EM Algorithm

2012-9-18

Yijuan Hu

Last lecture

- General optimization problems
  - Newton Raphson
  - Fisher scoring
  - Quasi Newton
- Nonlinear regression models
  - Gauss-Newton
- Generalized linear models
  - Iteratively reweighted least squares

Motivating example of EM algorithm

ABO blood groups

Genotype	Genotype Frequency	Phenotype
AA	p_{A}^2	**A**
AO	$2p_{\mathrm{A}}p_{\mathrm{O}}$	**A**
BB	p_{B}^2	**B**
BO	$2p_{\mathrm{B}}p_{\mathrm{O}}$	**B**
OO	p_{O}^2	**O**
AB	$2p_{\mathrm{A}}p_{\mathrm{B}}$	**AB**

- The genotype frequencies above assume Hardy-Weinberg equilibrium:
- For a random sample of n individuals, we observe their phenotype, but not their genotype.
- We wish to obtain the MLEs of the underlying allele frequencies p_{A}, p_{B}, and $p_{\mathrm{O}} = 1 - p_{\mathrm{A}} - p_{\mathrm{B}}$.

$$L(p_{\mathrm{A}}, p_{\mathrm{B}}) = (p_{\mathrm{A}}^2 + 2p_{\mathrm{A}}p_{\mathrm{O}})^{n_{\mathrm{A}}} \times (p_{\mathrm{B}}^2 + 2p_{\mathrm{B}}p_{\mathrm{O}})^{n_{\mathrm{B}}} \times (p_{\mathrm{O}}^2)^{n_{\mathrm{O}}} \times (2p_{\mathrm{A}}p_{\mathrm{B}})^{n_{\mathrm{AB}}}$$

We could, of course, form the likelihood function and find its maximum by Newton-Raphson.

$$\dot{l}(p_{\mathrm{A}}, p_{\mathrm{B}}) =$$

$$\ddot{l}(p_{\mathrm{A}}, p_{\mathrm{B}}) =$$

But there is an easier “allele counting” algorithm.

Let n_{A}, n_{B}, n_{O}, n_{AB} be the observed numbers of individuals with phenotypes A, B, O, AB, respectively.

Let n_{AA}, n_{AO}, n_{BB} and n_{BO} be the unobserved numbers of individuals with genotypes AA, AO, BB and BO, respectively. They satisfy $n_{\mathrm{AA}} + n_{\mathrm{AO}} = n_{\mathrm{A}}$ and $n_{\mathrm{BB}} + n_{\mathrm{BO}} = n_{\mathrm{B}}$.

1. Start with initial estimates $p^{(0)} = (p_{\mathrm{A}}^{(0)}, p_{\mathrm{B}}^{(0)}, p_{\mathrm{O}}^{(0)})$
2. Calculate the expected n_{AA} and n_{BB}, given observed data and $p^{(k)}$

$$n_{\mathrm{AA}}^{(k+1)} = \mathrm{E}(n_{\mathrm{AA}} | n_{\mathrm{A}}, p^{(k)}) = n_{\mathrm{A}} \frac{p_{\mathrm{A}}^{(k)} p_{\mathrm{A}}^{(k)}}{p_{\mathrm{A}}^{(k)} p_{\mathrm{A}}^{(k)} + 2 p_{\mathrm{O}}^{(k)} p_{\mathrm{A}}^{(k)}}, \qquad n_{\mathrm{BB}}^{(k+1)} = ?$$

3. Update $p^{(k+1)}$, imaging that $n_{\mathrm{AA}}^{(k+1)}$ and $n_{\mathrm{BB}}^{(k+1)}$ were actually observed

$$p_{\mathrm{A}}^{(k+1)} = (2 n_{\mathrm{AA}}^{(k+1)} + n_{\mathrm{AO}}^{(k+1)} + n_{\mathrm{AB}})/(2n), \qquad p_{\mathrm{B}}^{(k+1)} = ?$$

4. Repeat step 2 and 3 until the estimates converge

EM algorithm: Applications

Expectation-**M**maximization **algorithm** (*Dempster, Laird, & Rubin, 1977, JRSSB, 39:1–38*) is a general iterative algorithm for parameter estimation by maximum likelihood (optimization problems).

