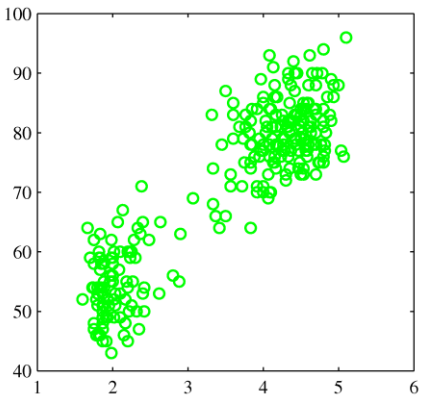
Expectation Maximization (EM) method & GMM

CE-477: Machine Learning - CS-828: Theory of Machine Learning
Sharif University of Technology
Fall 2024

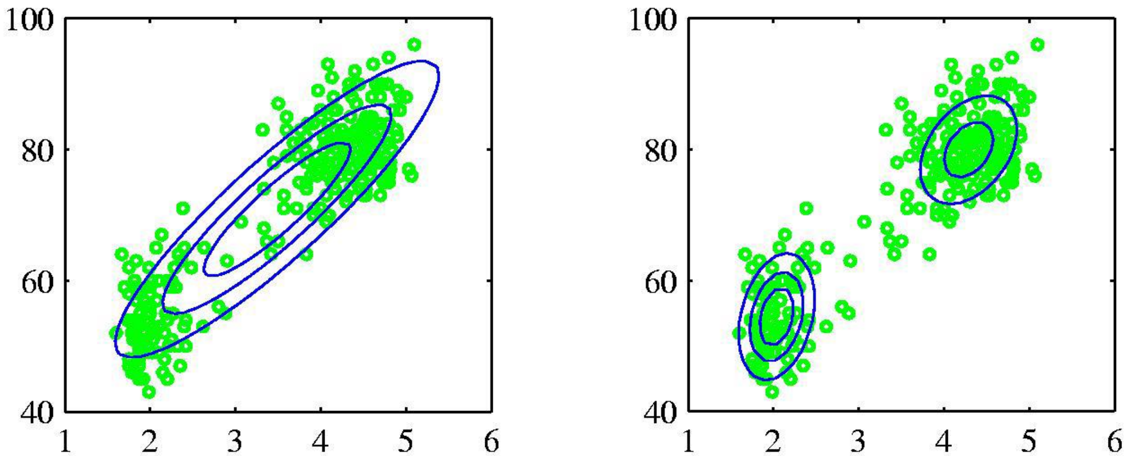
Fatemeh Seyyedsalehi

Gaussian mixture model (GMM)

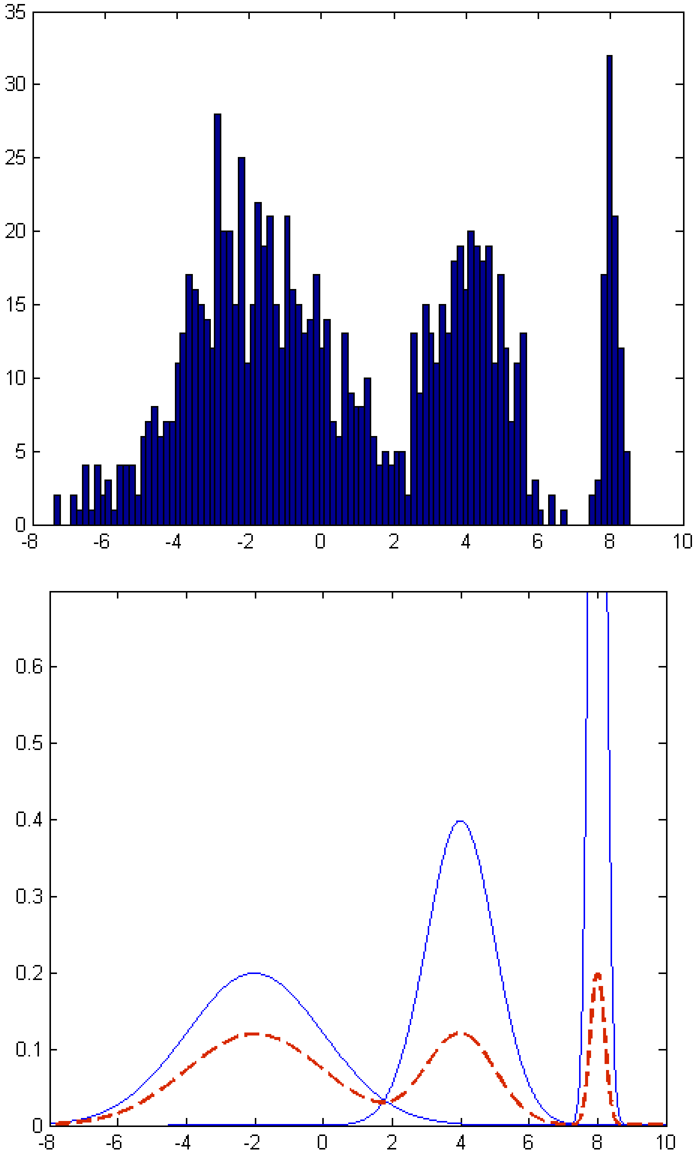
- Remember the density estimation problem from the observed data:



- Which densities are more fitted to the data?

