Statistics 580

The EM Algorithm

Introduction

The EM algorithm is a very general iterative algorithm for parameter estimation by maximum likelihood when some of the random variables involved are not observed i.e., considered missing or incomplete. The EM algorithm formalizes an intuitive idea for obtaining parameter estimates when some of the data are missing:

i. replace missing values by estimated values,

ii. estimate parameters.

iii. Repeat

- step (i) using estimated parameter values as true values, and
- step (ii) using estimated values as "observed" values, iterating until convergence.

This idea has been in use for many years before Orchard and Woodbury (1972) in their *missing information principle* provided the theoretical foundation of the underlying idea. The term EM was introduced in Dempster, Laird, and Rubin (1977) where proof of general results about the behavior of the algorithm was first given as well as a large number of applications.

For this discussion, let us suppose that we have a random vector $\mathbf{y}$ whose joint density $f(\mathbf{y};\boldsymbol{\theta})$ is indexed by a p-dimensional parameter $\boldsymbol{\theta} \in \Theta \subseteq R^p$. If the *complete-data* vector $\mathbf{y}$ were observed, it is of interest to compute the maximum likelihood estimate of $\boldsymbol{\theta}$ based on the distribution of $\mathbf{y}$. The log-likelihood function of $\mathbf{y}$

$$\log L(\boldsymbol{\theta};\mathbf{y}) = \ell(\boldsymbol{\theta};\mathbf{y}) = \log f(\mathbf{y};\boldsymbol{\theta}),$$

is then required to be maximized.

In the presence of missing data, however, only a function of the *complete-data* vector $\mathbf{y}$, is observed. We will denote this by expressing $\mathbf{y}$ as $(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, where $\mathbf{y}_{\text{obs}}$ denotes the observed but "incomplete" data and $\mathbf{y}_{\text{mis}}$ denotes the unobserved or "missing" data. For simplicity of description, assume that the missing data are missing at random (Rubin, 1976), so that

$$\begin{aligned} f(\mathbf{y};\boldsymbol{\theta}) &= f(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}};\boldsymbol{\theta}) \\ &= f_1(\mathbf{y}_{\text{obs}};\boldsymbol{\theta}) \cdot f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\boldsymbol{\theta}), \end{aligned}$$

where f_1 is the joint density of $\mathbf{y}_{\text{obs}}$ and f_2 is the joint density of $\mathbf{y}_{\text{mis}}$ given the observed data $\mathbf{y}_{\text{obs}}$, respectively. Thus it follows that

$$\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}}) = \ell(\boldsymbol{\theta};\mathbf{y}) - \log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\boldsymbol{\theta}),$$

where $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ is the observed-data log-likelihood.

EM algorithm is useful when maximizing ℓ_{obs} can be difficult but maximizing the complete-data log-likelihood ℓ is simple. However, since $\mathbf{y}$ is not observed, ℓ cannot be evaluated and hence maximized. The EM algorithm attempts to maximize $\ell(\boldsymbol{\theta}; \mathbf{y})$ iteratively, by replacing it by its conditional expectation given the observed data $\mathbf{y}_{\text{obs}}$. This expectation is computed with respect to the distribution of the complete-data evaluated at the current estimate of $\boldsymbol{\theta}$. More specifically, if $\boldsymbol{\theta}^{(0)}$ is an initial value for $\boldsymbol{\theta}$, then on the first iteration it is required to compute

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}) = E_{\boldsymbol{\theta}^{(0)}}\left[\ell(\boldsymbol{\theta}; \mathbf{y}) | \mathbf{y}_{\text{obs}}\right].$$

$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)})$ is now maximized with respect to $\boldsymbol{\theta}$, that is, $\boldsymbol{\theta}^{(1)}$ is found such that

$$Q(\boldsymbol{\theta}^{(1)}; \boldsymbol{\theta}^{(0)}) \geq Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)})$$

for all $\boldsymbol{\theta} \in \Theta$. Thus the EM algorithm consists of an E-step (Estimation step) followed by an M-step (Maximization step) defined as follows:

E-step: Compute $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ where

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}\left[\ell(\boldsymbol{\theta}; \mathbf{y}) | \mathbf{y}_{\text{obs}}\right].$$

M-step: Find $\boldsymbol{\theta}^{(t+1)}$ in Θ such that

$$Q(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$$

for all $\boldsymbol{\theta} \in \Theta$.

The E-step and the M-step are repeated alternately until the difference $L(\boldsymbol{\theta}^{(t+1)}) - L(\boldsymbol{\theta}^{(t)})$ is less than δ, where δ is a prescribed small quantity.

The computation of these two steps simplify a great deal when it can be shown that the log-likelihood is linear in the sufficient statistic for $\boldsymbol{\theta}$. In particular, this turns out to be the case when the distribution of the complete-data vector (i.e., $\mathbf{y}$) belongs to the exponential family. In this case, the E-step reduces to computing the expectation of the complete-data sufficient statistic given the observed data. When the complete-data are from the exponential family, the M-step also simplifies. The M-step involves maximizing the expected log-likelihood computed in the E-step. In the exponential family case, actually maximizing the expected log-likelihood to obtain the next iterate can be avoided. Instead, the conditional expectations of the sufficient statistics computed in the E-step can be directly substituted for the sufficient statistics that occur in the expressions obtained for the complete-data maximum likelihood estimators of $\boldsymbol{\theta}$, to obtain the next iterate. Several examples are discussed below to illustrate these steps in the exponential family case.

As a general algorithm available for complex maximum likelihood computations, the EM algorithm has several appealing properties relative to other iterative algorithms such as Newton-Raphson. First, it is typically easily implemented because it relies on complete-data computations: the E-step of each iteration only involves taking expectations over complete-data conditional distributions. The M-step of each iteration only requires complete-data maximum likelihood estimation, for which simple closed form expressions are already available. Secondly, it is numerically stable: each iteration is required to increase the log-likelihood $\ell(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ in each iteration, and if $\ell(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ is bounded, the sequence $\ell(\boldsymbol{\theta}^{(t)};\mathbf{y}_{\text{obs}})$ converges to a stationery value. If the sequence $\boldsymbol{\theta}^{(t)}$ converges, it does so to a local maximum or saddle point of $\ell(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ and to the unique MLE if $\ell(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ is unimodal. A disadvantage of EM is that its rate of convergence can be extremely slow if a lot of data are missing: Dempster, Laird, and Rubin (1977) show that convergence is linear with rate proportional to the fraction of information about $\boldsymbol{\theta}$ in $\ell(\boldsymbol{\theta};\mathbf{y})$ that is observed.

Example 1: Univariate Normal Sample

Let the complete-data vector $\mathbf{y}=(y_1,\ldots,y_n)^T$ be a random sample from $N(\mu,\sigma^2)$. Then

$$
\begin{aligned}
f(\mathbf{y};\mu,\sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}\right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-1/2\sigma^2\left(\sum y_i^2 - 2\mu\sum y_i + n\mu^2\right)\right\}
\end{aligned}
$$

which implies that $(\sum y_i,\, \sum y_i^2)$ are sufficient statistics for $\boldsymbol{\theta}=(\mu,\,\sigma^2)^T$. The complete-data log-likelihood function is:

$$
\begin{aligned}
\ell(\mu,\sigma^2;\mathbf{y}) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2} + \text{constant} \\
&= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}y_i - \frac{n\mu^2}{\sigma^2} + \text{constant}
\end{aligned}
$$

It follows that the log-likelihood based on complete-data is linear in complete-data sufficient statistics. Suppose $y_i, i=1,\ldots,m$ are observed and $y_i, i=m+1,\ldots,n$ are missing (at random) where y_i are assumed to be i.i.d. $N(\mu,\sigma^2)$. Denote the observed data vector by $\mathbf{y}_{\text{obs}}=(y_1,\ldots,y_m)^T$). Since the complete-data $\mathbf{y}$ is from the exponential family, the E-step requires the computation of

$$
E_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} y_i|\mathbf{y}_{\text{obs}}\right) \text{and}\, E_{\boldsymbol{\theta}}\left(\sum_{i=1}^{N} y_i^2|\mathbf{y}_{\text{obs}}\right),
$$

instead of computing the expectation of the complete-data log-likelihood function shown above. Thus, at the t^{th} iteration of the E-step, compute

$$
\begin{aligned}
s_1^{(t)} &= E_{\mu^{(t)},\sigma^{2(t)}}\left(\sum_{i=1}^{n} y_i|\mathbf{y}_{\text{obs}}\right) \qquad (1) \\
&= \sum_{i=1}^{m} y_i + (n-m)\,\mu^{(t)}
\end{aligned}
$$

since $E_{\mu^{(t)},\sigma^{2(t)}}(y_i)=\mu^{(t)}$ where $\mu^{(t)}$ and $\sigma^{2^{(t)}}$ are the current estimates of μ and σ^2, and

$$
\begin{aligned}
s_2^{(t)} &= E_{\mu^{(t)},\sigma^{2(t)}}\left(\sum_{i=1}^{n} y_i^2 | \mathbf{y}_{\text{obs}}\right) \qquad (2)\\
&= \sum_{i=1}^{m} y_i^2 + (n-m)\left[\sigma^{(t)^2} + \mu^{(t)^2}\right]
\end{aligned}
$$

since $E_{\mu^{(t)},\sigma^{2(t)}}(y_i^2) = \sigma^{2^{(t)}} + \mu^{(t)^2}$.

For the M-step, first note that the complete-data maximum likelihood estimates of μ and σ^2 are:

$$
\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2
$$

The M-step is defined by substituting the expectations computed in the E-step for the complete-data sufficient statistics on the right-hand side of the above expressions to obtain expressions for the new iterates of μ and σ^2 . Note that complete-data sufficient statistics themselves cannot be computed directly since $y_{m+1}, \ldots, y_n$ have not been observed. We get the expressions

$$
\mu^{(t+1)} = \frac{s_1^{(t)}}{n} \qquad (3)
$$

and

$$
\sigma^{2^{(t+1)}} = \frac{s_2^{(t)}}{n} - \mu^{(t+1)^2}. \qquad (4)
$$

Thus, the E-step involves computing evaluating (1) and (2) beginning with starting values $\mu^{(0)}$ and $\sigma^{2^{(0)}}$. M-step involves substituting these in (3) and (4) to calculate new values $\mu^{(1)}$ and $\sigma^{2^{(1)}}$, etc. Thus, the EM algorithm iterates successively between (1) and (2) and (3) and (4). Of course, in this example, it is not necessary to use of EM algorithm since the maximum likelihood estimates for (μ, σ^2) are clearly given by $\hat{\mu} = \sum_{i=1}^{m} y_i/m$ and $\hat{\sigma}^2 = \sum_{i=1}^{m} y_i^2/m - \hat{\mu}^2 \square$.

