

DTC Statistics module 2010

9:30-10:15 EM algorithm, simple case

10:15-11:30 Practice EM problem

11:30-12:00 EM algorithm, theory

14:00-15:00 Bayesian inference (separate pdf document)

15:00-17:00 Practice Bayes problems

Geoff Nicholls

nicholls@stats.ox.ac.uk

20/01/2010

Maximum Likelihood Estimation with missing data

Many statistical problems (in general, and in bioinformatics in particular) are difficult because we can't observe everything we would like to observe.

Example: The *ABO* Blood Group System

The *ABO* blood group system in humans reflects variation at a locus at which there are three alleles A , B , O . In this system, the alleles A and B are each dominant to O , and A and B are codominant to each other.

That is, there are 4 phenotypes, corresponding to genotypes $\{AA, AO\}$, $\{BB, BO\}$, $\{AB\}$, and $\{OO\}$.

Suppose we observe N_A , N_B , N_{AB} and N_O of the phenotypes A, B, AB and O, respectively. If the allele population frequencies of alleles A, B, O are p , q and r then, assuming Hardy-Weinberg equilibrium, the phenotypes above appear with frequencies $p^2 + 2pr$, $q^2 + 2qr$, $2pq$ and r^2 , respectively. There are effectively two unknown parameters here, since $p + q + r = 1$.

The observed data Y is (N_A, N_{AB}, N_B, N_O) and $\theta = (p, q, r)$. The likelihood is

$$L(\theta; Y) = \frac{N!}{N_A!N_B!N_{AB}!N_O!} (p^2 + 2pr)^{N_A} (q^2 + 2qr)^{N_B} (2pq)^{N_{AB}} (r^2)^{N_O}.$$

The likelihood L cannot be maximized analytically. Various numerical approaches are possible. We consider the use of the *EM algorithm*.

If we observed the full genotype frequencies N_{AA} , N_{AO} , N_{AB} , N_{BB} , N_{BO} and N_{OO} , the full-data likelihood is

$$L(\theta; Y, Z) \propto (p)^{2N_{AA}+N_{AO}+N_{AB}}(q)^{2N_{BB}+N_{BO}+N_{AB}}(r)^{2N_{OO}+N_{AO}+N_{BO}},$$

and the m.l.e.s are easily obtained,

$$\begin{aligned}\hat{p} &= \frac{2N_{AA} + N_{AO} + N_{AB}}{2N} \\ \hat{q} &= \frac{2N_{BB} + N_{BO} + N_{AB}}{2N} \\ \hat{r} &= \frac{2N_{OO} + N_{AO} + N_{BO}}{2N}.\end{aligned}$$

Here $Y = (N_A, N_B, N_{AB}, A_O)$, and the missing data is

$$Z = (N_{AA}, N_{BB})$$

(since $N_{AO} = N_A - N_{AA}$ and $N_{BO} = N_B - N_{BB}$).

There is a simple numerical algorithm which can be applied to find maximum likelihood estimates in many missing data problems, called the EM algorithm.

The EM Algorithm (simple form)

Input: Data $Y = y$ and a full-data likelihood function $L(\theta; y, z)$

Output: MLE

1. Start with initial guess $\hat{\theta}_0$ for θ and repeat these steps:
2. (The “Expectation” step) At step $k + 1$, using the current value, $\hat{\theta}_k$, of $\hat{\theta}$, calculate $\bar{Z}_{k+1} = E(Z \mid Y = y, \hat{\theta}_k)$, the expected value of the missing data, given parameter $\hat{\theta}_k$.
3. (The “Maximisation” step) Evaluate $\hat{\theta}_{k+1}$ as the m.l.e. in the complete data problem when the complete data is $(Y, Z) = (y, \bar{Z}_{k+1})$. (That is, assume the unobserved data was \bar{Z}_{k+1} , and calculate the m.l.e..)

So take $Y = (N_A, N_B, N_{AB}, N_O)$, and $Z = (N_{AA}, N_{BB})$. Then, if p , q and r are known,

$$\begin{aligned} E(N_{AA} | Y, \theta) &= \frac{p^2 N_A}{p^2 + 2pr}, \\ E(N_{BB} | Y, \theta) &= \frac{q^2 N_B}{q^2 + 2qr}, \end{aligned}$$

These equations allow us to implement the EM algorithm to find a maximum of the likelihood.

Example

In a sample of $N = 2128$ individuals from northeast Brazil, the numbers of each phenotype were $N_A = 725$, $N_B = 258$, $N_{AB} = 72$ and $N_O = 1073$.

We need to find some starting values for the EM algorithm. Any values will do. In this case we will start with:

$$\hat{p} = 1/3, \hat{q} = 1/3, \text{ and } \hat{r} = 1/3.$$

Applying the EM algorithm with these starting values, the m.l.e.s are

$$\hat{p} = 0.209, \quad \hat{q} = 0.081, \quad \hat{r} = 0.710.$$

```
%% EM example, simple case
```

```
%% Data
```

```
NA = 725;
```

```
NB = 258;
```

```
NAB = 72;
```

```
NO = 1073;
```

```
N=2128;
```

```
if (NA+NB+NAB+NO~=N), warning('wierdness'); end
```

```
%% EM
```

```
%Initialise
```

```
p=1/3; q=1/3; r=1-p-q;
```

```

%EM loop
notdone=true;
while (notdone),
    %E
    NAA=NA*p^2/(p^2+2*p*r);
    NAO=NA-NAA;
    NBB=NB*q^2/(q^2+2*q*r);
    NBO=NB-NBB;

    %M
    pp=(2*NAA+NAO+NAB)/(2*N);
    qq=(2*NBB+NBO+NAB)/(2*N);
    rr=1-pp-qq;

    notdone=any(abs([(p-pp)/p,(q-qq)/q])>eps);
    p=pp; q=qq; r=rr;
    display(sprintf(
        '%f %f %f %f %f %f %f',...
        NAA,NAO,NBB,NBO,p,q,r...
    ));
end
display(sprintf('Converged to MP\n'));

