Bayesian Inference in Phylogenetics using Nested Sampling

THE UNIVERSITY OF
AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
DEPARTMENT OF STATISTICS

Patricio Maturana Russel

A thesis submitted in fulfilment of the requirements for
the degree of *Doctor of Philosophy in Statistics*,
The University of Auckland, 2017.

Abstract

The Bayes factor is a common method for statistical model selection. The computation of such factor is based on the marginal likelihood, an integral that can be hard to estimate depending on the model complexity. The models employed in phylogenetic inference are of very high complexity. In this case, a direct computation of the Bayes factor is infeasible, and numerical methods or approximations are needed for its estimation. Model selection is an integral part of this field, but it tends to be obstructed by the requirements of the established marginal likelihood estimation methods, such as generalized steppingstone sampling. In this work, we introduce nested sampling to phylogenetics, a Bayesian algorithm which provides the means to estimate the marginal likelihood, and simultaneously sample from the posterior distribution. We study the behaviour of nested sampling for several statistical and phylogenetic scenarios and compare its performance to established estimation methods like steppingstone sampling. We introduce and discuss extensions to the initial algorithm, allowing for variable tree topology, estimating Bayes factor directly, and using importance sampling approaches to further improve its performance. Nested sampling has been shown to work in situations where most MCMC methods fail, e.g., if the true distribution is a mixture of quite distinct distributions. We show that the algorithm and its extensions offer a relatively cheap alternative to estimate, in a single run, the marginal likelihood together with its uncertainty, unlike established methods. It also permits us to sample from the posterior distribution at no extra cost. Overall, we establish nested sampling as a valuable alternative in Bayesian phylogenetics, in particular for model selection and parameter inference.

I would like to dedicate this thesis to my whole family

Acknowledgements

I would like to acknowledge all those who have been a part of this process, that started by me studying English in Chile and continuing in New Zealand. These years have been full of moments, and those moments full of people who have left their mark on me.

I would like to acknowledge especially my parents, partner, brothers and sister. Thanks to my PhD supervisors, Steffen and Brendon, who gave me all the necessary support and motivation in this journey. A special thanks to Mugdha who was always there as a proof-reader.

This work was funded by the CONICYT BECASCHILE/DOCTORADO 72110726.

Contents

List of Figures

List of Tables

Nomenclature

Models of evolution

JC69	Jukes and Cantor 1969
K80	Kimura 1980
F81	Felsenstein 1981
HKY85	Hasegawa and Kishino 1985
GTR	General time reversible

Marginal likelihood estimation methods

HM	Harmonic mean
PS	Path sampling
SS	Steppingstone sampling
GSS	Generalized steppingstone sampling
AIS	Annealed importance sampling
NS	Nested sampling
NIS	Nested importance sampling

Symbols

z	marginal likelihood
BF	Bayes factor

Chapter 1

Introduction

Using as a cornerstone the basic premise that the diversity of species can be explained by descent with modification [Darwin, 1859], phylogenetics is the study of evolutionary relationships among groups of organisms, based typically on molecular sequencing data. Among its multiple practical applications, we can highlight the identification of mutations likely to be associated with disease [Fleming et al., 2003], the reconstruction of ancient proteins [Chang et al., 2002] or the generation of legal evidence in a trial [Metzker et al., 2002].

Advancements in sequencing technologies have led to the accumulation of large datasets, allowing for the usage of increasingly complex statistical models to investigate patterns of evolution. The concurrent advancements in computer hardware and software provide the means to actually analyse these complex models. Reconstruction methods for phylogenetic trees come in various guises, from distance-based, over implicitly parametric (maximum parsimony) to explicitly parametric or model-based methods (maximum likelihood, Bayesian methods).

Distance-based methods first transform the sequence data into a distance matrix, representing the evolutionary distances between pairs of species. Such methods include least-squares [Cavalli-Sforza and Edwards, 1967], minimum evolution [Desper and Gascuel, 2002; Rzhetsky and Nei, 1992], and neighbour-joining [Saitou and Nei, 1987]. These methods have the advantage of being relatively fast and performing well when the divergence between the sequences is low. However, it has been shown that the transformation to distances loses information [Steel et al., 1988], and that this loss can lead to systematic error in the estimation [e.g., Gascuel and Steel, 2006].

Maximum parsimony [MP; Fitch, 1971b] minimizes the number of mutations needed to explain a site pattern on a given tree, and selects the tree with the overall smallest number of mutations. Its principal advantage is the speed for the analyses of hundreds of sequences. However, it possesses some well-known drawbacks. For instance, it does not take into account unseen events (e.g., if the ancestral nucleotide state of `G` was an `A`, MP assumes that the mutation occurred directly from `A` → `G`, rather than a transitory mutation, e.g., `T`, in between to mutate to `G`, such as `A` → `T` → `G`). So, it underestimates the actual number of mutations. The method makes implicit model assumptions as most non-parametric approaches do. This constitutes a weakness because it makes it difficult to incorporate a formal structure to explain the data. Further, MP is sensitive to long branch attraction which can lead to systematically inferring the wrong topology (branching pattern of the tree) when highly divergent species are included in the dataset [Felsenstein, 1978].

Maximum likelihood (ML) was introduced to phylogenetics by Felsenstein [1981]. Since then, it has become the most popular method of parametric inference thanks to efficient implementations like phyML [Guindon and Gascuel, 2002], RAxML [Stamatakis et al., 2005], or GARLI [Zwickl, 2006]. This method selects a "best" model from a predetermined set of models, making it possible to model, understand and assess evolutionary processes in more detail. Under this framework, model assumptions are explicit and consequently they can be evaluated and improved. This is a very useful property which MP lacks. The selection of the best model is assessed by maximizing the likelihood, i.e., the probability of observing the data given the model parameters. Thus, the model is selected according to how well it predicts the observed data. The point in the parameter space which maximizes the likelihood function is called the maximum likelihood estimate (MLE). If we are given the tree topology, then finding the MLE reduces to a continuous optimization problem for branch lengths and substitution parameters. This estimate is asymptotically unbiased, consistent and efficient. All of them are desirable statistical properties. However, if we include the topology into the optimization process, these asymptotic properties of the MLE do not apply anymore because, statistically, the tree is a model rather than a parameter [Yang, 1996c, 2006].

In practice, the tree topology is commonly found by using less computationally demanding methods, such as parsimony [Fitch, 1971b] or neighbor joining [Saitou and Nei, 1987], and then —implicitly assuming that this tree is the maximum likelihood tree for every candidate model of evolution—the MLE is calculated for each proposed model for this tree, including branch lengths. Thus, the models can be compared in order to select the best among the competing ones. Posada and Crandall [2001] argued, based on their simulations, that the initial phylogeny used to select the model of evolution does not have a major impact on model selection unless it is a randomly chosen tree. Once the model is selected, a search of the tree space is performed to find the MLE of the tree which includes topology, branch lengths and parameters related to the model of evolution. Finding this estimate involves a high computational effort and unfortunately the existing optimization algorithms are not guaranteed to solve the problem. However, the method works quite well for simple evolutionary models [Rogers and Swofford, 1998].

One maximum likelihood run is computationally intensive and it is only executed to identify a single point in the parameter space. One run does not provide any information about the uncertainty of the estimate. To account for it, Felsenstein [1985] proposed the use of bootstrapping techniques. The non-parametric bootstrap generates samples by drawing with replacement from the data until the sample has the same dimension as the original data. Then, the MLE is inferred from the bootstrap samples, which permits us to assess the uncertainty and robustness of our MLE. This method is widely used by the phylogenetic community, however, it has several limitations of which the users are not always aware. For instance, there is no consensus about the interpretation of bootstrap proportions [Yang, 2006; Yang and Rannala, 2005]; it cannot address systematic errors stemming from an inappropriate model choice [Holland, 2013]; and a more evident limitation is that it may require a huge computational cost.

Model choice under a ML framework is restricted to stationary, homogeneous, and reversible Markov processes to be able to execute the optimization. A violation of these assumptions further reduces the chance of finding the MLE. To summarize, ML is an often-used way of inferring a (potentially) best model for a given sequence alignment. However, its model choices are limited, and the computational cost of doing an inference are massive.

The increase in computational power and the development of Markov chain Monte Carlo (MCMC) approaches have led to the rise of Bayesian methods. The concept was introduced to phylogenetics in the 1990s [Larget and Simon, 1999; Rannala and Yang, 1996; Yang and Rannala, 1997], and has gained popularity because of its flexibility when dealing with complex models and large datasets, contrary to ML [Lartillot and Philippe, 2004; Nylander et al., 2004]. Its popularity was further increased by state-of-the-art implementations of the models in programs like MrBayes [Huelsenbeck and Ronquist, 2001], BEAST [Bouckaert et al., 2014; Drummond and Rambaut, 2007], or PhyloBayes [Lartillot et al., 2009]. The ultimate target of a Bayesian inference is the posterior distribution, which assesses the probability of a model given the data. This distribution is derived as the normalized product of the likelihood of the data given the model and the prior distribution for the model. The prior allows for the incorporation of information, available before collecting the data, into the inference, a feature which can greatly improve model fit. However, the dependency of the analysis on this distribution is also one of the major criticisms of Bayesian methods, since the use of prior information might be considered to imply subjective choices and lead consequently to biased inferences. To face these criticisms, there have been some attempts to permit the inclusion of objective or non-informative priors, such as Jeffreys [Jeffreys, 1946] or Reference priors [Berger and Bernardo, 1989; Bernardo, 1979]. These are rarely found in phylogenetics, because they require tractable likelihood functions, which are uncommon in the field. Nevertheless, there have been empirical attempts which have allowed to have reasonable non-informative priors in the field. Overall, the impact of priors on the inference has not been fully understood in phylogenetics and poses an interesting challenge.

Bayesian analysis often dispenses trying to find an explicit form of the posterior distribution. Instead, the posterior is approximated by MCMC methods. This procedure raises the query about the quality of the approximation. This issue is addressed in terms of the convergence of the chain towards the posterior distribution. In order for the MCMC to provide a reasonable approximation it might need to run for a large amount of time, which can lead to a high computational cost. However, this cost is offset by a wide variety of insights one can gain from the posterior distribution. For one, the resulting distribution provides a description of the parameter space and its uncertainty instead of a simple snapshot as provided by ML. The output is natural and easy to interpret, contrary to bootstrap proportions. And because no optimization is necessary, more complex models and datasets can be assessed [Holder and Lewis, 2003; Huelsenbeck et al., 2001; Yang and Rannala, 2012]. Finally, for comparable problems the richer output of a Bayesian method can be obtained in approximately the same time as the output of an ML method with bootstrapping estimation.

Similar to other fields, model selection plays an integral part in phylogenetic inference. A wide variety of different criteria are available for this task, such as the Likelihood Ratio Test (LRT), the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and Bayes factor (BF). Among the available methods, the hierarchical likelihood ratio tests (hLRTs) are the most popular. This method usually consists of a specific sequence of pairwise likelihood ratio tests performed until a final model cannot be rejected. A number of computational packages allow the pairwise comparison using this test, such as PAUP [Swofford, 2003], PAML [Yang, 2007] and the R package APE [Paradis et al., 2004]. Despite its popularity, Sanderson and Kim [2000] and Posada and Buckley [2004] pointed out that there are several reasons to explore new methods as alternatives.

For instance, a single optimal model might not exist, the probability of rejecting the null hypothesis when it is true (type I error) increases with the number of tests performed, the outcome of this method could be affected by the starting model, the best model could not be selected, it relies on large sample asymptotics, it does not allow the incorporation of phylogenetic uncertainty and the comparison of different topologies. Also, the LRT is limited to the comparison of nested models, cases in which pairs of models can be classified as submodel (its parameters are all also in the other model) or supermodel (it contains all the parameters of its submodel and a few more), and tends to favor parameter-rich models. This might have a negative effect on the analysis of real datasets, in particular when suitable models are not nested.

AIC [Akaike, 1974] is an asymptotically unbiased estimator of the expected relative Kullback-Leibler information quantity [Kullback and Leibler, 1951]. It depicts the amount of information lost when the proposal model is used to approximate the true model. Hence, the model with the smallest AIC is selected. On the other hand, BIC [Schwarz, 1978] is a rough approximation to the natural log-marginal likelihood (normalizing constant in the posterior distribution). Thus, the difference between two BIC values can be considered as a rough approximation of the Bayes factor. Unlike AIC, known for preferring complex models, BIC penalizes parameter rich models more severely [Kass and Raftery, 1995].

Posada and Buckley [2004] argued that BIC (and Bayesian methods) and AIC approaches have significant advantages compared with the hLRTs for model selection. These advantages include the use of these methods in nested or nonnested models, the inclusion of uncertainty in model selection and the allowance for model averaged-inference. Also, they are straightforward to calculate from the maximum likelihood estimate. Despite of their advantages over hLRTs, they also have some weak points. For instance, they are based on a point estimate and are not able to take into account uncertainty in topology [Bollback, 2002]. As with the LRT, they also rely on large sample asymptotics. Furthermore, BIC does not take into account the prior distribution and the same applies to AIC and LRT. Xie et al. [2011] described two primary reasons to consider the prior distribution in Bayesian model selection: first, the prior could prevent the model from fitting the data well, and second, it penalizes the inclusion of a new parameter when the marginal likelihood is used to assess model performance. An alternative which takes into account these considerations is the Bayes factor.

The Bayes factor [BF; Kass and Raftery, 1995] is the Bayesian analog of the likelihood ratio test under the classical paradigm. Unlike the model selection methods discussed above, which are based on a point estimate, BF accounts for the uncertainty in the parameters of the model. This method has the flexibility of not requiring alternative models to be nested, unlike LRT, and it embodies the desirable property of following Occam's razor. In other words, this model selection criterion prefers simpler models over complex models if they fit the data similarly well, or in a phylogenetic context, the tree that requires fewer mutations to explain the observed data. This criterion was used implicitly by Darwin, in the Origin of species, to construct evidence supporting his theory[1].

[1] Darwin studied pigeons under domestication to support his theory of common descent.

Another interesting feature is that Bayes factor can be used without conditioning on a specific topology [Holder et al., 2014; Huelsenbeck et al., 2004; Suchard et al., 2001; Wu et al., 2014]. It is also useful for guiding evolutionary model-building processes [Kass and Raftery, 1995], because it can deal with non-nested hypotheses, a fitting feature for evolutionary models. Furthermore, the Bayes factor assesses the evidence provided by the data in favor of a hypothesis, a natural scientific question which can be hard to answer under the frequentist paradigm.

The Bayes factor is based on the *marginal likelihood*, also called *evidence*, being obtained by integrating the product of the prior and the likelihood over the parameter space. In the case that the phylogeny is unknown, the sum over the tree space is included. In general, even for simple phylogenetic models, this quantity does not have an analytic solution, and has to be estimated. Different numerical methods have been proposed to estimate the Bayes factor directly. Suchard et al. [2001] used the Savage-Dickey ratio to approximate it for nested models. This method can only be applied to a family of models, a limitation shared by hLRTs. Huelsenbeck et al. [2004] used reversible jump Markov chain Monte Carlo, including all possible time-reversible models. Nevertheless, the comparison is restricted to the particular group of models included in the system. The incorporation of a new model would require a change on the design of the MCMC proposal which allows jumps between the models. Both methods of BF estimation allow the incorporation of the uncertainty about the topology.

Methods which aim to estimate the marginal likelihood have also been developed. These represent a more general alternative because the marginal likelihood of a model calculated today will be available for the comparison of other models in future studies. The most common of these methods is an importance-sampling approach known as the *harmonic mean* (HM) method [Newton and Raftery, 1994]. Even though its calculation is straightforward, this estimator has several

He believed that the great diversity of breeds had the rock-pigeon as common ancestor. This was a common opinion among naturalists at that time. According to his reasoning, if it were not true, they should have descended from at least 7 different aboriginal stocks. Also, they should have been selected and successfully domesticated by half civilized-man. Moreover, these species should have all become extinct or unknown. Using his own words: “so many strange contingencies seem to me improbable in the highest degree.”

widely known drawbacks. For instance, it is biased, overestimates the true value, in many situations its variance is not finite, and it leads to a strong overestimation of the marginal likelihood as the model increases its dimension, favoring parameter rich-models [Baele et al., 2016; Lartillot and Philippe, 2006; Xie et al., 2011]. An extension of the HM is the *inflated density ratio* [IDR; Arima and Tardella, 2014; Petris and Tardella, 2007]. This method relies on an auxiliary distribution, which is a perturbed version of the target distribution such that its total mass has a known functional relation to the marginal likelihood being estimated. Unlike HM, IDR avoids the infinite variance problem under fairly general conditions, but it keeps its simplicity.

Far more accurate than HM is *path sampling* (PS), or also known in physics as *thermodynamic integration* (TI), introduced into phylogenetics by Lartillot and Philippe [2006]. This method requires several Markov chains from a specific sequence of transition functions (probability distributions), which define a path between the prior and posterior distribution, in order to estimate the marginal likelihood. This leads to a much higher computational cost in comparison to HM. However, PS allows the direct BF estimation.

Another importance sampling approach is *steppingstone sampling* (SS), proposed by Xie et al. [2011]. This method performs similarly to TI, but requires slightly less computational effort. SS makes use of importance sampling to estimate a series of ratios of normalizing constants, taken from a sequence of transitional functions, like the ones defined for TI. It works in a very similar way to *annealed importance sampling* [AIS; Neal, 2001]. SS also allows for the direct estimation of the Bayes factor [Baele et al., 2013]. This method has gained popularity in phylogenetics due to its availability in many software packages (see Table 1.1). The performance of the method relies mainly on the path between the prior and the posterior. If these distributions are quite different, the method might require more samples to reduce its estimation error. Its extension, known as *generalized steppingstone sampling* [GSS; Fan et al., 2011], works by using a reference distribution to shorten the path, which leads to a more accurate estimate of the marginal likelihood. GSS has also been extended to account for uncertainty in topology [Baele et al., 2016; Holder et al., 2014].

GSS is a highly accurate method, widely used in phylogenetics. However, some

Software	Marginal likelihood estimation method
Bayou(R-package)	GSS
BEAST	HM, PS, SS, and GSS
MrBayes	HM and SS
PhyloBayes	PS under normal approximation
Phycas	HM, PS, and GSS
TESS(R-package)	PS and SS

Table 1.1: Marginal likelihood estimation methods currently implemented in Bayesian phylogenetic software packages.

considerations, vaguely discussed in the literature, should be taken into account when using it. First, the method requires a minimum number of specifications to yield reliable estimates, which have to be found by guesswork [Drummond and Bouckaert, 2015, Chapter 9]. This point is shared by PS, SS and AIS. On the other hand, its performance depends mainly on the reference distribution, which should optimally be an approximation of the posterior. In the case in which the posterior is a difficult distribution to sample from, the method could fail. This situation could occur, for instance, when the likelihood, a function that is not completely understood in phylogenetics, contains phase transitions (see Example 3.5.1 for more information). Therefore, in the meanwhile, methods of general applicability and more friendly for the users are needed.

A more general algorithm is *nested sampling* [NS; Skilling, 2006]. This method addresses many of the issues of the previously introduced methods while retaining their abilities. It does not depend on many specifications; it provides a measure of the uncertainty of the estimate in a single run; it can deal with phase transitions in the likelihood; it yields posterior samples. NS was developed in physics, but it has spread across different disciplines due to its properties. It has been used mainly for marginal likelihood estimation, but its applications also include parameter inference [e.g., Aitken and Akman, 2013; Pullen and Morris, 2014].

NS works by exploring the parameter space according to the prior when starting. Then, it progressively moves towards the areas of high likelihood. The sampled points can be reused as a posterior sample, if appropriate weights are assigned. The particular way of getting the points for the estimation might make

NS require many iterations to reach the areas which actually have a greater contribution to the marginal likelihood estimation. The trajectory explored by NS can be shortened by an importance sampling distribution. This approach is called *nested importance sampling* [NIS; Chopin and Robert, 2010; Skilling, 2006], requires less iterations and has a lower uncertainty in comparison to NS.

This thesis concerns principally the discussion of a variety of methods to estimate the Bayes factors in a phylogenetic context. The discussion is from a statistical point of view. Thus, first of all, we start giving a complete definition of phylogenetic models and their components in Chapter 2. This definition also includes a review of the prior distributions currently in use in Bayesian phylogenetic analysis. This chapter lays the groundwork of the phylogenetic concepts used throughout this thesis.

In Chapter 3, we describe and discuss a subset of methods to estimate the marginal likelihood and the Bayes factor. This includes the aforementioned HM, PS, SS, AIS, GSS, NS, and NIS. The NS method, originally developed for marginal likelihood estimation, is extended to allow the direct estimation of the Bayes factor. The methods are tested in statistical models, which represent challenging scenarios. They are also tested in a phylogenetic context assuming a fixed topology, and in the case of direct BF estimation.

Chapter 4 concerns the extension of NS to allow variable tree topology. The method is assessed in marginal likelihood estimation and also in parameter inference, especially for the tree topology. NS is also extended to allow the incorporation of an importance sampling distribution (NIS) for the tree topologies. This results in a decrease of the uncertainty related to the estimate and an acceleration in convergence.

Finally, in Chapter 5 we give a summary, conclusion, and discussion about the material presented in this thesis. We also discuss the potential future work that emerged from this thesis.

Chapter 2

Phylogenetic models

2.1 Introduction

A phylogenetic model is composed of two components: a tree and a model of evolution. The tree model describes the genealogical relationship among a group of sequences or organisms, whereas the substitution model describes the evolutionary process in sequence data along the branches of the tree. The latter is usually referred to as the model. Both play an important role in phylogenetic inference and the preference of one over another will depend on the goals of the study. For example, in the study of the evolution of biological molecules, the focus is on substitution patterns and therefore on the model of evolution.

Both evolutionary and tree models are characterized by a mathematical structure composed by parameters, which are the target of any statistical analysis. When the genealogical relationship among the taxa is unknown, the tree is considered as another parameter of the model. Under a Bayesian approach, the parameters are treated as random variables and thus are inferred as probability distributions, known as the posterior. This allows their estimation together with the corresponding uncertainty. However, this approach requires the inclusion of additional information about the parameters via prior distributions, before collecting the data.

The next two sections of this chapter aim to provide the basis of all the phylogenetic concepts required in the remaining chapters of this thesis. We describe the components of the phylogenetic model, the phylogeny and the evolutionary model, and their related concepts, followed by a discussion about the prior distributions used currently in Bayesian phylogenetic analysis.

2.2 Tree model

The tree model represents the evolutionary histories of a group of biological species or organisms. It provides a concise way to visualize and summarize how the taxa have evolved from a common ancestor. This concept was sketched by Darwin (Figure 2.1) in 1837 and subsequently developed formally and backed with evidence in *On the Origin of Species by Means of Natural Selection*. The generalization of this concept, embodying all living things together, even extinct, forms a vast evolutionary phylogeny which is known as the *Tree of Life*.

As any statistical model, the tree model is a simplification of reality and it is consequently subject to error. However, it seems a reasonable assumption since it is backed by many kinds of data that support an evolution that seems to be tree-like [Archie, 1989; Holland et al., 2004; Maturana R., 2017; Penny et al., 1982]. For instance, Archie [1989] analysed 28 datasets, based on quantitative morphological characters published at that time, by using randomization tests, and found that all of them contained phylogenetic information; Penny et al. [1982] analysed 5 sequence data from 11 species by using maximum parsimony, resulting in very similar trees from the individual and combined sequences, i.e., they contained similar evolutionary information supporting the existence of a phylogenetic tree. Even though in certain cases a network might be a better option [see, for instance, Hernández-López et al., 2013], the tree model is a practical and useful model.

2.2.1 Definition

A phylogenetic tree is a branching diagram that depicts the evolutionary relationship among a set of taxa Λ. Formally, a tree is a pair $\tau = (\mathcal{N}, \mathcal{V})$, where $\mathcal{N}$ is a set of *nodes*, and $\mathcal{V}$ is a set of *branches*. In the mathematical literature, they are known as *vertices* and *edges*, respectively. Each branch $t \in \mathcal{V}$ is a pair of nodes (v_i, v_{i+1}), with v_i and $v_{i+1} \in \mathcal{N}$. Every pair of nodes in $\mathcal{N}$ is connected

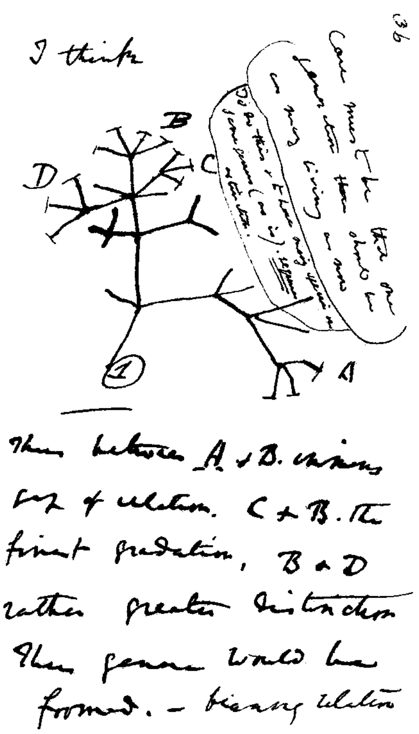

36
I think
Case must be that one generation then should be as many living as now
To do this & to have many species in same genus (as is) requires extinction.
B
C
D
1
A
Thus between A & B. immense gap of relation. C & B. the finest gradation, B & D rather greater distinction
Thus genera would be formed. — bearing relation

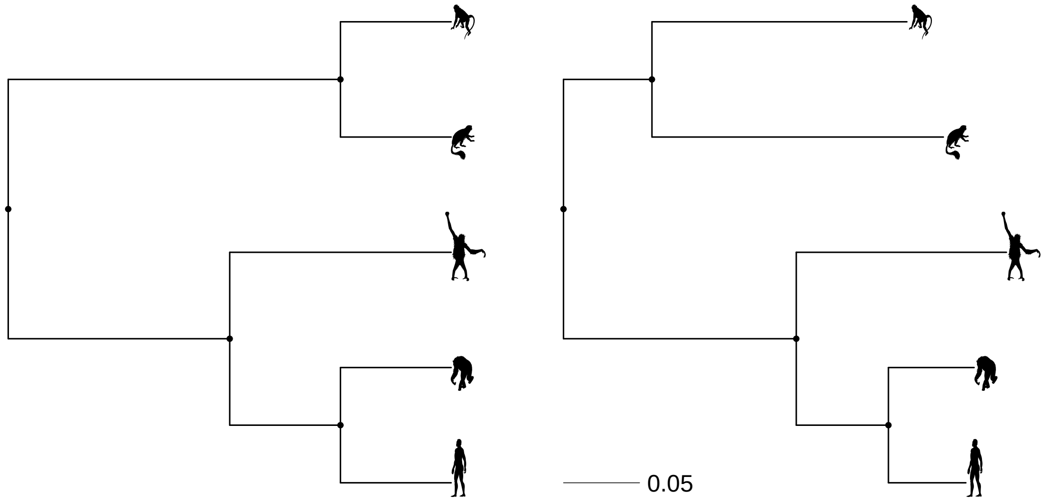
Figure 2.1: Charles Darwin's 1837 sketched visualizing the idea of descent from common ancestry. The upper text reads, "I think. Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is) requires extintion." The lower text reads, "Thus between A & B immense gap of relation. C & B the finest gradiation, B & D rather greater distinction."

by a unique path in τ, which does not admit cycles, in other words, every path $(v_1, v_2, \ldots, v_n)$, which is an ordered set of distinct nodes, fulfils the condition $v_1 \neq v_n$. The tree is a particular case of what, in mathematics, is defined as a *graph*.

2.2.2 Concepts

2.2.2.1 Topology and nodes

The order of the nodes, or the branching pattern, is called the *topology* of the tree and it is used to describe the shape of the tree regardless of the branch



(a) Cladogram

(b) Phylogram

Figure 2.2: The evolutionary relationship among 5 members of the primate family. From the top: macaque, guereza, orangutan, chimpanzee and human. These species are related via common ancestry which is explained by the tree. The branch lengths of the phylogram stand for the expected number of mutations per site.

lengths. The topology stands for the evolutionary relatedness of the sequences, organisms or species. A tree with only information about its topology is called a *cladogram*, whereas with the additional information about its branch lengths is called a *phylogram*. An example of these two trees is displayed in Figure 2.2. Both phylogenies show the evolutionary relationship among 5 members of the primate family. However, phylogeny 2.2a only contains this information, i.e., it is a cladogram, whereas the one in 2.2b includes information about the expected number of substitutions per site through the branch lengths, i.e., it is a phylogram. In the case that the branch lengths represent time, the tree is called an *ultrametric tree* or *chronogram*. In such a tree, the distance from the root (the ancestor of all the sequences) to the leaves are the same.

A node is classified according to its degree, which is defined as the number of branches connected to it. In this way, two kinds of nodes are distinguished: *external* and *internal*. An *external node* has degree equal to 1 and is called a *leaf*. It represents a taxon, typically an extant species (or sequences). On the other hand, an *internal node* has degree greater than 1 and represents a common ancestor for which generally there is no information available. In the case of having information (DNA or similar), it can also be considered as a taxon [Semple and Steel, 2003].

If an internal node has degree 3, excluding the root node, it is said to be resolved. When this is the case for all of them in a given tree and in addition the root node has degree 2, the tree is said to be fully resolved. This characteristic defines a special class of trees called *binary* or *bifurcating* trees. In this case, all speciation events produce just two descendants from one ancestor. This is the most common kind of topologies used in phylogenetics, as it is unlikely that an ancestral lineage diverges into more than 2 descendants simultaneously, even though it is still possible [Baum and Smith, 2012]. In this particular case in which a node has degree larger than three, i.e., more than two descendants, the node is called *polytomy* or *multifurcation*. A polytomy representing truly simultaneous divergence is called *hard polytomy*. However, polytomies most of the time represent uncertainty about the correct branching pattern in a tree, in which case they are called *soft polytomies*.

2.2.2.2 Rooted and unrooted trees

The common ancestor of all taxa in a tree is called the *root* and it is the most recent common ancestor (MRCA) of the phylogeny. If this is specified, the tree is called a *rooted tree*. In this case, time is represented by a single direction. For instance, the root in the trees displayed in Figure 2.2, which is the leftmost node, is the common ancestor of the 5 primates and establishes the direction of evolution. In a rooted binary tree, the root has degree 2, i.e., it produces two direct descendants.

On the other hand, if the root is not specified or unknown, the tree is called an *unrooted tree*. In this case, the tree is drawn without reference to the direction of time, i.e., ancestry cannot be identified among the internal nodes of the tree.

2.2.2.3 Clades and splits

Information about the direction of the evolutionary process in a phylogeny, granted by the root, provides the means to classify taxa. This information is of crucial importance in taxonomy, the research field that deals with description, identification and classification of living things. Actually, it is widely accepted that phylogeny should be the basis of taxonomic classifications [Yang, 2014]. The evolutionary direction allows us to identify the ancestors in the phylogeny and consequently define groups according to them. These groups form part of phylogenetic and taxonomic nomenclature. This is the case of a *monophyletic group* or better known in phylogenetics as a *clade*, which is the group composed of all the descendants from a common ancestor. The taxa which comprise this group are *closely related*. They have more recent common ancestors than any other outsider taxa. This concept is based on the time of divergence of the taxa and not on the sequence distance (sum of the branch lengths which connect the two taxa). For instance, two taxa which belong to the same clade will not necessarily have the smallest sequence distance. Two clades are said to be *sister clades* if they descend directly from the same ancestor. Analogously, if two species share a direct common ancestor, they are said to be *sister species*.

The use of these terms necessarily requires a rooted phylogeny. However, for the case of unrooted trees, groups can be still generated by cutting an internal branch which defines a *split* or *bipartition*. Formally, for a set of taxa Λ, a split $A|B$ is a partition of Λ, i.e., $A \cup B = \Lambda$ and $A \cap B = \emptyset$. The split divides the set of taxa into two mutually exclusive groups or *clans* where at least one of them must be a clade. In the case that the split is assertively made on the internal branch which should truly contain the root of the unrooted tree, the two partitions are clades.

2.2.2.4 Rooting techniques

Most of the tree reconstruction approaches do their searches over the unrooted tree space, producing consequently unrooted trees. This is only based on computational/mathematical convenience, which is granted by assuming reversibility on the model of evolution, an assumption examined below in Section 2.3.1.4. As was discussed above, the root provides information which is essential to answer questions related to the evolutionary process. However, in practice, the root can be identified and allocated in an unrooted tree in different ways. The most common method is known as *outgroup rooting*. The method consists of using an external group, which must be known to be outside of the group which is being analysed, the ingroup. Thus, the root is allocated between the outgroup and the ingroup.

Another method, known as *molecular clock rooting*, relies on the assumption that there is a constant rate of evolution over time and among evolutionary lineages in order to determine the root of the tree. This assumption, known as molecular clock [Zuckerkandl and Pauling, 1965], is valid when the taxa are closely related or in the presence of low divergent sequences. Another alternative to root a tree, referred as *mid-point rooting*, is to estimate an unrooted tree and then calculate the distance between each pair of sequences. This distance is calculated by summing the branch lengths which connect the sequences. Then, the root is allocated to the middle point of the path that connects the two sequences which have the biggest distance. This method also relies on the molecular clock assumption since it assumes that the two most divergent sequences have evolved at equal rates, but it is a weaker assumption than the previous method. Finally, a *Bayesian analysis* can be carried out to infer the root of a tree [Huelsenbeck et al., 2002]. This procedure works on a fixed unrooted tree where the root is allocated along its branches according to the posterior distribution of the root. The analysis can be performed by using either the outgroup method, or the molecular clock assumption. This alternative has the attribute of assessing the uncertainty associated with the potential roots.

Even though unrooted trees do not contain as much information as rooted trees, they are crucial in phylogenetic analysis. Most of the reconstruction approaches, such as parsimony, maximum likelihood, and Bayesian methods, work over the unrooted tree space. Therefore, unrooted trees play an essential role behind the scenes in the computation of rooted trees. They can also address questions about the relationships among the taxa or represent clusters of related sequences. They are also useful tools when we try to establish diversity among the taxa.

2.2.2.5 Consensus trees and networks

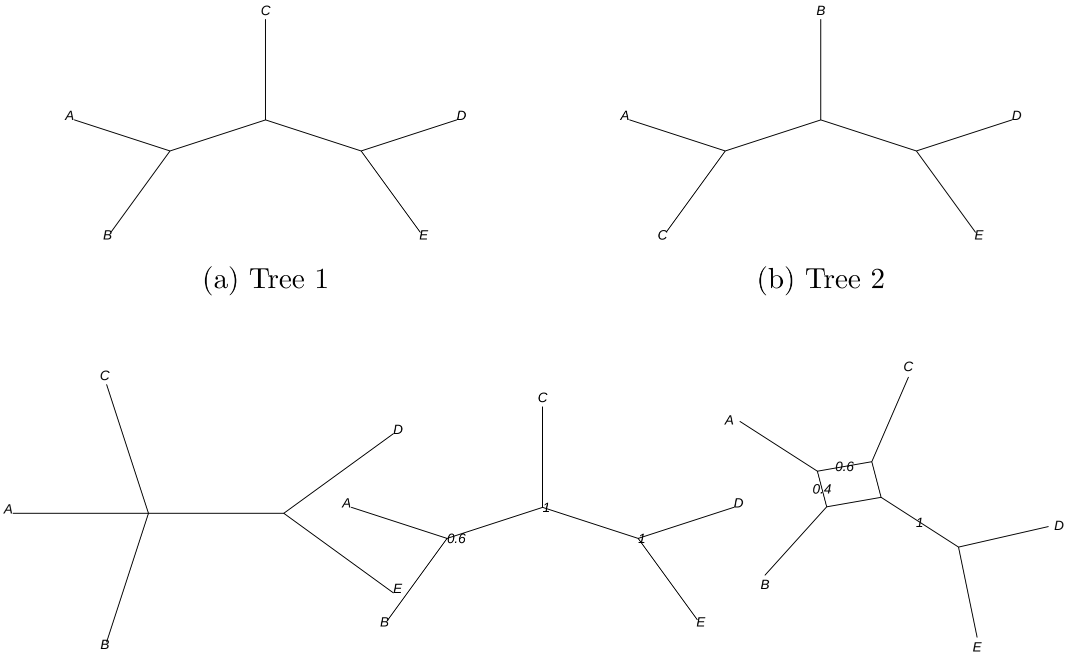
In the scenario of having a collection of phylogenetic trees, a natural situation in Bayesian analysis (which is usually based on a tree sample from the posterior distribution), or in bootstrap analysis, it is essential to have a mechanism which allows to summarize the topological information contained in it. One alternative is through a single representative tree, known as a *consensus tree*. This tree summarizes the common characteristics among the collection of trees.

Consensus trees can be defined in many ways (see Bryant [2003] for a complete review). One of the most popular is the *majority-rule consensus tree*. This tree contains only those clusters or splits which appear in at least $\vartheta\%$ of the times in the set of trees. This percentage varies usually between 50% and 100%. For a conservative $\vartheta = 100\%$, this approach is known as the *strict consensus tree*. Those splits which do not appear the minimum percentage of times are collapsed into a soft polytomy. This represents phylogenetic uncertainty about the relationship among those particular groups.

Another alternative is to summarize the collection of trees through a consensus network [Holland and Moulton, 2003; Holland et al., 2004]. This approach is a generalization of consensus trees. Instead of only depicting splits which appear at least half of the time in the set of trees, the consensus network allows us to visualize other competing splits with lower frequency according to certain threshold. Thus, it allows for the visualization of conflicting evolutionary hypothesis simultaneously. The conflicting splits are represented by high-dimensional hypercubes. In the particular case of two conflicting splits, they are represented by a box (see Figure 2.3e). Hence, the consensus network displays more information about the set of trees than the consensus trees.

To illustrate the methodology, consider a collection of 1,000 trees composed of two different phylogenies which are displayed in 2.3a and 2.3b. They appear with frequency 600 and 400, respectively. We identify two incompatible splits, i.e., they appear in only one of both trees, namely, $AB|CDE$ and $AC|BDE$. The set of trees is summarized by using the consensus trees and network discussed above.

Figure 2.3c displays the strict consensus tree. This contains the split $DE|ABC$,



(a) Tree 1

(b) Tree 2

(c) Strict

(d) Majority-rule

(e) Network

Figure 2.3: Tree 1 and Tree 2 compose a collection of trees. Their frequency is 600 and 400, respectively. This set of trees is summarized by the strict and middle-rule consensus tree and the consensus network.

since it appears in both trees, i.e., 100% of the time, but displays a polytomy for the subset (A, B, C) which is described in two different ways in the set of trees. Figure 2.3c shows the majority-rule consensus tree. This is identical to Tree 1 since it represents 60% of the set of trees, higher than the threshold of 50%. Figure 2.3e displays the consensus network. This shows clearly the two conflicting splits. The edge with the proportion 0.6 and its parallel edge stand for the split $AB|CDE$, whereas the edge with the proportion 0.4 and its parallel edge depicts the split $AC|BDE$.

Note that the consensus trees provide a summary of the collection of trees, but they lose information, as none of them provide information about the competing splits. On the other hand, the consensus network allows for the visualization of the two competing incompatible splits simultaneously.

2.3 Model of evolution

A model of evolution is a set of assumptions about the evolutionary process of a certain characteristic. It describes the different flows and probabilities of change from one state to another. Unlike non-parametric methods, such as parsimony, the model can incorporate explicit assumptions about the underlying evolutionary process, which can be tested and improved. This model is as important as the tree model. For instance, if the model of evolution assumed fits the data poorly, it could lead to incorrect phylogenies and inconsistency in the phylogenetic methods [Bruno and Halpern, 1999; Felsenstein, 1978; Gaut and Lewis, 1995; Hoff et al., 2016; Huelsenbeck and Hillis, 1993]. Hence, it is an essential part of phylogenetic inference.

Frequently in molecular phylogenetic analyses, a taxon is identified by DNA sequences which are composed by four basic molecules called *nucleotides*. These are adenine (A), cytosine (C), guanine (G), and thymine (T), which are linked together in a long chain. Thus, a DNA sequence is a character list with each state taking either A, C, G or T. According to their chemical structure, adenine and guanine are classified as *purines* (R), and cytosine and thymine as *pyrimidines* (Y). Each site in a sequence is assumed to be the outcome of the evolutionary process along the underlying tree. Thus, in order to reconstruct the tree the representative sequences are aligned to reflect the process, i.e., a position in one sequence is matched to another if the assumption can be made that they share a common ancestry. This assumption is known as homology. All alignment algorithms have been designed to produce homologous alignments. In this work, we assume that the sequences are perfectly aligned, that is, we do not consider the uncertainty in the alignment. Figure 2.4 shows an extract of 5 aligned mitochondrial DNA sequences from members of the primate family. Each position of the alignment, represented by the columns of the molecular data matrix, is called a *site*. In the example, 19 sites are displayed.

Statistically, substitution and mutation are synonymous. However, in biology a *mutation* is a random event whereas a *substitution* is the outcome of selective pressure. Mutations can be lost while substitutions are supposed to be fixed. There is a phenomenon where the number of substitutions appears more numerous

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Human	C	C	T	A	A	A	A	C	C	C	G	C	C	A	C	A	T	C	T
Chimpanzee	C	T	T	A	A	A	A	C	C	C	T	C	C	A	C	T	T	C	A
Orangutan	C	C	T	A	A	A	A	C	C	C	T	C	C	A	C	A	T	C	A
Guereza	C	T	C	A	A	A	A	C	C	C	G	C	A	A	C	C	T	C	C
Macaque	C	T	T	G	A	A	A	C	C	C	T	C	A	A	C	A	T	C	C

Figure 2.4: Extract of mitochondrial DNA for 5 species of primates. The rows of of this nucleotide matrix represent the species and each column a site.

in "young" sequence alignments, when it actually just shows a lot more mutations. Mutations in the purine group or in the pyrimidine group, i.e., `G`↔`A` or `T`↔`C`, are called *transitions*, whereas mutations between the groups, i.e., `G`↔`C`, `G`↔`T`, `T`↔`A` or `T`↔`G`, are called *transversions*. The sequences are subject to additional kinds of mutations, such as deletions and insertions. These are represented as gaps in an alignment. Most processes cannot distinguish between deletions nor insertions and most inference processes remove gappy sites and their neighbours to avoid bias in the inference due to alignment error.

Figure 2.5 shows an example of the evolution of 3 DNA sequences which contain 6 sites. The sites which have mutated are highlighted with a box around the corresponding nucleotide. The sequence on the top of the tree (`ACCTGG`) stands for the root from which the 3 sequences (named "Seq1", "Seq2" and "Seq3" in the Figure) have been generated. "Seq1" and "Seq2" share a common ancestor which has suffered a mutation in the second site (`C` → `T`). These sequences have kept this mutation but have also experienced their own in the sixth and third sites, respectively. "Seq3", which has descended directly from the root of the phylogeny, has also experienced substitutions in the first, third and fifth sites. Note that this evolutionary process could also contain *silent mutations*. For instance, a mutation could have occurred on the second position of sequence 3 of the type `C` → `T` → `C`, which is not visible.

The nucleotide substitution process in DNA sequences is generally described using Markov models. Throughout this section, we will discuss this kind of model, in particular, for DNA sequences. However, the extension to other character datasets, such as morphological, amino acids or any other molecular characteris-

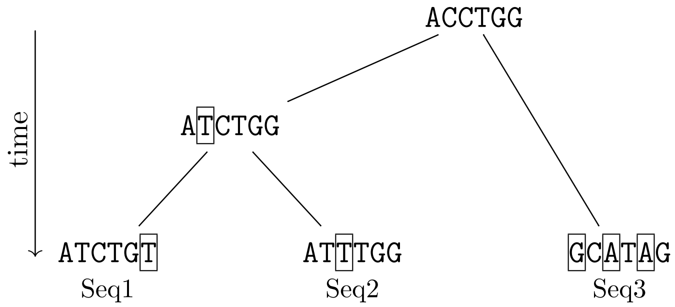


Figure 2.5: Evolution of 3 DNA sequences. The boxes around the nucleotides indicate where a mutation has happened.

tics, is analogous. Deletions and insertions are not accommodated by the models examined here.

2.3.1 Markov models of sequence evolution

Let $\widetilde{S}$ be a countable state space and let $\widetilde{s} = |\widetilde{S}|$ be its dimension. For instance, for DNA sequences $\widetilde{S} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, for RNA $\widetilde{S} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$, or for proteins $\widetilde{S}$ is composed of 20 amino acids. The number of mutations $N(t)$ at time t can be characterized by a Poisson process with mean μt ($\mu > 0$ and $t \geq 0$). The mutations occur with rate μ per unit of time t. Thus, the probability of k mutations at time t is given by

$$\mathbb{P}(N(t) = k) = \frac{(\mu t)^k}{k!} e^{-\mu t}, \quad k = 0, 1, \ldots$$

In a Poisson process, the time between mutation events follows an exponential distribution. If these times are ignored, a mutation event can be characterized by a discrete-time Markov chain $\{E_n\}_{n \in \mathbb{N}_0}$ with transition probabilities R_{xy}, which denote the probability of changing to state y given that the current state is x, with x and $y \in \widetilde{S}$. This value represents the probability of a single event. Given that redundant mutations are allowed, $R_{xx} > 0$. Thus, E_n denotes the state of the site of the sequence in the n^{th} substitution. For a DNA sequence, for instance, $E_3 = \mathtt{A}$ stands for the nucleotide at the third mutation.

Note that the Poisson process does not discriminate among the kind of mutations, but it considers all of them as equal. It only counts the number of mutations at time t. On the other hand, the E_n process stands for the current state after n mutations, but without considering the elapsed time t.

Assuming that E_n and $N(t)$ are independent, the process $\{X(t)\}_{t\geq 0}$, with

$$X(t) = E_{N(t)}, \tag{2.1}$$

defines a continuous-time Markov chain and generalizes the Poisson process. It is known as a uniform Markov jump process with subordinated chain $\{E_n\}$ and clock $N(t)$. This stochastic process embodies the current state through E_n, and the elapsed time through $N(t)$. The elements of its transition probability matrix $P_{xy}(t)$ are derived as follows:

$$\begin{aligned}
\mathbb{P}(X(t) = y|X(0) = x) &= \mathbb{P}(E_{N(t)} = y|E_0 = x) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(E_k = y, N(t) = k|E_0 = x) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(E_k = y|E_0 = x)\mathbb{P}(N(t) = k) \\
&= \sum_{k=0}^{\infty} R_{xy}^k \frac{(\mu t)^k}{k!} e^{-\mu t},
\end{aligned}$$

for $t \geq 0$, $x, y \in \widetilde{S}$, where R_{xy}^k is the corresponding element of $\mathbf{R}^k$ which denotes the k^{th} power of the matrix $\mathbf{R}$ and is known as the $k-$step transition matrix. Thus, the transition probability matrix is given by

$$\mathbf{P}(t) = \sum_{k=0}^{\infty} (\mathbf{R}^k) \frac{(\mu t)^k}{k!} e^{-\mu t}. \tag{2.2}$$

This probability matrix contains the probabilities of change at time t summed over all the possible number of mutation events k.

