Learning Theory Notes

immediate

June 5, 2022

Learning Limits

The Simulation Sample $\bar{B}$ Estimator

Our goal is to approximate a function that describes the reduction in pairwise diversity (relative to the neutral expectation) B due to BGS for some set of evolutionary parameters $\mathbf{x}$. The true reduction function is a conditional expectation $B(\mathbf{x}) = \mathbb{E}[T_2(\mathbf{x})|\mathbf{x}]$ taken over an infinite number of evolutionary replicates. Here $T_2(\mathbf{x})$ is the pairwise coalescent rate under BGS for parameters $\mathbf{x}$, with time scaled in $2N$ generation units.

We approximate this function $B(\mathbf{x})$ with $\widehat{B}(\mathbf{x})$ from evolutionary simulations using statistical learning. In each evolutionary simulation, we sample some parameter set $\mathbf{x}$ and evolve r independent populations forward in time and observe the resulting genealogy at a neutral site a given recombination distance away (this is one of the parameters). From this full population genealogy of $2N$ gametes, we estimate the reduction in neutral diversity as

$$\bar{B} = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{\pi}_i}{4N\mu} \tag{1}$$

where $\hat{\pi}$ is Tajima's estimator for pairwise diversity within a tree. For now, we can imagine letting the neutral mutations saturate as $\mu \to 1$ so we can ignore mutation rate hereafter. If we take expectation over the evolutionary process,

$$\begin{aligned}\mathbb{E}[\bar{B}] &= \frac{1}{4Nr}\sum_{i=1}^{r}\mathbb{E}[\hat{\pi}_i] &(2)\\ &= \frac{2T_2}{4N} &(3)\\ &= B &(4)\end{aligned}$$

thus, the estimated reduction in diversity from simulations $\bar{B}$ is unbiased, since $\mathbb{E}[\hat{\pi}] = 2T_2 = 4BN$.

The variance of the estimator $\bar{B}$

Using the law of total variance, we can write the variance of $\bar{B}$ as

$$\mathrm{Var}(\bar{B}) = \underbrace{\mathbb{E}(\mathrm{Var}(\bar{B}|\mathcal{G}))}_{\text{sampling noise}} + \underbrace{\mathrm{Var}(\mathbb{E}(\bar{B}|\mathcal{G}))}_{\text{evolutionary variance}} \tag{5}$$

where $\mathcal{G}$ is the genealogy at the focal neutral site affected by BGS at some segment. The distribution of genealogies $P(\mathcal{G})$ is unknown and quite complex under BGS, but we can use a simplifying assumption to get a rough estimate of the variance of $\bar{B}$.

Under strong BGS (e.g. $\mu/sN \gg 1$), the genealogy is often modeled as a neutral coalescent process with rescaled with an effective population size $N_e = BN$ (though see Cvijović et al. 2018). In the weak selection limit $s \to 0$, approximation becomes exact as BGS becomes effectively neutral. In both regimes, selection is treated as rescaling a neutral genealogical process to $N_e = BN$. We refer to this the **neutral-BGS model**, and it allows for us to approximate $\mathrm{Var}(\bar{B})$ with some $\mathrm{Var}(\bar{B}')$ where $\bar{B}'$ is the mean reduction if we averaged over neutral genealogies from a population of BN diploids, rather than averaged over exact selection simulations. Note that at $B = 1$, the neutral-BGS model becomes exact, as there is no reduction in diversity due to background selection and the only evolutionary variance is through the neutral coalescent process.

If the genealogies are neutral, we can calculate the variance of $\bar{B}$ using Tajima's equation for the variance of $\hat{\pi}$. For n samples and r evolutionary replicates,

$$\begin{aligned}
\mathrm{Var}(\bar{B}') &= \frac{1}{16N^2r^2}\sum_{i=1}^{r}\mathrm{Var}(\hat{\pi}_i) && (6)\\
&= \frac{\mathrm{Var}(\hat{\pi}_i)}{16N^2r} && (7)\\
&= \frac{1}{16N^2r}\left(\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2\right) && (8)\\
&= \frac{1}{16N^2r}\left(\frac{n+1}{3(n-1)}4BN + \frac{2(n^2+n+3)}{9n(n-1)}16B^2N^2\right) && (9)\\
&= \underbrace{\frac{n+1}{12(n-1)}\frac{B}{Nr}}_{\text{sampling noise}} + \underbrace{\frac{2(n^2+n+3)}{9n(n-1)}\frac{B^2}{r}}_{\text{evolutionary variance}} && (10)
\end{aligned}$$

