Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Inference for Λ-coalescents

Matthias Birkner[1,*], Jochen Blath[2,†]

submitted: February 27, 2007

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstraße 39
10117 Berlin
Germany

E-mail: birkner@wias-berlin.de

[2] Institut für Mathematik, Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
Germany

E-mail: blath@math.tu-berlin.de

No. 1211
Berlin 2007

2000 *Mathematics Subject Classification.* Primary: 92D15; Secondary: 60G09, 60G52, 60J75, 60J85.

Key words and phrases. Lambda-coalescent, inference, infinitely-many-sites model, mathematical population genetics, Fleming-Viot process, multiple collisions, frequency spectrum, Monte-Carlo simulation.

*Partially supported by EPSRC GR/R98563/01.

†Partially supported by EPSRC GR/R985603.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

One of the main problems in mathematical genetics is the inference of evolutionary parameters of a population (such as the mutation rate) based on the observed genetic types in a finite DNA sample. If the population model under consideration is in the domain of attraction of a classical Fleming-Viot process, then the standard means to describe the corresponding genealogy is Kingman's coalescent. For this process, powerful inference methods are well-established. An important feature of this class of models is, roughly speaking, that the number of offspring of each individual is small when compared to the total population size.

Recently, more general population models have been studied, in particular in the domain of attraction of so-called *generalised Lambda Fleming-Viot processes*, as well as their (dual) genealogies, given by the so-called *Lambda-coalescents*. Moreover, Eldon & Wakeley (2006) have provided evidence that such more general coalescents, which allow multiple collisions, might actually be more adequate to describe real populations with extreme reproductive behaviour, in particular many marine species.

In this paper, we extend methods of Ethier & Griffiths (1987) and Griffiths & Tavaré (1994) to obtain a likelihood based inference method for general Lambda-coalescents. In particular, we obtain a method to compute (approximate) likelihood surfaces for the observed type probabilities of a given sample. We argue that within the (vast) family of Lambda-coalescents, the parametrisable sub-family of Beta$(2-\alpha,\alpha)$-coalescents, where $\alpha \in (1,2]$, are of particular biological relevance. We apply our method in this case to simulated and real data (taken from Árnason (2004)).

We conclude that for populations with extreme reproductive behaviour, the Kingman-coalescent as standard model might have to be replaced by more general coalescents, in particular by Beta$(2-\alpha,\alpha)$-coalescents.

1 Introduction

1.1 Coalescent processes

For neutral population models of fixed population size in the domain of attraction of the classical Fleming-Viot process, such as the Wright Fisher model and the Moran model, the genealogy of a finite sample can be described by the now classical Kingman-coalescent, which we introduce briefly, followed by the more recently discovered and much more general Lambda-coalescents. For background on Fleming-Viot processes, see e.g. [EK86], [D93] and [DK99].

Kingman's coalescent. Let $\mathcal{P}_n$ be the set of partitions of $\{1,\dots,n\}$ and let $\mathcal{P}$ denote the set of partitions of $\mathbb{N}$. For each $n \in \mathbb{N}$, Kingman [K82] introduced the so-called *n-coalescent*, which is a $\mathcal{P}_n$-valued continuous time Markov process $\{\Pi_n(t), t \geq 0\}$, such that $\Pi_n(0)$ is the partition of $\{1,\dots,n\}$ into singleton block, and then each pair of blocks merges at rate one. Given there are b blocks at present, this means that the overall rate to see a merger between blocks is $\binom{b}{2}$. Note that only *binary mergers* are allowed. Kingman [K82] also showed that there exists a $\mathcal{P}$-valued Markov process $\{\Pi(t), t \geq 0\}$, which is now called the (standard) *Kingman-coalescent*, and whose restriction,

for each $n \in \mathbb{N}$, to the first n positive integers is the n-coalescent. To see this, note that the restriction of any n-coalescent to $\{1, \ldots, m\}$, where $1 \leq m \leq n$, is an m-coalescent. Hence the process can be constructed by an application of the standard extension theorem.

Lambda-coalescents. Pitman [P99] and Sagitov [S99] introduced and discussed coalescents which allow *multiple mergers*, i.e. more than just two blocks may merge at a time. Again, a coalescent with multiple mergers (which will be later called *Lambda-coalescent*) is a $\mathcal{P}$-valued Markov-process $\{\Pi(t), t \geq 0\}$, such that for each n, its restriction to the first n positive integers is a $\mathcal{P}_n$-valued Markov process (the "n-Lambda-coalescent") with the following transition rates. Whenever there are b blocks in the partition at present, each k-tuple of blocks (where $2 \leq k \leq b \leq n$) is merging to form a single block at rate $\lambda_{b,k}$, and no other transitions are possible. The rates $\lambda_{b,k}$ do neither depend on n nor on the structure of the blocks. Pitman showed that in order to be consistent, which means that for all $b, k \geq 2, b \geq k$,

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1},$$

such transition rates must necessarily satisfy

$$\lambda_{b,k} = \int_0^1 x^k (1-x)^{b-k} \frac{1}{x^2} \Lambda(dx), \tag{1}$$

for some finite measure Λ on the unit interval. Note that (1) sets up a one-to-one correspondence between coalescents with multiple collisions and finite measures Λ. Indeed, it is easy to see that the $\lambda_{b,k}$ determine Λ since they satisfy the conditions of Hausdorff's moment problem, which has a unique solution.

Due to the restriction property, the Lambda-coalescent on $\mathcal{P}$, with rates obtained from the measure Λ as described above, can be constructed from the corresponding n-Lambda-coalescents via extension. Sometimes, we use the shorthand notation Λ-coalescent.

Note that the family of Lambda-coalescents is rather large, and in particular cannot be parametrised by a few real variables. Important examples include $\Lambda = \delta_0$ (Kingman's coalescent) and $\Lambda = \delta_1$ (leading to star-shaped genealogies, i.e. one huge merger into one single block). Later, we will be concerned with two important parametric subclasses of Λ-coalescents, namely the so-called *Beta-coalescents*, where Λ has a Beta$(2-\alpha, \alpha)$-density for some $\alpha \in (1, 2]$, and simple linear combinations of atomic measures of the type $\Lambda = c_1\delta_0 + c_2\delta_y$ for some constants $c_1, c_2 > 0$ and $y \in (0, 1]$. To avoid trivialities, we will henceforth assume that $\Lambda \neq 0$.

Remark. An important difference between the classical Kingman-coalescent and coalescents which allow for multiple mergers should be pointed out here. Roughly speaking, a Kingman coalescent arises as limiting genealogy from a so-called Cannings population model ([C74], [C75]), if the individuals produce a number of offspring that is negligible when compared to the total population size (in particular, if the variance of the reproduction mechanism converges to a finite limit). More general coalescents with multiple mergers arise, once the offspring distribution is such that a non-negligible proportion, say $x \in (0, 1]$, of the population alive in the next generation goes back to a single reproduction event from a single ancestor. In this case, $x^{-2}\Lambda(dx)$ can be interpreted as the intensity at which we see such proportions x. Precise conditions on the underlying Cannings-models and the classification of the corresponding limiting genealogies can be found in [MS01]. □

Remark. In [EW06], Eldon and Wakeley assume that there are extreme reproductive events, which account for non-negligible proportions of the population in a single reproduction event, in the population dynamics of the Pacific Oyster. In fact, many marine species seem to exhibit such behaviour (see also [A04] and [BBB94]). This will be discussed in more detail in Subsection 7.2. □

"Coming down from infinity". Not all Lambda-coalescents seem to be reasonable models for biological populations, since some do not allow for a finite "time to the most recent common ancestor" (T_{MRCA}). This is equivalent with "coming down from infinity in finite time": it means that, given any initial partition in $\mathcal{P}$, and for all $\varepsilon > 0$, the partition $\Pi(\varepsilon)$ a.s. consists of finitely many blocks only. Schweinsberg [S03] proves that if either Λ has an atom at 0 or Λ has no atom at zero and

$$\sum_{b=2}^{\infty} \left[\sum_{k=2}^{b} (k-1) \binom{b}{k} \int_{[0,1]} x^{k-2}(1-x)^{b-k} \Lambda(dx) \right]^{-1} =: \lambda^* < \infty, \tag{2}$$

then the corresponding coalescent does come down from infinity (and if so, the time to come down to only one block has finite expectation).

An important example for a coalescent, which (only just) does not come down from infinity is the Bolthausen-Sznitman coalescent, where $\Lambda(dx) = dx$ on $[0,1]$.

Remark. It should be observed that all n-Lambda-coalescents (for finite n) do have an a.s. finite T_{MRCA}. □

Examples for coalescents which satisfy (2) are

$$\Lambda = c_1\delta_0 + c_2\delta_y, \quad c_1, c_2 \geq 0, y \in (0,1), c_1 + c_2 > 0, \tag{3}$$

(hence including Kingman's coalescent for $c_1 = 1, c_2 = 0$) and the so-called Beta$(2-\alpha, \alpha)$-coalescents with $\alpha \in (1,2]$, where

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}\, dx. \tag{4}$$

Note that the Bolthausen-Sznitman coalescent is the Beta-coalescent with $\alpha = 1$.

Remark. It is easy to interpret the behaviour of the population corresponding to the coalescent associated with (3). The first atom stands for a Kingman-component, i.e. essentially reproduction with finite variance. The second atom says that with rate c_2, a single individual can produce $100 \times y\%$ of the population currently alive in a single reproduction event. □

Populations with extreme reproductive behaviour. Recently, biologists studied the genetic variation of certain marine species with rather extreme reproductive behaviour, see, e.g. Arnason [A04] (Atlantic Cod) and [BBB94] (Pacific Oyster). Eldon and Wakely ([EW06]) investigated such populations and proposed more general coalescents than Kingman's coalescent as models for the genealogy of such populations. However, their model remains rather limited (based on the coalescents described by (3)), and their inference relies on *summary statistics*, in particular *segregating sites* and *singleton polymorphisms*. One critique is that there is no reason why there should be precisely one atom to the right of 0. Still, they conclude that there is evidence that more general coalescents need to be considered. They write:

"It may be that Kingman's coalescent applies only to a small fraction of species."

In this paper, we propose a new candidate as a null-model for the genealogy oft populations with extreme reproductive behaviour and provide some statistical evidence obtained both from simulated and real data, namely the Beta$(2-\alpha,\alpha)$-coalescent for some $\alpha \in (1,2]$, which then needs to be estimated from the data. A large part of what follows will be concerned with the question of how to estimate such an α.

1.2 Samples under the ∞-many sites model

Here, we intend to give the reader a hint about what the "data" in our infinitely-many-sites, or shorthand, infinite-sites, model look like. We will present a rigorous probabilistic basis on how such data may arise in the subsequent Section 2. Note that we will not work with original DNA sequence data here (i.e. sequences of the bases `A`, `T`, `C`, `G`), but with an extract of them, which contains much of the relevant information. The way of how to transform real sequence data under the infinite-sites model into data of the type presented below is, e.g., being discussed in [T01].

Let $n \in \mathbb{N}$ be the size of the sample, i.e. the number of sequences resp. alleles drawn from a large population. Let $i \in \{1,\dots,n\}$. Following [EG87] (or the overview article [T01]), we consider the i-th allele in an n-sample under the infinitely-many-sites model as a finite sequence of positive integers $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots)$, where each $x_{ij} \in \mathbb{N}_0$. It is common to think of $x_{i0}, x_{i1}, \dots$ as the most recently mutated site, the second most recently mutated site, etc., although the complete temporal order can in general not be reconstructed from the original sequence observations (however, this information will later be factored out by considering suitable equivalence classes, see below). An n-sample therefore consists of the sequences $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$.

We assume that the sequences follow these rules:

1) Coordinates within each sequence are distinct.

2) If for some $i, i' \in \{1,\dots,n\}$ and $j, j' \in \mathbb{N}_0$ we have $x_{ij} = x_{i'j'}$, then

$$x_{i,j+k} = x_{i',j'+k}, \quad k = 1, 2, \dots$$

3) There is at least one coordinate common to all n sequences.

Note that 1) – 3) imply that the observed types form a *rooted tree*.

Example: (taken from [T01]) A dataset which is consistent with the above rules.

allele 1 : (9,7,3,1,0)
allele 2 : (3,1,0)
allele 3 : (11,6,4,1,0)
allele 4 : (8,6,4,1,0)
allele 5 : (10,5,2,0)

allele 6 : (8,6,4,1,0)
allele 7 : (8,6,4,1,0) □

Note that the allelic type $(8,6,4,1,0)$ appears three times, i.e. has multiplicity 3. For notational convenience, our sequences all end in 0, this reflects the existence of a common "root". This is a little more information than originally contained in the infinite-sites data, in particular we assume a "known ancestral type". The labels of the mutations and the root are by no means required to be decreasing, this is just suitable convention. We will define an appropriate equivalence relation via bijections on $\mathbb{N}$ later.

Given a sample of size n, we will now write $(t, \mathbf{n})$ for the set consisting of the set of *different* types $t = (x_1, \ldots, x_d), d \leq n$, and the multiplicity vector $\mathbf{n}$. In the above example, we have

$$(t, \mathbf{n}) = \Big(\big((9,7,3,1,0),(3,1,0),(11,6,4,1,0),(8,6,4,1,0),(10,5,2,0)\big),(1,1,1,3,1)\Big).$$

If we take numbered samples into account, i.e. if we let $a_i \subset \{1, \ldots, n\}, i \in \{1, \ldots, d\}$ denote the set of the numbers of the sequences with type x_i, then one can also consider the set of types and ordered partitions $(t, \mathbf{a})$, where $\mathbf{a} = (a_1, \ldots a_d)$, in the above example given by

$$(t, \mathbf{a}) = \Big(\big((9,7,3,1,0),(3,1,0),(11,6,4,1,0),(8,6,4,1,0),(10,5,2,0)\big),\big(\{1\},\{2\},\{3\},\{4,6,7\},\{5\}\big)\Big).$$

The probabilistic mechanism behind these data and the necessary equivalence relation will be discussed in detail in Section 2.

