STAT 215A Fall 2017 Week 9

Rebecca Barter
10/20/2017

The Expectation-Maximization (EM) algorithm

The intuition behind the EM algorithm

The Expectation-Maximization algorithm

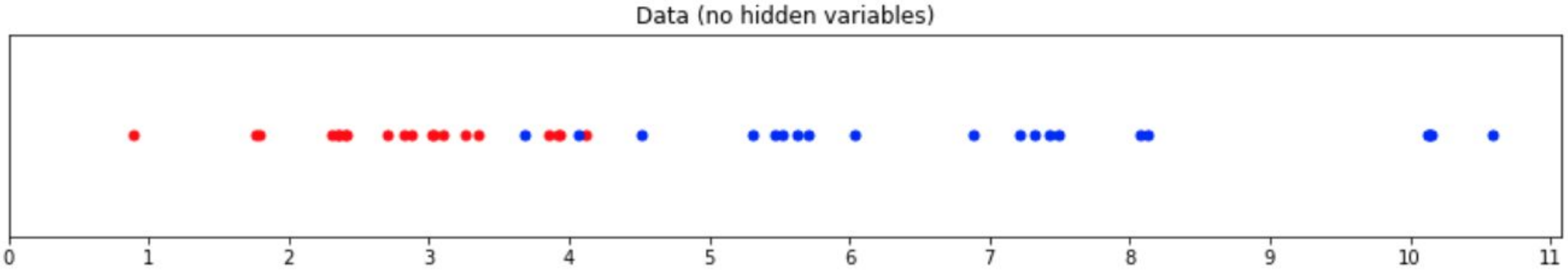
The material for this section came from this discussion on Stack Overflow (primarily the first answer by Alex Riley):

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

The Expectation-Maximization algorithm

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

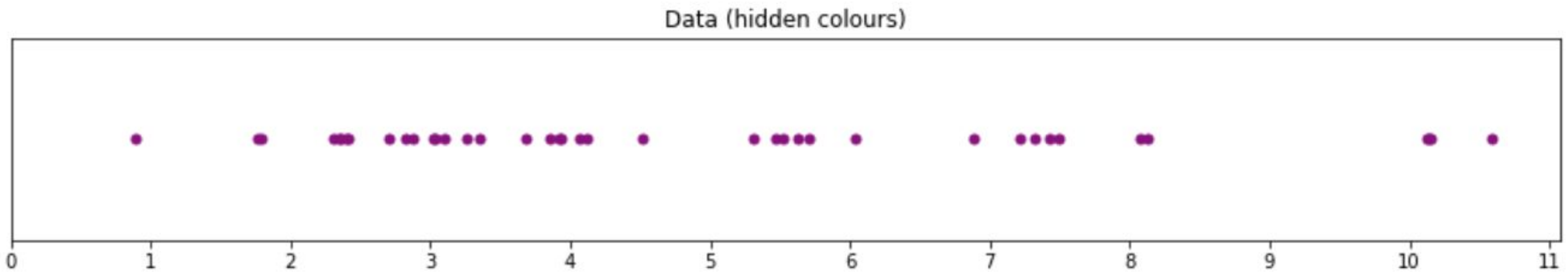
We want to know the mean and standard deviation of these Gaussians



