Machine Learning

Gaussian Mixture Model (GMM)

Dr. Mehran Safayani

safayani@iut.ac.ir

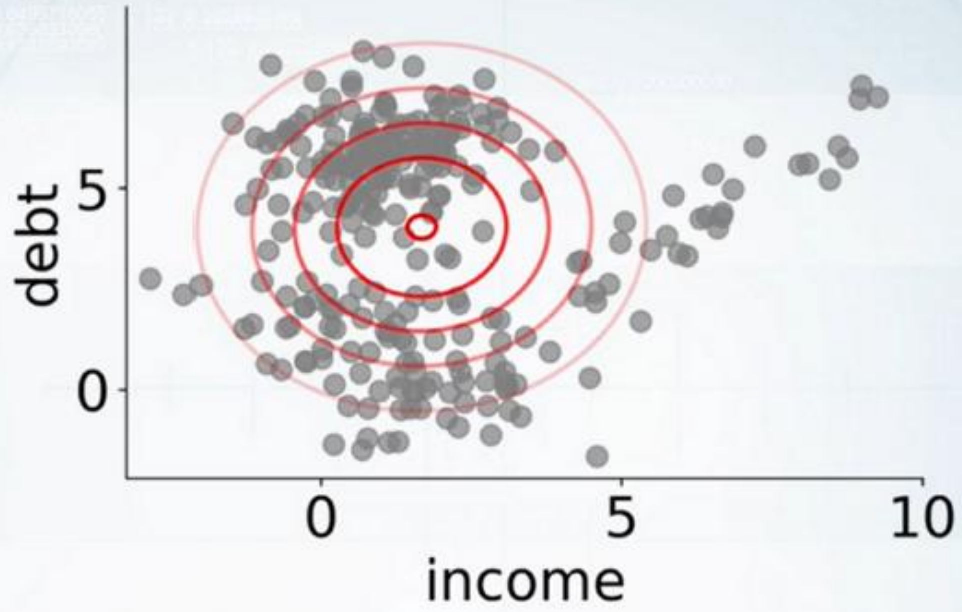
safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

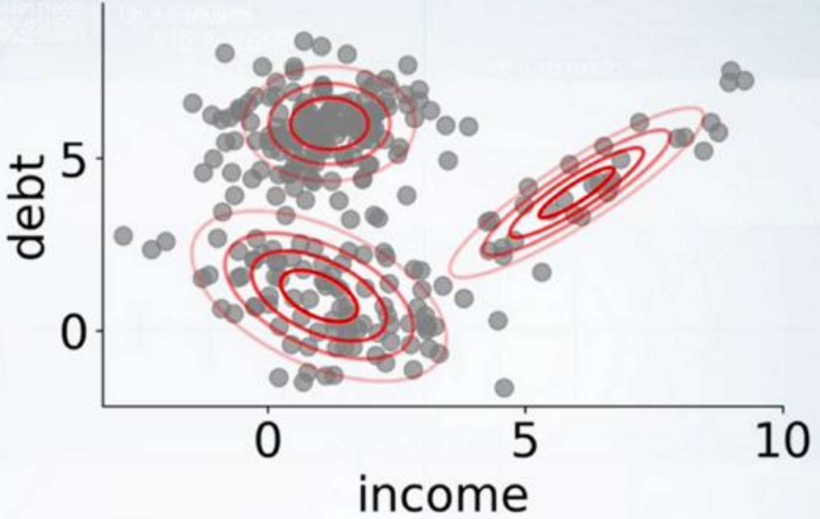
Probabilistic model of data



$$p(x \mid \theta) = \mathcal{N}(x \mid \mu, \Sigma)$$

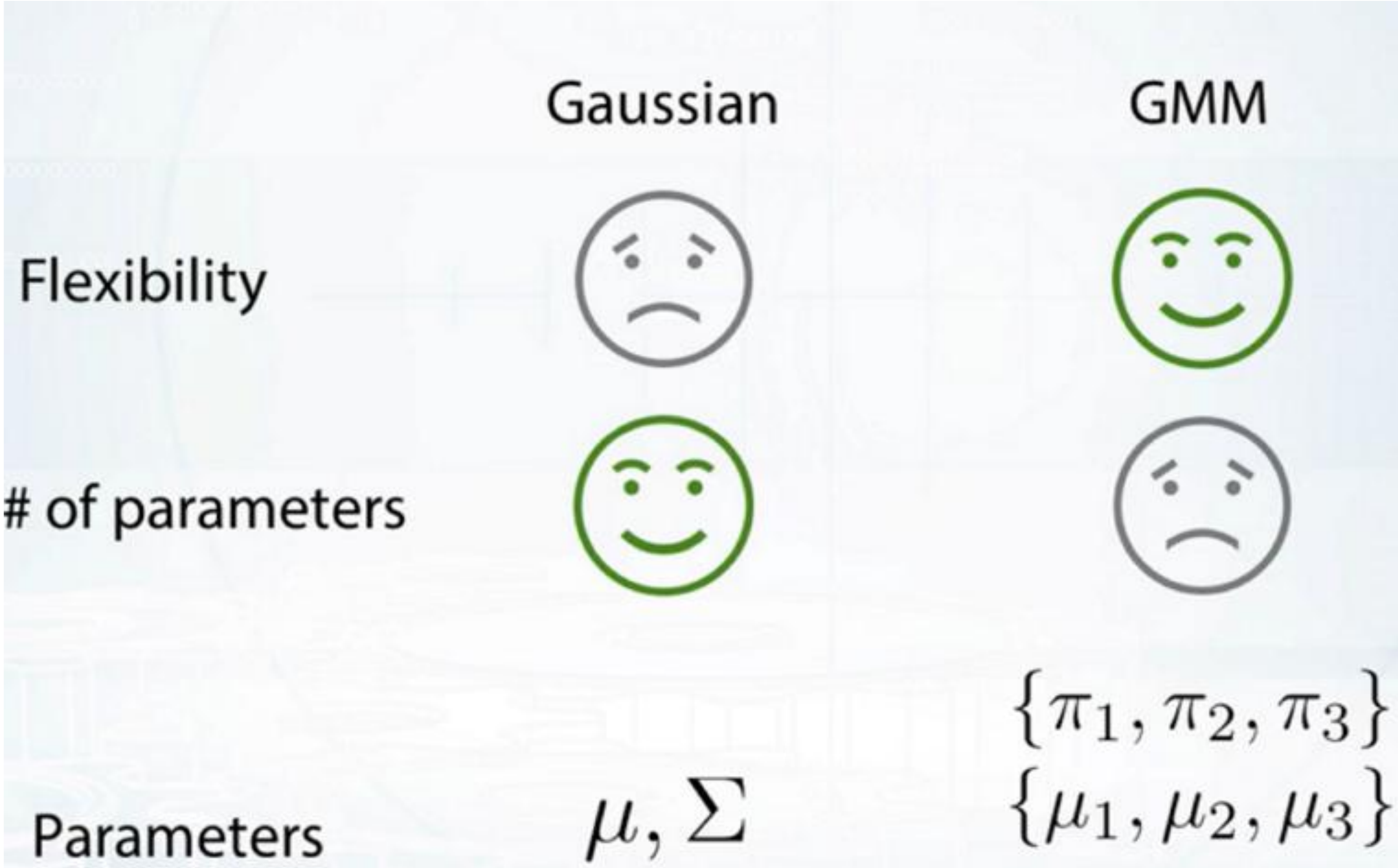
$$\theta = \{\mu, \Sigma\}$$

Gaussian Mixture Model (GMM)



$$p(x \mid \theta) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3)$$

$$\theta = \{\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3\}$$

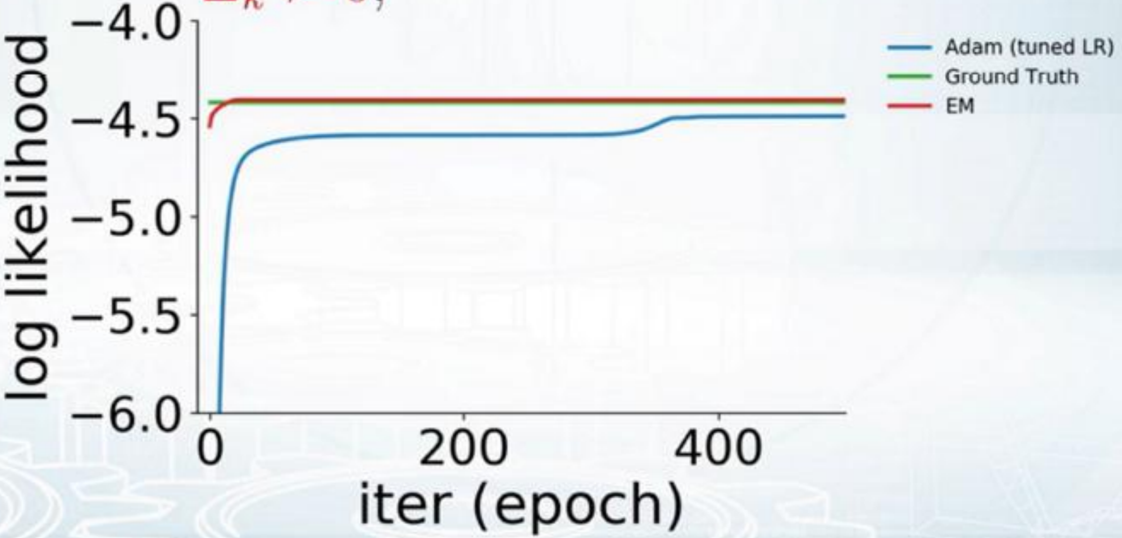

Gaussian
GMM
Flexibility
# of parameters
Parameters
μ, Σ
$\{\pi_1, \pi_2, \pi_3\}$
$\{\mu_1, \mu_2, \mu_3\}$