Now, let us look at the consistency of the estimator $\bar{B}$ (we drop the parameter set k for clarity) both in r (over evolutionary replicates) and in n (as the sample size increases). Let $\bar{B}_r$ be the estimator of B after r evolutionary replicates (conditioning on some n). By Chebyshev's inequality and for some $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{B}_r - B| \geq \epsilon\right) \leq \frac{\mathrm{Var}(\bar{B}_r)}{\epsilon^2} \tag{11}$$

and since $\lim_{r\to\infty}\mathrm{Var}(\bar{B}_r) = 0$,

$$\lim_{r\to\infty}\mathbb{P}\left(|\bar{B}_r - B| \geq \epsilon\right) = 0. \tag{12}$$

Thus, $\bar{B}_r \xrightarrow{p} B$ as $r \to \infty$ and $\bar{B}_r$ is consistent in r. Now, let us look at the consistency of $\bar{B}_n$ in sample size n. Note that $n \leq 2N$ (i.e. our sample size is bounded by the number of gametes in the population), so we imagine setting $n = 2N$ and taking the limit $N \to \infty$,

$$\lim_{N\to\infty} \mathbb{P}\left(|\bar{B}_N - B| \geq \epsilon\right) \leq \lim_{N\to\infty} \frac{\mathrm{Var}(\bar{B}_N)}{\epsilon^2} \tag{13}$$

$$\lim_{N\to\infty} \mathbb{P}\left(|\bar{B}_N - B| \geq \epsilon\right) \leq \frac{2B^2}{9r\epsilon^2}. \tag{14}$$

Thus, even if we sample the entire population, and let the population size $N \to \infty$, $\bar{B}_n$ is still an inconsistent estimator in n. Intuitively this is because the evolutionary variance is not decreased by taking more samples, *only by taking more evolutionary replicates*.

Below I work through where this bound came from. If we estimate pairwise diversity from the genealogical tree of the entire population such that $n = 2N$,

$$\mathrm{Var}(\bar{B}) = \frac{n+1}{12(n-1)}\frac{B}{Nr} + \frac{2(n^2+n+3)}{9n(n-1)}\frac{B^2}{r} \tag{15}$$

$$= \frac{2N+1}{12(2N-1)}\frac{B}{Nr} + \frac{2(4N^2+2N+3)}{18N(2N-1)}\frac{B^2}{r} \tag{16}$$

$$\approx \frac{B}{12Nr} + \frac{2N^2+N}{9N^2}\frac{B^2}{r} \tag{17}$$

If we take the infinite population size limit,

$$\lim_{N\to\infty} \mathrm{Var}(\bar{B}) = \frac{2B^2}{9r} \tag{18}$$

In practice, we use a different, better estimator of pairwise diversity than Tajima's π: branch statistic pairwise diversity (Ralph 2019, Ralph et al. 2020). This estimator removes mutation as a source of sampling noise and since $n = 2N$, our only source of variance in $\bar{B}$ is in the evolutionary process.

Partitioning the Prediction Error

To learn the function $B(\mathbf{x})$, we run evolutionary simulations under a training set of parameters, giving us $\{(\mathbf{x}_1, \bar{B}_1), (\mathbf{x}_2, \bar{B}_2), \ldots, (\mathbf{x}_n, \bar{B}_n)\}$, where each $\bar{B}_i$ is the observed reduction over r replicates. Since our goal is to approximate $\hat{B}(\mathbf{x})$ as closely as possible to $B(\mathbf{x})$ over the distribution of $\mathbf{x}$ most close to that across the genome, we find some $\hat{B}(\mathbf{x})$ that minimizes expected loss. An important metric, even if training is conducted with a different loss, is the mean squared error

$$MSE(\bar{B}, \hat{B}) = \mathbb{E}[(\bar{B}(\mathbf{x}) - \hat{B}(\mathbf{x}))^2]. \tag{19}$$

We can think of $\bar{B}(\mathbf{x})$ as a true reduction $B(\mathbf{x})$ plus a deviation ε due to the particular evolutionary replicates simulated, e.g. $\bar{B}(\mathbf{x}) = B(\mathbf{x}) + \varepsilon$.