The Expectation-Maximization algorithm

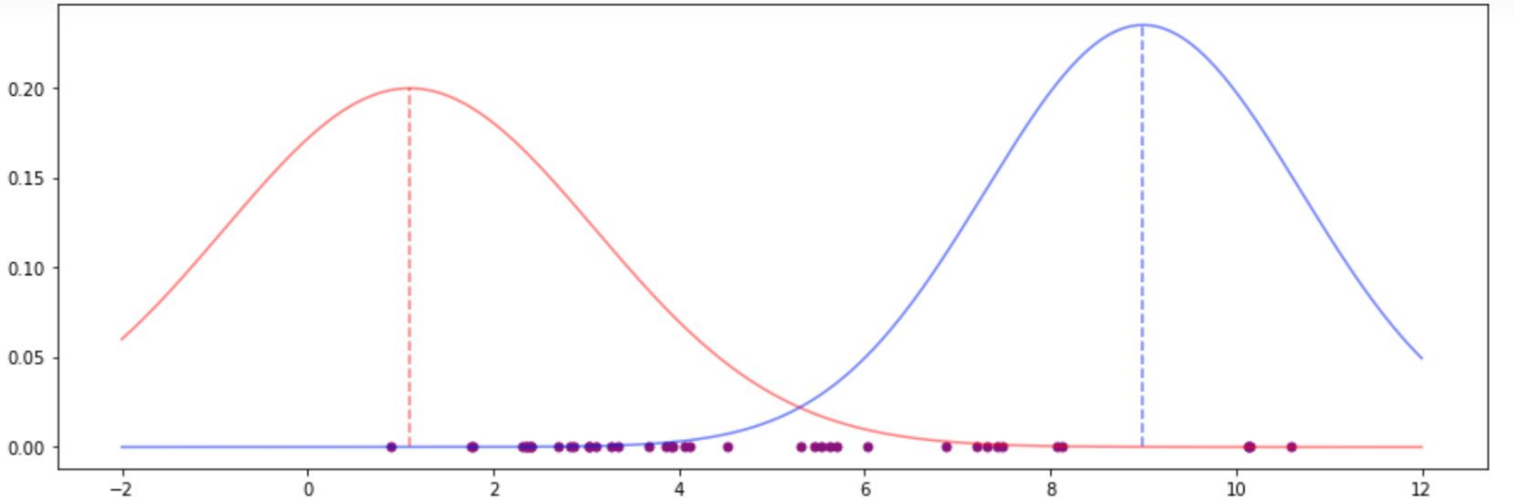
The problem: we don't actually know which data point came from which Gaussian.

If we knew that the problem would be easy!



The Expectation-Maximization algorithm

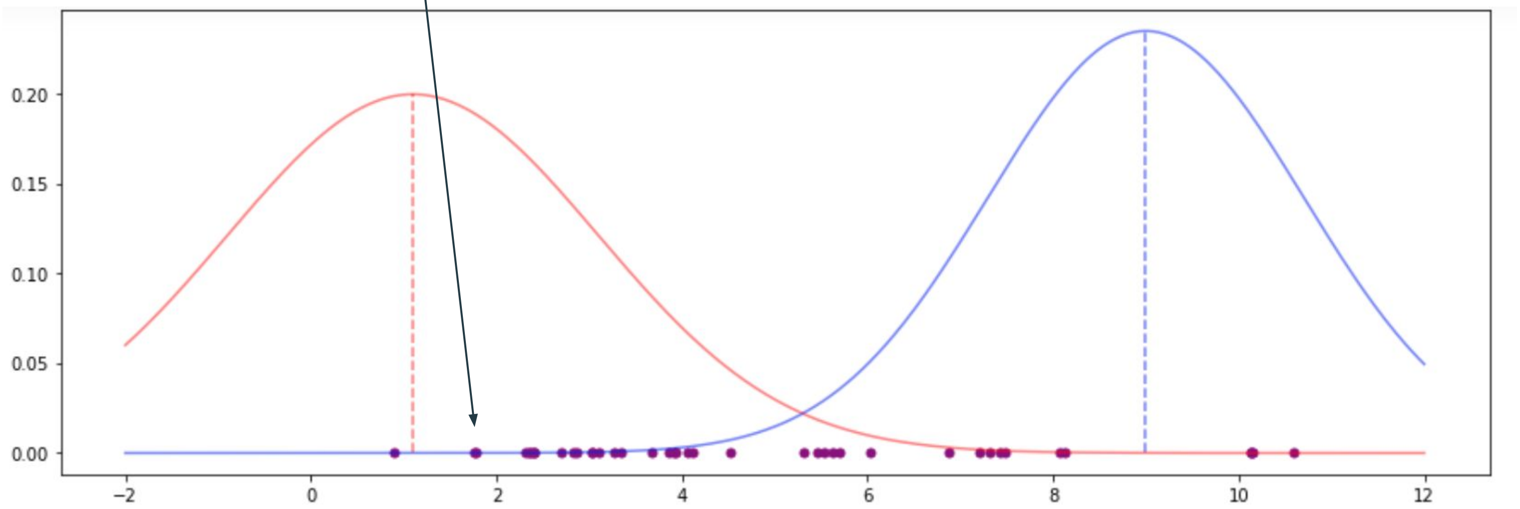
1. Start with initial estimates of the mean and standard deviation for each Gaussian



