EM Algorithm II

September 4, 2022

Review EM

$(Y_{\text{obs}}, Y_{\text{mis}}) \sim f(y_{\text{obs}}, y_{\text{mis}}|\theta)$, we observe Y_{obs} but not Y_{mis}.

Complete-data log likelihood: $l_C(\theta|Y_{\text{obs}}, Y_{\text{mis}}) = \log\{f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)\}$

Observed-data log likelihood: $l_O(\theta|Y_{\text{obs}}) = \log\left\{\int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\, dy_{\text{mis}}\right\}$

EM algorithm:

- **E step**: $h^{(k)}(\theta) \equiv \mathrm{E}\left\{l_C(\theta|Y_{\text{obs}}, Y_{\text{mis}})\middle|Y_{\text{obs}}, \theta^{(k)}\right\}$. Note, the expectation is taken with respect to $Y_{\text{mis}}|Y_{\text{obs}}$, so

$$\mathrm{E}\left\{l_C(\theta|Y_{\text{obs}}, Y_{\text{mis}})\middle|Y_{\text{obs}}, \theta^{(k)}\right\} = \int l_C(\theta|Y_{\text{obs}}, Y_{\text{mis}}) f(y_{\text{mis}}|Y_{\text{obs}}, \theta) dy_{\text{mis}}.$$

- **M step**: $\theta^{(k+1)} = \arg\max_\theta h^{(k)}(\theta)$

Ascent property: $l_O(\theta^{(k)}|Y_{\text{obs}})$ is non-decreasing along k. If you can calculate it, it is a good idea to monitor it for debugging purpose.

Issues:

1. Could trap in local maxima.
2. Slow convergence.

Standard errors for the EM estimates

Numerical approximation of the Hessian matrix

Note: $l(\theta) =$ observed-data log-likelihood

We estimate the gradient using

$$\{\dot{l}(\theta)\}_i = \frac{\partial l(\theta)}{\partial \theta_i} \approx \frac{l(\theta + \delta_i e_i) - l(\theta - \delta_i e_i)}{2\delta_i}$$

where e_i is a unit vector with 1 for the ith element and 0 otherwise.

In calculating derivatives using this formula, generally start with some medium size δ and then repeatedly halve it until the estimated derivative stabilizes.

We can estimate the Hessian by applying the above formula twice:

$$\{\ddot{l}(\theta)\}_{ij} \approx \frac{l(\theta + \delta_i e_i + \delta_j e_j) - l(\theta + \delta_i e_i - \delta_j e_j) - l(\theta - \delta_i e_i + \delta_j e_j) + l(\theta - \delta_i e_i - \delta_j e_j)}{4\delta_i\delta_j}$$

Standard errors for the EM estimates

Louis estimator

Louis TA (1982) *Finding the observed information matrix when using the EM algorithm*. **J R Statist Soc** 44: 226-233.

Problem setup:

$$l_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}) \equiv \log\left\{f(Y_{\mathrm{obs}}, Y_{\mathrm{mis}}|\theta)\right\}$$

$$l_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) \equiv \log\left\{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\; dy_{\mathrm{mis}}\right\}$$

$\dot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}),\ \dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$: gradients of $l_{\mathrm{C}},\ l_{\mathrm{O}}$

$\ddot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}),\ \ddot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$: second derivatives of $l_{\mathrm{C}},\ l_{\mathrm{O}}$

We can show that

$$(5)\qquad \dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) = \mathrm{E}\left\{\dot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\}$$

$$(6)\quad -\ddot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) = \mathrm{E}\left\{-\ddot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\} - \mathrm{E}\left\{\left[\dot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})\right]^{\otimes 2}\middle|Y_{\mathrm{obs}}\right\} + \left[\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})\right]^{\otimes 2}$$

Given MLE: $\hat{\theta} = \arg\max_{\theta} l_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$
Louis variance estimator is: $\left\{-\ddot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})\right\}^{-1}$ evaluated at $\theta = \hat{\theta}$

Note:

- All of the conditional expectations (right hand size of equation 6) can be computed in the EM algorithm using only $\dot{l}_{\mathrm{C}}$ and $\ddot{l}_{\mathrm{C}}$, which are first and second derivatives of the complete-data log-likelihood.
- Louis estimator should be evaluated at the last step of EM.

