1. **ML for general patterns of missing data**

Data are assumed MAR and the missing data parameters are distinct from the missing-data mechanism parameters. Therefore the missing data are ignorable in likelihood inference. In the following context, we consider how to obtain MLEs when the likelihood function cannot be factored into functions with distinctive parameters and explicit maximizers.

The ignorable likelihood

\[ L(\theta; Y_{obs}) = \int f(Y_{obs}, Y_{mis}; \theta) \, dY_{mis}. \]

When the \( L(\theta) \) is differentiable and unimodal, ML estimates can be found by solving the score equation

\[ D_l(\theta; Y_{obs}) = \frac{\partial \log L(\theta; Y_{obs})}{\partial \theta}. \]

When a closed-form solution of the above equation cannot be found, iterative methods can be applied.

1. Newton-Raphson (NR) algorithm. The iterative steps are

\[ \theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)} | Y_{obs}) \, D_l(\theta^{(t)}; Y_{obs}), \]

where \( I(\theta^{(t)} | Y_{obs}) = -\frac{\partial^2 l(\theta; Y_{obs})}{\partial \theta \partial \theta^T} \), is the observed information.
2. Scoring algorithm. The observed information is replaced by the expected information in the NR.

3. Other variants of NR such as the Quasi-Newton method.
2. **Introduction to the EM algorithm**

EM algorithm computes the MLEs for missing data problems when the complete-data likelihood has explicit maximizers. Compared with NR, EM is more stable that the likelihood function increases in the iterative steps but slower.

The idea was used several decades ago, for example, Buck’s regression imputation method for multivariate normal data with a general pattern. EM was formalized by Dempster, Laird and Rubin (1977) and began popular since then. Procedures for making inference in EM and variants, with faster convergence rates, of EM have been developed.

Each iteration of EM consists of two steps: M step and E step. In M step, update of the parameter estimates are obtained as if there were no missing data. In the subsequent E step, the sufficient statistics for the complete-data likelihood are imputed for the next iteration based on current parameter estimates.

Specifically, the E-step finds the expected complete-data loglikelihood function under current $\theta = \theta^{(t)}$

$$Q(\theta; \theta^{(t)}) = \int l(\theta; Y) f(Y_{mis} \mid Y_{obs}; \theta^{(t)}) dY_{mis} = E[l(\theta; Y) \mid Y_{obs}; \theta^{(t)}].$$

The M-step is achieved by finding

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

Therefore each iteration guarantees that $Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta^{(t)}; \theta^{(t)})$. 
3. Some examples

Example 8.1. Univariate normal data with missing values.

Example 8.2. The classical multinomial example.
$Y = (Y_1, Y_2, Y_3, Y_4) \sim \text{Multinomial}\{1/2 - \theta/2, \theta/4, \theta/4, 1/2\}$.

Example 8.3. Bivariate normal data with missing values on both variables.
4. Theory of the EM algorithm

The distribution of the complete data $Y$ can be factored as

$$f(Y; \theta) = f(Y_{\text{obs}}, Y_{\text{mis}}; \theta) = f(Y_{\text{obs}}; \theta) f(Y_{\text{mis}} \mid Y_{\text{obs}}; \theta)$$

The corresponding decomposition of the loglikelihood is

$$l(\theta; Y) = l(\theta; Y_{\text{obs}}, Y_{\text{mis}}) = l(\theta; Y_{\text{obs}}) + \log f(Y_{\text{mis}} \mid Y_{\text{obs}}; \theta).$$

When the missing-data mechanism is ignorable, the inference is based on the ignorable likelihood $l(\theta; Y_{\text{obs}})$. Given current estimate for $\theta$, say $\theta^{(t)}$, then

$$l(\theta; Y_{\text{obs}}) = l(\theta; Y) - \log f(Y_{\text{mis}} \mid Y_{\text{obs}}; \theta) = E\{l(\theta; Y) - \log f(Y_{\text{mis}} \mid Y_{\text{obs}}; \theta) \mid Y_{\text{obs}}; \theta^{(t)}\} \quad := Q(\theta; \theta^{(t)}) - H(\theta; \theta^{(t)}).$$

Note that for any $\theta$, $H(\theta; \theta^{(t)}) \leq H(\theta^{(t)}; \theta^{(t)})$. Therefore in EM algorithm,

$$l(\theta^{(t+1)}; Y_{\text{obs}}) - l(\theta^{(t)}; Y_{\text{obs}}) = \{Q(\theta^{(t+1)}; \theta^{(t)}) - Q(\theta^{(t)}; \theta^{(t)})\} - \{H(\theta^{(t+1)}; \theta^{(t)}) - H(\theta^{(t)}; \theta^{(t)})\} \geq 0.$$

A generalized EM (GEM) algorithm chooses $\theta^{(t+1)}$ so that $Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta^{(t)}; \theta^{(t)})$.

