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 $\mathbf{z}$ indicating the mixture component.

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 $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.

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 $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.
- For the mean $\boldsymbol{\mu}_k$, setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{21}$$

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 $\mathbf{z}$ indicating the mixture component.
- Recall the conditions that must be satisfied at a maximum of the likelihood function.
- For the mean $\boldsymbol{\mu}_k$, setting the derivatives of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}_k$ to zero yields

$$0 = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{20}$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{21}$$

- Note the natural appearance of the responsibility terms on the RHS.

- Multiplying by $\boldsymbol{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

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

- Multiplying by $\mathbf{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

  where

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

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

- Multiplying by $\boldsymbol{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

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

- 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 $\boldsymbol{\Sigma}_k^{-1}$, which we assume is non-singular, gives

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{22}$$

where

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

- 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.
- We find a similar result for the covariance matrix:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \boldsymbol{\mu}_k)(x_n - \boldsymbol{\mu}_k)^{\mathsf{T}} \; . \tag{24}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

- Maximizing it yields:

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

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

- Maximizing it yields:

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

- After multiplying both sides by π and summing over k, we get

$$\lambda = -N \tag{27}$$

- We also need to maximize $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the mixing coefficients π_k.
- Introduce a Lagrange multiplier to enforce the constraint $\sum_k \pi_k = 1$.

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

- Maximizing it yields:

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

- After multiplying both sides by π and summing over k, we get

$$\lambda = -N \tag{27}$$

- Eliminate λ and rearrange to obtain:

$$\pi_k = \frac{N_k}{N} \tag{28}$$

Solved...right?

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

Solved...right?

- 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:

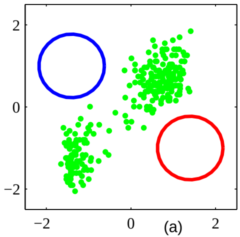
$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Solved...right?

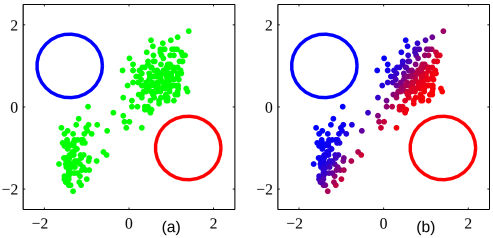
- So, we're done, right? We've computed the maximum likelihood solutions for each of the unknown parameters.
- Wrong!
- The responsibility terms depend on these parameters in an intricate way:

$$\gamma(z_k) \doteq p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

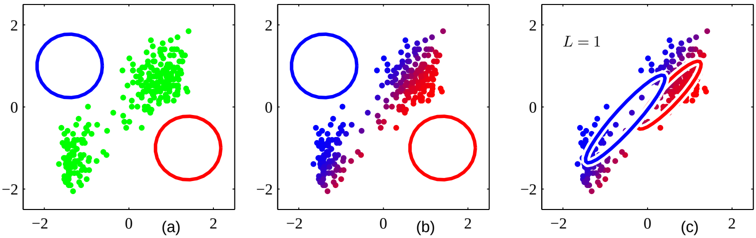
- But, these results do suggest an iterative scheme for finding a solution to the maximum likelihood problem.
  1. Chooce some initial values for the parameters, $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.
  2. Use the current parameters estimates to compute the posteriors on the latent terms, i.e., the responsibilities.
  3. Use the responsibilities to update the estimates of the parameters.
  4. Repeat 2 and 3 until convergence.



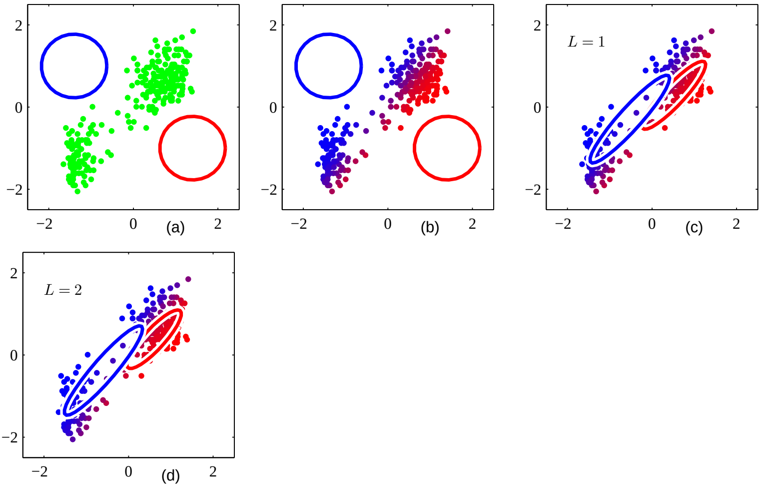
(a)