Example 2: Sampling from a Multinomial population

In the Example 1, "incomplete data" in effect was "missing data" in the conventional sense. However, in general, the EM algorithm applies to situations where the complete data may contain variables that are not observable by definition. In that set-up, the observed data can be viewed as some function or mapping from the space of the complete data.

The following example is used by Dempster, Laird and Rubin (1977) as an illustration of the EM algorithm. Let $\mathbf{y}_{\text{obs}} = (38, 34, 125)^T$ be observed counts from a multinomial population with probabilities: $(\frac{1}{2} - \frac{1}{2}\theta, \frac{1}{4}\theta, \frac{1}{2} + \frac{1}{4}\theta)$. The objective is to obtain the maximum likelihood estimate of θ. First, to put this into the framework of an incomplete data problem, define $\mathbf{y} = (y_1, y_2, y_3, y_4)^T$ with multinomial probabilities $(\frac{1}{2} - \frac{1}{2}\theta, \frac{1}{4}\theta, \frac{1}{4}\theta, \frac{1}{2}) \equiv (p_1, p_2, p_3, p_4)$. The $\mathbf{y}$ vector is considered complete-data. Then define $\mathbf{y}_{\text{obs}} = (y_1, y_2, y_3 + y_4)^T$. as the observed data vector, which is a function of the complete-data vector. Since only $y_3 + y_4$ is observed and y_3 and y_4 are not, the observed data is considered incomplete. However, this is not simply a missing data problem.

The complete-data log-likelihood is

$$\ell(\theta; \mathbf{y}) = y_1 \log p_1 + y_2 \log p_2 + y_3 \log p_3 + y_4 \log p_4 + \text{const.}$$

which is linear in y_1, y_2, y_3 and y_4 which are also the sufficient statistics. The E-step requires that $E_\theta(\mathbf{y}|\mathbf{y}_{\text{obs}})$ be computed; that is compute

$$\begin{aligned} E_\theta(y_1|\mathbf{y}_{\text{obs}}) &= y_1 = 38 \\ E_\theta(y_2|\mathbf{y}_{\text{obs}}) &= y_2 = 34 \\ E_\theta(y_3|\mathbf{y}_{\text{obs}}) &= E_\theta(y_3|y_3 + y_4) = 125(\frac{1}{4}\theta)/(\frac{1}{2} + \frac{1}{4}\theta) \end{aligned}$$

since, conditional on $(y_3 + y_4)$, y_3 is distributed as Binomial(125, p) where

$$p = \frac{\frac{1}{4}\theta}{\frac{1}{2} + \frac{1}{4}\theta}.$$

Similarly,

$$E_\theta(y_4|\mathbf{y}_{\text{obs}}) = E_\theta(y_4|y_3 + y_4) = 125(\frac{1}{2})/(\frac{1}{2} + \frac{1}{4}\theta),$$

which is similar to computing $E_\theta(y_3|\mathbf{y}_{\text{obs}})$. But only

$$y_3^{(t)} = E_{\theta^{(t)}}(y_3|\mathbf{y}_{\text{obs}}) = \frac{125(\frac{1}{4})\theta^{(t)}}{(\frac{1}{2} + \frac{1}{4}\theta^{(t)})} \tag{1}$$

needs to be computed at the t^{th} iteration of the E-step as seen below.

For the M-step, note that the complete-data maximum likelihood estimate of θ is

$$\frac{y_2 + y_3}{y_1 + y_2 + y_3}$$

(Note: Maximize

$$\ell(\theta; \mathbf{y}) = y_1 \log(\frac{1}{2} - \frac{1}{2}\theta) + y_2 \log \frac{1}{4}\theta + y_3 \log \frac{1}{4}\theta + y_4 \log \frac{1}{2}$$

and show that the above indeed is the maximum likelihood estimate of θ). Thus, substitute the expectations from the E-step for the sufficient statistics in the expression for maximum likelihood estimate θ above to get

$$\theta^{(t+1)} = \frac{34 + y_3^{(t)}}{72 + y_3^{(t)}}. \tag{2}$$

Iterations between (1) and (2) define the EM algorithm for this problem. The following table shows the convergence results of applying EM to this problem with $\theta^{(0)} = 0.50$.

Table 1. The EM Algorithm for Example 2 (from Little and Rubin (1987))

t	$\theta^{(t)}$	$\theta^{(t)} - \hat{\theta}$	$(\theta^{(t+1)} - \hat{\theta})/(\theta^{(t)} - \hat{\theta})$
0	0.500000000	0.126821498	0.1465
1	0.608247423	0.018574075	0.1346
2	0.624321051	0.002500447	0.1330
3	0.626488879	0.000332619	0.1328
4	0.626777323	0.000044176	0.1328
5	0.626815632	0.000005866	0.1328
6	0.626820719	0.000000779	.
7	0.626821395	0.000000104	.
8	0.626821484	0.000000014	.

Example 3: Sample from Binomial/ Poisson Mixture

The following table shows the number of children of N widows entitled to support from a certain pension fund.

Number of Children:	0	1	2	3	4	5	6
Observed # of Widows:	n_0	n_1	n_2	n_3	n_4	n_5	n_6

Since the actual data were not consistent with being a random sample from a Poisson distribution (the number of widows with no children being too large) the following alternative model was adopted. Assume that the discrete random variable is distributed as a mixture of two populations, thus:

Mixture of Populations:	Population A:	with probability ξ, the random variable takes the value 0, and
	Population B:	with probability $(1-\xi)$, the random variable follows a Poisson with mean λ

Let the observed vector of counts be $\mathbf{n}_{\text{obs}} = (n_0, n_1, \ldots, n_6)^T$. The problem is to obtain the maximum likelihood estimate of (λ, ξ). This is reformulated as an incomplete data problem by regarding the observed number of widows with no children be the sum of observations that come from each of the above two populations.

Define

$$\begin{aligned} n_0 &= n_A + n_B \\ n_A &= \# \text{ widows with no children from population A} \\ n_B &= n_o - n_A = \# \text{ widows with no children from population B} \end{aligned}$$

Now, the problem becomes an incomplete data problem because n_A is not observed. Let $\mathbf{n} = (n_A, n_B, n_1, n_2, \ldots, n_6)$ be the complete-data vector where we assume that n_A and n_B are observed and $n_0 = n_A + n_B$.

Then

$$\begin{aligned}
f(\mathbf{n};\xi,\lambda) &= k(\mathbf{n})\,\{P(y_0=0)\}^{n_0}\ \Pi_{i=1}^{\infty}\ \{P(y_i=i)\}^{n_i} \\
&= k(\mathbf{n})\left[\xi+(1-\xi)\,e^{-\lambda}\right]^{n_0}\left[\Pi_{i=1}^{6}\left\{(1-\xi)\,\frac{e^{-\lambda}\lambda^i}{i!}\right\}^{n_i}\right] \\
&= k(\mathbf{n})\left[\xi+(1-\xi)e^{-\lambda}\right]^{n_A+n_B}\left\{(1-\xi)e^{-\lambda}\right\}^{\sum_{i=1}^{6}n_i}\left[\Pi_{i=1}^{6}\left(\frac{\lambda^i}{i!}\right)^{n_i}\right].
\end{aligned}$$

where $k(\mathbf{n}) = \sum_{i=1}^{6} n_i/n_0!n_1!\ldots n_6!$. Obviously, the complete-data sufficient statistic is $(n_A, n_B, n_1, n_2, \ldots, n_6)$. The complete-data log-likelihood is

$$\begin{aligned}
\ell(\xi,\lambda;\mathbf{n}) &= n_0\log(\xi+(1-\xi)e^{-\lambda}) \\
&\quad +(N-n_0)\left[\log(1-\xi)-\lambda\right]+\sum_{i=1}^{6} i\,n_i\,\log\lambda + const.
\end{aligned}$$

Thus, the complete-data log-likelihood is linear in the sufficient statistic. The E-step requires the computing of

$$E_{\xi,\lambda}(\mathbf{n}|\mathbf{n}_{\text{obs}}).$$

This computation results in

$$E_{\xi,\lambda}(n_i|\mathbf{n}_{\text{obs}}) = n_i \quad \text{for} \quad i=1,...,6,$$

and

$$E_{\xi,\lambda}(n_A|\mathbf{n}_{\text{obs}}) = \frac{n_0\xi}{\xi+(1-\xi)\exp(-\lambda)},$$

since n_A is Binomial(n_0, p) with $p = \frac{p_A}{p_A+p_B}$ where $p_A = \xi$ and $p_B = (1-\xi)\,e^{-\lambda}$. The expression for $E_{\xi,\lambda}(n_B|\mathbf{n}_{\text{obs}})$ is equivalent to that for $E(n_A)$ and will not be needed for E-step computations. So the E-step consists of computing

$$n_A^{(t)} = \frac{n_0\xi^{(t)}}{\xi^{(t)}+(1-\xi^{(t)})\exp(-\lambda^{(t)})} \tag{1}$$

at the t^{th} iteration.

For the M-step, the complete-data maximum likelihood estimate of (ξ, λ) is needed. To obtain these, note that $n_A \sim \text{Bin}(N, \xi)$ and that $n_B, n_1, \ldots, n_6$ are observed counts for $i = 0, 1, \ldots, 6$ of a Poisson distribution with parameter λ. Thus, the complete-data maximum likelihood estimate's of ξ and λ are

$$\hat{\xi} \;=\; \frac{n_A}{N},$$

and

$$\hat{\lambda} = \sum_{i=1}^{6} \frac{i\, n_i}{n_B + \sum_{i=1}^{6} n_i}.$$

The M-step computes

$$\xi^{(t+1)} = \frac{n_A^{(t)}}{N} \tag{2}$$

and

$$\lambda^{(t+1)} = \sum_{i=1}^{6} \frac{i\, n_i}{n_B^{(t)} + \sum_{i=1}^{6} n_i} \tag{3}$$

where $n_B^{(t)} = n_0 - n_A^{(t)}$.

