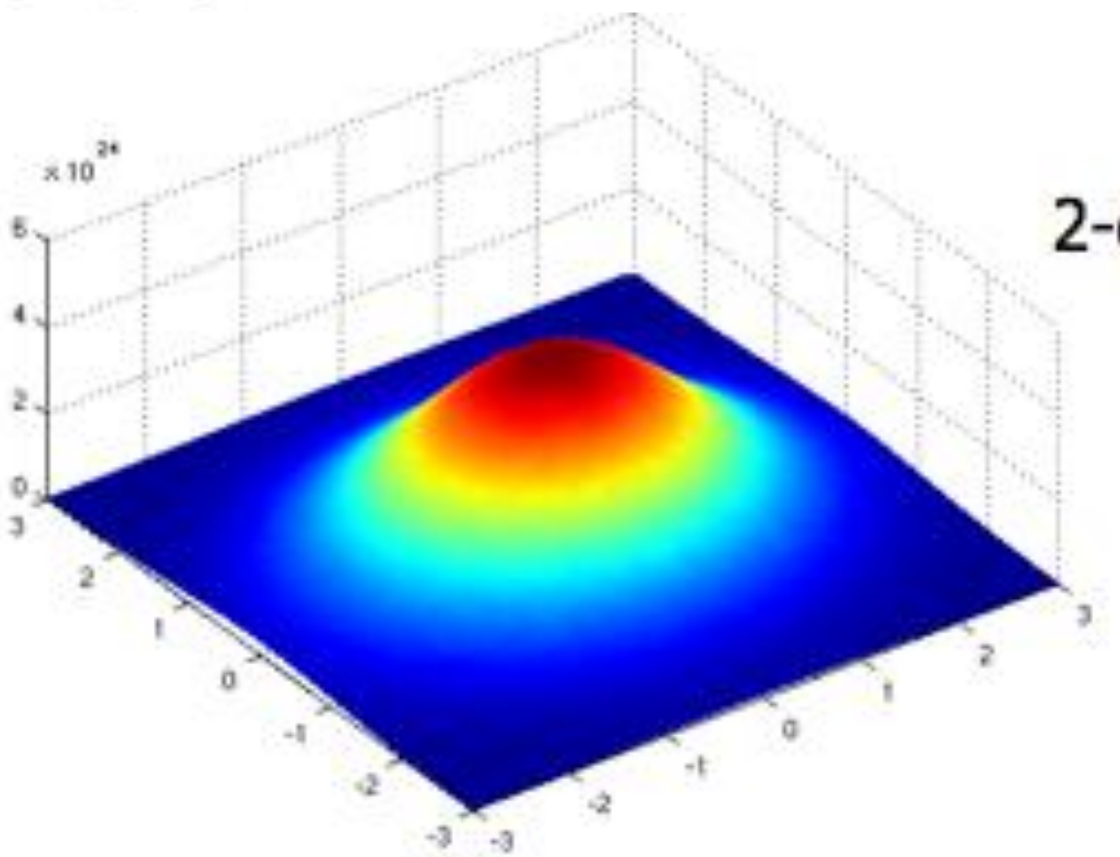
- Overview
- Gaussian Mixture Model
- The Expectation-Maximization Algorithm

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

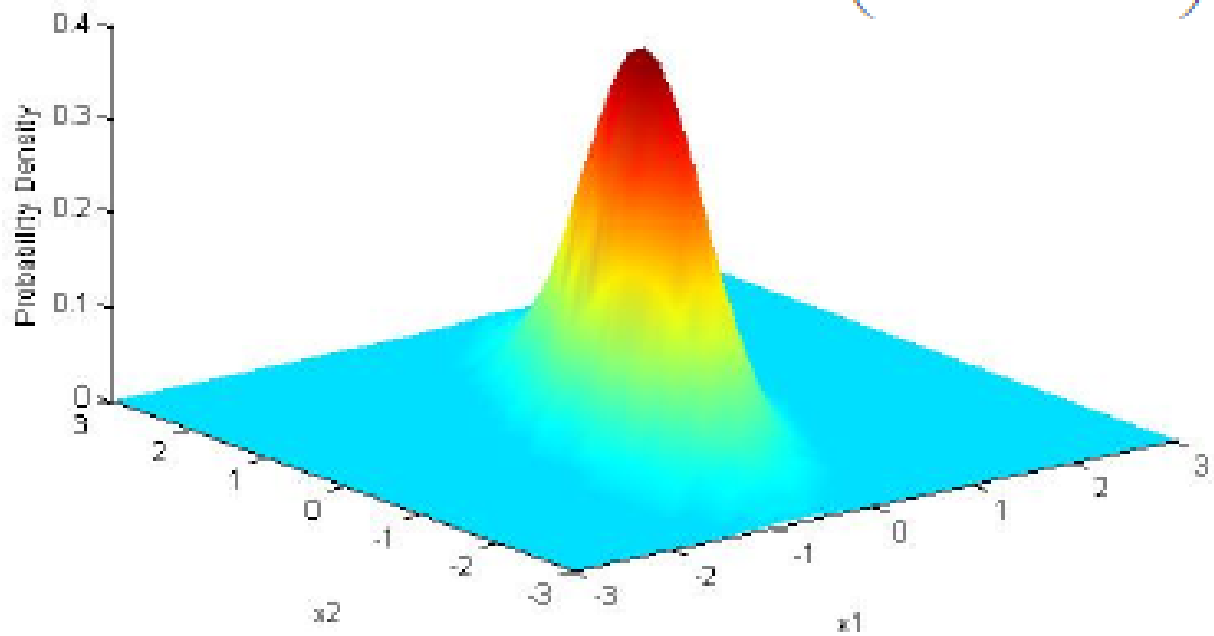
What is a Gaussian?

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

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

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

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



Mixture Models

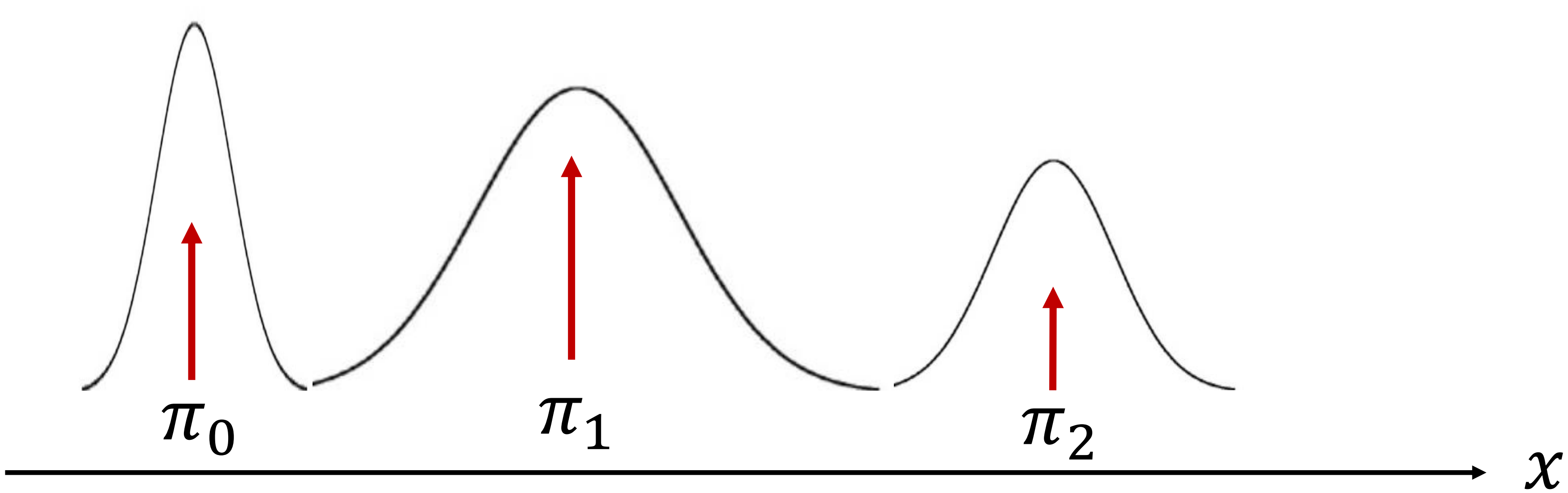
- Formally a Mixture Model is the weighted sum of a number of pdfs where the weights are determined by a distribution, π

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

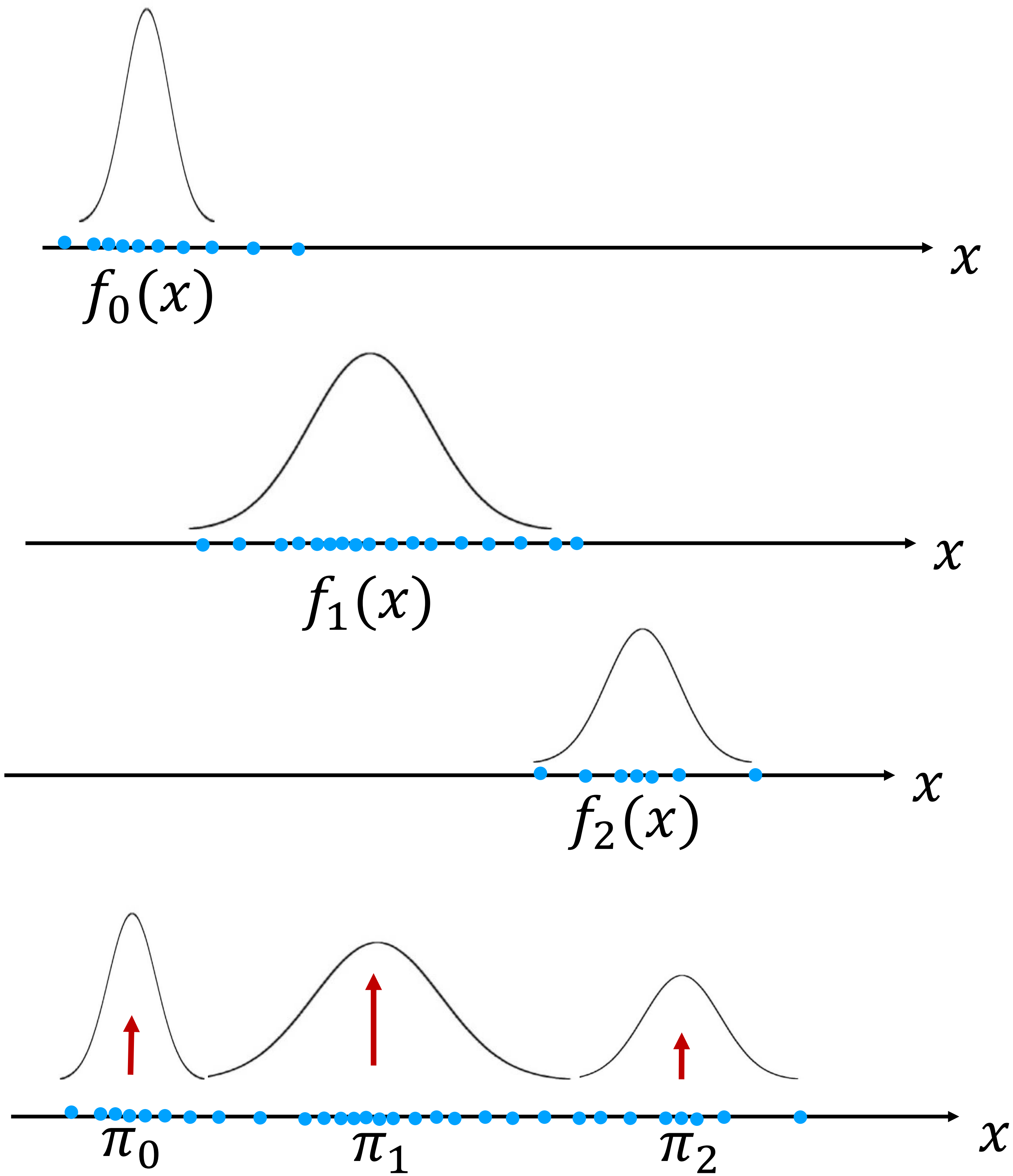
$$\text{where } \sum_{i=0}^{k} \pi_i = 1$$

