Importance Sampling Strategies for assigning Hybrid Alleles to Parental Populations

Robert Cope and Martin O'Hely
University of Queensland

Introduction

Consider two parental populations: P_1 and P_2. Within each there are population allele frequencies given by:

$$P_k = (f_{k1}, f_{k2}, ..., f_{kd}) \quad (k = 1, 2)$$

Alleles from each population leave that population to form a third population,P_H.

$$P_H = (f_{H1}, f_{H2}, ..., f_{Hd})$$

Each allele in P_H came either from P_1 with probability p_1 or from P_2 with probability p_2 (i.e. $f_{Hi} = p_1 f_{1i} + p_2 f_{2i}$).

From samples of N_k alleles in population P_k $(k = 1, 2, H)$, we assume that we can observe allele frequencies of $n_k = (n_{k1}, n_{k2}, ..., n_{kd})$ [1]. Now, given p_k $(k = 1, 2)$ and N_k $(k = 1, 2, H)$ we want to determine the probability of observing a particular configuration $(n_{k1}, n_{k2}, ..., n_{kd})$ $(k = 1, 2, H)$.

Analytical Method

The probability of coming to the observed configuration can be evaluated analytically by considering all possible ways that alleles in P_H could have come from P_1 and P_2. Let $n_{(H,k)i}$ be the number of alleles of type i in the Hybrid that have come from P_k, and m_{ki} be be the number of alleles of type i in P_k once the Hybrid alleles have been distributed to the Parental Populations (i.e. m_{ki} simulate initial parental configurations before hybridisation occurred):

$$m_{ki} = n_{ki} + n_{(H,k)i}.$$

The probability that the parental configuration $(m_{11}, \ldots, m_{1d}), (m_{21}, \ldots, m_{2d})$ leads to the observed configuration $(n_{11}, \ldots, n_{1d}), (n_{21}, \ldots, n_{2d}), (n_{H1}, \ldots, n_{Hd})$ is [2]

$$C(m_{11}, n_{(H,1)1}) \ldots C(m_{1d}, n_{(H,1)d}) \times$$
$$C(m_{21}, n_{(H,2)1}) \ldots C(m_{2d}, n_{(H,2)d}) \times$$
$$(p_1)^{n_{(H,1)1}+\cdots+n_{(H,1)d}}(p_2)^{n_{(H,2)1}+\cdots+n_{(H,2)d}}$$

We can then calculate the probability of the parental configurations under the multinomial-Dirichlet, before summing over all possible $n_{(H,k)i}$ to reach the desired probability.

Computational Methods

The problem with the analytical method is that the sum has many $((n_{H1}+1)\cdots(n_{Hd}+1))$ terms. This means that the algorithm for the analytical method is order $(N_H/d)^d$. It is preferable to generate a few parental configurations at random and average over the probabilities determined from these in a consistent way. This is done by assigning each allele i in the Hybrid to a parental population k.

The intuitive way of doing this involves iterating through the alleles in P_H and assigning them to P_k based upon p_k [3]. This can give terrible estimates, so other methods should be considered.

Three Candidate Methods

Three methods have been proposed to more efficiently assign the alleles in the Hybrid population to the parental populations.

Method 1

Send any allele i to population j with probability proportional to

$$\frac{p_k(n_{ki}+1)}{\left(\sum_{i=1}^{d} n_{ki}\right)+1},$$

provided there is an allele of type i in the Hybrid population. This method effectively picks both the population and the allele to be assigned both at the same time. This process continues (with new probabilities calculated each time) until there are no more alleles remaining in n_H.

Method 2

Pick a population with probability $\propto p_k$, then an allele $\propto n_{ki}+1$ for this fixed k.
This method selects first which population the allele will be assigned to, then which allele will be assigned. The allele is only assigned if there is an allele present in that position in n_H. This process continues (with new probabilities calculated each time) until there are no more alleles remaining in n_H.

Method 3

Pick an allele i from among those present in the Hybrid with probability $\propto n_{ki}$, then a population proportional to

$$\frac{p_k(n_{ki}+1)}{\left(\sum_{i=1}^{d} n_{ki}\right)+1}$$

(for this fixed i). This method selects first the allele which is to be allocated, and then the population which it is to be allocated to. This process continues (with new probabilities calculated each time) until there are no more alleles remaining in n_H.