The EM algorithm consists of iterating between (1), and (2) and (3) successively. The following data are reproduced from Thisted(1988).

Number of children	0	1	2	3	4	5	6
Number of widows	3,062	587	284	103	33	4	2

Starting with $\xi^{(0)} = 0.75$ and $\lambda^{(0)} = 0.40$ the following results were obtained.

Table 2. EM Iterations for the Pension Data

t	ξ	λ	n_A	n_B
0	0.75	0.40	2502.779	559.221
1	0.614179	1.035478	2503.591	558.409
2	0.614378	1.036013	2504.219	557.781
3	0.614532	1.036427	2504.704	557.296
4	0.614651	1.036747	2505.079	556.921
5	0.614743	1.036995	2505.369	556.631

A single iteration produced estimates that are within 0.5% of the maximum likelihood estimate's and are comparable to the results after about four iterations of Newton-Raphson. However, the convergence rate of the subsequent iterations are very slow; more typical of the behavior of the EM algorithm.

Example 4: Variance Component Estimation (Little and Rubin(1987))

The following example is from Snedecor and Cochran (1967, p.290). In a study of artificial insemination of cows, semen samples from six randomly selected bulls were tested for their ability to produce conceptions. The number of samples tested varied from bull to bull and the response variable was the percentage of conceptions obtained from each sample. Here the interest is on the variability of the bull effects which is assumed to be a random effect. The data are:

Table 3. Data for Example 4 (from Snedecor and Cochran(1967))

Bull(i)	Percentages of Conception	n_i
1	46,31,37,62,30	5
2	70,59	2
3	52,44,57,40,67,64,70	7
4	47,21,70,46,14	5
5	42,64,50,69,77,81,87	7
6	35,68,59,38,57,76,57,29,60	9
Total		35

A common model used for analysis of such data is the oneway random effects model:

$$y_{ij} = a_i + \epsilon_{ij}, \quad j = 1, ..., n_i, \quad i = 1, ..., k;$$

where it is assumed that the bull effects a_i are distributed as i.i.d. $N(\mu, \sigma_a^2)$ and the within-bull effects (errors) ϵ_{ij} as i.i.d. $N(0, \sigma^2)$ random variables where a_i and ϵ_{ij} are independent. The standard oneway random effects analysis of variance is:

Source	d.f.	S.S.	M.S.	F	E(M.S.)
Bull	5	3322.059	664.41	2.68	$\sigma^2 + 5.67\sigma_a^2$
Error	29	7200.341	248.29		σ^2
Total	34	10522.400			

Equating observed and expected mean squares from the above gives $s^2 = 248.29$ as the estimate of σ^2 and (664.41 - 248.29)/5.67 = 73.39 as the estimate of σ_a^2.

To construct an EM algorithm to obtain MLE's of $\theta = (\mu, \sigma_a^2, \sigma^2)$, first consider the joint density of $\mathbf{y}^* = (\mathbf{y}, \mathbf{a})^T$ where $\mathbf{y}^*$ is assumed to be complete-data. This joint density can be written as a product of two factors: the part first corresponds to the joint density of y_{ij} given a_i and the second to the joint density of a_i.

$$\begin{aligned} f(\mathbf{y}^*; \boldsymbol{\theta}) &= f_1(\mathbf{y}|\mathbf{a}; \boldsymbol{\theta}) f_2(\mathbf{a}; \boldsymbol{\theta}) \\ &= \Pi_i \, \Pi_j \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_{ij}-a_i)^2} \right\} \Pi_i \left\{ \frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{1}{2\sigma_a^2}(a_i-\mu)^2} \right\} \end{aligned}$$

Thus, the log-likelihood is linear in the following complete-data sufficient statistics:

$$\begin{aligned} T_1 &= \sum a_i \\ T_2 &= \sum a_i^2 \\ T_3 &= \sum_i \sum_j (y_{ij} - a_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum_i n_i (\bar{y}_{i.} - a_i)^2 \end{aligned}$$

Here complete-data assumes that both $\mathbf{y}$ and $\mathbf{a}$ are available. Since only $\mathbf{y}$ is observed, let $\mathbf{y}^*_{\text{obs}} = \mathbf{y}$. Then the E-step of the EM algorithm requires the computation of the expectations of T_1, T_2 and T_3 given $\mathbf{y}^*_{\text{obs}}$, i.e., $E_{\boldsymbol{\theta}}(T_i|\mathbf{y})$ for $i = 1, 2, 3$. The conditional distribution of $\mathbf{a}$ given $\mathbf{y}$ is needed for computing these expectations. First, note that the joint distribution of $\mathbf{y}^* = (\mathbf{y},\, \mathbf{a})^T$ is $(N + k)$-dimensional multivariate normal: $N(\boldsymbol{\mu}^*, \Sigma^*)$ where $\boldsymbol{\mu}^* = (\boldsymbol{\mu},\ \boldsymbol{\mu}_a)^T,\ \boldsymbol{\mu} = \mu \mathbf{j}_N,\ \boldsymbol{\mu}_a = \mu \mathbf{j}_k$ and Σ^* is the $(N + k) \times (N + k)$ matrix

$$\Sigma^* = \begin{pmatrix} \Sigma & \Sigma_{12} \\ \Sigma_{12}^T & \sigma_a^2 I \end{pmatrix}.$$

Here

$$\Sigma = \begin{bmatrix} \Sigma_1 & & & 0 \\ & \Sigma_2 & & \\ & & \ddots & \\ 0 & & & \Sigma_k \end{bmatrix},\ \Sigma_{12} = \sigma_a^2 \begin{bmatrix} \mathbf{j}_{n_1} & & & 0 \\ & \mathbf{j}_{n_2} & & \\ & & \ddots & \\ 0 & & & \mathbf{j}_{n_k} \end{bmatrix}$$

where $\Sigma_i = \sigma^2 I_{n_i} + \sigma_a^2 J_{n_i}$ is an $n_i \times n_i$ matrix. The covariance matrix Σ of the joint distribution of $\mathbf{y}$ is obtained by recognizing that the y_{ij} are jointly normal with common mean μ and common variance $\sigma^2 + \sigma_a^2$ and covariance σ_a^2 within the same bull and 0 between bulls. That is

$$\begin{aligned} \text{Cov}(y_{ij},\, y_{i'j'}) &= \text{Cov}(a_i + \epsilon_{ij},\, a_{i'} + \epsilon_{i'j'}) \\ &= \sigma^2 + \sigma_a^2 \quad \text{if } i = i^{'}, j = j^{'}, \\ &= \sigma_a^2 \quad \text{if } i = i^{'}, j \neq j^{'}, \\ &= 0 \quad \text{if } i \neq i^{'}. \end{aligned}$$

Σ_{12} is covariance of $\mathbf{y}$ and $\mathbf{a}$ and follows from the fact that $\text{Cov}(y_{ij},\, a_i) = \sigma_a^2$ if $i = i^{'}$ and 0 if $i \neq i^{'}$. The inverse of Σ is needed for computation of the conditional distribution of $\mathbf{a}$ given $\mathbf{y}$ and obtained as

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_1^{-1} & & & 0 \\ & \Sigma_2^{-1} & & \\ & & \ddots & \\ 0 & & & \Sigma_k^{-1} \end{bmatrix}$$

where $\Sigma_i^{-1} = \frac{1}{\sigma^2}\left[I_{n_i} - \frac{\sigma_a^2}{\sigma^2 + n_i \sigma_a^2} J_{n_i}\right]$. Using a well-known theorem in multivariate normal theory, the distribution of $\mathbf{a}$ given $\mathbf{y}$ is given by $N(\boldsymbol{\alpha},\, A)$ where $\boldsymbol{\alpha} = \boldsymbol{\mu}_a + \Sigma_{12}^{'} \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})$ and $A = \sigma_a^2 I - \Sigma_{12}^{'} \Sigma^{-1} \Sigma_{12}$. It can be shown after some algebra that

$$a_i|\mathbf{y} \overset{i.i.d}{\sim} N\left(w_i \mu + (1 - w_i)\bar{y}_{i.},\, v_i\right)$$

where $w_i = \sigma^2/(\sigma^2 + n_i\sigma_a^2)$, $\bar{y}_{i.} = (\sum_{j=1}^{n_i} y_{ij})/n_i$, and $v_i = w_i\sigma_a^2$. Recall that this conditional distribution was derived so that the expectations of T_1, T_2 and T_3 given $\mathbf{y}$ (or $\mathbf{y}^*_{\text{obs}}$) can be computed. These now follow easily. Thus the tth iteration of the E-step is defined as

$$\begin{aligned}
T_1^{(t)} &= \sum \left[w_i^{(t)}\mu^{(t)} + (1 - w_i^{(t)})\bar{y}_{i.}\right] \\
T_2^{(t)} &= \sum \left[w_i^{(t)}\mu^{(t)} + (1 - w_i^{(t)})\bar{y}_{i.}\right]^2 + \sum v_i^{(t)} \\
T_3^{(t)} &= \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 + \sum_i n_i \left[w_i^{(t)^2}(\mu^{(t)} - \bar{y}_{i.})^2 + v_i^{(t)}\right]
\end{aligned}$$

Since the complete-data maximum likelihood estimates are

$$\begin{aligned}
\hat{\mu} &= \frac{T_1}{k} \\
\hat{\sigma}_a^2 &= \frac{T_2}{k} - \hat{\mu}^2
\end{aligned}$$

and

$$\hat{\sigma}^2 = \frac{T_3}{N},$$

the M-step is thus obtained by substituting the expectations for the sufficient statistics calculated in the E-step in the expressions for the maximum likelihood estimates:

$$\begin{aligned}
\mu^{(t+1} &= \frac{T_1^{(t)}}{k} \\
\sigma_a^{2(t+1)} &= \frac{T_2^{(t)}}{k} - \mu^{(t+1)^2} \\
\sigma^{2(t+1)} &= \frac{T_3^{(t)}}{N}
\end{aligned}$$

Iterations between these 2 sets of equations define the EM algorithm. With the starting values of $\mu^{(0)} = 54.0$, $\sigma^{2^{(0)}} = 70.0$, $\sigma_a^{2^{(0)}} = 248.0$, the maximum likelihood estimates of $\hat{\mu} = 53.3184$, $\hat{\sigma}_a^2 = 54.827$ and $\hat{\sigma}^2 = 249.22$ were obtained after 30 iterations. These can be compared with the estimates of σ_a^2 and σ^2 obtained by equating observed and expected mean squares from the random effects analysis of variance given above. Estimates of σ_a^2 and σ^2 obtained from this analysis are 73.39 and 248.29 respectively.

