Applied Statistics: Lent Term

Lecture: EM algorithm

What is the EM algorithm?

The Expectation-Maximisation (EM) algorithm is a general iterative method for maximum likelihood (ML) estimation in a wide variety of situations, best described as incomplete data problems.

Usage of the EM algorithm:

The EM algorithm is used to tackle a wide variety of problems in many different fields. These problems are very broad and include problems that either would not usually be considered as or are not obvious incomplete data problems.

Examples:

Natural situations

1. Missing data problems
2. Grouped data problems
3. Truncated and censored (survival) data problems

Not obvious situations

1. Variance component estimation
2. Latent variable situations and random effects models
3. Mixture models

Areas of application:

1. Image analysis
2. Epidemiology and Medicine
3. Engineering
4. Genetics and Biology
Background:

The term EM algorithm was coined by Dempster, Laird and Rubin in their seminal paper read before the Royal Statistical Society in 1976 and published in JRSS B in 1977. In this paper, the authors presented a general framework and formulation for the EM algorithm, which up to then was lacking. Although the EM algorithm was used as early as 1926 by McKendrick (see Meng (1997)) when he analysed data on a cholera epidemic in an Indian village, it was only until the paper by Dempster, Laird and Rubin that the full generality of the algorithm was realised and the ideas synthesised. In their paper, basic properties of the EM algorithm were established and many examples were provided some in areas not up to then thought to be incomplete data problems.

The EM algorithm is closely related to the intuitive idea of

1. filling in missing values by estimated values;
2. estimating the parameters given the missing values;
3. re-estimating the missing values given the new parameter estimates;
4. re-estimating the parameters given the new missing values; and

continuing along these lines, iterating, until convergence of the parameter estimates.

In a strong sense the EM algorithm formalises this ad hoc idea of handling missing data. In fact, the ad hoc approach mentioned above is an EM algorithm under the condition that the "complete" log-likelihood is linear in the missing values, otherwise this approach may lead to biased estimates.

Formulation:

The basic idea behind the EM algorithm is quite natural and intuitive and can be stated as follows:

In an incomplete data problem, the observed likelihood (i.e. the likelihood constructed from the observed data) might be complicated or numerically infeasible to maximise, so instead we replace this likelihood by a likelihood, the complete likelihood, that is easier to maximise. To do this, we augment the observed data with manufactured data so as to create a complete likelihood that is more solvable. That is, we associate with the given incomplete data problem, a complete data problem for which ML estimation is computationally more tractable. For example, the complete data problem may produce closed form solutions for the maximum likelihood estimates (mle's) or may allow us to use standard statistical functions to obtain the mle's.
Each iteration of the EM algorithm consists of an E-step (expectation step) and an M-step (maximisation step).

The E-step calculates the conditional expectation of the complete log-likelihood given the observed data and the current estimates of the parameters.

The M-step requires performing maximum likelihood estimation for the parameters on the (conditional) expected log-likelihood found in the E-step.

Mathematically, let $Y$ be the random vector from density $g(y; \theta)$, where $\theta$ is a $p$-dimensional vector of parameters from the parameter space $\Omega$. Let $y$ be the observed data corresponding to $Y$. Let $Z$ be the random vector of unobservable data and let $X = (Y, Z)$ be the random vector of complete data. Represent the probability density function (pdf) of the random vector $X$ by $g_C(x; \theta)$.

The observed log-likelihood is given by

$$l(\theta | y) = \log L(\theta | y) = \log g(y; \theta)$$

and the complete log-likelihood is

$$l_c(\theta | x) = \log L_c(\theta | x) = \log g_C(x; \theta)$$

At the $(m+1)$th iteration,

Let $\theta^{(m)}$ denote the current estimate of $\theta$.

