arXiv:1703.05471v2 [q-bio.QM] 6 Dec 2017

Model selection and parameter inference in phylogenetics using Nested Sampling

Patricio Maturana R.[1*] Brendon J. Brewer[1] Steffen Klaere[1,2] Remco Bouckaert[3,4]

[1]Department of Statistics, University of Auckland, Auckland, New Zealand

[2]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[3] Center of Computational Evolution, University of Auckland, Auckland, New Zealand

[4]Max Planck Institute for the Science of Human History, Jena, Germany

Abstract

Bayesian inference methods rely on numerical algorithms for both model selection and parameter inference. In general, these algorithms require a high computational effort to yield reliable inferences. One of the major challenges in phylogenetics is the estimation of the marginal likelihood. This quantity is commonly used for comparing different evolutionary models, but its calculation, even for simple models, incurs high computational cost. Another interesting challenge relates to the estimation of the posterior distribution. Often, long Markov chains are required to get sufficient samples to carry out parameter inference, especially for tree distributions. In general, these problems are addressed separately by using different procedures. Nested sampling (NS) is a Bayesian computation algorithm which provides the means to estimate marginal likelihoods together with their uncertainties, and to sample from the posterior distribution at no extra cost. The methods currently used in phylogenetics for marginal likelihood estimation lack of practicality due to their dependence on many tuning parameters and their incapacity to provide a direct way to calculate the uncertainties associated with the estimates, unlike NS. In this paper, we introduce NS to phylogenetics. Its performance is analysed under different scenarios and compared to established methods. We conclude that NS is a competitive and attractive algorithm for phylogenetic inference. An implementation is available as a package for BEAST 2, available from `https://github.com/BEAST2-Dev/nested-sampling` under the LGPL licence.

Key Words: model selection, parameter inference, nested sampling, marginal likelihood

1 Introduction

Bayesian methods provide a comprehensive framework in which to explore parameter space, uncertainty and model-to-data fitness. The concept was introduced to phylogenetics in the 1990s (Rannala and Yang, 1996; Yang and Rannala, 1997; Larget and Simon, 1999), and has gained popularity because of its flexibility when dealing with complex models and large data sets, in contrast with maximum likelihood estimation. The increase in computational power led to the rise of Bayesian methods, as Markov Chain Monte Carlo (MCMC) approaches became feasible. Its popularity was further increased by state-of-the-art implementations of the models in programs like MrBayes (Huelsenbeck and Ronquist, 2001), BEAST (Drummond and Rambaut, 2007; Bouckaert et al., 2014), and PhyloBayes (Lartillot et al., 2009).

As in other fields, model selection plays an integral part in phylogenetic inference. A wide variety of different criteria are available for this task, with some of the most popular being the Likelihood Ratio Test (LRT), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Bayes Factors (BF), which

*Corresponding author. Address for correspondence: Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. Email address: p.russel@auckland.ac.nz (Patricio Maturana R.)

are ratios of marginal likelihoods. The latter is of particular interest as it provides many advantages over the other methods: i) It is a direct consequence of probability theory used as a theory of reasoning; ii) it allows comparison of nested and non-nested models; iii) it is not based on a point estimate in parameter space since it averages over parameter space; and iv) it embodies Occam's razor by involving the prior distribution in the model selection process. Marginal likelihoods penalize the inclusion of a new parameter when its value is unknown and some of the possible values do not fit the data well. However, the marginal likelihood is a difficult integral that depends on the complexity of the model. Finding ways to efficiently estimate this integral is one of the major challenges of the field.

A simple Monte Carlo method for estimating the marginal likelihood is the harmonic mean (Newton and Raftery, 1994). Despite its popularity, it is well-known to overestimate the real value, and in many situations its variance is infinite. Among the most accurate methods currently used in phylogenetics are path sampling (Lartillot and Philippe, 2006) and steppingstone sampling (Xie et al., 2011) which are much more accurate but have a high computational cost. The latter has gained popularity in recent years due to its implementation in different phylogenetic software packages (BEAST2; Bouckaert et al., 2014). However, these methods have a relatively large number of tuning parameters that need to be set prior to analysis, and there is no rigorous method of determining the values appropriate for the accurate estimation of the marginal likelihood. Also, these methods have problems dealing with some likelihood shapes (Skilling, 2006).

To efficiently deal with the above issues a generalised version of stepping stone sampling (GSS) has been proposed (Fan et al., 2011). This generalization also allows us to regard the phylogeny as an unknown parameter (Holder et al., 2014; Baele et al., 2016) incorporating the uncertainty in the tree topology in model selection.

