Extensions to EM

$y = (y_o, y_m) \sim f(y|\theta)$; we observe y_o but not y_m.

Complete data log likelihood: $l_c(\theta|y) = log f(y|\theta)$

Observed data log lik: $l(\theta|y_o) = \log \int f(y_o, y_m|\theta)\, \mathrm{d}y_m$

EM algorithm:

E step: $l^{(s)}(\theta) = \mathrm{E}\{l_c(\theta|y) \mid y_o, \hat{\theta}^{(s)}\}$

M step: $\hat{\theta}^{(s+1)} = \arg\max_\theta l^{(s)}(\theta)$

$l(\hat{\theta}^{(s)}|y_o)$ is non-decreasing.

Note: If you can calculate $l(\hat{\theta}^{(s)}|y_o)$, it is a good idea to monitor it for debugging purposes.

Problems:

1. Standard errors
2. Maximizing $l^{(s)}(\theta)$
3. Slow convergence

Good book: McLachlan and Krishnan (1997) The EM algorithm and extensions. Wiley.

SEM algorithm

Meng & Rubin (1991) JASA 96:899–909

EM defines a mapping, M: $\hat{\theta}^{(s+1)} = M[\hat{\theta}^{(s)}]$.

$$
\begin{aligned}
I_o(\theta|y_o) &= \frac{-\partial^2 l(\theta|y_o)}{\partial\theta \cdot \partial\theta} \\
I_o(\theta|y) &= \frac{-\partial^2 l_c(\theta|y)}{\partial\theta \cdot \partial\theta} \\
I_{oc}(\theta|y) &= \mathsf{E}\{I_o(\theta|y)|y_o, \hat{\theta}\} \\
\hat{V} &= \{I_o(\hat{\theta}|y_o)\}^{-1} \\
&= I_{oc}^{-1}\{I + DM(I - DM)^{-1}\} \\
(DM)_{ij} &= \frac{\partial M_j(\hat{\theta})}{\partial\theta_i} \approx \frac{\hat{\theta}_j^{(s+1)} - \hat{\theta}_j}{\hat{\theta}_i^{(s)} - \hat{\theta}_i}
\end{aligned}
$$

SEM algorithm (continued)

1. Run EM to get the MLE $\hat{\theta}$.
2. Pick a starting point, $\hat{\theta}^{(0)}$, some small distance from $\hat{\theta}$.
3. Repeat the following until convergence.
   (a) Calculate $\hat{\theta}^{(s)} = M[\hat{\theta}^{(s-1)}]$ using one step of EM.
   (b) For each $i = 1, \ldots, p$,
      i. Let $\hat{\theta}^{(s)}(i) = (\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \hat{\theta}_i^{(s)}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_p)$ (Replace the ith element of $\hat{\theta}$ with the ith element of $\hat{\theta}^{(s)}$.)
      ii. Perform one step of EM on $\hat{\theta}^{(s)}(i)$, to obtain $M[\hat{\theta}^{(s)}(i)]$.
      iii. Let $r_{ij}^{(s)} = \{M_j[\hat{\theta}^{(s)}(i)] - \hat{\theta}_j\}/\{\hat{\theta}_i^{(s)} - \hat{\theta}_i\}$

Notes:

1. The MLE $\hat{\theta}$ should be obtained at very low tolerance, (for example, $\epsilon = 10^{-12}$).
2. Allow the r_{ij} to converge at different rates. The element r_{ij} is taken to be the first value of $r_{ij}^{(s)}$ satisfying $|r_{ij}^{(s)} - r_{ij}^{(s-1)}| < \epsilon$, where (for example) $\epsilon = 10^{-6}$, and where the s at which we stop is allowed to depend on (i, j).

