Some Refined Versions of the EM Algorithm

Inchi Hu
Department of ISMT, HKUST

Visiting

Department of Mathematics
National Taiwan University

and

Institute of Statistical Science
Academia Sinica

1. Previous methods to speed up the EM algorithm:

- Aitkin Acceleration (Dempster et al, 1976, Louis, 1982)
- Combining with Newton-Raphson type algorithms (Lange, 1995)
- Combining with conjugate gradient methods (Jamshidian & Jennrich, 1993)

These algorithms increase the speed of EM algorithms at the expanse of a larger human investment for general user: more numerically complex implementation and more careful monitoring.

Because the rate of convergence is determined by the fraction of missing information, it is known that by augmenting less one can develop a faster algorithm but the resulting E-step and/or M-step may be more difficult to implement.

We will study how to search for an efficient augmentation scheme. By "efficient" we mean less augmentation while maintaining the simplicity and stability of the EM algorithm. The key idea is to introduce a working parameter to index a class of possible data augmentation schemes.

2.1 Example: t-model with known degrees of freedom

$$f_\nu(x\,|\mu,\Sigma) \propto |\Sigma|^{1/2}\{\nu + (x-\mu)^T\Sigma^{-1}(x-\mu)\}^{-(\nu+p)/2}$$

Fit $Y_{\text{obs}} = (y_1, \cdots, y_n)$ to the preceding model by finding the MLE, which has no general close form solution.

The standard augmentation is

$$t_p \equiv t_p(\mu,\Sigma,\nu) = \mu + \Sigma^{1/2}Z/\sqrt{q},\quad Z \sim N_p(0, I_p),$$

$$q \sim \chi^2_\nu/\nu,\quad Z \perp q.$$

The augmented likelihood is the product of

$$f(y\,|q,\mu,\Sigma) \propto q^{p/2}|\Sigma|^{-1/2}\exp\{-\frac{q}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\},$$

$$\text{and } f(q) \propto q^{\frac{\nu}{2}-1}e^{-\frac{\nu q}{2}}.$$

The E-step amounts to calculate

$$w_i^{(t+1)} = E\left(q_i \middle| y_i, \mu^{(t)}, \Sigma^{(t)}\right) = \frac{\nu + p}{\nu + d_i^{(t)}},$$

where $d_i^{(t)} = \left(y_i - \mu^{(t)}\right)^T \left(\Sigma^{(t)}\right)^{-1} \left(y_i - \mu^{(t)}\right)$.

Note: The expectation of Gamma(α, β) is $\alpha\beta$. Here we have

$$\alpha = \frac{\nu + p}{2}, \quad \beta = \frac{2}{\nu + \left(y_i - \mu^{(t)}\right)^T \left(\Sigma^{(t)}\right)^{-1} \left(y_i - \mu^{(t)}\right)}.$$

The M-step yields

$$\mu^{(t+1)} = \sum_i w_i^{(t+1)} y_i / \sum_i w_i^{(t+1)},$$

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_i w_i^{(t+1)} \left(y_i - \mu^{(t+1)}\right)\left(y_i - \mu^{(t+1)}\right)^T.$$

2.2 A working parameter

$$t_p(\mu, \Sigma, \nu) = \mu + \frac{\left|\Sigma\right|^{-a/2} \Sigma^{1/2} Z}{\sqrt{q(a)}}, \quad Z \sim N_p(0, I_p),$$

$$q(a) \sim \left|\Sigma\right|^{-a} \chi_\nu^2 / \nu, \quad Z \perp q(a).$$

A portion of the scale factor has been moved into the missing data. The standard augmentation corresponds to $a = 0$. What this accomplishes is to transform part of the M-step to E-step (or vice versa).