Convergence of the EM Algorithm

The EM algorithm attempts to maximize $\ell_{\text{obs}}(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$ by maximizing $\ell(\boldsymbol{\theta}; \mathbf{y})$, the complete-data log-likelihood. Each iteration of EM has two steps: an E-step and an M-step. The t^{th} E-step finds the conditional expectation of the complete-data log-likelihood with respect to the conditional distribution of $\mathbf{y}$ given $\mathbf{y}_{\text{obs}}$ and the current estimated parameter $\boldsymbol{\theta}^{(t)}$:

$$\begin{aligned}
Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) &= E_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta}; \mathbf{y})|\mathbf{y}_{\text{obs}}] \\
&= \int \ell(\boldsymbol{\theta}; \mathbf{y}) f(\mathbf{y}|\mathbf{y}_{\text{obs}}; \boldsymbol{\theta}^{(t)}) d\mathbf{y} \ ,
\end{aligned}$$

as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}_{\text{obs}}$ and fixed $\boldsymbol{\theta}^{(t)}$. The expectation is actually the conditional expectation of the complete-data log-likelihood, conditional on $\mathbf{y}_{\text{obs}}$.

The t^{th} M-step then finds $\boldsymbol{\theta}^{(t+1)}$ to maximize $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$ i.e., finds $\boldsymbol{\theta}^{(t+1)}$ such that

$$Q(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}),$$

for all $\boldsymbol{\theta} \in \Theta$. To verify that this iteration produces a sequence of iterates that converges to a maximum of $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$, first note that by taking conditional expectation of both sides of

$$\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}}) = \ell(\boldsymbol{\theta};\mathbf{y}) - \log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta}),$$

over the distribution of $\mathbf{y}$ given $\mathbf{y}_{\text{obs}}$ at the current estimate $\boldsymbol{\theta}^{(t)}$, $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ can be expressed in the form

$$\begin{aligned}
\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}}) &= \int \ell(\boldsymbol{\theta};\mathbf{y}) f(\mathbf{y}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta}^{(t)})d\mathbf{y} \;-\int \log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta}) f(\mathbf{y}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta}^{(t)})d\mathbf{y} \\
&= E_{\boldsymbol{\theta}^{(t)}}[\ell(\boldsymbol{\theta};\mathbf{y})|\mathbf{y}_{\text{obs}}] - E_{\boldsymbol{\theta}^{(t)}}[\log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta})|\mathbf{y}_{\text{obs}}] \\
&= Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})
\end{aligned}$$

where $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$ is as defined earlier and

$$H(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}[\log f_2(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}};\,\boldsymbol{\theta})|\mathbf{y}_{\text{obs}}].$$

The following Lemma will be useful for proving a main result that the sequence of iterates $\boldsymbol{\theta}^{(t)}$ resulting from EM algrithm will converge at least to a local maximum of $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$.

Lemma: For any $\boldsymbol{\theta} \in \Theta$,

$$H(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) \leq H(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)}).$$

Theorem: The EM algorithm increases $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$ at each iteration, that is,

$$\ell_{\text{obs}}(\boldsymbol{\theta}^{(t+1)};\mathbf{y}_{\text{obs}}) \geq \ell_{\text{obs}}(\boldsymbol{\theta}^{(t)};\mathbf{y}_{\text{obs}})$$

with equality if and only if

$$Q(\boldsymbol{\theta}^{(t+1)};\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)};\boldsymbol{\theta}^{(t)}).$$

This Theorem implies that increasing $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$ at each step leads to maximizing or at least constantly increasing $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{y}_{\text{obs}})$.

Although the general theory of EM applies to any model, it is particularly useful when the complete data $\mathbf{y}$ are from an exponential family since, as seen in examples, in such cases the E-step reduces to finding the conditional expectation of the complete-data sufficient statistics, and the M-step is often simple. Nevertheless, even when the complete data $\mathbf{y}$ are from an exponential family, there exist a variety of important applications where complete-data maximum likelihood estimation itself is complicated; for example, see Little & Rubin (1987) on selection models and log-linear models, which generally require iterative M-steps.

In a more general context, EM has been widely used in the recent past in computations related to Bayesian analysis to find the posterior mode of $\boldsymbol{\theta}$, which maximizes $\ell(\boldsymbol{\theta}|\mathbf{y}) + \log p(\boldsymbol{\theta})$ for prior density $p(\boldsymbol{\theta})$ over all $\boldsymbol{\theta} \in \Theta$. Thus in Bayesian computations, log-likelihoods used above are substituted by log-posteriors.

Extensions of the EM Algoirthm

Some Definitions and Notations

- `Regular Exponential Family (REF).`
  The joint density of an Exponential family may be written in the form :
  $$f(\boldsymbol{y};\boldsymbol{\theta}) = b(\boldsymbol{y}) \exp\left\{\boldsymbol{c}(\boldsymbol{\theta})^T \boldsymbol{s}(\boldsymbol{y})\right\} \Big/ a(\boldsymbol{\theta})$$
  where

  $\boldsymbol{s}(\boldsymbol{y})$ is a $k \times 1$ vector of sufficient statistics

  $\boldsymbol{c}(\boldsymbol{\theta})$ is a $k \times 1$ vector of parameters

  θ is a $d \times 1$ vector $\in \Omega$, a d dimensional convex set s.t. $f(\boldsymbol{y};\boldsymbol{\theta})$ is a p.d.f.

  $b(\boldsymbol{y})$ and $a(\boldsymbol{\theta})$ are scalars

  $\boldsymbol{c}(\boldsymbol{\theta})$ is called the `natural` or `canonical` parameter vector. If $k = d$ and the Jacobian of $\boldsymbol{c}(\boldsymbol{\theta})$, $\frac{\partial \boldsymbol{c}}{\partial \boldsymbol{\theta}}$ is a full rank $k \times k$ matrix, then $f(\boldsymbol{y};\boldsymbol{\theta})$ is said to belong to a Regular Exponential Family (REF). In this case
  $$f(\boldsymbol{y};\boldsymbol{\theta}) = b(\boldsymbol{y}) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y})\right\} \Big/ a(\boldsymbol{\theta})$$

- `Complete-data score vector`
  $$S(\boldsymbol{\theta};\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}}$$

- `Observed-data score vector`
  $$S_{\text{obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}}) = \frac{\partial \ell_{\text{obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}})}{\partial \boldsymbol{\theta}}$$
  Also can show that
  $$S_{\text{obs}}(\boldsymbol{\theta};\boldsymbol{y}_{\text{obs}}) = E_{\boldsymbol{\theta}}\Big[S(\boldsymbol{\theta};\boldsymbol{y})\Big|\boldsymbol{y}_{\text{obs}}\Big]$$
  assuming conditions for interchanging the operations of expectation and differentiation hold.

- `Complete-data Information Matrix`
  $$I(\boldsymbol{\theta};\boldsymbol{y}) = \frac{-\partial^2 \ell(\boldsymbol{\theta};\boldsymbol{y})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}$$

- `Complete-data Expected Information Matrix`

$$\mathcal{I}(\boldsymbol{\theta};\boldsymbol{y}) = E_{\boldsymbol{\theta}}\Big[I(\boldsymbol{\theta};\boldsymbol{y})\Big]$$

- `Observed-data Information Matrix`

$$I_{obs}\Big(\boldsymbol{\theta};\boldsymbol{y}_{obs}\Big) = \frac{-\partial^2\ell(\boldsymbol{\theta};\boldsymbol{y}_{obs})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}$$

- `Observed-data Expected Information Matrix`

$$\mathcal{I}_{obs}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = E_{\boldsymbol{\theta}}\Big[I(\boldsymbol{\theta};\boldsymbol{y}_{obs}\Big]$$

- `Conditional Expected Information Matrix`

$$\mathcal{I}_c(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = E_{\boldsymbol{\theta}}\Big[I(\boldsymbol{\theta};\boldsymbol{y})\Big|\boldsymbol{y}_{obs}\Big]$$

- `Missing Information Principle`

  Recall
  $$\ell_{obs}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = \ell(\boldsymbol{\theta};\boldsymbol{y}) - \log f_2(\boldsymbol{y}_{mis}|\boldsymbol{y}_{obs};\boldsymbol{\theta})$$
  Differentiating twice w.r.t. $\boldsymbol{\theta}$, we have

  $$I_{obs}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = I(\boldsymbol{\theta};\boldsymbol{y}) + \frac{\partial^2\log f_2(y_{mis}|\boldsymbol{y}_{obs};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}$$

  Now taking expectation over the conditional distribution $\boldsymbol{y}|\boldsymbol{y}_{obs}$:

  $$I_{obs}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = \mathcal{I}_c(\boldsymbol{\theta};\boldsymbol{y}_{obs}) - \mathcal{I}_{mis}(\boldsymbol{\theta};\boldsymbol{y}_{obs})$$

  where we denote the `missing information matrix` as

  $$\mathcal{I}_{mis}(\boldsymbol{\theta};\boldsymbol{y}_{obs}) = -E_{\boldsymbol{\theta}}\left\{\frac{\partial^2 f_2(\boldsymbol{y}_{mis}|y_{obs};\theta)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\middle|\boldsymbol{y}_{obs}\right\}$$

  In other words, the missing information principle asserts that

  `Observed Information` = `Complete Information` - `Missing Information`

Convergence Rate of EM

EM algorithm implicitly defines a mapping

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{M}(\boldsymbol{\theta}^{(t)}) \quad t = 0, 1, \ldots$$

where $\boldsymbol{M}(\boldsymbol{\theta}) = \Big(M_1(\boldsymbol{\theta}), \ldots, M_d(\boldsymbol{\theta})\Big)$. For the problem of maximizing $Q(\boldsymbol{\theta};\boldsymbol{\theta})$, it can be shown that $\boldsymbol{M}$ has a fixed point and since M is continuous and monotone $\boldsymbol{\theta}^{(t)}$ converges to a point $\boldsymbol{\theta}^* \in \Omega$.

