rijksuniversiteit groningen
faculteit Wiskunde en Natuurwetenschappen

Modelling macro-evolution: how long does speciation take?

Bachelor thesis Mathematics

July 2014

Student: Gerard Hekkelman

First Supervisor Mathematics: Prof. Dr. E.C. Wit

Second Supervisor Mathematics : Dr. W.P. Krijnen

Supervisor Centre for Evological and Evolutionary Studies: Prof. Dr. R.S. Etienne

Abstract

In this paper, we study the constant birth-death process and the protracted birth-death process which describes macro-evolution in biology. We derive some properties of these models, and analytically derive the likelihood of the constant birth-death model. For the protracted birth-death model, we derive an appromation of the likelihood using the concept of Gaussian processes. Using the exact likelihood of the constant birth-death process and the approximated likelihood of the protracted birth-death model, we infer a given phylogenetic tree and discuss the results.

Contents

Introduction

A phylogenetic tree (that is, the ancestry, predigree, genealogy) of a group of species represents the evolutionary relatedness between the species, and their common ancestry, for example between human and chimpansee. Currently, DNA data are routinely used to infer the phylogenetic tree. An example of an actual phylogenetic tree is given in figure 1.

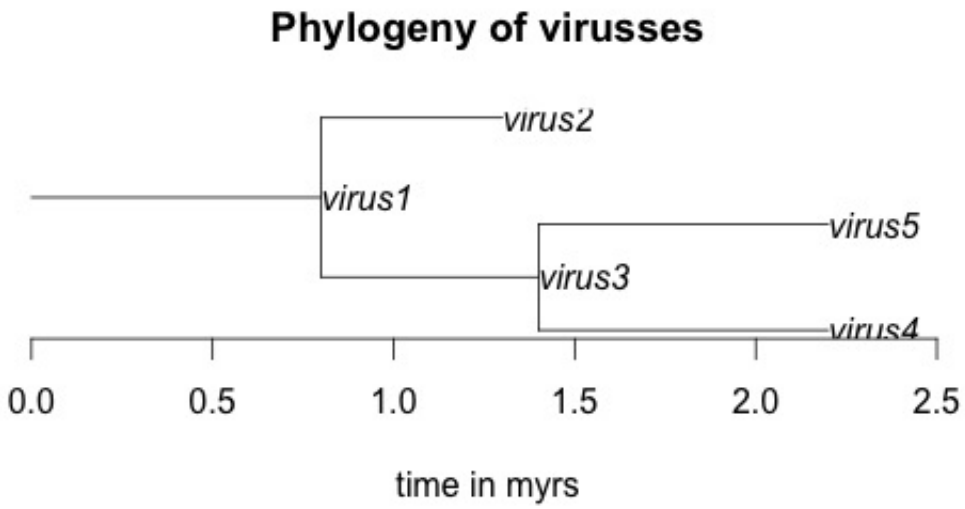

Figure 1: An example of a phylogenetic tree.

In this example, we can see that the group of virusses is represented in a tree form. Every branch represents a species, and each node represents a common ancestor for two species in the tree. For example, "virus2" and "virus5" have one common ancestor, which is "virus1". For these kind of phylogenies, we are interested in estimating the parameters which have their influence on the evolutionary process. However, the phylogenetic trees used as data only contain the genetic information of extant species. By this, we don't have information of the extant species which existed in the past, but are extinct today. Consider figure 1 again, the data we observe is given in figure 2.

These trees are called reconstructed phylogenies. We infere these phylogenies to obtain the parameters which cause the growth of a phylogeny. Sophisticated software, often using Bayesian approaches, has been developed for this. One component of this software is the probability of the phylogenetic tree given a model of species diversification (speciation and extinction). There is a lot of debate about the proper model of species diversification.

One example is a model which assumes that speciation and extinction are instantaneous events,

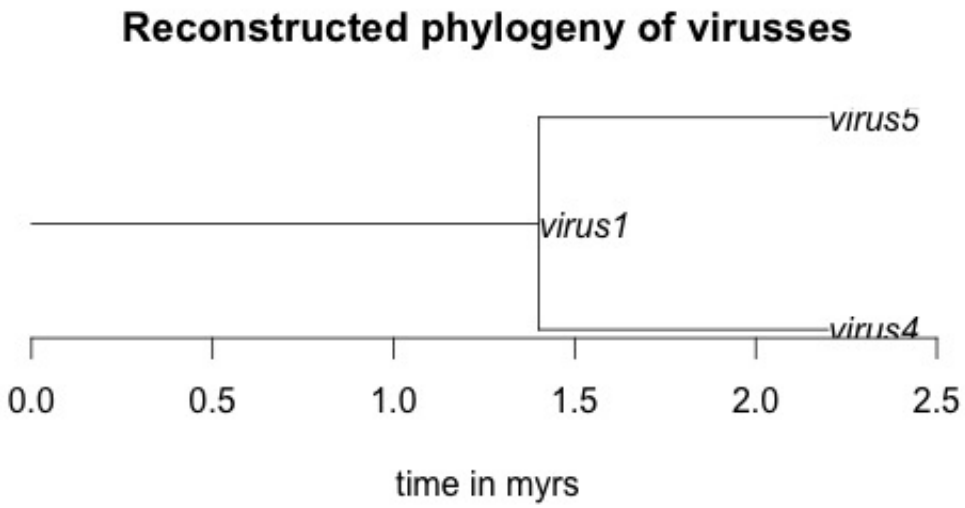

Figure 2: An example of a phylogenetic tree.

with occur with a constant rate through time. This model is called the constant birth-death model. Another example is a model that assumes that speciation takes time rather than being instantaneous (which all other models assume). This model, called the protracted birth-death model, allows us to estimate how long a speciation process takes to complete. However, the current mathematical representation of the protracted speciation model seems incoherent.

In this paper we first represent the constant birth-death model. We give the properties of the constant birth-death model and make simulation studies. Thereafter, we make inferences of this reconstructed phylogeny and obtain a maximum likelihood estimate of the diversification rate. Secondly, we introduce the protracted model and search for an approximation of the representation of the protracted model and a way to infer its parameters from a given phylogenetic tree.

Chapter 1

Birth-death models

The birth-death models are mathematical models of the dynamical process of speciation and extinction, which are used for informative thinking about macroevolutionary patterns. The pure birth-model and the pure death model are submodels of the birth-death models. We can simply model the growth of a clade through speciation to deduce the rate of speciation, discovering any irregularities. Also less obvious questions can be answered using the birth-death models.

The use of the birth-death models started back in 1924 to the work of the statistician Yule (Yule 1924), who modelled a clade growing according the pure birth process. Here, extinction does not occur.

The modern usage of the birth-death models started with the work of Raup and colleagues in 1973 (Raup et al 1973), when computers were starting to become a readily used tool (Nee 2006). They modelled a scenario where the probabilities of a species is either speciating or coming to extinct are equal, and the clade size was roughly constant. They discovered that this purely random process could produce trends and patterns resembling those in the fossil records.

We first investigate several models of macro-evolution, where the constant birth-death model and the protracted birth-death model are the most important ones in this paper.

1.1 Pure Birth Model

In this model, we assume that each species in a clade has a constant rate λ of producing a new species at any point in time and that extinction never occurs. We assume that the duration of speciation T_i is exponentially distributed for each species in the clade:

$$T_b(i) \sim \exp(\lambda), \quad \forall i = 1, 2, \ldots \tag{1.1}$$

Let N_g denote the number of species in the clade, we are interested in the probability density function of the number of species at a time t. The probability that at time t there are N_g species is given by the following stochastic differential equation (Yule 1924):

$$\frac{d}{dt}\Pr(N_g; t) = \lambda(N_g - 1)\Pr(N_g - 1; t) - \lambda N_g \Pr(N_g; t) \tag{1.2}$$

This equation is known as the master equation of the *pure birth model*, or *Yule process* named after its discoverer George Udny Yule. This equation will be solved analytically later in this paper. Instead of looking at the probability density function, we can also focus on the expected number of species in the clade at a time t:

Figure 1.1: George Udny Yule (1871-1951) [35].

Let $\mathrm{E}[N(t)]$ denote the expected number of species in the clade at time t and set $\mathrm{E}[N_0] = \mathrm{E}[N(0)]$ as the expected initial number of species. The expected number of species $\mathrm{E}[N(t)]$ grows exponentially over time with a rate λ, hence we have the ODE:

$$\frac{d}{dt}\mathrm{E}[N(t)] = \lambda \mathrm{E}[N(t)], \quad \mathrm{E}[N(0)] = \mathrm{E}[N_0] \tag{1.3}$$

The solution is straightforward:

$$\mathrm{E}[N(t)] = \mathrm{E}[N_0]\, e^{\lambda t} \tag{1.4}$$

As a consequence, a plot of $\log N(t)$ against the time should be linear and the slope of this plot provides an estimate of b, the per-capita birth rate.

1.2 Pure Death model

In this model, we assume that each species in a clade has a constant rate μ of going extinct in time t and that speciation never occurs. We assume that the duration of extinction T_i is exponentially distributed for each species in the clade:

$$T_d(i) \sim \exp(\mu), \quad \forall i = 1, 2, \ldots \tag{1.5}$$

Let N_g denote the number of species, we are interested in the probability density function of the number of species at a time t. The probability that at time t there are N_g species is similiar to the probability density function given in equation 1.2, and it is given implicitely by the following stochastic differential equation:

$$\frac{d}{dt}\Pr(N_g; t) = \mu(N_g + 1)\Pr(N_g + 1; t) - \mu N_g \Pr(N_g; t) \tag{1.6}$$

With initial condition:

$$\Pr(N_g = N_g(0); t = 0) = 1 \tag{1.7}$$

The equation 1.6 is the master equation of the *pure death model*. Again, this equation can be solved analytically. First, we look at the expected number of species in the clade at a time t:

Let $\mathrm{E}[N(t)]$ denote the expected number of species in the clade at time t and set $\mathrm{E}[N_0] = \mathrm{E}[N(0)]$ as the expected initial number of species. The expected number of species that survive $N(t)$ decays exponentially over time with a rate μ, hence we have the following ODE:

$$\frac{d}{dt}\mathrm{E}[N(t)] = -\mu\mathrm{E}[N(t)], \;\; \mathrm{E}[N(0)] = \mathrm{E}[N_0] \tag{1.8}$$

The solution is straightforward:

$$\mathrm{E}[N(t)] = \mathrm{E}[N_0]\, e^{-\mu t} \tag{1.9}$$

1.3 Birth-death model

In this model, we assume that each species in a clade speciate with a constant rate λ and that extinction occurs with a constant rate μ in time. We assume that the duration of speciation T_b and the duration of extinction T_d are exponetially distributed for each species:

$$T_b(i) \sim \exp(\lambda), \;\; \forall i = 1, 2, \ldots \tag{1.10}$$
$$T_d(i) \sim \exp(\mu), \;\; \forall i = 1, 2, \ldots \tag{1.11}$$

Let N_g denote the number of species, we are interested in the probability density function of the number of species at a time t. The probability that at time t there are N_g species is obtained by Kendall in 1948 (Kendall 1948):

$$\begin{aligned}\frac{d}{dt}\Pr(N_g; t) = {} & \lambda(N_g - 1)\Pr(N_g - 1; t) + \mu(N_g + 1)\Pr(N_g + 1; t) \\ & - (\lambda + \mu)N_g \Pr(N_g; t)\end{aligned} \tag{1.12}$$

With initial condition:

$$\Pr(N_g = N_g(0); t = 0) = 1 \tag{1.13}$$

This equation will be solved analytically later in this paper. Again, we look first to the expected number of species in the clade at a time t. We define the net rate of diversification as $r := \lambda - \mu$. The clades grow exponentially at a rate r, we therefore have the following ODE:

$$\frac{d}{dt}\mathrm{E}[N(t)] = (\lambda - \mu)\mathrm{E}[N(t)], \;\; \mathrm{E}[N(0)] = \mathrm{E}[N_0] \tag{1.14}$$

The solution is straight forward:

$$\mathrm{E}[N(t)] = \mathrm{E}[N_0]\, e^{(\lambda-\mu)t} \tag{1.15}$$

The semilog representation of the growth of a molecular phylogeny was expected to be linear with a constant slope r, however when the extinction rate is nonzero it is not. But, over much over the history the plot is expected to be linear with slope r (Nee 2006). Molecular phylogenies are completely based on data of extant species, and species that originated more recently in the past had less time to go extinct. Therefore, the extinction rate μ gets smaller as we approach the present. So as we approach the present, the slope of the semilog presentation is expected to

Figure 1.2: David George Kendall (1918-2007) [12].

increase and assymptotically approach λ. By this, we can estimate λ and μ seperately on intuition (Nee 2006). Notice that in molecular phylogenies, we can only see the history of extant species. Species which have become extinct are not represented in these phylogenies.

We have two surprises from the birth-death models (Nee 2006):

- We can estimate speciation and extinction rates from molecular phylogenies, even though they do not contain information from extant species.
- We can estimate per-species speciation and extinction rates even from fossil data that are not resolved to a level below that of the genus.

1.4 Protracted Pure Birth Model

In this model we make speciation a protracted process. That is, that each species in a clade produce new species with a rate λ_1, but these species are not fully completed. These incipient species become good species with a rate λ_2, and give rise to new incipient species by rate λ_3. This means that speciation in the protracted birth model takes one extra step with respect to the regular birth model.

To be more precise: in this model, we assume that each species in a clade speciate with a constant rate λ_1 and that speciation is completed by a constant rate λ_2. We assume that the duration of speciation T_b and the duration of completion T_c are exponetially distributed for each species:

$$T_b(i) \sim \exp(\lambda_1), \quad \forall i = 1, 2, \ldots \tag{1.16}$$
$$T_c(i) \sim \exp(\lambda_2), \quad \forall i = 1, 2, \ldots \tag{1.17}$$

Let N_g denote the number of good species (the completed species), and let N_i denote the incipient species (the incomplete species). Now, we are interested in the probability density function of the number of good species N_g and incipient species N_i at a time t. The probability that at time t there are N_g good species and N_i incipient species is implicitely obtained (Etienne, Rosindell 2012):

$$\begin{aligned}\frac{d}{dt}\Pr(N_g, N_i; t) &= \lambda_1 N_g \Pr(N_g, N_i - 1; t) + \lambda_3 (N_i - 1)\Pr(N_g, N_i - 1; t) \\ &+ \lambda_2 (N_i + 1)\Pr(N_g - 1, N_i + 1; t) - (\lambda_1 N_g + (\lambda_2 + \lambda_3) N_i)\Pr(N_g, N_i; t) \end{aligned} \tag{1.18}$$

With initial condition:

$$\Pr(N_g = N_g(0), N_i = 0; t = 0) = 1 \tag{1.19}$$

This model cannot be solved analytically, so we investigate the expected number of good and incipient species first. We obtain the following system of ODE's:

$$\frac{d}{dt}\begin{bmatrix}\mathrm{E}[N_g; t] \\ \mathrm{E}[N_i; t]\end{bmatrix} = \begin{bmatrix}0 & \lambda_2 \\ \lambda_1 & \lambda_3 - \lambda_2\end{bmatrix}\begin{bmatrix}\mathrm{E}[N_g; t] \\ \mathrm{E}[N_i; t]\end{bmatrix} \tag{1.20}$$

With initial condition:

$$\begin{bmatrix}\mathrm{E}[N_g; t = 0] \\ \mathrm{E}[N_i; t = 0]\end{bmatrix} = \begin{bmatrix}N_g(0) \\ 0\end{bmatrix} \tag{1.21}$$

The general solution of this system is given in (Etienne and Rosindell 2012). The case that $\lambda_1 = \lambda_3$ is much easier to solve, the solution of this system in the case is straight forward the solution of a linear system ODE:

$$\mathrm{E}[N_g; t] = \frac{N_g(0)}{1 + \frac{\lambda_1}{\lambda_2}}\exp(\lambda_1 t) + \frac{N_g(0)}{1 + \frac{\lambda_2}{\lambda_1}}\exp(-\lambda_2 t) \tag{1.22}$$

$$\mathrm{E}[N_i; t] = \frac{N_g(0)}{1 + \frac{\lambda_2}{\lambda_1}}(\exp(\lambda_1 t) - \exp(\lambda_2 t)) \tag{1.23}$$

1.5 Protracted Birth-Death Model

If we make speciation in the constant birth-death model protracted, analogously to the protracted birth model, we assume that species give birth to incipient species by a rate λ_1, incipient species reach completion by a rate λ_2, give birth to new incipient species by a rate λ_3 and extinction occurs for both species with a rate μ.[1]

$$\begin{aligned}\frac{d}{dt}\Pr(N_g, N_i; t) &= \lambda_1 N_g \Pr(N_g, N_i - 1; t) + \lambda_1 (N_i - 1)\Pr(N_g, N_i - 1; t) \\ &+ \lambda_2 (N_i + 1)\Pr(N_g - 1, N_i + 1; t) + \mu(N_g + 1)\Pr(N_g + 1, N_i; t) \\ &+ \mu(N_i + 1)\Pr(N_g, N_i + 1; t) \\ &((\lambda_1 + \mu) N_g + (\lambda_1 + \lambda_2 + \mu) N_i)\Pr(N_g, N_i; t) \end{aligned} \tag{1.24}$$

In this paper, we focus on obtaining an approximation of the the probability density function implicitely given in equation 1.24.

There are many other models describing evolutionary processes, However, they lie beyond our scope in this paper. A review of these models, and some biological properties are given in appendix A.1.

[1] Note that in the paper of Etienne and Rosindell, they assume that extinction rates may differ for good species with rate μ_1 and incipient species with rate μ_2. Here we assume $\mu = \mu_1 = \mu_2$.

Chapter 2

Inference

2.1 Inferences of Constant Birth-Death Models

Molecular systematics produces phylogenies that may have a temporal dimension, thus containing information about the tempo of the clade's evolution as well as the relationships among taxa (Nee 2001). We are particulary interested in extracting this information. We are able to study the rates of diversification in the clade. Bladwin and Sanderson (1998) used the simple Yule process to study the rate of diversification . The Yule Process is a fairly simple process, but we can use nice statistical approaches for obtaining results under this model.

For inference, we first must distinguish between actual and reconstructed phylogenies. We note four points (Nee et al. 1994):

- Both phylogenies have the same number of taxa at the present day
- At any point in the past, the number of lineages in the reconstructed phylogeny is less or equal than the number of lineages of the actual phylogeny.
- The number of lineages in the reconstructed phylogeny cannot decrease towards the present, this can happen in the actual phylogeny.
- The reconstructed phylogeny provides timings for when each pair of species has last shared a common ancestor and commences at that point in the past when all present-day species shared their most recent common ancestor.

For making the decision whether we have to investigate the causes of an apparently high diversification should be invesigated, it is desirable to whether or not the diversification really is remarkable in reference to some null model.

In a broad class of models, the number of progreny lineages of any particular ancestral lineages in a reconstructed phylogeny has a geometric distribution (Nee et al. 1994). The derivation of this will be given later in section 2.3.1. We are interested in how many lineages each ancestral lineage gives rise to and if the distribution of progeny lineages fits our geometric expectation.

Under the constant rate birth-death model, there are interesting properties for both the actual and the reconstructed phylogeny:

- If there was no extinction, the curves representing the number of lineages through time for the actual phylogeny and the reconstructed phylogeny are the same.
- The push of the past is observed as an apparently higher diversification rate at the beginning of the growth of the actual phylogeny. It is a result from the fact that we consider clades which have survived to the present day, and these are the ones which got a "flying start" most of the times.

- The pull of the present is the observed increase in the diversification rate in the recent past of the reconstructed phylogeny. It is a result of the fact that lineages who arose more recent in the past have had less time to go extinct.
- The slope of both phylogenies is $\lambda - \mu$ most of the time.
- The slope of the reconstructed phylogeny assymptotically approaches the birth rate.
- The pull of the present and the pull of the past increases as the fraction μ/λ approaches one.

Using reconstructed phylogenies, it is tempting to take a pure birth process intuitively, so $d = 0$, as a model for the data since there are no extinct species in the phylogeny. However, using likelihood plots in (Nee et al. 1994) show that we cannot exclude the possibility that it is actually nonzero. Using a likelihood surface apporach, we check in section 3.1.3 if this is indeed the case.

When only a sample of a clade has been used, so not the whole clade, creates an effect of slowdown in diversification. This effect becomes more pronounced the smaller the sample is (Nee et al. 1994). Lineages that have arisen in the recent past are likely to have fewer progeny than lineages which arose in a more distant past. So, they are less likely to have any progeny represented in the sample which causes the oberserved diversification slowdown.

2.2 Inference of the Yule Process

We first make the following assumptions, before we make inferences with the Yule Process:

- From the time of its origin with two lineages time t ago, the tree has grown according to a Yule Process with parameter λ
- the age of the clade, t, is a fixed variable.

Note that in this way, at each point in time each lineage has the same probability of giving birth to a new lineage. This probability is proportional to the parameter λ, controlling the rate of growth of the tree. Note also that the clade size N is not predetermined, it is a random variable.

We assume that our data consists if the length of time from each node to the present day , denoted by x_i. For example, a clade with four lineages at the end has 3 nodes. We let x_2 denote the time between the present and the last ancestor, which is also the age of this monophyletic clade. So we have $x_2 = t$. Moreover, we consider in general the following quantities:

$$s_r = \sum_{i=3}^{n} x_i \tag{2.1}$$

$$s = 2x_2 + \sum_{i=3}^{n} x_i = 2x_2 + s_r \tag{2.2}$$

We immediately that s represents the sum of all branch lengths in the tree, and s_r only represents the sum of the branch lengths, except the two basal branches.

To obtain the probability density function of the data, we have the following:

- We have n branches, from which the 2 base branches have the same length and are therefore the same. We therefore have $(n-1)!$ different permuations for the clade.
- We assume the branch lengths are independent exponential random variables

The probability for a branch i giving birth after a time x_i is:

$$\Pr(X_i \geq x_i) = \int_0^{x_i} \lambda \exp[-\lambda x_i] dx_i = \exp[-\lambda x_i] \tag{2.3}$$

The probability of j lineages in the tree at the time of a birth event is proportional to λj. Hence, for a tree which has 2 birth events after its first node and thus has 3 species at this moment. These events contribute therefore the term $2\lambda * 3\lambda$ tot the likelihood expression. In general: for n species we have $n-1$ birth events. So, we have a contribution of the term $(n-1)!\lambda^{n-2}$ to the likelihood expression.

The n branches $\mathbf{x}$, of which the two basal branches are the same, contribute a term $\exp[-\lambda s]$ to the likelihood by the combined probabilities of equation 2.3. Combining this and the latter, we obtain the likelihood for the data:

$$\Pr(\mathbf{x}, n; \lambda, t) = (n-1)!\lambda^{n-2} e^{-\lambda s} \tag{2.4}$$

Actually, we are only interested in the probability density function of $\mathbf{x}$ only. The probability of n lineages, given λ and t is obtained by the following:

- A clade starting with two species, has given birth to $n-2$ new species.
- Two species did not speciate before time t.
- There occured $n-1$ birth events.

