Stat 215A Week 9b

10/19/2018 - Zoe Vernon

Thanks to Rebecca Barter for sharing her slides

The Expectation-Maximization (EM) algorithm

Intuition behind the EM algorithm

The EM algorithm

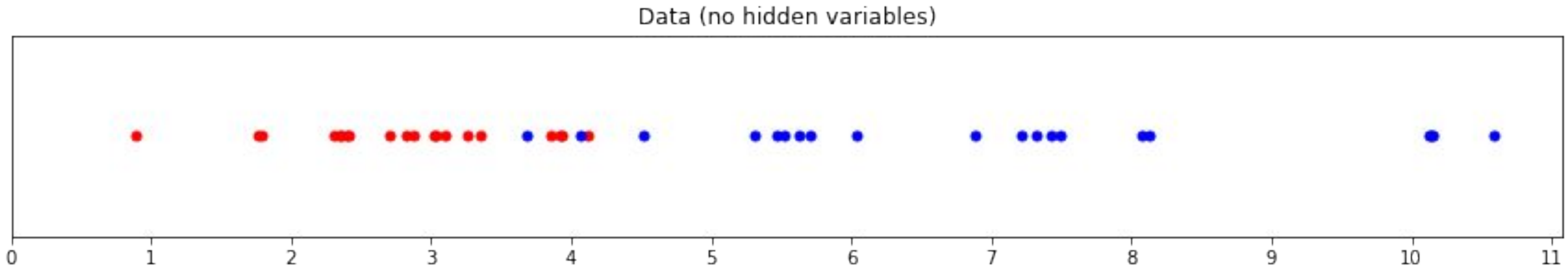
The material for this section came from this discussion on Stack Overflow (primarily the first answer by Alex Riley)

https://stackoverflow.com/questions/11808074/what-is-an-intuitive-explanation-of-the-expectation-maximization-technique

The EM algorithm

Suppose that we have some data samples from a mixture of two Gaussians.

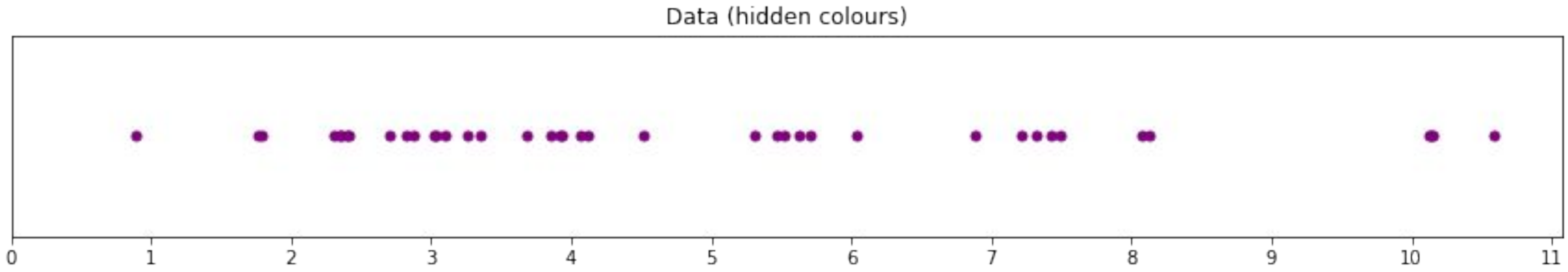
We want to know the mean and standard deviation of these Gaussians.