```

Iteration	\hat{p}	\hat{q}	\hat{r}
0	0.333	0.333	0.333
1	0.244	0.098	0.658
2	0.214	0.082	0.704
3	0.210	0.081	0.709
4	0.209	0.081	0.710
5	0.209	0.081	0.710

Iter	\hat{N}_{AA}	\hat{N}_{AO}	\hat{N}_{BB}	\hat{N}_{BO}
0				
1	241.667	483.333	86.000	172.000
2	113.386	611.614	17.833	240.167
3	95.574	629.426	14.147	243.853
4	93.364	631.636	13.911	244.089
5	93.096	631.904	13.891	244.109

The EM algorithm above is a special case. The formula for $\hat{\theta}_{k+1}$ is sometimes more complicated. See below.

The EM algorithm is a deterministic hill-climbing algorithm. It will always converge to a local maximum of the likelihood.

In complex problems where the likelihood has many peaks, it is prudent to start the algorithm from a number of different starting points, and to compare the value of the likelihood at each of the local maxima obtained.

Suppose θ is a parameter we don't know, $Y = 0, 1, 2, \dots$ and $Z = 0, 1, 2, \dots$ are discrete random variables with joint probability

$$\Pr(Y = y, Z = z | \theta) = f(y, z | \theta).$$

If we observe both Y and Z then the likelihood is

$$L(\theta; y, z) = f(y, z | \theta)$$

and the MLE is the θ -value that maximises $\ell(\theta; y, z) = \log(L(\theta; y, z))$,

$$\hat{\theta} = \arg \max_{\theta} \ell(\theta; y, z).$$

Let us suppose this is easy (it often is).

If we don't observe z we would marginalise the unknown variable z

$$g(y | \theta) = \sum_z f(y, z | \theta).$$

Now the likelihood is

$$L(\theta; y) = g(y|\theta)$$

and the MLE maximises $\ell(\theta; y) = \log(L(\theta; y))$,

$$\hat{\theta} = \arg \max_{\theta} \ell(\theta; y).$$

This second problem comes up whenever there are “missing data”. Because g depends on sums (or integrals), it may be hard to maximise.

If Y and Z were continuous r.v. then $f(y, z|\theta)$ would be a probability density and the likelihood would be $L(\theta; y) = g(y|\theta)$ with

$$g(y|\theta) = \int f(y, z|\theta) dz.$$

The missing-data problem is hard, but if we knew the values of the missing data, it would be an easy full-data problem.

EM Algorithm

(General form, for the more mathematically inclined)

Write Y for the observed data, Z for the missing data and $X = (Y, Z)$ for the complete data.

$$\begin{aligned}\log(g(y|\theta)) &= \log(f(y, z|\theta)) - \log\left(\frac{f(y, z|\theta)}{g(y|\theta)}\right) \\ &= E(\log(f(y, Z|\theta))|y, \theta') - E(\log(k(Z|y, \theta))|y, \theta') \\ &= Q(\theta, \theta') - C(\theta, \theta').\end{aligned}$$

Here $k(Z|y, \theta) = f(y, z|\theta)/g(y|\theta)$, and above true for any θ' .

The full EM algorithm (at step $k + 1$)

E-step calculate (or estimate) $Q(\theta, \hat{\theta}_k)$ (as function of θ).

M-step set $\hat{\theta}_{k+1} = \arg \max_{\theta} Q(\theta, \hat{\theta}_k)$.

Exercises

1. Show $g(y|\theta_{k+1}) \geq g(y|\theta_k)$, as follows.

(a) Why is $Q(\hat{\theta}_{k+1}, \hat{\theta}_k) \geq Q(\hat{\theta}_k, \hat{\theta}_k)$ (always)?

(b) Show that $C(\hat{\theta}_{k+1}, \hat{\theta}_k) \leq C(\hat{\theta}_k, \hat{\theta}_k)$ - assume

$$\sum_u \log(b(u))a(u) \leq \sum_u \log(a(u))a(u)$$

for a, b any probability mass functions.

2. Show that if $\theta_{k+1} = \theta_k$ then

$$\left. \frac{\partial Q(\theta, \theta_k)}{\partial \theta} \right|_{\theta=\theta_k} = 0,$$

and from this it follows

$$\left. \frac{\partial g(y|\theta)}{\partial \theta} \right|_{\theta=\theta_k} = 0.$$

3. Let $\bar{Z} = E(Z|y, \theta')$. Show that if f has the form

$$f(y, Z|\theta) = \exp(Z\psi - \kappa(\psi) + c(y, Z))$$

where $\psi = \psi(y, \theta)$ and $c(y, Z)$ doesn't depend on θ then

$$E(\log(f(y, Z|\theta))|y, \theta') = \log(f(y, \bar{Z}|\theta)) + \text{const wrt } \theta,$$

so the general form for the EM algorithm reduces to

E-step $\bar{Z}_{k+1} = E(Z|y, \theta_k)$

M-step $\theta_{k+1} = \arg \max_{\theta} \ell(\theta; y, \bar{Z}_{k+1})$.

Show that the blood group example has this form.

4. Express surprise this all works.