Proof of (5)

Proof: By the definition of $l_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$,

$$\begin{aligned}
\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) &= \frac{\partial \log \left\{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}\right\}}{\partial\theta} \\
&= \frac{\partial \left\{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}\right\} / \partial\theta}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \\
&= \frac{\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \qquad (7)
\end{aligned}$$

Multiplying and dividing the integrand of the numerator by $f(Y_{\text{obs}}, y_{\text{mis}}|\theta)$ gives (5),

$$
\begin{aligned}
i_{\text{O}}(\theta|Y_{\text{obs}}) &= \frac{\int \frac{f'(Y_{\text{obs}}, y_{\text{mis}}|\theta)}{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)} f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}}}{\int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}}} \\
&= \frac{\int \frac{\partial \log\{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\}}{\partial\theta} f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}}}{\int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}}} \\
&= \int i_{\text{C}}(\theta|Y_{\text{obs}}, Y_{\text{mis}}) \frac{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)}{\int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}}}\ dy_{\text{mis}} \\
&= \int i_{\text{C}}(\theta|Y_{\text{obs}}, Y_{\text{mis}}) f(y_{\text{mis}}|Y_{\text{obs}}, \theta)\ dy_{\text{mis}} \\
&= \text{E}\left\{ i_{\text{C}}(\theta|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}} \right\}.
\end{aligned}
$$

For above, we used the following results

$$
\begin{aligned}
\frac{\partial \log\{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\}}{\partial\theta} &= \frac{f'(Y_{\text{obs}}, y_{\text{mis}}|\theta)}{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)} \\
\int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\ dy_{\text{mis}} &= f(Y_{\text{obs}}|\theta) \\
\frac{f(Y_{\text{obs}}, y_{\text{mis}}|\theta)}{f(Y_{\text{obs}}|\theta)} &= f(y_{\text{mis}}|Y_{\text{obs}}, \theta)
\end{aligned}
$$

Proof of (6)

Note, expression (7) is

$$\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) = \frac{\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}$$

We take an additional derivative of $\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$ in expression (7) to obtain

$$\begin{aligned}
\ddot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) &= \frac{\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} - \left\{ \frac{\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \right\}^2 \\
&= \frac{\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} - \left\{ \dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) \right\}^{\otimes 2} \\
&= \frac{\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{f(Y_{\mathrm{obs}}|\theta)} - \left\{ \dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) \right\}^{\otimes 2} \qquad (8)
\end{aligned}$$

To see how the first term breaks down, we take an additional derivative of

$$\int f'(Y_{\rm obs}, y_{\rm mis}|\theta)\, dy_{\rm mis} \;= \int \frac{\partial \log\{f(Y_{\rm obs}, y_{\rm mis}|\theta)\}}{\partial\theta} f(Y_{\rm obs}, y_{\rm mis}|\theta)\, dy_{\rm mis}$$

to obtain

$$\begin{aligned}\int f''(Y_{\rm obs}, y_{\rm mis}|\theta)\, dy_{\rm mis} &= \int \frac{\partial^2 \log\{f(Y_{\rm obs}, y_{\rm mis}|\theta)\}}{\partial\theta\partial\theta'} f(Y_{\rm obs}, y_{\rm mis}|\theta)\, dy_{\rm mis} \\ &+ \int \left[\frac{\partial \log\{f(Y_{\rm obs}, y_{\rm mis}|\theta)\}}{\partial\theta}\right]^{\otimes 2} f(Y_{\rm obs}, y_{\rm mis}|\theta)\, dy_{\rm mis}\end{aligned}$$

Thus we express the first term in equation (8) to be

$$\mathrm{E}\left\{\ddot{l}_{\rm C}(\theta|Y_{\rm obs}, Y_{\rm mis})|Y_{\rm obs}\right\} + \mathrm{E}\left\{\left[\dot{l}_{\rm C}(\theta|Y_{\rm obs}, Y_{\rm mis})\right]^{\otimes 2}\middle|Y_{\rm obs}\right\}.$$

Convergence rate of EM

Let $I_{\mathrm{C}}(\theta)$ and $I_{\mathrm{O}}(\theta)$ denote the complete information and observed information, respectively.