1.3 Recursion for the type probabilities under Kingman's coalescent

It is in principle possible to compute the exact probabilities of a given type configuration $(t, \mathbf{n})$ via a recursive formula. The following recursion is due to Ethier and Griffiths, see [EG87] and [G89]. Let $p^0(t, \mathbf{n})$ be the probability of the ordered types t with multiplicities $\mathbf{n}$. Then, using standard coalescent arguments, considering the last event in the coalescent history, it is easy to arrive at the recursion

$$\begin{aligned} p^0(t,\mathbf{n}) = {} & \frac{1}{nr+\binom{n}{2}} \sum_{k:\, n_k \geq 2} \binom{n}{2} \frac{n_k - 1}{n-1} p^0(t, \mathbf{n} - \mathbf{e}_k) \\ & + \frac{r}{nr+\binom{n}{2}} \sum_{\substack{k: n_k = 1, x_{k0} \text{distinct} \\ s(\mathbf{x}_k) \neq \mathbf{x}_j \forall j}} p^0(s_k(t), \mathbf{n}) \\ & + \frac{r}{nr+\binom{n}{2}} \sum_{\substack{k: n_k = 1, \\ x_{k0} \text{ distinct}}} \sum_{j: s(\mathbf{x}_k) = \mathbf{x}_j} (n_j + 1) p^0(r_k(t), r_k(\mathbf{n} + \mathbf{e}_j)), \end{aligned} \tag{5}$$

where $\mathbf{e}_j$ denotes the j-th unit vector, $s_k(t)$ deletes first coordinate of the k-th sequence in t, $s(\mathbf{x}_k)$ removes the first coordinate from the sequence $\mathbf{x}_k$, $r_k(t)$ removes k-th sequence from t, and x_{k0} 'distinct' means that $x_{k0} \neq x_{ij}, \forall (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ and $(i,j) \neq (k,0)$. We have the boundary condition $p(\{(0)\},(1)) = 1$ for the root.

1.4 Inference for Kingman's coalescent

Efficient likelihood-based inference methods, some solving the above recursion (5) approximately via Monte Carlo methods, others using MCMC and importance sampling, have been developed since the beginning of the 90ies, see [EG87], [GT94a], [GT94b], [GT94c], [GT96b], [FKY99], [DIG04a], [SD00]). In [SD00], Stephens and Donnelly provide proposal distributions for importance sampling, which are optimal in some sense, and compare them to various other methods. Their importance sampling scheme seems, at present, to be the most efficient tool for inference for relatively large datasets.

Leaving our rather narrow, in particular neutral, non-spatial and constant-population-size framework, we would like to mention that there are also Kingman-coalescent based inference methods for models with non-constant population size (see, e.g. [KYF95], [KYF98]), spatial structure and migration (e.g. [DIG04b]) or recombination (e.g. [GM96]), which lead to interesting questions, such as how to distinguish between shallow genealogies, which might be the result of a recent increase in population size, or extreme offspring distributions, which then should be modelled using Lambda-coalescents.

1.5 Outline of the paper

In Section 2, we present in detail the probabilistic neutral coalescent model that gives rise to the data as presented in Subsection 1.2.

Section 3 contains the recursions for the type probabilities assuming a given underlying Λ-coalescent-tree. Moreover, introducing the block-counting process $\{Y_t, t \geq 0\}$ associated with a Λ-coalescent, and its time-reversal $\{\tilde{Y}_t, t \geq 0\}$, we derive a recursion for the distribution of the site-frequency spectrum. Note that our inference methods will be focused on the infinite-sites case.

In Section 4, we use alternative approaches to derive recursions also in the finite- and infinite-alleles cases. Indeed, we use Donnelly and Kurtz' [DK99] modified lookdown construction, assuming a given underlying generalised Λ-Fleming-Viot process, to obtain a recursion for the finite-alleles model. Moreover, we show that calculations based on the generator of the population model as in [DIG04a] also lead to these recursions in the finite-alleles model. Finally, we recall the recursion obtained by Möhle in [M06b] for the multiplicity vector $\mathbf{n}$ in the infinite-alleles model.

In Section 5, we derive proposal transitions for a Markov chain that we then use to obtain a Monte Carlo scheme for the type probabilities resp. likelihoods obtained in Section 3 under the Lambda-coalescent in the infinite-sites model.

In Section 6, we present some likelihood-surfaces, obtained from our Monte Carlo method when applied to simulated and real data. We claim that there is evidence that populations with extreme reproductive behaviour could be better modelled with Beta$(2-\alpha, \alpha)$-coalescents instead of Kingman's coalescent.

Section 7 contains a discussion of the biological and theoretical relevance of the Beta$(2-\alpha, \alpha)$ coalescent subfamily within the family of Λ-coalescents. We argue that they could be used as null-model in certain situations. We also discuss alternative approaches to inference questions as

derived in [EW06] and [BBS06].

Finally, in Section 8 (the Appendix), we present two algorithms to obtain samples of finite- and infinite alleles and infinite-sites data. Furthermore, we include all the original data, corresponding genetrees, likelihood-surfaces and standard deviations that lead to the statistical evidence in Section 6.

Acknowledgements

We wish to thank Bob Griffiths for many helpful discussions, Alison Etheridge for providing a stimulating environment and Matthias Steinrücken for various commments and help with the simulations.

2 Infinite sites data and Λ-coalescent trees

To obtain an n-sample under the infinite-sites model from a coalescent tree, we perform the following probabilistic experiment:

(i) Run an n-Λ-coalescent. Obtain a *rooted* coalescent tree.

(ii) On this rooted tree with n leaves (*numbered* from 1 to n), place mutations along the branches at rate r (note that this parameter is customarily called $\theta/2$).

(iii) *Label* these mutations randomly: Given there are s mutations in total, attach randomly (i.e. according to the uniform distribution) the labels from $1, \ldots, s$ to these mutations.

(iv) Turn this coalescent tree with *labelled* mutations and *numbered* leaves into a "genetree" by breaking edges at mutations, resulting in vertices of degree 2, and then moving the branching points inwards until they reach the nearest mutation. Ignore the lengths of the edges.

(v) A *type* is the sequence of labels of mutations observed following the path backwards from a leaf to the root. *Enumerate* the *different* types randomly to obtain a set of sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$, where $d \leq n$ is the number of different types.

(vi) Define an equivalence relation on the set of types by writing

$$(\mathbf{x}_1, \ldots, \mathbf{x}_d) \sim (\mathbf{y}_1, \ldots, \mathbf{y}_d)$$

if there is a bijection $\xi : \mathbb{N}_0 \to \mathbb{N}_0$ with $y_{ij} = \xi(x_{ij}), i \in 1, \ldots, d;\ j = 0, 1, \ldots$. Under "$\sim$", the concrete labels of mutations are irrelevant. Note that in what follows, we suppress the distinction between such an equivalence class, denoted by $[t]$, and a representative, denoted by t.

(vii) Let $A_i \subset \{1, \ldots, n\}$ be the random set of the numbers (being attached to leaves in Step 2) which have type $i \in \{1, \ldots, d\}$. We obtain a random pair $(T, \mathbf{A})$, where $\mathbf{A} = (A_1, \ldots, A_d)$ is an *ordered* random partition.

(viii) Finally, let

$$p(t, \mathbf{a}) := \Pr\{(T, \mathbf{A}) = (t, \mathbf{a})\}.$$

Note that, by the symmetry of the coalescent,

$$p\big(t,(a_1,\dots,a_d)\big) = p\big(t,(\pi(a_1),\dots,\pi(a_d))\big)$$

for any permutation $\pi \in S_n$.

We call such pairs $(T, \mathbf{A})$ a *numbered random sample configuration with ordered types*. Later, it will be useful to consider only the *frequencies* of the ordered types, i.e. define a map

$$\phi : (t, \mathbf{a}) \mapsto (t, \mathbf{n}),$$

where $\mathbf{n} = (n_1, \dots, n_d) := (\#a_1, \dots, \#a_d)$, i.e. $\sum_{i=1}^{d} n_i = n$. We denote its probability distribution by

$$\begin{aligned} p^0\big((t,\mathbf{n})\big) &:= p\big(\phi^{-1}(t,\mathbf{n})\big) \\ &= \frac{n!}{n_1!\cdots n_d!} p\big((t,\mathbf{a})\big) \end{aligned} \tag{6}$$

for any $(t, \mathbf{a}) \in \phi^{-1}(t, \mathbf{n})$ by the observation in Step 8.

For notational simplicity, we introduce the following slightly ambiguous operation. By $\mathbf{a} - \mathbf{e}_i$, we mean a partition obtained from $\mathbf{a}$ by removing one element from the *set* a_i (with implicit adjustments so that the result is a partition of $\{1, \dots, n-1\}$). Note that we will not be concerned with the fact which element we actually remove, since, by Step (viii) in the above mechanism, the type probability p will not depend on the actual choice. Similarly, by $\mathbf{a} - (k-1)\mathbf{e}_i$ we mean the partition obtained from $\mathbf{a}$ by removing k elements from a_i (certainly, this only makes sense if $\#a_i \geq k$). Finally, $\mathbf{a} + \mathbf{e}_i$ will be the partition obtained from $\mathbf{a}$ by adding an arbitrary element of $\mathbb{N}$ to the set a_i that is not yet contained in any other set $a_l, l = 1 \dots d$.

3 Genealogical tree probabilities for Λ-coalescents in the infinite-sites model

In this section, we obtain recursions for the probability of given type configurations of a sample based on the probabilistic model discussed above. These recursions lead to a Monte-Carlo method to compute the approximate likelihood of configurations.

We will distinguish two cases. In the first case, we will consider ordered labelled samples of type $(t, \mathbf{a})$, which take the full information contained in the partition $\mathbf{a}$ into account. In the second case, we restrict to numbered ordered configurations of the type $(t, \mathbf{n})$, which only count the multiplicities $\mathbf{n}$.

3.1 Ordered labelled samples

Let us derive the analogue of (5) for Λ-coalescents, again with mutation rate along branches being given by r. With similar (abuse of) notation as above, we have, for given $(t, \mathbf{a})$,

$$\begin{aligned} p(t,\mathbf{a}) = &\frac{1}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{i:\, n_i\geq 2} \sum_{k=2}^{n_i} \binom{n_i}{k} \lambda_{n,k}\, p(t, \mathbf{a}-(k-1)\mathbf{e}_i) \\ &+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1, x_{k0}\text{distinct} \\ s(\mathbf{x}_k)\neq \mathbf{x}_j \forall j}} p(s_k(t), \mathbf{a}) \\ &+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1, \\ x_{k0}\text{ distinct}}} \sum_{j: s(\mathbf{x}_k)=\mathbf{x}_j} p(r_k(t), r_k(\mathbf{a}+\mathbf{e}_j)), \end{aligned} \tag{7}$$

and the boundary condition for the root $p(\{0\},(1)) = 1$. Recursion (7) boils down to (5) in the case that $\Lambda = \delta_0$.

Proof. Similar to the Kingman-case by conditioning on the last event in the coalescent history, taking multiple mergers into account. □

3.2 Numbered ordered samples

Recall from (6), that

$$p^0(t,\mathbf{n}) = \frac{n!}{n_1!\cdots n_d!} p(t,\mathbf{a}). \tag{8}$$

Thus, for the types and multiplicities $(t, \mathbf{n})$, we obtain

$$\begin{aligned} p^0(t,\mathbf{n}) = &\frac{1}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{i:\, n_i\geq 2} \sum_{k=2}^{n_i} \binom{n_i}{k} \lambda_{n,k} \frac{n!}{n_1!\cdots n_d!} \frac{n_1!\cdots(n_i-k+1)!\cdots n_d!}{(n-k+1)!} p^0(t, \mathbf{n}-(k-1)\mathbf{e}_i) \\ &+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1, x_{k0}\text{distinct} \\ s(\mathbf{x}_k)\neq \mathbf{x}_j \forall j}} p^0(s_k(t), \mathbf{n}) \\ &+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1, x_{k0} \\ \text{distinct}}} \sum_{j: s(\mathbf{x}_k)=\mathbf{x}_j} \frac{n!}{n_1!\cdots n_d!} \frac{n_1!\cdots(n_j+1)!\cdots n_d!}{n!} p^0(r_k(t), r_k(\mathbf{n}+\mathbf{e}_j)). \end{aligned}$$

Since

$$\binom{n_i}{k} \frac{n!}{n_1!\cdots n_d!} \frac{n_1!\cdots(n_i-k+1)!\cdots n_d!}{(n-k+1)!} = \frac{n_i!}{k!(n_i-k)!} \frac{n!(n_i-k+1)!}{n_i!(n-k+1)!} = \binom{n}{k} \frac{n_i-k+1}{n-k+1},$$

rearrangement leads to

$$
\begin{aligned}
p^0(t,\mathbf{n}) = & \frac{1}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{i:\, n_i\geq 2} \sum_{k=2}^{n_i} \binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1} p^0(t,\mathbf{n}-(k-1)\mathbf{e}_i) \\
& + \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1,x_{k0}\text{distinct} \\ s(\mathbf{x}_k)\neq \mathbf{x}_j\,\forall j}} p^0(s_k(t),\mathbf{n}) \\
& + \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:n_k=1,x_{k0} \\ \text{distinct}}} \sum_{j:s(\mathbf{x}_k)=\mathbf{x}_j} (n_j+1) p^0(r_k(t), r_k(\mathbf{n}+\mathbf{e}_j)), \qquad (9)
\end{aligned}
$$

with the usual boundary condition for the root, i.e. $p^0(\{0\},(1)) = 1$.

Remark (Unrooted trees). To arrive at probabilities in the case of unrooted trees (corresponding to unknown ancestral types), simply sum over all choices of the root. □

3.3 The block-counting process and a recursion for the site frequency spectrum

In this section, we show how the so-called *block counting process*, which keeps track of the number of blocks of a coalescent-process, can be used to derive the site frequency spectrum for an n-sample in the infinite-sites model. The time-reversal of this process will later be useful in order to obtain urn-like algorithms to produce samples under the finite- and infinite-alleles model.

Definition 3.1 (block-counting process, skeleton chain)**.** *Let $\{\Pi_t\}_{t\geq 0}$ be a Λ-coalescent. We denote by $\{Y_t\}_{t\geq 0}$ the corresponding* block counting process*, i.e. $Y_t = \#$ of blocks of Π_t is a continuous-time Markov chain on $\mathbb{N}$ with jump rates*

$$
q_{ij} = \binom{i}{i-j+1}\lambda_{i,i-j+1}, \quad i > j \geq 1.
$$

The total jump rate while in i is of course $-q_{ii} = \sum_{j=1}^{i-1} q_{ij}$. We write

$$
p_{ij} := \frac{q_{ij}}{-q_{ii}} \qquad (10)
$$

for the jump probabilities of the skeleton chain*, noting that (p_{ij}) is a stochastic matrix.*

Note that in order to reduce i classes to j classes, an $i-j+1$-merger has to occur.

Definition 3.2 (Green's function of Y)**.** *Let*

$$
g(n,m) := \mathbb{E}_n\left[\int_0^\infty \mathbf{1}_{\{Y_s=m\}}\, ds\right] \quad \text{for } n \geq m \geq 2 \qquad (11)
$$

be the expected amount of time that Y, starting from n, spends in m.

