

Statistical Techniques Practice: EM Algorithm.

1. Some people can curl their tongue and others cannot. The ability to curl one's tongue (or not) is an inherited trait and it is known to be determined by variation in a single gene. The gene has two alleles, T and t , where T is dominant. Individuals with the genotype TT or Tt can curl their tongues whereas those with genotype tt cannot.

Write p for the population frequency of the allele T , and assume Hardy-Weinberg equilibrium, so that the frequencies of the three possible genotypes TT , Tt , and tt , are p^2 , $2p(1-p)$ and $(1-p)^2$, respectively.

Suppose N individuals are sampled from the population, and write N_T for the number of sampled individuals who can curl their tongue, and N_t for the number who can't. Based on this data we want to find the maximum likelihood estimate of p .

This problem is well suited to the EM algorithm. We can consider it as a missing data problem, with the complete data being the counts N_{TT} , N_{Tt} , and N_{tt} of each of the three genotypes defined above (note that we already know N_{tt}).

So, assuming we actually observed N_{TT} , N_{Tt} , and N_{tt} , the maximum likelihood estimate for p is just:

$$\hat{p} = \frac{2N_{TT} + N_{Tt}}{2N} \quad (1)$$

There is a natural interpretation of \hat{p} . What is it? (If you have time at the end of the question, and want the practice, derive the mle for the complete data problem.)

Now, if p is known, then the expected numbers of each genotype are:

$$E(N_{TT}|N_T, N_t, p) = \frac{p^2 N_T}{p^2 + 2p(1-p)}, \quad (2)$$

$$E(N_{Tt}|N_T, N_t, p) = \frac{2p(1-p)N_T}{p^2 + 2p(1-p)}. \quad (3)$$

Explain why this is so.

- (a) Use equations ?? – ?? to implement the EM algorithm using Matlab. The algorithm will converge to \hat{p} , the maximum likelihood estimate for the frequency of allele T in the population.
- (b) Use your code to find the m.l.e. for the frequency of the tongue-curling gene using the data collected in lectures.
- (c) The EM algorithm is usually used when it is impossible to maximize a likelihood function analytically. In the case of tongue curling, however, the result can be found directly. If you have time, write down the likelihood function for p . By taking the log of the likelihood function and differentiating, derive the formula for the m.l.e. of p .

For the tongue curling data, evaluate the mle with this formula, and compare with the result from running the EM algorithm.

- (d) If you are really keen, prove analytically that for this particular problem the EM algorithm will converge to the expression for the m.l.e.