One can show when the EM converges, the linear convergence rate, denoted as $(\theta^{(k+1)} - \hat{\theta})/(\theta^{(k)} - \hat{\theta})$ approximates $1 - I_{\mathrm{O}}(\hat{\theta})/I_{\mathrm{C}}(\hat{\theta})$.

This means that

- When missingness is small, EM converges quickly
- Otherwise EM converges slowly.

SEM algorithm

- **EM algorithm** does not generate asymptotic covariance matrix (standard errors) for parameters as a byproduct.
- The asymptotic covariance matrix for $\hat{\theta}$, denoted as V, can be found as $\left\{-\ddot{l}_{\mathrm{O}}(\hat{\theta}|Y_{\mathrm{obs}})\right\}^{-1}$. However, the derivations can be difficult to evaluate directly.
- In contrast, $-\ddot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$, and hence $I_{\mathrm{OC}} \equiv \mathrm{E}\left\{-\ddot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\}$ is relatively easier to evaluate.
- **Louis estimator** for covariance matrix, i.e.,

$$\mathrm{E}\left\{-\ddot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\} - \mathrm{E}\left\{\left[\dot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})\right]^{\otimes 2}\Big|Y_{\mathrm{obs}}\right\} + \left[\dot{l}_{\mathrm{O}}(\hat{\theta}|Y_{\mathrm{obs}})\right]^{\otimes 2}$$

  requires calculation of the conditional expectation of the square of the complete-data score function, which is specific to each problem.
- **S**upplemented **EM algorithm** (Meng & Rubin, 1991) obtains covariance matrix by using only the code for computing the complete-data covariance matrix, the code for EM itself, and code for standard matrix operations.

SEM algorithm (continued)

- EM defines a mapping, $M : \theta^{(k+1)} = M(\theta^{(k)})$, where $M(\theta) = (M_1(\theta), \ldots, M_p(\theta))$

- Let $\{DM\}_{ij} = (\partial M_j(\theta)/\partial\theta_i)|_{\theta=\hat{\theta}}$, which is a $p \times p$ matrix. We can show that

$$\theta^{(k+1)} - \hat{\theta} \approx DM(\theta^{(k)} - \hat{\theta}),$$

  which means DM is the rate of convergence of EM.

  Proof: Because $\theta^{(k+1)} = M(\theta^{(k)})$ and $\hat{\theta} = M(\hat{\theta})$, $\theta^{(k+1)} - \hat{\theta} = M(\theta^{(k)}) - M(\hat{\theta})$. By Taylor series expansion on the right hand side, we have $\theta^{(k+1)} - \hat{\theta} \approx DM(\theta^{(k)} - \hat{\theta})$.

- It has been shown that $V = I_{\mathrm{OC}}^{-1}(I - DM)^{-1}$. Here, $I_{\mathrm{OC}} \equiv \mathrm{E}\left\{-\ddot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\}$. This means,
  - The observed-data asymptotic variance can be obtained by inflating the complete-data asymptotic variance by the factor $(1 - DM)^{-1}$. So observed-data variance is larger (less information in the data).
  - Smaller missingness → smaller DM → less variance inflation and faster convergence.

SEM algorithm

SEM consists of three steps

1. The evaluation of I_{OC}
2. The evaluation of DM
3. The evaluation of V

Evaluation of $I_{\text{OC}} \equiv \mathrm{E}\left\{-\ddot{l}_{\text{C}}(\hat{\theta}|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}\right\}$

- Example 1 (Grouped Multinomial): $l_{\text{C}}(\theta|X) = \ (x_1 + y_4)\log\theta + (y_2 + y_3)\log(1-\theta)$.
- Example 2 (Normal Mixtures): $l_{\text{C}}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{p}|x, y) = \ \sum_{ij} y_{ij}\left\{\log p_j + \log\phi(x_i|\mu_j, \sigma_j)\right\}$
- $f(Y_{\text{obs}}, Y_{\text{mis}})$ belongs to exponential family: $l_{\text{C}}(\theta|X) = S(X)'\eta(\theta) - B(\theta)$

The l_{C}, $\dot{l}_{\text{C}}$ and $\ddot{l}_{\text{C}}$ are linear functions of x_1, $\sum_i y_{ij}$ and $S(X)$ (sufficient statistics).