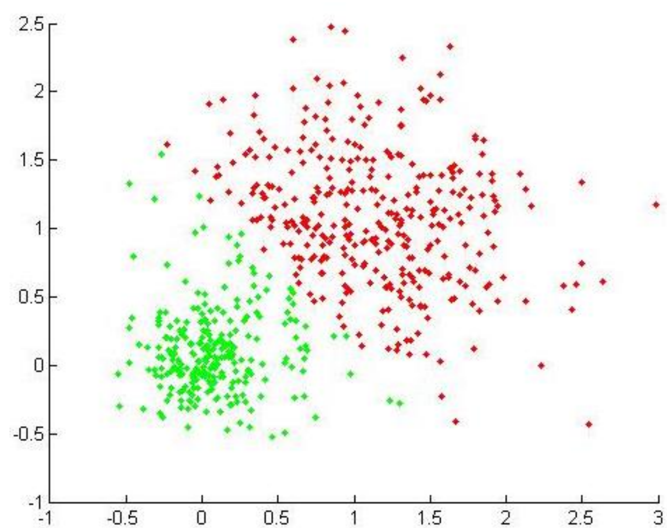


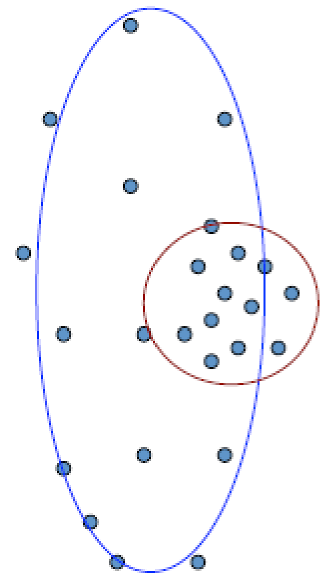
GMM: 1-D Example



Clustering view

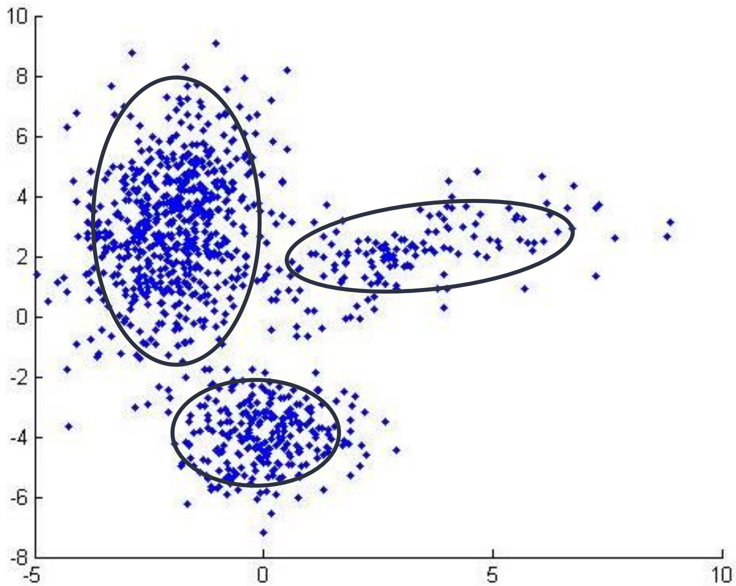
- Clusters may overlap
- Some clusters may be “wider” than others
- Can we model this explicitly?
- With what **probability** is a point from a cluster?





Gaussian mixture model (GMM)

- A generative story for data
- Each data point assumed to have been sampled from a generative process:
  - Choose component j with probability $P(z = j)$
    - A Multinomial distribution with parameters ϕ
  - Generate datapoint according to this component, i.e. $\mathrm{N}\left(\boldsymbol{x}\middle|\mu_j,\ \Sigma_j\right)$



- Framework for finding more complex probability distributions
  - We can learn new insight about the data

Gaussian mixture model (GMM)

- Therefore, we need to model the data by the joint distribution $P(\boldsymbol{x}, z)$.
- However, $z^{(i)}$ is a latent random variable, meaning that it's hidden or unobserved.
  - The observed data: $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$
- Consider the following marginal distribution:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{K} P(z = j; \boldsymbol{\theta})\, p(\boldsymbol{x}|z = j; \boldsymbol{\theta} \quad)$$

Gaussian mixture model (GMM)

- The likelihood of the observed data:
  - According to our generative story, parameters of our model include, $\boldsymbol{\theta}$ =$\{\mu, \Sigma, \phi\}$

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} p(\boldsymbol{x}^{(i)}|z^{(i)};\ \mu, \Sigma) p\left(z^{(i)}; \phi\right)$$

- Setting to zero the derivatives of this formula with respect to parameters
  - No closed form solution!