Consider the Taylor series expansion of

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{M}(\boldsymbol{\theta}^{(t)})$$

about $\boldsymbol{\theta}^*$ noting that $\boldsymbol{\theta}^* = M(\boldsymbol{\theta}^*)$:

$$\boldsymbol{M}(\boldsymbol{\theta}^{(t)}) = \boldsymbol{M}(\boldsymbol{\theta}^*) + (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*) \left. \frac{\partial \boldsymbol{M}(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

which leads to

$$\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^* = (\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*)DM$$

where $DM = \left. \frac{\partial \boldsymbol{M}(\boldsymbol{\theta}^t)}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is a $d \times d$ matrix.

Thus near $\boldsymbol{\theta}^*$, EM algorithm is essentially a linear iteration with the `rate matrix` DM.

Definition Recall that the rate of convergence of an iterative process is defined as

$$= \lim_{t \to \infty} \frac{\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|}$$

where $\| \cdot \|$ is any vector norm.

For the EM algorithm, the rate of convergence is thus r

$$r = \lambda_{max} = \text{largest eigen value of} DM$$

Dempster, Laird, and Rubin (1977) have shown that

$$DM = \mathcal{I}_{mis}\Big(\boldsymbol{\theta}^*; y_{obs}\Big)\mathcal{I}_c^{-1}\Big(\boldsymbol{\theta}^*; \boldsymbol{y}_{obs}\Big)$$

Thus the rate of convergence is the largest eigen value of $\mathcal{I}_{mis}\Big(\boldsymbol{\theta}^*; \boldsymbol{y}_{obs}\Big)\ \mathcal{I}_c^{-1}\Big(\boldsymbol{\theta}^*; \boldsymbol{y}_{obs}\Big)$.

Obtaining the covariance matrix of MLE from the EM Algorithm

In maximum likelihood estimation, the large-sample covariance matrix of the mle $\hat{\boldsymbol{\theta}}$ is usually estimated by the observed information matrix. When using the EM Algorithm for computing the maximum likelihood estimates, once convergence is reached we can evaluate $I(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_{obs})$ directly. However, this involves calculation of the second order derivatives of the observed-data log-likelihood $\ell_{obs}(\boldsymbol{\theta}; \boldsymbol{y}_{obs})$. This is not a viable option since, we have appealed to EM algoirthm expressly to avoid the complexity of evaluating $\ell_{obs}(\boldsymbol{\theta}; \boldsymbol{y}_{obs})$ itself. Thus we need to be able to approximate $I(\hat{\boldsymbol{\theta}}; y_{obs})$ by other methods if EM Algorithm is to be a useful alternative for maximum likelihood estimation using other iterative techniques. For this we need some more results:

We can use $\mathcal{I}_c(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_{obs}) = \mathcal{I}(\hat{\boldsymbol{\theta}}; y)$ to obtain the conditional expected information matrix in the REF case, because

$$\begin{aligned} \mathcal{I}_c\Big(\boldsymbol{\theta}; \boldsymbol{y}_{obs}\Big) &= E_{\boldsymbol{\theta}}\left\{ I(\boldsymbol{\theta}; \boldsymbol{y}) \Big| \boldsymbol{y}_{obs} \right\} \\ &= E_{\boldsymbol{\theta}}\left\{ I(\boldsymbol{\theta}; \boldsymbol{y}) \right\} = \mathcal{I}(\boldsymbol{\theta}; \boldsymbol{y}) \end{aligned}$$

as $I(\boldsymbol{\theta}, \boldsymbol{y})$ is not a function of $\boldsymbol{y}$ in the REF case.

That is $\mathcal{I}_c(\hat{\boldsymbol{\theta}}; \boldsymbol{y}_{obs})$ can be obtained by replacing the sufficient statistic $\boldsymbol{s}(\boldsymbol{y})$ by its conditional expectation evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ in $I(\boldsymbol{\theta}; \boldsymbol{y})$, the complete-data information matrix.

Example (continued from page 5)

$$
\begin{aligned}
\ell(\boldsymbol{\theta}; \boldsymbol{y}) &= y_1 \log\left(\frac{1-\theta}{2}\right) + y_2 \log\frac{\theta}{4} + y_3 \log\frac{\theta}{4} + y_4 \log 1/2 \\
\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\theta}} &= \frac{-y_1}{1-\theta} + \frac{y_2}{\theta} + \frac{y_3}{\theta} \\
\frac{\partial^2 \ell(\theta; y)}{\partial \theta^2} &= \frac{-y_1}{(1-\theta)^2} - \frac{y_2}{\theta^2} - \frac{y_3}{\theta^2} \\
I(\boldsymbol{\theta}; \boldsymbol{y}) &= \frac{y_1}{(1-\theta)^2} + \frac{y_2 + y_3}{\theta^2} \\
E[I(\boldsymbol{\theta}; \boldsymbol{y})|\boldsymbol{y}_{obs}] = \mathcal{I}_\rfloor(\boldsymbol{\theta}; \boldsymbol{y}_{obs}) &= \frac{38}{(1-\theta)^2} + \frac{34}{\theta^2} + \frac{125\theta}{(2+\theta)\theta^2} \\
\mathcal{I}_\rfloor(\hat{\theta}; \boldsymbol{y}_{obs}) &= \frac{38}{(1-\hat{\theta})^2} + \frac{34}{\hat{\theta}^2} + \frac{125}{(2+\hat{\theta})\hat{\theta}}
\end{aligned}
$$

Complete this computation using $\hat{\theta}$ from previous results. Also the convergence rate can be calculated simialry and shown to be .1328 which is the value obtained in the actual iteration.

Generalized EM Algorithm (GEM)

Recall that in the M-step we maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ i.e., find $\boldsymbol{\theta}^{(t+1)}$ s.t.

$$Q\Big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\Big) \geq Q\Big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\Big)$$

for all $\boldsymbol{\theta}$. In the generalized version of the EM Algorithm we will require only that $\boldsymbol{\theta}^{(t+1)}$ be chosen such that

$$Q\Big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}\Big) \geq Q\Big(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}\Big)$$

holds, i.e., $\boldsymbol{\theta}^{(t+1)}$ is chosen to increase $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ over its value at $\boldsymbol{\theta}^{(t)}$ at each iteration t. This is sufficient to ensure that

$$\ell\Big(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{y}\Big) \geq \ell\Big(\boldsymbol{\theta}^{(t)}; \boldsymbol{y}\Big)$$

at each iteration, so GEM sequence of iterates also converges to a local maximum.

GEM Algorithm based on a single N-R step

We use GEM-type algorithms when a global maximizer of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ does not exist in closed form. In this case, possibly an iterative method is required to accomplish the M-step, which might prove to be a computationally infeasible procedure. Since it is not essential to actually maximize Q in a GEM, but only increase the likelihood, we may replace the M-step with a step that achieves that. One possibilty of such a step is a single iteration of the

Newton-Raphson(N-R) algorithm, which we know is a descent method.

Let $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + a^{(t)}\boldsymbol{\delta}^{(t)}$

where $\boldsymbol{\delta}^{(t)} = -\left[\frac{\partial^2 Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \left[\frac{\partial Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})}{\partial\boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$

i.e., $\boldsymbol{\delta}^{(t)}$ is the N-R direction at $\boldsymbol{\theta}^{(t)}$ and $0 < a^{(t)} \le 1$. If $a^{(t)} = 1$ this will define an exact N-R step. Here we will choose $a^{(t)}$ so that this defines a GEM sequence. This will be achieved if $a^{(t)} < 2$ as $t \to \infty$.

General Mixed Model

The general mixed linear model is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}_{n\times 1}$ is an observed random vector, $\boldsymbol{X}$ is an $n \times p$, and $\boldsymbol{Z}_i$ are $n \times q_i$, matrices of known constants, $\boldsymbol{\beta}_{p\times 1}$ is a vector of unknown parameters, and $\boldsymbol{u}_i$ are $q_i \times 1$ are vectors of unobservable random effects.

$\boldsymbol{\epsilon}_{n\times 1}$ is assumed to be distributed n-dimensional multivariate normal $N(\mathbf{0},\ \sigma_{\mathbf{0}}^{\mathbf{2}}\boldsymbol{I}_{\mathbf{n}})$ and each $\boldsymbol{u}_i$ are assumed to have q_i-dimensional multivariate normal distributions $N_{q_i}(\mathbf{0},\ \sigma_{\mathbf{i}}^{\mathbf{2}}\boldsymbol{\Sigma}_{\mathbf{i}})$ for $i = 1,\ 2,\ \ldots,\ r$, independent of each other and of $\boldsymbol{\epsilon}$.

