Stat 543 2-11-05

<u>Recall</u> $k=4$ outcomes possible in each of $n=10$ trials

$$X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$$
$$X = (X_1, X_2, \ldots, X_{10})$$

$$Y_i = X_i \quad \text{for} \quad i = 1, 2, 4, 5, 7, 9, 10$$
$$Y_3 = (X_{31}, X_{33}, X_{32} + X_{34})$$
$$Y_6 = (X_{61}, X_{64}, X_{62} + X_{63})$$
$$Y_8 = (X_{83}, X_{84}, X_{81} + X_{82})$$
$$Y = (Y_1, Y_2, \ldots, Y_{10}) \qquad Y = S(X)$$

maximum likelihood based on Y is "hard" while maximum likelihood based on X is "easy"

$$\ell_X(p) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log(1 - p_1 - p_2 - p_3)$$

Given that Y = data in hand, this is

$$(2 + X_{81}) \log p_1 + (2 + X_{32} + X_{62} + X_{82}) \log p_2 + (2 + X_{63}) p_3 + (1 + X_{34}) \log(1 - p_1 - p_2 - p_3)$$

For any particular $p_0 = (p_{01}, p_{02}, p_{03}, p_{04})$

$$E_{p_0}\left[X_{81} \mid Y = \text{data in hand} \right] = \frac{p_{01}}{p_{01} + p_{02}}$$

$$E_{p_0}\left[X_{32} \mid Y = \text{data in hand}\right] = \frac{p_{02}}{p_{02}+p_{04}}$$

$$E_{p_0}\left[X_{62} \mid \qquad\qquad \right] = \frac{p_{02}}{p_{02}+p_{03}}$$

$$E_{p_0}\left[X_{82} \mid \qquad\qquad \right] = \frac{p_{02}}{p_{01}+p_{02}}$$

$$E_{p_0}\left[X_{63} \mid \qquad\qquad \right] = \frac{p_{03}}{p_{02}+p_{03}}$$

$$E_{p_0}\left[X_{34} \mid \qquad\qquad \right] = \frac{p_{04}}{p_{02}+p_{04}}$$

Based on these, perhaps a sensible estimation algorithm is the following

(0) pick some starting p_0 (say $\left(\frac{2}{7}, \frac{2}{7}, \frac{2}{7}, \frac{1}{7}\right)$ from "complete cases"/ where $Y_i = X_i$)

"E" step (1) Find $E_{p_0}\left[\, l_X(p) \mid Y = \text{data in hand} \,\right]$ for the current iterate p_0

Here this is

$$\underbrace{\left(2+\frac{p_{01}}{p_{01}+p_{02}}\right)}_{a(p_0)}\log p_1 + \underbrace{\left(2+p_{02}\left(\frac{1}{p_{02}+p_{04}}+\frac{1}{p_{02}+p_{03}}+\frac{1}{p_{01}+p_{02}}\right)\right)}_{b(p_0)}\log p_2$$
$$+\underbrace{\left(2+\frac{p_{03}}{p_{02}+p_{03}}\right)}_{c(p_0)}\log p_3 + \underbrace{\left(1+\frac{p_{04}}{p_{02}+p_{04}}\right)}_{d(p_0)}\log\left(1-p_1-p_2-p_3\right)$$

note $a(p_0)+b(p_0)+c(p_0)+d(p_0) = 10 = n$

this is the p_0 conditional expected X loglikelihood given Y = data in hand

---

(2) Maximize $E_{p_0}\left[\, l_X(p) \mid Y = \text{data in hand} \,\right]$

M-step

as a function of p — call the maximizer p^*

In our example the p_0 conditional expected log-likelihood is simple (i.e. of the form of log-likelihood for a complete data problem) and I know how to maximize it

$$p^* = \left(\frac{a(p_0)}{10}, \frac{b(p_0)}{10}, \frac{c(p_0)}{10}, \frac{d(p_0)}{10} \right)$$

---

③ replace current p_0 with p^* and return to ①

Iterate ①, ②, ③ until convergence

In our example this amounts to

$$\underbrace{p_0^{\ell+1}}_{\ell+1\text{ st iterate}} = \left(\frac{a(p_0^\ell)}{10}, \frac{b(p_0^\ell)}{10}, \frac{c(p_0^\ell)}{10}, \frac{d(p_0^\ell)}{10} \right)$$

and iterating to find a fixed point and hope that it is a solution to the original (Υ) likelihood equation

The general E-M algorithm is: for observable Y with nasty loglikelihood $l_Y(\theta)$, sometimes it is possible to dream up X s.t. $Y = S(X)$ and $\boxed{\text{optimization of } l_X(\theta)}$ is easy and computation of

$$\rightarrow E_{\theta_0}\left[l_X(\theta) \;\middle|\; Y = \text{data in hand} \right]$$

is feasible – If so I can

⓪ pick some starting θ_0

① Find

E-step

$$E_{\theta_0}\left[l_X(\theta) \mid Y = \text{data in hand} \right]$$

for current θ_0

M-step ② Optimize with θ^* the optimizer

③ Replace θ_0 with θ^* and return to ①

iterating to convergence

B+D give some arguments as to why E-M might work – see in particular Lemma 2.4.1 that says that at each step $l_y(\theta)$ never decreases –

The standard complaint w/ the E-M algorithm is that its convergence is often slow – you need a good starting point and where other methods are possible they may be better –