**Theorem 8.1.** Every GEM algorithm increases $l(\theta; Y_{\text{obs}})$ at each iteration with equality if and only if $Q(\theta^{(t+1)}; \theta^{(t)}) = Q(\theta^{(t)}; \theta^{(t)})$. 
Theorem 8.2. Suppose a sequence of EM iterates is such that

1. $D^{10}Q(\theta^{(t+1)}; \theta^{(t)}) = 0$, where $D^{10}$ means the first derivative with respect to the first argument.

2. $\theta^{(t)}$ converges to $\theta^*$, and

3. $f(Y_{mis} \mid Y_{obs}; \theta)$ is smooth in $\theta$. Then

$$D_t(\theta^*; Y_{obs}) = \frac{\partial}{\partial \theta} l(\theta; Y_{obs}) = 0,$$

which means that if $\{\theta^{(t)}\}$ converge, they converge to a stationary point of the loglikelihood.
5. **EM for Exponential families**

The complete data $Y$ have a distribution from a regular exponential family:

$$f(Y; \theta) = b(Y) \exp\{s(Y)\theta/a(\theta)\}.$$ 

E-step: calculate $s^{(t+1)} = E(s(Y) \mid Y_{\text{obs}}; \theta^{(t)})$.

M-step: solve equation $E(s(Y) \mid \theta) = s^{(t+1)}$.

The above equation often has an explicit solution.

**Example 8.4. ML estimation for a sample from the univariate $t$ distribution with known degrees of freedom.** Suppose that the observed data, $Y_{\text{obs}}$, consist of a random sample $X = (x_1, x_2, \ldots, x_n)$ from a Student’s $t$ distribution with center $\mu$, scale parameter $\sigma$, and known degrees of freedom $\nu$, with density

$$f(x_i; \theta) = \frac{\Gamma(\nu/2 + 1/2)}{(\pi \nu \sigma^2)^{1/2} \Gamma(\nu/2)} \frac{1}{\{1 + (x_i - \mu)^2/(\nu \sigma^2)\}^{(\nu+1)/2}}.$$ 

ML estimation requires an iterative algorithm. An augmented complete data set can be defined as $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{obs}} = X$ and $Y_{\text{mis}} = W = (w_1, w_2, \ldots, w_n)$ is a vector of unobserved positive quantities, such that pairs $(w_i, x_i)$ are independent across units $i$, with distribution

$$(x_i \mid w_i; \theta) \sim N(\mu, \sigma^2/w_i), \quad (w_i; \theta) \sim \chi^2_{\nu}/\nu.$$
6. Rate of convergence of EM

From \( l(\theta; Y_{obs}) = l(\theta; Y) - \log f(Y_{mis} \mid Y_{obs}; \theta) \), take differentiation twice on \( \theta \)

\[
I(\theta; Y_{obs}) = I(\theta; Y_{obs}, Y_{mis}) + \partial^2 \log f(Y_{mis} \mid Y_{obs}; \theta) / \partial \theta \partial \theta^T.
\]

Then taking expectations over the distribution of \( Y_{mis} \) given \( Y_{obs} \)

\[
I(\theta; Y_{obs}) = -D^{20} Q(\theta; \theta) + D^{20} H(\theta; \theta).
\]

Suppose the EM algorithm converges at \( \theta^* \), name

(a) The complete information: \( i_{\text{com}} = -D^{20} Q(\theta; \theta) \mid _{\theta=\theta^*} \).

(b) The observed information: \( i_{\text{obs}} = I(\theta; Y_{obs}) \mid _{\theta=\theta^*} \).

(c) The missing information: \( i_{\text{com}} = -D^{20} H(\theta; \theta) \mid _{\theta=\theta^*} \).

Therefore \( i_{\text{obs}} = i_{\text{com}} - i_{\text{mis}} \). Let

\[
DM = i_{\text{mis}}i_{\text{com}}^{-1} = I - i_{\text{obs}}i_{\text{com}}^{-1},
\]

where the matrix \( DM \) represents the fraction of missing information. It’s shown that when \( \theta(t) \) is near \( \theta^* \),

\[
| \theta^{(t+1)} - \theta^* | = \lambda | \theta^{(t)} - \theta^* |,
\]

where \( \lambda = DM \) for scalar \( \theta \) or the largest eigenvalue of \( DM \), for vector \( \theta \).

HW6: 7.16, 7.17, 8.6, 8.7, 8.8, 8.12. Due: 10/20

Biostat 2065: Analysis of Incomplete Data, 2005