
Extensions to EM

2012-9-20

Yijuan Hu

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

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

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

EM algorithm:

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

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

Issues:

1. Standard errors: Louis estimator.
2. Maximizing $h^{(k)}(\theta)$
3. Slow convergence

- **EM algorithm** does not generate asymptotic covariance matrix (standard errors) for parameters as a byproduct.

- **Louis estimator** for covariance matrix, i.e.,

$$E \left\{ -\ddot{l}_C(\hat{\theta}|Y_{\text{obs}}, Y_{\text{mis}}) | Y_{\text{obs}} \right\} = E \left\{ \left[\dot{l}_C(\hat{\theta}|Y_{\text{obs}}, Y_{\text{mis}}) \right]^{\otimes 2} | Y_{\text{obs}} \right\} + \left[\dot{l}_O(\hat{\theta}|Y_{\text{obs}}) \right]^{\otimes 2}$$

requires calculation of the conditional expectation of the square of the complete-data score function, which is specific to each problem.

- **Supplemented 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.

- EM defines a mapping, $M : \theta^{(k+1)} = M(\theta^{(k)})$, where $M(\theta) = (M_1(\theta), \dots, M_p(\theta))$
- 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 $\theta^{(k+1)} - \hat{\theta} \approx DM(\theta^{(k)} - \hat{\theta})$.

- The asymptotic covariance matrix for $\hat{\theta}$, denoted as V , can be found as $\{-\ddot{l}_O(\hat{\theta}|Y_{\text{obs}})\}^{-1}$. However, the derivations can be difficult to evaluate directly.
- In contrast, $-\ddot{l}_C(\hat{\theta}|Y_{\text{obs}}, Y_{\text{mis}})$, $I_{OC} \equiv E\{-\ddot{l}_C(\hat{\theta}|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}\}$ are easy to evaluate.
- It has been shown that

$$V = I_{OC}^{-1}(I - DM)^{-1},$$

which means, the observed-data asymptotic variance can be obtained by inflating the complete-data asymptotic variance by the factor $(1 - DM)$.

SEM consists of three parts

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

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

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

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

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

We easily obtain I_{OC} by substituting $E(\text{sufficient statistics} | Y_{obs}, \hat{\theta})$ at the last E step

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 $i = 1, \dots, p$,
 - i. Let $\theta^{(k)}(i) = (\hat{\theta}_1, \dots, \hat{\theta}_{i-1}, \theta_i^{(k)}, \hat{\theta}_{i+1}, \dots, \hat{\theta}_p)$ (Replace the i th element of $\hat{\theta}$ with the i th element of $\theta^{(k)}$)
 - ii. Perform one step of EM on $\theta^{(k)}(i)$ to obtain $M[\theta^{(k)}(i)]$
 - iii. Obtain $r_{ij}^{(k)} = \{M_j[\theta^{(k)}(i)] - \hat{\theta}_j\} / \{\theta_i^{(k)} - \hat{\theta}_i\}$ for $j = 1, \dots, p$

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) .
- When all elements in the i th row of DM are stable, there is no need to repeat step (b)

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

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

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

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	$\theta \sim \text{Beta}(\nu_1, \nu_2) : \pi(\theta) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \theta^{(\nu_1 - 1)} (1 - \theta)^{(\nu_2 - 1)}$
$l_C(\theta x_0, x_1, y_2, y_3)$	$(x_1 + y_3) \log \theta + y_2 \log(1 - \theta)$	$(x_1 + y_3 + \nu_1 - 1) \log \theta + (y_2 + \nu_2 - 1) \log(1 - \theta)$
$\omega_{12}^{(k)} = E(x_1 \theta^{(k)}, y_1)$	$\theta^{(k)} y_1 / (\theta^{(k)} + 2)$	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 (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:

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

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.

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 Conditional Maximization 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: $\theta = (\theta_1, \dots, \theta_T)$. For $t = 1, \dots, T$, obtain

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

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.