Decomposing according to the first jump of Y, we find the following set of equations for $g(n,m)$:

$$g(n,m) = \sum_{k=m}^{n-1} p_{nk} g(k,m), \quad n > m \geq 2, \tag{12}$$

$$g(m,m) = \frac{1}{-q_{mm}}, \quad m \geq 2. \tag{13}$$

Let us write $Y^{(n)}$ for the process starting from $Y_0^{(n)} = n$. Let $\tau := \inf\{t : Y_t^{(n)} = 1\}$ be the time required to come down to only one class, and let

$$\tilde{Y}_t^{(n)} := Y_{(\tau - t)-}^{(n)}, \quad 0 \leq t < \tau$$

be the time-reversed path, where we define $\tilde{Y}_t^{(n)} = \partial$, some cemetery state, when $t \geq \tau$.

Proposition 3.3 (Time-reversal). *With the above definitions, $\tilde{Y}^{(n)}$ is a continuous-time Markov chain on $\{2, \ldots, n\} \cup \{\partial\}$ with jump rates*

$$\tilde{q}_{ji}^{(n)} = \frac{g(n,i)}{g(n,j)} q_{ij}, \quad j < i \leq n,$$

and $\tilde{q}_{n\partial}^{(n)} = -q_{nn}$, where $g(n,m)$ is as in (11). The starting distribution *of $\tilde{Y}^{(n)}$ is given by*

$$\Pr\{\tilde{Y}_0^{(n)} = k\} = g(n,k) q_{k1},$$

for each k.

Proof. The result follows from Nagasawa's Formula, see e.g. [RW87], and the observation

$$\begin{aligned} \Pr\{\tilde{Y}_0^{(n)} = k\} &= \Pr_n \left\{ \tilde{Y}^{(n)} \text{ hits } k, \text{ jumps to 1 from there} \right\} \\ &= \Pr_n \left\{ \tilde{Y}^{(n)} \text{ hits } k \right\} \frac{q_{k1}}{-q_{kk}} \\ &= g(n,k) q_{k1}. \end{aligned}$$

Note that unless Λ is concentrated on $\{0\}$ (Kingman-case), the dynamics of $\tilde{Y}^{(n)}$ does depend on n. □

We now turn our attention to the "site frequency spectrum", in particular under the Beta-coalescent.

Definition 3.4 (Site frequency spectrum). *Consider a sample of size $n \in \mathbb{N}$ obtained in the infinite-sites model, assuming known ancestral types. Let $M_n(b), b \in \{0, \ldots, n\}$ denote the number of mutations which affect precisely b individuals out of the sample. The n-tuple*

$$(M_n(1), \ldots, M_n(n)), \quad M_n(b) \in \{1, \ldots, n\}, b \in \{1, \ldots, n\},$$

is called the (empirical) site frequency spectrum *of the sample. We denote by $\varphi_n(b)$ the probability to see a "typical mutation" b-times in a sample of size n.*

We now determine φ_n with the help of a recursion. Indeed, for $n \geq k > 1$, let $r_{nk}(b)$ be the probability that in an $n-\Lambda$-coalescent, conditioned that there are at some point in time exactly k branches, a given one of these k branches (e.g. the first, if we think of some ordering) subtends exactly b leaves. Obviously $r_{nn}(b) = \delta_{1b}$, and $r_{nk}(b) = 0$ if $b > n-(k-1)$. Decomposing according to the first jump of Y, starting from n, yields the recursion

$$r_{nk}(b) = \sum_{j=k}^{n-1} p_{nj} \frac{g(j,k)}{g(n,k)} \left[\mathbf{1}_{b>n-j} \frac{b-(n-j)}{j} r_{jk}(b-(n-j)) + \mathbf{1}_{b<j} \frac{j-b}{j} r_{jk}(b) \right]. \tag{14}$$

The idea is the following: Assume that the block counting process Y jumps from n down to j. Here, the factor $g(j,k)/g(n,k)$ accounts for the conditioning on hitting k. Then, thinking 'forwards in time from j lineages', either the $(n-j+1)$-split occurred to one of the then necessarily $b-(n-j)$ lineages subtended to the one we are interested in, or it occurs to one of the $j-b$ others. Note that when solving (14) numerically, we can do this separately for each k. Let

$$T_k := \int_0^\infty \mathbf{1}_{\{Y_s=m\}}\, ds$$

be the length of the time interval during which there are k lineages (possibly 0), and

$$\psi_n(b) = \text{expected total length of all branches with } b \text{ subtended leaves} \tag{15}$$

(in an n-Λ-coalescent). We arrive at the following result.

Theorem 3.5 (Distribution of the site frequency spectrum)**.** *Under the above assumptions, we have*

$$\psi_n(b) = \sum_{k=2}^{n-b+1} r_{nk}(b) k \mathbb{E}_n\big[T_k\big] = \sum_{k=2}^{n-b+1} r_{nk}(b) k g(n,k), \tag{16}$$

and the (normalised) site frequency spectrum *distribution is given by the weights*

$$\varphi_n(b) = \frac{\psi_n(b)}{\sum_{\ell=2}^n \ell \mathbb{E}_n\big[T_\ell\big]} = \frac{\sum_{k=2}^{n-b+1} r_{nk}(b) k g(n,k)}{\sum_{\ell=2}^n \ell g(n,\ell)}. \tag{17}$$

Remark. The above is a natural extension of the arguments in [GT98] to the multiple merger case. □

Example: The Beta-coalescent.
In the case when Λ has a Beta(a,b)-density for some $a, b > 0$, i.e.

$$\Lambda(dx) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} 1_{(0,1)}(x)\, dx, \tag{18}$$

the q_{ij} can be computed a little more explicitly:

$$\begin{aligned} \lambda_{n,k} &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{k+a-3}(1-x)^{n-k+b-1}\, dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(k+a-2)\Gamma(n-k+b)}{\Gamma(n-2+a+b)} \\ &= \frac{(a)_{k-2}(b)_{n-k}}{(a+b)_{n-2}}, \end{aligned}$$

where $(x)_i = x(x+1)\cdots(x+i-1)$, $(x)_0 = 1$, and we used $\Gamma(x+1) = x\Gamma(x)$. Thus

$$q_{ij} = \binom{i}{i-j+1}\lambda_{i,i-j+1} = \frac{i!}{(i-j+1)!(j-1)!}\frac{(a)_{i-j-1}(b)_{j-1}}{(a+b)_{i-2}}.$$

□

Note that asymptotic results for the site- (and also the allele-) frequency spectrum have been found by [BBS06], see Theorem 7.1.

4 Finite- and infinite alleles recursions

In this section, we briefly discuss methods to obtain recursions in the above spirit for the finite- and infinite alleles models from mathematical genetics. We will make use of the modified lookdown construction, a generator method, quote Möhle's recursion for the multiplicities in the infinite-alleles case and finally present some algorithms to generate samples under the respective models.

4.1 Finite-alleles I : an approach using the "modified lookdown construction"

We illustrate this method in the case of finitely many types, which we have not treated yet. We think of type changes, or mutations, occurring at rate r, and $P = (P_{ij})$ as the transition matrix on the finite type space E, where silent mutations are allowed (i.e. $P_{jj} \geq 0$). Here, we assume that the reader is familiar with the "modified lookdown construction" (*mld*) of the generalised Λ-Fleming-Viot process, see [DK99] for the general theory or [BBC05], Section 2, for a shorter description. This time, we interpret

$$\lambda_{n,k} = \int_{[0,1]} x^k(1-x)^{n-k}\frac{1}{x^2}\,\Lambda(dx), \quad n \geq k \geq 2 \tag{19}$$

as the rate with which one observes a particular resampling event involving exactly k among the first n levels in the *mld* construction. Suppose the system is in equilibrium. Consider the first n levels at time 0 and let τ_{-1} be the last instant before 0 when at least one of the types at levels $1,\ldots,n$ changes. Then, $-\tau_{-1}$ is exponentially distributed with rate

$$r_n = nr + \sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}. \tag{20}$$

Denote by q the distribution of the types of the first n levels in equilibrium in the modified lookdown construction. Later, due to exchangeability, we will merely be interested in the type frequency probability $p(\mathbf{n})$. Decomposing according to which event occurred at time τ_{-1}, we obtain

$$\begin{aligned} q\big((y_1,\ldots,y_n)\big) = {} & \frac{r}{r_n}\sum_{i=1}^{n}\sum_{z\in E} q\big((y_1,\ldots,y_{i-1},z,y_{i+1},\ldots,y_n)\big)P_{zy_i} \\ & + \frac{1}{r_n}\sum_{\substack{K\subset\{1,\ldots,n\}\\ |K|\geq 2}} \lambda_{n,|K|}1_{\{\text{all } y_j \text{ equal for } j\,\in\, K\}}q\big(\gamma_K(y_1,\ldots,y_n)\big), \end{aligned} \tag{21}$$

where $\gamma_K(y_1,\dots,y_n) \in E^{n-|K|+1}$ is that vector of types of length $n-|K|+1$ which $\big(\xi_1(\tau_{-1}-),\dots,\xi_{n-|K|+1}(\tau_{-1}-)\big)$ must be in order that a resampling event involving exactly the levels in K among levels $1,\dots,n$ generates $\big(\xi_1(\tau_{-1}),\dots,\xi_n(\tau_{-1})\big)=(y_1,\dots,y_n)$. Formally,

$$\gamma_K(y_1,\dots,y_n)_i = y_{i+\#((K\setminus\{\min K\})\cap\{1,\dots,i\})},\ 1\le i\le n-|K|+1.$$

We have the boundary condition $q\big((y_1)\big)=\mu(y_1),\ y_1\in E$. Note that, by exchangeability,

$$q\big((y_1,\dots,y_n)\big)=q\big((y_{\pi(1)},\dots,y_{\pi(n)})\big)$$

for any permutation π of $\{1,\dots,n\}$. So, the only relevant information is (of course) how many samples were of which type. For $\mathbf{n}=(n_1,\dots,n_d)\in\mathbb{Z}_+^d$ we write $\#\mathbf{n}:=n_1+\cdots+n_d$ for the 'length', and

$$\kappa(\mathbf{n})=\big(\underbrace{1,1,\dots,1}_{n_1},\underbrace{2,\dots,2}_{n_2},\dots,\underbrace{d,\dots,d}_{n_d}\big)\in E^{\#\mathbf{n}}$$

for a 'canonical representative' of the (absolute) type frequency vector $\mathbf{n}$. Put $\tilde q(\mathbf{n}):=q(\kappa(\mathbf{n}))$ and let

$$p(\mathbf{n}):=\binom{\#\mathbf{n}}{n_1,n_2,\dots,n_d}\tilde q(\mathbf{n}) \tag{22}$$

be the probability that in a sample of size $\#\mathbf{n}$, there are exactly n_j of type j, $j=1,\dots,d$. (21) translates into a recursion for p: (we abbreviate $n:=\#\mathbf{n}$, and write e_k for the k-th canonical unit vector of $\mathbb{Z}^d$)

$$\begin{aligned} p(\mathbf{n}) &= \frac{r}{r_n}\sum_{j=1}^d n_j\sum_{i=1}^d P_{ij}\binom{\#\mathbf{n}}{n_1,n_2,\dots,n_d}\tilde q(\mathbf{n}-e_j+e_i)\\ &\quad+\frac{1}{r_n}\sum_{j=1}^d\sum_{k=2}^{n_j}\binom{n_j}{k}\lambda_{n,k}\binom{\#\mathbf{n}}{n_1,n_2,\dots,n_d}\tilde q(\mathbf{n}-e_j+e_i).\end{aligned}$$

Note that

$$\begin{aligned} n_j\binom{\#\mathbf{n}}{n_1,n_2,\dots,n_d}\tilde q(\mathbf{n}-e_j+e_i) &= \big(n_i+1-\delta_{ij}\big)\binom{\#\mathbf{n}}{n_1,\dots,n_i+1,\dots,n_j-1,\dots,n_d}\tilde q(\mathbf{n}-e_j+e_i)\\ &= \big(n_i+1-\delta_{ij}\big)p(\mathbf{n}-e_j+e_i)\end{aligned}$$

and that (for $n_j\ge k$, otherwise the term is 0)

$$\begin{aligned}\binom{n_j}{k}\binom{\#\mathbf{n}}{n_1,n_2,\dots,n_d}\tilde q(\mathbf{n}-(k-1)e_j) &= \binom{n}{k}\frac{n_j-k+1}{n-k+1}\binom{n-k+1}{n_1,\dots,n_j-k+1,\dots,n_d}\tilde q(\mathbf{n}-(k-1)e_j)\\ &= \binom{n}{k}\frac{n_j-k+1}{n-k+1}p(\mathbf{n}-(k-1)e_j).\end{aligned}$$

Thus, the recursion for p is

$$p(\mathbf{n}) = \frac{r}{r_n}\sum_{j=1}^d\sum_{i=1}^d(n_i+1-\delta_{ij})P_{ij}p(\mathbf{n}-e_j+e_i) \tag{23}$$

$$+\frac{1}{r_n}\sum_{\substack{j=1\\ n_j\ge 2}}^d\sum_{k=2}^{n_j}\binom{n}{k}\lambda_{n,k}\frac{n_j-k+1}{n-k+1}p(\mathbf{n}-(k-1)e_j) \tag{24}$$

with boundary conditions $p(e_j) = \mu_j$.

Remark. In the Kingman-case, we have $\lambda_{n,k} = 1(n \geq 2 = k)$, $r_n = n\theta/2 + n(n-1)/2 = n(n-1+\theta)/2$ (and we assume $r = \theta/2$ as 'usual'), hence (23) becomes

$$\begin{aligned} p(\mathbf{n}) &= \frac{2}{n(n-1+\theta)} \frac{\theta}{2} \sum_{j=}^{d} \sum_{i=1}^{d} (n_i + 1 - \delta_{ij}) P_{ij} p(\mathbf{n} - e_j + e_i) \\ &\quad + \frac{2}{n(n-1+\theta)} \sum_{\substack{j=1 \\ n_j \geq 2}}^{d} \binom{n}{2} \frac{n_j - 1}{n-1} p(\mathbf{n} - e_j) \\ &= \frac{\theta}{n-1+\theta} \sum_{j=}^{d} \sum_{i=1}^{d} \frac{n_i + 1 - \delta_{ij}}{n} P_{ij} p(\mathbf{n} - e_j + e_i) \\ &\quad + \frac{n-1}{n-1+\theta} \sum_{\substack{j=1 \\ n_j \geq 2}}^{d} \frac{n_j - 1}{n-1} p(\mathbf{n} - e_j), \end{aligned}$$

which agrees with (3) in [DIG04a].

