EM Algorithm

August 24, 2020

Last lecture

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

Expectation–maximization (EM) algorithm

- 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: obtain parameters maximizing the expected log-likelihood from the E step.

Motivating example of EM algorithm

- 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 μ_1, μ_2, σ_1 and σ_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.

A sketch of an EM algorithm

Some notations: For person i, denote the height by x_i, and use Z_i to indicate gender (unobserved). Define π 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)

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

- 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,

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

Example results

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
λ_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: $\mu_1 = 179.6$, $\mu_2 = 161.5$, $\sigma_1 = 4.1$, $\sigma_2 = 3.5$, $\pi = 0.6$.

Another 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”.
- Data are available for n individuals. Observe phenotypes but not genotypes.
- We wish to obtain the MLEs of the underlying proportions p_{A}, p_{B}, and $p_{\mathrm{O}} = 1 - p_{\mathrm{A}} - p_{\mathrm{B}}$ (these are called “allele frequencies”).
- The likelihood is (from multinomial):

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

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

Motivating example (Allele counting algorithm)

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. Step step k, calculate the expected n_{AA} and n_{BB}, given observed data and $p^{(k)}$

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

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

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

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 finite mixtures model or HMMs.

Description of EM

Notations:

- Y_{obs}: observed data, for example, the heights.
- Y_{mis}: missing/latent data, for example, the genders.
- θ: parameters of interests.
- $f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$: complete data likelihood.
- $g(Y_{\text{obs}}|\theta)$: observe data likelihood, where $g(Y_{\text{obs}}|\theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta)\, dY_{\text{mis}}$
- $c(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$: conditional likelihood of the missing data, given observed data.

Description of EM (cont.)

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

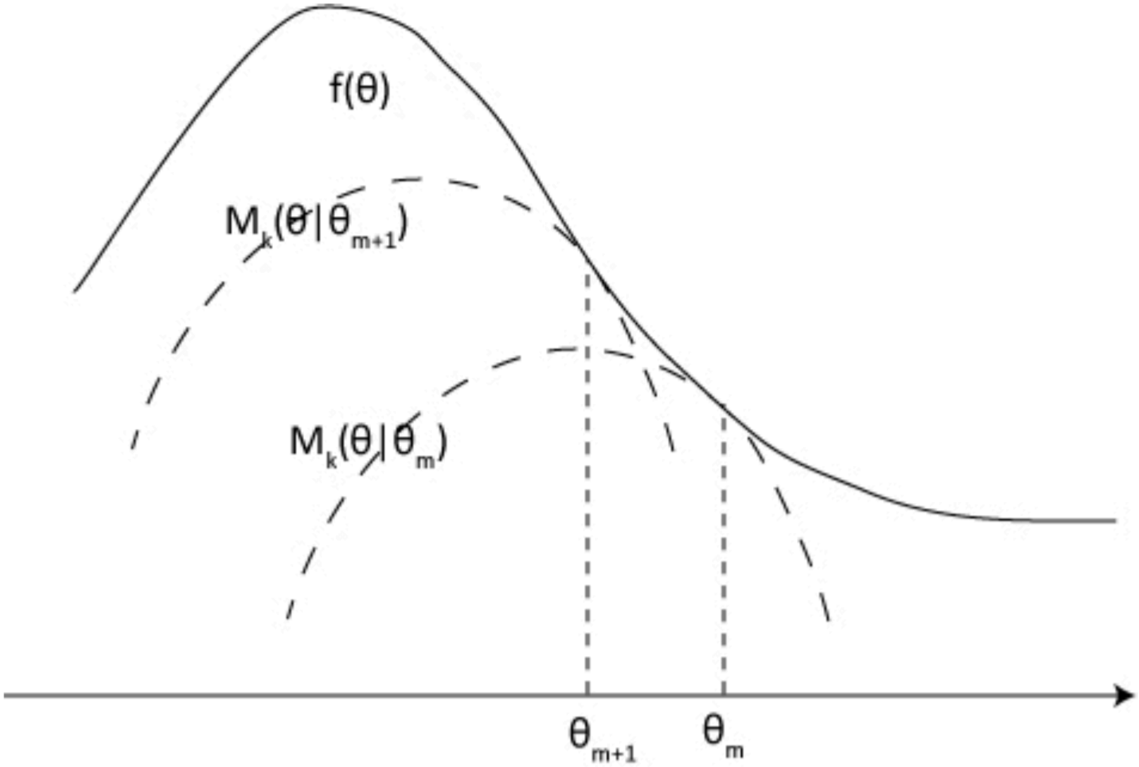
Nice properties (compared to Newton-Raphson):

1. simplicity of implementation
2. stable monotone convergence

Justification of EM

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_{mis} *given* Y_{obs}, under the current estimate of the parameters $\theta^{(k)}$.

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

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

which can be expanded to

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

By the non-negativity of the Kullback-Leibler divergence (the relative entropy), i.e.,

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

we have

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

Ascent property of EM (continued)

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.

Now we need to proof the “if and only if” part. 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) (both $\geq$ and $\leq$)

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

The Kullback-Leibler divergence 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 $\theta^{(k+1)} = \theta^{(k)}$. $\square$

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?

Example 1: Grouped Multinomial Data (continued)

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

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

Example 1: Grouped Multinomial Data (continued)

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 $x_1^{(k+1)}$

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

Example 1: Grouped Multinomial Data (continued)

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

Define indicator variable for observation i: $(y_{i1}, y_{i2}, \ldots, y_{iJ})$ follows a multinomial distribution (with trail number=1) and cell 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

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

Marginally, $x_i \sim \sum_{j=1}^{J} p_j \phi(x_i|\mu_j, \sigma_j)$. (Check this.)