Simulation

These methods (Method 0 along with Methods 1, 2 and 3) were simulated using C++ over a number of different inputs (n_k $(k = 1, 2, H)$). From each method for each iteration a log-probability was returned based upon the multinomial-Dirichlet distribution and the appropriate importance-sampling weights. The system was simulated with

- the set of data given in the example used in slides by Jean-Marie Cornuet.
- five sets of data generated from the negative-binomial distribution.
- systematically chosen sets of data where n_1 and n_2 have fixed numbers of empty lineages (i.e. $n_{ki} = 0$ $(k = 1, 2)$).
  An example of this type of data:

  $$\begin{aligned} n_1 &= (0, 0, 0, 12, 12) \\ n_2 &= (0, 0, 12, 12, 12) \\ n_H &= (10, 10, 10, 10, 10) \end{aligned}$$

  This situation has two common 0s between n_1 and n_2.

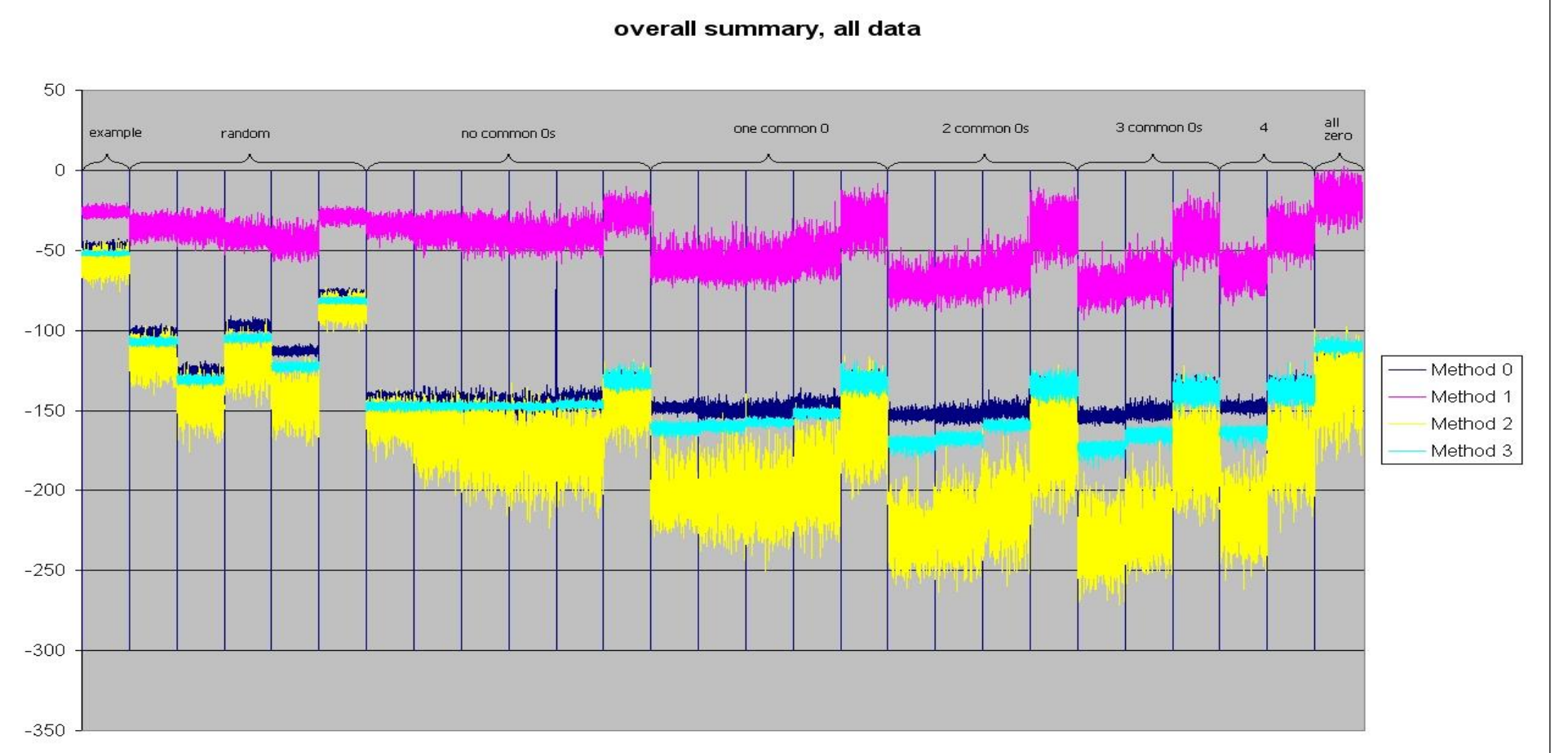
The results of these simulations give the possibility of observing:

- the variability of the generated log-probabilities for given input data, and
- the difference between the mean log-probabilities for each method and the analytical solution for given input data.

