Notes

David Helekal

July 14, 2020

Contents

1 Questions

- Multi+ The second term in equation 23 (i.e. what's the bifurcation rate in the forward process) is likely not correct. Clearly this is inversely proportional to the Neg of the lineage going extinct (or being birthed in the forward process), however presumably it should also be proportional to the Neg of the parent lineage? The reasoning being the larger the population size of the parent lineage the likelier it is for a bifurcation event to occur.
- Note: there are mistakes in section 2.4. The combination numbers need to be taken per subtree, and the bifurcation likelihood is just first iteration and needs further consideration.

2 Simulation

2.1 Coalescent Preliminaries

The coalescent is a CTMC defined on the set $\{1...n\}$, parametrised via the coalescent rate, in our case $1/Neg(t)$, where g is a scale parameter and $Neg(t)$ the population size at time t. The transition rates of the coalescent process are given by

$$\rho(j, j-1) = \binom{j}{2} \cdot \frac{1}{Neg(t)}$$

The waiting times in the homogenous case are exponentially distributed

$$P[W_j \leq s] = 1 - \exp\left(-s\frac{\binom{j}{2}}{Neg(t)}\right)$$

Furthermore, the waiting times for individual coalescent events, conditioned on being less than the time between two consecutive sampling events Δt are distributed as follows

$$P[W_j \leq s \mid W_j \leq \Delta t] = \frac{P[W_j \leq s]}{P[W_j \leq \Delta t]} \quad \forall s \leq \Delta t \tag{1}$$

In the inhomogenous case, the waiting times can be derived as follows: For an inhomogenous CTMC, let $E_j(t)$ be the total exit rate from state j at time t. By the markov property individual exit events from a given state only depend on the state and given time, i.e. they form a time-inhomogenous poisson process. As such the probability of no events in an interval $[t, t+s] \quad s \in \mathbb{R}^+$ is

$$\exp\left(-\int_t^{t+s} E_j(\tau)d\tau\right) = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right) \tag{2}$$

The waiting times are defined as

$$W_j(t) = \inf\{s : X(t+s) \neq j \mid X(t) = j\} \tag{3}$$

As such

$$W_j(t) > s \Rightarrow \forall \tau \in [t, t+s] X(\tau) = j \tag{4}$$

Furthermore the above relation doesn't hold iff an exit event has occured in the time interval $[t, t+s]$. As such:

$$P[W_j(t) > s] = P[\text{no exit events in } [t, t+s]] = \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

$$P[W_j(t) < s] = 1 - \exp\left(-\int_0^s E_j(t+\tau)d\tau\right)$$

In the case of phylodynamic coalescent this becomes

$$P[W_j(t) \leq s] = 1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{Neg(t+\tau)}d\tau\right) \tag{5}$$

Note, the waiting times a re still memoryless:

$$P\left[W_j(t) > s + u \mid W_j(t) > s\right] = P\left[W_j(t) > s + u \mid X(s) = j\right] \tag{6}$$

By markov property

$$P\left[W_j(t) > s + u \mid X(s) = j\right] = P\left[W_j(t + s) > u\right] \tag{7}$$

2.2 Homogenous case

The sampling process conditioned on sampling times follows a modified gillespie scheme. In order to facilitate the computation of the likelihoods of the individual simulated trees, it is preferred to avoid rejection sampling. As such we require sampling the conditional likelihood 1. This is achieved by inverse transform sampling. Let:

$$\begin{aligned} u \sim & U([0,1]) \\ T(u) : P[T(u) \leq s] = & \frac{P[T(u) \leq s]}{P[T(u) \leq \Delta t]} \quad \forall s \leq \Delta t \end{aligned} \tag{8}$$

Where $T(u)$ is assumed to be monotone increasing and invertible.

$$\begin{aligned} & P[T(u) \leq s] = P[u \leq T^{-}1(s)] \\ \Rightarrow \quad & P[u \leq T^{-1}(s)] = \frac{\int_0^s \lambda \exp(-\lambda t) dt}{\int_0^{\Delta t} \lambda \exp(-\lambda t) dt} \\ \Rightarrow \quad & T^{-1}(s) = \frac{1 - \exp(-\lambda s)}{1 - \exp(-\lambda \Delta t)} \end{aligned}$$