Gaussian mixture model (GMM)

- If we observed latent variables:
  - The observed data: $\{\boldsymbol{x}^{(i)}, z^{(i)}\}_{i=1}^{N}$
  - The maximum likelihood problem would be easy
    - GDA model

$$\ln p(\boldsymbol{X}, Z|\theta) = \sum_{i=1}^{N} \log p\left(\boldsymbol{x}^{(i)} \middle| z^{(i)}; \mu, \Sigma\right) + \log p(z^{(i)}; \phi)$$

- Then maximum likelihood estimation becomes nearly identical to what we had when estimating the parameters of the Gaussian discriminant analysis model,

Gaussian mixture model (GMM)

- If we observed latent variables:

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

Gaussian mixture model (GMM)

- However we don't observe $z^{(i)}$s:
  - The observed data: $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$
- Idea: using EM algorithm to solve the following problem
  - E-step: Tries to guess the value of $z^{(i)}$s

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

  - M-step: Update parameters of the model based on our guess

Gaussian mixture model (GMM)

- E step:

$$p(z^{(i)}=j|x^{(i)};\phi,\mu,\Sigma)=\frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{\sum_{l=1}^{k}p(x^{(i)}|z^{(i)}=l;\mu,\Sigma)p(z^{(i)}=l;\phi)}$$

- M step:

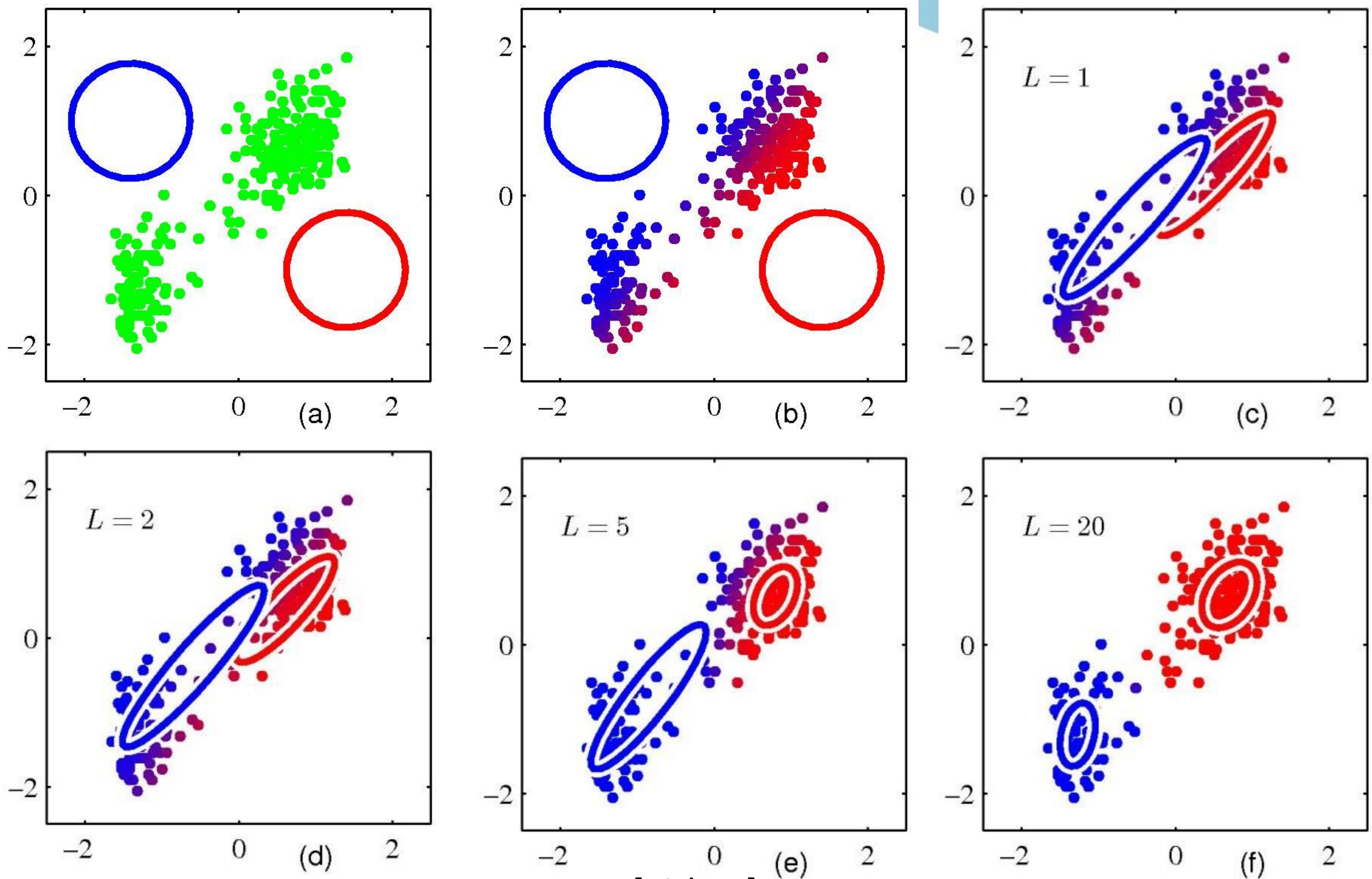
$$\begin{aligned}
\phi_j &:= \frac{1}{n}\sum_{i=1}^{n} w_j^{(i)}, \\
\mu_j &:= \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}}, \\
\Sigma_j &:= \frac{\sum_{i=1}^{n} w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}
\end{aligned}$$