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

What is **f** in GMM?



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



Why $p(x)$ is a pdf?

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

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

Why GMM?

It creates a new pdf for us to generate random variables. It is a generative model.

It clusters different components using a Gaussian distribution.

So it provides us the inferring opportunity. Soft assignment!!

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

Mixture Models are Generative

- Generative simply means dealing with joint probability $p(x, z)$

$$\mathrm{p(x)} = \pi_0 f_0(x) + \pi_1 f_1(x) + \cdots + \pi_k f_k(x)$$

Let's say $f(.)$ is a Gaussian distribution

$$\mathrm{p(x)} = \pi_0 N(X|\mu_0, \sigma_0) + \pi_1 N(X|\mu_1, \sigma_1) + \cdots + \pi_k N(X|\mu_k, \sigma_k)$$

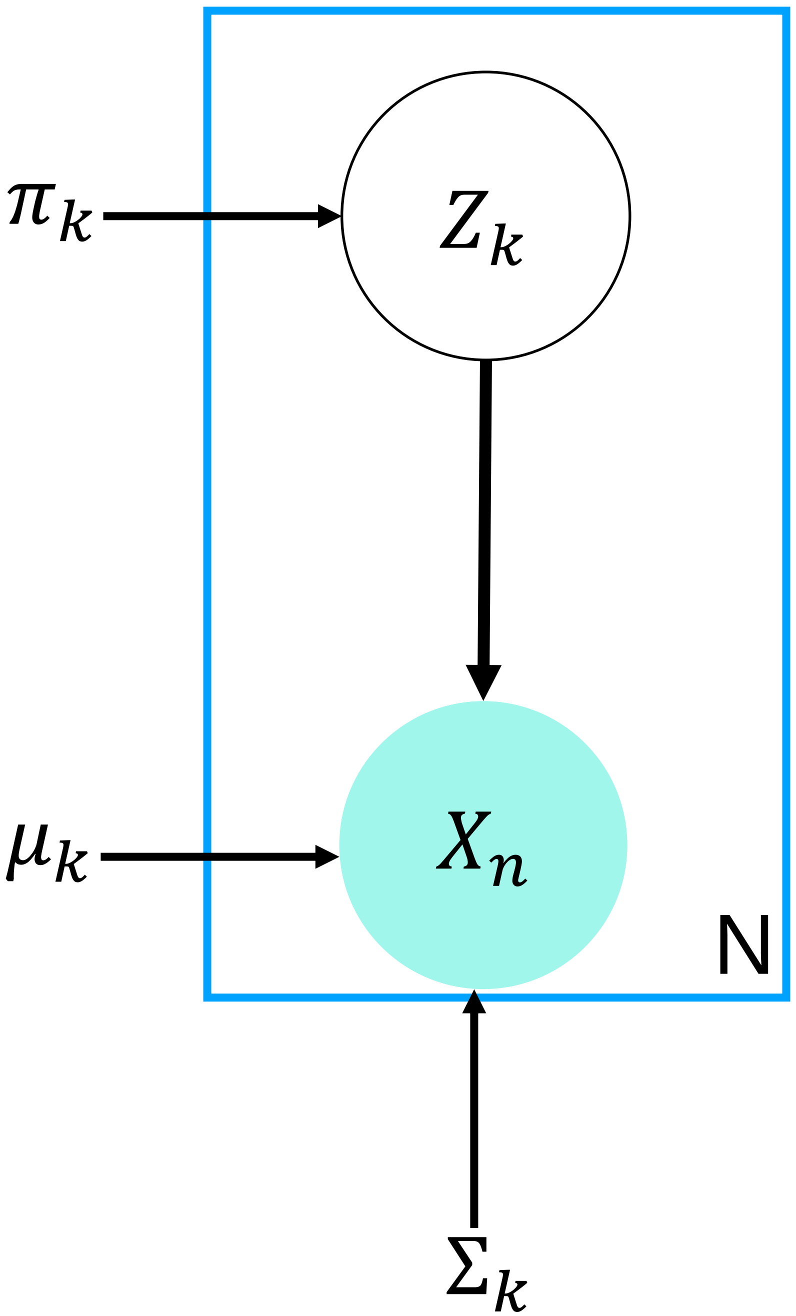
$$p(x) = \sum_k N(x|\mu_k, \sigma_k)\pi_k$$

$$p(x) = \sum_k p(x|z_k)p(z_k)$$

z_k is component k

$$p(x) = \sum_k p(x, z_k)$$

GMM with graphical model concept

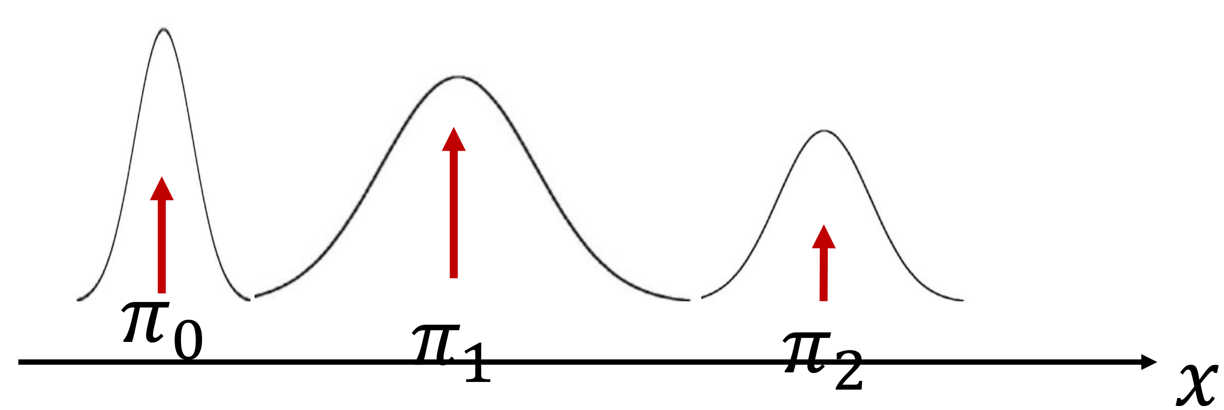


$$p(z_{nk}|\pi_k) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

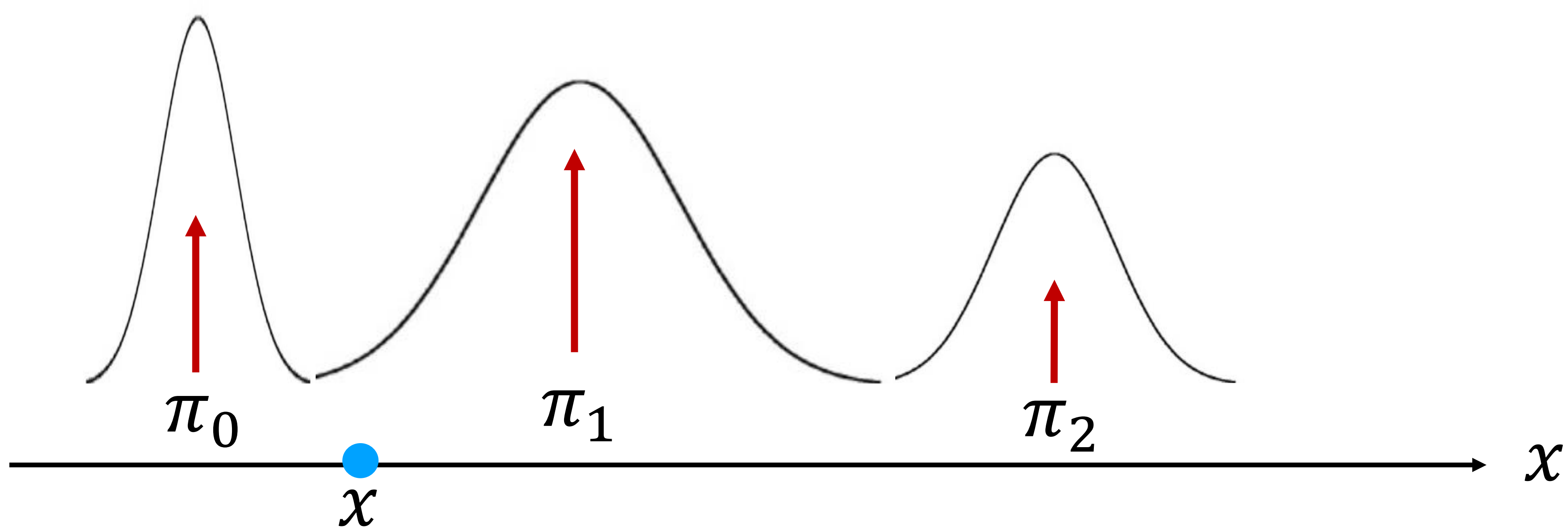
Z_k is the latent variable
1-of-K representation

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

Given $z, \pi, \mu,$ and Σ, what is the probability of x in component k



What is soft assignment?