We will show that the optimum value for a is $a_{\text{opt}} = 1/(\nu + p)$. Amazingly, the corresponding optimal EM algorithm differs from the standard algorithm by a trivial modification, i.e. by replacing the denominator n in estimating the covariance matrix with the sum of weights

$$\Sigma^{(t+1)} = \sum_i w_i^{(t+1)} (y_i - \mu^{(t+1)})(y_i - \mu^{(t+1)})^T / \sum_i w_i^{(t+1)}.$$

2.3 A simulation study

Generate 100 observation from

(a) $N(0, 1)$.

(b) $t_1(0, 1, \nu = 1)$ (i.e. standard Cauchy).

(c) a mixture of two-thirds $N(0, 1)$ and one third exponential with mean 3.

Fit $t_1(\mu, \Sigma, \nu)$ with $\nu = 1$ and 5 to each data set using the standard and the optimal algorithm. The simulation was repeated 1000 times. In all 6000 cases the optimal algorithm is faster than the standard algorithm. In 5997 cases the improvement was greater than 10% and often reaches as high as 50%.

A second simulation was run to investigate the improvement in higher dimensions. The improvement seems to be much more pronounced in higher dimensional problems.

3. A theorem concerning the speed of convergence:

Recall that the rate of convergence for the algorithm is

$$D\,M = I - I_{\text{obs}} I_{\text{aug}}^{-1}.$$

Consider the fundamental identity

$$Q(\phi'|\phi) = L(\phi') + H(\phi'|\phi), \text{ where}$$

$$Q(\phi'|\phi) = E[\log\{f(Y_{\text{aug}}|\phi')\}|Y_{\text{obs}}, \phi],$$

$$H(\phi'|\phi) = E[\log\{k(Y_{\text{aug}}|Y_{\text{obs}}, \phi')\}|Y_{\text{obs}}, \phi].$$

$$
\begin{aligned}
& D^{10}Q\left(\phi^{(t+1)}\middle|\phi^{(t)}\right) \\
= \; & D^{10}Q\left(\phi^{*}\middle|\phi^{*}\right) + \left(\phi^{(t+1)} - \phi^{*}\right)D^{20}Q\left(\phi^{*}\middle|\phi^{*}\right) + \\
+ \; & \left(\phi^{(t)} - \phi^{*}\right)D^{11}Q\left(\phi^{*}\middle|\phi^{*}\right) + \cdots .
\end{aligned}
$$

Thus

$$
0 = \left(\phi^{(t+1)} - \phi^{*}\right)D^{20}Q\left(\phi^{*}\middle|\phi^{*}\right) + \left(\phi^{(t)} - \phi^{*}\right)D^{11}Q\left(\phi^{*}\middle|\phi^{*}\right) + \cdots
$$

Note that $D^{11}Q\left(\phi^{*}\middle|\phi^{*}\right) = -D^{20}H\left(\phi^{*}\middle|\phi^{*}\right)$.

Hence $0 = DM \times D^{20}Q\left(\phi^{*}\middle|\phi^{*}\right) - D^{20}H\left(\phi^{*}\middle|\phi^{*}\right)$.

Consequently, $DM\, I_{\text{aug}} = I_{\text{mis}} = I_{\text{aug}} - I_{\text{obs}}$,

where $I_{\text{aug}} = -D^{20}Q\left(\phi^{*}\middle|\phi^{*}\right)$, $I_{\text{mis}} = -D^{20}H\left(\phi^{*}\middle|\phi^{*}\right)$, and $I_{\text{obs}} = -D^{20}L\left(\phi^{*}\right)$. 2

Define the speed of convergence to be

$$
S^{EM} = I_{\text{obs}} I_{\text{aug}}^{-1}.
$$

Then we have

Theorem 1. If $I_{\text{aug}}(a) \geq I_{\text{aug}}(a')$ then $s(a) \leq s(a')$.