We take the *complete data vector* to be $(\boldsymbol{y},\ \boldsymbol{u}_1,\ \ldots,\ \boldsymbol{u}_r)$ where $\boldsymbol{y}$ is the *incomplete* or the *observed data vector*. It can be shown easily that the covariance matrix of $\boldsymbol{y}$ is the $n \times n$ matrix $\boldsymbol{V}$ where

$$\boldsymbol{V} = \sum_{i=1}^{r} \boldsymbol{Z}_i\boldsymbol{Z}_i^T\sigma_i^2 + \sigma_0^2\boldsymbol{I}_n$$

Let $q = \sum_{i=0}^{r} q_i$ where $q_0 = n$. The joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}_1,\ \ldots,\ \boldsymbol{u}_r$ is q-dimensional multivariate normal $N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ where

$$\underset{q\times 1}{\boldsymbol{\mu}} = \begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \underset{q\times q}{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{V} & \left\{ \sigma_i^2\boldsymbol{Z}_i \right\}_{r\ \ i=1}^{\ \ r} \\ \left\{ \sigma_i^2\boldsymbol{Z}_i^T \right\}_{c\ \ i=1}^{\ \ r} & \left\{ \sigma_i^2\boldsymbol{I}_{q_i} \right\}_{d\ \ i=1}^{\ \ r} \end{bmatrix}$$

Thus the density function of $\boldsymbol{y},\ \boldsymbol{u},\ \ldots,\ \boldsymbol{u}_r$ is

$$f(\boldsymbol{y},\ \boldsymbol{u}_1,\ \boldsymbol{u}_2,\ \ldots,\ \boldsymbol{u}_r) = (2\pi)^{-\frac{1}{2}q}\ |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\ \exp(-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w})$$

where $\boldsymbol{w} = \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T,\ \boldsymbol{u}_1^T,\ \ldots,\ \boldsymbol{u}_r^T\right]$. This gives the complete data loglikelihood to be

$$l = -\frac{1}{2}q\log(2\pi) - \frac{1}{2}\sum_{i=0}^{r} q_i\log\sigma_i^2 - \frac{1}{2}\sum_{i=0}^{r}\frac{\boldsymbol{u}_i^T\boldsymbol{u}_i}{\sigma_i^2}$$

where $\boldsymbol{u}_0 = \boldsymbol{y} - \boldsymbol{X\beta} - \sum_{i=1}^{r} \boldsymbol{Z}_i \boldsymbol{u}_i = (\boldsymbol{\epsilon})$. Thus the sufficient statistics are: $\boldsymbol{u}_i^T \boldsymbol{u}_i \;\; i = 0, \ldots, r$, and $\boldsymbol{y} - \sum_{i=1}^{r} \boldsymbol{Z}_i \boldsymbol{u}_i$ and the maximum likelihood estimates (m.l.e.'s) are

$$\begin{aligned} \hat{\sigma}_i^2 &= \frac{\boldsymbol{u}_i^T \boldsymbol{u}_i}{q_i}, \qquad i = 0, 1, \ldots, r \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \boldsymbol{X})^{-} \boldsymbol{X}^T (\boldsymbol{y} - \sum_{i=1}^{r} \boldsymbol{Z}_i \boldsymbol{u}_i) \end{aligned}$$

Special Case: Two-Variance Components Model

The general mixed linear model reduces to:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1 \boldsymbol{u}_1 + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_{\mathbf{0}}^{\mathbf{2}} \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{u}_{\mathbf{1}} \sim \mathbf{N}(\mathbf{0}, \sigma_{\mathbf{1}}^{\mathbf{2}} \boldsymbol{I}_{\mathbf{n}})$$

and the covariance matrix of $\boldsymbol{y}$ is now

$$\boldsymbol{V} = \boldsymbol{Z}_1 \boldsymbol{Z}_1^T \sigma_1^2 + \sigma_0^2 \boldsymbol{I}_n$$

The complete data loglikelihood is

$$l = -\frac{1}{2} q \log(2\pi) - \frac{1}{2} \sum_{i=0}^{1} q_i \log \sigma_i^2 - \frac{1}{2} \sum_{i=0}^{1} \frac{\boldsymbol{u}_i^T \boldsymbol{u}_i}{\sigma_i^2}$$

where $\boldsymbol{u}_0 = \boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z}_1 \boldsymbol{u}_1$. The m.l.e.'s are

$$\begin{aligned} \hat{\sigma}_i^2 &= \frac{\boldsymbol{u}_i^T \boldsymbol{u}_i}{q_i} \qquad i = 0, 1 \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \boldsymbol{X})^{-} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{Z}_1 \boldsymbol{u}_1) \end{aligned}$$

We need to find the expected values of the sufficient statistics $\boldsymbol{u}_i^T \boldsymbol{u}_i, \;\; i = 0, 1$ and $\boldsymbol{y} - Z_1 \boldsymbol{u}_1$ conditional on observed data vector $\boldsymbol{y}$. Since $\boldsymbol{u}_i | \boldsymbol{y}$ is distributed as q_i-dimensional multivariate normal

$$N(\sigma_i^2 \boldsymbol{Z}_i^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}), \; \sigma_i^2 \boldsymbol{I}_{q_i} - \sigma_i^4 \boldsymbol{Z}_i^T \boldsymbol{V}^{-1} \boldsymbol{Z}_i)$$

we have

$$\begin{aligned} E(\boldsymbol{u}_i^T \boldsymbol{u}_i \mid \boldsymbol{y}) &= \sigma_i^4 (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} \boldsymbol{Z}_i \boldsymbol{Z}_i^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \quad + \; tr(\sigma_i^2 \boldsymbol{I}_{q_i} - \sigma_i^4 \boldsymbol{Z}_i^T \boldsymbol{V}^{-1} \boldsymbol{Z}_i) \\ E(\boldsymbol{y} - \boldsymbol{Z}_1 \boldsymbol{u}_1 \mid \boldsymbol{y}) &= \boldsymbol{X\beta} + \sigma_0^2 \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \end{aligned}$$

noting that

$$\begin{aligned} E(\boldsymbol{u}_0 \mid \boldsymbol{y}) &= \sigma_0^2 \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \\ E(\boldsymbol{u}_0^T \boldsymbol{u}_0 \mid \boldsymbol{y}) &= \sigma_0^4 (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} \boldsymbol{Z}_0 \boldsymbol{Z}_0^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \quad + \; tr(\sigma_0^2 \boldsymbol{I}_{q_i} - \sigma_0^4 \boldsymbol{Z}_0^T \boldsymbol{V}^{-1} \boldsymbol{Z}_0) \end{aligned}$$

where $\boldsymbol{Z}_0 = \boldsymbol{I}_n$.

From the above we can derive the following EM-type algorithms for this case:

Basic EM Algorithm

Step 1 (E-step) Set $\boldsymbol{V}^{(t)} = \boldsymbol{Z}_1\boldsymbol{Z}_1'\sigma_1^{2^{(t)}} + \sigma_0^{2^{(t)}}\boldsymbol{I}_n$ and for $i = 0,\ 1$ calculate

$$\begin{aligned}
\hat{s}_i^{(t)} &= E(\boldsymbol{u}_i^T\boldsymbol{u}_i|\boldsymbol{y}) \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)},\ \sigma_i^2=\sigma_i^{2^{(t)}}} \\
&= \sigma_i^{4^{(t)}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{V}^{(t)}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{V}^{(t)}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)}) \\
&\quad + tr(\sigma_i^{2^{(t)}}\boldsymbol{I}_{q_i} - \sigma_i^{4^{(t)}}\boldsymbol{Z}_i^T\boldsymbol{V}^{(t)^{-1}}\boldsymbol{Z}_i) \quad i = 0,\ 1 \\
\hat{\boldsymbol{w}}^{(t)} &= E(\boldsymbol{y} - \boldsymbol{Z}_1\boldsymbol{u}_1|\boldsymbol{y}) \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)},\ \sigma_i^2=\sigma_i^{2^{(t)}}} \\
&= \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \sigma_0^{2^{(t)}}\boldsymbol{V}^{(t)^{-1}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})
\end{aligned}$$

Step 2 (M-step)

$$\begin{aligned}
\sigma_i^{2^{(t+1)}} &= \hat{s}_i^{(t)}/q_i \qquad i = 0,\ 1 \\
\boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{w}}^{(t)}
\end{aligned}$$

ECM Algorithm

Step 1 (E-step) Set $\boldsymbol{V}^{(t)} = \boldsymbol{Z}_1\boldsymbol{Z}_1'\sigma_1^{2^{(t)}} + \sigma_0^{2^{(t)}}\boldsymbol{I}_n$ and, for $i = 0,\ 1$ calculate

$$\begin{aligned}
\hat{s}_i^{(t)} &= E(\boldsymbol{u}_i^T\boldsymbol{u}_i|\boldsymbol{y}) \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)},\ \sigma_i^2=\sigma_i^{2^{(t)}}} \\
&= \sigma_i^{4^{(t)}}(y-\boldsymbol{X}\boldsymbol{\beta}^{(t)})^T\boldsymbol{V}^{(t)^{-1}}\boldsymbol{Z}_i\boldsymbol{Z}_i\boldsymbol{V}^{(t)^{-1}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\
&\quad + tr(\sigma_i^{2^{(t)}}\boldsymbol{I}_{q_i} - \sigma_i^{4^{(t)}}\boldsymbol{Z}_i^T\boldsymbol{V}^{(t)^{-1}}\boldsymbol{Z}_i)
\end{aligned}$$

Step 2 (M-step)

Partition the parameter vector $\boldsymbol{\theta} = (\sigma_0^2,\ \sigma_1^2,\ \boldsymbol{\beta})$ as $\boldsymbol{\theta}_1 = (\sigma_0^2,\ \sigma_1^2)$ and $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$

CM-step 1

Maximize complete data log likelihood over $\boldsymbol{\theta}_1$

$$\sigma_i^{2^{(t+1)}} = \hat{s}_i^{(t)}/q_i \qquad i = 0,\ 1$$

CM-step 2

Calculate $\boldsymbol{\beta}^{(t+1)}$ as

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{w}}^{(t)}$$

where

$$\hat{\boldsymbol{w}}^{(t+1)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \sigma_0^{2^{(t+1)}}\boldsymbol{V}^{(t+1)^{-1}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^{(t)})$$

ECME Algorithm

Step 1 (E-step) Set $\boldsymbol{V}^{(t)} = \boldsymbol{Z}_1\boldsymbol{Z}_1'\sigma_1^{2^{(t)}} + \sigma_0^{2^{(t)}}\boldsymbol{I}_n$ and, for $i = 0,\ 1$ calculate

$$\begin{aligned} \hat{s}_i^{(t)} &= E(\boldsymbol{u}_i^T\boldsymbol{u}_i|\boldsymbol{y}) \mid \sigma_i^2 = \sigma_i^{2(t)} \\ &= \sigma_i^{4^{(t)}}\boldsymbol{y}^T\boldsymbol{P}^{(t)}\boldsymbol{Z}_i\boldsymbol{Z}_i^T\boldsymbol{P}^{(t)}\boldsymbol{y} + \ tr(\sigma_i^{2^{(t)}}\boldsymbol{I}_{q_i} - \sigma_i^{4^{(t)}}\boldsymbol{Z}_i^T\boldsymbol{V}^{(t)^{-1}}\boldsymbol{Z}_i) \end{aligned}$$

where $\boldsymbol{P}^{(t)} = \boldsymbol{V}^{(t)^{-1}} - \boldsymbol{V}^{(t)^{-1}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{(t)^{-1}}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{V}^{(t)^{-1}}$

Step 2 (M-step)

Partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_1 = (\sigma_0^2,\ \sigma_1^2)$ and $\boldsymbol{\theta}_2 = \boldsymbol{\beta}$ as in ECM.