Because we have n lineages and $n-1$ birth events, there are $n-1$ different possibilities of getting n lineages. This contributes a $(n-1)$ term to the likelihood. In the same reasoning as before, we have two branches which did not give birth, contributing the $\exp(-\lambda t)^2 = \exp(-2\lambda t)$ term to the likelihood. Finally, the $n-2$ branches who gave birth, contributed the $(1-\exp[-\lambda t])^{n-2}$ term to the likelihood. Combining this we get the probability of n lineages in the clade:

$$\Pr(n; \lambda, t) = (n-1)e^{-2\lambda t}(1-e^{-\lambda t})^{n-2} \tag{2.5}$$

Where the probability density functions obtained are the same as in (Nee 2001). Thus, the probability of $\mathbf{x}$ given n, λ and t is:

$$\Pr(\mathbf{x}; n, \lambda, t) = \frac{\Pr(\mathbf{x}, n; \lambda, t)}{\Pr(n; \lambda, t)} = \frac{(n-2)!\lambda^{n-2} e^{-\lambda s_r}}{(1-e^{-\lambda t})^{n-2})} \tag{2.6}$$

Remark Observe that:

$$\Pr(\mathbf{x}; n, \lambda, t) = \frac{(n-2)!\lambda^{n-2} e^{-\lambda s_r}}{(1-e^{-\lambda t})^{n-2})} \tag{2.7}$$

$$= (n-2)! \prod_{i=3}^{n} \frac{\lambda e^{-\lambda x_i}}{1-e^{-\lambda t}} \tag{2.8}$$

Therefore, this is also the probability density function of the order statistics of $n-2$ independent and identically distributed random variables, where the random variables are truncated exponentially distributed: i.e. they have the density:

$$\Pr(X_i = x_i; \lambda, t) = \frac{\lambda e^{-\lambda x_i}}{1-e^{-\lambda t}} \tag{2.9}$$

2.2.1 Confidence intervals for λ

Moran's Maximum Likelihood Estimation

Consider the likelihood function given in equation 2.4 again. For maximum likelihood estimation, we need both the outcome of our random variables n and $\mathbf{x}$. Taking the natural logarithm of equation 2.4, we obtain the loglikelihood:

$$\ell(\mathbf{x}, n; \lambda, t) = \log(n-1)! + (n-2)\log(\lambda) - \lambda s \tag{2.10}$$

Taking the partial derivative of 2.10 to λ yields us

$$\frac{\partial}{\partial \lambda}\ell(\mathbf{x}, n; \lambda, t) = \frac{n-2}{s} - \lambda \tag{2.11}$$

Where setting 2.11 equal to zero yields our maximum likelihood estimate of the parameter λ:

$$\hat{\lambda}_{\mathrm{KM}} = \frac{n-2}{s} \tag{2.12}$$

This estimator has been called the Kendall-Moran estimator, after those who have derived it first. To obtain the variance of this estimator, we look to the inverse of the Fisher information matrix $\mathcal{J}$, which is a scalar in this case:

$$\begin{aligned}\mathcal{J}(\lambda, n) &= -\mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2}\log\left(\Pr(\mathbf{x}, n; \lambda, t)\right)\right] \\ &= \frac{n-2}{\lambda^2} \\ \Rightarrow \mathrm{Var}(\hat{\lambda}_{KM}) &= \frac{\lambda^2}{n-2}\end{aligned} \tag{2.13}$$

Therefore, we can i see by the results in (Dobson and Barnett 2008):

$$\hat{\lambda}_{\mathrm{KM}} \sim \mathcal{N}\left(\lambda, \frac{\lambda^2}{n-2}\right) \tag{2.14}$$

By this result, we can obtain a two sided 95% confidence interval simply. By properties of the normal distribution, we know that $z_0.975 = 1.96 = -z_0.025$. Hence we obtain:

$$-1.96 < \frac{\hat{\lambda}_{\mathrm{KM}} - \lambda}{\mathcal{J}^{-\frac{1}{2}}} = \frac{\hat{\lambda}_{\mathrm{KM}} - \lambda}{\frac{\lambda}{\sqrt{n-2}}} < 1.96 \tag{2.15}$$

From this last equation 2.15, we obtain the 95% confidence interval for λ by basic calculus:

$$\frac{\hat{\lambda}_{KM}}{1 - \frac{1.96}{\sqrt{n-2}}} < \lambda < \frac{\hat{\lambda}_{KM}}{1 + \frac{1.96}{\sqrt{n-2}}} \tag{2.16}$$

Kendall's Maximum Likelihood Estimation

Kendall obtained in 1949 a different variance from equation 2.13,

$$\frac{\lambda^2}{2(e^{\lambda t} - 1)} \tag{2.17}$$

However, it is easy to see the relationship between the variances given in equations 2.13 and 2.17, since we can rewrite equation 2.17 as:

$$\frac{\lambda^2}{2(e^{\lambda t}-1)} = \frac{\lambda^2}{2e^{\lambda t}-2)} = \frac{\lambda^2}{2\mathrm{E}[n]-2} \tag{2.18}$$

Where we treat n now a random variable representing the population, because of the pure-birth assumption. Note that it bega with two ancestral lineages time t ago. To obtain a 95% confidence intervalm we proceed similiarly to the Kendall-Moran case. However, we obtain after some calculations:

$$\hat{\lambda}_K = \lambda\left(1 \pm \frac{1.96}{\sqrt{2e^{\lambda t}-2}}\right) \tag{2.19}$$

This equation can't be solved to λ analytically. So we can't derive any explicit confidence interval in this case. However, in any particular case we can obtain these intervals numerically (Nee 2001).

The variances of Kendall-Moran (equation 2.13) and Kendall (equation 2.17) differ by the assumptions they made for both situations. Moran was considering a population of processes that grew untill they reach exactly n lineages. In that case, n is a predetermined variable in the likelihood (2.4.) and the age of the clade t is a random variable. In this model, the branch lengths, the elements of $\mathbf{x}$, are exponentially distributed. In the Kendall model, the time t was fixed and the number of lineages n was a random variable. In this model the branch lengths are truncated exponentially distributed. Allthough the maximum likelihood estimates are the same for both models, the variances differ. The Kendall model seems to be the appropiated one for inference in this context, corresponding to our original specifications (Nee 2001).

However, if we want to take the Moran model then it is not necessary to make use of any approximations because we obtain an exact confidence interval. Notice that an exponential random variable with $\lambda = 0.5$ has a chi-squared distribution with two degrees of freedom. Let $\chi_{2n,\alpha}$ be the upper α point of the chi-squared distribution with $2n$ degress of freedom. Then the exact 95% confidence interval for λ under the Moran model is (Nee 2001):

$$\frac{\chi_{2(n-2),0.025}}{2s} < \lambda < \frac{\chi_{2(n-2),0.975}}{2s} \tag{2.20}$$

Paradis' Maximum Likelihood Estimation

Paradis (1997) suggested a third choice for the variance, where he uses the observed Fisher information instead of the expected information:

$$\frac{\hat{\lambda}_P^2}{n-2} \tag{2.21}$$

This variance yields another 95% confidence interval, which is straight forward:

$$\hat{\lambda}_P\left(1 - \frac{1.96}{\sqrt{n-2}}\right) < \lambda < \hat{\lambda}_P\left(1 + \frac{1.96}{\sqrt{n-2}}\right) \tag{2.22}$$

The analysis of Paradis differs from all the others. He assumes the branch lengths are exponentially distributed, so he is studying the same hypothetical population of processes as Moran. However, Paradis' maximum likelihood estimate of λ differs:

$$\hat{\lambda}_P = \frac{n-1}{\sum i = 2^n x_i} \tag{2.23}$$

The numerator is larger by one, and the denumerator smaller by x_2 in comparison with the maximum likelihood estimate of Kendall and Moran. This difference is the result of the use of a different likelihood than equation (2.4).

Hey's Maximum Likelihood Estimation

Hey (1992) ignores the length of time between the last node in the tree and the present. That is equivalent to substracting nx_n from s. For equation (2.4), this is only a appropiate likelihood corresponding to the Moran model if we assume that a speciation event occured at the present day, such that $x_n = 0$. If this is not suitable but one wished to use Moran's model, one has to use Hey's form. From now, we assume that $x_n = 0$ when discussing the Moran model.

Likelihood ratio analysis

Observe the likelihood ratio statistic, which is chi-squared distributed with one degree of freedom (Dobson and Barnett 2008):

$$W(\lambda_0) = 2[\ell(\hat{\lambda}) - \ell(\lambda_0)] \sim \chi^2(1) \tag{2.24}$$

Where $\ell(\lambda)$ is the log likelihood function given in equation (2.10). A 95% confidence interval is given by the set of points $\lambda \in \Lambda$ such that:

$$W(\lambda) < \chi^2_{0.95}(1) = 3.841 \tag{2.25}$$

Which can also be solved numerically (Nee 2001).

Summarizing the last sections:

- Kendall's model correspond to our original specifications of the correct probability model for inference, but the confidence interval it provids is a numerical approximation whose accuracy is unknown. We can get an exact interval, when we discard the information about λ in the clade size.
- Moran's model provides an exact confidence interval, but the model assumes a fixed clade size and a randomly varying clade age. This does not seem appropiate in the present context.
- The likelihood ratio test analysis and the Paradis variant falls outside the natural development of this topic, because we base our analysis on models in which clade size, age or both are fixed.

Because non of the confidence intervals presents an overwhelming case for itself, we compare their performances in simulations (Nee 2001). The following was observed:

- The Paradis variant was the least precise variant, since it delivered the largest confidence interval for an equivalent precision as the Moran and Kendall model.
- The truncated exponential model was also dropped for the same reason as the Paradis variant.
- By making the ratio of the variances of Moran and Kendall's models as a new parameter, we can naturally compare the results of both models for different values of this ratio. Due to the better performances, Moran's model was the best model.

2.3 Likelihood of the birth-death process

We set $t_0 = 0$ as the time where our phylogeny begins. Also, we set the time T as the time of the present day, and the time t as some arbitrary time between 0 and T. For the birth death process, we can distinguish four related processes (Nee et al. 1994). These processes are plotted in figure 2.1, where the blue line represents a time $t > 0$ and the red line represents a time $T > t$:

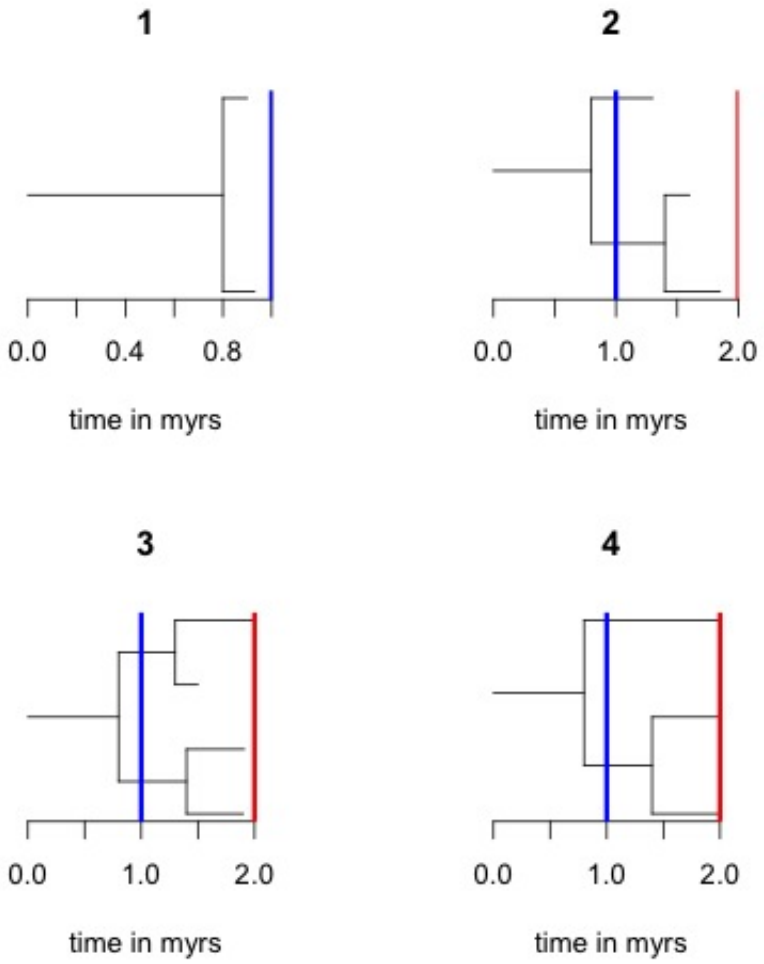

Figure 2.1: Four related processes within the constant birth-death process.

1. A simple birth death process which may or may not survive to time t.
2. A subset of the realizations of the first process and consists of those realizations which survive to time t between times 0 and T, but may or may not go extinct before the present day.
3. A subset of the second process which do survive to the present.
4. The reconstructed process, derived from the third by pruning the historical record of those lineages which do not have contemporary descendants. This process corresponds to a perfect phylogeny.

2.3.1 Probability functions of simulated trees.

We define $\Pr(i;t)$ as the probability that a process has i lineages at time t. For each of the four processes, we subsript the probabilities by 1 to 4 to clearify on which process we describe.

A crucial probability relevant to both paleontological and molecular phylogenetic data is the probability that a lineage that arose at some time t in the past, still has some descendants at the time T later (Kendall 1948):

$$P(t,T) = \frac{1-\frac{\mu}{\lambda}}{1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)(T-t)}} \tag{2.26}$$

The probability equation (2.26) is a crucial probabilty, because a lineage will not appear in a molecular phylogeny if it has no extant descendants.
We can estimate both composite parameters $a = \frac{\mu}{\lambda}$ and $r = b - d$, but in general the estimations of r are much more precise than the estimations of a (Foote 1988, Nee et al 1995a,). Together with the probability:

$$u_t = \frac{1-e^{-(\lambda-\mu)t}}{1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)t}} \tag{2.27}$$

We can express $\Pr_k(i;t)$ in terms of u_t and $P(t,T)$ for all four processes. We will show that these process all have a geometric distribution, except process 3, which has the distribution of two independent geometric random variables.

For the first process, we obtain that the probability of no descendants is equal to 1 minus the probability that a single lineage alive at time 0 has still some descendants at time t. Hence:

$$\underset{1}{\Pr}(i,t) = \begin{cases} 1-P(0,t) & \text{if } i=0 \\ P(0,t)(1-u_t)u_t^{i-1} & \text{if } i>0 \end{cases} \tag{2.28}$$

From the probability of the first process in equation 2.28, we immediately obtain the probability for process 2 by conditioning the probability on the event that the process survives untill time t:

$$\begin{aligned} \underset{2}{\Pr}(i,t) &= \Pr(i \text{ lineages}|\text{no extinction untill } t) \\ &= \frac{\Pr(i \text{ lineages} \wedge \text{no extinction untill } t)}{\Pr(\text{no extinction untill } t)} \\ &= \frac{\Pr_1(i,t|i>0)}{P(0,t)} \\ &= (1-u_t)u_t^{i-1} \end{aligned} \tag{2.29}$$

From the last probability equation 2.29, we obtain the conditional probability $\Pr_3(i,t;T)$ for a birth-death process that survives to T. To do this, we compound distribution 2.29 with the probability that at least one of the i lineages existing at time t has some descendants at time T (Nee et al. 1994):

$$\begin{aligned} \underset{3}{\Pr}(i,t;T) &= \frac{\Pr_2(i,t)(1-(1-P(t,T))^i)}{\sum_{j=1}^{\infty}\Pr_2(j,t)(1-(1-P(t,T))^i)} \\ &= \frac{(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)}{\sum_{j=1}^{\infty}(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)} \end{aligned} \tag{2.30}$$

This function is an ugly expression, but there is a simple underlying structure. To show this, we use the moment generating function (mgf) of random variables. To obtain our desired result, we first need the moment generating function of geometric random variables.

Property 1 *The moment generating function of a random variable X which is geometric distributed with parameter p, is given by:*

$$m_X(t) = \frac{pe^t}{1 - e^t(1-p)} \tag{2.31}$$

Proof

$$\begin{aligned}
m_X(t) &= \mathrm{E}\left[e^{Xt}\right] \\
&= \sum_{x=1}^{\infty} e^{xt} \Pr(X = x) \\
&= \sum_{x=1}^{\infty} (e^t)^x p(1-p)^{x-1} \\
&= e^t p \sum_{x=1}^{\infty} (e^t)^{x-1}(1-p)^{x-1} \\
&= e^t p \sum_{x=1}^{\infty} ((1-p)e^t)^{x-1} \\
&= e^t p \sum_{y=0}^{\infty} ((1-p)e^t)^{y} \\
&= e^t \frac{p}{1-(1-p)e^t}
\end{aligned}$$

As desired. ∎

Using the properties of the moment generating function, we observe the following for the probability density given for the third process:

Property 2 *The moment generating function of the number of lineages i in the third process, equals the moment generating function of the sum of two independent geometric variables G_1 and G_2, where*

$$\begin{aligned}
G_1 &\sim geo(u_t) \\
(G_2 + 1) &\sim geo(u_t(1 - P(t,T))
\end{aligned}$$

That is, that the pdf of G_2 is given by:

$$\Pr(G_2 = i) = (1 - u_t(1 - P(t,T))(u_t(1 - P(t,T)))^i, \quad i \geq 0$$

Proof We begin the proof with the moment generating function of G_2:

$$m_{G_2}(t) = \sum_{i=0}^{\infty} e^{it} \Pr(G_2 = i) \tag{2.32}$$

$$= \sum_{i=0}^{\infty} s^i (1 - u_t(1 - P(t,T))(u_t(1 - P(t,T))^i \tag{2.33}$$

$$= (1 - u_t(1 - P(t,T)) \sum_{i=0}^{\infty} (u_t(1 - P(t,T)s))^i \tag{2.34}$$

$$= \frac{1 - u_t(1 - P(t,T))}{1 - u_t(1 - P(t,T)s} \tag{2.35}$$

The moment generating of the the third process is given by:

$$\begin{aligned}
m_i(t) &= \sum_{i=1}^{\infty} e^{it} \Pr_3(i,t;T) \\
&= \sum_{i=1}^{\infty} s^i \frac{(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)}{\sum_{j=1}^{\infty}(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)} \\
&= \frac{s(1-u_t)}{\sum_{j=1}^{\infty}(1-u_t)u_t^{i-1}(1-(1-P(t,T))^i)} \sum_{i=1}^{\infty}(su_t)^{i-1}(1-(1-P(t,T))^i) \\
&= \frac{s(1-u_t)}{\kappa}\left[\sum_{i=1}^{\infty}(su_t)^{i-1} - \sum_{i=1}^{\infty}(su_t)^{i-1}(1-P(t,T))^i\right] \\
&= \frac{s(1-u_t)}{\kappa}\left[\frac{1}{1-su_t} - (1-P(t,T)\sum_{i=1}^{\infty}(su_t(1-P(t,T)))^{i-1}\right] \\
&= \frac{s(1-u_t)}{\kappa}\left[\frac{1}{1-su_t} - \frac{(1-P(t,T)}{su_t(1-P(t,T))}\right] \\
&= \frac{s(1-u_t)}{\kappa}\left[\frac{P(t,T)}{(1-su_t)(1-(1-P(t,T))su_t)}\right] \\
&= \left[\frac{s(1-u_t)}{1-u_ts}\right]\left[\frac{1}{\kappa}\frac{P(t,T)}{1-u_t(1-P(t,T))s}\right]
\end{aligned}$$

Where:

$$\begin{aligned}
\kappa &= \sum_{j=1}^{\infty}(1-u_t)u_t^{j-1}[1-(1-P(t,T))^i] \\
&= \sum_{j=1}^{\infty}(1-u_t)u_t^{j-1} - \sum_{j=1}^{\infty}(1-u_t)u_t^{j-1}[1-(1-P(t,T))^i] \\
&= 1-(1-u_t)(1-P(t,T))\sum_{j=1}^{\infty}u_t^{j-1}(1-P(t,T))^{i-1} \\
&= 1-(1-u_t)(1-P(t,T))\frac{1}{1-u_t(1-P(t,T))} \\
&= \frac{P(t,T)}{1-u_t(1-P(t,T))}
\end{aligned}$$

In that case:

$$m_i(t) = \left[\frac{s(1-u_t)}{1-u_ts}\right]\left[\frac{1-u_t(1-P(t,T))}{1-u_t(1-P(t,T))s}\right] \tag{2.36}$$

We recognize that the moment generating function is the product of two moment generating functions of variables G_1 and G_2. The first term is the moment generating function of $G_1 \sim \text{geo}(u_t)$ and the second term is the moment generating function of $(G_2+1) \sim \text{geo}(u_t(1-P(t,T))$. ∎

Thus, the number of lineages existing at time t for a birth-death process which will survive to a time T later can be treated as the sum of two independent variables, with a geometric distribution. Now we will see that process 4, which is reconstructed from this one, is again geometric distributed for $\Pr(i,t)$.

Let z_i be the probabilities of the third process, given by equation 2.30. Of the j lineages existing at time t, i will have some descendants at the time T. Here, i is binomial distributed with parameter $P(t,T)$ and $n > 0$, since at least one survives to T. So, we obtain for the reconstructed process:

$$\Pr_4(i,t;T) = \sum_{j=1}^{\infty} \frac{z_j}{1-(1-P(t,T))^j} \binom{j}{i} P(t,T)^i (1-P(t,T))^{j-i} \tag{2.37}$$

Which simplifies to (Nee et al. 1994):

$$\Pr_4(i,t;T) = \left(1 - u_t \frac{P(0,T)}{P(0,t)}\right) \left(u_t \frac{P(0,T)}{P(0,t)}\right)^{i-1}, \quad i > 0 \tag{2.38}$$

Which is a geometric distribution with parameter $u_t \frac{P(0,T)}{P(0,t)}$. Therefore, the first, second and fourth process have a geometric distributed number of lineages, the third process' number of lineages is distributed as the sum of two geometric variables.