4. Return to the t-model example:

For fixed a, the log-likelihood for (μ, Σ) based on $Y_{\mathrm{aug}}(a)$ is

$$L(\mu, \Sigma | Y_{\mathrm{aug}}(a)) = \frac{n}{2}\{a(p+\nu) - 1\} \log|\Sigma| -$$
$$- \frac{1}{2}|\Sigma|^a \sum_i q_i(a)\{\nu + (\bar{y}_w - \mu)^T \Sigma^{-1}(\bar{y}_w - \mu) + \mathrm{tr}(\Sigma^{-1} S_w)\},$$

where $S_w = \Sigma_i q_i(a)(y_i - \bar{y}_w)(y_i - \bar{y}_w)^T / \Sigma_i q_i(a)$,

$$\bar{y}_w = \sum_i q_i(a) y_i / \sum_i q_i(a).$$

To find the MLE of Σ, we need to solve the following score equation

$$\frac{\partial}{\partial \Psi} L(\mu, \Sigma | Y_{\mathrm{aug}}(a)) = -\frac{n}{2}\{a(p+\nu) - 1\}\mathcal{D}(\Sigma) -$$
$$- \frac{n}{2}|\Sigma|^a \bar{q}(a)\{\nu + (\bar{y}_w - \mu)(\bar{y}_w - \mu)^T + \mathcal{D}(S_w)\} +$$
$$+ \frac{n\,a}{2}\bar{q}(a)|\Sigma|^a \bar{q}(a)\{\nu + (\bar{y}_w - \mu)^T \Sigma^{-1}(\bar{y}_w - \mu) +$$
$$+ \mathrm{tr}(\Sigma^{-1} S_w)\}\mathcal{D}(\Sigma) = 0,$$

where $\bar{q}(a) = \sum_{i=1}^{n} q_i(a)/n$ and $\mathcal{D}(A) = 2A - \mathrm{diag}(A)$.

Evaluating the preceding score equation at the MLE of $\mu = \bar{y}_w$, we see that the MLE of Σ satisfies

$$\frac{a(p+\nu)-1}{|\Sigma|^a \bar{q}(a)}\Sigma + S_w = a\{\nu + \mathrm{tr}(\Sigma^{-1}S_w)\}\Sigma.$$

Solving this equation for general a is difficult if not impossible, but there are two values of a that make it trivial to solve. One is $a = 0$, which corresponds to the standard augmentation scheme. The other is $a_{\mathrm{opt}} = 1/(\nu + p)$, for which the solution is proportional to S_w. It is easy to verify that the proportional constant must be 1. Furthermore

$$I_{\mathrm{aug}}(a) - I_{\mathrm{aug}}(a_{\mathrm{opt}}) = (\nu + p)(a - \frac{1}{\nu + p})^2 C_n,$$

where C_n is a positive semidefinite matrix and does not depend on a. Clearly, $a_{\mathrm{opt}} = 1/(\nu + p)$ minimizes the augmented information.

5. Model reduction and conditional maximization:

The idea is to break the model into smaller models to match the 'data'. For example, we may be able to find a partition $\theta = (\theta_1, \theta_2)$ such that $L(\theta | Y_{\text{obs}})$ is easy to maximize for θ_1 given θ_2 and vice versa. In such cases, we can iterate between this two conditional (or constrained) maximization until converge to maximize $L(\theta | Y_{\text{obs}})$. In terms of the algorithm, the EM is such that

E-step: impute the augmented log-likelihood $L(\theta | Y_{\text{aug}})$ by

$$Q(\theta | \theta^{(t)}) = E[L(\theta | Y_{\text{aug}}) | Y_{\text{obs}}, \theta^{(t)}].$$

M-step: determine $\theta^{(t+1)}$ by maximizing the imputed log-likelihood $Q(\theta | \theta^{(t)})$,

$$Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}), \text{ for all} \theta \in \Theta.$$

We can replace the M-step by a set of CM steps: s th CM-step- determine $\theta^{(t+s/S)}$ by maximizing the constrained imputed log-likelihood

$$Q(\theta^{(t+s/S)} | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}),$$

for all $\theta \in \Theta_s^{(t)} = \{\theta \in \Theta : g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})\}$,

where $G^{[c]} = \left\{ g_s^{[c]}, s = 1, \cdots, S^{[c]} \right\}$ is the set of contraining functions.