This matrix (2.2) can be rewritten as

$$\mathbf{P}(t) = e^{-\mu t} \sum_{k=0}^{\infty} \frac{(\mu t \mathbf{R})^k}{k!} = e^{\mu t(\mathbf{R} - \mathbf{I}_{\widetilde{s}})} = e^{t\mathbf{Q}}, \tag{2.3}$$

where $\mathbf{I}_{\widetilde{s}}$ is the $\widetilde{s} \times \widetilde{s}$ identity matrix, and $e^{t\mathbf{Q}}$ is the exponential matrix, with $\mathbf{Q} = \mu(\mathbf{R}-\mathbf{I}_{\widetilde{s}})$. The exponential matrix is considered notoriously difficult for some Markov jump models with many states. The $\mathbf{Q}$ matrix is called the *instantaneous rate matrix* and plays an important role in evolution models (it will be discussed in detail in 2.3.1.1).

The $\mathbf{P}(t)$ matrix satisfies the conditions:

$$\begin{aligned} \sum_{j=1}^{\widetilde{s}} P_{ij}(t) &= 1, \quad \text{with} \quad P_{ij} > 0 \quad \text{for} \quad t \geq 0, && (2.4) \\ \mathbf{P}(t+\delta) &= \mathbf{P}(t)\mathbf{P}(\delta), \quad \text{with} \quad t, \delta \geq 0, && (2.5) \\ \mathbf{P}(0) &= \mathbf{I}_{\widetilde{s}}. && (2.6) \end{aligned}$$

Condition (2.4) states that the probability of staying or leaving the current state i is 1 a time t later. For this, the rows of $\mathbf{P}(t)$ represent the current state and the columns the next state. Condition (2.5), known as the Chapman-Kolmogorov equations, states that the probability at any time can be split up into 2 time intervals and calculated taking into account all the intermediate states at these times. This is due to the transition probabilities only depending on the elapsed time. Condition (2.6) depicts the transition probabilities at time 0 which would be the case of no evolution. Naturally, the probability of staying at the same state is 1 and leaving is 0 when time has not elapsed.

The construction of the Markov model has been presented in a general way. Thus, the procedure can be used for any kind of data with categorical variables. In practice, the features of the transition rate matrix $\mathbf{Q}$ are determined empirically. For instance, for DNA data, it is well known that transitions occur with more frequency than transversions. Therefore, we could introduce parameters in $\mathbf{Q}$ which take into account this feature and thus model separately these kind of mutations.

2.3.1.1 Instantaneous rate matrix

The $\mathbf{Q}$ matrix defined in (2.3) plays a key role in models of sequence evolution characterizing the process to be modeled. Its parametric structure defines the transition probabilities on $\mathbf{P}(t)$. Depending on the $\mathbf{Q}$ matrix, the model adopts a distinctive name in phylogenetics as is discussed below.

The off-diagonal elements q_{ij} represent the instantaneous substitution rate from state i to state j, in other words, how often the substitution from i to j occurs with respect to other possible substitutions. Its diagonal elements q_{ii} depict the total flow that leaves the state i, for this reason they are negative. Thus, they are defined as

$$q_{ii} = -\sum_{i \neq j}^{\widetilde{s}} q_{ij},$$

with $q_{ij} > 0$.

For DNA sequences, the nucleotide substitutions are modeled by a 4-state Markov process, which instantaneous transition rate matrix in its most general form is defined by

$$\begin{array}{cc} & \text{To} \\ & \begin{array}{cccc} \texttt{A} & \texttt{C} & \texttt{G} & \texttt{T} \end{array} \\ \mathbf{Q} = \text{From} \begin{array}{c} \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \end{array} & \begin{pmatrix} - & \mu r_{AC}\pi_C & \mu r_{AG}\pi_G & \mu r_{AT}\pi_T \\ \mu r_{CA}\pi_A & - & \mu r_{CG}\pi_G & \mu r_{CT}\pi_T \\ \mu r_{GA}\pi_A & \mu r_{GC}\pi_C & - & \mu r_{GT}\pi_T \\ \mu r_{TA}\pi_A & \mu r_{TC}\pi_C & \mu r_{TG}\pi_G & - \end{pmatrix} \end{array}, \tag{2.7}$$

where the diagonal elements are chosen to make the rows to sum to zero; the parameters $r_{ij} \geq 0$ for $i, j = \texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}$, with $i \neq j$; $\mu \geq 0$; and the equilibrium distribution is $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$ under the constraint $\sum_i \pi_i = 1$ with $\pi_i > 0$ for $i = \texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}$. The rows of $\mathbf{Q}$ represent the current state of the nucleotide whereas the columns represent the potential nucleotide at time t.

The factor μ, which represents the expected number of substitutions per unit of time, is called the *mean of the substitution rates* and is defined as

$$\mu = -\sum_{i=1}^{\widetilde{s}} \pi_i q_{ii}. \tag{2.8}$$

The *relative rate parameters* r_{ij} describe the relative rate of substitution with respect to any other possible substitution. These parameters weight the rate of change μ constituting the *rate parameter* $(\mu \times r_{ij})$. This product is a kind of breakdown of the average number of mutations according to each potential mutation.

It is important to mention that t and μ cannot be separated in the probability matrix $\mathbf{P}(t) = e^{t\mathbf{Q}}$. Note that t is multiplying each element of $\mathbf{Q}$, which are composed in part by μ. As a result, they cannot be estimated separately but only through their product, number of substitutions up to time t, unless a perfect molecular clock is assumed in which case $\mathbf{Q}$ is the rate of substitution per unit of time. This prevents us from knowing the exact cause of the branch length. For instance, a long branch could be due either to a long period of evolutionary time t or a high mean substitution rate μ. A common practice to deal with this problem is to set to 1 the mean substitution rate of change μ. This is done by scaling the $\mathbf{Q}$ matrix by μ^{-1}. This transformation does not alter the relation of the substitution rates and as a result, the branch length t turns into the expected number of substitutions per site along the branch of the tree model.

2.3.1.2 Stationary distribution

The Markov process defined above assumes that when t goes to infinity, the probability that a site is in state y is non-zero, and that it is independent of its initial state. Such a Markov process is called *ergodic*. In mathematical terms, there is a $\widetilde{s}$-dimensional vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\widetilde{s}})$, with $\pi_i > 0$ for all i, and $\sum_{i \in \widetilde{S}} \pi_i = 1$, such that

$$\lim_{t \to \infty} P_{xy}(t) = \pi_y,$$

for all $x, y \in \widetilde{S}$. This row vector $\boldsymbol{\pi}$ is said to be the stationary distribution and satisfies

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t),$$

for all $t > 0$. This is equivalent to satisfying to all y

$$\pi_y = \sum_{x \in \widetilde{S}} \pi_x P_{xy}(t).$$

This property means that if a state is sampled from the stationary distribution and the process is executed for time t, then the distribution of the final state will be equal to the stationary distribution. Equivalently, it can be read as the probability of starting from any state and ending in y is π_y. This property can also be expressed in terms of the stationary rate matrix as $\mathbf{0} = \boldsymbol{\pi}\mathbf{Q}$, where $\mathbf{0}$ is an $\widetilde{s}$-dimensional vector which contains only zeros.

2.3.1.3 Assumptions

The Markov process defined in (2.1), which models the mutation events, involves the following assumptions:

- *Memoryless.* At any site, the next state only depends on the current state.
- *Homogeneity.* Substitution rates remain constant over time, i.e., the substitution probabilities remain constant in different parts of the tree.
- *Stationarity.* The relative frequencies are in equilibrium, i.e., they remain constant over time.

Although these assumptions might not be valid in biological processes, in practice, these models have proved to be good tools to explain these processes. The assumptions are just the cost of a simplified representation of reality and obey mathematical convenience. Models which relax the last two are discussed in Section 2.3.5.

2.3.1.4 Time reversibility

Time reversibility means that the process will look the same whether it is considered forward or backward in time. It is an assumption commonly used in practice. Even though there is no biological argument for this assumption, it is supported by a practical reason: it reduces the complexity of the model, facilitating significant calculations. For instance, it makes easier to calculate the exponential matrix in the transitional probability matrix (2.3) and the likelihood on a tree, which under this assumption becomes independent of the location of the root. More advantages of time-reversibility are discussed in Keilson [1979].

The use of unrestricted models (without the reversibility restriction) complicates calculations and might not be worthwhile attempting. One of these complications is that they could involve complex numbers in the calculation of the probabilities. Yang [1994a] presented some cases where the unrestricted models yield similar estimated rate matrices to those obtained from the general time reversible models. Hence, the cost could be unjustified. Actually, Yang suggested that reversibility does not sacrifice biological reality.

Reversibility states that sampling x from the stationary distribution and going to state y is equivalent to sampling y from the stationary distribution and going to state x. In probabilistic terms, this assumption is given by

$$\pi_x P_{xy}(t) = \pi_y P_{yx}(t),$$

for $x, y \in \widetilde{S}$, and $t \geq 0$. This means that the probability of observing $X(t) = y$ and $X(0) = x$ is identical to the probability of observing $X(t) = x$ and $X(0) = y$. In a nucleotide sequence, the probability of observing the mutation $\mathtt{A} \rightarrow \mathtt{G}$ is equivalent to the probability of observing $\mathtt{G} \rightarrow \mathtt{A}$. In other words, the process of substitution between time 0 and t remains the same whether it is considered forward or backward in time. Note that time reversibility does not imply symmetry in the transitional probability matrix $\mathbf{P}(t)$. This restriction can be imposed by considering equal relative rate parameters in the instantaneous rate matrix. This is equivalent to considering

$$\pi_x q_{xy} = \pi_y q_{yx}.$$

For DNA sequences, the $\mathbf{Q}$ matrix defined in (2.7) is restricted to $r_{AC} = r_{CA}$, $r_{AG} = r_{GA}$, $r_{AT} = r_{TA}$, $r_{CG} = r_{GC}$, $r_{CT} = r_{TC}$, and $r_{GT} = r_{TG}$. Thus, the most general time-reversible model, known as GTR, can be represented by

$$\mathbf{Q} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \quad\texttt{A}\quad & \quad\texttt{C}\quad & \quad\texttt{G}\quad & \quad\texttt{T}\quad \end{array} \\ \begin{pmatrix} - & \mu r_{AC}\pi_C & \mu r_{AG}\pi_G & \mu r_{AT}\pi_T \\ \mu r_{AC}\pi_A & - & \mu r_{CG}\pi_G & \mu r_{CT}\pi_T \\ \mu r_{AG}\pi_A & \mu r_{CG}\pi_C & - & \mu r_{GT}\pi_T \\ \mu r_{AT}\pi_A & \mu r_{CT}\pi_C & \mu r_{GT}\pi_G & - \end{pmatrix} \end{array}. \tag{2.9}$$

Most of the most popular nucleotide substitution models rely on the time reversibility assumption. These models differ basically in the parametric structure of their instantaneous transition rate matrix. Naturally, they have chronologically evolved from the simplest to more general models due to the advancement of new methodologies and technology. Their evolution has been accompanied by the development of new statistical methods which have allowed their tractability. In general, they were initially developed for genetic distances for 2 sequences and then their application was extended to the evolutionary reconstruction for multiple sequences. Some of these models are discussed below.

2.3.1.5 The JC69 model

The JC69 model [Jukes and Cantor, 1969], or simply JC, is the simplest possible model of evolution. It assumes that each nucleotide has the same probability of mutation and when this occurs, the nucleotide changes to one of the three other possible nucleotides with equal probability. Even though its assumptions could be considered unrealistic, it was essential as a starting point to generate more complex models. As an anecdote, this model had to be included just as a part of a large empirical paper in order to be published and not rejected by the editors [Felsenstein, 2001]. This model has gradually been acquiring more complexity and thus flexibility generating new models.

This model sets the relative rate parameters to 1.0. In other words, it assumes that there is no difference between the occurrence of the different kinds of mutations ($r_{AC} = r_{AG} = r_{AT} = r_{CG} = r_{CT} = r_{GT} = 1.0$). Also, it assumes equal frequencies, i.e., the nucleotides occur with the same frequency ($\pi_A = \pi_C = \pi_G = \pi_T = 1/4$). These restrictions yield the following rate matrix

$$\mathbf{Q} = \begin{pmatrix} \text{-}3\mu/4 & \mu/4 & \mu/4 & \mu/4 \\ \mu/4 & \text{-}3\mu/4 & \mu/4 & \mu/4 \\ \mu/4 & \mu/4 & \text{-}3\mu/4 & \mu/4 \\ \mu/4 & \mu/4 & \mu/4 & \text{-}3\mu/4 \end{pmatrix},$$

which characterizes JC69. After some matrix algebra, the substitution probabilities of this model are given by

$$P_{xy}(t) = \begin{cases} 1/4 + 3/4\ e^{-\mu t}, & \text{if } x = y, \\ 1/4 - 1/4\ e^{-\mu t}, & \text{if } x \neq y. \end{cases}$$

The rate and the frequencies are usually combined into a single parameter $\alpha = \mu/4$, which reduces the instantaneous rate matrix to

$$\mathbf{Q} = \begin{pmatrix} \text{-}3\alpha & \alpha & \alpha & \alpha \\ \alpha & \text{-}3\alpha & \alpha & \alpha \\ \alpha & \alpha & \text{-}3\alpha & \alpha \\ \alpha & \alpha & \alpha & \text{-}3\alpha \end{pmatrix},$$

and consequently, the substitution probabilities to

$$P_{xy}(t) = \begin{cases} 1/4 + 3/4\ e^{-4\alpha t}, & \text{if } x = y, \\ 1/4 - 1/4\ e^{-4\alpha t}, & \text{if } x \neq y. \end{cases}$$

As it was discussed before, the mean substitution rate is usually set to 1. This transforms the branch length t into the number of substitutions per site. This is done by scaling $\mathbf{Q}$ by its inverse mean rate, which for this model is $(3\mu/4)^{-1}$. Thus, the average rate of substitutions at equilibrium is 1.

2.3.1.6 The K80 model

The K80 model [Kimura, 1980], also called K2P model, differs from the previous one by introducing a parameter to account for the kind of mutations. This is done by considering a different rate parameter for transitions and transversions ($r_{AG} = r_{CT}$ and $r_{AC} = r_{AT} = r_{CG} = r_{GT}$). This consideration is consistent with reality, since, as is well known, transitions occur more frequently than transversions, even though the latter can happen in more ways (G↔C, G↔T, T↔A or T↔G). The nucleotides are assumed equally frequent ($\pi_A = \pi_C = \pi_G = \pi_T = 1/4$).

The transversion rate is set to 1.0 in order to leave the transition rate measured with respect to it ($r_{AC} = r_{AT} = r_{CG} = r_{GT} = 1.0$). Thus, the rate matrix for this model is given by

$$\mathbf{Q} = \begin{pmatrix} \text{-}\mu(b+2)/4 & \mu/4 & b\ \mu/4 & \mu/4 \\ \mu/4 & \text{-}\mu(b+2)/4 & \mu/4 & b\ \mu/4 \\ b\ \mu/4 & \mu/4 & \text{-}\mu(b+2)/4 & \mu/4 \\ \mu/4 & b\ \mu/4 & \mu/4 & \text{-}\mu(b+2)/4 \end{pmatrix},$$

where $b = r_{AG} = r_{CT}$. This model can also be found in the literature with different parametrization. For instance, setting the transition rate to $\alpha = b\ \mu/4$ and the transversion rate to $\beta = \mu/4$, the rate matrix can be rewritten as

$$\mathbf{Q} = \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & -\alpha-2\beta \end{pmatrix}.$$

Note that if α/β, which represents the transition bias, is equal to 1, the model reduces to the JC69 model. This means that there is no preference for transitions.

The K80 model assigns different probabilities for transition and transversion mutations as well as for no substitutions. These probabilities are given by

$$P_{xy}(t) = \begin{cases} 1/4 + 1/4\ e^{-\mu t} + 1/2\ e^{-\mu t\left((b+1)/2\right)}, & \text{if } x = y, \\ 1/4 + 1/4\ e^{-\mu t} - 1/2\ e^{-\mu t\left((b+1)/2\right)}, & \text{if } x \neq y, \text{transition} \\ 1/4 - 1/4\ e^{-\mu t}, & \text{if } x \neq y, \text{transversion} \end{cases}$$

In order to make the expected flux of the model equal to 1, the instantaneous rate matrix must be scaled by $(\mu(b+2)/4)^{-1}$, which is the inverse mean rate.

2.3.1.7 The F81 model

The F81 model [Felsenstein, 1981] is another extension of the JC69 model. As its predecessor, this model considers equal relative parameter rates, but it differs by allowing the incorporation of different parameters for the frequencies. This flexibility is realistic since, in general, the overall DNA frequencies will not be equal, namely, the frequencies of `A` and `T` are approximately equal as well as the frequencies of `C` and `G`.

Its rate matrix is given by

$$\mathbf{Q} = \begin{pmatrix} \text{-}\mu(\pi_G+\pi_Y) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \text{-}\mu(\pi_T+\pi_R) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \text{-}\mu(\pi_A+\pi_Y) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & \text{-}\mu(\pi_C+\pi_R) \end{pmatrix},$$

where $\pi_Y = \pi_C + \pi_T$ and $\pi_R = \pi_A + \pi_G$ are pyrimidine and purine frequencies, respectively. In order to make the mean rate equal to 1, the instantaneous rate matrix must be multiplied by $\big(2\mu(\pi_A\pi_G + \pi_C\pi_T + \pi_Y\pi_R)\big)^{-1}$, which is the inverse mean rate.

After some matrix algebra, the probability matrix can be described by

$$P_{xy}(t) = \begin{cases} \pi_y + (1.0 - \pi_y)\, e^{-\mu t}, & \text{if } x = y, \\ \pi_y(1.0 - \ e^{-\mu t}), & \text{if } x \neq y. \end{cases}$$

2.3.1.8 The HKY85 model

The HKY85 model [Hasegawa and Kishino, 1984; Hasegawa et al., 1985], or simply HKY, generalizes the K80 and F81 model allowing a different relative rate for transitions and transversions and different frequencies. This model merges the attributes of both models into one. A model with the same features was independently implemented in PHYLIP computer package [Felsenstein, 1989]. It is called the F84 model and has a slightly different parametrization. This model is formally described in Kishino and Hasegawa [1989].

The instantaneous rate matrix for HKY85 is given by

$$\mathbf{Q} = \begin{pmatrix} \text{-}\mu(b\,\pi_G + \pi_Y) & \mu\pi_C & b\,\mu\pi_G & \mu\pi_T \\ \mu\pi_A & \text{-}\mu(b\,\pi_T + \pi_R) & \mu\pi_G & b\,\mu\pi_T \\ b\,\mu\pi_A & \mu\pi_C & \text{-}\mu(b\,\pi_A + \pi_Y) & \mu\pi_T \\ \mu\pi_A & b\,\mu\pi_C & \mu\pi_G & \text{-}\mu(b\,\pi_C + \pi_R) \end{pmatrix},$$

where $\pi_Y = \pi_C + \pi_T$ and $\pi_R = \pi_A + \pi_G$. By multiplying this matrix by its inverse rate mean $\left(2\mu(b(\pi_A\pi_G + \pi_C\pi_T) + \pi_Y\pi_R)\right)^{-1}$, the mean substitution rate is set to 1.

The HKY85 model assigns different probabilities to transitions, transversions, and no substitutions. These probabilities are given by

$$P_{xy}(t) = \begin{cases} \pi_y + \pi_y(1/\pi_y^* - 1)e^{-\mu t} + \big((\pi_y^* - \pi_y)/\pi_y^*\big)e^{-\mu t\rho}, & \text{if } x = y, \\ \pi_y + \pi_y(1/\pi_y^* - 1)e^{-\mu t} - \big(\pi_y/\pi_y^*\big)e^{-\mu t\rho}, & \text{if } x \neq y, \text{transition}, \\ \pi_y(1 - e^{-\mu t}), & \text{if } x \neq y, \text{transversion}, \end{cases}$$

where $\rho = 1 + \pi_y^*(b - 1)$, with $\pi_y^* = \pi_{\mathtt{A}} + \pi_{\mathtt{G}}$ if base y is a purine (A or G), and $\pi_y^* = \pi_{\mathtt{C}} + \pi_{\mathtt{T}}$ if base y is a pyrimidine (C or T) [Swofford et al., 1996].

2.3.1.9 The GTR model

The general time-reversible model (GTR) [Lanave et al., 1984; Tavaré, 1986], or also called the general reversible process model [REV; Yang, 1994a], is the most general model in its class. It allows the modeling of different relative transition and transversion rates and frequencies. Its instantaneous rate matrix is defined in (2.9). It was originally applied to sequence distance calculation and subsequently to substitution pattern estimation [Yang, 1994a]. Even though this model does not have explicit transition probabilities, unlike the other ones described above, they can be computed by numerical calculation of the eigenvalues and eigenvectors of its instantaneous rate matrix $\mathbf{Q}$ (see, for instance, Yang [1994a]).

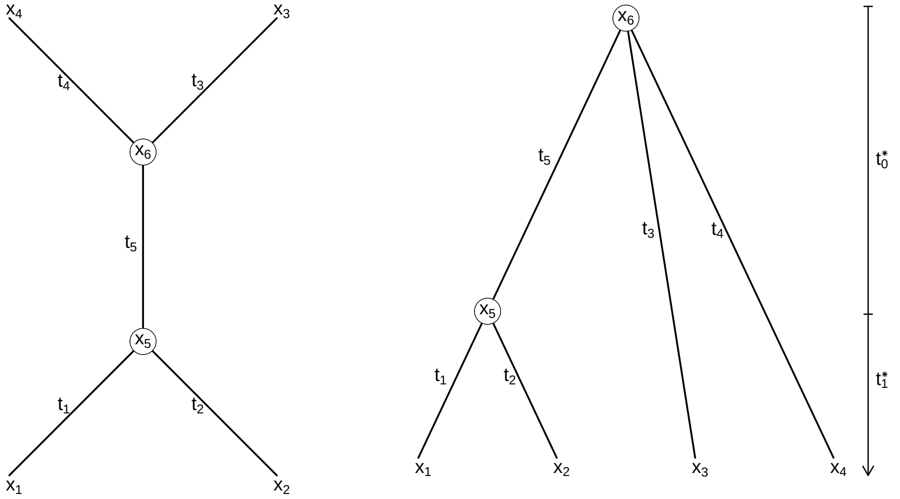
Allowing the six relative rates to vary and scaling the transition rate matrix, leads to a problem known as nonidentifiability [Zwickl and Holder, 2004]. This phenomenon happens when the model has two or more values for its parameters which produce the same distribution for the data. In other words, the data cannot discriminate between some parameter combinations. One common solution to this problem is to set one of the rates to 1. Thus, if the `TG` substitution rate is set to 1, the rest of rates are measured relative to it. For instance, a `AC` substitution rate value of 5.0 means that the `AC` mutations occurs 5 times more frequently than `TG` mutations. Another solution is to force the six rates to sum to 1 and as the previous approach, there are five free parameters but with a different interpretation. For instance, if the `TG` rate parameter is 0.20, then 20% of all mutations are expected to be between `T` and `G`. The former solution is known as 5RR and the latter one as ST1 [Zwickl and Holder, 2004].

2.3.2 Likelihood calculation

The Markov process defined previously describes an evolutionary course along a single edge, which represents a time of length t, of a phylogeny τ. In the same way, this process can be applied to each branch of a phylogeny in order to describe the evolution of multiple sequences along the full tree. The substitution process, defined as the random variable X, assigns to each vertex in τ an element of the set of states $\widetilde{S}$ following a Markov process on τ. Thus, the evolution of a character from one node to another can be described in probabilistic terms and the probability of the tree can be calculated.

The likelihood represents the probability of observing the data given the parameters and the model. Since the data are fixed in the analyses, the likelihood is considered as a function of the parameters given the data and the model. This explains why it does not integrate to 1 over the parameter space, but it does over the sample space.

The dataset consists of s aligned homologous sequences with m sites each and is represented by an $s \times m$ matrix $\boldsymbol{X} = \{x_{ij}\}$. Each column $\boldsymbol{x}_h$ of this matrix represents a site. For instance, the second site in the example presented in Figure 2.4 is $\boldsymbol{x}_2 = (\mathtt{C}, \mathtt{T}, \mathtt{C}, \mathtt{T}, \mathtt{T})$. Each element of this vector determines a character $\in \widetilde{S}$ on the leaves of the tree. The sites are usually assumed to be evolving independently of the remaining sites. The evolution in one lineage is also assumed independent of the other lineages.



(a) Unrooted.

(b) Rooted.

Figure 2.6: Trees for 4 species. Assuming reversibility, both trees have the same likelihood. The arrow on the right is used as reference in the molecular clock assumption, under which we have $t_1^* = t_1 = t_2$, $t_3 = t_4$, and $t_0^* = t_5 = t_3 - t_1$.

Assuming independence between sites, i.e., between the columns of $\boldsymbol{X}$, the likelihood for a given rooted tree is defined as

$$L(\boldsymbol{\theta}) = L(\boldsymbol{X}|\boldsymbol{\theta}, \tau) = \prod_{h=1}^{m} L(\boldsymbol{x}_h|\boldsymbol{\theta}),$$

which is the product of the probabilities for each site $L(\boldsymbol{x}_h|\boldsymbol{\theta})$. Therefore, the log-likelihood is given by

$$l(\boldsymbol{\theta}) = \sum_{h=1}^{m} \log L(\boldsymbol{x}_h|\boldsymbol{\theta}).$$

As the likelihood is extremely small in most cases in phylogenetics, it is preferable to work with its log version instead.

In general, the states of the leaves are all the available information about the evolutionary process of the taxa being analysed. In other words, the states of the internal nodes, which usually represent extinct ancestors, are commonly unknown. Therefore, to calculate the probability of a site $L(\boldsymbol{x}_h|\boldsymbol{\theta})$, it is necessary to sum over all possible states of the internal nodes of the tree.

We illustrate the calculation of the probability of a site in the 4 taxon phylogeny displayed in Figure 2.6a. But first, this unrooted tree needs to be rooted. By the time reversibility assumption, the root can be allocated arbitrarily at any internal node without affecting the likelihood, i.e., the latter is independent of the root position. This property is called the *pulley principle* [Felsenstein, 1981]. This principle allows to define a class of rooted trees which have the same likelihood and which are compatible with a single unrooted tree. This explains the reason behind the inability of likelihood-based methods in estimating rooted trees under the time reversibility assumption. In our example, we root the phylogeny at x_6 and the resulting tree is displayed in Figure 2.6b. Thus, the likelihood calculation procedure is described jointly for the unrooted and rooted trees and its extension to a larger phylogeny is analogous.

The vector $\boldsymbol{x}_h$, which represents the states of the leaves of the tree in the h^{th} site, contains the states x_1, x_2, x_3 and x_4. The internal states (x_5 and x_6), which represent the states of the ancestors, are unknown. Thus, to calculate the probability for $\boldsymbol{x}_h$ we must sum over all possible states of x_5 and x_6.

The calculation starts with the probabilities on the root which are given by the stationary distribution $\boldsymbol{\pi}$. From it, the probabilities are calculated along the tree. All of them are contained in the probability matrix $\mathbf{P}(t)$ which only varies according to the branch length t. Thus, the probability for a site is given by

$$\begin{aligned} L(\boldsymbol{x}_h|\boldsymbol{\theta}) &= \sum_{x_6}\sum_{x_5} \pi_{x_6} P_{x_6x_5}(t_5) P_{x_5x_1}(t_1) P_{x_5x_2}(t_2) P_{x_6x_3}(t_3) P_{x_6x_4}(t_4) \\ &= \sum_{x_6} \pi_{x_6} P_{x_6x_3}(t_3) P_{x_6x_4}(t_4) \sum_{x_5} P_{x_6x_5}(t_5) P_{x_5x_1}(t_1) P_{x_5x_2}(t_2), \end{aligned} \tag{2.10}$$

where $P_{x_ix_j}(t)$ are the entries of the probability matrix $\mathbf{P}(t) = e^{t\mathbf{Q}}$.

This calculation involves a very large number of terms which increases exponentially with the number of taxa. Nevertheless, there are some considerations that allow considerable time saving. First, the probability calculation can be carried out over descendant nodes and then over ancestral nodes. This is illustrated in (2.10) where the summation sign has been moved to the right which reduces the calculations. This simplification is known as the *pruning algorithm* [Felsenstein, 1973, 1981]. Second, if two or more sites have the same pattern, then they have the same probability. Therefore, their probability just needs to be calculated once and multiplied by their frequency. Finally, the probability matrix is calculated only once per branch and remains constant across sites.

Under the assumption of molecular clock, i.e., the substitution rates remain constant across the lineages in the phylogeny, the number of branch length parameters decreases. For a binary tree of s taxa, there will be $s-1$ internal nodes and consequently $s-1$ branch lengths (the height of each node). This assumption allows the identification of the root and measures the branch lengths according to it.

In our previous example, we identify two branch lengths which are depicted by the vertical line in the phylogeny displayed in Figure 2.6b. First t_0^*, which is defined from the functional root x_6 to the internal node x_5, and second t_1^*, which is defined from x_5 to its descendants x_1 or x_2. Considering these restrictions, the probability at a site assuming molecular clock becomes

$$L(\boldsymbol{x}_h|\boldsymbol{\theta}) = \sum_{x_6} \pi_{x_6} P_{x_6x_3}(t_0^* + t_1^*) P_{x_6x_4}(t_0^* + t_1^*) \sum_{x_5} P_{x_6x_5}(t_0^*) P_{x_5x_1}(t_1^*) P_{x_5x_2}(t_1^*).$$

2.3.3 Variable substitution rate across sites

The Markov models presented previously assume that all sites evolve at the same rate through the transition rate matrix **Q**. This assumption is violated in many real cases and it has been well documented in the literature [e.g., Hodgkinson and Eyre-Walker, 2011]. In these cases, its omission can lead to wrong conclusions [Gaut and Lewis, 1995; Sullivan and Swofford, 2001; Yang, 1996a]. For instance, Gaut and Lewis [1995] studied a simulated dataset in the presence of rate variation across sites and showed that maximum likelihood (ML) inference can be inconsistent under models which ignore this variability. As the length of the sequence increased, ML methods converged to the wrong tree.

Early methods dealt with this problem by keeping a portion of invariable sites and the rest varying at the same rate [e.g., Hasegawa and Kishino, 1984; Hasegawa et al., 1985]. This idea has been extended considering different categories or rate classes for the variable sites. This allows the use of different substitution rates for groups of sites according to their category. This method is known as the discrete-rate model.

As it is unknown to which category each site belongs, the probability for a given site is calculated as a weighted average which is composed by the product of the probability that the site belongs to the category and the probability of the site given the rate of that category. Thus, for k categories, where the rates are $r_1, \ldots, r_k$, with corresponding probabilities $p_1, \ldots, p_k$, respectively, the probability of a site is given by

$$L(\boldsymbol{x}_h|\boldsymbol{\theta}) = \sum_{i=1}^{k} p_i L(\boldsymbol{x}_h|r_i, \boldsymbol{\theta}),$$

where $\boldsymbol{x}_h$ is the site, $\boldsymbol{\theta}$ the parameter vector and $L(\boldsymbol{x}_h|r, \boldsymbol{\theta})$ the probability of observing the data $\boldsymbol{x}_h$ under the rate r. This latter term corresponds to the probability of the site under the one-rate model and its calculation is carried out by multiplying each branch length by the rate r. In the previous example (2.10), this would be

$$L(\boldsymbol{x}_h|r, \boldsymbol{\theta}) = \sum_{x_6} \pi_{x_6} P_{x_6x_3}(rt_3) P_{x_6x_4}(rt_4) \sum_{x_5} P_{x_6x_5}(rt_5) P_{x_5x_1}(rt_1) P_{x_5x_2}(rt_2).$$

Another similar approach is to assume that the rates over sites are random variables which follow a continuous distribution. Thus, the probability for a site is given by the integral of the likelihood of the site over the distribution of the rates, that is,

$$L(\boldsymbol{x}_h|\boldsymbol{\theta}) = \int g(r) L(\boldsymbol{x}_h|r, \boldsymbol{\theta}) \mathrm{d}r,$$

where $g(r)$ is the probability density of the rates r. Again, due to the lack of information about the relationship between the site and the rate, the likelihood for the site is calculated as its expected probability with respect to the rate. This is similar to the discrete rate model, but in the continuous case.

Yang [1993] proposed the gamma distribution to model the rate parameter and it has become the most used. The choice is based on practical rather than theoretical considerations; it had already been used in the study of evolutionary

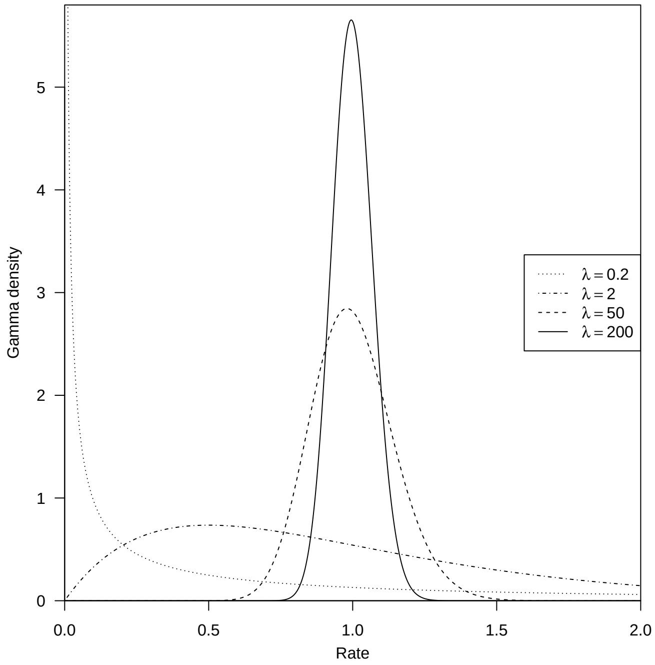


Figure 2.7: Gamma densities for different shape parameters.

divergence [e.g., Golding, 1983; Nei and Gojobori, 1986]. Its density is given by

$$g(r) = \frac{\beta^\lambda}{\Gamma(\lambda)} r^{\lambda-1} e^{-\beta r}, \quad 0 < r < \infty, \tag{2.11}$$

with $\lambda, \beta > 0$, shape and rate parameters, respectively, mean $\mathbb{E}[r] = \lambda/\beta$ and $\mathrm{Var}(r) = \lambda/\beta^2$. To reduce the number of parameters, it is usual to set $\lambda = \beta$. As a result, the expected value of r is 1 and its variance $1/\lambda$. Thus, the distribution is centered at 1 and shaped by λ, and is displayed in Figure 2.7 for different λ values. For small values (≤ 1), the distribution is positively skewed (L-shape). This reflects the common scenario [e.g., Yang, 1996a] where most sites have very low rates (no substitutions) or are practically invariable, with a few sites having very high rates (lots of substitutions). For large λ values, the probability mass is concentrated symmetrically around 1 and the distribution is approximately normal. This describes the situation of weak rate heterogeneity.

The use of the gamma parameter to scale the rates of the model is represented by adding the suffix "+Γ" to the corresponding model. For instance, the general time reversible model with a gamma stochastic component in its rates is represented by GTR+Γ.

The calculation of the probability for a site becomes exponentially expensive as the number of taxa increases. Note that this calculation involves an integral, and the function under the integral is polynomial in the number of taxa. Also, the calculation has to be applied to each different site in order to evaluate the likelihood function. This makes the approach impractical for a large number of taxa. A common practical solution is to discretize the gamma distribution. This approach is explained below.

2.3.3.1 Discrete gamma model

The most common method by far was proposed by Yang [1994b] and consists of a discretization of the continuous gamma density. The distribution is divided into k categories which have equal probabilities $1/k$. The rates are calculated as the expected values of the corresponding categories. Thus, the average rate for the i^{th} category is given by

$$\begin{aligned} r_i = \mathbb{E}\big[r|r \in [a,b]\big] &= \int_a^b r g(r)\mathrm{d}r \bigg/ \int_a^b g(r)\mathrm{d}r \\ &= k\frac{\beta^\lambda}{\Gamma(\lambda)} \int_a^b r^\lambda e^{-\beta/r}\mathrm{d}r, \end{aligned}$$

where a and b are the cutting points of this category such that $\mathbb{P}(a < r < b) = 1/k$. Like the continuous case, the parameters are usually set to $\lambda = \beta$, leaving only one free parameter. In general, 4-8 categories are enough to obtain a good approximation of the continuous function [Yang, 1994b].

After calculating the rates, the probability for each site can be obtained as in the discrete-rate model discussed before, that is,

$$L(\boldsymbol{x}_h|\boldsymbol{\theta}) = \frac{1}{k}\sum_{i=1}^{k} L(\boldsymbol{x}_h|r = r_i, \boldsymbol{\theta}).$$

The single rate model is obtained when $k = 1$, while the continuous gamma model is equivalent to $k = \infty$. The discrete gamma model is analogous to the discrete-rate model, but its rates are defined by the gamma distribution. The notation for this model with k rate categories is characterized by adding the suffix "$+\Gamma_k$" to the corresponding model. For instance, the JC69 model with a gamma parameter for the rates across sites, with k categories, is represented by JC69$+\Gamma_k$. A generalization of this approach was proposed by Mayrose et al. [2005], who instead of using a single gamma distribution, proposed a gamma mixture model. This approach allows greater flexibility in the shape of the distribution of the rates. The authors showed that this model better describes the evolutionary process for proteins.

2.3.4 Invariable sites

As was discussed previously, the assumption of equal rates of substitution across sites is often violated in practice. One alternative is to divide the sites into categories which are characterized by particular rates, namely, the discrete-rate model. Along these lines, Hasegawa and Kishino [1984] considered a category with invariable sites, i.e., a proportion of sites is assumed to have substitution rate equal to 0. Later, Gu et al. [1995] considered a proportion of invariable sites in addition to the gamma distribution for the rate across sites. This model assumes that a proportion of sites in the sequences are invariant whereas the remaining ones follow a gamma distribution. Such models are labeled by the suffix "Γ+I", where "I" denotes the invariant sites. For instance, a GTR model with a gamma parameter for the rates across sites and in addition, invariable sites, is represented by GTR$+\Gamma+$I. In short, it is known as a gamma-invariable mixture model and it has been widely used in phylogenetic analysis.

The probability density of the rates for Γ+I models is given by

$$g(r) = p_0 g_0(r) + (1 - p_0) g(r),$$

where g is the gamma distribution defined in (2.11), g_0 is the delta function at 0, i.e., $g_0 = 1$ if $r = 0$, otherwise $g_0 = 0$, and p_0 is the proportion of invariable sites. The other proportion of sites $1 - p_0$ are drawn from the gamma distribution. The mean and variance of r are given by

$$\mathbb{E}(r) = (1 - p_0)\frac{\lambda}{\beta} \qquad \text{and} \qquad \mathrm{Var}(r) = \frac{1}{\beta^2}\big(\lambda(1 - p_0)(1 + p_0\lambda)\big).$$

The mean is fixed at 1, thus $\beta = (1 - p_0)\lambda$ and the variance reduces to

$$\mathrm{Var}(r) = \frac{1 + p_0\lambda}{(1 - p_0)\lambda}.$$

Even though this kind of model is commonly used in phylogenetic analysis, there are some controversies about its use. Yang [2014] cataloged this model as pathological since the gamma parameter for $\lambda < 1$ already accommodates sites with very low rates [Golding, 1983]. As a result, there is a strong correlation between λ and p_0 which causes problems to get reliable estimates [Gu et al., 1995; Sullivan et al., 1999]. Another disadvantage is the dependence of p_0 on the number and divergence of the sequences in the data. The proportion of invariable sites p_0 is never larger than the observed proportion of constant sites, which decreases with the inclusion of highly divergent sequences. As a result, p_0 estimate tends to decrease too [Yang, 2014]. In fact, both parameters (λ and p_0) are highly sensitive to taxon sampling [Sullivan et al., 1999; Yang, 1996a]. In addition, Jia et al. [2014] found that p_0 is highly sensitive to the number of rate categories in the popular discrete gamma. In conclusion, Yang [2014] recommended avoiding Γ+I models and using Γ models instead. In fact, Jia et al. [2014] showed that discrete gamma models with between 6 and 10 rate categories is enough to achieve a balance between model complexity, computational cost, and parameter estimation in interspecific data. Additionally, Lemmon and Moriarty [2004] showed that GTR+Γ works quite well in terms of bipartition posterior probability, branch length and other model parameter estimates, under data simulated from a GTR+ Γ+I.

2.3.5 More complex models

In general the Markov models used to describe evolutionary processes assume stationarity, homogeneity and reversibility. The stochastic description is applied along each branch and thus along the entire phylogeny at each site. Therefore, the conditions are assumed on 2 levels: local and global. The first level refers to the assumptions in a branch, whereas the second level refers to the full phylogeny. For instance, homogeneity means that the instantaneous transition rate matrix is constant over a branch (local homogeneity) and over the entire tree, i.e., all the branches of the tree (global homogeneity). These assumptions can be violated, increasing the probability of incorrect phylogenetic results [Jayaswal et al., 2005]. Examples of these kinds of evolutionary processes abound in the literature; see for instance Jayaswal et al. [2005]; Yang [1994a].

Barry and Hartigan [1987] proposed a general Markov model (GMM), which is the most general model of evolution. This model does not make the assumptions of stationarity, homogeneity (local and global) and reversibility. Their proposed model considered different instantaneous rate matrices $\mathbf{Q}$, which do not require the reversibility assumption, for each branch of the phylogeny, assuming independent and identically distributed sites. If the evolutionary process is not stationary and homogeneous, the phylogeny obtained by using this general nonhomogeneous model may differ from that obtained by using standard time-reversible models [Jayaswal et al., 2005].

Sumner et al. [2012b] studied a case where the evolutionary process along a branch occurs according to two different GTR models. This process is represented in Figure 2.8, where the evolution occurs firstly according to $\mathbf{Q}_1$ and then finishes according to $\mathbf{Q}_2$. It emulates the case of violation of the assumption of homogeneity. They showed that a single GTR model is not able to estimate reasonably an average of the two instantaneous matrices involved in the data generating process, i.e., an approximation of the instantaneous rate matrix $\widehat{\mathbf{Q}}$, where $e^{(t_1+t_2)\widehat{\mathbf{Q}}} = e^{t_1\mathbf{Q}_1}e^{t_2\mathbf{Q}_2}$. The approximation becomes worse as the process becomes more heterogeneous, in other words, when the discrepancy between $\mathbf{Q}_1$ and $\mathbf{Q}_2$ increases. The biological reality is likely to be time-dependent and lineage-specific; thus a homogeneous Markov model is only a simplification of this reality. One would expect an "average" of the true heterogeneous effects from these models. This desirable property is not met by GTR, the most popular model of evolution used in phylogenetic analyses.

A nonhomogeneous process can be approximated by a homogeneous one if

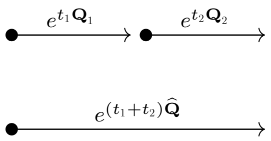


Figure 2.8: The upper evolutionary process is nonhomogeneous. Below, this process is approximated by a homogeneous one, where $e^{(t_1+t_2)\widehat{\mathbf{Q}}} = e^{t_1\mathbf{Q}_1}e^{t_2\mathbf{Q}_2}$.

the models possess the property of *multiplicative closure* [Sumner et al., 2012a]. A Markov model $\mathcal{M}$ is said to be multiplicative closed if and only if for all $e^{t_1\mathbf{Q}_1}, e^{t_2\mathbf{Q}_2} \in \mathcal{M}$, the condition $e^{t_1\mathbf{Q}_1}e^{t_2\mathbf{Q}_2} \in \mathcal{M}$ is met. In this case, the transition rates of the heterogeneous process can be approximated by a kind of average of their effects. The product of two GTR substitution matrices does not produce another GTR matrix, i.e., this model does not meet the multiplicative closure property. Sumner et al. [2012a] discussed about how to generate models which fulfill this mathematical property, a group referred as "*Lie Markov models*". They showed that GMM belongs to this class of models.

Other assumptions used in phylogenetic analyses are related to the likelihood calculation presented above in Section 2.3.2. It is often assumed that the sites are independent and identically distributed under the same Markov model. These are very usual assumptions, which reduce the calculations significantly, but do not necessarily conform to biological reality. Different techniques have been proposed in order to relax these assumptions.

With respect to the independence among sites, Yang [1995] and Felsenstein and Churchill [1996] relaxed this assumption by allowing correlated rates at adjacent sites. Thus, the sites in the sequence are assumed to evolve over time. Their models are extensions of the among-site rate variation models presented in Section 2.3.3, in which the rates differ among sites and the relative rates of change at the individual sites are unknown, but they are not correlated. The first proposal considers a discretized bivariate gamma distribution, leading to an *auto-discrete-gamma* model, to describe the autocorrelation of substitution rates at adjacent sites. The transition from one rate class to another along the sequence is described by a hidden Markov chain, which introduces the dependence. Thus, the correlation of the rates at sites is modeled. The second approach also makes use of a hidden Markov model in order to describe the belonging of a site to a given finite pool of categories, for which the probabilities are assumed known *a priori*. This Markov process introduces the dependence between adjacent rates in the clustering of the sites. The analyses of real data suggest a high degree of correlation among the rates, however, it seems that its omission does not affect the parameter estimates of the models significantly [Yang, 2014].

In these models, which include different rates among sites and eventually correlation between them, the relative rates are applied to the entire phylogeny. In other words, the evolutionary rates are allowed to vary along the molecule, but not along the tree, i.e., they are kept constant over time. For instance, slow and fast evolving sites keep their condition throughout every lineage of the phylogeny. However, this assumption might be violated in reality, where the rates can vary in time. The *covarion* models [for COncomitantly VARIable codON; Fitch, 1971a; Fitch and Markowitz, 1970] relax this assumption. These models allow each site to vary according its own evolutionary rate process by switching between being free to evolve in some taxa and being fixed in others [Penny et al., 2001]. Thus, covarion models allow the rate to vary among lineages as well. Extensions of these models have been studied in the literature, see for instance Galtier [2001]; Tuffley and Steel [1998].

With respect to the other assumption in the likelihood calculation, which says that the sites are identically distributed, it could be biologically unrealistic because it does not take into account compositional variation across the alignment. For instance, it is known that the third codon[1] evolves faster than the other two, or the assumption could be not met in the case of analysing multiple genes simultaneously. Two approaches can be considered to account for this variation among sites. If it is known a priori which sites are evolving in certain way, the data can be partitioned according to it. Then, the data can be analysed by using *partition* models, which use combined Markov processes applied to the different sections of the alignment. On the other hand, if it is not possible to make the partitions, a probability distribution can be assumed to model the heterogeneity of the data. These models are called *mixture* models. It is also possible to com-

[1] sequence of 3 nucleotides that translates to a particular genetic code in a RNA or DNA molecule.

bine both approaches into one model called a *mixed-effect* model. These models, which allow heterogeneity in the data, have been used widely in the literature, see for instance Fan et al. [2011]; Lartillot and Philippe [2004]; Nylander et al. [2004]; Yang [1996b]; Yang et al. [1995].