The Expectation-Maximization algorithm

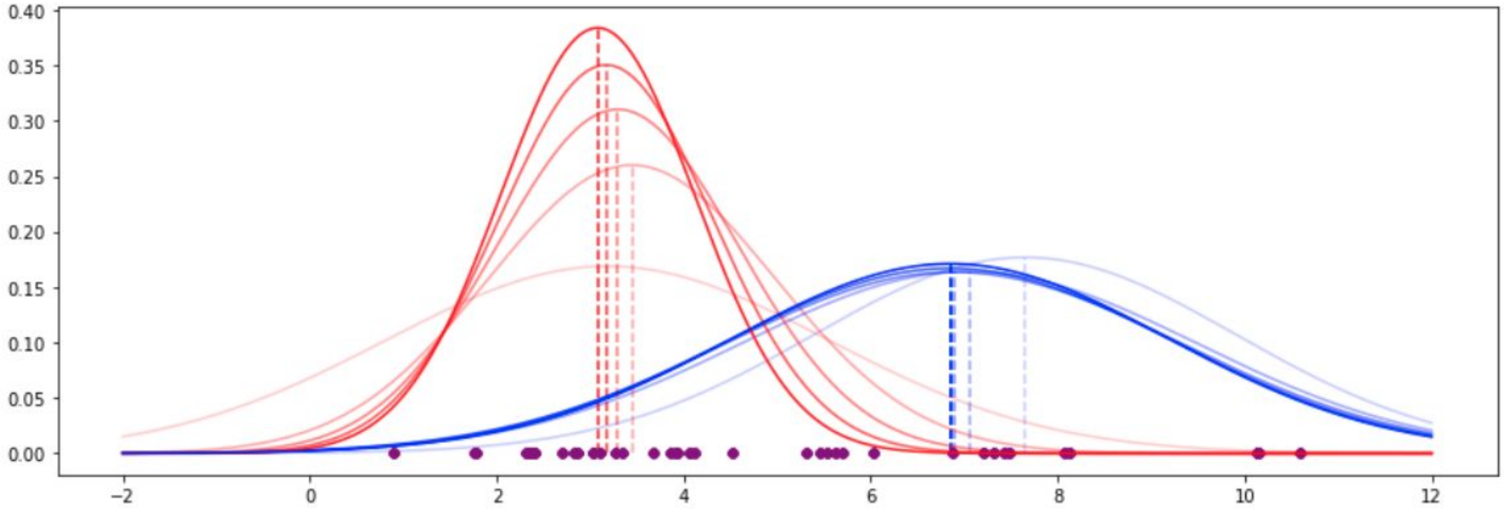
2. Compute the likelihood of each data point appearing under the current parameter guesses (using the density for each estimated Gaussian)

- the data point at 1.761 is more likely to be red (p=0.189) than blue (p=0.00003)



The Expectation-Maximization algorithm

3. Turn these two likelihood values into weights so that they sum to 1
4. Use these weights to re-estimate the mean and SD for each Gaussian
5. Repeat stages 2-4



The mathematical formulation of the EM algorithm

The Expectation-Maximization algorithm

The material for this section came from the following summary paper (by Alexis Roche 2003):

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

The Expectation-Maximization 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

The Expectation-Maximization 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

$$p(y|\theta) \ \text{wrt} \ \theta$$

The Expectation-Maximization 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

$$p(y|\theta) \;\; \text{wrt} \;\; \theta$$

Except in basic situations, there is no closed form solution to this problem.

EM provides a numerical approximation to the MLE.

The Expectation-Maximization algorithm

EM is a likelihood maximizer.

It iteratively maximizes successive local approximations of the likelihood function.

The Expectation-Maximization algorithm

EM is a likelihood maximizer.

It iteratively maximizes successive local approximations of the likelihood function.

There are two steps:

1. The **E-step**: approximate the likelihood function
2. The **M-step**: maximize this approximation with respect to θ

The Expectation-Maximization algorithm

In EM, we have a “latent” variable, Z, whose density depends on θ

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 of Jensen's inequality

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

EM as a consequence of Jensen’s inequality

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

Then taking any two values of the parameter vector θ and θ', we can show that

$$L(\theta) - L(\theta') = \log \frac{p(y|\theta)}{p(y|\theta')} \qquad \text{(by definition of } L\text{)}$$

EM as a consequence of Jensen's inequality

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

Then taking any two values of the parameter vector θ and θ', we can show that

$$\begin{aligned} L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} && \text{(by definition of } L\text{)} \\ &= \log \int \frac{p(z, y|\theta)}{p(y|\theta')}\, dz && \text{(since the marginal density of } y \text{ is the integral of the joint density of } z \text{ and } y\text{)} \end{aligned}$$

