Stat 111 Spring 2008 EM Algorithm

Suppose we observe data $y$ from a distribution $f_{y|\phi}(y | \phi)$ and we wish to estimate the unknown parameter $\phi$ (possibly a vector or matrix). The EM algorithm facilitates finding the posterior mode, or the value $\hat{\phi}$ which maximizes the posterior density $f_{\phi|y}(\phi | y)$. Note that this is equivalent to finding the maximum likelihood estimate for $\phi$ if we assume a prior density $p(\phi) \propto c$.

The EM algorithm works by introducing a vector of “missing” observations $\theta$, for which there are nice distributional forms for $\phi | \theta$ and for $\theta, \phi | y$. The key result for the EM algorithm is derived from the following equivalence:

$$\log f_{\theta,\phi|y}(\theta, \phi | y) = \log f_{\theta|y,\phi}(\theta | y, \phi) + \log f_{\phi|y}(\phi | y).$$

Each step of the algorithm updates a current estimate $\phi$ to a new value $\phi^{(t+1)}$ which increases the posterior density. That is, $f_{\phi|y}(\phi^{(t+1)} | y) \geq f_{\phi|y}(\phi^{(t)} | y)$, with equality only at a local mode. The new estimate requires taking an expectation with respect to $\theta$ given $y$ and the current estimate $\phi^{(t)}$. Applying this action to both sides of (1) yields:

$$\log f_{\phi|y}(\phi | y) = \int \left( \log f_{\theta,\phi|y}(\theta, \phi | y) - \log f_{\theta|y,\phi}(\theta | y, \phi) \right) f_{\theta|y,\phi}(\theta | \phi^{(t)}, y) d\theta. \quad (2)$$

$$= \int \log f(\theta, \phi | y) f(\theta | \phi^{(t)}, y) d\theta - \int \log f(\theta | y) f(\theta | \phi^{(t)}, y) d\theta$$

$$= Q(\phi | \phi^{(t)}) - R(\phi | \phi^{(t)}).$$

The left side remains $\log f_{\phi|y}(\phi | y)$, because this function does not depend on $\theta$.

The key result is as follows:

$$R(\phi | \phi^*) = \int \log f(\theta | y, \phi) f(\theta | \phi^*, y) d\theta < R(\phi^* | \phi^*), \quad \forall \phi \neq \phi^*. \quad (3)$$

Proof:

$$R'(\phi | \phi^*) = \int \left( \frac{\partial}{\partial \phi} \log f_{\theta|y,\phi}(\theta | \phi, y) \right) f_{\theta|y,\phi}(\theta | \phi^*, y) d\theta$$

$$= \int \frac{\partial}{\partial \phi} f_{\theta|y,\phi}(\theta | \phi, y) f_{\theta|y,\phi}(\theta | \phi^*, y) d\theta$$

$$R'(\phi^* | \phi^*) = \left[ \frac{\partial}{\partial \phi} \int f_{\theta|y,\phi}(\theta | \phi, y) d\theta \right]_{\phi=\phi^*} = \left[ \frac{\partial}{\partial \phi} \right]_{\phi=\phi^*} = 0.$$

This shows that $\phi^*$ is a local minimum or maximum of $R(\phi | \phi^*)$. It is not hard to show that it is a maximum.
**Implementation:**

The function $Q(\phi | \phi^*)$ is described as the expected value of the complete-data log-posterior density $f_{\theta, \phi|y}(\theta, \phi | y)$, taking expectation with respect to the density $f_{\theta|\phi,y}(\theta | \phi^*, y)$ and treating the result as a function of $\phi$.

Given that $R(\phi | \phi^{(t)}) \leq R(\phi^{(t)} | \phi^{(t)})$, it follows that the posterior density at $\phi^{(t+1)}$, for any value $\phi^{(t+1)}$ such that $Q(\phi^{(t+1)} | \phi^{(t)}) > Q(\phi^{(t)} | \phi^{(t)})$, will be larger than $f_{\phi|y}(\phi^{(t)} | y)$. This is true because

$$\log f_{\phi|y}(\phi^{(t)} | y) = Q(\phi^{(t)} | \phi^{(t)}) - R(\phi^{(t)} | \phi^{(t)}).$$

Changing $\phi^{(t)}$ to $\phi^{(t+1)}$ increases $Q$ and can only decrease $R$, so the difference increases.

The EM algorithm takes its name from the breakdown into two steps:

**E-step:** Compute the function $Q(\phi | \phi^{(t)}) = E(\log(f_{\theta,\phi|y}(\theta, \phi | y) | y, \phi^{(t)}))$.

**M-step:** Update $\phi^{(t)}$ to the value $\phi^{(t+1)}$ which maximizes the function $Q(\phi | \phi^{(t)})$.

In practice, it is only necessary to compute the expected sufficient statistics of $\log(f_{\theta,\phi|y}(\theta, \phi | y))$ (i.e., the functions of $\theta$ that involve $\phi$, and will therefore be needed for the M-step).