Example

Normal-Poisson model

$k_{ij} \sim \text{Poisson}(\lambda_i)$
$x_{ij}|k_{ij} \sim \text{N}(a + bk_{ij}, \sigma^2)$
(k_{ij}, x_{ij}) mutually independent
$i = 1, \ldots, G$; $j = 1, \ldots, n_i$

Complete-data log likelihood:

$$l_c(\theta|x,k) = -\sum_i n_i\lambda_i + \sum_i \log\lambda_i \sum_j k_{ij} \\ -(n/2)\log\sigma^2 - (1/2)\sum_{ij}(x_{ij} - a - bk_{ij})^2/\sigma^2$$

$$\begin{aligned}
\frac{\partial l_c}{\partial \lambda_i} &= -n_i + \sum_j k_{ij}/\lambda_i \\
\frac{\partial l_c}{\partial a} &= \sum_{ij}(x_{ij} - a - bk_{ij})/\sigma^2 \\
\frac{\partial l_c}{\partial b} &= \sum_{ij} k_{ij}(x_{ij} - a - bk_{ij})/\sigma^2 \\
\frac{\partial l_c}{\partial \sigma^2} &= -n/(2\sigma^2) + \sum_{ij}(x_{ij} - a - bk_{ij})^2/(2\sigma^4)
\end{aligned}$$

Example (continued)

$$
\begin{aligned}
-\frac{\partial^2 l_c}{\partial \lambda_i^2} &= \frac{\sum_j k_{ij}}{\lambda_i^2} \\
-\frac{\partial^2 l_c}{\partial a^2} &= 1/\sigma^2 \\
-\frac{\partial^2 l_c}{\partial a \partial b} &= \sum_{ij} k_{ij}/\sigma^2 \\
-\frac{\partial^2 l_c}{\partial a \partial \sigma^2} &= \sum_{ij} (x_{ij} - a - b k_{ij})/\sigma^4 \\
-\frac{\partial^2 l_c}{\partial b^2} &= \sum_{ij} k_{ij}^2/\sigma^2 \\
-\frac{\partial^2 l_c}{\partial b \partial \sigma^2} &= \sum_{ij} k_{ij}(x_{ij} - a - b k_{ij})/\sigma^4 \\
-\frac{\partial^2 l_c}{(\partial \sigma^2)^2} &= \sum_{ij} (x_{ij} - a - b k_{ij})^2/\sigma^6 - n/(2\sigma^4)
\end{aligned}
$$

Bayesian EM

Suppose we have prior $\pi(\theta)$ and wish to find the mode of the log posterior $= l(\theta) + \log \pi(\theta)$.

$$\begin{aligned} Q^{(s)}(\theta) &= \mathsf{E}\{l_c(\theta|y) + \log \pi(\theta) \mid y_o, \hat{\theta}^{(s)}\} \\ &= \mathsf{E}\{l_c(\theta|y) \mid y_o, \hat{\theta}^{(s)}\} + \log \pi(\theta) \\ &= l^{(s)}(\theta) + \log \pi(\theta) \end{aligned}$$

E step: Calculate $l^{(s)}(\theta) = \mathsf{E}\{l_c(\theta|y) \mid y_o \hat{\theta}^{(s)}\}$

M step: Find $\hat{\theta}^{(s+1)} = \arg\max_\theta Q^{(s)}(\theta)$

Note: $\log \mathsf{Pr}(\hat{\theta}^{(s)}|y_o)$ is non-decreasing.

Example

Consider $(y_1, y_2, y_3) \sim \text{MN}[n, ((1+2\theta)/2, (1-\theta)/2, \theta/4)]$.

Complete data:
$(y_{11}, y_{12}, y_2, y_3) \sim \text{MN}[n, (1/2, \theta/4, (1-\theta)/2, \theta/4)]$,
where $y_{11} + y_{12} = y_1$.

