12 The EM algorithm

We will first motivate the Expectation-Maximization algorithm (EM) with an example.

Example 26 (Gaussian mixture)**.** *Suppose we want to compute the MLEs and we have data $X_1, X_2, \ldots, X_n$ from a mixture of normal distribution:*

$$f(x|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*) = \pi^* f_1(x|\mu_1, \sigma_1^2) + (1 - \pi^*) f_2(x; \mu_2 \sigma_2^2)\,.$$

We are interested in the MLEs of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*$, which involves maximizing:

$$l(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*|X) = \sum_{i=1}^{n} \log f(x_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*)\,.$$

The idea behind the EM algorithm is to treat the problem as a "missing data/value problem", if all the data was available to us, it is assumed that solving the problem would be easier.

Suppose I have the information that $x)i$ is coming from which of the two populations. Thus, suppose the complete data was of the form

$$(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n)\,,$$

where each $Z_i = k$ means that X_i is from population k. If the whole data was available to us, then first note that the joint density is

$$f(x_i, z_i = k) = f(x_i|z_i = k) \Pr(Z_i = k)\,.$$

Suppose $\mathcal{D}_1 = \{i : 1 \leq i \leq n, z_i = 1\}$ and $\mathcal{D}_2 = \{i : 1 \leq i \leq n, z_i = 2\}$, with cardinality $d_!$ and d_2 respectively.. Then the likelihood from the full data is

$$\begin{aligned}
&L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*|X) \\
&= \prod_{i \in \mathcal{D}_1} f(x_i|z_i = 1) \Pr(Z_i = 1) \prod_{j \in \mathcal{D}_2} f(x_i|z_i = 2) \Pr(Z_i = 2) \\
&= \prod_{i \in \mathcal{D}_1} \left[\pi^* f_1(x_i|\mu_1, \sigma_1^2)\right] \prod_{j \in \mathcal{D}_2} \left[f_2(x_i|\mu_2, \sigma_2^2)(1 - \pi^*)\right] \\
&= (\pi^*)^{d_1} (1 - \pi^*)^{d_2} \prod_{i \in \mathcal{D}_1} \left[f_1(x_i|\mu_1, \sigma_1^2)\right] \prod_{j \in \mathcal{D}_2} \left[f_2(x_i|\mu_2, \sigma_2^2)\right]
\end{aligned}$$

$$\Rightarrow \log L = d_1 \log(\pi^*) + d_2 \log(1-\pi^*) + \sum_{i \in \mathcal{D}_1} \log f_1(x_i|\mu_1, \sigma_1^2) + \sum_{i \in \mathcal{D}_2} \log f_2(x_i|\mu_2, \sigma_2^2)$$

Differentiating with respect to π^*, we get

$$\frac{\partial l}{\partial \pi^*} = \frac{d_1}{\pi^*} - \frac{d_2}{1-\pi^*} \stackrel{set}{=} 0\,.$$

This gives that the MLE is

$$\hat{\pi}^* = \frac{d_1}{d_1 + d_2} = \frac{d_1}{n}\,.$$

You can also see that the MLEs for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ have all been isolated so that

$$\hat{\mu}_1 = \frac{1}{d_1} \sum_{i \in \mathcal{D}_1} X_i \quad , \quad \hat{\mu}_2 = \frac{1}{d_2} \sum_{i \in \mathcal{D}_2} X_i$$

$$\sigma_1^2 = \frac{1}{d_1} \sum_{i \in \mathcal{D}_1} (X_i - \hat{\mu}_1)^2 \quad , \quad \sigma_2^2 = \frac{1}{d_2} \sum_{i \in \mathcal{D}_2} (X_i - \hat{\mu}_2)^2\,.$$

So, if the complete data was available to me, I could easily find the MLE of all the 5 parameters. Unfortunately, the Zs are not available to me, and I have only observed the Xs. Thus, my likelihood is:

$$l(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*|X) = \sum_{i=1}^{n} \log f(x_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*) = \sum_{i=1}^{n} \log \sum_{k=1,2} f(x_i, z_i = k|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*)\,.$$

The EM algorithm will solve this problem.