Under certain conditions on $G^{[c]}$, Meng and Rubin (1993) proved that ECM algorithm retains the stable convergence properties of the EM algorithm. In addition they defined the multicycle ECM (MCECM) algorithm, which inserts an E-step before some of the CM-steps.

Liu and Rubin (1994) recognized that in some applications of the ECM algorithm the implementation of some of the CM-steps requires similar computation for maximizing the observed data log-likelihood and for maximizing the expected augmented data log-liklihood under the same constraint. In such cases, one is tempted to perform the former instead of the later, which give rise to the ECME algorithm whose CM-steps maximize either the imputed log-likelihood, $Q\left(\theta \left| \theta^{(t)} \right.\right)$, or the actual log-likelihood, $L\left(\theta \left| Y_{\text{obs}} \right.\right)$. They also provide numerical examples to demonstrate the gain in central processor unit (CPU) time, sometimes by a factor of 10.

6.1 Alternating Expectation-Conditional Maximization Algorithm (AECM)

A *cycle* consists of an E-step followed by an ordered set of CM-steps, the last of which will be followed immediately by a new E-step (which is itself the beginning of the next cycle). An *iteration* consists of one or more cycles.

$$\underbrace{\underbrace{\text{E-step}, \{\text{CM-step}, s = 1, \cdots, S^{[c_t+1]}\}}_{c_t+1\text{th cycle}}, \cdots, \underbrace{\text{E-step}, \{\text{CM-step}, s = 1, \cdots, S^{[c_{t+1}]}\}}_{c_{t+1}\text{th cycle}}}_{t\text{th iteration}}$$

The $(c+1)$th cycle of the AECM algorithm consists of

E-step: compute $Q^{[c+1]}(\theta \,|\, \theta^{[c]}) = E\left[L(\theta \,|\, Y_{\text{aug}}^{[c+1]}) \,|\, Y_{\text{obs}}, \theta^{[c]}\right]$

and

$S^{[c+1]}$ CM-steps: sth CM-step $(s = 1, \cdots, S^{[c+1]})$ calculates $\theta^{[c+s/S^{[c+1]}]}$ such that

$$Q^{[c+1]}(\theta^{[c+s/S^{[c+1]}]} \,|\, \theta^{[c]}) \geq Q^{[c+1]}(\theta \,|\, \theta^{[c]}), \text{ for all}$$

$$\theta \in \Theta_s^{[c+1]} = \left\{\theta \in \Theta : g_s^{[c+1]}(\theta) = g_s^{[c+1]}(\theta^{[c+(s-1)/S^{[c+1]}]})\right\}.$$

Specical cases of the AECM algorithms

Case	$Y_{\text{aug}}^{[c]}$	$C^{(t)}$	$S^{[c]}$	G
CM	Y_{obs}	1	S	G
EM	Y_{aug}	1	1	None
ECM	Y_{aug}	1	S	G
MCECM	Y_{aug}	C	1	$G_{(c \text{ modula } C)}$
ECME	$\begin{cases} Y_{\text{aug}} \text{ odd} \\ Y_{\text{obs}} \text{ even} \end{cases}$	2	$\begin{cases} S_1 \text{ odd} \\ S_2 \text{ even} \end{cases}$	$\begin{cases} G_1 \text{ odd} \\ G_2 \text{ even} \end{cases}$
SAGE	$Y_{\text{aug}}^{[c]}$	$C^{(t)}$	1	Partition of θ

6.2. Convergence results for AECM

Theorem 2: Any AECM sequence increases (or maintains) $L\left(\theta \middle| Y_{\text{obs}}\right)$ at every cycle and thus increases (or maintains) $L\left(\theta \middle| Y_{\text{obs}}\right)$ at every iteration.