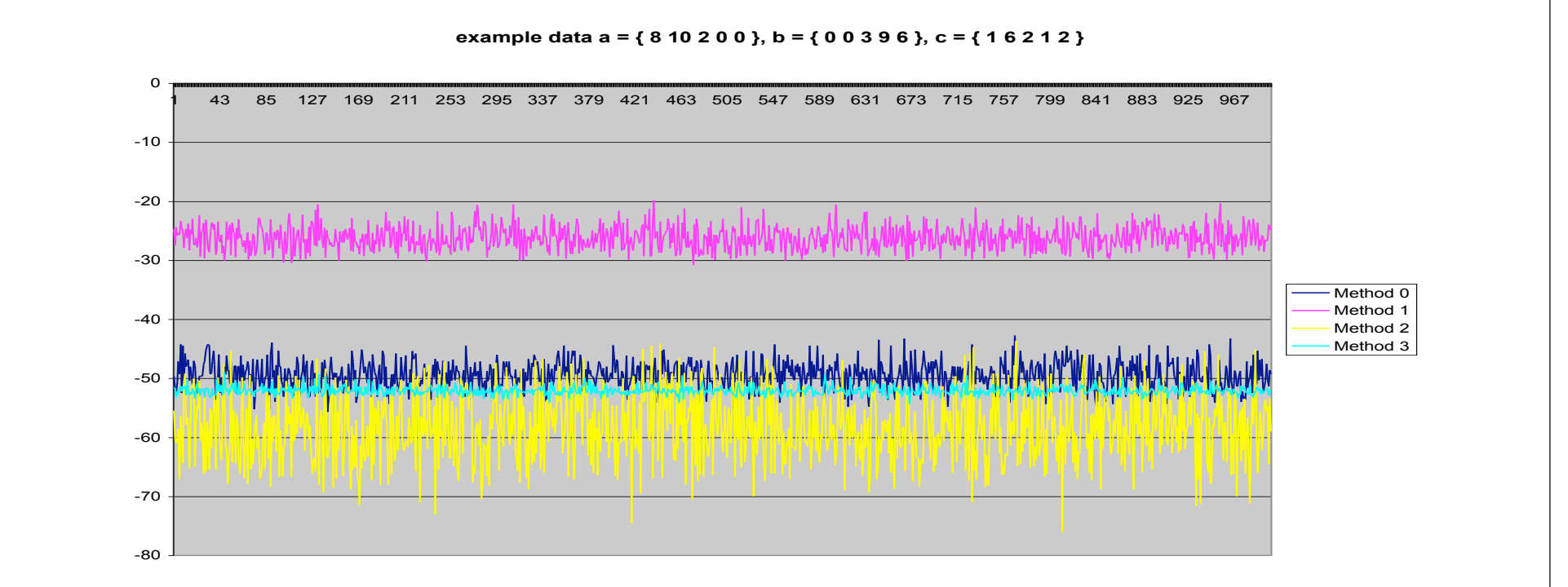
These describe, respectively, the efficiency and accuracy of the four methods.

Results

When considering the variation of the generated probabilities, the most useful results are in the form of figures showing the log-probabilities for all 4 methods over a number of iterations (Variance can be calculated as part of the simulation, but is easier to see/analyse via this sort of figure; narrower bands indicate higher efficiency).

