557: MATHEMATICAL STATISTICS II
THE EM ALGORITHM: GENETICS OF HUMAN BLOOD GROUPS

In human genetics, the **genotype** at a genomic locus is a pair of **alleles** corresponding to small segments of DNA lying on the two chromosomal strands. The **phenotype** is the physical presentation or trait arising from the genotype. At a certain locus that determines the phenotype of blood group, the relationship between genotype and phenotype is somewhat complex; there are

- three alleles (A, B and O) yielding **six** possible genotypes (ordering is not important)
- only **four** phenotypes (A,B,AB and O).

The relationship between phenotype and genotype in this case is determined by the following table. The third column, headed X, denotes a label for the genotype class. In simple experiments, however, only the phenotype may be observed; let $Y_1, \ldots, Y_n$ denote the recorded phenotype for each of the n data.

Genotype	Phenotype	X	Y
AA	A	1	1
AB	AB	2	3
AO	A	3	1
BB	B	4	2
BO	B	5	2
OO	O	6	4

Suppose that inference about the proportions of the three alleles A,B and O, denoted $\theta_A, \theta_B, \theta_O$ is required from a sample of size n of phenotype data. We formulate a data augmentation approach, and use the EM algorithm to perform maximum likelihood estimation. One independence assumption (based on so-called *Hardy-Weinberg equilibrium*) is needed; we assume that the probability of observing a genotype is the product of the individual allele probabilities. For example

$$\begin{aligned} \text{P(AA)} &= \theta_A \times \theta_A \\ \text{P(AB)} &= \theta_A \times \theta_B \end{aligned}$$

and so on.

Define the augmented data $X_1, \ldots, X_n$, where for $i = 1 \ldots, n$,

$$\Pr[X_i = j] = \Pr[i\text{th genotype is in class j}] \qquad j = 1, \ldots, 6$$

that is

$$\Pr[X_i = j] = \begin{cases} \theta_A^2 & j = 1 \\ \theta_A\theta_B & j = 2 \\ \theta_A\theta_O & j = 3 \\ \theta_B^2 & j = 4 \\ \theta_B\theta_O & j = 5 \\ \theta_O^2 & j = 6 \end{cases} \qquad \text{for } i = 1, \ldots, n$$

with $X_1, \ldots, X_n$ a random sample. This simplification yields a complete data likelihood

$$L(\underset{\sim}{\theta}|\underset{\sim}{x}, \underset{\sim}{y}) \equiv L(\underset{\sim}{\theta}|\underset{\sim}{x}) = \prod_{i=1}^{n} \left\{ \theta_A^{2I_{\{1\}}(x_i)+I_{\{2\}}(x_i)+I_{\{3\}}(x_i)} \theta_B^{I_{\{2\}}(x_i)+2I_{\{4\}}(x_i)+I_{\{5\}}(x_i)} \theta_O^{I_{\{3\}}(x_i)+I_{\{5\}}(x_i)+2I_{\{6\}}(x_i)} \right\}$$

say, where

$$n_j = \sum_{i=1}^{n} I_{\{j\}}(x_i) \qquad j = 1, \ldots, 6.$$

The complete data likelihood is a multinomial-type likelihood in $\underset{\sim}{\theta}$.

In the standard notation, for the EM steps, we have to

- E-step: compute
$$Q(\underset{\sim}{\theta}|\underset{\sim}{\theta}^{(r)}) = \mathrm{E}_{f_{\underset{\sim}{X}|\underset{\sim}{Y},\underset{\sim}{\theta}}}[\log L(\underset{\sim}{\theta}|\underset{\sim}{X},\underset{\sim}{Y})|\underset{\sim}{y},\underset{\sim}{\theta}^{(r)}]$$
taking the expectation over $X_1, \ldots, X_n$ etc.
- M-step: maximize $Q(\underset{\sim}{\theta}|\underset{\sim}{\theta}^{(r)})$ to get $\underset{\sim}{\theta}^{(r+1)}$.

Here the M-step is straightforward due to the multinomial likelihood. The E-step is also quite straightforward, but some steps need clarification.

The log complete data likelihood takes the form

$$\begin{aligned}
\log L(\underset{\sim}{\theta}|\underset{\sim}{x},\underset{\sim}{y}) \quad &= \quad \sum_{i=1}^{n} \left(2I_{\{1\}}(x_i) + I_{\{2\}}(x_i) + I_{\{3\}}(x_i)\right) \log \theta_A \\
&\quad + \sum_{i=1}^{n} \left(I_{\{2\}}(x_i) + 2I_{\{4\}}(x_i) + I_{\{5\}}(x_i)\right) \log \theta_B \\
&\quad + \sum_{i=1}^{n} \left(I_{\{3\}}(x_i) + I_{\{5\}}(x_i) + 2I_{\{6\}}(x_i)\right) \log \theta_O
\end{aligned}$$

which is linear and additive in the indicator functions.