Training GMM

$$\max_{\theta} \quad \prod_{i=1}^{N} p(x_i \mid \theta) = \prod_{i=1}^{N} \left(\pi_1 \mathcal{N}(x_i \mid \mu_1, \Sigma_1) + \ldots\right)$$

$$\text{subject to} \quad \pi_1 + \pi_2 + \pi_3 = 1;\ \pi_k \geq 0;\ k = 1, 2, 3.$$

$$\Sigma_k \succ 0;$$



Introducing latent variable

$$p(x \mid \theta) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3)$$

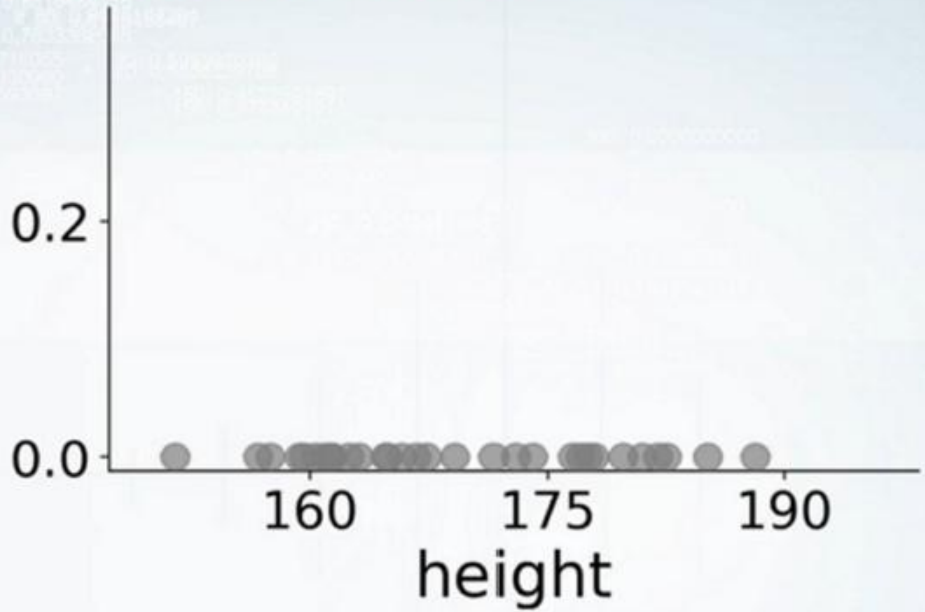
$$p(t = c \mid \theta) = \pi_c$$

$$p(x \mid t = c, \theta) = \mathcal{N}(x \mid \mu_c, \Sigma_c)$$

$$p(x \mid \theta) = \sum_{c=1}^{3} p(x \mid t = c, \theta) p(t = c \mid \theta)$$

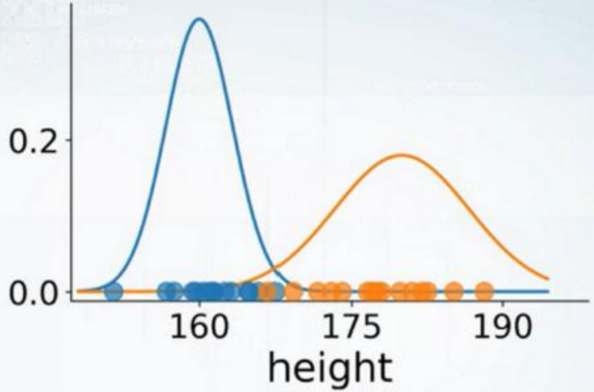
P(n,t]

Expectation Maximization



How to estimate parameter θ?

Expectation Maximization



How to estimate parameter θ?

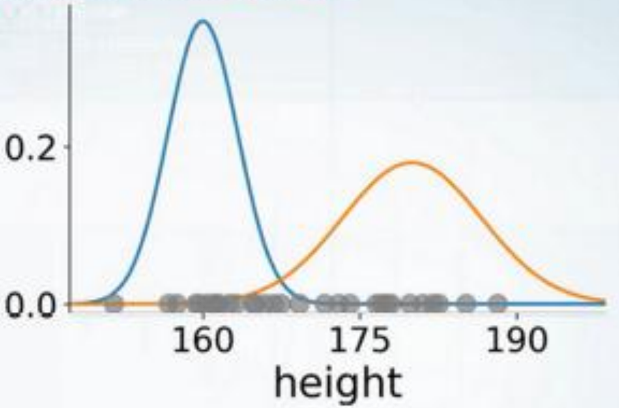
If sources t are known, easy:

$$p(x \mid t = 1, \theta) = \mathcal{N}(x \mid \mu_1, \sigma_1^2)$$

$$\mu_1 = \frac{\sum_{\text{blue } i} x_i}{\text{\# of blue points}} \qquad \sigma_1^2 = \frac{\sum_{\text{blue } i} (x_i - \mu_1)^2}{\text{\# of blue points}}$$

$$\mu_1 = \frac{\sum_i p(t_i = 1 \mid x_i, \theta)\, x_i}{\sum_i p(t_i = 1 \mid x_i, \theta)} \quad \sigma_1^2 = \frac{\sum_i p(t_i = 1 \mid x_i, \theta)\, (x_i - \mu_1)^2}{\sum_i p(t_i = 1 \mid x_i, \theta)}$$

Expectation Maximization



What if we don't know the sources?

Given: $p(x \mid t = 1, \theta) = \mathcal{N}(-2, 1)$

Find: $p(t = 1 \mid x, \theta)$

$$p(t = 1 \mid x, \theta) = \frac{p(x \mid t = 1, \theta)\, p(t = 1 \mid \theta)}{Z}$$

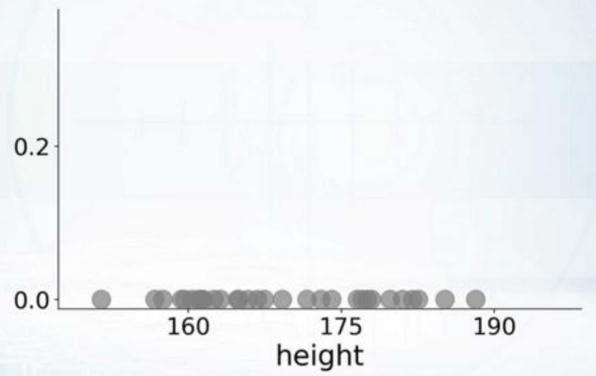
Chicken and egg problem

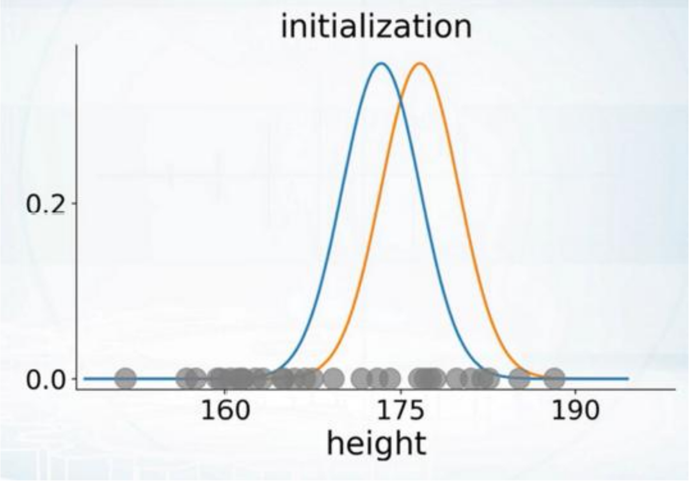
- Need Gaussian parameters to estimate sources
- Need sources to estimate Gaussian parameters

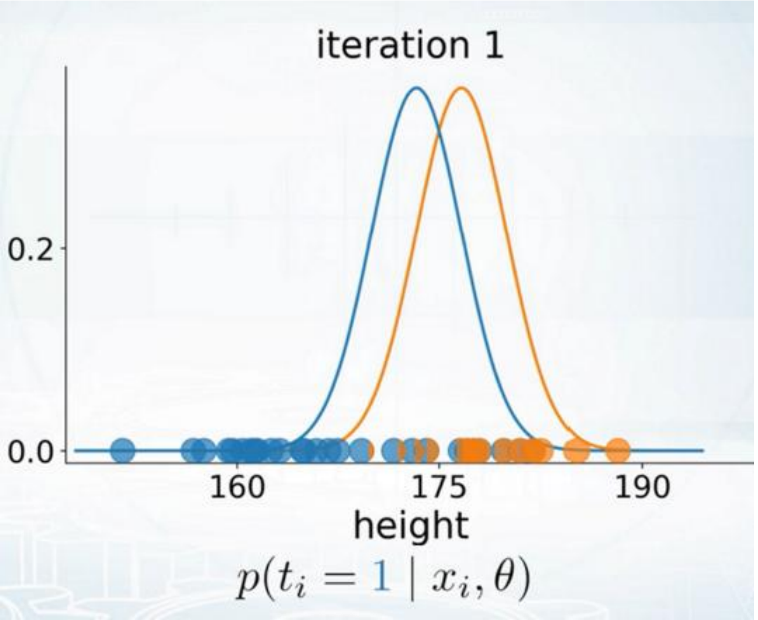
EM algorithm

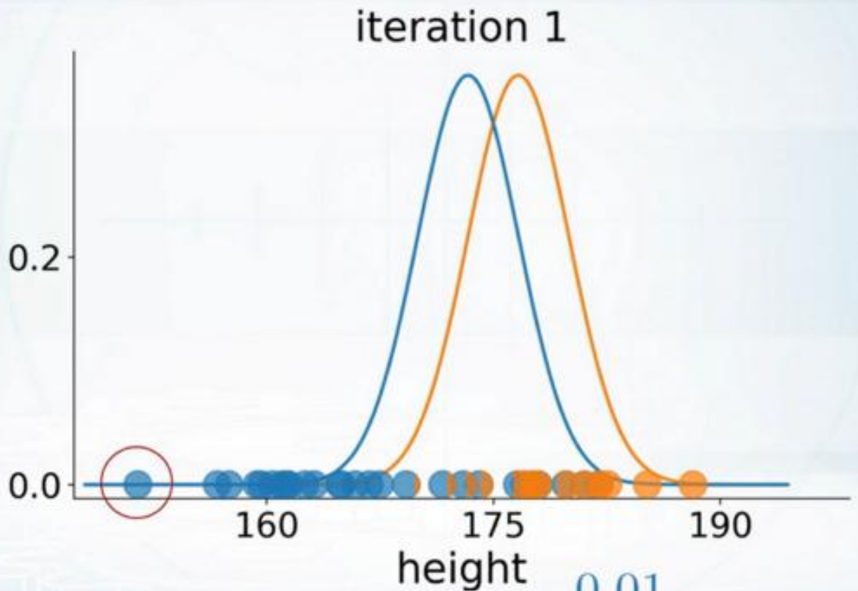
1. Start with 2 randomly placed Gaussians parameters θ
2. Until convergence repeat:
   a) For each point compute $p(t = c \mid x_i, \theta)$: does x_i look like it came from cluster c?
   b) Update Gaussian parameters θ to fit points assigned to them