In the E-step, compute the expectation of the complete log-likelihood given the observed data $Y$ and the current estimate $\theta^{(m)}$. That is,

$$Q(\theta, \theta^{(m)}) = \mathbb{E}[l_c(\theta | x) | y; \theta^{(m)}]$$

$$= \int l_c(\theta | x) f(z | y; \theta^{(m)}) dz$$

In the M-step, choose $\theta^{(m+1)}$ to be any value of $\theta$ that maximises $Q(\theta, \theta^{(m)})$. That is,

$$Q(\theta^{(m+1)}, \theta^{(m)}) \geq Q(\theta, \theta^{(m)}) \quad \forall \theta \in \Omega$$

The E- and M-steps are repeated until either

$$l(\theta^{(m+1)} | y) - l(\theta^{(m)} | y)$$

changes by at most a specified small amount, or if
\[ \| \theta^{(m+1)} - \theta^{(m)} \| \]
is sufficiently small or if
\[ |Q(\theta^{(m+1)}, \theta^{(m)}) - Q(\theta^{(m)}, \theta^{(m)})| \]
is sufficiently small.

* Dempster, Laird and Rubin (1977) have also defined a GEM (Generalised EM) algorithm that differs from the EM algorithm in the M-step. With the GEM algorithm \( \theta^{(m+1)} \) is chosen not to globally maximise \( Q(\theta^m, \theta^{(m)}) \), but rather to ensure that \( Q(\theta^{(m+1)}, \theta^{(m)}) \) is greater than \( Q(\theta^m, \theta^{(m)}) \).

**Examples:**

In this section, we illustrate the use of the EM algorithm for three different types of models. Note however, that the first two examples can be solved directly quite easily and so there really is no need for the EM algorithm in these cases.

1. Genetic Linkage Model (Rao (1973); Dempster, Laird and Rubin (1977))

This example has been used on numerous occasions to demonstrate the EM algorithm.

Suppose that 197 animals are distributed into four categories as follows:
\[ y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34) \]
and \( y \) is postulated to have arisen from a multinomial distribution with cell probabilities
\[ \left( \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4} \right) \]

Augment the observed data by splitting the first cell into two cells with probabilities 1/2 and \( \theta/4 \) respectively.

The complete data vector is given now by
\[ x = (x_1, x_2, x_3, x_4, x_5) \]

where \( x_1 + x_2 = 125 \), \( x_3 = y_2 \), \( x_4 = y_3 \) and \( x_5 = y_4 \).

Now the observed likelihood is
The complete likelihood is

$$L_c(\theta \mid x) \propto (1 - \theta)^{(y_2 + y_3)} \theta^{x_2 + y_4}$$

$$l_c(\theta \mid x) = \log L_c(\theta \mid x) = \text{const} + (y_2 + y_3) \log(1 - \theta) + (x_2 + y_4) \log \theta,$$

which gives a "simplification" in the functional form.

For this genetic linkage model, in the E-step we compute

$$Q(\theta, \theta^{(m)}) = \mathbb{E}\left[(x_2 + y_4) \log \theta + (y_2 + y_3) \log(1 - \theta) \mid y; \theta^{(m)}\right] + \text{const}$$

$$= \mathbb{E}(x_2 \mid y; \theta^{(m)}) \log \theta + y_4 \log \theta + (y_2 + y_3) \log(1 - \theta) + \text{const}$$

Now,

$$X_2 \mid Y; \theta^{(m)} \sim \text{Binomial} \left\{ \frac{\theta^{(m)}}{\theta^{(m)} + 2}, \frac{4}{1 + \theta^{(m)}} \right\}$$

$$\therefore \mathbb{E}(X_2 \mid Y; \theta^{(m)}) = \frac{125 \theta^{(m)}}{\theta^{(m)} + 2}$$

So,

$$Q(\theta, \theta^{(m)}) = \frac{125 \theta^{(m)}}{\theta^{(m)} + 2} \log \theta + y_4 \log \theta + (y_2 + y_3) \log(1 - \theta) + \text{const}$$

