EM Algorithm

The EM algorithm is an iterative method for finding the mode of $p(\phi|y)$ when it is hard to maximize $p(\phi|y)$ directly but it is easy to work with $p(\gamma|\phi, y)$ and $p(\phi|\gamma, y)$.

For example, let ϕ be the model parameters and γ be missing data. In this case we can (1) replace missing values by their expectations given the guessed parameters, (2) estimate the parameters assuming that the missing data are given by their estimated values, (3) reestimate the missing values assuming the new parameter values are correct, (4) reestimate parameters... until convergence.

EM Algorithm

In general the EM algorithm is used to obtain the maximum of a marginal density. The "missing data" correspond usually to nuisance parameters or latents variables that are introduced in order to simplify calculations.

EM steps: finding the *expectation* of the missing data, and *maximizing* the functions to estimate the parameters assuming the missing data were observed.

Implementation

1. Start with a crude parameter estimate ϕ^0
2. For $t = 1, 2, \ldots$
   (a) E-step: find

$$E_{old}(\log(p(\gamma, \phi|y)) = \int \log(p(\gamma, \phi|y))p(\gamma|\phi^{old}, y)d\gamma,$$

   where the expectation averages over the conditional posterior distribution of γ, given the current estimate, $\phi^{old} = \phi^{t-1}$

   (b) M-step: Let ϕ^t be the value of ϕ that maximizes $E_{old}(\log p(\gamma, \phi|y))$. For the GEM algorithm, we only need to find a value of ϕ such that $E_{old}(\log p(\gamma, \phi|y))$ is increased.

Note that

$$\log p(\phi|y) = \log p(\gamma, \phi|y) - \log p(\gamma|\phi, y)$$

The expression

$$E_{old}(\log p(\gamma, \phi|y)) = \int (\log p(\gamma, \phi|y)) p(\gamma|\phi^{old}, y) d\gamma$$

is usually called $Q(\phi|\phi^{old})$ or $Q(\phi, \phi^{old})$ in the EM literature.

By construction, $p(\phi|y)$ increases in each step of the EM algorithm. Why? (hmk).

The EM algorithm converges to a local mode of the posterior density (except in some very special cases).

Finding multiple modes. Start the iterations at many points throughout the parameter space.

Debugging. Compute $p(\phi^t|y)$ at each iteration and check that it increases monotonically.

Rate of convergence. Depends on the proportion of "information" about ϕ in $p(\gamma, \phi|y)$ that is missing from $p(\phi|y)$. We can have slow convergence if the proportion of missing information is high.

Example 1

$y_1, \ldots, y_n \sim N(\mu, \sigma^2)$, consider the prior $p(\mu) = N(\mu_0, \tau_0^2)$ and $p(\log \sigma) \propto 1$. Then

$$\log p(\mu, \sigma|y) = -\frac{1}{2\tau_0^2}(\mu - \mu_0)^2 - (n+1)\log\sigma - \frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2 + k$$

So, for the E-step we need to calculate $E_{old}(\log\sigma)$ and $E_{old}(1/\sigma^2)$. The first is actually irrelevant for the M-step. The second is obtained by noting that

$$p(\sigma^2|\mu, y) \propto IG(\sigma^2|n/2, 1/2\sum_i (y_i - \mu)^2)$$

Thus $1/\sigma^2 \sim G(n/2, 1/2\sum_i (y_i - \mu)^2)$ and its expectation is $n/\sum_i (y_i - \mu)^2)$.

The expression for $Q(\mu|\mu^{old})$ is

$$-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2-\frac{1}{2}\frac{n}{\sum_i(y_i-\mu^{old})^2}\sum_i(y_i-\mu)^2+k$$

This has the form of a normal likelihood with a normal prior. So this is proportional to a normal posterior. Thus the maximum is obtained at the posterior mode, which is given by

$$\mu^{new}=\frac{\frac{\mu_0}{\tau_0^2}+\frac{n\overline{y}}{\frac{1}{n}\sum_i(y_i-\mu^{old})^2}}{\frac{1}{\tau_0^2}+\frac{n}{\frac{1}{n}\sum_i(y_i-\mu^{old})^2}}$$

Example 2

Example (Tanner, 1993): Genetic Linkage Model. Suppose 197 animals ($\boldsymbol{y}$) are distributed into four categories

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

with cell probabilities

$$\left(\frac{1}{2} + \frac{\phi}{4}, \frac{1}{4}(1-\phi), \frac{1}{4}(1-\phi), \frac{\phi}{4}\right).$$

Augment the observed data by splitting the first cell into two cells with probabilities $1/2$ and $\phi/4$. The augmented data $\gamma = \boldsymbol{x}$ are given by $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ such that $x_1 + x_2 = 125$, and $x_i = y_{i-1}$ for $i = 3, 4, 5$. In this case we have $\theta = (\gamma, \phi) = (\boldsymbol{x}, \phi)$.