GMM EM example

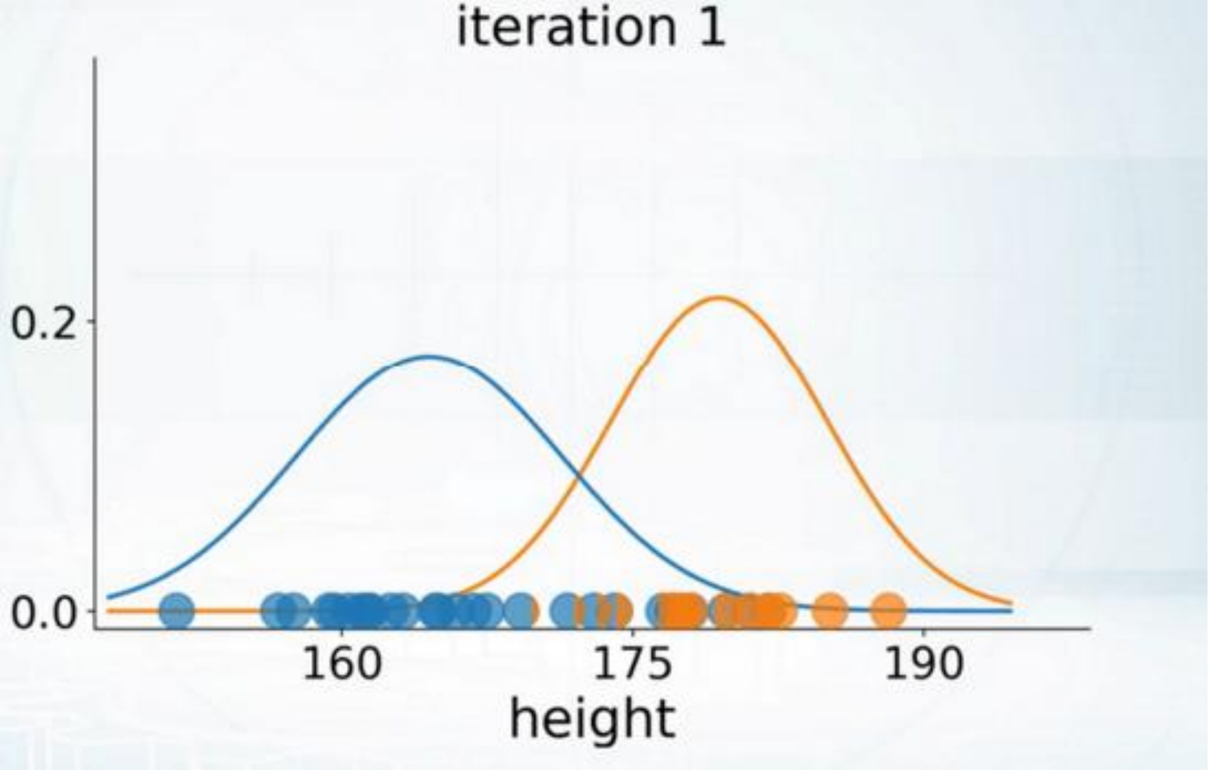


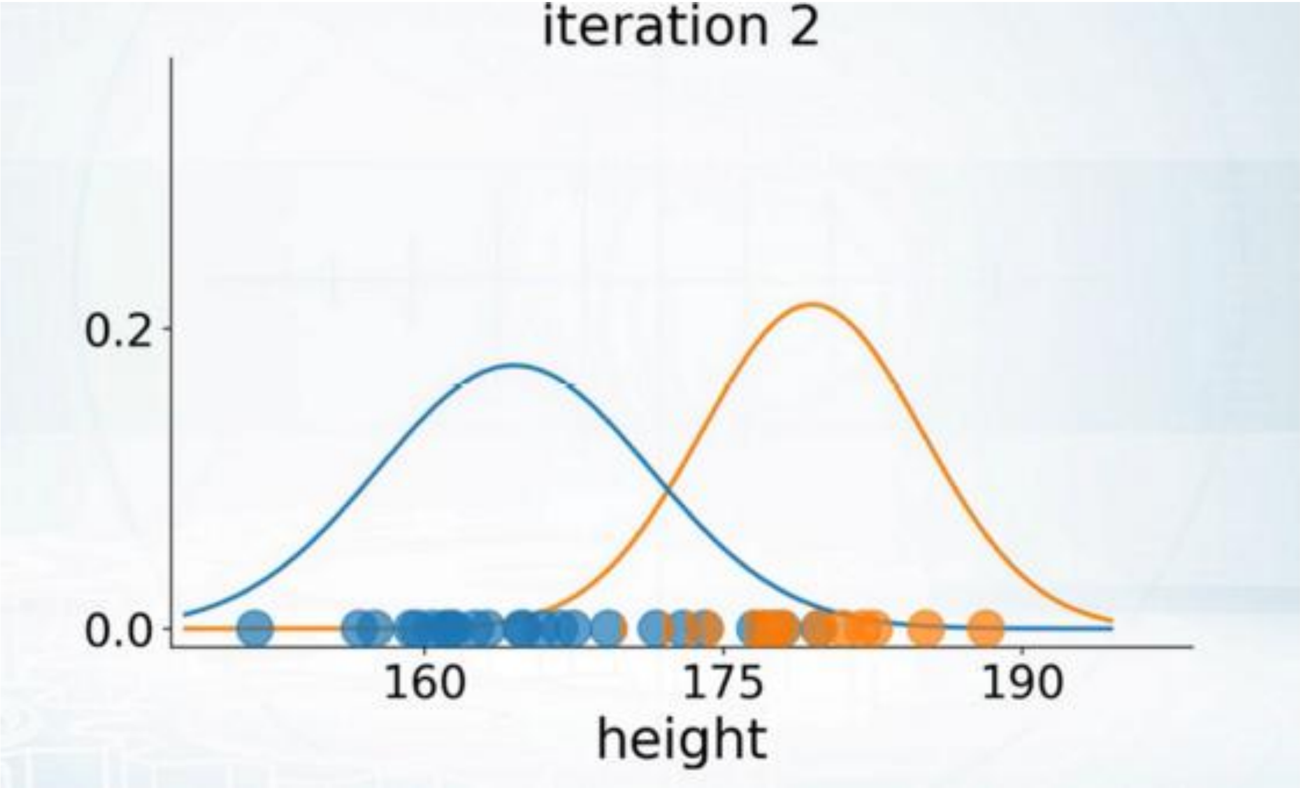



iteration 1
0.2
0.0
160
175
190
height
$p(t_i = 1 \mid x_i, \theta)$

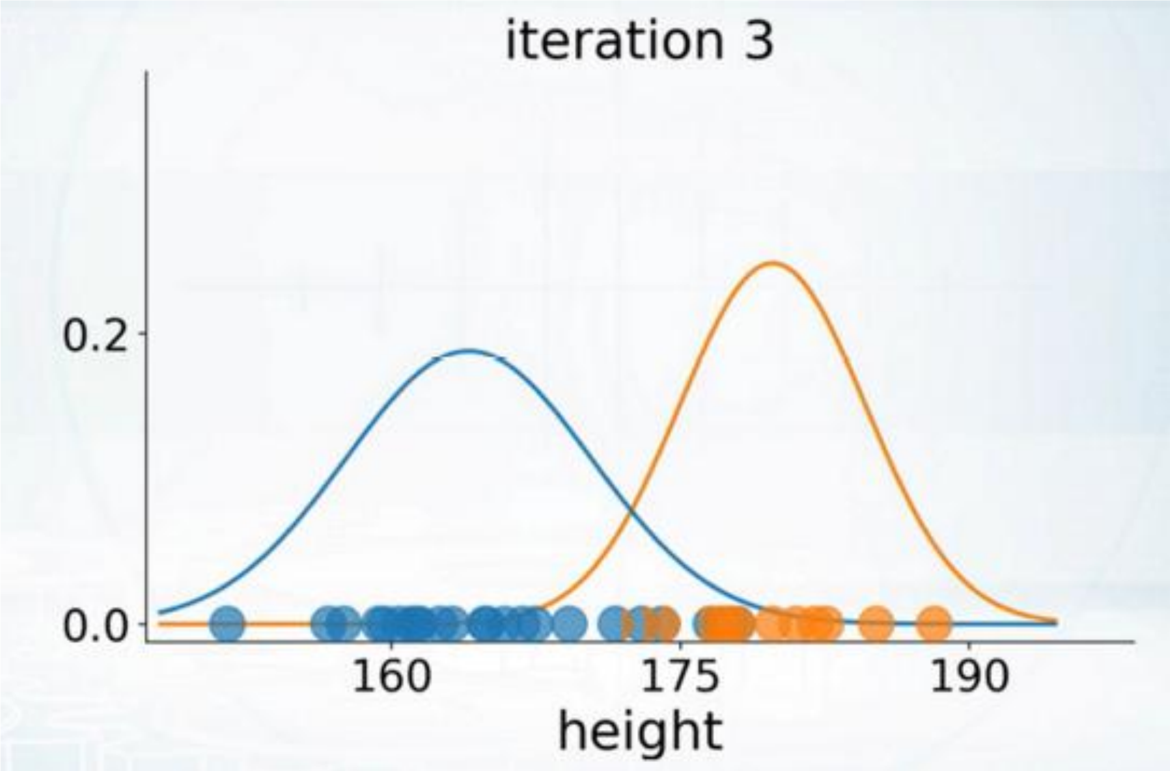


$$p(t_i = 1 \mid x_i, \theta) \approx \frac{0.01}{0.01 + 0.00002} \approx 1$$

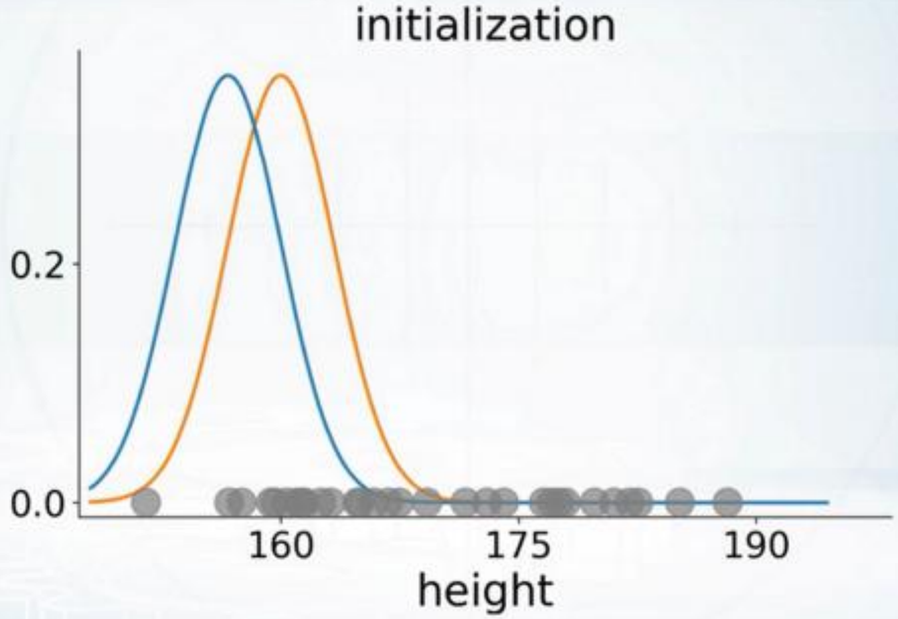

iteration 1
0.2
0.0
160
175
190
height

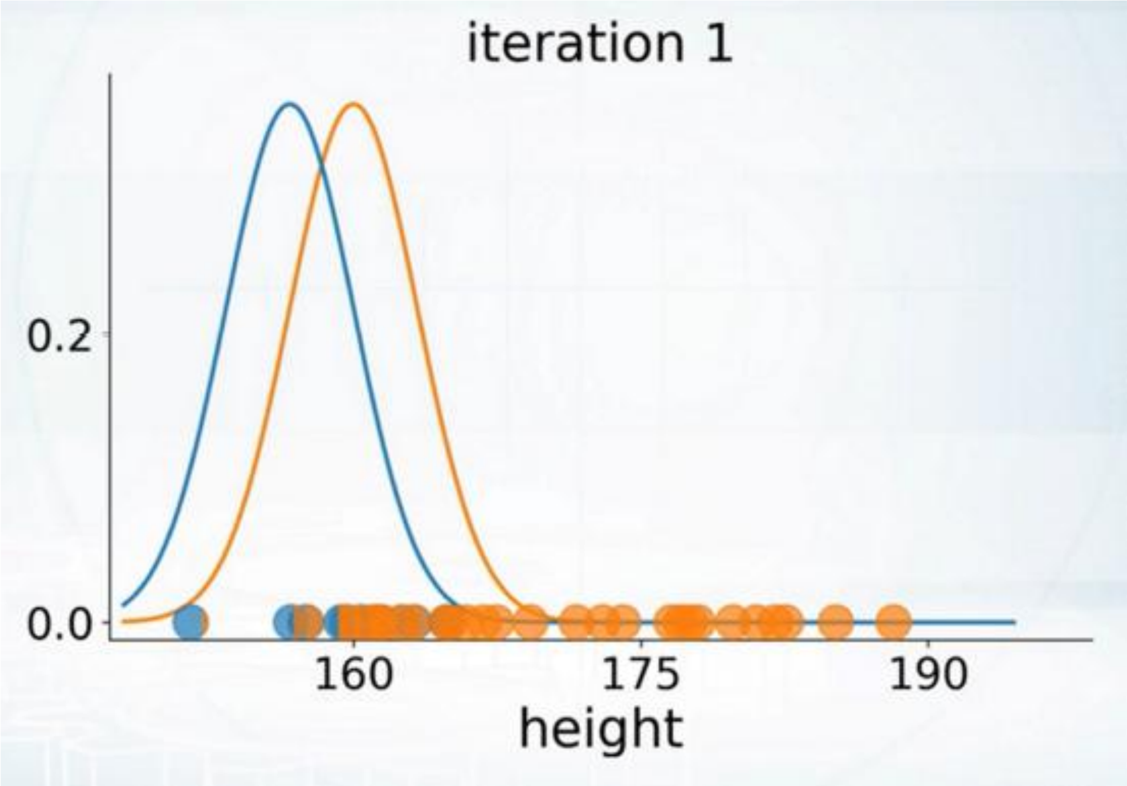


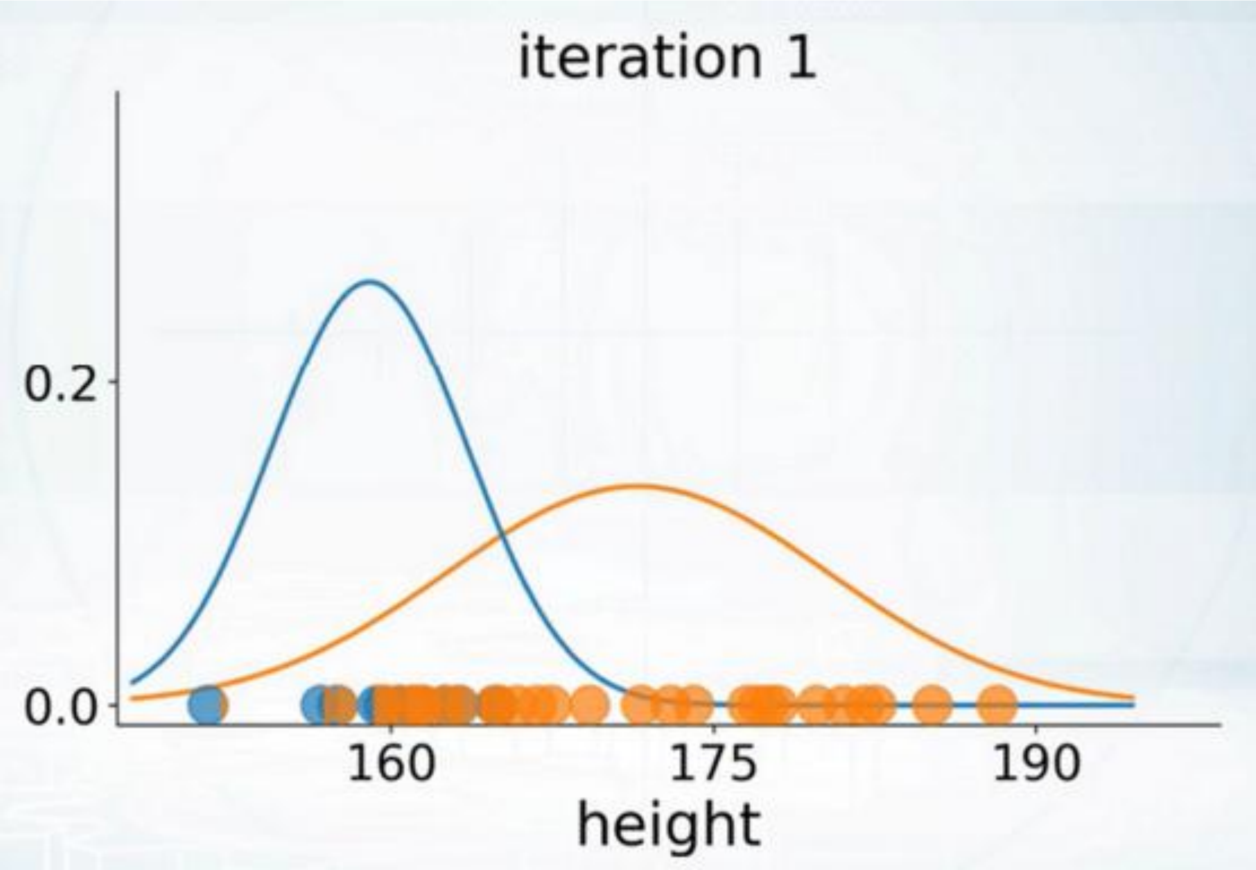
iteration 2
0.2
0.0
160
175
190
height

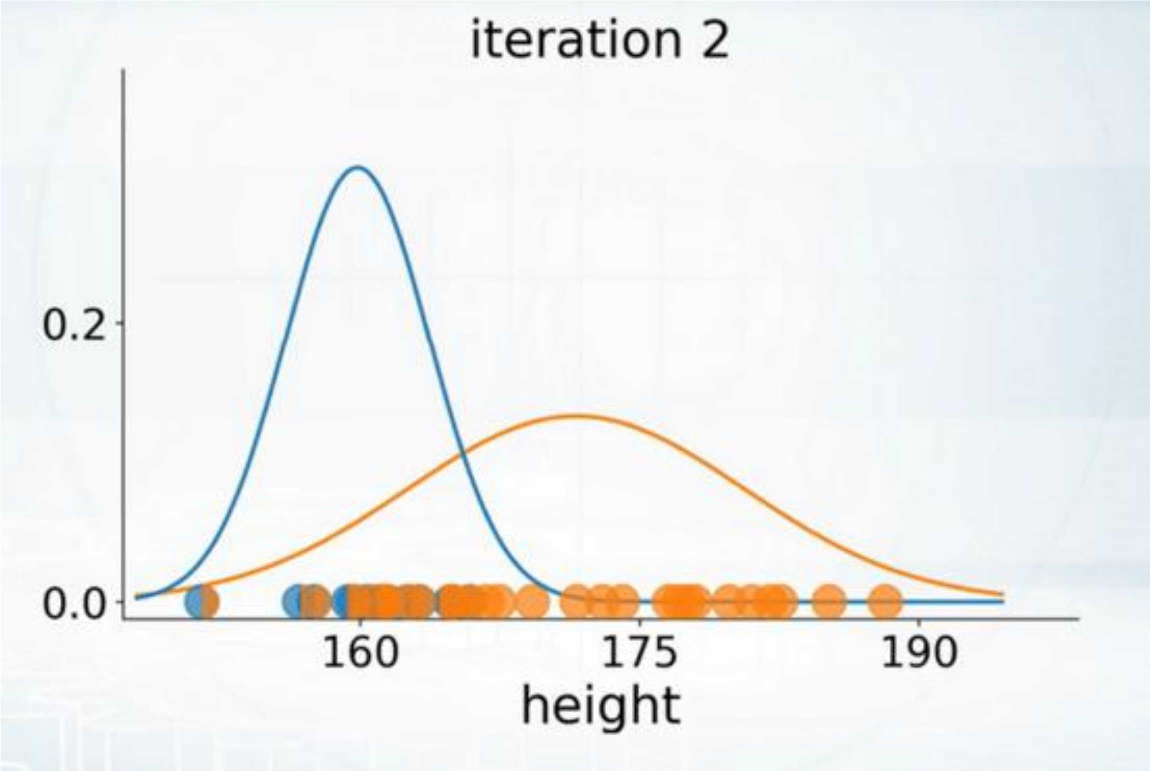

iteration 3
0.2
0.0
160
175
190
height

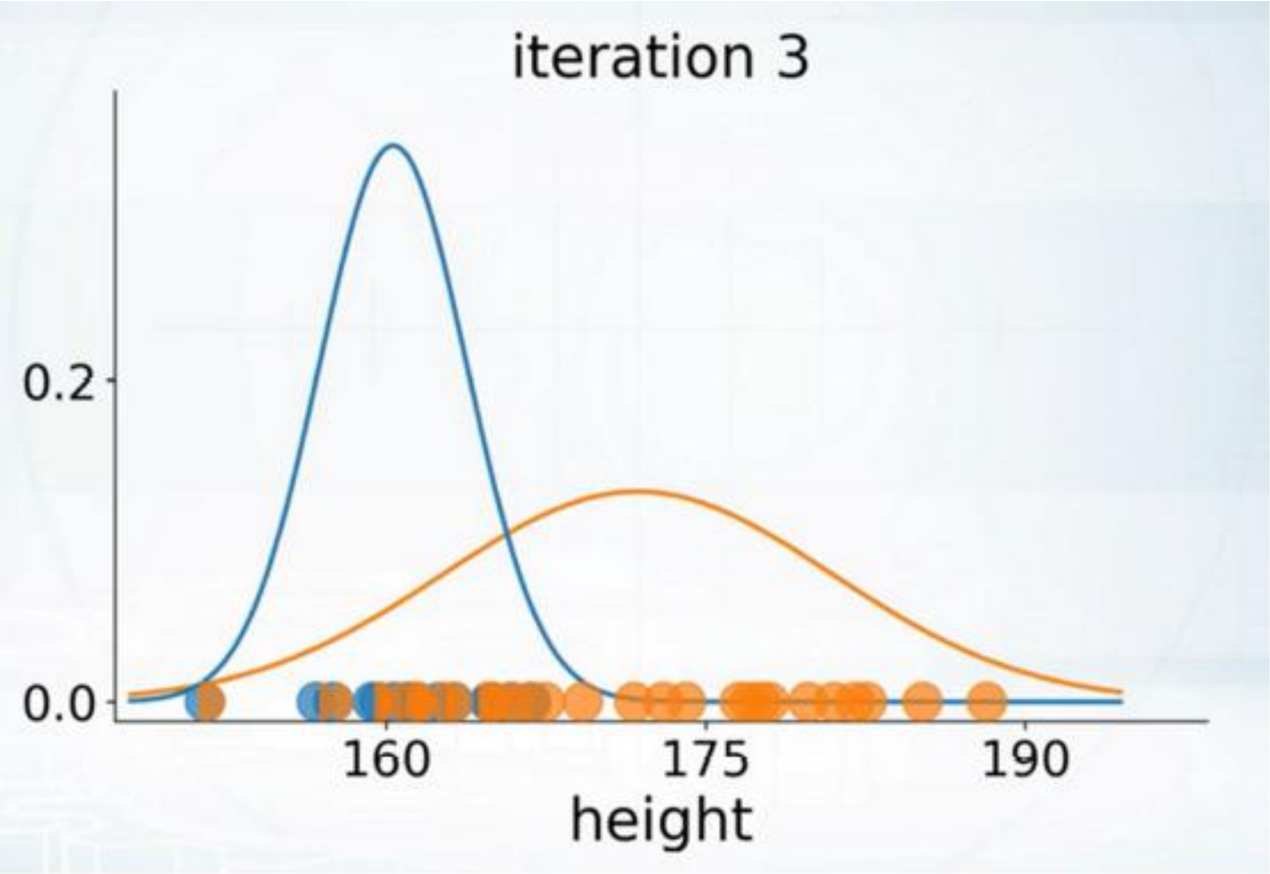
GMM EM local maximum example




iteration 1
0.2
0.0
160
175
190
height

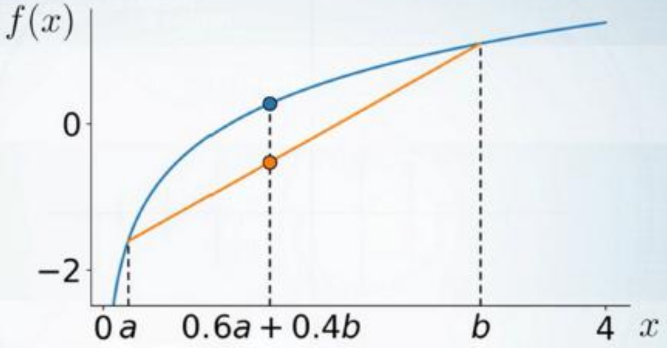

iteration 1
0.2
0.0
160
175
190
height


iteration 2
0.2
0.0
160
175
190
height


iteration 3
0.2
0.0
160
175
190
height

General form of Expectation Maximization

Concave functions



Def.: $f(x)$ is concave if