Recall that we evaluate $\mathrm{E}(\text{sufficient statistics}|Y_{\text{obs}}, \theta^{(k)})$ at every E step.

We easily obtain I_{OC} by plugging in $\mathrm{E}(\text{sufficient statistics}|Y_{\text{obs}}, \hat{\theta})$ at the last E step, no additional coding.

SEM algorithm (continued)

Evaluation of $DM = \{r_{ij}\}$

For a scalar θ, we can use the sequence $\theta^{(k)}$ to obtain DM.

For a vector θ, we cannot do so, because $\theta_i^{(k+1)} - \hat{\theta}_i \approx \sum_j DM_{ij}(\theta_j^{(k)} - \hat{\theta}_j)$.

Each DM_{ij} is the component-wise rate of convergence of the following “forced EM”

1. Run EM to get the MLE $\hat{\theta}$
2. Pick a starting point, $\theta^{(0)}$, some small distance from $\hat{\theta}$ but not equal to $\hat{\theta}$ in any component
3. Repeat the following until $r_{ij}^{(k)}$ is stable
   (a) Calculate $\theta^{(k)} = M(\theta^{(k-1)})$ using one step of EM
   (b) For each $i = 1, \ldots, p$,
      i. Let $\theta^{(k)}(i) = (\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \theta_i^{(k)}, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_p)$ (Replace the ith element of $\hat{\theta}$ with the ith element of $\theta^{(k)}$)
      ii. Perform one step of EM on $\theta^{(k)}(i)$ to obtain $M[\theta^{(k)}(i)]$
      iii. Obtain $r_{ij}^{(k)} = \left\{M_j[\theta^{(k)}(i)] - \hat{\theta}_j\right\} / \left\{\theta_i^{(k)} - \hat{\theta}_i\right\}$ for $j = 1, \ldots, p$

SEM algorithm (continued)

Note:

- The MLE $\hat{\theta}$ should be obtained at very low tolerance (e.g., $\epsilon = 10^{-12}$)
- The final r_{ij} is taken to be the first value of $r_{ij}^{(k)}$ satisfying $|r_{ij}^{(k)} - r_{ij}^{(k-1)}| < \epsilon$, where k can be different for different (i, j).

MCEM algorithm

EM algorithm: the analytical integration of the likelihood required for the E-step can be difficult.

Monte **C**arlo **EM** algorithm (Wei & Tanner, 1990) replaces the analytical integration in the E-step by a Monte Carlo integration procedure with MCMC sampling techniques such as the Gibbs or the Metropolis Hastings algorithm.

- **MCE step:** Simulate a sample $Y_{\text{mis},1}, \ldots, Y_{\text{mis},m}$ from $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})$ and calculate

$$h^{(k)}(\theta) = m^{-1} \sum_{j=1}^{m} l_{\text{C}}(\theta|Y_{\text{obs}}, Y_{\text{mis},j})$$

- **M step:** $\theta^{(k+1)} = \arg\max_\theta h^{(k)}(\theta)$

Choose m to guarantee convergence

Wei & Tanner recommend starting with small value of m and then increasing m as $\theta^{(k)}$ moves closer to the true maximizer.

Bayesian EM

Suppose we have prior $\pi(\theta)$ and wish to find the mode of

$$\text{log posterior} = l_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) + \log \pi(\theta).$$

- **E step**: $h^{(k)}(\theta) \equiv \mathrm{E}\left\{ l_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}) + \log \pi(\theta) \middle| Y_{\mathrm{obs}}, \theta^{(k)} \right\}$
$= \mathrm{E}\left\{ l_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}) \middle| Y_{\mathrm{obs}}, \theta^{(k)} \right\} + \log \pi(\theta)$
- **M step**: $\theta^{(k+1)} = \arg\max_{\theta} h^{(k)}(\theta)$

Bayesian EM: Example

Observed data:

$$(y_1, y_2, y_3) \sim \text{multinomial}\left\{n;\ \frac{2+\theta}{4}, \frac{1-\theta}{2}, \frac{\theta}{4}\right\}$$

Complete data:

$$(x_0, x_1, y_2, y_3) \sim \text{multinomial}\left\{n;\ \frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{2}, \frac{\theta}{4}\right\}$$

where $x_0 + x_1 = y_1$.