It is useful when

- some of the random variables involved are not observed, i.e., considered missing or incomplete.
- direct maximizing the target likelihood function is difficult, but one can introduce (missing) random variables so that maximizing the complete-data likelihood is simple.

Typical problems include:

- Filling in missing data in a sample
- Discovering the value of latent variables
- Estimating parameters of HMMs
- Estimating parameters of finite mixtures

Description of EM

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

It can be difficult to find MLE $\hat{\theta} = \arg\max_\theta g(Y_{\text{obs}}|\theta) = \arg\max_\theta \int f(Y_{\text{obs}}, y_{\text{mis}}|\theta)\, dy_{\text{mis}}$

But it could be easy to find $\hat{\theta}_{\text{C}} = \arg\max_\theta f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$, if we had observed Y_{mis}.

- **E step**: $h^{(k)}(\theta) \equiv \text{E}\left\{\log f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)\middle| Y_{\text{obs}}, \theta^{(k)}\right\}$
- **M step**: $\theta^{(k+1)} = \arg\max_\theta h^{(k)}(\theta)$;
  can also use one-step Newton-Raphson iteration (EM extensions later)

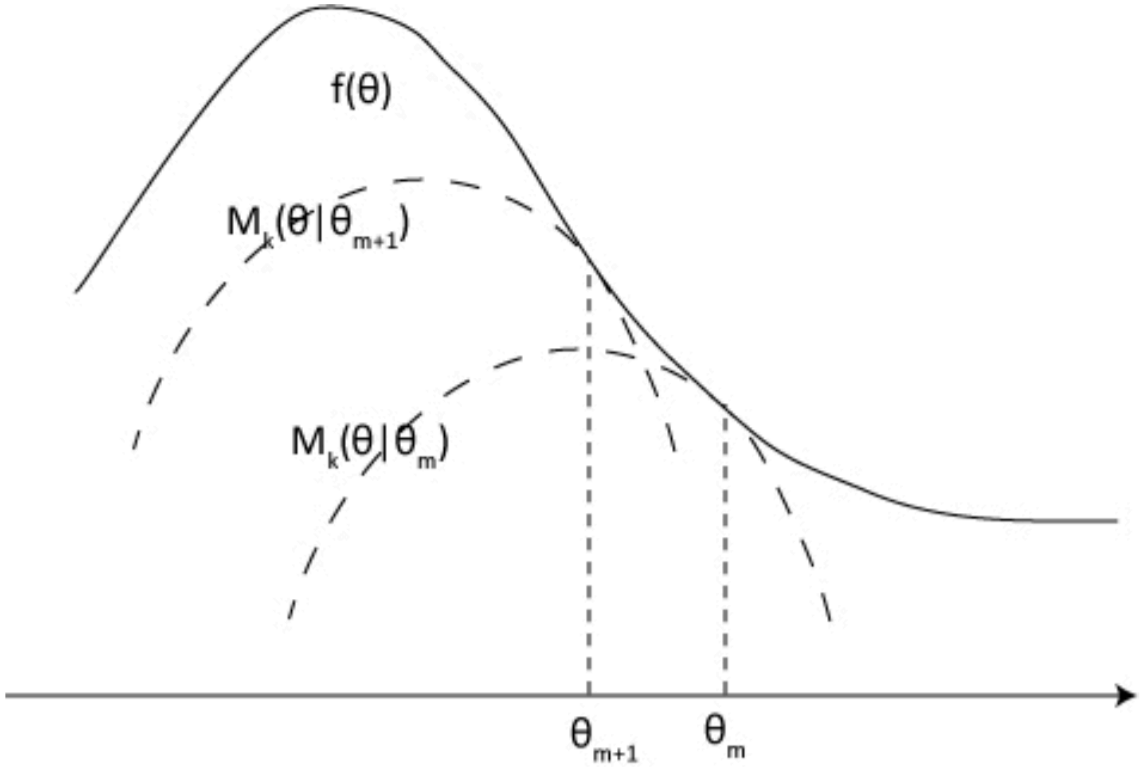
Nice properties (compared to Newton-Raphson):

1. simplicity of implementation
2. stable monotone convergence

Justification of EM

The E-step creates a surrogate function by identifying a complete-data log-likelihood function and evaluating it respect to the observed data.

The M-step maximizes the surrogate function.