Conditional on $\underset{\sim}{Y}$ and $\underset{\sim}{\theta}$, some expectations can be written down automatically. For example

$$\begin{aligned}
\mathrm{E}_{f_{X_i|Y_i,\underset{\sim}{\theta}}}[I_{\{j\}}(X_i)|Y_i = 3, \underset{\sim}{\theta}] \quad &= \quad \begin{cases} 1 & j = 2 \\ 0 & j \neq 2 \end{cases} \\
\mathrm{E}_{f_{X_i|Y_i,\underset{\sim}{\theta}}}[I_{\{j\}}(X_i)|Y_i = 4, \underset{\sim}{\theta}] \quad &= \quad \begin{cases} 1 & j = 6 \\ 0 & j \neq 6 \end{cases}
\end{aligned}$$

as by definition $Y = 3 \Longrightarrow X = 2$ and $Y = 4 \Longrightarrow X = 6$. For the remaining conditional expectations, we have by Bayes theorem

$$\mathrm{E}_{f_{X_i|Y_i,\underset{\sim}{\theta}}}[I_{\{j\}}(X_i)|Y_i = 1, \underset{\sim}{\theta}] = \begin{cases} \dfrac{\theta_A^2}{\theta_A^2 + 2\theta_A\theta_O} & j = 1 \\[2ex] \dfrac{2\theta_A\theta_O}{\theta_A^2 + 2\theta_A\theta_O} & j = 3 \\[2ex] 0 & \text{otherwise} \end{cases}$$

as if $Y = 1$, then either $X = 1$ or $X = 3$, with conditional probability for each determined by noting that

$$\Pr[X = 1|Y = 1] = \frac{\Pr[X = 1, Y = 1]}{\Pr[Y = 1]} = \frac{\Pr[X = 1, Y = 1]}{\Pr[X = 1, Y = 1] + \Pr[X = 3, Y = 1]} = \frac{\text{P(AA)}}{\text{P(AA)} + \text{P(AO)}}$$

Similarly,

$$\mathrm{E}_{f_{X_i|Y_i;\underset{\sim}{\theta}}}[I_{\{j\}}(X_i)|Y_i = 2, \underset{\sim}{\theta}] = \begin{cases} \dfrac{\theta_B^2}{\theta_B^2 + 2\theta_B\theta_O} & j = 4 \\[2ex] \dfrac{2\theta_B\theta_O}{\theta_B^2 + 2\theta_B\theta_O} & j = 5 \\[2ex] 0 & \text{otherwise} \end{cases}$$

Thus $Q(\underset{\sim}{\theta}|\underset{\sim}{\theta}^{(r)})$ takes the form

$$Q(\underset{\sim}{\theta}|\underset{\sim}{\theta}^{(r)}) = \alpha_A^{(r)} \log \theta_A + \alpha_B^{(r)} \log \theta_B + \alpha_O^{(r)} \log \theta_O$$

where

$$\begin{aligned} \alpha_A^{(r)} &= \frac{2n_1\theta_A^{(r)2}}{\theta_A^{(r)2} + 2\theta_A^{(r)}\theta_O^{(r)}} + n_3 + \frac{2n_1\theta_A^{(r)}\theta_O^{(r)}}{\theta_A^{(r)2} + 2\theta_A^{(r)}\theta_O^{(r)}} \\ \alpha_B^{(r)} &= n_3 + \frac{2n_2\theta_B^{(r)2}}{\theta_B^{(r)2} + 2\theta_B^{(r)}\theta_O^{(r)}} + \frac{2n_2\theta_B^{(r)}\theta_O^{(r)}}{\theta_B^{(r)2} + 2\theta_B^{(r)}\theta_O^{(r)}} \\ \alpha_O^{(r)} &= \frac{2n_1\theta_A^{(r)}\theta_O^{(r)}}{\theta_A^{(r)2} + 2\theta_A^{(r)}\theta_O^{(r)}} + \frac{2n_2\theta_B^{(r)}\theta_O^{(r)}}{\theta_B^{(r)2} + 2\theta_B^{(r)}\theta_O^{(r)}} + 2n_4. \end{aligned}$$

and $n_1, \ldots, n_4$ are the observed counts for phenotypes A,B,AB and O. By the results for the multinomial likelihood, we can maximize $Q(\underset{\sim}{\theta}|\underset{\sim}{\theta}^{(r)})$ analytically to get

$$\theta_A^{(r+1)} = \frac{\alpha_A^{(r)}}{\alpha_A^{(r)} + \alpha_B^{(r)} + \alpha_O^{(r)}} \qquad \theta_B^{(r+1)} = \frac{\alpha_B^{(r)}}{\alpha_A^{(r)} + \alpha_B^{(r)} + \alpha_O^{(r)}} \qquad \theta_O^{(r+1)} = \frac{\alpha_O^{(r)}}{\alpha_A^{(r)} + \alpha_B^{(r)} + \alpha_O^{(r)}}$$