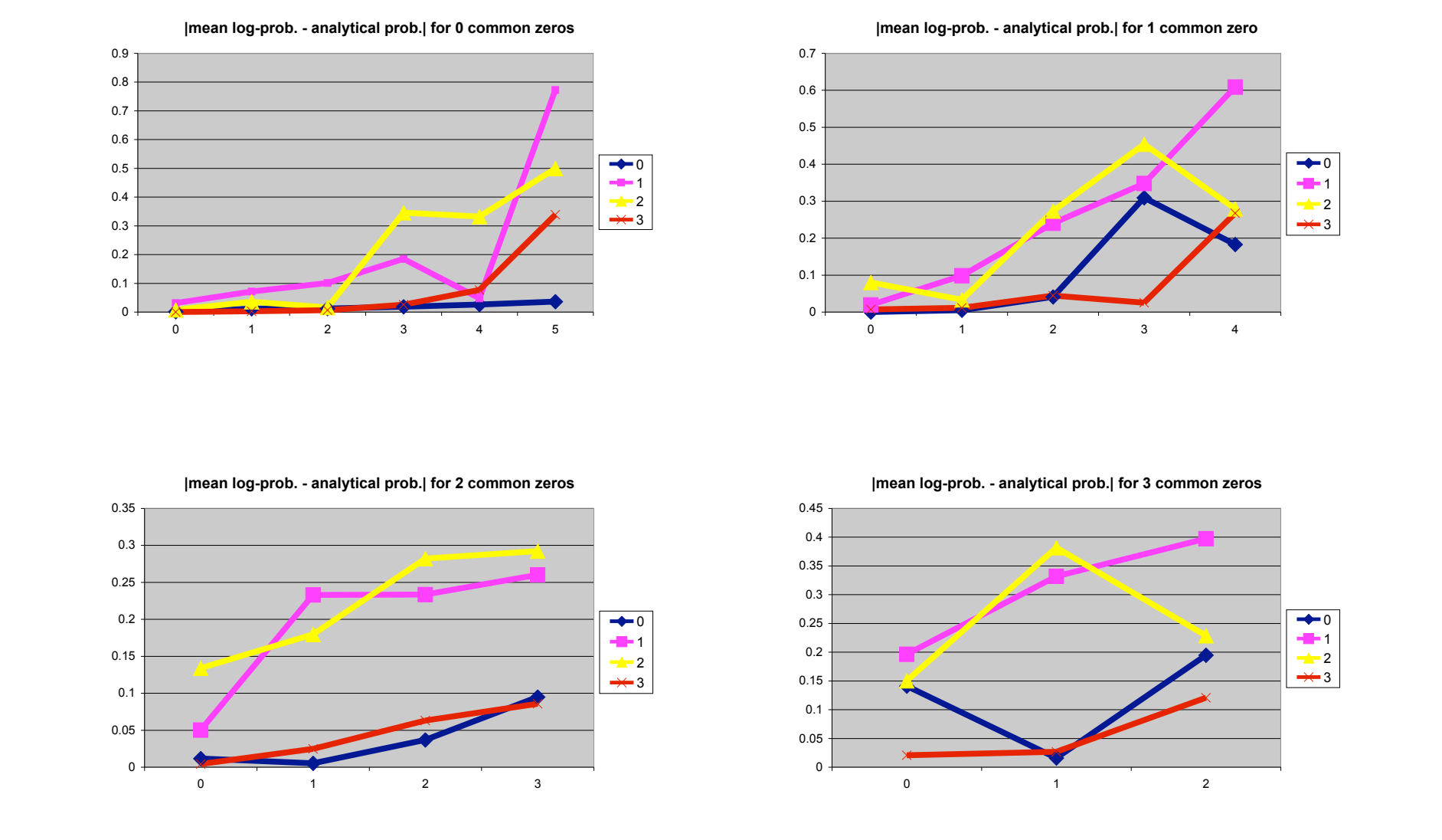


log-probabilities for the 4 methods for 1000 iterations of all data (each band is a different set of data)



log-probabilities for the 4 methods for 1000 iterations of the example data

It should be noted that these log-probabilities are weighted, and hence their average is calculated via the sum of their Importance Sampling weights, so all averages end up very close to each other/the analytical solution.
The difference between the analytical probability and the simulated mean probability for each method can also be considered.



Differences between mean log-prob. and analytical probability for systematic data [4]

Analysis

It is clear from both observing the figures showing variance in log-probability, and the differences between the analytical solution and the mean solutions, that in the vast majority of cases Methods 0 and 3 are clearly superior to Methods 1 and 2. The fact that Method 0 is so good is somewhat surprising considering that it is the simplest method.

Since Methods 0 and 3 are superior, it is possible that the lower variance of these methods can be attributed to the process of first selecting the allele i from n_H to be allocated, as Methods 0 and 3 both define which allele will be allocated before choosing which population to allocate to.

In general, the fewer zeroes n_1 and n_2 have, the closer the mean log-probabilities of Methods 0 and 3 are to the analytical solution. Method 3 with the very nearly full n_1 and n_2 is particularly close.

In terms of variance, Method 3 is generally slightly better than Method 0. The variance in the negative binomial datasets is very similar, but the difference is somewhat noticeable in the Cornuet's example data and in the mostly-full input data, except when n_1 and n_2 are completely different in the non-common-zero lineages (these bands are obvious on the overall summary figure).

Notes

1. $n_{k,i}$ is the number of type i alleles observed in a sample of n_k individuals from population k $(k = 1, 2, H)$.
2. $C(\alpha, \beta)$ is the binomial coefficient, the number of distinct ways of choosing β items from among α items.
3. This method is referred to henceforth as Method 0.
4. each of these plots represents a number of sets of data. e.g. the top left plot has no lineages i where both n_{1i} and n_{2i} are 0. The x axis shows the number of lineages i where n_{1i} and n_{2i} are both non-zero (in this simulation, the value 12 was used).

Acknowledgments

- Thanks to Martin O'Hely for supervising the project. Also for the `lmulti_d()` and `separate_inf()` functions which were used.
- Also requiring acknowledgment are the slides by Jean-Marie Cornuet defining the 3 methods of allocating the Hybrid alleles to parental populations.

For further information, please contact Robert Cope at
s4097764@student.uq.edu.au