The EM Algorithm Suppose, in general, I have a vector of parameters θ, and I have observed the marginal data $X_1, \ldots, X_n$ from the complete data (X_i, Z_i). The objective function is to maximize is

$$l(\theta|X) = \log \int f(x, z|\theta) d\nu_z\,.$$

The EM algorithm iterates through the following: Consider a starting value θ_0. Then for any $k+1$ iteration

1. **E-Step**: Compute

$$q(\theta; \theta_{(k)}) = \mathrm{E}_{Z|y}\Big[\log f(\mathbf{x}, \mathbf{z}|\theta) \mid Y = y\Big]$$

where the expectation is computed with respect to the conditional distribution of Z given $Y = y$ for the current iterate $\theta_{(k)}$.

2. **M-Step**: Compute
$$\theta_{k+1} = \arg\max_{\theta \in \Theta} q(\theta; \theta_k) \, .$$

Proof of EM algorithm convergence. The EM algorithm works because it is a special case of the MM algorithm. The objective function is

$$f(\theta) = \log f(\mathbf{x}|\theta) \, .$$

The minorizing function is $\tilde{f}(\theta|\theta_{(k)} = q(\theta|\theta_{(k)})$+constants. Let

$$\tilde{f}(\theta|\theta_{(k)}) = \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + \log f(\mathbf{x}|\theta_{(k)}) - \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz \, .$$

Naturally, we can see that at $\theta = \theta_{(k)}$, $\tilde{f}(\theta|\theta_{(k)}) = f(\theta_{(k)}$. We will now show that minorizing property.

$$\begin{aligned}
&\tilde{f}(\theta|\theta_{(k)}) \\
&= \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + \log f(\mathbf{x}|\theta_{(k)}) - \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz \\
&= \int_z \log \left\{ \frac{f(\mathbf{x}, \mathbf{z}|\theta) \, f(x|\theta_{(k)})}{f(x, z|\theta_{(k)})} \right\} f(z|x, \theta_{(k)}) \\
&= \int_z \log \left\{ \frac{f(\mathbf{x}, \mathbf{z}|\theta) \, f(x|\theta_{(k)})}{f(x, z|\theta_{(k)})} \right\} f(z|x, \theta_{(k)}) + \log f(x|\theta) - \log f(x|\theta) \\
&= \int_z \log \left\{ \frac{f(\mathbf{x}, \mathbf{z}|\theta) \, f(x|\theta_{(k)})}{f(x, z|\theta_{(k)})} f(x|\theta) \right\} f(z|x, \theta_{(k)}) + \log f(x|\theta)
\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
&\leq \log \int_z \left\{ \frac{f(\mathbf{x}, \mathbf{z}|\theta) \, f(x|\theta_{(k)})}{f(x, z|\theta_{(k)})} f(x|\theta) \right\} f(z|x, \theta_{(k)}) + \log f(x|\theta) \\
&= \log \int_z \frac{f(z|x, \theta)}{f(z|x, \theta_{(k)})} f(z|x, \theta_{(k)}) + \log f(x|\theta) \\
&= \log \int_z f(z|x, \theta) dz + \log f(x|\theta) \\
&= \log f(x|\theta) \, .
\end{aligned}$$

Thus, $\tilde{f}(\theta|\theta_{(k)})$ is a minorizing function, and the next iterate is

$$\theta_{(k+1)} = \arg\max_{\theta} \tilde{f}(\theta|\theta_{(k)}) = q(\theta|\theta_{(k)})$$

□

Back to the example: To implement the EM algorithm, we first need to find $q(\theta|\theta_{(k)})$. First note that the conditional distribution of $Z|X$ is

$$\Pr(Z = c|X = x) = \frac{f_c(x|\mu_c, \sigma_c^2)\pi_c}{\sum_{j=1,2} f_j(x|\mu_j, \sigma_j^2)\pi_j}\,.$$

So for any kth iterate with current step $\theta_{(k)} = (\mu_{1,k}, \mu_{2,k}, \sigma_{1,k}^2, \sigma_{2,k}^2, \pi_k^*)$, we have

$$\Pr(Z = c|X = x, \theta_{(k)}) = \frac{f_c(x|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}}\,.$$