for any a, b, α : $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha) f(b)$

$$0 \leq \alpha \leq 1$$

Jensen's inequality

If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha) f(b)$

Then $\alpha_1 + \alpha_2 + \alpha_3 = 1;\ \alpha_k \geq 0.$

$$f(\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3) \geq \alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)$$

$$f(\underbrace{\alpha_1 a_1 + \alpha_2 a_2 + \alpha_3 a_3}_{\mathbb{E}_{p(t)} t}) \geq \underbrace{\alpha_1 f(a_1) + \alpha_2 f(a_2) + \alpha_3 f(a_3)}_{\mathbb{E}_{p(t)} f(t)}$$

$$p(t = a_1) = \alpha_1,$$
$$p(t = a_2) = \alpha_2,$$
$$p(t = a_3) = \alpha_3$$

Jensen's inequality

If $f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha) f(b)$

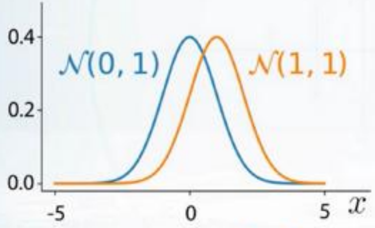
Then Jensen's inequality:

$$f\left(\mathbb{E}_{p(t)} t\right) \geq \mathbb{E}_{p(t)} f(t)$$

Kullback–Leibler divergence

Parameters difference: 1

