EM: Motivating example

ABO blood groups

Genotype	Phenotype	Gen freq
AA	**A**	p_A^2
AO		$2p_Ap_O$
BB	**B**	p_B^2
BO		$2p_Bp_O$
OO	**O**	p_O^2
AB	**AB**	$2p_Ap_B$

The genotype frequencies above assume Hardy-Weinberg equilibrium. Imagine we sample n individuals (at random) and 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.

We observe n_A, n_B, n_O, n_{AB}, the numbers of individuals with each of the four phenotypes.

We could, of course, form the likelihood function and find its maximum. (There are two free parameters.) But long ago, RA Fisher (or others?) came up with the following (iterative) "allele counting" algorithm.

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.

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

Here's the algorithm:

1. Start with initial estimates $\hat{p}^{(0)} = (\hat{p}_A^{(0)}, \hat{p}_B^{(0)}, \hat{p}_O^{(0)})$.
2. Calculate the expected numbers of individuals in each of the genotype classes, given the observed numbers of individuals in each phenotype class and given the current estimates of the allele frequencies. For example:

$$\begin{aligned} n_{AA}^{(s)} &= \mathsf{E}(n_{AA} | n_{AA}, \hat{p}^{(s-1)}) \\ &= n_A \hat{p}_A^{(s-1)} / (\hat{p}_A^{(s-1)} + 2\hat{p}_O^{(s-1)}) \end{aligned}$$

3. Get new estimates of the allele frequencies, imagining that the expected n's were actually observed.

$$\begin{aligned} \hat{p}_A^{(s)} &= (2\, n_{AA}^{(s)} + n_{AO}^{(s)} + n_{AB})/n \\ \hat{p}_B^{(s)} &= (2\, n_{BB}^{(s)} + n_{BO}^{(s)} + n_{AB})/n \\ \hat{p}_O^{(s)} &= (n_{AO}^{(s)} + n_{BO}^{(s)} + 2\, n_O)/n \end{aligned}$$

4. Repeat steps (2) and (3) until the estimates converge

EM algorithm

Consider $Z \sim f(z|\theta)$ where $Z = (Z_p, Z_m)$
and $f(z_p|\theta) = \int f(z_p, z_m|\theta) dz_m$.
We observe Z_p but not Z_m.
We wish to find the MLE $\hat{\theta} = \arg\max_\theta f(Z_p|\theta)$.

In many cases, this can be quite difficult directly, but if we had observed Z_m, it would be easy to find

$$\hat{\theta}_C = \arg\max_\theta f(Z_p, Z_m|\theta)$$

E step: $l^{(s)}(\theta) = \mathrm{E}\{\log f(Z_p, Z_m|\theta) \mid Z_p, \hat{\theta}^{(s)}\}$
M step: $\hat{\theta}^{(s+1)} = \arg\max_\theta l^{(s)}(\theta)$

Nice property: the sequence $l[\hat{\theta}^{(s)}]$ is non-decreasing.

Exponential family: $l(\theta|x) = T(x)'\eta(\theta) - B(\theta)$.
$T(x)$ are the sufficient statistics. Suppose $x = (y, z)$ where y is observed and z is missing.

E step: Calculate $W^{(s)} = \mathrm{E}\{T(x)|y, \hat{\theta}^{(s-1)}\}$
M step: Determine $\hat{\theta}^{(s)}$ solving $\mathrm{E}\{T(x)|\theta\} = W^{(s)}$.

Refs: Dempster et al (1977) JRSS B 39:1–38
Wu (1983) Ann Stat 11:95–103

Example 1: Normal mixtures

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

(I put the SD rather than the variance here because the power ^{2} will really get in the way later.)

Let

$$y_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is drawn from N}(\mu_j, \sigma) \\ 0 & \text{otherwise} \end{cases}$$

so that $\sum_j y_{ij} = 1$.

(x_i) is the observed data; (x_i, y_i) is the *complete* data.