The compositional variation can also occur across the lineages, for instance, when distantly related species are compared. In this situation, it is impossible to approximate the evolutionary process by any time-reversible Markov model [Jermiin et al., 2008]. Actually, this variation can be modeled basically by assigning different Markov processes along the phylogeny, which can be time reversible models (local assumptions), but the evolutionary process as a whole is not reversible (global assumption). Different approaches have been implemented to model heterogeneity across the sequences with the aim of finding a trade-off between simplicity and accuracy, see for instance Blanquart and Lartillot [2006]; Foster [2004]; Heaps et al. [2014]; Yang and Roberts [1995].

Models which take into account compositional heterogeneity, either across sites and/or lineages, are often referred as *heterogeneous models*, whereas those ones which make use of a single Markov process are called *homogeneous models*. A practical discussion about the analysis of molecular data and the use of these kind of models can be found in Moran et al. [2015].

2.4 Prior distributions

A model is a structure composed of parameters which aims to describe the data-generating process. As was discussed above, the models differ in which aspects of the evolutionary process they allow to vary. This variation is described by the parameters. The way of dealing with these parameters leads to two statistical approaches: *frequentist* and *Bayesian*. The frequentist approach assumes fixed but unknown values for the parameters and its inferences about them are valid in repeated sampling, whereas the Bayesian approach considers them as random variables, which allows the use of probability distributions to represent their uncertainty. For the former, an optimization algorithm is required to estimate the parameters. However, for the latter, it is compulsory to add extra information about them via prior distributions before starting the analysis.

The prior distribution represents our knowledge about the parameters before collecting and analysing the data. This is a natural way of incorporating available information about them before the study, but it could also represent our ignorance. In these cases, the prior is called *informative* and *non-informative*, respectively. The combination of this information, through the prior distribution, and the information contained in the data, through the likelihood, composes the posterior distribution which is the core of Bayesian inference.

The inclusion of researcher's belief into the analysis through the prior distributions is one of the main sources of criticism in Bayesian analysis, because it allows the researcher include subjective information which could potentially lead to biased conclusions. However, there are priors which allow the representation of our knowledge more objectively. Furthermore, it is possible to incorporate reasonably our lack of it (a good review is given in Kass and Wasserman [1996]), i.e., we can consider priors with a small effect on the conclusions of the study. Many of these priors consist in analytical approaches, for instance, Jeffrey's priors [Jeffreys, 1946], which are based on the Fisher information. However, they are rarely used in phylogenetics because they require tractable likelihood functions, which might not be feasible for complex models or multiple sequences [Wang and Yang, 2014]. Nevertheless, empirical attempts have been carried out to generate these kinds of priors. These aim to let the results of the analysis be dominated by the likelihood and consequently reduce the impact of the prior on the posterior distribution. In this section, we discuss mainly these kinds of prior distributions and those ones used so far in phylogenetics.

2.4.1 Tree

The most common practice is the use of a discrete uniform prior distribution over all the topologies [Holder et al., 2014; Huelsenbeck et al., 2004; Huelsenbeck and Ronquist, 2001; Jow et al., 2002; Suchard et al., 2001]. This is calculated by taking the inverse of the number of possible trees. For instance, for the unrooted binary tree case, each phylogeny τ has a prior probability given by

$$\mathbb{P}(\tau) = \frac{1}{\prod_{i=3}^{s}(2i-5)} = \frac{2^{s-3}(s-3)!}{(2s-5)!},$$

where s is the number of taxa.

However, this apparently natural way of representing ignorance about the tree topologies indirectly introduces information about other aspects of the tree. Pickett and Randle [2005] pointed out that this prior distribution induces a non-uniform prior distribution over clades. In fact, it tends to assign high prior probabilities to small and large clades. The authors exemplified this problem considering a set of 5 taxa (A-E) in the rooted tree case. The total number of trees is 105. The clade containing taxa A, B and C has a prior probability of $9/105 = 0.086$ and the one which contains taxa A and B has a prior probability of $15/105 = 0.14$. The smallest clade has highest probability. To sum up, the uniform prior assigns unequal probabilities on the clades. This phenomenon is didactically illustrated by Alfaro and Holder [2006] who considered the example of wanting to estimate the number of beans in a jar. In this case, we could assign equal probability to each possible value. However, we could not expect that this prior distribution assigned equal probabilities to hypothesis such as "the number of beans divisible by two" or "the number of beans divisible by three", even though it could be desirable to express our ignorance about them. Analogously, the discrete uniform prior assigns non-equal probabilities over the different clades, preventing us from introducing our lack of knowledge about them. In this respect, Steel and Pickett [2006] pointed out that no matter what kind of prior is assigned to the tree topology, it is impossible to assign uniform priors on clades. In practice, if the data are informative enough, these unequal clade probabilities are not a real problem [Brandley et al., 2006].

2.4.2 Branch length

The branch lengths are parameters of vital importance in phylogenetic inference. The inadequate incorporation of prior information (often unintentionally) can lead to, for instance, problems in the tree topology estimation. In this regard, Yang and Rannala [2005] showed that the posterior distribution for the tree topologies is sensitive to branch length prior specifications, in particular, to those for the internal branch lengths.

It is usual to consider independent and identically distributed priors on the branch lengths, even though they are not independent. One of these priors is the Uniform$(0, m)$ distribution, where m is the upper bound. This alternative is not recommended since it tends to overestimate the branch lengths and consequently the tree lengths (sum of all branch lengths) [Rannala et al., 2011; Wang and Yang, 2014]. For instance, Wang and Yang [2014] showed that for $s = 100$ and $m = 100$, the 95% prior interval for the tree length is 9850 ± 794.2, which is an unreasonable assumption since branch lengths are frequently small and consequently the tree length too. Moreover, this prior distribution with $m = 10$ or 100, often called non-informative, causes inflated clade probabilities [Yang and Rannala, 2005]. The exponential prior with a rate parameter λ is another alternative commonly used [see, for instance, Holder et al., 2014]. The default λ value in MrBayes [Huelsenbeck and Ronquist, 2001] is 10 which leads to a prior branch length mean of 0.1. However, this distribution also leads to an overestimation [Brown et al., 2010; Marshall, 2010; Rannala et al., 2011; Wang and Yang, 2014]. The posterior distributions for the branch lengths and tree length are extremely sensitive to the specification of λ. Therefore, this prior distribution is also not recommended.

An alternative to deal with the problems discussed above is to incorporate a hyperparameter in the exponential distribution. Suchard et al. [2001] assigned an inverse-gamma as a hyperparameter on the mean. This proposal is given by

$$\begin{aligned} t_i|\mu^* &\sim \text{Exp}(1/\mu^*), \quad \text{for} \quad i = 1, \ldots, 2s-3, \\ \mu^* &\sim \text{Inverse-Gamma}(\alpha^*, \beta^*), \end{aligned}$$

with

$$f(t_i|\mu^*) = \frac{1}{\mu^*} e^{-t_i/\mu^*} \quad \text{and} \quad f(\mu^*) = \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \mu^{*-\alpha^*-1} e^{-\beta^*/\mu},$$

where α^* and β^* are the shape and rate parameters. The authors recommended $\alpha^* = 2.1$ and $\beta^* = 1.1$, which makes μ have an expectation of 1.0 and variance of 10 and consequently a prior mean branch length of 1.0. These specifications lead to fairly diffuse priors for the branch lengths. Since this prior mean seems large, Rannala et al. [2011] suggested another parametrization of the inverse-gamma distribution with $\alpha^* = 3$ and $\beta^* = 0.2$, which leads to a prior mean for the branch length of 0.1. These specifications could prevent overestimation. The authors showed that the exponential prior with a hyperparameter on its mean performs much better than the uniform and exponential priors.

A more robust alternative is the compound Dirichlet prior proposed by Rannala et al. [2011]. This is constructed by first defining a prior distribution on the total tree length and then partitioning it into the branch lengths according to a Dirichlet distribution. The authors proposed a gamma distribution for the tree length T, defined by

$$f(T) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} T^{\alpha_T - 1} e^{-\beta_T T},$$

with mean α_T/β_T and variance α_T/β_T^2, where α_T and β_T are the shape and rate parameters, respectively. Following the suggestion of Yang and Rannala [2005], in order to reduce the high posterior probabilities for trees and clades, Rannala et al. defined different priors for the $s-3$ internal and s external branch lengths. Thus, the i^{th} branch length is defined as $t_i = v_i \times T$, for $i = 1, \ldots, (2s-3)$, where $\boldsymbol{v} = \{v_i\}$, which describes the distribution from which branch lengths are chosen, such that they have a fixed sum, has the Dirichlet distribution

$$f(\boldsymbol{v}) = \frac{\Gamma(s\alpha_v + (s-3)\alpha_v c)}{\Gamma(\alpha_v)^s \Gamma(\alpha_v c)^{s-3}} \prod_{j=1}^{s} v_j^{\alpha_v - 1} \prod_{h=1}^{s-3} v_h^{\alpha_v c - 1},$$

with the subscripts j and h representing external and internal branches (in the original article these appear incorrectly in the other way around), respectively, $v_j, v_h > 0$ with

$$\sum_{i=1}^{2s-3} v_i = \sum_{j=1}^{s} v_j + \sum_{h=1}^{s-3} v_h = 1,$$

and c the ratio of the mean internal/external branch lengths, which in general should be less than 1.0 in order to assign smaller values for the internal branches as Yang and Rannala [2005] suggested to control high posterior probabilities for trees and clades.

Thus, the joint prior distribution for the $2s-3$ branch lengths is given by

$$f(t) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T T} T^{\alpha_T - 1} \frac{\Gamma(s\alpha_v + (s-3)\alpha_v c)}{\Gamma(\alpha_v)^s \Gamma(\alpha_v c)^{s-3}} \prod_{j=1}^{s} t_j^{\alpha_v - 1} \prod_{h=1}^{s-3} t_h^{\alpha_v c - 1}$$
$$T^{-\alpha_v s - \alpha_v c(s-3)+1},$$

where $T = \sum_{i=1}^{2s-3} t_i$ is the tree length.

For the particular case of considering a uniform Dirichlet for the proportions v_i, i.e., $\alpha_v = 1$, and no distinction between internal an external branch lengths ($c = 1$), the joint prior distribution for the $2s-3$ branch lengths is given by

$$f(t) = \frac{\beta_T^{\alpha_T}}{\Gamma(\alpha_T)} e^{-\beta_T} T^{\alpha_T - 1} (2s-4)! T^{-2s+4}.$$

Yang and Rannala [2005] showed that this approach, with $\alpha_T = \beta_T = 1$, $\alpha_v = 1$ and $c = 1$, performs much better than the other prior distributions. Wang and Yang [2014] tested this prior for different parameterizations of the gamma distribution assigned to the tree length and keeping the remaining specifications fixed, namely, β_T equals to 0.1, 0.01 and 0.0001, and $\alpha_T = \alpha_v = c = 1$. Their results confirmed the initial conclusions about the robustness of the compound Dirichlet prior for the branch lengths with a gamma distribution for the tree length, and also the unreasonable large tree lengths yielded by the uniform and exponential prior distributions.

Alternatively, Yang and Rannala [2005] also considered an inverse-gamma distribution for the tree length. This distribution also has a robust behavior, in other words, it did not affect the posterior distribution of the tree length drastically.

2.4.3 Relative rates

The substitution rate matrix $\mathbf{Q} = q_{ij}$, with $i, j = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$, contains the instantaneous rate of substitutions from nucleotide i to nucleotide j, for $i \neq j$. Each element consists of the product of the relative rate r_{ij}, the frequency of the state

π_j and the mean instantaneous substitution rate μ. The instantaneous transition rate matrix $\mathbf{Q}$ for the GTR model is defined in (2.9) where, for instance, $q_{AC} = r_{AC}\pi_C\mu$. It is usual to rescale this matrix by the factor $1/\mu$ (μ is defined in (2.8)), so that the mean rate of change is $\mu = 1$. Therefore, we only estimate the relative rates (r_{ij}), to which we need to allocate prior distributions, and not the absolute rates (μr_{ij}). Additionally, this allows us to interpret the branch length as the expected number of substitution per site.

Zwickl and Holder [2004] discussed and tested different prior distributions on the relative rate parameters in the GTR model. The proposals were focused on two parameterizations of it: 5RR and ST1. The former approach sets one of the substitution rates to 1.0 arbitrarily, and allows the rest of the parameters to vary with respect to the one that is fixed. Usually, r_{TC} is the one chosen as reference. The latter forces the rates to sum to 1.0.

The authors found that a uniform prior on 5RR parametrization tends to overestimate the true value. In contrast, an exponential prior with a hyperparameter on its mean, that is,

$$r_{ij}|\phi \sim \text{Exp}(\phi) \quad \text{with} \quad \phi \sim \text{Exp}(1),$$

where

$$f(r_{ij}|\phi) = \phi e^{-r_{ij}\phi} \quad \text{and} \quad f(\phi) = e^{-\phi},$$

appears to perform well.

For the ST1 parameterization, the authors proposed a Dirichlet distribution. They tested a Dirichlet$(\alpha, \alpha, \alpha, \alpha, \alpha, \alpha)$, with $\alpha = 1.0$ and 0.5, and both distributions performed similarly well. The latter is the Jeffrey's prior when the likelihood is a multinomial distribution with 6 categories [Zwickl and Holder, 2004]. Both distributions have equal expectations of 0.17 for all the substitution rates, but they differ in their variance. For the former it is 0.07 and for the latter 0.17. This means that the first prior penalizes, in a greater degree, the models of evolution in which the rates are not similar.

For evolutionary models with fewer relative rate parameters alternative admissible prior distributions are the exponential with a hyperparameter or a gamma distribution [Wang and Yang, 2014].

2.4.4 Frequencies

The most common prior assigned on the nucleotide equilibrium frequencies is naturally a Dirichlet$(\alpha_A, \alpha_C, \alpha_G, \alpha_T)$ distribution. Usually, it is simply a Dirichlet(1,1,1,1) (see, for instance, Suchard et al. [2001]). A common practice, aiming to reduce the number of free parameters, is the use of empirical frequencies, which are based on a simple average. However, they might be quite different from the estimated ones and thus have different effects on the likelihood values [Yang, 1994a].

2.4.5 Heterogeneous substitution rate among sites

The most common method to introduce different heterogeneous substitution rates among sites is via a discrete gamma distribution [Yang, 1994b]. As was discussed previously, the gamma distribution used in this approach usually depends only on its shape parameter λ. Therefore, this parameter needs a prior distribution.

One common practice is to consider a Uniform$(0, m)$ distribution (see, for instance, Huelsenbeck et al. [2004]; Jow et al. [2002]). To the best of our knowledge, its impact has not been studied yet, in contrast to its performance as prior distribution for the branch length and relative rate parameters. Interestingly, it has been implemented in MrBayes with a default value $m = 200$, which indicates almost a homogeneous model.

The exponential distribution is perhaps the most common prior for this parameter. This prior, with mean 1.0, is used by default in MrBayes and BEAST. It seems more reasonable than the uniform since most estimates of λ from real datasets are small. In fact, less than 1.0 [Yang, 1996a].

A more flexible alternative is the gamma distribution which actually embodies the exponential prior. This better accommodates the inclusion of additional information into the model. Its density is given by

$$f(\lambda) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}e^{-\lambda/\beta}, \quad \text{with} \quad \lambda > 0,$$

where α and β are the shape and scale parameters, respectively.

2.4.6 Invariable sites

The invariable site parameter p_0 is the proportion of sites, i.e., it lies between 0 and 1. The most natural prior distribution for such parameters is the Beta distribution. It permits great flexibility, containing the Uniform(0,1) distribution as a special case, where $\alpha = \beta = 1$. The uniform is the most popular option in the phylogenetic literature [e.g., Holder et al., 2014] and the default choice in MrBayes. Yang [2014] pointed out that it would be a better option to consider a prior which accommodates the correlation between p_0 and the gamma shape parameter λ.

2.5 Summary

This chapter gives an overview of some of the main components and concepts involved in Bayesian phylogenetic analysis, which will be used throughout this thesis. This was carried out mainly by defining and discussing the parts which compose a phylogenetic model and the prior distributions for its parameters.

A phylogenetic model is actually composed of two elements: a tree model and a model of evolution. The tree model deals with the evolutionary relationship among the taxa. It is formally defined as a set of nodes and branches which connect the nodes without loops. This model describes the relatedness among the taxa through the branching pattern, called topology, and their distance via the branch lengths.

The direction of the evolutionary process on a phylogeny is granted by the root, which is the most recent common ancestor to all the taxa at the leaves. For the case in which this is specified, the tree is referred to as a rooted tree. However, many algorithms, including likelihood-based methods, produce unrooted trees, i.e., without specifying the root. This is due to the mathematically convenient assumption of reversibility on the model of evolution. In practice, these trees are rooted using different mechanisms, such as outgroup rooting, molecular clock rooting, mid-point rooting or Bayesian analysis.

The model of evolution explains the mutation process for a sequence site. It is usually defined by a Markov model which relies on 3 assumptions: memorylessness, homogeneity and stationarity. This model is characterized by its instantaneous transition rate matrix, which defines the rates of flow of the possible substitutions and consequently their probabilities. The different evolutionary models used in phylogenetics differ basically in the parametric structure of this matrix. Depending on it, the model adopts a distinctive name. A common practice is to assume a symmetrical structure for it, i.e., equal rates to describe the substitution flow from one state to another and for its inverse. This assumption generates a class of models called time reversible models. Some of them were discussed in this chapter for the particular case of DNA sequences, but their extension to other data is analogous.

The Markov process is applied from node to node along the phylogeny. This is carried out for each site of the sequences which are usually assumed to have evolved independently. To make it more flexible and in agreement with reality, the transition rates can be scaled by a random variable, usually following a gamma distribution, in order to allow the modeling of different rates across sites. In addition, a proportion of invariable sites can also be considered. Therefore, the substitution process defined on the phylogeny allows the calculation of the tree likelihood. We have presented this calculation in a four taxon case, which allowed to illustrate the methodology that can be easily extended to larger phylogenies. Additionally, we presented its calculation under the molecular clock assumption, i.e., when the lineages are assumed to have evolved at equal rates.

The models of evolution have also been extended to allow the modeling of more complex evolutionary processes. One aspect is the relaxation of the assumptions of homogeneity and reversibility. Allowing different Markov processes along the phylogeny and across segments of sites can deal efficiently with the modeling of groups of taxa with different evolutionary histories and sites with nonhomogeneous behaviors, respectively. The extensions of these models also include the relaxation of the independence among sites.

Both components of the phylogenetic model attempt to describe the data generating process through their parametric structure, which involves generally several parameters. In Bayesian statistics, these models require additional specifications about their parameters, before collecting and analysing the data. These specifications are incorporated into the model through prior distributions. If the phylogeny is unknown, it is treated as any other parameter of the model. We have critically discussed about the use of prior distributions in phylogenetics for their respective parameters.

Chapter 3

Estimating the marginal likelihood

3.1 Introduction

In phylogenetics, as in any other field, there are a great variety of models, aiming to describe the evolutionary processes that generate the data to be analysed. Hence, the natural and crucial question is: which of the available models offers the best description? Ideally, the chosen model should be as simple as possible, but without sacrificing the predictive power. It has to be noted that even though the inclusion of a parameter increases the fit, it also adds uncertainty to the estimates; and that simplistic models could lead to erroneous inferences. A Bayesian model selection criterion which takes these considerations into account is the Bayes factor (BF). This criterion performs model selection through the comparison of a quantity that is unique to each model, called the *marginal likelihood* (a component of the Bayes' theorem). This chapter principally concerns the estimation of the marginal likelihood and the Bayes factor.

Bayes' theorem, in a phylogenetic context, is given by

$$p(\boldsymbol{\theta}|\boldsymbol{X}, M, \tau) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M, \tau)\pi(\boldsymbol{\theta}|M, \tau)}{m(\boldsymbol{X}|M, \tau)}, \tag{3.1}$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $\boldsymbol{X}$ is the data, M is the substitution model,

$\tau \in \mathcal{T}$ is a given tree, $p(\boldsymbol{\theta}|\boldsymbol{X}, M, \tau)$ is the posterior probability distribution of $\boldsymbol{\theta}$, $L(\boldsymbol{X}|\boldsymbol{\theta}, M, \tau)$ is the likelihood function, $\pi(\boldsymbol{\theta}|M, \tau)$ is the prior distribution of $\boldsymbol{\theta}$, and $m(\boldsymbol{X}|M, \tau)$ is the marginal likelihood defined by

$$\int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M, \tau)\pi(\boldsymbol{\theta}|M, \tau)\mathrm{d}\boldsymbol{\theta}. \tag{3.2}$$

The parameter vector $\boldsymbol{\theta}$ contains all the model parameters, such as branch lengths, frequencies, gamma shape parameter and rate parameters. The tree topology τ is assumed to be fixed throughout this chapter. Thus, the parameter space Θ is continuous. Variable tree topology case is addressed in Chapter 4.

The prior distribution represents our previous knowledge of the parameters which is updated after taking into account the data and is reflected in the posterior distribution. The likelihood function represents the probability of the data given the parameters and the phylogenetic models (M, τ). The marginal likelihood is the probability of the data under the model and plays a key role in model selection. Indeed, this quantity is used to select among models. Because of this, it is also called *evidence* [MacKay, 2002]. To understand its role, note that the posterior distribution for the model M is given by

$$p(M|\boldsymbol{X}, \tau) = \frac{m(\boldsymbol{X}|M, \tau)\pi(M|\tau)}{m(\boldsymbol{X}|\tau)},$$

where $m(\boldsymbol{X}|M, \tau)$ is the marginal likelihood as defined in 3.2, $\pi(M|\tau)$ is the prior probability for the model, and $m(\boldsymbol{X}|\tau)$ is the probability of the data given the tree. The marginal likelihood will also be denoted by "z" henceforth. The assumed phylogeny τ will be omitted in the definition of the marginal likelihood estimation methods.

The Bayesian comparison of two models M_0 and M_1 can be carried out by comparing their posterior probabilities. This comparison is done through the ratio of their probabilities which represents the plausibility of one model over another and is defined as follows:

$$\frac{p(M_1|\boldsymbol{X}, \tau)}{p(M_0|\boldsymbol{X}, \tau)} = \frac{m(\boldsymbol{X}|M_1, \tau)}{m(\boldsymbol{X}|M_0, \tau)} \frac{\pi(M_1|\tau)}{\pi(M_0|\tau)},$$
$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}.$$

$2\log\left(\mathrm{BF}_{10}\right)$	BF_{10}	Evidence against M_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Table 3.1: Categories for Bayes factor interpretation taken from Kass and Raftery [1995].

The ratio of marginal likelihoods, the first ratio on the right side, is called the Bayes factor [Kass and Raftery, 1995]. If we have no preference for any model, i.e., each model is assigned the same prior probability, the priors are canceled and the posterior odds is only given by ratio of likelihoods, which are marginal likelihoods. A qualitative interpretation of this quantity is given in Table 3.1.

The Bayes factor (BF) is of particular interest as it provides many advantages over other methods of model selection: it allows comparison of nested and non-nested models, it is not based on a point estimate in parameter space since it averages over parameter space, and it embodies Occam's razor, i.e., it implicitly penalizes for parameter-rich models. Here lies its importance. Its properties have made it become a standard approach for performing model selection in a Bayesian phylogenetic framework. However, it is based on the marginal likelihood, a difficult multidimensional integral.

The marginal likelihood is often ignored at the inferential stage, but it plays a key role in model selection: it is a measure of the goodness of fit. In fact, it is the probability of the data given the model, i.e., it is by definition a measure of model fit. Thus, models with larger marginal likelihoods fit the data better and consequently are preferred over models with smaller marginal likelihoods. However, it is a difficult multidimensional integral of the prior distribution times the likelihood function over the parameter space. This quantity acts as the normalization constant in the posterior distribution, making it a probability density function. MCMC methods used for parameter estimation within a model use only ratios of posterior densities, and are therefore unable to measure its normalization in general.

Unlike maximum likelihood, which represents the model fit at a single point, this quantity stands for an average of how well the model fits the data. By being an average of the likelihood function with respect to the prior, the model with the greatest evidence might be different from the model with the highest likelihood because the prior could downweight some regions of parameter space. Also, the marginal likelihood is sensitive to the size of the region over which the likelihood is high. As a result, both methods could favour different models. Despite its important role in model selection, the marginal likelihood is usually analytically intractable and has to be approximated by numerical methods.

In phylogenetics, the marginal likelihood is in general not analytically available due to the complexity of the models. However, many techniques to estimate it have been studied and are available in computational packages (see Table 1.1). The most common ones are harmonic mean and steppingstone sampling. Many of the estimation methods have also been extended to estimate the Bayes factor directly. This alternative has the potential of leading to lower uncertainty [Baele et al., 2013].

This chapter focuses on the estimation of the marginal likelihood and the direct estimation of the Bayes factor. A subset of all available numerical methods is discussed including the uncertainty associated with them. The estimation process is carried out for a given topology. The direct Bayes factor estimate allows the direct comparison of two distinct phylogenies. The methodologies are applied in 4 scenarios. First, a statistical model that includes a phase transition is analysed. This phenomenon causes problems for many algorithms. The methods that reach success in this scenario are assessed in terms of their uncertainty. Second, these methods are assessed in a variant of the statistical model that even causes problems to those methods that use reference distributions to make more efficient the estimation process. Third, nested sampling (NS) is used to carry out model selection in a dataset, which contains 10 taxa. Then, estimation methods are assessed in their performance under different prior specifications. Finally, a simulation study in the four taxon case is carried out to assess direct Bayes factor estimation.

3.2 Importance sampling approaches

Importance sampling is a general technique to study the properties of a particular distribution based on samples of a reference distribution. This technique uses a trial density which should be easier to sample from than the original one for estimating the expectation of a function. Thus, it can be used in the context of marginal likelihood estimation, which can be seen as computing an expected value.

Noting this, the marginal likelihood defined in (3.2) can be expressed equivalently as

$$\begin{aligned} z &= \int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M)\frac{\pi(\boldsymbol{\theta}|M)}{g(\boldsymbol{\theta})}g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_g\big[L(\boldsymbol{X}|\boldsymbol{\theta}, M)w(\boldsymbol{\theta})\big], \end{aligned}$$

where $g(\boldsymbol{\theta})$ is the importance sampling distribution, $\mathbb{E}_g[\cdot]$ denotes the expectation with respect to g, and $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|M)/g(\boldsymbol{\theta})$ is the importance weight. The weights calibrate the estimation due to the fact that the sampling is from another distribution.

For a completely known trial density $g(\boldsymbol{\theta})$, i.e., $\int_{\Theta} g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$, the Monte Carlo estimator is given by

$$\widehat{z} = \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)w(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ are drawn from $g(\boldsymbol{\theta})$.

More generally, for an unnormalized $g(\boldsymbol{\theta})$, the marginal likelihood can be written as

$$\begin{aligned} z &= \frac{\int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M)w(\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} w(\boldsymbol{\theta})g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{\mathbb{E}_g\big[L(\boldsymbol{X}|\boldsymbol{\theta}, M)w(\boldsymbol{\theta})\big]}{\mathbb{E}_g\big[w(\boldsymbol{\theta})\big]}. \end{aligned} \tag{3.3}$$

Its corresponding Monte Carlo estimator is

$$\widehat{z} = \frac{\sum_{i=1}^{n} L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)w(\boldsymbol{\theta}_i)}{\sum_{i=1}^{n} w(\boldsymbol{\theta}_i)}.$$

The performance of the method relies directly on the choice of the importance function. First, the accuracy of $\widehat{z}$ depends on the variability of the importance weights, and these depend directly on the relationship between the target and importance distributions. If the latter emphasizes points that have more impact on the marginal likelihood, the error can be reduced. And second, the importance function can speed up as well as slow down the estimation process. A poor choice could require a huge number of samples. In general, the importance sampling function should have heavier tails than the target distribution, the posterior in this case, to get reliable estimates [MacKay, 2002].

3.2.1 Arithmetic mean

If the prior is used as the importance distribution, the method is usually known as *arithmetic mean* [AM; Lartillot and Philippe, 2006; Xie et al., 2011]. It is worth mentioning that the mean of the likelihoods of parameter values sampled from the posterior [Aitkin, 1991] is also sometimes called arithmetic mean [Baele and Lemey, 2014]. We refer to the former as the arithmetic mean in this thesis.

AM produces an unbiased estimate of the marginal likelihood. However, the region of high likelihood is usually very concentrated making the method require an enormous sampling effort to estimate the marginal likelihood. In practice, the sample drawn from the prior is unlikely to contain enough points from this area, leading to poor estimates [Lartillot and Philippe, 2006]. As a result, this method is rarely used, especially in phylogenetics.

3.2.2 Harmonic mean

Harmonic mean [HM; Newton and Raftery, 1994] is the most common and simplest method to estimate the marginal likelihood and has been used extensively in phylogenetics [see for instance Jia et al., 2014; Nylander et al., 2004; Pagel et al., 2004]. Its popularity can be explained by the simplicity of its estimation and that it only requires samples from the posterior, which can be recycled from a standard Bayesian phylogenetic analyses. Therefore, its calculation does not entail a significant extra cost. Also, this method has been implemented in many phylogenetic software packages, most notably, Mr BAYES [Huelsenbeck and Ronquist, 2001] and BEAST [Drummond and Rambaut, 2007]. However, it is well-known that it suffers from several disadvantages, which will be discussed below.

HM utilizes the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, M)$ as importance sampling distribution $g(\boldsymbol{\theta})$ in equation (3.3). This yields the identity

$$z = \mathbb{E}_p\left[\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}\right]^{-1},$$

where $\mathbb{E}_p[\cdot]$ is the expectation over the posterior p. Equivalently, this result can be obtained through the harmonic mean identity [Raftery et al., 2007]. Its Monte Carlo approximation is given by

$$\widehat{z} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}\right)^{-1} \tag{3.4}$$

where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ are drawn from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, M)$.

An approximation of its variance can be derived using the delta method [Oehlert, 1992]. For this, consider $W = 1/z$ and $f(W) = 1/W$. The method states that the variance of the latter transformation is given by $\text{Var}[f(W)] \approx (f'(\mathbb{E}[W]))^2 \times \text{Var}[W]$, where $f'(\mathbb{E}[W])$ is the first derivative of f with respect to $\mathbb{E}[W]$. Thus, the variance for the HM estimate has the form of $\text{Var}[\widehat{z}] = \text{Var}[f(\widehat{W})]$. The expected value of each value $L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)^{-1}$ is given by

$$\mathbb{E}_p[L(\boldsymbol{X}|\boldsymbol{\theta}, M)^{-1}] = \int_\Theta \frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)}\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{z}\mathrm{d}\boldsymbol{\theta} = \frac{1}{z}.$$

and consequently, the expected value of W has the following form

$$\begin{aligned}\mathbb{E}[\widehat{W}] = \mathbb{E}\left[\frac{1}{\widehat{z}}\right] &= \int_\Theta \frac{1}{n}\sum_{i=1}^{n}\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}p(\boldsymbol{\theta}|\boldsymbol{X}, M)\mathrm{d}\boldsymbol{\theta}\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_\Theta \frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}p(\boldsymbol{\theta}|\boldsymbol{X}, M)\mathrm{d}\boldsymbol{\theta}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[L(\boldsymbol{X}|\boldsymbol{\theta}, M)^{-1}] \\
&= \frac{1}{z}.
\end{aligned}
$$

Hence, assuming that the points have been independently sampled, the variance can be approximated by

$$
\begin{aligned}
\operatorname{Var}[\widehat{z}] &\approx z^4 \cdot \operatorname{Var}\left[\frac{1}{\widehat{z}}\right] \\
&= z^4 \cdot \operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}\right] \\
&= \frac{z^4}{n} \cdot \operatorname{Var}\left[\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)}\right] \\
&= \frac{z^4}{n}\left(\mathbb{E}_p\left[\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^2}\right] - \mathbb{E}_p\left[\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)}\right]^2\right) \\
&= \frac{z^4}{n}\left(\frac{1}{z}\int_{\Theta}\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)}\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} - \frac{1}{z^2}\right). \qquad (3.5)
\end{aligned}
$$

However, the integral in this expression might not be finite because the likelihood in general takes small values with respect to the prior. This would not happen if the likelihood were more diffuse than the prior, a condition that is rarely met in practice.

Not only does the HM estimator often have infinite variance, but has several other drawbacks, which have been widely documented in the literature [Baele, Lemey, Bedford, Rambaut, Suchard, and Alekseyenko, 2012; Baele, Lemey, and Suchard, 2016; Baele, Lemey, and Vansteelandt, 2013; Lartillot and Philippe, 2006; Newton and Raftery, 1994; Xie, Lewis, Fan, Kuo, and Chen, 2011]. One of them can be noticed directly from the structure of its estimate defined in (3.4), which reveals that the estimate is extremely sensitive to points with small likelihood values. This is a potential cause of instability in the estimate.

Another drawback is its lack of sensitivity to prior distributions. The HM just takes into account the prior knowledge through the posterior samples, but usually, the posterior is much narrower than the prior distribution. Consequently, the prior distribution is not well represented in the posterior sample, leading to the estimate being relatively insensitive to prior specifications. The consequences of this feature of the HM can be noticed considering the following example. In the case of quite informative data, two Bayesian models that differ in their priors might have similar HM estimates. This makes the method fail to fairly penalize the complexity of a model [Xie et al., 2011], which is induced by the prior. In a model selection context, the capacity of penalizing unnecessary complexity is crucial and HM might not provide the means to include it.

Another consequence of the usual relationship between the posterior and the prior (the former is much more constrained than the latter) is that the HM tends to overestimate the marginal likelihood. This phenomenon is due to underrepresentation of those areas of low likelihood in the posterior sample and is exacerbated as the model becomes higher in dimension. Thus, the method can fail in high dimensional problems, a common situation in phylogenetics.

Finally, the last shortcoming listed in this thesis is related to the sample size required by the HM method. Even though the law of large numbers guarantees that the estimate is consistent, in practice, the number of required samples to obtain an estimate close to the true value is non-viable. This is caused by the almost null representation in the posterior sample of those areas of low likelihood values. Even a posterior sample of an astronomical size would not be sufficient for the method to work [Lartillot and Philippe, 2006].

To sum up, HM is poor in accuracy, has an unstable behavior, is insensitive to prior specifications, overestimates the true value, and may not work in high dimensional models. Therefore, since it yields unreliable estimates, it is not recommended. In his blog, Neal [2008] has described HM as the "worst Monte Carlo method ever". Many of these drawbacks discussed above are illustrated in the Application section 3.5.

3.2.2.1 Direct Bayes factor estimation

It is possible to estimate directly the Bayes factor based on the HM structure. For this, consider two models M_0 and M_1 to be compared, with marginal likelihoods z_0 and z_1, respectively. Also, $\boldsymbol{\theta}_0 \in \Theta_0$ and $\boldsymbol{\theta}_1 \in \Theta_1$ are their respective parameter vectors which are merged into $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \in \Theta$. For instance, if $\boldsymbol{\theta}_0 = (a, b, c)$ and $\boldsymbol{\theta}_1 = (c, d)$, then $\boldsymbol{\theta} = (a, b, c, d)$. This is the most general case, when the models are defined on different parameter spaces, but the methodology is also valid for the particular case when the parameter space is the same for both models.

The Bayes factor can be expanded as follows

$$\begin{aligned} \mathrm{BF}_{10} = \frac{z_1}{z_0} &= \frac{\int_{\Theta_1} L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1|M_1)\mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_0} L(\boldsymbol{X}|\boldsymbol{\theta}_0, M_0)\pi(\boldsymbol{\theta}_0|M_0)\mathrm{d}\boldsymbol{\theta}_0} \\ &= \frac{\int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\pi(\boldsymbol{\theta}|M_0)\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta}|M_0)\pi(\boldsymbol{\theta}|M_1)\mathrm{d}\boldsymbol{\theta}}. \end{aligned}$$

Note that the integrals in the numerator and denominator in the last identities are equivalent since the prior distributions must be proper. The subscripts in $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ have been omitted since $\boldsymbol{\theta}$ is the common parameter vector. The parameters in $\boldsymbol{\theta}$ which do not correspond to the model have no participation on the computation of the likelihood and prior functions. We define an instrumental function

$$g_1(\boldsymbol{\theta}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1|M_1)}{z_1}\pi(\boldsymbol{\theta}_0|M_0),$$

which is a probability density function on Θ, that is, $g_1(\boldsymbol{\theta}) > 0$ and $\int_{\Theta} g_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = 1$. Equivalently, it could be defined based on the posterior of model 0 and the prior distribution of model 1. This density is used as an importance sampling function in the Bayes factor yielding, after some calculations, the following

$$\begin{aligned} \mathrm{BF}_{10} &= \frac{\int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\pi(\boldsymbol{\theta}|M_0)}{g_1(\boldsymbol{\theta})} g_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta}|M_0)\pi(\boldsymbol{\theta}|M_1)}{g_1(\boldsymbol{\theta})} g_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{1}{\int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)} g_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\ &= \mathbb{E}_{g_1}\left[\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}\right)^{-1}\right]^{-1}. \end{aligned}$$

This procedure can be seen as the identity studied by Meng and Wong [1996].

An estimator of Bayes factor is

$$\widehat{\mathrm{BF}}_{10} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_1)} \right)^{-1},$$

constructed using samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from $g_1(\boldsymbol{\theta})$. This function can be seen as a pseudo-posterior distribution and $L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)/L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)$ as a pseudo-likelihood function from a HM perspective. This estimate is also valid when both models are defined on the same parameter space. In the particular case that both models have the same priors, the auxiliary function is given by

$$g_1(\boldsymbol{\theta}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta})}{z_1}.$$

An approximate of the variance of this estimator can be obtained using the delta method [Oehlert, 1992] similarly to the approximated variance for the HM case presented before. Now, consider $W = \mathrm{BF}_{10}^{-1} = z_0/z_1$ and $f(W) = 1/W = \mathrm{BF}_{10}$. The expected value of the inversed pseudo-likelihood function is given by

$$\begin{aligned} \mathbb{E}_{g_1}\left[\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)} \right] &= \int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)}{z_1} \pi(\boldsymbol{\theta}|M_0) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{z_0}{z_1}, \end{aligned}$$

and the expected value of $\widehat{W}$ assumes the form

$$\begin{aligned} \mathbb{E}_{g_1}[\widehat{W}] &= \int_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_1)} g_1(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{n} \sum_{i=1}^{n} \int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_1)} g_1(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{g_1}\left[\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)} \right] \\ &= \frac{z_0}{z_1}. \end{aligned}$$

Thus, an estimate for the variance is given by the following derivation

$$\begin{aligned}
\mathrm{Var}[\widehat{\mathrm{BF}}_{10}] &= \mathrm{Var}[f(\widehat{W})] \\
&\approx f'(\mathbb{E}[\widehat{W}])^2 \cdot \mathrm{Var}[\widehat{W}] \\
&= \left(\frac{z_1}{z_0}\right)^4 \cdot \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M_1)}\right] \\
&= \left(\frac{z_1}{z_0}\right)^4 \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}\right] \\
&= \left(\frac{z_1}{z_0}\right)^4 \frac{1}{n}\mathrm{Var}\left[\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}\right] \\
&= \left(\frac{z_1}{z_0}\right)^4 \frac{1}{n}\left\{\mathbb{E}_{g_1}\left[\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}\right)^2\right] - \mathbb{E}_{g_1}\left[\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}\right]^2\right\} \\
&= \left(\frac{z_1}{z_0}\right)^4 \frac{1}{n}\left\{\int_{\Theta}\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}\right)^2 g_1(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \left(\frac{z_0}{z_1}\right)^2\right\} \\
&= \left(\frac{z_1}{z_0}\right)^4 \frac{1}{n}\left\{\frac{1}{z_1}\int_{\Theta_0} L(\boldsymbol{X}|\boldsymbol{\theta}_0, M_0)^2\pi(\boldsymbol{\theta}|M_0)\mathrm{d}\boldsymbol{\theta}_0 \times \right. \\
&\qquad\qquad \left. \int_{\Theta_1}\frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)}\pi(\boldsymbol{\theta}|M_1)\mathrm{d}\boldsymbol{\theta}_1 - \left(\frac{z_0}{z_1}\right)^2\right\}.
\end{aligned}$$

These calculations are valid if and only if the points are truly independent from the posterior distribution.

Unfortunately, this variance may not be finite. The integral that embodies the inversed likelihood for model 1 multiplied by the prior could be infinite, unless the likelihood was more diffuse than the prior. In general, however, this does not happen.

The variance of the estimation of the BF via two independent HM estimates involves two potential sources of infinity. This is because the approximation of this quantity is equal to the sum of the two variances defined in (3.5), each one being potentially infinite. On the other hand, the direct BF estimation via HM involves only one potential source of infinity. Hence, this method could be a valid alternative if, at most, one of the estimated variances is infinite. In this case, the one with estimated infinite variance should not be allocated in model 1.

3.3 Path sampling approaches

Path sampling methods are based on a series of transitional distributions which connect the prior with the posterior, or in their generalized form, a reference distribution to the posterior. These distributions form a path which serves to refine the marginal likelihood estimate. A particular case is bridge sampling, which only uses a single transitional function. In general, these methods yield reliable estimates and are far more stable and accurate than importance sampling methods. Their positive attributes have allowed them to be implemented in different popular software packages such as BEAST [Drummond and Rambaut, 2007], gaining popularity.

3.3.1 Path sampling

Path sampling (PS) or *thermodynamic integration* (TI) is a natural generalization of importance sampling [Gelman and Meng, 1998]. It was introduced into phylogenetics by Lartillot and Philippe [2006], although it was originally proposed by Gelman and Meng [1998] to estimate the ratio of marginal likelihoods, but its connections could be traced back further [Ogata, 1989]. Also, it was proposed independently by Friel and Pettitt [2008], who called it *method of power posteriors*. This method is far more accurate than importance sampling methods, but it requires a much higher computational cost. However, it provides more reliable estimates across diverse scenarios, most notably, in high dimensional problems, which is a typical situation in phylogenetics.

Consider two unnormalized density functions $q_0(\boldsymbol{\theta})$ and $q_1(\boldsymbol{\theta})$ defined on the same parameter space Θ. The full probability densities are given by

$$p_j(\boldsymbol{\theta}) = \frac{q_j(\boldsymbol{\theta})}{z_j}, \quad j = 0, 1.$$

In a Bayesian context, $z_j = \int_\Theta q_j(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is the normalizing constant or marginal likelihood, $q_j(\boldsymbol{\theta}) = L(\boldsymbol{X}|\boldsymbol{\theta}, M_j)\pi(\boldsymbol{\theta}|M_j)$ and $p_j(\boldsymbol{\theta}) = p_j(\boldsymbol{\theta}|\boldsymbol{X}, M_j)$ is the posterior.

In order to perform a numerical evaluation of the log-ratio $\log(z_1/z_0)$, it is always possible to construct a continuous and differentiable path $\big(q_\beta(\boldsymbol{\theta})\big)_{0\leq\beta\leq1}$ to link the unnormalized functions. This concept defines a set of transitional distributions which form the path as

$$p_\beta(\boldsymbol{\theta}) = \frac{q_\beta(\boldsymbol{\theta})}{z_\beta}, \quad 0 \leq \beta \leq 1,$$

where $z_\beta = \int_\Theta q_\beta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. In a marginal likelihood estimation context, the most popular is the geometric path, also known as *power posterior density*, and is defined by

$$p_\beta(\boldsymbol{\theta}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta}. \tag{3.6}$$

Note that for $\beta = 0$ the power posterior is equivalent to the prior distribution and for $\beta = 1$ is equivalent to the posterior distribution. Their corresponding normalizing constants are $z_0 = 1$ and $z_1 = z$, respectively. Figure 3.1 displays how the geometric path between a Uniform$(-6, 6)$ and N$(0, 1)$ distribution, as prior and posterior respectively, looks like for 6 transitional distributions with β values uniformly spaced. This is assuming a likelihood N$(\mu, 1)$ with a single observation $x = 0$.

Assuming the legitimacy of interchange of integration with differentiation, the basic identity for PS is given by

$$\begin{aligned}\frac{\partial \log z_\beta}{\partial\beta} &= \mathrm{E}_{p_\beta}\left[\frac{\partial \log q_\beta(\boldsymbol{\theta})}{\partial\beta}\right] \\ &= \mathrm{E}_{p_\beta}\big[U(\boldsymbol{\theta}, \beta)\big],\end{aligned} \tag{3.7}$$

where E_{p_β} depicts the expectation with respect to the sampling distribution $p_\beta(\boldsymbol{\theta})$ and $U(\boldsymbol{\theta}, \beta) = \frac{\partial}{\partial\beta}\log q_\beta(\boldsymbol{\theta})$. For the power posterior case, we have that $U(\boldsymbol{\theta}, \beta) = U(\boldsymbol{\theta}) = \log L(\boldsymbol{X}|\boldsymbol{\theta}, M)$. This function no longer depends on β.

The aim is to compute the ratio of normalizing constants, which under the

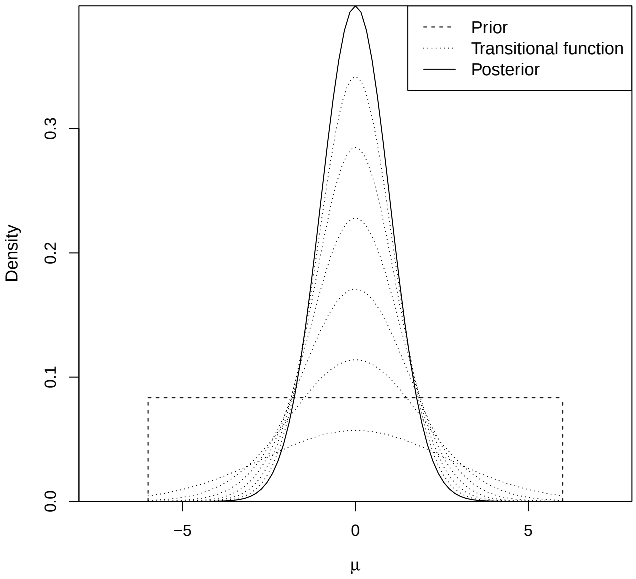


Figure 3.1: Geometric path between a prior Uniform$(-6, 6)$ and posterior N$(0, 1)$. The likelihood is N$(\mu, 1)$ with one observation $x = 0$.

log-transformation, is given by

$$\log z = \log\left(\frac{z_1}{z_0}\right) = \log z_1 - \log z_0.$$

Considering the previous results, this log-ratio can be expressed equivalently as

$$\log z = \int_0^1 \frac{\partial \log z_\beta}{\partial \beta} \mathrm{d}\beta = \int_0^1 \mathrm{E}_{p_\beta}\big[U(\boldsymbol{\theta}, \beta)\big] \mathrm{d}\beta. \tag{3.8}$$

PS relies on this integral to calculate the marginal likelihood. Its key idea is that for any value β between 0 and 1, a Markov chain Monte Carlo (MCMC) can be run to approximate the expected value by using q_β as sampling distribution. This process produces a sample from p_β. Thus, $\mathrm{E}_{p_\beta}\big[U(\boldsymbol{\theta}, \beta)\big]$ can be approximated taking the average over this sample as

$$\widehat{U}(\boldsymbol{\theta}_\beta) = \frac{1}{n}\sum_{i=1}^{n} U(\boldsymbol{\theta}_\beta^i, \beta), \tag{3.9}$$

where $\boldsymbol{\theta}_\beta \sim p_\beta$ and $\boldsymbol{\theta}_\beta^i$ is a sample point from it for $i = 1, \ldots, n$. This procedure can be done for a series of $K + 1$ values of β and then the expectation for each sample can be calculated. The generated set of expectations is used to approximate the integral defined in (3.8).