$l_c(\theta) = (y_{12} + y_3) \log \theta + y_2 \log(1-\theta)$

$w_{12}^{(s)} = \text{E}(y_{12} | \hat{\theta}^{(s)}, y_1) = \hat{\theta}^{(s)} y_1 / (\hat{\theta}^{(s)} + 2)$

$\hat{\theta}^{(s+1)} = (w_{12}^{(s)} + y_3) / (w_{12}^{(s)} + y_2 + y_3)$

Consider a Beta(ν_1, ν_2) prior on θ:

$$\pi(\theta) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \theta^{\nu_1 - 1} \theta \nu_2 - 1$$

$Q^{(s)}(\theta) = (w_{12}^{(s)} + y_3 + \nu_1 - 1) \log \theta + (y_2 + \nu_2 - 1) \log(1-\theta)$

$w_{12}^{(s)}$ is as before

$\hat{\theta}^{(s+1)} = (w_{12}^{(s)} + y_3 + \nu_1 - 1) / (w_{12}^{(s)} + y_2 + y_3 + \nu_1 + \nu_2 - 2)$

EM gradient algorithm

Generalized EM:

E step: $l^{(s)}(\theta)$ as before

M step: Choose $\hat{\theta}^{(s+1)}$ such that $l^{(s)}(\hat{\theta}^{(s+1)}) \geq l^{(s)}(\hat{\theta}^{(s)})$ (don't necesssarily maximize $l^{(s)}$, just increase it.)

This retains the ascent property of EM.

EM gradient algorithm:
In the M step, do one step of Newton-Raphson:

$$\hat{\theta}^{(s+1)} = \hat{\theta}^{(s)} + \alpha^{(s)}\delta^{(s)}$$

$$\delta^{(s)} = -\left(\frac{\partial l^{(s)}(\hat{\theta}^{(s)})}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial l^{(s)}(\hat{\theta}^{(s)})}{\partial\theta}\right)$$

$\alpha^{(s)} = 1$; do step-halving to ensure $l^{(s)}(\hat{\theta}^{(s+1)}) \geq l^{(s)}(\hat{\theta}^{(s)})$

Example

Let $x_1, \ldots, x_p \sim$ indep gamma$(\theta_1, 1)$,
so $f(x_i) = x_i^{\theta_i - 1} e^{-x_i} / \Gamma(\theta_i)$.

Let $y_1, \ldots y_p$ be defined by $y_i = x_i / \sum x_i$, with $y_i \geq 0$; $\sum y_i = 1$.

Then $y \sim$ Dirichlet$(\theta_1, \ldots, \theta_p)$, with density

$$f(y|\theta) = \frac{\Gamma(\sum \theta_i)}{\prod \Gamma(\theta_i)} \prod y_i^{\theta_i - 1}$$

Suppose we observe $y_1, \ldots y_n \sim$ iid Dirichlet(θ), and wish to obtain the MLE of θ by the EM algorithm, with the corresponding x's serving as the complete data.

Complete data likelihood:
$l_c(\theta) = \sum_j (\theta_j - 1) \sum_i \log x_{ij} - n \sum_j \log \Gamma(\theta_j)$

$w_{ij}^{(s)} = \mathsf{E}(\log x_{ij} \mid y_{ij}, \hat{\theta}_j^{(s)})$

$l^{(s)}(\theta) = \sum_j (\theta_j - 1) \sum_i w_{ij}^{(s)} - n \sum_j \log \Gamma(\theta_j)$

$\partial l^{(s)} / \partial \theta_j = \sum_i w_{ij}^{(s)} - n\psi(\theta_j)$, where $\psi(\theta) = \Gamma'(\theta)/\Gamma(\theta)$.

$\partial^2 l^{(s)} / \partial \theta_j^2 = -n\psi'(\theta)$

ECM algorithm

Meng & Rubin (1993) Biometrika 80:267–278

Replace the M step with a series of conditional maximization (CM) steps.

$\hat{\theta}^{(s)} =$ current iterate

E step: as before $\rightarrow l^{(s)}(\theta)$

CM steps: for $t = 1, \ldots, T$
$\hat{\theta}^{(s+t/T)} = \arg\max_\theta l^{(s)}(\theta)$
such that $g_t(\theta) = g_t(\hat{\theta}^{(s+(t-1)/T)})$

Important special case:
Partition θ into T bits: $\theta = (\theta_1, \ldots, \theta_T)$.
Let $g_t(\theta) = \theta_t$.
$\rightarrow$ maximize w.r.t the bits of θ, one at a time.