4.2 Finite-alleles II: the generator approach

An alternative method to obtain the recursion for the type probabilities in the finite-sites case is by using a generator approach, see [DIG04a]. Let $f \in C_2$ and $\Delta_d = \{(x_1, \ldots, x_d) : x_i \geq 0, x_1 + \cdots + x_d = 1\}$ and consider the mutation operator

$$Bf(x_1, \ldots, x_d) = r \sum_{i=1}^{d} \Big(\sum_{j=1}^{d} x_j P_{ji} - x_i P_{ij} \Big) \frac{\partial f}{\partial x_i}(x_1, \ldots, x_d)$$

For the resampling operator, we distinguish the Kingman- and non-Kingman components. First, assume $\Lambda(0) = 0$ (non-Kingman). Consider

$$\begin{aligned} R_1 f(x_1, \ldots, x_d) = \sum_{i=1}^{d} \int x_i \Big(& f\big((1-r)x_1, \ldots, (1-r)x_{i-1}, (1-r)x_i + r, (1-r)x_{i+1}, \ldots, (1-r)x_d\big) \\ & - f(x_1, \ldots, x_d) \Big) r^{-2} \Lambda(dr). \end{aligned} \tag{25}$$

For the Kingman-part ($\Lambda = \delta_0$) of the resampling operator, we have

$$R_2 f(x_1, \ldots, x_d) = \frac{1}{2} \sum_{i,j=1}^{d} x_i (\delta_{ij} - x_j) \frac{\partial^2 f}{\partial x_i \partial x_j}(x_1, \ldots, x_d).$$

Finally, for general Λ and $a \geq 0$, write

$$R = R_1 + a R_2,$$

where R_1 uses Λ', $\Lambda'(\cdot) := \Lambda(\cdot \cap (0,1])$, $a = \Lambda(0)$. Now, let $X(t) = (X_1(t), \ldots, X_d(t))$ be the stationary process with generator $L = B + R$ (see [BLG03]). Write $X = X(0)$. Let $\mathbf{n} = (n_1, \ldots, n_d)$, $n =$

$n_1 + \cdots + n_d$. Then,

$$\mathbb{E}\Big[\prod_{i=1}^{d} X_i^{n_i}\Big]$$

is the probability of observing in a sample of size n from the equilibrium population type i precisely n_i times in a particular order (e.g. first n_1 samples of type 1, next n_2 samples of type 2, etc.). Put

$$f_{\mathbf{n}}(\mathbf{x}) := \mathbf{x}^{\mathbf{n}} := \prod_{i=1}^{d} x_i^{n_i}.$$

Then,

$$g(\mathbf{n}) := \binom{n}{n_1 \dots n_d} \mathbb{E}\big[f_{\mathbf{n}}(X)\big]$$

is the probability of observing type i exactly n_i times, $i = 1, \dots, d$, without regard of the order. In equilibrium, we have $\mathbb{E}Lf(X) = 0$. Note that

$$\begin{aligned} Bf_{\mathbf{n}}(x_1, \dots, x_d) &= r \sum_{i=1}^{d} \Big(\sum_{j=1}^{d} x_j P_{ji} - x_i P_{ij} \Big) n_i f_{\mathbf{n}-\mathbf{e}_i}(x_1, \dots, x_d) \\ &= r \sum_{i,j=1}^{d} n_i P_{ji} f_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j}(\mathbf{x}) - rn f_{\mathbf{n}}(\mathbf{x}) \end{aligned}$$

and

$$\begin{aligned} f_{\mathbf{n}}((1-r)\mathbf{x} + r\mathbf{e_i}) &= (1-r)^{n-n_i} \prod_{j \neq i}^{d} x_j^{n_j} \times \big((1-r)x_i + r\big)^{n_i} \\ &= (1-r)^{n-n_i} \prod_{j \neq i}^{d} x_j^{n_j} \times \sum_{k=0}^{n_i} \binom{n_i}{k} r^k (1-r)^{n_i-k} x_i^{n_i-k} \\ &= \sum_{k=0}^{n_i} \binom{n_i}{k} r^k (1-r)^{n-k} \Big(x_i^{n_i-k} \prod_{j \neq i}^{d} x_j^{n_j} \Big), \end{aligned}$$

so the term inside the integral in the expression (25) for R_1 can be written as

$$\begin{aligned} &\sum_{i=1}^{d} \sum_{k=0}^{n_i} \binom{n_i}{k} r^k (1-r)^{n-k} x_i^{n_i-k+1} \prod_{j \neq i}^{d} x_j^{n_j} - \sum_{k=0}^{n} \binom{n}{k} r^k (1-r)^{n-k} \prod_{\ell=1}^{d} x_\ell^{n_\ell} \\ &= \sum_{i:n_i \geq 2}^{d} \sum_{k=2}^{n_i} \binom{n_i}{k} r^k (1-r)^{n-k} x_i^{n_i-k+1} \prod_{j \neq i}^{d} x_j^{n_j} - \sum_{k=2}^{n} \binom{n}{k} r^k (1-r)^{n-k} \prod_{\ell=1}^{d} x_\ell^{n_\ell} \end{aligned}$$

(the terms with $k = 0$ and $k = 1$ cancel since $x_1 + \cdots + x_d = 1$ and $n_1 + \cdots + n_d = n$). Recalling $\lambda_{n,k} = \int r^{k-2}(1-r)^{n-k} \Lambda(dr)$ we obtain

$$R_1 f_{\mathbf{n}}(\mathbf{x}) = \sum_{i:n_i \geq 2}^{d} \sum_{k=2}^{n_i} \binom{n_i}{k} \lambda_{n,k} f_{\mathbf{n}-(k-1)\mathbf{e}_i}(\mathbf{x}) - \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} f_{\mathbf{n}}(\mathbf{x}). \tag{26}$$

Furthermore

$$\begin{aligned} R_2 f_{\mathbf{n}}(\mathbf{x}) &= \frac{1}{2}\sum_{i,j=1}^{d} x_i(\delta_{ij}-x_j)n_i(n_j-\delta_{ij})f_{\mathbf{n}-\mathbf{e}_i-\mathbf{e}_j}(\mathbf{x}) \\ &= \sum_{i=1}^{d}\frac{n_i(n_i-1)}{2}f_{\mathbf{n}-\mathbf{e}_i}(\mathbf{x}) - \sum_{i,j=1}^{d}\frac{n_i(n_j-\delta_{ij})}{2}f_{\mathbf{n}}(\mathbf{x}) \\ &= \sum_{i=1}^{d}\frac{n_i(n_i-1)}{2}f_{\mathbf{n}-\mathbf{e}_i}(\mathbf{x}) - \frac{n(n-1)}{2}f_{\mathbf{n}}(\mathbf{x}). \end{aligned} \tag{27}$$

Combining the terms from R_1 and R_2 (using (26) and (27) above, and replacing Λ by Λ' in (25)), we have

$$Rf_{\mathbf{n}}(\mathbf{x}) = \sum_{i:n_i\geq 2}^{d}\sum_{k=2}^{n_i}\binom{n_i}{k}\lambda_{n,k}f_{\mathbf{n}-(k-1)\mathbf{e}_i}(\mathbf{x}) - \sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}f_{\mathbf{n}}(\mathbf{x}).$$

Thus we obtain from $\mathbb{E}\,Lf_{\mathbf{n}}(X)=0$:

$$z_n\mathbb{E}f_{\mathbf{n}}(X) = r\sum_{i,j=1}^{d} n_iP_{ji}\mathbb{E}f_{\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j}(X) + \sum_{i:n_i\geq 2}^{d}\sum_{k=2}^{n_i}\binom{n_i}{k}\lambda_{n,k}\mathbb{E}f_{\mathbf{n}-(k-1)\mathbf{e}_i}(X),$$

where

$$z_n = rn + \sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}.$$

Multiply with $\binom{n}{n_1\ldots n_d}$ to obtain

$$\begin{aligned} z_n g(\mathbf{n}) &= r\sum_{i,j=1}^{d}(n_j+1-\delta_{ij})P_{ji}g(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) \\ &\quad + \sum_{i:n_i\geq 2}^{d}\sum_{k=2}^{n_i}\underbrace{\binom{n_i}{k}\frac{n!}{n_1!\ldots n_d!}\frac{n_1!\ldots(n_i-k+1)!\ldots n_d!}{(n-k+1)!}}_{=\frac{n_i!}{k!(n_i-k)!}\frac{n!}{n_i!}\frac{(n_i-k+1)!}{(n-k+1)!}=\frac{n!}{k!(n-k)!}\frac{n_i-k+1}{n-k+1}}\lambda_{n,k}\,g(\mathbf{n}-(k-1)\mathbf{e}_i) \\ &= r\sum_{i,j=1}^{d}(n_j+1)P_{ji}g(\mathbf{n}-\mathbf{e}_i+\mathbf{e}_j) \\ &\quad + \sum_{i:n_i\geq 2}^{d}\sum_{k=2}^{n_i}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1}\,g(\mathbf{n}-(k-1)\mathbf{e}_i) \end{aligned}$$

which agrees with (23) after dividing by z_n.

4.3 Infinite-alleles: Möhle's recursion

Here, one assumes that every mutation, which occurs along the coalescent tree with rate $r>0$, leads to an entirely new type, no other information is being retained. If we take a sample of $n\in\mathbb{N}$ genes,

it is natural to ask for the probability $p(\mathbf{n})$ to sample a specific, non-ordered allele configuration $\mathbf{n} = (n_1, \dots, n_k)$, where $k \leq n$ is the number of different types and $n_i, i \in \{1, \dots, k\}$ is the number of times that type i is being observed. Using coalescent arguments, it is possible obtain the following recursion, see [M06b], Theorem 3.1.

Theorem 4.1 (Möhle (2006)). *The probability of a non-ordered allele configuration* $\mathbf{n} = (n_1, \dots, n_k)$ *satisfies the recursion given by* $p(1) = 1$ *and*

$$p(\mathbf{n}) = \frac{nr}{\sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} + nr} \sum_{\substack{j=1 \\ n_j = 1}}^{k} \frac{1}{k} p(\tilde{\mathbf{n}}_j) + \frac{1}{\sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} + nr} \sum_{i=2}^{n} \sum_{\substack{j=1 \\ n_j \geq i}}^{k} \lambda_{n,i} \frac{n_j - i + 1}{n - i + 1} p(\mathbf{n} - (i-1)\mathbf{e}_j), \tag{28}$$

with $n = \sum_j n_j \geq 2$, $r = \theta/2$, *and* $\tilde{\mathbf{n}}_j = (n_1, \dots, n_{j-1}, n_{j+1}, \dots, n_k)$. *As before,* $\mathbf{e}_j$ *denotes the unit vector in* $\mathbb{R}^k$.

In the Kingman-case, this recursion can be solved explicitly and leads to an alternative formulation of the famous *Ewens sampling formula*, see [E79]. It seems that the only other case in which an explicit solution is known is the case $\Lambda = \delta_1$, in which the genealogy is star-shaped.

5 A Monte Carlo method for the computation of the likelihoods in the infinite-sites model

We first derive a simple Monte-Carlo approximation of the exact sampling probabilities in the infinite-sites model by simulating a Markov chain backwards along the sample paths of the coalescent (essentially based on [GT94b], see also [T01]).

5.1 An unbiased estimator for $p^0(t, \mathbf{n})$

First, we recall a suitable notion of tree complexity.

Definition 5.1 (Tree complexity). *Given ordered types and frequencies* $(t, \mathbf{n})$, *we define the* tree complexity *of* $(t, \mathbf{n})$ *as*

$$c[(t, \mathbf{n})] = \sum_{i=1}^{d} n_i + \sum_{i=1}^{d} \#\mathbf{x}_i \in \mathbb{N},$$

where, for $1 \leq i \leq d$, $\#\mathbf{x}_i$ *denotes the length of the sequence* $\mathbf{x}_i$ *(exclusive of the root).*

Note that the tree complexity is the sum of the sample size and the number of segregating sites. This definition transfers in the obvious way also to the pair of ordered types and partitions $(t, \mathbf{a})$. It is clear that the tree complexity is independent of the choice of a representative t from the equivalence class $[t]$ and hence well-defined. If $c[(t, \mathbf{n})] = 1$, the minimum for a non-vanishing tree, then the tree consists only of its root with multiplicity one, i.e. $(t, \mathbf{n}) = (\{0\}, (1)) =: t_0$. We write

$$(t', \mathbf{n}') \prec (t, \mathbf{n})$$

if $(t', \mathbf{n}')$ can be reached from $(t, \mathbf{n})$ by either removing one mutation or a coalescence event, see below. In this case, $c[(t', \mathbf{n}')] < c[(t, \mathbf{n})]$. Hence observe that the recursions (5) and (9) are proper recursions in the sense that they strictly decrease the tree complexity in each step.

The following lemma is an appropriate version of the corresponding Lemma 6.1 in [T01].

Lemma 5.2. *Let $\{X_k, k \geq 0\}$ be a Markov chain on the space of ordered types with corresponding frequencies, denoted by $(\mathcal{T}, \mathcal{N})$, and with transitions $Q = (q_{(t,\mathbf{n}),(t',\mathbf{n}')})$ such that the hitting time*

$$\tau = \inf\big\{k \geq 0 \,:\, X_k = (\{0\}, (1))\big\}$$

for any given initial state $(t, \mathbf{n})$ in $(\mathcal{T}, \mathcal{N})$ is bounded by some constant $0 \leq K_1(t, \mathbf{n}) < \infty$. Let $f : (\mathcal{T}, \mathcal{N}) \to [0, \infty)$ be a measurable function and define

$$u_{(t,\mathbf{n})}(f) = \mathbb{E}_{(t,\mathbf{n})} \prod_{k=0}^{\tau} f(X_k) \tag{29}$$

for all $X_0 = (t, \mathbf{n}) \in (\mathcal{T}, \mathcal{N})$, so that

$$u_{(\{0\},(1))}(f) = f\big(\{0\}, (1)\big).$$

Then

$$u_{(t,\mathbf{n})}(f) = f\big((t, \mathbf{n})\big) \sum_{\substack{(t,\mathbf{n}) \in (\mathcal{T},\mathcal{N}) \\ (t',\mathbf{n}') \prec (t,\mathbf{n})}} q_{(t,\mathbf{n}),(t',\mathbf{n}')} u_{(t',\mathbf{n}')}(f) \tag{30}$$

for all $(t, \mathbf{n}) \in (\mathcal{T}, \mathcal{N}) \backslash (\{0\}, (1))$. Conversely, the unique solution of (30) is given by (29).