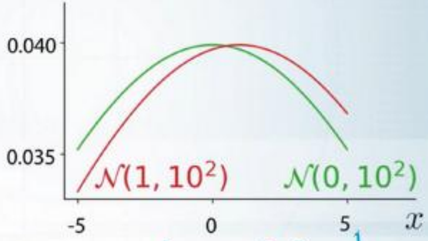
$$\mathcal{KL}(q_1 \parallel p_1) = 0.5$$



$|\mu_1 - \mu_2| = 1$

Parameters difference: 1

$$\mathcal{KL}(q_2 \parallel p_2) = 0.005$$



$|\mu_1 - \mu_2| = 1$

$$\mathcal{KL}(q \parallel p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

1. $\mathcal{KL}(q \parallel p) \neq \mathcal{KL}(p \parallel q)$
2. $\mathcal{KL}(q \parallel q) = 0$
3. $\mathcal{KL}(q \parallel p) \geq 0$

Proof: $-\mathcal{KL}(q \parallel p) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right)$

$$\leq \log(\mathbb{E}_q \frac{p}{q}) = \log \int \cancel{q(x)} \frac{p(x)}{\cancel{q(x)}} dx = 0$$

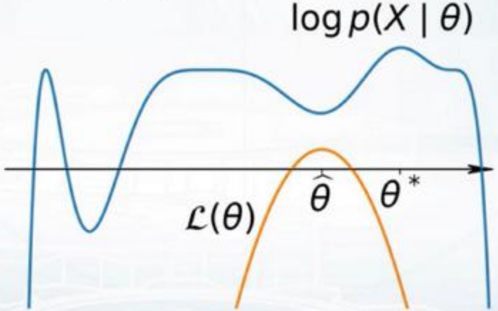
General form of Expectation Maximization

$$p(x_i \mid \theta) = \sum_{c=1}^{3} p(x_i \mid t_i = c, \theta)\, p(t_i = c \mid \theta)$$

$$\begin{aligned}\max_{\theta}.\ \log p(X \mid \theta) &= \log \prod_{i=1}^{N} p(x_i \mid \theta) \\ &= \sum_{i=1}^{N} \log p(x_i \mid \theta)\end{aligned}$$