What is the probability of a datapoint x in each component?

How many components we have here? 3

How many probability distributions? 3

What is the sum value of the 3 probabilities for each datapoint? 1

How to calculate the probability of datapoints in the first component (inferring)?

$$\mathrm{p(x)} = \pi_0 N(X|\mu_0, \sigma_0) + \pi_1 N(X|\mu_1, \sigma_1) + \pi_2 N(X|\mu_2, \sigma_2)$$

Let's calculate the responsibility of the first component among the rest for one point *x*

Let's call that τ_0

$$\tau_0 = \frac{N(X|\mu_0, \sigma_0)\pi_0}{N(X|\mu_0, \sigma_0)\pi_0 + N(X|\mu_1, \sigma_1)\pi_1 + N(X|\mu_2, \sigma_2)\pi_2}$$

$$\tau_0 = \frac{p(x|z_0)p(z_0)}{p(x|z_0)p(z_0) + p(x|z_1)p(z_1) + p(x|z_1)p(z_1)}$$

$$\tau_0 = \frac{p(x, z_0)}{\sum_{k=0}^{k=2} p(x, z_k)} = \frac{p(x, z_0)}{p(x)} = p(z_0|x)$$

Given a datapoint x, what is probability of that datapoint in component 0

If I have 100 datapoints and 3 components, what is the size of τ? 100X3

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.

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

Mixtures of Gaussians

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

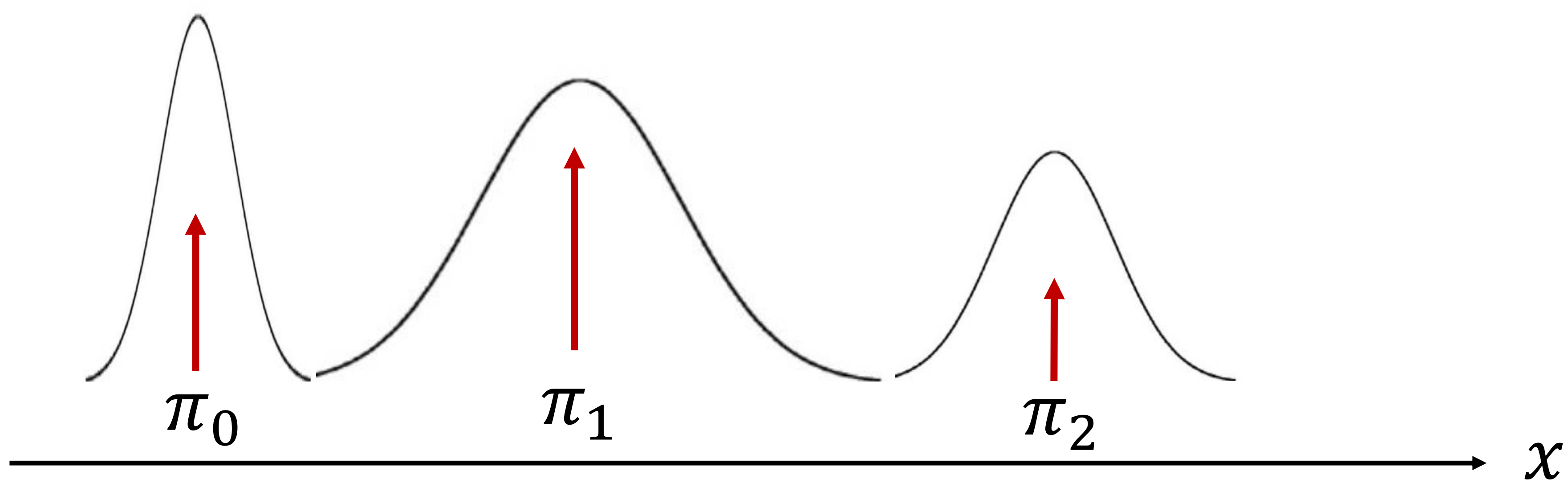
AND

Picking data from that specific mixture component = $\mathrm{p}(x|z)$

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

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

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



What are GMM parameters?

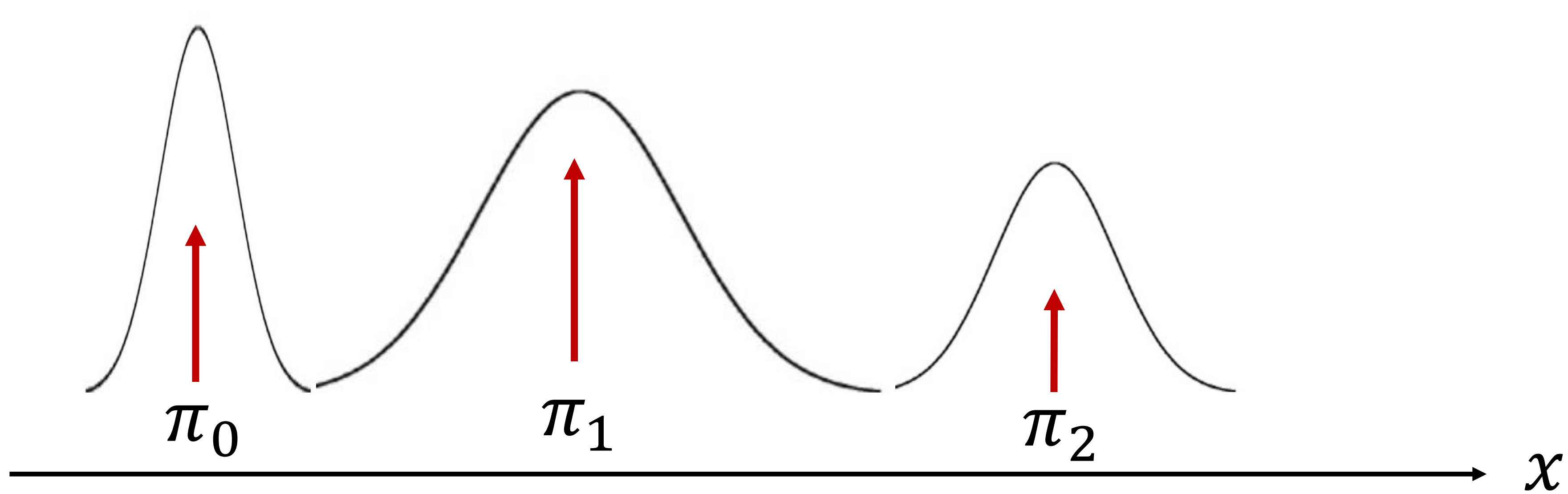
Mean μ_k Variance σ_k Size π_k

Marginal probability distribution

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

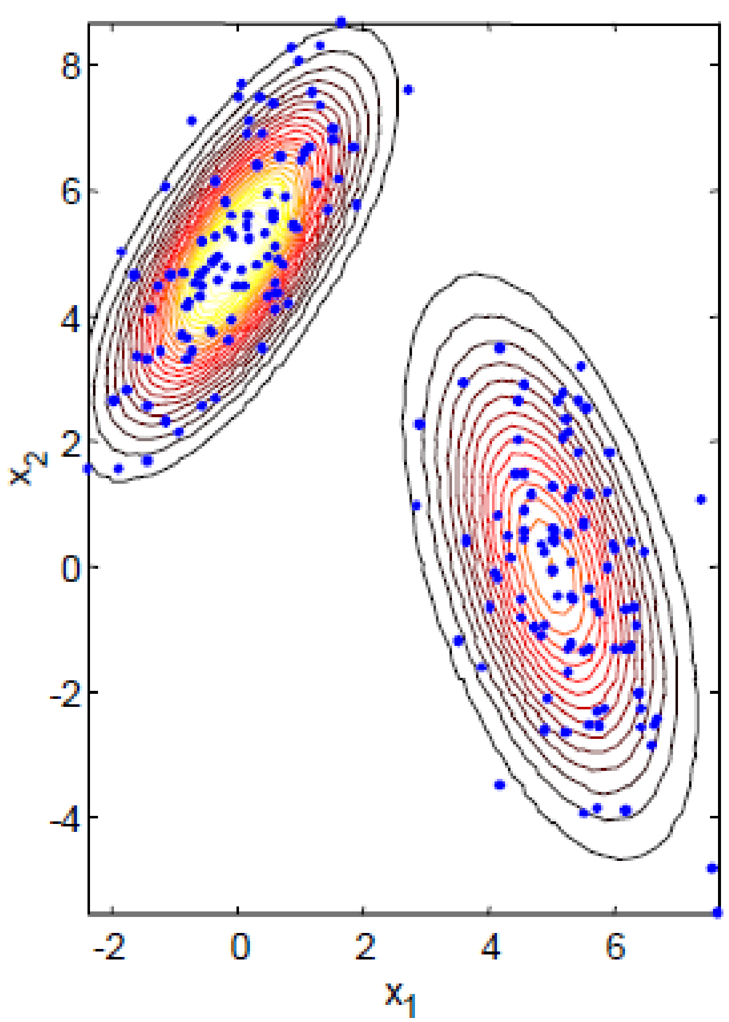
$p(z_k|\theta) = \pi_k$ Select a mixture component with probability π