For the M-step,

$$\frac{\partial Q(\theta, \theta^{(m)})}{\partial \theta} = 0 \Rightarrow \frac{125 \theta^{(m)}}{(\theta^{(m)} + 2)} \frac{1}{\theta} + \frac{y_4}{\theta} - \frac{(y_2 + y_3)}{(1 - \theta)} = 0$$

$$\therefore \theta^{(m+1)} = \frac{159 \theta^{(m)} + 68}{197 \theta^{(m)} + 144},$$

which can be solved iteratively. However, if we set $\theta^{(m+1)} = \theta^{(m)} = \theta^*$ say, then...
\[ \theta^* = \frac{159 \theta^* + 68}{197 \theta^* + 144} \]

\[ \vdots \]

\[ \theta^* = 0.627 \]

2. Censored (survival) data:

Let \( y_1, \ldots, y_m \) be the uncensored observed survival times and let \( y_{m+1}, \ldots, y_{m+n} \) be the censored survival times (i.e. all we know is that the unobserved survival times \( y_{m+1}, \ldots, y_{m+n} \) are greater than or equal to \( y_{m+1}, \ldots, y_{m+n} \)). Let the uncensored survival time data come from an exponential distribution with rate parameter \( \theta \). Denote this exponential density by \( g(y; \theta) \).

Then the observed likelihood is

\[
L(\theta \mid y) = \prod_{i=1}^{m} g(y_i; \theta) \prod_{j=m+1}^{m+n} S(y_j),
\]

where

\[
S(y) = \int_{y}^{\infty} g(u; \theta) du
\]

is the survivor function.

For each of the censored survival times \( y_{m+1}, \ldots, y_{m+n} \) there is the associated unobserved, uncensored survival times \( x_{m+1}, \ldots, x_{m+n} \).

Let \( x = (y_1, \ldots, y_m, x_{m+1}, \ldots, x_{m+n}) \) represent the complete data vector. Then the complete log-likelihood is

\[
L_c(\theta \mid x) = \prod_{i=1}^{m} g(y_i; \theta) \prod_{j=m+1}^{m+n} g(x_j; \theta)
\]

Therefore,

\[
l_c(\theta \mid x) = \sum_{i=1}^{m} \log \left\{ g(y_i; \theta) \right\} + \sum_{j=m+1}^{m+n} \log \left\{ g(x_j; \theta) \right\}
\]

At the \((k+1)\)th iteration, the E-step computes

\[
Q(\theta, \theta^{(k)}) = \sum_{i=1}^{m} \log \ g(y_i; \theta) + \sum_{j=m+1}^{m+n} E \left\{ \log \ g(x_j; \theta) \mid y; \theta^{(k)} \right\}
\]
Now,  
$$g(y; \theta) = \theta \exp(-\theta y)$$  
and  
$$S(y; \theta) = \exp(-\theta y)$$  
Therefore,  
$$Q(\theta, \theta^{(k)}) = m \log \theta - \theta \sum_{i=1}^{m} y_i + n \log \theta - \sum_{j=m+1}^{m+n} \theta \mathbb{E}[X_j | X_j \geq y_j; \theta^{(k)}]$$  
Now,  
$$\mathbb{E}[X_j | X_j \geq y_j; \theta^{(k)}] = y_j + \frac{1}{\theta^{(k)}}$$  
Therefore,  
$$Q(\theta, \theta^{(k)}) = m \log \theta - \theta \sum_{i=1}^{m} y_i + n \log \theta - \theta \sum_{j=m+1}^{m+n} \left(y_j + \frac{1}{\theta^{(k)}}\right)$$  
For the M-step,  
$$\frac{\partial Q(\theta, \theta^{(k)})}{\partial \theta} = 0 \Rightarrow m + n - \sum_{i=1}^{m} y_i - \frac{n}{\theta^{(k)}} = 0$$  
$$\Rightarrow \theta^{(k+1)} = \frac{(m + n)\theta^{(k)}}{\theta^{(k)} \sum_i y_i + n},$$  
which can be solved iteratively. In fact, if we set \(\theta^{(k+1)} = \theta^{(k)} = \theta^*\) say, then  
$$\theta^* = \frac{m}{\sum_i y_i}$$
3. Normal mixture model with two components:

Let \( y_1, \ldots, y_n \sim g(y; \theta) \), where

\[
g(y; \theta) = \pi g_1(y; \mu_1, \sigma_1^2) + (1 - \pi) g_2(y; \mu_2, \sigma_2^2)
\]

\( \theta = (\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \) and \( g_1(.) \) and \( g_2(.) \) are normal densities with different mean and variance parameters.

Now the observed likelihood from \( y = (y_1, \ldots, y_n) \) is given by

\[
L(\theta | y) = \prod_{i=1}^n \left[ \pi g_1(y_i; \mu_1, \sigma_1^2) + (1 - \pi) g_2(y_i; \mu_2, \sigma_2^2) \right]
\]

Define \( z_i = (z_{i1}, z_{i2}) \) where

\[
z_{i1} = \begin{cases} 1 & \text{if } y_i \in 1\text{st component} \\ 0 & \text{otherwise} \end{cases}
\]

and

\[
z_{i2} = \begin{cases} 1 & \text{if } y_i \in 2\text{nd component} \\ 0 & \text{otherwise} \end{cases}
\]

But \( z_{i1} = 1 - z_{i2} \), therefore we just need one of these component indicator variables, say \( z_{i1} \). In fact, call this indicator \( z_i \), for simplicity. Now define \( z = (z_1, \ldots, z_n) \).

Augment \( y \) with \( z \) to create \( x = (y, z) \). Then the complete likelihood is

\[
L_c(\theta | x) = \prod_{i=1}^n \left( \pi g_1(y_i; \mu_1, \sigma_1^2) \right)^{z_{i1}} \left( (1 - \pi) g_2(y_i; \mu_2, \sigma_2^2) \right)^{1 - z_{i1}}
\]

Therefore,

\[
l_c(\theta | x) = \sum_{i=1}^n \left[ z_i \left( \log \pi + \log g_1(y_i; \mu_1, \sigma_1^2) - \log(1 - \pi) - \log g_2(y_i; \mu_2, \sigma_2^2) \right) \right]
\]

\[+ \log(1 - \pi) + \log g_2(y_i; \mu_2, \sigma_2^2) \]
For the E-step, at the (m+1)th iteration, we would need to compute

\[
Q(\theta, \theta^{(m)}) = E c(\theta | x) | y; \theta^{(m)}]
\]

\[
= \sum_{i=1}^{n} \{E[z_i | y; \theta^{(m)}] \left[ \log \frac{\pi}{1-\pi} + \log \frac{g_1(y_i; \mu_1, \sigma_1^2)}{g_2(y_i; \mu_2, \sigma_2^2)} \right] + \log(1-\pi) + \log g_2(y_i; \mu_2, \sigma_2^2) \}
\]

Now,

\[
E[Z_i | Y; \theta^{(m)}] = \frac{\pi^{(m)} g_1(y_i; \mu_1^{(m)}, \sigma_1^{(m)})}{\pi^{(m)} g_1(y_i; \mu_1^{(m)}, \sigma_1^{(m)}) + (1-\pi^{(m)}) g_2(y_i; \mu_2^{(m)}, \sigma_2^{(m)})} = \omega_i^{(m)}
\]

Therefore,

\[
Q(\theta, \theta^{(m)}) = \sum_{i=1}^{n} \{\omega_i^{(m)} \left[ \log \frac{\pi}{1-\pi} + \log \frac{g_1(y_i; \mu_1, \sigma_1^2)}{g_2(y_i; \mu_2, \sigma_2^2)} \right] + \log(1-\pi) + \log g_2(y_i; \mu_2, \sigma_2^2) \}
\]