$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} p(x_i, t_i = c \mid \theta) \geq \mathcal{L}(\theta)$$



$$\log p(X \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta)$$

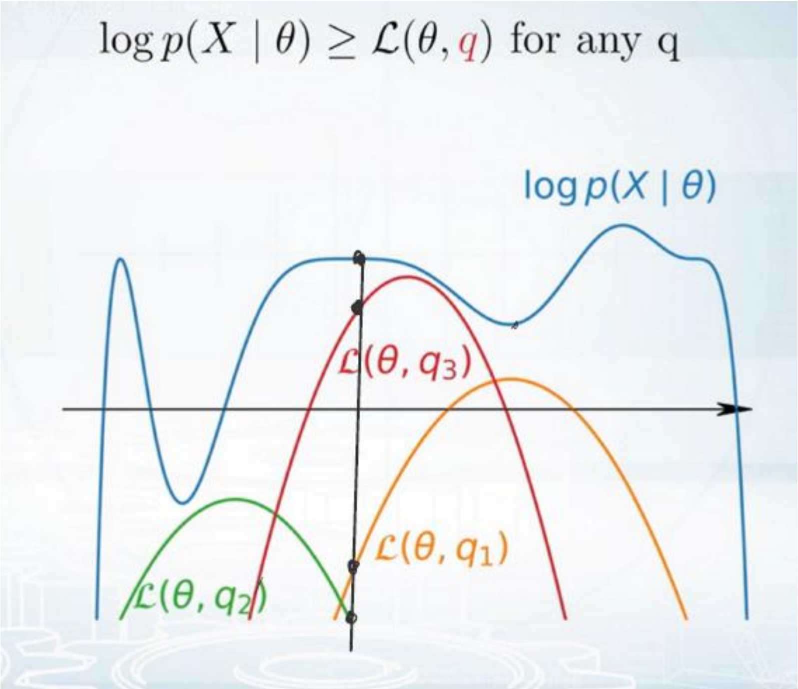
$$= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta)$$

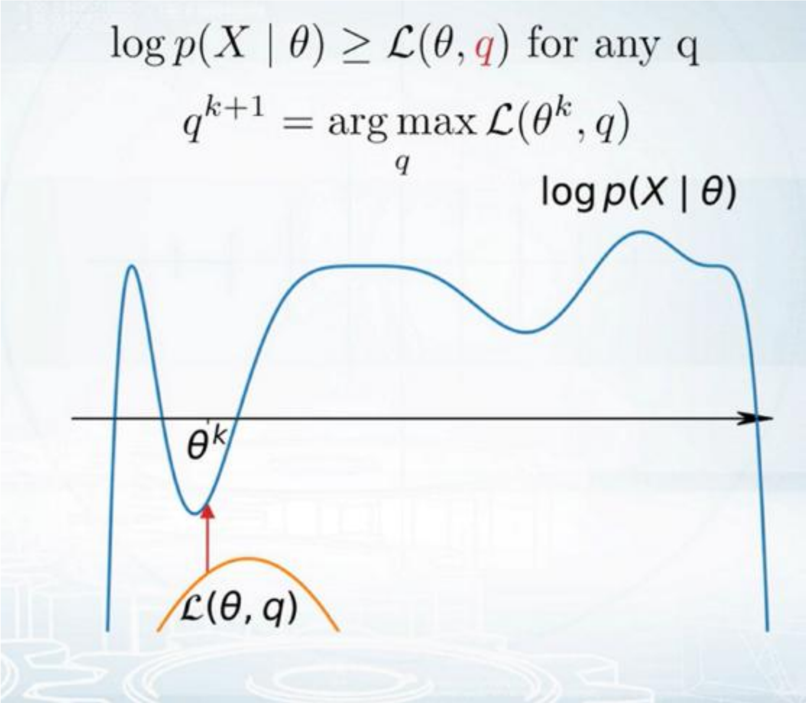
$$\geq \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i, t_i = c \mid \theta)}{q(t_i = c)}$$

Jensen's inequality

$$\boxed{\log \left(\sum_{c} \alpha_c v_c \right) \geq \sum_{c} \alpha_c \log(v_c)}$$

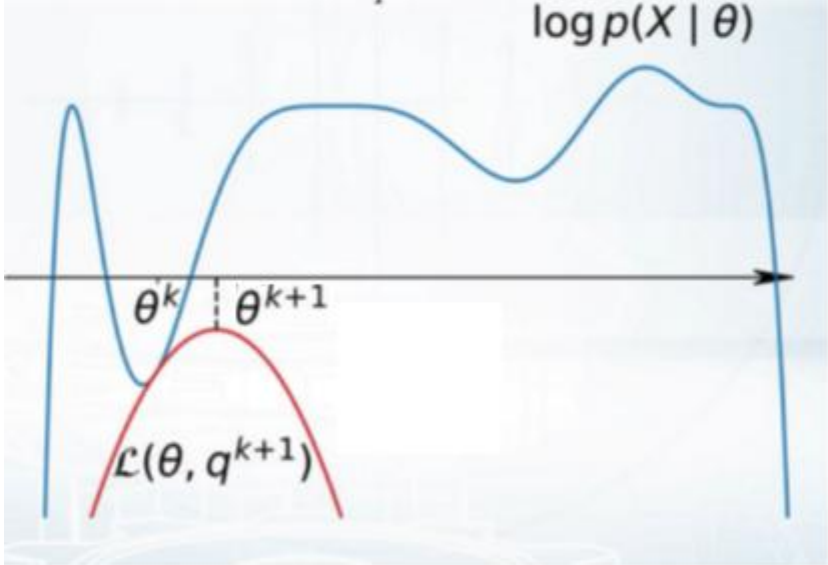
$$
\begin{aligned}
\log p(X \mid \theta) &= \sum_{i=1}^{N} \log p(x_i \mid \theta) \\
&= \sum_{i=1}^{N} \log \sum_{c=1}^{3} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c \mid \theta) \\
&\geq \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i, t_i = c \mid \theta)}{q(t_i = c)} \\
&= \mathcal{L}(\theta, q)
\end{aligned}
$$


$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$ for any q
$\log p(X \mid \theta)$
$\mathcal{L}(\theta, q_3)$
$\mathcal{L}(\theta, q_1)$
$\mathcal{L}(\theta, q_2)$


$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$ for any q
$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$
$\log p(X \mid \theta)$
θ^k
$\mathcal{L}(\theta, q)$

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

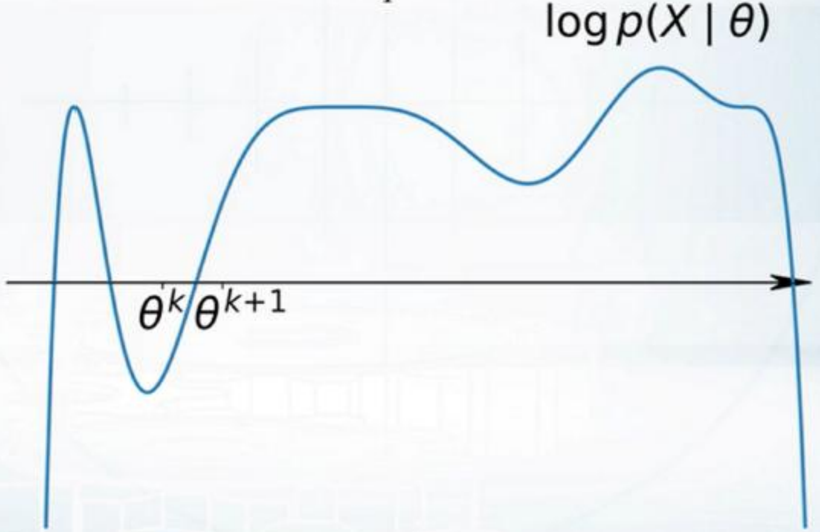
$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$

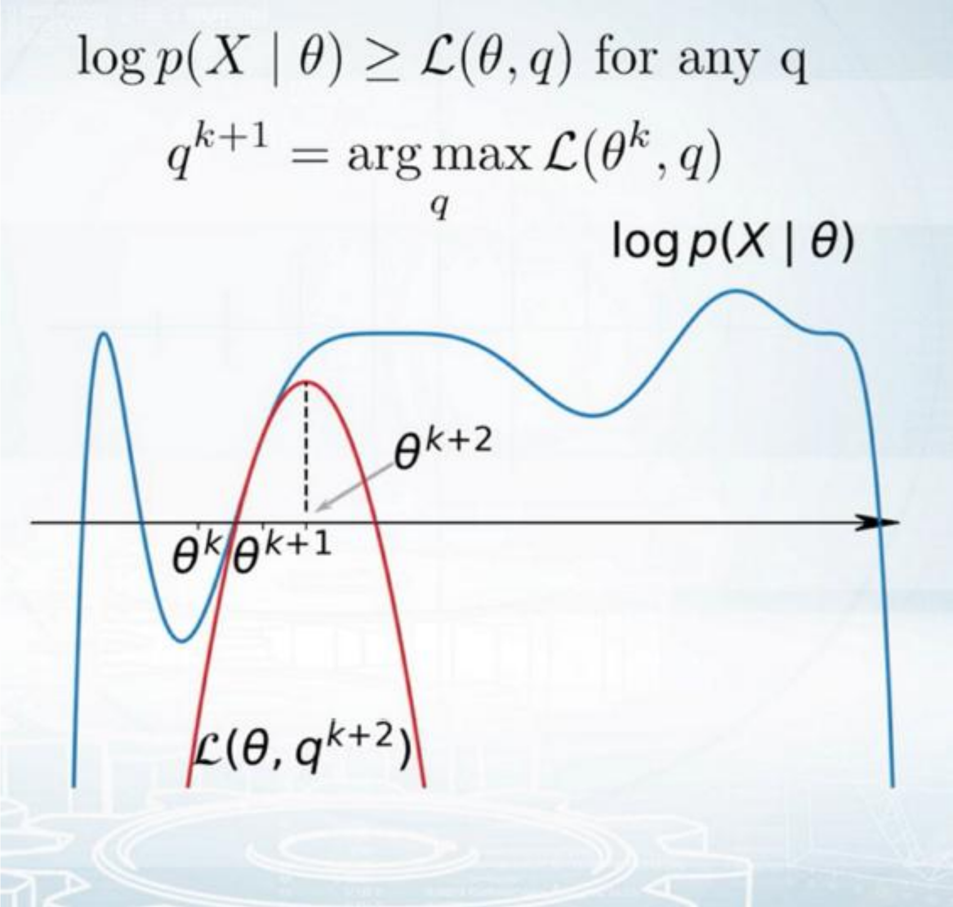


$$\theta^{k+1} = \arg\max_{\theta} \mathcal{L}(\theta, q^{k+1})$$

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$





$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$ for any q
$q^{k+1} = \arg\max_q \mathcal{L}(\theta^k, q)$
$\log p(X \mid \theta)$
θ^{k+2}
θ^k
θ^{k+1}
$\mathcal{L}(\theta, q^{k+2})$

