Lecture 20 : Infinite-sites model

MATH285K - Spring 2010 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 1.4].

1 Infinite-sites model

We consider a sequence-based model of evolution similar to the infinite-alleles model. Imagine that each individual has an infinitely long DNA sequence. As in the infinite-alleles model, mutations occur in each individual at rate $\theta/2$. Any time a mutation occurs, it creates a state change in a new position of the sequence. One formulation is as follows:

DEF 20.1 (Infinite-sites model) *For a partition Π on n samples, we let $|\Pi|$ be the number of sets in Π. Consider the following algorithm with n and θ as input:*

- *Set $\Pi = \{\{1\}, \ldots, \{n\}\}$.*
- *Repeat until $|\Pi| = 1$:*
  - *Setting $k := |\Pi|$, after an exponential time with parameter $\binom{k}{2} + k\frac{\theta}{2}$ (going backwards in time):*
    - *With probability $\frac{\theta}{\theta+k-1}$, generate a mutation at new position in a uniformly random lineage.*
    - *With probability $\frac{k-1}{\theta+k-1}$, merge two uniformly random lineages.*

Each sample inherits a sequence mutated at the sites encountered along the way to the root.

In fact, there are two cases: whether or not the ancestral sequence is assumed known.

2 Segregating sites

A natural way to estimate θ in this model is to consider the number of segregating sites S_n in the sample, that is, the number of positions at which the sequences differ. We let T_{tot} be the total length of the coalescent tree.

LEM 20.2 *We have*

$$\mathbb{E}[S_n] = \theta h_n,$$

where $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

Proof: Conditioning on T_{tot}, S_n is Poisson with mean $\frac{\theta}{2} T_{\text{tot}}$

$$\mathbb{E}[S_n] = \mathbb{E}[\mathbb{E}[S_n \,|\, T_{\text{tot}}]] = \mathbb{E}\left[\frac{\theta}{2} T_{\text{tot}}\right] = \frac{\theta}{2}\mathbb{E}[T_{\text{tot}}].$$

To compute the expectation of T_{tot}, we divide the process into stages with j lineages for $j = n, \ldots, 2$ which last t_j respectively where t_j is exponential with parameter $\binom{j}{2}$ independently of the other stages. Then

$$\mathbb{E}[T_{\text{tot}}] = \sum_{j=2}^{n} j\mathbb{E}[t_j] = \sum_{j=2}^{n} \frac{2}{j-1} = 2\sum_{i=1}^{n-1} \frac{1}{i}.$$

■

LEM 20.3 *We have*

$$\text{Var}[S_n] = \theta h_n + \theta^2 g_n,$$

where $g_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$.

Proof: Conditioning again on T_{tot}, S_n is Poisson with mean (and variance) $\frac{\theta}{2} T_{\text{tot}}$

$$\begin{aligned}
\text{Var}[S_n] &= \mathbb{E}[\text{Var}[S_n \,|\, T_{\text{tot}}]] + \text{Var}[\mathbb{E}[S_n \,|\, T_{\text{tot}}]] \\
&= \mathbb{E}\left[\frac{\theta}{2} T_{\text{tot}}\right] + \text{Var}\left[\frac{\theta}{2} T_{\text{tot}}\right] \\
&= \frac{\theta}{2}\mathbb{E}[T_{\text{tot}}] + \frac{\theta^2}{4}\text{Var}[T_{\text{tot}}].
\end{aligned}$$

By the same reasoning as in the previous proof,

$$\text{Var}[S_n] = \sum_{j=2}^{n} j^2 \text{Var}[t_j] = \sum_{j=2}^{n} \frac{4}{(j-1)^2} = 4\sum_{i=1}^{n-1} \frac{1}{i^2}.$$

■

Therefore:

THM 20.4 (Watterson's estimator) *The estimator*

$$\theta_W = \frac{S_n}{h_n},$$

is unbiased for θ*. Its variance is*

$$\text{Var}[\theta_W] = \theta \frac{1}{h_n} + \theta^2 \frac{g_n}{h_n^2},$$

which converges to 0*.*

3 Cramer-Rao Bound

To get a sense of how much better we can do when estimating θ, we consider an hypothetical likelihood approach. We first recall the Cramer-Rao bound—for simplicity we restrict ourselves to the discrete case with a single parameter.

THM 20.5 (Cramer-Rao bound) *Let X be a random vector on a finite state space S with law p_θ. Assume that p_θ is differentiable with respect to θ for all $x \in S$. If $\phi(X)$ is an unbiased estimator of θ, then*

$$\mathrm{Var}_\theta[\phi(X)] \geq \frac{1}{I_X(\theta)},$$

where

$$I_X(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right],$$

is the Fisher information.

