Zig-zag sampling for genealogies with dormancy

Edoardo Botta
University of Warwick, Department of Statistics

Motivation

Many species of plants show dormancy effects in their evolutionary process. In other words, in such populations seeds can delay the germination and enter a dormant state if the environmental conditions are not favorable, by forming a so-called "seedbank", which can account for up to 80% of the total population. [6] The seedbank coalescent [2] is one of the main model for genealogies that show dormancy effects but the complexity of the sample space poses challenges at inference time. The aim of this work is to explore the application of an efficient MCMC algorithm based on the zig-zag process to an inference problem under the seedbank coalescent model.

Seedbank coalescent

The seedbank coalescent describes the genealogy of an observed sample that has evolved by following a Wright-Fisher with seedbank model. It is a continuous-time Markov process defined on the state space of marked partitions

$$\mathcal{P}_k^{p,s} = \left\{(\zeta, \boldsymbol{u}) : \zeta \in \mathcal{P}_k, \boldsymbol{u} \in \{s,p\}^{|\zeta|}\right\} \tag{1}$$

where $\mathcal{P}_k$ is the space of partitions of $[k]$. This space is constructed by attaching a label (p or s) to every element of every partition in $\mathcal{P}_k$, depending on whether the corresponding ancestral line is currently respectively in the active (plant) or dormant (seed) state.

Dynamics

The model is characterized by two parameters, that arise from the underlying evolutionary process:

- c is defined as the number of seeds produced at each generation
- K is defined as the ratio of active to dormant population size

Given the above, the dynamics of the seedbank coalescent with values in $\mathcal{P}_k^{p,s}$ are described by the following transition rates:

$$\pi \to \pi' \text{ at rate } \begin{cases} 1, & \text{if } \pi' \text{ is obtained from } \pi \text{ by merging two lineages} \\ c, & \text{if } \pi' \text{ is obtained from } \pi \text{ by changing a p to s} \\ cK, & \text{if } \pi' \text{ is obtained from } \pi \text{ by changing a s to p} \end{cases}.$$

MCMC sampling with the zig-zag process

In the most general characterization of the zig-zag process [1], sampling is performed by enhancing the state space with velocities $\boldsymbol{v}$, each one corresponding to a parameter. The process follows deterministic dynamics with constant velocities in between flipping points for time intervals whose lengths follow a Poisson process with inhomogenous rates. Once a flipping event is triggered, the sign of one of the velocities is flipped and the deterministic dynamics resume with the new velocities. This process converges to the target distribution $\pi(\boldsymbol{x}, \boldsymbol{v})$ if the flip rates of the components are set to be

$$\lambda_i(\boldsymbol{x}, \boldsymbol{v}) = (-v_i \partial_i \log(\pi(\boldsymbol{x}, \boldsymbol{v})))^+, \tag{2}$$

where $(x)^+ \coloneqq \max\{x, 0\}$.

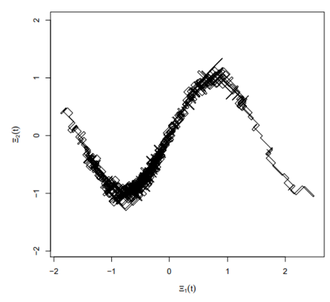

Figure 1. 2D S-shaped density [1]

Geometric embedding

We encode a tree as a tuple $(E_n, \boldsymbol{t})$ [3] and in this section consider a fixed E_n, where E_n is a ranked topology that represents the sequence of events that happened in a tree and $\boldsymbol{t}$ is a vector of the time interval lengths between two consecutive events. In the setting of the seedbank coalescent, the events can be of three types:

- Merger between lineage A and lineage B is encoded as $\{A, B\}$.
- Lineage A moving to the dormant population is encoded as $\{A, \mathcal{S}\}$.
- Lineage A moving to the active population is encoded as $\{A, \mathcal{W}\}$.