A more general technique for the estimation of the marginal likelihood is nested sampling (NS; Skilling, 2006). This method requires less tuning and can deal with partly convex likelihood functions (in the NS sense, explained below). Its main feature is the reduction of the multidimensional integral over parameter space to a one-dimensional integral of the likelihood as a function of the enclosed prior probability. This technique, and several variants (e.g. Feroz et al., 2009; Brewer et al., 2011; Handley et al., 2015) have been successfully applied to fields like astronomy (Mukherjee et al., 2006; Brewer and Donovan, 2015) and systems biology (Aitken and Akman, 2013; Pullen and Morris, 2014) and have shown great promise in parameter inference and model selection.

In this paper, we assess the merits of nested sampling in phylogenetic inference, and compare it to established methods. Firstly, a reasonably big dataset, which contains several partitions and many parameters, is analysed in order to show the consistent marginal likelihood estimates and parameter inferences yielded by NS. Secondly, two datasets, which have become standard phylogenetic datasets for the analysis of MCMC method performance, are analysed for marginal likelihood estimation and parameter inference.

2 Bayesian inference

Let $\boldsymbol{\theta}$ be the vector of parameters, $\boldsymbol{X}$ the data, and M the model (assumed throughout). Bayes' theorem is given by

$$f(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{f(\boldsymbol{X}|M)}. \tag{1}$$

The prior distribution $\pi(\boldsymbol{\theta}|M)$ represents our previous knowledge of the parameters which is updated after taking into account the data; the updated knowledge is reflected in the posterior probability distribution $f(\boldsymbol{\theta}|\boldsymbol{X}, M)$. The likelihood function $L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ represents the probability of the data given the parameters and the model. The marginal likelihood $f(\boldsymbol{X}|M)$ 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 the *evidence* (MacKay, 2002). To understand its role, note that the posterior distribution for the model

M_j is given by

$$f(M_j|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|M_j)f(M_j)}{f(\boldsymbol{X})}, \quad j = 0, 1,$$

where $f(\boldsymbol{X}|M_j)$ is the marginal likelihood as defined in (1), $f(M_j)$ is the prior probability for the model, and $f(\boldsymbol{X})$ is the probability of the data. The marginal likelihood will also be denoted by "$\mathcal{Z}$" henceforth. The Bayesian comparison of two models M_0 and M_1 can be carried out by comparing their posterior probabilities. This comparison is often through the ratio of their probabilities, which represents the plausibility of one model over another and is defined as follows:

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

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 the marginal likelihoods. Here lies the importance of the latter.

Although the marginal likelihood is, in general, ignored in parameter inference, it plays a key role in model selection: it is a measure of the goodness of fit. Indeed, it is the probability of the data given the model, i.e., it is by definition a measure of model fit. The marginal likelihood acts as the normalisation constant in the posterior distribution making it a probability density function. Thus, this quantity is a multidimensional integral of the prior distribution times the likelihood function over the parameter space. MCMC methods used for parameter estimation within a model use only ratios of posterior densities, and are therefore unable to measure its normalisation 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 down-weight 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.

2.1 Estimation of marginal likelihoods

Typically, phylogenetic models involve a high level of complexity, making it difficult to calculate the marginal likelihood. Suchard et al. (2001) proposed the Savage-Dickey ratio to estimate Bayes factors for nested models (Verdinelli and Wasserman, 1995). Huelsenbeck et al. (2004) used reversible jump Markov chain Monte Carlo including all possible time-reversible models. Nevertheless, these methods are restricted to a particular group of models. Other alternatives have been proposed to allow a more general comparison of models. Among them, the harmonic mean (HM) is the most popular to estimate the marginal likelihood (Newton and Raftery, 1994), an importance-sampling approach. Its popularity is due to its simplicity, it only requires samples from the posterior distribution. However, the HM estimator often has infinite variance, overestimates the true value of the marginal likelihood, and does not work in high dimensions, the usual case in phylogenetics (Newton and Raftery, 1994; Lartillot and Philippe, 2006; Xie et al., 2011; Baele et al., 2013a, 2012, 2016).