	No Prior	$\theta \sim$ Beta$(\nu_1, \nu_2) : \pi(\theta) = \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)}\theta^{(\nu_1-1)}(1-\theta)^{(\nu_2-1)}$
$l_C(\theta \vert x_0, x_1, y_2, y_3)$	$(x_1 + y_3)\log\theta + y_2\log(1-\theta)$	$(x_1 + y_3 + \nu_1 - 1)\log\theta + (y_2 + \nu_2 - 1)\log(1-\theta)$
$\omega_{12}^{(k)} = \mathrm{E}(x_1 \vert \theta^{(k)}, y_1)$	$\theta^{(k)}y_1/(\theta^{(k)}+2)$	Same as left
$\theta^{(k+1)}$	$(\omega_{12}^{(k)} + y_3)/(\omega_{12}^{(k)} + y_2 + y_3)$	$(\omega_{12}^{(k)} + y_3 + \nu_1 - 1)/(\omega_{12}^{(k)} + y_2 + y_3 + \nu_1 + \nu_2 - 2)$

Generalized EM algorithm

Generalized EM (GEM)

- **E step**: evaluate $h^{(k)}(\theta)$ as before
- **M step**: Choose $\theta^{(k+1)}$ such that $h^{(k)}(\theta^{(k+1)}) \geq h^{(k)}(\theta^{(k)})$
  (do not necessarily maximize $h^{(k)}(\theta)$, just increase it.)

Note: This retains the ascent property of EM.

EM gradient algorithm (Lange, 1995): a class of GEM

- **M step**: Do one step of Newton-Raphson:

$$\theta^{(k+1)} = \theta^{(k)} + \alpha^{(k)} d^{(k)}, \text{ where } d^{(k)} = -\left\{ \frac{\partial^2 h^{(k)}(\theta)}{\partial\theta\partial\theta'} \right\}^{-1} \left\{ \frac{\partial h^{(k)}(\theta)}{\partial\theta} \right\} \Bigg|_{\theta=\theta^{(k)}}.$$

  Start with $\alpha^{(k)} = 1$; do step-halving until $h^{(k)}(\theta^{(k+1)}) \geq h^{(k)}(\theta^{(k)})$

Lange pointed out that one step Newton-Raphson saves us from performing iterations within iterations and yet still displays the same local rate of convergence as a full EM algorithm that maximizes $h^{(k)}(\theta)$ at each iteration.

ECM algorithm

EM is unattractive if maximizing complete-data log likelihood $h^{(k)}(\theta)$ is complicated.

In many cases, maximizing $h^{(k)}(\theta)$ is relatively simple when conditional on some of the parameters being estimated.

Expectation **C**onditional **M**aximization **algorithm** (Meng & Rubin, 1993) replaces a (complicated) M step with a sequence of conditional maximization (CM) steps.

- **E step:** evaluate $h^{(k)}(\theta)$ as before
- **CM step:** Partition θ into T parts: $\theta = (\theta_1, \ldots, \theta_T)$. For $t = 1, \ldots, T$, obtain

$$\theta_t^{(k+1)} = \arg\max_{\theta_t} h^{(k)}(\theta_1^{(k+1)}, \ldots, \theta_{t-1}^{(k+1)}, \theta_t, \theta_{t+1}^{(k)}, \ldots, \theta_T^{(k)})$$

Note:

- Sharing all appealing convergence properties of EM, such as ascent property
- Typically need more E- and M- iterations but can be faster in total computer time.

ECM algorithm: example

Complete data:

$$y_1, \ldots, y_n \sim \text{gamma}(\alpha, \beta) \;\text{ with density }\; f(y|\alpha,\beta) = \frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

Observed data:

$$y_\text{O} = \;\text{ censoring of the complete data}$$

Complete-data log-likelihood

$$l_\text{C}(\alpha,\beta|y_1,\ldots,y_n) = (\alpha-1)\sum_i \log y_i - \sum_i y_i/\beta - n\Big\{\alpha \log\beta + \log\Gamma(\alpha)\Big\}$$

Define $\bar{y} = n^{-1}\sum_i y_i$, $\bar{g} = n^{-1}\sum_i \log y_i$, and $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$

E step:

$$\omega^{(k)} \equiv \mathrm{E}(\bar{y}|y_{\mathrm{O}}, \alpha^{(k)}, \beta^{(k)})$$
$$\tau^{(k)} \equiv \mathrm{E}(\bar{g}|y_{\mathrm{O}}, \alpha^{(k)}, \beta^{(k)})$$

Now, directly maximizing the Q-function is difficult, because of the $\alpha \log \beta$ term. We can, however, update α and β one by one, assuming the other is know.