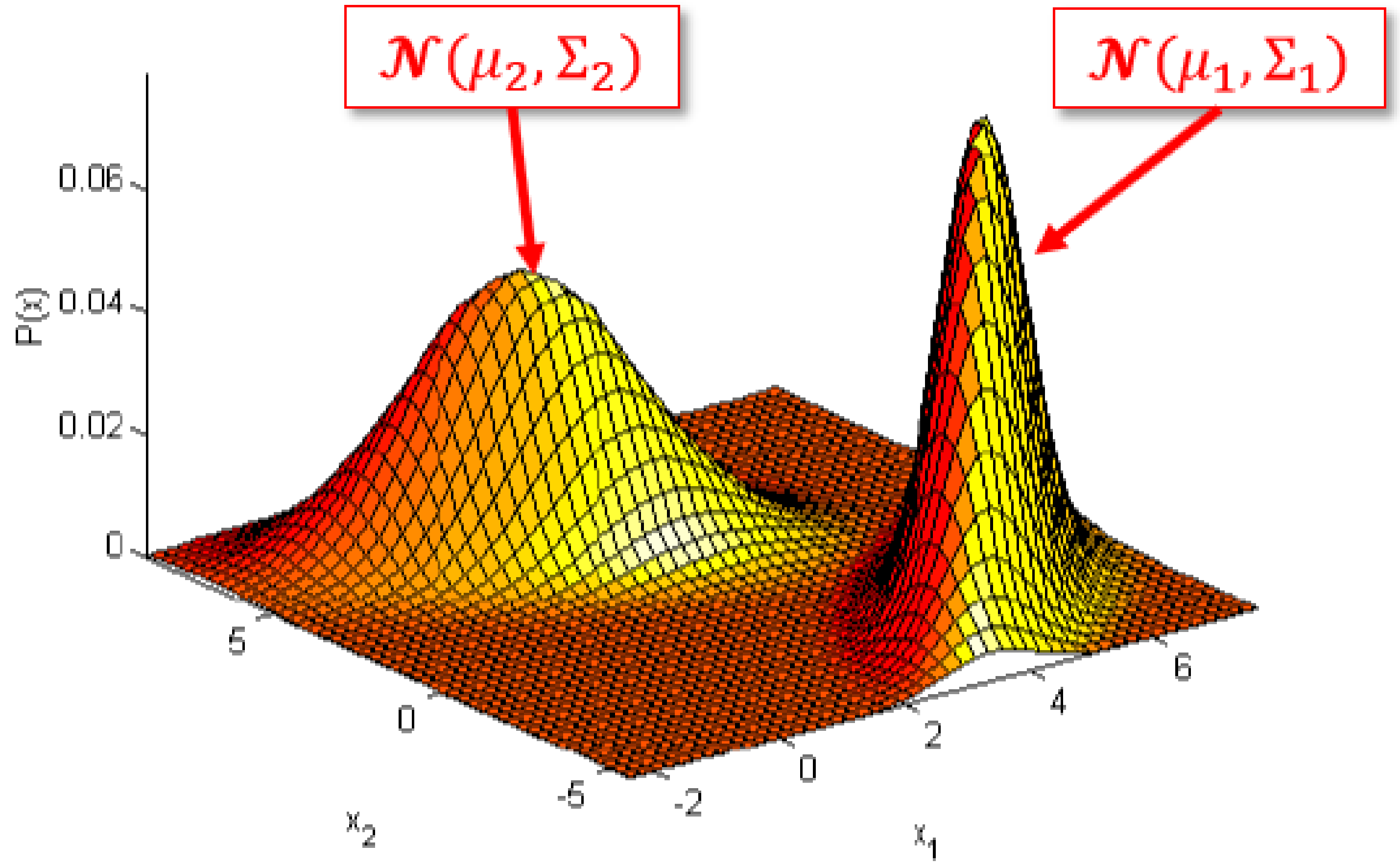
$p(x|z_k, \theta) = N(x|\mu_k, \sigma_k)$ Sample from that component's Gaussian



How about GMM for multimodal distribution?

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





Gaussian Mixture Model

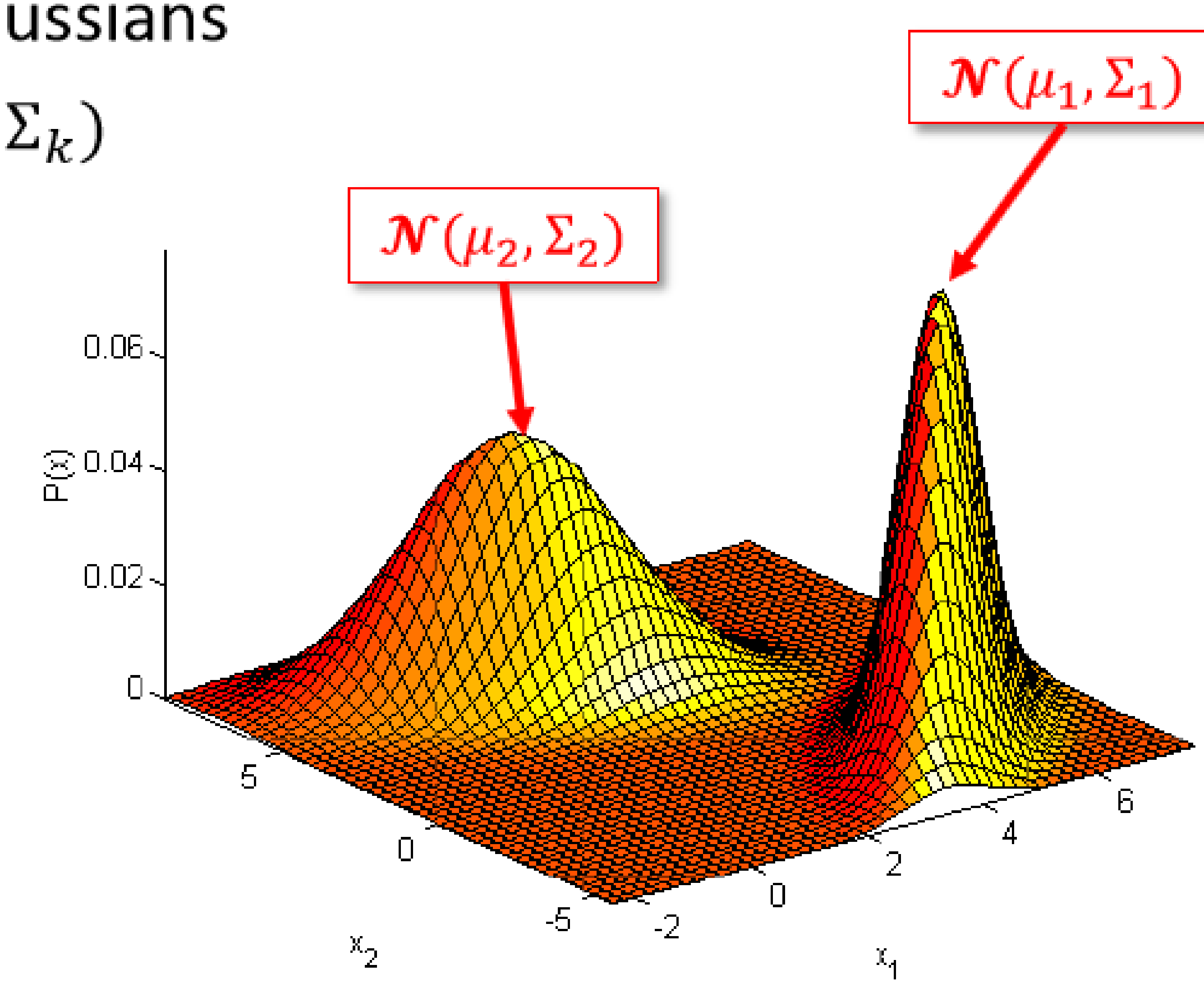
- A density model $p(X)$ may be multi-modal: model it as a mixture of uni-modal distributions (e.g. Gaussians)

- Consider a mixture of K Gaussians
  - $p(X) = \sum_{k=1}^{K} \pi_k \boldsymbol{\mathcal{N}}(X|\mu_k, \Sigma_k)$

mixing proportion

mixture Component

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



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

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

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

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.

Well, we don't know $\pi_k, \mu_k, \Sigma_\mathrm{k}$
What should we do?

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

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

Let's identify a likelihood function, why?

Because we use likelihood function to optimize the probabilistic model parameters!

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

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

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

- 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.

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

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

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

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

Maximum Likelihood of a GMM

- Optimization of covariance

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

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

Maximum Likelihood of a GMM

- Optimization of mixing term

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

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

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

MLE of a GMM

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

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

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

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

Not a closed form solution!!
τ is not known exactly
What next?

Outline

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

EM for GMMs

- E-step: Evaluate the Responsibilities

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

EM for GMMs

- M-Step: Re-estimate Parameters

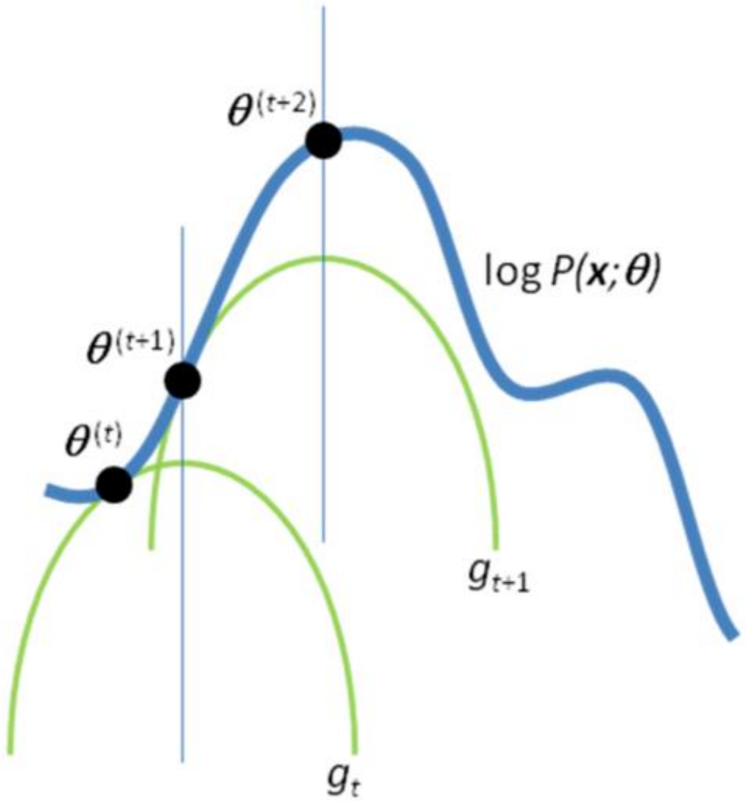
$$\boxed{\mu_k^{new} = \frac{\sum_{n=1}^{N} \tau(z_{nk}) x_n}{N_k}}$$