For the M-step,

\[
\frac{\partial Q}{\partial \pi} = 0 \Rightarrow \pi^{(m+1)} = \frac{\sum_{i=1}^{n} \omega_i^{(m)} \pi^{(m)}}{n}
\]

\[
\frac{\partial Q}{\partial \mu_1} = 0 \Rightarrow \mu_1^{(m+1)} = \frac{\sum_{i=1}^{n} \omega_i^{(m)} y_i}{n \pi^{(m+1)}}
\]

\[
\frac{\partial Q}{\partial \mu_2} = 0 \Rightarrow \mu_2^{(m+1)} = \frac{\sum_{i=1}^{n} (1-\omega_i^{(m)}) y_i}{n(1-\pi^{(m+1)})}
\]

\[
\frac{\partial Q}{\partial \sigma_1^2} = 0 \Rightarrow \sigma_1^{2(m+1)} = \frac{\sum_{i=1}^{n} (y_i - \mu_1^{(m+1)})^2 \omega_i^{(m)}}{n \pi^{(m+1)}}
\]

and

\[
\frac{\partial Q}{\partial \sigma_2^2} = 0 \Rightarrow \sigma_2^{2(m+1)} = \frac{\sum_{i=1}^{n} (y_i - \mu_2^{(m+1)})^2 (1-\omega_i^{(m)})}{n(1-\pi^{(m+1)})}
\]

which can be solved iteratively.
EM algorithm for the regular exponential family:

Observe that the complete data pdf's in Examples 1 and 2 are from the regular exponential family. In this family, a general formulation of the EM algorithm can be derived.

Let \( X^T = (Y^T, Z^T) \) be distributed (wlog) from

\[
g_c(X; \theta) = b(X) \exp(\theta^T t(X))/a(\theta),
\]

where \( \theta \) is the vector of natural parameters and \( t(X) \) is the vector of sufficient statistics.

The \( Q \)-function is given by

\[
Q(\theta, \theta^{(m)}) = \int \log g_c(X; \theta) f(Z | Y; \theta^{(m)}) dZ \\
= \int \log b(X) f(Z | Y; \theta^{(m)}) dZ + \theta^T \int t(X) f(Z | Y; \theta^{(m)}) dZ - \log a(\theta) \\
= \int \log b(X) f(Z | Y; \theta^{(m)}) dZ + \theta^T E[t(X) | Y; \theta^{(m)}] - \log a(\theta)
\]

Note that \( \int \log b(X) f(Z | Y; \theta^{(m)}) dZ \) does not depend on \( \theta \).

The E-step requires the computing of \( E[t(X) | Y; \theta^{(m)}] = t^{(m)} \), say.

The M-step consists of maximising \( Q(\theta, \theta^{(m)}) \) or \( -\log a(\theta) + \theta^T t^{(m)} \)

\[
\Rightarrow \frac{\partial Q(\theta, \theta^{(m)})}{\partial \theta} = 0 \Rightarrow \frac{\partial \log a(\theta)}{\partial \theta} = t^{(m)}
\]

Now recall that,

\[
a(\theta) = \int b(X) \exp(\theta^T t(X)) dX
\]

and so,

\[
\frac{\partial \log a(\theta)}{\partial \theta} = E[t(X) | \theta]
\]

Therefore, maximising \( Q(\theta, \theta^{(m)}) \) is equivalent to solving

\[
E[t(X) | \theta] = t^{(m)}
\]
EM algorithm for the finite mixture problem:

We now generalise Example 3.

Let \( X^T = (Y^T, Z^T) \) be the complete random data vector. \( Y \) is the observed data vector and \( Z \) the unobserved data vector.

Now assume that \( Y \) come from a finite mixture distribution with \( k \) components and \( k \) densities \( \{ g_j(\cdot; \psi_j) : j = 1, \ldots, k \} \). Let \( \{ \pi_j : j = 1, \ldots, k ; \sum \pi_j = 1 \} \) be the probabilities of a \( y_i \) belonging to the different components. Let \( \theta^T = (\pi^T, \psi^T) \).

