Expectation-Maximization for GMMs
Jason Corso

(If equation fonts are garbled in your reader, please use Adobe Reader; not sure why this happened...)}
Expectation-Maximization for GMMs

- **Expectation-Maximization** or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables \( z \) indicating the mixture component.

Recall the conditions that must be satisfied at a maximum of the likelihood function.

For the mean \( \mu_k \), setting the derivatives of \( \ln p(X | \pi, \mu, \Sigma) \) w.r.t. \( \mu_k \) to zero yields

\[
0 = -N \sum_{n=1}^{N} \pi_k N(x|\mu_k, \Sigma_k) \sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j) \Sigma_k (x_n - \mu_k) (20)
\]

\[
= -N \sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k) (21)
\]

Note the natural appearance of the responsibility terms on the RHS.
Expectation-Maximization for GMMs

- Expectation-Maximization or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $z$ indicating the mixture component.

- Recall the conditions that must be satisfied at a maximum of the likelihood function.

$$0 = -\sum_{n=1}^{N} \pi_k \sum_{j=1}^{K} \pi_j \sum_{k=1}^{K} (x_n - \mu_k) \Sigma_k$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \sum_{k=1}^{K} (x_n - \mu_k) \Sigma_k$$

Note the natural appearance of the responsibility terms on the RHS.
Expectation-Maximization or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $z$ indicating the mixture component.

Recall the conditions that must be satisfied at a maximum of the likelihood function.

For the mean $\mu_k$, setting the derivatives of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t. $\mu_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k) \quad (20)$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k) \quad (21)$$
Expectation-Maximization or EM is an elegant and powerful method for finding MLE solutions in the case of missing data such as the latent variables $z$ indicating the mixture component.

Recall the conditions that must be satisfied at a maximum of the likelihood function.

For the mean $\mu_k$, setting the derivatives of $\ln p(X|\pi, \mu, \Sigma)$ w.r.t. $\mu_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \Sigma_k (x_n - \mu_k)$$

(20)

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k)$$

(21)

Note the natural appearance of the responsibility terms on the RHS.
Multiplying by $\Sigma_k^{-1}$, which we assume is non-singular, gives

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$  \hspace{1cm} (22)$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$  \hspace{1cm} (23)$$
Multiplying by $\Sigma_k^{-1}$, which we assume is non-singular, gives

$$
\mathbf{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n
$$

(22)

where

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

(23)

We see the $k^{th}$ mean is the weighted mean over all of the points in the dataset.
Multiplying by $\Sigma_k^{-1}$, which we assume is non-singular, gives

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

where

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

We see the $k^{th}$ mean is the weighted mean over all of the points in the dataset.

Interpret $N_k$ as the number of points assigned to component $k$. 
- Multiplying by $\Sigma_k^{-1}$, which we assume is non-singular, gives

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$  \hspace{1cm} (22)

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$  \hspace{1cm} (23)

- We see the $k^{\text{th}}$ mean is the weighted mean over all of the points in the dataset.
- Interpret $N_k$ as the number of points assigned to component $k$.
- We find a similar result for the covariance matrix:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T.$$  \hspace{1cm} (24)
We also need to maximize $\ln p(X|\pi, \mu, \Sigma)$ with respect to the mixing coefficients $\pi_k$. 

Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$\ln p(X|\pi, \mu, \Sigma) + \lambda (K \sum_k \pi_k - 1)$ \hspace{1cm} (25)

Maximizing it yields:

$0 = 1 \sum_n k \gamma(z_{nk}) + \lambda$ \hspace{1cm} (26)

After multiplying both sides by $\pi$ and summing over $k$, we get

$\lambda = -N$ \hspace{1cm} (27)

Eliminate $\lambda$ and rearrange to obtain:

$\pi_k = N_k / N$ \hspace{1cm} (28)
We also need to maximize $\ln p(X|\pi, \mu, \Sigma)$ with respect to the mixing coefficients $\pi_k$.

Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

(25)
We also need to maximize $\ln p(X|\pi, \mu, \Sigma)$ with respect to the mixing coefficients $\pi_k$.

Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

Maximizing it yields:

$$0 = \frac{1}{N_k} \sum_{n=1}^{N_k} \gamma(z_{nk}) + \lambda$$
We also need to maximize $\ln p(X|\pi, \mu, \Sigma)$ with respect to the mixing coefficients $\pi_k$.

Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

(25)

Maximizing it yields:

\[
0 = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) + \lambda
\]

(26)

After multiplying both sides by $\pi$ and summing over $k$, we get

\[
\lambda = -N
\]

(27)
We also need to maximize $\ln p(X|\pi, \mu, \Sigma)$ with respect to the mixing coefficients $\pi_k$.

Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

$$\ln p(X|\pi, \mu, \Sigma) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right)$$  \hspace{1cm} (25)

Maximizing it yields:

$$0 = \frac{1}{N_k} \sum_{n=1}^{N_k} \gamma(z_{nk}) + \lambda$$  \hspace{1cm} (26)

After multiplying both sides by $\pi$ and summing over $k$, we get

$$\lambda = -N$$  \hspace{1cm} (27)

Eliminate $\lambda$ and rearrange to obtain:

$$\pi_k = \frac{N_k}{N}$$  \hspace{1cm} (28)
Solved...right?

- So, we’re done, right? We’ve computed the maximum likelihood solutions for each of the unknown parameters.
So, we’re done, right? We’ve computed the maximum likelihood solutions for each of the unknown parameters.

Wrong!
Solved...right?

- So, we’re done, right? We’ve computed the maximum likelihood solutions for each of the unknown parameters.

- Wrong!

- The responsibility terms depend on these parameters in an intricate way:

\[
\gamma(z_k) \equiv p(z_k = 1|x) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
\]
Solved...right?

- So, we’re done, right? We’ve computed the maximum likelihood solutions for each of the unknown parameters.
- Wrong!
- The responsibility terms depend on these parameters in an intricate way:

\[
\gamma(z_k) = p(z_k = 1|x) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
\]

- But, these results do suggest an iterative scheme for finding a solution to the maximum likelihood problem.
  1. Choose some initial values for the parameters, \(\pi, \mu, \Sigma\).
  2. Use the current parameters estimates to compute the posteriors on the latent terms, i.e., the responsibilities.
  3. Use the responsibilities to update the estimates of the parameters.
  4. Repeat 2 and 3 until convergence.
(a) $L = 1$
(b) $L = 2$
(c) $L = 5$
(d) $L = 20$
(a) $L = 1$

(b) $L = 2$

(c) $L = 5$

(d) $L = 20$
Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.

- Care must be taken to avoid singularities in the MLE solution.
- There will generally be multiple local maxima of the likelihood function and EM is not guaranteed to find the largest of these.
Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.

- Care must be taken to avoid singularities in the MLE solution.
- There will generally be multiple local maxima of the likelihood function and EM is not guaranteed to find the largest of these.
Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.
- Care must be taken to avoid singularities in the MLE solution.
Some Quick, Early Notes on EM

- EM generally tends to take more steps than the K-Means clustering algorithm.
- Each step is more computationally intense than with K-Means too.
- So, one commonly computes K-Means first and then initializes EM from the resulting clusters.
- Care must be taken to avoid singularities in the MLE solution.
- There will generally be multiple local maxima of the likelihood function and EM is not guaranteed to find the largest of these.
Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covariances, and the mixing coefficients).

1. Initialize the means, $\mu_k$, the covariances, $\Sigma_k$, and mixing coefficients, $\pi_k$. Evaluate the initial value of the log-likelihood.
Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covariances, and the mixing coefficients).

1. Initialize the means, $\mu_k$, the covariances, $\Sigma_k$, and mixing coefficients, $\pi_k$. Evaluate the initial value of the log-likelihood.

2. **E-Step** Evaluate the responsibilities using the current parameter values:

$$
\gamma(z_k) = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
$$
Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covariances, and the mixing coefficients).

1. Initialize the means, \( \mu_k \), the covariances, \( \Sigma_k \), and mixing coefficients, \( \pi_k \). Evaluate the initial value of the log-likelihood.

2. **E-Step** Evaluate the responsibilities using the current parameter values:

\[
\gamma(z_k) = \frac{\pi_k \mathcal{N}(x | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x | \mu_j, \Sigma_j)}
\]

3. **M-Step** Update the parameters using the current responsibilities

\[
\mu_{k}^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n
\]

\[
\Sigma_{k}^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_{k}^{\text{new}})(x_n - \mu_{k}^{\text{new}})^T
\]

\[
\pi_{k}^{\text{new}} = \frac{N_k}{N}
\]

where

\[
N_k = \sum_{n=1}^{N} \gamma(z_{nk})
\]
4. Evaluate the log-likelihood

\[
\ln p(X|\mu^{\text{new}}, \Sigma^{\text{new}}, \pi^{\text{new}}) = \sum_{n=1}^{N} \ln \left[ \sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}(x_n|\mu_k^{\text{new}}, \Sigma_k^{\text{new}}) \right]
\] (33)

