Chapter 1. The EM Algorithm

1 Introduction

The expectation maximization, EM, algorithm for finding maximum likelihood estimates (MLE's) is a powerful numerical technique useful in contexts ranging from standard incomplete data problems (e.g. censored and truncated), to iteratively reweighted least squares analysis and empirical Bayes models. Several analyses of its properties with examples and cautions for its use have appeared in the literature (See Dempster, Laird and Rubin, 1977, and the references thereof).

The primary conceptual power of this iterative algorithm lies in converting a maximization problem involving a complicated likelihood, into a sequence of **pseudo-complete** problems, where at each step the updated parameters can be obtained in a closed form (or at least in a straightforward manner). Unlike Newton-Raphson or Fletcher-Powell techniques, no gradients or curvature matrices need to be derived.

2 The ML Problem and EM Algorithm

Suppose we have a model for the complete data $\mathbf{x}$, with associated density $f(\mathbf{x}|\theta)$, where $\theta \in \Theta \subset R^p$ is the unknown parameter and $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{X}$. We write $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis})$, where $\mathbf{x}_{obs}$ represents the observed part of $\mathbf{x}$ and $\mathbf{x}_{mis}$ denotes the missing values. The objective of the EM algorithm is to find the MLE for θ based on the observed data $\mathbf{x}_{obs}$. Instead of observing $\mathbf{x}$, one observes the value of a measurable function $Y(\mathbf{x}) = \mathbf{x}_{obs} = y \in Y$. The MLE $\hat{\theta}$ is to be found using the data y.

The EM method is only attractive in situations where finding the complete data MLE and either the observed or the expected information matrix would be straightforward, but the problem based on the incomplete data Y requires an iterative solution. Typically the complete data are from an exponential family. The EM algorithm starts with an initial estimate $\theta^{(0)}$. Letting $\theta^{(v)}$ be the estimate of θ at the vth iteration, iteration $(v+1)$ of EM is as follows.

E step: Find the expected complete-data log-likelihood f θ were $\theta^{(v)}$:

$$\lambda^*(\mathbf{y}, \theta|\theta^{(v)}) = \log\{f(y|\theta)\} = \log\left\{\int_R f_{\mathbf{X}}(\mathbf{x}|\theta^{(v)})d\mu(\mathbf{x})\right\}, \tag{1}$$

where $R = \{\mathbf{x} : y(\mathbf{x}) = y\}$, and $\mu(\mathbf{x})$ is a dominating measure.

M step: Determine $\theta^{(v+1)}$ by maximizing this expected log- likelihood:

$$\lambda^*(y, \theta^{(v+1)}|\theta^{(v)}) \geq \lambda^*(y, \theta|\theta^{(v)}), \quad \text{for all } \theta.$$

Let

$$\begin{aligned} \lambda(\mathbf{x}, \theta) &= \log\{f(\mathbf{x}|\theta)\} \\ \lambda^*(y, \theta) &= \log\{f(y|\theta)\} = \log\left\{\int_R f_{\mathbf{X}}(\mathbf{x}|\theta)d\mu(\mathbf{x})\right\}, \end{aligned}$$

Now, instead of maximizig λ^* directly, the EM algorithm proceeds by using an initial estimate $\theta^{(0)}$ and solving the pseudo-complete data problem:

$$\text{maximize}_{\theta \in \Theta} E_{\theta^{(0)}}[\lambda(\mathbf{X}, \theta)|\mathbf{X} \in R].$$

The maximizing value for this pseudo-complete data problem is called $\theta^{(1)}$ and the iteration is continued until $||\theta^{(v+1)} - \theta^{(v)}||$ is sufficiently small or some other convergence criterion is satified. The M step is standard maximum likelihood estimation for complete-data problems, and the E step is usually available from standard complete-data theory of conditional distributions. The EM algorithm just described implicitly defines a mapping $\theta \to M(\theta)$ from the parameter space Θ to itself such that

$$\theta^{(v+1)} = M(\theta^{(v)}), \quad \text{for } v = 0, 1, \ldots.$$

3 Examples of EM

3.1 A Multinormial Example

This example has frequently appeared in the literature of the EM algorithm since it was first used in DLR to introduce EM. Rao (1965, pp. 368-369) presents data in which 197 animals are distributed multinomially into four categories, so that the observed data consist of