Then the observed likelihood is

\[
L(\theta \mid y) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j g_j(y_i; \psi_j),
\]

which is difficult to maximise.

Define \( Z^T = (z_1^T, \ldots, z_n^T) \), where \( z_i^T = (z_{i1}, \ldots, z_{ik}) \) and \( z_{ip} = 1_{\{y_i = \text{pth component}\}} \)

Thus,

\[
g_c(x; \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{z_{ip}} g_j(y_i; \psi_j)^{z_{ip}} \tag{1}
\]

and

\[
l_c(\theta \mid x) = \log g_c(x; \theta) = \sum_{i=1}^{n} z_i^T \{ v(\pi) + u_i(\psi) \},
\]

where

\[
v(\pi)^T = (\log \pi_1, \ldots, \log \pi_k)
\]

\[
u_i(\psi)^T = (\log g_1(y_i; \psi_1), \ldots, \log g_k(y_i; \psi_k))
\]

In the E-step, we compute

\[
Q(\theta, \theta^{(m)}) = \sum_{i=1}^{n} w_i(\theta^{(m)})^T v(\pi) + \sum_{i=1}^{n} w_i(\theta^{(m)})^T u_i(\psi),
\]

where \( w_i(\theta^{(m)}) = E[z_i \mid y_i; \theta^{(m)}] \) and the jth component of \( w_i \) is
In the M-step, we now simply have to maximise the factorised likelihood (1) or equivalently the logarithm of this likelihood and then replace \( z_i \) by \( w_i(\theta^{(m)}) \) at the (m+1)th step. Alternatively, just maximise \( Q(\theta, \theta^{(m)}) \) at the (m+1)th iteration.

**Properties of the EM algorithm:**

1. **Stability/Monotonicity** - Each iteration of the EM algorithm, leads to an observed likelihood that is greater than or equal to the previous observed likelihood. That is, 
   \[ l(\theta^{(m+1)} \mid y) \geq l(\theta^{(m)} \mid y). \] (Left as an exercise to prove. Hint: Write the observed log-likelihood in terms of the complete log-likelihood and the conditional log-likelihood. Take expectations with respect to the conditional distribution and given \( \theta^{(m)} \). Apply Jensen’s inequality and also use concavity of the logarithmic function to obtain the required result.)

2. **Under suitable regularity conditions, if \( \theta^{(m)} \)'s converge then they converge to a stationary point of \( l(\theta \mid y) \). If they are multiple stationary points (local maxima or saddle points) then the algorithm may not converge to the global maximum. Therefore, it is good practice to try different starting values.

3. **Dempster, Laird and Rubin (1977)** showed that the EM algorithm converges at a linear rate, with the rate depending on the proportion of information about \( \theta \) in the observed density. As a consequence the EM algorithm can be quite slow. Hybrids of the EM algorithm (e.g. GEM using a Newton-Raphson step) have been proposed to accelerate convergence.

**Standard errors for the parameters:**

The EM algorithm does not automatically give standard errors for the parameter estimates, and this was one of the drawbacks initially. However, if we approximate the standard errors by choosing the appropriate elements of the inverted observed information matrix, then Louis (1982) derived the following result called the Missing Information Principle:

\[
l(\theta; y) = I_c(\theta; y) - I_m(\theta; y),
\]

where
\[
1(\theta; y) = \frac{\partial^2 \log L(\theta \mid y)}{\partial \theta \partial \theta^\top}
\]

\[
I_c(\theta; y) = \mathbb{E}\left\{ \frac{\partial^2 \log L_c(\theta \mid x)}{\partial \theta \partial \theta^\top} \mid y; \theta \right\} = \mathbb{E}\left\{ \frac{\partial^2 \log L_c(\theta \mid x)}{\partial \theta \partial \theta^\top} \right\}_{y; \theta}
\]

\[
I_m(\theta; y) = \mathbb{E}\left\{ \frac{\partial^2 \log g_m(x \mid y; \theta)}{\partial \theta \partial \theta^\top} \mid y; \theta \right\} = \mathbb{E}\left\{ \frac{\partial^2 \log g_m(x \mid y; \theta)}{\partial \theta \partial \theta^\top} \right\}_{y; \theta}
\]

and \( g_m(x \mid y; \theta) \) is the conditional distribution of \( X \mid Y \).

