PB HLTH C240C / STAT C245C

Assignment #3

Due: Thursday, October 30th, 11:59pm

Maximum likelihood estimation of the ABO blood group allele frequencies using the EM algorithm

The *ABO blood groups* were the first to be discovered and are important in assuring safe blood transfusions (Cf. Landsteiner, 1930 Nobel Prize in Physiology and Medicine, `nobelprize.org/nobel_prizes/medicine/laureates/1930/landsteiner-bio.html`).

As indicated in Table 1, the ABO blood groups are characterized by the presence or absence of *antigens* on the surface of red blood cells and *antibodies* in serum. The ABO locus has three *alleles*, **A**, **B**, and **O**, leading to $3^2 = 9$ *phased genotypes*, $3 + 3 \times 2/2 = 6$ *unphased genotypes*, and only four *phenotypes*, the blood groups A, B, AB, and O.

Let $\pi = (\pi_A, \pi_B, \pi_O)$ denote the ABO allele frequencies in a well-defined population of interest. Under the assumption of *Hardy-Weinberg equilibrium* (HWE), the maternal and paternal alleles are independent, i.e., genotype frequencies are products of allele frequencies. Let $Y = (Y_A, Y_B, Y_{AB}, Y_O)$ denote the ABO phenotype counts for a random sample of n individuals from the population of interest.

The objective of this assignment is to apply the EM algorithm to derive maximum likelihood estimates of the ABO allele frequencies for a dataset from the classical article of Clarke *et al.* (1959). Specifically, consider the following ABO phenotype counts for a sample of $n = 521$ duodenal ulcer patients (Clarke *et al.*, 1959, Table III): $Y_A = 186$, $Y_B = 38$, $Y_{AB} = 13$, and $Y_O = 284$. For simplicity, you may assume that the $n = 521$ patients are a random sample from a well-defined population, with Hardy-Weinberg equilibrium at the ABO locus.

Table 1: *ABO blood groups.* Phenotypes, genotypes, and genotype frequencies under Hardy-Weinberg equilibrium.

Phenotype ABO blood group	**Antigens**	**Antibodies**	**Unphased genotype**	**Unphased genotype frequency (HWE)**
A	A	Anti-B	AA, AO	$\pi_A^2 + 2\pi_A\pi_O$
B	B	Anti-A	BB, BO	$\pi_B^2 + 2\pi_B\pi_O$
AB	A and B	Neither	AB	$2\pi_A\pi_B$
O	Neither	Anti-A and anti-B	OO	π_O^2

Question 1. Conditional distribution of multinomial counts.
Suppose $X = (X_k : k = 1, \ldots, K)$ is a random variable with Multinomial$(n, \pi = (\pi_k : k = 1, \ldots, K))$ distribution, that is,

$$\Pr(X = x) = \frac{n!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} \pi_k^{x_k}, \tag{1}$$

where $x = (x_k : k = 1, \ldots, K) \in \mathbb{N}^K$, with $\sum_k x_k = n$, and $\pi = (\pi_k : k = 1, \ldots, K) \in [0, 1]^K$, with $\sum_k \pi_k = 1$.

Derive the conditional distribution of X_k given $X_k + X_{k'}$, where k, $k' \in \{1, \ldots, K\}$, $k \neq k'$.

In particular, show that the conditional expected value of X_k given $X_k + X_{k'}$ is

$$\mathrm{E}[X_k | X_k + X_{k'}] = (X_k + X_{k'}) \frac{\pi_k}{\pi_k + \pi_{k'}}. \tag{2}$$

Question 2. Derivation of EM algorithm.
Derive the *expectation-maximization algorithm* (EM) for *maximum likelihood estimation* (MLE) of the ABO allele frequencies $\pi = (\pi_\mathrm{A}, \pi_\mathrm{B}, \pi_\mathrm{O})$, based on ABO phenotype counts from a random sample of n individuals from a well-defined population, with Hardy-Weinberg equilibrium at the ABO locus.

Specifically, define the observed incomplete and unobserved complete data structures, provide the incomplete and complete data log-likelihood functions, supply the main EM Q-function, and derive explicit solutions for the E- and M-steps.

Question 3. Software implementation of EM algorithm.
Write an R function implementing the EM algorithm for maximum likelihood estimation of the ABO allele frequencies $\pi = (\pi_\mathrm{A}, \pi_\mathrm{B}, \pi_\mathrm{O})$.

Arguments to this function should include: the phenotype counts, starting values for the allele frequencies, stopping criteria; it should return candidate MLE for the allele frequencies and the corresponding value of the observed data log-likelihood.

Question 4. Application of EM algorithm.
Apply the EM algorithm to derive maximum likelihood estimates of the ABO allele frequencies $\pi = (\pi_\mathrm{A}, \pi_\mathrm{B}, \pi_\mathrm{O})$ for the Clarke *et al.* (1959) dataset.

Trace the progress of the EM algorithm by providing a table of candidate MLE for the allele frequencies and corresponding values of the observed data log-likelihood at each iteration.

Also provide graphical summaries of these results.

Comment on the EM algorithm's performance in terms of sensitivity to starting values, convergence, and any other features you deem relevant.

Compare the results from your implementation of the EM algorithm to those from one of the R optimization functions (e.g., `optim`).

References

Clarke, C. A., Evans, D. A., McConnell, R. B., and Sheppard, P. M. (1959). Secretion of blood group antigens and peptic ulcer. *British Medical Journal*, **1**(5122), 603–607.