Ascent property of EM

Theorem: At each iteration of the EM algorithm,

$$\log g(Y_{\text{obs}}|\theta^{(k+1)}) \geq \log g(Y_{\text{obs}}|\theta^{(k)})$$

and the equality holds if and only if $\theta^{(k+1)} = \theta^{(k)}$.

Proof: The definition of $\theta^{(k+1)}$ gives

$$\text{E}\{\log f(Y_{\text{obs}}, Y_{\text{mis}}|\theta^{(k+1)})|Y_{\text{obs}}, \theta^{(k)}\} \geq \text{E}\{\log f(Y_{\text{obs}}, Y_{\text{mis}}|\theta^{(k)})|Y_{\text{obs}}, \theta^{(k)}\},$$

which can be expanded to

$$\text{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k+1)})|Y_{\text{obs}}, \theta^{(k)}\}+\log g(Y_{\text{obs}}|\theta^{(k+1)}) \geq \text{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})|Y_{\text{obs}}, \theta^{(k)}\}+\log g(Y_{\text{obs}}|\theta^{(k)}). \quad (1)$$

By the non-negativity of the Kullback-leibler information, i.e.,

$$\int p(x) \log \frac{p(x)}{q(x)} \geq 0, \quad \text{for densities } p(x), q(x),$$

we have

$$\text{E}\left[\log \frac{c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})}{c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k+1)})}\middle| Y_{\text{obs}}, \theta^{(k)}\right] = \int \log \frac{f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})}{f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k+1)})} f(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})\, dy_{\text{mis}} \geq 0. \quad (2)$$

Combining (1) and (2) yields

$$\log g(Y_{\text{obs}}|\theta^{(k+1)}) \geq \log g(Y_{\text{obs}}|\theta^{(k)}),$$

thus we partially proved the theorem. If the equality holds, i.e.,

$$\log g(Y_{\text{obs}}|\theta^{(k+1)}) = \log g(Y_{\text{obs}}|\theta^{(k)}), \tag{3}$$

by (1) and (2),

$$\text{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k+1)})|Y_{\text{obs}}, \theta^{(k)}\} = \text{E}\{\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)})|Y_{\text{obs}}, \theta^{(k)}\}.$$

The Kullback-leibler information is zero if and only if

$$\log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k+1)}) = \log c(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(k)}). \tag{4}$$

Combining (3) and (4), we have

$$\log f(Y|\theta^{(k+1)}) = \log f(Y|\theta^{(k)}).$$

The uniqueness of θ leads to $\theta^{(k+1)} = \theta^{(k)}$.

Example 1: Grouped Multinomial Data

Suppose $Y = (y_1, y_2, y_3, y_4)$ has a multinomial distribution with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

Then the probability for Y is given by

$$L(\theta|Y) \equiv \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}.$$

If we use **Newton-Raphson** to directly maximize $f(Y, \theta)$, we need

$$\dot{l}(\theta|Y) = \frac{y_1/4}{1/2 + \theta/4} - \frac{y_2 + y_3}{1-\theta} + \frac{y_4}{\theta}$$
$$\ddot{l}(\theta|Y) = -\frac{y_1}{(2+\theta)^2} - \frac{y_2 + y_3}{(1-\theta)^2} - \frac{y_4}{\theta^2}$$

The probability of the first cell is a trouble-maker!

How to avoid?

Suppose $Y = (y_1, y_2, y_3, y_4)$ has a multinomial distribution with cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

Define the complete-data: $X = (x_0, x_1, y_2, y_3, y_4)$ to have a multinomial distribution with probabilities

$$\left(\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right),$$

and to satisfy

$$x_0 + x_1 = y_1$$

Observed-data log likelihood

$$l(\theta|Y) \equiv \; y_1 \log\left(\frac{1}{2} + \frac{\theta}{4}\right) + (y_2 + y_3)\log(1-\theta) + y_4 \log\theta$$

Complete-data log likelihood

$$l_C(\theta|X) \equiv \; (x_1 + y_4)\log\theta + (y_2 + y_3)\log(1-\theta)$$

E step: evaluate

$$x_1^{(k+1)} = \mathrm{E}(x_1|Y, \theta^{(k)}) = y_1 \frac{\theta^{(k)}/4}{1/2 + \theta^{(k)}/4}$$