The EM algorithm

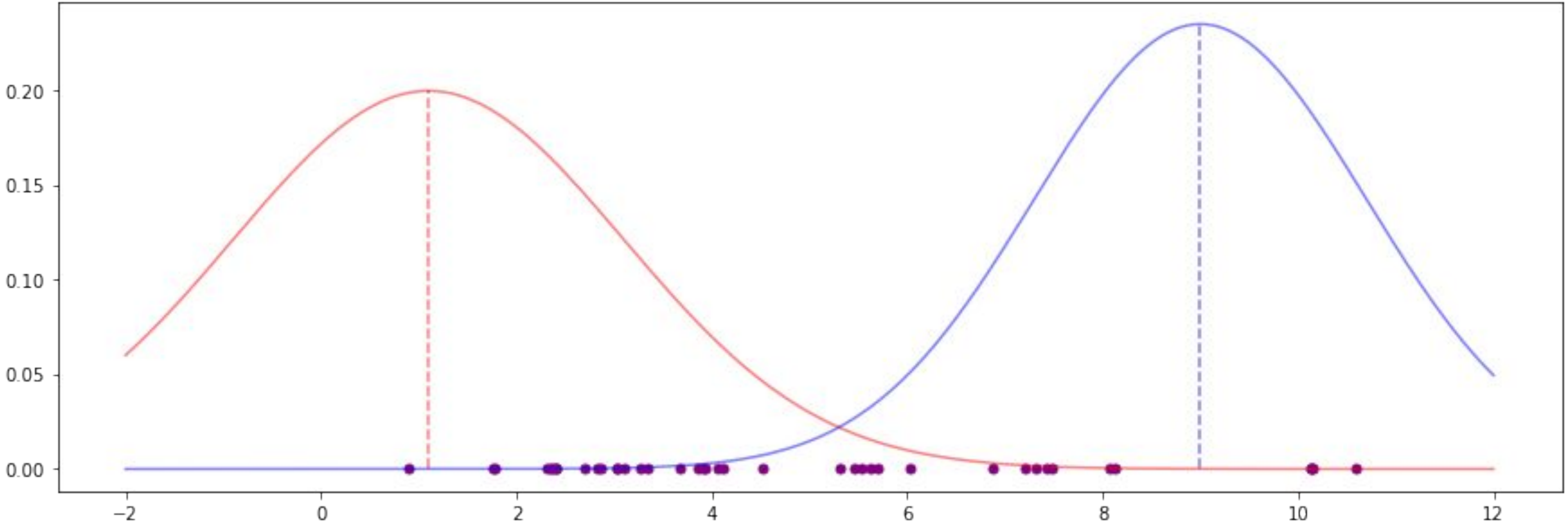
However, we don't actually know which data point came from which Gaussian.

If we knew that then the problem would be easy!



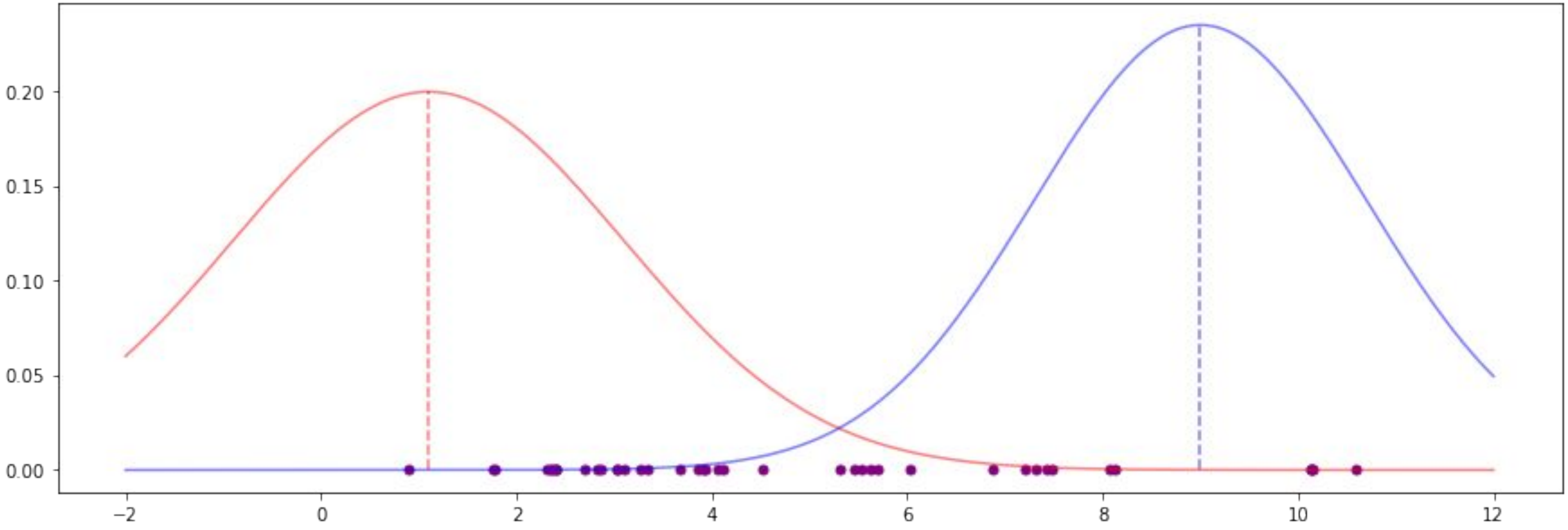
The EM algorithm

Step 1: Start with initial estimates of the mean and standard deviation for each Gaussian.