Proof: By the construction of the AECM algorithm, we have

$$Q^{[c+1]}\left(\theta^{[c+s/S^{[c+1]}]} \middle| \theta^{[c]}\right) \geq Q^{[c+1]}\left(\theta^{[c+(s-1)/S^{[c+1]}]} \middle| \theta^{[c]}\right)$$

and thus $Q^{[c+1]}\left(\theta^{[c+1]} \middle| \theta^{[c]}\right) \geq Q^{[c+1]}\left(\theta^{[c]} \middle| \theta^{[c]}\right)$, which implies that $L\left(\theta^{[c+1]} \middle| Y_{\text{obs}}\right) \geq L\left(\theta^{[c]} \middle| Y_{\text{obs}}\right)$.

Theorem 3: In addition to regularity conditions (6)-(10) in Wu's (1983), suppose that

(a) all the CM steps are unique and

(b) the AECM iteration mapping, $M_{(t)}^{\mathrm{AECM}} : \theta^{(t)} \longrightarrow \theta^{(t+1)}$, does not depend on t.

Then all the limit points of an AECM sequence $\{\theta^{(t)}, t \geq 0\}$ are stationary points of $L(\theta \,|Y_{\mathrm{obs}})$.

Theorem 4: Suppose that the AECM iteration mapping is a composition of C fixed cycle mappings, all the CM steps satisfy the Lagrange multiplier equations and $\theta^{[c+s/S^{[c+1]}]} \longrightarrow \theta^{*}$ as $c \longrightarrow \infty$. Then the matrix rate of convergence of the AECM iteration is

$$\mathrm{DM}^{\mathrm{AECM}} = \prod_{c=1}^{C} \{I - I_{\mathrm{obs}}(I_{\mathrm{aug}}^{[c]})^{-1}(I - \prod_{s=1}^{S^{[c]}} P_s^{[c]})\},$$

where

$$I_{\mathrm{aug}}^{[c]} = -E\,[\frac{\partial^2\{\log f(Y_{\mathrm{aug}}^{[c]}|\theta)\}}{\partial\theta\,\partial\theta^T}\Big|Y_{\mathrm{obs}}, \theta\,]\Big|_{\theta=\theta^*},$$

$$P_s^{[c]} = \nabla_s^{[c]}\{(\nabla_s^{[c]})^T(I_{\mathrm{aug}}^{[c]})^{-1}\nabla_s^{[c]}\}^{-1}(\nabla_s^{[c]})^T(I_{\mathrm{aug}}^{[c]})^{-1},$$

$$\nabla_s^{[c]} = \nabla g_s^{[c]}(\theta^*).$$

7. Example: t-model with unknown degrees of freedom

We first reduce the model by breaking the parameter space into two parts to use a two-cycle algorithm. In the first cycle, we update (μ, Σ) given the degrees of freedom ν and, in the second cycle, we update ν given (μ, Σ).

$$\Theta^{[c]} = \left\{\theta = (\mu, \Sigma, \nu) \in \Theta : \nu = \nu^{[c-1]}\right\} \text{ for } c \text{ odd.}$$

$$\Theta^{[c]} = \left\{\theta \in \Theta : (\mu, \Sigma) = (\mu^{[c-1]}, \Sigma^{[c-1]})\right\} \text{ for } c \text{ even.}$$

Given ν, there are two choices for the data augmentation which results in CM-steps that are simple to compute: The first one is the standard augmentation

$$\left\{(y_i, q_i), i = 1, \cdots, n\right\}$$

and the second is the optimal augmentation

$$\left\{(y_i, q_i(a_{\text{opt}}^{[c]})), i = 1, \cdots, n\right\},$$

where $q_i(a) = |\Sigma|^{-a} q_i$ and $a_{\text{opt}}^{[c]} = 1/(\nu^{[c]} + p)$. Because $\nu^{[c]}$ changes with c, the optimal augmentation is not fixed but rather is a function of the cycle and iteration.