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

A genetic model for the population specifies a probability model for the complete data $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)$ is a multinomial distribution with cell probabilities

$$(1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4), \quad 0 \leq \theta \leq 1.$$

Based on this theory, we can model $\mathbf{y}$ as a four-category multinomial population where the cell probabilities are

$$(1/2 + \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4), \quad 0 \leq \theta \leq 1.$$

Thus

$$f(\mathbf{y}|\theta) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{\theta}{4})^{y_1}(\frac{1}{4}-\frac{\theta}{4})^{y_2}(\frac{1}{4}-\frac{\theta}{4})^{y_3}(\frac{\theta}{4})^{y_4}.$$

The objective here is to find the ML estimate for θ based on the observed counts $\mathbf{y}$.

Suppose that if the complete data $\mathbf{x}$ were observed. Then

$$f(\mathbf{x}|\theta) = \frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!x_2!x_3!x_4!x_5!}(\frac{1}{2})^{x_1}(\frac{\theta}{4})^{x_2}(\frac{1}{4}-\frac{\theta}{4})^{x_3}(\frac{1}{4}-\frac{\theta}{4})^{x_4}(\frac{\theta}{4})^{x_5}. \tag{2}$$

Note that the integral (1) consists in this case of summing (2) over the (x_1, x_2) pairs $(0, 125)$, $(1, 124), \ldots, (125, 0)$, while simply substituting $(18, 20, 34)$ for (x_3, x_4, x_5). The MLE of θ based on $\mathbf{x}$ would be immediate:

$$\theta^* = (x_2 + x_5)/(x_2 + x_3 + x_4 + x_5). \tag{3}$$

Also, note that the log-likelihood $L(\theta|\mathbf{x})$ is linear in $\mathbf{x}$, so in the E step we only need to replace the missing values by their corresponding conditional expectations. Or, estimate x_1 and x_2 when $x_1 + x_2 = y_1 = 125$. Thus at the $(v+1)$th iteration, we have for the E step

$$x_2^{(v)} = 125\frac{\theta^{(v)}/4}{1/2 + \theta^{(v)}/4}; \tag{4}$$

and for the M step, from (3), we have

$$\theta^{(v+1)} = (x_2^{(v)} + 34)/(x_2^{(v)} + 34 + 18 + 20). \tag{5}$$

The EM algorithm for this example is defined by cycling back and forth between (4) and (5). Formulas (4) and (5) together define the mapping $\theta^{(v+1)} = M(\theta^{(v)})$. In fact, by substituting $x_2^{(v)}$ from (4) into (5), and letting $\theta^{(v+1)} = \theta^{(v)} = \theta^*$, we can explicitly solve a quadratic equation for the MLE of θ:

$$\theta^* = (15 + \sqrt{53,809})/394 \cong 0.6268214980.$$

Starting from an initial value of $\theta^{(0)} = 0.5$, the algorithm moved for eight steps as following:

$$0.608247423, 0.624321051, 0.626488879, 0.626777323, 0.62677323, 0.626815632, 0.626820719,$$

0.626821395, 0.626821484.

Since the complete-data density is

$$f(\mathbf{x}|\theta) \propto \theta^{x_2+x_5}(1-\theta)^{x_3+x_4},$$

the complete-data variance for $\theta - \theta^*$ is simply the ordinary binomial variance $\theta(1-\theta)/n$, where $n = x_2 + x_3 + x_4 + x_5$. Thus, the asymptotic standard error of $\theta - \theta^*$ based on the observed data $\mathbf{y}$ is

$$\left[\frac{\theta^*(1-\theta^*)}{n^*(1-r)}\right]^{1/2} \cong 0.051,$$

where $n^* = x_2^* + 72 = 101.83$ is the expectation of n given the MLE of θ. Here $\theta^*(1-\theta^*)/n^*$ is the observed Fisher information of complete data and r is the limit of $] \ (\theta^{(v+1)} - \theta^*)/(\theta^{(v)} - \theta^*)$. Details on the asymptotic standard error will be given in Section 4.