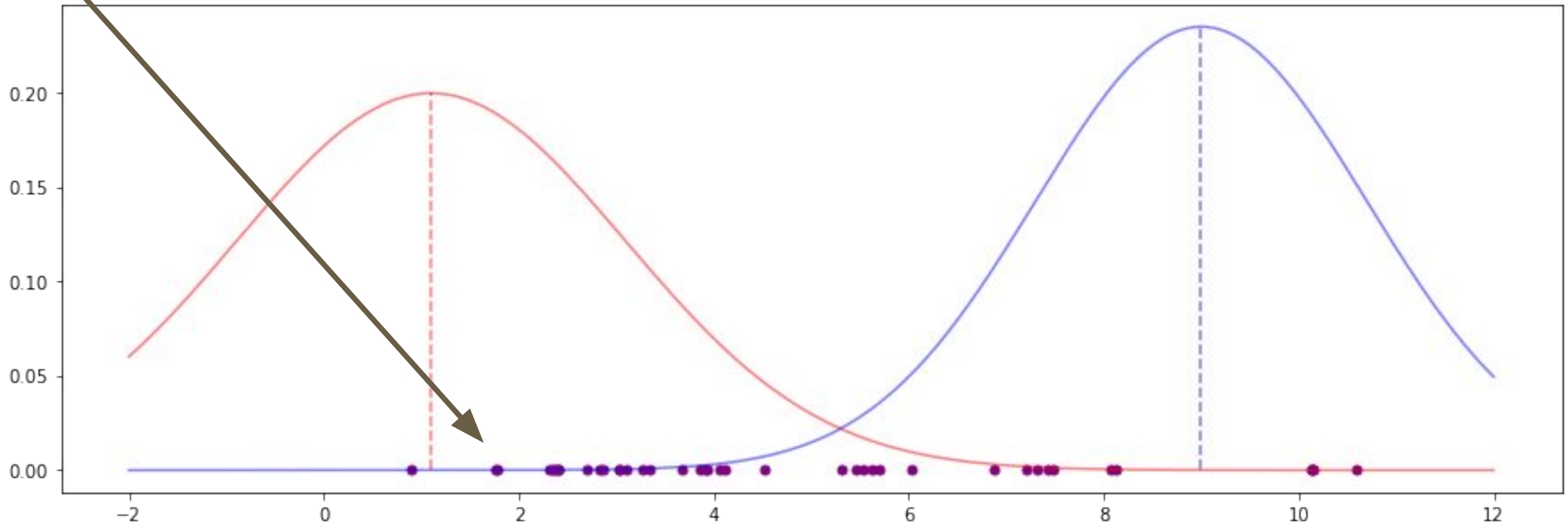
The EM algorithm

Step 2: Compute the likelihood of each data point appearing under the current parameter guesses (using the density for each estimated mean and standard error)



The EM algorithm

Step 2: Consider the point 1.761. The value of the red density at that point is p = 0.189 and the value of the blue density at that point is p = 0.00003. The point is more likely to come from the red density.



The EM algorithm

Step 3: Turn these two likelihood values into weights so that the weights sum to 1.