There are two useful augmentation schemes when conditioning on $(\mu\,,\,\Sigma)$: the standard augmentation $\left\{(y_i,\,q_i),\,i\,=\,1,\,\cdots,\,n\,\right\}$ and no augmentation $\left\{y_i,\,i\,=\,1,\,\cdots,\,n\,\right\}$.

The odd cycles are conditional on the current iterate $\nu^{[c]}$ and are implemented exactly as in the example of t-model with known ν. Regardless of the data augmentation, the even cycles require numerical optimization of $Q^{[c+2]}\left(\theta\,\middle|\,\theta^{[c+1]}\right)$, where c is even.

When using the standard augmentation

$$Y_{\mathrm{aug}}\,=\,\left\{(y_i,\,q_i),\,i\,=\,1,\,\cdots,\,n\,\right\},$$

we first execute an E-step which sets

$$w_i^{[c+2]}\,=\,\left(\nu^{[c]}\,+\,p\,\right)/\left(\nu^{[c]}\,+\,d_i^{[c+1]}\right)$$

for each i, and then set $\nu^{[c+2]}$ equal to the solution of the equation

$$-\,\phi\left(\frac{\nu}{2}\right)+\log\left(\frac{\nu}{2}\right)+\phi\left(\frac{\nu^{[c]}\,+\,p}{2}\right)-\log\left(\frac{\nu^{[c]}\,+\,p}{2}\right)+$$
$$+\,\frac{1}{n}\sum_{i=1}^{n}\left(\log w_i^{[c+2]}\,-\,w_i^{[c+2]}\right)+1\,=\,0,$$

where ϕ is the digamma function.

Likewise, when using $Y_{\text{aug}} = \{y_i, i = 1, \cdots, n\}$, ν is updated by setting $\nu^{[c+2]}$ equal to the solution of the equation

$$-\phi\left(\frac{\nu}{2}\right) + \log\left(\frac{\nu}{2}\right) + \phi\left(\frac{\nu + p}{2}\right) - \log\left(\frac{\nu + p}{2}\right) + \\ + \frac{1}{n}\sum_{i=1}^{n}\left(\log \tilde{w}_i^{[c+2]}(\nu) - \tilde{w}_i^{[c+2]}\right)(\nu) + 1 = 0,$$

where $\tilde{w}_i^{[c+2]} = (\nu + p)/(\nu + d_i^{[c+1]})$.

In the following table, we list four algorithms according to their respective model reduction and data augmentation scheme.

Algorithm	Model reduction (CM-steps)			
	Update μ and Σ		Update ν	
	Data augmentation (E-step)			
	$\{y_i, q_i\}$	$\{y_i, q_i(a)\}$	$\{y_i, q_i\}$	$\{y_i\}$
MCECM	X		X	
ECME	X			X
AECM1		X	X	
AECM2		X		X

To compare the performance of the four algorithms described in the preceding table, we fitted a 10-dimensional t-model to 100 observations generated from $t_{10}(0, V, \nu = 1)$, where $V > 0$ was selected at the outset of the simulation as a non-diagonal matrix.

The simulation results shows that AECM1 is 8-12 times faster than the MCECM or ECME algorithms. Note that the cost per iteration is less for AECM1 and for the MCECM algorithm than for the ECME algorithm, mainly because the difference in computation required for respective equation solving. Moreover, AECM2 was only slightly more efficient than AECM1 in terms of the number of iterations required, and less efficient in terms of actual computer time.

A simulation study is also performed on data with missing values. The conclusion is: for fitting a t-model with unknown degrees of freedom, we recommend AECM 1 unless there are many missing values among the y_i in which case AECM2 may be somewhat more efficient than AECM1.