Complete data log likelihood

$$l(\mu, \sigma, p|x, y) = \sum_i \sum_j y_{ij} \left\{ \log p_j + \log f(x_i|\mu_j, \sigma) \right\}$$

Sufficient statistics

$$S_{1j} = \sum_i y_{ij} \qquad S_{2j} = \sum_i y_{ij} x_i \qquad S_{3j} = \sum_i y_{ij} x_i^2$$

Example 1 (continued)

E step

$$
\begin{aligned}
w_{ij}^{(s)} &= \mathsf{E}[y_{ij}|x_i, \hat{p}^{(s-1)}, \hat{\mu}^{(s-1)}, \hat{\sigma}^{(s-1)}] \\
&= \mathsf{Pr}[y_{ij} = 1|x_i, \hat{p}^{(s-1)}, \hat{\mu}^{(s-1)}, \hat{\sigma}^{(s-1)}] \\
&= \frac{\hat{p}_j^{(s-1)} \; f(x_i|\hat{\mu}_j^{(s-1)}, \hat{\sigma}^{(s-1)})}{\sum_j \hat{p}_j^{(s-1)} \; f(x_i|\hat{\mu}_j^{(s-1)}, \hat{\sigma}^{(s-1)})}
\end{aligned}
$$

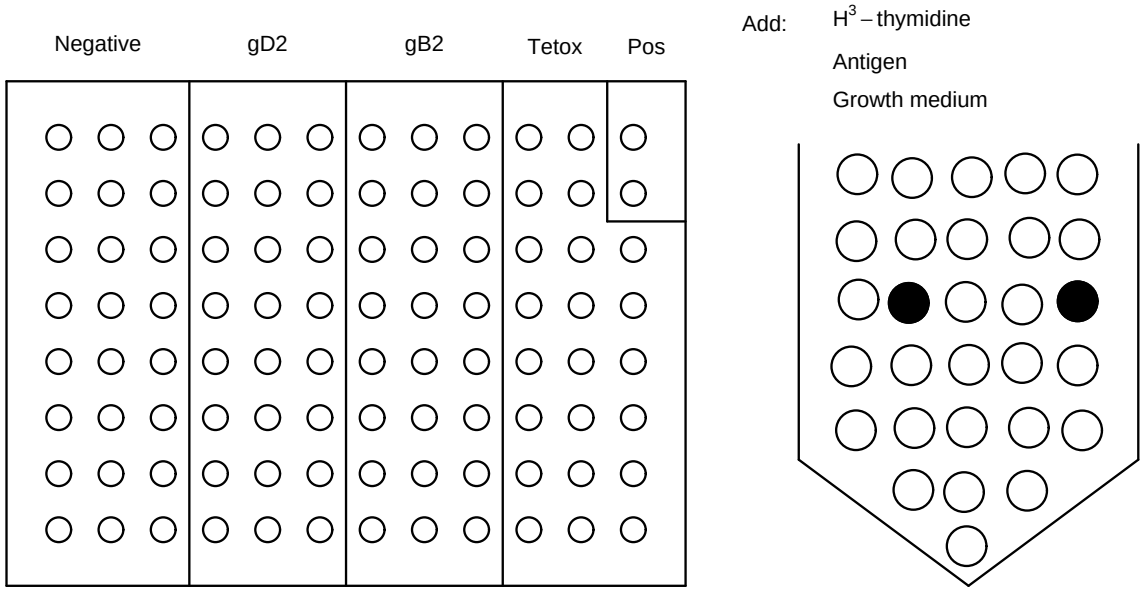
$$
S_{1j}^{(s)} = \sum_i w_{ij}^{(s)} \quad S_{2j}^{(s)} = \sum_i w_{ij}^{(s)} x_i \quad S_{3j}^{(s)} = \sum_i w_{ij}^{(s)} x_i^2
$$

M step

$$
\begin{aligned}
\hat{p}_j^{(s)} &= S_{1j}^{(s)}/n \\
\hat{\mu}_j^{(s)} &= S_{2j}^{(s)}/S_{1j}^{(s)} \\
\hat{\sigma}^{(s)} &= \sqrt{\sum_j \left\{ S_{3j}^{(s)} - [S_{2j}^{(s)}]^2 S_{1j}^{(s)} \right\} /n}
\end{aligned}
$$

Example 2: normal-Poisson mixture

Biological context



We consider an assay for estimating the frequency of T-cells responding to a set of antigens.

In each well of a microtiter plate, a sample sample of blood cells are placed with antigen, tritiated thymidine and growth medium. A small number of the T cells may respond to the antigen by replicating; in doing so, they will take up some of the H^3-thymidine into their DNA. A scintillation counter is used to measure the radioactivity of each well, which indicates the amount of H^3-thymidine taken up and hence the number of T cells in the well that responded to the antigen.