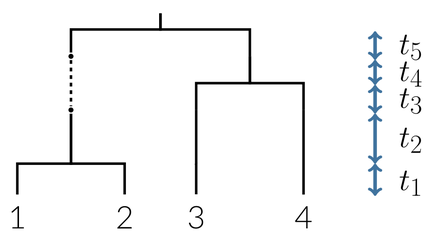

Figure 2. Sample tree with 4 nodes

The realization in figure (2) is encoded as $(E_4, \boldsymbol{t})$, where

- $\boldsymbol{t} = (t_1, \dots, t_5) \in \mathbb{R}_+^5$
- $E_4 = (\{1,2\}, \{\{1,2\}, \mathcal{S}\}, \{3,4\}, \{\{\{1,2\}, \mathcal{S}\}, \mathcal{W}\}, \{\{\{\{1,2\}, \mathcal{S}\}, \mathcal{W}\}, \{3,4\}\})$.

In the space of pairs $(E_n, \boldsymbol{t})$, the algorithm follows zig-zag dynamics until it hits a boundary point ($t_i = 0$ for some i) [5]. Here, one of two moves is triggered:

- Cross boundary by swapping the order of events $i-1$ and i if $E_{n,i-1} \notin E_{n,i}$
- Bounce back otherwise

Adding/Removal of dormancy periods

In this section, we consider fixed holding times and introduce jumps in the current topology by adding or removing a dormancy period. This is done in the form of Metropolis-Hastings proposals following a Poisson clock independent of the rest of the process. The proposal mechanisms is as follows:

- Pick uniformly an active lineage or a dormancy period that is not overlapping with other events
- If the chosen element is an active lineage, propose the introduction of a dormancy period with uniformly sampled start and end (Figure 3b)
- Otherwise, propose the removal of the dormancy period (Figure 3a)

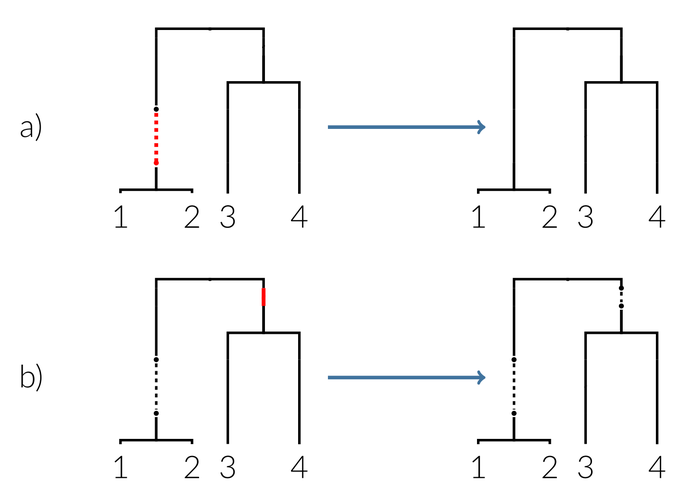

Figure 3. Sample proposals. Left to right respectively current state and proposed state

Seedbank coalescent posterior

In practice, one observes intervals of DNA with sequences of mutations (blue dots) and has to condition and integrate over all possible genealogies that may have generated them. MCMC is currently the only way to do that efficiently.

Figure 4. Sample data

Let

- L be the number of events under the topology E_n,
- S be the total number of dormancy periods in the tree,
- $n_{i,1}$, $n_{i,2}$ be the number of lineages respectively in active and dormant population at the start of interval $t_i \;\; \forall i \in \{1, 2, \dots, L\}$.

The likelihood of the tree $(E_n, \boldsymbol{t})$ can be written as a product of L independent terms

$$p(E_n, \boldsymbol{t} \mid c, K) = \exp\left\{-\sum_{i=1}^{L}\left(\binom{n_{i,1}}{2} + cn_{i,1} + cKn_{i,2}\right) t_i\right\} c^{2S} K^S. \tag{3}$$