8.1 Image Reconstruction under Poisson Model

$$Y_n \sim \text{Poisson}\Big(\sum_{k=1}^{p} a_{nk}\lambda_k + r_n\Big),$$

independent for $n = 1, \cdots, N$, where k $(= 1, \cdots, p)$ indexes pixels, n $(= 1, \cdots, N)$ indexes the bins (e.g. a pair of detectors), a_{kn} is the conditional probability that an emission from pixel k will be detected by bin n, r_n is the background rate and $\{a_{nk}, r_n\}$ are known. We would like to estimate the emission density $\lambda = (\lambda_1, \cdots, \lambda_p)$ by maximum likelihood.

Standard EM implementation for this problem is to use:

$$Y_{\text{aug}}^{\text{EM}} = \{Y_{nk}, R_n, 1 \le k \le p, 1 \le n \le N\},$$

where $Y_{nk} \sim \text{Poisson}(a_{nk}\lambda_k)$, $R_n \sim \text{Poisson}(r_n)$ and all the Poisson variables are independent. Clearly the preceding augmentation is a massive one over the observed data and thus leads to a slow EM algorithm.

8.2 Another augmentation

Given λ_j, $j \neq k$, we can write Y_n as

$$Y_n \sim \text{Poisson}\{a_{nk}\lambda_k + r_n + \Lambda_{nk}(\lambda)\},$$

independent for $n = 1, \cdots, N$, where $\Lambda_{nk}(\lambda) = \Sigma_{j \neq k} a_{nj}\lambda_j$ is known. We can consider the following augmentation.

$$Y_{\text{aug}}^{[k]}(\beta_k) = \{Z_{nk}(\beta_k), B_{nk}(\beta_k), 1 \leq n \leq N\},$$

where $Z_{nk}(\beta_k) \sim \text{Poisson}\{a_{nk}(\lambda_k + \beta_k)\}$,

$$B_{nk}(\beta_k) \sim \text{Poisson}\{r_n - a_{nk}\beta_k + \Lambda_{nk}(\lambda)\},$$

and β_k is a non-negative working parameter such that the Poisson parameter for B_{nk} is non-negative.

8.3 The E-step and the CM-step

From the decomposition $Y_n = Z_{nk}(\beta_k) + B_{nk}(\beta_k)$, if $Z_{nk}(\beta_k)$ were observed then the MLE for λ_k would be

$$\hat{\lambda}_k = \max(\frac{Z_{\cdot k}}{a_{\cdot k}} - \beta_k, 0),$$

where $Z_{\cdot k} = \Sigma_{n=1}^{N} Z_{nk}(\beta_k)$ and $a_{\cdot k} = \Sigma_{n=1}^{N} a_{nk}$.

This defines the M-step of the EM algorithm for the reduced model, or equivalently the CM step of the ECM algorithm for updating λ_k given λ_j, $j \neq k$.

Since $z_{\cdot k}$ is unobserved, and it is the augmented data sufficient statistics for λ, the E-step replaces $z_{\cdot k}$ in the preceding equation concerning $\hat{\lambda}_k$ by its conditional expectation given $Y_{\text{obs}} = \{y_n, 1 \leq n \leq N\}$ and the most up-to-date estimate of λ, say, $\lambda^{[c]}$,

$$\begin{aligned} z_{\cdot k}^{[c+1]} &= E\left(z_{\cdot k} \middle| Y_{\text{obs}}, \lambda_k^{[c]}\right) \\ &= \left(\lambda_k^{[c]} + \beta_k\right) \sum_{n=1}^{N} \frac{a_{nk} y_n}{a_{nk} \lambda_k^{[c]} + r_n + \Lambda_{nk}\left(\lambda^{[c]}\right)}. \end{aligned}$$

Once we have completed an E-step and an M-step for λ_k, which corresponds to the kth cycle of the AECM algorithm, we move on to the next cycle for λ_{k+1}, and so forth until we complete an iteration of AECM when all of $\lambda = (\lambda_1, \cdots, \lambda_p)$ is updated.