M step: maximize complete-data log likelihood with x_1 replaced by $x_1^{(k+1)}$

$$\theta^{(k+1)} = \frac{x_1^{(k+1)} + y_4}{x_1^{(k+1)} + y_4 + y_2 + y_3}$$

We observe $Y = (125, 18, 20, 34)$ and start EM with $\theta^{(0)} = 0.5$.

k	Parameter update $\theta^{(k)}$	Convergence to $\hat{\theta}$ $\theta^{(k)} - \hat{\theta}$	Convergence rate $(\theta^{(k)} - \hat{\theta})/(\theta^{(k-1)} - \hat{\theta})$
0	.500000000	.126821498	
1	.608247423	.018574075	.1465
2	.624321051	.002500447	.1346
3	.626488879	.000332619	.1330
4	.626777323	.000044176	.1328
5	.626815632	.000005866	.1328
6	.626820719	.000000779	.1328
7	.626821395	.000000104	
8	.626821484	.000000014	
$\hat{\theta}$	.626821498	Stop	

Example 2: Normal mixtures

Consider $x_1, \ldots, x_n \sim \sum_{j=1}^{J} p_j \phi(x_i|\mu_j, \sigma_j)$, where $\phi(.|\mu, \sigma)$ is the normal density.

Suppose that $(y_{i1}, y_{i2}, \ldots, y_{iJ})$ follows a multinomial distribution with cel probabilities $\boldsymbol{p} = (p_1, p_2, \ldots, p_J)$. Clearly, $\sum_j y_{ij} = 1$. Given $y_{ij*} = 1$ and $y_{ij} = 0$ for $j \neq j*$, we assume

$$x_i \sim N(\mu_{j*}, \sigma_{j*}).$$

You can check, marginally, $x_i \sim \sum_{j=1}^{J} p_j \phi(x_i|\mu_j, \sigma_j)$.

$\{x_i\}_i$ is the observed data; $\{x_i, y_{i1}, \ldots, y_{iJ}\}_i$ is the complete data.

This is the clustering/finite mixture problem in which EM is typically used for.

Observed-data log likelihood

$$l(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{p}|x) \equiv \sum_i \log \left\{ \sum_{j=1}^{J} p_j \phi(x_i|\mu_j, \sigma_j) \right\}$$

Complete-data log likelihood

$$l_C(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{p}|x, y) \equiv \sum_{ij} y_{ij} \left\{ \log p_j + \log \phi(x_i|\mu_j, \sigma_j) \right\}$$

Example 2: Normal mixtures (continued)

Complete-data log likelihood:

$$l_C(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{p}|x, y) \equiv \sum_{ij} y_{ij}\{\log p_j - (x_i - \mu_j)^2/(2\sigma_j^2) - \log \sigma_j\}$$

E step: evaluate for $i = 1, \ldots, n$ and $j = 1, \ldots, J$,

$$\begin{aligned}
\omega_{ij}^{(k)} &\equiv \mathrm{E}(y_{ij}|x_i, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \boldsymbol{p}^{(k)}) \\
&= \Pr(y_{ij} = 1|x_i, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \boldsymbol{p}^{(k)}) \\
&= \frac{p_j^{(k)} f(x_i|\mu_j^{(k)}, \sigma_j^{(k)})}{\sum_j p_j^{(k)} f(x_i|\mu_j^{(k)}, \sigma_j^{(k)})}
\end{aligned}$$

M step: maximize complete-data log likelihood with y_{ij} replaced by ω_{ij}

$$p_j^{(k+1)} = n^{-1} \sum_i \omega_{ij}^{(k)}$$

$$\mu_j^{(k+1)} = \sum_i \omega_{ij}^{(k)} x_i \Big/ \sum_i \omega_{ij}^{(k)}$$

$$\sigma^{(k+1)} = \sqrt{\sum_j \left\{ \sum_i \omega_{ij}^{(k)} x_i^2 - \left(\sum_i \omega_{ij}^{(k)} x_i \right)^2 \sum_i \omega_{ij}^{(k)} \right\} \Big/ n}$$