CM steps

$$\text{Given } \alpha^{(k)}, \ \beta^{(k+1)} = \omega^{(k)}/\alpha^{(k)}$$
$$\text{Given } \beta^{(k+1)}, \ \alpha^{(k+1)} = \psi^{-1}(\tau^{(k)} - \log \beta^{(k+1)})$$

ECME algorithm

ECM Either **algorithm** (Liu & Rubin, 1994) is a generalization of the ECM algorithm. It replaces some CM-steps of ECM, which maximize the conditional expected complete-data log likelihood, with steps that maximize the corresponding conditional observed-data log likelihood.

- **E step:** evaluate $h^{(k)}(\theta)$ as before
- **CM step:** Partition θ into T parts: $\theta = (\theta_1, \ldots, \theta_T)$.
  For $t = 1, \ldots, T$, obtain **either**

$$\theta_t^{(k+1)} = \arg\max_{\theta_t} h^{(k)}(\theta_1^{(k+1)}, \ldots, \theta_{t-1}^{(k+1)}, \theta_t, \theta_{t+1}^{(k)}, \ldots, \theta_T^{(k)})$$

or

$$\theta_t^{(k+1)} = \arg\max_{\theta_t} l_O(\theta_1^{(k+1)}, \ldots, \theta_{t-1}^{(k+1)}, \theta_t, \theta_{t+1}^{(k)}, \ldots, \theta_T^{(k)} \mid Y_{\text{obs}})$$

Note:

- Share with both EM and ECM their stable monotone convergence and simplicity of implementation
- Converge substantially faster than either EM or ECM, measured by either the number of iterations or actual computer time.

Example: Mixed-effects model

For a longitudinal dataset of $i = 1, \ldots, N$ subjects, each with $t = 1, \ldots, n_i$ measurements of the response, a simple linear mixed effect model is given by

$$Y_{it} = X_i\beta + b_i + \epsilon_{it}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_i \sim N_{n_i}(0, \sigma_\epsilon^2 I_{n_i}), \quad b_i, \epsilon_i \text{ independent}$$

Observed-data log-likelihood

$$l(\beta, \sigma_b^2, \sigma_\epsilon^2 | Y_1, \ldots, Y_N) \equiv \sum_i \left\{ -\frac{1}{2}(Y_i - X_i\beta)' \Sigma_i^{-1} (Y_i - X_i\beta) - \frac{1}{2} \log |\Sigma_i| \right\},$$

where $\{\Sigma_i\}_{tt} = \sigma_b^2 + \sigma_\epsilon^2$ and $\{\Sigma_i\}_{tt'} = \sigma_b^2$ for $t' \neq t$.

- In fact, this likelihood can be directly maximized for $(\beta, \sigma_b^2, \sigma_\epsilon^2)$ by using Newton-Raphson or Fisher scoring.
- **Note:** Given $(\sigma_b^2, \sigma_\epsilon^2)$ and hence Σ_i, we obtain β that maximizes the likelihood by solving

$$\frac{\partial l(\beta, \sigma_b^2, \sigma_\epsilon^2 | Y_1, \ldots, Y_N)}{\partial \beta} = \sum_i X_i' \Sigma_i^{-1} (Y_i - X_i\beta) = 0,$$

$$\Rightarrow \quad \beta = \left(\sum_{i=1}^N X_i' \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^N X_i' \Sigma_i^{-1} Y_i.$$

Example: Mixed-effects model (continued)

Complete-data log-likelihood: b_i are treated as missing data

Let $\epsilon_i = Y_i - X_i\beta - b_i$. We know that

$$\begin{pmatrix} b_i \\ \epsilon_i \end{pmatrix} = N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_\epsilon^2 I_{n_i} \end{pmatrix}\right\}$$

$$l_C(\beta, \sigma_b^2, \sigma_\epsilon^2 | \epsilon_1, \ldots, \epsilon_N, b_1, \ldots, b_N) \equiv \sum_i \left\{ -\frac{1}{2\sigma_b^2} b_i^2 - \frac{1}{2}\log\sigma_b^2 - \frac{1}{2\sigma_\epsilon^2}\epsilon_i'\epsilon_i - \frac{n_i}{2}\log\sigma_\epsilon^2 \right\}$$