EM as a consequence of Jensen's inequality

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

Then taking any two values of the parameter vector θ and θ', we can show that

$$\begin{aligned} L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} && \text{(by definition of } L) \\ &= \log \int \frac{p(z,y|\theta)}{p(y|\theta')}\, dz && \text{(since the marginal density of } y \text{ is the integral of the joint density of } z \text{ and } y) \\ &= \log \int \frac{p(z,y|\theta)}{p(z,y|\theta')}\, p(z|y,\theta')\, dz && \text{(since } P(A, B) = P(A \mid B)\, P(B)) \end{aligned}$$

EM as a consequence of Jensen's inequality

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

Then taking any two values of the parameter vector θ and θ', we can show that

$$
\begin{aligned}
L(\theta) - L(\theta') &= \log \frac{p(y|\theta)}{p(y|\theta')} && \text{(by definition of } L\text{)} \\
&= \log \int \frac{p(z,y|\theta)}{p(y|\theta')}\, dz && \text{(since the marginal density of } y \text{ is the integral of the joint density of } z \text{ and } y\text{)} \\
&= \log \int \frac{p(z,y|\theta)}{p(z,y|\theta')}\, p(z|y,\theta')\, dz && \text{(since } P(A, B) = P(A \mid B)\, P(B)\text{)} \\
&= \log \int \frac{p(z|\theta)}{p(z|\theta')}\, p(z|y,\theta')\, dz && \text{(using } p(z,y|\theta) = p(z|\theta)p(y|z)\text{)}
\end{aligned}
$$

EM as a consequence of Jensen's inequality

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

Then taking any two values of the parameter vector θ and θ', we can show that

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

EM as a consequence of Jensen's inequality

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

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

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

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 $L(\theta)$ in that

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

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

EM as a consequence of Jensen's inequality

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

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

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

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

Iterating such a process defines the EM algorithm

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 $Q(\theta, \theta')$ wrt θ is equivalent to maximizing $Q(\theta \mid \theta')$

(i.e. we can ignore the second term in the equation above)

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

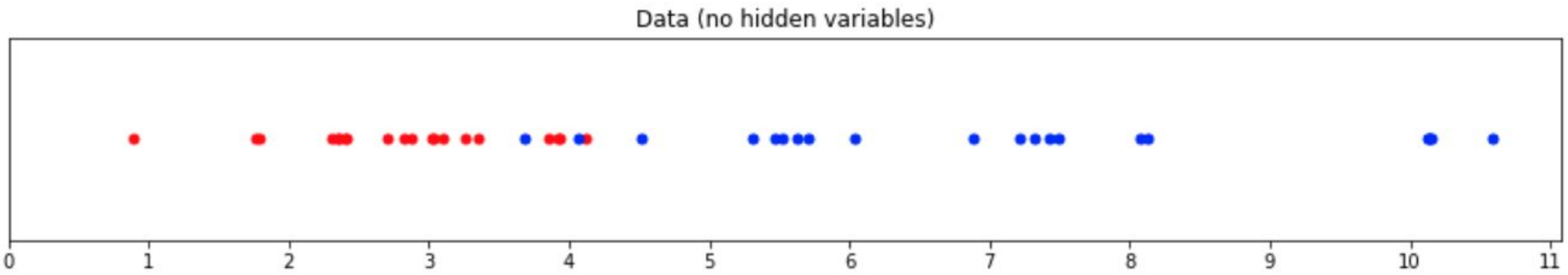
$$\theta_{n+1} = \arg\max_{\theta} Q(\theta|\theta_n)$$

A worked example: Gaussian mixtures

Returning back to our Gaussian Mixture example

The materials for this section can be found mostly on the wikipedia page

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm#Gaussian_mixture



Gaussian Mixture example

Consider the mixture model

$$X_i | (Z_i = 1) \sim \mathcal{N}_d(\boldsymbol{\mu}_1, \Sigma_1) \text{ and}$$
$$X_i | (Z_i = 2) \sim \mathcal{N}_d(\boldsymbol{\mu}_2, \Sigma_2)$$

Let Z_i be the indicator that observation i comes from group 1:

$$\mathrm{P}(Z_i = 1) = \tau_1 \text{ and}$$
$$\mathrm{P}(Z_i = 2) = \tau_2 = 1 - \tau_1$$

Z_i is unobserved...

