

---

# EM Algorithm

---

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

- An iterative algorithm for maximizing likelihood when the model contains unobserved latent variables.
- Was initially invented by computer scientist in special circumstances.
- Generalized by Arthur Dempster, Nan Laird, and Donald Rubin in a classic 1977 *JRSSB* paper, which is widely known as the “DLR” paper.
- The algorithm iterate between **E-step** (expectation) and **M-step** (maximization).
- E-step: create a function for the expectation of the log-likelihood, evaluated using the current estimate for the parameters.
- M-step: compute parameters maximizing the expected log-likelihood found on the E step.

- Assume people's height (in cm) follow normal distributions with different means for male and female:  $N(\mu_1, \sigma_1^2)$  for male, and  $N(\mu_2, \sigma_2^2)$  for female.
- We observe the heights for 5 people (don't know the gender): 182, 163, 175, 185, 158.
- We want to estimate  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$  and  $\sigma_2$ .

This is the typical “**two-component normal mixture model**”, e.g., data are from a mixture of two normal distributions. The goal is to estimate model parameters.

We could, of course, form the likelihood function (multiplication of Normal densities) and find its maximum by Newton-Raphson.

Some notations: For person  $i$ , denote his/her height by  $x_i$ , and use  $Z_i$  to indicate gender. Define  $p_i$  be the proportion of male in the population.

Start by choosing reasonable initial values. Then:

- In the E-step, compute the probability of each person being male or female, given the current model parameters. We have (after some derivation)

$$\lambda_i^{(k)} \equiv E[Z_i | \mu_1^{(k)}, \mu_2^{(k)}, \sigma_1^{(k)}, \sigma_2^{(k)}] = \frac{\pi^{(k)} \phi(x_i; \mu_1^{(k)}, \sigma_1^{(k)})}{\pi^{(k)} \phi(x_i; \mu_1^{(k)}, \sigma_1^{(k)}) + (1 - \pi^{(k)}) \phi(x_i; \mu_2^{(k)}, \sigma_2^{(k)})}$$

- In the M-step, update parameters and group proportions by considering the probabilities from E-step as weights. They are basically weighted average and variance. For example,

$$\mu_1^{(k+1)} = \frac{\sum_i \lambda_i^{(k)} x_i}{\sum_i \lambda_i^{(k)}}, \mu_2^{(k+1)} = \frac{\sum_i (1 - \lambda_i^{(k)}) x_i}{\sum_i (1 - \lambda_i^{(k)})}, p_i = \sum_i \lambda_i^{(k)} / 5$$

We choose  $\mu_1 = 175$ ,  $\mu_2 = 165$ ,  $\sigma_1 = \sigma_2 = 10$  as initial values.

- After first iteration, we have after E-step

Person	1	2	3	4	5
$x_i$ : height (cm)	179	165	175	185	158
$\lambda_i$ : prob. male	0.79	0.48	0.71	0.87	0.31

The estimates for parameters after M-step are (weighted average and variance):  
 $\mu_1 = 176$ ,  $\mu_2 = 167$ ,  $\sigma_1 = 8.7$ ,  $\sigma_2 = 9.2$ ,  $\pi = 0.63$ .

- At iteration 15 (converged), we have:

Person	1	2	3	4	5
Height (cm)	179	165	175	185	158
Prob. male	9.999968e-01	4.009256e-03	9.990943e-01	1.000000e+00	2.443061e-06

The estimates for parameters are:  $\mu_1 = 179.6$ ,  $\mu_2 = 161.5$ ,  $\sigma_1 = 4.1$ ,  $\sigma_2 = 3.5$ ,  
 $\pi = 0.6$ .

## ABO blood groups

Genotype	Genotype Frequency	Phenotype
AA	$p_A^2$	A
AO	$2p_Ap_O$	A
BB	$p_B^2$	B
BO	$2p_Bp_O$	B
OO	$p_O^2$	O
AB	$2p_Ap_B$	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  $p_O = 1 - p_A - p_B$ . The likelihood is (from multinomial):

$$L(p_A, p_B) = (p_A^2 + 2p_Ap_O)^{n_A} \times (p_B^2 + 2p_Bp_O)^{n_B} \times (p_O^2)^{n_O} \times (2p_Ap_B)^{n_{AB}}$$

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  $n_{AA} + n_{AO} = n_A$  and  $n_{BB} + n_{BO} = n_B$ .