EM algorithm

- An iterative algorithm in which each iteration is guaranteed to improve the log-likelihood function

- General algorithm for finding ML estimation when the data is incomplete (missing or unobserved data).
  - EM finds the maximum likelihood parameters in cases where the models involve unobserved variables Z in addition to unknown parameters $\boldsymbol{\theta}$ and known data observations X.

EM & GMM: Example



[Bishop]

EM theoretical foundation: Objective function

- ELBO: Evidence Lower Bound
  - For any choice for distribution $Q(z)$, ELBO gives a lower bound for $\log p(x,\theta)$.

$$\log p(x,\theta) = \log \sum_{z} p(x,z;\theta)$$

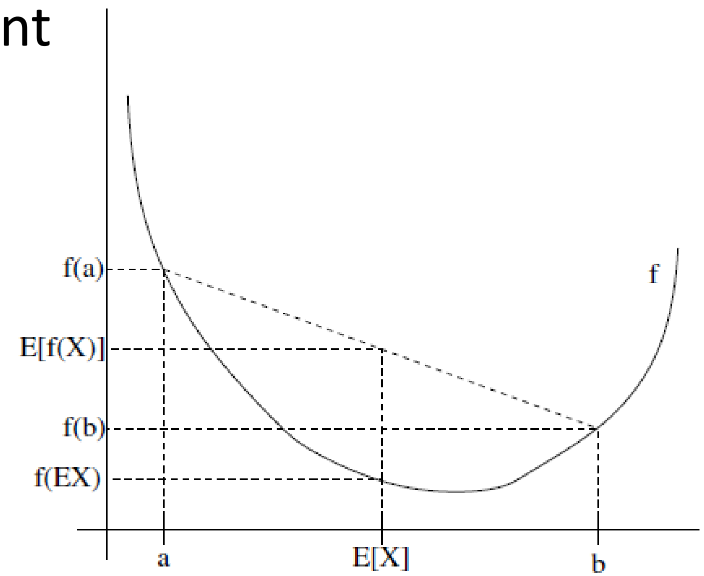
$$= \log \textstyle\sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}$$

Jensen's inequality

Jensen's inequality

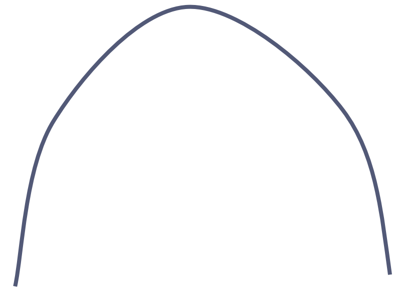
- For a convex function f and a random variable x
  - Equality holds only when x is constant

$$E[f(x)] \geq f(E[x])$$



- For the concave function f

$$E[f(x)] \leq f(E[x])$$



EM theoretical foundation: Objective function

- ELBO: Evidence Lower Bound
  - According to the Jensen's inequality, equality holds only when $\frac{p(x,z;\theta)}{Q(z)}$ is constant.

$$\frac{p(x,z;\theta)}{Q(z)} = c$$

  - As $Q(z)$ is a PDF,

$$\sum_z Q(z) = 1 \rightarrow \sum_z p(x,z;\theta) * \frac{1}{c} = 1 \rightarrow c = \sum_z p(x,z;\theta)$$

Therefore:

$$Q(z) = \frac{p(x,z;\theta)}{\sum_z p(x,z;\theta)} = \frac{p(x,z;\theta)}{p(x;\theta)} = p(z|x;\theta)$$

ELBO equality condition

EM theoretical foundation: Objective function

- As ELBO is a lower bound for the likelihood function of the data, i.e. $\log p(x,\theta)$, we can maximize it to find an approximation for the maximum of $\log p(x,\theta)$.