Far more accurate than the HM method is path sampling (PS), also known as thermodynamic integration, proposed in phylogenetics by Lartillot and Philippe (2006). The method requires several Markov chains from transition distributions which form a path between the prior and the posterior distribution. These transition functions are defined by the "power posterior"

$$p_\beta = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{\mathcal{Z}_\beta}, \quad \text{for} \quad 0 \leq \beta \leq 1,$$

where $\mathcal{Z}_\beta$ is the normalizing constant of the unnormalized power posterior density $L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)$. Similarly, a path between the posterior of two models could be defined to estimate the Bayes factor directly. Note that for $\beta = 0$ the power posterior is equivalent to the prior distribution and for $\beta = 1$ is equivalent to the posterior distribution. In the latter case, $\mathcal{Z}_1 = \mathcal{Z}$ is the marginal likelihood. PS relies on the identity

$$\log \mathcal{Z} = \int_0^1 \mathbb{E}_{p_\beta}\big[\log L(\boldsymbol{X}|\boldsymbol{\theta}, M)\big]\mathrm{d}\beta,$$

where the expected value is with respect to the power posterior distribution p_β. PS uses a series of β values which define the transition distributions. For each value, a Markov chain is required to estimate the expected values and consequently the integral over β. Clearly, the increase in accuracy comes at the cost of increased complexity.

Another importance sampling approach is stepping-stone sampling (SS) proposed by Xie et al. (2011). SS relies on transition distributions like PS in order to define an equivalence between the marginal likelihood and the telescope product of ratios of normalizing constants given by

$$\mathcal{Z} = \prod_{k=1}^{K} \frac{\mathcal{Z}_{\beta_k}}{\mathcal{Z}_{\beta_{k-1}}},$$

where $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-1} < \beta_K = 1$. Each ratio $\mathcal{Z}_{\beta_k}/\mathcal{Z}_{\beta_{k-1}}$ is estimated by importance sampling. The performance of this method is similar to PS. However, SS requires slightly less computational effort: it does not need posterior samples and in general requires a smaller number of transition distributions to reduce its discretization bias than PS (Xie et al., 2011). SS also allows us to estimate the Bayes factor directly defining a path between the posterior of both models (Baele et al., 2013a). The extended version of SS, generalized steppingstone sampling (GSS; Fan et al., 2011), uses a reference distribution to shorten the distance between the prior and posterior distribution. This strategy can potentially lead to a more efficient estimation process. Like in its predecessor, the geometric path is often used to connect these densities, which is defined by

$$p_\beta = \frac{\big(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\big)^\beta \pi_0(\boldsymbol{\theta}|M)^{1-\beta}}{\mathcal{Z}_\beta}, \quad \text{for} \quad 0 \leq \beta \leq 1,$$

where π_0 is the reference distribution. When $\beta = 0$ and $\beta = 1$, the power posterior is equivalent to the reference and posterior distribution, respectively. GSS is more accurate than the original version when the reference distribution approximates the posterior distribution reasonably well. This method has also been extended, allowing the tree topology to be variable (Holder et al., 2014; Baele et al., 2016). This allows the user to accommodate phylogenetic uncertainty in model selection. GSS has the potential of leading to remarkable improvements in comparison to the original SS and PS, that is, 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, in which cases PS/SS overestimate the true marginal likelihood (Baele et al., 2016).

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 commonly used 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, or for big datasets. Actually, flat priors are most often incorrectly used and constitute improper priors, challenging MCMC sampling (Baele et al., 2013b). Secondly, the path described by the β-values has to be defined. 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 concentrate the computational effort in that place. For example Lepage et al. (2007) used a sigmoidal function to estimate

the Bayes factor using PS; Friel and Pettitt (2008) proposed $\beta_k = x_k^4$ in PS, 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 $\alpha = 0.3$. 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 tuning parameters that might be unsuitable.

In this context, GSS requires 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. Unlike SS or PS, the β-values do not necessarily need to follow any particular distribution to effectively control the uncertainty of the estimate, 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 (Xie et al., 2011; Baele et al., 2013a; Arima and Tardella, 2014; Baele and Lemey, 2014). However, these methods only work when the shape of the likelihood, as a function of the cumulative prior probabilities (see Figure 1), is concave. Partly convex likelihood functions might make them require impractical computational effort or make them fail outright (Skilling, 2006). 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, for instance, to cope with partly convex likelihood functions. More importantly, unlike PS, SS and GSS, NS requires less problem-specific tuning.

3 Nested Sampling

Here we explain nested sampling in more detail than the original paper (Skilling, 2006) and give details on application to phylogenetic inference. The marginal likelihood or evidence, in a simplified notational version, is given by

$$\mathcal{Z} = \int_{\Theta} \pi(\boldsymbol{\theta}) L(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{2}$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector, $L(\boldsymbol{\theta})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is the prior distribution. All the conditionals have been omitted, that is, $L(\boldsymbol{\theta})$ is written as the likelihood function instead of $L(\boldsymbol{\theta}|\boldsymbol{X}, M)$ and $\pi(\boldsymbol{\theta})$ is written as the prior instead of $\pi(\boldsymbol{\theta}|M)$.