1. Start with initial estimates  $p^{(0)} = (p_A^{(0)}, p_B^{(0)}, p_O^{(0)})$
2. Calculate the expected  $n_{AA}$  and  $n_{BB}$ , given observed data and  $p^{(k)}$

$$n_{AA}^{(k+1)} = E(n_{AA}|n_A, p^{(k)}) = n_A \frac{p_A^{(k)} p_A^{(k)}}{p_A^{(k)} p_A^{(k)} + 2p_O^{(k)} p_A^{(k)}}, \quad n_{BB}^{(k+1)} = ?$$

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

$$p_A^{(k+1)} = (2n_{AA}^{(k+1)} + n_{AO}^{(k+1)} + n_{AB}^{(k+1)})/(2n), \quad p_B^{(k+1)} = ?$$

4. Repeat step 2 and 3 until the estimates converge



**Expectation-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

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

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  $\hat{\theta}_C = \arg \max_{\theta} f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ , if we had observed  $Y_{\text{mis}}$ .

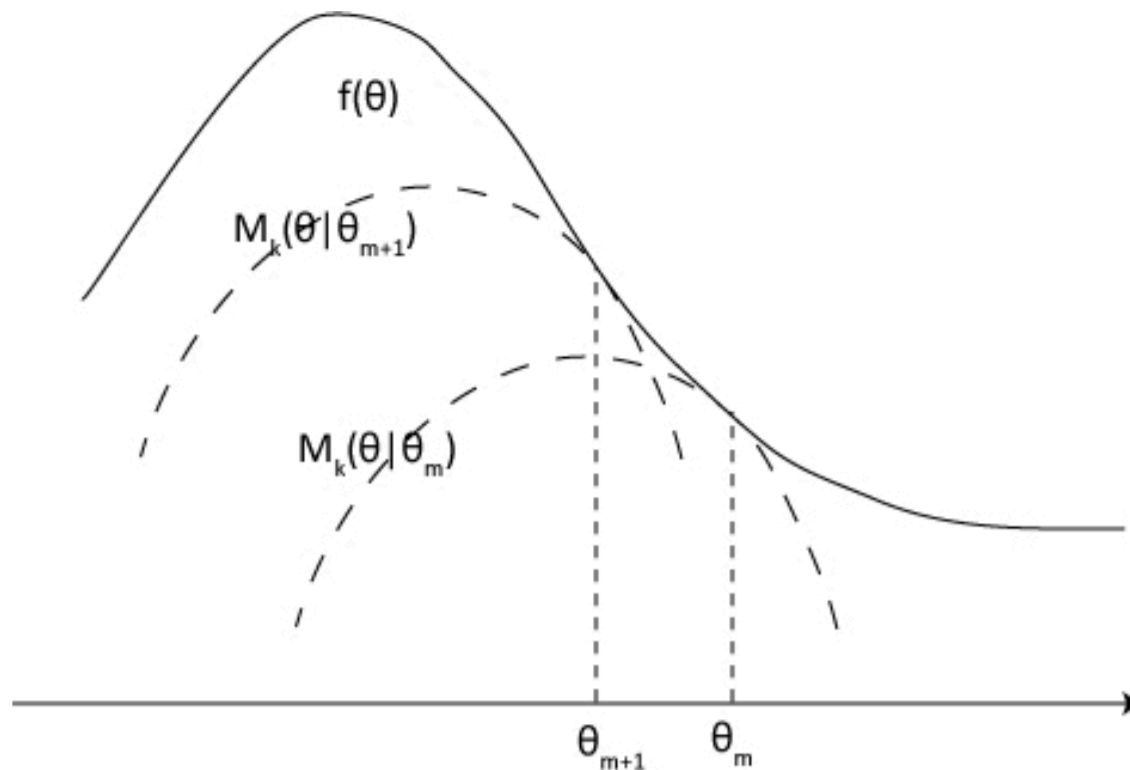
- **E step:**  $h^{(k)}(\theta) \equiv \mathbb{E} \left\{ \log f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \middle| Y_{\text{obs}}, \theta^{(k)} \right\}$
- **M step:**  $\theta^{(k+1)} = \arg \max_{\theta} h^{(k)}(\theta);$

**Nice properties** (compared to Newton-Raphson):

1. simplicity of implementation
2. stable monotone convergence

The E-step creates a surrogate function (often called the “**Q function**”), which is the expected value of the log likelihood function, with respect to the conditional distribution of  $Y_{\text{mis}}$  given  $Y_{\text{obs}}$ , under the current estimate of the parameters  $\theta^{(k)}$ .

