Module 3.2 - Missing Data and EM Algorithm

(Fundamentals of) Statistical Genetics

Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Outline

- Allele frequency estimation - the ABO-blood type example
- Missing data issue
- An approximated solution
- Iterative algorithms
  - The Newton-Raphson algorithm
  - The EM algorithm

The ABO-Blood Type Example I

- The ABO-gene or ABO-locus is on chromosome 9
- It has 3 alleles (antigens) (A, B, O)
- And it determines 4 blood type (A, B, AB, O):

genotype	phenotype
AA AO	A
BB BO	B
AB	AB
OO	O

A, B are dominant to O.
O is recessive to A, B.
A, B are co-dominant.

The ABO-Blood Type Example II

- E.g. in a large random sample obtained from Berlin (Bernstein 1925, Sham's book page 44):
  - $n_A = 9123$ blood type A
  - $n_B = 2987$ blood type B
  - $n_{AB} = 1269$ blood type AB
  - $n_O = 7725$ blood type O

- How to estimate the allele frequencies of alleles A, B and O (or even just to estimate the genotype frequencies of genotypes AA, AO, etc)?

The Missing Data Problem

- 6 possible genotypes, but only 4 phenotypically distinguishable.
  - Phenotype may be directly measured (e.g. blood type, height etc).
  - Genotype may not be observed.

- Problem with the direct counting method: missing data w.r.t. the count of each of the 6 genotypes: $n_{AA}, n_{AO}, n_{BB}, n_{BB}, n_{AB}, n_{OO}$!
  - $n_A = 9123 = n_{AA} + n_{AO}$: Among 9123 blood type A individuals, some have genotype AA and the others have genotype AO.
  - $n_B = 2987 = n_{BB} + n_{BO}$: Among 2987 blood type B individuals, some have genotype BB and the others have genotype BO.

- So we need a more statistical approach.

Notation and Likelihood I

- Let

$$p = freq(\text{allele } A),$$
$$q = freq(\text{allele } B),$$
$$1 - p - q = freq(\text{allele } O).$$

- Assuming HWE, frequencies of 6 genotypes:

$$freq(AA) = p^2, \quad freq(AO) = 2p(1 - p - q),$$
$$freq(BB) = q^2, \quad freq(BO) = 2q(1 - p - q),$$
$$freq(AB) = 2pq, \quad freq(OO) = (1 - p - q)^2.$$

- Therefore, frequencies of 4 phenotypes:

$$freq(A) = p^2 + 2p(1 - p - q),$$
$$freq(B) = q^2 + 2q(1 - p - q),$$
$$freq(AB) = 2pq,$$
$$freq(O) = (1 - p - q)^2.$$

Notation and Likelihood II

- The log-likelihood is

$$\begin{aligned} lnL(p,q) \sim\ & n_A ln(p^2 + 2p(1-p-q)) \\ & + n_B ln(q^2 + 2q(1-p-q)) \\ & + n_{AB} ln(2pq) \\ & + n_O ln((1-p-q)^2) \end{aligned}$$

- Take the partial derivatives of this log-likelihood function, set them to be 0 and solve the equations: MLEs.

No closed formed solutions!

Approximation I

Bernstein (1925) gave an approximate solution based on groupings of phenotypes.

- Expected frequencies of blood type A or O:

$$freq(A, O) = p^2 + 2p(1 - p - q) + (1 - p - q)^2 = (1 - q)^2.$$

- Expected frequencies of blood type B or O:

$$freq(B, O) = q^2 + 2q(1 - p - q) + (1 - p - q)^2 = (1 - p)^2.$$

- Let $n = n_A + n_B + n_{AB} + n_O$.

Approximation II

- Observed frequencies of blood type A or O:

$$(n_A + n_O)/n = 16848/21104 = 0.7983.$$

- Observed frequencies of blood type B or O:

$$(n_B + n_O)/n = 10712/21104 = 0.5076.$$

- The approximate estimates would be:

$$(1-q)^2 = 0.7983 \Rightarrow \tilde{q} = 1 - \sqrt{0.7983} = 0.106506,$$

$$(1-p)^2 = 0.5076 \Rightarrow \tilde{p} = 1 - \sqrt{0.5076} = 0.287552.$$

- Alternatively, the Newton-Raphson and EM algorithms can be used.

The Newton-Raphson Algorithm I

- A numerical iterative approach to obtain the maximum (or the minimum) a function: $f(\vec{\theta})$, $(\vec{\theta} \in R^n)$, e.g.

$$f(\vec{\theta}) = lnL(p, q) = f(p, q)$$

- It is based on the first derivatives (gradient vector), e.g.

$$f'(\vec{\theta}) = f'(p, q) = \begin{bmatrix} \frac{\partial f(p,q)}{\partial p} \\ \frac{\partial f(p,q)}{\partial q} \end{bmatrix}$$

and the second derivatives (Hessian matrix), e.g.

$$f''(\vec{\theta}) = f''(p, q) = \begin{bmatrix} \frac{\partial^2 f(p,q)}{\partial p^2} & \frac{\partial^2 f(p,q)}{\partial p \partial q} \\ \frac{\partial^2 f(p,q)}{\partial q \partial p} & \frac{\partial^2 f(p,q)}{\partial q^2} \end{bmatrix}$$

The Newton-Raphson Algorithm II

- The algorithm is such:
  - Choose a starting value, $\vec{\theta}^{(0)}$.
  - For $k = 1, 2, ...$ the updating function is

$$\vec{\theta}^{(k)} = \vec{\theta}^{(k-1)} - [f''(\vec{\theta}^{(k-1)})]^{-1} f'(\vec{\theta}^{(k-1)})$$

  - Under certain conditions, $\{\vec{\theta}^{(k)}\}$ converges to the value that maximizes (or minimizes) the function.