Complete data:

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

Observed data:

$$y_O = \text{censoring of the complete data}$$

Complete-data log-likelihood

$$l_C(\alpha, \beta | y_1, \dots, y_n) = (\alpha - 1) \sum_i \log y_i - \sum_i y_i / \beta - n \{ \alpha \log \beta + \log \Gamma(\alpha) \}$$

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:**

$$\omega^{(k)} \equiv E(\bar{y} | y_O, \alpha^{(k)}, \beta^{(k)})$$

$$\tau^{(k)} \equiv E(\bar{g} | y_O, \alpha^{(k)}, \beta^{(k)})$$

• **CM steps**

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

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

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

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

For $t = 1, \dots, T$, obtain **either**

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

or

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

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.

For a longitudinal dataset of $i = 1, \dots, N$ subjects, each with $t = 1, \dots, n_i$ 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

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

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

$$\begin{aligned} \frac{\partial l(\beta, \sigma_b^2, \sigma_\epsilon^2 | Y_1, \dots, Y_N)}{\partial \beta} &= \sum_i X_i' \Sigma_i^{-1} (Y_i - X_i\beta) = 0, \\ \Rightarrow \beta &= \left(\sum_{i=1}^N X_i' \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^N X_i' \Sigma_i^{-1} Y_i. \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} \sigma_b^2 &= N^{-1} \sum_{i=1}^N b_i^2 \\ \sigma_\epsilon^2 &= \left(\sum_{i=1}^N n_i \right)^{-1} \sum_{i=1}^N \epsilon_i' \epsilon_i \\ \beta &= \left(\sum_{i=1}^N X_i' X_i \right)^{-1} \sum_{i=1}^N X_i' (Y_i - b_i). \end{aligned}$$

E step: to evaluate

$$E\left(b_i^2 \mid Y_i, \beta^{(k)}, \sigma_b^{2(k)}, \sigma_\epsilon^{2(k)}\right)$$

$$E\left(\epsilon_i' \epsilon \mid Y_i, \beta^{(k)}, \sigma_b^{(k)}, \sigma_\epsilon^{2(k)}\right)$$

$$E\left(b_i \mid Y_i, \beta^{(k)}, \sigma_b^{(k)}, \sigma_\epsilon^{2(k)}\right)$$

We use the relationship

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

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

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

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

$$E(b_i \mid Y_i) = 0 + \sigma_b^2 e_{n_i}' \Sigma_i^{-1} (Y_i - X_i \beta)$$

$$\text{Var}(b_i \mid Y_i) = \sigma_b^2 - \sigma_b^2 e_{n_i}' \Sigma_i^{-1} \sigma_b^2 e_{n_i}.$$

Similarly, We use the relationship

$$E(\epsilon'_i \epsilon_i | Y_i) = E(\epsilon'_i | Y_i)E(\epsilon_i | Y_i) + \text{Var}(\epsilon_i | Y_i).$$

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

$$\begin{aligned} E(\epsilon_i | Y_i) &= 0 + \sigma_\epsilon^2 \Sigma_i^{-1} (Y_i - X_i \beta) \\ \text{Var}(\epsilon_i | Y_i) &= \sigma_\epsilon^2 I_{n_i} - \sigma_\epsilon^4 \Sigma_i^{-1}. \end{aligned}$$

M step of standard EM algorithm

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

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

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

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

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

Convergence accelerated: In some of ECME's M steps, the observed-data likelihood is being conditionally maximized, rather than an approximation to it (expected complete-data log likelihood).

Aitken's acceleration method (Louis, 1982)

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

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

Now

$$\begin{aligned} \theta^{(k+h+2)} - \theta^{(k+h+1)} &= M(\theta^{(k+h+1)}) - M(\theta^{(k+h)}) && M : \text{mapping defined by EM} \\ &\approx J(\theta^{(k+h)}) [\theta^{(k+h+1)} - \theta^{(k+h)}] && J : \text{Jacobian of } M \\ &\approx J(\theta^{(k)}) [\theta^{(k+h+1)} - \theta^{(k+h)}] \\ &\approx \{J(\theta^{(k)})\}^{h+1} [\theta^{(k+1)} - \theta^{(k)}] \end{aligned}$$

Thus

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

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

The algorithm:

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

Louis (1982) showed

$$(I - \hat{J})^{-1} = I_{OC} (I_O)^{-1}$$

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

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

Monte Carlo 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 $Y_{\text{mis},1}, \dots, Y_{\text{mis},m}$ from $f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})$ and calculate

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

- **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.

-
- 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.