The M-step maximizes the surrogate function.



**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

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

which can be expanded to

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

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

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

we have

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

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.,

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

by (1) and (2),

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

The Kullback-leibler information is zero if and only if

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

Combining (3) and (4), we have

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

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

## Example 1: Grouped Multinomial Data

— 13/35 —

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

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

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

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

**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	.5000000000	.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	

Consider a  $J$ -group normal mixture, where  $x_1, \dots, x_n \sim \sum_{j=1}^J p_j \phi(x_i | \mu_j, \sigma_j)$ .  $\phi(\cdot | \mu, \sigma)$  is the normal density. This is the clustering/finite mixture problem in which EM is typically used for.

Define indicator variable for observation  $i$ :  $(y_{i1}, y_{i2}, \dots, y_{iJ})$  follows a multinomial distribution (with trial number=1) and cell probabilities  $\mathbf{p} = (p_1, p_2, \dots, p_J)$ . 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,  $x_i \sim \sum_{j=1}^J p_j \phi(x_i | \mu_j, \sigma_j)$ .

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

### Observed-data log likelihood

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

### Complete-data log likelihood

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

**Complete-data log likelihood:**

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

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

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

**M step:** maximize complete-data log likelihood with  $y_{ij}$  replaced by  $\omega_{ij}$

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

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

$$\sigma_j^{(k+1)} = \sqrt{\sum_i \omega_{ij}^{(k)} (x_i - \mu_j^{(k)})^2 / \sum_i \omega_{ij}^{(k)}}$$