The Newton-Raphson Algorithm III

- A few notes on the Newton-Raphson algorithm.
  - The staring value, $\vec{\theta}^{(0)}$, is important: the algorithm is not guaranteed to converge from all starting values, particularly in regions where the matrix $-[f''(\vec{\theta}^{(k-1)})]$ is no positive definite.
    (Staring values may be obtained from some crude parameter estimates.)
  - The advantage of Newton's method is: once the iterates are close to the solution, convergence is extremely fast.
  - If the iterations do not converge: they typically move off quickly toward the edge of the parameter space.
    (The remedy can be trying again with a new staring point.)
  - The computational load can be heavy, if the number of parameters is large, because of the inverse of the Hessian matrix.

The EM Algorithm I

- The Expectation-Maximization (EM) algorithm is a numerical iterative method for finding the Maximum Likelihood Estimates (MLE) of parameters.
- EM algorithms are often used in situations where the problem of estimation can be solved much easier if certain additional pieces of data are available.

The EM Algorithm II

- The ABO-blood problem can be formulated as such incomplete data or missing data problem:
  - Some of the counts of the 6 genotypes are missing:
    among blood type A: $n_A = n_{AA} + n_{AO}$,
    among blood type B: $n_B = n_{BB} + n_{BO}$.
  - Complete data:
    n_{AA}, n_{AO}, n_{BB}, n_{BO}, n_{AB}, n_{OO}.
  - Observed data:
    $n_A = n_{AA} + n_{AO}$, $n_B = n_{BB} + n_{BO}$, $n_{AB} = n_{AB}$, $n_O = n_{OO}$.
  - Missing data:
    n_{AA} or n_{AO}, n_{BB} or n_{BO}.

- Parameters of interest: $p = freq(\text{allele } A)$ and $q = freq(\text{allele } B)$.

The EM Algorithm III

- E-step: the expected value of the log likelihood is calculated (when the log likelihood is linear w.r.t. to the missing data as in this case, then essentially, the missing data are imputed), assuming some initial values for the parameters, e.g. given the initial parameter values $p^{(0)}$, $q^{(0)}$:

$$E[n_{AA}] = \frac{freq(AA)}{freq(AA)+freq(AO)} n_A = \frac{p^{(0)}p^{(0)}}{p^{(0)}p^{(0)}+2p^{(0)}(1-p^{(0)}-q^{(0)})} n_A = n_{AA}^{(0)}$$

$$E[n_{AO}] = n_A - n_{AA} = \frac{freq(AO)}{freq(AA)+freq(AO)} n_A = \frac{2p^{(0)}(1-p^{(0)}-q^{(0)})}{p^{(0)}p^{(0)}+2p^{(0)}(1-p^{(0)}-q^{(0)})} n_A = n_{AO}^{(0)}$$

$$E[n_{BB}] = \frac{freq(BB)}{freq(BB)+freq(BO)} n_B = \frac{q^{(0)}q^{(0)}}{q^{(0)}q^{(0)}+2q^{(0)}(1-p^{(0)}-q^{(0)})} n_B = n_{BB}^{(0)}$$

$$E[n_{BO}] = n_B - n_{BB} = \frac{freq(BO)}{freq(BB)+freq(BO)} n_B = \frac{2q^{(0)}(1-p^{(0)}-q^{(0)})}{q^{(0)}q^{(0)}+2q^{(0)}(1-p^{(0)}-q^{(0)})} n_B = n_{BO}^{(0)}$$

The EM Algorithm IV

- M-step: MLE can then be calculated based on
  imputed missing data + observed data = complete data

e.g. MLE of the parameters of interest, p and q, given the imputed missing data $(n_{AA}^{(0)}, n_{AO}^{(0)}, n_{BB}^{(0)}, n_{BO}^{(0)})$, and the observed data (n_{AB}, n_{OO}):

$$p^{(1)} = \frac{2n_{AA}^{(0)} + n_{AO}^{(0)} + n_{AB}}{2n}$$

$$q^{(1)} = \frac{2n_{BB}^{(0)} + n_{BO}^{(0)} + n_{AB}}{2n}$$

where $n = n_A + n_B + n_{AB} + n_O$, the total number of individuals in the sample. $p^{(1)}$ and $q^{(1)}$ are improved estimates of the parameters!

The EM Algorithm V

- Use $p^{(1)}$ and $q^{(1)}$ to perform the E-step again, and then perform the M-step to obtain improved estimates, $p^{(2)}$ and $q^{(2)}$.
- Continue until convergence: the changes in parameter estimates $(p^{(k)} - p^{(k-1)}, q^{(k)} - q^{(k-1)})$ are negligible for the purpose of the study.
- A couple of comments on the EM algorithm:
  - Under regular conditions, the algorithm converges to a local mode of the posterior density.
  - The rate at which the EM algorithm converges depends on the proportion of missing "information".

Exercises

- Implement the two algorithms.