Introducing mutations

We assume different mutation rates between active and dormant population. Given the length of an edge l_γ, we model the number of mutations as $m_\gamma \sim \text{Pois}\left(\frac{\theta_j l_\gamma}{2}\right)$, where θ_j is the state-dependent mutation rate.[4] The resulting posterior distribution of the parameters of a tree is

$$\begin{aligned}\pi(E_n, \boldsymbol{t}, c, K, \theta_1, \theta_2 | D_n) \propto & \prod_{\gamma \in F_n : p_\gamma \neq W} p_1(m_\gamma | l_\gamma, \theta_1) \prod_{\gamma \in F_n : p_\gamma = W} p_2(m_\gamma | l_\gamma, \theta_2) \\ & \cdot p(E_n, \boldsymbol{t} | c, K) \pi_0(\theta_1) \pi_0(\theta_2) \pi_0(c) \pi_0(K),\end{aligned} \tag{4}$$

where $p_j(m_\gamma | \cdot)$ are Poisson kernels and $\pi_0(\cdot)$ are priors on the model's parameters.

Posterior sampling

The distribution (4) can be sampled from by simulating a zig-zag process with rates:

$$\lambda_K(E_n, \boldsymbol{t}, c, K) = \left[v_K\left(c\sum_{i=1}^{L} n_{i,2} t_i - \frac{S}{K} - \partial_K \log(\pi_0(K))\right)\right]^+$$

$$\lambda_c(E_n, \boldsymbol{t}, c, K) = \left[v_c\left(\sum_{i=1}^{L}(n_{i,1} + Kn_{i,2}) t_i - \frac{2S}{c} - \partial_c \log(\pi_0(c))\right)\right]^+$$

$$\lambda_{\theta_1}(E_n, \boldsymbol{t}, \theta_1, \theta_2) = \left[v_{\theta_1}\left(\sum_{\gamma \in F_n : p_\gamma \neq \mathcal{W}}\left(\frac{l_\gamma}{2} - \frac{m_\gamma}{\theta_1}\right) - \partial_{\theta_1} \log(\pi_0(\theta_1))\right)\right]^+$$

$$\lambda_{\theta_2}(E_n, \boldsymbol{t}, \theta_1, \theta_2) = \left[v_{\theta_2}\left(\sum_{\gamma \in F_n : p_\gamma = \mathcal{W}}\left(\frac{l_\gamma}{2} - \frac{m_\gamma}{\theta_2}\right) - \partial_{\theta_2} \log(\pi_0(\theta_2))\right)\right]^+$$

$$\begin{aligned}\lambda_i(E_n, \boldsymbol{t}, c, K, \theta_1, \theta_2) = \Bigg[v_i\Bigg(& -\sum_{\gamma \in F_n : t_i \in \gamma}\left(\frac{m_\gamma}{l_\gamma} - \left(\frac{\theta_1}{2}\mathbb{1}(p_\gamma \neq \mathcal{W}) + \frac{\theta_2}{2}\mathbb{1}(p_\gamma = \mathcal{W})\right)\right) \\ & + \binom{n_{i,1}}{2} + cn_{i,1} + cKn_{i,2}\Bigg)\Bigg]^+ \quad \forall i \in \{1, 2, \dots, L\}\end{aligned}$$

Simulating the holding times is one of the major computational challenges, since it involves solving an intractable integral. By establishing constant computational upper bounds for the rates, we can in practice resort to Poisson thinning.

References

[1] Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The zig-zag process and super-efficient sampling for bayesian analysis of big data. *Annals of Statistics*, 47, 07 2016.

[2] Jochen Blath, Adrián González Casanova, Noemi Kurt, and Maite Wilke-Berenguer. A new coalescent for seed-bank models. *The Annals of Applied Probability*, 26(2):857–891, 2016.

[3] Alex Gavryushkin and Alexei J. Drummond. The space of ultrametric phylogenetic trees. *Journal of Theoretical Biology*, 403:197–208, 2016.

[4] R. C. Griffiths and Simon Tavare. Ancestral Inference in Population Genetics. *Statistical Science*, 9(3):307 – 319, 1994.

[5] Jere Koskela. Zig-zag sampling for discrete structures and nonreversible phylogenetic mcmc. *Journal of Computational and Graphical Statistics*, 0(0):1–11, 2022.

[6] Jay Lennon, Frank Hollander, Maite Wilke Berenguer, and Jochen Blath. Principles of seed banks: complexity emerging from dormancy, 11 2020.