Defining $y \triangleq T^{-1}(s)$, we obtain the transform:

$$T(y) = \frac{-1}{\lambda} \log[1 - y(1 - \exp(-\lambda \Delta t))] \tag{9}$$

The corresponding pdf evaluated at u is

$$f_{\mathbf{T}(u)}(T(u)) = \lambda \left(\frac{1}{1 - \exp(\lambda \Delta t)} - u \right) \tag{10}$$

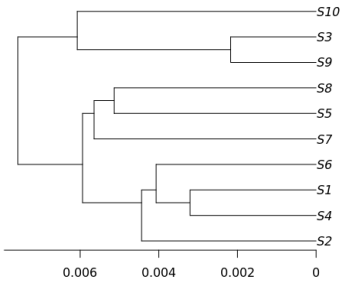


Figure 1: An example simulated coalescent tree

```
f <- (sampling_times, Ne): //Sampling times in descending order
  extant_lineages <- 1
  future_lineages <- length(sampling_times)-1
  t <- sampling_times[1]
  idx <- 1
  log_lh <- 0

  while extant_lineages > 1 or future_lineages > 0:
    if extant_lineages < 2:
      idx++
      t <- sampling_times[idx]
      extant_lineages++
      future_lineages--
    else:
      delta_t <- t-sampling_times[idx+1]
      rate <- binom(extant_lineages,2)/Ne

      p_coal <- 1-exp(-rate*delta_t)
      r_c ~ U[0,1]

      if r_c < p_coal:

        log_lh += log(p_coal)

        coalesce_lineages
        extant_lineages--

        r_w ~ U[0,1]
        w_t <- (-1/rate)*log(1-r_w*(1-exp(-rate*delta_t)))
        t <- t+w_t
```

```
        cond_lh <- rate*(1/(1-exp(-rate*delta_t)) - r_w)
        log_lh += log(cond_lh)

      else:
        log_lh += log(1-p_coal)
        idx++
        t <- sampling_times[idx]
        extant_lineages++
        future_lineages--

  return: coalescent_times, log_lh
```

2.3 Inhomogenous Case

In the inhomogenous case, the scheme is similar, with the key difference that the sampling times now follow a much more complex distribution. We proceed with a modified conditional sampling scheme as in 8. To obtain draws $w_j(t)$, draws from standard exponential w_j are rescaled, akin to algorithm described in [5] Pg 98.

2.3.1 Waiting times distribution

Consider the time interval $[t_i, s_i]$ with $s_j = min\{s \in S : s > t_i\}$. Define $\Delta t_i \triangleq s_i - t_i$. The probability that no coalescent events happens within this interval is

$$P[W_j(t_i) > \Delta t_i] = \exp\left(-\int_0^{\Delta t_i} \frac{\binom{j}{2}}{Neg(t+\tau)} d\tau\right)$$

analogously, the probability of waiting times conditioned on that the waiting time is less than Δt_i is:

$$P[W_j(t_i) < s \mid W_j(t_i) < \Delta t_i] = \frac{1 - \exp\left(-\int_0^s \frac{\binom{j}{2}}{Neg(t+\tau)} d\tau\right)}{1 - P[W_j(t_i) > \Delta t_i]} \tag{11}$$

2.3.2 Sampling

In order to sample $W_j(t_i)$ we proceed with an inverse transform sampling scheme, derived from the base samples W_j. First, assume W_j are distributed according to

$$P[W_j < s \mid W_j < u] = \frac{1 - \exp\left(-\frac{\binom{j}{2}}{Neg}\right)}{1 - P[W_j > u]} \tag{12}$$

Where u is chosen such that

$$P[W_j > u] = P[W_j(t_i) > \Delta t_i] \tag{13}$$

Then the function

$$F(W_j; t_i): \quad P[F(W_j; t_i) < s \mid W_j < u] = P[W_j(t_i) < s \mid W_j(t_i) < \Delta t_i] \tag{14}$$