The discretization of the interval $[0, 1]$, domain of β, implies sampling different power posterior densities induced by β_k, with $k = 0, \ldots, K$. In particular, we have that $\beta_0 = 0$ and $\beta_K = 1$. These samples can be obtained by the *quasistatic* method which yields a great accuracy [Lartillot and Philippe, 2006]. The method consists of equilibrating a MCMC under $\beta = 0$, then smoothly increasing the value of β, by adding a constant $\Delta\beta$, until $\beta = 1$ is reached. Along the process, a certain number (N) of points $\boldsymbol{\theta}_\beta$ are stored before each update of β. Using the trapezoidal rule, an estimate of $\log z$ is given by

$$\begin{aligned} \widehat{\log z} &= \sum_{k=0}^{K-1} (\beta_{k+1} - \beta_k) \frac{\widehat{U}(\boldsymbol{\theta}_{\beta_{k+1}}) + \widehat{U}(\boldsymbol{\theta}_{\beta_k})}{2} \qquad (3.10) \\ &= \frac{1}{2} \left\{ \beta_1 \widehat{U}(\boldsymbol{\theta}_{\beta_0}) + \sum_{k=1}^{K-1} (\beta_{k+1} - \beta_{k-1}) \widehat{U}(\boldsymbol{\theta}_{\beta_k}) + (\beta_K - \beta_{K-1}) \widehat{U}(\boldsymbol{\theta}_{\beta_K}) \right\}, \end{aligned}$$

for $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$. The procedure of moving from $\beta = 0$ to $\beta = 1$ is known as *annealing* scheme. Equivalently, one could start from $\beta = 1$ and progressively decrease to $\beta = 0$. This procedure is known as *melting* scheme.

Lartillot and Philippe [2006] proposed to spread the β values regularly spaced between 0 and 1. But since often most of the variability of the expected values is concentrated for β near 0, some authors have proposed to place more computational effort in that place. This phenomenon is illustrated in Figure 3.2 where can be noticed that most of the $\widehat{U}$ variability is concentrated for $\beta < 0.2$. In this context, Lepage et al. [2007] used a sigmoidal function to estimate the Bayes factor; Friel and Pettitt [2008] proposed $\beta_k = x_k^4$, where x-values are equally spaced between 0 and 1; and Xie et al. [2011] advocated spreading the values according to evenly spaced quantiles of a Beta$(\alpha, 1)$, with $\alpha = 0.3$, and showed that the efficiency of PS can be dramatically improved. This approach is displayed in Figure 3.2.

The precision of PS is directly proportional to the number of transitional dis-

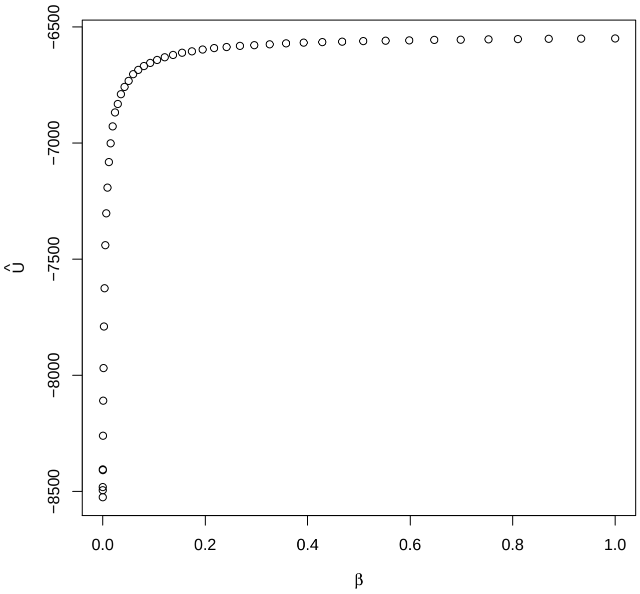


Figure 3.2: Estimated expectations $\widehat{U}$ over p_β for different β values which together makes up the PS estimate. These estimates have been taken from the data analysed in example 3.5.2.1.

tributions connecting the prior with the posterior, in other words, as K increases, the accuracy increases. This also reduces the bias introduced by the numerical approximation. This bias can also be reduced by using more accurate numerical integration techniques.

The Monte Carlo standard deviation is given by the square root of the variance given by

$$\mathrm{Var}\left[\widehat{\log z}\right] = \frac{1}{4}\Bigg\{\beta_1^2\mathrm{Var}\left[\widehat{U}(\boldsymbol{\theta}_{\beta_0})\right] + \sum_{k=1}^{K-1}(\beta_{k+1}-\beta_{k-1})^2\mathrm{Var}\left[\widehat{U}(\boldsymbol{\theta}_{\beta_k})\right] + \\ (\beta_K-\beta_{K-1})^2\mathrm{Var}\left[\widehat{U}(\boldsymbol{\theta}_{\beta_K})\right]\Bigg\}$$

where the variances are defined as

$$\mathrm{Var}\left[\widehat{U}(\boldsymbol{\theta}_\beta)\right] = \frac{1}{n}\mathrm{Var}\left[U(\boldsymbol{\theta},\beta)\right]$$

$$= \frac{1}{n}\left(\mathbb{E}_{p_\beta}\left[U(\boldsymbol{\theta},\beta)^2\right] - \mathbb{E}_{p_\beta}\left[U(\boldsymbol{\theta},\beta)\right]^2\right).$$

The first expectation can be approximated by

$$\mathbb{E}_{p_\beta}\left[U(\boldsymbol{\theta},\beta)^2\right] \approx \frac{1}{n}\sum_{i=1}^{n} U(\boldsymbol{\theta}_\beta^i,\beta)^2 = \frac{1}{n}\sum_{i=1}^{n} \log L(\boldsymbol{X}|\boldsymbol{\theta}_\beta^i, M)^2$$

where $\boldsymbol{\theta}_\beta^i$ are sampling points from p_β. The last identity corresponds to the power posterior case. The approximation for the second expectation is given in (3.9).

The total uncertainty estimation has two sources of error: the sampling error, explained above, and the error induced by the discretization of the integral defined in (3.8). In general, the later is ignored since it diminishes proportionally as the number of transitional distributions increases.

PS performance can be improved by shortening the path between the distributions. This can be done by replacing the prior by a reference distribution which should optimally approximate the posterior [Lefebvre et al., 2010]. This extension has the potential of improving its performance significantly [Arima and Tardella, 2014]. The same concept is used by generalized steppingstone sampling which is discussed in 3.3.3.

3.3.1.1 Direct Bayes factor estimation

When the difference between the logarithm of the marginal likelihoods of two models is small compared to these two values, the Bayes factor estimate could be poor [Lartillot and Philippe, 2006]. In this situation, it would be preferable to estimate the difference directly, unless the precision of each marginal likelihood estimate is very high. So, instead of defining a path between the prior and the posterior, it could be better to define a path directly between the two models. This scheme is known as *model-switch* path sampling or thermodynamic integration.

Consider the two models M_0 and M_1 defined on the parameter space Θ_0 and Θ_1, respectively. Also, consider $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \in \Theta$, where $\boldsymbol{\theta}_0 \in \Theta_0$ and $\boldsymbol{\theta}_1 \in \Theta_1$. These two latter parameter vectors have been merged into one, as defined in Section 3.2.2.1. This is the most general case discussed by Lartillot and

Philippe when both models (or part of them) are defined on different parameter spaces. Analogously to the power posterior defined previously in (3.6), now the unnormalized transitional distributions q_β and the normalized density p_β are given by

$$\begin{aligned} q_\beta(\boldsymbol{\theta}) &= \big(L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta}|M_0)\big)^{1-\beta}\big(L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\big)^{\beta} \\ p_\beta &= q_\beta/z_\beta \end{aligned}$$

with $0 \leq \beta \leq 1$, and normalizing constant z_β. Note that when $\beta = 0$, this function represents the case of model 0 and when $\beta = 1$ the case of model 1. Thus, this power posterior density bridges the two models. Under this scheme, for U in (3.7) it then holds that

$$U(\boldsymbol{\theta}, \beta) = \log L(\boldsymbol{X}|\boldsymbol{\theta}, M_1) + \log \pi(\boldsymbol{\theta}|M_1) - \log L(\boldsymbol{X}|\boldsymbol{\theta}, M_0) - \log \pi(\boldsymbol{\theta}|M_0)$$

Once this quantity and the power posterior are defined, the log-Bayes factor estimate is equivalent to that one described in (3.10).

3.3.2 Steppingstone sampling

Steppingstone sampling [SS; Xie et al., 2011] is a method to estimate the marginal likelihood which gathers ideas of importance sampling and path sampling. This method uses importance sampling method to estimate each element of a telescope product of ratios of normalizing constants of the transitional distributions. This approach has the advantage of requiring fewer path steps than PS to estimate accurately the marginal likelihood and yielding a less-biased estimator. Since its publication, it has been implemented in many phylogenetic software packages (see Table 1.1).

Consider the power posterior density defined in (3.6). The marginal likelihood can be expanded as a product of K ratios of normalizing constants of their transitional distributions as follows

$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}}\frac{z_{\beta_2}}{z_{\beta_1}}\cdots\frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}}\frac{z_{\beta_K}}{z_{\beta_{K-1}}} = \prod_{k=1}^{K}\frac{z_{\beta_k}}{z_{\beta_{k-1}}} = \prod_{k=1}^{K} r_k, \tag{3.11}$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$ and $r_k = z_{\beta_k}/z_{\beta_{k-1}}$.

The direct estimate of this ratio of marginal likelihoods z_1/z_0 is not easy because the distributions involved in the numerator and denominator (posterior and prior, respectively) are, in general, quite different. To solve this problem, SS expands this ratio in a series of ratios. These ratios are individually easier to estimate because the involved distributions on the top and bottom are now quite similar. In this situation importance sampling method works well.

SS estimates each ratio r_k by importance sampling using $p_{\beta_{k-1}}$ as importance sampling distribution. This is a suitable distribution because it has heavier tails than p_{β_k} which leads to an efficient estimate of r_k. In this manner, it avoids estimating from the posterior distribution, making it slightly less expensive computationally than PS for the same number K of path steps. Each ratio is estimated based on the identity

$$\begin{aligned} r_k = \frac{z_{\beta_k}}{z_{\beta_{k-1}}} &= \int_\Theta \frac{q_{\beta_k}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})} p_{\beta_{k-1}}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_{p_{\beta_{k-1}}}\left[\frac{q_{\beta_k}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})}\right]. \end{aligned} \tag{3.12}$$

Thus, an unbiased MC estimator of r_k is given by

$$\begin{aligned} \widehat{r}_k &= \frac{1}{n}\sum_{i=1}^{n} \frac{q_{\beta_k}(\boldsymbol{\theta}^i_{\beta_{k-1}})}{q_{\beta_{k-1}}(\boldsymbol{\theta}^i_{\beta_{k-1}})} \\ &= \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k-\beta_{k-1}}, \end{aligned}$$

where $\boldsymbol{\theta}^1_{\beta_{k-1}}, \ldots, \boldsymbol{\theta}^n_{\beta_{k-1}}$ are drawn from $p_{\beta_{k-1}}$ and $k = 1, \ldots, K$. To improve its numerical stability, it has been proposed to factorize by the largest sampled likelihood $L^k_{\max}$. Thus, an estimate of z is given by

$$\widehat{z} = \prod_{k}^{K} \widehat{r}_k = \prod_{k=1}^{K} \frac{1}{N}\left(L^k_{\max}\right)^{\beta_k-\beta_{k-1}} \sum_{i=1}^{n}\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)}{L^k_{\max}}\right)^{\beta_k-\beta_{k-1}}.$$

In a log scale, this could be written as

$$\begin{aligned}\log \widehat{z} =& - K\log N + \sum_{k=1}^{K}(\beta_k - \beta_{k-1})\log L^k_{\max} \\ &+ \sum_{k=1}^{K}\log\sum_{i=1}^{n}\exp\left\{(\beta_k - \beta_{k-1})(\log L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M) - \log L^k_{\max})\right\}.\end{aligned}$$

Although $\widehat{z}$ is unbiased, the log transformation introduces a bias which can be alleviated by increasing K.

The MC sampling error of each ratio $\widehat{r}_k$ is given by $\mathrm{SD}[\widehat{r}_k] = \mathrm{Var}[\widehat{r}_k]^{1/2}$ with estimated variance

$$\begin{aligned}\mathrm{Var}[\widehat{r}_k] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{q_{\beta_k}(\boldsymbol{\theta}^i_{\beta_{k-1}})}{q_{\beta_{k-1}}(\boldsymbol{\theta}^i_{\beta_{k-1}})}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\frac{q_{\beta_k}(\boldsymbol{\theta}^i_{\beta_{k-1}})}{q_{\beta_{k-1}}(\boldsymbol{\theta}^i_{\beta_{k-1}})}\right] \\ &= \frac{1}{n}\mathrm{Var}\left[\frac{q_{\beta_k}(\boldsymbol{\theta})}{q_{\beta_{k-1}}(\boldsymbol{\theta})}\right] \\ &\approx \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{q_{\beta_k}(\boldsymbol{\theta}^i_{\beta_{k-1}})}{q_{\beta_{k-1}}(\boldsymbol{\theta}^i_{\beta_{k-1}})} - \widehat{r}_k\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k-\beta_{k-1}} - \widehat{r}_k\right)^2,\end{aligned}$$

as $\widehat{r}_k$ is an unbiased estimator of r_k. Hence, the standard deviation for the SS estimate is given by $\mathrm{SD}[\widehat{z}] = \mathrm{Var}[\widehat{z}]^{1/2}$, where

$$\begin{aligned}\mathrm{Var}[\widehat{z}] &= \mathrm{Var}\left[\prod_{k=1}^{K}\widehat{r}_k\right] \\ &= \prod_{k=1}^{K}\left(\mathrm{Var}[\widehat{r}_k] + \mathbb{E}[\widehat{r}_k]^2\right) - \prod_{k=1}^{K}\mathbb{E}[\widehat{r}_k]^2 \\ &\approx \prod_{k=1}^{K}\left(\mathrm{Var}[\widehat{r}_k] + \widehat{r}_k^2\right) - \prod_{k=1}^{K}\widehat{r}_k^2\end{aligned}$$

$$= \prod_{k=1}^{K} \left(\mathrm{Var}[\widehat{r}_k] + \widehat{r}_k^2\right) - \widehat{z}^2.$$

Based on the delta method, the sampling variance of the estimated log-marginal likelihood can be approximated as

$$\begin{aligned}\mathrm{Var}[\log \widehat{z}] &\approx \sum_{k=1}^{K} \frac{1}{\widehat{r}_k^2} \mathrm{Var}[\widehat{r}_k] \\ &= \frac{1}{n^2} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k - \beta_{k-1}}}{\widehat{r}_k} - 1 \right)^2\end{aligned}$$

All this is valid only if the samples are truly independent from their respective densities p_β.

SS possesses important characteristics which have allowed it to gain popularity in the phylogenetic community. The most important advantage is that it requires a lower number of transitional distributions than PS to yield a reliable estimate. Furthermore, SS seems to be less sensitive to the choice of β values than PS. For a moderately positively skewed distribution of β, there is no significant difference in root mean square error (RMSE) between the methods. However, SS outperforms PS significantly when the β values tend to be equally spaced [Xie, Lewis, Fan, Kuo, and Chen, 2011].

3.3.2.1 Direct Bayes factor estimation

Similarly to model-switch PS defined in Section 3.3.1.1, SS can be adapted to directly estimate the Bayes factor [Baele et al., 2013]. In general, the model-switch SS estimate has a lower variance than the one obtained via the ratio of two independent SS estimates, especially if the two priors are the same. This method is carried out by defining a path that directly connects the two models in the space of unnormalized densities. The two competing models can be defined on different parameter spaces. Again, the path can be geometrically constructed as

$$q_\beta(\boldsymbol{\theta}) = \big(L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta}|M_0)\big)^{1-\beta}\big(L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\big)^{\beta},$$

where $0 \leq \beta \leq 1$ and $\boldsymbol{\theta} \in \Theta$. The full density is $p_\beta(\boldsymbol{\theta}) = q_\beta(\boldsymbol{\theta})/z_\beta$, where z_β is its normalizing constant.

Each ratio in (3.11) is defined following the identity (3.12) which yields

$$r_k = \mathbb{E}_{p_{\beta_{k-1}}}\left[\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta}|M_0)}\right)^{\beta_k - \beta_{k-1}}\right],$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$ and $k = 1, \ldots, K$. A natural estimator of this ratio is

$$\widehat{r}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_1)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_0)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_0)}\right)^{\beta_k - \beta_{k-1}},$$

where $\boldsymbol{\theta}^i_{\beta_{k-1}}$ are samples from the power posterior distribution $p_{\beta_{k-1}}$, for $i = 1, \ldots, n$. This density has the potential of being a good importance distribution because it is slightly different from p_{β_k}. Numerical stability can be improved by factoring out the largest sampled term

$$\eta_k = \max_{1 \leq i \leq n}\left\{\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_1)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_0)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_0)}\right\}.$$

Thus, an estimate of $\log r$ is given by

$$\begin{aligned}\log \widehat{\mathrm{BF}}_{10} &= \sum_{k=1}^{K}\log \widehat{r}_k \\ &= -K\log n + \sum_{k=1}^{K}(\beta_k - \beta_{k-1})\log \eta_k \\ &\quad + \sum_{k=1}^{K}\log\sum_{i=1}^{n}\left(\frac{1}{\eta_k}\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_1)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M_0)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M_0)}\right)^{\beta_k - \beta_{k-1}}.\end{aligned}$$

The log-transformation induces a bias which can be minimized by increasing K. In addition, Baele et al. [2013] showed that doing this reduces the bidirectional error. The authors also noticed that the number of transitional distributions (K) has a higher impact on this error than the number of MCMC iterations (n) run per distribution.

3.3.3 Generalized steppingstone sampling

The *generalized steppingstone sampling* method [GSS; Fan et al., 2011] is a generalization of SS and is potentially more efficient. Inspired by the geometric path taken by Lefebvre et al. [2010] in a path sampling context, its main idea is to find a path shorter than the one between the posterior and prior by replacing the latter by a reference distribution. This strategy has the potential of leading to remarkable improvements in comparison to the original SS and PS, namely, less tuning parameters, lower variance, avoidance of numerical instabilities, reduction in the computational time, and it is more accurate in case of very diffusive priors [Baele et al., 2016]. This method also has the potential of dealing well with partially convex likelihoods (as functions of the cumulative prior probabilities, see more details in Section 3.4), unlike its predecessor and many other methods, but if and only if adequate reference distributions are used.

Consider the unnormalized density function q_β, the normalized density p_β with its normalizing constant z_β:

$$\begin{aligned} q_\beta(\boldsymbol{\theta}) &= \big(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\big)^\beta \pi_0(\boldsymbol{\theta}|M)^{1-\beta}, \\ p_\beta(\boldsymbol{\theta}) &= \frac{q_\beta(\boldsymbol{\theta})}{z_\beta}, \\ z_\beta &= \int_\Theta q_\beta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \end{aligned}$$

where $0 \leq \beta \leq 1$ and $\pi_0(\boldsymbol{\theta}|M)$ is the reference distribution. This is assumed to be proper, i.e., $\int_\Theta \pi_0(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} = 1$. When $\beta = 0$, p_β is equivalent to the posterior distribution and when $\beta = 1$, it is equivalent to the reference distribution. Note that when the prior is used as reference distribution, the original SS is recovered. The method bypasses sampling near the prior, which could be difficult when this is diffuse (a common case), through the insertion of the auxiliary distribution.

Similar to the original SS method, the marginal likelihood is expanded as the product of ratios of normalizing constants r_k, as it is defined in (3.11). Now, following the identity (3.12), each ratio is given by

$$r_k = \mathbb{E}_{p_{\beta_{k-1}}}\left[\frac{\big(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\big)^{\beta_k}\pi_0(\boldsymbol{\theta}|M)^{1-\beta_k}}{\big(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\big)^{\beta_{k-1}}\pi_0(\boldsymbol{\theta}|M)^{1-\beta_{k-1}}}\right]$$

$$= \mathbb{E}_{p_{\beta_{k-1}}}\left[\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{\pi_0(\boldsymbol{\theta}|M)}\right)^{\beta_k-\beta_{k-1}}\right]$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$ and $k = 1, \ldots, K$. An estimator of this ratio is given by

$$\widehat{r}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}{\pi_0(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}\right)^{\beta_k-\beta_{k-1}},$$

where $\boldsymbol{\theta}^i_{\beta_{k-1}}$ are drawn from $p_{\beta_{k-1}}$, for $i = 1, \ldots, n$. Numerical stability can be improved by factoring out the largest sampled term

$$\eta_k = \max_{1\leq i\leq n}\left\{\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}{\pi_0(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}\right\}.$$

Thus, taking the logarithm and summing over all K ratios, an estimator for $\log z$ is given by

$$\begin{aligned}\log\widehat{z} =& \sum_{k=1}^{K}\log\widehat{r}_k\\ =& -K\log n + \sum_{k=1}^{K}(\beta_k-\beta_{k-1})\eta_k\\ &+ \sum_{k=1}^{K}\log\sum_{i=1}^{n}\left(\frac{1}{\eta_k}\frac{L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)\pi(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}{\pi_0(\boldsymbol{\theta}^i_{\beta_{k-1}}|M)}\right)^{\beta_k-\beta_{k-1}}.\end{aligned} \tag{3.13}$$

The choice of the reference distribution is essential to guarantee a good performance of the method. Lefebvre et al. [2010] showed that this should be close to the posterior in the Kullback-Leibler sense. In this sense, the prior distribution is a poor option since the prior and the posterior are generally quite different. In theory, the optimal choice of reference distribution is the posterior, but it is not possible in practice because it is not completely known. In fact, it is possible to show that replacing the reference distribution by the posterior in (3.13) yields the equality $\log\widehat{z} = \log z$. Instead, these authors proposed the use of posterior samples to construct a density approximation of the posterior distribution. Following this line, Fan et al. [2011] proposed the use of matching the moments, for instance, the marginal posterior sample mean and variance, to parameterize a reference distribution for a parameter or block of parameters. This procedure leads to much more stable and efficient estimators. In practice, the priors can be parametrized by posterior samples following this approach. Baele et al. [2016] proposed to use kernel density estimation to construct this working distribution, in particular, the normal kernel.

This generalization of SS also leads to fewer tuning parameters. The original version defines a path between the prior and the posterior. The prior is, in general, very different from the posterior, which is much constrained than the former. This causes the estimated ratios near the prior to present more variability. The consensus solution is to put more effort in that area. In other words, the β values should follow a right skewed distribution which has to be specified. On the other hand, GSS defines a path between a reference and the posterior distribution. Optimally, these distributions should be similar. As a result, the estimated ratios near the reference distribution will not have as much variability as when the prior is used. The distributions involved in the numerator and denominator in each ratio will not be so different, yielding less variability in their estimation. Thus, the β values do not require to be positively skewed to increase the precision, but these can be equally spaced [Fan et al., 2011]. This means one less tuning parameter.

Finally, GSS requires less computational effort to achieve the same accuracy as PS and SS. This is a consequence of using the reference distribution which approximates the posterior. As a result, the path between these distributions is shorter than the one between the prior and the latter. Thus, it can be described in less steps leading to require less β values to achieve the same precision as PS and SS. This path also avoids numerical instabilities as the estimator moves towards the working distribution (in the case of melting scheme) and not the prior which can be very diffusive in practice. Consequently, GSS drastically reduces computation time [Baele et al., 2016].

3.3.4 Annealed importance sampling

Annealed importance sampling [AIS; Neal, 2001] is a method to estimate expectations, especially for those cases in which these are with respect to complex distributions, such as the marginal likelihood. Actually, the method allows the sampling from this distribution and consequently can produce an estimate of an expected value. It is especially suitable to deal in the case that multimodality may be a problem, but it allows general scenarios of estimation. The method uses an annealing scheme to adaptively define an importance sampling distribution to approximate the posterior. The author proposed the power posterior distribution as annealing scheme. Thus, the method produces a sample of points with corresponding weights which are averaged to estimate the marginal likelihood.

In a general case, suppose that an expectation of some function of $\boldsymbol{\theta}$ with respect to a probability density $p_{\beta_K}(\boldsymbol{\theta})$ is required. Also, consider the sequence of distributions $p_{\beta_0}(\boldsymbol{\theta})$ up to $p_{\beta_{K-1}}(\boldsymbol{\theta})$ with their corresponding proportional functions $q_\beta(\boldsymbol{\theta}) \propto p_\beta(\boldsymbol{\theta})$ and normalizing constants z_β. Again, these functions define a path between $p_{\beta_0}(\boldsymbol{\theta})$ and $p_{\beta_K}(\boldsymbol{\theta})$. This sequence can be specially constructed to solve a particular problem. In the same way as for the other path sampling methods described before, the geometric sequence is often a useful option and is defined as

$$q_\beta(\boldsymbol{\theta}) = q_0(\boldsymbol{\theta})^{1-\beta} q_K(\boldsymbol{\theta})^\beta,$$

where $0 \le \beta \le 1$. The unnormalized densities q_0 and q_K are a distribution easy to sample from and the one of interest, respectively.

AIS operates as the usual importance sampling method generating a sample of points with their corresponding weights to approximate the target distribution p_{β_K}. This is done via a Markov chain transition kernel, e.g., by using Metropolis-Hasting or Gibbs updates, at $\boldsymbol{\theta}_{\beta_{k-1}}$ (starting value). The algorithm to generate the n^{th} point, for $i = 1, \ldots, n$, is given by

- Sample $\boldsymbol{\theta}^i_{\beta_0}$ from p_{β_0}
- Sample $\boldsymbol{\theta}^i_{\beta_1}$ from p_{β_1} at $\boldsymbol{\theta}^i_{\beta_0}$

$\vdots$

- Sample $\boldsymbol{\theta}^i_{\beta_{K-2}}$ from $p_{\beta_{K-2}}$ at $\boldsymbol{\theta}^i_{\beta_{K-3}}$
- Sample $\boldsymbol{\theta}^i_{\beta_{K-1}}$ from $p_{\beta_{K-1}}$ at $\boldsymbol{\theta}^i_{\beta_{K-2}}$

The last sampling point $\boldsymbol{\theta}^i_{\beta_{K-1}}$ is a point from the target distribution if we consider its corresponding weight

$$\begin{aligned} w_i &= \frac{q_{\beta_1}\left(\boldsymbol{\theta}^i_{\beta_0}\right)}{q_{\beta_0}\left(\boldsymbol{\theta}^i_{\beta_0}\right)} \frac{q_{\beta_2}\left(\boldsymbol{\theta}^i_{\beta_1}\right)}{q_{\beta_1}\left(\boldsymbol{\theta}^i_{\beta_1}\right)} \cdots \frac{q_{\beta_{K-1}}\left(\boldsymbol{\theta}^i_{\beta_{K-2}}\right)}{q_{\beta_{K-2}}\left(\boldsymbol{\theta}^i_{\beta_{K-2}}\right)} \frac{q_{\beta_K}\left(\boldsymbol{\theta}^i_{\beta_{K-1}}\right)}{q_{\beta_{K-1}}\left(\boldsymbol{\theta}^i_{\beta_{K-1}}\right)} \\ &= \prod_{k=1}^{K} \frac{q_{\beta_k}\left(\boldsymbol{\theta}^i_{\beta_{k-1}}\right)}{q_{\beta_{k-1}}\left(\boldsymbol{\theta}^i_{\beta_{k-1}}\right)}, \end{aligned}$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$. Thus, this procedure produces independent samples from the target distribution by using a sequence of points from an approximating distribution which are weighted to compensate the use of the wrong sampling distribution. The average of these importance weights converges to the ratio of normalizing constants and defines an estimator

$$z = \frac{z_{\beta_K}}{z_{\beta_0}} \approx \frac{1}{n} \sum_{i=1}^{n} w_i, \tag{3.14}$$

where $z_{\beta_K} = \int q_{\beta_K}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ and $z_{\beta_0} = \int q_{\beta_0}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. The method is still valid even when the starting values $\boldsymbol{\theta}^i_{\beta_0}$ at each iteration are not independent [Neal, 2001]. Even in the case that MCMC steps are far from equilibrium, AIS is unbiased, though poor mixing will not make the method outperform standard importance sampling [Murray, 2007].

The ratio defined in (3.14) will be an estimate of the marginal likelihood when q_{β_0} and q_{β_K} are proportional to the prior and posterior, respectively. In this case, the geometric sequence of unnormalized densities is given by

$$q_\beta(\boldsymbol{\theta}) = L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M),$$

leading consequently to weights

$$w_i = \prod_{k=1}^{K} L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k - \beta_{k-1}}.$$

Alternatively, q_{β_0} could be chosen such that it defines a shorter path to the posterior. Like in the GSS case, we could use a reference distribution which approximates the posterior. This is potentially a more efficient way to estimate the marginal likelihood. In this case, we will refer to the method as GAIS in order to avoid confusion. We use AIS when q_{β_0} is proportional to the prior. Analogously, the Bayes factor can be estimated defining a path between the unnormalized posterior distributions of the two models.

For a sample of independent points from the transition distributions, an estimate of the variance is

$$\begin{aligned}
\mathrm{Var}[\widehat{z}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} w_i\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[w_i] \\
&= \frac{1}{n}\mathrm{Var}[w] \\
&\approx \frac{1}{n^2}\sum_{i=1}^{n}(w_i - \widehat{z})^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n} w_i^2 - \frac{\widehat{z}^2}{n}.
\end{aligned}$$

Based on the delta method, an estimate of the variance of the log-marginal likelihood is given by

$$\mathrm{Var}[\log \widehat{z}] \approx \frac{1}{\widehat{z}^2}\mathrm{Var}[\widehat{z}].$$

3.3.4.1 Link to SS

Even though AIS and SS were developed independently and published in 2001 and 2011, respectively, they are closely related. Both methods can be seen as using the same telescope product of ratios of normalizing constants to estimate the marginal likelihood. SS relies directly on the identity

$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}} \frac{z_{\beta_2}}{z_{\beta_1}} \dots \frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}} \frac{z_{\beta_K}}{z_{\beta_{K-1}}} = \prod_{k=1}^{K} \frac{z_{\beta_k}}{z_{\beta_{k-1}}} = \prod_{k=1}^{K} r_k,$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$. This method works by estimating each ratio $r_k = z_{\beta_k}/z_{\beta_{k-1}}$ via importance sampling. Thus, the estimate for the marginal likelihood is given by

$$\widehat{z}_{\mathrm{SS}} = \prod_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k - \beta_{k-1}}.$$

On the other hand, AIS can be seen as an estimate of the same product of ratios r_k. But, instead of estimating each ratio separately, it uses estimates of the whole product through importance sampling and then it averages them. Hence, the estimate is given by

$$\widehat{z}_{\mathrm{AIS}} = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{K} L(\boldsymbol{X}|\boldsymbol{\theta}^i_{\beta_{k-1}}, M)^{\beta_k - \beta_{k-1}}.$$

For $K = 1$, both methods are reduced to the arithmetic mean. When only one point is sampled from each transition distribution ($n = 1$), both methods are equivalent.

We could also see SS from an AIS point of view. If we expanded SS estimate, we would notice that SS estimates the whole telescope ratio at once many times (n^K) and then takes its average in the same way as AIS. This is done by reutilizing each sample point at least once, unlike AIS which uses each exactly once in the estimation process. The former is an average of n^K points whereas the latter is only composed of n. Actually, a set of elements which composes the SS estimate is the AIS estimate. We could also see SS estimate as an average of all the possible AIS estimates obtained by combining the samples points.

To understand this, consider the case that $n = 3$ and $K = 2$. The elements that do not depend on these specifications have been omitted for simplicity. In this particular case, AIS estimate is given by

$$\widehat{z}_{\text{AIS}} = \frac{1}{2}\Big(L\big(\boldsymbol{\theta}^1_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^1_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^2_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^2_{\beta_1}\big)^{\beta_2-\beta_1} + \\ L\big(\boldsymbol{\theta}^3_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^3_{\beta_1}\big)^{\beta_2-\beta_1}\Big),$$

and the SS estimate

$$\widehat{z}_{\text{SS}} = \frac{1}{3}\Big(L\big(\boldsymbol{\theta}^1_{\beta_0}\big)^{\beta_1-\beta_0} + L\big(\boldsymbol{\theta}^2_{\beta_0}\big)^{\beta_1-\beta_0} + L\big(\boldsymbol{\theta}^3_{\beta_0}\big)^{\beta_1-\beta_0}\Big)\times \\ \frac{1}{3}\Big(L\big(\boldsymbol{\theta}^1_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^2_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^3_{\beta_1}\big)^{\beta_2-\beta_1}\Big).$$

To see this estimate as an average of the whole ratio of normalizing constants, note that this last expression can be expanded as follows

$$\frac{1}{9}\Big(L\big(\boldsymbol{\theta}^1_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^1_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^1_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^2_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^1_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^3_{\beta_1}\big)^{\beta_2-\beta_1} + \\ L\big(\boldsymbol{\theta}^2_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^1_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^2_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^2_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^2_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^3_{\beta_1}\big)^{\beta_2-\beta_1} + \\ L\big(\boldsymbol{\theta}^3_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^1_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^3_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^2_{\beta_1}\big)^{\beta_2-\beta_1} + L\big(\boldsymbol{\theta}^3_{\beta_0}\big)^{\beta_1-\beta_0} L\big(\boldsymbol{\theta}^3_{\beta_1}\big)^{\beta_2-\beta_1}\Big).$$

From here, one sees that the diagonal terms of the broken equation above are actually the elements which compose AIS estimate. One also sees that the sample points are reutilized multiple times. For instance, $\boldsymbol{\theta}^3_{\beta_0}$ is used 3 times in the first line of the equation. In general, each point is used N times.

The methods use the sample points in a different way to produce an estimate of the marginal likelihood. This difference leads to different uncertainties in the estimates. This aspect is studied in example 3.5.1. The results show that GSS yields a lower uncertainty than GAIS under different specifications, even though they are performed by using the same sample points.

Although PS and SS yield accurate estimates of the marginal likelihood, they require several specifications depending on the problem. Firstly, an annealing schedule (a number of β values) is required. A common practice is to try with different numbers until the estimate is stable. This procedure is described in Drummond and Bouckaert [2015] as follows: "*run the path sampling analysis* *with a low number of steps (say 10) first, then increase the number of steps (with say increments of 10, or doubling the number of steps) and see whether the marginal likelihood estimates remain unchanged*". This could be impractical in some situations, for instance, when flat priors are used, which would increase the number of steps. Secondly, the path described by the β values has to be defined. However, there is a consensus on putting more effort in the area around 0 see [see Friel and Pettitt, 2008; Lepage et al., 2007; Xie et al., 2011]. Finally, these methods require a number of samples from the power posterior for each β value. Thus, the main problem is that optimal specifications vary from case to case. The popularity of SS is due to its implementation in popular software such as MrBayes [Huelsenbeck and Ronquist, 2001] or BEAST [Drummond et al., 2012]. However the mentioned specifications have to be defined by the user, or use some predetermined parameters that might be unsuitable.

In this context, GSS, and also GAIS, require potentially less tuning parameters for an appropriate reference distribution. Firstly, it requires an annealing/melting scheme (a number of β values). The estimation can start from either the prior or posterior distribution. However, the β values do not require to follow any particular distribution to control effectively the uncertainty of the estimate as in PS or SS, because of the similarity of the reference and posterior distributions [Fan et al., 2011]. Thus, the values can be equally spaced between 0 and 1. Also, GSS does not need as many transitional distributions as its original version and it is more robust to prior specifications, i.e., the prior does not have a huge effect on the method performance. Finally, the method requires a number of samples from each transitional distribution.

PS and SS have usually been presented as methods of general applicability [Arima and Tardella, 2014; Baele and Lemey, 2014; Baele et al., 2013; Xie et al., 2011]. However, these methods only work when the shape of the likelihood, as a function of the cumulative prior probabilities (see Figure 3.3), is concave. Partly convex likelihood functions might need impractical computational effort or make them fail outright. This phenomenon has been well studied in statistical physics and is known as *phase transition*, a situation where a slight change in β leads to a big change in the power posterior distribution. Thus, the transition distributions are unable to mix between different phases of the likelihood function, resulting in a poor estimate. A more general method is nested sampling [Skilling, 2006], an algorithm that measures the relationship between likelihood values and the prior distribution, and uses this to compute the marginal likelihood. This characteristic allows it to cope with partly convex likelihood functions. More importantly, unlike PS, SS, AIS and GSS, NS requires less problem-specific tuning.

3.4 Nested sampling

Nested sampling [NS; Skilling, 2006] is a more general technique for the estimation of the marginal likelihood. It requires less tuning, yields a measure of the uncertainty of the estimate in a single run, and can deal with partly convex likelihood (as a function of the cumulative prior probabilities) shapes. Its main feature is the reduction of the multidimensional integral to a one-dimensional integral over the parameter space. This technique, and several variants [e.g., Brewer et al., 2011; Feroz et al., 2009; Handley et al., 2015] have been successfully applied to fields like astronomy [Brewer and Donovan, 2015; Mukherjee et al., 2006] or systems biology [Aitken and Akman, 2013; Pullen and Morris, 2014] and shown great promise in parameter inference and model selection.

To understand the key idea of the method, consider that for any positive random variable Y, its expected value can be written as

$$\mathbb{E}[Y] = \int_0^\infty \big(1 - F_Y(y)\big)\mathrm{d}y, \tag{3.15}$$

where F_Y is the distribution function of Y. This integral depicts the area between the distribution function of Y and 1. Similarly, the likelihood function $L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ can be seen as a positive random variable L which is a function of $\boldsymbol{\theta}$, i.e., $L(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ follows the prior distribution $\pi(\boldsymbol{\theta}|M)$. Thus, the marginal likelihood can be seen as the expected value of the likelihood function. NS takes advantage of this approach and the property (3.15) to transform the multidimensional integral defined in (3.2) into a one dimensional integral as follows

$$\mathbb{E}_\theta\big[L(\boldsymbol{X}|\boldsymbol{\theta}, M)\big] \equiv \mathbb{E}_L\big[L\big] = \int_0^\infty \big(1 - F_L(l)\big)\mathrm{d}l, \tag{3.16}$$

where $\mathbb{E}_\theta[\cdot]$ and $\mathbb{E}_L[\cdot]$ stand for the expectation with respect to the densities of θ and L respectively, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|M)$, $L = L(\boldsymbol{\theta}) = L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ and $F_L(l)$ is the cumulative distribution function of the likelihood defined by

$$F_L(l) = \underset{L(\boldsymbol{X}|\boldsymbol{\theta},M)<l}{\int \cdots \int} \pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}.$$

Considering $\xi(l) = 1 - F_L(l) = p$, the proportion of prior mass with likelihood greater than l, and taking its inverse, the evidence given in (3.16) is redefined as

$$z = \int_0^1 \xi^{-1}(p)\mathrm{d}p. \tag{3.17}$$

This is the integral used by nested sampling, and is displayed in Figure 3.3. ξ^{-1} is the likelihood function but with a different domain. Its arguments are cumulative prior probabilities, unlike $L(\boldsymbol{\theta})$ which has parameter vectors as argument. Its codomain is naturally composed of likelihood values. Mathematically speaking, $\xi^{-1}(p) = l$ is that likelihood l such that $\mathbb{P}(L(\boldsymbol{\theta}) > l) = p$. For instance, $\xi^{-1}(0.95) = 0.02$ can be read as 95% of draws $\boldsymbol{\theta}$ from the prior will have likelihoods greater than 0.02. In general, ξ^{-1} is highly right skewed, concentrating its mass near 0 because the posterior is usually located in a small area of the prior. Note that this function is a monotonically decreasing function which reaches its highest point at $p = 0$ and its lowest point at $p = 1$ (see Figure 3.3).

The following example in the discrete case helps to understand how NS works. Consider the two-dimensional parameter space on the 4×4 grid displayed in Figure 3.4 (on the left). The values of each quadrant stand for the likelihood. Also, each cell is assumed to have equal prior probability 1/16. NS sorts the likelihood values in an increasing order and relates them with their cumulative prior mass. Actually, this is what the integral 3.17 represents. Figure 3.4 (on the right) shows this procedure (from right to left). The relationship between the values can be read as 38% of the points drawn from the prior will have a likelihood greater or equal than 11 or 6% will be greater or equal to 42. The area covered by the bars is equivalent to the marginal likelihood, i.e., the sum of the heights times the widths of each bar or equivalently, the likelihood times the

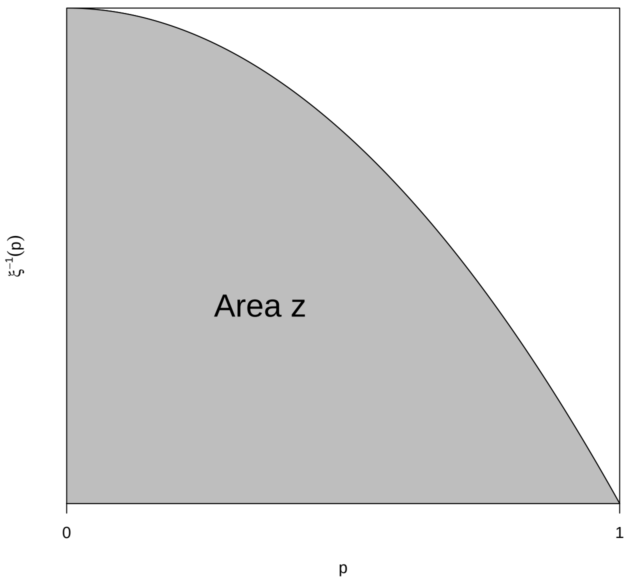


Figure 3.3: Sorted likelihood function with respect to the prior distribution with area z. ξ^{-1} is the likelihood function with cumulative prior probabilities as its argument.

prior. Thus, its calculation is given by

$$\begin{aligned} z &= \frac{1}{16}(1+2+3+4+5+6+7+8+9+10+11+13+38+40+41+42) \\ &= 15. \end{aligned}$$

In this toy example is quite easy to associate the likelihood to the prior mass and consequently estimate the marginal likelihood. However, this is rarely possible in a multidimensional and continuous parameter space, but feasible in principle.

In general, if a decreasing sequence of p-values and an increasing sequence of L-values ($\xi^{-1}(p)$) is available, the marginal likelihood can be approximated numerically by the basic standard quadrature method

$$\widehat{z} = \sum_{i=1}^{k} w_i L_i, \tag{3.18}$$

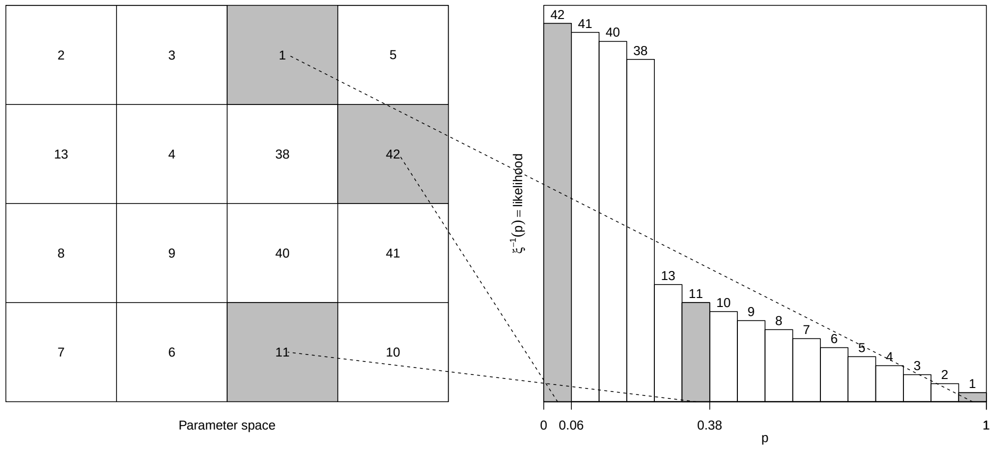


Figure 3.4: On the left: likelihood values in the parameter space. On the right: sorted likelihood.

where $w_i = p_{i-1} - p_i$ (or $w_i = (p_{i-1} - p_{i+1})/2$ for the trapezoidal rule) and $L_i = \xi^{-1}(p_i)$. How to generate these sequences is described below.

Sequence of L-values

Nested sampling maintains a set of N *active points* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ (with respective associated likelihood values $L(\boldsymbol{\theta}_1), \ldots, L(\boldsymbol{\theta}_N)$) to generate the ith likelihood value required in (3.18). Initially they are drawn from the prior distribution, $\pi(\boldsymbol{\theta})$. From this set, the method requires selecting the point $\boldsymbol{\theta}_l$, where $l \in \{1, \ldots, N\}$, with the lowest likelihood value. This value contributes to the estimation as a summand in (3.18). Then, the point $\boldsymbol{\theta}_l$ is discarded from the active points and replaced by a new point $\boldsymbol{\theta}$ sampled from the prior, but constrained to have a greater likelihood value than the point being replaced, i.e., $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_l)$. This procedure shrinks the parameter space according to the likelihood restriction. The process is repeated until a given stopping rule is satisfied (more information on this will follow later). Thus, a sequence of increasing likelihood values $(L_1, \ldots, L_k)$ and *discarded points* $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ are generated. The discarded points are the ones that contribute to the estimate of the marginal likelihood through their respective likelihoods.

Sequence of p-values

The discarded points generate an increasing sequence of likelihoods, which are known precisely. An important insight of Skilling [2006] is that the corresponding p values, while they cannot be measured precisely, can be estimated from the nature of the NS procedure. Nested sampling explores the prior distribution geometrically as follows

$$p_1 = u_1, \quad p_2 = u_1 u_2, \quad \dots \quad , \quad p_k = \prod_{i=1}^{k} u_i,$$

where $u_i = p_i/p_{i-1} \in [0,1]$, for $i = 1, \dots, k$, with $p_0 = 1$. This variable follows a Beta$(N,1)$ distribution. This is because at the i^{th} iteration, NS takes N points $x_1^i, \dots, x_N^i$ which follows a Uniform$(0, p_{i-1})$. These values are cumulative probabilities and consequently have a uniform distribution. Their maximum value is p_i which is related to the minimum likelihood value (note that $\xi^{-1}(p_i)$ is a non-increasing function). Since the distribution of x_j^i/p_{i-1} is a Uniform$(0,1)$, for $j = 1, \dots, N$, their maximum value p_i/p_{i-1} follows a Beta$(N,1)$ distribution. Skilling [2006] defined two schemes for estimating the p-values: *stochastic* and *deterministic*.