Summary of Expectation Maximization

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q) \text{ for any q}$$

Variational lower bound

E-step

$$q^{k+1} = \arg\max_{q} \mathcal{L}(\theta^k, q)$$

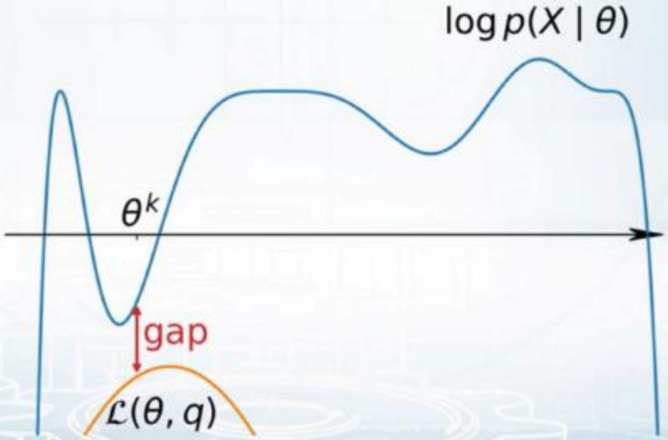
M-step

$$\theta^{k+1} = \arg\max_{\theta} \mathcal{L}(\theta, q^{k+1})$$

E-step details

$$\log p(X \mid \theta) \geq \mathcal{L}(\theta, q)$$

E-step: $\max_q \mathcal{L}(\theta^k, q)$



$$GAP = \log p(X|\theta) - L(\theta, q) = \sum_{i=1}^{N} \log p(x_i|\theta) - \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(m_i, t_i = c|\theta)}{q(t_i = c)}$$

$$\sum_{i=1}^{N} \log p(x_i|\theta) \cdot \sum_{c=1}^{3} q(t_i = c) - \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i, t_i = c|\theta)}{q(t_i = c)}$$

$$= - \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c)(\log p(x_i|\theta) - \log \frac{p(x_i, t_i = c|\theta)}{q(t_i = c)})$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{p(x_i|\theta) q(t_i = c)}{p(x_i, t_i = c|\theta)}$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{3} q(t_i = c) \log \frac{q(t_i = c)}{p(t_i = c|x_i, \theta)} = \sum_{i=1}^{N} \sum_{c=1}^{3} KL(q(t_i = c)||p(t_i = c|x_i, \theta))$$

Estep:

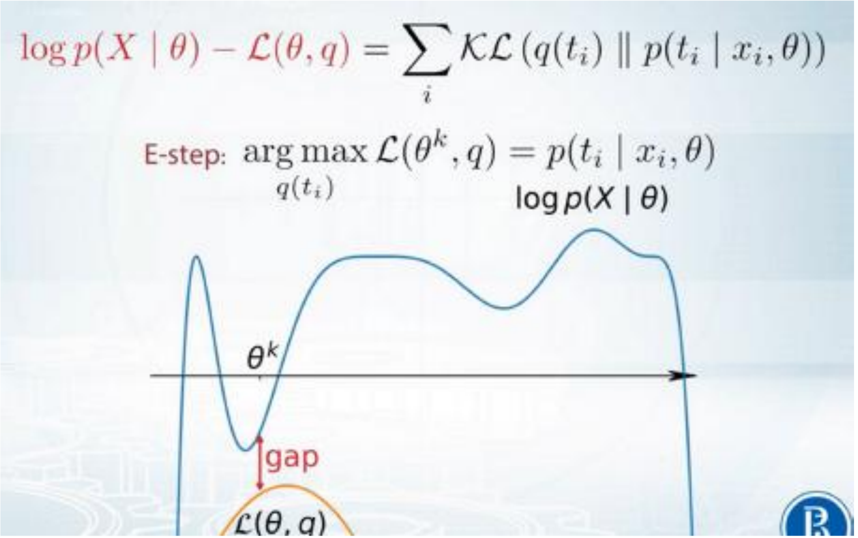
$$GAP = \log p(X|\theta) - L(\theta, q) = \log p(X|\theta) - ELBO$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{3} KL(q(t_i = c)||p(t_i = c|x_i, \theta)) >> 0$$

We maximize ELBO or minimize KL

$$\log p(X|\theta) = \sum_{i=1}^{N} \sum_{c=1}^{3} KL(q(t_i = c)||p(t_i = c|x_i, \theta)) + ELBO = GAP + ELBO$$

Solution: $q(t_i = c) = p(t_i = c|x_i, \theta)$


$\log p(X \mid \theta) - \mathcal{L}(\theta, q) = \sum_i \mathcal{KL}\left(q(t_i) \parallel p(t_i \mid x_i, \theta)\right)$
E-step: $\arg\max_{q(t_i)} \mathcal{L}(\theta^k, q) = p(t_i \mid x_i, \theta)$
$\log p(X \mid \theta)$
θ^k
gap
$\mathcal{L}(\theta, q)$

M-step details

$$\mathcal{L}(\theta, q) = \sum_i \sum_c q(t_i = c) \log \frac{p(x_i, t_i = c \mid \theta)}{q(t_i = c)}$$

$$= \sum_i \sum_c q(t_i = c) \log p(x_i, t_i = c \mid \theta)$$

$$\cancel{- \sum_i \sum_c q(t_i = c) \log q(t_i = c)}$$

$$= \mathbb{E}_q \log p(X, T \mid \theta) + \text{const}$$

Const w.r.t. θ

$$= \boxed{\mathbb{E}_q \log p(X, T \mid \theta)} + \text{const}$$

(Usually) concave function w.r.t. θ, easy to optimize

Expectation Maximization algorithm

For k = 1, …

E-step

$$q^{k+1} = \arg\min_q \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

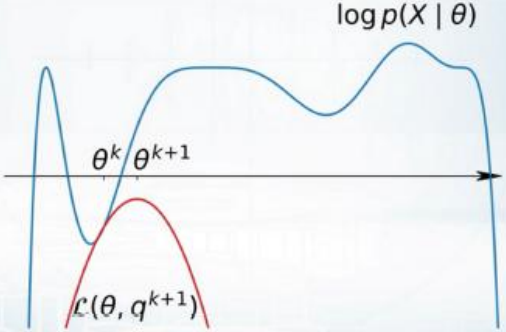
$$\Leftrightarrow$$

$$q^{k+1}(t_i) = p(t_i \mid x_i, \theta^k)$$

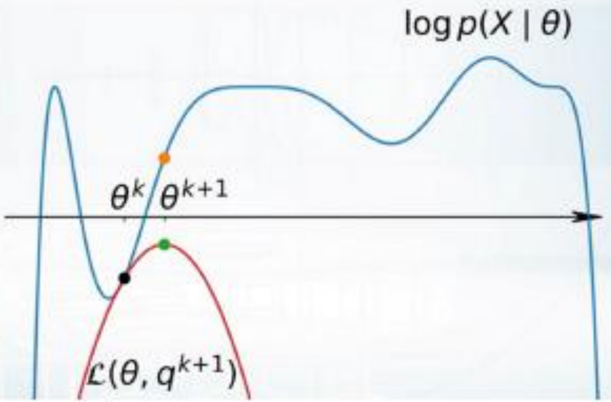
M-step

$$\theta^{k+1} = \arg\max_\theta \mathbb{E}_{q^{k+1}} \log p(X, T \mid \theta)$$

Convergence guaranties



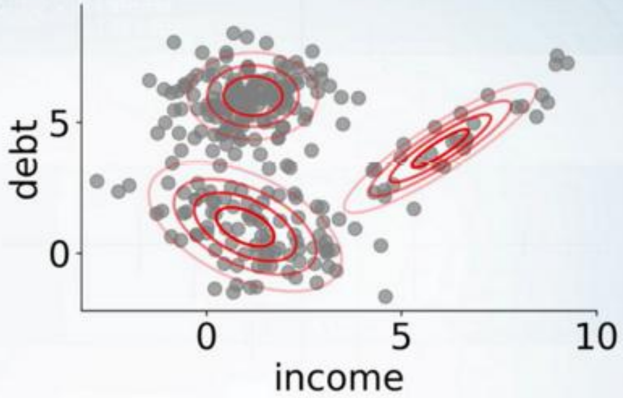
$$\log p(X \mid \theta^{k+1}) \geq \mathcal{L}(\theta^{k+1}, q^{k+1}) \geq \mathcal{L}(\theta^{k}, q^{k+1}) = \log p(X \mid \theta^{k})$$



$$\log p(X \mid \theta^{k+1}) \geq \log p(X \mid \theta^{k})$$

- On each iteration EM doesn't decrease the objective (good for debugging!)
- Guarantied to converge to a local maximum (or saddle point)

Gaussian Mixture Model revisited



$$p(x \mid \theta) = \pi_1 \mathcal{N}(x|\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x|\mu_2, \Sigma_2) + \pi_3 \mathcal{N}(x|\mu_3, \Sigma_3)$$

$$\theta = \{\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3\}$$

E-step

EM: For each point compute

$$q(t_i) = p(t_i \mid x_i, \theta)$$

GMM: For each point compute

$$p(t_i \mid x_i, \theta)$$

M-step

EM: Update parameters to maximize

$$\max_{\theta} \mathbb{E}_q \log p(X, T \mid \theta)$$

GMM: Update Gaussian parameters to fit points assigned to them

$$\mu_1 = \frac{\sum_i p(t_i = 1 \mid x_i, \theta)\, x_i}{\sum_i p(t_i = 1 \mid x_i, \theta)}$$

Applying EM on Gaussian Mixtures

In this section, we will use an example of Gaussian Mixture to demonstrate the application of EM algorithm.

Suppose we have some data $\mathbf{x} = x^{(1)}, \ldots, x^{(m)}$, which some from K different Gaussian distributions (K mixtures). We will use the following notations:

- μ_k: the mean of the k^{th} Gaussian component
- Σ_k: the covariance matrix of the k^{th} Gaussian component
- ϕ_k: the multinomial parameter of a specific datapoint belonging to the k^{th} componenet.
- $z^{(i)}$: the latent variable (multinomial) for each $x^{(i)}$

We also assume that the dimension of each $x^{(i)}$ is n.

