Statistical Methods

immediate

July 9, 2022

Likelihood

The likelihood in Eyalshiv et al. (2016) has the form,

$$\log \mathcal{L} = \sum_{v \in \mathcal{V}} \sum_{i \neq j \in \mathcal{S}} \log(P(O_{i,j}(v)|\theta)) \tag{1}$$

where $\mathcal{V}$ is the set of putatively neutral sites, $\mathcal{S}$ is the set of samples, and θ are the BGS parameters. The indicator variable $O_{i,j}(v)$ is 1 if samples i and j are different at site v, and zero otherwise. Thus as they specify in the paper,

$$P(O_{i,j}(v)|\theta) = \begin{cases} \pi(v|\theta), & O_{i,j}(v) = 1 \\ 1 - \pi(v|\theta), & O_{i,j}(v) = 0 \end{cases}. \tag{2}$$

The total size of the set of samples $\mathcal{S}$ is $n_{\mathcal{S}} = |\mathcal{S}|$. Assuming all sites are biallelic, we can simplify the calculation of the likelihood by noting that for a site v's vector of allele counts $[c_1, c_2]$, the total number of pairwise combinations with the same alleles is

$$n_s(v) = \binom{c_1}{2} + \binom{c_2}{2} \tag{3}$$

and the number of different pairwise combinations is

$$n_d(v) = n_T - n_s(v) \tag{4}$$

where $n_T = n_{\mathcal{S}}(n_{\mathcal{S}} - 1)/2$ is the total number of pairwise combinations across the sample set $\mathcal{S}$. Furthermore, note that we can partition the set $\mathcal{V}$ of neutral sites into polymorphic ($\mathcal{P}$) and fixed sites ($\mathcal{F}$), i.e. $\mathcal{V} = \mathcal{P} \cup \mathcal{F}$ and $\mathcal{P} \cap \mathcal{F} = \varnothing$. For all $v \in \mathcal{F}$, $n_d(v) = 0$ and $n_s(v) = n_T$.

Then,

$$\log \mathcal{L} = \sum_{v \in \mathcal{V}} [\log(\pi(v|\theta)) n_{\mathrm{D}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{S}}(v)] \tag{5}$$

$$= \sum_{v \in \mathcal{P}} [\log(\pi(v|\theta)) n_{\mathrm{D}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{S}}(v)] + \sum_{v \in \mathcal{F}} \log(1 - \pi(v|\theta)) n_T \tag{6}$$

$$\tag{7}$$

1 B Scores