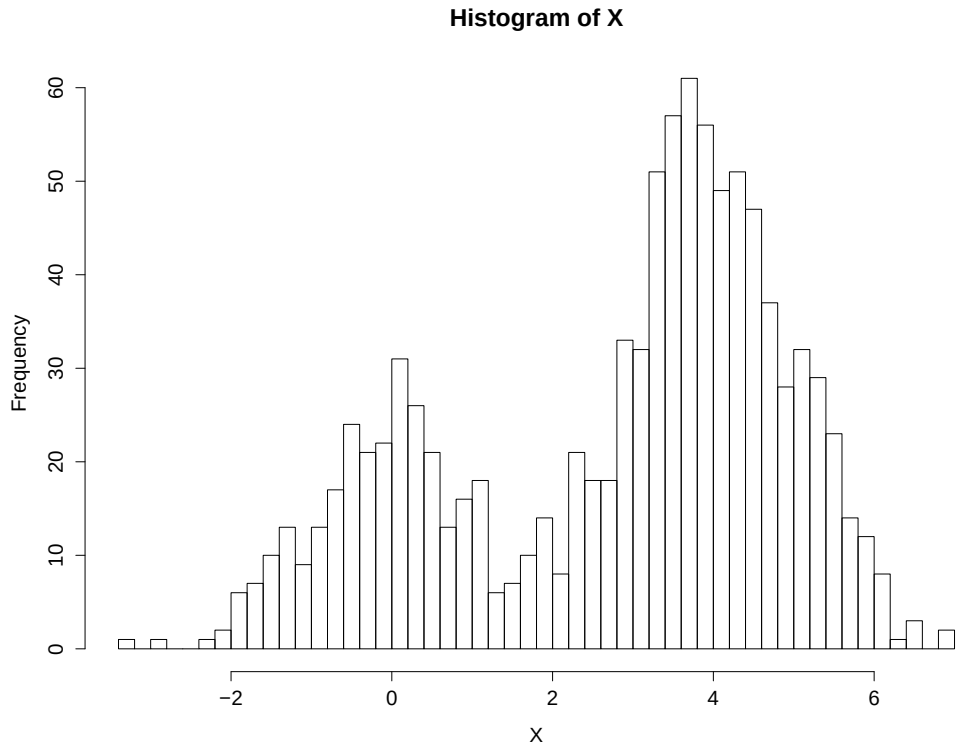
Example 2: Normal mixtures in R


```
### two component EM
### pN(0,1)+(1-p)N(4,1)

EM_TwoMixtureNormal = function(p, mu1, mu2, sd1, sd2, X, maxiter=1000, tol=1e-5)
{
    diff=1
    iter=0

    while (diff>tol & iter<maxiter) {

        ## E-step: compute omega:
        d1=dnorm(X, mean=mu1, sd=sd1)   # compute density in two groups
        d2=dnorm(X, mean=mu2, sd=sd2)
        omega=d1*p/(d1*p+d2*(1-p))

        ## M-step: update p, mu and sd
        p.new=mean(omega)
        mu1.new=sum(X*omega) / sum(omega)
        mu2.new=sum(X*(1-omega)) / sum(1-omega)
        resid1=X-mu1
        resid2=X-mu2;
```

```
        sd1.new=sqrt(sum(resid1^2*omega) / sum(omega))
        sd2.new=sqrt(sum(resid2^2*(1-omega)) / sum(1-omega))

        ## calculate diff to check convergence
        diff=sqrt(sum((mu1.new-mu1)^2+(mu2.new-mu2)^2
                                +(sd1.new-sd1)^2+(sd2.new-sd2)^2))

        p=p.new;
        mu1=mu1.new;
        mu2=mu2.new;
        sd1=sd1.new;
        sd2=sd2.new;

        iter=iter+1;

        cat("Iter", iter, ": mu1=", mu1.new, ", mu2=",mu2.new, ", sd1=",sd1.new,
            ", sd2=",sd2.new, ", p=", p.new, ", diff=", diff, "\n")
    }

}
```

Example 2: Normal mixtures in R (continued)


```
> ## simulation
> p0=0.3;
> n=5000;
> X1=rnorm(n*p0);                 # n*p0 indiviudals from N(0,1)
> X2=rnorm(n*(1-p0), mean=4)      # n*(1-p0) individuals from N(4,1)
> X=c(X1,X2)                      # observed data
> hist(X, 50)
```