In this problem, $\{x_i\}$ are the observed data; $\{x_i, y_{i1}, \ldots, y_{iJ}\}$ are the complete data.

Observed-data log likelihood (have a sum within log, trouble)

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

Complete-data log likelihood (with known group assignments, easy)

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

Practice to derive the above.

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 $j = 1, \ldots, J$,

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

This is the posterior probability for observation i being in group j. From this, we can get the Q function. (Try it.)

Note: it's easy to get Q function in this case, because l_C is linear to the data, Here we only need to evaluate $E[y_{ij}|x_i, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k)}, \boldsymbol{p}^{(k)}]$ and plug in to get $E[l_C]$. In some cases, we need to evaluate other expectations in order to get a Q function (see the mixed effect model example later).

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

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

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

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

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

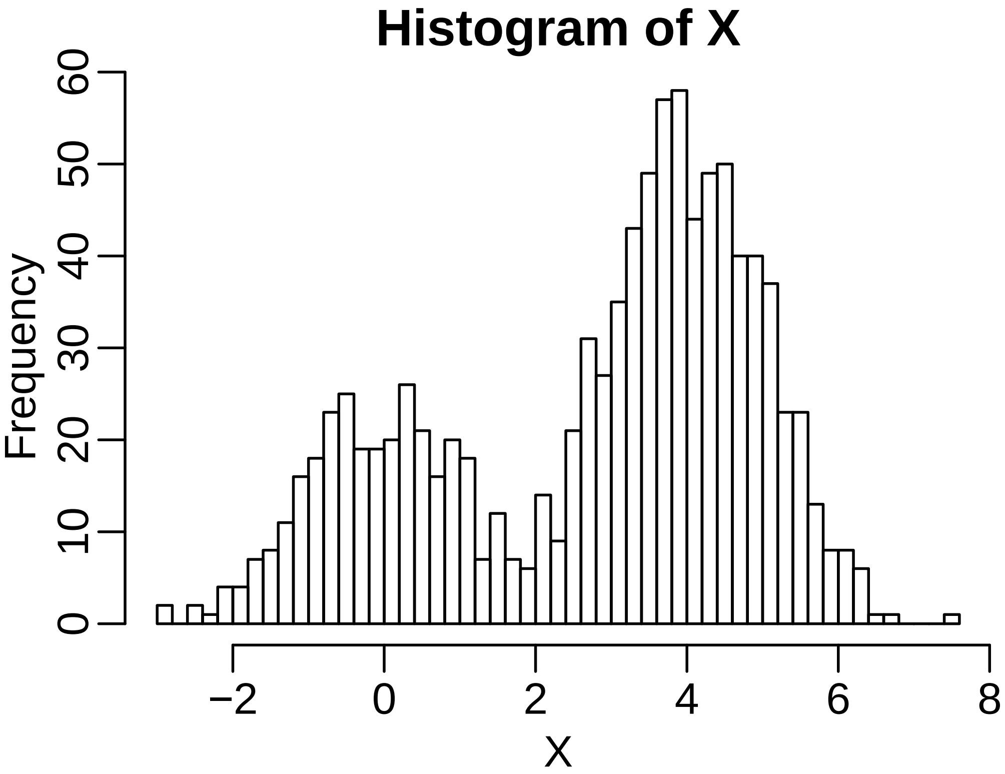
$$\sigma^{(k+1)} = \sqrt{\sum_j \left\{ \sum_i \omega_{ij}^{(k)} x_i^2 - \left(\sum_i \omega_{ij}^{(k)} x_i \right)^2 \sum_i \omega_{ij}^{(k)} \right\} \Big/ n}$$

Example 2: two-component 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
```

In class practice: Poisson mixture

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 $\boldsymbol{p} = (p_1, p_2, \dots, p_J)$.

Now the observed-data log likelihood

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

Complete-data log likelihood

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

Derivate the EM iterations!

Example 3: Mixed-effects model (time permitting)

Mixed effect model is often used in clustered data and repeated measurements, such as longitudinal data.

For a dataset of $i = 1, \ldots, N$ subjects, each with n_i observations. let Y_i be the outcome ($n_i \times 1$), X_i be the “fixed effect” design matrix ($n_i \times p$), and Z_i be the “random effect” design matrix ($n_i \times q$), . The linear mixed effect model is given by

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

- b_i is a vector of random effect coefficients, which cannot be “estimated” (because they don’t exist). It is characterized by its variance D.
- The model parameters are (β, D, σ^2)
- The **Observed-data log-likelihood** is

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

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

- This likelihood can be directly maximized for (β, D, σ^2), but difficult.
  - Since there are some constraints on the parameters (σ^2 needs to be positive, D needs to be positive definite), this needs to be maximized by restricted maximum likelihood (REML).
- This can be fit by EM, treating b_i's as missing data.
  - Given (D, σ^2) and hence Σ_i, we obtain β that maximizes the likelihood by solving

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

    This gives (weighted least square estimate):

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

Example 3: Mixed-effects model (continued)

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,

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

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

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

Example 3: Mixed-effects model (continued)

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

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

Example 3: Mixed-effects model (continued)

Similarly, We use the relationship

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

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

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

M step

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

Issues

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 could run into problems when the components of θ are of very different magnitudes.

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

Issues (continued)

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