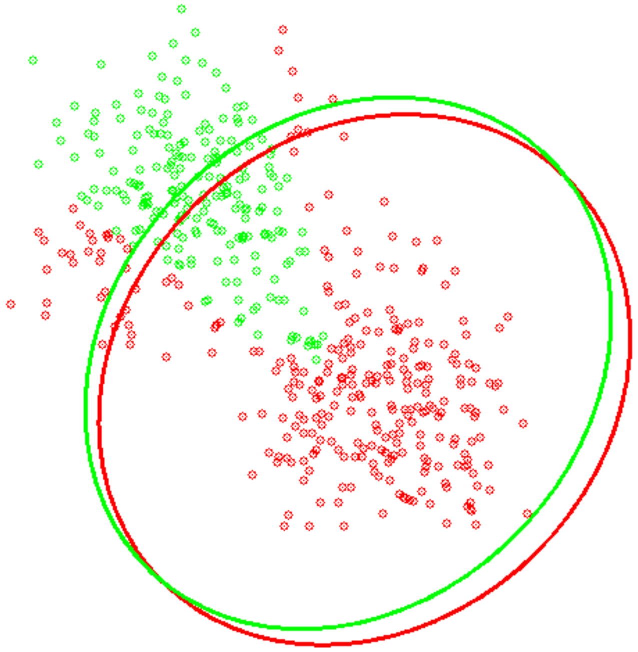
$$\boxed{\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \tau(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T}$$

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

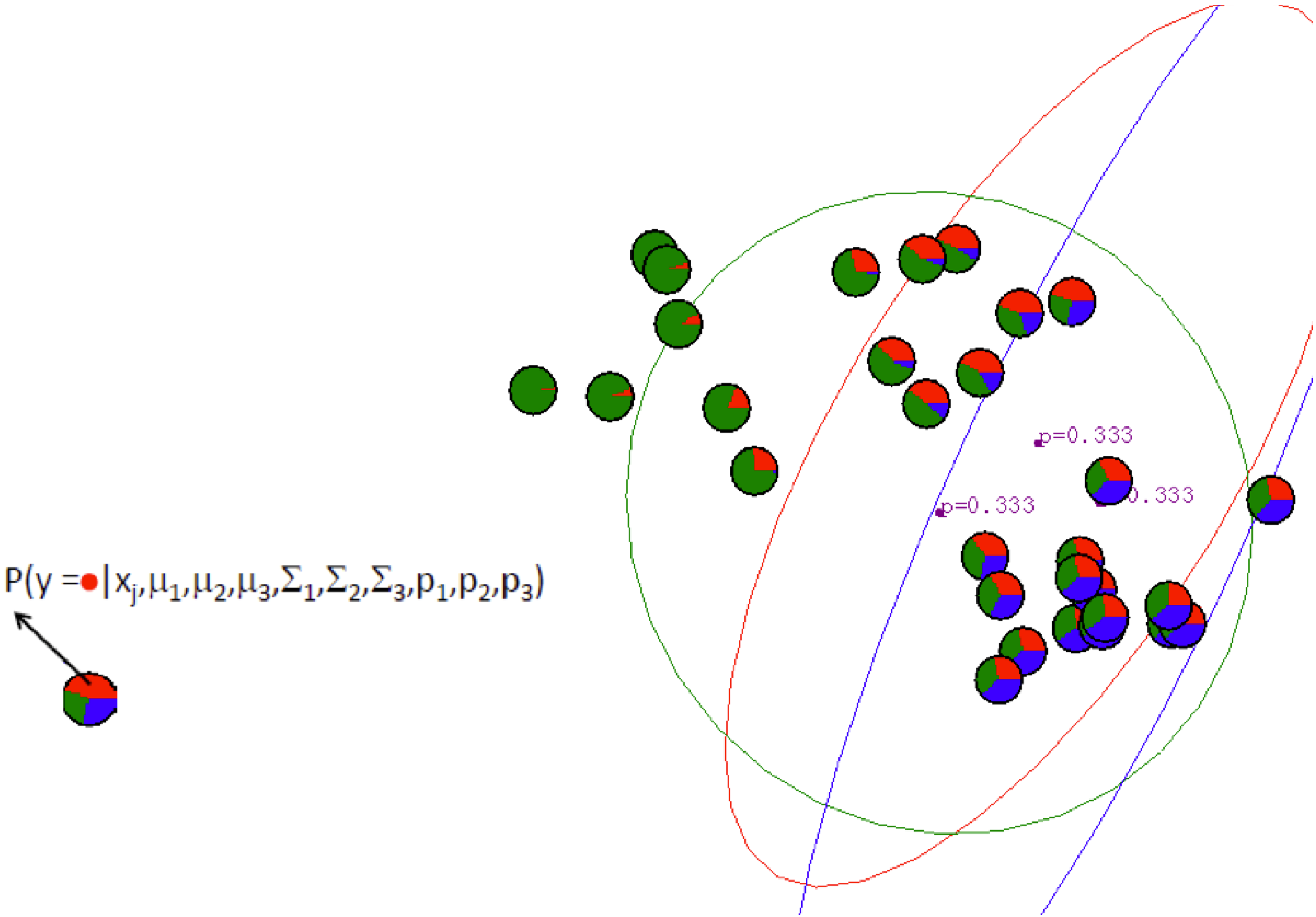
Expectation Maximization

- 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.



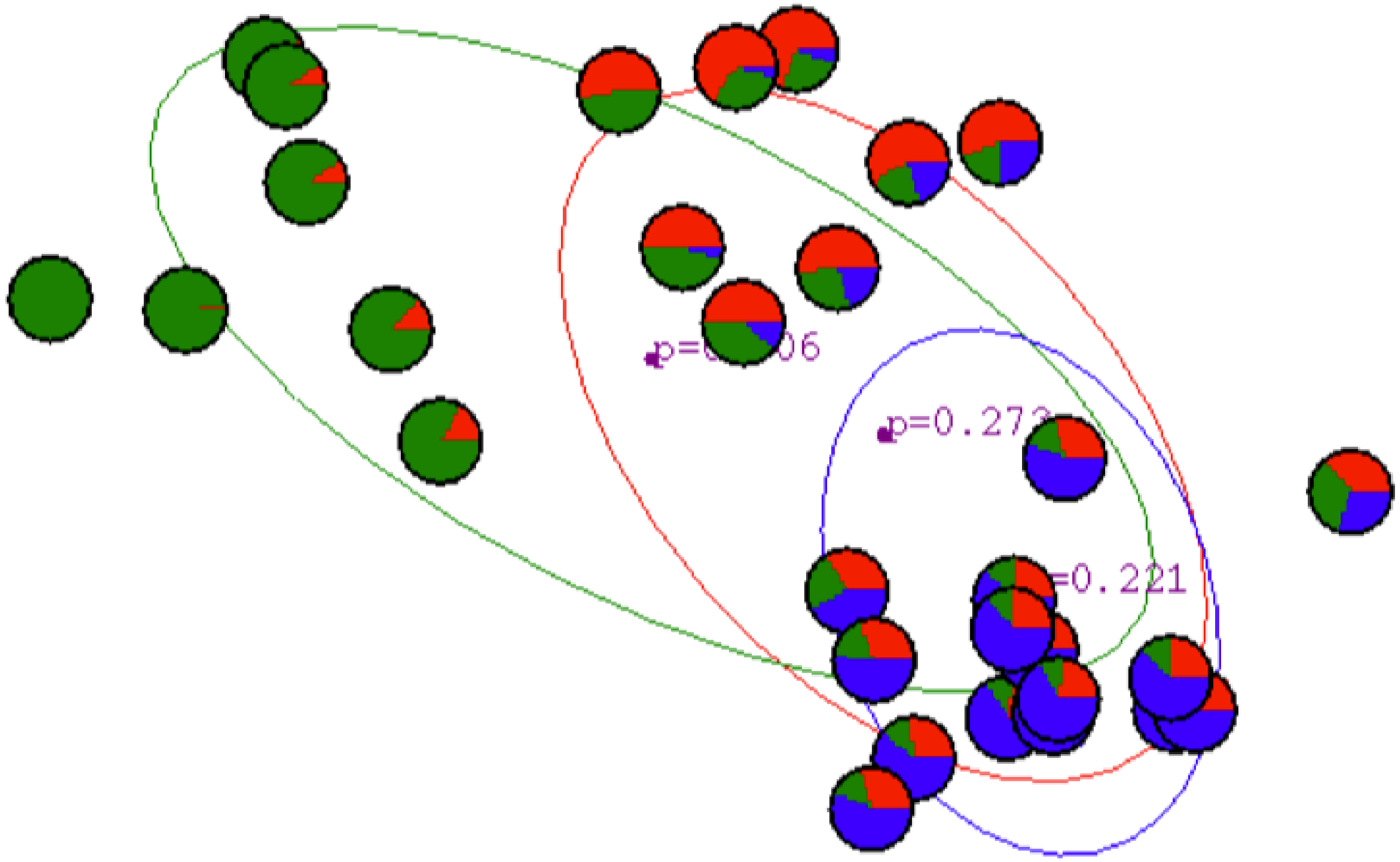


EM for Gaussian Mixture Model:



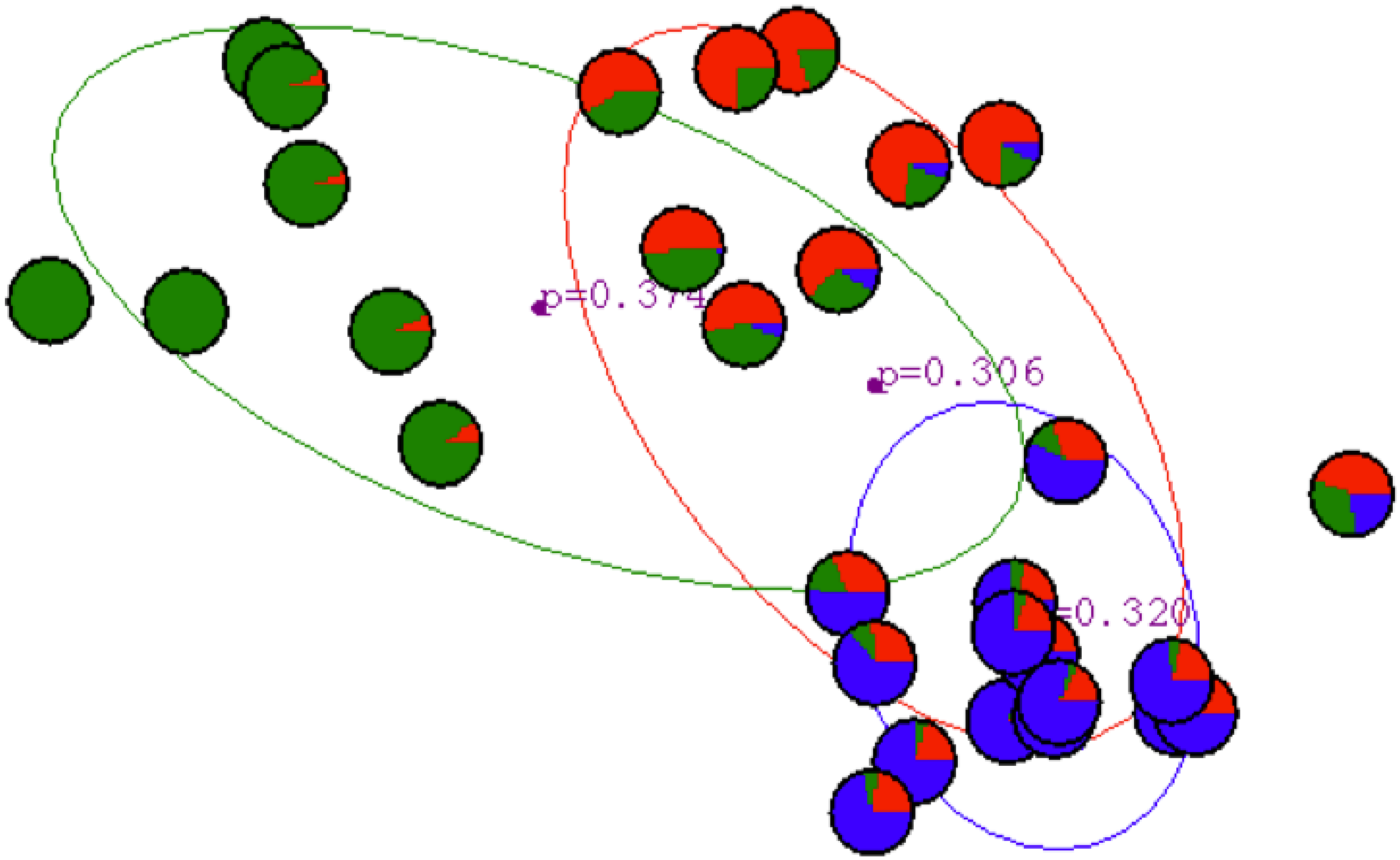
EM for Gaussian Mixture Model: Example

After 1st iteration



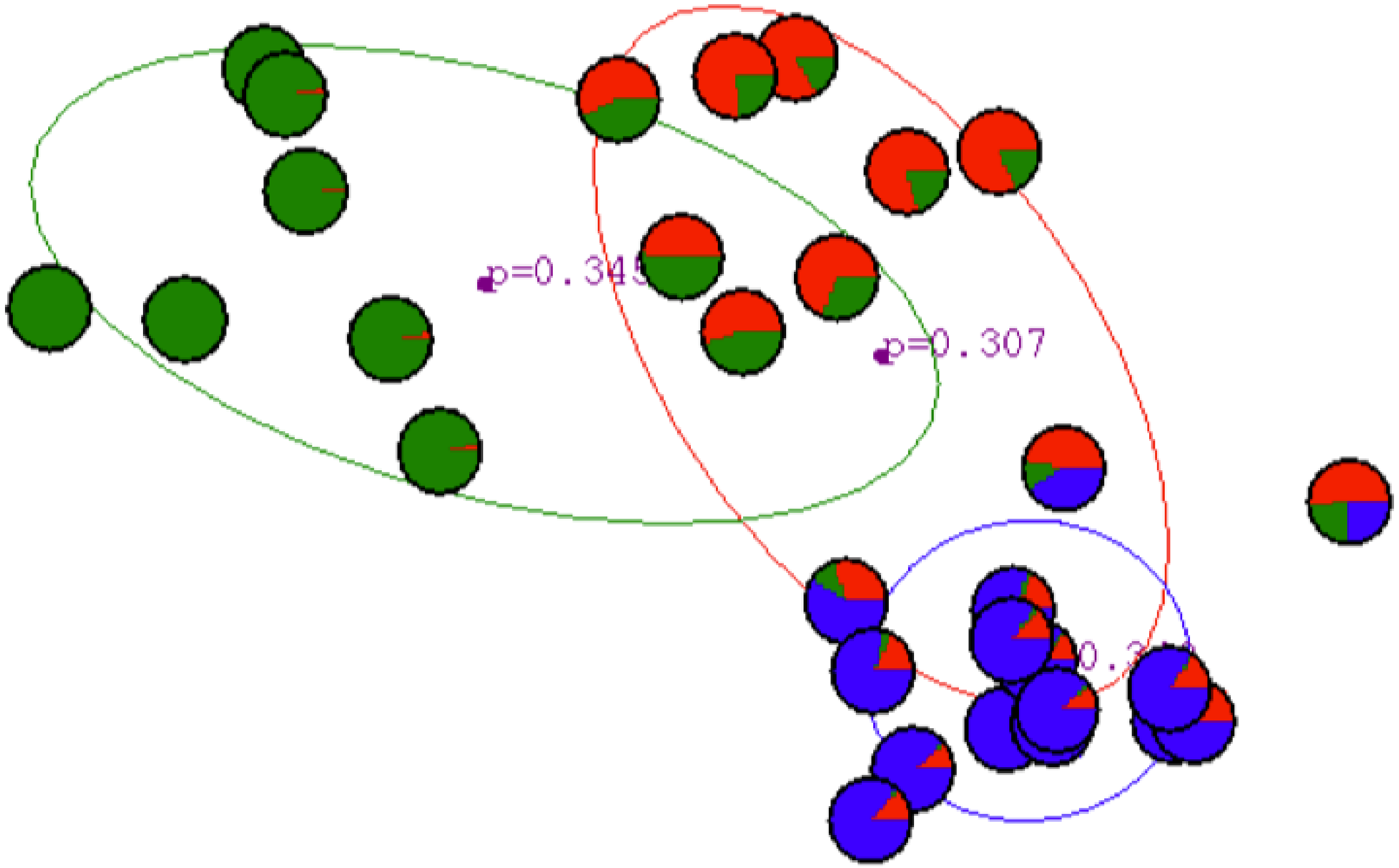
EM for Gaussian Mixture Model: Example

After 2nd iteration



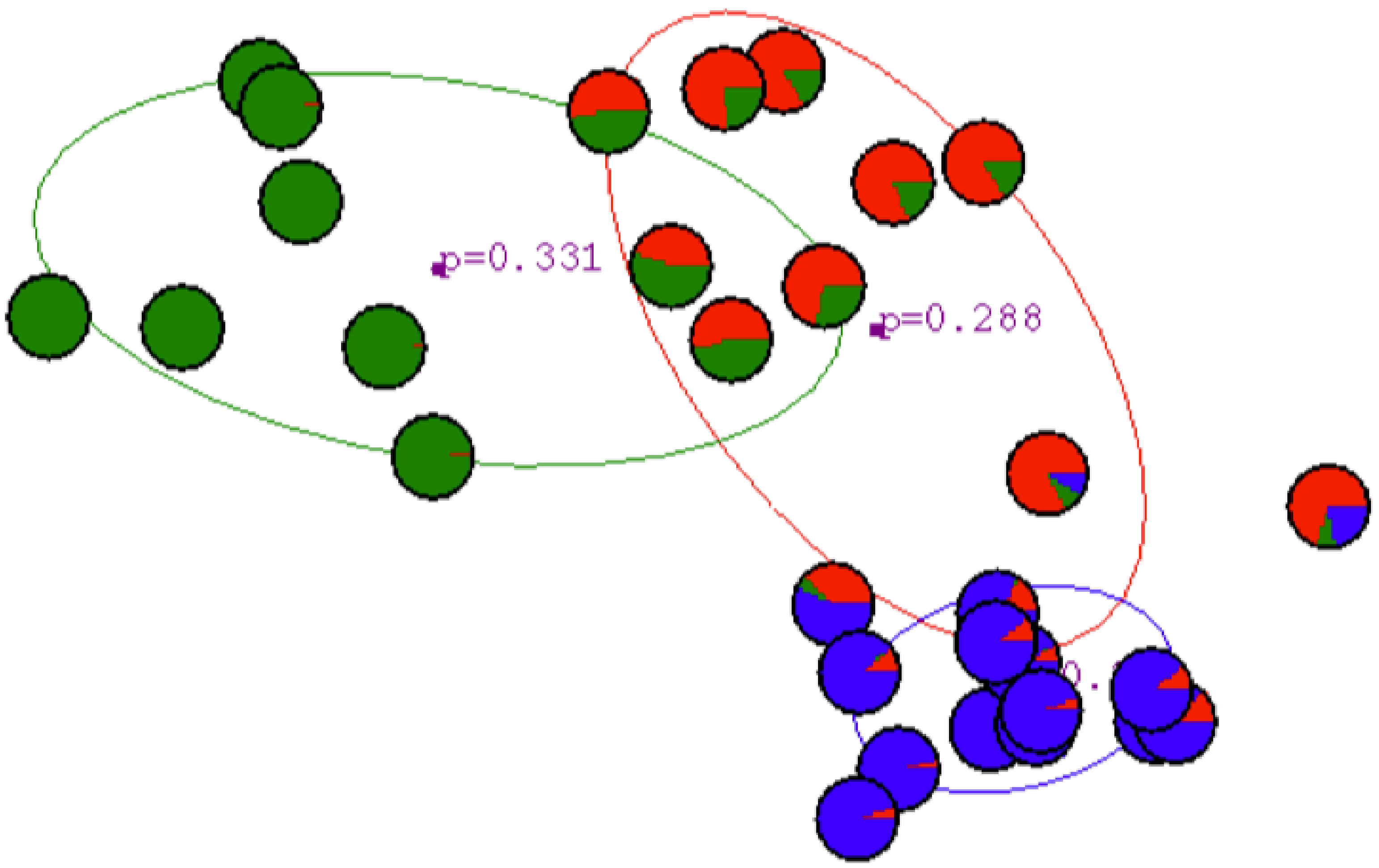
EM for Gaussian Mixture Model: Example

After 3rd iteration



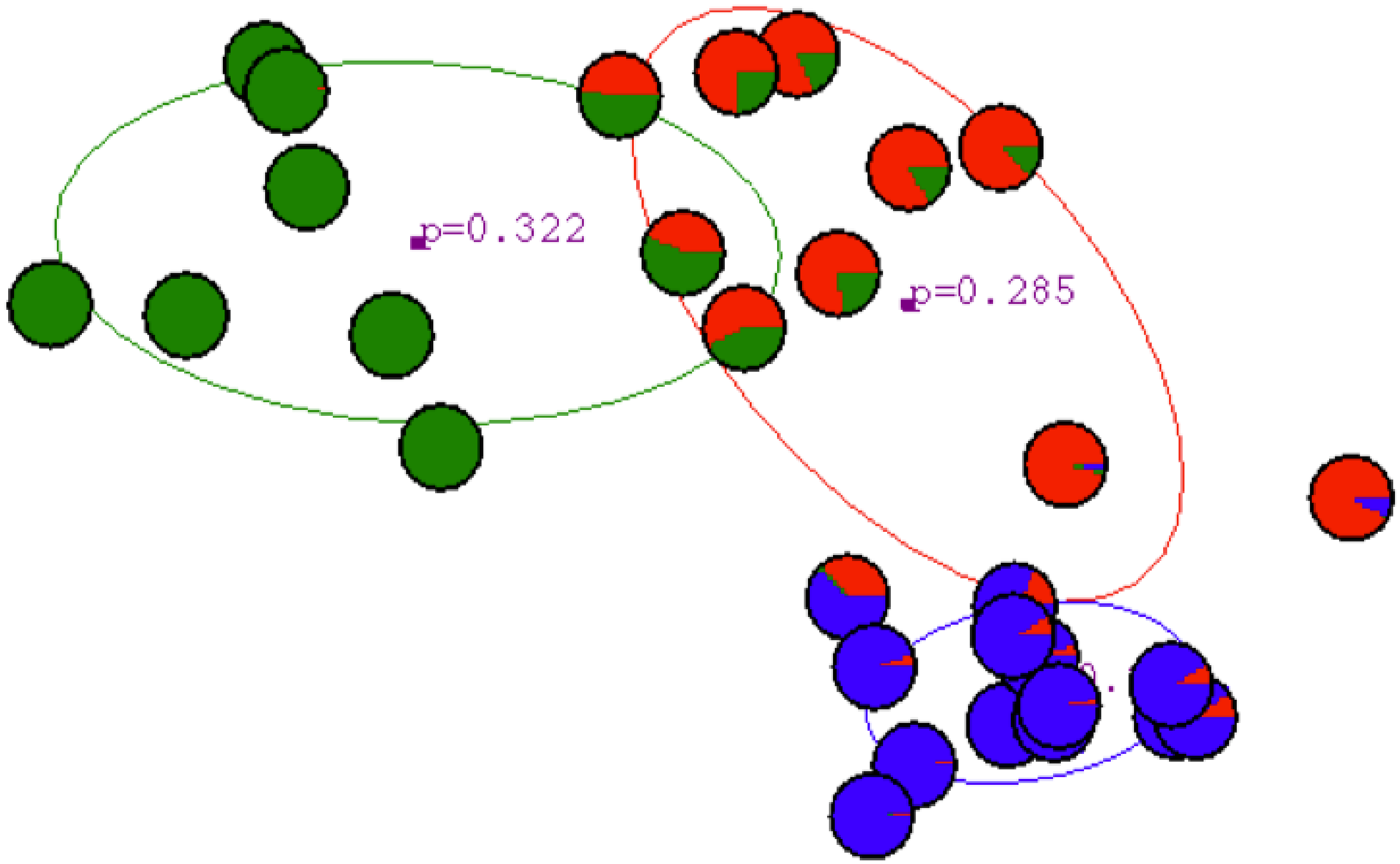
EM for Gaussian Mixture Model: Example