Gaussian Mixture example

The aim is to estimate the unknown parameters

$$\theta = \left(\boldsymbol{\tau}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2\right)$$

To do this, we will use the EM algorithm

1. **E-step:** given the current estimate of the parameters, calculate the following conditional expectation:

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}\left[\log L(\theta;\mathbf{x},\mathbf{Z})\right]$$

2. **M-step**: find the argmax of $Q(\theta|\theta^{(t)})$ for each component of θ

Gaussian Mixture example: the E-step

The E-step involves performing the following calculation:

Given our current estimate of the parameters, the **conditional distribution** of Z_i is determined by Bayes theorem to be the proportional height of the normal density weighted by $\boldsymbol{\tau}$

$$T_{j,i}^{(t)} := \mathrm{P}(Z_i = j | X_i = \mathbf{x}_i ; \theta^{(t)}) = \frac{\tau_j^{(t)} \ f(\mathbf{x}_i ; \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}{\tau_1^{(t)} \ f(\mathbf{x}_i ; \boldsymbol{\mu}_1^{(t)}, \Sigma_1^{(t)}) + \tau_2^{(t)} \ f(\mathbf{x}_i ; \boldsymbol{\mu}_2^{(t)}, \Sigma_2^{(t)})}$$

Gaussian Mixture example: the E-step

The E-step involves performing the following calculation:

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}[\log L(\theta;\mathbf{x},\mathbf{Z})] \\
&= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}[\log \prod_{i=1}^{n} L(\theta;\mathbf{x}_i,\mathbf{z}_i)] \\
&= \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta^{(t)}}[\sum_{i=1}^{n} \log L(\theta;\mathbf{x}_i,\mathbf{z}_i)] \\
&= \sum_{i=1}^{n} \mathrm{E}_{\mathbf{Z}|\mathbf{X};\theta^{(t)}}[\log L(\theta;\mathbf{x}_i,\mathbf{z}_i)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2} P(Z_i = j | X_i = \mathbf{x}_i;\theta^{(t)}) \log L(\theta_j;\mathbf{x}_i,\mathbf{z}_i) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2} T_{j,i}^{(t)} \left[\log \tau_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j) - \tfrac{d}{2}\log(2\pi)\right]
\end{aligned}
$$

Gaussian Mixture example: the M-step

The M-step involves performing the following optimizations:

$$\boldsymbol{\tau}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} Q(\theta|\theta^{(t)})$$
$$= \arg\max_{\boldsymbol{\tau}} \left\{ \left[\sum_{i=1}^{n} T_{1,i}^{(t)} \right] \log \tau_1 + \left[\sum_{i=1}^{n} T_{2,i}^{(t)} \right] \log \tau_2 \right\}$$

Which can be shown to yield

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)}}{\sum_{i=1}^{n} (T_{1,i}^{(t)} + T_{2,i}^{(t)})} = \frac{1}{n} \sum_{i=1}^{n} T_{j,i}^{(t)}$$

Gaussian Mixture example: the M-step

The M-step involves performing the following optimizations:

$$
\begin{aligned}
(\boldsymbol{\mu}_1^{(t+1)}, \Sigma_1^{(t+1)}) &= \underset{\boldsymbol{\mu}_1, \Sigma_1}{\arg\max}\, Q(\theta|\theta^{(t)}) \\
&= \underset{\boldsymbol{\mu}_1, \Sigma_1}{\arg\max} \sum_{i=1}^{n} T_{1,i}^{(t)} \left\{ -\tfrac{1}{2}\log|\Sigma_1| - \tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1) \right\}
\end{aligned}
$$

Which can be shown to yield

$$\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{1,i}^{(t)}} \text{ and } \Sigma_1^{(t+1)} = \frac{\sum_{i=1}^{n} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^\top}{\sum_{i=1}^{n} T_{1,i}^{(t)}}$$

Gaussian Mixture example: the M-step

The M-step involves performing the following optimizations:

$$
\begin{aligned}
(\boldsymbol{\mu}_1^{(t+1)}, \Sigma_1^{(t+1)}) &= \arg\max_{\boldsymbol{\mu}_1, \Sigma_1} Q(\theta|\theta^{(t)}) \\
&= \arg\max_{\boldsymbol{\mu}_1, \Sigma_1} \sum_{i=1}^{n} T_{1,i}^{(t)} \left\{ -\tfrac{1}{2} \log|\Sigma_1| - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1) \right\}
\end{aligned}
$$

And by symmetry

$$
\boldsymbol{\mu}_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{2,i}^{(t)}} \text{ and } \Sigma_2^{(t+1)} = \frac{\sum_{i=1}^{n} T_{2,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})^\top}{\sum_{i=1}^{n} T_{2,i}^{(t)}}
$$