3.2 A Univariate Contaminated Normal Example

Finite mixtures are a common modeling technique. Suppose that an observable $\mathbf{y}$ is represented as n observations $\mathbf{y} = (y_1, \ldots, y_n)$. Suppose further that there exists a finite set of R states, and that each y_i is associated with an unobserved state. Thus, there exists an unobserved vector $\mathbf{q} = (q_1, \ldots, q_n)$, where q_i is the indicator vector of length R whose components are all zero except for one equal to unity indicating the unobserved state associated with y_i. Defining the complete data to be $\mathbf{x} = (\mathbf{y}, \mathbf{q})$.

A natural way to conceptualize mixture specifications is to think of the marginal distribution of the indicators $\mathbf{q}$, and then to specify the distribution of $\mathbf{y}$ given $\mathbf{q}$. We then assume there is a set of R densities $u(\ldots|r, \phi)$ for $r = (1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)$ such that the y_i given q_i are conditionally independent with densities $u(\ldots|q_i, \phi)$. Finally, denoting

$$U(y_i|\phi) = (\log u(y_i|(1, 0, \ldots, 0), \phi), \log u(y_i|(0, 1, \ldots, 0), \phi) \ldots, \log u(y_i|(0, \ldots, 1), \phi))$$

and

$$V(\phi) = (\log v((1, 0, \ldots, 0)|\phi), \log v((0, 1, \ldots, 0)|\phi), \ldots, \log v((0, 0, \ldots, 1)|\phi)),$$

we can express the complete-data log-likelihood as

$$\log f(\mathbf{x}|\phi) = \sum_{i=1}^{n} \mathbf{q}_i U(y_i|\phi) + \sum_{i=1}^{n} \mathbf{q}_i^T V(\phi) \tag{6}$$

Since the complete-data log-likelihood is linear in the components of each $\mathbf{q}_i$, the E-step of the EM algorithm requires us to estimate the components of $\mathbf{q}_i$ given the observed $\mathbf{y}$ and the current fitted parameters. These estimated components of $\mathbf{q}_i$ are simply the current conditional probabilities that y_i belongs to each of the R states. In many examples, the ϕ parameters of $u(\ldots|\phi)$ and $v\ldots|\phi)$ are unrelated, so that the first and the second terms in (6) may be maximized separately. The M-step is then equivalent to the complete-data maximization for the problem except that each observation y_i contributes to the log-likelihood associated with each of the R states, with weights given by the R estimated components of $\mathbf{q}_i$, and the counts in the R states are the sums of the estimated components of the $\mathbf{q}_i$.

We now use the case of equal, known variances to illustrate the above algorithm. Data $Y_1, \ldots, Y_n$ are observed and are known to be i.i.d. from a mixture of two normal distributions. Assuming $\sigma^2 = 1$, the density of Y is

$$f(y|\mu_0, \mu_1, \pi) = (1-\pi)\phi(y-\mu_0) + \pi\phi(y-\mu_1), \quad -\infty < \mu_0, \mu_1 < \infty, \quad 0 \leq \pi \leq 1/2.$$

Let

$$\begin{aligned} h(q) &= \pi, \qquad \text{if } q = (0,1), \\ &= 1-\pi, \qquad \text{if } q = (1,0). \end{aligned}$$

Then $\mathbf{x} = (x_1, \ldots, x_n)$ can be considered as an independent sample from a population such that

$$x_i|\theta, q_i \overset{ind}{\sim} N(q_i^T(\mu_0, \mu_1), 1),$$

where the q_i are unobserved iid random variables with density $h(q_i)$.