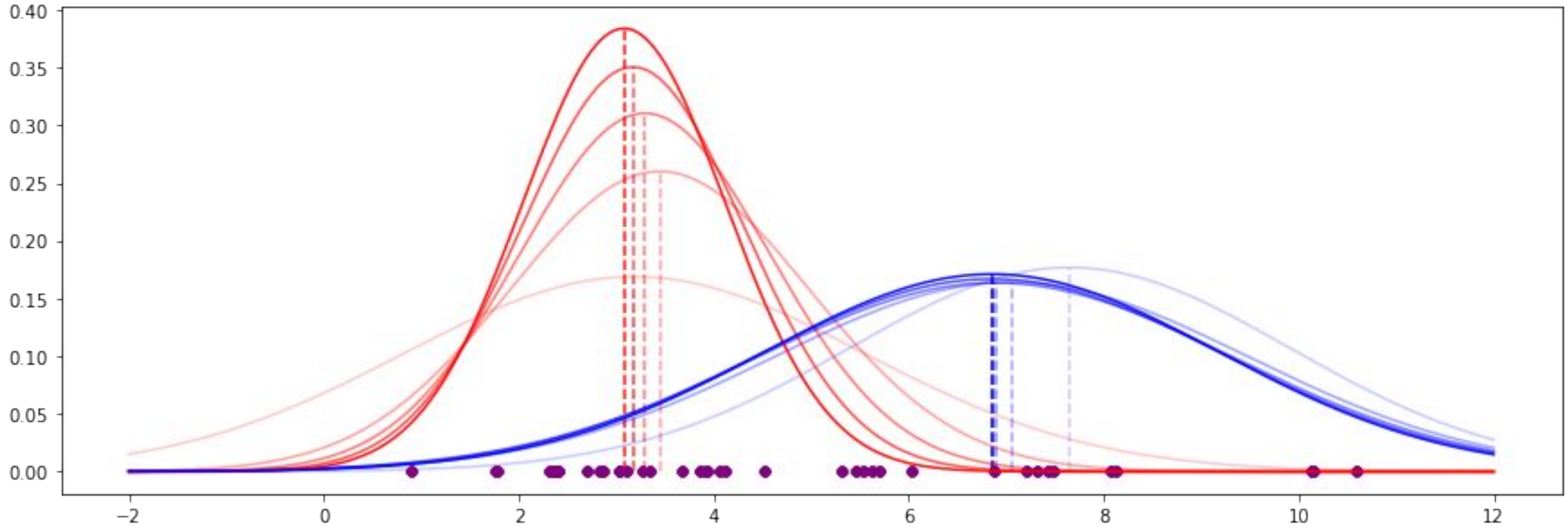
For each data point these weights are

$$\text{weight}_{\text{red}}(x) = \frac{\text{red likelihood at } x}{\sum_x \{\text{red likelihood at } x + \text{blue likelihood at } x\}}$$

$$\text{weight}_{\text{blue}}(x) = \frac{\text{blue likelihood at } x}{\sum_x \{\text{red likelihood at } x + \text{blue likelihood at } x\}}$$

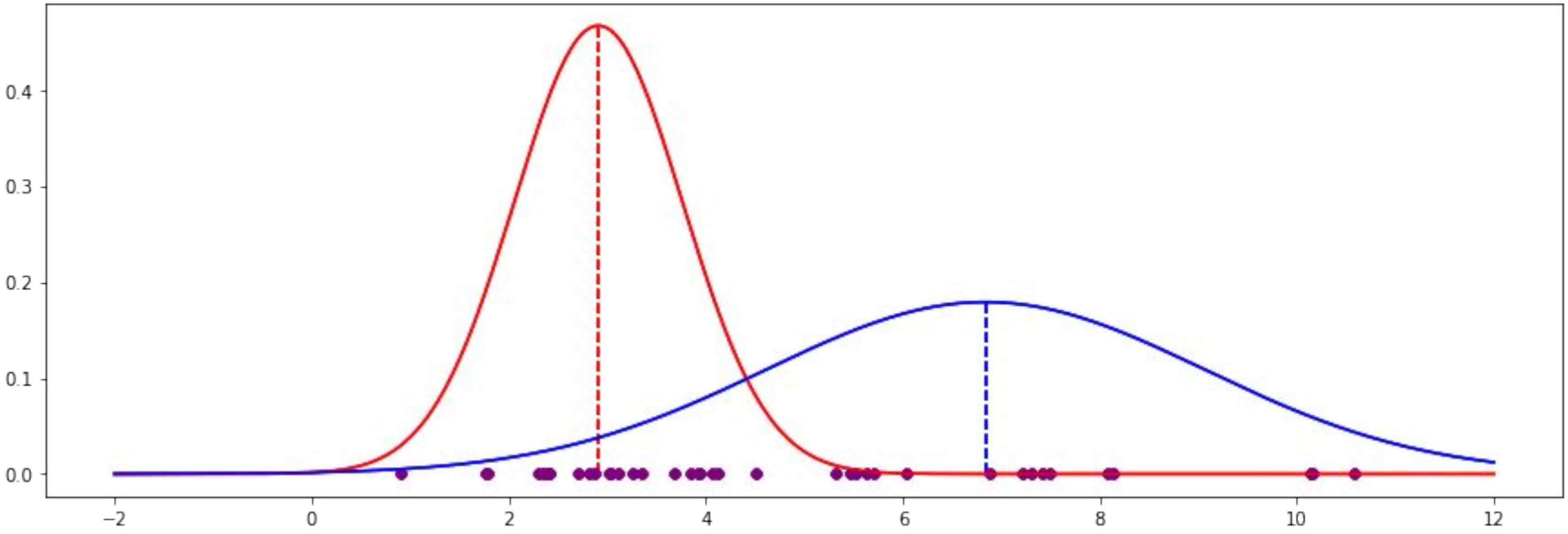
The EM algorithm

Step 4: Weight each data point by red weights to re-estimate red mean and SD.
Weight each data point by blue weights and to re-estimate blue mean and SD.
Step 5: repeat steps 2 through 4.



