ABSTAT17

IGC, April 10-13, 2017

Project: EM algorithm[1]

Suppose you are given a sample of m male twin pairs, f female twin pairs, and d opposite sex twin pairs. Estimate the probability p that a twin pair is identical and the probability q that a child is male.

Here $y = (m, f, d)$ is the vector of observed data and $\theta = (p, q)$ is the parameter vector. If we knew exactly which pairs of same-sex twins were identical, then it would be easy to estimate p and q. Thus, we postulate complete data $x = (m1, m2, f1, f2, d)$, with $m1$ representing the number of male identical twin pairs and $m2$ the number of male non-identical twin pairs. Note that $m1$ and $m2$ are observations of the random variables $M1$ and $M2$, respectively. $f1$ and $f2$ are defined similarly.

Note that $m_1 + m_2 = m$, $f_1 + f_2 = f$ e $m + f + d = N$. Algorithm EM should be used in order to estimate p and q.

1. The complete likelihood function is:

$$L(\theta|x) = \frac{N!}{m_1!m_2!f_1!f_2!d!}(pq)^{m_1}[(1-p)q^2]^{m_2}[p(1-q)]^{f_1}[(1-p)(1-q)^2]^{f_2}$$
$$[(1-p)2q(1-q)]^d,$$

   since the identical twins have the same genetic code. Identify the distribution of the population.

2. Present an expression for the logarithm of $L(\theta|x)$.

[1] Exercise based on http://www.leg.ufpr.br/ paulojus/EM/EM-Exemplos-Lange.pdf

3. **Step E**

   Considering that the expected values for M_1 e M_2, at iteration k, are:

   $$m_1^k = \mathrm{E}(M_1|y,\theta^k) = m\frac{p^k q^k}{p^k q^k + (1-p^k)(q^k)^2},\ m_2^k = \mathrm{E}(M_2|y,\theta^k) = m - m_1^k\,,$$

   determine f_1^k, f_2^k.

4. **Step M**

   Find the estimators of p and q at iteration $k+1$.

5. Develop a script for the EM algorithm and find estimates for p and q. The data available are: $m = 39, f = 31, d = 30$.

6. Prepare some slides including:

   - Description of the problem;
   - Description of the method (workflow);
   - Eventually some parts of the script;
   - The answer to the problem;
   - Compare the estimated frequencies to the true frequencies, which are: $p = 0.4$ and $q = 0.5$.