- Expectation-maximization (EM) method:
  - A coordinate ascent algorithm to maximize ELBO
  - **E-step**

$$Q^{t+1} = argmax_Q \, ELBO(Q,\theta^t)$$

  - **M-step**

$$\theta^{t+1} = argmax_\theta \, ELBO(Q^{t+1},\theta)$$

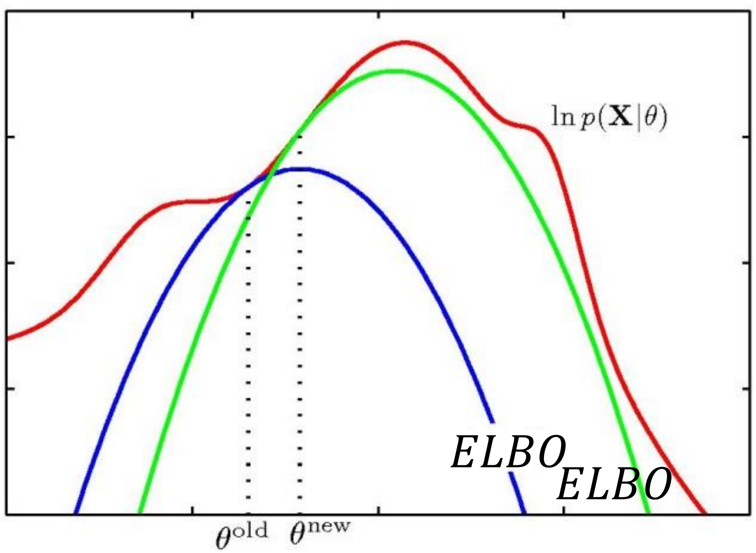
EM Convergence

- The likelihood function is increased when EM progresses

$$\log p(x,\theta^{t+1}) \geq ELBO(x, Q^{t+1}, \theta^{t+1})$$
$$\geq ELBO(x,\ Q^{t+1}, \theta^{t}) = \log p(x,\theta^{t})$$

M step of the iteration t

E step of iteration t



[Bishop]

EM theoretical foundation: Objective function

- Expectation-maximization (EM) method:
  - A coordinate ascent algorithm to maximize ELBO
  - **E-step**

$$Q^{t+1} = argmax_Q \; ELBO(Q, \theta^t)$$

ELBO is maximized when equals to $\log p(x, \theta)$, which holds when:

$$Q^{t+1} = p(z|x; \theta^t)$$

Therefore, in the E-step we only need to set Q^t as the posterior probability function on $p(z|x; \theta^t)$.

EM theoretical foundation: Objective function

- Expectation-maximization (EM) method:
  - A coordinate ascent algorithm to maximize ELBO
  - **M-step**

$$\theta^{t+1} = argmax_{\theta}\, ELBO(Q^{t+1}, \theta)$$

$$ELBO(Q^{t+1}, \theta) = \sum_{z} Q^{t+1}(z) \log \frac{p(x, z; \theta)}{Q^{t+1}(z)}$$

$$= \sum_{z} Q^{t+1}(z) \log p(x, z|\theta) - \sum_{z} Q^{t+1}(z) \log Q^{t+1}(z)$$

$$= E_{Q^{t+1}}[\log p(x, z|\theta)] + H\big(Q^{t+1}(z)\big)$$

We only need to maximize it

Independent of θ

Expectation-maximization (EM) method

X: observed variables

Z: unobserved variables

$\boldsymbol{\theta}$: parameters

Expectation step (E-step): Given the current parameters, find soft completion of data using probabilistic inference

Maximization step (M-step): Treat the soft completed data as if it were observed and learn a new set of parameters

Choose an initial setting $\boldsymbol{\theta}^0, t = 0$

Iterate until convergence:

E Step: Use X and current $\boldsymbol{\theta}^t$ to calculate $P(Z|X, \boldsymbol{\theta}^t)$

M Step: $\boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} E_{Z\sim P(Z|X,\boldsymbol{\theta}^t)}[\log p(X, Z|\boldsymbol{\theta})]$

$t \leftarrow t + 1$

expectation of the log-likelihood evaluated using the current estimate for the parameters $\boldsymbol{\theta}^t$

$$E_{Z\sim P(Z|X,\boldsymbol{\theta}^{\text{old}})}[\log p(X, Z|\boldsymbol{\theta})]$$
$$= \sum_Z P(Z|X, \boldsymbol{\theta}^{\text{old}}) \times \log p(X, Z|\boldsymbol{\theta})$$

Gaussian mixture model (GMM)

- E step:

$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

- M step:

$$\begin{aligned}
&\sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}
\end{aligned}$$

Gaussian mixture model (GMM)

- M step:

$$\nabla_{\mu_l} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

$$= -\nabla_{\mu_l} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)$$

$$= \frac{1}{2} \sum_{i=1}^{n} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l$$

$$= \sum_{i=1}^{n} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)$$

$$\mu_l := \frac{\sum_{i=1}^{n} w_l^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_l^{(i)}},$$

Gaussian mixture model (GMM)