5. Check for convergence of either the parameters of the log-likelihood. If the convergence is not satisfied, set the parameters:

\[
\mu = \mu^{\text{new}} (34)
\]
\[
\Sigma = \Sigma^{\text{new}} (35)
\]
\[
\pi = \pi^{\text{new}} (36)
\]
and goto step 2.
4. Evaluate the log-likelihood

\[
\ln p(X|\mu^{\text{new}}, \Sigma^{\text{new}}, \pi^{\text{new}}) = \sum_{n=1}^{N} \ln \left[ \sum_{k=1}^{K} \pi_{k}^{\text{new}} N(x_n|\mu_{k}^{\text{new}}, \Sigma_{k}^{\text{new}}) \right]
\]  

(33)

5. Check for convergence of either the parameters of the log-likelihood. If the convergence is not satisfied, set the parameters:

\[
\begin{align*}
\mu &= \mu^{\text{new}} \tag{34} \\
\Sigma &= \Sigma^{\text{new}} \tag{35} \\
\pi &= \pi^{\text{new}} \tag{36}
\end{align*}
\]

and goto step 2.
A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\theta$, and so the log-likelihood function is

$$\ln p(X|\theta) = \ln \left[ \sum_{Z} p(X, Z|\theta) \right]$$

(37)
A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\theta$, and so the log-likelihood function is

\[
\ln p(X|\theta) = \ln \left[ \sum_Z p(X, Z|\theta) \right]
\]  

(37)

- Note how the summation over the latent variables appears inside of the log.
  - Even if the joint distribution $p(X, Z|\theta)$ belongs to the exponential family, the marginal $p(X|\theta)$ typically does not.
A More General View of EM

- The goal of EM is to find maximum likelihood solutions for models having latent variables.
- Denote the set of all model parameters as $\theta$, and so the log-likelihood function is

$$\ln p(X | \theta) = \ln \left[ \sum_{Z} p(X, Z | \theta) \right]$$  \hspace{1cm} (37)

- Note how the summation over the latent variables appears inside of the log.
  - Even if the joint distribution $p(X, Z | \theta)$ belongs to the exponential family, the marginal $p(X | \theta)$ typically does not.
- If, for each sample $x_n$ we were given the value of the latent variable $z_n$, then we would have a complete data set, $\{X, Z\}$, with which maximizing this likelihood term would be straightforward.
However, in practice, we are not given the latent variables values.
However, in practice, we are not given the latent variables values.

So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
However, in practice, we are not given the latent variables values.

So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.

In the E-Step, we use the current parameter values $\theta^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(Z|X, \theta^{\text{old}})$. 

Then, in the M-step, we revise the parameters to $\theta^{\text{new}}$ by maximizing this function:

$$
\theta^{\text{new}} = \arg \max \theta Q(\theta, \theta^{\text{old}})
$$

Note that the log acts directly on the joint distribution $p(X, Z|\theta)$ and so the M-step maximization will likely be tractable.
However, in practice, we are not given the latent variables values.
So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.
In the E-Step, we use the current parameter values $\theta^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(Z|X, \theta^{\text{old}})$.
This posterior is used to define the expectation of the complete-data log-likelihood, denoted $Q(\theta, \theta^{\text{old}})$, which is given by

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (38)$$
However, in practice, we are not given the latent variables values.

So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.

In the E-Step, we use the current parameter values $\theta^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(Z|X, \theta^{\text{old}})$.

This posterior is used to define the expectation of the complete-data log-likelihood, denoted $Q(\theta, \theta^{\text{old}})$, which is given by

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (38)$$

Then, in the M-step, we revise the parameters to $\theta^{\text{new}}$ by maximizing this function:

$$\theta^{\text{new}} = \operatorname{arg\,max}_\theta Q(\theta, \theta^{\text{old}}) \quad (39)$$
However, in practice, we are not given the latent variables values.

So, instead, we focus on the expectation of the log-likelihood under the posterior distribution of the latent variables.

In the E-Step, we use the current parameter values $\theta^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(Z|X, \theta^{\text{old}})$.

This posterior is used to define the expectation of the complete-data log-likelihood, denoted $Q(\theta, \theta^{\text{old}})$, which is given by

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (38)$$

Then, in the M-step, we revise the parameters to $\theta^{\text{new}}$ by maximizing this function:

$$\theta^{\text{new}} = \arg \max_\theta Q(\theta, \theta^{\text{old}}) \quad (39)$$

Note that the log acts directly on the joint distribution $p(X, Z|\theta)$ and so the M-step maximization will likely be tractable.