Example 2: Normal mixtures in R (continued)

```
> ## initial values for EM
> p=0.5
> mu1=quantile(X, 0.1);
> mu2=quantile(X, 0.9)
> sd1=sd2=sd(X)

> c(p, mu1, mu2, sd1, sd2)
0.5000000 -0.3903964  5.0651073  2.0738555  2.0738555


> EM_TwoMixtureNormal(p, mu1, mu2, sd1, sd2, X)
Iter 1: mu1=0.8697, mu2=4.0109, sd1=2.1342, sd2=1.5508, p=0.3916, diff=1.7252
Iter 2: mu1=0.9877, mu2=3.9000, sd1=1.8949, sd2=1.2262, p=0.3843, diff=0.4345
Iter 3: mu1=0.8353, mu2=4.0047, sd1=1.7812, sd2=1.0749, p=0.3862, diff=0.2645
Iter 4: mu1=0.7203, mu2=4.0716, sd1=1.6474, sd2=0.9899, p=0.3852, diff=0.2070
...
Iter 44: mu1=-0.0048, mu2=3.9515, sd1=0.9885, sd2=1.0316, p=0.2959, diff=1.9e-05
Iter 45: mu1=-0.0048, mu2=3.9515, sd1=0.9885, sd2=1.0316, p=0.2959, diff=1.4e-05
Iter 46: mu1=-0.0049, mu2=3.9515, sd1=0.9885, sd2=1.0316, p=0.2959, diff=1.1e-05
Iter 47: mu1=-0.0049, mu2=3.9515, sd1=0.9885, sd2=1.0316, p=0.2959, diff=8.7e-06
```

Example 3: Mixed-effects model

For a longitudinal dataset of $i = 1, \ldots, N$ subjects, each with n_i measurements of the phenotype, the linear mixed effect model is given by

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad b_i \sim N_q(0, D), \quad \epsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i}), \quad b_i, \epsilon_i \text{ independent}$$

Observed-data log-likelihood

$$l(\beta, D, \sigma^2 | Y_1, \ldots, Y_N) \equiv \sum_i \left\{ -\frac{1}{2}(Y_i - X_i\beta)'\Sigma_i^{-1}(Y_i - X_i\beta) - \frac{1}{2}\log|\Sigma_i| \right\},$$

where $\Sigma_i = Z_i D Z_i' + \sigma^2 I_{n_i}$.

- In fact, this likelihood can be directly maximized for (β, D, σ^2) by using Newton-Raphson or Fisher scoring.
- Given (D, σ^2) and hence Σ_i, we obtain β that maximizes the likelihood by solving

$$\frac{\partial l(\beta, D, \sigma^2 | Y_1, \ldots, Y_N)}{\partial \beta} = \sum_i X_i'\Sigma^{-1}(Y_i - X_i\beta) = 0,$$

which implies

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

Complete-data log-likelihood

Note the equivalence of (ϵ_i, b_i) and (Y_i, b_i) and the fact that

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

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

The parameter that maximizes the complete-data log-likelihood is obtained as, conditional on other parameters,

$$D = N^{-1} \sum_{i=1}^{N} b_i b_i'$$

$$\sigma^2 = \left(\sum_{i=1}^{N} n_i \right)^{-1} \sum_{i=1}^{N} \epsilon_i' \epsilon_i$$

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

E step: to evaluate

$$\mathrm{E}\left(b_i b_i' \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)}\right)$$
$$\mathrm{E}\left(\epsilon_i' \epsilon \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)}\right)$$
$$\mathrm{E}\left(b_i \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)}\right)$$

We use the relationship

$$\mathrm{E}(b_i b_i' \mid Y_i) = \mathrm{E}(b_i \mid Y_i)\mathrm{E}(b_i' \mid Y_i) + \mathrm{Var}(b_i \mid Y_i).$$

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

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} = N\left\{\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} Z_i D Z_i' + \sigma^2 I_{n_i} & Z_i D \\ D Z_i' & D \end{pmatrix}\right\},$$

Let $\Sigma_i = Z_i D Z_i' + \sigma^2 I_{n_i}$. We known that

$$\mathrm{E}(b_i \mid Y_i) = 0 + D Z_i' \Sigma_i^{-1}(Y_i - X_i\beta)$$
$$\mathrm{Var}(b_i \mid Y_i) = D - D Z_i' \Sigma_i^{-1} Z_i D.$$