The parameter that maximizes the l_C is obtained as, given the complete data

$$\sigma_b^2 = N^{-1}\sum_{i=1}^N b_i^2$$

$$\sigma_\epsilon^2 = \left(\sum_{i=1}^N n_i\right)^{-1} \sum_{i=1}^N \epsilon_i'\epsilon_i$$

$$\beta = \left(\sum_{i=1}^N X_i'X_i\right)^{-1} \sum_{i=1}^N X_i'(Y_i - b_i).$$

Example: Mixed-effects model (continued)

E step: to evaluate

$$\begin{aligned}
&\mathrm{E}\left(b_i^2 \mid Y_i, \beta^{(k)}, \sigma_b^{2(k)}, \sigma_\epsilon^{2(k)}\right)\\
&\mathrm{E}\left(\epsilon_i'\epsilon \mid Y_i, \beta^{(k)}, \sigma_b^{(k)}, \sigma_\epsilon^{2(k)}\right)\\
&\mathrm{E}\left(b_i \mid Y_i, \beta^{(k)}, \sigma_b^{(k)}, \sigma_\epsilon^{2(k)}\right)
\end{aligned}$$

We use the relationship

$$\mathrm{E}(b_i^2 \mid Y_i) = \{\mathrm{E}(b_i \mid Y_i)\}^2 + \mathrm{Var}(b_i \mid Y_i).$$

Thus we need to calculate $\mathrm{E}(b_i \mid Y_i)$ and $\mathrm{Var}(b_i \mid Y_i)$. Recall the conditional distribution for multivariate normal variables

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} = N\left\{\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 e_{n_i} e_{n_i}' + \sigma_\epsilon^2 I_{n_i} & \sigma_b^2 e_{n_i} \\ \sigma_b^2 e_{n_i}' & \sigma_b^2 \end{pmatrix}\right\}, \qquad e_{n_i}' = (1, 1, \ldots, 1)$$

Let $\Sigma_i = \sigma_b^2 e_{n_i} e_{n_i}' + \sigma_\epsilon^2 I_{n_i}$. We known that

$$\begin{aligned}
\mathrm{E}(b_i \mid Y_i) &= 0 + \sigma_b^2 e_{n_i}' \Sigma_i^{-1}(Y_i - X_i\beta)\\
\mathrm{Var}(b_i \mid Y_i) &= \sigma_b^2 - \sigma_b^2 e_{n_i}' \Sigma_i^{-1} \sigma_b^2 e_{n_i}.
\end{aligned}$$

Example: Mixed-effects model (continued)

Similarly, We use the relationship

$$\mathrm{E}(\epsilon_i'\epsilon_i \mid Y_i) = \mathrm{E}(\epsilon_i' \mid Y_i)\mathrm{E}(\epsilon_i \mid Y_i) + \mathrm{Var}(\epsilon_i \mid Y_i).$$

We can derive

$$\begin{pmatrix} Y_i \\ \epsilon_i \end{pmatrix} = N\left\{\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 e_{n_i}e_{n_i}' + \sigma_\epsilon^2 I_{n_i} & \sigma_\epsilon^2 I_{n_i} \\ \sigma_\epsilon^2 I_{n_i} & \sigma_\epsilon^2 I_{n_i} \end{pmatrix}\right\}.$$

Let $\Sigma_i = \sigma_b^2 e_{n_i}e_{n_i}' + \sigma_\epsilon^2 I_{n_i}$. Then we have

$$\mathrm{E}(\epsilon_i \mid Y_i) = 0 + \sigma_\epsilon^2\Sigma_i^{-1}(Y_i - X_i\beta)$$
$$\mathrm{Var}(\epsilon_i \mid Y_i) = \sigma_\epsilon^2 I_{n_i} - \sigma_\epsilon^4\Sigma_i^{-1}.$$