Implementation of the EM algorithm is straighforward. Let

$$\begin{aligned} w_i &= w(y_i, \mu_0, \mu_1, \pi) = E(q_i|y_i, \mu_0, \mu_1, \pi) \\ &= \frac{\pi\phi(y_i - \mu_1)}{\pi\phi(y_i - \mu_1) + (1-\pi)\phi(y_i - \mu_0)}, \end{aligned}$$

and

$$w_i^{(v)} = w(y_i, \mu_0^{(v)}, \mu_1^{(v)}, \pi^{(v)}).$$

Then the parameters are updated by

$$\begin{aligned}
\pi^{(v+1)} &= \frac{1}{n}\sum_{i=1}^{n} w_i^{(v)}, \\
\mu_0^{(v+1)} &= \{n(1-\pi^{(v+1)})\}^{-1}\sum_{i=1}^{n}(1-w_i^{(v)})y_i \\
\mu_1^{(v+1)} &= \{n\pi^{(v+1)}\}^{-1}\sum_{i=1}^{n} w_i^{(v)} y_i.
\end{aligned}$$

We now use the example in Little and Rubin (1987) to illustrate the idea of treating mixture models as incomplete data problems. Suppose $y_1, \ldots, y_n$ represent an independent sample from the univariate contaminated normal model

$$f(y|\mu,\sigma^2) = (1-\pi)N(\mu,\sigma^2) + \pi N(\mu,\sigma^2/\lambda), \tag{7}$$

where $0 < \pi < 1$ and $\lambda > 0$ are both unknown. As a contrast to the above setting, the means in these two normal distribution are the same. The objective here is to find the MLE $\theta^* = (\mu^*, \log\sigma^{2*})$ for $\theta = (\mu, \log\sigma^2)$.

This problem can be treated as an incomplete-data problem although there are no missing data in the usual sense. Let

$$\begin{aligned}
h(q) &= 1-\pi, \quad \text{if } q = 1, \\
&= \pi, \qquad \text{if } q = \lambda, \\
&= 0, \qquad \text{otherwise;}
\end{aligned}$$

then $\mathbf{x} = (x_1, \ldots, x_n)$ can be considered as an independent sample from a population such that

$$x_i|\theta, q_i \overset{ind}{\sim} N(\mu, \sigma^2/q_i),$$

where the q_i are unobserved iid random variables with known density $h(q_i)$. In this setting we can apply the EM algorithm to compute θ^*, treating $\mathbf{Q} = (q_1, \ldots, q_n)$ as the missing data $\mathbf{x}_{mis}$, $\mathbf{Y}$ as the observed data $\mathbf{x}_{obs}$, and $(\mathbf{Y}, \mathbf{Q})$ as the complete data $\mathbf{X}$.

Implementing EM is quite straightforward. Since the complete-data log-likelihood is linear in the q_i, the E step reduces to finding the conditional expectation

$$w_i^{(v+1)} = w_i(\theta^{(v+1)}) = E(q_i|y_i, \theta^{(v)}), \tag{8}$$

where $\theta^{(v)}$ is the current estimate of θ. A simple application of Bayes's theorem yields

$$w_i(\theta) = E(q_i|y_i, \theta)$$

$$= \frac{1-\pi+\pi\lambda^{3/2}\exp\{(1-\lambda)z_i^2/2\}}{1-\pi+\pi\lambda^{1/2}\exp\{(1-\lambda)z_i^2/2\}}, \tag{9}$$

where $z_i^2 = (x_i-\mu)^2/\sigma^2$. The M step is also trivial since the complete-data MLE of $\theta = (\mu, \log\mu^2)$ is in closed form:

$$\mu^* = \sum_{i=1}^n q_i y_i / \sum_{i=1}^n q_i \tag{10}$$

and

$$\log\sigma^{2*} = \log\left\{\frac{1}{n}\sum_{i=1}^n q_i(y_i-\mu^*)^2\right\}. \tag{11}$$

Therefore the new estimate $\theta^{(v+1)} = (\mu^{(v+1)}, \log\sigma^{2(v+1)})$ Can be easily obtained from (10) and (11) with q_i replaced by $w_i^{(v)}$ from (8) and (9).

