Lilelihood of observed fitness given selection and epistasis

Moi Exposito-Alonso

2017-11-08

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. For simplicity we use an haploid model, which would be similar to the diploid homozygotes case (also the case of highly self-fertilizing species):

	B	b
A	$[(1+s_A)(1+s_B)]^\epsilon$	$[(1+s_A)(1-s_B)]^e$
a	$[(1-s_A)(1+s_B)]^e$	$[(1-s_A)(1-s_B)]^\epsilon$

We could then expect that the observed fitness y of one of N number of haplotype of p loci is expressed as:

$$y = \Big[\prod_{i=1}^{p}(1+s_i \odot X_i)\Big]^e$$

The probabilistic model

We can assume the data y_h follows a Gamma distribution (typically with α and $beta$ parameters or $\alpha = k$ and $\theta = 1/\beta$),

$$y \sim \Gamma(\alpha, \beta)$$

The first moment is:

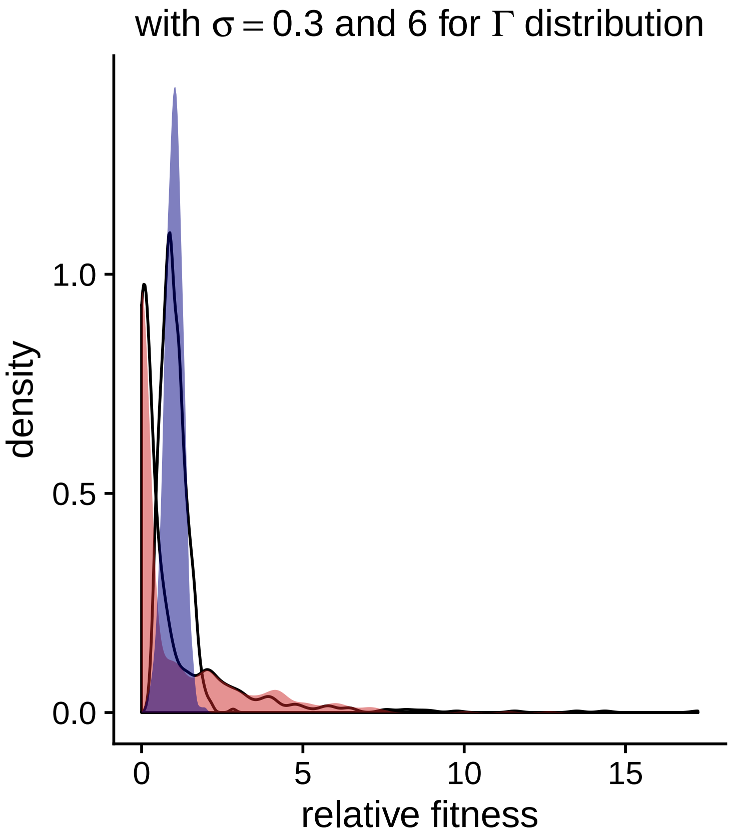
$$E[y] = \alpha/\beta = 1$$

Thus, $\alpha = \beta$ so we use instead σ. This special type of Γ distribution will always generate a distribution whose mean is in 1, as any relative fitness situation. What contains the information is the shape of the gamma distribution, hence our use of σ.

The second moment is then:

$$Var(y) = \alpha/\beta^2 = \sigma^{-1}$$

In order to show that this distribution is appropriate, here are two relative fitnesses, in Madrid under drought and high water conditions. Only by moving σ we obtain almot perfect fit.


with σ = 0.3 and 6 for Γ distribution

Then, we express σ in terms of selection coefficients and the epistatic term:

$$Var(y) = E[(y - E[y])]^2 = E[y^2] - E[y]^2 = E[y^2]$$

$$= N^{-1} \ sum\Big\{ \prod_{i=1}^{p} (1 + s_i \odot X_i)^{2e} \Big\} - 1 = \frac{1}{\sigma}$$

$$\sigma = \frac{N}{sum\Big\{ \prod_{i=1}^{p} (1 + s_i \odot X_i)^{2e} \Big\} - 1}$$

The probability density function of Gamma in terms of k and θ is:

$$f(y) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} e^{-\frac{y}{\theta}}$$

In log form this can be written as:

$$\ln f(y) = (k-1)\ln(y) - \frac{y}{\theta} - \ln\Gamma(k) - k\ln\theta$$

What in terms of σ can be written as:

$$\ln f(y) = (\sigma - 1)\ln(y) - y\sigma - \ln\Gamma(\sigma) + \sigma\ln(\sigma)$$

The derivative of this function, needed as gradient for the optimization is:

Then we can generate the full log likelihood in terms of the selection coefficients, epistatic term, and genotypes by substituting σ in the above equation (not shown).

We can express this likelihood for the sake of simplicity as:

$$-\ln \ell(\vec{s}, e) = -\ln f(y_h \ ; \sigma)$$

If we wanted to apply this to dense data, we can use some regulaization. $L_1 = \lambda||w||_1$, $L_2 = \lambda||w||_2^2$ norms, or a combination, $L_{12} = \lambda(m||w||_1 + (1-m)||w||_2^2), m \in [0,1]$ typically used from regression approaches could be used. A plane L_0 norm equal to the total number of non-zero coefficients could be adapted to the percentage of non-zero values thus $L_0 \in [0,1]$, which in log would be of the same scale as the likelihood function.

$$-\ln \ell(\vec{s}, e) \quad + \ln(L_0)$$

Extension where e is a random variable (needs to be worked out)

An important extension of this method would be to instead of estimating a common epistatic term for the whole genome, we infer its ditribution.

To do that we integrate the likelihood over the distribution of e given some parameters θ_e of such distribution with known probability denstiy function.

$$\ln \ell(\vec{s}, \theta_e) = \int_{-\inf}^{+\inf} \ln \ell(\vec{s}, e) f(e; \theta_e) de$$

Or for feasibility, the e distribution can be bined in equal sizes and approximate:

$$\ln \ell(\vec{s}, \theta_e) \approx \sum_{m \in e \ bins} \ln \ell(\vec{s}, e) P(e; \theta_e)$$

The distribution of e should be optimally a conjugate of the likelihood.