2.3.2 Probability Density Function of a Simulated Tree

We now look more closely to a birth-death process which generated the distribution given in equation 2.38. We let $n(t)$ be the number of lineages at time t. We suppose that we grow a reconstructed phylogenetic tree, starting at time $t = 0$, thus $n(0) = 1$. Each lineage a time t give rise to a daughter lineage at a rate $\lambda \Pr(t,T)$. So, after a small amount of time dt, we have:

$$n(t) = \begin{cases} n(t)+1 & \text{with probability } n(t)\lambda P(t,T)dt \\ n(t) & \text{with probability } 1 - n(t)\lambda P(t,T)dt \end{cases} \tag{2.39}$$

We extend the probability model given in equation 2.39 by the following. Given that we have n lineages at a time t_n, we denote the time untill the next lineage as τ. We then have:

$$\begin{aligned} \Pr(\tau > t + dt; 0 < t < T - t_n) &= \Pr(\tau > t; 0 < t < T - t_n) \Pr(\text{no lineage in } dt) \\ &= \Pr(\tau > t; 0 < t < T - t_n)(1 - \lambda n(t) P(t,T) dt) \end{aligned} \tag{2.40}$$

Which is equivalent to:

$$\frac{d}{dt} \Pr(\tau > t; 0 < t < T - t_n) = -\lambda n(t) P(t,T) \Pr(\tau > t; 0 < t < T - t_n) \tag{2.41}$$

Since we know the number of lineages $n(t)$ at time t, we can treat it as a fixed variable. The solution of this ODE is then straight forward:

$$\Pr(\tau > t; 0 < t < T - t_n) = \exp\left(-\lambda n \int_{t_n}^{t_n + t} P(s,T) ds\right) \tag{2.42}$$

Where the last integral needs some computations to get the solution:

$$
\begin{aligned}
\int_{t_n}^{t_n+t} P(s,T)ds &= \int_{t_n}^{t_n+t} \frac{1-\frac{\mu}{\lambda}}{1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)(T-s)}}ds \\
&= \int_{t_n}^{t_n+t} \frac{\lambda-\mu}{\lambda-\mu e^{-(\lambda-\mu)(T-s)}}ds \\
&\text{substitute } u(t) = (\lambda-\mu)t, \text{ to obtain:} \\
&= \int_{u(t_n)}^{u(t_n+t)} \frac{1}{\lambda-\mu e^{u-(\lambda-\mu)T}}du \\
&= \int_{u(t_n)}^{u(t_n+t)} \frac{1}{\lambda-\kappa e^{u}}du \\
&\text{substitute } v(u) = \lambda-\kappa e^{u} \text{ to obtain} \\
&= \int_{v(u(t_n))}^{v(u(t_n+t))} \frac{1}{w}\frac{1}{w-\lambda}dw \\
&= \int_{v(u(t_n))}^{v(u(t_n+t))} \left(-\frac{1}{\lambda w}+\frac{1}{\lambda(w-\lambda)}\right)dv \\
&= -\frac{1}{\lambda}\log(w)+\frac{1}{\lambda}\log(\lambda(w-\lambda))\Big|_{v(u(t_n))}^{v(u(t_n+t))} \\
&= t-\frac{\mu}{\lambda}t-\frac{1}{\lambda}\log\left(\frac{1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n-t))}{1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n))}\right)
\end{aligned} \tag{2.43}
$$

Which yields us the solution of the ODE:

$$
\Pr(\tau > t; 0 < t < T-t_n) = \exp(-n(\lambda-\mu)t)\left(\frac{1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n-t))}{1-\frac{\mu}{\lambda}\exp(-(\lambda-\mu)(T-t_n))}\right)^n \tag{2.44}
$$

From the last equation 2.45, we can obtain the probability density function of the waiting time for a birth, t (Nee et al. 1994):

$$
\Pr(t; t_n, T, \lambda, \mu) = n(\lambda-\mu)e^{-n(\lambda-\mu)t}\frac{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)(T-t_n-t)}\right)^{n-1}}{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)(T-t_n)}\right)^{n}} \tag{2.45}
$$

From this probability density function in equation 2.45, we obtain the likelihood of a tree which has grown according to a birth-death process. Let x_i denote the time between the i-th birth event and the present day T. Then, the time between two lineages is $t_i = x_i - x_{i+1}$ and $x_i = T - t_n$. Suppose we have a tree which has N lineages at the present day. Let the vector $\mathbf{x}$ contain all the times between lineages, which starts from the time that the first species gave birth to two new species. Therefore, $\mathbf{x} = (x_2, x_3, \ldots, x_N)$. We obtain the likelihood of the tree by multiplying all independent probabilities of the times between lineages:

$$
\begin{aligned}
L(\mathbf{x}|\lambda,\mu) &= \prod_{i=1}^{N-1} \Pr(t_i; t_n, T, \lambda, \mu) \\
&= \prod_{i=1}^{N-1} n(\lambda-\mu)e^{-n(\lambda-\mu)(x_i-x_{i+1})}\frac{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)x_{i+1}}\right)^{n-1}}{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)x_i}\right)^{n}}
\end{aligned} \tag{2.46}
$$

2.3.3 Maximum Likelihood Estimation

To find the maximum likelihood estimator of the parameters of the birth-death process, we need the likelihood obtain in equation 2.46. Instead of maximizing the likelihood, we maximize the log-likelihood:

$$\ell(\mathbf{x}|\lambda,\mu) = \log\left(\prod_{i=1}^{N-1} n(\lambda-\mu)e^{-n(\lambda-\mu)(x_i-x_{i+1})}\frac{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)x_{i+1}}\right)^{n-1}}{\left(1-\frac{\mu}{\lambda}e^{-(\lambda-\mu)x_i}\right)^{n}}\right)$$
$$= \sum_{i=2}^{N-1}\{\log(n) + log(\lambda-\mu) - n(\lambda-\mu)(x_i - x_{i+1})\}$$
$$+ \sum_{i=2}^{N-1}\left\{(n-1)\log(1-\frac{\lambda}{\mu}e^{-(\lambda-\mu)x_{i+1}}) - n\log(1-\frac{\lambda}{\mu}e^{-(\lambda-\mu)x_i})\right\} \quad (2.47)$$

The normal approach, i.e. taking partial derrivatives and setting equal to zero is not applicable in this case, since this equation can't be solved analytically. Therefore, we obtain maximum likelihood estimates of the parameters λ and μ numerically. This will be done in chapter 3.

2.4 Approximating the Likelihood of the Protracted Birth-Death Model

Instead of finding the likelihood of a tree which has grown according to a protracted birth-death model, we make an approximation using Gaussian processes. A protracted birth-death tree which has grown according to a Gaussian process is not a useful tree for the estimation, since the tree its lineages might not be a natural number in this process. However, treating a tree which has grown according to a protracted birth-death process, can be inferred by an approximation with a birth-death process. We first give the definition of a Gaussian process:

Definition A real-valued stochastic process, $\{X_t; t \in T\}$, where T is an index set, is a *Gaussian process* if all the finite-dimensional distributions have a multivariate normal distribution. That is, for any choice of distinct values $t_1, \ldots, t_k \in T$, the random vector $\mathbf{X} = (X_{t_1}, \ldots, X_{t_k})'$ has a multivariate normal distribution with mean $\mu = \mathrm{E}[\mathbf{X}]$ and covariance matrix $\Sigma = \mathrm{cov}(X, X)$, which has the probability density function:

$$f_{\mathbf{X}}(\mathbf{X};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{X}-\mu)\right] \quad (2.48)$$

Which is denoted by:

$$\mathbf{X} \sim N(\mu, \Sigma) \quad (2.49)$$

The mean and covariance of a Gaussian process are defined by:

$$\mu(t) = \mathrm{E}[X_t], \quad \gamma(s,t) = \mathrm{cov}(X_s, X_t), \quad t, s \in T \quad (2.50)$$

2.4.1 Approximation of the Process

We let N_t denote the state at time t, which consists of the number of good species $N_g(t)$ and the number of incipient species $N_i(t)$ at time t. Suppose that we take a step in time dt, such that the next event occurs. For the protracted birth-death process, five different event may occur:

1. A good species giving birth to an incipient species, with a rate $\lambda_1 N_g(t)dt$.
2. A good species going extinct, with a rate $\mu N_g(t)dt$.
3. An incipient species reaching completion, with a rate $\lambda_2 N_i(t)dt$
4. An incipient species giving birth to a new incipient species, with a rate $\lambda_3 N_i(t)dt$
5. An incipient species going extinct, with a rate $\mu N_i(t)dt$.

The Mean and Covariance of the Process

We look to the difference between the state of today $N_t = [N_g(t), N_i(t)]$ and the state of the next event N_{t+dt}, which is denoted by $\Delta N_t = N_{t+dt} - N_t$. To avoid confusion with the mean μ and the death rate μ, we let the vector $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \mu)$ denote the parameters. Doing so, we have the following expectation:

$$\begin{aligned}
\mathrm{E}[\Delta N_t | N_t] &= \begin{bmatrix} -\mu N_g(t)dt + \lambda_2 N_i(t)dt \\ -\mu N_i(t)dt - \lambda_2 N_i(t)dt + \lambda_1 N_g(t)dt + \lambda_3 N_i(t)dt \end{bmatrix} \\
&= \begin{bmatrix} -\mu N_g(t) + \lambda_2 N_i(t) \\ +\lambda_1 N_g(t) - (\lambda_2 + \lambda_3 + \mu) N_i(t) \end{bmatrix} dt \\
&= \mu \left\{ N_t, \Lambda \right\} dt
\end{aligned} \tag{2.51}$$

We again look to the difference difference between the state of today ΔN_t. As we have earlier defined the expectation of the development in the protracted birth-death process, given the last state of the process, we do exactly the same for the conditional covariance. In a similiar way as for the expectation, we obtain the following conditional covariance:

$$\mathrm{cov}(\Delta N_t | N_t) = \begin{bmatrix} \mathrm{Var}\left(\Delta N_g(t+dt) | N_t\right) & \mathrm{cov}(\Delta N_g(t+dt), \Delta N_i(, t+dt) | N_t) \\ \mathrm{cov}(\Delta N_g(t+dt), \Delta N_i(, t+dt) | N_t) & \mathrm{Var}\left(\Delta N_i(t+dt) | N_t\right) \end{bmatrix} \tag{2.52}$$

We obtain the entries of the matrix part by part. We start with $\Sigma_{1,1}$:

$$\begin{aligned}
\mathrm{Var}\left(\Delta N_g(t+dt) | N_t\right) &= \mathrm{E}[(\Delta N_g(t+dt))^2 | N_t] - (\mathrm{E}[\Delta N_g(t+dt) | N_t])^2 \\
&= \mu N_g(t)dt + \lambda_2 N_i(t)dt - (\mu N_g(t)dt + \lambda_2 N_i(t)dt)^2 \\
&= \mu N_g(t)dt + \lambda_2 N_i(t)dt - (\mu N_g(t) + \lambda_2 N_i(t))^2 dt^2 \\
&\propto \mu N_g(t)dt + \lambda_2 N_i(t)dt
\end{aligned} \tag{2.53}$$

$\Sigma_{1,2} = \Sigma_{2,1}$ is given by:

$$\begin{aligned}
\mathrm{cov}(\Delta N_g(t+dt), \Delta N_i(t+dt) | N_t) &= \mathrm{E}[\Delta N_g(t+dt) \cdot \Delta N_i(t+dt) | N_t] \\
&\quad - \mathrm{E}[\Delta N_g(t+dt) | N_t] \cdot \mathrm{E}[\Delta N_i(t+dt) | N_t] \\
&= -\lambda_2 N_i(t)dt - (-\mu N_g(t)dt + \lambda_2 N_i(t)dt) \\
&\quad \times (-\mu N_i(t)dt - \lambda_2 N_i(t)dt + \lambda_1 N_g(t)dt + \lambda_3 N_i(t)dt) \\
&= -\lambda_2 N_i(t)dt - (-\mu N_g(t) + \lambda_2 N_i(t)) \\
&\quad \times (-\mu N_i(t) - \lambda_2 N_i(t) + \lambda_1 N_g(t) + \lambda_3 N_i(t))dt^2 \\
&\propto -\lambda_2 N_i(t)dt
\end{aligned} \tag{2.54}$$

$\Sigma_{2,2}$ is given by:

$$\begin{aligned}
\mathrm{Var}\left(\Delta N_i(t+dt) | N_t\right) &= \mathrm{E}[(\Delta N_i(t+dt))^2 | N_t] - (\mathrm{E}[\Delta N_i(t+dt) | N_t])^2 \\
&= \lambda_1 N_g(t)dt + (\lambda_2 + \lambda_3 + \mu) N_i(t)dt \\
&\quad - (-\mu N_i(t)dt - \lambda_2 N_i(t)dt + \lambda_1 N_g(t)dt + \lambda_3 N_i(t)dt)^2 \\
&= \lambda_1 N_g(t)dt + (\lambda_2 + \lambda_3 + \mu) N_i(t)dt \\
&\quad - (-\mu N_i(t) - \lambda_2 N_i(t) + \lambda_1 N_g(t) + \lambda_3 N_i(t))^2 dt^2 \\
&\propto \lambda_1 N_g(t)dt + (\lambda_2 + \lambda_3 + \mu) N_i(t)dt
\end{aligned} \tag{2.55}$$

We therefore obtain the covariance matrix Σ as:

$$\operatorname{cov}(\Delta N_t|N_t) = \begin{bmatrix} \mu N_g(t) + \lambda_2 N_i(t) & -\lambda_2 N_i(t) \\ -\lambda_2 N_i(t) & \lambda_1 N_g(t) + (\lambda_2 + \lambda_3 + \mu) N_i(t) \end{bmatrix} dt = \Sigma\{N_t, \Lambda\} \cdot dt \quad (2.56)$$

If we assume that the number good species and incipient species is nonnegative, the matrix $\Sigma\{N_t, \Lambda\}$ is nonsingular in most of the situations because the determinant is nonzero. Observe that the determinant of Σ is given by the following expression:

$$\begin{aligned} \det(\Sigma) &= \begin{vmatrix} \mu N_g(t) + \lambda_2 N_i(t) & -\lambda_2 N_i(t) \\ -\lambda_2 N_i(t) & \lambda_1 N_g(t) + (\lambda_2 + \lambda_3 + \mu) N_i(t) \end{vmatrix} \\ &= \mu N_g(t) \left(\lambda_1 N_g(t) + [\lambda_2 + \lambda_3 + \mu] N_i(t)\right) + \lambda_2 N_i(t) \left(\lambda_1 N_g(t) + (\lambda_3 + \mu) N_i(t)\right) \end{aligned} \quad (2.57)$$

The determinant is nonzero, except if:

1. All species are extinct.
2. Only one of the parameters is nonzero.
3. All good species have become extinct, and $\lambda_2 = 0$ or $\lambda_3 = \mu = 0$.
4. There are no incipient species left, $\mu = 0$ or $\lambda_1 = 0$.

Because the matrix Σ is in general nonsingular, we can find the square root of the matrix by diagonalization. That is:

$$B(X_t|\Lambda) = V D^{\frac{1}{2}} V^{-1} \quad (2.58)$$

Where D is the matrix consisting of the eigenvalues of Σ, and V is the matrix consisting of the eigenvalues of Σ.

Defining the New Process

Now we have obtained the expressions of $\mu(N_t, \Lambda)$ and $\Sigma(N_t, \Lambda)$, we can obtain a Gaussian process for our protracted birth-death model. We first need to define the concept of a Wiener Process:

Definition A statistical process $\{W_t; t \in T\}$ in continuous time is called a *Wienerprocess* or *Brownian Motion* if it has the following properties:

1. $W_0 = 0$
2. The function $t \to W_t$ is everywhere continuous almost surely.
3. W_t has independent increments with $W_t - W_s \sim N(0, I(t-s))$ for $0 \leq s < t$.

We define the new process Y_t by means of a Wiener Process $\{W_t; t \in T\}$:

$$dY_t = \mu(Y_t; \Lambda)dt + B(Y_t; \Lambda)dW_t \quad (2.59)$$

When we look more closely to equation 2.59, we see that the conditional expectation and the conditional variance of dY_t given Y_t are the same for the protracted birth-death process. Therefore, we conclude that this newly defined process is an approximation of the original protracted birth-death process. However, notice that is it not exactly the same process. We can even see that this construction of the approximation yields us a Gaussian process:

$$\begin{aligned} \mathrm{E}[dY_t|Y_t] &= \mu(Y_t; \Lambda)dt \\ \operatorname{cov}[dY_t|Y_t] &= B(Y_t; \Lambda)dt B(Y_t; \Lambda)' = \Sigma \cdot dt \\ dY_t &\sim N(\mu dt, \Sigma dt) \end{aligned}$$

Therefore, the process given by $Y_{k+1} = Y_k + dY_k$ is a Gaussian process. This means that we can approximate the protracted birth-death model, which is originally a Poisson process, by means of a Gaussian process, given by $\{Y_t, t \in T\}$.

2.4.2 Approximation of the likelihood

Suppose that we have a tree which has grown according to a protracted birth-death process. For this model, we assume that we use actual phylogenies for inferences, instead of the reconstructed phylogenies. Doing so, we can list the number of good species and incipient species on a time t in a vector, called N_t. At a time $t + \Delta t$ later, the next event occurs in our protracted birth-death process. We are interested in the difference between N_t and $N_{t+\Delta t}$:

$$\Delta N_t = N_{t+\Delta t} - N_t \tag{2.60}$$

This difference can be seen as a development in the process. Normally, the difference is obviously a vector containing whole numbers. We can view these differences however, as an outcome in a Gaussian process. Intuitively, the set $\{\Delta N_t | t \in T\}$ can be treated a Gaussian process. Furthermore, by the construction of our approximation dY_t, we can even claim that all these variables are indepent multivariate normal random variables. This is the case, because the approximation dY_t is constructed by means of a Wienerprocess which has independent increments.

Since the likelihood of a development ΔN_t in the protracted process is now approximated by the probability density of the multivariate normal distribution, we have that for each event time $t \in T$:

$$\begin{aligned} f_{\Delta N_t}(\Delta N_t; \mu_t, \Sigma_t) &\approx \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma_t|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\Delta N_t - \mu_t)' \Sigma_t^{-1} (\Delta N_t - \mu_t)\right] \\ &= dmnorm(\Delta N_t; \Sigma_t, \mu_t) \end{aligned} \tag{2.61}$$

By the independent increments of the Wienerprocess, we therefore obtain the following loglikelihood of the actual phylogeny:

$$\ell(\{N_t, t \in T\}; \Lambda) \approx \sum_{t \in T} \log\left(dmnorm(\Delta N_t; \Sigma_t, \mu_t)\right) \tag{2.62}$$

Thismaximum likelihood estimate can't be solved analytically for $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \mu)$. The inferences will therefore be computed numerically in chapter 3.

Chapter 3

Simulation studies

In this chapter, we investigate the simulations of the constant birth-death model and the protracted birth-death model. We first review and proof basic properties of the models.

3.1 Simulation of a Birth-Death Model

The simulation of the birth-death model has been done numerically in R. In order to do so, we use a recursive algorithm. First, we give an outline of important analytical properties of the algorithm.

3.1.1 Basic Properties

The first event

In a birth-death model, we suppose that speciation and extinction are processes that proceed simulataneously. The process that finished first, is the process that we observe. Let $T_b(1)$ denote the time that speciation of the first species in a phylogeny would occur, and $T_d(1)$ the time that extinction of this species would occure. Furthermore, we assume that $T_b(1) \sim \exp(\lambda)$ and $T_d(1) \sim exp(\mu)$ are independent random variables. We now consider the time that the first of these two events occurs: $T(1) = \min(T_b(1), T_d(1))$. First, we state an useful theorem.

Property 3 *Suppose $X_1, \ldots, X_n$ are independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$. Then the minimum of the random variables, $X = \min\{X_1, \ldots, X_n\}$, is exponentially distributed with parameter $\lambda = \sum_{i=1}^{n} \lambda_i$.*

Proof We make use of unicity of the surivival function for random variables. Recall that the survival function of an exponentially distributed random variable Y with parameter θ, is defined by:

$$S(y;\theta) = \Pr(Y \geq y) = \exp(-\theta y) \tag{3.1}$$

Now we are going to make use of the fact, that for a certain $a > 0$: $X = \min\{X_1, \ldots, X_n\} > a$ if and only if $X_i > a$ for all $i = 1 \ldots n$. By independency of the random variables, we obtain the survival function of X:

$$\begin{aligned}
\Pr(X = \min\{X_1, \ldots, X_n\} > a) &= \prod_{i=1}^{n} \Pr(X_i > a) \\
&= \prod_{i=1}^{n} \exp(-a\lambda_i) \\
&= \exp(-a \sum_{i=1}^{n} \lambda_i)
\end{aligned}$$

So by the unicity of the survival function, $X \sim \exp(\sum_{i=1}^{n} \lambda_i)$. ∎

As a direct consequence of this theorem, the time untill the first event (speciation or extinction) occurs, $T = \min\{T_b, T_d\}$ is exponentially distributed with parameter $\lambda + \mu$. We now make the following statement for the probability of speciation:

Property 4 *Suppose that the speciation event time is exponentially distributed with parameter λ, and the extinction event time is exponentially distributed with parameter μ. Then, given that speciation or extinction has occured: the probability of a speciation event is $\frac{\lambda}{\lambda+\mu}$.*

Proof Speciation occurs if the speciation process completes before the extinction process, that is when $T_b < T_d$.The probability of speciation in the first event is straight fowardly obtained:

$$\begin{aligned}
\Pr(\text{birth occurs}) &= Pr(T_b < T_d) \\
&= \int_0^\infty \Pr(T_b < T_d | T_d = x) f_{T_d}(x) dx \\
&= \int_0^\infty \Pr(T_b < x) f_{T_d}(x) dx \\
&= \int_0^\infty (1 - \exp(-\lambda x)) \mu \exp(-\mu x) dx \\
&= \mu \int_0^\infty \{\exp[-\mu x] - \exp[-(\lambda+\mu)x]\} dx \\
&= \mu \int_0^\infty \{\exp[-\mu x] - \exp[-(\lambda+\mu)x]\} dx \\
&= \lim_{t\to\infty} \mu \left[\frac{\exp[-(\lambda+\mu)x]}{\lambda+\mu} - \frac{\exp[-\mu x]}{\mu} \right] \\
&= \mu \left[\frac{1}{\mu} - \frac{1}{\lambda+\mu} \right] \\
&= \frac{\lambda}{\lambda+\mu}
\end{aligned}$$

As desired. ∎

The Second Event

Suppose now that for our first species in the tree, speciation has occured. We get two new species named "A" and "B", in which again the speciation process and extiction process develop simultaneously. Likewise to the first eventtime, we denote $T(2)$ as the second eventtime, which is given by the minimum of the four events which are developing at the moment. That is, the two speciation processes and the two extinction processes developing for each species. So, $T(2) = \min\{T_{A,b}, T_{A,d}, T_{B,b}, T_{B,d}\}$. Again as a consequence of property 3, we have that

$$T(2) = \min\{T_{A,b}, T_{A,d}, T_{B,b}, T_{B,d}\} \sim \exp[2(\lambda+\mu)] \tag{3.2}$$

Now we know the distribution of the second event time, we can investigate in which the probability of which branch the event took place. Logically, the probability of an event occuring for a certain species is for both branches the same. So, we obtain probabilities:

$$\begin{aligned}
\Pr(\min\{T_{A,b}, T_{A,d}\} < \min\{T_{B,b}, T_{B,d}\}) &= \Pr(\min\{T_{A,b}, T_{A,d}\} > \min\{T_{B,b}, T_{B,d}\}) \\
&= 0.5
\end{aligned}$$

for both species. This probability can also be obtained analytically. Let $T_A = \min\{T_{A,b}, T_{A,d}\}$ and $T_B = \min\{T_{B,b}, T_{B,d}\}$, then we obtain:

$$\begin{aligned}
\Pr(T_A < T_B) &= \int_0^\infty \Pr(T_A < T_B | T_B = x) f_{T_B}(x) dx \\
&= \int_0^\infty \Pr(T_A < x) f_{T_B}(x) dx \\
&= \int_0^\infty \{1 - \exp[-(\lambda+\mu)x]\}(\lambda+\mu)\exp[-(\lambda+\mu)x] dx \\
&= (\lambda+\mu)\int_0^\infty \{\exp[-(\lambda+\mu)x] - \exp[-2(\lambda+\mu)x]\} dx \\
&= (\lambda+\mu)\left[\frac{1}{\lambda+\mu} - \frac{1}{2(\lambda+\mu)}\right] \\
&= 1 - \frac{1}{2} \\
&= 0.5
\end{aligned}$$

Obviously, given that an event has occured for a particular species, the probability of speciation is the same as for the first speciation event. So, $\Pr(\text{speciation}) = \frac{\lambda}{\lambda+\mu}$.