Under a flat prior we have

$$p(\phi|\boldsymbol{y}) \propto (2+\phi)^{y_1}(1-\phi)^{y_2+y_3}\phi^{y_4}$$

while the augmented posterior is given by

$$p(\phi|\boldsymbol{x}) \propto \phi^{x_2+x_5}(1-\phi)^{x_3+x_4}$$

Then

$$Q(\phi|\phi^{old}) = E[(x_2+x_5)\log(\phi) + (x_3+x_4)\log(1-\phi)|\phi^{old}, \boldsymbol{y})],$$

where $p(x_2|\boldsymbol{y}, \phi)$ is the binomial distribution with $n = 125$ and $p = \phi/(\phi+2)$. Then

$$Q(\phi|\phi^{old}) = (E(x_2|\phi^{old}, \boldsymbol{y}) + x_5)\log(\phi) + (x_3+x_4)\log(1-\phi).$$

For the M-step we have that

$$\phi^t = \frac{E(x_2|\phi^{t-1}, \boldsymbol{y}) + x_5}{E(x_2|\phi^{t-1}, \boldsymbol{y}) + x_3 + x_4 + x_5} \quad E(x_2|\phi^{t-1}, \boldsymbol{y}) = 125\frac{\phi^{t-1}}{2+\phi^{t-1}}$$

Starting at $\phi^0 = 0.5$ the algorithm converges to $\phi^* = 0.6268$, after 4 iterations.

Example 3

Consider a normal hierarchical model with J classes. We can obtain the joint posterior mode of $\theta_j, j = 1, \ldots, J, \mu, \sigma$ and τ using iterative conditional modes. This consists of maximizing the full conditional of each parameter and then cycle.

If we focus on μ, σ and τ we set $\gamma = (\theta_1, \ldots, \theta_J) = \theta$ and $\phi = (\mu, \log\sigma, \log\tau)$. Then,

$$\begin{aligned} p(\mu, \log\sigma, \log\mu|y) &= \frac{p(\theta, \mu, \log\sigma, \log\tau|y)}{p(\theta|\mu, \log\sigma, \log\tau, y)} \\ &\propto \frac{\tau \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_{ij}|\theta_j, \sigma^2)}{\prod_{j=1}^{J} N(\theta_j|\hat{\theta}_j, V_{\theta_j})} \end{aligned}$$

which holds for every θ. Setting $\theta = \hat{\theta}$ we have,

$$p(\mu, \log\sigma, \log\mu|y) \propto \tau \prod_{j=1}^{J} N(\hat{\theta}_j|\mu, \tau^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} N(y_{ij}|\hat{\theta}_j, \sigma^2) \prod_{j=1}^{J} V_{\theta_j}^{1/2}$$

Notice that $\hat{\theta}_j$ and V_{θ_j} are the mean and variances of the full conditionals of θ_j and are functions of μ, σ and τ.

$$\begin{aligned}\log p(\theta, \mu, \log\sigma, \log\tau|y) &= -n\log\sigma - (J-1)\log\tau - \\ &\quad \frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j - \mu)^2 \\ &\quad -\frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2 + const\end{aligned}$$

- *E-step*

$$\begin{aligned}E_{old}((\theta_j - \mu)^2) &= E((\theta_j - \mu)^2|\mu^{old}, \sigma^{old}, \tau^{old}, y) \\ &= [E_{old}(\theta_j - \mu)]^2 + var_{old}(\theta_j) \\ &= (\hat{\theta}_j - \mu)^2 + V_{\theta_j} \\ E_{old}((y_{ij} - \theta_j)^2) &= (y_{ij} - \hat{\theta}_j)^2 + V_{\theta_j}\end{aligned}$$

- *M-step.*

$$\begin{aligned}
\mu^{new} &= \frac{1}{J}\sum_{j=1}^{J}\hat{\theta}_j \\
\sigma^{new} &= \left(\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}[(y_{ij}-\hat{\theta}_j)^2 + V_{\theta_j}]\right)^{1/2} \\
\tau^{new} &= \left(\frac{1}{J-1}\sum_{j=1}^{J}[(\hat{\theta}_j-\mu_{new})^2 + V_{\theta_j}]\right)^{1/2}
\end{aligned}$$