After 4th iteration



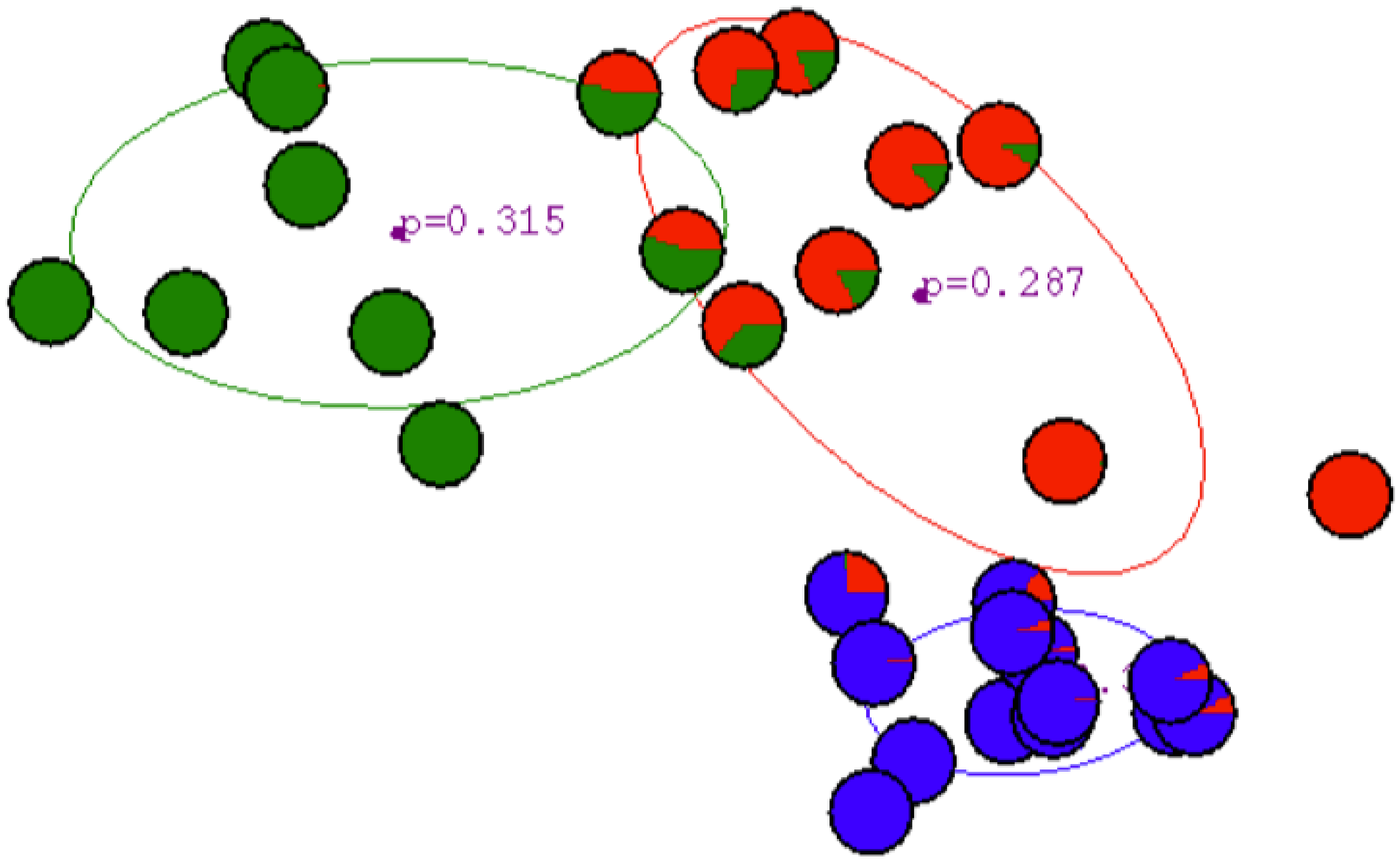
EM for Gaussian Mixture Model: Example

After 5th iteration



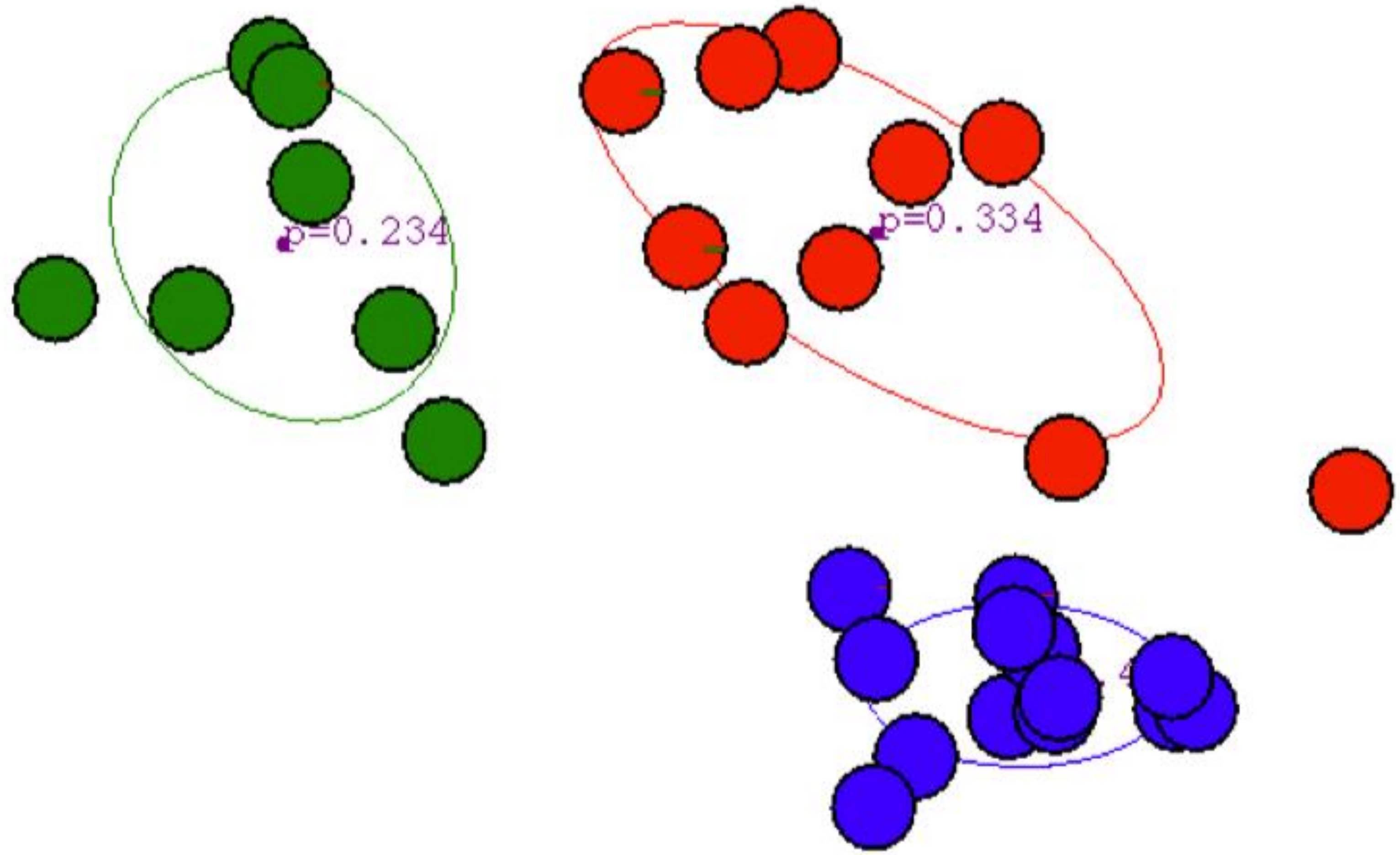
EM for Gaussian Mixture Model: Example

After 6th iteration



EM for Gaussian Mixture Model: Example

After 20th iteration



Demo

- Demo link: https://lukapopijac.github.io/gaussian-mixture-model/

EM Algorithm for GMM (matrix form)