8.4 Select the value of working parameters

Now let's consider the choice of the value of working parameters $\{\beta_k, 1 \leq k \leq p\}$. Fessler and Hero (1994) used $\beta_k = z_k$

$$z_k = \min_{n:a_{nk}>0} (r_n / a_{nk}), \quad k = 1, \cdots, p.$$

This value of β_k produces a faster algorithm than ECM without introducing the working parameter (which is the same as setting $\beta_k = 0$). But it can still be improved on.

It is easy to verify that the augmented data Fisher information at MLE $\lambda_k^* > 0$ is

$$I_{\text{aug}}^{[k]}(\beta_k) = \frac{a_{\cdot k}}{\lambda_k^* + \beta_k}.$$

The value of β_k that minimize the preceding Fisher information under the constraint

$$\beta_k \leq \frac{r_n + \Lambda_{nk}(\lambda)}{a_{nk}}, \quad \text{for all } n : a_{nk} > 0$$

is

$$\beta_k^{\text{opt}} = \min_{n:a_{nk}>0} \left\{ \frac{r_n + \Lambda_{nk}(\lambda)}{a_{nk}} \right\}.$$

It can be verified that in this example, the Newton-Raphson algorithm, the EM algorithm with $\beta_k = 0$, and Fessler and Hero's algorithm with $\beta_k = z_k$ all have the following form of mapping

$$\lambda_k^{(t+1)} = \max[\lambda_k^{(t)} - \{a_{\cdot k} - e_k(\lambda^{[c]})\} S_k(\lambda^{[c]}), 0],$$

where $e_k(\lambda) = \Sigma_{n=1}^{N} \frac{a_{nk} y_n}{\Sigma_{k=1}^{p} a_{nk}\lambda_k + r_n}$, and for the Newton-Raphson, EM and Fessler and Hero's (1994) algorithm

$$S_k^{\mathrm{NR}}(\lambda) = \left\{ \sum_{i=1}^{N} a_{nk}^2 y_n / \left(\sum_{k=1}^{p} a_{nk}\lambda_k + r_n \right)^2 \right\}^{-1},$$

$$S_k^{\mathrm{EM}}(\lambda) = \lambda_k / a_{\cdot k},$$

$$S_k^{\mathrm{FH}}(\lambda) = (\lambda_k + z_k) / a_{\cdot k}.$$

For the optimal algorithm,

$$\begin{aligned} S_k^{\mathrm{OPT}}(\lambda) &= \frac{1}{a_{\cdot k}} \min_{n:a_{nk}>0} \left\{ \frac{\Lambda_{nk}(\lambda)}{a_{nk}} + \frac{r_n}{a_{nk}} + \lambda_k \right\} \\ &\geq \frac{1}{a_{\cdot k}} \min_{n:a_{nk}>0} \left\{ \frac{\Lambda_{nk}(\lambda)}{a_{nk}} \right\} + \frac{z_k + \lambda_k}{a_{\cdot k}}, \end{aligned}$$

and thus

$$S_k^{\mathrm{OPT}}(\lambda) > S_k^{\mathrm{FH}}(\lambda) > S_k^{\mathrm{EM}}(\lambda).$$

Also note that

$$
\begin{aligned}
& S_k^{\mathrm{NR}}(\lambda^*) \\
\geq\ & \Big[\max_{n:a_{nk}>0}\Big\{\frac{a_{nk}}{\sum_{k=1}^{p} a_{nk}\lambda_k^* + r_n}\Big\}\sum_{n=1}^{N}\frac{a_{nk}y_n}{\sum_{k=1}^{p} a_{nk}\lambda_k^* + r_n}\Big]^{-1} \\
\geq\ & S_k^{\mathrm{OPT}}(\lambda^*).
\end{aligned}
$$

In other words, the optimal algorithm is the closest to the Newton-Raphson algorithm near convergence, but with more stable behavior than the Newton-Raphson algorithm in the early iterations because of the guaranteed monotone convergence in likelihood. The key idea here is to search for algorithms within the class of algorithms with monotone convergence in likelihood , that have as large a step size as possible. The AECM formulation provides a more flexible setting for this search than did its predecessors.