Proof: Differentiating in

$$1 = \sum_{x\in S} p_\theta(x),$$

and

$$\theta = \sum_{x\in S} \phi(x)p_\theta(x),$$

gives respectively

$$0 = \sum_{x\in S}\frac{\partial}{\partial\theta}p_\theta(x) = \sum_{x\in S}\frac{\frac{\partial}{\partial\theta}p_\theta(x)}{p_\theta(x)}p_\theta(x) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p_\theta(x)\right], \tag{1}$$

and

$$1 = \sum_{x\in S}\phi(x)\frac{\partial}{\partial\theta}p_\theta(x) = \mathbb{E}_\theta\left[\phi(X)\frac{\partial}{\partial\theta}\log p_\theta(X)\right]. \tag{2}$$

Using (1) into (2) and applying Cauchy-Schwarz, we get

$$\begin{aligned} 1 &= \mathbb{E}_\theta\left[(\phi(X) - \mathbb{E}_\theta[\phi(X)])\frac{\partial}{\partial\theta}\log p_\theta(X)\right] \\ &\leq \sqrt{\mathbb{E}_\theta\left[(\phi(X) - \mathbb{E}_\theta[\phi(X)]^2)\right]}\sqrt{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^2\right]}. \end{aligned}$$

To see the second expression for the Fisher information, note that

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(p_\theta(x)\frac{\partial}{\partial\theta}\log p_\theta(x)\right) &= p_\theta(x)\frac{\partial^2}{\partial\theta^2}\log p_\theta(x) + \frac{\partial}{\partial\theta}p_\theta(x)\frac{\partial}{\partial\theta}\log p_\theta(x) \\ &= p_\theta(x)\frac{\partial^2}{\partial\theta^2}\log p_\theta(x) + p_\theta(x)\left(\frac{\partial}{\partial\theta}\log p_\theta(x)\right)^2.\end{aligned}$$

Summing over x and using (1) gives the result. ■

Imagine that we had access to the number of segregating sites s_j generated when the number of lineages was j. (In reality, we do not.) Clearly, by the description of the infinite-sites model above, s_j is geometric (shifted) and the log-likelihood of the data is

$$\begin{aligned}\ell_n(\theta) &= \log\prod_{j=2}^{n}\left(\frac{\theta}{\theta+j-1}\right)^{s_j}\left(\frac{j-1}{\theta+j-1}\right) \\ &= \log(n-1)! + S_n\log\theta - \sum_{j=2}^{n}(s_j+1)\log(\theta+j-1).\end{aligned}$$

To compute the Fisher information note

$$-\frac{\partial^2}{\partial\theta^2}\ell_n(\theta) = \frac{S_n}{\theta^2} - \sum_{j=2}^{n}\frac{s_j+1}{(\theta+j-1)^2}.$$

Note that s_j+1 is geometric with mean one over the success probability, that is,

$$\mathbb{E}_\theta[s_j+1] = \frac{\theta+j-1}{j-1},$$

and $\mathbb{E}_\theta[S_n] = \theta h_n$. Hence

$$\begin{aligned}I_X(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell_n(\theta)\right] \\ &= \frac{h_n}{\theta} - \sum_{j=2}^{n}\frac{1}{(j-1)(\theta+j-1)} \\ &= \frac{1}{\theta}\sum_{i=1}^{n-1}\left[\frac{1}{i} - \frac{\theta}{i(\theta+i)}\right] \\ &= \frac{1}{\theta}\sum_{i=1}^{n-1}\frac{1}{\theta+i}.\end{aligned}$$

Note that

$$\frac{\mathrm{Var}[\theta_W]}{1/I_X(\theta)} \to 1,$$

as $n \to \infty$ for fixed θ. In other words, in the large sample limit even if we were given knowledge of the s_j's we couldn't expect to obtain an unbiased estimator of θ much better than θ_W. However, for fixed n as $\theta \to \infty$,

$$\frac{\mathrm{Var}[\theta_W]}{1/I_X(\theta)} \to \frac{(n-1)g_n}{h_n^2},$$

which can mean that θ_W may have a much larger variance.

On the basis of this observation, it has been propsed that ML approaches may offer better estimates. Using the Splits-Equivalence Theorem, we can get a rough estimate of the coalescent tree as well as of the branches on which mutations have occured (assume that we know the ancestral sequence—for instance by looking at closely related species). However, determining the s_j's would require knowing exactly where on the branches these mutations have occured. A likelihood approach can integrate over all these possibilities. In fact, as in the case of the infinite-alleles model, recursions for the likelihood can be derived, although no analytic solution is known. See [Tav04].

Further reading

The material in this section was taken from Section 1.4 of the excellent monograph [Dur08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[Tav04] Simon Tavaré. Ancestral inference in population genetics. In *Lectures on probability theory and statistics*, volume 1837 of *Lecture Notes in Math.*, pages 1–188. Springer, Berlin, 2004.