The goal is: $max_{\mu,\Sigma,\phi}\ ln\ p(\mathbf{x}; \mu, \Sigma, \phi)$. Therefore this follows exactly the EM framework.

E step

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

M step

We will write down the lower bound and get derivatives for each of the three parameters.

$$\begin{aligned}
&\sum_i^m \sum_j^K q_i\left(z^{(i)} = j\right) \ln \frac{p\left(x^{(i)}, z^{(i)} = j; \mu, \Sigma, \phi\right)}{q_i\left(z^{(i)} = j\right)} \\
&= \sum_i^m \sum_j^K q_i\left(z^{(i)} = j\right) \ln \frac{p\left(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma\right) p\left(z^{(i)} = j; \phi\right)}{q_i\left(z^{(i)} = j\right)}
\end{aligned}$$

Note that:

- $x^{(i)} | z^{(i)} = j; \mu, \Sigma \sim \mathcal{N}(\mu_j, \Sigma_j)$
- $z^{(i)} = j; \phi \sim Multi(\phi)$

We can then leverage these probability distributions and continue

$$ll := \sum_{i}^{m} \sum_{j}^{K} w_j^{(i)} ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} exp\Big(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \Big) \phi_j}{w_j^{(i)}}$$

Now, we need to maximize this lower bound for each of the three parameters. Many of the derivative on vector/matrix are based on Matrix Cookbook

Derivative of μ_j

$$\nabla_{\mu_j} ll = \nabla_{\mu_j} \sum_i^m w_j^{(i)} ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right)\right)\phi_j}{w_j^{(i)}}$$

$$= \nabla_{\mu_j} \sum_i^m w_j^{(i)} \left[\ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_j|}} \phi_j}{w_j^{(i)}} + \ln \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right)\right) \right]$$

$$= \nabla_{\mu_j} \sum_i^m w_j^{(i)} \left[\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \right]$$

$$= -\frac{1}{2} \sum_i^m w_j^{(i)} \nabla_{\mu_j} \left[\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \right]$$

$$\left[\text{For } f(x) = x^T A x : \nabla_x f(x) = (A + A^T) x\right]$$

$$= \frac{1}{2} \sum_i^m w_j^{(i)} \nabla_{(x^i - \mu_j)} \left[\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \right]$$

$$= \frac{1}{2} \sum_i^m w_j^{(i)} \left[\left(\Sigma_j^{-1} + \left(\Sigma_j^{-1}\right)^T\right) \left(x^{(i)} - \mu_j\right) \right]$$

$$\nabla_{\mu_j} ll = 0$$

$$\sum_i^m w_j^{(i)} \left[\Sigma_j^{-1} \left(x^{(i)} - \mu_j \right) \right] = 0$$

$$\sum_i^m w_j^{(i)} \left(x^{(i)} - \mu_j \right) = 0$$

$$\sum_i^m w_j^{(i)} x^{(i)} = \sum_i^m w_j^{(i)} \mu_j$$

$$\mu_j = \frac{\sum_i^m w_j^{(i)} x^{(i)}}{\sum_i^m w_j^{(i)}}$$

Derivative of Σ_j

$$= \sum_i^m w_j^{(i)} \nabla_{\Sigma_j} \left[\ln \frac{1}{\sqrt{|\Sigma_j|}} - \frac{1}{2} \left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \right]$$

$$= -\frac{1}{2} \sum_i^m w_j^{(i)} \left[\frac{\partial \ln |\Sigma_j|}{\partial \Sigma_j} + \frac{\partial}{\partial \Sigma_j} \left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \right]$$

First, we consider the derivative of the first term in the square bracket:

$$\begin{aligned} \frac{\partial \ln |\Sigma_j|}{\partial \Sigma_j} &= \frac{1}{|\Sigma_j|} \frac{\partial |\Sigma_j|}{\partial \Sigma_j} \\ &= \frac{1}{|\Sigma_j|} |\Sigma_j| \left(\Sigma_j^{-1}\right)^Y \\ &= \Sigma_j^{-1} \end{aligned}$$

Then, we do the second term

$$\frac{\partial}{\partial \Sigma_j} \left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) = -\Sigma_j^{-1} \left(x^{(i)} - \mu_j\right) \left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1}$$

Combined these results back and set it to zero, we have:

$$\nabla_{\Sigma_j} ll = -\frac{1}{2}\sum_i^m w_j^{(i)}\left[\Sigma_j^{-1} - \Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T\Sigma_j^{-1}\right]$$
$$= -\frac{1}{2}\sum_i^m w_j^{(i)}\left[I - \Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T\right]\Sigma_j^{-1} \overset{\text{get}}{=} 0$$

Rearrange the equation and we have:

$$\sum_i^m w_j^{(i)}\left[\Sigma_j - \left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T\right] = 0$$
$$\sum_i^m w_j^{(i)}\Sigma_j = \sum_i^m w_j^{(i)}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T$$

$$\Sigma_j = \frac{\sum_i^m w_j^{(i)}\left(x^{(i)}-\mu_j\right)\left(x^{(i)}-\mu_j\right)^T}{\sum_i^m w_j^{(i)}}$$

Derivative of ϕ_j

This is relatively simpler but we need to apply Lagrange multipliers because $\sum_j \phi_j = 1$.

$$
\begin{aligned}
ll &= \sum_i^m \sum_l^k w_l^{(i)} \ln \frac{\frac{1}{\sqrt{(2\pi)^n |\Sigma_l|}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_l\right)^T \Sigma_l^{-1} \left(x^{(i)} - \mu_l\right)\right) \phi_l}{w_l^{(i)}} \\
&= \sum_i^m \sum_l^k w_l^{(i)} \ln \phi_l
\end{aligned}
$$

We need to construct Lagrangian, with λ as the Lagrange multiplier:

$$
\mathcal{L}(\phi) = ll + \lambda \left(\sum_l^k \phi_l - 1 \right)
$$

We will take derivative on $\mathcal{L}$ and set it to zero:

$$\begin{aligned}\frac{\partial \mathcal{L}(\phi)}{\partial \phi_j} &= \frac{\partial}{\partial \phi_j}\left[l + \lambda\left(\sum_l^k \phi_l - 1\right)\right] \\ &= \sum_i w_j^{(i)} \frac{1}{\phi_j} + \lambda \overset{\text{set}}{=} 0\end{aligned}$$

Rearrange and we will have $\phi_j = -\frac{\sum_i w_j^{(i)}}{\lambda}$. Recall that $\sum_j \phi_j = 1$, we have:

$$\begin{aligned}\sum_j \phi_j = \sum_j -\frac{\sum_i w_j^{(i)}}{\lambda} &= 1 \\ \lambda &= -\sum_j \sum_i w_j^{(i)} \\ &= -\sum_j \sum_i p\left(z^{(i)} = j \mid x^{(i)}\right) \\ &= -\sum_i 1 = -m\end{aligned}$$

Finally, we have:

$$\phi_j = \frac{\sum_i w_j^{(i)}}{m}$$

K-Means from GMM perspective

From GMM to K-means:

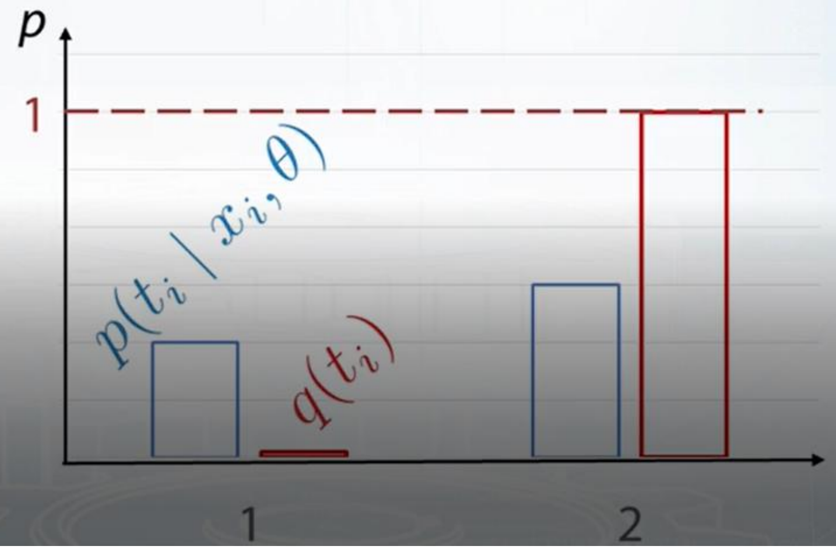
- Fix covariances to be identical $\Sigma_c = I$
- Fix weights to be uniform $\pi_c = \dfrac{1}{\text{\# of Guassians}}$

$$p(x_i \mid t_i = c, \theta) = \frac{1}{Z} \exp\left(-0.5\|x_i - \mu_c\|^2\right)$$

E-step

$$q^{k+1} = \arg\min_{q \in Q} \mathcal{KL}\left[q(T) \parallel p(T \mid X, \theta^k)\right]$$

Where Q is the set of delta-functions



E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\max_c p(t_i = c \mid x_i, \theta) = \arg\min_c \|x_i - \mu_c\|^2$$

$$p(t_i \mid x_i, \theta) = \frac{1}{Z} p(x_i \mid t_i, \theta)\, p(t_i \mid \theta)$$

$$= \frac{1}{Z} \exp\left(-0.5\|x_i - \mu_c\|^2\right) \pi_c$$

E-step

$$q^{k+1}(t_i) = \begin{cases} 1 & \text{if } t_i = c_i \\ 0 & \text{otherwise} \end{cases}$$

$$c_i = \arg\min_c \|x_i - \mu_c\|^2$$

Exactly like in K-Means!

$$\max_{\mu} \sum_{i=1}^{N} \mathbb{E}_{q(t_i)} \log P(x_i, t_i \mid \mu)$$

$$\mu_c = \frac{\sum_{i=1}^{N} \left(q(t_i = c) \cdot x_i \right)}{\sum_{i=1}^{N} q(t_i = c)} = \frac{\sum_{i:\, c_i^* = c} x_i}{\# i : c_i^* = c}$$

$$q(t_i) = \begin{cases} 1, & t_i = c_i^* \\ 0, & t_i \neq c_i^* \end{cases}$$

Reference

- Bayesian Methods for Machine Learning, HSE university, Coursera