- *Stochastic:* the u_i values are generated randomly from the Beta$(N,1)$ distribution, for $i = 1, \dots, k$.
- *Deterministic:* the u_i values are fixed by using their expectations as follows:
  - Considering its *arithmetic mean*, $u_i = N/(N+1)$, approximate p-values would be given by
  $$p_i = \left(\frac{N}{N+1}\right)^i.$$
  - Considering its *geometric mean*, $u_i = e^{-1/N}$, the estimated prior mass would be
  $$p_i = e^{-i/N}.$$

Thus, a sequence of p values can be generated and used in (3.18). The use of the geometric mean seems more reasonable given that the prior mass exploration is geometric. This scheme is considered for our examples, and is the one recommended by most authors. However, the arithmetic mean allows nested sampling to be connected to rare event simulation [Walter, 2017], and allows for an alternative version of NS with unbiased estimates of z.

Sampling

The highest cost of nested sampling is in sampling from the restricted prior distribution (due to the condition that the likelihood needs to increase). Skilling [2006] suggested to use a Metropolis-Hastings algorithm as usual, to explore the prior with the additional condition of rejecting the proposal points which do not fulfill the likelihood restriction. As a starting value, a point from the sequence of active points can randomly be selected at each iteration of NS, as all of them meet the likelihood condition by definition. Several other efficient methods have also been proposed [Brewer et al., 2011; Feroz et al., 2009; Mukherjee et al., 2006]. We use Skilling's method to generate the restricted prior samples in our application with proposal moves as described in Brewer and Donovan [2015], which are discussed in Section 4.4.

The amount of information provided by the data about the parameters plays a key role in NS. It helps to define the uncertainty of the estimate and also a termination criterion. This concept is explained below.

Information

The idea of how much we have learned from the data is quantified through the notion of entropy. The measure of information [Knuth and Skilling, 2012; Sivia and Skilling, 2006] is given by the negative relative entropy

$$H = \int P(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) \mathrm{d}\boldsymbol{\theta},$$

where $P(\boldsymbol{\theta})$ is the posterior distribution. This quantity represents the amount of information in the posterior with respect to the prior, after acquiring the data. By definition, it can be seen as the expected value $H = \mathbb{E}_P[\log(P(\boldsymbol{\theta})/\pi(\boldsymbol{\theta}))]$. Its approximation is given by

$$H \approx \sum_i \frac{w_i L_i}{z} \log\left(\frac{L_i}{z}\right)$$

with $w_i = p_{i-1} - p_i$ [Sivia and Skilling, 2006]. The following property of expected values is useful to understand the use of this concept. If G_Y is the geometric mean of Y, we have that

$$\log \mathrm{G}_Y = \mathbb{E}[\log Y] \Leftrightarrow \mathrm{G}_Y = e^{\mathbb{E}[\log Y]}. \tag{3.19}$$

According to this property, e^{-H} is a measure of central tendency or a typical value of $\pi(\boldsymbol{\theta})/P(\boldsymbol{\theta})$. This value can be seen as the bulk of the posterior mass that occupies the prior. This idea helps to define a termination condition for nested sampling which will be described later.

Note that a prior distribution which is consistent with the likelihood function, namely, they support the same parameter values, has a lower information than a likelihood function which is in contradiction with the prior, i.e., their mass is concentrated in different places. In other words, if the previous belief changes a lot after acquiring the data, more information has been gained from the data.

Uncertainty

The numerical uncertainty associated with the NS estimation of z comes from two sources: i) approximating the prior volume ($w_i = p_{i-1} - p_i$), and ii) the error imposed by the integration rule. However, the total uncertainty is usually dominated by the first. Actually, the second is at most $\mathcal{O}(N^{-1})$ and $\mathcal{O}(N^{-2})$ for the simple standard quadrature and trapezoidal methods, respectively, and thus negligible in comparison to the first source [Skilling, 2006].

In a such a way, the uncertainty in $\log \widehat{z}$ depends directly on the uncertainty in $\sum_{i=1}^{k} \log p_i$. Noting that $-\log p_i \sim \mathrm{Exp}(N)$, and that consequently $-\sum_{i=1}^{k} \log p_i \sim \mathrm{Gamma}(k, N)$, where k is the number of iterations required by NS, we have that $\mathrm{dev}\left[\sum \log p_i\right] = \sqrt{k}/N$. Skilling [2006] argued that NS requires around $N \times H$ steps to reach the posterior, therefore its uncertainty can be approximated as

$$\operatorname{dev}\big[\log z\big] = \sqrt{\frac{H}{N}}. \tag{3.20}$$

The asymptotic variance of the nested sampling approximation grows linearly with the dimension of $\boldsymbol{\theta}$ and its distribution is asymptotically Gaussian [Chopin and Robert, 2010].

Another way of calculating the uncertainty is by replicating the NS estimates for different p-sequences, i.e., using the stochastic approach, but keeping the same likelihood sequence. Thus, a distribution of $\log \widehat{z}$ can be inferred. Note that this represents a marginal computational cost since most of it is spent by generating the likelihood sequence. This strategy can also be used similarly for parameter inference.

Figure 3.5 shows the NS algorithm. The routine is repeated until a given stopping criterion is satisfied. However, there is no rigorous criterion that guarantees that we have found most of the bulk of z. Nevertheless, some termination conditions have been proposed [Skilling, 2006] and are described below.

Termination

The loop could continue until the potential maximum new contribution $L_i w_i$ represents a small fraction γ of the accumulated evidence, that is,

$$\max\big(L(\boldsymbol{\theta}_1), L(\boldsymbol{\theta}_2), \ldots, L(\boldsymbol{\theta}_N)\big) w_i < \gamma z_{i-1}.$$

The algorithm stops when the potential maximum new contribution is not significant.

Another criterion is based on the concept of information defined before. Typically, the likelihood values L start dominating the prior mass w, so the contribution Lw increases at the beginning until the prior mass dominates this quantity. After reaching a maximum, theses values start to decrease. The peak of this function is reached in the region of $p \approx e^{-H}$, when most of the posterior mass in the prior has been found. Given that $p_i \approx e^{-i/N}$, a natural termination condition

```
Nested sampling algorithm:
1.  Sample $N$ points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ from the prior $\pi(\boldsymbol{\theta})$;
2.  Initialize $z_0 = 0$ and $p_0 = 1$;
3.  Repeat for $i = 1, \dots, k$;
    i)   out of the $N$ live points, take the one with the lowest
         likelihood which we call $\boldsymbol{\theta}_l$ with corresponding likelihood
         $L_i = L(\boldsymbol{\theta}_l)$;
    ii)  set $p_i = \exp(-i/N)$;
    iii) set $w_i = p_{i-1} - p_i$ (or $w_i = (p_{i-1} - p_{i+1})/2$ for the trape-
         zoidal rule);
    iv)  update $z_i = w_i L_i + z_{i-1}$; and
    v)   update the set of active points $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ replacing $\boldsymbol{\theta}_l$
         by drawing a new point $\boldsymbol{\theta}$ from the prior distribution
         restricted to $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_l)$.
```

Figure 3.5: Description of NS algorithm. It iterates until a stopping criterion is satisfied.

to estimate the log-evidence would be to stop the loop when i/N significantly exceeds H, i.e., when the posterior mass has been explored completely. In practice, NS can be stopped when the number of iterations exceeds $2\times N\times H$, where H is estimated inside the loop at each iteration.

There is no guarantee *in general* that these termination conditions will work perfectly. L might start increasing at a greater rate in the future, overwhelming the points that currently have high weights. In specific cases where the maximum likelihood value is known or can be roughly anticipated, it is possible to be confident that this will not happen.

Posterior samples

NS yields posterior samples at no extra cost, if we assign appropriate weights to the discarded output points. In each iteration, NS has taken out a point from the active points generating a sequence of discarded points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k$. These discarded points have contributed to the estimation of the marginal likelihood with their respective weights wL which are proportional to the posterior distribution, i.e., prior multiplied by likelihood. Thus, the sequence of discarded points can be sampled according to these weights in order to get a posterior sample. The effective sample size is related to the entropy of the posterior weights as

$$\mathcal{M} = \exp\left(- \sum_{i=1}^{k} \tilde{p}_i \log \tilde{p}_i \right), \quad \text{where} \quad \tilde{p}_i = \frac{w_i L_i}{z}.$$

3.4.1 Direct Bayes factor estimation

The NS algorithm can also be extended to estimate the Bayes factor directly. Assume, one wants to compare models M_0 and M_1 defined on the parameter spaces Θ_0 and Θ_1, respectively. The marginal likelihood is denoted by z_j for M_j, with $j = 0, 1$. The Bayes factor has the form of

$$\mathrm{BF}_{10} = \frac{\int_{\Theta_1} L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1|M_1)\mathrm{d}\boldsymbol{\theta}_1}{\int_{\Theta_0} L(\boldsymbol{X}|\boldsymbol{\theta}_0, M_0)\pi(\boldsymbol{\theta}_0|M_0)\mathrm{d}\boldsymbol{\theta}_0},$$

where $\boldsymbol{\theta}_0 \in \Theta_0$ and $\boldsymbol{\theta}_1 \in \Theta_1$. We merge these parameter vectors into one as $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \in \Theta$, such that, for instance, if $\boldsymbol{\theta}_0 = (a, b, c)$ and $\boldsymbol{\theta}_1 = (c, d)$, then $\boldsymbol{\theta} = (a, b, c, d)$. It is feasible to rewrite the Bayes factor expression as

$$\begin{aligned}
\mathrm{BF}_{10} &= \frac{1}{z_0} \int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\pi(\boldsymbol{\theta}|M_0)\mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{z_0} \int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)\pi(\boldsymbol{\theta}|M_0)}{g_0(\boldsymbol{\theta})} g_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_1)}{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)} g_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\Theta} \Psi(\boldsymbol{\theta}) g_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

where g_0 is an instrumental density defined by

$$g_0(\boldsymbol{\theta}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_0, M_0)\pi(\boldsymbol{\theta}_0|M_0)}{z_0}\pi(\boldsymbol{\theta}_1|M_1)$$

and $\Psi(\boldsymbol{\theta}) = L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)/L(\boldsymbol{X}|\boldsymbol{\theta}_0, M_0)$. The parameters in $\boldsymbol{\theta}$ which do not correspond to the model have no participation on the computation of the likelihood and prior functions. In the simplest case that both models are defined on the same parameter space and with same priors, the instrumental function is given by

$$g_0(\boldsymbol{\theta}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M_0)\pi(\boldsymbol{\theta})}{z_0}.$$

Now, the Bayes factor can be seen as the expected value of a pseudo-likelihood Ψ with respect to a pseudo-prior g_0. In other words, it has been rewritten in a way NS can deal with. Actually, any extension of NS can be applied to this expression.

Noting that the ratio of likelihoods Ψ is a positive random variable (see property (3.15)), BF_{10} can take the form of

$$\begin{aligned}\mathbb{E}_\theta\big[\Psi(\boldsymbol{\theta})\big] \equiv \mathbb{E}_\Psi\big[\Psi\big] &= \int_0^\infty \big(1 - F(\psi)\big)\mathrm{d}\psi \\ &= \int_0^1 \xi^{-1}(p)\mathrm{d}p\end{aligned}$$

where E_θ and E_Ψ depict expected values with respect to the densities of $\boldsymbol{\theta}$ and Ψ respectively, $\boldsymbol{\theta} \sim g_0(\boldsymbol{\theta})$, $F(\psi)$ is the distribution function of Ψ given by

$$F(\psi) = \underset{\Psi(\boldsymbol{\theta})<\psi}{\int\cdots\int} g_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

and ξ^{-1} is the inversed function of $\xi(\psi) = 1 - F(\psi) = p$, i.e., $\xi^{-1}(1 - F(\psi)) = \psi$.

With the pseudo-likelihood Ψ and the pseudo-prior g_0 defined, NS follows its normal course. It starts with a set of N active points sampled from g_0. It initializes $\mathrm{BF}_0 = 0$ and $p_0 = 1$. At the i^{th} iteration, the point with the lowest ratio of likelihoods $\boldsymbol{\theta}_l$ contributes to the estimation as $\mathrm{BF}_i = (p_{i-1} - p_i)\Psi(\boldsymbol{\theta}_l) + \mathrm{BF}_{i-1}$, where $p_i = \exp(-i/N)$. Then, it is replaced by a new point $\boldsymbol{\theta}$ sampled from g_0 restricted to $\Psi(\boldsymbol{\theta}) > \Psi(\boldsymbol{\theta}_l)$. The algorithm is repeated until a given stopping criterion is satisfied.

When both models are defined on the same parameter space and considering same priors, the algorithm yields posterior samples for model 1 at no extra cost. The procedure is analogous to that for the original algorithm defined above.

Analogously to the original algorithm, the standard deviation of the estimate depends on the information and the number of active points. This is

$$\mathrm{SD}[\log(\mathrm{BF}_{10})] = \sqrt{\frac{H}{N}}.$$

The pseudo-posterior distribution for this case is given by

$$P(\boldsymbol{\theta}) = \frac{\Psi(\boldsymbol{\theta})g_0(\boldsymbol{\theta})}{\mathrm{BF}_{10}} = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1|M_1)}{z_1}\pi(\boldsymbol{\theta}_0|M_0) = g_1(\boldsymbol{\theta}).$$

Thus, the information takes the form of

$$H = \int_\Theta g_1(\boldsymbol{\theta}) \log\left(\frac{g_1(\boldsymbol{\theta})}{g_0(\boldsymbol{\theta})}\right)\mathrm{d}\boldsymbol{\theta}.$$

For the most general case, when the models are defined on different parameter spaces, the information is given by

$$H = H_1 + \int_{\Theta_0} \pi(\boldsymbol{\theta}_0|M_0) \log\left(\frac{\pi(\boldsymbol{\theta}_0|M_0)}{p(\boldsymbol{\theta}_0|\boldsymbol{X}, M_0)}\right)\mathrm{d}\boldsymbol{\theta}_0,$$

where H_1 is the information for model 1. The second term is the Kullback-Leibler divergence $\mathrm{KL}(\pi_0, p_0)$ between the prior and posterior of model 0. This quantity can be seen as the capacity of the posterior to approximate the prior. However, in general, the posterior is much more constrained than the prior. The prior could allow for values that are hard to generate by the posterior. Thus, the integral may be undefined.

For the particular case where both models are defined on the same parameter space with the same priors, the information is given by

$$H = H_1 - \int_\Theta p_1(\boldsymbol{\theta}|\boldsymbol{X}, M_1) \log\left(\frac{p_0(\boldsymbol{\theta}|\boldsymbol{X}, M_0)}{\pi(\boldsymbol{\theta})}\right)\mathrm{d}\boldsymbol{\theta}.$$

The calculation of the Bayes factor by using independent marginal likelihood estimates via NS leads to the following variance

$$\mathrm{Var}\left[\log\left(\frac{z_1}{z_0}\right)\right] = \mathrm{Var}[\log z_1] + \mathrm{Var}[\log z_0] = \frac{H_1}{N} + \frac{H_0}{N}. \tag{3.21}$$

This variance differs only in the second term in comparison to our proposal. For the case that both models have the same parametric support and priors, the direct estimation of the Bayes factor could lead to a smaller uncertainty (see example 3.5.2.2).

3.4.2 Nested importance sampling

Nested importance sampling (NIS) is an extension of NS which gathers the idea of importance sampling methods making use of an auxiliary function to potentially make the estimation process more efficient. This idea was briefly discussed by Skilling [2006] and subsequently in more depth by Chopin and Robert [2010] and Feroz et al. [2013].

The marginal likelihood can be rewritten as

$$\begin{aligned} z &= \int_\Theta \frac{L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \\ &= \int_\Theta \Psi(\boldsymbol{\theta}) g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{3.22}$$

where $\Psi(\boldsymbol{\theta}) = L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})/g(\boldsymbol{\theta})$, with $g(\boldsymbol{\theta})$ the importance sampling density. The functions $g(\boldsymbol{\theta})$ and $\Psi(\boldsymbol{\theta})$ play the role of a pseudo-prior and pseudo-likelihood, respectively. Ψ can be seen as a positive random variable since it is a product of positive functions. Thus, considering the properties of this kind of variables (see property (3.15)), the integral can be seen as follows

$$\begin{aligned} \mathbb{E}_\theta\big[\tilde{\omega}(\boldsymbol{\theta})\big] \equiv \mathbb{E}_\Psi\big[\Psi\big] &= \int_0^\infty \big(1 - F(\psi)\big)\mathrm{d}\psi \\ &= \int_0^1 \xi^{-1}(p)\mathrm{d}p, \end{aligned}$$

where $\mathbb{E}_\theta()$ and $\mathbb{E}_\Psi$ stand for expectations with respect to the densities of $\boldsymbol{\theta}$ and Ψ respectively, $\boldsymbol{\theta} \sim g(\boldsymbol{\theta})$, ψ is a value from the random variable Ψ, $F(\psi)$ is the cumulative distribution function of the pseudo-likelihood defined as

$$F(\psi) = \int \cdots \int_{\Psi(\boldsymbol{\theta})<\psi} g(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

and $\xi(\psi) = 1 - F(\psi) = p$ with inverse function $\xi^{-1}(p) = \psi$. Note that ξ^{-1} has as domain cumulative pseudo-prior probabilities and as codomain pseudo-likelihood values.

After defining the pseudo-prior and the pseudo-likelihood functions, NIS works in the same way as NS. In each iteration it requires the identification of $\boldsymbol{\theta}_l$ with the lowest pseudo-likelihood value $\Psi(\boldsymbol{\theta}_l)$. Then, a new point $\boldsymbol{\theta}$ has to be sampled from the pseudo-prior $g(\boldsymbol{\theta})$ restricted to have a greater pseudo-likelihood value, i.e., $\Psi(\boldsymbol{\theta}) > \Psi(\boldsymbol{\theta}_l)$. The rest of the algorithm does not change.

The performance on NIS depends on the choice of the auxiliary distribution $g(\boldsymbol{\theta})$. A good choice could dramatically improve the performance in comparison to the original algorithm. This would yield an estimate with lower uncertainty and with less computational effort. On the other hand, a bad choice could lead to the opposite scenario. Theoretically, the optimal choice of the auxiliary density is the posterior distribution. In this case, it is straightforward to show that the integral defined in (3.22) is reduced to the marginal likelihood z. Or equivalently, it is easy to show that the information H, defined as

$$H = \int P(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right) \mathrm{d}\boldsymbol{\theta}, \tag{3.23}$$

is 0 when the $g(\boldsymbol{\theta})$ is the posterior distribution $P(\boldsymbol{\theta})$. Consequently, the uncertainty of the estimate, $\mathrm{SD}\big[\log z\big] = \sqrt{H/N}$, is equal to 0. However, this density is not feasible in practice. When the importance sampling function is the prior, it is easy to see that the integral (3.22) becomes the definition of the marginal likelihood, so the original NS algorithm is recovered. However, this is a poor choice since the prior and posterior are generally very different.

In practice, the auxiliary density should approximate the posterior distribution as much as possible. For the path sampling case, Lefebvre et al. [2010] suggested to use posterior samples to calibrate an auxiliary density. For instance, Fan et al. [2011] proposed to parameterize the prior density by marginal posterior samples for the generalized steppingstone sampling method. In practical terms, this choice represents a viable and efficient alternative in a nested importance sampling context (see example 3.5.1.1 and 3.5.2.1). The variable tree topology case in which a reference distribution for phylogenies is required is discussed in the next chapter.

Even though NIS requires an additional computational effort of sampling the posterior in order to parametrize the reference distribution, this effort is compensated by allowing an easier sampling distribution and requiring less iterations to get a stable estimate.

NS requires independent points from the prior restricted to the likelihood. This is particularly hard for non-informative priors or when this is in contradiction with the likelihood. In these cases, low prior volume is assigned to where the likelihood is densely concentrated. Thus, the points generated from the prior will be situated in areas with high likelihood values with a small probability. This problem is exacerbated over time, when the sampling distribution gets more constrained. On the other hand, NIS requires independent points from the instrumental distribution restricted to the pseudo-likelihood. This sampling distribution is potentially much easier and more efficient to sample from. Note that the pseudo-prior is ideally an approximation of the posterior and consequently provides information about the posterior range. This gives us an idea of the length of the proposals to get a good acceptance probability. Thus, we expect the sampling to be much easier.

In these situations NS will require a high number of iterations to find the posterior mass and then yield an estimate of the marginal likelihood. On the other hand, NIS will require much lower number of iterations for this task. This can be understood by noting that the integral for z is dominated by wherever the bulk of the posterior mass is to be found [Skilling, 2006], and it is actually where the pseudo-prior distribution optimally allocates its probabilities. In other words, the posterior probability mass has already been found at the beginning of the estimation process because NIS starts with an approximation of it. Thus, NIS would require less iterations than NS to completely explore the posterior bulk, independently of the kind of priors used.

Another positive attribute of NIS is its potential of possessing a lower uncertainty than NS. For both methods, the standard deviation of the estimates depends directly on the information H, which is a kind of measurement of the dissimilarity between the prior and the posterior in the NS case, or the reference distribution and the posterior in the NIS case. The higher the information is, the higher the uncertainty. In general, the prior is much more different from the posterior, which generates higher information in the NS case. In other words, after acquiring the data, our prior beliefs will change a lot, which can be seen as a high amount of information gained from the data. On the other hand, the information associated with NIS is with respect to the reference distribution, which is an approximation of the posterior. Therefore, this artificial information gained from the data is naturally smaller than in the NS case. It can be noticed from (3.23) that the information tends to 0 as the approximation of the reference distribution to the posterior improves. Thus, NIS uncertainty is potentially smaller than NS uncertainty.

Finally, NIS provides posterior samples at no extra cost, like NS. The procedure is analogous to the original algorithm. The samples can be drawn from the discarded points according to the corresponding pseudo-posterior weights $\Psi_i g_i$ with $i = 1, \ldots, k$. Note that these weights are indeed proportional to posterior weights because they are equivalent to $L_i \pi_i$ (likelihood × prior).

3.5 Application

The methods described previously are performed in two different contexts. First, they are applied to two statistical scenarios where many methods are unable to estimate the normalizing constant accurately. Second, the methods are evaluated in two phylogenetic situations: i) a real dataset is analysed to select among diverse evolutionary models and then, with the chosen model, we test the sensitivity of the estimation methods to different prior specifications; and ii) a simulated dataset in the four taxon case is analysed in order to detect the true phylogeny by using direct Bayes factor estimation.

3.5.1 Statistical example

Skilling [2006] pointed out that path sampling does not work when the likelihood, as a function of the cumulative prior probabilities, is not concave (see Figure 3.8). Such a problem seems like it ought to be easy to solve, but it can be intractable for PS. Actually, these problems are intractable for most known methods based on power posteriors in their simple form (path connecting the prior and posterior), including steppingstone sampling or annealed importance sampling. In this situation, a slight change in β (see the definition of the power posterior density in (3.6)) leads to a big change in the distribution, phenomenon known as *phase transition*. We study Skilling's example referred as Model 1 and also a variant of it, named Model 2. The latter model presents problems to even generalized forms of power posterior methods, but not for nested sampling.

3.5.1.1 Model 1

Consider the d-dimensional parameter vector $\boldsymbol{\theta}$ with a uniform prior in the unit cube $[-0.5, 0.5]^d$, and likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{d} \frac{1}{v\sqrt{2\pi}} \exp\left(- \frac{\theta_i^2}{2v^2} \right) + 100 \prod_{i=1}^{d} \frac{1}{u\sqrt{2\pi}} \exp\left(- \frac{(\theta_i - \mu)^2}{2u^2} \right). \tag{3.24}$$

Skilling [2006] considered the following values: $d = 20$, $\mu = 0$, $v = 0.1$ and $u = 0.01$. The likelihood function is the sum of two Gaussians (3.24): the first is a Gaussian of width 0.1 and the second Gaussian has a factor of 100 and a width 0.01 that is superposed on the first one. The likelihood function is a relatively flat density with a spike in its center. An unscaled representation of this function in one-dimension is shown in Figure 3.6. Independently of the dimension d, the center peak μ and the variances u and v, the marginal likelihood is 101. Skilling assessed this problem analytically whereas we do it in a practical way.

We assess the marginal likelihood estimation by using the methods presented in this chapter: HM, PS, SS, AIS, GSS, GAIS, NS and NIS. The power posterior methods in their simple form are assessed under an annealing ($\mathrm{PS_a}$, $\mathrm{SS_a}$ and $\mathrm{AIS_a}$) and a melting ($\mathrm{PS_m}$, $\mathrm{SS_m}$ and $\mathrm{AIS_m}$) scheme. Their generalizations are evaluated under the latter (GSS and GAIS). The estimation process is replicated 1,000 times for each method. In addition, we assess the uncertainty of the methods which succeed in this situation.

To be fair in the comparison, we use around 100,000 samples for each method. For the HM, after a burn-in period of 500, a sample is taken every 100 iterations in a Markov chain of length 10,000,000. The slice sampling algorithm [Neal, 2003]

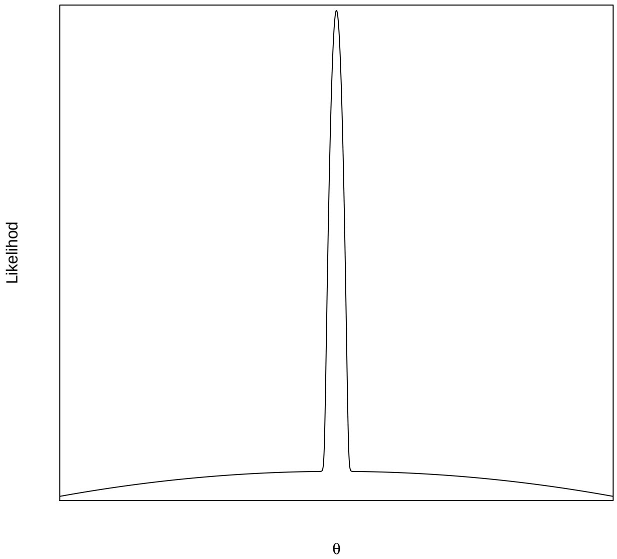


Figure 3.6: Likelihood function with two phases: a plateau and a spike. The plot is not–scaled but it represents, didactically, the likelihood function in one dimension of the model defined in (3.24).

is used for this procedure. Thus, a total of 100,000 points are used per estimate.

For TI, 50 transition distributions and 2,000 samples from each of them are taken. After a burn-in period of 500 in the prior or the posterior ($\beta = 0$ and $\beta_1 = 1$, respectively), every point is taken after 100 iterations using slice sampling [Neal, 2003]. The final point at each level is taken as starting value in the sampling of the next transition density. This yields a total of 100,000 samples for PS. The β values follow a Beta$(0.3, 1)$ distribution [Xie et al., 2011]. SS and AIS follow the same design, but they do not require samples from the posterior. For GSS and GAIS, we consider 48 transitional distributions and 2,000 posterior samples to calibrate the reference distribution which is defined below. Their β values are uniformly spaced. Thus, SS, AIS and their generalizations are assessed under the same number of sample points. This number is slightly less than the total number of samples used for HM and PS.

The performance of NS and NIS is assessed for 1300 and 4200 active points, respectively. These values have been calculated by running NS methods and seeing if they converge before 100,000 samples. Thus, these specifications make them require similar computational effort to the other methods evaluated. A more formal way of finding the number of iterations required by NS is explained below. Each proposal value from the restricted prior was drawn using the Metropolis algorithm [Metropolis et al., 1953]. After randomly selecting a starting point that meets the likelihood restriction, 100 iterations are run before choosing the new value. For NIS, GSS and GAIS, the reference distribution is a uniform, like the prior, but parametrized by a posterior sample. This is done by calculating the absolute maximum value x_{M} of 1,000 posterior samples and then using it as the upper and in its negative form as the lower bound. Thus, the reference distribution is a uniform prior in the unit cube $[-x_{\mathrm{M}}, x_{\mathrm{M}}]^{20}$. This is updated in each of the 1000 replications.

Figure 3.7 shows the box plots for the 1,000 estimates of the log-marginal likelihood. The horizontal dotted line depicts the true value $\log(101) = 4.62$. HM overestimates the true value because the posterior distribution is dominated by the spike, ignoring the flatter Gaussian. Under the annealing scheme, PS, SS and AIS underestimate the evidence due to the nature of the power posteriors. For some transition distributions, the power posterior consists of a mixture of the narrow and the broad Gaussians where both components should contribute to $\mathbb{E}_{p_\beta}\big[\log L(\boldsymbol{X}|\boldsymbol{\theta}, M)\big]$. However, the sampler gets trapped in the wide component and cannot find the narrow component because its volume is so small. Something similar happens with these methods under the melting scheme, but in this case the sampler gets trapped in the spike making it unable to mix both components well. As a result, the true value is underestimated and overestimated under the annealing and melting schemes, respectively.

On the contrary, GSS, GAIS, NS and NIS work perfectly. For GSS and GAIS, the reference distribution encapsulates the spike and concentrates its probability mass around it. Thus the probability of going from the spike to the reference distribution (under the melting scheme) increases, allowing the mixing of the two components of the likelihood. On the other hand, NS uses the likelihood contours regardless of their shape, allowing it to deal well with phase transitions. Figure 3.8 shows the relationship between the prior mass and the likelihood function estimated by NS. The phase transition is clearly identified. NIS works similarly, but

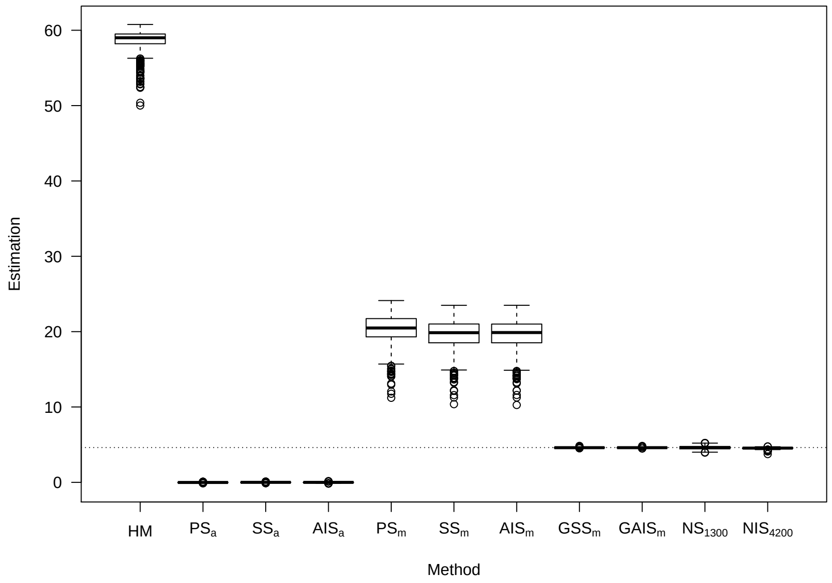


Figure 3.7: Performance of HM, PS, SS, AIS, GSS, GAIS, NS and NIS for the partly convex likelihood function in the statistical example. The subscripts "a" and "m" depict the annealing and melting scheme used by the power posterior methods, respectively. The subscripts in NS methods depict the number of active points. The horizontal dotted line stands for the true log-marginal likelihood value.

it starts the exploration from the pseudo-likelihood which makes it require less iterations and yield lower uncertainty. The exploration of the parameter space is displayed in Figure 3.9. On the bottom right part of the graph, we can see that the spike of the likelihood is found at the very beginning with only few iterations. Even though it can be noted from Figure 3.7 that the NS and NIS estimates have higher variabilities than the generalized power posterior methods, they only need one single run to estimate also their uncertainty, unlike GSS or GAIS.

GSS and GAIS uncertainties

GSS and GAIS proved to be very accurate methods in this situation. However, it is not possible to detect a difference in performance between them from Figure 3.7 due to the scale of the y-axis, so we assess it separately in this section. In

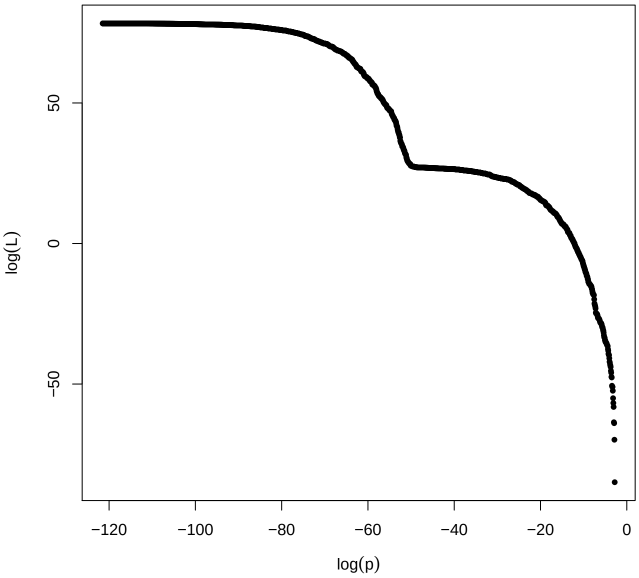


Figure 3.8: Estimated log-prior mass vs log-likelihood under NS with only 16 active points.

particular, we evaluate their performance in terms of their uncertainty under the same computational effort. The marginal likelihood is estimated 500 times for 4, 8, 16, 32 and 64 transitional distributions (K) with 1,000 samples (n) from each of them. The same samples are used for both methods at each replication. Then, the standard deviations of the estimates are calculated per each method.

The results are visualized in Figure 3.10. GSS yields clearly a lower uncertainty than GAIS for all the different specifications, even though the same amount of information is used. The difference between the standard deviations becomes much smaller as the number of transitional distributions increases.

GSS has as tuning parameters the number of transition distributions K and the samples from each of them n. These parameters have a different impact on the uncertainty of the estimate. To study this, 500 log-marginal likelihoods are estimated under different conditions by using GSS and then their standard deviations are calculated. This procedure is carried out for all the combinations of the following paired specifications: for K and n equal to 4, 8, 16, 32, 64, and 100, 200,

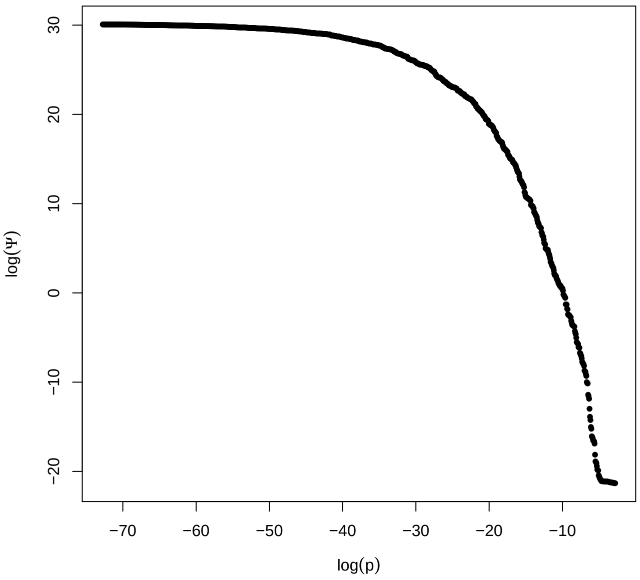


Figure 3.9: Estimated log-pseudo-prior mass vs log-pseudo-likelihood under NIS with only 16 active points.

400, and 1000, respectively. The results are presented in Figure 3.11. Evidently, the increase in K makes the standard deviation decrease more significantly than N. For instance, the pairs $(K = 8, n = 100)$ and $(K = 4, n = 200)$ have exactly the same computational effort, but they possess different uncertainties. The first pair has a standard deviation approximately 38% lower than the second one. The first pair even has a lower standard deviation than the pair $(K = 4, n = 400)$ which has twice the computational effort. This pattern is observed for all the combinations analysed, but the standard deviation differences decrease as the computational effort increases.

NS, NIS and GSS uncertainties

NS, NIS and GSS can deal efficiently with the phase transition of this statistical model. The two approaches carry out the estimation in a very different way by construction. Putting them on the same ground, we study their performance according to the number of likelihood evaluations required to obtain a determined

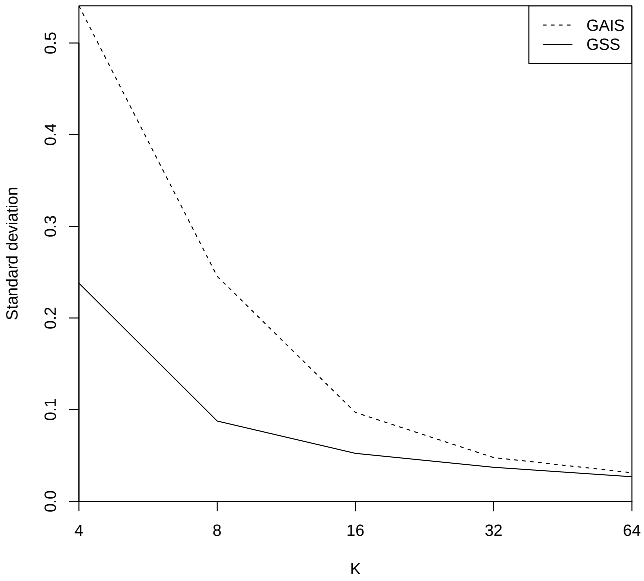


Figure 3.10: Standard deviation of 500 log-marginal likelihood estimates for 4, 8, 16, 32 and 64 transitional distributions (K) with 100 samples (N) from each of them. A continuous line is used to illustrate the trend. GSS has a lower standard deviation than AIS despite the equal computational effort. The uncertainty and the difference between the methods become smaller as the number of transitional distributions increases.

standard deviation. The comparison is carried out by using GSS as reference. We note that GSS outperforms GAIS in this settings, so it is excluded.

NS starts the exploration from the Uniform prior until it finds the spike. This is done without being helped by any reference distribution, unlike NIS and GSS. It can be noted from Figure 3.8 that NS requires around 100 times the number of active points (N) iterations to explore the full parameter space. To understand this, note that $-i/N$ is the estimated log-prior mass explored at the i^{th} iteration which is represented at the horizontal axis of the plot. So, to equate this quantity to -100, the position where the function gets stabilized from the right to the left in Figure 3.8, we need to consider $i = 100 \times N$ iterations. For instance, if we use 2 active points, 200 iterations will be enough to reach the $\log(p) = -100$ area in the horizontal axis of the plot. The larger the number of active points, the

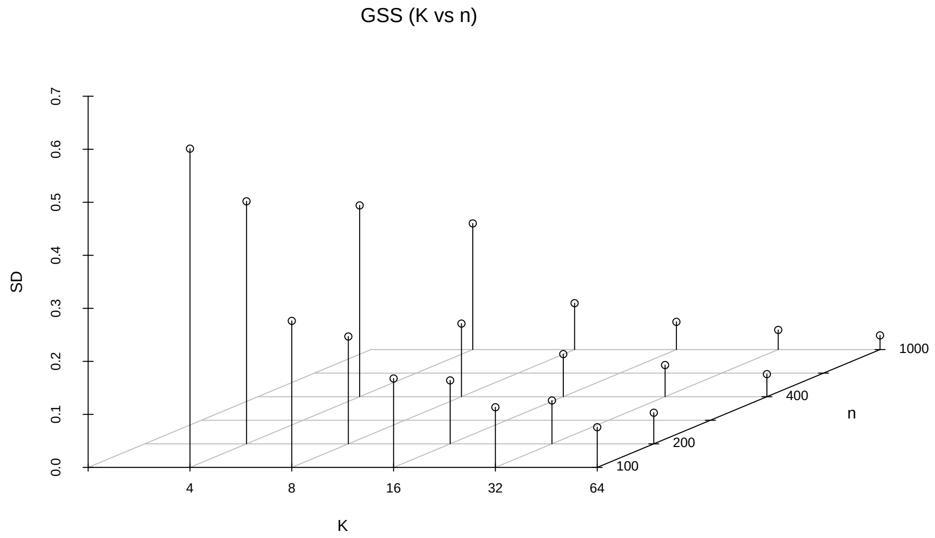


Figure 3.11: Standard deviation of 500 log-marginal likelihood estimates for different number of transition distributions K and samples from each of them n. It seems that K has a higher impact on decreasing the uncertainty than n.

larger the number of iterations required by NS to reach that area. Analogously, Figure 3.9 shows that NIS requires at most $70 \times N$ iterations to converge. This is absolutely linked to the reference distribution used. Having all this information, i.e., the number of iterations and active points, and the posterior samples required by NIS, we can calculate the number of likelihood evaluations required by NS methods.

We estimate the log-marginal likelihood by using GSS for K = 4, 8, 16, 32, 64 transitional distributions with 100 samples from each of them. We replicate this analysis 500 times and then calculate the standard deviation for the different specifications. This allows us to relate GSS uncertainty with the number of likelihood evaluations.

Taking the GSS standard deviations as references we calculate the number of likelihood evaluations required by NS and NIS to equate them. Then, we calculate the odds of likelihood evaluations. The results are displayed in Figure 3.12. It is clear that NS requires much more likelihood evaluations to equate GSS and NIS. For instance, it needs 148 times more evaluations than GSS to get a standard deviation of 0.076. NIS also requires more computational effort, but much less

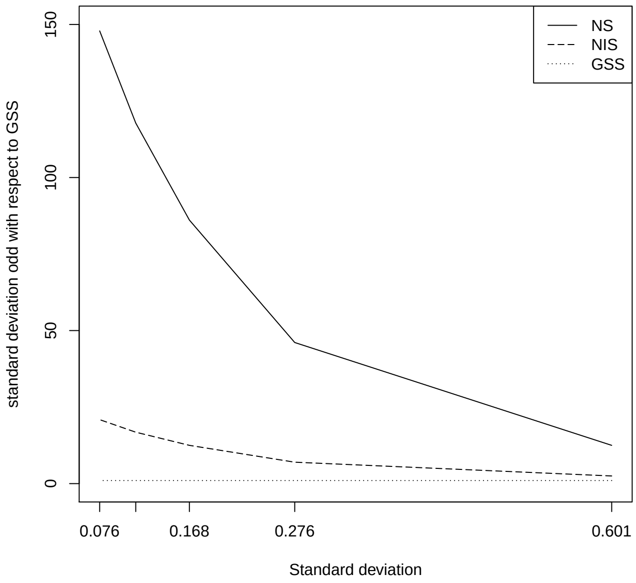


Figure 3.12: Ratio of NS and NIS standard deviations with respect to the one obtained by using GSS.

than NS. In this case, it needs 21 times more evaluations than GSS to get that standard deviation. The difference of likelihood evaluations between the methods decreases as the standard deviation increases. We do have to highlight that even though NS and NIS require more likelihood evaluations to obtain a similar GSS uncertainty, they only require one single run to yield an estimate of their uncertainty. In contrast, GSS uncertainty is estimated, in practice, by replicating several times the estimation process, which can be carried out only after the tuning parameters are found to yield reliable marginal likelihood estimates.

3.5.1.2 Model 2

In the previous model, GSS performs very well, despite not using an optimal reference distribution. A tentative option had been the use of a Normal distribution. We test this alternative with 100 transitional distribution and 100 samples from each of them. The estimation is replicated 1,000 times and the results are displayed in Figure 3.13. Even though the estimates are fairly close to the true

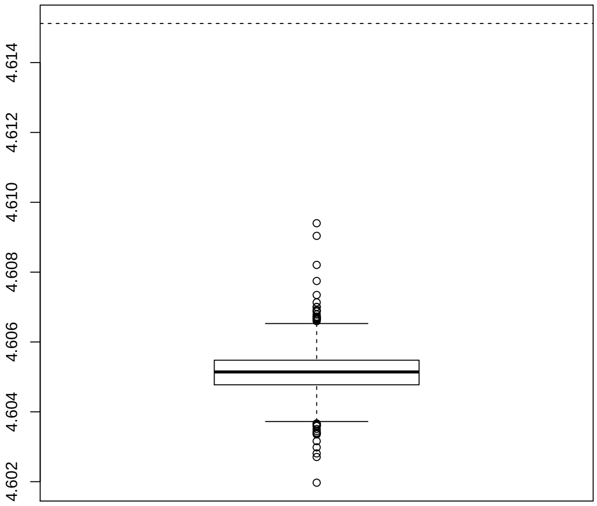


Figure 3.13: GSS estimates using 100 transitional distributions and 100 samples from each of them. A normal distribution is used as reference distribution. The dashed line at the top stands for the true value.

value, 0.01 away from it on average, GSS underestimates it. This can be explained by the fact that the reference distribution restricts the exploration of the parameter space to its domain, around the peak. In general, this distribution is approximately centered around 0 with a standard deviation of 0.01. So, 99.7% of the samples are between -0.03 and 0.03. However, the plateau Gaussian allocates its probability between -0.3 and 0.3 which is in part outside the range of the reference distribution and eventually will be unexplored by it. As a result, the reference distribution leaves some areas of low probability without consideration in this model, which explains the underestimation.

This phenomenon leads to the natural question: what would happen if the unexplored area had more probability? In order to study this situation, we consider Skilling's model but change the weight of the narrow Gaussian. As before, consider the d-dimensional parameter vector $\boldsymbol{\theta}$ with a uniform prior in the unit cube $[-0.5, 0.5]^d$, but now a likelihood given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{d} \frac{1}{v\sqrt{2\pi}} \exp\left(- \frac{\theta_i^2}{2v^2} \right) + \prod_{i=1}^{d} \frac{1}{u\sqrt{2\pi}} \exp\left(- \frac{(\theta_i - \mu)^2}{2u^2} \right), \tag{3.25}$$

where $v = 0.1$, $u = 0.01$, $\mu = 0$, and $d = 20$. This likelihood has a flat density with a spike in its center. In Skilling's model, the spike is 100 times over the flat distribution, but in our model the spike is not as tall as in his model. Thus, the tails contain more probability. Independently of μ, d, u, and v, the marginal likelihood is 2.

We assess the marginal likelihood estimation by using GSS, NS and NIS. For the first method, we use 1,000 posterior samples to parameterize the normal distribution, 100 transitional distributions and 100 samples from each of them. This yields a total of 10,000 samples to calculate the GSS estimate. This is without considering the initial posterior samples. We use slice sampling to generate the samples. For NS, we use 99 active points which makes it require around 10,000 samples as well. This number is calculated as for the previous model. NIS is performed by arbitrarily using 99 active points (following NS settings) and collecting 1,000 posterior samples to calibrate its reference distribution. Also, we include the NS estimate with only a single point to evaluate its performance in the simplest condition. The estimations are replicated 1,000 times and the results are shown in Figure 3.14.

The approximation of the reference distribution for this model is determined by the starting point in the Markov chain. If it is around 0, the distribution will be approximately a N(0, 0.01), but if it is a little distant from its center, let's say outside the approximated interval $[-0.023, 0.023]$, the distribution will be approximately a N(0, 0.1). Actually, they are the narrow and the plateau normal distribution which compose the likelihood. In our analysis, we use 0 as starting value as it is the center of the target distribution.

GSS underestimates the true value which is $\log(2) = 0.693$. Its estimates are around 0 with a standard deviation of 0.001. Its reference distributions are centered approximately around 0 with a standard deviation of 0.01. Consequently, they restrict their samples to the interval $[-0.03; 0.03]$ excluding those areas which

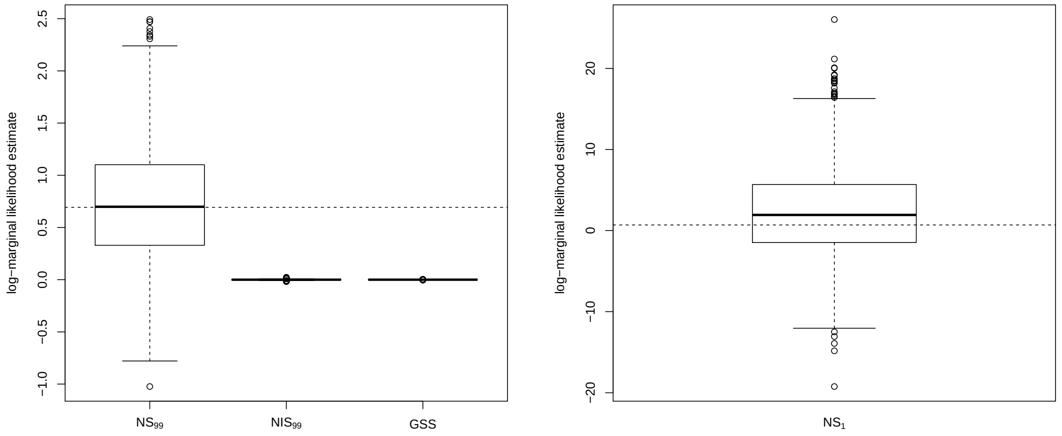


Figure 3.14: GSS, NS and NIS log-marginal likelihood estimates. The first two methods are assessed under approximately the same number of sampled points. The subscript of NS methods depicts the number of active points. The right plot shows NS performance with a single active point. The horizontal dashed lines stand for the true value.

now have a significant amount of probability mass. This is clearly illustrated in one dimension in Figure 3.15 (right graph). These significant areas are excluded from the marginal likelihood estimation. This is the reason of the underestimation which is now much more severe than in the previous model. NIS possesses the same problem, since it also relies on a reference distribution. On the other hand, NS estimates, with 99 active points, are around the true value with a sample standard deviation of 0.57. Even in its simplest case, with a single active point, its estimates are around the true value with a sample standard deviation of 5.85. NS also provides accurate posterior samples which cannot be obtained by conventional MCMC methods (see Figure 3.15), such as Metropolis-Hasting or slice sampling.

3.5.2 Phylogenetic analysis

We present the performance of the methods discussed in this chapter with two phylogenetic examples. Firstly, we carry out model selection by using NS for a

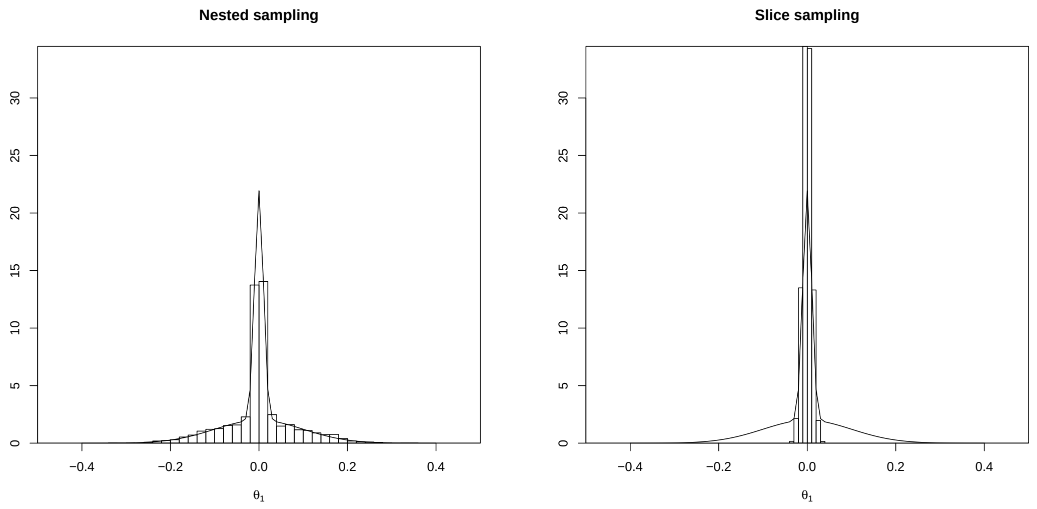


Figure 3.15: The histograms stand for 10,000 posterior samples for the first component of $\boldsymbol{\theta}$ by using nested sampling on the left, and a slice sampling on the right. This behavior is similar in all the components of $\boldsymbol{\theta}$. The continuous lines depict the true marginal density of Model 2.

real dataset of 10 green plants assuming a known phylogeny. Then, HM, PS, SS, AIS, GSS, NS and NIS are evaluated in terms of their sensitivity under different prior specifications. Furthermore, NS and NIS uncertainties are compared.

Finally, we present an example to illustrate the case of direct Bayes factor estimation. For this, a simulated dataset for the four taxon case is used to select between two phylogenies. The true tree is allocated in the Felsenstein zone [Huelsenbeck and Hillis, 1993] generating a good scenario to assess method performances. Model selection is carried out by using independent and direct Bayes factor estimation via HM, SS and NS.

3.5.2.1 Marginal likelihood estimation

We studied a dataset previously analysed by Xie et al. [2011]. This dataset has 10 species of green plant and its phylogeny, shown in Figure 3.16, is uncontroversial. The DNA sequences are from the chloroplast-encoded large subunit of the RuBisCO gene (*rbc*L). We use NS to estimate the marginal likelihood for 6 different evolutionary models and select the best among them. For this, we consider

50 active points for each estimation and use the termination criterion based on the amount of information H. Finally, we assess the sensitivity of the reviewed methods under 3 different priors: diffusive, in contradiction and in agreement with the likelihood function.

Model selection

The 6 models of DNA evolution to be compared are: JC69, JC69+Γ_4, HKY85, HKY85+Γ_4, GTR and GTR+Γ_4. Regarding the prior distributions, we consider the following hierarchical structure for the branch lengths:

$$\begin{aligned} t_i|\mu &\sim \text{Exp}(1/\mu), \quad \text{for} \quad i = 1, \ldots, 17, \\ \mu &\sim \text{Inverse-Gamma}(\alpha^*, \beta^*), \end{aligned}$$

with μ a scale/mean parameter. Rannala et al. [2011] suggested $\alpha^* = 3$ and $\beta^* = 0.2$ in order to prevent an overestimation of the branch lengths. The simplest model JC69 only has these free parameters. For the relative rates we use

$$\begin{aligned} r_i|\phi &\sim \text{Exp}(\phi), \\ \phi &\sim \text{Exp}(1), \end{aligned}$$

where ϕ is a rate parameter, $i = 1$ for the HKY85 models and $i = 1, 2, 3, 4, 5$ for the GTR models. The remaining parameters for each model are set to 1. A Dirichlet(1,1,1,1) is selected for the joint prior of the four nucleotide frequencies $(\pi_A, \pi_C, \pi_G, \pi_T)$ in the HKY85 and GTR models. Finally, we use a Gamma(α, β) for the shape parameter (λ) of the four-category discrete gamma distribution of rates across sites. We use $\alpha = 1$ and $\beta = 1000$, which is equivalent to an Exponential(0.001), as non-informative prior for λ in the JC69+Γ_4, HKY85+Γ_4 and GTR+Γ_4 models. The densities are fully defined in Section 2.4.

The results are displayed in Table 3.2. The GTR+Γ_4 model has the best fit among the models compared. This conclusion is fairly reliable since this model is around 15 standard deviations from the next competing model HKY85+Γ_4. These results also show the importance of considering heterogeneous substitution

Model	$\log \widehat{z}$	SD
JC69	-7260.19	0.76
JC69+Γ_4	-6903.39	0.84
HKY85	-7042.32	0.89
HKY85+Γ_4	-6618.76	0.97
GTR	-6982.79	0.99
GTR+Γ_4	-6603.67	1.04

Table 3.2: Log-marginal likelihood and standard deviation estimates under different substitution models for the green plant rbcL data. GTR+Γ_4 has the highest marginal likelihood.

rates among sites. For instance, the log-marginal likelihood increases drastically from GTR to GTR+Γ_4. We observe the same situation for JC69 and HKY85 models.

Taking advantage of the posterior samples yielded by NS analysis for the GTR+Γ_4 model, we estimate the branch lengths by using the marginal posterior means. They are displayed in Figure 3.16. The proportions of the tree are very similar to the one estimated by maximum likelihood presented in Xie et al. [2011].

Sensitivity

Xie et al. [2011] assessed the performance of HM, PS and SS under different prior specifications for a given phylogeny. The marginal likelihood itself is sensitive to the prior, in that it can depend strongly on the prior, so one would expect the methods are too. In their study, they showed that PS and SS are sensitive to prior specifications, unlike HM. In our study, we follow the same experimental design by using the same different priors for the shape parameter of the discrete gamma distribution of rates across sites, but considering a distinct evolutionary model and prior distributions for the remaining parameters. We estimate the marginal likelihood for the selected GTR+Γ_4 model through HM, PS, SS, AIS, GSS, NS and NIS.

The models are just differentiated by 3 prior distributions placed on the shape parameter of the discrete gamma distribution of rates across sites. The remaining

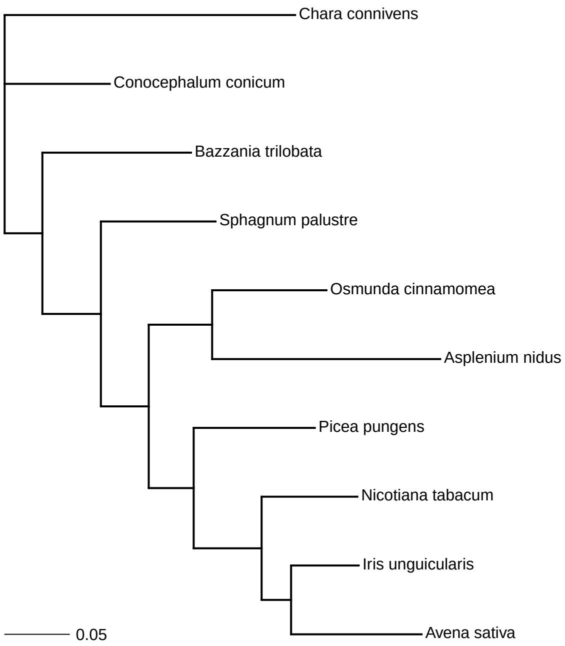


Figure 3.16: Tree topology for the rbcL data. The branch lengths are estimated by using the posterior means under the GTR+Γ_4 model.

priors are the same used in the model selection analysis presented before. The 3 priors are the following: Exponential(0.001) is the "vague" prior which has variance 1 million; Gamma(10, 0.026) is the "good" prior which is centered around its posterior mean, 0.26; and Gamma(148, 0.00676) is the "wrong" prior which is centered arbitrarily at 1. The names "good" and "wrong" are just labels which are related to the relationship between the information contained in the data and in the prior. The "wrong" prior is in contradiction with the likelihood function (i.e., the prior density and the likelihood function peak in different regions), unlike the "good" prior. While the "vague" and the "wrong" priors seem qualitatively different, whether a prior appears vague or wrong can depend on the coordinate system. In any case, the "good" prior should lead to a better marginal likelihood. Figure 3.17 illustrates the idea of the 3 different priors.

For HM, after a burn-in period of 2,000 cycles, we take a sample of 50,000 points. The samples are taken every 100 cycles. For PS, we use 50 transition distributions characterized by the β values which are chosen according to the evenly

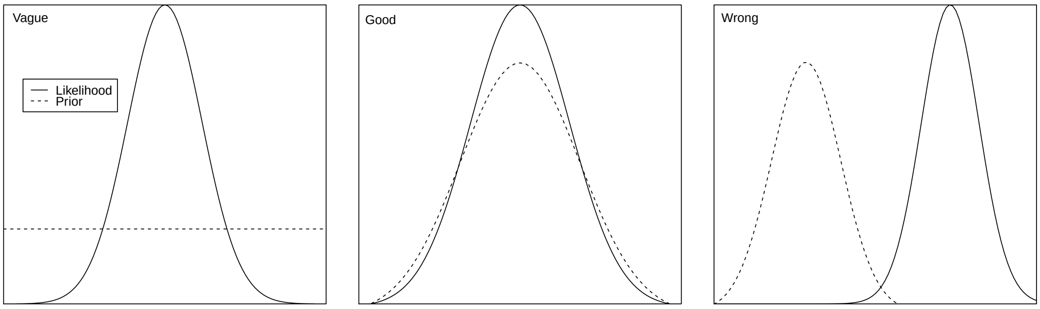


Figure 3.17: Three possible priors (dotted lines) in comparison with the likelihood function (solid line). The "vague" is a flat density, the "good" is consistent with the likelihood whereas the "wrong" is in contradiction with it.

spaced quantiles of a Beta$(0.3, 1.0)$ distribution. For each of them, following a burn-in period of 2,000 cycles, a chain is run for 100,000 in parallel. Samples are taken every 100 cycles after the burn-in period. This gives a sample of 1,000 for each transition distribution and 50,000 in total. Similarly, but without considering the posterior distribution, SS and AIS just use 49 transition distributions making a total of 49,000 samples. For GSS, we follow the same design as for SS and AIS, but the β values are equally spaced. The parametrization of its reference distributions is described below together with those used by NIS.

The tuning parameters used in the power posterior methods could not be the optimal ones, but it is the way to guarantee a reliable estimate. Thus, the main purpose of these analysis, to test sensitivity of the methods to prior specifications, can be carried out. On the other hand, we use 50 active points for NS and NIS which are enough to yield reliable estimates in this case.

For NIS and GSS, the prior is parametrized by a pilot posterior sample of 1,000 points and used as an importance sampling distribution, as proposed by Fan et al. [2011]. The marginal posterior sample means $(\widehat{\mu}_\theta)$ and variances $(\widehat{\sigma}^2_\theta)$ are matched to the respective prior parameters in order to calibrate the reference distributions. The prior for each branch length is equivalent to a Gamma$(1, \mu)$ distribution which is used as the density to be reparametrized. Hence, the mean $\mathrm{E}[t] = \tilde{\alpha}\tilde{\beta}$ and variance $\mathrm{Var}[t] = \tilde{\alpha}\tilde{\beta}^2$ (from $t \sim$ Gamma$(\tilde{\alpha}, \tilde{\beta})$) are matched to the posterior mean μ_t and variance σ^2_t, respectively, leading to a Gamma$(\widehat{\mu}^2_t/\widehat{\sigma}^2_t, \widehat{\sigma}^2_t/\widehat{\mu}_t)$ as reference distribution. For its hyperparameter μ, Inverse-Gamma$(\widehat{\mu}^2_\mu/\widehat{\sigma}^2_\mu{+}2, \widehat{\mu}_\mu(\widehat{\mu}^2_\mu/\widehat{\sigma}^2_\mu{+}1))$, since $\mathbb{E}[\mu] = \beta^*/(\alpha^* - 1)$ and $\mathrm{Var}[\mu] = \beta^{*^2}/\big((\alpha^* - 1)^2(\alpha^* - 2)\big)$. For each relative rate parameter, the procedure is analogous to the one described for the branch lengths. The reference is a $\mathrm{Gamma}(\widehat{\mu}_r^2/\widehat{\sigma}_r^2, \widehat{\sigma}_r^2/\widehat{\mu}_r)$ distribution, where $\widehat{\mu}_r$ and $\widehat{\sigma}_r^2$ are the mean and variance from the posterior sample of the corresponding relative rate, respectively. Similarly, for its hyperparameter, the reference is a $\mathrm{Gamma}(\widehat{\mu}_\phi^2/\widehat{\sigma}_\phi^2, \widehat{\sigma}_\phi^2/\widehat{\mu}_\phi)$ distribution, where $\widehat{\mu}_\phi$ and $\widehat{\sigma}_\phi$ are the mean and variance of the posterior sample of the hyperparameter ϕ, respectively. For the frequencies, a $\mathrm{Dirichlet}(\alpha_A, \alpha_C, \alpha_G, \alpha_T)$ is used. The means $(\widehat{\mu}_A, \widehat{\mu}_C, \widehat{\mu}_G, \widehat{\mu}_T)$ and variances $(\widehat{\sigma}_A^2, \widehat{\sigma}_C^2, \widehat{\sigma}_G^2, \widehat{\sigma}_T^2)$ of the sampled base frequencies are used to define $\alpha_A = \widehat{m}\widehat{\mu}_A$, $\alpha_C = \widehat{m}\widehat{\mu}_C$, $\alpha_G = \widehat{m}\widehat{\mu}_G$, $\alpha_T = \widehat{m}\widehat{\mu}_T$, with

$$\widehat{m} = \frac{\sum_{i\in\mathcal{B}} \widehat{\mu}_i^2(1 - \widehat{\mu}_i)^2}{\sum_{i\in\mathcal{B}} \widehat{\sigma}_i^2\widehat{\mu}_i(1 - \widehat{\mu}_i)} - 1, \quad \text{for} \quad \mathcal{B} = \{A, C, G, T\},$$

where $\widehat{m}$ represents the least square estimate of the sum of all parameters [Fan et al., 2011]. Finally, for the gamma shape parameter a $\mathrm{Gamma}(\widehat{\mu}_\lambda^2/\widehat{\sigma}_\lambda^2, \widehat{\sigma}_\lambda^2/\widehat{\mu}_\lambda)$ is chosen, since $\mathbb{E}[\lambda] = \alpha\beta$ and $\mathrm{Var} = \alpha\beta^2$.

The marginal likelihood estimates under the 3 different priors are displayed in Table 3.3. As expected, HM is not able to discriminate between the “vague” and the “good” priors. Their influence is annulled by the sharp posterior distribution from where the samples were taken to calculate the estimate. On the other hand, PS, SS, AIS and GSS are capable to ponder these distributions adequately making them sensitive to prior specifications. Like these power posterior methods, NS and NIS also prove to be in accordance with them and consequently sensitive to prior specifications.

Despite the high precision, it is noteworthy to highlight some disadvantages of methods based on power posteriors in their simple form. The “vague” prior makes these methods, in their simple form like SS and AIS, require a high number of β values. For instance, 30 values are enough for the other priors, but this number is not enough for the “vague” prior and in practice it must be estimated by guesswork. This could make these methods impractical in many situations, e.g. when just extremely diffuse prior distributions are considered. In this analysis, we can rely on their estimates because they have been compared to those obtained by using the other methods. However, these make the power posterior methods

Method	Prior model		
	Vague	Good	Wrong
HM	-6560.6	(-0.1)	(-35.7)
PS	(-8.8)	-6594.9	(-59.4)
SS	(-8.6)	-6594.7	(-59.4)
AIS	(-8.4)	-6594.9	(-59.0)
GSS	(-8.3)	-6594.7	(-59.1)
NS	(-9.3)	-6594.4	(-59.0)
NIS	(-8.5)	-6594.8	(-59.4)

Table 3.3: Estimated log-marginal likelihood values for the GTR+Γ_4 model under 3 different priors. The highest marginal likelihood value is displayed whereas the difference with the other models are shown in parenthesis.

depend on several replications in order to be confident about the reliability of the estimate when they are used as the only method of estimation.

On the other hand, for NS we only need to specify the number of active points and steps to generate the samples (also required by posterior sample methods) in order to produce an estimate of the marginal likelihood. From the single run, we also obtain their uncertainties which are displayed in Table 3.4. Even though NIS requires posterior samples to calibrate its reference distributions, like GSS, it produces lower uncertainties than NS. It also has a more stable behavior in terms of its uncertainty along the 3 different priors.

The precision of their estimates depends exclusively on the number of active points. The higher the number the more accurate the estimate. A pilot NS run can also be used to get a controlled uncertainty in a new estimate. For instance, for the "good" prior the estimated information is $H = 44.72$, so if we need to derive the number N of active points to get an uncertainty of at most 0.5, we need to solve $\sqrt{H/N} = 0.5$, which calculates the standard deviation. Thus, NS requires $N \geq 179$ active points to produce the desired uncertainty. Analogously, NIS can be specified to produce an estimate with a determined precision.

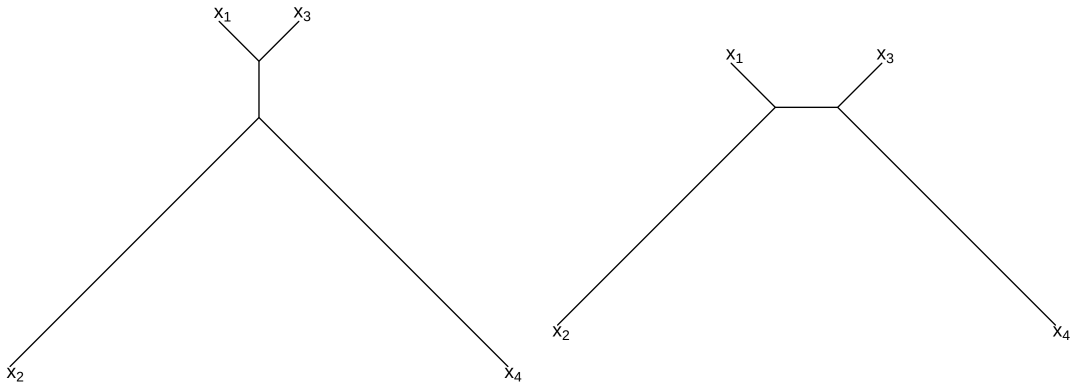
	Prior		
Method	Vague	Good	Wrong
NS	1.04	0.95	1.25
NIS	0.29	0.30	0.28

Table 3.4: Estimated standard deviation for NS and NIS estimates under the 3 different priors. In terms of their uncertainty, NIS outperforms NS in all the scenarios.

3.5.2.2 Bayes factor estimation

Long branch attraction is a systematic error which happens when two distantly related taxa are incorrectly inferred to be closely related [Bergsten, 2005]. In such cases, the rate of evolution has been accelerated for certain branches in different clades. Several methods, most prominently maximum parsimony, tend to group these branches together in a clade based on their amount of evolutionary change instead of its descendant relationship. This makes them fail in identifying the true phylogeny. Even under maximum likelihood or Bayesian approaches this phenomenon can arise, for instance, when an inadequate evolutionary model is used. The long branch attraction problem has been widely studied [Felsenstein, 1978; Gaut and Lewis, 1995; Huelsenbeck and Hillis, 1993]. In this context, we assess HM, SS and NS performance by using marginal likelihood and direct Bayes factor estimation in a simulation study for the 4 taxon case.

Consider the unrooted tree for 4 taxa given in Figure 3.18b named as Tree 1. The two short external branches are separated by a short internal branch whereas the long branches are in different clades. There are evident unequal rates of change along different branches. This is the typical problem of long-branch attraction. A dataset that consists of four taxa with 10,000 sites is generated from this phylogeny. The sequences are evolved along the branches on this fixed tree. The short branches have length 0.02 ($t_1 = t_3 = t_5 = 0.02$) and the long ones 0.74



(a) Tree 0.

(b) Tree 1.

Figure 3.18: Two phylogenies in the 4 taxon case. The data has been generated from Tree 1.

$(t_2 = t_4 = 0.74)$. The rate matrix is

$$\mathbf{R}^* = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} - & & & \\ 1 & - & & \\ 7 & 2 & - & \\ 2 & 8 & 1 & - \end{pmatrix} \end{array}$$

The rates strongly favour transitions over transversions. The base frequencies are equal ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$). Also, a variable rate across sites has been considered which follows a Gamma distribution with shape 0.5 and scale parameter 2.0. This leads to a mean of 1.0 and variance of 2.0.

Tree 1 (Fig. 3.18b), the model from where the data is generated, is compared to Tree 0 (Fig. 3.18a). Many methods incorrectly infer Tree 0. They group the long branches together into a clade because just by chance the same mutation will occur on both terminal branches and is considered a shared trait.

We carry out model selection via Bayes factor. This (BF_{10}) will show evidence in favour of Tree 1 if it is positive. The strength of the evidence can be categorized according to Table 3.1. The GTR+Γ_4 model is used for the analysis of the dataset with the following priors: $(v_1, v_2, v_3, v_4, v_5) \sim \text{Dirichlet}(1, 1, 1, 1, 1)$ and

Method	$\log \widehat{z}_1$	$\log \widehat{z}_0$	dif	$\log\left(\widehat{\mathrm{BF}}_{10}\right)$
HM	-33785.20	-33790.04	4.84	5.16
SS	-33824.78	-33829.15	4.37	4.46
NS	-33824.11	-33827.19	3.08	4.04

Table 3.5: Independent and direct Bayes factor estimation via HM, SS and NS. “dif” depicts the independent Bayes factor estimate, i.e., the difference between the log-marginal likelihood for Tree 1 and Tree 0.

$T \sim \text{Gamma}(1,1)$ which combined generate the prior for the branch lengths $t_i = v_i T$, for $i = 1,2,3,4,5$, where T is the tree length, $\text{Dirichet}(1,1,1,1,1,1)$ for the joint prior on the six GTR relative rates, $\text{Dirichet}(1,1,1,1)$ for joint prior on the four nucleotide frequencies, and $\text{Gamma}(1,1)$ for the shape parameter of the four-category discrete gamma distribution of rates across sites. We assume same priors for both phylogenies.

For HM, 50,000 samples are considered. For SS, we consider 500 transitional densities with 100 points from each of them. Thus, both methods are using the same number of samples. For NS, we use 100 active points and the termination criterion based on the information H.

The results presented in Table 3.5 show that all the methods infer correctly Tree 1. According to the criterion given in Table 3.1, all of them detect at least strong evidence against Tree 0, with exception of the direct Bayes factor estimate via HM, which yields very strong evidence against Tree 0. This method also produces much higher marginal likelihoods than SS and NS. However, the difference of these estimates, the log-Bayes factor, is coincidently quite close to those yielded by SS and NS. Even though the marginal likelihood estimate produced by NS for Tree 0 is slightly higher than the SS, the resulting evidence against this model falls into the same category, strong evidence. In general, these methods produce similar results.

The amount of information H for Tree 1 and Tree 0 are 45.98 and 44.25, respectively. The amount of information in the direct BF estimation is 12.37. The information values involved in the single estimates are higher than the one for the direct estimate. Having these values, we calculate their respective standard deviations for the BF estimate. Following equation (3.21), the standard deviation for the BF estimate using independent NS estimates is $\sqrt{(45.98 + 44.25)/100} = 0.95$. On the other hand, the direct Bayes factor estimation via NS yields 0.35 which is a lower uncertainty using the same number of active points.

3.6 Discussion and conclusions

Bayes factor (BF) is one of the most popular methods for model selection under a Bayesian perspective. This quantity depends on a ratio of multidimensional integrals called marginal likelihoods. In general, but especially in phylogenetics, these integrals involve high complexity requiring numerical approximations. This chapter presented several methods to estimate the Bayes factor in phylogenetics. The discussion was mainly centered in the independent estimation of the marginal likelihood and the direct estimation of the Bayes factor, given an evolutionary tree.

Among the methods of estimation, we discussed the harmonic mean (HM), path sampling (PS), steppingstone sampling (SS), generalized steppingstone sampling (GSS), annealed importance sampling (AIS), nested sampling (NS) and nested importance sampling (NIS). We evaluated their performance in statistical settings and in phylogenetic scenarios. Especially, we assessed the performance of NS algorithms in phylogenetics.

Even though AIS and SS were developed independently from different backgrounds, they are closely related. Indeed, they are equivalent when only one transition distribution ($K = 1$) or one single sample ($n = 1$) are considered. For the former case, both methods reduce to the arithmetic mean. In a more general scenario, we have shown that the SS estimate contains the AIS estimate as one of its multiple components. Despite their similarity, they differ in their performance. For instance, their estimates have different uncertainties in practice. We illustrated this aspect in a statistical model with a phase transition presented in 3.5.1.1. Their generalizations GSS and GAIS, under exactly the same computational effort, yielded different standard deviations. GSS had a lower uncertainty under a variety of specifications. More study is needed to be able to generalize this observation.

Many of the methods currently used in phylogenetics, such as HM, PS and SS, do not work in some scenarios. The problems related to HM are well documented, unlike those related to power posterior methods. In fact, these latter are often presented as methods of general applicability [Arima and Tardella, 2014; Baele and Lemey, 2014; Baele et al., 2013; Xie et al., 2011]. However, as Skilling [2006] pointed out for the case of PS and AIS, they fail at mixing between the different phases of the likelihood, leading to poor estimates. Example 3.5.1.1 illustrated this problem and showed that SS fails in this situation. On the other hand, it was shown that the generalized versions of SS and AIS can deal efficiently with this likelihood shape by using an adequate reference distribution. Similarly, NS versions also had no problems in this situation. These algorithms work by tracking down the likelihood contours independently of its shape.

However, GSS and NIS performances are significantly determined by the reference distribution which is in practice constructed by posterior samples. In the case that this distribution cannot be accurately estimated, the methods fail. Example 3.5.1.2 illustrated this situation. In this scenario the posterior was dominated by a spike around 0, but its tails had areas of significant probability which could not be accessed via conventional MCMC methods. As a result, the posterior samples were concentrated around its mean making the reference distribution be constrained to the center of the posterior distribution (see Figure 3.15). This prevented GSS and NIS from exploring those areas of the parameter space which could contribute with probability to the marginal likelihood estimation. This explains why these methods underestimated the true value. On the other hand, NS estimates were located around the true value. This even happened in the simplest condition, that is, with one active point ($N = 1$). The method works by starting the estimation from the prior toward those areas of high likelihood. These areas are encapsulated according to the likelihood contours, not its shape. Thus, NS was also able to yield posterior samples which are displayed in Figure 3.15. In contrast to standard MCMC methods, the NS sample points were drawn from both Gaussian components of the posterior distribution.

In general, the likelihood surface can be very complex, especially in phylogenetics. However, its nature in this field is still not sufficiently understood. To provide an understanding of it, the work has been mainly oriented on finding multiple likelihood maxima [Rogers and Swofford, 1999; Steel, 1994]. For instance, Chor et al. [2000] described some situations, even biologically reasonable, where the sequence data can lead to continuum of points, all of them with the same maximum likelihood value. Another approach has been to investigate its shape in the one dimensional case [Fukami and Tateno, 1989]. In this regard, Dinh and Matsen [2015] proved that under simple evolutionary models (K80) the one dimensional likelihood function may have multiple stationary points which indicates that under advanced evolutionary models the shape can be more complex than it is typically assumed.

The lack of understanding of the likelihood surface in phylogenetics is a potential problem for statistical inference. We have shown that HM, even methods which are believed to be of general applicability, such as PS and SS, can fail in estimating the marginal likelihood in presence of phase transitions in the likelihood. The presence of this phenomenon has also been noticed in phylogenetics. Mossel and Steel [2007] illustrated this situation in the case of estimating the root state on a phylogenetic tree without enough information. They also indicated the relevance to biology of these phase transitions, especially in recovering information deep within the tree, from the character states observed at the leaves. We have also shown that the calibration of the reference distribution can be badly affected by a poor approximation of the posterior. This makes methods of high precision, such as GSS or NIS, fail. Examples of complex posterior shapes can be found in the literature, see for instance Nylander et al. [2004]. Hence, statistical inferential methods must be general enough to deal with these, usually, hidden behaviors in the parameter space. Especially, methods of marginal likelihood estimation must be able to deal efficiently with these latent obstacles which can lead to poor inferences.

In example 3.5.2.1, we have assessed the performance of the different methods presented in this chapter in a phylogenetic context under a fixed topology. The dataset contains 10 species of green plant and was previously used by Xie et al. [2011]. We used NS to select among 6 evolutionary models. The marginal likelihood estimates with their corresponding uncertainties provided all the information required to carry out model selection. Then, having chosen GTR+Γ as evolutionary model, we tested the sensitivity of the estimation methods to 3 different prior specifications on the shape parameter of the discrete gamma distribution of rates across sites. These were: flat, in contradiction and in agreement with the likelihood. In concordance with our expectations, we found that HM is insensitive to these specifications, unlike PS and SS. Furthermore, we found that GSS, AIS, NS and NIS are sensitive to this kind of priors.

All the methods discussed in this chapter allow direct Bayes factor estimation. In general, this is a more efficient way of comparing two models in terms of uncertainty [Baele et al., 2013; Lartillot and Philippe, 2006], because otherwise the estimation must be carried out by estimating independently the marginal likelihoods, which is consequently affected by two sources of error. For this purpose, we have given an expression for its direct estimation via HM. This estimator has the characteristic of reducing the potential sources of infinity of its variance to only one. This represents a good alternative over independent HM estimates when one of them has infinite variance. In addition, we have extended NS for the direct BF estimation. In the case that the models are defined on the same parameter space, this method could potentially yield a much lower uncertainty than the BF calculation by using independent estimates via the original algorithm. This extension also represents an efficient alternative to the direct BF estimation via SS, which is extremely sensitive to its specifications in terms of bidirectional error [Baele et al., 2013].

We showed the performance of the direct Bayes factor estimation methods in model selection through an example. This example (3.5.2.2) is a simulation study that illustrated the case of topological comparison of two small trees of four taxa. The methods were performed in the critical situation of long branch attraction. All of them had a good performance detecting the correct phylogeny. Again, HM overestimated the marginal likelihoods, but the resulting Bayes factor was coincidently similar to those obtained via SS or NS. Its direct estimate also was higher than the rest of the estimates. Our proposal, the direct BF estimate by using NS, yielded a much lower uncertainty than the calculation through independent estimates via NS. It is worthwhile to highlight that these uncertainties were calculated in a single run.

The applications presented in this chapter are limited to the case of fixed topology. However, it could be of interest to incorporate phylogenetic uncertainty into the model selection process. This extension is addressed in the next chapter.

Chapter 4

Variable tree topology nested sampling

4.1 Introduction

Phylogenetic inference is commonly carried out under the frequentist approach by finding a preliminary phylogeny using a fast tree-building approach, such as neighour-joining or parsimony, as a first step. Then, the evolutionary model is chosen based on this initial fixed topology. Finally, having chosen the model, the parameter space is explored in order to find the maximum likelihood estimate. This includes topology, branch lengths, and the parameters associated with the evolutionary model. This practice has as an objective the reduction of computational time [Minin et al., 2003]. However, this model selection procedure leaves out phylogenetic uncertainty.

For the model selection stage, there are many competing statistical tools, but in particular the Bayes factor is one of the techniques that has become a standard method in phylogenetics. It allows comparison of nested and non-nested models, assessment of conclusions to prior distributions, and accommodation of topological uncertainty. A set of algorithms to estimate this quantity through either marginal likelihood or direct estimation was discussed in the previous chapter. We included the most popular methods used in phylogenetics, such as harmonic mean [HM; Newton and Raftery, 1994], thermodynamic integration [TI; Lartillot and Philippe, 2006], steppingstone sampling [SS; Xie et al., 2011] and generalized steppingstone sampling [GSS; Fan et al., 2011]. Additionally, we evaluated nested sampling [NS; Skilling, 2006] in a phylogenetic context. So far, these methods were limited to the fixed topology case.

However, it might be unwise to base model comparison only on the merits of a single phylogeny, even if it is the optimal one. This is a common practice under the frequentist approach. Ideally, we would like to study molecular evolution while at the same time incorporating inherent phylogenetic uncertainty. Actually, this is one of the attributes of Bayes factor and should be exploited. On the other hand, systematics, in general, use Bayesian inference methods to estimate the tree topology and thus the evolutionary model should be selected by considering phylogenetic uncertainty. Following this line, GSS has been extended to allow variable tree topology [Holder et al., 2014] and also within a coalescent-based framework [Baele et al., 2016].

In Chapter 3, nested sampling was introduced as a reliable method to estimate the marginal likelihood, but it was also discussed how it can be utilized to sample the posterior distribution. In this chapter, we take advantage of its properties to use it in model selection and parameter inference allowing variable tree topology. We also describe how an importance sampling distribution can be utilized in this algorithm in order to improve its performance. The necessary elements required to apply NS methods in this situation are discussed throughout this chapter. The algorithms are performed in different situations, which are illustrated in the application section.

4.2 Total marginal likelihood

The definition (3.2) of the marginal likelihood presented in Chapter 3 applies to a continuous parameter space Θ. This is the case when the phylogeny is fixed. However, when the tree topology is unknown, this is treated as any other parameter. To account for this, the parameter space is additionally composed by a discrete part. In this case, Bayes' theorem is given by

$$p(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau | \boldsymbol{X}, M) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau, M)\pi(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M)}{m(\boldsymbol{X}|M)}, \tag{4.1}$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector composed by elements such as frequencies, gamma shape parameter and rates parameters, $\boldsymbol{t}_\tau \in \mathcal{V}_\tau$ is the set of branch lengths of a tree topology $\tau \in \mathcal{T}$, $\boldsymbol{X}$ is the sequence data, M is the substitution model, $p(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|\boldsymbol{X}, M)$ is the posterior distribution, $L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau, M)$ is the likelihood function, $\pi(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M)$ is the prior distribution and $m(\boldsymbol{X}|M)$ is the marginal likelihood, which is defined as

$$m(\boldsymbol{X}|M) = \sum_{\tau \in \mathcal{T}} \int_{\mathcal{V}_\tau} \int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau, M)\pi(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{t}_\tau. \tag{4.2}$$

The marginal likelihood now incorporates the sum over the tree parameter space and is known in this case as *total marginal likelihood*. This quantity can be seen as an average of all possible marginal likelihoods corresponding to the different topologies weighted by their tree prior probabilities. Hereafter, it will also be denoted by "z".

The posterior probability for a model M is given by

$$p(M|\boldsymbol{X}) = \frac{m(\boldsymbol{X}|M)\pi(M)}{m(\boldsymbol{X})},$$

where $m(\boldsymbol{X}|M)$ is the marginal likelihood defined in (4.2) and plays the role of likelihood in this posterior probability, $\pi(M)$ is the prior probability for the model, and $m(\boldsymbol{X})$ is the probability of the data. This posterior probability for the evolutionary model takes into account the phylogenetic uncertainty, since it averages over the whole tree parameter space. This quantity is particularly useful in the comparison of two models, which can be carried out by means of the ratio of their posterior probabilities. This is

$$\frac{p(M_1|\boldsymbol{X})}{p(M_0|\boldsymbol{X})} = \frac{m(\boldsymbol{X}|M_1)}{m(\boldsymbol{X}|M_0)} \frac{\pi(M_1)}{\pi(M_0)}$$

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds.}$$

The ratio of marginal likelihoods is called the Bayes factor [Kass and Raftery, 1995]. This quantity depicts the strength of the evidence of model 1 over model 0. A qualitative interpretation for the evidence is given in Table 3.1.

Nested sampling (NS) is a Bayesian algorithm which is mainly addressed to estimate marginal likelihoods but it also provides the means to sample the posterior distribution. It transforms the multi-dimensional integral that defines the marginal likelihood, into a one-dimensional integral which relates prior mass to likelihood values. The estimation process is carried out by exploring the parameter space from the prior distribution toward those areas of high likelihood values. This algorithm is defined and explained in detail in Section 3.4. That definition of the algorithm is also valid for this case of variable tree topology, which is the central theme of this chapter. However, the extension of NS methods to this case brings with it new challenges into Bayesian phylogenetics, such as, sampling a dynamic tree distribution, or the construction of a reference distribution for the tree topologies for nested importance sampling (NIS). These subjects are addressed below.

4.3 Nested importance sampling

NS starts the estimation process from the prior distribution and over time encapsulates those areas of high likelihood values. This could be computationally expensive, especially in cases where the prior is vague or in contradiction with the likelihood. In these scenarios in particular, an importance sampling distribution can be used to make the estimation process more efficient. The resulting algorithm is called nested importance sampling [NIS; Chopin and Robert, 2010; Skilling, 2006] and has the potential of requiring less computational effort to accurately estimate the marginal likelihood than NS.

Equivalently to definition (4.2), the total marginal likelihood can be reformulated by adding an auxiliary distribution as follows

$$z = \sum_{\tau \in \mathcal{T}} \int_{\mathcal{V}_\tau} \int_{\Theta} \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau, M)\pi(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M)}{g(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M)} g(\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau|M) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{t}_\tau \tag{4.3}$$

$$= \sum \int \Psi(\boldsymbol{\eta}) g(\boldsymbol{\eta}|M) \mathrm{d}\boldsymbol{\eta},$$

where $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{t}_\tau, \tau)$ and $\Psi(\boldsymbol{\eta}) = L(\boldsymbol{X}|\boldsymbol{\eta}, M)\pi(\boldsymbol{\eta}|M)/g(\boldsymbol{\eta}|M)$, with g an importance sampling density. The functions Ψ and g play the role of pseudo-likelihood and pseudo-prior, respectively.

The description of the NIS algorithm given in Section 3.4.2 is still valid when investigating variable tree topologies. Extending to variable tree topologies involves making NIS work on a parameter space that has both continuous and discrete states. This extension requires the construction of an importance sampling distribution for tree topologies. As was discussed in the previous chapter, the reference distribution should be an approximation of the posterior [Lefebvre et al., 2010]. Along these lines, Fan et al. [2011] proposed to calibrate an independent reference distribution (for a parameter or block of parameters), for instance, the prior, by matching its mean and variance to the corresponding ones obtained from a pre-existing posterior sample. This is easily done for the continuous parameters (see its calculation in Example 3.5.2.1). However, such an approach represents a challenge when it comes to tree topologies. In the next section, we propose an efficient alternative to deal with this issue.

4.3.1 Tree reference distribution

Fan et al. [2011] proposed, in the context of GSS, to generate the reference distribution g by parameterizing the prior according to a pilot posterior sample. This is done by matching the prior mean and variance with the ones obtained from the posterior sample in order to get the parameter values for the reference distribution. This procedure was used for GSS, AIS, and NIS in the Application section 3.5 presented in the previous chapter and it proved to be an efficient way of generating reference distributions. However, this method can only be used for the continuous parameters $\boldsymbol{\theta}$ and $\boldsymbol{t}_\tau$, and it is not applicable for the tree reference distribution, which is a nominal variable.

The tree reference distribution should be optimally a good approximation of the marginal posterior of tree topologies, like for the rest of parameters. This distribution should be able to assign probability to even those topologies which were not visited in the posterior sample. In this sense, the sample relative frequencies, typically used to approximate the posterior, are not adequate. One alternative proposed by Holder et al. [2014], it is the use of a reference distribution which assigns high probability to those trees containing splits considered important in the posterior sample. The authors showed its good performance in a 6 taxon example in the context of GSS. Here, we propose the use of conditional clade distributions (CCDs) to approximate the tree posterior distribution [Höhna and Drummond, 2012; Larget, 2013].

4.3.1.1 Conditional clade probability distribution

The simple relative frequency method assigns probability zero to those trees which were not observed. Among them, we find trees which contain impossible phylogenetic combinations, and consequently with fair probability zero, but also those ones which contain highly probable clades. This method allocates them all in the same category. This phenomenon raised the idea of defining clade probabilities in order to estimate tree probabilities more accurately.

The conditional clade probability distribution [CCD; Larget, 2013] provides an accurate estimation of the true posterior distribution and can be used to calculate the probability of any tree, whether or not it was included in the posterior sample. The method is based on the assumption of conditional independence of separated subtrees which is used to estimate the posterior probability of trees by using conditional clade probability distributions. Roughly speaking, this method considers the idea that clades in different parts of the phylogeny are approximately independent.

To understand how this method works, consider a dataset which contains 6 hypothetical taxa: A, B, C, D, E, and F. Also, a posterior sample which only contains two different trees (Tree 1 and Tree 2) which are displayed in Figure 4.1. The probability of the Tree 1 (τ_1) can be fully described by its clades (C), as the probability of the intersection of all its clades, that is,

$$\begin{aligned}\mathrm{P}(\tau_1) &= \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_3 \cap \mathrm{C}_4 \cap \mathrm{C}_5)\\ &= \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4)\mathrm{P}(\mathrm{C}_5|\mathrm{C}_4 \cap \mathrm{C}_2)\mathrm{P}(\mathrm{C}_3|\mathrm{C}_2 \cap \mathrm{C}_4 \cap \mathrm{C}_5),\end{aligned}$$

where C_1 has been omitted since it is the universe, the set of all taxa in the tree. The latter expression is obtained by grouping the clades with their sister clades, but omitting in the notation those clades which are a single taxon.

Conditional independence in this instance is described as the probability of that any given clade with a set of taxa, a, only depends on the next smallest clade with a given set of taxa, b, where $a \in b$ and $a \neq b$. For instance, C_5 only depends on C_4 and not on the rest (Figure 4.1). Assuming this conditional independence, the probability of τ_1 can be approximated as follows

$$\mathrm{P}(\tau_1) \approx \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4)\mathrm{P}(\mathrm{C}_5|\mathrm{C}_4)\mathrm{P}(\mathrm{C}_3|\mathrm{C}_2).$$

These probabilities are calculated by using the corresponding proportions from the posterior sample. Note that, $\mathrm{P}(\mathrm{C}_i|\mathrm{C}_j) = \mathrm{P}(\mathrm{C}_i \cap \mathrm{C}_j)/\mathrm{P}(\mathrm{C}_j)$. Analogously, the probability of Tree 2 (τ_2) is approximated by

$$\mathrm{P}(\tau_2) \approx \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4)\mathrm{P}(\mathrm{C}_7|\mathrm{C}_4)\mathrm{P}(\mathrm{C}_6|\mathrm{C}_2).$$

More generally, all these probabilities are of the form

$$\mathrm{P}(\text{left subclade} \cap \text{right subclade}|\text{clade}).$$

Note that $\mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4) = \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4|\mathrm{C}_1)$, since C_1 is certain for any tree with these 6 taxa; or $\mathrm{P}(\mathrm{C}_5|\mathrm{C}_4) = \mathrm{P}(\mathrm{C}_5 \cap \mathsf{F}|\mathrm{C}_4)$, because the occurrence of clade C_5, given the restricted universe C_4, implies that there is no other possible option for its sister clade as it is the single taxon F.