The EM algorithm

After 20 iterations...



Mathematical formulation of the EM algorithm

The EM algorithm

The material for this section came from the following summary paper

https://arxiv.org/pdf/1105.1476.pdf

The EM algorithm

EM is a general theory for calculating maximum likelihood estimates (MLE)

Let ***Y*** be a random variable with density $p(y|\theta)$

- ❑ $\boldsymbol{\theta}$ is an unknown parameter vector

The EM algorithm

EM is a general theory for calculating maximum likelihood estimates (MLE)

Let ***Y*** be a random variable with density $p(y|\theta)$

- $\boldsymbol{\theta}$ is an unknown parameter vector

Given observed data, ***y***, our aim is to maximize the likelihood function $p(y|\theta)$ with respect to $\boldsymbol{\theta}$

The EM algorithm

EM is a general theory for calculating maximum likelihood estimates (MLE)

Let ***Y*** be a random variable with density $p(y|\theta)$

- **θ** is an unknown parameter vector

Given observed data, ***y***, our aim is to maximize the likelihood function $p(y|\theta)$ with respect to **θ**

In many situations, there is no closed form solution to this problem. EM provides a numerical approximation to the MLE.

The EM algorithm

EM is a likelihood maximizer.

It iteratively maximizes successive local approximations of the likelihood function.

The EM algorithm

EM is a likelihood maximizer.

It iteratively maximized successive local approximations of the likelihood function.

These are the two steps

1. **E-step**: approximate the likelihood function
2. **M-step**: maximize this approximation with respect to $\boldsymbol{\theta}$

The EM algorithm

In EM, we have a latent (unobserved) variable, ***Z***, who density depends on $\boldsymbol{\theta}$.

In a mixture model, we assume that we first sample ***z***, and then we sample the observables ***y*** from a distribution that depends on ***z***

$$p(z, y|\theta) = p(z|\theta)p(y|z)$$

EM as a consequence on Jensen's inequality

Let's define $L(\theta) \equiv \log p(y|\theta)$

Then taking any two value of the parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we can show that

$$L(\theta) - L(\theta') = \log \frac{p(y|\theta)}{p(y|\theta')}$$

(by definition of ***L***)

EM as a consequence on Jensen's inequality

Let's define $L(\theta) \equiv \log p(y|\theta)$

Then taking any two value of the parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we can show that

$$\begin{aligned} L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} \\ &= \log \int \frac{p(z, y|\theta)}{p(y|\theta')} \, dz \end{aligned}$$

(by definition of ***L***)

(since the marginal density of ***y*** is the integral of the joint density of ***z*** and ***y***)

EM as a consequence on Jensen's inequality

Let's define $L(\theta) \equiv \log p(y|\theta)$

Then taking any two value of the parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we can show that

$$L(\theta) - L(\theta') = \log \frac{p(y|\theta)}{p(y|\theta')}$$

(by definition of ***L***)

$$= \log \int \frac{p(z, y|\theta)}{p(y|\theta')} \, dz$$

(since the marginal density of ***y*** is the integral of the joint density of ***z*** and ***y***)

$$= \log \int \frac{p(z, y|\theta)}{p(z, y|\theta')} p(z|y, \theta') \, dz$$

(since ***P(A, B) = P(A|B)P(B)***)

EM as a consequence on Jensen's inequality

Let's define $L(\theta) \equiv \log p(y|\theta)$

Then taking any two value of the parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we can show that

$$\begin{aligned} L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} \\ &= \log \int \frac{p(z,y|\theta)}{p(y|\theta')}\, dz \\ &= \log \int \frac{p(z,y|\theta)}{p(z,y|\theta')}\, p(z|y,\theta')\, dz \\ &= \log \int \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dz \end{aligned}$$

(by definition of ***L***)