Similarly, We use the relationship

$$\mathrm{E}(\epsilon_i'\epsilon_i \mid Y_i) = \mathrm{E}(\epsilon_i' \mid Y_i)\mathrm{E}(\epsilon_i \mid Y_i) + \mathrm{Var}(\epsilon_i \mid Y_i).$$

We can derive

$$\begin{pmatrix} Y_i \\ \epsilon_i \end{pmatrix} = N\left\{\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} Z_iDZ_i' + \sigma^2 I_{n_i} & \sigma^2 I_{n_i} \\ \sigma^2 I_{n_i} & \sigma^2 I_{n_i} \end{pmatrix}\right\}.$$

Let $\Sigma_i = Z_iDZ_i' + \sigma^2 I_{n_i}$. Then we have

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

M step

$$\begin{aligned} D^{(k+1)} &= N^{-1}\sum_{i=1}^{N} \mathrm{E}(b_ib_i' \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)})) \\ \sigma^{2(k+1)} &= \left(\sum_{i=1}^{N} n_i\right)^{-1} \sum_{i=1}^{N} \mathrm{E}(\epsilon_i'\epsilon_i \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)}) \\ \beta^{(k+1)} &= \left(\sum_{i=1}^{N} X_i'X_i\right)^{-1} \sum_{i=1}^{N} X_i'\mathrm{E}(Y_i - Z_ib_i \mid Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)}). \end{aligned}$$

1. Stopping rules

- $|l(\theta^{(k+1)}) - l(\theta^{(k)})| < \epsilon$ for m consecutive steps, where $l(\theta)$ is observed-data log-likelihood

  This is **bad**! $l(\theta)$ may not change much even when θ does.

- $||\theta^{(k+1)} - \theta^{(k)}|| < \epsilon$ for m consecutive steps

  This runs into problems when the components of θ are of quite different magnitudes.

- $|\theta_j^{(k+1)} - \theta_j^{(k)}| < \epsilon_1(|\theta_j^{(k)}| + \epsilon_2)$ for $j = 1, \ldots, p$

  In practice, take

$$\epsilon_1 = 10^{-8}$$
$$\epsilon_2 = 10\epsilon_1 \text{ to } 100\epsilon_1$$

2. Local vs. global max

- There may be multiple modes
- EM may converge to a saddle point
- **Solution**: Multiple starting points

3. Starting points

- Use information from the context
- Use a crude method (such as the method of moments)
- Use an alternative model formulation

4. Slow convergence

- EM can be painfully slow to converge near the maximum
- **Solution**: Switch to another optimization algorithm when you get near the maximum

5. Standard errors

- Numerical approximation of the Hessian matrix
- Louis (1982), Meng and Rubin (1991)

Note: $l(\theta) =$ observed-data log-likelihood

We estimate the gradient using

$$\{\dot{l}(\theta)\}_i = \frac{\partial l(\theta)}{\partial \theta_i} \approx \frac{l(\theta + \delta_i e_i) - l(\theta - \delta_i e_i)}{2\delta_i}$$

where e_i is a unit vector with 1 for the ith element and 0 otherwise.

In calculating derivatives using this formula, I generally start with some medium size δ and then repeatedly halve it until the estimated derivative stabilizes.

We can estimate the Hessian by applying the above formula twice:

$$\{\ddot{l}(\theta)\}_{ij} \approx \frac{l(\theta + \delta_i e_i + \delta_j e_j) - l(\theta + \delta_i e_i - \delta_j e_j) - l(\theta - \delta_i e_i + \delta_j e_j) + l(\theta - \delta_i e_i - \delta_j e_j)}{4\delta_i \delta_j}$$

Louis estimator (1982)

$$l_C(\theta|Y_{obs}, Y_{mis}) \equiv \log\{f(Y_{obs}, Y_{mis}|\theta)\}$$

$$l_O(\theta|Y_{obs}) \equiv \log\left\{\int f(Y_{obs}, y_{mis}|\theta)\, dy_{mis}\right\}$$

$$\dot{l}_C(\theta|Y_{obs}, Y_{mis}),\ \dot{l}_O(\theta|Y_{obs}) = \text{gradients of } l_C,\ l_O$$

$$\ddot{l}_C(\theta|Y_{obs}, Y_{mis}),\ \ddot{l}_O(\theta|Y_{obs}) = \text{second derivatives of } l_C,\ l_O$$