Example 2 (continued)

Model

Let k_{ij} denote the (unobserved) number of responding cells in well j of antigen group i, where $i = 1, ..., G$ and $j = 1, \ldots, n_i$.

Let x_{ij} be the log scintillation count for well j in group i.

We assume

- (x_{ij}, k_{ij}) are mutually independent
- $k_{ij} \sim$ Poisson(λ_i)
- $x_{ij}|k_{ij} \sim \mathsf{N}(a + bk_{ij}, \sigma)$

We seek the MLEs of $(\lambda_1, \ldots, \lambda_G, a, b, \sigma)$.

Sufficient statistics

$$\sum_i k_{ij} \qquad \sum_i k_{ij}^2 \qquad \sum_i x_{ij}^2 \qquad \sum_i x_{ij}k_{ij}$$

Example 2 (continued)

E step

$$w_{ij}^{(s)} = \mathsf{E}[k_{ij}|x_{ij}, \hat{\lambda}^{(s-1)}, \hat{a}^{(s-1)}, \hat{b}^{(s-1)}, \hat{\sigma}^{(s-1)}]$$

$$= \frac{\sum_k k \,\mathsf{Pr}(k|\hat{\lambda}_i^{(s-1)}) \,\mathsf{Pr}(x_{ij}|k, \hat{a}^{(s-1)}, \hat{b}^{(s-1)}, \hat{\sigma}^{(s-1)})}{\sum_k \mathsf{Pr}(k|\hat{\lambda}_i^{(s-1)}) \,\mathsf{Pr}(x_{ij}|k, \hat{a}^{(s-1)}, \hat{b}^{(s-1)}, \hat{\sigma}^{(s-1)})}$$

$$\tau_{ij}^{(s)} = \mathsf{E}[k_{ij}^2|x_{ij}, \hat{\lambda}^{(s-1)}, \hat{a}^{(s-1)}, \hat{b}^{(s-1)}, \hat{\sigma}^{(s-1)}]$$

M step

$$\hat{\lambda}_j^{(s)} = \sum_i w_{ij}^{(s)}/n_i$$

$$\hat{b}^{(s)} = \frac{\sum_{ij} y_{ij} w_{ij}^{(s)} - \left(\sum_{ij} y_{ij}\right)\left(\sum_{ij} w_{ij}^{(s)}\right)/n}{\sum_{ij} \tau_{ij}^{(s)} - \left(\sum_{ij} w_{ij}^{(s)}\right)^2/n}$$

$$\hat{a}^{(s)} = \{\sum_{ij} y_{ij} - \hat{b}^{(s)} \sum_{ij} w_{ij}^{(s)}\}/n$$

$$\hat{\sigma}^{(s)} = \sqrt{\{\sum_{ij} y_{ij}^2 + n(\hat{a}^{(s)})^2 + (\hat{b}^{(s)})^2 \sum_{ij} \tau_{ij}^{(s)} - 2\hat{a}^{(s)} \sum_{ij} y_{ij}}$$

$$-2\hat{b}^{(s)} \sum_{ij} y_{ij} w_{ij}^{(s)} + 2\hat{a}^{(s)}\hat{b}^{(s)} \sum_{ij} w_{ij}^{(s)}\}/n$$

Issues

I. Stopping rules

1. $|l(\hat{\theta}^{(s+1)}) - l(\hat{\theta}^{(s)})| < \epsilon$ for m consecutive steps.

   This is **bad**! l may not change much even when θ does.

2. $||\hat{\theta}^{(s+1)} - \hat{\theta}^{(s)}|| < \epsilon$ for m consecutive steps.

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

3. $|\hat{\theta}_j^{(s+1)} - \hat{\theta}_j^{(s)}| < \epsilon_1(|\hat{\theta}_j^{(s)}| + \epsilon_2)$ for $j = 1, \ldots, p$

   In practice, take

   $$\epsilon_1 = \sqrt{\text{machine } \epsilon} \approx 10^{-8}$$
   $$\epsilon_2 = 10\epsilon_1 \text{ to } 100\epsilon_1$$

Issues (continued)

II. Local vs global max

- There may be **many** modes
- EM may converge to a saddle point

Solution: **Many** starting points

III. Starting points

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

IV. Slow convergence

The EM algorithm can be painfully slow to converge near the maximum.

Solution: Switch to another optimization algorithm when you get near the maximum.

V. Standard errors

- Numerical approximation of the Fisher information (*ie*, the Hessian)
- Louis (1982), Meng and Rubin (1991)