Now we illustrate the SEM algorithm proposed in Meng (1991) for finding the variance of the above estimate. A simulation is performed with $\mu = 0$, $\sigma^2 = 1$, $\lambda = 2$, $\pi = 0.1$, and the sample size 100. We first define r_{ij} which is the (i,j)th element of DM. Define $\theta^{(v)}(i)$ to be

$$\theta^{(v)}(i) = (\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(v)}, \theta_{i+1}^*, \ldots, \theta_d^*),$$

that is, only the ith component in $\theta^{(v)}(i)$ is active in the sense that the other components are fixed at their MLE's. Then

$$\begin{aligned} r_{ij} &= \frac{\partial M_j(\theta^*)}{\partial\theta_i} \\ &= \lim_{\theta_i\to\theta_i^*} \frac{M_j(\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i, \theta_{i+1}^*, \ldots, \theta_d^*) - M_j(\theta_i^*)}{\theta_i - \theta_i^*} \\ &= \lim_{t\to\infty} \frac{M_j(\theta_i^{(v)}) - \theta_i^*}{\theta_i^{(v)} - \theta_i^*} = \lim_{t\to\infty} r_{ij}^{(v)} \end{aligned}$$

3.3 Factor Analysis

In our final class of examples, interest focuses on the dependence of p observed variables on $q < p$ unobserved "latent" variables or "factors." When both sets of variables are continuous and the observed variables are assumed to have a linear regression on the factors, the model is commonly called factor analysis. Our discussion using the EM algorithm applies when the variables are normally distributed.

More specifically, let $\mathbf{y}$ be the $n \times p$ observed data matrix and $\mathbf{z}$ be the $n \times q$ unobserved factor-score matrix. Then $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, where the rows of $\mathbf{x}$ are independently and identically distributed. The marginal distribution of each row of $\mathbf{z}$ is normal with mean $(0, \ldots, 0)$, variance $(1, \ldots, 1)$ and correlation $\mathbf{R}$. The conditional distribution of the ith row of $\mathbf{y}$ given $\mathbf{z}$ is normal with mean $\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{z}_i$ and residual covariance $\boldsymbol{\tau}^2 = diag(\tau_1^2, \ldots, \tau_p^2)$, where $\mathbf{z}_i$ is the ith row of $\mathbf{z}$. Note that given the factors the variables are independent. The parameters $\boldsymbol{\phi}$ thus consist of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}^2$. The regression coefficient matrix $\boldsymbol{\beta}$ is commonly called the factor-loading matrix and the residual variances $\boldsymbol{\tau}^2$ are commonly called the uniquenesses.

Two cases are defined by further restrictions on $\boldsymbol{\beta}$ and/or $\mathbf{R}$. In the first case, $\boldsymbol{\beta}$ is unrestricted and $\mathbf{R} = \mathbf{I}$. In the second case, restrictions are placed on $\boldsymbol{\beta}$ (a *priori* zeroes), and the requirement that $\mathbf{R} = \mathbf{I}$ is possibly relaxed so that some of the correlations among the factors are to be estimated. See Joreskog (1969) for examples and discussion of these models. It is sometimes desirable to place a prior distribution on the uniquenesses to avoid the occurrence of zero estimates.

If the factors were observed, the computation of the maximum-likelihood estimates of $\boldsymbol{\phi}$ would follow from the usual least-squares computations based on the sums, sums of squares, and sum of cross-products of $\mathbf{x}$. Let $(\bar{\mathbf{y}}, \bar{\mathbf{z}})$ be the sample mean vector and

$$\begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{yz} \\ \mathbf{C}_{zy} & \mathbf{C}_{zz} \end{bmatrix}$$

be the sample cross-products matrix of $\mathbf{x}$. Then the maximum-likelihood estimate of $\boldsymbol{\alpha}$ is simply $\bar{\mathbf{y}}$ while the maximum-likelihood estimates of the factor loadings and uniqueness for the jth variable follow from the regression of that variable on the factors. Note that the calculations of these parameters may involve different sets of factors for different observed variables reflecting that a *prior* zeroes in $\boldsymbol{\beta}$. The matrix $\mathbf{R}$ is estimated from $\mathbf{C}_{zz}$ (and $\bar{\mathbf{z}}$); if restrictions are placed on $\mathbf{R}$, special complete-data maximum likelihood techniques may have

to be used. We have thus described the M-step of the algorithm, namely, the computation of the maximum-likelihood estimate of ϕ from complete data.

The E-step of the algorithm requires us to calculate the expected value of $\mathbf{C}_{zz}$ and $\mathbf{C}_{zy}$ given the current estimated ϕ ($\bar{\mathbf{z}}$ is always estimated as $\mathbf{o}$). The computation is again a standard least-squares computation: we estimate the regression coefficients of the factors on the variables assuming the current estimated ϕ found from the M-step. Thus the resultant EM-algorithm consists of back and forth least-squares calculations on the cross-products matrix of all variables.