Multicycle ECM: Perform additional E steps

ECME algorithm: Replace some CM steps with a full M step

Example

Complete data:

$y_1, \ldots, y_n \sim$ iid gamma(α, β), with density

$$f(y|\alpha, \beta) = \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

$y_o =$ censoring of the complete data.

$$l_c(\alpha, \beta|y) = (\alpha - 1)\sum \log y_i - \sum y_i/\beta \\ -n\{\alpha \log \beta + \log \Gamma(\alpha)\}$$

Given α, $\hat{\beta} = \bar{y}/\alpha$.

Given β, $\hat{\alpha} = \psi^{-1}(\bar{g} - \log \beta)$,

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ and $\bar{g} = \sum \log y_i/n$

E step:

$w^{(s)} = \mathsf{E}(\bar{y}|y_o, \hat{\alpha}^{(s)}, \hat{\beta}^{(s)})$

$\tau^{(s)} = \mathsf{E}(\bar{g}|y_o, \hat{\alpha}^{(s)}, \hat{\beta}^{(s)})$

CM steps:

$\hat{\beta}^{(s+1)} = w^{(s)}/\hat{\alpha}^{(s)}$

$\hat{\alpha}^{(s+1)} = \psi^{-1}[\tau^{(s)} - \log \hat{\beta}^{(s+1)}]$

Accelerated EM

Aitken's acceleration method

[Louis (1982) JRSSB 44:226-233]

Suppose $\hat{\theta}^{(s)} \to \hat{\theta}$ as $s \to \infty$

Then $\hat{\theta} = \hat{\theta}^{(s)} + \sum_{h=1}^{\infty} [\hat{\theta}^{(s+h)} - \hat{\theta}^{(s+h-1)}]$

Now

$$\begin{aligned} \hat{\theta}^{(s+h)} - \hat{\theta}^{(s+h-1)} &= M[\hat{\theta}^{(s+h-1)}] - M[\hat{\theta}^{(s+h-2)}] \\ &\approx J(\hat{\theta}^{(s+h-2)}) \, [\hat{\theta}^{(s+h-1)} - \hat{\theta}^{(s+h-2)}] \\ &\approx J(\hat{\theta}^{(s)}) \, [\hat{\theta}^{(s+h-1)} - \hat{\theta}^{(s+h-2)}] \end{aligned}$$

Thus

$$\begin{aligned} \hat{\theta} &\approx \hat{\theta}^{(s)} + \sum_{h=0}^{\infty} \{J(\hat{\theta}^{(s)})\}^h [\hat{\theta}^{(s+1)} - \hat{\theta}^{(s)}] \\ &= \hat{\theta}^{(s)} + \{I - J(\hat{\theta}^{(s)})\}^{-1} \, [\hat{\theta}^{(s+1)} - \hat{\theta}^{(s)}] \end{aligned}$$

The algorithm:

1. From $\hat{\theta}^{(s)}$ produce $\hat{\theta}^{(s+1)}$ using EM
2. Estimate $J(\hat{\theta}^{(s)})$ by $\hat{J}$ (see below)
3. Compute $\hat{\theta}_{\star}^{(s+1)} = \hat{\theta}^{(s)} + (I - \hat{J})^{-1} \, (\hat{\theta}^{(s+1)} - \hat{\theta}^{(s)})$
4. Use $\hat{\theta}_{\star}^{(s+1)}$ in step 1.

Producing $\hat{J}$

From Louis (1982), we have $(I - \hat{J})^{-1} = I_c I_o^{-1}$
where I_c is the complete data information matrix and I_o is the observed data information matrix.

Also from Louis (1982), we have

$$I_o \approx \mathsf{E}\{\partial^2 l_c/(\partial\theta\partial\theta')\} - \mathsf{E}\{(\partial l_c/\partial\theta)(\partial l_c/\partial\theta)'\}$$