The k-th Event

Intuitively, we observe analytical structures in the first two stages of the phylogenetic tree. Therefore, we present some general results.

Property 5 *The k-th event time, which is the minimum of k speciation times and k extinction times, is exponentially distributed with parameter* $k(\lambda+\mu)$.

Proof Since the k-th eventtime is the minimum of k independent exponential random variables with parameter λ and k independent exponential random variables with parameter μ. By property 3, the minimum of these $2k$ independent exponential random variables with parameter $k\lambda + k\mu = k(\lambda+\mu)$. ∎

Property 6 *Suppose at the time of the k-th event, we have* $m > 0$ *extant species. Then, the probability that an extinction or speciation event has occured for a certain species, is for each species the same:* $\frac{1}{m}$.

Proof We consider a certain fixed species j. The probabilty that the event has occured for this particular species, is the probability that the minimum of its speciation and extinction time in branch j is smaller than all other speciation and extinction times, developing for the other $m-1$ species. Let $T_j = \min(T_{j,b}, T_{j,d})$ denote this time and let T_{rest} be the minimum of the other $m-1$ speciation times and the $m-1$ extinction times. Thus, T_{rest} is exponentially distributed with parameter $(m-1)(\lambda+\mu)$ and T_j is exponentially distributed with parameter $\lambda+\mu$. We thus obtain:

$$\begin{aligned}
\Pr(T_j < T_{\text{rest}}) &= \int_0^\infty \Pr(T_j < T_{\text{rest}} | T_{\text{rest}} = x) f_{<T_{\text{rest}}}(x) dx \\
&= \int_0^\infty (1 - \exp[-(\lambda+\mu)x])\,((m-1)(\lambda+\mu)\exp[-(m-1)(\lambda+\mu)x])\,dx \\
&= (m-1)(\lambda+\mu)\int_0^\infty (\exp[-(m-1)(\lambda+\mu)x] - \exp[-m(\lambda+\mu)x])\,dx \\
&= (m-1)(\lambda+\mu)\left[\frac{1}{(m-1)(\lambda+\mu)} - \frac{1}{m(\lambda+\mu)}\right] \\
&= 1 - \frac{m-1}{m} \\
&= \frac{1}{m}
\end{aligned}$$

As desired. ∎

As shown in the first part, the probability of a speciation event given that speciation or extinction has occured, is $\frac{\lambda}{\lambda+\mu}$.

Algorithm birth-death process

In appendix A.3, the source code for a constant birth-death simulation is given. During the simulation, we refer a node as an event time and a branch as a time interval in which speciation and extinction processes are developing.

In this recursive algorithm, we do the following:

1. Set the speciation and extinction rates.
2. - Draw a event time $T \sim \exp(\lambda + \mu)$.
   - Draw a uniform random variable $b \sim U(0, 1)$.
3. - If $b <$ Pr(birth), we have a birth event. Draw two new pairs of random variables T and b and store them in the tree. After that, jump to the next event in order of time.
   - If not, we have a death event. Go to the next event in order of time. If there is no next event in the tree, stop the simulation.
4. Repeat step 4 untill the tree has fully developed, or the maximum amount of time has been reached.

An example of a simulation is plotted below:

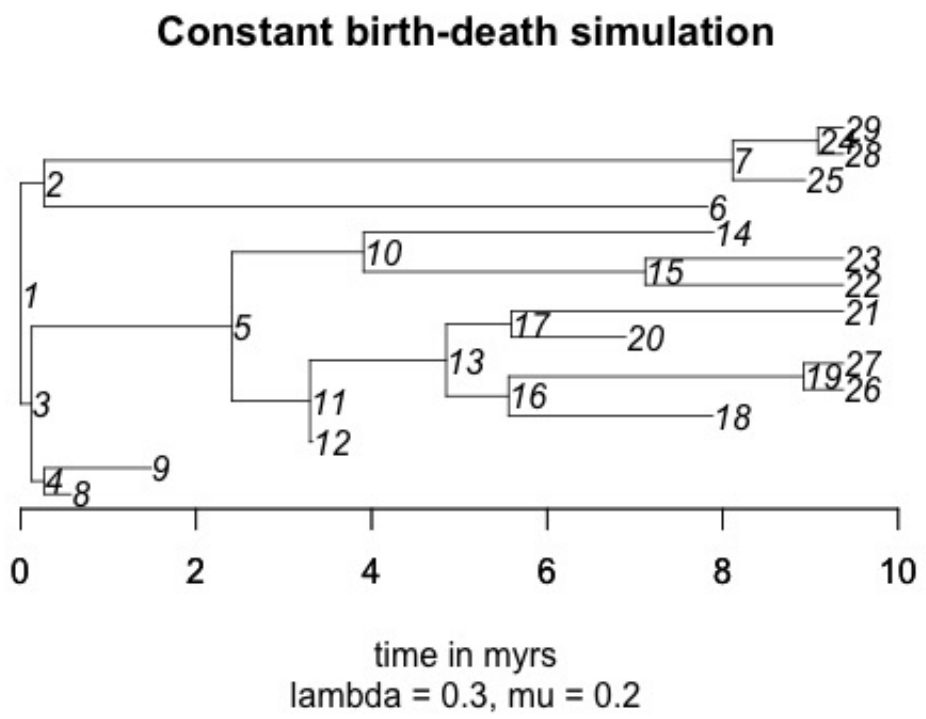

Figure 3.1: A plot of a constant birth-death simulation.

The data frame belonging to plot 3.1 is given in tabel 3.1.

	id	ancestor	origin	eventtime	branchtime	birth
1	1	0	0.00000	0.64537	0.64537	TRUE
2	3	1	0.64537	0.77253	0.12716	TRUE
3	2	1	0.64537	0.91345	0.26808	TRUE
21	4	3	0.77253	0.91731	0.14478	TRUE
41	8	4	0.91731	1.21859	0.30127	FALSE
5	9	4	0.91731	2.13321	1.21589	FALSE
31	5	3	0.77253	3.05408	2.28155	TRUE
411	11	5	3.05408	3.94290	0.88881	TRUE
412	12	11	3.94290	3.97149	0.02859	FALSE
32	10	5	3.05408	4.55314	1.49905	TRUE
51	13	11	3.94290	5.49617	1.55327	TRUE
52	16	13	5.49617	6.20631	0.71014	TRUE
6	17	13	5.49617	6.23669	0.74052	TRUE
71	20	17	6.23669	7.54231	1.30562	FALSE
511	15	10	4.55314	7.76335	3.21021	TRUE
311	6	2	0.91345	8.47504	7.56159	FALSE
61	18	16	6.20631	8.53278	2.32647	FALSE
413	14	10	4.55314	8.54152	3.98839	FALSE
4	7	2	0.91345	8.75956	7.84611	TRUE
7	19	16	6.20631	9.56840	3.36209	TRUE
62	25	7	8.75956	9.58521	0.82564	FALSE
53	24	7	8.75956	9.73031	0.97075	TRUE
8	21	17	6.23669	10.01926	3.78257	FALSE
711	22	15	7.76335	10.01926	2.25591	TRUE
81	23	15	7.76335	10.01926	2.25591	TRUE
611	26	19	9.56840	10.01926	0.45086	TRUE
73	27	19	9.56840	10.01926	0.45086	TRUE
63	28	24	9.73031	10.01926	0.28894	TRUE
72	29	24	9.73031	10.01926	0.28894	TRUE

Table 3.1: Data frame of the simulated tree.

3.1.2 Results of simulation

Different Parameter Settings

During the simulation we consider three situations for the constant birth-death model.

1. the birth rate is larger than the death rate: $\lambda > \mu$.
2. the birth rate is smaller than the death rate $\lambda < \mu$
3. the birth rate equals the death rate $\lambda = \mu$

Case 1. In simulation studies, we discover fast growing trees when we set $\lambda > \mu$. An illustrating example of such simulation is given in table 3.1 with figure 3.1 . Setting $\lambda > \mu$ yields us a probability bigger than 0.5 in equation 4. Thus, the probability of a speciation event is larger than the probability of an extinction event in any node. The greater the relative difference between λ and μ gets, the higher the probability the phylogeny expands faster. By growing multiple trees, we find patterns which agree on this. But, even if $\lambda > \mu$, the process still can go extinct. This happens when all its lineages become to extinct.

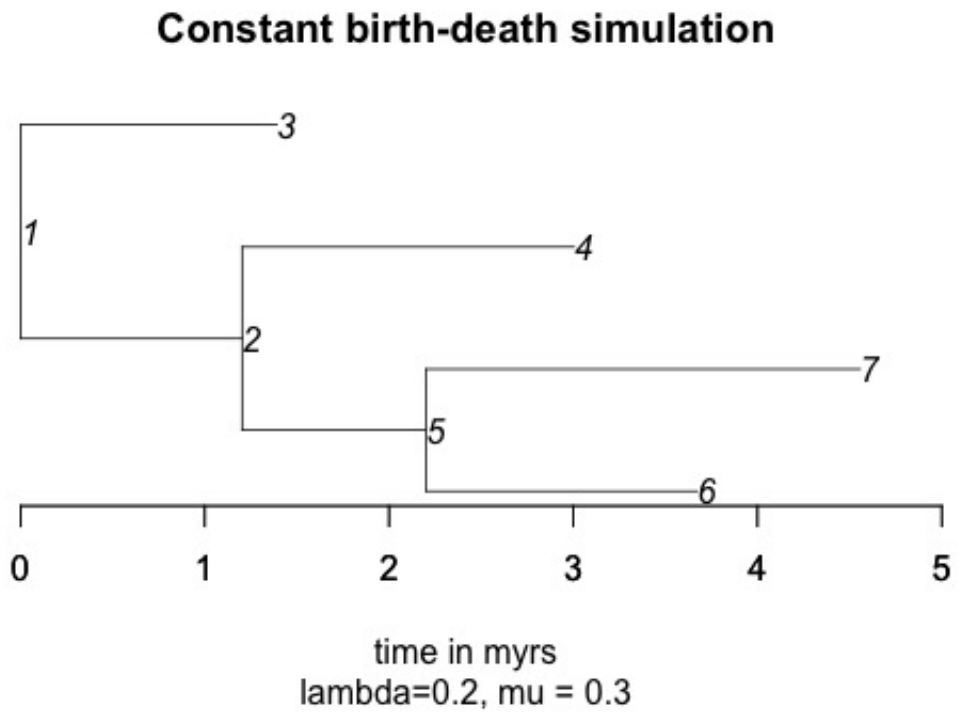

Figure 3.2: Plot of a constant birth-death simulation, with a birth rate smaller than the death rate. The complete tree has become extinct.

Case 2. Setting $\lambda < \mu$, yields a probability lower than 0.5 in equation 4. Thus, the probability of a speciation event is lower than the probability of an extinction event in any node in the phylogeny. By simulating multiple trees, we find trees which often become extinct during the first event. Bigger trees become more rare the greater the difference between λ and μ get. An illustrative example of such tree is given in figure 3.2. Notice that the tree completely became extinct.

Case 3. When the birth rate equals the death rate, $\lambda = \mu$, it yields us the same probability for speciation events and extinction events. So, the probability of a speciation event in a node is $\frac{1}{2}$

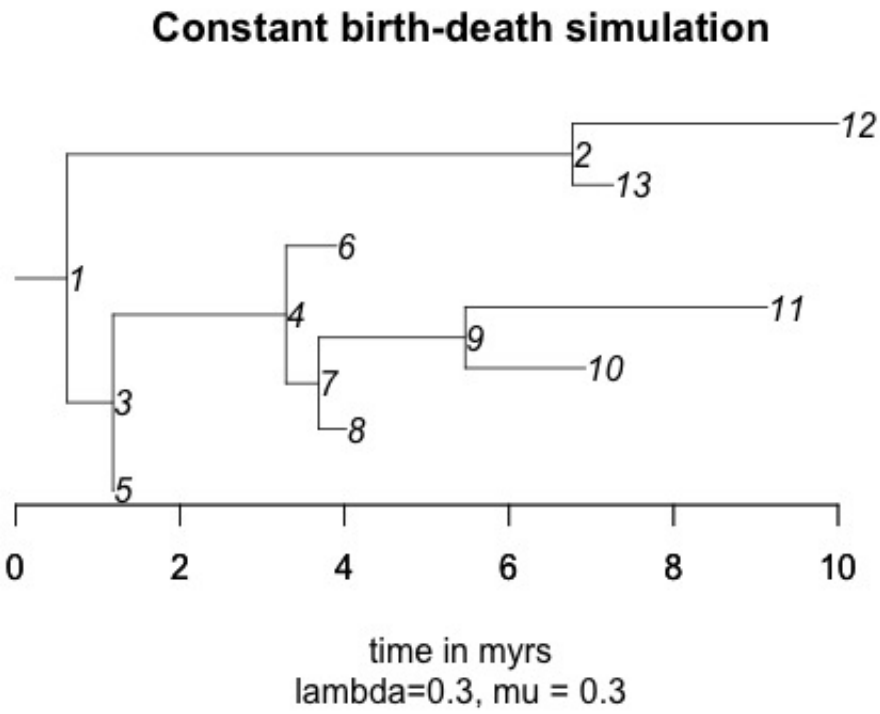

Figure 3.3: A plot of a constant birth-death simulation, with parameters $\lambda = \mu$ and $t_{\max} = 10$.

by equation 4. In this particular case, we see a wide range of types of trees in simulation studies. We see trees which become to extinct in the one hand, and trees which grow tremendously on the other hand. The simulations strongly suggests that this case produces a completely random process. An example of such tree is given in figure 3.3. This case is completely random, due to the same probability for speciation and extinction events. Therefore, I invite the readers to use the algorithm in appendix A.3 to make a few simulation themselves.

3.1.3 Inferences

For making inferences, we must have reconstructed phylogenies. We use the algorithm introduced in subsection 3.1.1, to obtain complete phylogenies. The phylogenies we use, however, are the reconstructed ones. That means, that all extinct species are removed from the phylogeny, as well as their original birth events. To obtain such a reconstructed phylogeny, we use the following algorithm:

1. Set a time where to "cut" the phylogeny. This time will be considered as the time today. For our inference purposes, we set the "cut time" as the time when the actual phylogeny reaches N lineages.
2. Collect all extinct species from the cutted phylogeny.
3. Draw the species which became extinct most recently.
4. - If its ancestor has another offspring which still is extant, we remove the birth event and the extinct species from the phylogeny. The offspring species is then the same species as its ancestor.

- If its ancestor has another offspring species which became also extinct, we remove both offspring species from the phylogeny. The event of the ancestor is changed from "birth" to "death".

5. Repeat steps $2-4$ untill there are no extinct species left.

This algorithm is found under the "pruning algorithm", which can be found in appendix A.3. An example of such a pruning process is graphically given below in figure 3.4.

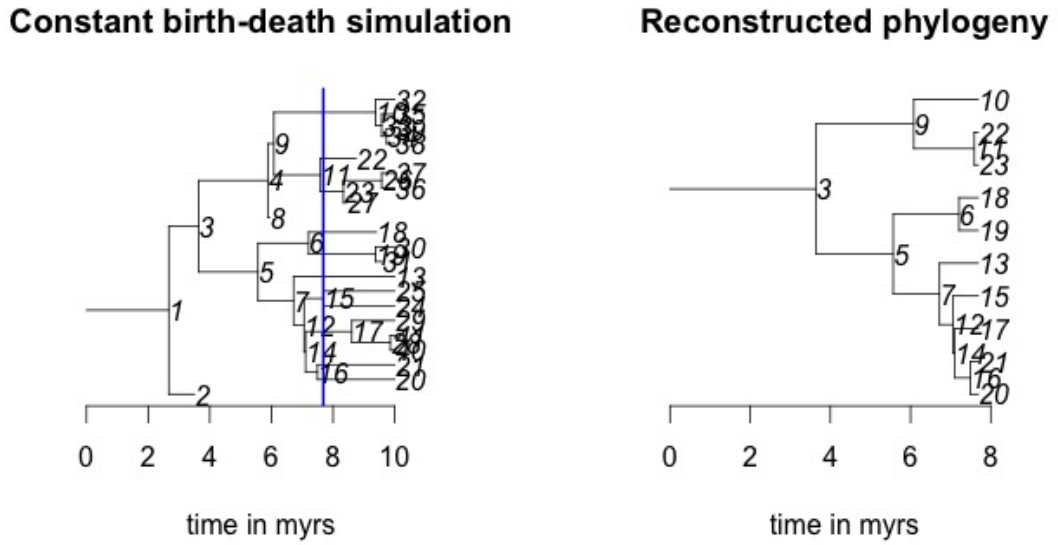

Figure 3.4: A plot of a constant birth-death process where the actual tree was cutted at the time that it reaches $N = 10$ lineages, which is given by the blue line.

Now, we want to use the reconstructed phylogeny to obtain estimates of λ and μ and compare it to the original values. Using the log-likelihood function obtained in equation 2.47, we can make contourplots of the maximum likelihood estimators for every single tree. First we have a look at the contour plots of trees, where we have cut at the moment when we had $N = (10, 25, 50, 100)$ lineages and compare the results. Thereafter, we will grow $10,000$ different trees to make a histogram and boxplot for both estimates of λ and μ.

Contourplots For Differently Sizes Reconstructed Phylogenies

The results for several values of N are given below. In the figure we can see the results for varying numbers of lineages N. The blue dot is the true values of the simulation, which is in this case $\lambda = 0.4$ and $\mu = 0.1$. The red dot is the maximum likelihood estimate of λ and μ.
To obtain the maximum likelihood estimator in this procedure, we create a grid for λ and μ and use the observed values in the tree to obtain the likelihoods for different values of λ and μ. The pair of values (λ, μ) which have the highest likelihood of the grid, is chosen as the maximum likelihood estimate. The contourplots for reconstructed phylogenies of size $N = 10, 25, 50, 100$ are given in figure 3.5.

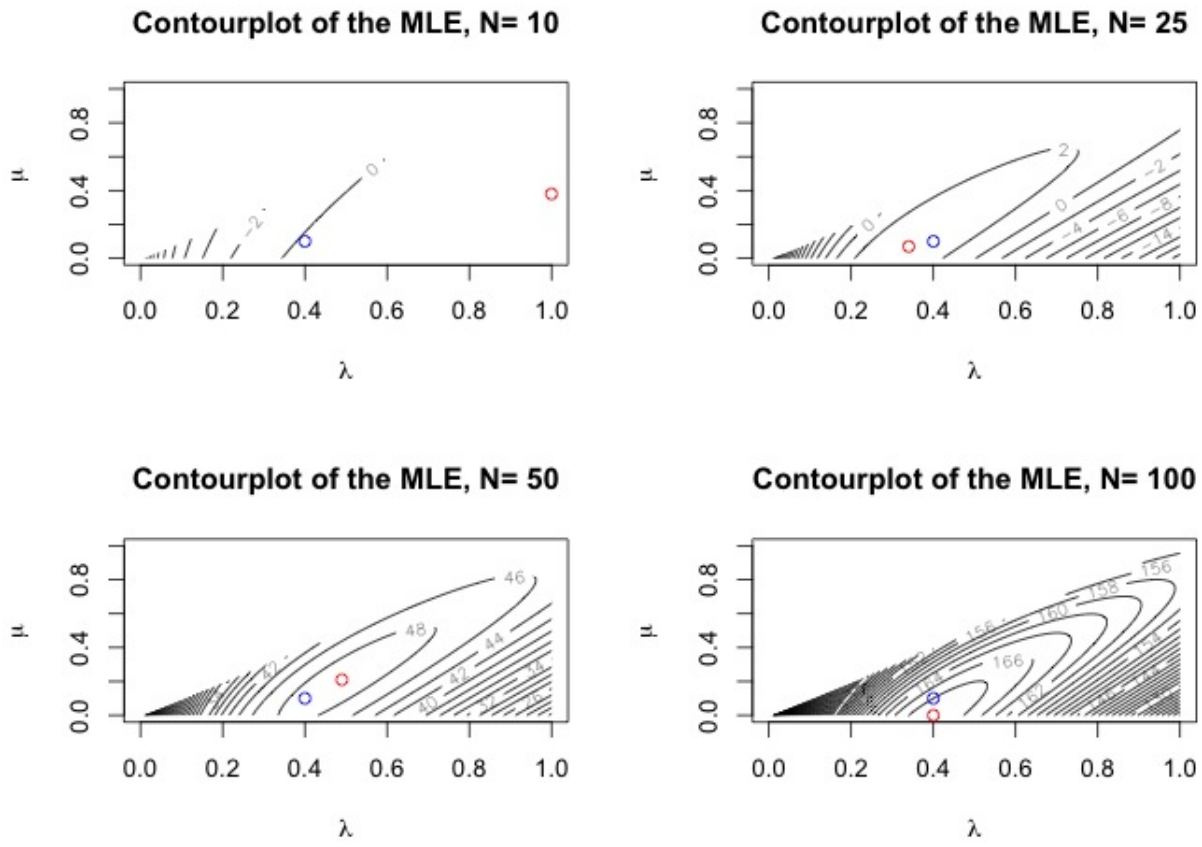

Figure 3.5: Contour plot of loglikelihood of the constant birth-death process with reconstruced clade size $N = 10, 25, 50, 100$.

We observe the following:

- the results that the maximum likelihood estimates using $N = 10$ are from poor quality, due to the big contour size.
- The bigger the number of lineages in the reconstructed phylogeny gets, the better the accuracy of the maximum likelihood procedure becomes. We can see this by the improving quality of the contour plots. The smaller the surface between highest contour line is, the better the estimation is.
- The maximum likelihood estimate $\hat{\mu}$ equals zero for the last estimations. We will see that this happens more often in the analysis of $10,000$ trees.

