3
A General View to EM

In this section, we present a general view of the EM algorithm that recognizes the key role played by latent variables. We discuss this approach first of all in an abstract setting, and then for illustration we consider once again the case of Gaussian mixtures.

The EM Algorithm: General Case

The Training Objective. The goal of the EM algorithm is to find maximum likelihood solution for models having latent variables. We denote the observed data by $\boldsymbol{X}$, and similarly we denote the latent variables by $\boldsymbol{Z}$. The set of all model parameters is denoted by $\boldsymbol{\theta}$, and so the log likelihood function is given by

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

Note that our discussion will apply equally well to continuous latent variables simply by replacing the sum over $\boldsymbol{Z}$ with an integral.

Suppose that, for the observation $\boldsymbol{X}$, we were told the corresponding value of the latent variable $\boldsymbol{Z}$. We shall call $\{\boldsymbol{X}, \boldsymbol{Z}\}$ the *complete* data set, and we shall refer to the actual observed data $\boldsymbol{X}$ as *incomplete*. *The likelihood function for the complete data set simply takes the form* $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$*, and we shall suppose that maximization of this complete-data log likelihood function is straightforward.*

Maximising the Likelihood of Incomplete Data. In practice, we are not given the complete data set $\{\boldsymbol{X}, \boldsymbol{Z}\}$, but only the incomplete data $\boldsymbol{X}$. Our state of knowledge of the values of the latent variables in $\boldsymbol{Z}$ is given only by the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$. Because we cannot use the complete-data log likelihood, we consider instead its *expected* value under the posterior distribution of the latent variable, which corresponds to the E step of the EM algorithm. In the subsequent M step, we maximize this expectation as a function of model parameters. If the current estimate for the parameters is denoted $\boldsymbol{\theta}^{\text{old}}$, then a pair of successive E and M steps gives rise to a revised estimate $\boldsymbol{\theta}^{\text{new}}$. The algorithm is initialized by choosing some starting value for the parameters $\boldsymbol{\theta}^0$. The expectation and maximisation steps are iterated until a convergence condition is met.

The EM Algorithm. More specifically, 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 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$. We then use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value $\boldsymbol{\theta}$. This expectation, denoted $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, is given by

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

In the *M step*, we determine the revised parameter estimate θnew by maximising this function

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

Note that in the definition of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, the logarithm acts directly on the joint distribution $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$, and so the corresponding M-step maximization will, by supposition, be tractable.

To summarise, the general EM algorithm is as follows:

- Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
- While the convergence is not met:
  - **E step:** Evaluate $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$
  - **M Step:** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by
    $\boldsymbol{\theta}^{\text{new}} \leftarrow \arg\max_{\boldsymbol{\theta}} \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}_{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})}$
  - $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$

The incomplete-data log likelihood is guaranteed to increase in each cycle of the EM algorithm (we don't prove it here).

The Hard-EM Algorithm. In the Hard-EM Algorithm, there is no expectation in over the latent variables in the definition of the Q function. Instead, only the most probable value for the latent variable is chosen to define the Q function. In summary, this algorithm is as follows:

- Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$
- While the convergence is not met:
  - **E step:** Set $\boldsymbol{Z}^* \leftarrow \arg\max_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$
  - **M Step:** Set $\boldsymbol{\theta}^{\text{new}} \leftarrow \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{X}, \boldsymbol{Z}^*|\boldsymbol{\theta})$
  - $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$

It can be shown that the incomplete-data log likelihood is guaranteed to increase in each cycle of the Hard-EM algorithm (we don't prove it here).

EM for GMMs: Revisited

The Complete Data Likelihood. Assume that in addition to the observed data set $\boldsymbol{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, we were also given the values of the corresponding latent variables $\boldsymbol{Z} := \{\boldsymbol{z}_1, .., \boldsymbol{z}_N\}$, then

$$\begin{aligned}\ln p(X, Z|\mu, \Sigma, \varphi) &= \ln \prod_{n=1}^{N}\prod_{k=1}^{K} \varphi^{z_{n,k}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{n,k} \ln \varphi_k + z_{n,k} \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\end{aligned}$$

where $\boldsymbol{z}_n := (z_{n1}, \ldots, z_{nk})$ is the cluster assignment vector for the nth datapoint in which $z_{nk} = 1$ if this datapoint belongs to the cluster k and zero otherwise. Note that only one element of the cluster assignment vector is 1, and the rest are zero.

Maximising the complete data likelihood to find the parameters of the model is straightforward:

- The mixing components: $\varphi_k = \dfrac{N_k}{N}$ where $N_k := \sum_{n=1}^{N} z_{nk}$

- The mean parameters: $\boldsymbol{\mu}_k = \dfrac{1}{N_k}\sum_{n=1}^{N} z_{nk}\boldsymbol{x}_n$

- The covariance matrices: $\boldsymbol{\Sigma}_k = \dfrac{1}{N_k}\sum_{n=1}^{N} z_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T$

The Q function. In practice we are not given the values for the latent variables, and we need to resort to the EM algorithm to find the best values for the parameters and the latent variables.

$$\begin{aligned}Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &:= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_{nk}|\boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}) \left[z_{n,k} \ln \varphi_k + z_{n,k} \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\right] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}) \ln \varphi_k + p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}}) \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)\end{aligned}$$

where $p(z_{nk} = 1|\boldsymbol{x}_n, \boldsymbol{\theta}^{\text{old}})$ is indeed the responsibility that we defined before:

$$\gamma(z_{nk}) := p(z_{nk}=1|\boldsymbol{x}_n,\boldsymbol{\theta}^{\text{old}}) = \frac{\varphi_k^{\text{old}}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{\text{old}},\boldsymbol{\Sigma}_k^{\text{old}})}{\sum_j \varphi_j^{\text{old}}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{\text{old}},\boldsymbol{\Sigma}_j^{\text{old}})}$$

Maximising the Q function, leads to the following updated parameters:

- The mixing components: $\varphi_k^{\text{new}} = \frac{N_k}{N}$ where $N_k := \sum_{n=1}^{N}\gamma(z_{nk})$
- The mean parameters: $$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n$$
- The covariance matrices: $$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T$$

The EM Algorithm for GMMs. The EM Algorithms thus is as follows:

- Choose an initial setting for the parameters: $\boldsymbol{\theta}^{\text{old}} = (\boldsymbol{\varphi}^{\text{old}}, \boldsymbol{\mu}_1^{\text{old}}, \dots, \boldsymbol{\mu}_K^{\text{old}}, \boldsymbol{\Sigma}_1^{\text{old}}, \dots, \boldsymbol{\Sigma}_K^{\text{old}})$
- While the convergence is not met:
  - **E step:** Set $\forall n, \forall k : \gamma(z_{nk})$ based on $\boldsymbol{\theta}^{\text{old}}$
  - **M Step:** Set $\boldsymbol{\theta}^{\text{new}}$ based on $\forall n, \forall k : \gamma(z_{nk})$
  - $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$

This is exactly the EM Algorithm that we saw in the previous chapter.