Is given by the inverse with respect to s of

$$G(s; t_i) = \int_0^s \frac{Neg}{Neg(t_i + \tau)} d\tau \tag{15}$$

Which exists for any biologically sensible choice of $Neg(t)$

2.3.3 Likelihood

Let $\{t_i\}_{i \in S \subset \mathbb{N}}$ denote the times of events in increasing order. Using notational convention from [1], let $t_Y \triangleq \{t_i\}_{i \in Y}$ denote times of coalescent events and $t_I \triangleq \{t_i\}_{i \in I}$ denote times of sampling events, where Y, I are disjoint partitions of the index set S with the property that $S = Y \cup I$. The likelihood of a particular genealogy is then given by:

$$\mathcal{L}(g \mid Neg) = \prod_{i \in S \setminus 1} \left(\mathbb{1}_Y(i) \frac{\binom{k_i}{2}}{Neg(t_i)} + \mathbb{1}_I(i) \right) \exp\left(- \int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{Neg(\tau)} d\tau \right) \tag{16}$$

The log-likelihood is:

$$\log \mathcal{L}(g \mid Neg) = - \sum_{i \in S \setminus 1} \int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{Neg(\tau) d\tau} + \sum_{i \in Y} \log \frac{\binom{k_i}{2}}{Neg(t_i)} \tag{17}$$

2.3.4 *Skygrid* and other families of population functions

Skygrid is an approach introduced in [3]. It considers a family of time-dependent effective population size functions specified as follows. Let $T \subset \mathbb{R}$ be the interval under consideration in coalescent time. Consider an arbitrary population size function F:

$$F: \quad T \subset \mathbb{R} \to \mathbb{R}^+ \tag{18}$$

Given a mutually disjoint family of time intervals $\{D_i\}_{i \in N} \subseteq T$, $D_j = [d_{j-1}, d_j)$, with $\bigcup_{i \in N} D_i = T$, the *skygrid* family of functions is then given by

$$F_{skygrid} \triangleq \left\{ Neg \in F \mid \forall i \in N, \quad \forall t \in D_i, \quad Neg(t) = c_i, \quad c_i \in \mathbb{R}^+ \right\} \tag{19}$$

This can be easily extended to any function G

$$G(t) \triangleq \sum_{i \in N} \mathbb{1}_{D_i}(t) g_i(t) \tag{20}$$

Where $g_i(t)$ are arbitrary positive integrable functions. Such formulation makes computation of likelihood straightforward.

$$\begin{aligned}\log \mathcal{L}\left(g \mid G(t)\right) = &- \sum_{i \in S \setminus 1} \binom{k_i}{2} \sum_{j \in N} \mathbb{1}_{D_j}(t_{i-1}) \int_{t_{i-1}}^{\min\{d_j, t_i\}} g_j^{-1}(\tau) d\tau \\ &+ \sum_{i \in Y} \log \binom{k_i}{2} - \sum_{i \in Y} \log \sum_{j \in N} \mathbb{1}_{D_j}(t_i) g_j(t_i)\end{aligned} \tag{21}$$

2.4 Multistrain+Inhomogenous Case

In this case, coalescent nodes have an added colour property, and each colour coalesces according to a colour specific, time dependent case. Nodes of non-identical colour can coalesce iff at least one of them is the last remaining node of a given colour.

2.4.1 Model

Given M colours, M population size functions $\{Neg_j(t)\}_{1 \le j \le M}$, and initial population size N, Let $Y(t)$ be a CTMC with the state space:

$$\Sigma = \left\{\mathbf{s} \in \mathbb{Z}_+^N : |\mathbf{s}| \le N, |\mathbf{s}| \ge 1\right\} \tag{22}$$

. and the transition rates

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} \quad \binom{s_j}{2} Neg_j^{-1}(t) + \delta_{s_j,1} Neg_j(t)^{-1} \sum_{i \ne j} s_i Neg_i(t) \tag{23}$$