Remark. If the transitions $Q = (q_{(t',\mathbf{n}'),(t,\mathbf{n})})$ are only positive if $c[(t', \mathbf{n}')] < c[(t, \mathbf{n})]$, then

$$\tau = \inf\big\{k \geq 0 \,:\, X_k = (\{0\}, (1))\big\}$$

is always bounded from above by the tree complexity of the initial state. □

Proof. Note that by the boundedness of τ, the expected value remains finite for each initial condition. Now, compute

$$\begin{aligned}
u_{(t,\mathbf{n})}(f) &= \mathbb{E}_{(t,\mathbf{n})} \prod_{k=0}^{\tau} f(X_k) \\
&= f(t, \mathbf{n})\, \mathbb{E}_{(t,\mathbf{n})} \prod_{k=1}^{\tau} f(X_k) \\
&= f(t, \mathbf{n})\, \mathbb{E}_{(t,\mathbf{n})} \Big[\mathbb{E}_{(t,\mathbf{n})} \prod_{k=1}^{\tau} f(X_k) \mid X_1 \Big] \\
&= f(t, \mathbf{n})\, \mathbb{E}_{(t,\mathbf{n})} \Big[\mathbb{E}_{X_1} \prod_{k=0}^{\tau} f(X_k) \Big] \\
&= f(t, \mathbf{n})\, \mathbb{E}_{(t,\mathbf{n})} \big[u_{X_1}(f) \big] \\
&= f(t, \mathbf{n}) \sum_{\substack{(t',\mathbf{n}') \in (\mathcal{T},\mathcal{N}) \\ (t',\mathbf{n}') \prec (t,\mathbf{n})}} q_{(t,\mathbf{n}),(t',\mathbf{n}')} u_{(t',\mathbf{n}')}(f),
\end{aligned}$$

as required. □

The result provides a simulation method for solving recursions of type (30): simulate a trajectory of the chain X starting at $(t,\mathbf{n})$ until it hits the root $(\{0\},(1))$ at time τ, compute the value of the product $\prod_{k=0}^{\tau} f(X_k)$ and repeat this many times. Averaging these values provides an unbiased and consistent estimate of $u_{(t,\mathbf{n})}(f)$ in terms of an approximation of the expected value $\mathbb{E}_{(t,\mathbf{n})}\prod_{k=0}^{\tau} f(X_k)$ by the strong law of large numbers. Lemma 5.2 states that this expectation is a solution to the recursion in question.

Corollary 5.3. *For ordered types and frequencies $(t,\mathbf{n})$, define*

$$u_{(t,\mathbf{n})}(f) = p^0(t,\mathbf{n})$$

and for $c[(t,\mathbf{n})] > 1$, put

$$f(t,\mathbf{n}) = \frac{1}{r_n}\left(\sum_{\substack{k:n_k=1,x_{k0}\ distinct\\ s_k(\mathbf{x}_k)\neq\mathbf{x}_j\forall j}} r \;+\; \sum_{\substack{k:n_k=1,\\ x_{k0}\ distinct}}\ \sum_{j:s_k(\mathbf{x}_k)=\mathbf{x}_j} r(n_j+1) + \sum_{i:\,n_i\geq 2}\sum_{k=2}^{n_i}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1}\right) \tag{31}$$

where

$$r_n\,(=r_n(r,\Lambda)) = rn + \sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}. \tag{32}$$

Furthermore, let

$$u_{(\{0\},(1))}(f) = f\big(\{0\},(1)\big) = 1. \tag{33}$$

Consider a Markov-Chain $\{X_l = (t(l),\mathbf{n}(l))\}$ on $(\mathcal{T},\mathcal{N})$ with transitions

$$(t,\mathbf{n}) \to \begin{cases} (s_k(t),\mathbf{n}) & \text{with probability } \frac{r}{r_n f(t,\mathbf{n})} \quad \text{if} \quad n_k=1, x_{k0}\ \text{distinct},\ s(\mathbf{x}_k)\neq\mathbf{x}_j\ \forall j,\\ (r_k(t), r_k(\mathbf{n}+\mathbf{e}_j)) & \text{with probability } \frac{r(n_j+1)}{r_n f(t,\mathbf{n})} \quad \text{if} \quad n_k=1, x_{k0}\ \text{distinct},\ s(\mathbf{x}_k)=\mathbf{x}_j,\\ (t,\mathbf{n}-(k-1)\mathbf{e}_i) & \text{with probability } \frac{1}{r_n f(t,\mathbf{n})}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1} \quad \text{if} \quad 2\leq k\leq n_i. \end{cases}$$

Then,

$$p^0(t,\mathbf{n}) = \mathbb{E}_{(t,\mathbf{n})}\prod_{l=0}^{\tau} f(t(l),\mathbf{n}(l)).$$

Proof. This is the immediate application of Lemma 5.2, noting that, as in last remark, starting from $(t,\mathbf{n})$, the stopping time τ is bounded by $c[(t,\mathbf{n})] < \infty$. □

Simulating independent copies and taking the average now yields an unbiased estimator of $p^0(t,\mathbf{n})$. Note that a similar result holds for the recursion w.r.t. $(t,\mathbf{a})$.

5.2 Simulation of likelihood surfaces with pre-specified driving values.

It is actually possible to obtain simultaneous likelihoods for a variety of values for r,Λ, using a single realization of the Markov-chain X only. First, we need to extend Lemma 5.2 as in Subsection 6.2 in [T01].

Lemma 5.4. *Let $\{X_k, k \geq 0\}$ be a Markov chain with state space $(\mathcal{T},\mathcal{N})$ and with transitions $Q = (q_{(t,\mathbf{n}),(t',\mathbf{n}')})$ such that the hitting time*

$$\tau = \inf\big\{k \geq 0 \,:\, X_k = (\{0\},(1))\big\}$$

for any given initial state $(t,\mathbf{n})$ in $(\mathcal{T},\mathcal{N})$ is bounded by some constant $0 \leq K_2(t,\mathbf{n}) < \infty$. Let $g : (\mathcal{T},\mathcal{N}) \times (\mathcal{T},\mathcal{N}) \to [0,\infty)$ be a measurable function and define

$$u_{(t,\mathbf{n})}(g) = \mathbb{E}_{(t,\mathbf{n})} \prod_{k=0}^{\tau-1} g(X_k, X_{k+1}) \tag{34}$$

for all $X_0 = (t,\mathbf{n}) \in (\mathcal{T},\mathcal{N})$), with $u_{(\{0\},(1))}(g) = 1$. Then, for all $(t,\mathbf{n}) \in (\mathcal{T},\mathcal{N})\backslash(\{0\},(1))$,

$$u_{(t,\mathbf{n})}(g) = \sum_{\substack{(t,\mathbf{n})\in(\mathcal{T},\mathcal{N})\\(t',\mathbf{n}')\prec(t,\mathbf{n})}} g\big((t,\mathbf{n}),(t',\mathbf{n}')\big) q\big((t,\mathbf{n}),(t',\mathbf{n}')\big) u_{(t',\mathbf{n}')}(g) \tag{35}$$

and this set of equations has the unique solution (34).

Proof. Similar to the proof of Lemma 5.2. □

We follow the spirit of Proposition 5.3 and rewrite (9) to be of the form (35). To this end, define $p^0_{(r,\Lambda)}(t,\mathbf{n})$ to be the probability of observing the unordered, labelled tree $(t,\mathbf{n})$ if the underlying mutation rate is r and the genealogy is governed by a Λ-coalescent.

Corollary 5.5. *Let (r,Λ) and $(r^*,\Lambda^*) \in \mathbb{R}_+ \times \mathcal{M}([0,1])$ be given. For ordered types and frequencies $(t,\mathbf{n})$, define $f_{(r,\Lambda)}(t,\mathbf{n})$ through (31) - (33) and similarly $f_{(r^*,\Lambda^*)}(t,\mathbf{n})$. Consider a Markov-Chain $\{X_l = (t(l),\mathbf{n}(l))\}$ on $(\mathcal{T},\mathcal{N})$ with transitions $q_{(r^*,\Lambda^*)}$ given by*

$$(t,\mathbf{n}) \to \begin{cases} (s_k(t),\mathbf{n}) & \text{with probability} \quad \frac{r^*}{r^*_n f_{(r*,\Lambda*)}(t,\mathbf{n})} \quad \text{if} \quad n_k = 1, x_{k0} \text{ distinct, } s(\mathbf{x}_k) \neq \mathbf{x}_j \,\forall j,\\ (r_k(t), r_k(\mathbf{n}+\mathbf{e}_j)) & \text{with probability} \quad \frac{r^*(n_j+1)}{r^*_n f_{(r*,\Lambda*)}(t,\mathbf{n})} \quad \text{if} \quad n_k = 1, x_{k0} \text{ distinct, } s(\mathbf{x}_k) = \mathbf{x}_j,\\ (t,\mathbf{n}-(k-1)\mathbf{e}_i) & \text{with probability} \quad \frac{1}{r^*_n f_{(r*,\Lambda*)}(t,\mathbf{n})}\binom{n}{k}\lambda^*_{n,k}\frac{n_i-k+1}{n-k+1} \quad \text{if} \quad 2 \leq k \leq n_i. \end{cases}$$

Then, defining

$$g_{(r,\Lambda),(r^*,\Lambda^*)}((t,\mathbf{n}),(t',\mathbf{n}')) = f_{(r,\Lambda)}(t,\mathbf{n}) \frac{q_{(r,\Lambda)}\Big((t,\mathbf{n}),(t',\mathbf{n}')\Big)}{q_{(r^*,\Lambda^*)}\Big((t,\mathbf{n}),(t',\mathbf{n}')\Big)},$$

one has

$$p^0_{(r,\Lambda)}(t,\mathbf{n}) = \mathbb{E}^{(r^*,\Lambda^*)}_{(t,\mathbf{n})} \prod_{k=0}^{\tau-1} g_{(r,\Lambda),(r^*,\Lambda^*)}(X_k, X_{k+1}), \tag{36}$$

provided that the parameters (r,Λ), (r^*,Λ^*) fulfil the condition

$$f_{(r,\Lambda)}(t,\mathbf{n}) q_{(r,\Lambda)}\Big((t,\mathbf{n}),(t',\mathbf{n}')\Big) > 0 \;\Rightarrow\; q_{(r^*,\Lambda^*)}\Big((t,\mathbf{n}),(t',\mathbf{n}')\Big) > 0. \tag{37}$$

Again, this gives rise to a simulation algorithm, this time based on (r^*, Λ^*) rather than the "target" (r, Λ).

Proof. We may rewrite (9) as

$$p^0_{(r,\Lambda)}(t, \mathbf{n}) = \sum_{\substack{(t',\mathbf{n}'):\\ (t',\mathbf{n}') \prec (t,\mathbf{n})}} f_{(r,\Lambda)}(t, \mathbf{n}) q_{(r,\Lambda)}\Big((t, \mathbf{n}), (t', \mathbf{n}')\Big) p^0_{(r,\Lambda)}(t', \mathbf{n}') \tag{38}$$

for the obvious choice for $q_{(r,\Lambda)}$. Furthermore, using (37), (38) may be recast as

$$p^0_{(r,\Lambda)}(t, \mathbf{n}) = \sum_{\substack{(t',\mathbf{n}'):\\ (t',\mathbf{n}') \prec (t,\mathbf{n})}} f_{(r,\Lambda)}(t, \mathbf{n}) \frac{q_{(r,\Lambda)}\Big((t, \mathbf{n}), (t', \mathbf{n}')\Big)}{q_{(r^*,\Lambda^*)}\Big((t, \mathbf{n}), (t', \mathbf{n}')\Big)} q_{(r^*,\Lambda^*)}\Big((t, \mathbf{n}), (t', \mathbf{n}')\Big) p^0_{(r,\Lambda)}(t', \mathbf{n}'), \tag{39}$$

hence

$$p^0_{(r,\Lambda)}(T, \mathbf{n}) = \sum_{\substack{(t',\mathbf{n}'):\\ (t',\mathbf{n}') \prec (t,\mathbf{n})}} g_{(r,\Lambda),(r^*,\Lambda^*)}((t, \mathbf{n}), (t', \mathbf{n}')) q_{(r^*,\Lambda^*)}\Big((t, \mathbf{n}), (t', \mathbf{n}')\Big) p^0_{(r,\Lambda)}(t', \mathbf{n}'), \tag{40}$$

so that Lemma 5.4 may directly be applied to equation (40) and the Markov chain $X_l = (t(l), \mathbf{n}(l))$ with driving values r^* and $(\lambda^*_{n,k})_{2 \le k \le n}$ (coming from Λ^*) and transitions as above. Thus we arrive at the representation

$$p^0_{(r,\Lambda)}(t, \mathbf{n}) = \mathbb{E}^{(r^*,\Lambda^*)}_{(t,\mathbf{n})} \prod_{k=0}^{\tau-1} g_{(r,\Lambda),(r^*,\Lambda^*)}(X_k, X_{k+1}),$$

as required. □

With this result, many estimators for $p^0_{(r,\Lambda)}(t, \mathbf{n})$ for various values of (r, Λ), respecting the absolute continuity condition (37), can be obtained by simulating just one realization of the Markov chain with driving values (r^*, Λ^*). This seems computationally much more efficient than using different driving values. However, one should be aware that one obtains correlated estimates and that the variance of the estimator for $p^0_{(r,\Lambda)}(t, \mathbf{n})$ depends on (r^*, Λ^*).

Remark. There are obvious improvements of this method. Combining likelihoods in approximately optimal linear combinations of the (r_i, Λ_i) leads to a further reduction in variance (see [T01] for details). More advanced techniques such as a sophisticated *importance sampling* in the spirit of [SD00] or *bridge sampling* are under investigation in an ongoing research project.

6 Illustration using artificial and real data

In this section, we apply the simulation algorithm suggested by Corollaries 5.3 and 5.5 to two kinds of data, namely randomly generated type configurations, where the underlying genealogy has been obtained from a Lambda-coalescent and the mechanism described in Section 2 resp. the algorithm described in Subsection 8.3, and to real data, namely random sub-samples drawn from Arnason's Atlantic Cod data [A04]. The latter is necessary, since our method, at present, can only deal reasonably well with samples of size up to $n = 100$, whereas the size of Arnason's data is about $n = 1000$.

To simulate a random sample, we used the `R` program `simbeta.R`. This produces data of the form as in Subsection 1.2.

Such data can easily be visualised using the program `treepic` from Bob Griffith's `genetree` software suite.

The program `bgt0.3` is an implementation of the above Monte Carlo method and is, together with the technical report [B06] documenting the program, available from Matthias Birkner.

The resulting likelihood-surfaces can be seen below. Although our methods are not yet very sophisticated and we were subject to limited computing resources, at a first glance it seems to be possible to reject the "Kingman line" $\alpha = 2$ in Figure 6.1 for artificially generated data obtained from an underlying Beta-coalescent with $\alpha = 1.5$ and mutation rate 2.