- M step:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j .$$

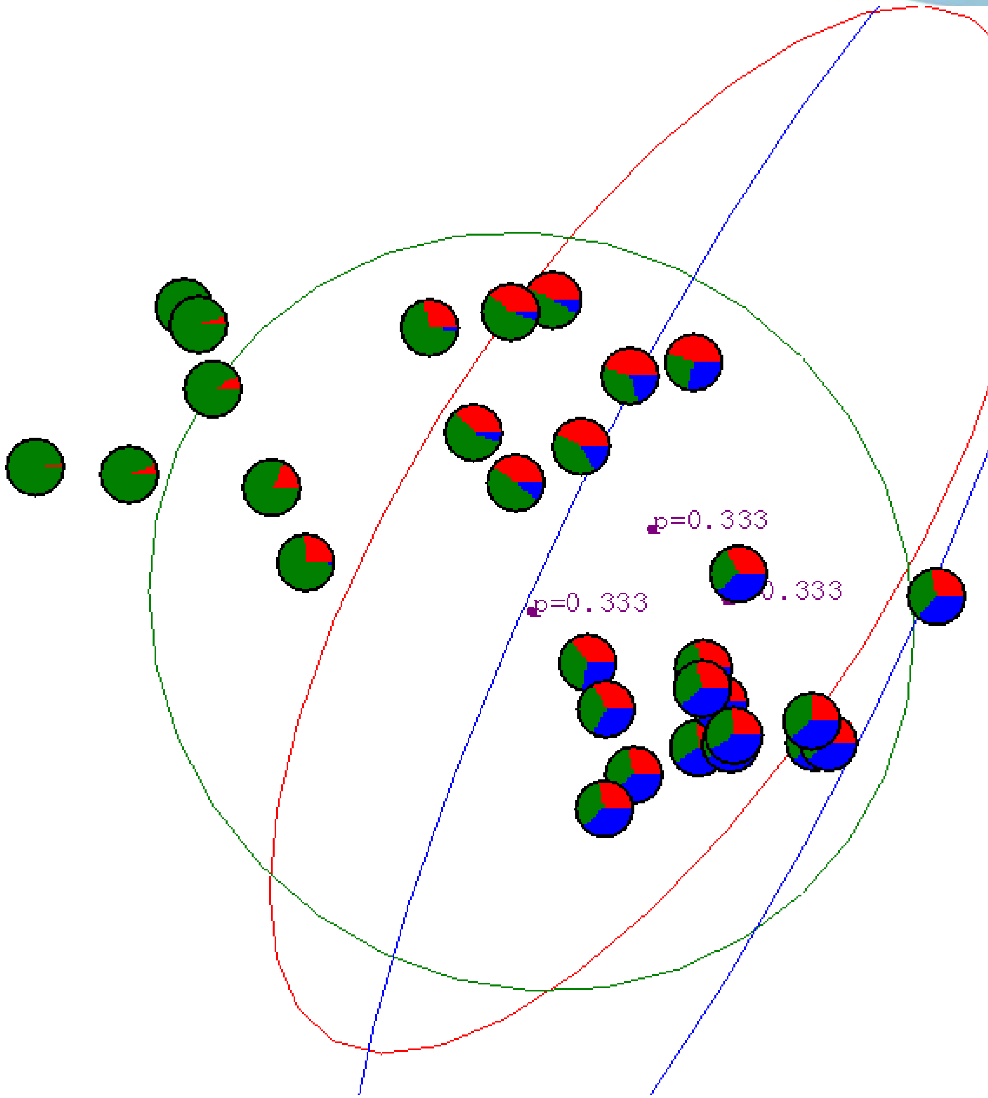
$$\mathcal{L}(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta\left(\sum_{j=1}^{k} \phi_j - 1\right),$$

$$\frac{\partial}{\partial \phi_j}\mathcal{L}(\phi) = \sum_{i=1}^{n} \frac{w_j^{(i)}}{\phi_j} + \beta$$

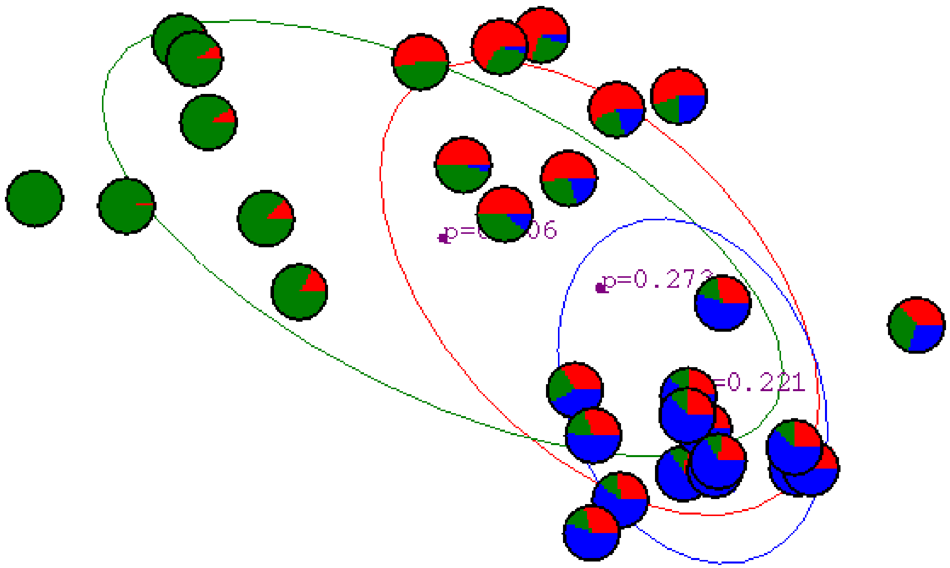
$$\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)}}{-\beta}$$