4 Discussion

Suppose we study multiple linear regression with missing values. An obvious practical approach is to guess at suitable values for the missing quantities and then to analyze the data as if they were complete. Typically, these guessed values are called *imputed values* if the guessing is done objectively. Now the natural approach is to replace each missing quantity by its mean value conditional on the values of all known variables.

Finally, we consider a procedure for extracting the observed information matrix when the EM algorithm is used to find maximum likelihood estimates in incomplete data problems. Let $S(\mathbf{x},\theta)$ and $S^*(\mathbf{y},\theta)$ be the gradient vectors of $\lambda(\mathbf{x},\theta)$ and $\lambda^*(\mathbf{y},\theta)$ respectively and $B(\mathbf{x},\theta)$ and $B^*(\mathbf{y},\theta)$ be the negatives of the associated second derivatives. Then, by straightforward differentiation,

$$S^*(\mathbf{y},\theta) = \int_R f^{(1)}(\mathbf{x}|\theta)d\mu(\mathbf{x}) / \int_{\mathbf{R}} \mathbf{f}(\mathbf{x}|\theta)\mathbf{d}\mu(\mathbf{x}).$$

Multiply and divide the integrand in the numerator by $f(\mathbf{x}|\theta)$, we have

$$S^*(\mathbf{y},\theta) = E_\theta \left[S(\mathbf{x},\theta)|\mathbf{x}\in R\right]. \tag{12}$$

Recall that $S^*(\mathbf{y},\theta^*) = 0$. Taking additional derivative, we have

$$-B^*(\mathbf{y},\theta) = \frac{\int_R f^{(2)}(\mathbf{x}|\theta)d\mu(\mathbf{x})}{\int_{\mathbf{R}} \mathbf{f}(\mathbf{x}|\theta)\mathbf{d}\mu(\mathbf{x})} - S^*(\mathbf{y},\theta)S^{*T}(\mathbf{y},\theta).$$

By multiplying and dividing by $f(\mathbf{x}|\theta)$ as before we obtain

$$-B^*(\mathbf{y},\theta) = E_\theta \left[f^{(2)}_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{x}|\theta)\right] - S^*(\mathbf{y},\theta)S^{*T}(\mathbf{y},\theta),$$

which can be rewritten as

$$B^*(\mathbf{y},\theta) = E_\theta\left[B(\mathbf{x},\theta)|\mathbf{x}\in R\right] - E_\theta\left[S(\mathbf{x},\theta)S^T(\mathbf{x},\theta)|\mathbf{x}\in R\right] + S^*(\mathbf{y},\theta^*)S^{*T}(\mathbf{y},\theta^*). \quad (13)$$

Efron and Hinkley (1978) define $B^*(\mathbf{y},\theta)$ as the observed information and show that in most case it is a more appropriate measure of information than the *a priori* expectation $E_\theta B^*(\mathbf{Y},\theta)$. It is certainly easier to compute. The first term in the right hand side of (13) is the conditional expected full data observed information matrix, while the last two produce the expected information for the conditional distribution of $\mathbf{x}$ given $\mathbf{x}\in R$. The last term does not need to be evaluated since $S^*(\mathbf{y},\theta^*)=0$.

For the first normal mixture model, we now use the above derivation to get the covariance matrix of obtained estimate. Note that

$$\begin{aligned}
\lambda(\mathbf{x},\theta) &= q\log(\pi) + (1-q)\log(1-\pi) + q\log(\phi(X-\mu_1)) + (1-q)\log(\phi(X-\mu_1)) \\
S(Y_i,\theta) &= (1-q_i)(Y_i-\mu_0) \\
&= q_i(Y_i-\mu_1) \\
&= \frac{q_i}{\pi} - \frac{1-q_i}{1-\pi} \\
B(\mathbf{x},\theta) &= \begin{pmatrix} 1-q_i & 0 & 0 \\ 0 & q_i & 0 \\ 0 & 0 & q_i/\pi^2 + (1-q_i)/(1-\pi)^2 \end{pmatrix}, \\
S^*(X_i,\theta) &= \begin{pmatrix} (1-w_i)(X_i-\mu_0) \\ w_i(X_i-\mu_1) \\ \frac{w_i}{\pi} - \frac{1-w_i}{1-\pi} \end{pmatrix}.
\end{aligned}$$