**Practice:** When all groups share the same variance ( $\sigma^2$ ), what's the M-step update for  $\sigma^2$ ?

## Example 2: Normal mixtures in R

— 20/35 —

```
### 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;
```

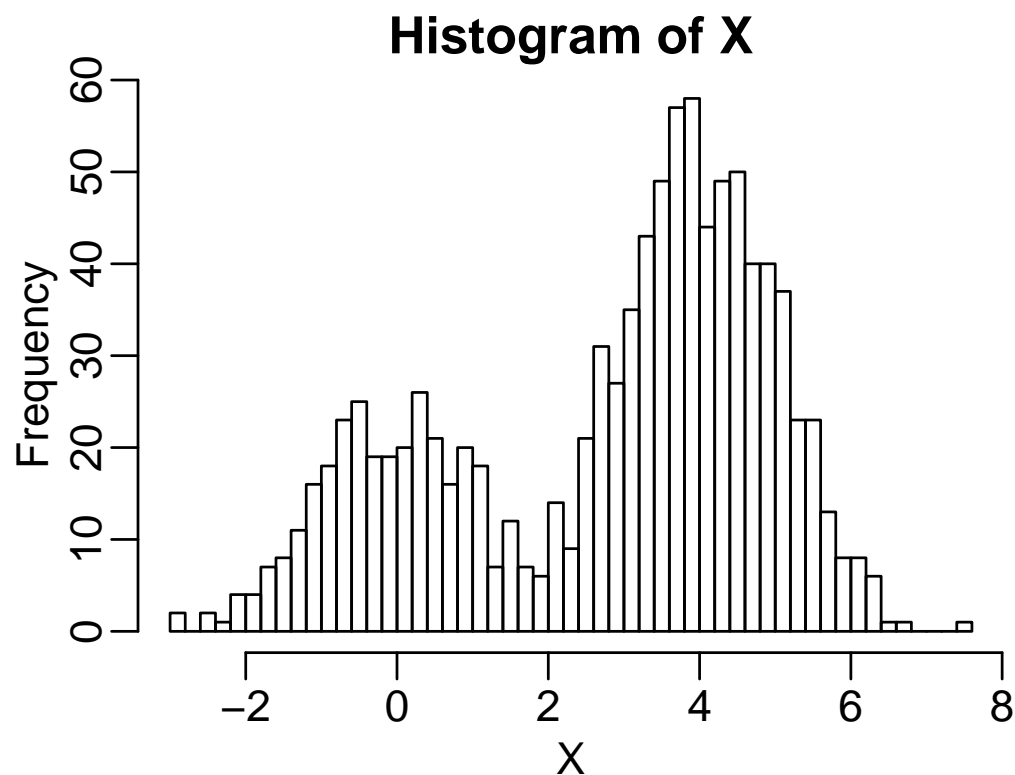
```
iter=iter+1;
```

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

```
}
```

```
}
```

```
> ## simulation
> p0=0.3;
> n=5000;
> X1=rnorm(n*p0);           # n*p0 individuals 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)
```



```
> ## 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
```



Using the same notations as in Normal mixture model. now assume the data is from a mixture of Poisson distributions.

Consider  $x_1, \dots, x_n \sim \sum_{j=1}^J p_j \phi(x_i | \lambda_j)$ , where  $\phi(.|\lambda)$  is the Poisson density. Again use  $y_{ij}$  to indicate group assignments,  $(y_{i1}, y_{i2}, \dots, y_{iJ})$  follows a multinomial distribution with cell probabilities  $\mathbf{p} = (p_1, p_2, \dots, p_J)$ .

**Now the observed-data log likelihood**

$$l(\lambda, \mathbf{p} | x) \equiv \sum_i \log \left\{ \sum_{j=1}^J p_j (x_i \log \lambda_j - \lambda_j) \right\}$$

**Complete-data log likelihood**

$$l_C(\lambda, \mathbf{p} | x, y) \equiv \sum_{ij} y_{ij} \{ \log p_j + (x_i \log \lambda_j - \lambda_j) \}$$

Derivate the EM iterations!

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

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

### Observed-data log-likelihood

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

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

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

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

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  $(\epsilon_i, b_i)$  and  $(Y_i, b_i)$  and the fact that

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

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

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

$$\begin{aligned} D &= N^{-1} \sum_{i=1}^N b_i b_i' \\ \sigma^2 &= \left( \sum_{i=1}^N n_i \right)^{-1} \sum_{i=1}^N \epsilon_i' \epsilon_i \\ \beta &= \left( \sum_{i=1}^N X_i' X_i \right)^{-1} \sum_{i=1}^N X_i' (Y_i - Z_i b_i). \end{aligned}$$

**E step:** to evaluate

$$E(b_i b_i' | Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)})$$

$$E(\epsilon_i' \epsilon | Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)})$$

$$E(b_i | Y_i, \beta^{(k)}, D^{(k)}, \sigma^{2(k)})$$

We use the relationship

$$E(b_i b_i' | Y_i) = E(b_i | Y_i) E(b_i' | Y_i) + \text{Var}(b_i | Y_i).$$

Thus we need to calculate  $E(b_i | Y_i)$  and  $\text{Var}(b_i | Y_i)$ . 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 know that

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

$$\text{Var}(b_i | Y_i) = D - D Z_i' \Sigma_i^{-1} Z_i D.$$

Similarly, We use the relationship

$$E(\epsilon_i' \epsilon_i | Y_i) = E(\epsilon_i' | Y_i)E(\epsilon_i | Y_i) + \text{Var}(\epsilon_i | Y_i).$$

We can derive

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

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

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

**M step**

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

## 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  $\theta$  does.

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

This could run into problems when the components of  $\theta$  are of very different magnitudes.

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

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  $i$ th element and 0 otherwise.

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

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

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



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

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

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

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

We can prove that

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

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

- **MLE:**  $\hat{\theta} = \arg \max_{\theta} l_O(\theta|Y_{\text{obs}})$
- **Louis variance estimator:**  $\left\{ -\ddot{l}_O(\theta|Y_{\text{obs}}) \right\}^{-1}$  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:* By the definition of  $l_O(\theta|Y_{\text{obs}})$ ,

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

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

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

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

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

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

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

to obtain

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

Thus we express the first term to be

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

Let  $I_C(\theta)$  and  $I_O(\theta)$  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  $1 - I_O(\hat{\theta})/I_C(\hat{\theta})$ . (later)

This means that

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