$$\phi_j := \frac{1}{n}\sum_{i=1}^{n} w_j^{(i)}.$$

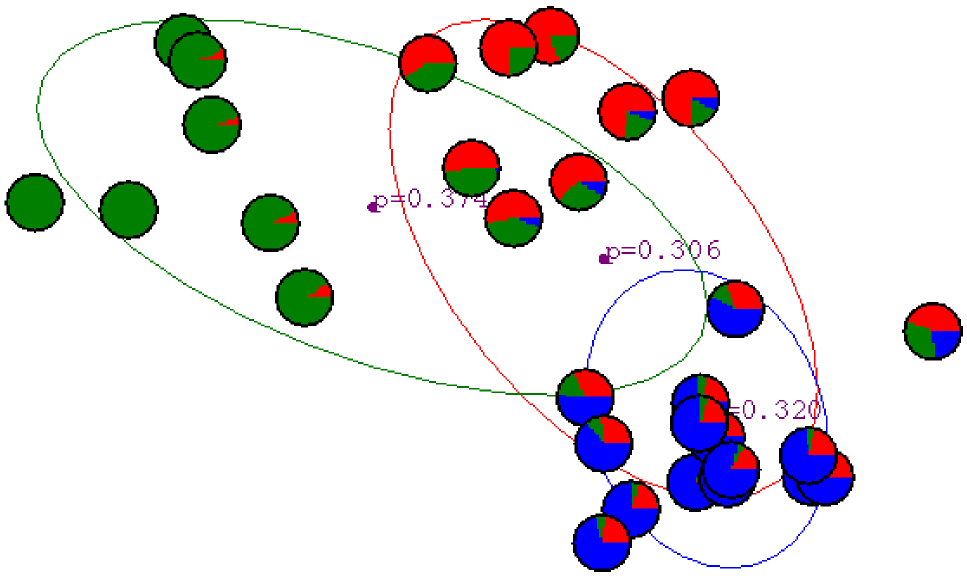
GMM and EM



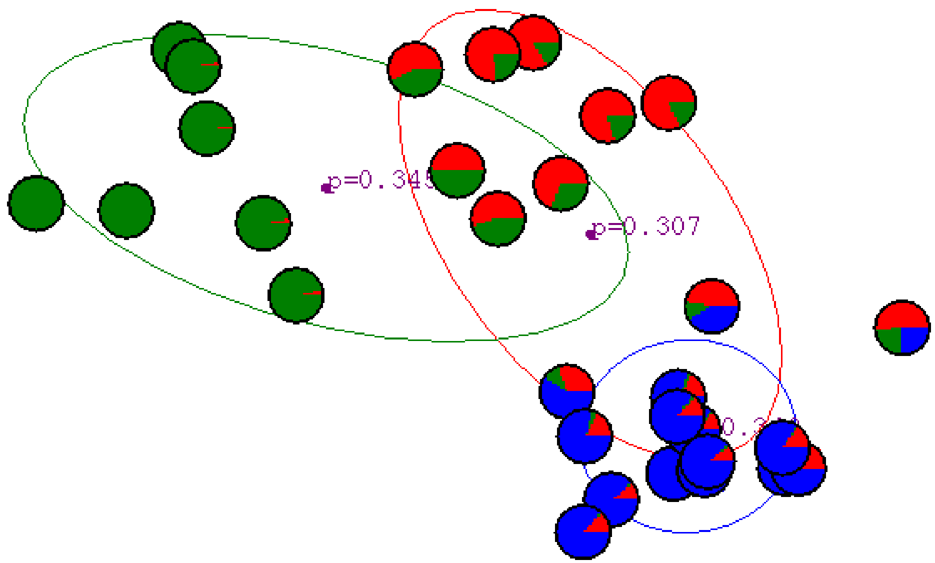
GMM and EM