M step of standard EM algorithm

$$\sigma_b^{2(k+1)} = N^{-1}\sum_{i=1}^{N}\mathrm{E}(b_i^2 \mid Y_i, \beta^{(k)}, \sigma_b^{2(k)}, \sigma_\epsilon^{2(k)})) \tag{1}$$

$$\sigma_\epsilon^{2(k+1)} = \left(\sum_{i=1}^{N} n_i\right)^{-1}\sum_{i=1}^{N}\mathrm{E}(\epsilon_i'\epsilon_i \mid Y_i, \beta^{(k)}, \sigma_b^{2(k)}, \sigma_\epsilon^{2(k)}) \tag{2}$$

$$\beta^{(k+1)} = \left(\sum_{i=1}^{N} X_i'X_i\right)^{-1}\sum_{i=1}^{N} X_i'\mathrm{E}(Y_i - b_i \mid Y_i, \beta^{(k)}, \sigma_b^{2(k)}, \sigma_\epsilon^{2(k)}). \tag{3}$$

ECME algorithm: Mixed-effects model (continued)

M step of ECME algorithm

- Partition the parameter vector $(\beta, \sigma_b^2, \sigma_\epsilon^2)$ as β and $(\sigma_b^2, \sigma_\epsilon^2)$
- First maximize complete-data log-likelihood over $(\sigma_b^2, \sigma_\epsilon^2)$, given by (1) and (2)
- Given $(\sigma_b^{2(k+1)}, \sigma_\epsilon^{2(k+1)})$, we can calculate $\Sigma_i^{(k+1)} = \sigma_b^{2(k+1)} e_{n_i} e'_{n_i} + \sigma_\epsilon^{2(k+1)} I_{n_i}$ and obtain β that maximizes the **observed**-data log likelihood

$$\beta^{(k+1)} = \left(\sum_{i=1}^{N} X_i' \left\{ \Sigma_i^{(k+1)} \right\}^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i' \left\{ \Sigma_i^{(k+1)} \right\}^{-1} Y_i.$$

Accelerated EM

Aitken's acceleration method (Louis, 1982)

Suppose $\theta^{(k)} \to \hat{\theta}$, as $k \to \infty$. Then

$$\hat{\theta} = \theta^{(k)} + \sum_{h=0}^{\infty} \left[\theta^{(k+h+1)} - \theta^{(k+h)}\right].$$

Now

$$\begin{aligned} \theta^{(k+h+2)} - \theta^{(k+h+1)} &= M(\theta^{(k+h+1)}) - M(\theta^{(k+h)}) \\ &\approx J(\theta^{(k+h)}) \left[\theta^{(k+h+1)} - \theta^{(k+h)}\right] \\ &\approx J(\theta^{(k)}) \left[\theta^{(k+h+1)} - \theta^{(k+h)}\right] \\ &\approx \left\{J(\theta^{(k)})\right\}^{h+1} \left[\theta^{(k+1)} - \theta^{(k)}\right] \end{aligned}$$

M : mapping defined by EM

J : Jabobian of M

Thus

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

by which we can produce the effect of an infinite number of iterations by the following algorithm

Accelerated EM

The algorithm:

1. From $\theta^{(k)}$, produce $\theta^{(k+1)}$ using EM
2. Estimate $\left(I - J(\theta^{(k)})\right)^{-1}$ by $\left(I - \hat{J}\right)^{-1}$ (see below)
3. Compute $\theta_*^{(k+1)} = \theta^{(k)} + \left(I - \hat{J}\right)^{-1}\left[\theta^{(k+1)} - \theta^{(k)}\right]$
4. Use $\theta_*^{(k+1)}$ in step 1.

Louis (1982) showed

$$\left(I - \hat{J}\right)^{-1} = I_{\mathrm{OC}}\left(I_{\mathrm{O}}\right)^{-1}$$

where $I_{\mathrm{OC}} = \mathrm{E}\left\{-\ddot{l}_{\mathrm{C}}(\hat{\theta}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\}$ and I_{O} can be obtained by the Louis formula.

References

- Dempster, A.P., Laird, N., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1–38.
- Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society B*, 57, 425–437.
- Liu, C. and Rubin, D.B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633–648.
- Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society B*, 44, 98–130.
- Meng, X. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899–909.
- Meng, X. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267–278.
- Wei, G. C. G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor mans data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699–704.