The above can be calculate explicitly for any data points x. So

$$\begin{aligned}
q(\theta|\theta_{(k)}) &= \mathrm{E}_{Z|x}\left[\log f(x, z|\theta) \mid X = x, \theta_{(k)}\right] \\
&= \mathrm{E}_{Z|x}\left[\sum_{i=1}^{n} \log f(x_i, z_i|\theta) \mid X = x, \theta_{(k)}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}_{Z_i|x_i}\left[\log f(x_i, z_i|\theta) \mid X = x_i, \theta_{(k)}\right] \\
&= \sum_{i=1}^{n}\sum_{c=1}^{2} \log\left\{f_c(x_i|\mu_c, \sigma_c^2)\pi_c\right\} \frac{f_c(x_i|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x_i|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}}\,.
\end{aligned}$$

This is the E-step. To implement the E-step we need to update

$$\gamma_{i,c,k} = \frac{f_c(x_i|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x_i|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}}\,.$$

This completes the E-step. We move on to the M-step. To complete the M-step

$$\theta_{(k+1)} = \arg\max q(\theta|\theta_{(k)})\,.$$

$$L(\theta) = \sum_{i=1}^{n}\sum_{c=1}^{2} \left\{\log f_c(x_i|\mu_c, \sigma_c^2) + \log \pi_c\right\} \gamma_{i,c,k}$$

$$= \sum_{i=1}^{n}\sum_{c=1}^{2}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_c^2 - \frac{(X_i-\mu_c)^2}{2\sigma_c^2} + \log\pi_c\right]\gamma_{i,c,k}$$

$$= \text{const} - \frac{1}{2}\sum_{i=1}^{n}\sum_{c=1}^{2}\log\sigma_c^2\gamma_{i,c,k} - \sum_{i=1}^{n}\sum_{c=1}^{2}\frac{(X_i-\mu_c)^2}{2\sigma_c^2}\gamma_{i,c,k} + \sum_{i=1}^{n}\sum_{c=1}^{2}\log\pi_c\,\gamma_{i,c,k}\,.$$

Taking derivatives and setting to 0, we get For any c,

$$\frac{\partial L}{\partial \mu_c} = \sum_{i=1}^{n}\frac{(x_i-\mu_c)\,\gamma_{i,c,k}}{\sigma_c^2} \overset{set}{=} 0 \;\Rightarrow\; \mu_{c,(k+1)} = \frac{\sum_{i=1}^{n}\gamma_{i,c,k}\;x_i}{\sum_{i=1}^{n}\gamma_{i,c,k}} \tag{1}$$

$$\frac{\partial L}{\partial \sigma_c^2} = -\frac{1}{2}\sum_{i=1}^{n}\frac{\gamma_{i,c,k}}{\sigma_c^2} + \sum_{i=1}^{n}\frac{(x_i-\mu_c)^2}{2\sigma_c^4}\gamma_{i,c,k} \overset{set}{=} 0 \;\Rightarrow\; \sigma^2_{c,(k+1)} = \frac{\sum_{i=1}^{n}\gamma_{i,c,k}(x_i-\mu^2_{c,(k+1)})}{\sum_{i=1}^{n}\gamma_{i,c,k}} \tag{2}$$

For π_c note that the optimization requires a constraint, since $\sum_c \pi_c = 1$. So we will use Lagrange multipliers. The objective function is

$$\begin{aligned}
&L(\theta) - \lambda\left(\sum_{c=1}^{c}\pi_c - 1\right)\\
\Rightarrow &\frac{\partial L}{\partial \pi_c} = \sum_{i=1}^{n}\frac{\gamma_{i,c,k}}{\pi_c} - \lambda \overset{set}{=} 0\\
\Rightarrow &\pi_c = \sum_{i=1}^{n}\frac{\gamma_{i,c,k}}{\lambda}\\
\Rightarrow &\sum_c \pi_c = \sum_c\sum_{i=1}^{n}\frac{\gamma_{i,c,k}}{\lambda}\\
\Rightarrow &1 = \frac{1}{\lambda}\sum_{i=1}^{n}\\
\Rightarrow &\lambda = n\\
\Rightarrow &\pi_{c,(k+1)} = \pi_{c,(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_{i,c,k}\,.
\end{aligned} \tag{3}$$

Thus equations (1) and (3) provide the iterative updates for the parameters.

References