Estimating allele probabilities for the ABO blood type system using Gibbs sampling

An individual's blood type in the ABO blood type system is one of A, B, O, or AB. The blood type is determined by the genotype of the individual at a certain locus called the ABO locus, which is a certain position on a particular chromosome. This locus has three alleles: A, B and O. Every individual has two copies of each chromosome in each of the normal cells: one chromosome which derives from the DNA of the father of the individual and one which derives from the DNA of the mother. Hence there are six possible genotypes (unordered pairs of alleles) at the ABO locus, namely AA, AO, BB, BO, OO, and AB. The ABO blood type is the observable characteristic or phenotype of the individual. The phenotype is determined by the genotype. The alleles A and B are dominant and the allele O is recessive. This implies that individuals with AA or AO genotype have type-A blood; individuals with BB or BO genotype have type-B blood; individuals with OO genotype have type-O blood and individuals with AB genotype have type-AB blood.

Consider a population which is in genetic equilibrium (the Hardy-Weinberg equilibrium). Let p be the population relative frequency of allele A, q be the population relative frequency of allele B and r be the population relative frequency of allele O. Then the probabilities of the genotypes (for a randomly selected individual) are given in table 1. Notice that probability of a heterozygous individual (an individual whose alleles are different) is twice the product of the allele probabilities. E.g., the genotype AO can result from inheriting allele A from the mother and allele O from the father or inheriting allele O from the mother and allele A from the father. A homozygous individual has received identical alleles from the mother and the father, and therefore the probability is the corresponding allele probability squared.

In a certain population the observed blood type frequencies for type A, B, O and AB were, respectively

$$(y_1, y_2, y_3, y_4) = (182, 60, 176, 17).$$

We want to estimate the allele probabilities p, q and r based on this data. Notice that we have the constraints $0 < p, q, r < 1$ and $p+q+r = 1$. We will take p and q as the parameters, and then r is short hand for the expressions $r = 1-p-q$. If p and q were known, then the random variables corresponding to the data would have the multinomial distribution

$$Y = (Y_1, Y_2, Y_3, Y_4) \sim \text{Mult}(n, (p^2 + 2pr, q^2 + 2qr, r^2, 2pq)),$$

Table 1: Genotypes, phenotypes and probabilities for the ABO locus.

Genotype	phenotype	probability
AA	A	p^2
AO	A	$2pr$
BB	B	q^2
BO	B	$2qr$
OO	O	r^2
AB	AB	$2pq$

where $n = 435$.

As the prior, we will use the Dirichlet distribution

$$f(p,q) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}\, p^{a-1}q^{b-1}(1-p-q)^{c-1}, \qquad \text{with}$$
$$0 < p < 1, \quad 0 < q < 1, \quad p+q < 1,$$

where $a, b, c > 0$ are known constants.

Your task is to implement Gibbs sampling for the posterior distribution of (p, q).

Guidance:

- Define two latent variables Z_1 and Z_2 which are the frequencies of the AA and BB genotypes. Conditional on the parameters, the complete data likelihood is then

  $$(Z_1, Y_1 - Z_1, Z_2, Y_2 - Z_2, Y_3, Y_4) \sim \text{Mult}(n, (p^2, 2pr, q^2, 2qr, r^2, 2pq)).$$

  Derive the needed posterior full conditional distributions. (The correct results are certain binomial distributions and certain scaled beta distributions.)

- Implement Gibbs sampling and draw a scatter plot of the posterior distribution, when $a = b = c = 1$. Also, estimate posterior means for the parameters p and q.