To see how this procedure works, 500 observations were generated using $\mu_0=2$, $\mu_1=0$ and $\pi=0.5$. From these data, we have $\hat{\mu}_0=1.970$, $\hat{\mu}_1=-0.042$, $\hat{\pi}=0.483$,

$$I_Y = \begin{pmatrix} 241.7 & 0 & 0 \\ 0 & 258.3 & 0 \\ 0 & 0 & 2002.2 \end{pmatrix}$$

and I_Y is symmetric with lower triangle:

$$\begin{pmatrix} 136.9 & 0 & 0 \\ -20.6 & 159.7 & 0 \\ 218.5 & 219.1 & 1131.2 \end{pmatrix}.$$

I_Y can be invereted to find the covariance matrix of $(\hat{\mu}_0,\hat{\mu}_1,\hat{\pi})$.

At the end of Section 3.1, the asymptotic standard error of $\theta - \theta^*$ based on the observed data $\mathbf{y}$ is given as

$$\left[\frac{\theta^*(1-\theta^*)}{n^*(1-r)}\right]^{1/2} \cong 0.051,$$

where $n^* = x_2^* + 72 = 101.83$ is the expectation of n given the MLE of θ. Here r is the limit of the ratios of successive deviations which are the limit of $(\theta^{(v+1)} - \theta^*)/(\theta^{(v)} - \theta^*)$. We have $r = 0.1328$ which seems to be the limit of $.1465, .1346, .1330, .1328, .1328, .1328, .1328, \ldots$. Recall that the EM algorithm just defines a mapping $\theta \to M(\theta)$ from the parameter space of θ, Θ, to itself such that

$$\theta^{(v+1)} \to M(\theta^{(v)}).$$

If $\theta^{(v)}$ converges to some point θ^* and $M(\theta)$ is continuous, then θ^* must satisfy

$$\theta^* = M(\theta^*).$$

Therefore, in the neighborhood of θ^*, by a Taylor series expansion, we have

$$\theta^{(v+1)} - \theta^* \approx (\theta^{(v)} - \theta^*) M^{(1)}(\theta^*).$$

Thus in the neighborhood of θ^*, the EM algorithm is essentially a linear iteration with rate $M^{(1)}(\theta^*)$.

We now derive that the large sample variance matrix of $\theta - \theta^*$ based on $\mathbf{y}$, V, can be found as the inverse of the observed information matrix

$$V = I_o^{-1}(\theta^*|\mathbf{y}), \tag{14}$$

where $I_o(\theta^*|\mathbf{y})$ is the negative second derivative of the log-likelihood of θ given $\mathbf{y}$,

$$I_o(\theta^*|\mathbf{y}) = -\frac{d^2 \log f(\mathbf{y}|\theta)}{d\theta^2}. \tag{15}$$

Suppose that θ^* is a consistent estimate of θ. Recall that $S^*(\mathbf{y}, \theta^*) = 0$. Using the mean value theorem, we have

$$\begin{aligned} 0 &= S^*(\mathbf{y}, \theta^*) = S^*(\mathbf{y}, \theta) + (\theta^* - \theta)(-B^*(\mathbf{y}, \theta^{**})), \\ \sqrt{n}(\theta^* - \theta) &= \frac{S^*(\mathbf{y}, \theta)/\sqrt{n}}{n^{-1}B^*(\mathbf{y}, \theta^{**})}, \end{aligned}$$

where θ^{**} lies between θ and θ^*. By the law of large numbers,

$$\frac{B^*(\mathbf{y}, \theta^{**})}{n} \to I(\theta).$$

Observe that $E(\ln f(y|\theta))^{(1)} = 0$ and by the central limit theorem

$$\frac{1}{\sqrt{n}} S^*(\mathbf{y}, \theta^{**}) \rightarrow N(0, E\{(\ln f(y|\theta))^{(1)}\}^2.$$