2.4.2 Likelihood

To derive the likelihood of a genealogy given a colouring with non-zero likelihood and growth functions $\mathbf{Neg} := \{Neg_j(t)\}_{1 \le j \le M}$, first assume event notation as used in 2.3.3.
Define the set of bifurcation events:

$$\Omega \triangleq \{\omega_j\}_{1 \le j \le K} \subset S \setminus 1 \tag{24}$$

With the property that:

$$\forall i < j \quad \omega_i < \omega_j$$

Next define the subtrees of omega

$$\begin{aligned}\mathbf{W}' &\triangleq \{W_j'\}_{1 \le j \le K} \\ W_j' &\triangleq \{i \in S \mid t_i \text{is a descendant of } \omega_j\}\end{aligned} \tag{25}$$

Finally, we define the subtrees corresponding to individual lineages, associated with a particular Neg_j, denoted by W:

$$\begin{aligned} \mathbf{W} &\triangleq \{W_j\}_{1\leq j\leq K} \\ W_j &\triangleq W_j \setminus \bigcup_{i<j}(W_i \cup \omega_i) \end{aligned} \tag{26}$$

The total likelihood is equal to:

$$\mathcal{L}(\mathbf{g} \mid \Omega, \mathbf{Neg}) = \prod_{j=1}^{K} \mathcal{L}(\mathbf{g} \cap W_j \mid Neg_j) \prod_{j=1}^{K} \mathcal{L}(\mathbf{g} \cap \omega_j \mid \mathbf{Neg}) \tag{27}$$

This can be understood as the product of the likelihoods of individual subtrees and the product of the likelihoods of the individual bifurcation events.
The first term in equation 27 can then be expanded as:

$$\prod_{j=1}^{K} \mathcal{L}(\mathbf{g} \cap W_j \mid Neg_j) = \prod_{j=1}^{K} \prod_{i \in S \cap W_j} \left(\mathbb{1}_Y(i) \frac{\binom{k_i}{2}}{Neg(t_i)} + \mathbb{1}_I(i) \right) \exp\left(- \int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{Neg(\tau)} d\tau \right) \tag{28}$$

The second term:

$$\prod_{j=1}^{K} \mathcal{L}(\mathbf{g} \cap \omega_j \mid \mathbf{Neg}) = \prod_{i \in \Omega} \left(\frac{k_i}{Neg(t_i)} \right) \exp\left(- \int_{t_{i-1}}^{t_i} \frac{k_i}{Neg(\tau)} d\tau \right) \tag{29}$$

2.4.3 Inference

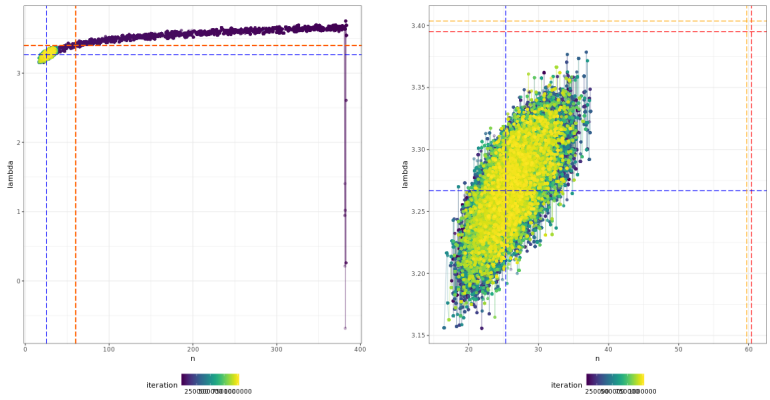
Due to the variable nature of the dimensionality of the parameter space, it will be necessary to use Reversible Jump MCMC (rjMCMC) [2, 4]. Additionally, one will have to consider a suitable prior for bifurcation points. This will likely have to be computed as a function of leaf colouring, possiblt integrating covariate information (geographic location?).