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 $\boldsymbol{\mu}_j$, covariances Σ_j and mixing coefficients $\boldsymbol{\pi}_j$, and evaluate the initial value of the log likelihood.
2. E step. Evaluate the responsibilities using the current parameter values

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

Book : C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

EM for GMMs

- M-Step: Re-estimate Parameters

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

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

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

EM Algorithm for GMM (matrix form)

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

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

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

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

4. Evaluate log likelihood

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

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)

$$\theta^{new} = \text{argmax}_\theta \sum_Z p(Z|X, \theta^{old}) \ln p(X, Z|\theta) = argmax_\theta \, \mathbb{E}[\ln(p(x, z|\theta)]$$

4. Check for convergence of params or likelihood

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

Jensen’s inequality

$$\begin{aligned}
\ell(\theta; x) &= \log p(x \mid \theta) \\
&= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)}
\end{aligned}$$

Will lead to maximizing this

Maximizing this

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

$$= \sum_z q(z \mid x) \log p(x,z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x)$$

$$= \langle \ell_c(\theta; x,z) \rangle_q + H_q$$

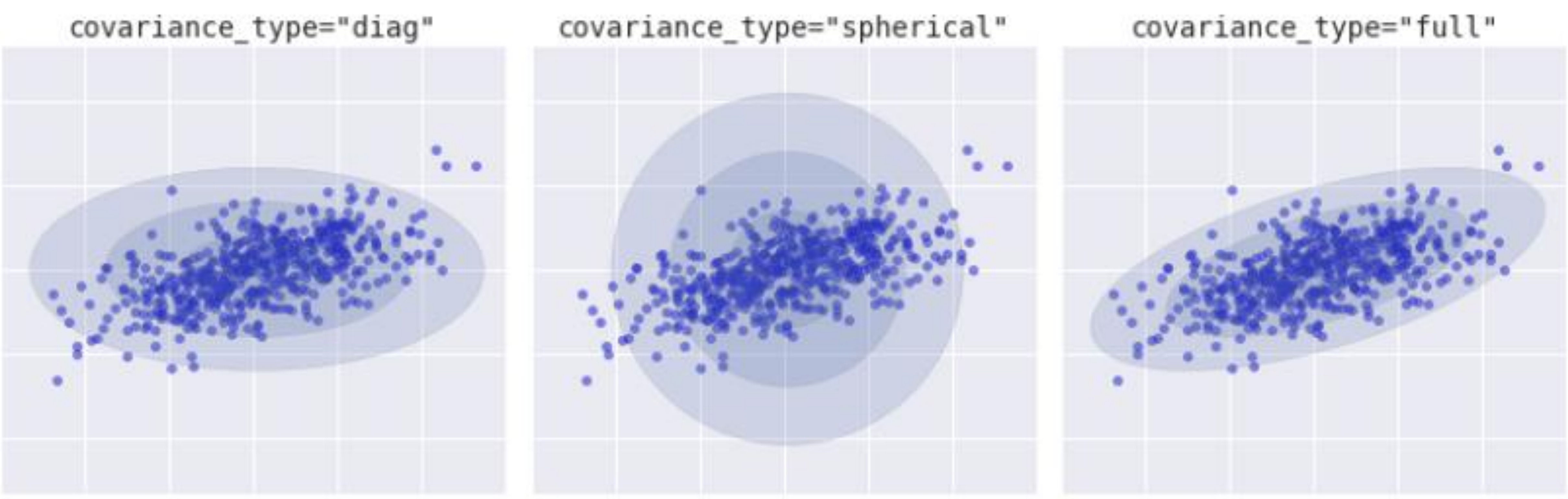
The first term is the expected complete log likelihood and the second term, which does not depend on θ, is the entropy.

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

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

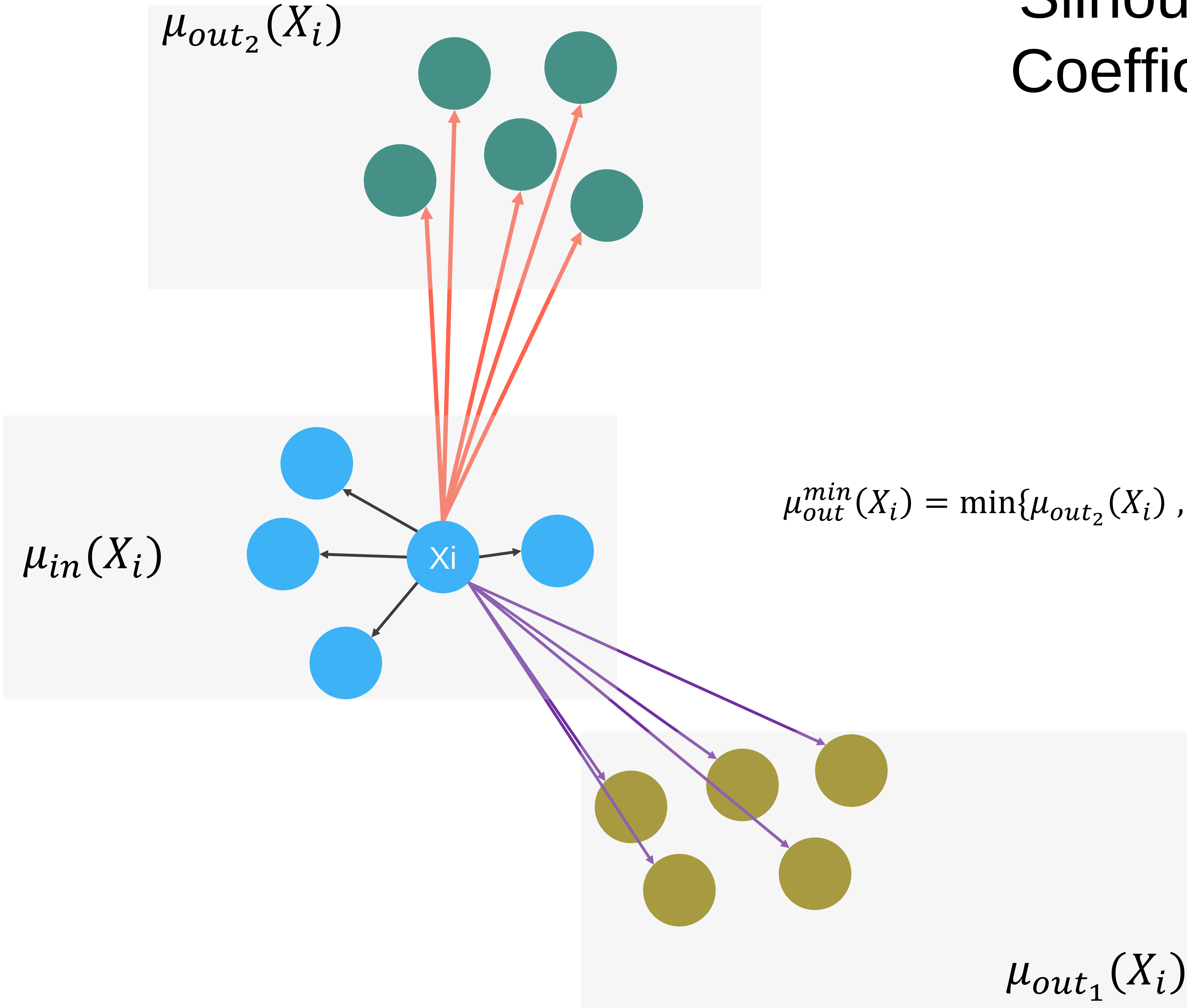
EM for Gaussian Mixture Model: Example

covariance_type="diag“ or "spherical“ or “full”



Source: Python Data Science Handbook by Jake VanderPlas

Silhouette Coefficient



$$\mu_{out}^{min}(X_i) = \min\{\mu_{out_2}(X_i)\ ,\mu_{out_1}(X_i)\}$$

Silhouette Coefficient

Define the silhoutte coefficient of a point $\mathbf{x}_i$ as

$$s_i = \frac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\left\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\right\}}$$

where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from $\mathbf{x}_i$ to points in its own cluster $\hat{y}_i$:

$$\mu_{in}(\mathbf{x}_i) = \frac{\sum_{\mathbf{x}_j \in C_{\hat{y}_i}, j \neq i} \delta(\mathbf{x}_i, \mathbf{x}_j)}{n_{\hat{y}_i} - 1}$$

and $\mu_{out}^{\min}(\mathbf{x}_i)$ is the mean of the distances from $\mathbf{x}_i$ to points in the closest cluster:

$$\mu_{out}^{\min}(\mathbf{x}_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{\mathbf{y} \in C_j} \delta(\mathbf{x}_i, \mathbf{y})}{n_j} \right\}$$

The Silhouette Coefficient for clustering C: $SC = \frac{1}{n} \sum_{i=1}^{n} s_i.$

SC close to 1 implies a good clustering (Points are close to their own clusters but far from other clusters)

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
- Silhouette Coefficient