# Multilocus multiplicative fitness: A likelihood approach

*Moi Exposito-Alonso*

*2017-11-11*

### Genotypes and selection model

Given two loci, A and B, with alleles A and a, and B and b, there must be nine possible gammetes. If the A allele and B allele are selected with coefficients $s_A$ and $s_B$, and we assume no dominance and no epistasis, we can write the average fitness per genotype very simply as the haploid model below:

| | B | b |
|---|---|---|
| A | $(1+s_A)(1+s_B)$ | $(1+s_A)(1-s_B)$ |
| a | $(1-s_A)(1+s_B)$ | $(1-s_A)(1-s_B)$ |

From a previous experiment, we have measurements of absolute fitness (percentage of survival x number of offspring) denoted as $y_{abs}$ of length $N \times r$, where $N = 515$ is the number of different genotypes and $r = 5$ is the number of replicates per genotype. Here we prefer to use relative fitness, taking as reference fitness the average value of the population, $y = y_{abs}/\bar{y}_{abs}$, so the mean of $y$ will be 1.

We have whole-genome information for those $N$ genotypes as biallelic SNPs, a total of $p$ loci. The genotypic information is represented in a genome matrix $X$ of $N$ rows and $p$ columns. The genotypes are represented as -1 for homozygotes of the reference allele and +1 for homozygote of the alternative allele (there are no heterozygotes in the dataset but they could be represented as 0).

Then we can propose that the observed fitness $y$ is a multiplicative function of the selection coefficients at every locus:

$$y = w(s, X) = \prod_{i=1}^{p} (1 + s_i \odot X_i)$$

In this function, $X_i$ works as a design matrix, and $\odot$ denotes element-wise multiplication. When the SNP is the alternative, the fitness reported by that snp is $= 1 + s \times 1$, whereas if it is the reference SNP it would be the opposite, $= 1 + s \times (-1)$, reducing in this case the mean fitness of 1.

### The probabilistic model

#### Observed fitness

We can assume that the relative fitness $y$ follows an exponential distribution. This is most likely under very selective environments where the mode of fitness is 0.

$$y \sim \text{Exp}(\lambda)$$

The moments:

$$E[y] = \lambda^{-1}$$

$$var(y) = E[y^2] - E[y] = \lambda^2$$

Lambda can be then written in terms of the fitness function dependent on selection coefficients and genotypes:

$$\lambda = \frac{1}{\prod_{i=1}^{p}(1 + s_i \odot X_i)}$$

The probability density function of the exponential distribution is:

$$f(x) = \lambda \times e^{-\lambda y}$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^n \exp\left(-\lambda n \overline{x}\right)$$

(The final expression allows me to work directly with means of fitness per genotype instead of with each replicate.)

$$L(\lambda) = \prod_{i=1}^{p}(1 + s_i \odot X_i)^{-n} \exp\left(-\prod_{i=1}^{p}(1 + s_i \odot X_i) n \overline{x}\right)$$

The derivative of the likelihood function (i.e. gradient in optimization) is :

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln(L(\lambda)) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(n \ln(\lambda) - \lambda n \overline{x}\right) = \frac{n}{\lambda} - n\overline{x} \begin{cases} > 0, & 0 < \lambda < \frac{1}{\overline{x}}, \\ = 0, & \lambda = \frac{1}{\overline{x}}, \\ < 0, & \lambda > \frac{1}{\overline{x}}. \end{cases}$$

**Finally full likelihood function**

This is the function to be minimized through a numerical optimization algorithm, the L-BFGS.

$$-\ln \ell(\vec{s}) = -\sum_{h \ \in \ haps.} \ln P(y_h | \vec{s}, X_h)$$

And the gradient to inform the optimization would be:

$$gr(\vec{s}) = N^{-1} \sum_{h \ \in \ haps.} P(y_h | \vec{s}, X_h)$$

The greatest difficulty here is that the algorithm needs to search a $p$ dimension space ($p$ equals to number of SNPs).

$$\frac{n}{\lambda} - n\overline{x}$$