Lecture 19 : Chinese restaurant process

MATH285K - Spring 2010 *Lecturer: Sebastien Roch*

References: [Dur08, Chapter 1.3].

Previous class

Recall Ewens' sampling formula (ESF).

THM 19.1 (Ewens' sampling formula) *Letting $\|\mathbf{a}\| = \sum_{i=1}^n i a_i$, in a sample of size n we have*

$$q(\mathbf{a}) = \mathbb{1}\{\|\mathbf{a}\| = n\} \frac{n!}{\theta_{(n)}} \prod_{i=1}^n \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!},$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$.

1 Proof of ESF

We begin with a proof of the ESF. This is essentially Kingman's proof (understanding ESF was the main motivation for his introduction of the coalescent).

Proof: Let Π be the partition generated by the infinite-alleles model on n sample. We call each set in Π a *cluster*. Assume there are k clusters. Looking backwards in time, to obtain Π it must be that each cluster undergoes a sequence of coalescences followed by a single mutation: more precisely, a cluster of size λ_i goes through $\lambda_i - 1$ coalescences then a single mutation. Once this happens, we say that the cluster has *merged*. Moreover, no coalescence event can occur between two unmerged clusters—we call such an event *invalid*.

We now compute the probability that these events occur. The probability that a coalescence occurs within a cluster with j lineages remaining when the total number of lineages remaining is i is given by

$$\frac{i-1}{\theta+i-1} \frac{\binom{j}{2}}{\binom{i}{2}} = \frac{j(j-1)}{i(\theta+i-1)}.$$

Similarly, the probability that a mutation occurs in a cluster with 1 lineage remaining when the total number of lineages reamining is i is given by

$$\frac{\theta}{\theta + i - 1}\frac{1}{i} = \frac{\theta}{i(\theta + i - 1)}.$$

Note that in both expressions, the denominator depends only on i and the numerator depends only on j.

Hence, the probability that a particular sequence of valid events occurs must be

$$\frac{\prod_{i=1}^{k} \theta \lambda_i!(\lambda_i - 1)!}{n!\theta_{(n)}}.$$

The particular sequence of valid events is not important to us so we multiply by the number of ways of choosing $\lambda_1, \ldots, \lambda_k$ positions among the n time slots

$$\frac{\theta^k \prod_{i=1}^{k} \lambda_i!(\lambda_i - 1)!}{n!\theta_{(n)}} \times \frac{n!}{\lambda_1! \cdots \lambda_k!} = \frac{\theta^k \prod_{i=1}^{k} (\lambda_i - 1)!}{\theta_{(n)}}.$$

(Note that we ignore those events that are valid but irrelevant like the coalescence between a merged cluster and an unmerged cluster. The properties of exponentials allow us to do so.)

Since we are looking for the probability of the allele frequencies, we multiply by the number of ways of choosing the elements in each cluster—not distinguishing between clusters of the same size—to get

$$\begin{aligned} q(\mathbf{a}) &= \frac{\theta^k \prod_{i=1}^{k} (\lambda_i - 1)!}{\theta_{(n)}} \times \frac{n!}{\lambda_1! \cdots \lambda_k!} \times \frac{1}{a_1! \cdots a_n!} \\ &= \frac{n! \theta^{\sum_i a_i} \prod_{i=1}^{k} \left(\frac{1}{\lambda_i}\right)}{\theta_{(n)} a_1! \cdots a_n!} \\ &= \frac{n!}{\theta_{(n)}} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{a_i} \frac{1}{a_i!}. \end{aligned}$$

■

2 A better estimator for θ

Implicit in Kingman's proof of ESF is the following urn process due to Hoppe.

DEF 19.2 (Hoppe's urn) *An urn contains a black ball and a certain number of balls of other colors. At each time step, a ball is picked at random, the black having weight θ and all other balls having weight* 1*. If the black ball is picked, a ball of a new color is added to the urn (and the black ball is replaced). If a ball of another color is chosen, a new ball of the same color is added to the urn (and the ball picked is returned to the urn). The process starts with a single black ball. We stop when we have n non-black balls.*

From the proof of the ESF given above, the distribution of the frequencies of the non-black balls is given by Ewens' formula. (Just look at the process going backwards in time.)

To illustrate the use of Hoppe's urn, we consider a different estimator of θ. A natural quantity which is influenced by the mutation rate is the total number of alleles K_n in the sample.

THM 19.3 (Watterson's estimator) *We have*

$$\mathbb{E}[K_n] \sim \theta \log n \qquad\qquad \mathrm{Var}[K_n] \sim \theta \log n.$$

Proof: By Hoppe's urn, K_n—the number of non-black colors in the urn—is a sum of independent Bernoulli variables with mean $\frac{\theta}{\theta+i-1}$, $i = 1, \ldots, n$. Therefore,

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}.$$

and

$$\mathrm{Var}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \frac{i-1}{\theta + i - 1} = \sum_{i=1}^{n} \frac{\theta(i-1)}{(\theta + i - 1)^2}.$$

Using the fact that $\frac{i-1}{\theta+i-1} \to 1$ and the approximation of $\sum_i \frac{\theta}{\theta+i-1}$ by an integral which scales as $\log(n+\theta) - \log\theta$, the result follows. ∎

Note that the variance of $K_n / \log n$ is roughly $\theta / \log n$ and therefore converges to 0 as $n \to \infty$. In fact, an appropriate version of the Central Limit Theorem shows that K_n (standardized) converges in law to a Gaussian.

3 Chinese restaurant process

Although K_n may appear to be a rather naive estimator (not to mention it converges painfully slowly), we will show here that one cannot much better.

We expand Hoppe's urn by tracking the time at which each ball enters the urn. It turns out to be useful to represent the state of the system by the cycle decomposition of a permutation. We start with the empty permutation. At time i:

- If the black ball is chosen (including at the very first step) a new cycle is created including only i.
- If a ball of another color is chosen, say with index j, we include i in the cycle of j to the left of j.

This is called the *Chinese restaurant process*.

THM 19.4 *If the permutation π generated by the chinese restaurant process has k cycles then its probability is $\frac{\theta^k}{\theta_{(n)}}$.*

Proof: Noting that the process can be reversed in a unique way and using an argument similar to the proof of the ESF, the result follows. ■
When $\theta = 1$, we get a uniformly random permutation and Ewens' formula gives a formula for the number of permutations with k cycles.

Recall:

DEF 19.5 (Sufficient statistic) *If a statistic X is such that the conditional dsitribution of the data given X does not depend on the parameter θ, we say that X is* sufficient.

THM 19.6 *K_n is a sufficient statistic for θ.*

Proof: Since the probability in Theorem 19.4 depends only on k, we have that

$$\mathbb{P}[K_n = k] = \frac{\theta^k}{\theta_{(n)}} |S_n^k|,$$

where $|S_n^k|$ is the number of permutations of n elements with k cycles. Using ESF, we get

$$\mathbb{P}[\mathbf{a} \mid K_n = k] = \mathbb{1}\{|\mathbf{a}| = k\} \frac{n!}{|S_n^k|} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{a_j} \frac{1}{a_j!},$$

which does not depend on θ. ■

Further reading

The material in this section was taken from Section 1.3 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.