Histograms and Boxplots For Differently Sizes Reconstructed Phylogenies

In figure 3.6 - 3.9 we see the results of $10,000$ simulations for differently sized reconstructed trees, namely $N = 10, 25, 50, 100$. By doing this we want to investigate how many species we need in general for a reconstructed phylogeny, to make good inferences about the true value of λ and μ. In our simulation, we choose the parameter values $\lambda = 0.4$ and $\mu = 0.1$. Doing so, we obtain the following boxplots and histograms:

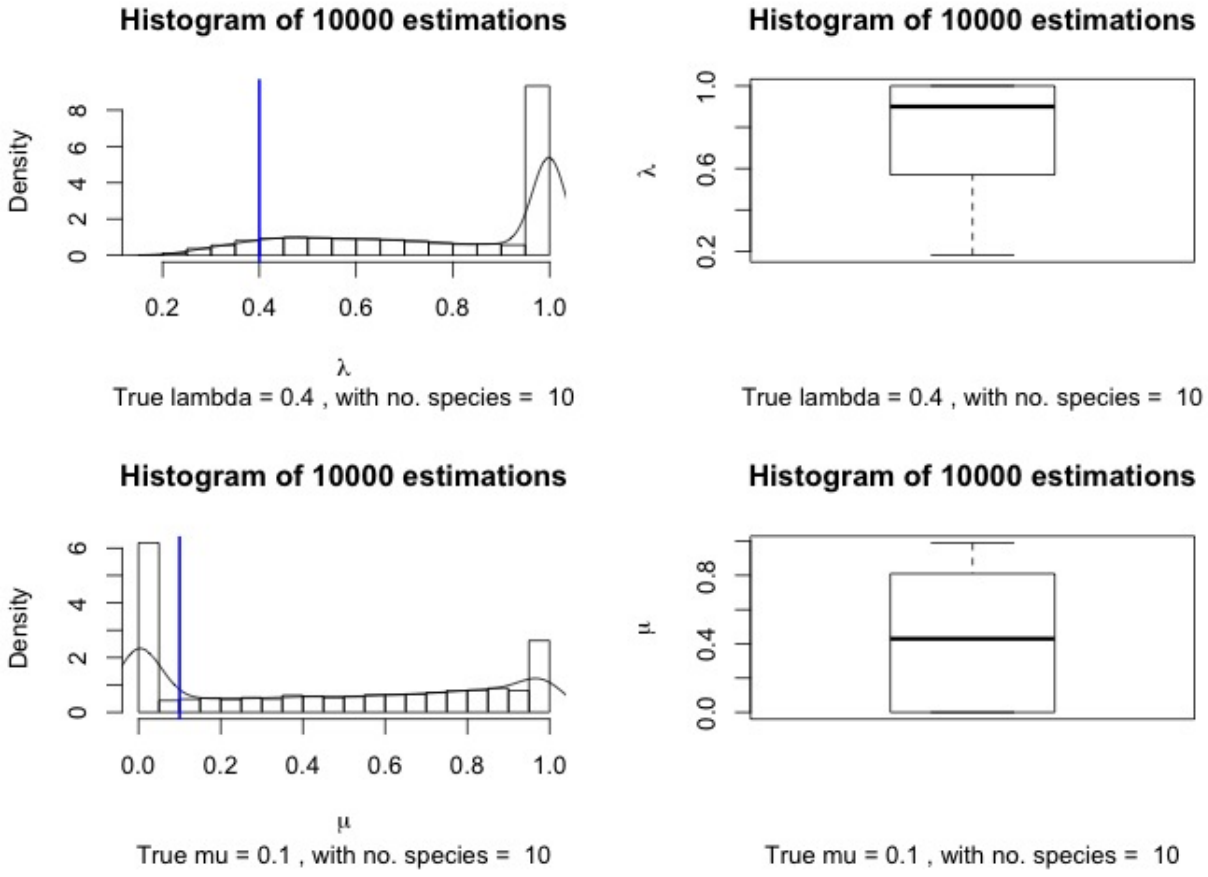

Figure 3.6: A plot of a constant birth-death process with reconstruced clade size $N = 10$.

We observe the following for $N = 10$:

- λ is most often estimated near the value 1 according to the histogram.
- The estimates $\hat{\lambda}$ of λ are in this case very poor, the spread of the estimates is according to the boxplot very large and the median lies far from the true value.
- μ is most often estimated near the value 0 according to the histogram.
- The estimates $\hat{\mu}$ of μ are in this case very poor, the spread of the estimates is according to the boxplot very large and the median lies far from the true value.

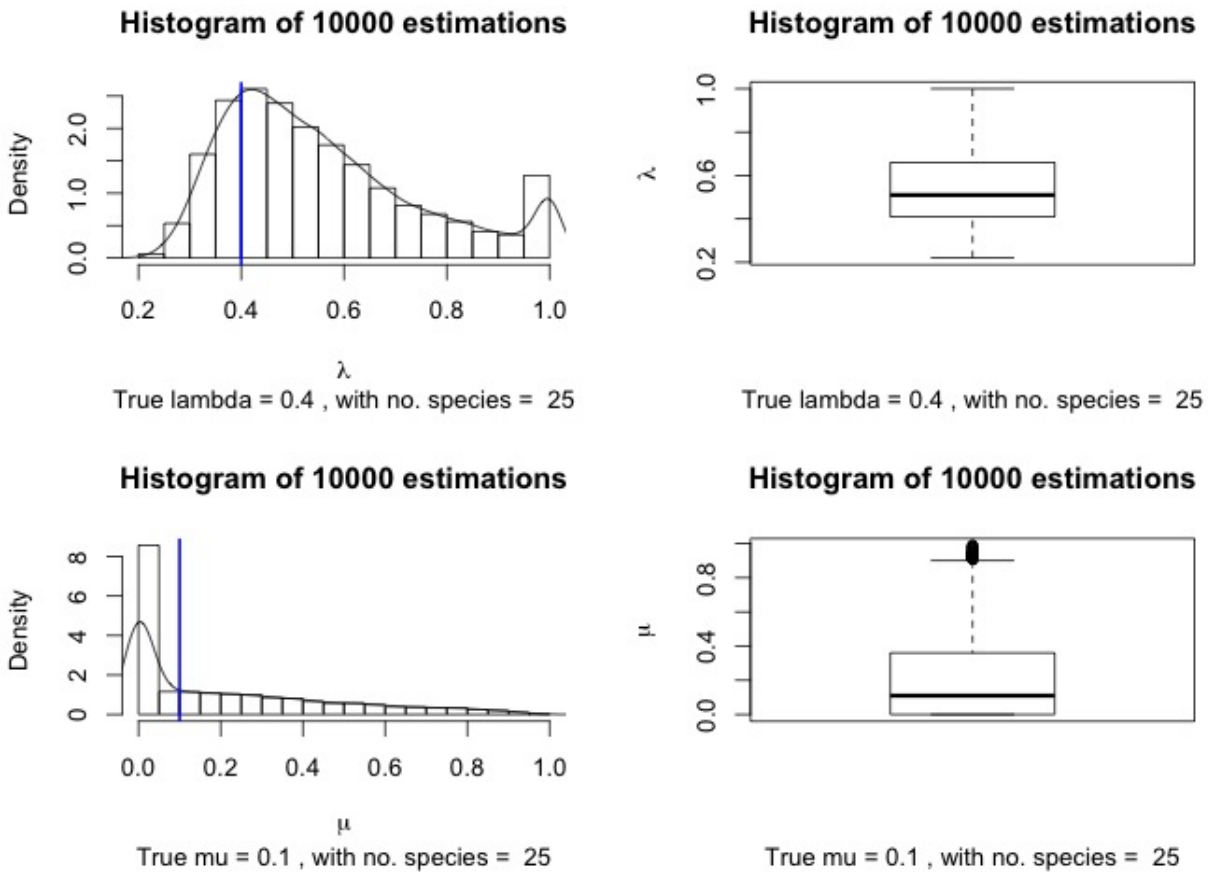

Figure 3.7: A plot of a constant birth-death process with reconstruced clade size $N = 25$.

We observe the following for $N = 25$:

- λ is most often estimated near the value 0.4 according to the histogram. Also, the values bigger than 0.4 are more often the estimate than values smaller than 0.4.
- The estimates $\hat{\lambda}$ of λ are in this case quite good, the spread of the estimates is according to the boxplot not so large and the median lies not far from the true value.
- μ is most often estimated near the value 0 according to the histogram.
- The estimates $\hat{\mu}$ of μ are in this case good, the spread of the estimates is according to the boxplot not so large and the median close to the true value.

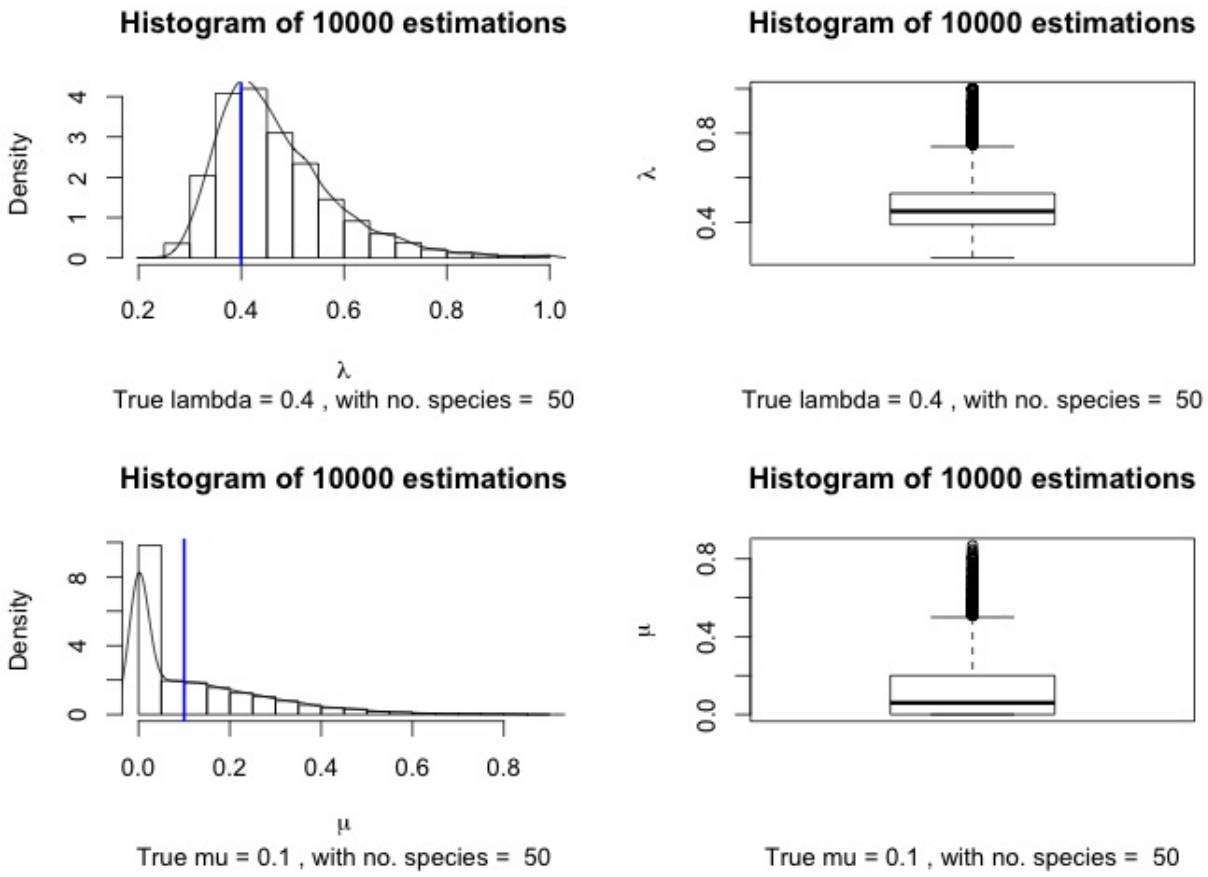

Figure 3.8: A plot of a constant birth-death process with reconstruced clade size $N = 50$.

We observe the following for $N = 50$:

- λ is most often estimated near the value 0.4 according to the histogram. Also, the values bigger than 0.4 are more often the estimate than values smaller than 0.4.
- The estimates $\hat{\lambda}$ of λ are in this case good, the spread of the estimates is according to the boxplot small and the median lies close to the true value.
- μ is most often estimated near the value 0 according to the histogram.
- The estimates $\hat{\mu}$ of μ are in this case good, the spread of the estimates is according to the boxplot small and the median close to the true value.

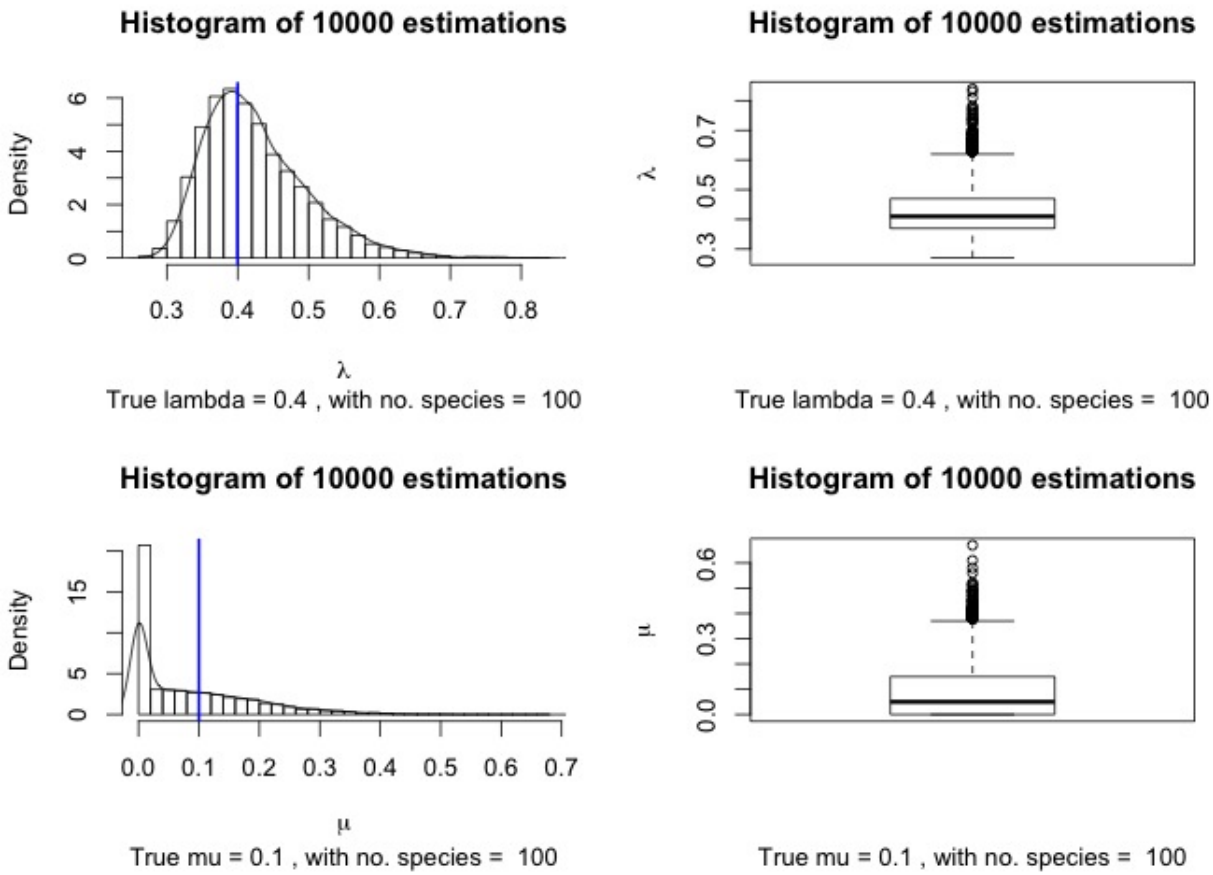

Figure 3.9: A plot of a constant birth-death process with reconstruced clade size $N = 100$.

We observe the following for $N = 100$:

- λ is most often estimated near the value 0.4 according to the histogram. Also, the values bigger than 0.4 are more often the estimate than values smaller than 0.4.
- The estimates $\hat{\lambda}$ of λ are in this case very good, the spread of the estimates is according to the boxplot very small and the median lies very close to the true value.
- μ is most often estimated near the value 0 according to the histogram.
- The estimates $\hat{\mu}$ of μ are in this case quite good, the spread of the estimates is according to the boxplot very small and the median quite close to the true value. However, the big number of estimates of $\mu = 0$ cause a shift of the median of the estimates. Therefore, the estimate of μ is not as good as the estimate of λ.

3.2 Simulation of a Protracted Birth-Death Model

We make simulations of a protracted birth-death process by means of the Gillespie Algortihm (Gillespie 1976). For fulfilling the algortihm, we first point out the basic properties of the protracted birth-death model. After this, we introduce an algorithm to obtain a phylogeny.

3.2.1 Basic Properties

In a protracted birth-death model, we suppose that speciation, completion and extinction are processes that proceed simulataneously. The process that finished first, is the process that we observe. We first give an outline of important properties of the protracted birth-death process.

The Event Times

We first make observations for incipient species and good species seperately. Let T_g denote the event time of a good species, and T_i denote the event time of an incipient species. The first one can be seen as the minimum time when the birth of an incipient species or death occurs. The second one is the minimum time when the completion of the species, a birth of another incipient species or death occurs. For these event times of these species we observe the following:

Property 7 *The event times are exponentially distributed, i.e.:*

$$T_g \sim \exp(\lambda_1 + \mu)$$
$$T_i \sim \exp(\lambda_2 + \lambda_3 + \mu)$$

Proof This follows directly from property 4. ∎

Now suppose we have at a time t exactly N_g good species and N_i incipient species. Then we directly know by theorem 4 and lemma 7 that the time untill the next event, denote T_e, is also expenontially distributed, i.e.:

$$T_e \sim \exp(N_g(\lambda_1 + \mu) + N_i(\lambda_2 + \lambda_3 + \mu)) \tag{3.3}$$

In Gillespie's algorithm, they work with random variables which are standard uniformly distributed. Thus, to draw a new event time by means of a uniform random variable, we use the following theorem.

Theorem 1 *Let $X \sim U(0,1)$, and let $R = N_g(\lambda_1 + \mu) + N_i(\lambda_2 + \lambda_3 + \mu)$. Then:*

$$Y = -\frac{1}{R}\log(f(X)) \sim \exp(R) \tag{3.4}$$

Proof We simply constuct the cumulative density function of Y:

$$\begin{aligned}\Pr(Y \leq y) &= \Pr(-\frac{1}{R}\log(f(X)) \leq y) \\ &= \Pr(\log(f(X)) \geq -Ry) \\ &= Pr(X \leq -e^{Ry}) \\ &= 1 - e^{Ry}\end{aligned}$$

Which is the cumulative density function of an exponetial random variable with parameter R. ∎

By the result of theorem 1, we can use uniform random variables in the Gillespie algorithm to determine our next event times, which is simply the old event time plus the time untill the next event:

$$t_{i+1} = t_i - \frac{1}{R}\log\left(f(X)\right), \quad X \sim U(0,1) \tag{3.5}$$

The Events

We first determine the different actions which can occur in the protracted birth model. We assume that we have N_g good species and N_i incipient species. Then the different possble actions of the model are:

1. A good species giving birth to an incipient species, with a rate $\lambda_1 N_g$.
2. A good species going extinct, with a rate μN_g.
3. An incipient species reaching completion, with a rate $\lambda_2 N_i$
4. An incipient species giving birth to a new incipient species, with a rate $\lambda_3 N_i$
5. An incipient species going extinct, with a rate μN_i.

Suppose now that we have drawn a new event time, we are now interested in the probabilities that a certain event occurs. We denote the events by the corresponding numbers in enumeration given before. Let E denote the random event. We obtain straight fowardly by using the R defined earlier:

$$\Pr(E=1)=\frac{\lambda_1 N_g}{R},\quad \Pr(E=2)=\frac{\mu N_g}{R},\quad \Pr(E=3)=\frac{\lambda_2 N_i}{R}$$
$$\Pr(E=4)=\frac{\lambda_4 N_i}{R}\quad \Pr(E=5)=\frac{\mu N_i}{R} \tag{3.6}$$

Now we draw again a uniform random variable Y $U(0,1)$. We sum up the probabilities given in 3.6 beginning at $i=1$ untill the summation exceeds the random variable Y. The last i added to the summation is then event occuring at the next event time. So, mathematically denoted:

$$\min\left(i\in\{1,2,3,4,5\}|\sum_{j=1}^{i}\Pr(E=i)>Y,\quad Y\sim U(0,1)\right) \tag{3.7}$$

Suppose now that we have drawn an event. Then all the species which may make such event, are equally likely to be the actual species which is contributed to the event. This is a result of the memoryless property of the exponential random variables:

Property 8 *Let $X\sim\exp(\theta)$, then X is memoryless, i.e.:*

$$\Pr(X>s+t;X>t)=\Pr(X>s) \tag{3.8}$$

Proof

$$\begin{aligned}\Pr(X>s+t;X>t)&=\frac{\Pr(X>s+t)}{\Pr(X>t)}\\&=\frac{\exp(-\theta(s+t))}{\exp(-\theta t)}\\&=\exp(-\theta s)\\&=\Pr(X>s)\end{aligned}$$

As desired. ∎

Thus, in our simulation: If we know what event will happen, the species which is contributed to this event can be sampled uniformly from all possible species which can be contributed to this event. Thus, if we know that we have a birth of an incipient species by an incipient species: every incipient species that already existed has the same probability of being the ancestor of the newborn incipient species. And second, if we have the birth of an incipient species by a good

species, every good species that already existed has the same probability of being the ancestor of the newborn incipient species. Notice that the birth of a good species is actually a completion event, its ancestor was already determined when it was an incipient species. This sampling idea for ancestor is a reversed reasoning of the memoryless property of exponential variables. So, given that we have a birth of a new incipient species caused by an good species, the probability that a certain good species is the ancestor of the new incipient species is $\frac{1}{N_g}$. The same holds for incipient species giving birth to a new incipient species, with equal probabilities $\frac{1}{N_i}$.

Algorithm

As said before, the algorithm is constructed as a Gillespie algorithm. We give pseudo-code of the algorithm in this section. In Appendix A.4 the original code can be found.

1. Set $\lambda_1, \lambda_2, \lambda_3$ and μ. Also set the maximum time T. The initial number of good species $N_g = 1$ and the number of incipient species $N_i = 0$.
2. Calculate the parameter $R = N_g(\lambda_1 + \mu) + N_i(\lambda_2 + \lambda_3 + \mu)$. Update the probabilities of the events.
3. Draw two uniform random variables X and Y.
4. Obtain the next event time by equation 3.5.
5. Determine the event by equation 3.7
6. If the event is
   - $E = 1$, draw a random ancestor from the list of good species. Set the origin time of the incipient species as t, the rest is not determined. Update the list of incipient species.
   - $E = 2$, draw a random species from the list of good species. Set the extinction time of the species as t, the rest is not determined. Update the list of good species.
   - $E = 3$, draw a random species from the list of incipient species. Set the completion time of the species as t, the rest is not determined. Update the list of good species and the list of incipient species.
   - $E = 4$, draw a random ancestor from the list of incipient species. Set the origin time of the new incipient species as t, the rest is not determined. Update the list of incipient species.
   - $E = 5$, draw a random species from the list of incipient species. Set the extinction time of the species as t, the rest is not determined. Update the list of incipient species.
7. If the number of incipient species or the number of good species is bigger than zero, and the time is smaller than the maximum time, repeat steps $2 - 6$.

An example of a protracted birth-death process is given in table 3.2.1, given below.

As stated before, we use the developments in the process in time to infer the phylogenetic tree. Thus, the data we use in our loglikelihood function is given in table 3.2.1.

Ng	Ni	time	id
1	0	0.00000	1
1	1	0.65638	-2
1	2	0.80777	-3
1	3	0.82079	-4
1	4	0.84062	-5
1	3	0.87014	-3
1	4	1.06914	-6
2	3	1.12314	2
3	2	1.21148	5
4	1	1.33315	6
3	1	1.56176	2
3	2	1.85875	-7
3	3	2.09510	-8

Table 3.2: Data frame of a Protracted birth-death process. The first column denotes the number of good species, the second colum the number of incipient species, the third colum the event time and the fourth column the identity of the species which is subject to an event.

dNg	dNi	time	id
0	1	0.65638	-2
0	1	0.80777	-3
0	1	0.82079	-4
0	1	0.84062	-5
0	-1	0.87014	-3
0	1	1.06914	-6
1	-1	1.12314	2
1	-1	1.21148	5
1	-1	1.33315	6
-1	0	1.56176	2
0	1	1.85875	-7
0	1	2.09510	-8

Table 3.3: Data of the Protracted birth-death process in table 3.2.1. Every row represents the difference between the states after an event.