(since the marginal density of ***y*** is the integral of the joint density of ***z*** and ***y***)

(since ***P(A, B) = P(A|B)P(B)***)

(using $p(z,y|\theta) = p(z|\theta)p(y|z)$)

EM as a consequence on Jensen's inequality

Let's define $L(\theta) \equiv \log p(y|\theta)$

Then taking any two value of the parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we can show that

$$
\begin{aligned}
L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} && \text{(by definition of } \boldsymbol{L}\text{)} \\
&= \log \int \frac{p(z,y|\theta)}{p(y|\theta')}\, dz && \text{(since the marginal density of } \boldsymbol{y} \text{ is the integral of the joint density of } \boldsymbol{z} \text{ and } \boldsymbol{y}\text{)} \\
&= \log \int \frac{p(z,y|\theta)}{p(z,y|\theta')} p(z|y,\theta')\, dz && \text{(since } \boldsymbol{P(A, B) = P(A|B)P(B)}\text{)} \\
&= \log \int \frac{p(z|\theta)}{p(z|\theta')} p(z|y,\theta')\, dz && \text{(using } p(z,y|\theta) = p(z|\theta)p(y|z)\text{)} \\
&\geq \underbrace{\int \log \frac{p(z|\theta)}{p(z|\theta')} p(z|y,\theta')\, dz}_{\text{Call this } Q(\theta,\theta')} && \text{(by Jensen's inequality)}
\end{aligned}
$$

EM as a consequence of Jensen's inequality

$$L(\theta) - L(\theta') \quad \geq \underbrace{\int \log \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dz}_{\text{Call this } Q(\theta,\theta')}$$

***Q*(θ, θ′)** is thus an auxiliary function for the log-likelihood ***L*(θ)** in that

1. The increase in likelihood when moving from **θ** to **θ′** is always greater than ***Q*(θ, θ′)**
2. ***Q*(θ′, θ′) = 0**

EM as a consequence of Jensen's inequality

$$L(\theta) - L(\theta') \geq \underbrace{\int \log \frac{p(z|\theta)}{p(z|\theta')} p(z|y,\theta')\, dz}_{\text{Call this } Q(\theta,\theta')}$$

***Q*(θ, θ')** is thus an auxiliary function for the log-likelihood ***L*(θ)** in that

1. The increase in likelihood when moving from **θ** to **θ'** is always greater than ***Q*(θ, θ')**
2. ***Q*(θ', θ') = *0***

Starting from an initial guess **θ'**, we are guaranteed to increase the likelihood value if we can find a **θ** such that ***Q*(θ, θ') > *0***

EM as a consequence of Jensen's inequality

$$L(\theta) \equiv \log p(y|\theta)$$

$$Q(\theta,\theta') \equiv \int \log \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dz$$

Using EM, we will maximize $Q(\theta^{t+1},\theta^{t})$ instead of the difference in likelihood functions $L(\theta^{t+1}) - L(\theta^{t})$

EM as expectation-maximization

We can decompose ***Q*** into the following difference:

$$Q(\theta, \theta') = Q(\theta|\theta') - Q(\theta'|\theta')$$

where

$$Q(\theta|\theta') \equiv \int \log p(z|\theta)\, p(z|y, \theta')\, dx \;\; \equiv \mathrm{E}[\log p(Z|\theta)|y, \theta']$$

EM as expectation-maximization

We can decompose ***Q*** into the following difference:

$$Q(\theta, \theta') = Q(\theta|\theta') - Q(\theta'|\theta')$$

where

$$Q(\theta|\theta') \equiv \int \log p(z|\theta)\, p(z|y, \theta')\, dx \;\; \equiv \mathrm{E}[\log p(Z|\theta)|y, \theta']$$

For a fixed θ' maximizing ***Q(θ, θ')*** wrt θ is equivalent to maximizing ***Q(θ | θ')*** (i.e. we can ignore the second term).

EM as expectation maximization

Given a current parameter estimate θ_n

E-step: form the auxiliary function $Q(\theta|\theta_n)$ which involves computing the posterior distribution of the unobserved variable

$$Q(\theta|\theta_n) \equiv \int \log p(z|\theta)\, p(z|y,\theta_n)\, dx \quad \equiv \mathrm{E}[\log p(Z|\theta)|y,\theta_n]$$

M-step: update the parameter estimate by maximizing the auxiliary function

$$\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n)$$

A worked example