CM-step 1

Maximize complete data log likelihood over $\boldsymbol{\theta}_1$

$$\sigma_i^{2^{(t+1)}} = \hat{s}_i^{(t)}/q_i \qquad i = 0,\ 1$$

CM-step 2

Maximize the observed data log likelihood over $\boldsymbol{\theta}$ given $\boldsymbol{\theta}_1^{(t)} = (\sigma_0^{2^{(t)}},\ \sigma_0^{2^{(t)}})$:

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^T\boldsymbol{V}^{(t+1)^{-1}}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{V}^{(t+1)^{-1}})\boldsymbol{y}$$

(Note: This is the WLS estimator of $\boldsymbol{\beta}$.)

Example of Mixed Model Analysis using the EM Algorithm

The first example is an evaluation of the breeding value of a set of five sires in raising pigs, taken from Snedecor and Cochran (1967). (The data is reported in Appendix 4.) The experiment was designed so that each sire is mated to a random group of dams, each mating producing a litter of pigs whose characteristics are criterion. The model to be estimated is

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \epsilon_{ijk}, \tag{4}$$

where α_i is a constant associated with the i-th sire effect, β_{ij} is a random effect associated with the i-th sire and j-th dam, ϵ_{ijk} is a random term. The three different initial values for (σ_0^2, σ_1^2) are $(1, 1)$, $(10, 10)$ and $(.038, .0375)$; the last initial value corresponds to the estimates from the SAS ANOVA procedure.

Table 1: **Average Daily Gain of Two Pigs of Each Litter (in pounds)**

Sire	Dam	Gain
1	1	2.77
1	1	2.38
1	2	2.58
1	2	2.94
2	1	2.28
2	1	2.22
2	2	3.01
2	2	2.61
3	1	2.36
3	1	2.71

Sire	Dam	Gain
3	2	2.72
3	2	2.74
4	1	2.87
4	1	2.46
4	2	2.31
4	2	2.24
5	1	2.74
5	1	2.56
5	2	2.50
5	2	2.48

```
###########################################################
#     Classical EM algorithm for Linear  Mixed Model      #
###########################################################
em.mixed <- function(y, x, z, beta, var0, var1,maxiter=2000,tolerance = 1e-0010)
        {
        time <-proc.time()
        n <- nrow(y)
        q1 <- nrow(z)
        conv <- 1
        L0 <- loglike(y, x, z, beta, var0, var1)
        i<-0
        cat("  Iter.       sigma0                 sigma1        Likelihood",fill=T)
        repeat {
                if(i>maxiter) {conv<-0
                        break}

        V <- c(var1) * z %*% t(z) + c(var0) * diag(n)
        Vinv <- solve(V)
        xb <- x %*% beta
        resid <- (y-xb)
        temp1 <- Vinv %*% resid
        s0 <- c(var0)^2 * t(temp1)%*%temp1 + c(var0) * n - c(var0)^2 * tr(Vinv)
        s1 <- c(var1)^2 * t(temp1)%*%z%*%t(z)%*%temp1+ c(var1)*q1 -
                                                    c(var1)^2 *tr(t(z)%*%Vinv%*%z)
        w <- xb + c(var0) * temp1
        var0 <- s0/n
        var1 <- s1/q1
        beta <- ginverse( t(x) %*% x) %*% t(x)%*% w
        L1 <- loglike(y, x, z, beta, var0, var1)

        if(L1 < L0) { print("log-likelihood must increase, llikel <llike0, break.")
                             conv <- 0
                             break
                                    }
        i <- i + 1
        cat("  ", i,"  ",var0,"  ",var1,"  ",L1,fill=T)
        if(abs(L1 - L0) < tolerance) {break}  #check for convergence
        L0 <- L1
        }
list(beta=beta, var0=var0,var1=var1,Loglikelihood=L0)

}
```

```
##########################################################
#  loglike calculates the LogLikelihood for Mixed Model #
##########################################################
loglike<- function(y, x, z, beta, var0, var1)
        {
        n<- nrow(y)

        V <- c(var1) * z %*% t(z) + c(var0) * diag(n)

        Vinv <- ginverse(V)
        xb <- x %*% beta
        resid <- (y-xb)
        temp1 <- Vinv %*% resid
        (-.5)*( log(det(V)) + t(resid) %*% temp1 )
}
```

```
> y <- matrix(c(2.77, 2.38, 2.58, 2.94, 2.28, 2.22, 3.01, 2.61,
+  2.36, 2.71, 2.72, 2.74, 2.87, 2.46, 2.31, 2.24,
+  2.74, 2.56, 2.50, 2.48),20,1)
> x1 <- rep(c(1,0,0,0,0),rep(4,5))
> x2 <- rep(c(0,1,0,0,0),rep(4,5))
> x3 <- rep(c(0,0,1,0,0),rep(4,5))
> x4 <- rep(c(0,0,0,1,0),rep(4,5))
> x <- cbind(1,x1,x2,x3,x4)
> x
        x1 x2 x3 x4
 [1,] 1  1  0  0  0
 [2,] 1  1  0  0  0
 [3,] 1  1  0  0  0
 [4,] 1  1  0  0  0
 [5,] 1  0  1  0  0
 [6,] 1  0  1  0  0
 [7,] 1  0  1  0  0
 [8,] 1  0  1  0  0
 [9,] 1  0  0  1  0
[10,] 1  0  0  1  0
[11,] 1  0  0  1  0
[12,] 1  0  0  1  0
[13,] 1  0  0  0  1
```

```
[14,] 1  0  0  0  1
[15,] 1  0  0  0  1
[16,] 1  0  0  0  1
[17,] 1  0  0  0  0
[18,] 1  0  0  0  0
[19,] 1  0  0  0  0
[20,] 1  0  0  0  0
```

```
> beta <- lm(y~ x1 + x2 + x3 +x4)$coefficients
>  beta
               [,1]
(Intercept)  2.5700
         x1  0.0975
         x2 -0.0400
         x3  0.0625
         x4 -0.1000

> z=matrix(rep( as.vector(diag(1,10)),rep(2,100)),20,10)
> z
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    1    0    0    0    0    0    0    0    0     0
 [2,]    1    0    0    0    0    0    0    0    0     0
 [3,]    0    1    0    0    0    0    0    0    0     0
 [4,]    0    1    0    0    0    0    0    0    0     0
 [5,]    0    0    1    0    0    0    0    0    0     0
 [6,]    0    0    1    0    0    0    0    0    0     0
 [7,]    0    0    0    1    0    0    0    0    0     0
 [8,]    0    0    0    1    0    0    0    0    0     0
 [9,]    0    0    0    0    1    0    0    0    0     0
[10,]    0    0    0    0    1    0    0    0    0     0
[11,]    0    0    0    0    0    1    0    0    0     0
[12,]    0    0    0    0    0    1    0    0    0     0
[13,]    0    0    0    0    0    0    1    0    0     0
[14,]    0    0    0    0    0    0    1    0    0     0
[15,]    0    0    0    0    0    0    0    1    0     0
[16,]    0    0    0    0    0    0    0    1    0     0
[17,]    0    0    0    0    0    0    0    0    1     0
[18,]    0    0    0    0    0    0    0    0    1     0
[19,]    0    0    0    0    0    0    0    0    0     1
[20,]    0    0    0    0    0    0    0    0    0     1

> tolerance <- 1e-0010
```

```
> maxiter <- 2000
> seed <- 100

> tr <- function(x) sum(diag(x))

> pig.em.results=em.mixed(y,x,z,beta,1,1)
```

```
 Iter.      sigma0                 sigma1          Likelihood
  1    0.355814166666667    0.672928333333333    1.79926992591149
  2    0.161289219777595    0.41563069748673   7.67656899233908
  3    0.0876339412658286    0.251467232868861    12.1211098467902
  4    0.0572854577134676    0.154608860144254    15.1706421424132
  5    0.0442041832136993    0.0994160507019009    17.130593350043
  6    0.0383642480366208    0.0681788894372488    18.3694784469532
  7    0.0356787493219611    0.0501555498139391    19.1464492643216
  8    0.0344485615271845    0.0393230630005528    19.630522654304
  9    0.0339421060204835    0.032463488015722    19.9344237098546
  10    0.0338157550885801    0.0278799360094821    20.1296516102853
  11    0.0338906702246361    0.0246617461868549    20.2590878975877
  12    0.0340677808426562    0.0223026501468333    20.3478401280992
  13    0.0342914060268899    0.02050908367209    20.4106731999847
  14    0.0345307406853334    0.0191033399491361    20.4564548122603
  15    0.0347693177408093    0.0179733335898586    20.4906651701558
  16    0.0349988121264555    0.0170456362588657    20.5167959381214
  17    0.0352154250729573    0.0162704665032569    20.5371387472463
  18    0.0354178131324059    0.0156130259538466    20.5532397057213
  19    0.035605923319033    0.0150483214110959    20.5661685104734
  20    0.0357803477632565    0.0145579670979211    20.5766822585211
  .              .                  .                       .
  .              .                  .                       .
  .              .                  .                       .

  56    0.0381179518848995    0.00973330752818472    20.6382666130708
  57    0.0381395936235214    0.00969817892685448    20.6383975558995
  58    0.0381603337241741    0.00966462573503992    20.6385178920802
  59    0.0381802168890155    0.00963256091234115    20.6386285603282
  60    0.0381992850741614    0.00960190350873251    20.6387304072052
  61    0.0382175776978136    0.00957257813571635    20.6388241972305
  62    0.0382351318295782    0.00954451449226923    20.6389106217546
  63    0.0382519823629303    0.0095176469390277    20.6389903067599
  64    0.0382681621725531    0.00949191411504429    20.639063819736
  65    0.0382837022580806    0.00946725859219654    20.639131675748
  66    0.0382986318756009    0.00944362656297278    20.6391943428055
  67    0.038312978658123    0.00942096755790665    20.6392522466191
  68    0.0383267687260784    0.00939923418940295    20.6393057748245
  69    0.0383400267888113    0.0093783819191014    20.6393552807376
  70    0.0383527762379075    0.00935836884627387    20.6394010866997
  71    0.0383650392331249    0.00933915551505174    20.6394434870619
  72    0.0383768367816026    0.00932070473854103    20.639482750852
  73    0.0383881888109615    0.00930298143810976    20.6395191241599
```