Louis (1982) showed that

\[
I_m(\theta; y) = \text{cov}\left\{ \frac{\partial \log L_c(\theta \mid x)}{\partial \theta} \mid y; \phi \right\}_{\hat{\theta}}
\]

and that

\[
I_m(\hat{\theta}; y) = \text{cov}\left\{ \frac{\partial \log L_c(\theta \mid x)}{\partial \theta} \mid y; \hat{\theta} \right\}_{\hat{\theta}} = \mathbb{E}\left\{ \frac{\partial \log L_c(\theta \mid x) \partial \log L_c(\theta \mid x)}{\partial \theta \partial \theta^\top} \right\}_{\hat{\theta}}
\]

since

\[
\left. \frac{\partial \log L(\theta \mid y)}{\partial \theta} \right|_{\hat{\theta}} = \mathbb{E}\left\{ \frac{\partial \log L_c(\theta \mid x)}{\partial \theta} \mid y; \hat{\theta} \right\}_{\hat{\theta}}
\]

and

\[
\left. \frac{\partial \log L(\theta \mid y)}{\partial \theta} \right|_{\hat{\theta}} = 0
\]

Therefore

\[
1(\hat{\theta}; y) = I_c(\hat{\theta}; y) - \text{cov}\left\{ \frac{\partial \log L_c(\theta \mid x)}{\partial \theta} \mid y; \hat{\theta} \right\}_{\hat{\theta}}
\]

\[
= I_c(\hat{\theta}; y) - \mathbb{E}\left\{ \frac{\partial \log L_c(\theta \mid x) \partial \log L_c(\theta \mid x)}{\partial \theta \partial \theta^\top} \right\}_{\hat{\theta}} ... (2)
\]

Therefore using (2) we can get the standard errors of the parameter estimates (by inverting).
Returning to Example 1 (Genetic Linkage),

\[
\frac{\partial \log L_c(\theta \mid x)}{\partial \theta} = \frac{x_2 + y_1}{\theta} - \frac{y_2 + y_3}{1 - \theta}
\]

\[
- \frac{\partial^2 \log L_c(\theta \mid x)}{\partial \theta \partial \theta^T} = \frac{x_2 + y_1}{\theta^2} + \frac{(y_2 + y_3)}{(1 - \theta)^2}
\]

Therefore,

\[
I_c(\hat{\theta}; y) = \frac{\text{E}[x_2 \mid y; \hat{\theta}] + y_1}{\hat{\theta}^2} + \frac{(y_2 + y_3)}{(1 - \hat{\theta})^2}
\]

\[
= \frac{125 \hat{\theta}}{\hat{\theta} + 2} + \frac{y_4}{\hat{\theta}^2} + \frac{y_2 + y_3}{(1 - \hat{\theta})^2} = 435.5
\]

and

\[
\text{cov} \left( \frac{\partial \log L_c(\theta \mid x)}{\partial \theta} \mid y; \hat{\theta} \right)_{\theta = \hat{\theta}} = \text{var}(x_2 \mid y; \hat{\theta}) = \frac{125}{\hat{\theta}^2} \left( \frac{\hat{\theta}}{\hat{\theta} + 2} \right) \left( \frac{2}{\hat{\theta} + 2} \right) = 57.8
\]

Thus,

\[
I(\hat{\theta}; y) = 435.5 - 57.8 = 377.7
\]

and the standard error of \( \hat{\theta} \) is equal to \( \sqrt{1/377.7} = 0.05 \).

References:


