Politecnico di Torino

Department of Applied Science and Technology

Master Degree in

Physics of Complex Systems

Master degree thesis

Bayesian variable selection for environment-dependent phylogenetic models of diversification

Internal supervisor

Andrea Antonio Gamba

Candidate

Mattia Tarabolo

External supervisors

Hélène Morlon

Julien Clavel

Abstract

Understanding how past environmental changes have influenced the diversification of species is key for predicting the impact of current and future environmental changes on biodiversity, and the associated human, social and economic impact. Various environment-dependent phylogenetic comparative methods, that allow testing whether and how past. These methods build upon classical birth-death models of cladogenesis used to study speciation and extinction dynamics from phylogenies of extant species, where the evolutionary rates correspond to speciation and extinction rates. Even though several recent studies have fitted these models to comparative phylogenetic data, providing estimates of the association between evolutionary rates and environmental variables, the phylogenetic methods already developed have several limitations. The most limiting factor is that they were implemented in a maximum likelihood rather than a bayesian framework, which precludes the development of more complex models. In particular, the maximum likelihood approach allows to test only the effect of one environmental variable at a time, due to the problem of overparametrization. environmental changes influenced evolutionary rates have recently been developed. We propose to overcome this problem using a Bayesian implementation, and we will show how this approach actually outperform the Maximum likelihood implementation even when using a single environmental dependency. Implementing these models in a Bayesian framework allow to use Bayesian Variable selection techniques, which overcome the problem of overparametrization through the use of informative priors. We will present in details the method, and we will propose a simple implementation through Monte Carlo Markov Chain sampling.

Table of Contents

Chapter 1

Introduction

1.1 What is macroevolution?

The theory of evolution, which is the study of changes in the heritable characteristics of biological populations over successive generations, has interested scientists and philosophers well before Darwin, even though the idea of evolution through natural selection was firstly proposed by the latter together with Wallace [1]. After their groundbreaking work,the so-called *modern synthesis* reconciled natural selection and Mendelian inheritance theories into a unified framework [2, 3].

Later on the evolutionary theory further developed, and scientists started to distinguish between *macroevolution* and *microevolution*. The two terms were first introduced by Philiptschenko [4], who defined macroevolution as evolution above the species level in the Linnaean hierarchy [5, 6, 7] and microevolution as evolution beneath the species level. The modern definition of macroevolution, firstly proposed by Stanley [8], is instead evolution that is guided by sorting of interspecific variation, as opposed to sorting of intraspecific variation in microevolution [3, 9]. As a consequence of these definitions macroevolution can be safely thought as evolution on a grand scale.

1.2 Phylogenetic approaches for studying diversification

Among all the macroevolutionary processes, *diversification* is key to understanding how biodiversity changed over time, and what are the drivers of these changes. The term describes the balance between *speciation* and *extinction*, where the former is a fundamental evolutionary process by which populations evolve to become distinct species.

Even though diversification is fundamental in studying biodiversity, it turned out to be difficult to explain. Indeed, the speciation and extinction processes occur on extremely large timescales, and inferring their variations can be hard. This is mostly because our knowledge of past biodiversity is limited. Several models were developed

for that purpose, in order to estimate diversification rates from available data. While this can be done using fossil data for some groups such as planktonic foraminifers, planktonic diatoms, bivalves, gastropods and mammals, for which fossil availability is sufficient to apply statistical methods [10, 11, 12, 13], it is not feasible for most of the other groups. The lack of fossil records for most of the groups pushed scientists to develop alternative approaches to analyze diversification. The pioneering papers of Hey, Nee, May and Harvey [14, 15, 16, 17] paved the way for the development of new methods based on the study of *reconstructed phylogenies*, which are branching trees describing the evolutionary relationships among extant species. Reconstructed phylogenies can be inferred from molecular DNA sequences of extant species, thus being collected more easily than fossil occurrences. Phylogenetic methods have therefore become predominant to study biodiversity [18, 19, 20, 21, 22, 23].

1.3 Environmental changes and diversification

The links between environmental changes and diversification dynamics was apparent since the 19th century [24, 25]. Over the 20th century a variety of hypothesis regarding the drivers of diversification were proposed. Among them, the two most important are the *Red Queen hypothesis* and the *Court Jester hypothesis*. The former, firstly proposed by Van Valen [26], assumes that the major drivers of biodiversity changes interspecific interaction (e.g. antagonistic such as competition and parasitism or mutulalistic such as plant/pollinisators etc.), called *biotic interactions*, while the latter, firstly proposed by Barnosky [27] in contrast with the Red Queen, assumes that the major driving forces of diversification are environmental changes, also called *abiotic forces*.

In order to test these hypotheses various observations were made, finding support for both the Red Queen [28] and the Court Jester [29, 30, 31, 32]. However testing the relative importance of the two hypotheses is not trivial, and requires to develop models which include an explicit dependency of the diversification processes from the biotic and abiotic forces.

This project will focus on the environment-dependent birth-death model, developed firstly by Condamine, Rolland and Morlon [33]. This model builds on the time-dependent birth-death model developed by Morlon, Parsons and Plotkin [34], assuming the diversification dynamics to be governed by a time-continuous birth-death process, in which the speciation rate corresponds to the birth rate of the process and the extinction rate to the death rate. The model can accommodate any dependency of the diversification rates from the environmental curves, as well as any curve describing biotic interactions. In particular, the model computes the likelihood of observing a reconstructed phylogeny assuming a particular functional form of the diversification rates. The computation of the likelihood and of the *Maximum Likelihood* (ML) estimates of the rates was implemented in the *R* package *RPANDA* [35], allowing to test whether environmental changes had a significant influence on the diversification dynamics, as well as to quantify the direction and magnitude of this potential influence.

1.4 Project objectives and outline

The environment-dependent birth-death model can in principle accommodate any dependency of the diversification rates from the environmental and biotic variables, including even a multivariate dependency. Nevertheless, the Maximum Likelihood implementation does not allow to test more than one dependency at the time, since it is subject to the risk of overparametrization when including many dependencies with a small signal.

Until now the influence of different environmental variables on the speciation and extinction rates was tested only with likelihood-based approaches using one dependency at a time [36]. This method limits the possible dependencies which can be tested, while it is likely that the diversification dynamics was altered by a combination of environmental factors. This project will focus on developing new methods which can test multiple environmental dependencies while avoiding the problem of overparametrization. This can be done by implementing the environment-dependent birth-death model within a Bayesian framework, in order to employ so-called *Bayesian variable selection* techniques [37]. These methods assign specific priors to the parameters of the model which allows to correctly identify signal overcoming the risk of overparametrization. In particular we used the horseshoe prior, which belongs to the family of adaptive shrinkage methods. The technique consists in placing hierarchical priors on the parameters to be estimated, which have tall spikes at zero, yielding the shrinkage of noise parameters (i.e. negligible dependencies), and heavy tails, allowing signals (i.e. significant dependencies) of potentially strong positive or negative intensity.

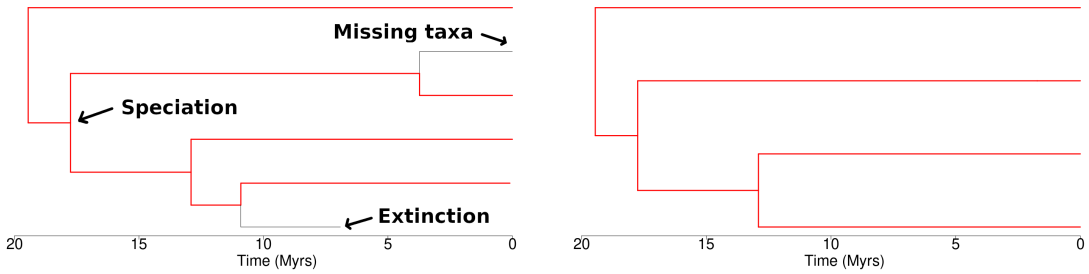
In chapter 2 we present in details the methods used in the project. In particular, in chapter 2.1 we discuss the environment-dependent birth death model, and in chapter 2.2 we show how to implement this model in a bayesian framework, discussing how the Markov Chain Monte Carlo sampling technique can be implemented in order to sample the posterior distribution of the model. In chapters 2.3, 2.4 and 2.5 we present the Bayesian Variable Selection technique, which employs the bayesian framework to fit the model with multiple environmental dependencies. In particular we consider the use of the so-called horseshoe prior, and we propose a reparametrization to enhance the efficiency of the Monte Carlo sampling. In chapter 3 we presents the numerical results obtained from the implementation of the aforementioned methods. In particular, in chapter 3.1 we show how the likelihood computation was drastically reduced optimizing the code available in the R package RPANDA, while in chapter 3.2 we illustrate the Bayesian implementation of the model. The new method can recover on average the correct parameters from simulated phylogenies, and is more precise than the Maximum Likelihood implementation.

Chapter 2

Methods

2.1 The environment-dependent birth-death model

The environment-dependent birth-death [33] model was built upon the time-dependent birth death model [34] simply assuming any non-negative dependency of the speciation and extinction rates from the environmental curves. The model does not care about the mechanisms producing speciation and extinction events, but only about the diversification dynamics.



(a) Complete phylogeny (b) Reconstructed phylogeny

Figure 2.1: Examples of (a) a complete phylogeny and (b) the corresponding reconstructed phylogeny of a clade. The extinction events, as well as the missing extant taxa which are not sampled, are not present in the reconstructed phylogeny, which is inferred from molecular data of extant species.

One can assume that a clade, a monophyletic group of species, of which the reconstructed phylogeny is analysed evolved according to a continuous birth-death process, in which the birth process corresponds to speciation and the death process to extinction of a species. The process generates a complete phylogeny with speciation and extinction events, as shown in figure 2.1a. Nonetheless the reconstructed phylogeny, as shown in figure 2.1b, does not include extinction events since it is inferred from molecular data of extant species, without including fossil occurrences. The model should therefore take this into account. In addition, when examining r;constructed phylogenies it is not always possible to sample all of the

extant species. The model considers that only a fraction $f \leq 1$ of them is sampled.

Given k environmental curves $E_1(t), ..., E_k(T)$, the speciation rate $\lambda(t)$ and the extinction rate $\mu(t)$ can take any non-negative functional form which links them to the environmental variables

$$\lambda(t) = \tilde{\lambda}\left(t, E_1(t), ..., E_k(t)\right) \tag{2.1}$$

$$\mu(t) = \tilde{\mu}\left(t, E_1(t), ..., E_k(t)\right) \tag{2.2}$$

where $\tilde{\lambda}(x) \geq 0 \ \forall x$ and $\tilde{\mu}(x) \geq 0 \ \forall x$. In order to easily define a reference frame, time is measured from the present to the past, where $t = 0$ corresponds to the present and t increases into the past. If the reconstructed phylogeny has n sampled extant species, it will be characterized by n branching times $t_1, ..., t_n$ at which speciation occurred, such that $t_1 > t_2 > ... > t_n > 0$. t_1 is called *stem age* and corresponds to the time of origination of the ancestral species, while t_2 is called *crown age* and corresponds to the time of origination of the most recent common ancestor of the clade (figure 2.2 shows a schematic illustration of the notations used).

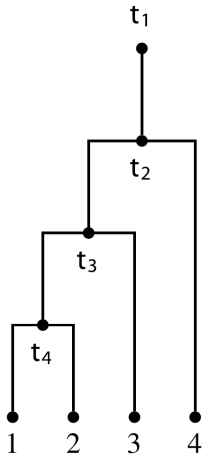


Figure 2.2: Schematic view of the notation used by the environment-dependent birth-death model. The phylogeny is characterized by the stem age t_1 and the branching times t_2, t_3, t_4. t_1 is the origination time of the ancestral species, while t_2 is the time of origination of the most recent common ancestor of extant species, therefore corresponding to the time at which the ancestral species split into two new species. Adapted from [34].

Under these assumptions, the probability of observing such a reconstructed phylogeny conditioned on the stem age (presence of at least one descendant in the sample) is given by

$$\mathcal{L}(t_1, ..., t_n | \lambda(t), \mu(t)) = \frac{f^n \Psi(t_2) \prod_{i=2}^{n} \lambda(t_i) \Psi(t_i)}{1 - \Phi(t_1)} \tag{2.3}$$

where $\Psi(t)$ is the probability that a species alive at time $t > 0$ leaves exactly one descendant species at the present in the reconstructed phylogeny

$$\Psi(t) = e^{\int_0^t ds(\lambda(s) - \mu(s)))} \left[1 + f \int_0^t ds \lambda(s) e^{\int_0^s du(\lambda(u) - \mu(u))}\right]^{-2} \tag{2.4}$$

and $\Phi(t)$ is the probability that a species alive at time $t > 0$ has no descendant in the sample

$$\Phi(t) = 1 - \frac{e^{\int_0^t ds(\lambda(s)-\mu(s))}}{\frac{1}{f} + \int_0^t ds\lambda(s)e^{\int_0^s du(\lambda(u)-\mu(u))}} \tag{2.5}$$

If the stem age t_1 of the reconstructed phylogeny is not known, one can condition on the crown age (presence of at least two descendant in the sample originated by a speciation event at time t_2)

$$\mathcal{L}(t_1, ..., t_n|\lambda(t), \mu(t)) = \frac{f^n \Psi(t_2) \prod_{i=2}^n \lambda(t_i)\Psi(t_i)}{\lambda(t_2)\left(1 - \Phi(t_1)\right)^2} \tag{2.6}$$

As already stressed out the model is very flexible and can include any dependency of the diversification rates from the environmental variables.

The most straightforward way to study the influence of environmental changes to the diversification dynamics of a clade of which the reconstructed phylogeny is known would be to use a Maximum Likelihood based approach, as done by Lewitus and Morlon in [36]. They tested one environmental dependency at a time using both a linear functional form

$$\lambda(t) = \max\left(0, \lambda_0 + \theta E(t)\right) \tag{2.7}$$
$$\mu(t) = \max\left(0, \mu_0 + \nu E(t)\right) \tag{2.8}$$

and an exponential form

$$\lambda(t) = \lambda_0 e^{\theta E(t)} \tag{2.9}$$
$$\mu(t) = \mu_0 e^{\nu E(t)} \tag{2.10}$$

The parameters λ_0, θ, μ_0 and ν are then inferred as the Maximum Likelihood point estimates. Even though they were able to determine which environmental variable most influenced the diversification of Cetaceans and Ruminants the method is limited by the use of a single environmental dependency at a time. This requires to have a certain degree of prior knowledge about the potential factors which influenced speciation and extinction events. If one has little knowledge about that trying to fit a multivariate dependency including as many environmental variables as possible can lead to overparametrization. In addition, it is likely that a combination of environmental factors (and biotic factors) rather than a single factor played a role in shaping biodiversity.

2.2 Bayesian implementation of the model

The problem of determining the influence of environmental variables on the diversification of a clade can be assessed using a Bayesian implementation. Bayes' theorem states that given a model with unknown parameters θ the *posterior probability* $p(\theta|X)$ of the parameters θ after having observed an outcome X is

$$p(\theta|X) \propto \mathcal{L}(X|\theta)p(\theta) \tag{2.11}$$

where $\mathcal{L}(X|\theta)$ is the likelihood probability of observing an outcome X assuming it to be generated by the parameters θ, and $p(\theta)$ is the *prior probability* encoding our a priori believes about the parameters. Bayesian theory suggests to use the posterior distribution rather than the Maximum Likelihood point estimate to do inferences. Once the posterior distribution of the parameters space is known, one can infer any feature such as means, variances and marginal distributions. Using a distribution over the whole parameter space has the advantage of offering a simple interpretation of the error associated to a point estimate of the parameters.

Nevertheless, using a Bayesian implementation does not come without any challenge. Computing the posterior distribution is way more complex than simply computing the likelihood due to the presence of the prior. Even when the posterior is computable, if the parameters space is too large it would be impossible to compute averages, variances and marginal distributions. In most of the cases one have to resort to numerical approximations to compute them. *Markov Chain Monte Carlo* (MCMC) simulations are by far the most used tool. MCMC is a technique designed to sample draws from a general distribution. In particular, samples are drawn sequentially from a random walk (also called *Markov Chain*, hence the name of the sampling method) which has the desired distribution as its stationary distribution. A Markov chain is a sequence of random variables $\{\theta_1, ..., \theta_n\}$ in which the probability distribution of any random variable θ_i, denoted as $p_i(\theta_i|\theta_1, ..., \theta_{i-1})$, depends only on the value of the previous random variable θ_{i-1}

$$p_i(\theta_i|\theta_1, ..., \theta_{i-1}) = T_i(\theta_i|\theta_{i-1}) \tag{2.12}$$

where $T_i(\theta_i|\theta_{i-1})$ is called *transition distribution*. If the transition probability distribution is suitably constructed the Markov Chain will converge to a unique stationary distribution which is the desired one [38, 39].

MCMC can therefore be used to sample any posterior distribution $p(\theta|X)$. Among the various methods used to construct a Markov chain which converges to $p(\theta|X)$ the *Metropolis-Hastings* (MH) algorithm [40, 41] is the most general to implement. It employs a proposal step followed by an acceptance-rejection procedure which ensures the convergence of the chain to the desired distribution. The two steps for drawing the new parameter θ_i given the already sampled parameters $\{\theta_0, ..., \theta_{i-1}\}$, where θ_0 is an appropriately chosen starting point (such that $p(\theta_0|X) > 0$), are the following:

1. Sample a *proposal* θ^* from a *proposal distribution* $J_i(\theta^*|\theta_{i-1})$
2. Compute the *acceptance probability*

$$\alpha(\theta^*, \theta_{i-1}) = \min\left(1, \frac{p(\theta^*|X)/J_i(\theta^*|\theta_{i-1})}{p(\theta_{i-1}|X)/J_i(\theta_{i-1}|\theta^*)}\right) \tag{2.13}$$

and set

$$\theta_i = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*, \theta_{i-1}) \\ \theta_{i-1} & \text{with probability } 1 - \alpha(\theta^*, \theta_{i-1}) \end{cases} \tag{2.14}$$

Among the different proposal functions, the most simple is the *sliding window proposal*. It consists in proposing a new parameter θ^* drawn from a symmetric distribution $J_i(\theta^*|\theta_{i-1})$ centered in θ_{i-1}, such that

$$\theta^* = \theta_{i-1} + \delta\theta^* \qquad \text{where} \qquad p(\delta\theta^*) = p(-\delta\theta^*) = \tilde{J}(\delta\theta^*) \tag{2.15}$$

Various symmetric distributions can be used, such as a uniform distribution, a Gaussian distribution, a Laplace distribution and many others. Each of them will be characterized by a *scale hyperparameter* γ defining how broad they are (e.g. $\mathcal{U}(-\gamma, \gamma)$ for a uniform distribution, $\mathcal{N}(0, \gamma^2)$ for a normal distribution, Laplace$(0, \gamma)$ for a Laplace distribution[1]). The larger γ is, the more distant from θ_{i-1} the proposed value θ^* will be on average.

Once the Markov Chain $\{\theta_0, ..., \theta_n\}$ has been simulated one can infer the expected value of any quantity $f(\theta|X)$ approximating it as

$$\mathbb{E}[f(\theta|X)] = \int d\theta\, f(\theta|X)p(\theta|x) \approx \frac{1}{n}\sum_{i=0}^{n} f(\theta_i) \tag{2.16}$$

In this way one can compute any quantity of interest. The posterior probability distribution and the marginal posterior probability distributions can also be approximated taking histograms of the sample.

Using the Markov Chain to perform inference needs however some precautions. First of all, when initializing the chain in a starting point θ_0 one has to wait some iterations for the chain to reach stationarity. The first first part of the chain, called *burn-in* needs thus to be discarded. In order to check whether the chain has converged one needs to compute k independent Markov Chains and remove the burn-in from each of them. *Convergence* is reached when each chain has reached stationarity converging to a common distribution(see figure 2.3 for an example). A quantitative convergence analysis can be done computing the *Gelman-Rubin convergence diagnostic* [42, 43, 38] for any scalar quantity of interest ψ, which is computed as follows. Each chain is split into two halves, therefore obtaining $m = 2k$ chains. Let n be the length of each chain after splitting. The values of the scalar are labeled $\psi_{i,j}$ where $i = 1, ..., n$ and $j = 1, ..., m$. The single-chain averages $\overline{\psi}_{\cdot j}$ and variances s_j^2 and the all-chains

[1]Given a random variable X, we will use the following notation. $X \sim \mathcal{U}(a, b)$ for a uniform random variable in the interval $[a, b]$, $X \sim \mathcal{N}(\mu, \sigma^2)$ for a normal random variable with mean μ and variance σ^2 and $X \sim$ Laplace(μ, r) for a Laplace random variable with location parameter μ and scale parameter r. In the case of the sliding window proposal function we can therefore use the notation $\delta\theta^* \sim \mathcal{S}(\gamma)$ where $\mathcal{S}(\gamma)$ is a generic symmetric distribution centered in zero and with scale parameter γ.

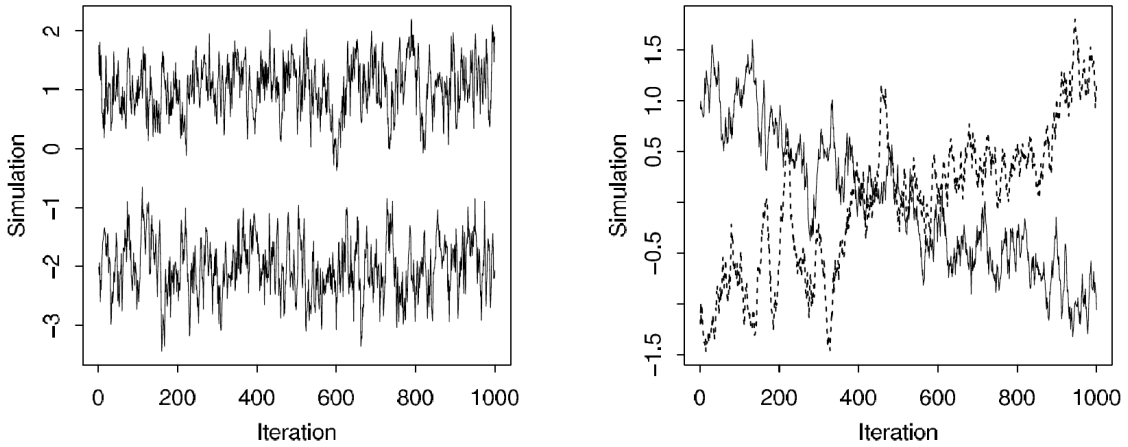


Figure 2.3: Examples of two challenges when assessing convergence. (a) In the left figure, the two chains are stationary. However they have reached different stationary distributions, meaning that none of them have converged. (b) In the right figure, the two chains seem to cover a common distribution. However none of them has reached a stationary state, meaning that they have not converged. When assessing convergence one has to pay attention to both of the two aspects. Adapted from [38].

average $\overline{\psi}_{\cdot\cdot}$ are defined as

$$\overline{\psi}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{i,j} \tag{2.17}$$

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_{i,j} - \overline{\psi}_{\cdot j}\right) \tag{2.18}$$

$$\overline{\psi}_{\cdot\cdot} = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\psi_{i,j} = \frac{1}{m}\sum_{j=1}^{m}\overline{\psi}_{\cdot j} \tag{2.19}$$

The mean within chain empirical variance W and the between chains empirical variance B are computed as

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2 \tag{2.20}$$

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\psi_{\cdot j} - \overline{\psi}_{\cdot\cdot}\right)^2 \tag{2.21}$$

The marginal posterior variance can therefore be estimated from all the chains combined as a weighted average of W and B

$$\widehat{\mathrm{Var}}^{+}(\psi|X) = \frac{n-1}{n}W + \frac{1}{n}B \tag{2.22}$$

If the chains have converged, then both estimates are unbiased. Otherwise the first method will underestimate the variance, since the individual chains have not

had time to range all over the stationary distribution, and the second method will overestimate the variance assuming the starting points to be overdispersed. In the limit $n \to \infty$, when the chains have converged, $\widehat{\mathrm{Var}}^{+}(\psi|X) \to W$ and this converges to the exact marginal posterior variance. The Gelman-Rubin convergence diagnostic of the scalar quantity ψ is defined as

$$\widehat{R}(\psi) = \sqrt{\frac{\widehat{\mathrm{Var}}^{+}(\psi|X)}{W}} \tag{2.23}$$

and it is expected to converge to one when the chains have converged. The diagnostic is also called *potential scale reduction factor*, since it is an estimate of the factor by which the scale of the current distribution for ψ might be reduced if the simulations were continued in the limit $n \to \infty$.
Another aspect one needs to pay attention to when using Markov Chains generated through a Metropolis-Hastings algorithm is *mixing*. Due to the acceptance-rejection procedure an *average acceptance rate* $\overline{\alpha}$ can be defined as

$$\overline{\alpha} = \frac{n_{\text{accepted}}}{n_{\text{tot}}} \tag{2.24}$$

where n_{accepted} and n_{tot} are respectively the number of accepted proposals and the total number of proposals. The average acceptance rate should lay in the range $[0.1, 0.6]$. Having a bigger acceptance means that the proposals are accepted almost every time, thus the chain makes each time only small steps and converges slowly to the stationary distribution. If instead the acceptance is too small the chains is stuck and it takes many iterations to jump in another state. Bad acceptance rates can be fixed performing some preliminary iterations, called *warm-up* or *adaptation phase*, in which the proposal function is tuned in order to reach an optimal acceptance. This can be done, for example, modifying the hyperparameter γ of the sliding window proposal in order to reach an optima average jump distance. Yang and Rodriguez [44] studied the efficiency of different proposal distributions and proposed a method to tune the scale hyperparameter γ. Let $\overline{\alpha}$ be the average acceptance rate of a chain using a scale hyperparameter γ. The optimal scale γ_{opt} is

$$\gamma_{\text{opt}} = \gamma \, \frac{\tan\left(\frac{\pi}{2}\overline{\alpha}\right)}{\tan\left(\frac{\pi}{2}\overline{\alpha}_{\text{opt}}\right)} \tag{2.25}$$

where $\overline{\alpha}_{\text{opt}}$ is the optimal average acceptance rate for the particular proposal distribution (i.e. $\overline{\alpha}_{\text{opt}} \approx 0.44$ for the uniform, normal and Laplace distributions).
Finally, after the proposal function has been tuned and the burn-in period has been discarded, one has to check that a sufficient number of independent samples has been drawn. In general, a Markov Chain sample of a scalar quantity ψ is autocorrelated, since the value drawn at step t depends on the value drawn at step $t-1$, which in turn depends on the value drawn at step $t-2$, and so on. It is possible to define an *Effective Sample Size* for the quantity ψ, denoted as $\mathrm{ESS}(\psi)$, which estimates the number of independent draws of ψ. As explained in [38] and [45], for m chains each of length n, it can be defined as the ratio between the total number of steps mn and

the sum of all the autocorrelations

$$\mathrm{ESS}(\psi) = \frac{mn}{\sum_{t=-\infty}^{+\infty} \rho_t(\psi)} = \frac{mn}{1 + 2\sum_{t=1}^{+\infty} \rho_t(\psi)} \tag{2.26}$$

where $\rho_t(\psi)$ is the autocorrelation of ψ at lag t, defined as

$$\rho_t(\psi) = \frac{1}{\sigma_\psi^2} \int d\psi\, \psi_0 \psi_t\, p(\psi|X) \tag{2.27}$$

with $p(\psi|X)$ being the marginal posterior distribution of ψ and σ_ψ^2 its variance. As explained in [38] a good estimator for $\mathrm{ESS}(\psi)$ is constructed as follows. The autocorrelation is estimated as

$$\widehat{\rho}_t(\psi) = 1 - \frac{V_t(\psi)}{2\,\widehat{\mathrm{Var}}^{+}(\psi|X)} \tag{2.28}$$

where $\widehat{\mathrm{Var}}^{+}(\psi|X)$ is obtained from equation (2.22) and $V_t(\psi)$ is the *variogram* at each lag t

$$V_t(\psi) = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2 \tag{2.29}$$

For large values of t the estimate of the correlation is too noisy. An estimator of $\mathrm{ESS}(\psi)$ can therefore be constructed truncating the sum in the denominator at the first positive odd integer T for which $\widehat{\rho}_T(\psi) + \widehat{\rho}_{T+2}(\psi) < 0$

$$\widehat{\mathrm{ESS}}(\psi) = \frac{mn}{1 + 2\sum_{t=1}^{T} \widehat{\rho}_t(\psi)} \tag{2.30}$$

Convergence and Effective Sample Size should be assessed for any scalar quantity of interest. A good practice rule to stop the simulation is therefore to monitor both the Gelman-Rubin diagnostic $\widehat{R}(\psi)$ and the Effective Sample Size $\widehat{\mathrm{ESS}}(\psi)$ for any quantity ψ of interest, and terminate the MCMC chain only when the diagnostics have reached a desirable value for all the quantities. As mentioned in [38] one should be satisfied when $\widehat{R}(\psi) < 1.1$ and $\widehat{\mathrm{ESS}}(\psi) > 5m$.

The environment-dependent birth-death model can be easily implemented in a Bayesian framework, using a MCMC sampling to infer the speciation and extintion rates from phylogenetic data. We tested the efficiency of the Bayesian implementation and we compared it to the Maximum Likelihood approach, which was already implemented in the R package RPANDA [35].
In particular, we have tested the model using a single environmental dependency with an exponential functional form for the speciation rate

$$\lambda(t) = \lambda_0 e^{\theta_T\, T(t)} \tag{2.31}$$

and either a constant extinction rate

$$\mu(t) = \mu_0 \tag{2.32}$$

or a costant *turnover*[2], which is the ratio between the extinction and the speciation rates

$$\mu(t) = \mu_0, \lambda(t) \tag{2.33}$$

The environmental curve used is the average global change in surface air temperature $T(t)$, inferred from deep-sea benthic foraminifer oxygen isotope δ^{18}O by [46]. The posterior distribution of the parameter space $\{\lambda_0, \mu_0, \theta_T\}$, having observed a phylogenetic tree with branching times $\{t_1, ..., t_n\}$, is therefore

$$p(\lambda_0, \mu_0, \theta_T | t_1, ..., t_n) \propto \mathcal{L}(t_1, ..., t_n | \lambda_0, \mu_0, \theta_T)\, p(\lambda_0)\, p(\mu_0)\, p(\theta_T) \tag{2.34}$$

where $\mathcal{L}(t_1, ..., t_n | \lambda_0, \mu_0, \theta_T)$ is the likelihood of observing such a phylogeny assuming that it was generated by the parameters λ_0, μ_0 and θ_T, given by equation (2.3) or (2.6), while $p(\lambda_0)$, $p(\mu_0)$ and $p(\theta_T)$ are respectively the priors of λ_0, μ_0 and θ_T. We used three weakly informative prior classes to compare the Bayesian implementation with the Maximum Likelihood implementation:

- **Uniform prior class**: λ_0 and μ_0 have a uniform prior in the range $[0, a]$, while θ_T has a uniform prior in the range $[-b, b]$

$$\lambda_0 \sim \mathcal{U}(0, a) \tag{2.35}$$
$$\mu_0 \sim \mathcal{U}(0, a) \tag{2.36}$$
$$\theta_T \sim \mathcal{U}(-b, b) \tag{2.37}$$

- **Normal prior class**: λ_0 and μ_0 have an half-normal prior with variance a^2, while θ_T has a normal prior with mean 0 and variance b^2

$$\lambda_0 \sim \mathcal{HN}(a^2) \tag{2.38}$$
$$\mu_0 \sim \mathcal{HN}(a^2) \tag{2.39}$$
$$\theta_T \sim \mathcal{N}(0, b^2) \tag{2.40}$$

- **Exponential prior class**: λ_0 and μ_0 have an exponential prior with rate a, while θ_T has a Laplace prior with location parameter 0 and scale parameter b

$$\lambda_0 \sim \text{Exp}(a) \tag{2.41}$$
$$\mu_0 \sim \text{Exp}(a) \tag{2.42}$$
$$\theta_T \sim \text{Laplace}(0, b) \tag{2.43}$$

2.3 Bayesian Variable Selection

The use of weakly-informative priors such as the ones described in the previous section still suffers of overparametrization. If one wants to include multiple environmental dependencies, the model should use informative priors which control for

[2]Usually the turnover is denoted ε. Nevertheless, we decided to adopt the notation μ_0 for simplicity, so that the same symbol can be used for the constant extinction rate of the turnover.

overparametrization, reducing the number of possible parameters that can explain the model and correctly identifying signal out of noise.

The problem of fitting the environment-dependent birth-death model with multiple environmental dependencies, of which only a small subset has played an influential role in shaping biodiversity of the observed clade, can be reconducted to the problem of Variable Selection, which is determining the subset of variables that played an influential role in the model. Several methods have been developed to perform Variable Selection. For example, some of them test one variable at a time using stepwise selection methods based on information criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC), or deviance information criterion (DIC). These methods have however some limitations, such as their poor performance in terms of ability to selecting the correct variables with finite samples[47]. Other approaches which can potentially test multiple variables at a time are the penalised likelihood methods, such as LASSO and Ridge penalties. However these methods do not allow for quantifying the uncertainty of the selected variables, and can fail when highly correlated variables are considered [48].

Bayesian Variable Selection methods employ instead a fully Bayesian approach specifying informative priors which limits the number of influential variables. The problem can be simplified as trying to explain an observed variable X with a large number p of explanatory variables $\{E_1, ..., E_p\}$. The goal is to correctly select a small subset of the explanatory variables $\{E_{i_1}, ..., E_{i_k}\}$ whilst controlling for overparametrization. We can formalize the problem as follows. The outcome X is randomly generated according to the likelihood

$$\mathcal{L}(X|\theta_1 E_1 + ... + \theta_p E_p) \tag{2.44}$$

where $\{\theta_1, ..., \theta_p\}$ are correlation parameters, accounting for the strength of the influence of the explanatory variables. The Variable Selection problem consists in determining which of the θ_is are equal to zero. This can be done in a fully Bayesian framework assigning special priors to them.
Several Bayesian Variable Selection techniques were developed, as reviewed in [37]. Among the various methods, one of the first is the *indicator model selection*. It uses an auxiliary indicator variable I_i which can only take values $I_i = 0$ when E_i does not explain the model and $I_i = 1$ when E_i explains the model. Several variants of the method were developed, differing in the way they model the relation between θ_i and I_i and how they assign priors to them. Some examples are the method of Kuo and Mallick [49], which considers the indicators and the correlation factors to be independent and place independent priors on them, or the *discrete mixture* approach, also called *Stochastic Search Variable Selection* (SSVS) [50, 51, 52], which employs a discrete mixture prior for θ_i, where the first one (for $I_i = 0$) is a spike centered in zero and with small variance, while the second one (for $I_i = 1$) is a broad distribution centered in zero and with large variance. An alternative approach, which is the one we used in this project, is to directly assign a prior on θ_i which can control for overparametrization. The method is called *adaptive shrinkage* since it employs *shrinkage priors* [53, 54] which shrink toward zero the θ_is for which no evidence of non-zero values is present in the data, while they have almost no shrinkage effect

on data-supported non-zero correlation parameters. The method is adaptive since the level of sparsness (the number of non-zero correlation parameters) is directly estimated from the data. In contrast with the other method, adaptive shrinkage does not employ indicator variables I_i to determine whether an explanatory variable should be included in the model. A user-defined threshold c should be set, so that E_i is included if and only if $|\theta_i| > c$.

2.4 The horseshoe hyperprior

Even though discrete mixtures are the correct representation of variable selection problems, since they assign a positive prior probability to $\theta_i = 0$, they have several computational difficulties which limit their use in practice. Adaptive shrinkage priors, on the other hand, are more easily implemented. These methods assign normal scale mixture priors to the correlation coefficients

$$\theta_i|\tau_i \sim \mathcal{N}(0, \tau_i^2) \tag{2.45}$$

where, depending on the method, different *hyperpriors* are assigned to the *hyperparameters* τ_is. For example, the exponential mixing of the *LASSO* model [53] $\tau_i^2 \,|\, \tau \sim \text{Exp}(\tau)$ implies independent Laplacian priors for θ_i, while the inverse-gamma mixing of the *Relevance Vector Machine* model [54] $\tau_i^2 \,|\, a, b \sim \text{IG}(a, b)$ implies Student-t priors. The hyperparameters τ, a and b can either be estimated from the data or within a fully Bayesian framework, assigning to them appropriate hyperpriors (for example as in the *Bayesian LASSO* model [55]). One can therefore define the *shrinkage coefficient*

$$\kappa_i = \frac{1}{1+\tau_i^2}, \qquad 0 \leq \kappa_i \leq 1 \tag{2.46}$$

which measures the amount of shrinkage of correlation parameter θ_i. A value of κ_i close to zero means that θ_i has not been shrunk, while a value close to one means that θ_i has been shrunk almost completely. Using tranformation (2.46) together with the specific hyperprior for τ_i, one can find the hyperprior of κ_i. A good shrinkage prior would be concentrated around the two extreme values $\kappa_i = 0$ and $\kappa_i = 1$, in order to correctly distinguish noise from signal and identify the relevant variables. Nevertheless, the Laplacian and Student-t priors do not have these features, as shown in figure 2.4. Being peaked between zero and one, the Laplacian prior tends to over-shrink the large signal parameters and yet under-shrink noise. The Student-t prior, on the other hand, tends to over-shrink the signal parameters. An elegant solution was proposed by Carvalho, Polson and Scott [56, 57]. They considered an half-Cauchy scale mixing with both a *global shrinkage hyperparameter* τ and local shrinkage hyperparameters ε_i

$$\theta_i \,|\, \varepsilon_i, \tau \sim \mathcal{N}(0, \varepsilon_i^2\, \tau^2) \tag{2.47}$$

$$\varepsilon_i \sim \text{C}^+(0, 1) \tag{2.48}$$

$$\tau \sim \text{C}^+(0, 1) \tag{2.49}$$

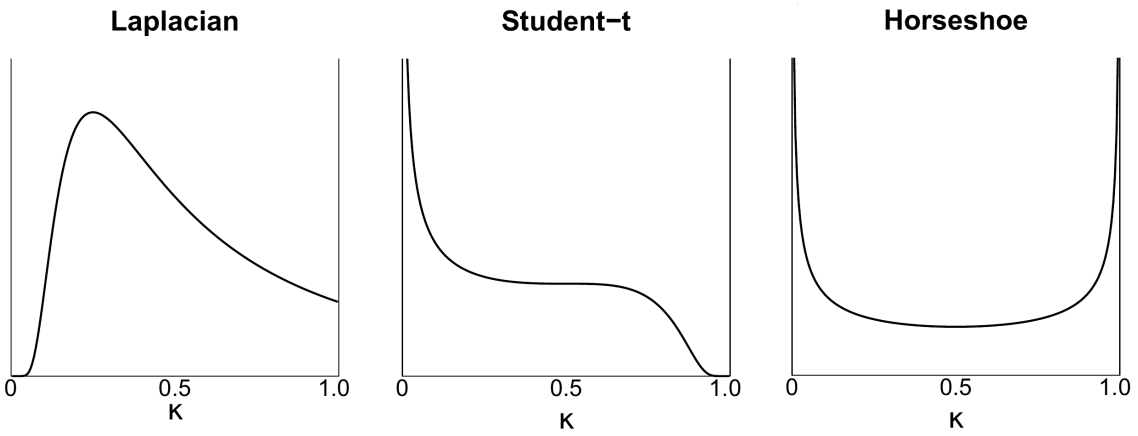


Figure 2.4: Examples of three hyperpriors for the shrinkage coefficient κ_i. (a) The Laplacian hyperprior is concentrated in between zero and one rather than at the extremes. The shrinkage of noise, as well as the detection of signal, will be thus inefficient. (b) The Student-t hyperprior, in the central figure, is concentrated in zero but has a vanishing probability at one. The shrinkage of noise would be efficient, however the correct detection of signal would be inefficient. (c) The Horseshoe hyperprior has both the desired features. It is concentrated in zero and one, while it has a low probability in the middle, yielding to an efficient shrinkage of the noise and an efficient detection of the signal. Adapted from [56].

where $\mathrm{C}^+(a, b)$ is the half-Cauchy distribution with location parameter a and scale parameter b. The resulting prior is called *Horseshoe prior*, and does not have a closed-form expression, even though tight bounds are obtained in [57].
The name of the distribution comes from the shape of the hyperprior of the shrinkage coefficients κ_i, which are now defined as

$$\kappa_i = \frac{1}{1 + \varepsilon_i^2 \tau^2} \tag{2.50}$$

Figure 2.4 compares the Horseshoe hyperprior with the Laplacian and the Student-t hyperpriors. We can see that the Horseshoe possesses all the features which define a well behaving shrinkage prior, since it diverges in the boundaries zero and one and has a low probability density in the middle. Figure 2.5 compares the Horseshoe with the Laplacian and the Student-t distributions. The Horseshoe prior behaves essentially as $\log(1 + 2/\theta_i^2)$. It possesses two interesting features, which makes it an interesting alternative to perform Bayesian Variable Selection. Its infinitely tall spike at zero yields noise parameters to be shrunk toward zero. Yet, its flat, Cauchy-like tails allow signals of potentially strong positive or negative intensity to remain un-shrunk. This interesting features are a result of the clear separation between the global and local shrinkage effects. The global shrinkage parameter tries to estimate the overall sparsity level, while the local shrinkage parameters are able to distinguish the non-zero signal correlation parameters from the zero noise correlation parameters. Heavy tails for the local shrinkage hyperparameters are fundamental in this process, allowing the estimates of signal parameters θ_i to escape the strong "gravitational pull" towards zero exercised by τ. Put another way, the horseshoe

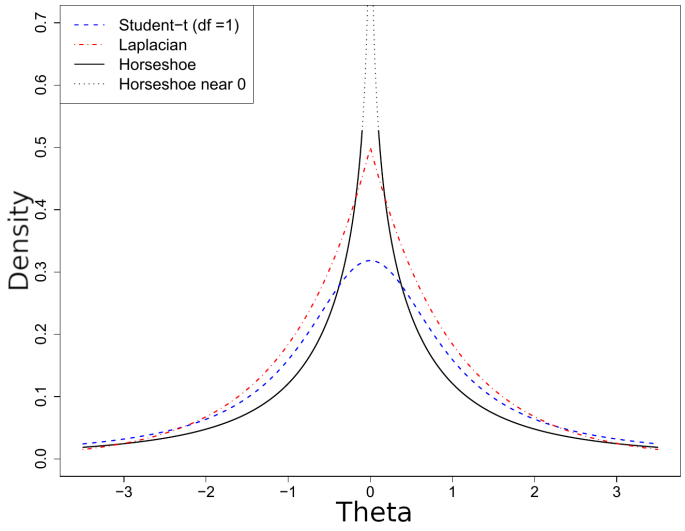


Figure 2.5: Comparison of the Horseshoe prior distribution (black solid line) for the correlation parameters θ_i with the Laplacian (red dashed line) and the Student-t (blue dashed line) priors. The Horseshoe infinitely tall spike at zero shrunk the noise parameters toward zero, while its Cauchy-like tails allow signal to remain unshrunk. The Laplacian and Student-t, on the other hand, have a finite shrinkage effect due to their bounded distributions. Adapted from [56].

has the freedom to shrink globally (via τ) and yet act locally (via ε_i). This is not possible under the Laplacian and Student-t priors, whose shrinkage profiles force a compromise between shrinking noise and determining signals.
In order to define a threshold rule to discriminate between signal and noise, we used the *shrinkage weights* w_i proposed by Carvalho, Polson and Scott [57]

$$w_i = 1 - \kappa_i = 1 - \frac{1}{1 + \varepsilon_i^2 \, \tau^2} \tag{2.51}$$

and we inferred its value $\widehat{w}_i$ from the posterior, for example using the Maximum A Posteriori (MAP) value, the mean or the median. An estimator $\widehat{w}_i$ lower than 0.5 means that the variable E_i should not be included in the model, while a value greater than 0.5 means that the variable E_i should be included in the sample.

$$\begin{cases} \widehat{w}_i < 0.5 & E_i \text{ noise} \\ \widehat{w}_i \geq 0.5 & E_i \text{ signal} \end{cases} \tag{2.52}$$

We used the Horseshoe prior to fit the environment-dependent birth-death model with multiple dependencies avoiding the risk of overparametrization. Let's consider a set of p environmental curves $\{E_1(t), ..., E_p(t)\}$. We considered, for simplicity, an exponential speciation rate with multiple environmental dependencies

$$\lambda(t) = \lambda_0 \exp\left(\theta_1 \, E_1(t) + ... + \theta_p \, E_p(t)\right) \tag{2.53}$$

and a constant extinction rate

$$\mu(t) = \mu_0 \tag{2.54}$$

where $\{\theta_1, ..., \theta_p\}$ are the correlation parameters of the environmental variables and the speciation rate. We place the Horseshoe prior only on these correlation parameters. The Bayesian model is therefore

$$t_1, ..., t_n \sim \mathcal{L}(t_1, ..., t_n \,|\, \lambda_0, \mu_0, \theta_1, ..., \theta_p) \tag{2.55}$$
$$\theta_i \,|\, \varepsilon_i, \tau \sim \mathcal{N}(0, \varepsilon_i^2\, \tau^2) \tag{2.56}$$
$$\varepsilon_i \sim \mathrm{C}^+(0, 1) \tag{2.57}$$
$$\tau \sim \mathrm{C}^+(0, 1) \tag{2.58}$$

while we can assign uninformative priors to λ_0 and μ_0, such as uniform, exponential or half-normal. The influence of each environmental variable $E_i(t)$, as well as the magnitude and the direction of this influence, can thus be tested by sampling the posterior distribution and inferring the posterior estimates for the shrinkage weights $\widehat{w}_i$ and the correlation factors $\widehat{\theta}_i$.

2.5 Model reparametrization and MCMC implementation

In order to compute the posterior estimates of the shrinkage weights and the correlation factors we need to perform a MCMC sampling of the posterior distribution. Nevertheless, the complex hierarchical structure of the Horseshoe prior makes the sampling almost impossible. At each MCMC step the acceptance probability of the correlation parameters θ_i strongly depends on the current value of the local shrinkage hyperparameters ε_i and of the global shrinkage hyperparameter τ. The bigger the product $\varepsilon_i^2\, \tau^2$ is, the broader the normal distribution $\mathcal{N}(0, \varepsilon_i^2\, \tau^2)$ and, in turn, the greater is the probability of accepting a proposal θ_i^* for the correlation parameter. The issue hides in the strong correlation between the correlation parameters and the shrinkage hyperparameters.

In order to obtain a more efficient sampler for the Horseshoe, we reparametrized the model in such a way that the sampling is performed on independent random variables. In order to do so, we can use a well known property of normal random variables. Given a normal random variable X with mean μ and variance σ^2, it can always be obtained from a unit normal random variable Y through the transformation

$$X = \sigma Y + \mu \qquad \text{where} \qquad \begin{cases} X \sim \mathcal{N}(\mu, \sigma^2) \\ Y \sim \mathcal{N}(0, 1) \end{cases} \tag{2.59}$$

Then the Horseshoe model can be reparametrized as

$$t_1, ..., t_n \sim \mathcal{L}(t_1, ..., t_n \,|\, \lambda_0, \mu_0, \theta_1, ..., \theta_p) \tag{2.60}$$
$$\tilde{\theta}_i \sim \mathcal{N}(0, 1) \tag{2.61}$$
$$\varepsilon_i \sim \mathrm{C}^+(0, 1) \tag{2.62}$$
$$\tau \sim \mathrm{C}^+(0, 1) \tag{2.63}$$

where

$$\theta_i \,|\, \tilde{\theta}_i, \varepsilon_i, \tau = \varepsilon_i\, \tau\, \tilde{\theta}_i \tag{2.64}$$

In addition, in order to efficiently sample the shrinkage hyperparameters, which have heavy-tailed half-Cauchy hyperpriors, we can use another reparametrization called *inverse transform sampling*, which allows to sample random variables with invertible Cumulative Distribution Function (CDF) from uniform random variables. The half-Cauchy distribution with location parameter $a = 0$ and scale parameter b

$$p_X(x) = \frac{2}{\pi b} \frac{1}{1 + x^2/b^2} \qquad x > 0 \tag{2.65}$$

has CDF

$$F_X(x) = \int_0^x dx\, p_X(x) = \frac{2}{\pi} \tan^{-1}\left(\frac{x}{b}\right) \tag{2.66}$$

The *probability integral transform* (see [58], Theorem 2.1.10 p.54) states that the random variable $u = F_X(X)$ has a uniform distribution

$$u = F_X(X) = \frac{2}{\pi} \tan^{-1}\left(\frac{X}{b}\right) \sim \mathcal{U}(0, 1) \tag{2.67}$$

Then, inverting equation (2.67) we can sample an half-Cauchy random variable directly from a uniform random variable

$$X = F_X^{-1}(u) = b \tan\left(\frac{\pi}{2} u\right) \qquad \text{where} \qquad \begin{cases} X \sim \mathrm{C}^+(0, b) \\ u \sim \mathcal{U}(0, 1) \end{cases} \tag{2.68}$$

Inserting the reparametrization into the model we obtain

$$t_1, ..., t_n \sim \mathcal{L}(t_1, ..., t_n \,|\, \lambda_0, \mu_0, \theta_1, ..., \theta_p) \tag{2.69}$$
$$\tilde{\theta}_i \sim \mathcal{N}(0, 1) \tag{2.70}$$
$$u_i \sim \mathcal{U}(0, 1) \tag{2.71}$$
$$u \sim \mathcal{U}(0, 1) \tag{2.72}$$

where

$$\theta_i \,|\, \tilde{\theta}_i, \varepsilon_i, \tau = \varepsilon_i \, \tau \, \tilde{\theta}_i \tag{2.73}$$
$$\varepsilon_i \,|\, u_i = \tan\left(\frac{\pi}{2} u_i\right) \tag{2.74}$$
$$\tau \,|\, u = \tan\left(\frac{\pi}{2} u\right) \tag{2.75}$$

With this reparametrization we implemented a simple Metropolis-Hastings sampler with pre-adaptation to tune the proposal function. We used a simple exponential prior for the parameters λ_0 and μ_0.

Chapter 3

Results

3.1 Likelihood speed up

In order to implement the environment-dependent birth-death model in a Bayesian framework we need to perform a MCMC sampling of the posterior. This requires to compute the likelihood several times, which will take too much time due to the computational burden of the integrals in equation (2.4) and (2.5). The computational time can be reduced using a sufficiently large approximation interval dt for the calculation of the integrals by piece-wise approximation. This has already been implemented in the package RPANDA [35], where a value $dt = 10^{-3}$ was suggested as a safe approximation. However the execution time is still too big to efficiently implement the model within a Bayesian framework. In order to decrease it we avoided some unnecessary calculations, dramatically reducing the computation of the likelihood. In this chapter we will refer to the implementation of the package RPANDA as "old likelihood" and to the new implementation as "new likelihood".

In order to benchmark the new likelihood and quantitatively test its increase in performances we considered the environment-dependent birth death-model with multiple environmental dependencies. In particular we considered five environmental curves: the average global change in surface air temperature $T(t)$, the average sea level $h_{\text{sea}}(t)$, the benthic foraminifera isotopic signature $\delta^{13}C(t)$ (all of them obtained from [9]), the average carbon dioxide concentration $\text{CO2}(t)$ and the Silica weathering ratio $\text{Si}(t)$. We considered an exponential speciation rate and a constant extinction rate

$$\lambda(t) = \lambda_0 \exp\left(\theta_T T(t) + \theta_h h_{\text{sea}}(t) + \theta_\delta \delta^{13}C(t) + \theta_{CO2}\text{CO2}(t) + \theta_{\text{Si}}\text{Si}(t)\right) \tag{3.1}$$

$$\mu(t) = \mu_0 \tag{3.2}$$

We considered three integral approximation intervals $dt = 10^{-2}$, $dt = 10^{-3}$ and $dt = 10^{-4}$. For each of them we computed, with both the old and the new methods, the likelihood of the model for 1000 randomly generated parameter sets $\{\lambda_0, \mu_0, \theta_T, \theta_h, \theta_\delta, \theta_{CO2}, \theta_{\text{Si}}\}$. We then compared the execution time and the log-likelihood of the two methods. The results are shown in figure 3.1. As we can see, the computation time is drastically reduced with an approximation interval for the

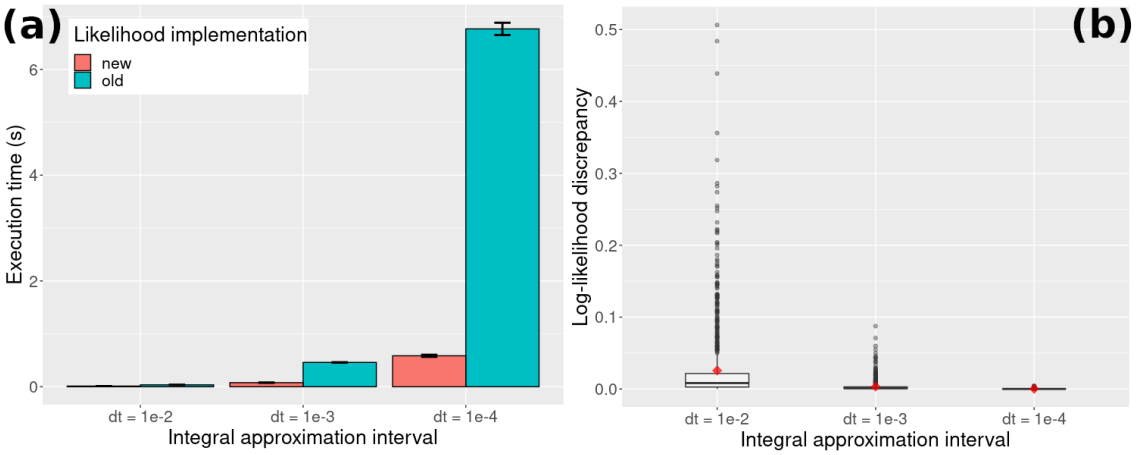


Figure 3.1: Comparison of the new implementation of the computation of the likelihood with the old implementation of the package RPANDA. (a) In the left panel, the average computational time is compared. As we can see the new likelihood is dramatically quicker than the old one for $dt = 10^{-4}$. (b) In the right panel is shown the discrepancy between the log-likelihood computed by the new implementation and the one computed by the old implementation. $dt = 10^{-4}$ has almost no discrepancy, being therefore a safe choice. Red dot represent mean values, while the bottom, middle and upper lines of the boxes represent respectively the lower quartile (Q_1), the median and the upper quartile (Q_3). Black points represent outliers which fall outside the range $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$, which is specified by the vertical black line.

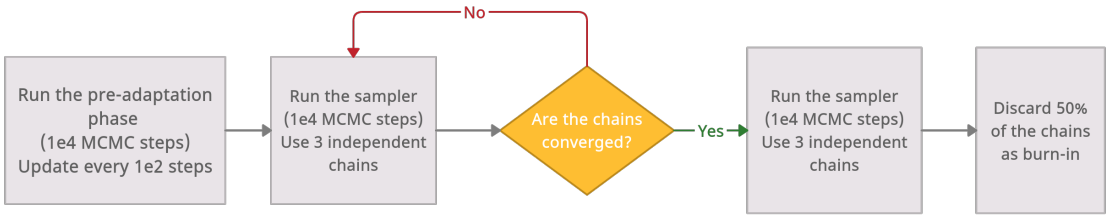


Figure 3.2: Flowchart of the feedback process implemented to automatically generate converged MCMC chains. Three independent chains are generated in parallel. First of all, the pre-adaptation phase is run for each chain using 10^4 MCMC steps. The scale parameters γ_i of the proposal functions for each parameter are updated every 10^2 steps using equation (2.25). Once the pre-adaptation phase is terminated, the chains are sampled using a MH algorithm with the optimised proposal functions. Convergence is checked every 10^4 MCMC steps, checking that $\widehat{R}(\psi) < 1.05$ and $\widehat{\mathrm{ESS}}(\psi) > 200$ for any quantity of interest ψ. Once the chains have converged, 10^4 more steps are computed, and finally 50% of the chains is discarded as burn-in.

integral $dt = 10^{-4}$. If we then look at the discrepancy between the values of the likelihood computed with the old and the new method, we can see that this approximation interval corresponds to almost no discrepancy. We have therefore decided to use $dt = 10^{-4}$ as a safe approximation interval in the Bayesian implementation. The new likelihood computation ensures a sufficiently small computation time, making possible to perform the MCMC sampling.

3.2 Bayesian implementation

We implemented the model in a Bayesian framework using Metropolis-Hastings MCMC sampling of the posterior distribution (2.34), adapted from the algorithm developed by Maliet, Hartig and Morlon [59]. In order to be sure to always sample converged chains we developed an automated flowchart which monitors convergence during the simulation and stops it as soon as convergence is reached. Figure 3.2 shows a schematic representation of the procedure. Each run simulates three independent chains in parallel, in order to apply equations (2.23) and (2.30) to assess convergence. During each MCMC step we update all the parameters one at a time, each time carrying out the procedure of proposing a new parameter and accepting (or rejecting) the new proposal based on the acceptance rate. Then, if one has a p-dimensional parameter space, this implies to compute the likelihood p times per MCMC step. This procedure makes the convergence and the mixing of the chains faster, at the price of increasing the computation time per MCMC step. Before sampling the chains we apply a pre-adaptation phase using 10^4 MCMC steps, in which the proposal functions for each parameter are tuned. In particular we update the scale parameters γ_i of the proposal functions for parameters θ_i every 10^2 MCMC steps using the update rule (2.25). Once the adaptation phase is concluded, the tuned proposal functions are used to sample the three chains. We assess conver-

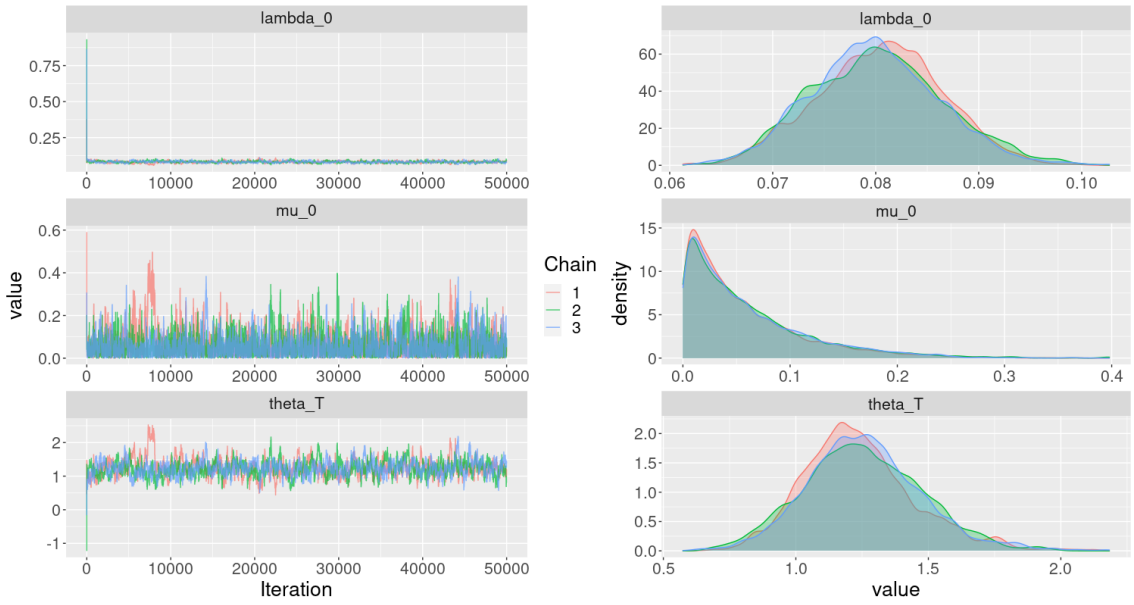


Figure 3.3: Traces (left panels) and marginal posterior densities (right panels) of the three parameters λ_0, μ_0 and θ_T. The trace plots clearly show the burn-in period at the beginning of the chains. Th marginal posterior densities are computed discarding the first half of the chains as burn-in.

Table 3.1: Settings of the different prior classes specified in equations (2.35)-(2.43).

Prior class	a	b
Uniform	5	5
Normal	2	2
Exponential	4	2

gence each 10^4 MCMC steps, checking that

$$\begin{cases} \widehat{R}(\psi) < 1.05 \\ \widehat{\text{ESS}}(\psi) > 200 \end{cases} \tag{3.3}$$

for any quantity of interest ψ. Once convergence is reached for any ψ, 10^4 more MCMC steps are simulated, and finally half of the chain is discarded as burn-in. In order to quantitatively test the Bayesian implementation we confronted it with the Maximum Likelihood implementation already implemented in the RPANDA package. We used the simple model described in section 2.2 with an exponential dependency of the speciation rate from the average temperature (2.31) and either a constant extinction rate (2.32) or a constant turnover (2.33). We used the three different prior classes described in the aforementioned section, setting $a = b = 5$ for the uniform class, $a = b = 2$ for the normal class and $a = 4$ and $b = 2$ for the exponential class. The settings are schematized in table 3.1. We randomly simulated 100 phylogenetic trees for both the constant extinction rate model and the constant turnover model using the function `sim_env_bd` of the RPANDA package, using the same parameters for all of them. The setup of the simulations is resumed

Table 3.2: Settings of the parameters used in the simulations of the phylogenetic trees.

	Constant extinction rate	Constant turnover
λ_0	0.10	0.08
μ_0	0.06	0.06
θ_T	1.30	1.10

in table 3.2. For each simulated tree we sampled the posterior probability density using the aforementioned Metropolis-Hastings MCMC sampling procedure and the three different prior classes. We checked convergence of the three parameters λ_0, μ_0, θ_T.

Figure 3.3 shows the traces and the marginal posterior densities of the three parameters λ_0, μ_0 and θ_T for a tree generated with a constant turnover, considering exponential priors. Traces and marginal posterior densities are represented for each of the three parallel independent chains. The marginal posterior distributions were computer disregarding the first halves of the chains as burn-in. The burn-in period at the beginning of the chains is well visible. All the three chains seem to have converged to the same posterior distribution and are stationary.

We then compared the ability of the Bayesian implementation to correctly infer the model parameters with the Maximum Likelihood implementation. We used the posterior means $\widehat{\lambda}_0 = \mathbb{E}(\lambda_0|t_1, ..., t_n)$, $\widehat{\mu}_0 = \mathbb{E}(\mu_0|t_1, ..., t_n)$ and $\widehat{\theta}_T = \mathbb{E}(\theta_T|t_1, ..., t_n)$ as estimators of our parameters, and we compared them to the Maximum Likelihood estimators. In order to reduce the correlation between samples and reduce the memory usage of the algorithm we saved the sampled parameters only every 10^2 MCMC steps. The procedure is called *thinning*. The results of the comparison are shown in figure 3.4. As we can see, the Bayesian implementation correctly estimates the parameters with less variance than the Maximum Likelihood method. However, in the case of a constant turnover both the Bayesian and the Maximum Likelihood implementation fail to correctly identify the extinction parameter μ_0. This can be due to the well known difficulty of fitting the extinction process from reconstructed phylogenies, which lack extinction events not including fossils.

We then analysed the dependencies of the accuracy in the estimates from the size of the simulated trees, which is the number of tips (extant species). Figure 3.5 shows the results. We can see that the accuracy strongly depends on the tree size, in particular for the extinction parameter μ_0. Using the Bayesian implementation can then help in estimating the uncertainty of the inferred parameters even with small phylogenetic trees, since it provides the full conditional posterior ditribution rather than a simple point estimates as for the Maximum Likelihood implementation.

Finally, in order to assess the ability of the Bayesian implementation to correctly fit the model with only two environmental variables, we randomly simulated 100 phylogenetic trees using an exponential speciation rate depending on the temperature

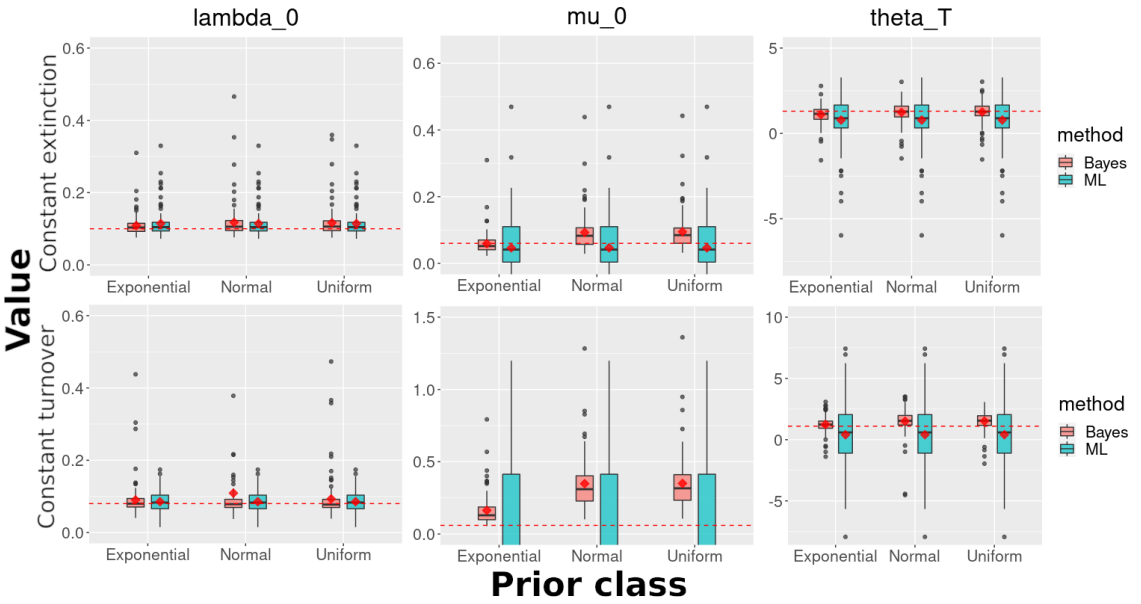


Figure 3.4: Comparison of the efficiency of the Bayesian implementation (denoted "Bayes" in the caption) and the Maximum Likelihood implementation ("ML" in the caption). Upper panels correspond to a constant extinction rate, while lower panels correspond to a constant turnover. Plots for μ_0 were bounded to $\mu_0 > 0$ in order to appreciate the difference between the two methods. Horizontal red dashed lines correspond to the true value of the parameters used in the simulation setup.

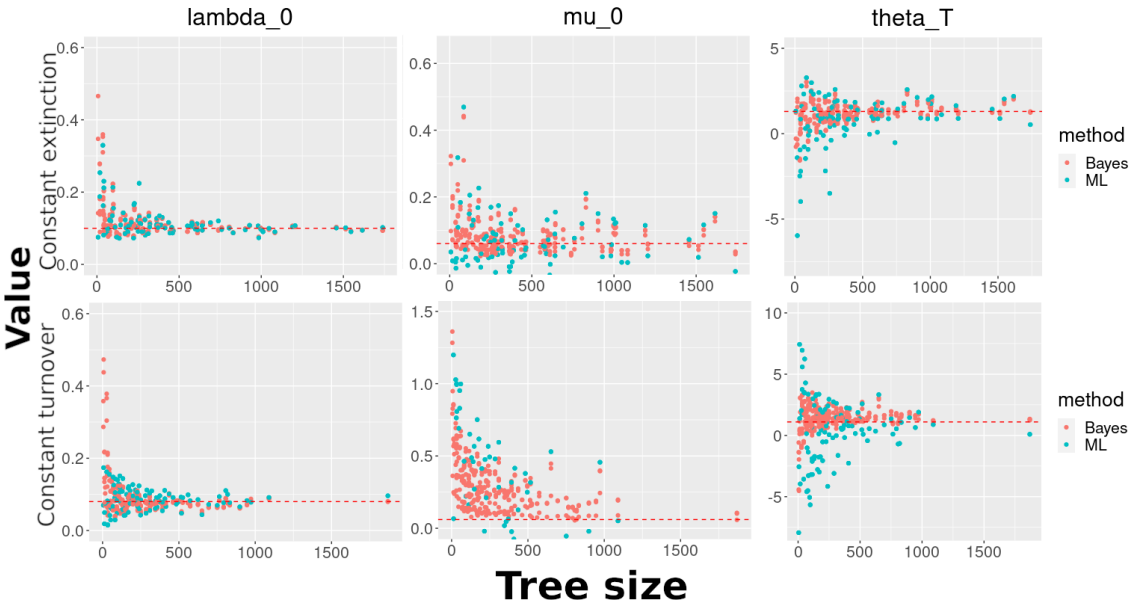


Figure 3.5: Upper panels correspond to a constant extinction rate, while lower panels correspond to a constant turnover. Plots for μ_0 were bounded to $\mu_0 > 0$ in order to appreciate the difference between the two methods. Horizontal red dashed lines correspond to the true value of the parameters used in the simulation setup.

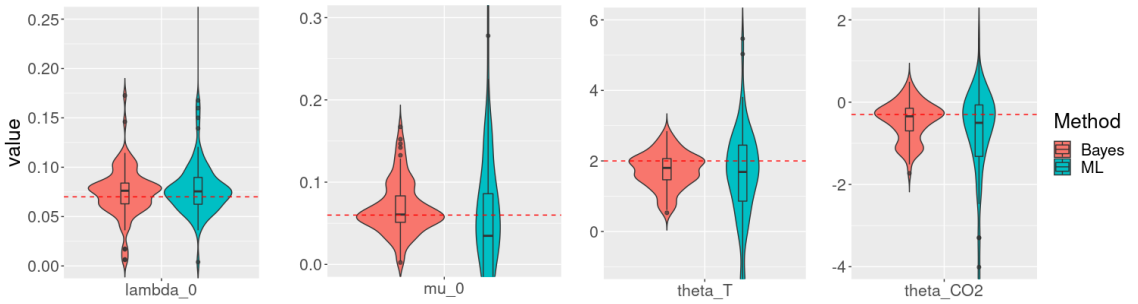


Figure 3.6: Comparison of the efficiency of the Bayesian implementation (denoted "Bayes" in the caption) and the Maximum Likelihood implementation ("ML" in the caption) using a constant extinction rate and an exponential extintion rate depending from the temperature and the carbon dioxide. Densities and summary statistics of the estimates of each parameter are shown, together with the true values of the parameters used in the simulation setup (red dashed lines).

and the carbon dioxide

$$\lambda(t) = \lambda_0 \exp(\theta_T T(t) + \theta_{\mathrm{CO2}} \mathrm{CO2}(t)) \tag{3.4}$$

and a constant extinction rate

$$\mu(t) = \mu_0 \tag{3.5}$$

We considered a strong positive dependency from the temperature and a small negative dependency from the carbon dioxide, setting the parameters as $\lambda_0 = 0.07$, $\mu_0 = 0.06$, $\theta_T = 2$ and $\theta_{\mathrm{CO2}} = -0.3$. This will avoid to confound the effect of the two environmental variables and better infer their correlation parameters. We then inferred the parameters using the procedure explained before, with an exponential prior family

$$\lambda_0 \sim \mathrm{Exp}(4) \tag{3.6}$$
$$\mu_0 \sim \mathrm{Exp}(4) \tag{3.7}$$
$$\theta_T \sim \mathrm{Laplace}(0, 2) \tag{3.8}$$
$$\theta_{\mathrm{CO2}} \sim \mathrm{Laplace}(0, 2) \tag{3.9}$$

We checked convergence of the four parameters λ_0, μ_0, θ_T and θ_{CO2}. The estimated parameters are shown in figure 3.6 and are compared with the ones obtained from a Maximum Likelihood approach. As we can see the Bayesian implementation is slightly better than the Maximum Likelihood approach, since it has less variance in the estimates of the correlation coefficients θ_T and θ_{CO2}. The difference however is not as sharp as in the single environmental dependency model. This is due to the use of uninformative priors. The strength of a Bayesian approach is that it can employ informative priors to better estimate multivariate models without encountering the problem of overparametrization.

3.3 Bayesian Variable Selection

The Bayesian Variable Selection technique employing horseshoe priors, described in sections 2.3 - 2.5, can be implemented within the automated flowchart described in

section 3.2. Nonetheless, the number of parameters to be sampled is way larger than the simple environment-dependent birth-death model with a single environmental dependency. We cannot therefore update one parameter at a time for each MCMC sample, since this would require to compute the likelihood too many times and slowing down the sampler. A solution to the problem which still ensures fast enough convergence and mixing of the chains is to group the parameters into a fixed number of blocks and update one block at a time for each MCMC step. The blocking procedure can either be defined *a priori* or randomly.

We implemented the Bayesian Variable Selection procedure using random blocking within the automated flowchart schematized in figure 3.2. In order to meet a good compromise between mixing and computational time, we decided to take four blocks. We considered for simplicity only five environmental variables, which are again the average global change in surface air temperature $T(t)$, the average sea level $h_{\text{sea}}(t)$, the benthic foraminifera isotopic signature $\delta^{13}C(t)$ (all of them obtained from [9]), the average carbon dioxide concentration $\text{CO2}(t)$ and the Silica weathering ratio $\text{Si}(t)$. In addition, we slightly modified the algorithm to sample from the reparametrized model of equations (2.69) - (2.75). In particular, while the sampling is done on the parameters $\{\lambda_0, \mu_0, \tilde{\theta}_1, ..., \tilde{\theta}_p, u_1, ..., u_p, u\}$, at each updating step the original parameters $\{\lambda_0, \mu_0, \theta_1, ..., \theta_p, \varepsilon_1, ..., \varepsilon_p, \tau\}$, are computed and the acceptance probability is obtained from them.
Since the objective of the sampling is to infer the parameters of the model λ_0, μ_0, θ_T, θ_h, θ_δ, θ_{CO2} and θ_{Si} and the corresponding shrinkage weights, in order to determining the environmental variables which played a role in shaping biodiversity of the observed clade, we monitored only the convergence of these quantities through their Gelman-Rubin convergence diagnostic and their Effective Sample Size.

Unfortunately, due to the short length of the internship, we were not able to test the implementation with simulated phylogenetic trees, in order to asses its ability to correctly identifying the environmental dependency used to generate the trees. The code is available at [60], and it will be tested in the future.

Chapter 4

Discussion

The study of biodiversity and how it is shaped by abiotic and biotic factors, such as human induced climate change, is key to understand how our actions will affect the wellness of wildlife, and to create more efficient preservation policies. At present, one of the most used tools in the study of biodiversity is inference models on reconstructed phylogenetic trees, such as the environment-dependent birth-death model. Yet, these methods are often implemented in a Maximum Likelihood framework, and have some serious limitations in fitting multivariate environmental dependencies. In this report we presented an efficient Bayesian implementation of the environment-dependent birth death model, which can give a better interpretation of inference estimates through the use of the full marginal posterior distributions instead of simple point estimates, such as the one obtained in Maximum Likelihood based approaches.

We tested the accuracy of the Bayesian implementation using both a single environmental dependency and two environmental dependencies, of which one has a strong effect while the other has a small effect. The Bayesian implementation has shown to outperform the Maximum Likelihood implementation, being more accurate and more precise. However, directly fitting the model with uninformative priors has similar problems as a Maximum Likelihood implementation. If one wants to consider multivariate environment-dependent models, it should include informative priors such as the Horseshoe.

Even though we were not able to test the Bayesian Variable Selection implementation, the code has been made available on [60] for future work. Once the procedure has been tested, improvements of the sampling algorithm can be made. For example, instead of using the suggested reparametrization one can use a slice sampling algorithm, as done by Silvestro et al. in [61].

In cocnlusion, being able to implement the model in a Bayesian framework can help future researches to obtain better inferences from reconstructed phylogenies.

Bibliography

[1] Charles R. Darwin and Alfred R. Wallace. "On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection". In: *Journal of the Proceedings of the Linnean Society of London. Zoology* 3 (1858), pp. 46–50.

