Learning Theory Notes

immediate

May 28, 2022

Learning Limits

The Simulation Sample $\bar{B}$ Estimator

Our goal is to approximate the function $B = f(X)$ with some learned function $\widehat{B} = \hat{f}(X)$ from evolutionary simulations. In each evolutionary simulation, we sample some evolutionary parameters X_k and evolve r independent populations forward in time under these parameters and observe some genealogy. From this genealogy, we estimate the reduction in neutral diversity as:

$$\bar{B}_k = \frac{1}{r} \sum_{i=1}^{r} \frac{\hat{\pi}_{k,i}}{4N\mu} \tag{1}$$

there $\hat{\pi}$ is Tajima's estimator for pairwise diversity within a tree. When we estimate $\hat{\pi}$ from using branch-mode tree statistics from an observed genealogy, $\mu \to 1$, so we can ignore mutation. If we take expectation over the evolutionary process,

$$\begin{aligned}
\mathbb{E}(\bar{B}_k) &= \frac{1}{4Nr} \sum_{i=1}^{r} \mathbb{E}(\hat{\pi}_{k,i}) && (2)\\
&= \frac{T_k^{(2)}}{4N} && (3)\\
&= \frac{4B_k N}{4N} && (4)\\
&= B_k && (5)\\
& && (6)
\end{aligned}$$

thus, the estimated reduction in diversity from simulations $\bar{B}$ is unbiased, since $\mathbb{E}(\hat{\pi}_k) = 2T_k^{(2)} = 4B_k N$. For independent evolutionary replicates, we calculate the variance of this estimator using Tajima's equation for the variance of $\hat{\pi}$ as

$$\mathrm{Var}(\bar{B}_k) = \frac{1}{16N^2r^2}\sum_{i=1}^{r}\mathrm{Var}(\hat{\pi}_{k,i}) \tag{7}$$

$$= \frac{\mathrm{Var}(\hat{\pi}_{k,i})}{16N^2r} \tag{8}$$

$$= \frac{1}{16N^2r}\left(\frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2\right) \tag{9}$$

$$= \frac{1}{16N^2r}\left(\frac{n+1}{3(n-1)}4BN + \frac{2(n^2+n+3)}{9n(n-1)}16B^2N^2\right) \tag{10}$$

$$= \underbrace{\frac{n+1}{12(n-1)}\frac{B}{Nr}}_{\text{sampling noise}} + \underbrace{\frac{2(n^2+n+3)}{9n(n-1)}\frac{B^2}{r}}_{\text{evolutionary variance}} \tag{11}$$

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 $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{B}_r - B| \geq \epsilon\right) \leq \frac{\mathrm{Var}(\bar{B}_r)}{\epsilon^2} \tag{12}$$

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

$$\lim_{r\to\infty}\mathbb{P}\left(|\bar{B}_r - B| \geq \epsilon\right) = 0. \tag{13}$$

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 $N \to \infty$,

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

$$\lim_{N\to\infty}\mathbb{P}\left(|\bar{B}_N - B| \geq \epsilon\right) \leq \frac{2B^2}{9r\epsilon^2}. \tag{15}$$

$$\tag{16}$$

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

In our case, we estimate pairwise diversity from the genealogical tree of the entire population, so $n = 2N$

$$\mathrm{Var}(\bar{B}_k) = \frac{n+1}{12(n-1)}\frac{B}{Nr} + \frac{2(n^2+n+3)}{9n(n-1)}\frac{B^2}{r} \tag{17}$$

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

$$\approx \frac{B}{12Nr} + \frac{2N^2+N}{9N^2}\frac{B^2}{r} \tag{19}$$

If we take the infinite population size limit,

$$\lim_{N\to\infty} \mathrm{Var}(\bar{B}_k) = \frac{2B^2}{9r} \tag{20}$$