Moreover, the likelihood-surface for Arnason's real data looks qualitatively different from the one associated with a Kingman-coalescent, in particular has a maximum sufficiently far away from the "Kingman line" $\alpha = 2$, although this time the surface is more flat than in the artificially generated data with $\alpha < 2$.

6.1 Likelihood-surfaces for randomly generated data

We consider the log-likelihood surfaces for type configuration under the Beta$(2-\alpha, \alpha)$-coalescent as a function of $\alpha \in (1, 2]$ and mutation rate $\theta \in (0, 5]$.

Figure 6.1 shows the log-likelihood surfaces (actually on a 50 by 50 grid) for type-configurations drawn under the infinite-sites model, where the first sample (of size $n = 50$) has been obtained from a Kingman-coalescent with mutation rate 2, and the second sample has been obtained from a Beta-coalescent with $\alpha = 1.25$ and mutation rate 2.

6.2 Likelihood-surfaces of samples taken from for Arnason's data

Again, we consider the log-likelihood surfaces for a type configuration under the Beta$(2-\alpha, \alpha)$-coalescent as a function of $\alpha \in (1, 2]$ and mutation rate $\theta \in (0, 2]$.

Figure 6.2 shows the two likelihood surfaces corresponding to two independent samples of 50 and 117 sequences drawn from a slightly modified set of Arnasons's Atlantic Cod data [A04]. It seems reasonable to reject the Kingman-hypothesis.

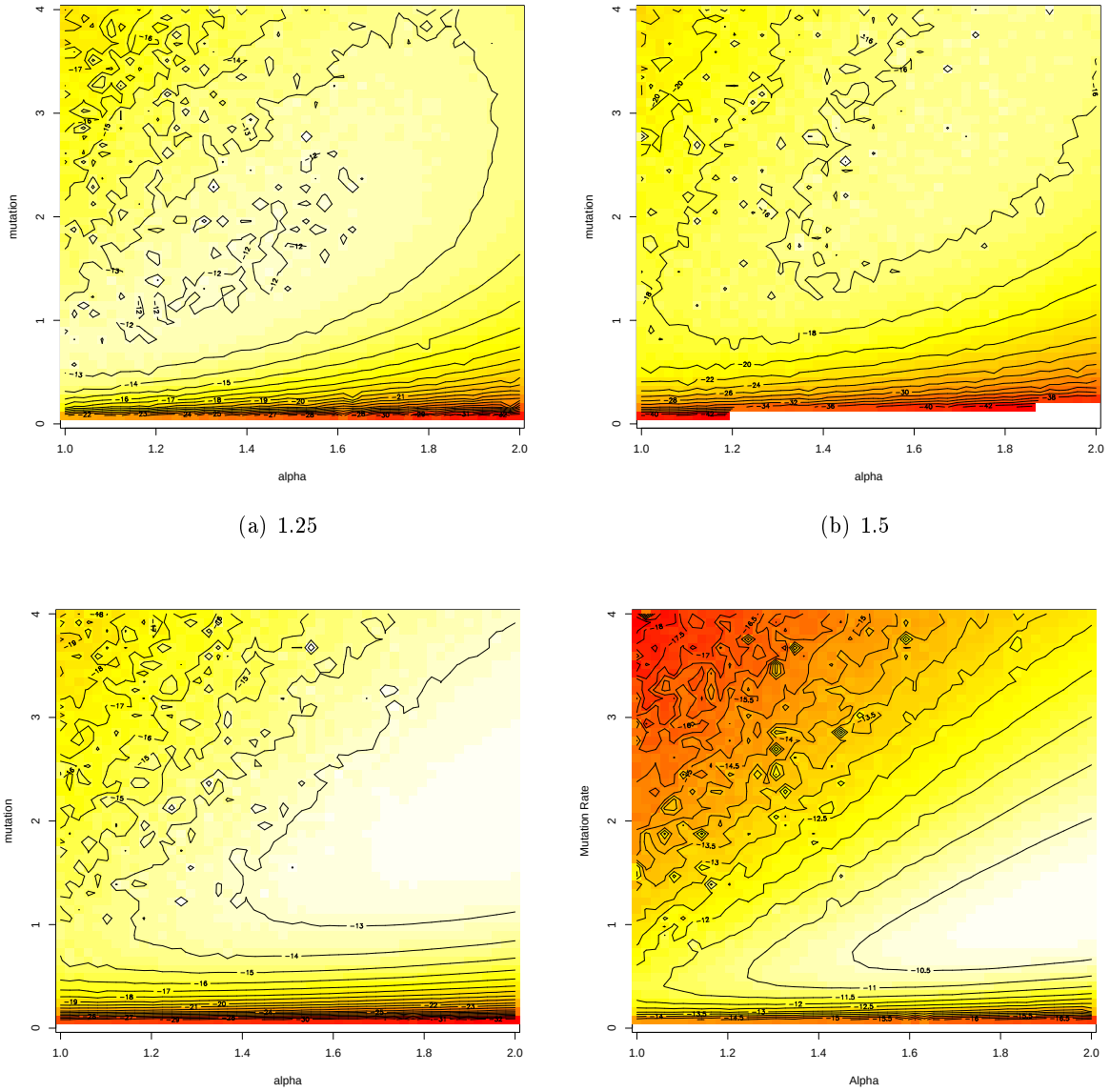
7 Discussion

7.1 Relation with existing models and asymptotic results

The results obtained in this paper should be compared with results in the following two papers.

Berestycki, Berestycki and Schweinsberg (2006). In [BBS06], Berestycki, Berestycki and Schweinsberg obtain asymptotic results for the site frequency spectrum in the infinite-sites case and the allele frequency spectrum in the infinite alleles case, if the underlying genealogy is assumed to be driven by a Beta$(2-\alpha, \alpha)$-coalescent, where $\alpha \in (1, 2]$.

Figure 1: Likelihood-surfaces for $\alpha = 1.25, 1.5, 1.75$ and 2 (Kingman case).

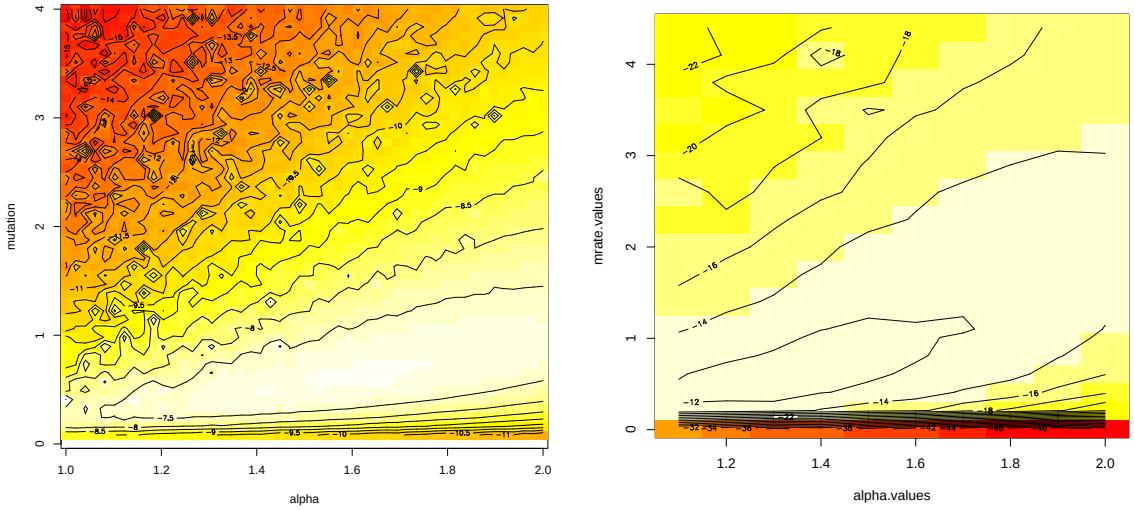


(a) 1.25 (b) 1.5

(c) 1.75 (d) 2

More precisely, in the infinite alleles model, they consider an allelic partition, i.e. a division of the sample into groups of individuals having the same allele at the observed locus. For a sample of size n, one is interested in the number of groups, denoted by $N(n)$, as well as the sizes of the groups. We denote by $N_k(n)$ the number of blocks in the allelic partition of size k. In the infinite sites model, one considers the number $M(n)$, the total number of mutations, and $M_k(n)$, the number of mutations affecting precisely k individuals in the sample (assuming known ancestral type). With this notation, $(N_1(n), \ldots, M_n(n))$ is called the "allele frequency spectrum" and $(M_1(n), \ldots, M_n(n))$ is called the "site

Figure 2: Likelihood-surfaces obtained from Arnason's Atlantic cod data



frequency spectrum".

If we assume that the data is being generated from a Λ-coalescent, and mutations are distributed along the branches at rate θ according to either the infinite alleles or infinite site model, one has the following asymptotic result.

Theorem 7.1 (BBS06). *Assume that Λ has Beta$(2-\alpha, \alpha)$ distribution with $\alpha \in (1,2)$. Let $k \in \mathbb{N}$. Then,*

$$\frac{M_k(n)}{n^{2-\alpha}} \to \alpha(\alpha-1)^2\theta\frac{\Gamma(k+\alpha-2)}{k!}$$

and

$$\frac{N_k(n)}{n^{2-\alpha}} \to \alpha(\alpha-1)^2\theta\frac{\Gamma(k+\alpha-2)}{k!}$$

in probability as $n \to \infty$.

Hence, at least for large sample sizes, the empirical frequency spectrum could be used as a statistic in order validate or overturn the underlying model. However, sample sizes for real data are typically rather small.

Note that in the Kingman case, i.e. $\Lambda = \delta_0$, the famous *Ewens sampling formula* gives the exact distribution of the allele frequency spectrum, namely

$$\mathbb{P}\{N_1(n) = a_1, \ldots, N_n(n) = a_n\} = p(a_1, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{i=1} \frac{\theta^{a_i}}{i^{a_i} a_i!}.$$

However, the case of the Λ-coalescent so far has proved to be inaccessible to explicit solutions up to very few special cases, i.e. $\Lambda = \delta_0$ (Kingman) and $\Lambda = \delta_1$ (star-shaped), see [M06b] (and

Subsection 4.3) for more detail.

Eldon and Wakely (2006). In [EW06], the authors discuss scaling relations between mutation and reproduction in simple population models, where individuals can potentially have very many offspring. The paper treats inference questions based on the number of segregating sites and singletons (i.e. summary statistics) under a rather restrictive coalescent model. In particular, they focus on the Lambda-coalescents presented in (3), where

$$\Lambda(dx) = c_1\delta_0(dx) + c_2\delta_y(dx), \quad c_1, c_2 \geq 0, y \in (0,1],$$

i.e. a genealogy with a Kingman component (the atom in 0) and a reproduction mechanism, in which a single particle can produce a fraction of c_2-many offspring when compared to the total population size. More precisely, they consider a model with fixed population size N, which is a generalisation of a Moran model with fixed inter-generation times in the following sense. At each time step, exactly one individual reproduces (uniformly chosen among the living) and is the parent of $U-1$ new individuals ($U \in \{1,\dots,N\}$). The parent persists, while the offspring replace $U-1$ individuals who die. The other $N-U$ individuals simply persist until the next time step when they might be chosen to die or reproduce.

The offspring mechanism (i.e. the distribution of U) is as follows. Fix $\gamma \geq 0$ and $\psi \in [0,1]$. Then,

$$\mathbb{P}\{U = 2\} = 1 - N^{-\gamma},$$

and

$$\mathbb{P}\{U \approx N\psi\} = N^{-\gamma}.$$

Depending on the choice of ψ, γ, this model leads to a Λ-coalescent with only two atoms, one in 0 (leading to a Kingman-component) and one in $y = \psi$.

Remark. A note on *Type-III survivorship curves.* A survivorship curve in population dynamics is a plot of the life expectancy l_x on a logarithmic scale against the age x. If the mortality does not depend on the actual age, one expects a straight decreasing line. This is called a "type-II-survivorship curve". If the mortality is convex and decreasing, this corresponds to a high mortality early in life and is called "type-III-survivorship curve" ("type-I" now being the obvious notion for the concave case). [EW06] mention "type-III" curves as a situation in which Λ-coalescents might be useful from a modelling perspective. However, we need an additional effect, namely extremely high variation in the reproduction mechanism, which is, strictly speaking, not part of the "type-III" behaviour, in order to arrive at genealogies with multiple collisions. □

7.2 Biological relevance of Beta-coalescents?

Evidence from the Pacific oyster data treated in [EW06] suggests that populations with rather extreme reproductive behaviour should be described by genealogies in which multiple mergers are allowed. However, their proposed Lambda-coalescents seem to be too restrictive. Why should a single individual produce either 2 offspring or *exactly* $\psi*100$ % of the population alive in the next generation and nothing in between? Still, the authors obtain evidence that the simple Kingman case might not be adequate. So an important question is:

What is the right (family of) distribution(s) on the offspring proportions?

To discuss this, it would be useful to find out which reasonable/natural population models actually imply genealogies driven by certain kinds of Λ-coalescents.

A Cannings-model with extremely heavy tails. In [BBC05], it has been pointed out that the Beta$(2-\alpha, \alpha)$-density arises naturally if the approximating models are self-similar continuous state branching processes, time-changed and renormalised to have mass one. It is possible to argue that this renormalising might be rather unnatural. However, there are also reasonable Cannings-models which converge into the Beta-Coalescent genealogy.

Indeed, from the modelling perspective, so-called Cannings models are popular in mathematical population genetics. A Cannings model is a population model with discrete non-overlapping generations and a fixed finite total population size N. At each generation m, the distribution of the offspring of the k-th particle is given by an *exchangeable* random vector

$$(\nu_1, \ldots, \nu_N), \quad N \in \mathbb{N},$$

independent of the generation number m and in between generations. In [S03], the following mechanism is being investigated. Consider a model in which the number of offspring for the individuals are independent (hence no fixed population size), but in each generation only N of the offspring are chosen at random for survival. We assume further that if X is the number of offspring of an individual, then

$$P\{X \geq k\} \sim Ck^{\alpha}$$

for some $\alpha > 0$ and $C > 0$. Schweinsberg shows that, depending on the value of α, the limit may be Kingman's coalescent, in which case each pair of ancestral lines merges at rate one, a coalescent with multiple collisions, or a coalescent with simultaneous multiple collisions. We are most interested in the case that the limit is a coalescent with multiple collisions. It turns out that if $\alpha \in (1, 2)$, the limit is a Λ-coalescent wit Beta$(2-\alpha, \alpha)$-density. Here, the Beta-coalescent appears naturally from a Cannings-model, if we consider the population to consist of the N survivors for each generation. Note the fixed inter-generation times.