74	0.0383991142368419	0.00928595249632906	20.6395528322763
75	0.0384096310253713	0.00926958662222157	20.6395840816114
76	0.0384197562510039	0.00925385422762105	20.6396130614187
77	0.0384295061501315	0.00923872731357883	20.6396399453462
78	0.0384388961708265	0.00922417936586811	20.6396648928339
79	0.0384479410190403	0.00921018525873869	20.6396880503735
80	0.038456654701554	0.00919672116616442	20.639709552647
81	0.0384650505659457	0.00918376447990421	20.6397295235547
.	.	.	.
.	.	.	.
.	.	.	.
148	0.0386757695291326	0.00886369156913502	20.6399971471733
149	0.038676564371102	0.00886250240210069	20.6399973283659
150	0.0386773329964768	0.00886135258462483	20.6399974978113
151	0.0386780762792084	0.00886024079675071	20.6399976562748
152	0.0386787950635179	0.00885916576395544	20.6399978044714
153	0.0386794901649452	0.00885812625551144	20.6399979430694
154	0.0386801623713602	0.00885712108291183	20.6399980726932
155	0.0386808124439345	0.00885614909835692	20.6399981939264
156	0.0386814411180777	0.00885520919329904	20.6399983073142
157	0.0386820491043391	0.0088543002970433	20.6399984133665
158	0.0386826370892749	0.00885342137540194	20.6399985125595
159	0.0386832057362845	0.00885257142939978	20.6399986053387
160	0.0386837556864153	0.0088517494940289	20.6399986921201
161	0.0386842875591385	0.00885095463705022	20.639998773293
162	0.038684801953096	0.00885018595784018	20.6399988492209
163	0.0386852994468205	0.0088494425862806	20.6399989202439
164	0.0386857805994294	0.00884872368168998	20.6399989866798
165	0.0386862459512932	0.00884802843179444	20.6399990488258
166	0.0386866960246808	0.00884735605173682	20.6399991069597
167	0.03868713132438	0.0088467057831223	20.6399991613413
168	0.0386875523382973	0.00884607689309907	20.6399992122135
169	0.0386879595380351	0.00884546867347273	20.6399992598033
170	0.0386883533794493	0.00884488043985296	20.639999304323
171	0.0386887343031859	0.00884431153083127	20.6399993459712
.	.	.	.
.	.	.	.
.	.	.	.
200	0.0386957009460699	0.00883391223498804	20.6399998631143
201	0.0386958413121088	0.00883370281147353	20.6399998687721
202	0.0386959770906578	0.00883350023633904	20.6399998740661
203	0.0386961084319343	0.00883330428508165	20.6399998790199
204	0.0386962354812184	0.00883311474059363	20.6399998836552
205	0.0386963583790162	0.00883293139291657	20.6399998879925
206	0.0386964772612182	0.00883275403900379	20.6399998920511
207	0.0386965922592513	0.00883258248249084	20.6399998958488
208	0.038696703500227	0.0088324165334737	20.6399998994025
209	0.0386968111070837	0.00883225600829448	20.6399999027278
210	0.0386969151987245	0.00883210072933436	20.6399999058394
211	0.0386970158901508	0.00883195052481344	20.639999908751
212	0.0386971132925907	0.00883180522859742	20.6399999114756
213	0.0386972075136236	0.00883166468001066	20.6399999140251
214	0.0386972986573006	0.0088315287236556	20.6399999164108

```
  215   0.0386973868242609   0.00883139720923821   20.6399999186432
  216   0.0386974721118441   0.00883126999139926   20.6399999207322
  217   0.0386975546141991   0.00883114692955128   20.639999922687
  218   0.0386976344223887   0.00883102788772094   20.6399999245163
  219   0.0386977116244922   0.00883091273439672   20.639999926228
   .            .                   .                   .
   .            .                   .                   .
   .            .                   .                   .
  243   0.0386989678273622   0.00882903917935159   20.6399999460984
  244   0.0386990015024108   0.00882898895948373   20.6399999464241
  245   0.0386990340785396   0.00882894037866968   20.6399999467289
  246   0.0386990655916262   0.00882889338338294   20.6399999470141
  247   0.038699096076376   0.00882884792184704   20.639999947281
  248   0.0386991255663601   0.00882880394397821   20.6399999475308
  249   0.038699154094053   0.00882876140132994   20.6399999477646
  250   0.0386991816908679   0.00882872024703932   20.6399999479833
  251   0.0386992083871918   0.00882868043577519   20.639999948188
  252   0.0386992342124192   0.00882864192368795   20.6399999483796
  253   0.0386992591949842   0.00882860466836104   20.6399999485588
  254   0.0386992833623921   0.00882856862876405   20.6399999487266
  255   0.0386993067412498   0.00882853376520732   20.6399999488836
  256   0.0386993293572951   0.00882850003929805   20.6399999490305
  257   0.0386993512354253   0.00882846741389785   20.639999949168
  258   0.0386993723997246   0.00882843585308171   20.6399999492966
  259   0.0386993928734904   0.00882840532209825   20.639999949417
  260   0.0386994126792596   0.00882837578733138   20.6399999495297
  261   0.0386994318388329   0.00882834721626309   20.6399999496351
  262   0.0386994503732993   0.00882831957743758   20.6399999497338

> pig.em.results
$beta:
        [,1]
[1,]  2.5700
[2,]  0.0975
[3,] -0.0400
[4,]  0.0625
[5,] -0.1000

$var0:
           [,1]
[1,] 0.03869945

$var1:
           [,1]
[1,] 0.00882832

$Loglikelihood:
      [,1]
[1,] 20.64
```

References

Aitkin, M. and Wilson G.T. (1980) "Mixture Models, Outliers and the EM Algorithm," *Technometrics*, 22, 325–331.

Baker S.G. (1990) "A simple EM Algorithm for Capture-Recapture Data with Categorical Covariates," *Biometrics*, 46, 1193–1200.

Callanan, T. and Harville, D.A. (1991) "Some New Algorithms for Computing REML of Variance Components," *Journal of Statistical Computation and Simulation*, 38, 239–259.

Davenport, J.W., Pierce, M.A, and Hathaway, R.J. (1988) " A Numerical Comparison of EM and Quasi-Newton Type Algorithms for MLE for a Mixture of Normals," *Computer Science and Statistics: Proc. of the 20th Symp. on the Interface*, 410–415.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) " Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B, Methodological*, 39, 1–38.

Fessler, J.A. and Hero, A.O. (1994) "Space-alternating Generalized Expaectation–Maximization Algorithm," *IEEE Trans. on Signal Processing*, 42,2664–2677.

Jamshidian, M. and Jennrich, R.I.(1993) "Conjugate Gradient Acceleration of the EM Algorithm," *Journal of the American Statistical Association*, 88, 221–228.

Laird, N.M. (1982) " Computation of Variance Components using the EM Algorithm ," *Journal of Statistical Computation and Simulation*, 14, 295–303.

Laird, N.M., Lange, N. and Stram, D. (1987) " Maximum Likelihood Computation with Repeated Measures," *Journal of the American Statistical Association*, 83, 97–105.

Lindstrom and Bates, D. (1988) " Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated Measure Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*—, New York: John Wiley.

Liu, C.H. and Rubin, D.B. (1994) "The ECME Algorithm: a simple Expansion of EM and ECM with Faster Monotone Convergence," *Biometrika*, 81, 633–648.

Louis, T.A. (1982) " Finding the Observed Information Matrix when using the EM Algorithm," *Journal of the Royal Statistical Society, Series B, Methodological*, 44, 226–233.

Meng, X.L. (1994) "On the Rate of Convergence of the ECM Algorithm," *Annals of Statistics*, 22, 326–339.

Meng, X.L. 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.L. and Rubin, D.B. (1993) "Maximum Likelihood Estimation via the ECM Algorithm: a general framework," *Biometrika*, 80, 267–278.

Meilijson, I. (1989) "A Fast Improvement to the EM Algorithm on its Own Terms," *Journal of the Royal Statistical Society, Series B, Methodological*, 44, 226–233.

Orchard, T. and Woodbury, M.A. (1972) "A Missing Information Principle: Theory and Applications," *Proc. of the 6th Symp. on Math. Stat. and Prob. Vol.1*, 697–715.

Rubin, D.B. (1976) "Inference with Missing Data," *Biometrika*, 63, 581–592.

Rubin, D.B. (1991) "EM and Beyond," *Psychometrika*, 56, 241–254.

Tanner, M.A. and Wong, W.H. (1987) "The Calculation of Posterior Distributions by Data Augmentation (with discussion)," *Journal of the American Statistical Association*, 82, 805–811.

Titterington, D.M. (1984) "Recursive Parameter Estimation using Incomplete Data," *Journal of the Royal Statistical Society, Series B, Methodological*, 46, 257–267.

Wei, G.C.G. and Tanner, M.A. (1990) "A Monte Carlo Implementation of the EM algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.

Wu, C.F.J. (1983) "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.