3.2.2 Approximation of the Process

Before we go any further, we check wheter we can really estimate the protracted birth-death process with Gaussian processes. In order to do so, we made one simulation of a protracted birth-death process, and let Gaussian process develop simultaneosly. The obtain a 95% confidence interval per event time of the protracted birth-process. We observe, that the protracted birth-death process, which is the black line, moves between the red and blue dots, which are respectively 2.5% and 97.5

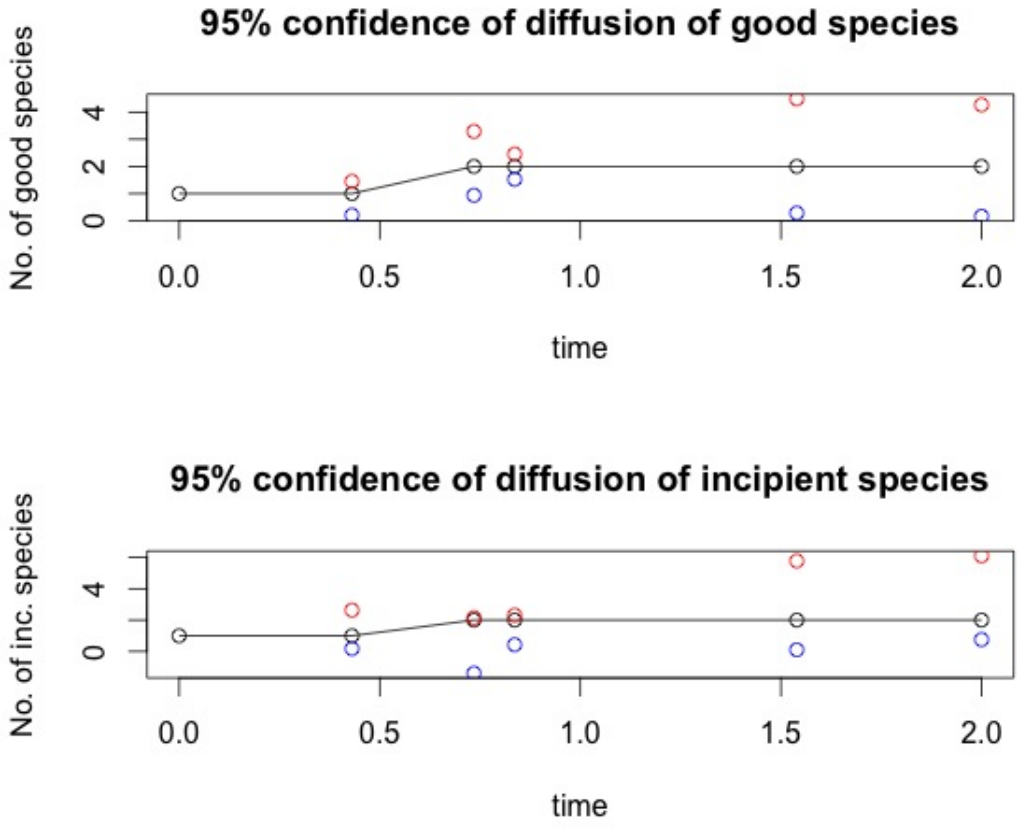

Figure 3.10: A plot of the protracted birth-death process (the black line), in a 95% confidence region for Gaussian Processes.

3.2.3 Inferences

Using the likelihood obtained in equation 2.62, we simulate $1,000$ trees and obtain a maximum likelihood estimate of the parameters, given the data. Since the loglikelihood can't be solved analytically, we use the *optim* algorithm in R to optimize the loglikelihood by changing the parameters. By repeating this for $1,000$ different trees, we obtain the following histogram with estimates of Λ.

We can see in this figure that the true values of Λ are close to the tops of the histograms, except for λ_1. This can be the case, because we have grown differently sized trees. Therefore we also obeserve trees which have grown to a fixed total number of species $n = N_g + N_i$, for $n = 10, 25, 50, 100$. We observe in these histograms, that an increasing tree size results into a biases for all of the four maximum likelihood estimates (figure 3.12, figure 3.13, figure 3.14 and figure 3.15.

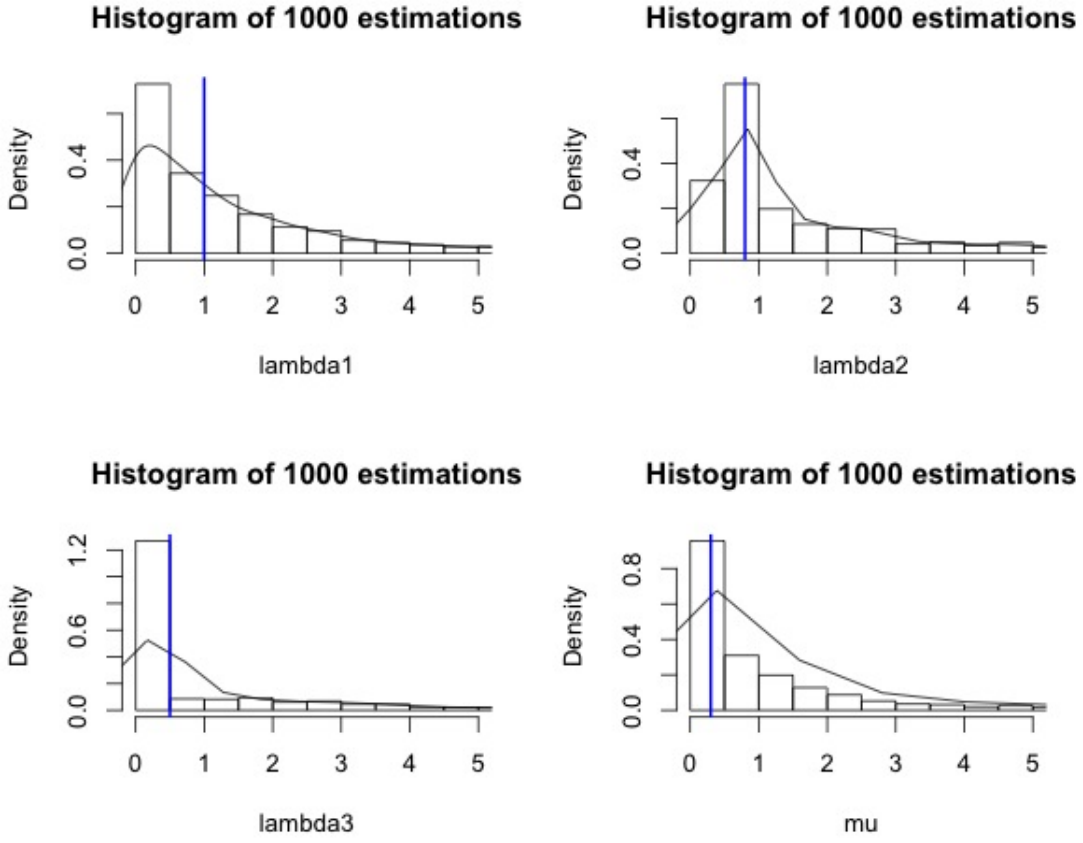

Figure 3.11: A histogram of $1,000$ trees of the four maximum likelihood estimates of the protracted birth-death process .

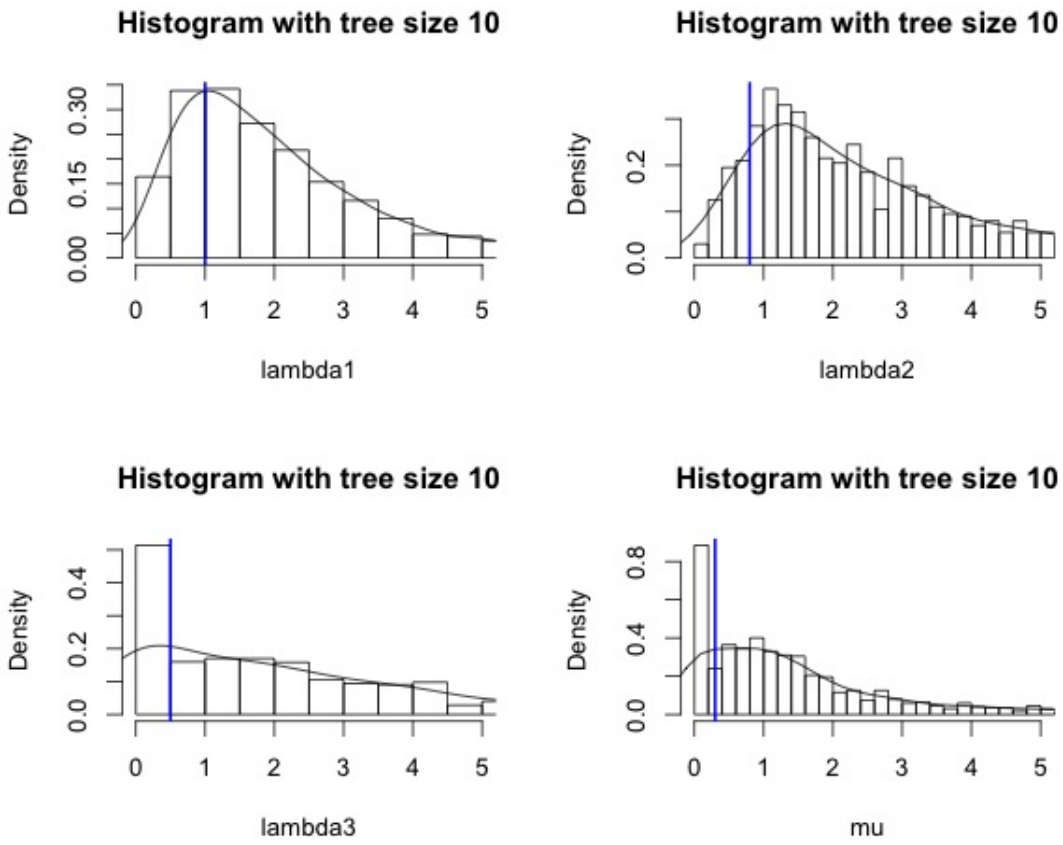

Figure 3.12: A histogram of $1,000$ trees with tree size $n = 10$, for the maximumlikelihood estimates of Λ.

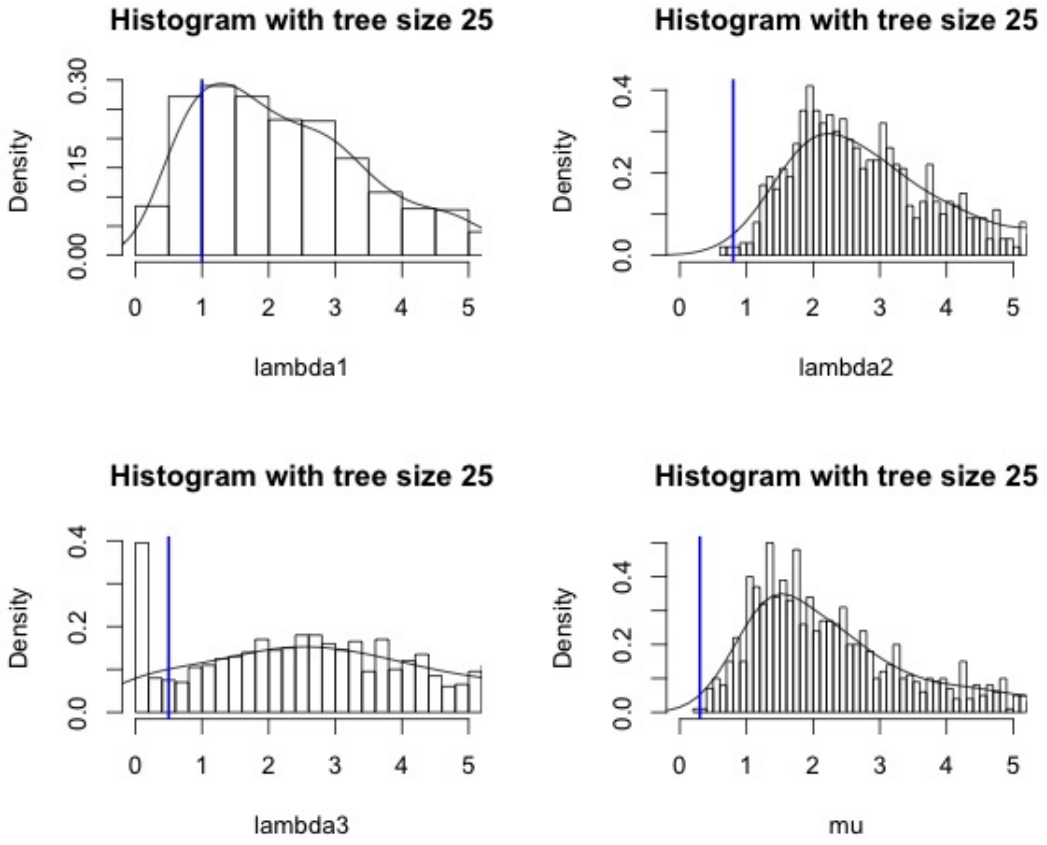

Figure 3.13: A histogram of $1,000$ trees with tree size $n = 25$, for the maximumlikelihood estimates of Λ.

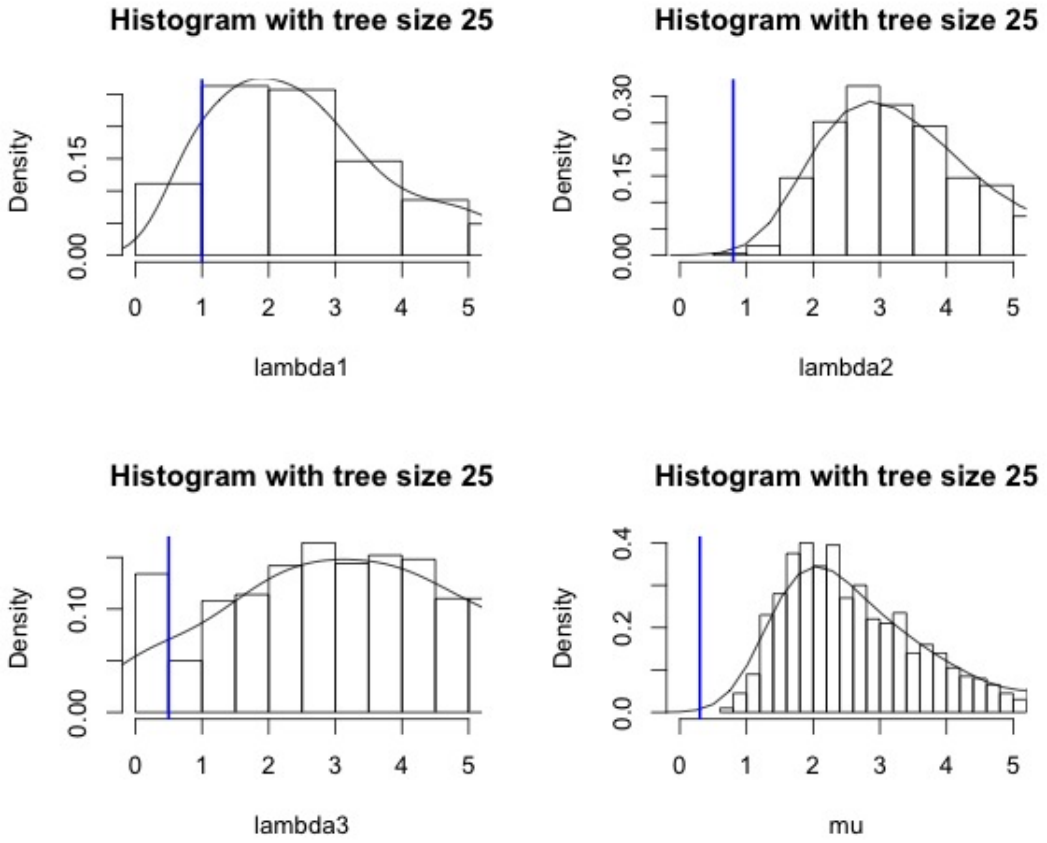

Figure 3.14: A histogram of $1,000$ trees with tree size $n = 50$, for the maximumlikelihood estimates of Λ.

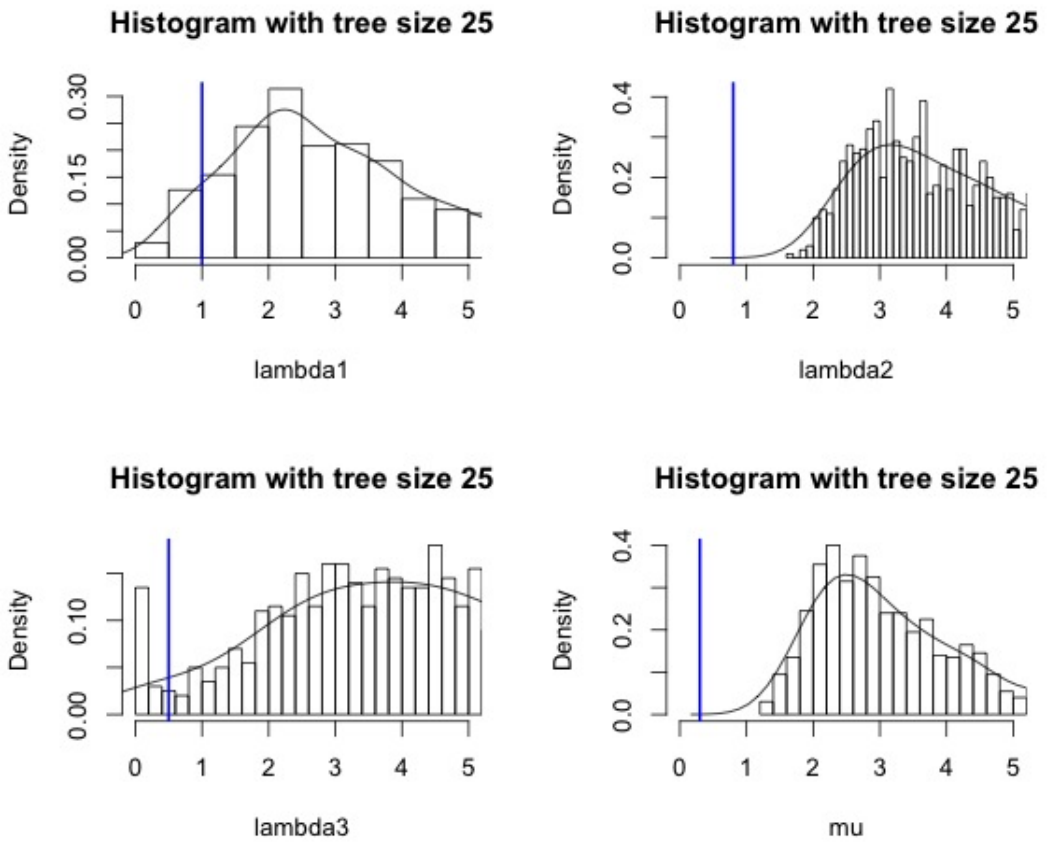

Figure 3.15: A histogram of $1,000$ trees with tree size $n = 100$, for the maximumlikelihood estimates of Λ.

Chapter 4

Conclusion and Discussion

4.1 Conclusion

We can observe for the constant birth-death model, that:

- the maximum likelihood estimation of λ and μ are good for the cases when the number of lineages in the reconstructed phylogeny are greater of equal to $N = 25$. Using only 10 lineages, we obtain a very poor and often wrong estimates of the birth- and death rate.
- we see in these procedures that the estimate of μ is often zero. This result is the same for different values of μ, what can be easily checked by the algorithm given in Appendix A.3.

Thus, we can use reconstructed phylogenies to make estimates about birth and death rates. However, in the reconstructed phylogenies we obtain in real life, we observe the following:

- For a constant birth-death process, we expect the logarithm of number of lineages to be linear mostly through time. In the beginning of the process, we deal with the "push of the past", which means that the number of lineages are greater than expected in the beginning of the phylogeny. We can explain this by the fact that the success of a phylogeny is often relied on a "flying start".
- At the end of the process, the number of lineages accelerate because the species had less time to go extinct. So in the graph, we observe an accelerating number of phylogenies. This phenomenon is known as the "pull of the present". However, in real reconstructed phylogenies from genetic data, we observe the number of lineages to slow down instead of accelerating. This can be explained, however, that we only have a sample of a phylogeny.

For the protracted birth-death model, the protracted birth-death process can be explained as a Gaussian process. We checked this by means of a 95% confidence interval for the Gaussian process. Since the protracted birth-death process remains within this confidence interval, this justifies our procedure.

An approximation of the likelihood using Gaussian processes gives us maximum likelihood estimates of the parameters $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \mu)$, but the results show that these estimates are biased for large sets of data. However, for smaller data sets and mixed data sets, the estimates of these parameters are close to the true values of the parameters. If we look more closely to λ_2, we see that the parameter is often estimated near its true value. This means, since λ_2 is responsible for the duration of speciation, that we can estimate the duration of speciation using smaller data sets.

4.2 Discussion and Impossible Improvements

The constant birth-death model may not be a sufficient model to describe evolutionay processes. The protracted birth-death model can explain the diversification slowdowns (Etienne and Rosindell 2012). But, the exact likelihood of the protracted birth-death model has not been obtained yet.

Using the approximation of the loglikelihood of the protracted birth-death process, we were able to obtain maximum likelihood estimates of good quality for small phylogenies. The bias in the estimates using bigger phylogenies, can be explained by the continuity assumption of the Gaussian process. The original process is not a continuous process, but the procedure we use to approximate is continuous. Also, remember that this maximum likelihood procedure is in fact "an approximation of an approximation". The accuracy of this method is therefore, lower than a procedure which used direct approximations.

Further, we can extend the Gaussian process approximation to reconstructed phylogenies. This well be a further improvement to the existing approximation. Secondly, and most difficult, is finding the exact likelihood of the general protracted birth-death process.

Bibliography

[1] J.C. Avise, D. Walker, G.C. Johns. 1998. *Speciation durations and Pleistocene effects on vertebrate phylogeography. Proc. R. Soc. Lond. Biol. Sci.*, 265:1707-1712

[2] B.G. Baldwin, M.J. Sanderson. 1998. *Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc. Natl. Acad. Sci. USA*, 95:9402-9406

[3] A.J. Dobson and A.G. Barnett. 2008. *An Introduction to Generalized Linear Models* (3rd ed.). Chapman & Hall/CRC, pp. 77-78

[4] R.S. Etienne and J. Rosindell. 2012. Prolonging the Past Counteracts the Pull of the Present: Protracted Speciation Can Explain Observed Slowdowns in Diversification *Syst. Biology*, 61(2) : 204-213

[5] M. Foote. 1988. Survivorship analysis of Cambrian and Ordovician trilobites. *Paleobiology.* 14:258-71

[6] S.J.. Gould, D.M. Raup, J.J. Sepkoski Jr., T.J.M. Schopff, D.S. Simberloff 1977. The shape of evolution: a comparison of real and random clades. *Paleobiology.* 3:23-40

[7] S.J.. Gould, N.L. Gillinsky, R.Z. German. 1987. Asymmetry of lineages and the direction of evolutionary time. *Science.* 236:1437-41

[8] D.T.. Gillespie. 1976. A General Method for Numerically Simulating the Stochastic Time Evolution of Cuopled Chemical Reactions *J. Comput. Phys.*, 22 : 403-434

[9] J. Hey. 1992. Using phylogenetic trees to study speciation and extinction. *Evolution*, 46 : 627-40

[10] S.P. Hubbell. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography* Princeton University Press, Princeton.