Figure 4.1 also displays a third tree (τ_3) which has not been observed in the posterior sample. However, this tree shares clades with τ_1 and τ_2. For instance, C_4 is in τ_1 and τ_2, and C_7 is only contained by τ_2. Similar to the previous calculations, the probability of τ_3 is approximated by

$$\mathrm{P}(\tau_3) \approx \mathrm{P}(\mathrm{C}_2 \cap \mathrm{C}_4)\mathrm{P}(\mathrm{C}_3|\mathrm{C}_2)\mathrm{P}(\mathrm{C}_7|\mathrm{C}_4).$$

This approximation is composed by elements which constitute the calculation of the probability of τ_1 and τ_2. Even though τ_3 was not observed in the posterior

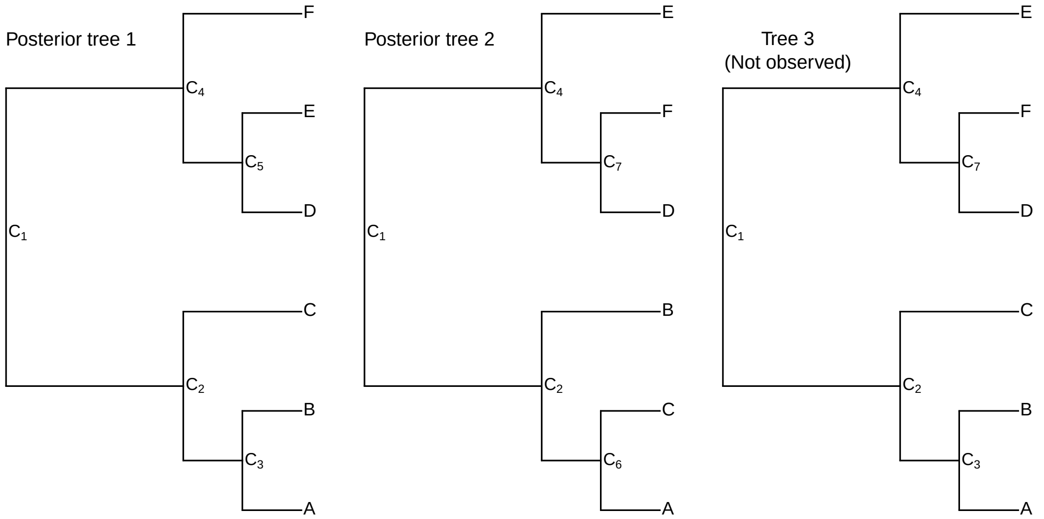


Figure 4.1: Trees for 6 hypothetical taxa: A, B, C, D, E, and F. Tree 1 and Tree 2 were visited in the posterior sample, unlike Tree 3.

sample, its probability can still be calculated mixing the corresponding elements from $P(\tau_1)$ and $P(\tau_2)$ approximations.

In general terms, this is how CCD works. The method has been described for rooted trees, but its application to unrooted trees is invariant to the location of the root. A computational algorithm to calculate it efficiently is described in detail in Larget [2013].

4.3.2 Branch length reference distribution

Strictly speaking, each tree has a unique set of branches. Consequently, different branch length reference distributions could be defined for each topology, analogously as proposed by Fan et al. [2011] for the continuous parameters. Those trees with low frequency in the posterior sample should be excluded and their prior distribution should be considered as importance sampling density instead, similarly to those which were not observed. This could prevent choosing a reference distribution in disagreement with the posterior caused by being constructed based on a non-representative sample.

An alternative, which does not exclude non-observed trees, is the use of branch length reference distributions according to subtrees. This method generates auxiliary densities similarly to the previous proposal but for subtrees. This has the potential of allowing the use of reference distributions even for some branches of unobserved trees. In the CCD example, Tree 3 is not visited in the posterior sample, but the reference distributions for the branch lengths of its subtrees, which contain C_2 and C_4, can be calibrated by using the corresponding branch lengths of posterior samples τ_1 and τ_2, respectively. For the remaining branch lengths, which cannot be matched to τ_1 or τ_2, the prior can be used.

More generally, for those subtrees which are contained in different trees of the posterior sample, a common branch length reference distribution can be constructed according to the posterior samples. The construction of these distributions should start from smallest to the biggest subtree. Thus, if a subtree contains another subtree with a reference distribution for its branch lengths already constructed, we only need to construct the reference distribution for its remaining branch lengths. For the same reasons explained for the method above, those subtrees with low frequency in the posterior sample should be excluded.

In general, each topology would need its own branch length distribution. Considering the size of the space of topologies, this seems to be a huge endeavor. In fact, exploring the space of tree topologies is the most complex process that MCMC methods have to deal with, within a phylogenetic context. Not using a reference distribution for the branch lengths seems to be a better strategy considering that the number of possible specifications outweighs the benefits.

4.4 Sampling

Skilling [2006] proposed a Metropolis-Hastings algorithm to sample the points required by NS. This can be done by choosing one of the surviving points randomly, namely, those that meet the likelihood restriction, as a starting value for performing a Metropolis-Hastings algorithm. However, an additional likelihood restriction is applied to this algorithm: the proposal point must have a likelihood higher than the lowest one associated with the active points in the previous iteration. This procedure is used in this work for the continuous parameters and also for the tree topologies. Special care must be taken in case of using NS to explore just the tree topologies for a fixed mutation model (Cf. p96-98 of Murray [2007]).

For the continuous parameters, the proposal moves can be generated as in any other statistical model where NS is applied. In this thesis, we use proposal moves for single parameters as proposed in Brewer and Foreman-Mackey [2016]. The new proposal is generated by adding a perturbation to the current value. This perturbation is composed by the product of a measure of the prior width (or the reference distribution in the case of NIS) and the random component $10^{1.5-3|t|}x$, where t has a student-t distribution with 2 degrees of freedom and x a standard normal distribution. Thus, the proposal distribution has heavy tails, a condition which makes the proposals as efficient as slice sampling [Neal, 2003], at least in simple experiments [Brewer and Foreman-Mackey, 2016].

For the tree topologies, as far as we know, the particular kind of sampling required by NS is a new way of exploring the tree parameter space and thus makes the tree proposals face new challenges. These proposals, in general, deal with the most difficult parameter space to sample from and affect the efficiency of the sampling directly. Lakner et al. [2008] assessed the performance of many of these mechanisms in standard Bayesian MCMC analysis. However, this study was limited to the sampling of the posterior distribution, a case in which the target distribution is static over time. Unlike these proposal mechanisms in standard MCMC methods, in NS these mechanisms have to deal with a variable target distribution over time. This is a new scenario for them and represents a challenge. NS compresses the prior at each iteration, making it vary generally at an estimated constant rate. Therefore, the proposals have to explore a quite wide area at the beginning, but becomes constrained over time. Tree proposal mechanisms should ideally be able to adapt to this sampling characteristic.

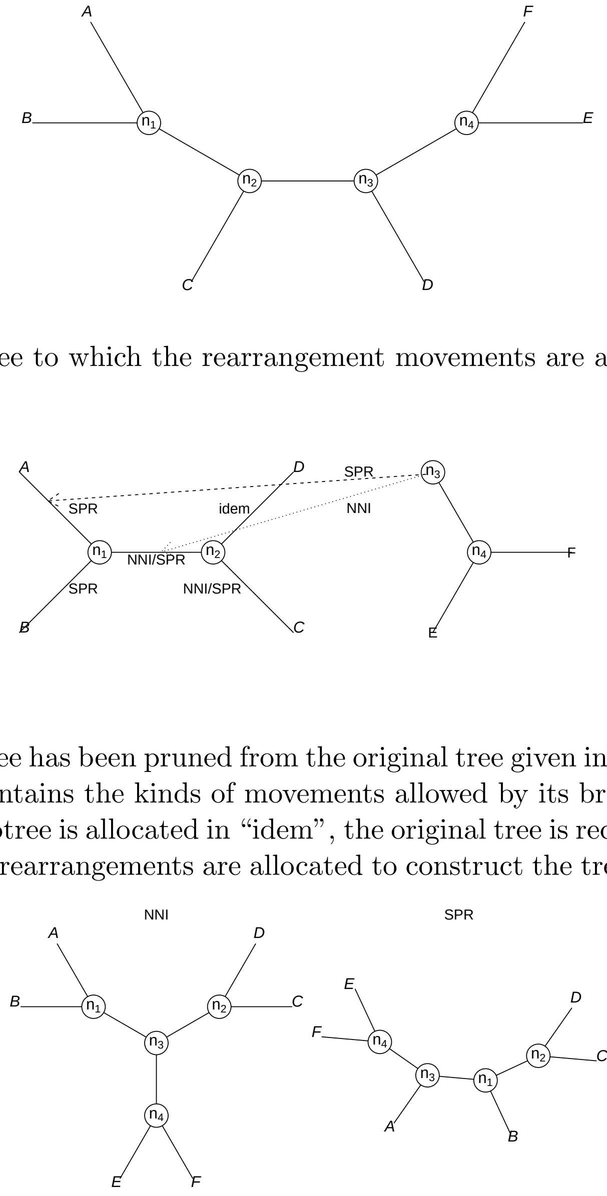
In practice, a uniform prior distribution is frequently assigned over the tree parameter space, which is quite huge even for a few taxa. Initially, bold moves would allow a better exploration requiring less steps than conservative ones. However, the acceptance probability would decrease drastically over time due to the fact that the target distribution becomes constrained. On the other hand, conservative moves would require more steps to explore the prior distribution at the beginning, but later on, the acceptance probability would be higher than for bold moves.

The boldest and the most conservative tree proposals examined by Lakner et al. [2008] are *Random Subtree Pruning and Regrafting* (rSPR) and *Stochastic Nearest Neighbour Interchange* (stNNI) methods, respectively. These methods work with branch rearrangement moves, i.e., two subtrees simply trade places. rSPR prunes a random subtree and regrafts it randomly into the remaining subtree. stNNI works similarly, but the picked subtree is allocated randomly in one of its nearest neighbors of the remaining tree. The Hastings ratio associated with both proposals is one [Lakner et al., 2008]. These tree rearrangements are illustrated in Figure 4.2.

To understand how these tree proposal mechanisms work, consider the 6 hypothetical taxa A, B, C, D, E and F, which are related according to the phylogenetic structure displayed in Figure 4.2a. We prune this tree generating the subtree (E,F), which is shown on the right in Figure 4.2b. The remaining tree is shown on the left, where the tree rearrangements allowed by its branches are indicated. For instance, NNI allows to reallocate the subtree in the two adjacent branches, which are indicated in the remaining tree. This movement is a particular case of SPR and thus both methods are indicated in these branches (NNI/SPR). As an example, we reallocate the subtree in the regrafting points which are pointed out by the arrows. For NNI, the regrafting point is in the branch connecting node 1 and 2, whereas for SPR, it is in the branch which leads to taxon A. The resulting trees are displayed in Figure 4.2c. The branch lengths are kept intact in our tree proposals.

The behavior of the proposal mechanisms in a nested sampling context is illustrated in Figure 4.3. This information has been obtained from the green plant rbcL data analysed in 4.6.2. At early iterations, both methods yield similar acceptance probabilities. However, the discrepancies are clear over time. From around the 500th iteration, rSPR starts to have a lower acceptance probability than stNNI. Even though their performance are different in terms of the acceptance probability, they behave very similarly in terms of convergence. We carried out some analysis (not shown), without finding significant differences.

Ideally, the proposal mechanism should take into account this dynamical behaviour of the target distribution over time. But instead of trying to adapt the



(a) Tree to which the rearrangement movements are applied.

(b) The right subtree has been pruned from the original tree given in (a). The remaining one, on the left, contains the kinds of movements allowed by its branches, to wit, NNI and SPR. If the subtree is allocated in "idem", the original tree is recovered. The arrows indicate where the rearrangements are allocated to construct the trees displayed below.

(c) Trees obtained by using NNI and SPR rearrangements. The regrafting points are indicated by the arrows in (b).

Figure 4.2: Tree rearrangement procedures in the 6 taxon case.

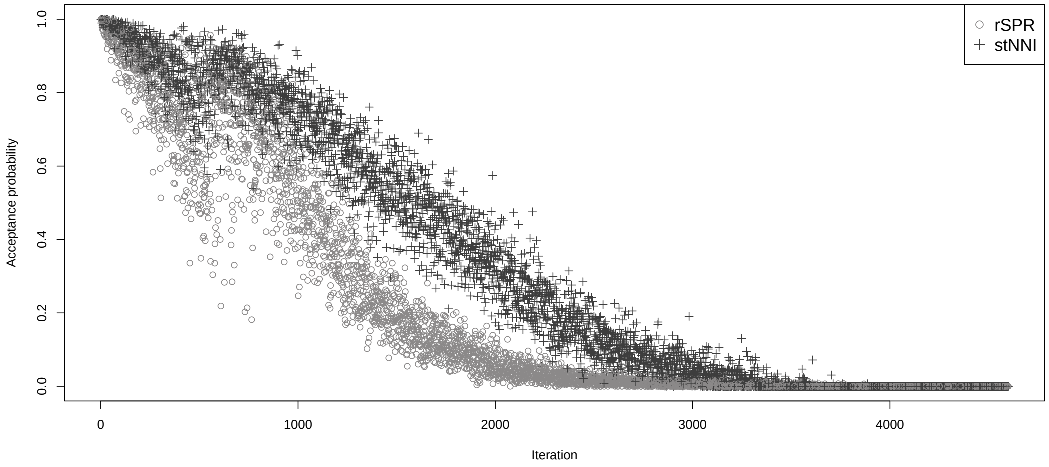


Figure 4.3: Comparison of rSPR and stNNI in terms of the acceptance probability in NS. This performance corresponds to green plant rbcL data analysed in Example 4.6.2.

proposal according to the restricted prior at each iteration, we propose to use a mix between bold and conservative moves generated by rSPR and the stNNI methods, respectively. Allowing bold movements over time could prevent the proposal tree from being stuck in a tree island, an isolated area of the tree topology space with a high probability mass. On the other hand, the conservative moves allow the exploration of the near areas of the current tree state, making the acceptance probability increase and thus allowing a good mixing. This approach can be seen as the recommended heavy-tailed proposals in the continuous case [Brewer and Foreman-Mackey, 2016].

4.5 Posterior samples

Nested sampling discards a point from the set of active points at each iteration. Thus, at the end of the estimation process, this method produces a sequence of discarded points with corresponding posterior weights $\tilde{p}_i = L_i w_i / z$ (see Section 3.4 for more details). The posterior sample can be drawn from this sequence according to these weights. This procedure can also be applied likewise in NIS.

NS offers an efficient way to carry out parameter inference. First, the posterior samples can be obtained at no extra cost from the marginal likelihood estimation process. Second, unlike MCMC approaches, NS does not require a burn-in period, i.e., the sequence of samples generated before the Markov chain had reached stationarity and subsequently discarded. These methods may require a long time to converge and thus represent a significant computational cost. The length of this period is not obviously determined and might be affected by many factors, such as proposal mechanisms [e.g., Lakner et al., 2008] and/or starting values. A common practice is to discard the first 10% of the chain as burn-in period [Huelsenbeck et al., 2004; Rambaut et al., 2014], but it might not be enough in some situations. Alternatively, a look at the trace plot might lead to a more informed when it comes to the burn-in. But again, the length of the trace plot could affect the assessment. Finally, NS allows the sampling of complex posterior shapes which could cause problems to standard MCMC methods. See, for instance, the example given in 3.5.1.2 in which some areas of the distribution are almost impossible to be reached by the standard methods.

4.6 Application

NS has proven to be a competent algorithm to estimate the marginal likelihood in the fixed topology case. We now assess it accommodating phylogenetic uncertainty in model selection and parameter inference. Firstly, an analysis of a dataset containing 3 species of primates is performed in order to assess the reliability of NS. Secondly, the analysis of the green plant rbcL data in model selection and sensitivity presented in the previous chapter is extended to the variable tree topology case. Additionally, NS is assessed in parameter inference, compared to NIS, and evaluated under the simplest specification in marginal likelihood estimation. Finally, a dataset containing 47 species of the Laurasetherian group is analysed to assess NS performance in parameter inference.

Specifications

We use Metropolis-Hastings to sample from the restricted prior distribution. The prior mass is estimated by using the geometric mean, the deterministic approach. We use 50% of rSPR and 50% of stNNI rearrangements as tree proposals in NS, NIS and GSS. The reference distribution for the tree topology is constructed by using the CCD method for NIS and GSS. Regarding the prior distributions, we consider the following hierarchical structure for the branch lengths:

$$\begin{aligned} t_i|\mu &\sim \text{Exp}(1/\mu), \quad \text{for} \quad i = 1, \dots, n, \\ \mu &\sim \text{Inverse-Gamma}(\tilde{\alpha}, \tilde{\beta}), \end{aligned}$$

where n is the number of branches for the particular tree topology, $\tilde{\alpha} = 3$ and $\tilde{\beta} = 0.2$. The JC69 model only has these free parameters. For the rate parameters we use

$$r_i|\phi \sim \text{Exp}(\phi), \;\; \text{with} \;\; \phi \sim \text{Exp}(1),$$

where $i = 1$ for the HKY85 models and $i = 1, 2, 3, 4, 5$ for the GTR models. A Dirichlet(1,1,1,1) is selected for the joint prior of the four nucleotide frequencies in the HKY85 and GTR models and a Gamma(α, β) distribution for the gamma shape parameter in the HKY85 + Γ_4, JC69 + Γ_4, and GTR + Γ_4 models. The parameters of the latter vary depending on the problem and are defined where applicable. Finally, we use a discrete uniform for the tree topologies.

4.6.1 Three primates

We analyse a small dataset, assuming the GTR+Γ_4 model, to evaluate NS performance in a controlled situation. The dataset comprises 15,727 sites from the mitochondrial DNA of 3 primates: human, chimpanzee and orangutan. This is a subset of a dataset which was previously analysed in the literature [Roos et al., 2011]. We use a Gamma(0.5, 1) distribution as prior for the shape parameter of the gamma distribution of rates across sites. The analysis is carried out under the molecular clock assumption.

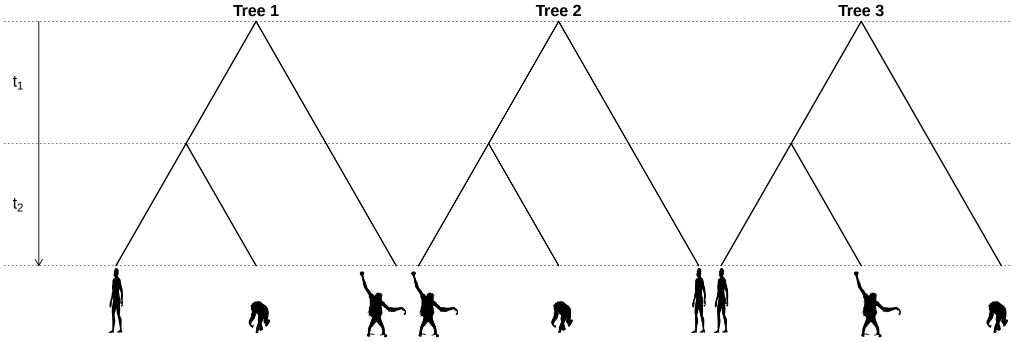


Figure 4.4: Three possible rooted trees for 3 primates species under molecular clock assumption. According to Tree 1, the species, from left to right, are: human, chimpanzee and orangutan.

Even though there is no analytical form of the total marginal likelihood for this simple example, this can be estimated by averaging all the individual marginal likelihood estimates of each possible topology. In general, this is not viable since the number of trees increases drastically as a function of the number of taxa. For 3 species, there are only 3 possible rooted trees. Figure 4.4 shows them with labels corresponding to the example. This allows us to calculate their individual marginal likelihoods and thus, estimate the total marginal likelihood. Therefore, we can evaluate and compare it with the estimate obtained by using its direct estimation via NS.

In this example, the estimated marginal likelihoods from the generalized steppingstone sampling of each phylogeny are used to calculate the total marginal likelihood. This estimate is considered as reference value and used to compare with the NS estimate. The 3 marginal likelihoods are estimated by using 500 samples from each of 200 transitional distributions. These specifications yield fairly reliable estimates with a very small variation (analysis not shown). Thus, the total marginal likelihood is calculated as a weighted average as follows

$$\widehat{z} = \sum_{i=1}^{3} \pi(\tau_i|M)\,\widehat{z}_{\tau_i},$$

where $\widehat{z}_{\tau_i}$ is the marginal likelihood estimate for its respective tree τ_i, with prior probability $\pi(\tau_i|M) = 1/3$ for the given substitution model $M = \text{GTR+}\Gamma_4$, with

$\log \widehat{z}_{\tau_1}$	$\log \widehat{z}_{\tau_2}$	$\log \widehat{z}_{\tau_3}$	$\log \widehat{z}$
-33067.81	-33468.71	-33468.41	-33068.91

Table 4.1: Estimated log-marginal likelihoods per phylogeny by using generalized steppingstone sampling. These values are used to estimate the log total marginal likelihood $\log \widehat{z}$.

$i = 1, 2, 3$. The results are presented in Table 4.1.

The log-total marginal likelihood has also been directly estimated via NS using 100 active points. The estimate is -33068.64 with estimated standard deviation 0.60. This estimate is consistent with the reference value estimated by using the 3 individual marginal likelihoods via GSS. Its uncertainty could be decreased by increasing the number of active points. The evidence is in favor of Tree 1 which depicts a well known phylogenetic relationship among these species [Roos et al., 2011].

4.6.2 Green plant rbcL

We study a dataset previously used by Xie et al. [2011] which contains 10 species of green plant. The DNA sequences are from the chloroplast-encoded large subunit of the RuBisCO gene (*rbc*L). This dataset was already used in the previous chapter to evaluate mainly NS performance, and also other methods, in model selection in the fixed topology case. The analysis also included an evaluation of the sensitivity of the methods to prior specifications, as in Xie et al. [2011]. Here, we extend these analysis to the variable tree topology case. We also assess NS ability to generate posterior samples for the tree topologies under different models of evolution. Furthermore, we compare NS to NIS, and evaluate its performance by using a single active point, which represents the case of its most basic specification.

4.6.2.1 Model selection

We compare 6 models of DNA evolution: JC69, JC69+Γ_4, HKY85, HKY85+Γ_4, GTR and GTR+Γ_4. We consider a broad Gamma$(1, 1000)$ distribution as

Model	$\log \widehat{z}$	SD
JC69	-7273.22	0.92
JC69+Γ_4	-6917.18	1.00
HKY85	-7057.45	1.04
HKY85+Γ_4	-6632.30	1.10
GTR	-6996.82	1.11
GTR+Γ_4	-6617.81	1.17

Table 4.2: Log-marginal likelihood and standard deviation estimates under different substitution models for the green plant rbcL data. GTR+Γ_4 model has the highest value.

prior for the gamma distribution shape parameter. The total marginal likelihood estimation is carried out for each model via NS with 50 active points.

Table 4.2 shows the log-evidence estimates and their respective standard deviations for the 6 different evolutionary models. The uncertainties are relatively small with respect to the marginal likelihood differences between the models. Thus, for any model comparison, the decision in favor of the model with higher marginal likelihood is quite reliable. The GTR+Γ_4 has the highest log-evidence value and consequently it fits the data better than the other models. Interestingly, the HKY85+Γ_4 has an evidence significantly higher than the GTR model, which has 4 more rate parameters but it does not include a variable rate across sites. This shows the importance of including a parameter that models the variability across sites for this dataset. The marginal likelihood estimates have been additionally compared to GSS in order to corroborate the model selection analysis. All the estimate values are quite similar and thus, both methods are in accordance (results not shown).

These results are consistent with those obtained in Section 3.5.2.1 (see Table 3.2). In that analysis, we calculated the marginal likelihood for the same models, but considering a fixed topology. If we ranked the evolutionary models according to their marginal likelihoods, we would obtain the same order as that obtained in this analysis, considering variable tree topologies. This might be explained by the fact that the phylogeny used in that analysis was the maximum posterior tree, an optimal one, and as Posada and Crandall [2001] have shown in the maximum likelihood approach, model selection, for the fixed tree topology case, is affected only when the phylogeny is randomly chosen. In this analysis the standard deviation estimates are higher because the tree topology is included as a parameter, which introduces more variability.

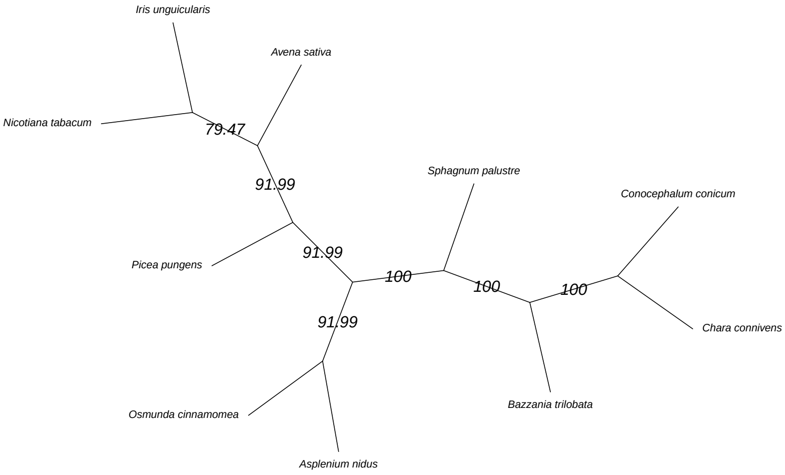
4.6.2.2 Phylogenetic inference

In this section, we assess the ability of nested sampling to sample the posterior distribution of the parameters involved in a phylogenetic analysis. In particular, we evaluate the posterior distribution for the tree topologies and the gamma shape parameter. As described in Section 3.4, a posterior sample is easily obtained from the discarded points generated at each iteration of NS, weighted by their contribution in the estimated evidence. To this end, we can simply continue from the model selection results in the previous section.

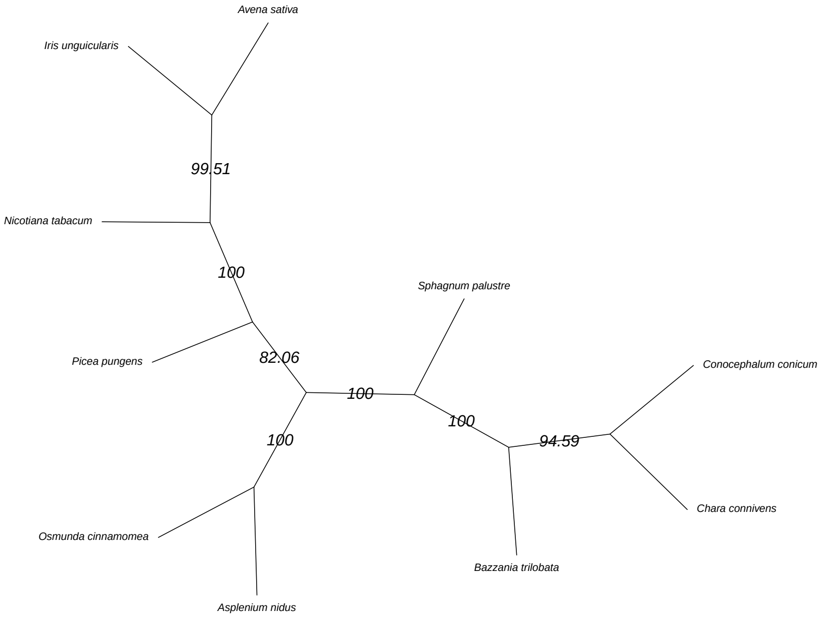
Phylogenetic reconstruction

The model selection procedure carried out previously via nested sampling left available samples from the tree posterior distribution for the different models of evolution evaluated in that analysis. We use them to study the capacity of these models to reconstruct the tree topology. We carry out this analysis by using consensus networks [Holland et al., 2004], method implemented in phangorn [Schliep, 2011].

The simplest model, JC69, is not able to recover the topology yielded by the more complex models. Its consensus network is displayed in Figure 4.5a. The subtree, which contains Avena sativa, Iris unguicularis and Nicotiana tabacum, groups these two last taxa into a single clade, unlike for the rest of the models. The rest of the phylogeny is equivalent to that obtained from more complex models. This behavior is repeated for JC69+Γ_4 (consensus network not shown). The rest of the models, that is, HKY, HKY+Γ_4, GTR and GTR+Γ_4, yields the same consensus network. The one for the latter model is shown in Figure 4.5b. All the consensus networks show no conflicting splits, so they are reduced to trees.



(a) JC69 model.



(b) GTR+Γ_4 model.

Figure 4.5: Consensus networks for the rbcL data calculated from NS posterior samples. They do not have conflicting splits, so they are reduced to trees.

GTR+Γ model

Model selection analysis identified GTR+Γ_4 as the evolutionary model best suited for these data. From it, we make use of NS posterior samples to carry out parameter inference. In particular, we focus on the tree topology and the shape parameter of the gamma distribution for the rates across sites. We also compare the posterior tree sample to another one which was obtained independently via MCMC.

Furthermore, we assess the NS tree posterior sample by calculating its effective sample size (ESS). This is a measure of the number of uncorrelated points which are needed to obtain the same information about a parameter as the one obtained from the MCMC. It is a very useful tool to assess the adequacy of posterior samples taken via MCMC analysis. However, its calculation is not direct for the tree sample. Lanfear et al. [2016] proposed to estimate it by randomly selecting a focal tree (from the posterior sample) and calculating the path differences [Steel and Penny, 1993] of the tree sample with respect to it. Thus, the topologies are converted into a continuous parameter. Then, the ESS is calculated from these distances. Replicating this procedure many times and registering the ESS values, provides the means to calculate a confidence interval.

We do have to mention that the analyses proposed by Lanfear et al. were addressed for the evaluation of tree samples obtained from standard MCMC methods. To wit, these analyses treat the sample as a time series. However, in NS case the sample points are not indexed by time, thus the appropriateness of these analyses could be argued. Nevertheless, we still use the ESS to assess the posterior sample with respect to this particular parameter.

NS analysis was performed with 50 active points and thus 4500 points were required to get a stable marginal likelihood estimate. For this single run, the maximum number of representative samples was 814. This represents 18% of the samples from the restricted prior required by the algorithm.

These sampled trees are shown in the first row of Table 4.3, with their respective proportions. This table also includes 1,000 posterior samples obtained independently via MCMC. The proportions are fairly similar. Both methods are in agreement inferring the posterior tree distribution. Tree 2, maximum posterior

Tree	1	2	3	4	5
NS	0.045	0.942	0.001	0.007	0.004
MCMC	0.042	0.945	0.000	0.009	0.004

Table 4.3: Proportion comparison of a tree posterior sample obtained via NS and MCMC methods.

tree topology, is equivalent to the consensus network displayed in Figure 4.5b. This is the one used in the previous chapter in the analysis under the fixed topology (see Figure 3.16). This also was used by Xie et al. [2011], but in their analysis was obtained by maximum likelihood method under the same evolutionary model.

In order to evaluate the quality of tree posterior samples obtained using NS, we calculate a confidence interval for its effective sample size (ESS). For this, we use 1,000 replications to generate it with a 95% of confidence. As a result, for the 814 posterior samples of tree topologies, the confidence interval for its ESS is $[652 - 725]$. According to the usual criterion if ESS > 200 [Drummond et al., 2006; Lanfear et al., 2016], the sample contains enough uncorrelated points to estimate the posterior tree distribution, validating NS as an algorithm to perform parameter inference.

Figure 4.6 displays the marginal posterior distribution for the shape parameter of the gamma distribution of the rates across sites. This sampling distribution is quite symmetric and centered around 0.26. This justifies the "good" prior defined and used in the sensitivity analysis.

4.6.2.3 Sensitivity

Following the analysis carried out by Xie et al. [2011] to show that TI and SS are sensitive to prior distributions in the fixed tree topology case, unlike HM, we estimate the total marginal likelihood for the selected GTR$+\Gamma_4$ model under three different priors. The models are just differentiated by the prior distributions placed on the shape parameter of the discrete gamma distribution of rates across sites. These priors distributions are Exp(0.001) as the "vague", Gamma(10,0.026) as the "good", and Gamma(148, 0.00676) as the "wrong". This analysis is the extension of the study carried out in Example 3.5.2.1, previous chapter, where

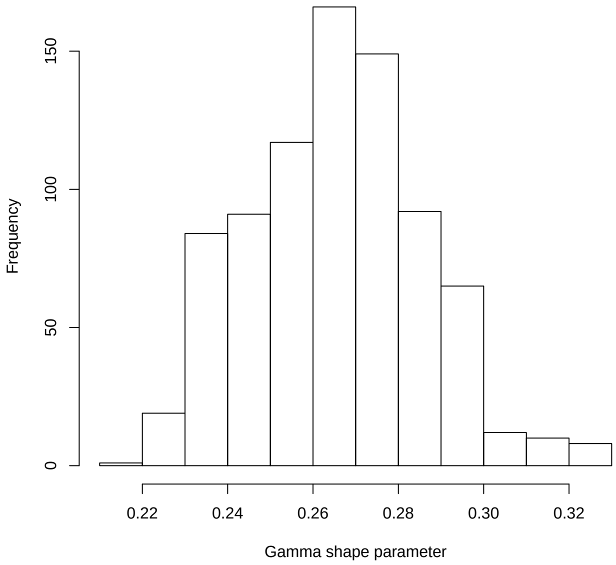


Figure 4.6: Marginal posterior distribution for the shape parameter of the gamma distribution of the rate across sites.

more information about these priors can be found.

For each of the three priors, we carry out marginal likelihood estimation allowing variable tree topology via HM, GSS, NS and NIS. For estimating HM, we use a single-chain MCMC sampler of 50,000 posterior samples. For GSS, we use 100 transitional distributions with 500 samples from each of them. The path connecting the distributions follows an annealing scheme. For NS and NIS, we consider 50 active points. NS posterior samples are used to calibrate the reference distributions required by NIS and GSS [Fan et al., 2011]. For the continuous parameters, the construction of the reference distributions is defined in Example 3.5.2.1. This is with the exception of the branch lengths for which we have not considered importance sampling distributions. For the tree reference distribution, we use the CCD method described above.

Table 4.4 shows the log-marginal likelihood estimates for the different estimation methods under the 3 prior distributions. As expected, HM does not discriminate between the “vague” and the “good” priors. These priors are dominated

Method	Prior model		
	Vague	Good	Wrong
HM	-6559.51	(-1.03)	(-35.94)
GSS	(-8.55)	-6609.04	(-59.37)
NS	(-8.42)	-6609.39	(-58.23)
NIS	(-8.36)	-6609.21	(-58.86)

Table 4.4: Estimated log-total marginal likelihood values for the GTR+Γ_4 model under 3 different priors for the gamma shape parameter. The highest one for each method is displayed whereas the difference with the other models are shown in parenthesis.

by the area of highest likelihood, unlike the “wrong” prior, which prevents the exploration of that place. On the other hand, GSS is sensitive to the prior specification. This method takes into account the prior information which is reflected in the estimate. Like this power posterior method, NS and NIS are sensitive to the prior choice, producing estimates for the “wrong”, “vague” and “good” priors in this increasing order.

It is interesting to note that HM estimates are similar for the fixed and variable topology case. For the former, see Table 3.3 in the previous chapter. This phenomenon could be explained by the extremely sharp posterior tree distribution in the latter case. Most of the probability is concentrated in the tree used in the fixed topology case. Thus, most of the posterior samples are from the maximum posterior tree and they constitute the samples used in the fixed tree case. These tree posterior probabilities are displayed in Table 4.3.

4.6.2.4 NS vs NIS

We take advantage of the previously executed analysis of sensitivity to prior specifications in order to compare NS and NIS performance. The comparison is carried out in terms of their uncertainty and convergence. Both methods were performed with 50 active points, but NIS required additionally the posterior samples to calibrate its reference distributions.

Table 4.5 shows the information and standard deviation of the estimates for

	NS		NIS	
Prior	H	SD	H	SD
Vague	68.00	1.17	27.86	0.75
Good	58.96	1.09	27.45	0.74
Wrong	95.53	1.38	26.83	0.73

Table 4.5: Information (H) and standard deviation (SD) estimates for NS and NIS. The analysis is carried out for the 3 prior specifications used in the sensitivity analysis: vague, good and wrong.

the different models. For NS, the information, a measure of how much we have learned from the data, varies for the different priors. As expected, the wrong prior has the highest one because the posterior distribution is much more different from the prior. In other words, this is the case in which our prior beliefs have changed more significantly after acquiring the data. The opposite case occurs with the "good" prior. These values affect proportionally the standard deviation of the estimates, which can consequently be ordered in an increasing order according to the model as good, vague and wrong. For NIS, the information values are smaller and more similar for the different priors and thus the standard deviations are too. This is due to the similarity between the posteriors and the reference distributions, which are independent of the priors. Unlike NS, NIS performance does not depend on the prior specifications but only on the reference distributions, making it a more stable method.

Another positive characteristic of NIS is its efficiency in terms of uncertainty. Using the same number of active points, NIS uncertainties are much lower than the ones obtained via NS. This happens because the uncertainties are proportional to the information values, which are lower for NIS due to the similarity between the posteriors and the reference distributions. In order to equate its uncertainty, NS would require more active points. For the good, vague and wrong prior cases, it would require approximately 108, 121 and 180 active points to yield NIS uncertainty, respectively. These values were calculated by equating NS standard deviation $\sqrt{H/N}$ to the estimated NIS uncertainty, and solving for N. On the other hand, the number of iterations required by NS to converge is directly proportional to the number of active points. For instance, if we consider the usual termination criterion that states that the algorithm stops when the number of iterations exceeds $2 \times N \times H$, the number of iterations increases from 5,896 to 12,736 for the good prior; from 6,800 to 16,456 for the vague prior; and from 9,553 to 34,391 for the wrong prior. Using the same criteria, NIS requires less than 2,790 iterations for all the models. Thus, increasing of the number of active points to equate NIS in terms of uncertainty makes NS require a higher number of iterations to converge and thus increase its computational cost.

Finally, NIS converges to the estimate faster than NS under the same number of active points. Figure 4.7 shows the evolution of the log-estimate over time for the wrong prior considering 50 active points for each method. It can be noticed that NIS starts much closer to the marginal likelihood value than NS. This happens because it starts the exploration of the parameter space from the reference distribution, which is similar to the posterior. Since z is dominated by the latter, whose mass has been located by the reference distribution, NIS requires less time to estimate it. This behaviour is repeated for the rest of the prior models, thus they are not included here. To summarize, NIS requires many less iterations to converge than NS.

The good performance of NIS, in comparison to NS, is solely granted by the reference distributions. Thus, those for the continuous parameters, but most importantly for our analysis, the one for the tree topologies, constructed by the CCD method, proved to be efficient methods to narrow the parameter space in a NS context.

4.6.2.5 NS with a single active point

It is well known that under inadequate specifications, or even when the reference distribution is based in a poor posterior sample as was shown in Example 3.5.1.2, GSS performs poorly. This opens the question: how does nested sampling perform in an analogous situation? Trying to answer this query, we test NS in the simplest condition, that is, by using only a single active point. The idea is to assess NS under the poorest specifications in a phylogenetic context by comparing their estimates with an accurate one. We carry out the analysis by replicating the estimation process via NS 500 times, for which we consider the good prior, and

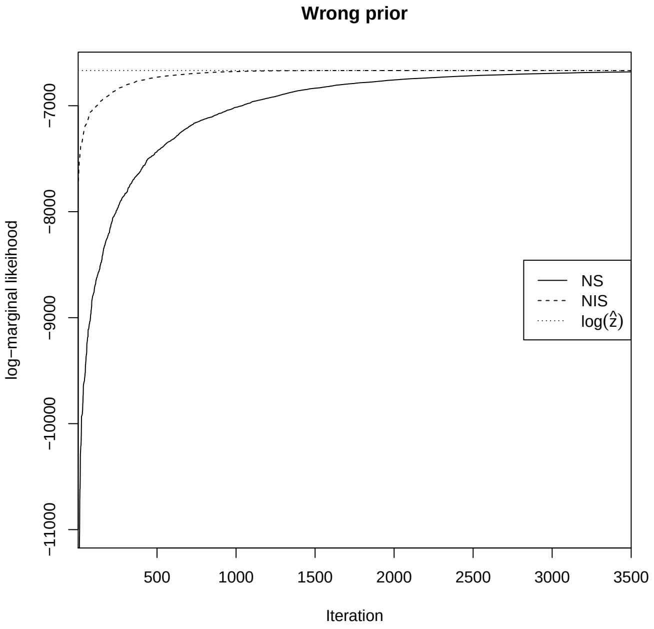
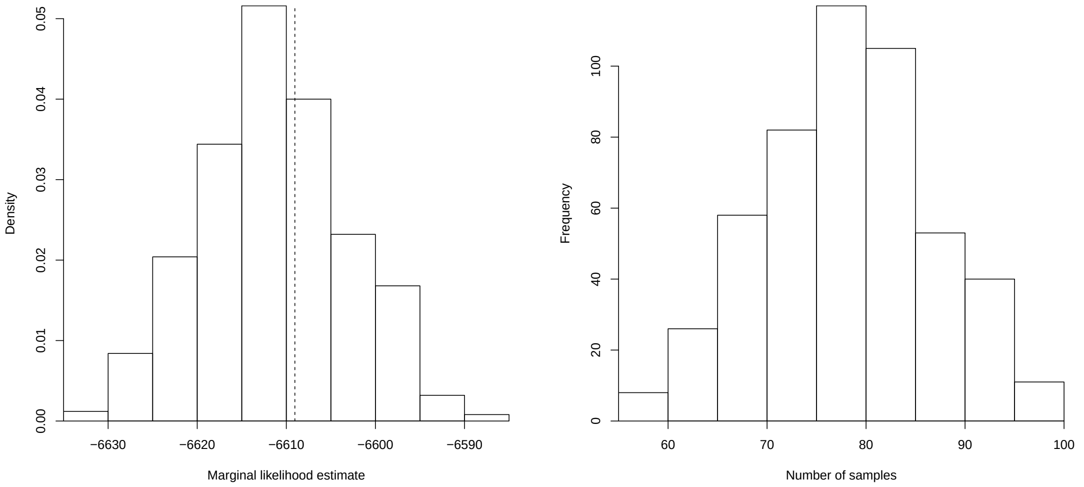


Figure 4.7: Evolution of NS and NIS estimates over time for the wrong prior. The dotted line stands for the GSS estimate.

registering the estimates. The results are shown in Figure 4.8a.

The vertical dotted line stands for a consensus marginal likelihood estimate calculated by averaging GSS, NIS and NS estimates displayed in Table 4.4. NS estimates, with a single active point, are around this consensus value. The number of samples to get these estimates are displayed in Figure 4.8b. These correspond to the minimum number of samples from which the estimate does not increase significantly, namely, $|\widehat{z}_i - \widehat{z}_{i-1}| < 10^{-4}$. NS required around 80 sampled points on average, and in general, less than 100. This computational effort represents a small part of what GSS requires to yield an estimate around the true value. Note that the rule of thumb to estimate the tree topology is an effective sample size of at least 200 samples. This number allows to estimate appropriately the tree reference distribution. Just then, GSS can start working properly, but under unknown specifications, which must be found via guesswork. On the other hand, NS, even with a single active point, yields estimates around the true value.

Another important feature of NS to highlight and be exploited is its ability of



(a) Estimates.

(b) Required samples.

Figure 4.8: Nested sampling performance by using a single active point. The vertical line in (a) stands for a consensus marginal likelihood value. Figure (b) depicts the number of samples required to get a stable estimate.

producing simultaneously an estimate of the standard deviation associated with the marginal likelihood estimate, even in these conditions. Therefore, we can construct a confidence interval for the estimate in a single run. From the 500 replications, we build 3 different intervals which vary in their widths. These are composed by the estimate plus or minus one, two and three standard deviations. These intervals contain in 60.8%, 93.8% and 99.8% of the times the consensus value, respectively. This suggests that a model selection analysis can be carried out with NS with a single active point by using wide intervals. If two intervals (from two models) overlapped, the analysis should be redone with more precision, i.e., with more active points. This offers a very cheap alternative to carrying out model selection via NS.

4.6.3 Laurasiatherian data

Laurasiatheria is a superorder of placental mammals which originated on the northern supercontinent of Laurasia 99 million years ago. We chose a dataset provided with the R-package phangorn [Schliep, 2011]. The dataset consists of 47

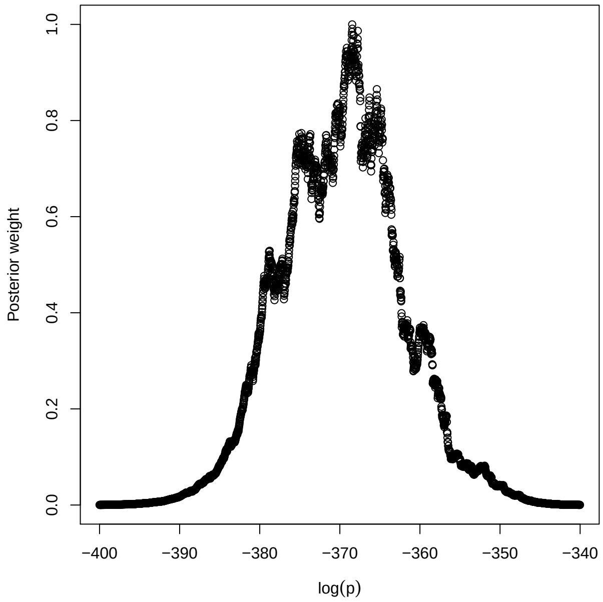
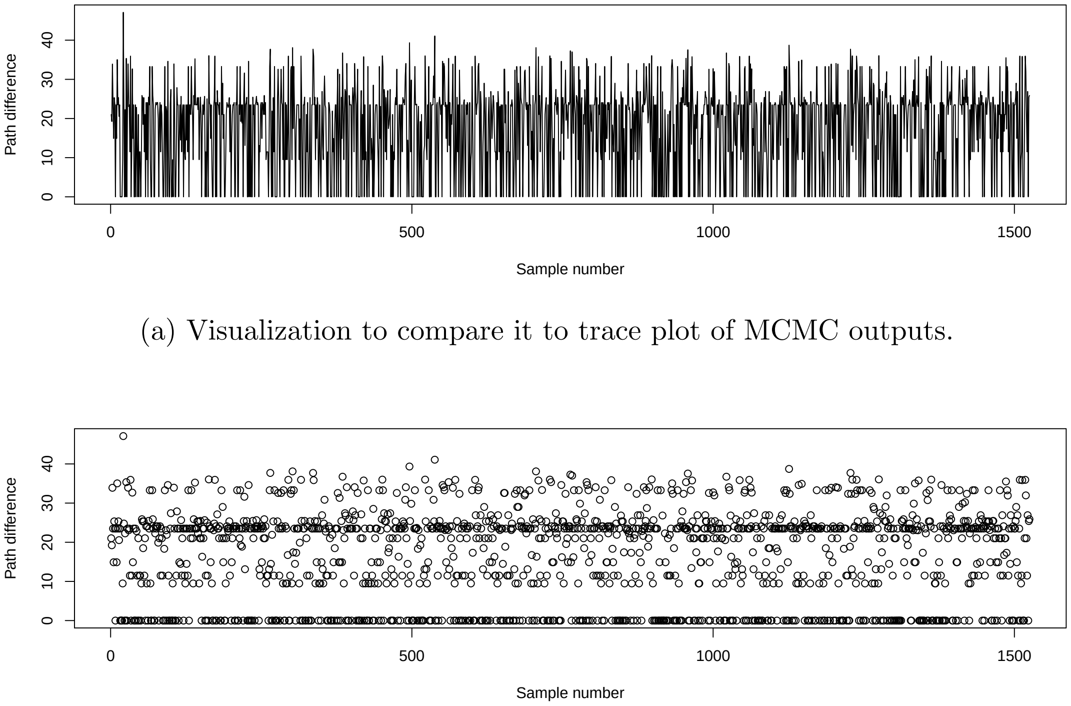


Figure 4.9: Posterior weights for the Laurasiatherian data. The NS posterior sample is taken according to these weights.

aligned mitochondrial RNA sequences of a total length of 3,179 sites.