Extremely heavy tails vs. Selection?
We have just seen that extremely heavy tails in the reproduction mechanism can account for "shallow" genealogies. However, it is important to point out that also other driving forces in population genetics can account for such behaviour. So far, we have only discussed *neutral* population models, in which there are not beneficial / deleterious mutations. Durrett and Schweinsberg show (see [DS05]) that genealogies with so-called *selective sweeps*, i.e. beneficial mutations, which quickly perpetrate large parts of a population on a different time scale, can be modelled via Λ-coalescents. Our empirical likelihood-surfaces are rather shallow. So how should one be able to check whether a neutral model is adequate, or whether other effects like selection should be taken into account, too? A very recent study on the HIV envelope gene, see [EHP06], for example, claim that the rapid turnover of genetic diversity in HIV-1 is due to strong purifying selection.

In order to judge whether effects due to selection overlap or disguise the possibly heavy tails of the reproduction mechanism, it would be *very* useful to study multi-locus data. Unfortunately, such datasets seem to be hard to obtain – maybe someone should fund a study!

7.3 Outlook

Although the preliminary numerical results presented here seem to be promising, we cannot yet treat large datasets. One line to attack this might be to try to extend Stephens and Donnelly's [SD00] importance sampling techniques or de Iorio and Griffiths [DIG04a] Monte-Carlo techniques to our setting.

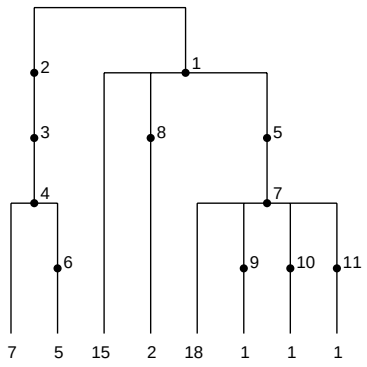
There are other obvious extensions, such as incorporating more general models with varying populations size, distinguish between the effects of selection and an extreme reproductive mechanism etc., which will also be part of the ongoing research project. In order to understand the effects of selection, a study incorporating multi-locus data would be very useful.

8 Appendix

8.1 Underlying datasets and genetrees

The Kingman-case likelihood-surface in Figure 2(d) has been computed using the following data:

(4, 3, 2, 0)
(1, 0)
(7, 5, 1, 0)
(6, 4, 3, 2, 0)
(8, 1, 0)
(9, 7, 5, 1, 0)
(10, 7, 5, 1, 0)
(11, 7, 5, 1, 0)

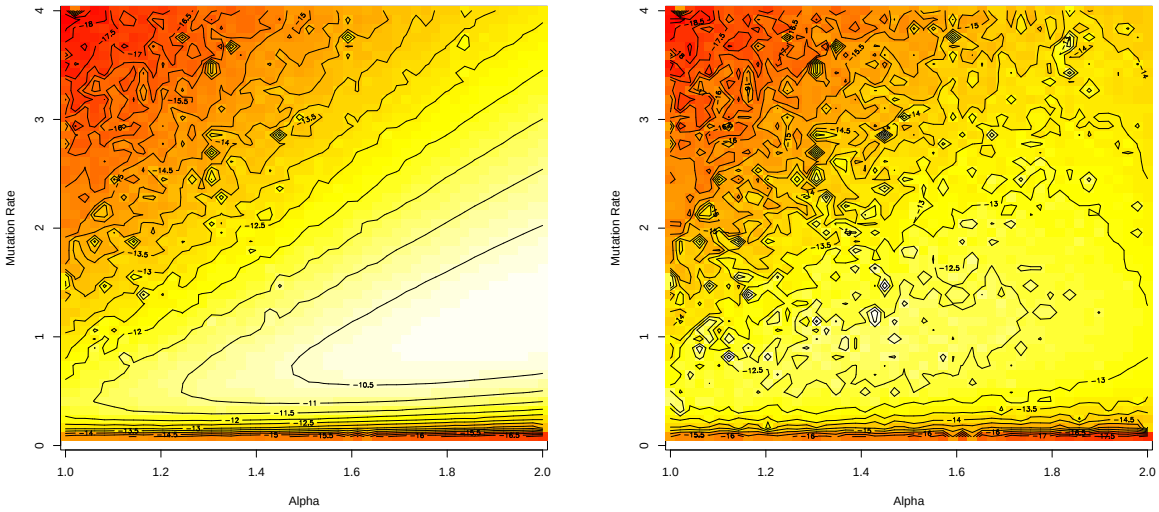


with multiplicities $\mathbf{n} = (7, 15, 18, 5, 2, 1, 1, 1)$, $n = 50$.

Data have been generated with `simbeta`, using the function `sample.ims`, where the underlying parameters were $\alpha = 1.25, n = 50, m = 2, r = 10^7$. The corresponding likelihood-surface and standard-deviation are plotted below.

The likelihood-surface in Figure 2(c) has been computed using the following data:

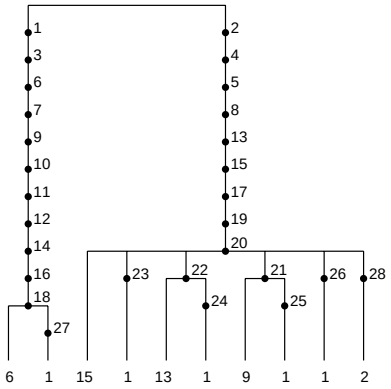
Figure 3: Likelihood and standard deviation for the `a2-m2-n50` tree with 10^7 runs



(a) Likelihood

(b) Deviation

(20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(18, 16, 14, 12, 11, 10, 9, 7, 6, 3, 1, 0)
(21, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(22, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(23, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(24, 22, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(25, 21, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(26, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)
(27, 18, 16, 14, 12, 11, 10, 9, 7, 6, 3, 1, 0)
(28, 20, 19, 17, 15, 13, 8, 5, 4, 2, 0)

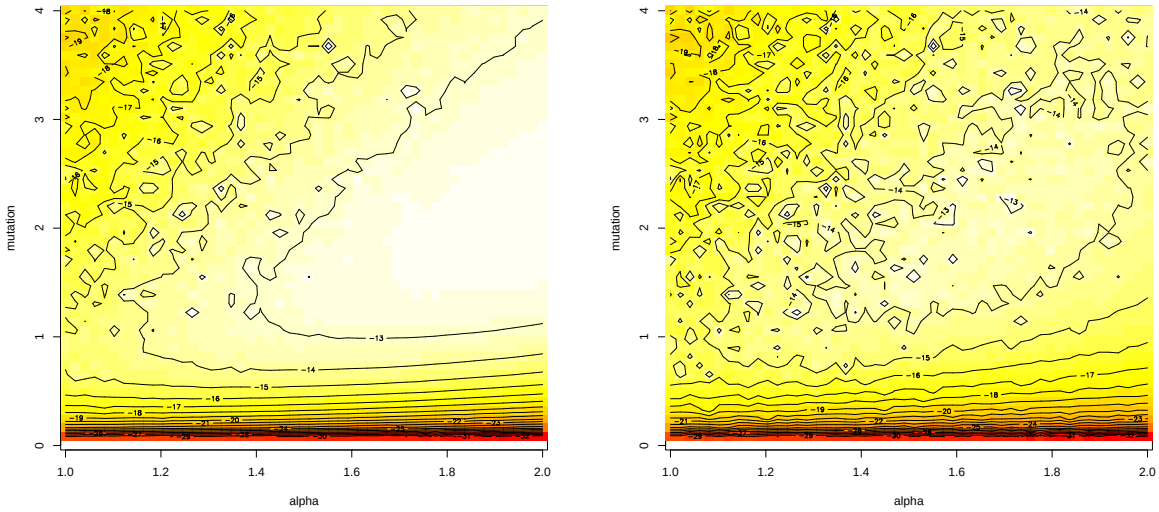


with multiplicities $\mathbf{n} = (15, 6, 9, 13, 1, 1, 1, 1, 1, 2)$, $n = 50$.

Data have been generated with `simbeta`, using the function `sample.ims`, where the underlying parameters were $\alpha = 1.75, n = 50, m = 2, r = 10^7$. The corresponding likelihood-surface and standard-deviation are plotted below.

The likelihood-surface in Figure 2(b) has been computed using the following data:

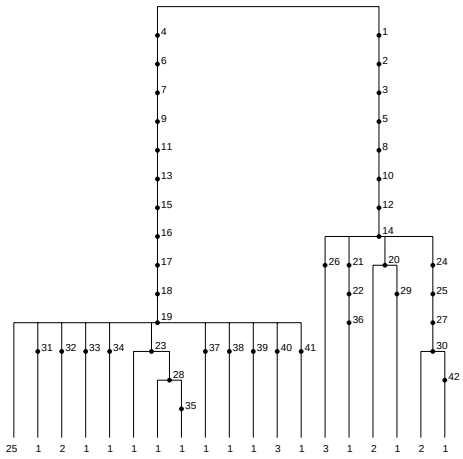
Figure 4: Likelihood and standard deviation for the `a1.75-m2-n50` tree with 10^7 runs



(a) Likelihood

(b) Deviation

(19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(26, 14, 12, 10, 8, 5, 3, 2, 1, 0)
(20, 14, 12, 10, 8, 5, 3, 2, 1, 0)
(36, 22, 21, 14, 12, 10, 8, 5, 3, 2, 1, 0)
(23, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(30, 27, 25, 24, 14, 12, 10, 8, 5, 3, 2, 1, 0)
(28, 23, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(29, 20, 14, 12, 10, 8, 5, 3, ,2 1, 0)
(31, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(32, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(33, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(34, 19, 18, 17, 16, 15, 13, 11 9, 7, 6, 4, 0)
(35, 28, 23, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(37, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(38, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(39, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(40, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(41, 19, 18, 17, 16, 15, 13, 11, 9, 7, 6, 4, 0)
(42, 30, 27, 25, 24, 14, 12, 10, 8, 5, 3, 2, 1, 0)

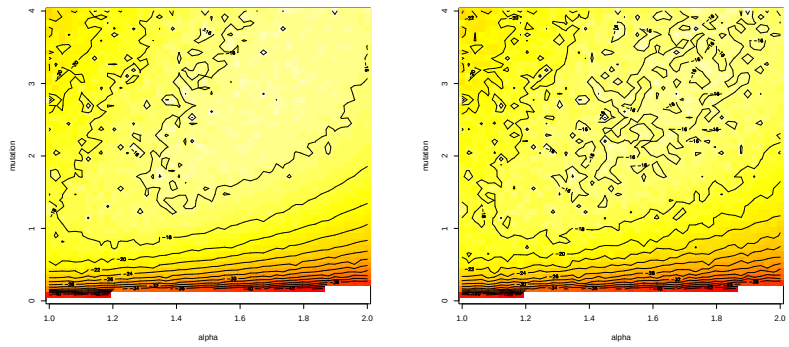


with multiplicities $\mathbf{n} = (25, 3, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 3, 1, 1)$, $n = 50$.

Data have been generated with `simbeta`, using the function `sample.ims`, where the underlying parame-

ters were $\alpha = 1.5, n = 50, m = 2, r = 10^7$. The corresponding likelihood-surface and standard-deviation are plotted below.

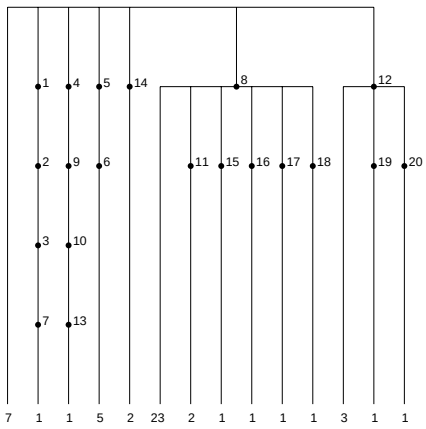
Figure 5: Likelihood and standard deviation for the `a1.5-m2-n50`



(a) Data (b) Deviation

The likelihood-surface in Figure 2(a) has been computed using the following data:

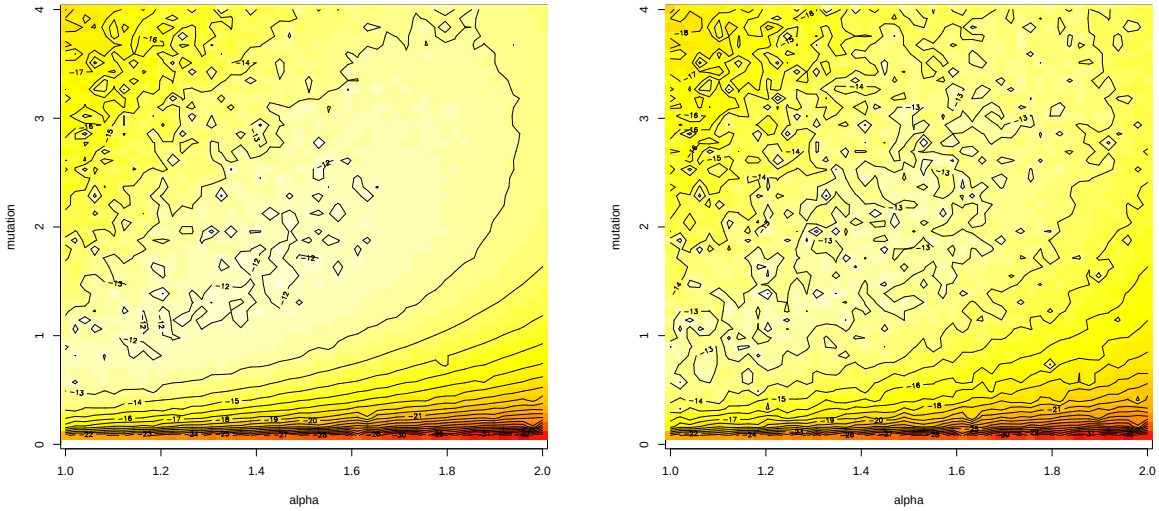
(0)
(7, 3, 2, 1, 0)
(13, 10, 9, 4, 0)
(6, 5, 0)
(8, 0)
(11, 8, 0)
(12, 0)
(14, 0)
(15, 8, 0)
(16, 8, 0)
(17, 8, 0)
(18, 8, 0)
(19, 12, 0)
(20, 12, 0)



with multiplicities $\mathbf{n} = (7, 1, 1, 5, 23, 2, 3, 2, 1, 1, 1, 1, 1, 1)$, $n = 50$.

Data have been generated with `simbeta`, using the function `sample.ims`, where the underlying parameters were $\alpha = 1.25, n = 50, m = 2, r = 10^7$. The corresponding likelihood-surface and standard-deviation are plotted below.