$$MSE(\bar{B}, \hat{B}) = \mathbb{E}[(\bar{B}(\mathbf{x}) - \hat{B}(\mathbf{x}))^2] \tag{20}$$

$$= \mathbb{E}[(B(\mathbf{x}) + \varepsilon - \hat{B}(\mathbf{x}))^2] \tag{21}$$

$$= \underbrace{\mathrm{Var}(\hat{B}(\mathbf{x}))}_{\text{training variance}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{evolutionary variance}} + \underbrace{\mathbb{E}[\hat{B}(\mathbf{x}) - B(\mathbf{x})]^2}_{\text{bias}^2} \tag{22}$$

At $B = 1$, the evolutionary variance is exactly the neutral variance given by $\mathrm{Var}(\varepsilon) = {}^{2B^2}/_{9r}$. Thus, MSE loss has an absolute bound at $B = 1$,

$$MSE(\bar{B}, \hat{B}|B = 1) \geq \frac{2}{9r} \tag{23}$$

For $B < 1$, we can write the evolutionary variance as some deviation away from the coalescent noise under the neutral-BGS model $\mathrm{Var}(\varepsilon) = \mathrm{Var}(\bar{B}') + e$.

Multiplicative Error Model

The predicted reduction for segment i from the machine learned model $\hat{B}_i$ can be partitioned as $\hat{B}_i = B_i + \epsilon_i$, where ϵ_i is the cumulative error due to bias, irreducible evolutionary variance, and training variance. Then, our total predicted reduction at some focal site v is:

$$\hat{B}(v) = \prod_i^n (B_i + \epsilon_i) \tag{24}$$

We want to understand the bias and variance of this product estimator based on the total learning error. Taking logs,

$$\hat{B}(v) = \prod_i^n (B_i + \epsilon_i) \tag{25}$$

$$\log(\hat{B}(v)) = \log\left(\prod_i^n (B_i + \epsilon_i)\right) \tag{26}$$

$$= \sum_i^n \log(B_i + \epsilon_i). \tag{27}$$

$$\tag{28}$$

If we approximate the expected log with a Taylor series,

$$\mathbb{E}[\log(B_i + \epsilon_i)^2] \approx \log(B_i)^2 + \frac{2\log(B_i)}{B_i}\mathbb{E}[\epsilon_i] + \frac{(1 - \log(B_i))}{B_i^2}\mathbb{E}[\epsilon_i^2] \tag{29}$$

$$\approx \log(B_i)^2 + \frac{2\log(B_i)}{B_i}b + \frac{(1 - \log(B_i))}{B_i^2}(\sigma^2 + b^2) \tag{30}$$

Then,

$$\mathbb{E}[\log(\hat{B}(v))] = \sum_i^n \mathbb{E}[\log(B_i + \epsilon_i)] \tag{31}$$

$$\approx \sum_i^n \log(B_i) + \sum_i^n \frac{b}{B_i} - \sum_i^n \frac{\sigma^2 + b^2}{2B_i^2} \tag{32}$$

$$\tag{33}$$

Bias:

$$\mathbb{E}[\log(\hat{B}(v))] - \mathbb{E}[\log(B(v))] = \sum_i^n \mathbb{E}[\log(B_i + \epsilon_i)] - \sum_i^n \log(B_i) \tag{34}$$

$$\approx \sum_i^n \log(B_i) + \sum_i^n \frac{b}{B_i} - \sum_i^n \frac{\sigma^2 + b^2}{2B_i^2} - \sum_i^n \log(B_i) \tag{35}$$

$$\approx \sum_i^n \frac{b}{B_i} - \sum_i^n \frac{\sigma^2 + b^2}{2B_i^2} \tag{36}$$

By Jensen's inequality $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ for some convex function f,

Likelihood

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

$$\log \mathcal{L} = \sum_{v \in V} \sum_{i \neq j \in S} \log(P(O_{i,j}(v)|\theta)) \tag{37}$$

where V is the collection of neutral sites, 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,

$$P(O_{i,j}(v)|\theta) = \begin{cases} \pi(v|\theta), & O_{i,j}(v) = 1 \\ 1 - \pi(v|\theta), & O_{i,j}(v) = 0 \end{cases}. \tag{38}$$

For S samples, there are $\binom{|S|}{2}$ pairwise comparisons that we can partition into the different and same classes; we assign their counts to $n_{\mathrm{D}}(v)$ and $n_{\mathrm{S}}(v)$. Then,

$$\log \mathcal{L} = \sum_{v \in V} (\log(\pi(v|\theta)) n_{\mathrm{S}}(v) + \log(1 - \pi(v|\theta)) n_{\mathrm{D}}(v)) \tag{39}$$

Since B is discretized into genomic windows of length K basepairs, $\pi(v|\theta)$ is constant in a window.