This dataset presents two interesting and challenging characteristics which provide a good scenario to assess the performance of nested sampling. Firstly, the tree space contains around 3.5×10^{68} possible phylogenies that NS has to potentially explore. Secondly, the posterior distribution is not dominated by a single tree topology. This represents a challenge for MCMC methods.

Assuming a GTR$+\Gamma_4$ model and a Gamma$(2, 2.5)$ distribution as a prior for the gamma shape parameter, we use NS with 50 active points in order to sample the posterior distribution. Actually, we are only interested in studying the tree posterior distribution. We assess this sample according to two analyses proposed by Lanfear et al. [2016]. For the first analysis, we calculate the path difference [Steel and Penny, 1993] between the tree samples and a focal tree. We consider the tree with the highest frequency as the focal tree. Then, these differences are visualized in a trace plot. For the second analysis, we estimate a confidence interval for the effective sample size (ESS) as described in the previous



(a) Visualization to compare it to trace plot of MCMC outputs.

(b) Visualization to highlight the behaviour of path differences.

Figure 4.10: Path differences in nested sampling tree posterior sample with respect to the maximum posterior tree for the Laurasiatherian data.

example.

We again point out that these analyses were proposed for the evaluation of tree samples obtained from standard MCMC methods. However, in the NS case the posterior samples are not indexed by time. For instance, the trace plot does not convey the same information for NS compared to standard MCMC methods. Nevertheless, it provides an insight into the tree posterior sample and is therefore of use to us.

NS yielded, in this single run, a sample of 1,525 points. It required around 20,000 points to explore the parameter space to estimate the marginal likelihood. The posterior weights of the discarded points are shown in Figure 4.9. This gives a quick picture of the posterior density. The posterior is concentrated in a small area of the prior (in around e^{-340} of its mass). This might be mainly caused by the uniform prior assigned to the huge tree space. The trace plot based on the path differences is shown in Figure 4.10a. It shows that several trees were visited in the NS posterior sample and that this sample is not only dominated by the maximum posterior tree, since the trace does not stay continuously at 0. The average path difference with respect to the maximum posterior tree is around 17. Note that two or more different tree topologies can have the same path difference with respect to the focal tree, which could potentially make the trace seem at times constant despite having undergone a change in state.

To accommodate the fact that the samples are not indexed by time, we propose a simple visualization of the NS path differences. Figure 4.10b shows a gap between 0 and the next smallest value (9.49), a characteristic which cannot be seen in the trace plot.

To estimate the confidence interval for the effective sample size (ESS) for the tree topologies, we replicated 1,000 times its estimation under randomly chosen focal trees. The 95% confidence interval is $[1494, 1525]$. A rule of thumb suggests 200 to be a lower limit to accurately infer the tree posterior distribution [Drummond et al., 2006; Lanfear et al., 2016]. Therefore, this interval indicates the adequacy of the posterior sample obtained from NS.

4.7 Discussion

In this chapter, we have extended nested sampling algorithms to the variable tree topology case. Their use involves new challenges in Bayesian phylogenetic analysis. One of them is the construction of tree proposal mechanisms, which have to deal with the dynamic behaviour of the target distribution over time, unlike standard MCMC methods. Another one, which is not exclusive to NS but to those methods which make use of importance sampling distributions, is the construction of a reference distribution for tree topologies. This represents the main challenge for the implementation of NIS. This chapter addressed mainly these themes and the exploitation of NS in its full potential. The methodologies have been tested through 3 phylogenetic examples of different complexities.

The inclusion of the tree topology as a parameter in the phylogenetic analysis brings with it the annexation of a huge parameter space which has to be explored. This can be a challenge for NS, which could require many iterations to reach those areas of high likelihood values that represent the most significant areas of contribution to the marginal likelihood estimation. The use of an importance sampling approach can deal efficiently with this problem. We have proposed the use of conditional clade distributions (CCD) in order to generate a reference distribution for the tree topology. This method relies on the principle of conditional independence of separated subtrees. Its logic has been explained through a 6 taxon example. We have used this importance sampling distribution for NIS and GSS methods.

The tree posterior distribution is usually the target of Bayesian phylogenetic inference. The mechanisms to explore this parameter space have been widely studied. However, in NS sampling, the target distribution is the prior distribution, which gets constrained at each iteration according to the likelihood. In other words, the target distribution has a dynamical behaviour over time, which represents a new challenge, especially for the tree proposals. Instead of adapting the proposals at each iteration, we have proposed the use of a mix between bold and conservative branch rearrangements, namely, Random Subtree Prunning and Regrafting (rSPR) and Stochastic Nearest Interchange (stNNI), respectively. We have illustrated its performance through a didactical example.

To test the consistency of nested sampling in the variable tree topology case, we applied it in a manageable case where the total marginal likelihood can be estimated by brute force, i.e., as an average of all the marginal likelihoods of the possible topologies. For this, we presented a 3 taxon example of the primate family under the molecular clock assumption. We estimated the total marginal likelihood through the individual marginal likelihood estimates via GSS. The specifications considered were enough to get reliable estimates and thus, to have a comparable total marginal likelihood value. NS was consistent with this estimate.

We have analysed a 10 taxon dataset of green plants in order to evaluate NS in different scenarios. Firstly, we carried out model selection among 6 models of evolution. We used 50 active points to estimate the marginal likelihood with their corresponding uncertainties. This information allowed us to choose among the models with certain degree of confidence. The selected model was GTR+Γ. Secondly, taking advantage of the previous analysis, we tested the ability of the evolutionary models of reconstructing the tree topology. We found that the simplest models JC69 and JC69+Γ yielded a clade in disagreement with more complex models. For the selected model, we showed that the tree posterior samples obtained via NS are similar to those obtained via an MCMC method. According to the effective sample size criteria, this sample was adequate to estimate the tree posterior distribution. Thirdly, we tested the sensitivity of NS to prior specifications and compared it to the established methods, such as, HM and GSS. For this, we allocated three different prior distributions on the gamma shape parameter. NS was able to detect and estimate correctly the marginal likelihood, like GSS, but in contrast to HM. Fourthly, we compared NIS to NS. We found that NIS performance is independent of the prior distribution, unlike NS. It also yields lower uncertainty. Also, we showed that NS would require a much higher computational cost to equate NIS in terms of uncertainty. This application allowed us to evaluate the CCD method as a tool to build a reference distribution for the tree topologies. It proved to be an efficient method. Finally, we tested NS under the simplest specification, i.e., by using a single active point. Even in this scenario, the method yields estimates around the consensus estimate. The effort required in this analysis represents a small fraction of what GSS would require to start to yield estimates around the true value. In addition, as NS is able to estimate the uncertainty related to the marginal likelihood estimate and thus produce a confidence interval, we tested this ability in terms of containing a fairly accurate estimate obtained from the previous analysis. We found that in 500 replications, the intervals of the form $\log \widehat{z} \pm 3 \times \mathrm{SD}(\widehat{\log z})$ contained the accurate estimate in all but a single case. This makes NS quite a cheap alternative to carry out model selection, unlike established methods.

Finally, we tested NS as a means of sampling the tree posterior distribution in a bigger phylogeny. The data contained 47 species of the Laurasiatherian group. The posterior was concentrated in a small area of the prior distribution, but NS was able to reach it. Using 50 active points, NS yielded 1,525 posterior samples which were adequate according to ESS criteria. The sample consisted of a variety of trees without being dominated by the one with maximum frequency.

We analysed the tree posterior samples obtained via NS using the proposals made by Lanfear et al. [2016], but we were cautious about their applications.

This is because these proposals are only valid for analysing samples obtained via standard MCMC methods, in which case the samples are indexed by time. In the NS case, this is not met, however, these analyses can still give an insight into the tree posterior distribution. To accommodate the non-temporal link of samples, we propose an analog visualization of the path difference for the NS case.

The gain of including topological uncertainty in Bayesian model selection is still an open question. A common practice in phylogenetics is to select a phylogeny by using any fast tree building approach, then select an evolutionary model for the previously chosen phylogeny, and finally the parameter space is explored to get estimates of the involved parameters, including the phylogeny. Thus, the natural question that emerges is: Is model selection topology dependent on the initial phylogeny? In this regard, Posada and Crandall [2001] argued, based on their simulations, that the initial phylogeny does not affect model selection unless it is a random chosen tree. Abdo et al. [2005] also showed, in a decision theory framework, the little effect of the topology on model selection when an optimal tree is used, even though they did not consider properly phylogenetic uncertainty in their study. On the other hand, the topology dependence in parameter inference is well known. Sullivan et al. [1996] showed that the transition/transversion rate is topology dependent when among-site rate variation is accommodated. Yang [1994b] also found that there is variation in estimates of the gamma distribution shape parameter across topologies. This relationship between the phylogeny and the parameter estimates could potentially lead to topology dependent in model selection. All these studies have been addressed in a frequentist context. However, it could be interesting to study the topology dependence in model selection in a Bayesian context.

Chapter 5

Conclusion

Bayesian methods have become a feasible and efficient alternative across different disciplines, including phylogenetics. Especially in this field, these tools have opened up the possibility of using more complex models and analysing large datasets. However, this development has brought with it new challenges, some of them belonging solely to phylogenetics. This thesis aimed to discuss, analyse, compare, and extend Bayesian statistical methods, mainly in model selection, in a phylogenetic context.

Commonly, statistical models are composed by a unique inseparable parametric structure. But this is not the case in phylogenetics. In this field, the model is composed of two different elements: a tree and a model of evolution. The tree stands for the evolutionary relationship among the analysed taxa. This is characterized by the branching pattern, called topology, which shapes the relationship, and the branch lengths, which may represent the amount of sequence divergence or time elapsed. On the other hand, the model of evolution describes the changes between nucleotides, or any other kind of data, over evolutionary time. Time reversible Markov models are usually considered for this model. The models of evolution only vary in the parametric structure of their transition rate matrix, adopting distinctive names.

Bayesian analyses require the explicit definition of prior distributions for each parameter. These should reflect our knowledge before collecting the data. In phylogenetics, the model involves, in general, several parameters. The tree has the topology and the branch lengths; on the other hand, the evolutionary model may contain frequencies, relative rates, heterogeneous substitution rate among sites, and a proportion of invariable sites. In the case of wanting the prior to reflect our complete ignorance about the values of the parameter, i.e., letting the analysis be dominated by the likelihood, non-informative priors can be used. Against the natural tendency of using a uniform distribution for this purpose, it has been found that this does not work properly for many phylogenetic parameters. In practice, these kind of distributions have been defined via trial and error in phylogenetics, since, most of the time, the posterior does not have an explicit form. We discussed the development of these kinds of distributions in Section 2.4.

Having a set of available models, it is necessary to determine which model is best suited to the data. This procedure should be routine in any phylogenetic analysis [Posada and Crandall, 2001]. There are many statistical tools to select among models. In the context of Bayesian analysis, the Bayes factor has become a standard method for this purpose. This quantity is composed by the product of the ratio of the marginal likelihoods and prior probabilities of the models. In the case of equal prior probability for the models, the Bayes factor (BF) is just the ratio of marginal likelihoods. However, this quantity is, most of the time, a very difficult multidimensional integral, namely, the integral of the likelihood times the prior distribution over the parameter space. When the phylogeny is unknown, this has to be considered as any other parameter, adding more complexity to the calculation. Thus, the BF requires of a numerical approximation to be calculated.

Many methods have been proposed to estimate the Bayes factor. They work by either estimating the marginal likelihoods, which is the most common practice, or estimating it directly. In Chapter 3, we discussed several marginal likelihood estimation methods, which also allow direct Bayes factor estimation. Among them: harmonic mean (HM), path sampling (PS), steppingstone sampling (SS), generalized steppingstone sampling (GSS), annealed importance sampling (AIS), nested sampling (NS), and nested importance sampling (NIS).

HM has been the most popular method used in phylogenetics, most likely due to its simplicity, to wit, it only requires samples from the posterior. However, its drawbacks are well known and widely documented. For instance, it overestimates the true value, which has been confirmed in our analysis, may have infinite variance, is insensitive to prior specifications, among many others. We included its inability of mixing phase transitions in the likelihood to its list of disadvantages.

We have also extended the HM method for estimating the Bayes factor directly. We provided an approximation of its variance, finding that this one contains only one potential source of infinity. Thus, in the case that there is only one model with associated infinite variance, this can be allocated appropriately in the calculation in order to avoid the problem of infinite variance of the estimate. Therefore, the direct estimation of BF via HM could outperform the one obtained via independent HM estimates in terms of uncertainty.

PS and SS have a high accuracy and are available in different phylogenetic software packages. However, they depend on many tuning parameters. In practice, the optimal values for these parameters vary from problem to problem. For instance, vague priors might make these methods require more transitional distributions. The problem is that these specifications have to be set up by the user. A bad choice could lead to poor estimates. A common practice is to try different values until the estimate shows signs of stability. This could be impractical in many situations, such as the analysis of large datasets. This disadvantage is also shared by AIS.

Another disadvantage of these methods is their inability of dealing with phase transitions in the likelihood. Actually, they would require non-viable computational effort to be able to explore the parameter space. This difficulty can be overcome by using an importance sampling distribution. However, the quality of the estimation will be determined by this distribution.

GSS makes use of that reference distribution in order to shorten the path between the prior and the posterior. This makes it require less computational effort to accurately estimate the marginal likelihood than SS. This approach dispenses with the distribution of the β values which characterize the transitional distributions, due to the similarity of the reference distribution and the posterior. Thus, it requires less tuning parameters than SS. As was discussed above, GSS is able to deal with phase transitions in the likelihood as long as the reference distribution is a good approximation of the posterior, which could be very difficult in some cases.

One interesting finding in this work was the relationship between AIS and SS. Both methods were developed independently in different years, 2001 and

2011, respectively, but they are closely related. These methods make use of a telescope product to estimate the marginal likelihood, but they differ in the way of how it is used. AIS estimates it directly, whereas SS estimates each factor separately. For some particular specifications, they reduce to the same estimator. In practical terms, they differ in the uncertainty associated with their estimates. In our analysis, we found that SS has a lower uncertainty. However, more studies are required to generalize this observation.

NS is another method that we have introduced to phylogenetics for model selection and parameter inference, and also extended to allow variable tree topology. We showed that this method offers a viable alternative for Bayesian phylogenetic analysis. However, its implementation brought new challenges to Bayesian phylogenetic methods, which were discussed throughout this thesis, especially in Chapter 4.

NS relies on a property of positive random variables in order to redefine the marginal likelihood into a one dimensional integral. Thus, the prior mass is linked to the likelihood values. The estimation is carried out by sampling the prior distribution, which becomes constrained toward those areas of high likelihood values at each iteration. This is an unusual scenario for standard MCMC methods used in phylogenetics, which often deal with a static target distribution, and represents a real challenge, especially for the tree topology. For this parameter, we proposed the usual Metropolis algorithm with proposals composed by a mix between conservative and bold branch-rearrangement of the tree topology. In our analysis, this approach worked quite well.

The incorporation of the tree topology into the marginal likelihood estimation increases the parameter space to be explored enormously. This could make NS require a high number of iterations to reach the areas where the posterior distribution is located, often in a very small portion of the parameter space. Mainly to solve this problem, we proposed NIS, which makes use of a reference distribution to shorten the exploration. The challenge of this method is principally the definition of this distribution for the tree topologies, which should optimally be an approximation of the posterior. For this, we proposed the use of conditional clade distributions [Larget, 2013]. In our analysis, in which we also used it for GSS, it performed very well, making NIS outperform NS in terms of uncertainty and convergence.

Even though GSS, and other power posterior methods, are quite accurate, they require several specifications, which affect and determine directly the estimate. In practice, they vary from problem to problem and the optimal values must be found by guesswork. NS does not suffer from this problem, being one of its main advantages over these methods. This algorithm only requires the specification of the number of active points, and also the number of steps to generate the new point at each iteration, but which is also required by any other method relying on MCMC. Even under the poorest specification, which corresponds to the case of a single active point, the NS estimate will be around the true value. This advantage is also valid for NIS.

Another good property of NS is its ability to generate posterior samples at no extra cost. These can even be obtained from complex posterior distributions, for instance, in the presence of phase transitions in the likelihood, in which case standard MCMC methods face difficulties. Actually, this is a critical situation for methods which rely on an importance sampling distribution, such as GSS or NIS. In this case, they may fail in the marginal likelihood estimation, due to the practical impossibility of sampling the posterior and consequently of constructing their reference distribution adequately.

The property of dealing with phase transitions in sampling the posterior distribution, and consequently in estimating the marginal likelihood, is granted by the way NS explores the parameter space. This is carried out by sampling the prior distribution according to the likelihood contours, independently of the likelihood shape, unlike standard MCMC methods which depend on this latter. This difference makes NS be a more general method than established ones in estimating marginal likelihoods and sampling the posterior.

The quality of an estimation is given by the degree of its uncertainty, which should therefore accompany any estimate. This is one of the main attributes of NS. This algorithm is also able to provide an estimate of its uncertainty in a single run. This is an important difference in comparison to the established methods, which have to be executed many times in order to provide a measure of their uncertainty. Even though it could be argued that NS yields high uncertainties, even in simple problems, this can be quantified in order to have an idea of the quality of the estimate.

We have also extended NS to estimate the Bayes factor directly. This approach does not depend on any questionable annealing scheme, unlike, for instance, SS. This estimate can potentially have a lower uncertainty than the one obtained by using independent estimates via NS with the same number of active points. We illustrated its performance in an example. Furthermore, this method can yield, at no extra cost, posterior samples from one of the models. As Lartillot and Philippe [2006] pointed out, this approach can be mainly useful when the difference between the log-marginal likelihoods of the models is small with respect to these values. In such case, the individual estimates should be highly accurate to obtain a good BF estimate.

In this thesis, we have mainly introduced NS to phylogenetics and studied its performance. The method proved to be an efficient way to carry out marginal likelihood estimation. Actually, it represents a relatively cheap alternative for the estimation. The method also proved to be more general than others considered in this thesis. One could argue that extreme situations, as the ones presented in this work (phase transitions), are not found in phylogenetics, however, the shape of the likelihood is something not well understood in the field. There have been some attempts, but they have not been definitive. Meanwhile, methods of general applicability should be used, either for estimating the marginal likelihood or in sampling the posterior distribution.

Model selection is a key part of phylogenetic inference, but it should not be the last assessment of the selected model. After model comparison by using any criterion, such as the Bayes factor, it is recommendable to evaluate model misspecifications. In other words, how well the model fits the data. For this purpose, for instance, Bollback [2002] proposed the analysis of the posterior predictive distribution. This is carried out by comparing a single test statistic, calculated from the data, to the distribution of the same measure obtained from simulations from the posterior distribution. If the test indicates that the model does not fit the data well, it could be that some relevant characteristics of the evolutionary process have not been taken into account. A model, despite of being the best among a pool of models, may not describe the information contained in the data about the underlying evolutionary process effectively. Thus, it is also necessary to test its adequacy, even though it is rarely carried out [Gatesy, 2007].

Recently, we have implemented the nested sampling algorithm in the NS package for BEAST2 [Bouckaert et al., 2014], available from `https://github.com/BEAST2-Dev/nested-sampling` under the LGPL licence. The NS package allows for phylogenetic inference under any of the models available to BEAST.

5.1 Future work

The work developed throughout this thesis has given rise to different questions and potential future works. Some of them are discussed below.

Uniform prior on the gamma shape parameter

The most tentative way of incorporating our ignorance with respect to a parameter, but many times erroneously, is by using a uniform prior. As discussed before in Chapter 2, this distribution may add information about the parameter, most of the time, without this being the primary intention. For instance, it assigns different probabilities on clades in the tree topology case; it has a huge impact on the posterior distribution in the case of branch lengths, assigning ridiculous probabilities to big tree length values.

Its impact has been discussed and well documented for some parameters in phylogenetics. However, to the best of our knowledge, its impact on the shape parameter of the gamma distribution for the rates across sites has not been yet studied. Even though the exponential is the most popular prior for this parameter, the use of the uniform can be found in the literature. Therefore, the impact of this distribution on the inferences about this parameter should be understood and documented.

NS direct Bayes factor estimation

The use of a reference distribution can save considerable computational time in estimating the marginal likelihood. This works by shortening the path between the prior and the posterior distributions. For instance, this is how GSS and NIS works mainly. In the case of estimating the Bayes factor directly, the path connects the posterior distributions of the models instead. In this context, it would be worthwhile to study how to incorporate, if it is possible, a reference distribution in order to shorten the path between these posterior distributions in the NS estimate.

Tree proposals in NS

NS brings with it new challenges for MCMC methods in phylogenetics. The target distribution has a dynamical behaviour over time, namely, it is the prior distribution which gets constrained toward those areas of high likelihood values at each iteration. Thus, the proposal mechanisms have to be able to deal with this characteristic, which is a new scenario for them, especially for the tree topologies.

As Lakner et al. [2008] showed, the tree proposals play a key role in the performance of the MCMC methods. Actually, they must deal with the most difficult parameter space exploration in most Bayesian MCMC phylogenetic analysis. The authors studied two kind of proposal mechanisms: *branch-change proposals* and *branch-rearrangement proposals*, finding that the latter perform better. In this thesis, we took these results into account and proposed the use of a mix of two proposals which are in this latter category. Our scheme is composed by a mix between conservative and bold tree proposals. This scheme performed quite well in the examples. However, it could be worthwhile to study all branch-rearrangement proposals studied in Lakner et al. [2008], even branch-change proposals, in a NS context. Their performance could be different, since they were only studied in standard MCMC context, where the target distribution is static. Hopefully, they could help NS performance in terms of convergence.

Effect of phylogenetic uncertainty in model selection

One of the positive features of Bayesian phylogenetic methods is the possibility of incorporating uncertainty in tree topology simultaneously into the analysis. This is one of the limitations of the frequentist approach, which bypasses this by fixing the topology in model selection, and then carrying out the parameter estimation. Supporting this practice, some studies have shown that the initial topology has a small effect on model selection when this is not randomly chosen [Abdo et al., 2005; Posada and Crandall, 2001]. However, to the best of our knowledge, this is an aspect not well understood in a Bayesian context.

Systematics often aim to infer phylogenies, so it seems natural to consider a model of evolution which has been selected according to its performance in the tree parameter space. Bayesian methods allow to take into account phylogenetic uncertainty, but this might represent a significant increase in computational cost. Therefore, it would be interesting to carry out model selection in different scenarios with different levels of complexity, and study the gaining of accommodating phylogenetic uncertainty in order to determine if it is worth the computational effort.

More complex models

In this work, the evaluation of NS in phylogenetics has been restricted to homogeneous models, namely, those ones which assume a single reversible Markov process along the phylogeny and across sites. These models might fail in the description of more complex evolutionary processes. In these situations, more general Markov models, assigned across the lineages and/or sites in order to model compositional variation in the data, can be utilized. The extension of NS to these scenarios is part of our future work.

R package

All the analyses carried out in this thesis were implemented and performed in R [R Core Team, 2015]. This included the implementation of the marginal likelihood estimation methods, such as harmonic mean, path sampling, annealed importance sampling, steppingstone sampling, generalized steppingstone sampling, nested sampling, and nested importance sampling. The conditional clade probability distribution was also implemented, which allows us to summarize tree distributions. Part of the future work is to make all these methods available in an R-package.

Appendix 1

The marginal likelihood estimation analyses carried out in this work were all performed in R Core Team [2015]. This includes the calculations and figures.

All the marginal likelihood estimation methods used in the phylogenetic analyses were implemented in such a way that they work for any nucleotide dataset. These implementations allow to use fix and variable tree topologies. The prior distributions on the branch length, mean branch length, relative rate, gamma shape parameters allow to specify different values for their parameters.

In these implementations, we took advantage of some functions already implemented in phangorn [Schliep, 2011]. In particular, we used the pml function to evaluate the likelihood at single points, rSPR to generate the tree proposals and consensusNet to produce the consensus networks.

In the case of the direct Bayes factor estimation in the phylogenetic analysis, we implemented some relevant functions in Rcpp [Eddelbuettel and François, 2011]. This R-package allows to build C++ functions which can be easily called from R. In that analysis, we implemented these kind of functions for the likelihood function, the proposal distributions and the nested sampling loop. Until the completion of the R-package, code can be provided upon request.

References

Abdo, Z., Minin, V. N., Joyce, P., Sullivan, J., 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol. Biol. Evol. 22 (3), 691–703. 164, 173

Aitken, S., Akman, O., 2013. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. BMC Syst. Biol. 7 (1), 72. 9, 89

Aitkin, M., 1991. Posterior Bayes factors. J. Roy. Stat. Soc. B Met. 53 (1), 111–142. 62

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19 (6), 716–723. 5

Alfaro, M. E., Holder, M., 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Evol. Syst. 37 (1), 19–42. 48

Archie, J., 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38 (3), 239–252. 12

Arima, S., Tardella, L., 2014. Inflated density ratio (IDR) method for estimating marginal likelihoods in Bayesian phylogenetics. In: Chen, M., Kuo, L., Lewis, P. O. (Eds.), Bayesian phylogenetics: methods, computational algorithms, and applications. Chapman and Hall/CRC, New York, Ch. 3, pp. 25–58. 8, 74, 88, 128

Baele, G., Lemey, P., 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. In: Chen, M., Kuo, L., Lewis, P. O.

(Eds.), Bayesian phylogenetics: methods, computational algorithms, and applications. Chapman and Hall/CRC, New York, Ch. 4, pp. 59–94. 62, 88, 128

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., Alekseyenko, A. V., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29 (9), 2157–2167. 64

Baele, G., Lemey, P., Suchard, M. A., 2016. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol. 65 (2), 250–264. 8, 64, 80, 82, 132

Baele, G., Lemey, P., Vansteelandt, S., 2013. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. BMC Bioinformatics 14 (1), 85. 8, 60, 64, 78, 79, 88, 128, 130

Barry, D., Hartigan, J. A., 1987. Statistical analysis of hominoid molecular evolution. Stat. Sci. 2 (2), 191–210. 43

Baum, D. A., Smith, S. D., 2012. Tree thinking: an introduction to phylogenetic biology. Roberts and Company Publishers, Greenwood Village, Colorado. 15

Berger, J. O., Bernardo, J. M., 1989. Estimating a product of means: Bayesian analysis with reference priors. J. Amer. Statist. Assoc. 84 (405), 200–207. 4

Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21 (2), 163–193. 124

Bernardo, J. M., 1979. Reference Posterior Distributions for Bayesian Inference. J. Roy. Stat. Soc. B Met. 41 (2), 113–147. 4

Blanquart, S., Lartillot, N., 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol. Biol. Evol. 23 (11), 2058–2071. 46

Bollback, J. P., 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19 (7), 1171–1180. 6, 170

Bouckaert, R., Heled, J., Khnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., Drummond, A. J., 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Comput. Biol. 10 (4), 1–6. 4, 171

Brandley, M. C., Leach, A. D., Warren, D. L., McGuire, J. A., 2006. Are unequal clade priors problematic for Bayesian phylogenetics? Syst. Biol. 55 (1), 138–146. 48

Brewer, B. J., Donovan, C. P., 2015. Fast Bayesian inference for exoplanet discovery in radial velocity data. Mon. Not. R. Astron. Soc. 448 (4), 3206–3214. 89, 94

Brewer, B. J., Foreman-Mackey, D., 2016. DNest4: Diffusive nested sampling in C++ and Python. ArXiv preprint arXiv:1606.03757. 140, 143

Brewer, B. J., Prtay, L. B., Csnyi, G., 2011. Diffusive nested sampling. Stat. Comput. 21 (4), 649–656. 89, 94

Brown, J. M., Hedtke, S. M., Lemmon, A. R., Lemmon, E. M., 2010. When trees grow too long: Investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. 59 (2), 145–161. 49

Bruno, W. J., Halpern, A. L., 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16 (4), 564–566. 20

Bryant, D., 2003. A classification of consensus methods for phylogenetics. In: Bioconsensus (Piscataway, NJ, 2000/2001). Vol. 61 of DIMACS Ser. Discrete Math. Theoret. Comput. Sci. Amer. Math. Soc., Providence, Rhode Island, pp. 163–183. 18

Cavalli-Sforza, L. L., Edwards, A. W. F., 1967. Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19 (3), 233–257. 1

Chang, B. S. W., Jnsson, K., Kazmi, M. A., Donoghue, M. J., Sakmar, T. P., 2002. Recreating a functional ancestral archosaur visual pigment. Mol. Biol. Evol. 19 (9), 1483–1489. 1

Chopin, N., Robert, C., 2010. Properties of nested sampling. Biometrika 97 (3), 741–755. 10, 96, 101, 134

Chor, B., Hendy, M. D., Holland, B. R., Penny, D., 2000. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. Mol. Biol. Evol. 17 (10), 1529–1541. 129

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life, 1st Edition. John Murray, London. 1, 6, 12

Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9 (5), 687–705. 1

Dinh, V., Matsen, F. A., 2015. The shape of the one-dimensional phylogenetic likelihood function. ArXiv preprint arXiv:1507.03647. 129

Drummond, A., Bouckaert, R., 2015. Bayesian evolutionary analysis with BEAST. Cambridge University Press, Cambridge. 9, 87

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4 (5), e88. 152, 161

Drummond, A. J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7 (1), 214. 4, 63, 69

Drummond, A. J., Suchard, M. A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29 (8), 1969–1973. 88

Eddelbuettel, D., François, R., 2011. Rcpp: Seamless R and C++ integration. J. Stat. Softw. 40 (8), 1–18. 175

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., Lewis, P. O., 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28 (1), 523–532. 8, 46, 80, 82, 88, 102, 121, 122, 132, 135, 138, 153

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22 (3), 240–249. 36

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27 (4), 401–410. 2, 20, 124

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376. 2, 32, 36

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39 (4), 783–791. 3

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5 (2), 164–166. 32

Felsenstein, J., 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. 53 (4), 447–455. 29

Felsenstein, J., Churchill, G. A., 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13 (1), 93–104. 44

Feroz, E., Hobson, M. P., Bridges, M., 2009. Multinest: an efficient and robust Bayesian inference tool for cosmology amd particle physics. Mon. Not. R. Astron. Soc. 398 (4), 1601–1614. 89, 94

Feroz, F., Hobson, M. P., Cameron, E., Pettitt, A. N., 2013. Importance nested sampling and the Multinest algorithm. ArXiv preprint arXiv:1306.2144. 101

Fitch, W. M., 1971a. Rate of change of concomitantly variable codons. J. Mol. Evol. 1 (1), 84–96. 45

Fitch, W. M., 1971b. Toward defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20 (4), 406–416. 2, 3

Fitch, W. M., Markowitz, E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4 (5), 579–593. 45

Fleming, M. A., Potter, J. D., Ramirez, C. J., Ostrander, G. K., Ostrander, E. A., 2003. Understanding missense mutations in the BRCA1 gene: An evolutionary approach. Proc. Natl. Acad. Sci. USA 100 (3), 1151–1156. 1

Foster, P. G., 2004. Modeling compositional heterogeneity. Syst. Biol. 53 (3), 485–495. 46

Friel, N., Pettitt, A. N., 2008. Marginal likelihood estimation via power posteriors. J. Roy. Stat. Soc. B 70 (3), 589–607. 69, 72, 88

Fukami, K., Tateno, Y., 1989. On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. J. Mol. Evol. 28 (5), 460–464. 129

Galtier, N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18 (5), 866–873. 45

Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Mol. Biol. Evol. 23 (11), 1997–2000. 2

Gatesy, J., 2007. A tenth crucial question regarding model use in phylogenetics. Trends. Ecol. Evol. 22 (10), 509–510. 171

Gaut, B., Lewis, P. O., 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12 (1), 152–162. 20, 37, 124

Gelman, A., Meng, X.-L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. 13 (2), 163–185. 69

Golding, G. B., 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. 1 (1), 125–142. 39, 42

Gu, X., Fu, Y. X., Li, W. H., 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12 (4), 546. 41, 42

Guindon, S., Gascuel, O., 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. Mol. Biol. Evol. 19 (4), 534–543. 2

Handley, W. J., Hobson, M. P., Lasenby, A. N., 2015. POLYCHORD: next-generation nested sampling. Mon. Not. R. Astron. Soc. 453 (4), 4384–4398. 89

Hasegawa, M., Kishino, H., 1984. A new molecular clock of mitochondrial DNA and the evolution of hominoids. P. Jpn. Acad. B-Phys. 60 (4), 95–98. 32, 37, 41

Hasegawa, M., Kishino, H., Yano, T. A., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22 (2), 160–174. 32, 37

Heaps, S. E., Nye, T. M., Boys, R. J., Williams, T. A., Embley, T. M., 2014. Bayesian modelling of compositional heterogeneity in molecular phylogenetics. Stat. Appl. Genet. Mol. Biol. 13 (5), 589–609. 46

Hernández-López, A., Chabrol, O., Royer-Carenzi, M., Merhej, V., Pontarotti, P., Raoult, D., 2013. To tree or not to tree? Genome-wide quantification of recombination and reticulate evolution during the diversification of strict intracellular bacteria. Genome Biol. Evol. 5 (12), 2305–2317. 12

Hodgkinson, A., Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12 (11), 756–766. 37

Hoff, M., Orf, S., Riehm, B., Darriba, D., Stamatakis, A., 2016. Does the choice of nucleotide substitution models matter topologically? BMC Bioinformatics 17 (1), 143. 20

Höhna, S., Drummond, A. J., 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Syst. Biol. 61 (1), 1–11. 136

Holder, M., Lewis, P. O., 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. 4 (4), 275–284. 5

Holder, M., Lewis, P. O., Swofford, D. L., Bryant, D., 2014. Variable tree topology stepping-stone marginal likelihood estimation. In: Chen, M., Kuo, L., Lewis, P. O. (Eds.), Bayesian phylogenetics: methods, computational algorithms, and applications. Chapman and Hall/CRC, New York, Ch. 5, pp. 95–111. 7, 8, 47, 49, 54, 132, 136

Holland, B., Moulton, V., 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. In: Benson, G., Page, R. (Eds.), Algorithms in Bioinformatics, WABI 2003. Springer-Verlag, Berlin, pp. 165–176. 18

Holland, B. R., 2013. The rise of statistical phylogenetics. Aust. NZ J. Stat. 55 (3), 205–220. 3

Holland, B. R., Huber, K. T., Moulton, V., Lockhart, P. J., 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. Mol. Biol. Evol. 21 (7), 1459–1461. 12, 18, 149

Huelsenbeck, J. P., Bollback, J. P., Levine, A. M., Olmstead, R., 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51 (1), 32–43. 17

Huelsenbeck, J. P., Hillis, D. M., 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42 (3), 247–264. 20, 117, 124

Huelsenbeck, J. P., Larget, B., Alfaro, M. E., 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21 (6), 1123–1133. 7, 47, 53, 144

Huelsenbeck, J. P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8), 754–755. 4, 47, 49, 63, 88

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., Bollback, J. P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294 (5550), 2310–2314. 5

Jayaswal, V., Jermiin, L. S., Robinson, J., 2005. Estimation of phylogeny using a general Markov model. Evol. Bioinform. Online 1, 62–80. 43

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lon. Ser-A 186 (1007), 453–461. 4, 47

Jermiin, L. S., Jayaswal, V., Ababneh, F., Robinson, J., 2008. Phylogenetic model evaluation. In: Keith, J. M. (Ed.), Bioinformatics, 1st Edition. Vol. 452 of Methods in Molecular Biology. Humana Press, Totowa, New Jersey, Ch. 16, pp. 331–364. 46

Jia, F., Lo, N., Ho, S. Y. W., 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. PLoS One 9 (5), e95722. 42, 62

Jow, H., Hudelot, C., Rattray, M., Higgs, P., 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Mol. Biol. Evol. 19 (9), 1591–1601. 47, 53

Jukes, T. H., Cantor, C. R., 1969. Evolution of protein molecules. In: Munro, H. N. (Ed.), Mammalian protein metabolism. Academic Press, New York, pp. 21–123. 29

Kass, R. E., Raftery, A. E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90 (430), 773–795. 6, 7, 59, 134

Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. J. Am. Stat. Assoc. 91 (435), 1343–1370. 47

Keilson, J., 1979. Markov chain models - rarity and exponentiality. Springer-Verlag, New York. 28

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16 (2), 111–120. 31

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29 (2), 170–179. 33

Knuth, K. H., Skilling, J., 2012. Foundations of inference. Axioms 1 (1), 38–73. 94

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86. 5

Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B., Ronquist, F., 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57 (1), 86–103. 140, 141, 144, 172

Lanave, C., Preparata, G., Sacone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20 (1), 86–93. 33

Lanfear, R., Hua, X., Warren, D. L., 2016. Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome Biol. Evol. 8 (8), 2319–2332. 151, 152, 159, 161, 163

Larget, B., 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. Syst. Biol. 62 (4), 501–511. 136, 138, 168

Larget, B., Simon, D. L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16 (6), 750–759. 4

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25 (17), 2286–2288. 4

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21 (6), 1095–1109. 4, 46

Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55 (2), 195–207. 8, 62, 64, 65, 69, 72, 74, 130, 131, 170

Lefebvre, G., Steele, R., Vandal, A. C., 2010. A path sampling identity for computing the Kullback-Leibler and J divergences. Comput. Stat. Data An. 54 (7), 1719 – 1731. 74, 80, 81, 102, 135

Lemmon, A. R., Moriarty, E. C., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53 (2), 265–277. 42

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24 (12), 2669–2680. 72, 88

MacKay, D. J. C., 2002. Information theory, inference and learning algorithms. Cambridge University Press, New York. 58, 62

Marshall, D. C., 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59 (1), 108–117. 49

Maturana R., P., 2017. Bayesian support for Evolution: detecting phylogenetic signal in a subset of the primate family. ArXiv preprint arXiv:1709.04588. 12

Mayrose, I., Friedman, N., Pupko, T., 2005. A gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21 (Suppl. 2), 151–158. 41

Meng, X.-l., Wong, W. H., 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat. Sinica 6 (4), 831–860. 66

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21 (6), 1087–1092. 107

Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptak, R. G., Gibbs, R. A., Hillis, D. M., 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proc. Natl. Acad. Sci. USA 99 (22), 14292–14297. 1

Minin, V., Abdo, Z., Joyce, P., Sullivan, J., 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52 (5), 674. 131

Moran, R. J., Morgan, C. C., O' Connell, M. J., 2015. A guide to phylogenetic reconstruction using heterogeneous models–a case study from the root of the placental mammal tree. Computation 3 (2), 177–196. 46

Mossel, E., Steel, M., 2007. How much can evolved characters tell us about the tree that generated them? In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press, New York, Ch. 14, pp. 384–412. 129

Mukherjee, P., Parkinson, D., Liddle, A. R., 2006. A nested sampling algorithm for cosmological model selection. Astrophys. J. Lett. 638 (2), L51–L54. 89, 94

Murray, I., 2007. Advances in Markov chain Monte Carlo methods. Ph.D. thesis, The University of London. 84, 140

Neal, R., 2003. Slice sampling. Ann. Stat. 31 (3), 705–767. 105, 106, 140

Neal, R., 2008. The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever., Radford Neal's blog, August 17. 65

Neal, R. M., 2001. Annealed importance sampling. Statistics and Computing 11 (2), 125–139. 8, 83, 84

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3 (5), 418–426. 39

Newton, M. A., Raftery, A. E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Roy. Statist. Soc. Ser. B 56 (1), 3–48. 7, 62, 64, 131

Nylander, J. A. A., Ronquist, F., Huelsenbeck, J. P., Nieves-Aldrey, J., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53 (1), 47–67. 4, 46, 62, 129

Oehlert, G. W., 1992. A note on the delta method. Am. Stat. 46 (1), 27–29. 63, 67

Ogata, Y., 1989. A Monte Carlo method for high dimensional integration. Numer. Math. 55 (2), 137–157. 69

Pagel, M., Meade, A., Crandall, K., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53 (4), 571–581. 62

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20 (2), 289–290. 5

Penny, D., Foulds, L. R., Hendy, M. D., 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297 (5863), 197–200. 12

Penny, D., McComish, B. J., Charleston, M. A., Hendy, M. D., 2001. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. J. Mol. Evol. 53 (6), 711–723. 45

Petris, G., Tardella, L., 2007. New perspectives for estimating normalizing constants via posterior simulation. Technical report, Università di Roma "La Sapienza". 8

Pickett, K. M., Randle, C. P., 2005. Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. Mol. Phylogenet. Evol. 34 (1), 203 – 211. 48

Posada, D., Buckley, T. R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53 (5), 793–808. 5, 6

Posada, D., Crandall, K. A., 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50 (4), 580–601. 3, 148, 164, 166, 173

Pullen, N., Morris, R. J., 2014. Bayesian model comparison and parameter inference in systems biology using nested sampling. PLoS One 9 (2), e88419. 9, 89

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 173, 175

Raftery, A. E., Newton, M. A., Satagopan, J. M., Krivitsky, P. N., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M. (Eds.), Bayesian Statistics. Vol. 8. Oxford University Press, Oxford, pp. 1–45. 63

Rambaut, A., Suchard, M. A., Xie, D., Drummond, A. J., 2014. Tracer v1.6. URL http://beast.bio.ed.ac.uk/Tracer 144

Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43 (3), 304–311. 4

Rannala, B., Zhu, T., Yang, Z., 2011. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29 (1), 325–335. 49, 50, 118

Rogers, J. S., Swofford, D. L., 1998. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide sequences. Syst. Biol. 47 (1), 77–89. 3

Rogers, J. S., Swofford, D. L., 1999. Multiple local maxima for likelihoods of phylogenetic trees: a simulation study. Mol. Biol. Evol. 16 (8), 1079–1085. 128

Roos, C., Zinner, D., Kubatko, L. S., Schwarz, C., Yang, M., Meyer, D., Nash, S. D., Xing, J., Batzer, M. A., Brameier, M., Leendertz, F. H., Ziegler, T., Perwitasari-Farajallah, D., Nadler, T., Walter, L., Osterholz, M., 2011. Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. BMC Evol. Biol. 11 (1), 77. 145, 147

Rzhetsky, A., Nei, M., 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9 (5), 945–967. 1

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425. 1, 3

Sanderson, M. J., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49 (4), 817–29. 5

Schliep, K., 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27 (4), 592–593. 149, 158, 175

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464. 6

Semple, C., Steel, M., 2003. Phylogenetics. Oxford lecture series in mathematics and its applications. Oxford University Press, Oxford. 15

Sivia, D. S., Skilling, J., 2006. Data analysis: a Bayesian tutorial. Oxford University Press. 94, 95

Skilling, J., 2006. Nested sampling for general Bayesian computation. Bayesian Analysis 1 (4), 833–860. 9, 10, 89, 93, 94, 95, 96, 101, 103, 104, 105, 128, 132, 134, 139

Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21 (4), 456–463. 2

Steel, M., 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43 (4), 560–564. 128

Steel, M., Pickett, K. M., 2006. On the impossibility of uniform priors on clades. Mol. Phylogenet. Evol. 39 (2), 585 – 586. 48

Steel, M. A., Hendy, M. D., Penny, D., 1988. Loss of information in genetic distances. Nature 336 (6195), 118. 1

Steel, M. A., Penny, D., 1993. Distributions of tree comparison metrics - some new results. Syst. Biol. 42 (2), 126–141. 151, 159

Suchard, M. A., Weiss, R. E., Sinsheimer, J. S., 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18 (6), 1001–1013. 7, 47, 49, 53

Sullivan, J., Holsinger, K. E., Simon, C., 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. 42 (2), 308–312. 164

Sullivan, J., Swofford, D., Naylor, G., 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16 (10), 1347–1356. 42

Sullivan, J., Swofford, D. L., 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50 (5), 723–729. 37

Sumner, J., Fernández-Sánchez, J., Jarvis, P., 2012a. Lie Markov models. J. Theor. Biol. 298, 16–31. 44

Sumner, J. G., Jarvis, P. D., Fernández-Sánchez, J., Kaine, B. T., Woodhams, M. D., Holland, B. R., 2012b. Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol. 61 (6), 1069–1074. 43

Swofford, D. L., 2003. PAUP: phylogenetic analysis using parsimony (and other methods) 4.0.b5. Sinauer Associates. 5

Swofford, D. L., Olsen, G. J., Waddell, P. J., Hillis, D. M., 1996. Phylogenetic inference. In: Hillis, D. M., Moritz, C., Mable, B. K. (Eds.), Molecular systematics (2nd ed.). Sinauer Associates, Sunderland, Massachusetts, pp. 407–514. 33

Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86. 33

Tuffley, C., Steel, M., 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147 (1), 63 – 91. 45

Walter, C., 2017. Point process-based Monte Carlo estimation. Stat. Comput. 27 (1), 219–236. 94

Wang, Y., Yang, Z., 2014. Priors in Bayesian phylogenetics. In: Chen, M., Kuo, L., Lewis, P. O. (Eds.), Bayesian phylogenetics: methods, computational algorithms, and applications. Chapman and Hall/CRC, New York, Ch. 2, pp. 5–23. 47, 49, 51, 53

Wu, R., Chen, M., Kuo, L., Lewis, P. O., 2014. Consistency of marginal likelihood estimation when topology varies. In: Chen, M., Kuo, L., Lewis, P. O. (Eds.), Bayesian phylogenetics: methods, computational algorithms, and applications. Chapman and Hall/CRC, New York, Ch. 6, pp. 113–127. 7

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., Chen, M.-H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60 (2), 150–160. 6, 8, 62, 64, 65, 72, 75, 78, 88, 106, 117, 119, 128, 129, 132, 147, 152

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10 (6), 1396–1401. 38

Yang, Z., 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39 (1), 105–111. 28, 33, 43, 53

Yang, Z., 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39 (3), 306–314. 40, 53, 164

Yang, Z., 1995. A space-time process model for the evolution of DNA sequences. Genetics 139 (2), 993–1005. 44

Yang, Z., 1996a. Among-site rate variation and its impact on phylogenetic analyses. Trends. Ecol. Evol. 11 (9), 367 – 372. 37, 39, 42, 53

Yang, Z., 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42 (5), 587–596. 46

Yang, Z., 1996c. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42 (2), 294–307. 2

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford. 3

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24 (8), 1586–1591. 5

Yang, Z., 2014. Molecular evolution: a statistical approach. Oxford University Press, Oxford. 16, 42, 45, 54

Yang, Z., Lauder, I. J., Lin, H. J., 1995. Molecular evolution of the hepatitis B virus genome. J. Mol. Evol. 41 (5), 587–596. 46

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14 (7), 717–724. 4

Yang, Z., Rannala, B., 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54 (3), 455–470. 3, 48, 49, 50, 51

Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13 (5), 303–314. 5

Yang, Z., Roberts, D., 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12 (3), 451–458. 46

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. In: Bryson, V., Vogel, H. J. (Eds.), Evolving Genes and Proteins. Academic Press, New York, pp. 97–166. 17

Zwickl, D., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis, The University of Texas at Austin. 2

Zwickl, D. J., Holder, M. T., 2004. Model parameterization, prior distributions, and the general time-reversible model in Bayesian phylogenetics. Syst. Biol. 53 (6), 877–888. 33, 34, 52