Note that

$$\begin{aligned} E(\ln f(y|\theta))^{(2)} &= \int \frac{f^{(2)}(y|\theta) - [f(y|\theta)]^2}{f^2(y|\theta)} f(y|\theta) dy \\ &= \int f^{(2)}(y|\theta) dy - E\{(\ln f(y|\theta))^{(1)}\}^2 = -I(\theta). \end{aligned}$$

This function can be very difficult to evaluate directly. In contrast, the complete-data observed information matrix,

$$I_o(\theta^*|\mathbf{x}) = -\frac{d^2 \log f(\mathbf{x}|\theta)}{d\theta^2}, \tag{16}$$

whose inverse gives the complete-data variance, is often a simple function, as is its expectation over the conditional distribution $f(\mathbf{x}_{mis}|\mathbf{y}, \theta)$ evaluated at $\theta = \theta^*$:

$$I_{oc} = E[I_o(\theta^*|\mathbf{x})|\mathbf{y}, \theta]|_{\theta=\theta^*}. \tag{17}$$

From the factorization

$$f(\mathbf{x}|\theta) = f(\mathbf{y}|\theta) f(\mathbf{x}_{mis}|\mathbf{y}, \theta),$$

it follows that the log-likelihood of θ given $\mathbf{y}$ is

$$\lambda(\theta|\mathbf{y}) = \lambda(\theta|\mathbf{x}) - \log f(\mathbf{x}_{mis}|\mathbf{y}, \theta). \tag{18}$$

Equation (18) implies, after taking second derivatives, averaging over $f(\mathbf{x}_{mis}|\mathbf{y}, \theta)$, and evaluating at $\theta = \theta^*$, that

$$I_o(\theta^*|\mathbf{y}) = I_{oc} - I_{om}, \tag{19}$$

where the matrix

$$I_{om} = E\left[-\frac{d^2 \log f(\mathbf{x}_{mis}|\mathbf{y}, \theta)}{d\theta^2}|\mathbf{y}, \theta\right]\Big|_{\theta=\theta^*}, \tag{20}$$

can be viewed as the missing information.

Equation (19) can be written as

$$I_o(\theta^*|\mathbf{y}) = (1 - I_{om} I_{oc}^{-1}) I_{oc}. \tag{21}$$

Equation (21) is useful because Dempster et al.(1977) showed that $M^{(1)}(\theta^*) = I_{om} I_{oc}^{-1}$. Then we have

$$V = I_{oc}^{-1}(1 - M^{(1)}(\theta^*))^{-1}.$$

We now compute $M^{(1)}(\theta^*)$ for the multinomial example presented in Section 3. Note that

$$\frac{d^2 \log f(\mathbf{x}|\theta)}{d\theta^2} = -\frac{x_2 + x_5}{\theta^2} + \frac{x_3 + x_4}{(1-\theta)^2}.$$

Hence

$$\begin{aligned} I_{oc} &= E[I_o(\theta^*|\mathbf{x})|\mathbf{y}, \theta]|_{\theta=\theta^*} \\ &= E[-\frac{x_2 + x_5}{\theta^{*2}} + \frac{x_3 + x_4}{(1-\theta)^{*2}}|\mathbf{y}, \theta^*] \\ &= -[E(x_2|\theta^*, \mathbf{y}) + y_4]/\theta^{*2} - (y_2 + y_3)/(1-\theta^*)^2. \end{aligned}$$

Note that the conditional distribution of X_2 given Y_1 is binomially distributed with parameter $\theta/(2+\theta)$ and hence

$$\frac{d^2 \log f(\mathbf{x}_{mis}|\mathbf{y}, \theta)}{d\theta^2} = -\frac{x_2}{\theta^2} + \frac{y_1}{(2+\theta)^2}.$$

We have

$$I_{om} = -E(x_2|\theta^*, \mathbf{y})/\theta^{*2} + y_1/(2+\theta^*)^2.$$

Substituting θ^* as given in Section 3, we find $M^{(1)}(\theta^*) = I_{om} I_{oc}^{-1} = 0.1327778$ which is close to 0.1328 derived before.