Example: Data from Clarke *et. al.* (1959)
We have $n_1 = 186$, $n_2 = 38$, $n_3 = 13$ and $n_4 = 284$ for the numbers of A,B,AB and O phenotypes in a sample of $n = 521$. Starting the iterative procedure at $\underset{\sim}{\theta}^{(0)} = (1/3, 1/3, 1/3)^\mathsf{T}$ yields the following first ten iterations:

r	$\theta_A^{(r)}$	$\theta_B^{(r)}$	$\theta_O^{(r)}$
1	0.25047985	0.06110045	0.68841971
2	0.21845436	0.05049394	0.73105170
3	0.21418233	0.05016173	0.73565593
4	0.21366195	0.05014667	0.73619139
5	0.21359944	0.05014547	0.73625508
6	0.21359196	0.05014535	0.73626270
7	0.21359106	0.05014533	0.73626361
8	0.21359095	0.05014533	0.73626372
9	0.21359094	0.05014533	0.73626373
10	0.21359094	0.05014533	0.73626373

indicating that convergence to the maximum value is fairly rapid.

The EM Algorithm: Censored Data

Suppose that $Y_1, \ldots, Y_n$ are the realized failure times of electronic components, and that in addition there are m additional components that are censored at times $t_{n+1}, \ldots, t_{n+m}$. Denote by $X_{n+1}, \ldots, X_{n+m}$ the unobserved failure times of these m components (so that we observe only that $X_{n+j} > t_{n+j}$ for $j = 1, \ldots, m$).

Under the assumption that the data are $Exponential(\theta)$ distributed, we may carry out inference about θ using the EM algorithm. We have the complete data likelihood as

$$L(\theta|\underset{\sim}{x}, \underset{\sim}{y}) = \prod_{i=1}^{n} \theta e^{-\theta y_i} \times \prod_{i=n+1}^{n+m} \theta e^{-\theta x_i} = \theta^{n+m} \exp\left\{-\theta\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{n+m} x_i\right]\right\}$$

so that

$$\log L(\theta|\underset{\sim}{x}, \underset{\sim}{y}) = (n+m)\log\theta - \theta\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{n+m} x_i\right].$$

Bearing in mind the constraint that $X_{n+j} > t_{n+j}$, we note that for $i = n+1, \ldots, n+m$, in the Exponential model that exhibits the lack of memory property

$$\mathrm{E}_{f_{X_i|\underset{\sim}{Y},\theta}}[X_i|\underset{\sim}{y}, \theta] = t_i + \frac{1}{\theta}$$

Thus

$$Q(\theta|\theta^{(r)}) = (n+m)\log\theta - \theta\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{n+m} t_i + \frac{m}{\theta^{(r)}}\right]$$

which is readily maximized to yield

$$\theta^{(r+1)} = \frac{n+m}{\sum\limits_{i=1}^{n} y_i + \sum\limits_{i=n+1}^{n+m} t_i + \dfrac{m}{\theta^{(r)}}}$$

For the following data

$$3.479 \quad 0.57 \quad 1.067^\star \quad 1.736^\star \quad 0.156^\star \quad 0.265 \quad 0.044^\star \quad 0.595 \quad 4.515^\star \quad 1.617$$

where the $\star$ superscript indicates censored values, we have $n = m = 5$. If $\theta^{(0)} = 1$, we have

r	$\theta^{(r)}$	r	$\theta^{(r)}$
1	0.525137	11	0.356170
2	0.424376	12	0.356114
3	0.387227	13	0.356086
4	0.370989	14	0.356072
5	0.363370	15	0.356065
6	0.359677	16	0.356061
7	0.357858	17	0.356060
8	0.356956	18	0.356059
9	0.356506	19	0.356058
10	0.356282	20	0.356058

indicating that convergence to the maximum value is slower than in earlier examples. Note that in the exponential model, the maximum likelihood estimate is available directly as

$$L(\theta|\underset{\sim}{y}, \underset{\sim}{t}) = \theta^n \exp\left\{-\theta\left[\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{n+m} t_i\right]\right\} \quad \therefore \quad \widehat{\theta}(\underset{\sim}{y}, \underset{\sim}{t}) = \frac{n}{\sum\limits_{i=1}^{n} y_i + \sum\limits_{i=n+1}^{n+m} t_i} = \frac{5}{6.525 + 7.518} = 0.356058.$$