[2] Julian Huxley. *Evolution: The Modern Synthesis*. Allen & Unwin London, 1942.

[3] Stephen J. Gould. *The Structure of Evolutionary Theory*. Belknap Press of Harvard University Press, 2002.

[4] Jur Philiptschenko. *Variabilitat Und Variation*. Gebrüder Borntraege, Berlin, 1927.

[5] Carl Linnaeus. *Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera, & species*. Haak, Leiden, 1735.

[6] Carl Linnaeus. *Species Plantarum*. Stockholm, Sweden, 1735.

[7] Carl Linnaeus. *Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera, & species*. 10th edition. Haak, Leiden, 1758.

[8] Steven M. Stanley. "A theory of evolution above the species level". In: *Proceedings of the National Academy of Sciences* 72.2 (1975), pp. 646–650. DOI: 10.1073/pnas.72.2.646.

[9] Michael Hautmann. "What is macroevolution?" In: *Palaeontology* 63.1 (2020), pp. 1–11. DOI: https://doi.org/10.1111/pala.12465.

[10] "Stochastic Models of Phylogeny and the Evolution of Diversity". In: *The Journal of Geology* 81.5 (1973), pp. 525–542. ISSN: 00221376, 15375269. URL: http://www.jstor.org/stable/30060095 (visited on 06/06/2022).

[11] Michael Foote et al. "Rise and Fall of Species Occupancy in Cenozoic Fossil Mollusks". In: *Science* 318.5853 (2007), pp. 1131–1134. DOI: 10.1126/science.1146303.

[12] Thomas H. G. Ezard et al. "Interplay Between Changing Climate and Species; Ecology Drives Macroevolutionary Dynamics". In: *Science* 332.6027 (2011), pp. 349–351. DOI: 10.1126/science.1203060.

[13] Tiago B. Quental and Charles R. Marshall. "How the Red Queen Drives Terrestrial Mammals to Extinction". In: *Science* 341.6143 (2013), pp. 290–292. DOI: 10.1126/science.1239431.

[14] Jody Hey. "Using phylogenetic trees to study speciation and extinction". In: *Evolution* 46.3 (1992), pp. 627–640. DOI: https://doi.org/10.1111/j.1558-5646.1992.tb02071.x.

[15] Sean Nee et al. "Extinction rates can be estimated from molecular phylogenies". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1307 (1994), pp. 77–82. DOI: 10.1098/rstb.1994.0054.

[16] Sean Nee, Robert Mccredie May, and Paul H. Harvey. "The reconstructed evolutionary process". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1309 (1994), pp. 305–311. DOI: 10.1098/rstb.1994.0068.

[17] Paul H. Harvey, Robert M. May, and Sean Nee. "Phylogenies without fossils". In: *Evolution* 48.3 (1994), pp. 523–529. DOI: https://doi.org/10.1111/j.1558-5646.1994.tb01341.x.

[18] Arne O. Mooers and Stephen B. Heard. "Inferring Evolutionary Process from Phylogenetic Tree Shape". In: *The Quarterly Review of Biology* 72.1 (1997), pp. 31–54. DOI: 10.1086/419657.

[19] A.O. Mooers et al. "Some models of phylogenetic tree shape". In: *Reconstr. Evol. New Math. Comput. Adv.* (2007), pp. 149–170.

[20] Robert E. Ricklefs. "Estimating diversification rates from phylogenetic information". In: *Trends in Ecology & Evolution* 22.11 (2007), pp. 601–610. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2007.06.013.

[21] Matthew W. Pennell and Luke J. Harmon. "An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology". In: *Annals of the New York Academy of Sciences* 1289.1 (2013), pp. 90–105. DOI: https://doi.org/10.1111/nyas.12157.

[22] R. Alexander Pyron and Frank T. Burbrink. "Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses". In: *Trends in Ecology & Evolution* 28.12 (2013), pp. 729–736. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2013.09.007.

[23] Hélène Morlon. "Phylogenetic approaches for studying diversification". In: *Ecology Letters* 17.4 (2014), pp. 508–525. DOI: https://doi.org/10.1111/ele.12251.

[24] Alexander von Humboldt. *Ansichten der Natur: mit wissenschaftliehen Erläuterungen*. Stuttgart, J.G. Cotta, 1849.

[25] Alfred R. Wallace. *Tropical nature, and other essays*. London, Macmillan and co, 1878.

[26] Lee Van Valen. "A new evolutionary law". In: *Evolutionary theory* 1 (1973), pp. 1–30. ISSN: 0093-4755.

[27] Anthony D. Barnosky. "Distinguishing the effects of the Red Queen and Court Jester on Miocene mammal evolution in the northern Rocky Mountains". In: *Journal of Vertebrate Paleontology* 21.1 (2001), pp. 172–185. DOI: 10.1671/0272-4634(2001)021[0172:DTEOTR]2.0.CO;2.

[28] "Red Queen: from populations to taxa and communities". In: *Trends in Ecology & Evolution* 26.7 (2011), pp. 349–358. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2011.03.016.

[29] J. John Sepkoski. "A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions". In: *Paleobiology* 10.2 (1984), pp. 246–267. DOI: 10.1017/S0094837300008186.

[30] Peter J. Mayhew, Gareth B. Jenkins, and Timothy G. Benton. "A long-term association between global temperature and biodiversity, origination and extinction in the fossil record". In: *Proceedings of the Royal Society B: Biological Sciences* 275.1630 (2008), pp. 47–53. DOI: 10.1098/rspb.2007.1302.

[31] R.A. Spicer and J.L. Chapman. "Climate change and the evolution of high-latitude terrestrial vegetation and floras". In: *Trends in Ecology & Evolution* 5.9 (1990), pp. 279–284. ISSN: 0169-5347. DOI: https://doi.org/10.1016/0169-5347(90)90081-N.

[32] Clément Coiffard et al. "Rise to dominance of angiosperm pioneers in European Cretaceous environments". In: *Proceedings of the National Academy of Sciences* 109.51 (2012), pp. 20955–20959. DOI: 10.1073/pnas.1218633110.

[33] Fabien L. Condamine, Jonathan Rolland, and Hélène Morlon. "Macroevolutionary perspectives to environmental change". In: *Ecology Letters* 16.s1 (2013), pp. 72–85. DOI: https://doi.org/10.1111/ele.12062.

[34] Hélène Morlon, Todd L. Parsons, and Joshua B. Plotkin. "Reconciling molecular phylogenies with the fossil record". In: *Proceedings of the National Academy of Sciences* 108.39 (2011), pp. 16327–16332. DOI: 10.1073/pnas.1102543108.

[35] Hélène Morlon et al. "RPANDA: an R package for macroevolutionary analyses on phylogenetic trees". In: *Methods in Ecology and Evolution* 7.5 (2016), pp. 589–597. DOI: https://doi.org/10.1111/2041-210X.12526.

[36] Eric Lewitus and Hélène Morlon. "Detecting Environment-Dependent Diversification From Phylogenies: A Simulation Study and Some Empirical Illustrations". In: *Systematic Biology* 67.4 (Dec. 2017), pp. 576–593. ISSN: 1063-5157. DOI: 10.1093/sysbio/syx095.

[37] R. B. O'Hara and M. J. Sillanpää. "A review of Bayesian variable selection methods: what, how and which". In: *Bayesian Analysis* 4.1 (2009), pp. 85–117. DOI: 10.1214/09-BA403.

[38] Andrew Gelman et al. *Bayesian Data Analysis*. 3rd. Chapman and Hall/CRC, 2013. DOI: https://doi.org/10.1201/b16018.

[39] Siddhartha Chib. "Chapter 57 - Markov Chain Monte Carlo Methods: Computation and Inference". In: ed. by James J. Heckman and Edward Leamer. Vol. 5. Handbook of Econometrics. Elsevier, 2001, pp. 3569–3649. DOI: 10.1016/S1573-4412(01)05010-3.

[40] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[41] Wilfred K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (Apr. 1970), pp. 97–109. ISSN: 0006-3444. DOI: 10.1093/biomet/57.1.97.

[42] Andrew Gelman and Donald B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[43] Stephen P. Brooks and Andrew Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (1998), pp. 434–455. DOI: 10.1080/10618600.1998.10474787.

[44] Ziheng Yang and Carlos E. Rodríguez. "Searching for efficient Markov chain Monte Carlo proposal kernels". In: *Proceedings of the National Academy of Sciences* 110.48 (2013), pp. 19307–19312. DOI: 10.1073/pnas.1311790110.

[45] Aki Vehtari et al. "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)". In: *Bayesian Analysis* 16.2 (2021), pp. 667–718. DOI: 10.1214/20-BA1221.

[46] Thomas Westerhold et al. "An astronomically dated record of Earth's climate and its predictability over the last 66 million years". In: *Science* 369.6509 (2020), pp. 1383–1387. DOI: 10.1126/science.aba6853.

[47] Arnab Kumar Maity, Sanjib Basu, and Santu Ghosh. "Bayesian criterion-based variable selection". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 70.4 (2021), pp. 835–857. DOI: https://doi.org/10.1111/rssc.12488.

[48] Annette Spooner et al. "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction". In: *Scientific Reports* 10.20410 (2020). DOI: 10.1038/s41598-020-77220-w.

[49] Lynn Kuo and Bani Mallick. "Variable Selection for Regression Models." In: *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)* 60.1 (1998), pp. 65–81.

[50] T. J. Mitchell and J. J. Beauchamp. "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032. DOI: 10.1080/01621459.1988.10478694.

[51] Edward I. George and Robert E. McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889. DOI: 10.1080/01621459.1993.10476353.

[52] P. J. Brown, M. Vannucci, and T. Fearn. "Multivariate Bayesian Variable Selection and Prediction". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60.3 (1998).

[53] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[54] Michael E. Tipping. "Sparse Bayesian Learning and the Relevance Vector Machine". In: *J. Mach. Learn. Res.* 1 (2001), pp. 211–244. ISSN: 1532-4435. DOI: 10.1162/15324430152748236.

[55] Trevor Park and George Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686. DOI: 10.1198/016214508000000337.

[56] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "Handling Sparsity via the Horseshoe". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Ed. by David van Dyk and Max Welling. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 2009, pp. 73–80. URL: https://proceedings.mlr.press/v5/carvalho09a.html.

[57] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), pp. 465–480. ISSN: 0006-3444. DOI: 10.1093/biomet/asq017.

[58] George Casella and Roger L. Berger. *Statistical inference*. Australia: Thomson Learning, 2002.

[59] Odile Maliet, Florian Hartig, and Hélène Morlon. "A model with many small shifts for estimating species-specific diversification rates". In: *Nature Ecology & Evolution* 3.1086-1092 (2019). DOI: 10.1038/s41559-019-0908-0.

[60] Mattia Tarabolo. *VarSel*. 2022. URL: https://github.com/Mattiatarabolo/VarSel.git.

[61] Daniele Silvestro et al. "Bayesian estimation of multiple clade competition from fossil data". In: *Evolutionary Ecology Research* 18.1 (2017), pp. 41–59.