[11] D.G. Kendall. 1948 On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* 35: 6-15.

[12] Picture of Kendall: `http://en.wikipedia.org/wiki/David_George_Kendall#mediaviewer/File:David_Kendall.jpg` (26-6)

[13] J.K. Kitchell, N. Macleod. 1988. Macroevolutionary interpretations of symmetry and synchroneity in the fossil record. *Science* 240:1190–9

[14] M.A.. McPeek. 2008. The ecological dynamics of clade diversification and community assembly. *Am. Nat* 172 : 270-284

[15] D. Moen and H. Morlon. 2014. Why does Diversification Slow Down? *Trends in Ecology & Evolution* 29 : 190-197

[16] A.O. Mooers, S.B. Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.*, 72: 31–55

[17] A.O. Mooers, L.J. Harmon, M.G.B. Blum, D.H.J. Wong, S.B. Heard. 2007. Some models on phylogenetic tree shape. In: *Reconstr. Evol. New Math. Comput. Adv.* Oxford University Press, Oxford, pp. 149-170

[18] P.A.P. Moran. 1958. Random processes in genetics *Proc. Camb. Philos. Soc.* 54 : 60-71

[19] H. Morlon. 2014. Phylogenetic Approaches for Studying Diversification. *Ecology Letters*, 17: 508–525

[20] S. Nee, E.C. Holmes, R.M. May and P.H. Harvey. 1994. Extinction rates can be estimated from molecular phylogenies *Phil. Trans. R. Soc. Lond. B*, 344 : 77-82

[21] S. Nee, E.C. Holmes, R.M. May, P.H. Harvey. 1995a. Estimating extinction from molecular phylogenies. In *Extinction rates*. Oxford University Press, Oxford, pp. 164-82.

[22] S. Nee, E.C. Holmes, R.M. May, P.H. Harvey. 1995b. The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond*. B 344 : 77-82.

[23] S. Nee. 2001. Inferring Speciation Rates From Phylogenies. *Evolution*, 55(4), pp. 661–668.

[24] S. Nee. 2006. Birth-Death Models in Macroevolution. *Annu. Rev. Ecol. Evol. Syst*. 37:1–173.

[25] E. Paradis. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. *Proc. R. Soc. Lond. B*. 264:1141–1147.

[26] A.L. Pigot, A.B. Phillermore, I.P.F. Owens, C.D.L. Orme. 2010. The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Syst. Biol.*. 59:660–673.

[27] I. Pinelis. 2003. Evolutionary models of phylogenetic trees. *Proc. R. Soc. London Ser. B*. 270:1425–31.

[28] IA. Purvis, C.D.L. Orme, N.H Toomey, P.N. Pearson. 2009. Temporal patterns in diversification rates In: *Speciat. Patterns Divers*. Cambrigde University Press, Cambridge, 278-300.

[29] D.L. Rabosky. 2006. Likelihood Methods for Detecting Temporal Shifrs in Diversification Rates. *Evolution*, **60**(6), pp. 1152-1164.

[30] D.M. Raup, S.J. Gould, T.J.M Schopff, D.S. Simberloff. 1973. Stochastic models of phylogeny and the evolution of diversity. *J. Geol*. 81:525–42.

[31] J.B. Slowinski, C. Guyer. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *Am. Nat.*, 142, 1019 -1024.

[32] T. Stadler. 2012. How Can We Improve Accuracy of Macroevolutionary Rate Estimates? *Syst. Biology*, **62**(2), pp. 321-329

[33] M.D. Uhen. 1996. . An evaluation of clade-shape statistics using simulations and extinct families of mammals. *Paleobiology*, 22:8-22

[34] G.U. Yule. 1924. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis,FRS. *Philos. Trans. R. Soc. London Ser*. B 213:21-87

[35] Picture of Yule: `http://www.apprendre-math.info/history/photos/Yule_2.jpeg` (26-6)

Appendix A

R codes

A.1 Review: Other Models and Biological Properties

There are various other models for describing evolutionary processes. We describe the main importances of these models, which is an overview of the work of Morlon (2014). Thereafter, we describe the diversification slowdown in reconstructed phylogenies, based on the work of Moen and Morlon (2014).

A.1.1 Moran Process

Suppose that we require a model for clades that have reached a limit of diversity. A useful model for this was introduced by Moran in 1958. At each point in time, each lineage has a probability of going extinct and when a lineage does go extinct, it is replaced by a progeny of other lineage chosen at random, such that the clade size remains constant. In simulation studies of Raup et al. (1973) and Gould et al. (1977) they decided to:

- Set $\lambda = \mu$ whenever the diversity in the clade hits the limit, to hold it at this point.
- Set $\lambda > \mu$ whevever the diversity gets below the limit, to get it back to the top.
- Set $\lambda < \mu$ whenever the diversity gets above the limit, to drop the diversity back to the plateau.

The pure brith process ($\mu = 0$) is statistically superior description of the data than the Moran process (Hey 1992). However, Nee argues that we should keep the Moran process as a component of our arsenal (Nee 2001).

A.1.2 Random Walk

The difference between the Moran process and a birth-death process with $\lambda = \mu$, is that the total size of the clade is in the Moran process constant and in the constant birth-death process it follows a random walk.

Definition (Simple random walk) Let $Z_1, Z_2, \ldots$ be independent random variables where $\Pr(Z_i = 1) = \Pr(Z_i = -1) = 0.5$, for all $i = 1, \ldots, n$. Set $S_0 = 0$ and define $S_n = \sum_{i=1}^{n} Z_j$. Then the series $\{S_n\}$ is called a *simple random walk.*

Obviously, if we let:

- k denote the k-th event (which is birth or death)

- $Z_k = 1$ denote the whether a species has speciated and $Z_k = -1$ when a species has become extinct
- $S_n = \sum_{k=1}^{n} Z_k$ the number of species in the phylogeny

Then S_n follows a random walk.

A.1.3 Shapes of clades

From birth-death processes with equal probabilities λ and μ, which is a random walk, we obtain that the form of this clade is symmetric with respect to time. However, clades arising early in time tended to be bottom-heavy (Gould et al. 1987). Bottom-heavy means an strong increase in the beginning of the clade and ending with a weaker decrease to extinction. The more recent clades tended to be symmetrical on anverage. This means that diversification is asymmetrical with respect to time, unlike the random walk model. Kitchell & Macleod (1988) critized this results, because the used statistic (the center of gravity) is quite likely to arouse by randomness only. However, this critique is not entirely definitive by a performed regression analysis. However, the result seems to be extremely sensitive to the data used, in a reanalysis exactly the opposite pattern was obtained (Uhen 1996).

Analogous to the analysis of clade shapes is the investigation of tree balance. There are various statistics and null models for studying the tree belance, the only null model meaningful in a macroevolutionary context is the "equal rates Markov" (EMR) model. The usual description of the EMR model is equivalent to the pure birth process, each lineage is as likely to give rise to a new lineage as any other. So whevener a new lineage appears it arises from an existing lineage, chosen at random.

The analysis of the tree shape consists of (Nee 2006):

- defining a tree balance statistic
- calculating the distribution of the statistic under the EMR model
- comparing the statistic's value in a real tree with this distribution to see if it is usual

The general result is that real trees are more unbalanced than expected under the EMR model (Mooers & Heard 1997, Pinelis 2013). The conclusion is that there is heterogeneity among lineages in their propensity to diversify.

If a tree has grown according to a pure birth process, then the number of daughter lineages for each parental lineage is expected to have a one parameter geometric distribution. For this result, the birth rate does not have to be constant (Nee et al. 1994). The birth rate only has to be the same for each lineage over time.

A.1.4 Sampling and paraphyly

Thus far, we have assumed that we have all the members of the clade in our analysis. However, this is in general not the case which means that our analysis is only a sample of the true clade. Paleontological data is often incomplete as well and the exclusion of a lineage and all its descendants from a clade on the basis of the possession of some caracteristic is know as a paraphyly. Molecular phylogenies that are based on a simple random sample of a clade that has grown according to a birth or birth-death process, will give the missleading impression that the rate of cladogenisis has been slowing down instead of accelerating. This effect arises because lineages that have arisen in the recent past are likely to have fewer progeny than older lineages, and therefore less likely to have any progeny lineages in the random sample. This effect is striking to phylogenetic trees of virusses which have been spreading exponentially. Sampling has no qualitive effect on trees that have grown according to a Moran process (Nee 2001).

A.1.5 Diversification models

Diversification, the balance between speciation and extinction is central to one of the most fundamental questions in ecology: 'How is biodiversity generated and maintained?' Diversification is a key to understand how biodiversity varies over geological time scales (citaties), how it is distributed across the Earth's surface, the tree of life and ecological communities . Diversification is also a primary predictor of three fundamental patterns in macrobiology (Morlon 2014).

- the species abundance distribution, which describes how individuals are partioned among species
- the species-area relationship, which describes how species richness increase with geographical area
- the distance-decay relationship, which describes how community similarity declines with geographical distance (citaties)

Diversification rates are thus some of the most important parameters in marcroevolution, macroecology and community ecology. Diversification is particularly hard to study, because speciation and extinction processes typically happen on a time scale of thousands to millions of years. While estimating diversification rates from fossil data is feasible for some groups, it is not feasible for the majority for extant groups on earth.

The shortage of the fossil record led to the development of alternative approaches to study diversification, themselves inspired from paleontological models. Phylogenies are branching trees that represent the evolutionary relationship among species and they contain information about past diversification events. The phylogenetic trees of extant species, referred as 'reconstructed phylogenies' can be inferred using molecular data. These trees can be used along with various statistical models to draw inferences about diversity dynamics. Phylogenetic methods have become a prevailing approach for studying diversification (Morlon 2014).

Phylogenetic approaches for studying diversification focus on two main apsects of phylogenetic trees:

- branching times: which require phylogenetic branch lengths to be in units relative to time
- topology: the shape of the phylogenetic tree

Phylogenetic approaches to understanding diversication rely on a common principle: comparing empirical phylogenies to phylogenies obtain under various models of diversification. These models can be classified into models where:

- Species are the unit of diversification, without any reference to individuals, population sizes or geographical ranges.
- The dynamics of individuals, population sizes or speices ranges are considered explicitely.

The various models can further be classified with respect to whether diversification is assumed:

- time-constant or time-varying
- homogeneous or varying across lineages
- instantaneous or protracted

We now give some examples of different models of diversification.

Homogeneous, time-constant diversification

In the simplest model, also referred as the equal rates model, diversification is modelled as a birth-death process in which species either give birth to new species, or become to extinct with constant rates λ and μ respectively. We assume λ and μ to be constant across all species, such that all species have equal diversification rates. Under this model, $\lambda > \mu$ typically and the clade diversity increases exponentially through time. Equal rates models are still widely used, serving in particular as null models of diversification.

Time-varying diversification

Time-constant models are useful, but there are many reasons why diversification rates can vary over time. A straightforward and widespread approach to account for time variation in diversification rates is to assume functional dependence of speciation and extinction rates with time. These time-dependent models allow a quantitative estimation of how diversification varied through time.

Environmental dependence

The main drivers of temporal variations in diversification rates are modifications in the abiotic and biotic environment (Morlon 2014). In this type of modelling, we assume a functional dependence of diversification rates on the environment.

Diversity dependence

Diversification rates can also depend on the growth of the clade itself. As diversification proceeds and species accumulate, they fil geographical space and niche space, potentially decreasing opportunities for speciation and increasing extinction risk.

Equilibrium diversity

In this model, diversity reaches a 'dynamic equilibrium' where immigration and extinction are balanced. Similiary, when *in situ* speciation (as opposed to immigration) plays a key role in the assemply of biotas, clades undergoing diversity-dependent diversification may eventually reach a 'diversity limit', called the 'carrying capacity' denoted by K.

Waxing and waning of diversity

From fossil records, we know that clades wax and wane (expand and decline). Such waxing-waning diversity dynamics can be represented by time-varying models by any model which has a positve net diversification rate at the beginning of the clade's history followed by a negative net diversification rate.

Protracted speciation

In species-based models, speciation is typically modelled as an instantaneous event. In reality, speciation requires reproductive isolation and may take millions of years to complete (Avise et al. 1998). The duration of speciation can have a significant impact on species richness patterns. Accounting for the duration of speciation, or 'protracted speciation', substantially modifies the expected shape of reconstructed phylogenies (Purvis et al. 2009; Etienne & Rosindell 2012).

Clade-specific diversification

The difference is species richness across groups of organisms is often too large to be explained solely by stochastically driven variation. Detecting rare shifts, that is specific subclades in which diversification has been 'abnormally' fast or slow, can help in understanding the potential causes of this heterogeneity.

Character-dependent diversification

If there are no *a priori* reasons to believe that specific characters influence diversification, one can test for a correlation between the diversification rate and an 'average' trait of individual clades (Slowinski & Guyer 1993). Such an average may be obtained by taking a mean of trait values across extant species, or an average over the full history of the clade based on ancestral trait values.
In these modelsm lineages are characterised by an evolving trait; they follow a birth-death process, in which speciation and extinction rates at any given time depend on the value of the trait at that time (Mooers et al. 2007).

Age dependence

A specific character that may influece species' rates of speciation and extinction is their age: species can be more or less likely to go extinct or speicate the older they get (Mooers et al. 2007). Species age is a non-inherited deterministic trait, and thus cannot be modelled as a trait evolving stochastically along phylogenetic branches, as is done in the character-dependent approaches.

The Neutral Theory of Biodiversity (NTB)

The NTB of Hubbell (2001) is an individual-based model, in which two principal hyptheses are

- a metacommunity of constant size (the zero-sum assumption)
- an ecological equivalence between individuals (the neutrality assumption)

When an individual dies in the metacommunity with probability ν, the individual is replaced by an individual from an entirely new species. That is, there is a speciation event (this form of speciation is typically referred to as point mutation mode of speciation, and ν is the per individual speciation rate). Alternatively, the individual is replaced by an offspring of the metacommunity with probabilty $1 - \nu$. Therefore, each individual has at each time step an equal probability of giving rise to an entirely new species, and the probability for a-species speciating is proportional to its abundance in the metacommunity.

Geographical speciation

Pigot et al. (2010) developed, a spatially explict model that considers the geographical context of speciation and extinction which is another type of neutral model. This model allows one to account for the fact that geographical isolation, which reduces gene flow between populations, often is an essential element of speciation. Range boundaries evolve under a Brownian process; extinction arises when range size drifts to zero, and speciation occurs via vicariance or peripatry.

Ecological differentiation

In this model, each patch in the metacommunity occupies a random position along an environmental gradient, and each species is characterised by its optimal position on the gradient. The dynamices of a given species in a given path follows a logistic equation in which carrying capacity decreases as the species gets further away grom its optimal position on the gradient. New species arise at a constant per-species rate with small abundances in each patch, and their characteristic is determined by a normal deviation from that of their progenitor. McPeek (2008) proposed a metacommunity model covering speciation dynamics ranging from ecological equivalence to ecological divergence.

A.1.6 Diversification slowdowns

Instead of the acceleration diversification of the birth-death model, studies of phylogenetic diversification often show evidence for slowdowns in diversification rates over the history of clades (Moen

and Morlon 2014). Recent studies seeking biological explanations to this pattern and have inverstigated the role of niche differentiation, adaptive radiation and ecological limits to diversity.

Metodological biases can lead to this results, however recent studies avoiding such biases have still found strong support for slowdowns (Moen and Morlon 2014). Many recent papers have emphasized the role of competition for limited resources, adaptive radiation and ecological limits on the number of species within a clade. The dominant explanation in literature for diversification slowdowns is that they result from the influence of competition for limited resources or niches on diversification (Moen and Morlon 2014). This explanation falls under the category of diversity-dependent explanations. Authors have hypothesized that speciation rate would slow down after the inital rapid speciation and niches are filled. However, examples of phylogenetic analysis in greater Antillian Anolis Lizards, show that niche differences did not lead to equilibrium diversities (Moen and Morlon 2014). We now state other possible causes of diversification slowdowns.

Geography of diversification

Most speciation, results from geographical isolation of populations with a lack of gene flow. A common cause of this event is a geographic barrier (i.e. vicariance), such as a river, mountain etcetera. These events are more likely to strike in large areas, or large range sizes. When such event happens, the species range size gets smaller and thus the extinction rates will increase. Recent studies lend support to the role of geography in explaining diversification slowdowns. Studies of island species have shown that the probability of speciation increases as the island size increases. Large, single-island assemblages show slowdowns, whereas clades in arcipelagoes of small isolated islands show constant net diversification rates (Moen and Morlon 2014). So, the geographic context of diversification alone can lead to diversification slowdowns.

Environment-driven pulses of high speciation rate

Speciation rates might increase during periods of rapid environmental or geological change, and decrease after such periods ends. Such changes can lead to speciation, for example by population isolations or the creation of an environmental gradient along which speciation occurs due to climatic specialization. The slowdown in this scenario arises from the slowing of extrinsic factors that lead to speciation.

Failure to keep pace with a changing environment

An inability to keep pace with a changing biotic or abiotic environment may lead to diversification slowndowns. This scenario falls into the category of time-dependent explanations, with diversification slowing down such that the net diversification rates from positive to negative. (i.e. from an increasing diversity to a decreasing diversity in the history of the clade). This could happen by extinction rates increasing above speciation rates, the speciation rates decreasing under the extinction rates or both cases. In this scenario, clades in both expanding and declining phase will show diversification slowdowns. This might explain the observed pattern.

Protracted speciation

Under this concept, there is a positive amount of time between the initial divergence of populations and when they have achieved reproductive isolation or when gene flow completely stops. The time at which sister species are inferred to have split in a phylogeny will date to the original population split (Moen and Morlon 2014). The presence of incipient species in a clade, will exclude the most recent branching points of a phylogeny. This might explain the observed diversification slowdowns because branching events near the tips (the near past) will be excluded. However, many analyses have found slowdowns after removing the most recent branch lengths. Protracted speciation cannot explain this slowdown (Moen and Morlon 2014).

A.2 Birth-Death Simulation Functions

```
# Open phylogenetics package
library(ape)


# Plotting instruments for first tree
traverse <- function(a){
  txt1 <- ""
  if (!is.na(a)) {
    theRow <- treeDF[treeDF$id == a ,]
    children <- treeDF[treeDF$ancestor == a ,]

    if (nrow(children) > 0) {

      for (i in 1:nrow(children)) {
        row <- children[i,]
        id <- as.numeric(row["id"])
        minTime <- as.numeric(row["branchtime"])
        if (i==1) {
          txt1 <- "("
        }

        #txt1 <- paste(txt1,traverse(id), id,":",minTime)
        txt1 <- paste(txt1,traverse(id), id, ":",minTime)

        if (i == nrow(children)) {
          txt1 <- paste(txt1,")")
        } else {
          txt1 <- paste(txt1, ",")
        }
      }
    }
  }
  return (paste(txt1))
}

buildtree <- function(a) {
  row <- treeDF[treeDF$id == a ,]
  minTime <- as.numeric(row["branchtime"])
  txt <- paste(a,":",minTime)
  return(paste(traverse(a),txt,";"))
}

# Plotting instruments for pruned tree

traverse_prune <- function(a){
  txt1 <- ""
  if (!is.na(a)) {
    theRow <- hist_tree[hist_tree$id == a ,]
    children <- hist_tree[hist_tree$ancestor == a ,]

    if (nrow(children) > 0) {
```

```
      for (i in 1:nrow(children)) {
        row <- children[i,]
        id <- as.numeric(row["id"])
        minTime <- as.numeric(row["branchtime"])
        if (i==1) {
          txt1 <- "("
        }

        txt1 <- paste(txt1,traverse_prune(id), id, ":",minTime)

        if (i == nrow(children)) {
          txt1 <- paste(txt1,")")
        } else {
          txt1 <- paste(txt1, ",")
        }
      }
    }
  }
  return (paste(txt1))
}

buildprunedtree <- function(a) {
  row <- hist_tree[hist_tree$id == a ,]
  minTime <- as.numeric(row["branchtime"])
  txt <- paste(a,":",minTime)
  return(paste(traverse_prune(a),txt,";"))
}
```

