

Stat 471: Lecture 27

Expectation-Maximization Algorithm II.

Moo K. Chung
mchung@stat.wisc.edu

November 12, 2003

1. Multinomial distribution.

$$X = (X_1, \dots, X_n) \sim M(n, p_1, p_2, \dots, p_{k-1}, p_k)$$

is multinomial if

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k},$$

where $\sum_{i=1}^k x_i = n$ and $\sum_{i=1}^k p_i = 1$. When $k = 2$,

$$P(X_1 = x_1, X_2 = n - x_1) = \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n-x_1}$$

is Binomial $B(n, p_1)$. From symmetry all marginal is binomial, i.e. $X_i \sim B(n, p_i)$. We can trivially show

$$(X_1 + X_2, X_3, \dots, X_n) \sim M(n, p_1 + p_2, p_3, \dots, p_n)$$

and

$$(X_2, \dots, X_k) | X_1 = x_1 \sim M(n - x_1, \frac{p_2}{1 - p_1}, \dots, \frac{p_k}{1 - p_1})$$

2. From Rao C.R. (Linear statistical inference and its applications, 1973), the EM algorithm applied to the genetic problem (see Robert and Casella for the detailed discussion), where observations $x = (x_1, x_2, x_3)$ are from the multinomial $M(n, \frac{1}{2} + \frac{p}{4}, \frac{1-p}{2}, \frac{p}{4})$. Find the maximum likelihood estimator (MLE) of p . The observed likelihood is

$$L(p|x) \propto (2+p)^{x_1} (1-p)^{x_2} p^{x_3}.$$

MLE can be computed if one is willing to do algebraic manipulation. Instead we augment data by additional data such that

$$x^c = (z_1, z_2, x_2, x_3) \sim M(n, \frac{1}{2}, \frac{p}{4}, \frac{1-p}{2}, \frac{p}{4})$$

with $x_1 = z_1 + z_2$. The complete likelihood function is

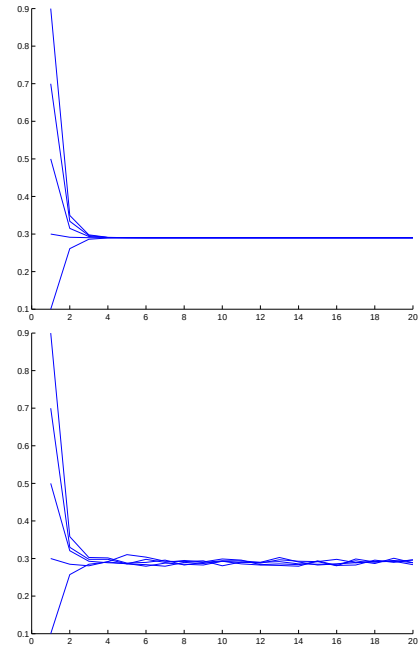
$$L(p|x^c) \propto p^{z_2+x_3} (1-p)^{x_2}.$$

This is easier to manipulate than $L(p|x)$. The conditional expected complete log-likelihood $Q(p|p_0, x)$ is

$$\mathbb{E}[\log L(p|X^c) | p_0, x] \propto \mathbb{E}_{p_0}[(Z_2 + x_3) \log p + x_2 \log(1-p)]$$

We can also show that $Z_2 | X_1 = x_1 \sim B(x_1, \frac{p}{2+p})$ (check it yourself). So the E-step is

$$Q(p|p_0, x) = \left(\frac{p_0}{2+p_0} x_1 + x_3 \right) \log p + x_2 \log(1-p).$$



The M-step is also easy. Computing $\partial Q / \partial p = 0$, we get

$$\hat{p}_1 = \frac{\frac{p_0}{2+p_0} x_1 + x_3}{\frac{p_0}{2+p_0} x_1 + x_2 + x_3}.$$

This gives the recursive formula to find $\hat{p}_i$.

```
clear p
x1=35;x2=50;x3=16;
p(1)=0.9;
for i=2:20
    pr= p(i-1)/(2+p(i-1))
    p(i) = (pr*x1 + x3)/(pr*x1+x2+x3);
end;
hold on; plot(p)
```

3. Monte-Carlo EM method. We estimate

$$\mathbb{E}[Z_2 | p_0, x] = \frac{1}{m} \sum_{i=1}^m B_i,$$

where $B_i \sim B(x_1, \frac{p_0}{2+p_0})$.

```
for i=2:20
    pr= p(i-1)/(2+p(i-1));
    ex= mean(binornd(x1,pr,10,1));
    p(i) = (ex + x3)/(ex + x2 + x3);
end;
```