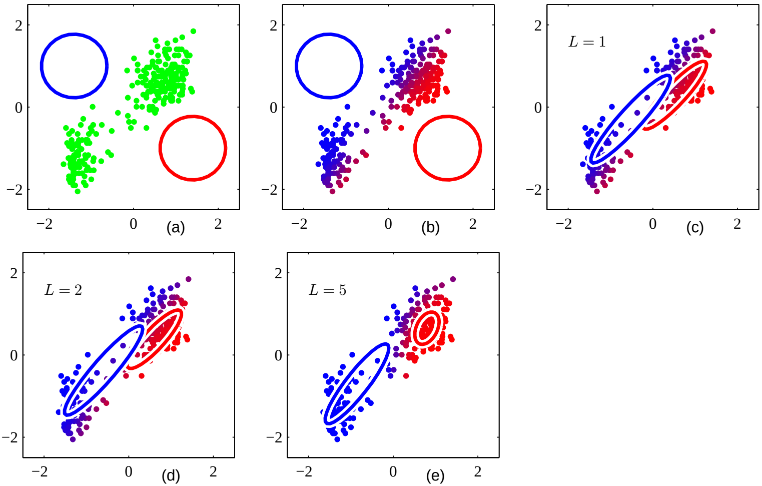
2
0
−2
−2
0
2
(a)
2
0
−2
−2
0
2
(b)



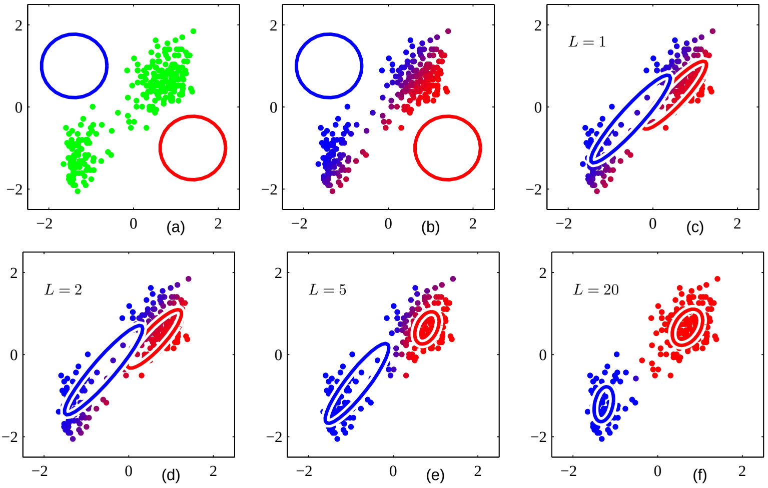
2
0
−2
−2
0
2
(a)
(b)
(c)
$L = 1$



(a)
(b)
(c)
(d)
L = 1
L = 2
−2
0
2



(a)
(b)
(c)
(d)
(e)
$L=1$
$L=2$
$L=5$



(a)
(b)
(c)
(d)
(e)
(f)
$L = 1$
$L = 2$
$L = 5$
$L = 20$

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.

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.

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 covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. 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 covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. Evaluate the initial value of the log-likelihood.
2. **E-Step** Evaluate the responsibilities using the current parameter values:

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Given a GMM, the goal is to maximize the likelihood function with respect to the parameters (the means, the covarianes, and the mixing coefficients).

1. Initialize the means, $\boldsymbol{\mu}_k$, the covariances, $\boldsymbol{\Sigma}_k$, and mixing coefficients, $\boldsymbol{\pi}_k$. Evaluate the initial value of the log-likelihood.
2. **E-Step** Evaluate the responsibilities using the current parameter values:

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

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

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{29}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^{\mathsf{T}} \tag{30}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{31}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{32}$$

4. Evaluate the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}^{\text{new}}, \boldsymbol{\Sigma}^{\text{new}}, \boldsymbol{\pi}^{\text{new}}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right) \right] \quad (33)$$

4. Evaluate the log-likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}^{\text{new}}, \boldsymbol{\Sigma}^{\text{new}}, \boldsymbol{\pi}^{\text{new}}) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}\right) \right] \quad (33)$$

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

$$\boldsymbol{\mu} = \boldsymbol{\mu}^{\text{new}} \quad (34)$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\text{new}} \quad (35)$$

$$\boldsymbol{\pi} = \boldsymbol{\pi}^{\text{new}} \quad (36)$$

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 $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

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 $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

- Note how the summation over the latent variables appears inside of the log.
  - Even if the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ belongs to the exponential family, the marginal $p(\mathbf{X}|\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$, and so the log-likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right] \tag{37}$$

- Note how the summation over the latent variables appears inside of the log.
    - Even if the joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ belongs to the exponential family, the marginal $p(\mathbf{X}|\boldsymbol{\theta})$ typically does not.
- If, for each sample $\mathbf{x}_n$ we were given the value of the latent variable $\mathbf{z}_n$, then we would have a complete data set, $\{\mathbf{X}, \mathbf{Z}\}$, 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 $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.

- 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 $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the expectation of the complete-data log-likelihood, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

- 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 $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the expectation of the complete-data log-likelihood, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

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

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

- 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 $\boldsymbol{\theta}^{\text{old}}$ to find the posterior distribution of the latent variables given by $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- This posterior is used to define the expectation of the complete-data log-likelihood, denoted $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, which is given by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{38}$$

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

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

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