We can prove that

$$(5) \quad \dot{l}_O(\theta|Y_{obs}) = E\left\{\dot{l}_C(\theta|Y_{obs}, Y_{mis})|Y_{obs}\right\}$$

$$(6) \quad -\ddot{l}_O(\theta|Y_{obs}) = E\left\{-\ddot{l}_C(\theta|Y_{obs}, Y_{mis})|Y_{obs}\right\} - E\left\{\left[\dot{l}_C(\theta|Y_{obs}, Y_{mis})\right]^{\otimes 2}\Big|Y_{obs}\right\} + \left[\dot{l}_O(\theta|Y_{obs})\right]^{\otimes 2}$$

- **MLE**: $\hat{\theta} = \arg\max_\theta \dot{l}_O(\theta|Y_{obs})$
- **Louis variance estimator**: $\left\{-\ddot{l}_O(\theta|Y_{obs})\right\}^{-1}$ evaluated at $\theta = \hat{\theta}$
- **Note**: All of the conditional expectations can be computed in the EM algorithm using only $\dot{l}_C$ and $\ddot{l}_C$, which are first and second derivatives of the complete-data log-likelihood. Louis estimator should be evaluated at the last step of EM.

Proof of (5)

Proof: By the definition of $l_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$,

$$\begin{aligned} i_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) &= \frac{\partial \log\left\{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}\right\}}{\partial\theta} \\ &= \frac{\partial \int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}/\partial\theta}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \\ &= \frac{\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}. \end{aligned} \tag{7}$$

Multiplying and dividing the integrand of the numerator by $f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)$ gives (5),

$$\begin{aligned} i_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) &= \frac{\int \frac{f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)}{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)} f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \\ &= \frac{\int \frac{\partial \log\{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\}}{\partial\theta} f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} \\ &= \int i_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}) \frac{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}\, dy_{\mathrm{mis}} \\ &= \mathrm{E}\left\{ i_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}) | Y_{\mathrm{obs}} \right\}. \end{aligned}$$

Proof of (6)

Proof: We take an additional derivative of $\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})$ in expression (7) to obtain

$$\begin{aligned}\ddot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}}) &= \frac{\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} - \left\{\frac{\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}\right\}^2 \\ &= \frac{\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}}{\int f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}} - \left\{\dot{l}_{\mathrm{O}}(\theta|Y_{\mathrm{obs}})\right\}^{\otimes 2}.\end{aligned}$$

To see how the first term breaks down, we take an additional derivative of

$$\int f'(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}} = \int \frac{\partial \log\{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\}}{\partial\theta} f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}$$

to obtain

$$\begin{aligned}\int f''(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}} &= \int \frac{\partial^2 \log\{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\}}{\partial\theta\partial\theta'} f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}} \\ &\quad + \int \left[\frac{\partial \log\{f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\}}{\partial\theta}\right]^{\otimes 2} f(Y_{\mathrm{obs}}, y_{\mathrm{mis}}|\theta)\, dy_{\mathrm{mis}}\end{aligned}$$

Thus we express the first term to be

$$\mathrm{E}\left\{\ddot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})|Y_{\mathrm{obs}}\right\} + \mathrm{E}\left\{\left[\dot{l}_{\mathrm{C}}(\theta|Y_{\mathrm{obs}}, Y_{\mathrm{mis}})\right]^{\otimes 2}\middle|Y_{\mathrm{obs}}\right\}.$$

Convergence rate of EM

Let $I_{\mathrm{C}}(\theta)$ and $I_{\mathrm{O}}(\theta)$ denote the complete information and observed information, respectively.

One can show when the EM converges, the linear convergence rate, denoted as $(\theta^{(k+1)} - \hat{\theta})/(\theta^{(k)} - \hat{\theta})$ approximates $1 - I_{\mathrm{O}}(\hat{\theta})/I_{\mathrm{C}}(\hat{\theta})$. (later)

This means that

- When missingness is small, EM converges quickly
- Otherwise EM converges slowly.