A.3 Code for Constant Birth-Death Simulations

```
MLE <- NULL
estimations <- NULL
count <- 0
lambda <- 0.4
mu <- 0.1
t_max <-  12

# Print development of tree?
print_dev <- FALSE

# Plot function procedure wanted?
plot_proc <- FALSE

# Plot histogram of MLE?
plot_mle <- TRUE

# No. of requested trees
NIT <- 10000
# No. of extant species we want for the pruned tree
samplesize <- 25


while(count<NIT){
  treeDF<-NULL
  while(length(treeDF[,1])<2){
    #Set initial values
    id <- 1
    t <- 0
    cuttime <- Inf
    nl <- 1
    LTT <- NULL

    # Set number of digits
    options(digits=5)


    # Draw first event
    e_t <- rexp(1,lambda+mu)
    b <- (runif(1)< lambda/(lambda + mu))

    # Fill list with branch numbers
    listid <- c(1)

    # Create empty tree
    treeDF <- NULL


    # Create set with branches
    pos <- data.frame(id = id, ancestor = 0, origin = t, eventtime = t+e_t, branchtime = e_t ,
                      birth = b)
```

```
# Update time
t <- t + e_t

# Draw random branch with removal
b_n <- sample(listid,1)
listid <- listid[!listid==b_n]


# Add drawn branch to tree
newbranch <- pos[pos$id == b_n,]
treeDF <-  rbind(treeDF,newbranch)
pos <- pos[!pos$id == b_n,]


if(newbranch[,6]==TRUE){

  while(t < t_max) {
    if(print_dev == TRUE){
    perc <- 100 - (t_max - t)/t_max * 100
    print(perc)
    }

    if(newbranch[,6]==TRUE){
      # Update list of possible branchnumbers
      listid <- c(listid,c(id+1,id+2))

      # Draw new event
      k <- length(listid)
      e_t <- rexp(1,k*(lambda+mu))
      b <- (runif(1) < lambda / (lambda + mu))

      # Update the event time and branch time
      pos[,4] <- pos[,4] + e_t
      pos[,5] <- pos[,5] + e_t

      # Create two new possible branches
      id <- id +1
      row1 <- data.frame(id= id,ancestor= b_n, origin= t,   eventtime = t+e_t,
                         branchtime = e_t, birth =  NA)
      id <- id + 1
      row2 <- data.frame(id= id,ancestor= b_n, origin= t, eventtime = t+e_t,
                         branchtime = e_t, birth =  NA)

      # Add branches to List of possible branches
      pos <- rbind (pos,row1,row2)

      # Choose random branch
      if(k>1){
        b_n <- sample(listid,1)
      }
      if(k==1){
        b_n <- listid
      }
```

```
  # Remove chosen branch from the possible list
  listid <- listid[!listid==b_n]
  # Update k
  k <- length(listid)

  # Retrieve new branch from the list
  newbranch <- pos[pos$id == b_n,]
  # Add birth/death event to branch
  newbranch[,6] <- b

  # Add branch to existing tree
  treeDF <- rbind(treeDF,newbranch)

  # Remove used branch from possible branches
  pos <- pos[!pos$id == b_n,]

  # Update time
  t <- t + e_t

  # Check the number of species in the tree, if it equals the requested sample
  # size, store the time.
  if(length(pos[,1]) == (samplesize - 1)){
    cuttime <- t
  }


}else{

  # If there is a branch left
  if(k>0){
    # Draw new event
    k <- length(listid)
    e_t <- rexp(1,k*(lambda+mu))
    b <- (runif(1) < lambda / (lambda + mu))

    # Update event time
    pos[,4] <- pos[,4] + e_t
    pos[,5] <- pos[,5] + e_t

    if(k>1){
      b_n <- sample(listid,1)
    }else{
      b_n <- listid
    }

    listid <- listid[!listid==b_n]

    # Update k
    k <- length(listid)

    # Retrieve new branch from the list
    newbranch <- pos[pos$id == b_n,]
    # Add birth/death event to branch
    newbranch[,6] <- b
```

```
          # Add branch to existing tree
          treeDF <- rbind(treeDF,newbranch)

          # Remove used branch from possible branches
          pos <- pos[!pos$id == b_n,]

          # Update time
          t <- t + e_t

          #If there are no branches left, end process
        }else{
          t <- t_max
        }

      }
    }
  }
  if(min(dim(pos))==0 ){
    #print("Tree has reached extinction.")
    #print("----------------------------")
  }else{
    #print("Tree has reached max. time.")
    #print("----------------------------")
    pos[,6] <- TRUE
    treeDF <- rbind(treeDF,pos)
    treeDF[treeDF$eventtime>t_max,][4] <- t_max
    treeDF[,5] <- treeDF[,4] - treeDF[,3]
  }

}

# Prune the tree if the number of species equaled the sample size
if(cuttime < max(treeDF[,4])){

  # Obtain the tree  existing untill time t
  hist_tree <- treeDF[treeDF$origin<cuttime,]


  # Collection of dead branches of history tree
  POB <- hist_tree[hist_tree$birth==FALSE,]
  POB <- POB[POB$eventtime < cuttime,]


  # Repeat process untill all dead branches are removed from the tree
  while(min(dim(POB)) >0){

    # Get the time when the first extinct species was born.
    times <- POB[,3]
    maxtime <- as.numeric(max(times))
    times <- times[!times==maxtime]
```

```
# Get the most recent dead branch. If both brother and sister species became
# extinct, take one of the two with the lowest id, which we call the brother.
x <- POB[POB$origin==maxtime,]
i <- as.numeric(min(x[1]))
x<- POB[POB$id==i,]



# Get the sister corresponding to the brother species (highest id).
y <- hist_tree[hist_tree$origin==maxtime,]
y <- y[!y$id==i,]
j <- as.numeric(y[1])



# Get the ancestor of the brother and sister.
a <- as.numeric(y[2])
z <- hist_tree[hist_tree$id==a,]


# If both brother and sister species became extinct, also the ancestor species
#became extinct in our phylogeny. We remove the childen species from our phylogeny,
# and set the ancestor species as an extinct species.
if(y[6]==FALSE & as.numeric(y[4]) < cuttime){
  # Erase birth event ancestor as a death event
  hist_tree[hist_tree$id==a,][6] <- FALSE
  # Remove the children from the possible branches
  POB <- rbind(hist_tree[hist_tree$id==as.numeric(a),],POB)
  POB <-  POB[!POB$id==as.numeric(i),]
  POB <-  POB[!POB$id==as.numeric(j),]
  # Remove the children out of the tree
  hist_tree <- hist_tree[!hist_tree$id == as.numeric(i),]
  hist_tree <- hist_tree[!hist_tree$id == as.numeric(j),]
}

# If a species will become extinct after cut time, set the species as alive.
if(y[6]==FALSE & as.numeric(y[4])>cuttime){
  hist_tree[hist_tree$origin==maxtime & hist_tree$id == as.numeric(j),][6] <- TRUE



  # If the sister species survived, it will take the origin of its ancestor since
  # its brother did not survive. The branch time of the sister equals now the branch
  # time of the ancestor, plus the original branch time of the sister. The ancestor
  # does not appear in our phylogeny, and is therefore removed from the data.
}
if(y[6] == TRUE){
  # take origin value of ancestor
  new_or <- as.numeric(hist_tree[hist_tree$id==a,3])
  hist_tree[hist_tree$origin==maxtime & hist_tree$id == as.numeric(j),][3] <- new_or

  # take branch time = branchtime + branchtime ancestor
  new_branchtime <- as.numeric(hist_tree[hist_tree$id == as.numeric(j),][5]) +
    as.numeric(hist_tree[hist_tree$id==a,][5])
```

```
    hist_tree[hist_tree$id == as.numeric(j),][5] <- new_branchtime


    # take ancestor value of the ancestor
    new_anc <- as.numeric(hist_tree[hist_tree$id==a,][2])
    hist_tree[hist_tree$id == as.numeric(j),][2] <- new_anc

    # Remove the ancestor, which is now part of the species
    hist_tree <- hist_tree[!hist_tree$id==a,]
    POB <- POB[!POB$id == as.numeric(i),]

    # remove the dead branch
    hist_tree <- hist_tree[!hist_tree$id==i,]
  }


}

# If the event time of some species exceed the time value, set these values to the
# time value. Update the branch times.
diff <- cuttime - hist_tree[hist_tree$eventtime > cuttime,][4]
hist_tree[hist_tree$eventtime > cuttime,][5] <- hist_tree[hist_tree$eventtime >
                                                    cuttime,][5] + diff
hist_tree[hist_tree$eventtime > cuttime,][4] <- cuttime

# Set all events to birth
hist_tree[,6] <- TRUE


# Plot the graph when the number of branches in the tree exceeds 1.

if(min(dim(hist_tree))>1){

  #Retrieve first species appearing in the phylogeny.
  k <- min(hist_tree[,1])

  # Plot the graph
  if(plot_proc == TRUE){

    #  Plot the LTT
    times <- hist_tree[,4]
    times <- times[!times == t_max]
    time <- 0
    nl <- 1
    LTT <- data.frame(t = time, no.lineages = nl )
    while(length(times)>0){
      time <- min(times)
      maxtime <- max(times)
      times <- times[!times == maxtime]
      times <- times[!times == time]
      if(hist_tree[hist_tree$eventtime == time,][6] == TRUE){
        nl <- nl + 1
      }else{
        nl <- nl - 1
```

```
      }
      newltt <- data.frame(t = time, no.lineages = nl )
      LTT <- rbind(LTT,newltt)
    }

    #  Plot the original LTT
    times <- treeDF[,4]
    times <- times[!times >= t_max]
    maxtime <- max(times)
    times <- times[!times >= maxtime]
    time <- 0
    nl <- 1
    aLTT <- data.frame(t = time, no.lineages = nl )

    while(length(times)>0){
      time <- min(times)
      times <- times[!times == time]
      if(treeDF[treeDF$eventtime == time,6] == TRUE){
        nl <- nl + 1
      }
      if(treeDF[treeDF$eventtime == time,6] == FALSE){
        nl <- nl - 1
      }
      newltt <- data.frame(t = time, no.lineages = nl )
      aLTT <- rbind(aLTT,newltt)

    }

  }

}


if(length(hist_tree[,1])>1){
  # Inferences part

  # Retrieve the time elapsed from the root note to the present.
  tto <- cuttime - hist_tree[,4]
  tto <- tto[!tto==0]

  # Create loglikelihood
  loglikelihood <- function(lambda,mu){
    N <- length(tto)
    n <- 2
    sum <- 0
    while(n <= N-1){
      add <- log(n) + log(lambda-mu) -n*(lambda-mu)*(tto[n]-tto[n+1]) +
        (n-1)*log(1-(mu/lambda)*exp(-(lambda-mu)*tto[n+1])) -
        n*log(1-(mu/lambda)*exp(-(lambda-mu)*tto[n]))
      sum <-sum + add
      n <- n+1
    }
    return(sum)
  }
```

```
Loglikelihood <- function(MU){
  lambda <- MU[,1]
  mu <- MU[,2]
  N <- length(tto)
  n <- 2
  sum <- 0
  while(n < N){
    add <- log(n) + log(lambda-mu) -n*(lambda-mu)*(tto[n]-tto[n+1]) +
      (n-1)*log(1-(mu/lambda)*exp(-(lambda-mu)*tto[n+1])) -
      n*log(1-(mu/lambda)*exp(-(lambda-mu)*tto[n]))
    sum <-sum + add
    n <- n+1
  }
  return(sum)
}



# Prepare contour plot, where the real values are lambda = 0.3 and mu = 0.1.
lambdag <- seq(0,1,0.01)
mug <- seq(0,1,0.01)
X <- as.matrix(expand.grid(lambdag, mug))
colnames(X) <- c("lambda", "mu")
z <- outer(lambdag,mug,loglikelihood)

# Determine maximum likelihood
max <- max(z, na.rm = TRUE)

mval <- which(z==max)


MLE <- X[mval,]
true <- data.frame(lambda = lambda, mu = mu)

# Plot of the process
if(plot_proc == TRUE){
  par(mfrow=c(2,2))


  # Plot the tree
  graph_tree <- plot(read.tree(text=buildtree(1)), type = "phylogram",
                     use.edge.length = TRUE, node.pos = NULL, show.tip.label = TRUE,
                     show.node.label = TRUE, edge.color = "black", edge.width = 1,
                     edge.lty = 1, font = 3, cex = par("cex"), adj = NULL, srt = 0,
                     no.margin = FALSE, root.edge = TRUE, label.offset = 0,
                     underscore = FALSE, x.lim = NULL, y.lim = NULL,
                     direction = "rightwards", lab4ut = "horizontal",
                     tip.color = "black")

  axis(1,labels=TRUE)
  title("Constant birth-death simulation", xlab="time in myrs")
  abline(v = cuttime, col = "blue", lwd = 2)
```

```
        # Plot pruned tree
        graph_pruned_tree <- plot(read.tree(text=buildprunedtree(k)), type = "phylogram",
                                  use.edge.length = TRUE, node.pos = NULL,
                                  show.tip.label = TRUE, show.node.label = TRUE,
                                  edge.color = "black", edge.width = 1, edge.lty = 1,
                                  font = 3, cex = par("cex"), adj = NULL, srt = 0,
                                  no.margin = FALSE, root.edge = TRUE, label.offset = 0,
                                  underscore = FALSE, x.lim = NULL, y.lim = NULL,
                                  direction = "rightwards", lab4ut = "horizontal",
                                  tip.color = "black")
        axis(1,labels=TRUE)
        title("Reconstructed phylogeny", xlab="time in myrs")

        # Plot Reconstructed phylogeny LTT
        plot(LTT[,1],log(LTT[,2]),main = "LTT plot of constant birth-death simulation",
             xlab = "time", ylab = " log(no. lineages)", col = "red")

        # Plot Original LTT
        points(aLTT[,1],log(aLTT[,2]), col = "green")
        # Expected LTT plot
        abline(a=0,b=lambda-mu, col="black")

        # Plot the maximum likelihood estimate together with the contour plot
        contour(lambdag,mug,z,nlevels=samplesize,
      main=paste("Contourplot of the MLE, N=",samplesize),
    xlab = expression(lambda), ylab= expression(mu))
        points(MLE[1],MLE[2],col="red")
        points(lambda,mu,col="blue")
       }

     }else{
       MLE <- NULL
     }
   }
   if(cuttime > max(treeDF[,4])){
     MLE <- NULL
   }
   estimations <- rbind(estimations,MLE)
   count <-length(estimations[,1])
   print(paste("Treenumber =",count))
   }

 if(plot_mle == TRUE){
   par(mfrow=c(2,2))
   hist(estimations[,1],main=paste("Histogram of", NIT, "estimations"),
        xlab= expression(lambda), sub = paste("True lambda =", lambda,",",
                                              "with no. species = ", samplesize),
        prob=T, breaks = 25)
   lines(density(estimations[,1],na.rm=T))
   abline(v = lambda, col = "blue", lwd = 2)
   boxplot(estimations[,1], main=paste("Histogram of", NIT, "estimations"),
           ylab = expression(lambda), sub = paste("True lambda =", lambda,",",
                                                  "with no. species = ", samplesize))
   hist(estimations[,2],main=paste("Histogram of", NIT, "estimations"),
```

```
      xlab=expression(mu), prob=T, sub = paste("True mu =", mu,",", "with no. species = ",
                          samplesize), breaks =25)
  lines(density(estimations[,2],na.rm=T))
  abline(v = mu, col = "blue", lwd = 2)
  boxplot(estimations[,2], main = paste("Histogram of", NIT, "estimations"),
          ylab = expression(mu), sub = paste("True mu =", mu,",", "with no. species = ",
                                samplesize))
}
```

A.4 Code for Protracted Birth-Death Simulations

```
# Protracted Tree


library(mnormt)
mle <- NULL

NIT <- 1
samplesize <- 10


teller <- 0

# SIZE ON?
SIZE <- FALSE

# PLOT PROCEDURE?
procedure <- TRUE

# PlOT MLE?
plotmle <- FALSE

while(teller < NIT){
dt <- NULL

# Set intial no. of species
Ng <- 1
Ni <- 0

# Set initial values
t <- 0
T <- 2
i <- 1
ii <- 1
list.incip <- NULL
list.good <- i
signal <- 1

# Create tree dataframe
tree <- data.frame(id =ii,
                   anc = 0,
                   origin = t,
                   completion = t,
                   extinction = Inf)


# Set rates
l1 <- 1
l2 <- 0.8
l3 <- 0.5
m <- 0.3
```

```
# Set first event
Nt <- NULL
Nt <- data.frame(good.spec = Ng, inc.spec = Ni, time = t, id =ii)

while(t<T & (Ng + Ni)>0 & signal>0){
  R <- (l1 + m)*Ng + (l2 + l3 + m)*Ni

  f <- runif(1)

  t <- t - log(f)/R

  if(t > T){
    t <- T
  }

  # Probabilities
  k <- matrix(NA,nrow=5,ncol = 1)
  # Pr(good -> inc)
  k[1,] <- l1*Ng / R
  # Pr(good -> death)
  k[2,] <- m*Ng / R
  # Pr(inc -> good)
  k[3,] <- l2*Ni /R
  # Pr(inc -> inc)
  k[4,] <- l3*Ni / R
  # Pr(inc -> death)
  k[5,] <- m*Ni / R

  # Obtain the event
  g <- runif(1)
  prj <- j
  j<-1
  som <- k[1,]
  while(som < g){
    j <- j + 1
    som <- som + k[j,]
  }
  # The event is thus j.

  # Good -> Inc
  if(j == 1){
    Ni <- Ni + 1
    i <- i + 1
    ii <- -i

    if(length(list.good)>1){
    ance <- sample(x = list.good[!list.good == ii], size = 1)
    }else{
      ance <- list.good
    }
    or <- t
    branch <- data.frame(id =ii,
                         anc = ance,
                         origin = or,
```

```
                      completion = NA,
                      extinction = Inf)
  tree <- rbind(tree,branch)

  list.incip <-  append(list.incip, ii)
}
# Good -> death
if(j == 2){
  Ng <- Ng - 1
  if(length(list.good)>1){
  ii <- sample(x = list.good, size =1)
  }else{
    ii <- as.numeric(list.good)
  }

  tree[tree$id == ii,5] <- t

  list.good <- list.good[!list.good == ii]
}
# Inc -> Good
if(j == 3){
  Ng <- Ng + 1
  Ni <- Ni - 1
  if(length(list.incip)>1){
  ii <- -( sample(x = list.incip, size = 1))
  }else{
    ii <- - as.numeric(list.incip)
  }
  co <- t
  tree[tree$id == (-ii),1] <- ii
  tree[tree$id == ii,4] <- t
  tree[tree$anc == (-ii),2] <- ii



  list.incip <- list.incip[!list.incip == (-ii)]
  list.good <- append(list.good, ii)

}
# Inc -> Inc
if(j == 4){
  Ni <- Ni + 1
  i <- i + 1
  ii <- -i

  ance <- sample(x = list.incip[!list.incip == ii], size = 1)
  or <- t
  branch <- data.frame(id =ii,
                      anc = ance,
                      origin = or,
                      completion = NA,
                      extinction = Inf)
  tree <- rbind(tree,branch)
  list.incip <- append(list.incip, ii)
```

```
  }
  # Inc -> death
  if(j == 5){
    Ni <- Ni - 1
    if(length(list.incip)>1){
    ii <- sample(list.incip, size = 1)
    }else{
      ii <- as.numeric(list.incip)
    }
    tree[tree$id == ii,5] <- t

    list.incip <- list.incip[!list.incip == ii]
  }

  newNt <- data.frame(good.spec = Ng, inc.spec = Ni, time = t, id = ii)
  Nt <- rbind(Nt,newNt)

  if(Ng + Ni == samplesize & SIZE == TRUE){
    signal <- 0
}
}
# determine the dt's
if(signal == 0 | SIZE == FALSE){

times <- Nt[,3]
dt <- NULL
for(i in 1:(length(times) - 1)){
  dt[i] <- Nt[(i+1),3] - Nt[i,3]
}

good <- Nt[,1]
inc <- Nt[,2]
dgood <- NULL
dinc <- NULL
for(i in 1:(length(good)-1)){
  dgood[i] <- good[i+1] - good[i]
  dinc[i] <- inc[i+1] - inc[i]
}
X <- cbind(dgood,dinc)

# Inference part

lambda <- c(l1,l2,l3,m)

expected <- function(j,LAM){
  lambda1 <- LAM[1]
  lambda2 <- LAM[2]
  lambda3 <- LAM[3]
  mu <- LAM[4]
  E <- matrix(NA,nrow=2,ncol =1)
  E[1,1] <- -mu*Nt[j,1] + lambda2*Nt[j,2]
  E[2,1] <- -mu*Nt[j,2] - lambda2 * Nt[j,2] + lambda1 * Nt[j,1] + lambda3*Nt[j,2]
  E <- E * dt[j]
  return(E)
```

```
}

covariance <- function(j,LAM){
  lambda1 <- LAM[1]
  lambda2 <- LAM[2]
  lambda3 <- LAM[3]
  mu <- LAM[4]
  V <-matrix(NA,nrow=2,ncol=2)
  V[1,1] <- mu*Nt[j,1] + lambda2*Nt[j,2]
  V[1,2] <- -lambda2*Nt[j,2]
  V[2,1] <- -lambda2*Nt[j,2]
  V[2,2] <- lambda1*Nt[j,1] + (lambda2 + lambda3+ mu)*Nt[j,2]
  V <- V*dt[j]
  return(V)
}

loglik <- function(j,LAM){
  u <- X[j,]
  ll <-dmnorm(x = u,
              mean = t(expected(j,LAM)),
              varcov = covariance(j,LAM),
              log = T)
  id <- j
  return(ll)
}

name <- function(i){
  nam <- paste("loglik",i, sep = ".")
}



N <- length(dt)
listl <- NULL
for(i in 1:N){
nam <- name(i)

assign(nam,loglik)

listl <- append(listl,nam)

}


total <- function(LAM){
  totlik <- NULL
  for(i in 1:N){
    totlik[i] <- get(listl[i])(i,LAM)
  }
  return(sum(totlik))
}

mle_ <- optim(par = lambda, fn = total, lower = c(0.0001,0.0001,0.0001,0.0001), method = "L-BFGS-B", co
MLE <- c(as.numeric(unlist(mle_)[1]),as.numeric(unlist(mle_)[2]),as.numeric(unlist(mle_)[3]),as.numeric
```

```
mle <- rbind(mle,MLE)
teller <- teller + 1
print(teller)
}
}

if(plotmle == TRUE){
par(mfrow=c(2,2))
hist(mle[,1], main=paste("Histogram with tree size", 25),
     xlab= "lambda1",
     prob=T, breaks = 0.1*teller, xlim = c(0,5))
lines(density(mle[,1],na.rm=T, n= NIT))
abline(v = l1, col = "blue", lwd = 2)
hist(mle[,2], main=paste("Histogram with tree size", 25),
     xlab= "lambda2", xlim = c(0,5),
     prob=T, breaks = teller)
lines(density(mle[,2],na.rm=T,n= NIT))
abline(v = l2, col = "blue", lwd = 2)
hist(mle[,3], main=paste("Histogram with tree size", 25),
     xlab= "lambda3",xlim = c(0,5),
     prob=T, breaks = teller)
lines(density(mle[,3],na.rm=T,n= NIT * 5))
abline(v = l3, col = "blue", lwd = 2)
hist(mle[,4], main=paste("Histogram with tree size",25),
     xlab= "mu",xlim = c(0,5),
     prob=T, breaks = teller)
lines(density(mle[,4],na.rm=T, n = NIT))
abline(v = m, col = "blue", lwd = 2)
}

if(procedure == TRUE){
  # Diffusion plots.

  QT <- NULL
  Y1 <- NULL
  Y2 <- NULL

  j <- 1
  while(j < length(Nt[,1])){


    k <- 0
    difflist <- NULL
    while(k <= 100){
      k <- k+ 1
      x <- rmnorm(n = 1,mean = expected(j,lambda),varcov =covariance(j,lambda))
      difflist <- rbind(difflist,x )
    }
    q11 <- Nt[j+1,1] + quantile(difflist[,1], 0.025)
    q12 <- Nt[j+1,1] + quantile(difflist[,1], 0.975)
    q21 <- Nt[j+1,2] + quantile(difflist[,2], 0.025)
    q22 <- Nt[j+1,2] + quantile(difflist[,2], 0.975)

    y1 <- c(Nt[j+1,3],q11,q12)
```

```
    y2 <- c(Nt[j+1,3],q21,q22)

    Y1 <- rbind(Y1,y1)
    Y2 <- rbind(Y2,y2)
    j <- j + 1
  }


  par(mfrow=c(2,1))
  plot(NA,main = "95% confidence of diffusion of good species",
       xlim = c(0,max(Nt[,3])), xlab = "time", ylim = c(min(Y1[,2]),max(Y1[,3])),
       ylab = "No. of good species")
  points(Nt[,3], Nt[,1])
  lines(Nt[,3], Nt[,1])
  points(Y1[,1], Y1[,2], col = "blue")
  points(Y1[,1], Y1[,3], col = "red")

  plot(NA,main = "95% confidence of diffusion of incipient species",
       xlim = c(0,max(Nt[,3])), xlab = "time", ylim = c(min(Y2[,2]),max(Y2[,3])),
       ylab = "No. of inc. species")
  points(Nt[,3], Nt[,1])
  lines(Nt[,3], Nt[,1])
  points(Y2[,1], Y2[,2], col = "blue")
  points(Y2[,1], Y2[,3], col = "red")
}
```