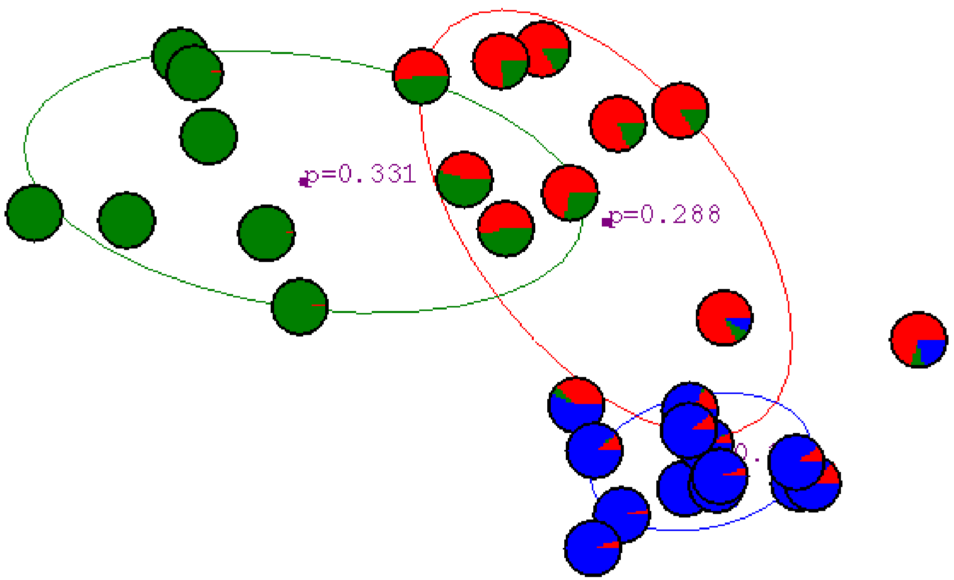
GMM and EM



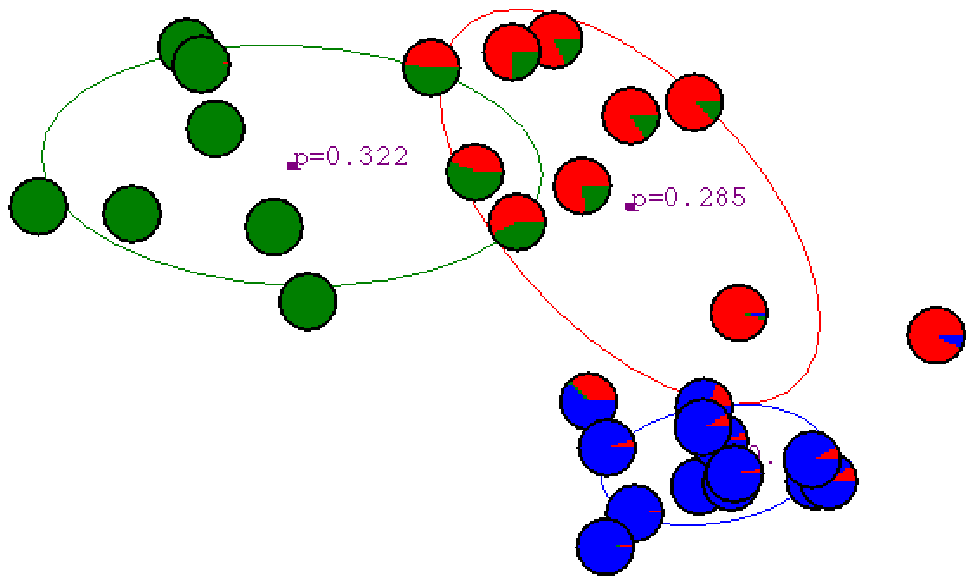
GMM and EM



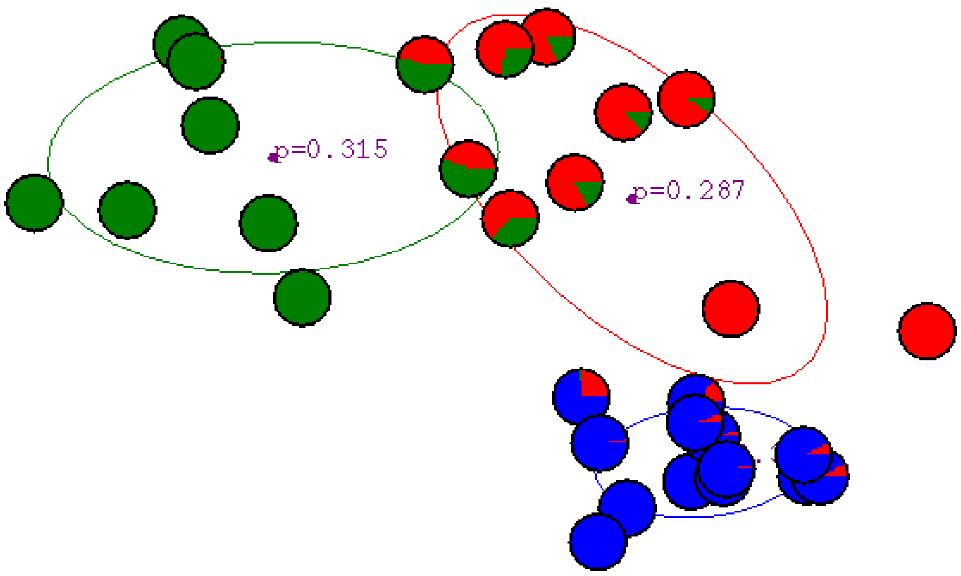
GMM and EM



GMM and EM



GMM and EM



GMM and EM