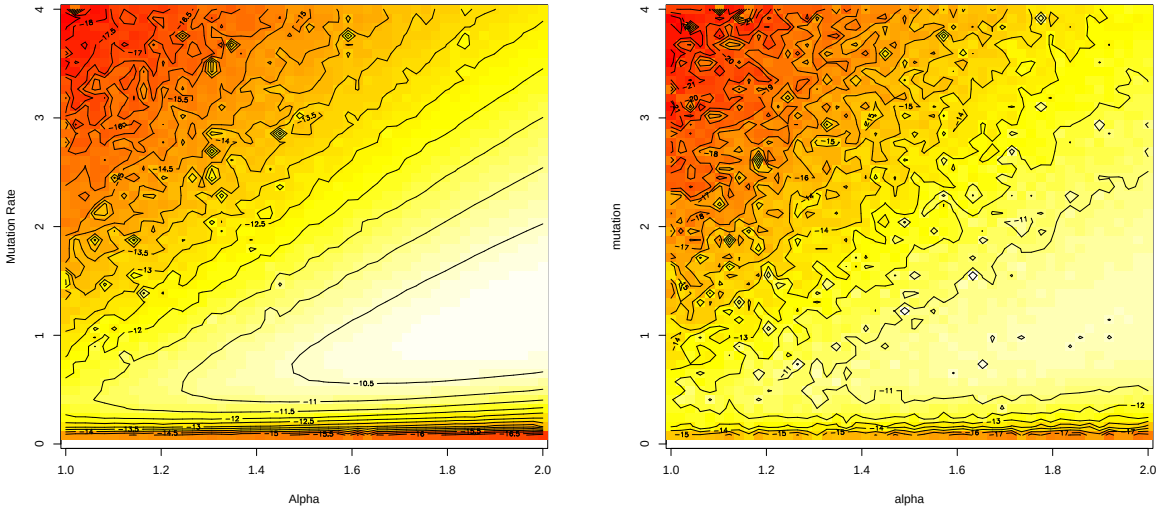
Figure 6: Likelihood and standard deviation for the `a1.25-m2-n50`



(a) Data (b) Deviation

8.2 Convergence of approximate likelihoods

Finally, an important question which needs to be addressed is whether the 10^7 runs used in the Monte-Carlo approximation are sufficient to produce useful results. Compare the likelihood-surfaces obtained after 10^7 runs with the one obtained with 10^3 runs (using the same dataset).



(c) 10^7 (d) 10^3

The improvement is significant. The likelihood-surface looks nearly differentiable around the maximum,

but still rough at the edges.

8.3 Generating samples in the finite and infinite alleles case

Let E be a countable (possibly finite) 'type space', $P = (P_{xy})$ a(n irreducible) stochastic matrix on E with unique equilibrium μ, $r \geq 0$. On each lineage, mutations occur at rate r, given a mutation takes place, the current type is changed according to a P-step.

Type distribution in n-sample arises as follows: genealogy is Λ-coalescent. Given genealogy, give MRCA type according to μ, then run tree-indexed continuous-time MC with generator $r(P - I)$. We use the inherent symmetries and only record $\mathbf{n} = (n_x)_{x \in E} \in \mathbb{Z}_+^E$ (with $\#\mathbf{n} := \sum_{x \in E} n_x = n$), i.e. the frequencies of the types, but not the order of the sample. We write $\tilde{q}_k^{(n)} := -\tilde{q}_{kk}^{(n)}$ for the total jump rate of $\tilde{Y}^{(n)}$ in state k. We assume $n \geq 3$ and obtain the following algorithm.

Algorithm 1.

(i) Draw K according to $\mathcal{L}(\tilde{Y}_0^{(n)})$, i.e. $\Pr(K = k) = g(n,k)q_{k1}$.
Begin with $\eta = K\delta_X$, where $X \sim \mu$.

(ii) Draw $U \sim \mathrm{Unif}([0,1])$.

If $U \leq \frac{kr}{kr + \tilde{q}_k^{(n)}}$:

Replace one of the present types by a P-step from it, i.e. replace $\eta := \eta - \delta_x + \delta_y$ with probability $\frac{\eta_x}{\#\eta} P_{xy}$ (for $x \neq y$).

If $U > \frac{kr}{kr + \tilde{q}_k^{(n)}}$:

If $\#\eta = n$: Stop.

Otherwise, pick $J \in \{k+1, \ldots, n\}$ with $\Pr(J = j) = \tilde{q}_{\#\eta,j}^{(n)} / \tilde{q}_{\#\eta}^{(n)}$.
Choose one of the present types (according to their present frequency), and add $J - \#\eta$ copies of this type, i.e. replace $\eta := \eta + (J - \#\eta)\delta_x$ with probability $\frac{\eta_x}{\#\eta}$.

Repeat.

Note: Ordered sample can (in principle) be obtained from a realization of η by random reordering (assuming $E = \{1, \ldots, d\}$): pick uniformly one of the $\binom{\#\eta}{\eta_1 \ldots \eta_d}$ possible reorderings.

Remark (Infinitely many sites).
Can be done analogously, one has to adapt the 'mutation step' accordingly. See e.g. function `sample.ims` in file `simbeta0.1.R` for an implementation in R. □

In the case of parent-independent mutation, i.e. $P_{ij} = P_j$ for all i, j, it is possible to simulate "backwards in time". Indeed, in order to simulate a sample one follows lineages backwards. "Active" ancestral lineages are lost either by (possibly multiple) coalescence or when hitting their 'defining' mutation.

Fix n, the required sample size. Along the way, we need ξ, a $\mathbb{Z}_+^E$-valued variable (variable in the sense of computer programming), and ζ, a variable with values in $\cup_{j=1}^n \mathbb{N}^j$.

$|\zeta|$ is the current number of 'active lineages', $\zeta(i)$ records how many leaves are presently subtended to i-th lineage (we think of an arbitrary ordering). ξ records the types already 'generated' by now inactive lineages.

$\xi := \mathbf{0}$, $\zeta := (1, 1, \ldots, 1) \in \mathbb{N}^n$.

Algorithm 2.

While there are $\ell = |\zeta| > 1$ lineages: Draw U

if $U \leq \frac{\ell r}{\ell r + q_\ell}$:

Pick $L \sim \mathrm{Unif}(\{1, \ldots, \ell\})$, inactivate L-th lineage, i.e.
$\xi := \xi + \zeta(L)\delta_X$ $(X \sim P_\cdot)$, $\zeta := \big(\zeta(1), \ldots, \zeta(L-1), \zeta(L+1), \ldots, \zeta(\ell)\big)$

else (i.e. when $U > \frac{\ell r}{\ell r + q_\ell}$):

Pick J according to $\Pr(J = j) = \frac{q_{\ell j}}{q_\ell}$.
Merge $\ell - J + 1$ randomly chosen lineages to one, i.e.
draw $S \subset \{1, \ldots, \ell\}$ with $|S| = \ell - J + 1$, then pick $L' \in S$ uniformly.
Put $\zeta(L') := \sum_{i \in S} \zeta(i)$,
then remove entries in $S \setminus \{L'\}$ from ζ.

Finally, when $|\zeta| = 1$, put $\xi := \xi + \zeta(1)\delta_X$ $(X \sim P_\cdot)$.

Remark (Infinitely many alleles).
The same algorithm can be used to generate a sample from the family size spectrum in the infinitely many alleles model. Instead of ξ, we use a variable $\bar{\xi}$ with values in $\mathbb{Z}_+^n$ (idea: finally, $\bar{\xi}(k) =$# families with k members. Just before removing lineage L in the algorithm above, increase $\bar{\xi}(\zeta(L))$ by one. □

References

[A04] Árnason, E.: Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy, *Genetics* **166**, 1871–1885 (2004).

[BBS05] Berestycki, N.; Berestycki, J.; Schweinsberg, J.: Small time behaviour of Beta-coalescents. Preprint, (2005).

[BBS06] Berestycki, N.; Berestycki, J.; Schweinsberg, J.: Beta-coalescents and continuous stable random trees. *to appear in: Ann. Probab.*, (2006).

[BLG03] Bertoin, J.; Le Gall, J.-F.: Bertoin, J.; Le Gall, J.-F.: Stochastic flows associated to coalescent processes. Probab. Theory Related Fields 126 (2003), no. 2, 261–288.

[BLG05] Bertoin, J.; Le Gall, J.-F.: Stochastic flows associated to coalescent processes. II. Stochastic differential equations. Ann. Inst. H. Poincaré Probab. Statist. 41, no. 3, 307–333, (2005)

[BLG06] Bertoin, J.; Le Gall, J.-F.: Stochastic flows associated to coalescent processes III: Infinite population limits. Illinois J. Math., to appear.

[BBC05] Birkner, M.; Blath, J.; Capaldo, M.; Etheridge, A.; Möhle, M.; Schweinsberg, J.; Wakolbinger, A.: Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.* **10**, 303–325, (2005).

[B06] Birkner, M.: Using Beta-Genetree, technical report, 2007.

[BBB94] Boom, J. D. G.; Boulding, E. G.; Beckenbach, A. T.: Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. *Can. J. Fish. Aquat. Sci.* 51:1608Ű1614, (1994).

[C74] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Prob.* **6**, 260–290, (1974).

[C75] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Prob.* **7**, 264–282, (1975).

[D93] Dawson, D.: Lecture Notes, Ecole d'Eté de Probabilités de Saint-Flour XXI, Berlin, Springer, (1993).

[DIG04a] De Iorio, M. and Griffiths, R. C.: Importance sampling on coalescent histories I. *Adv. Appl. Prob.* **36**, 417-433, (2004).

[DIG04b] De Iorio, M. and Griffiths, R. C.: Importance sampling on coalescent histories II: Subdivided population models. *Adv. Appl. Prob.* **36**, 434-454, (2004).

[DK96] Donnelly, P.; Kurtz, T.: A countable representation of the Fleming-Viot measure-valued diffusion. Ann. Probab. 24, no. 2, 698–742, (1996)

[DK99] Donnelly, P.; Kurtz, T.: Particle representations for measure-valued population models. Ann. Probab. 27, no. 1, 166–205, (1999)

[DS05] Durrett, R.; Schweinsberg, J.: A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Proc. Appl.* **115**, 1628–1657 (2005)

[E79] Ewens, W.J.: Mathematical population genetics. Berlin: Springer-Verlag, (1979).

[EG87] Ethier, S.; Griffiths, R.C.: The infinitely-many-sites model as a measure-valued diffusion. *Ann. Probab.* **15**, no. 2, 515–545, (1987).

[EK86] Ethier, S.; Kurtz, T.: Markov Processes: Characterization and Convergence. Wiley, (1986).

[EHP06] Edwards, C. T. T.; Holmes, E. C.; Pybus, O. G.; Wilson, D. J.; Viscidi, R. P.; Abrams, E. J.; Phillips, R. E.; Drummond, A. J.: Evolution of Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection. *Genetics* **174**: 1441–1453, (2006).

[EW06] Eldon, B.; Wakeley, J.: Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* "in press", (2006).

[FKY99] Felsenstein, J., Kuhner, M. K., Yamato, J. and Beerli, P.: Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. IMS Lecture Notes - Monograph Series, **33**, 163–185, (1999).

[G89] Griffiths, R. C.: Genealogical-tree probabilities in the infinitely-many-site model. J. Math. Biol. 27 (1989), no. 6, 667–680.

[GM96] Griffiths, R.C. and Majoram, P.: Ancestral Inference from samples of DNA sequences with recombination, *J. of Comp. Biol.* **3**, 479–502, (1996).

[GT94a] Griffiths, R.C. and Tavaré, S.: Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* **46**, 131–159, (1994).

[GT94b] Griffiths, R.C. and Tavaré, S.: Ancestral Inference in population genetics. *Statistical Science* **9**, 307–319, (1994).

[GT94c] Griffiths, R.C. and Tavaré, S.: Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society London*, Series B, **344**, 403–410, (1994).

[GT96a] Griffiths, R.C. and Tavaré, S.: Monte Carlo inference methods in population genetics. Monte Carlo and quasi-Monte Carlo methods. *Math. Comput. Modeling* 23 (1996), no. 8-9, 141–158.

[GT96b] Griffiths, R.C. and Tavaré, S.: Markov chain inference methods in population genetics. *Math. Comput. Modeling* **23**, 8/9, 141-158, (1996).

[GT96c] Griffiths, R.C. and Tavaré, S.: Markov chain inference methods in population genetics. *Math. Comput. Modeling* 23, 8/9, 141-158.

[GT97] Griffiths, R.C. and Tavaré, S.: Computational Methods for the coalescent. Progress in Population Genetics and Human Evolution, 165–182, Springer, (1997).

[GT98] Griffiths, R. C.; Tavaré, S.: The age of a mutation in a general coalescent tree. *Comm. Statist. Stochastic Models* **14**, 273–29, (1998).

[GT99] Griffiths, R. C.; Tavaré, S.: The ages of mutations in gene trees. *Ann. Appl. Probab.* **9**, no. 3, 567–590, (1999).

[K82] Kingman, J. F. C.: The coalescent. *Stoch. Proc. Appl.* **13**, 235–248, (1982).

[KYF95] Kuhner, M. K., Yamato, J. and Felsenstein, J.: Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421–1430, (1995).

[KYF98] Kuhner, M. K., Yamato, J. and Felsenstein, J.: Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429–434, (1998).

[M05] Möhle, M.: Simulation algorithms for integrals of a class of sampling distributions arising in population genetics. *J. Stat. Comp. Simul.* 75, 731-749 (2005)

[M06a] Möhle, M.: On the number of segregating sites for populations with large family sizes. *J. Appl. Prob.* "in press", (2006).

[M06b] Möhle, M.: On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12**, "in press", (2006).

[M07] Möhle, M.: On a class of non-regenerative sampling distributions. *To appear in: Combin. Probab. Comput. 16,* (2007)

[MS01] Möhle, M; Sagitov, S.: A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29**, 1547–1562, (2001).

[P99] Pitman, J.: Coalescents with multiple collisions. *Ann. Probab.* **27** (4), 1870-1902, (1999).

[R87] Ripley, B.D.: Stochastic Simulation. Wiley, (1987)

[RW87] Rogers, L.C.G.; Williams, D.: Diffusions, Markov Processes and Martingales. Vol. 1, 2nd ed., Wiley, (1994)

[S99] Sagitov, S.: The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** (4) 1116–1125, (1999).

[S00] Schweinsberg, J.: A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab.* **5**, 1–11, (2000).

[S00b] Schweinsberg, J.: Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5**, Paper no. 12, 50 pp., (2000).

[S03] Schweinsberg, J.: Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch. Proc. Appl.* **106**, 107–139, (2003).

[SD00] Stephens, M; Donnelly, P.: Inference in molecular population genetics. *J. Roy. Stat. Soc.* B. **62**, 605–655 (2000).

[T01] Tavaré, S.: Ancestral Inference in Population Genetics. Springer Lecture Notes **1837**, 2001.