3 Simulation

3.1 Exponetial Growth

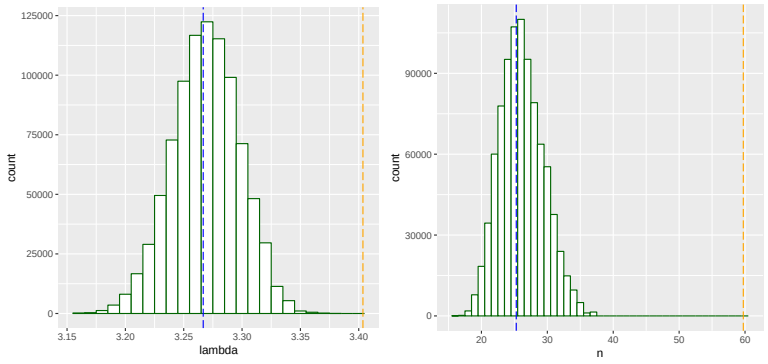
An example of a population under exponential growth, $Neg(t) = N * \exp(-\lambda t)$ consisting of 100 sampling events between 0 and 10 years before present has been simulated with a rate parameter λ and final size parameter N drawn from a uniform densities on intervals $[0.1, 10]$, and $[1, 100]$ respectively.
A Metropolis-Hastings MCMC scheme was then used to infer the parameters λ and N. A zero-centred laplace distribution with rate equal to one was chosen as the prior for λ, whereas for N the exponential distribution with rate one was used.

One million iterations were used and the first One hundred thousand discarded as burn in time. To further validate the fit, both a maximum likelihood (MLE) and maximum *a-posteriori* (MAP) estimates were computed and plotted against the posterior marginals inferred by the MCMC.



(a) (b)

Figure 2: Trace plots for the markov chain. Red lines denote true parameter values. MLE marked by blue lines. MAP marked by orange lines. 2a Shows the entire trace of the chain. 2b Shows the trace with the first 100000 iterations discarded



(a) (b)

Figure 3: Histograms of the posterior marginals. MLE marked by blue line. MAP marked by orange line. 3a λ marginal 3b N marginal

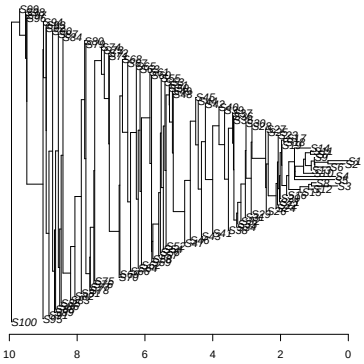


Figure 4: The simulated genealogy used for this example.

4 Previous Work

A framework utilising sampling intensity in order to extract more information is proposed in [6]

5 Bibliography

References

[1] Alexei J. Drummond et al. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data". In: *Genetics* 161.3 (July 1, 2002). Publisher: Genetics Section: INVESTIGATIONS, pp. 1307–1320. ISSN: 0016-6731, 1943-2631. URL: https://www.genetics.org/content/161/3/1307 (visited on 07/02/2020).

[2] Y. Fan and S. A. Sisson. "Reversible jump Markov chain Monte Carlo". In: *arXiv:1001.2055 [stat]* (Jan. 12, 2010). arXiv: 1001.2055. URL: http://arxiv.org/abs/1001.2055 (visited on 07/13/2020).

[3] Mandev S. Gill et al. "Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci". In: *Molecular Biology and Evolution* 30.3 (Mar. 1, 2013). Publisher: Oxford Academic, pp. 713–724. ISSN: 0737-4038. DOI: 10.1093/molbev/mss265. URL: https://academic.oup.com/mbe/article/30/3/713/1041171 (visited on 06/30/2020).

[4] Peter J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4 (Dec. 1, 1995). Publisher: Oxford Academic, pp. 711–732. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.711. URL: https://academic.oup.com/biomet/article/82/4/711/252058 (visited on 07/14/2020).

[5] Jotun. Hein, Mikkel H. Schierup, and Carsten. Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory:* Oup Oxford, Dec. 9, 2004. ISBN: 978-0-19-154615-0. URL: https://www.dawsonera.com:443/abstract/9780191546150.

[6] Kris V. Parag, Louis du Plessis, and Oliver G. Pybus. "Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences". In: *Molecular Biology and Evolution* (). DOI: 10.1093/molbev/msaa016. URL: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa016/5719057 (visited on 06/19/2020).