This definition applies for a continuous parameter space Θ. This is the case when the phylogeny is fixed. When the tree topology is unknown, the parameter space is additionally composed by a discrete part. In this case, the marginal likelihood incorporates the sum over the tree parameter space and is known as *total marginal likelihood*. Strictly speaking, its definition is given by

$$\mathcal{Z} = \sum_{\tau \in \mathcal{T}} \int_{\mathcal{V}_\tau} \int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, \nu_\tau, \tau, M) \pi(\boldsymbol{\theta}, \nu_\tau, \tau|M) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\nu_\tau,$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter vector composed by elements such as frequencies, gamma parameter and rates parameters, $\nu_\tau \in \mathcal{V}_\tau$ is the set of branch lengths of $\tau \in \mathcal{T}$ which is the tree topology, $\boldsymbol{X}$ is the molecular data and M is the substitution model. For simplicity, NS will be described for definition (2), but its generalization to the case of variable tree topology is analogous.

To understand the key idea of NS, consider that for any positive random variable Y, its expected value can

be written as

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

which depicts the area between the distribution function of Y and 1, where F is the cumulative distribution function of Y. Similarly, the likelihood function $L(\boldsymbol{\theta})$ can be seen as a positive random variable where $\boldsymbol{\theta}$ follows the prior distribution $\pi(\boldsymbol{\theta})$ and the evidence as the expected value of the likelihood function. Thus, nested sampling takes advantage of this property to transform the multi-dimensional integral defined in (2) into a one-dimensional integral as follows

$$\mathbb{E}_\theta\big[L(\boldsymbol{\theta})\big] \equiv \mathbb{E}_\lambda\big[\lambda\big] = \int_0^\infty \big(1 - F(\lambda)\big)\mathrm{d}\lambda, \tag{3}$$

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

$$F(\lambda) = \int_{L(\boldsymbol{\theta})<\lambda} \cdots \int \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Considering $\xi(\lambda) = 1 - F(\lambda)$, the proportion of prior mass with likelihood greater than λ, and taking its inverse, the evidence given in (3) is redefined as

$$\mathcal{Z} = \int_0^1 L(\xi)\mathrm{d}\xi.$$

This is the integral used by nested sampling, and is displayed in Figure 1. In general, this function concentrates its mass near zero because the posterior is located in a small area of the prior. We use the "overloaded" notation, where the same letter L represents the likelihood function over different domains: $L(\boldsymbol{\theta})$ has the parameter vector $\boldsymbol{\theta}$ as argument, and $L(\xi)$ has the prior mass ξ (scalar) as argument. Note that $L(\xi)$ is a monotonically decreasing function which reaches its highest point at $\xi = 0$ and its lowest point at $\xi = 1$ (see Figure 1). $L(0.9) = 0.3$ means that 90% of the draws $\boldsymbol{\theta}$ from the prior distribution will have likelihoods greater than 0.3. If a set of points on the $L(\xi)$ curve can be obtained, the integral can be approximated numerically by the basic standard quadrature method

$$\mathcal{Z} \approx \sum_{i=1}^{k} w_i L_i, \tag{4}$$

where $w_i = \xi_{i-1} - \xi_i$ (or $w_i = (w_{i-1} - w_{i+1})/2$ for the trapezoidal rule) and $L_i = L(\xi_i)$. For a decreasing sequence of ξ-values and an increasing sequence of L-values the evidence can be estimated. How to generate these sequences is described below.

3.1 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 (4). 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 (4). 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.

Figure 1: Association between the cumulative prior mass and the likelihood function. Nested sampling estimates the gray area which is the marginal likelihood. In general, a small area of the prior concentrates high likelihood values causing the area to be concentrated around $\xi \approx 0$.

3.2 Sequence of ξ-values

The discarded points generate an increasing sequence of likelihoods, which are known precisely. An important insight of Skilling (2006) is that the corresponding ξ values, while they cannot be measured precisely, can be estimated from the nature of the NS procedure. The uncertainty of the NS estimate is mainly due to these approximations.

Nested sampling explores the prior distribution geometrically as follows

$$\xi_0 = 1, \;\; \xi_1 = t_1, \;\; \xi_2 = t_1 t_2, \quad \ldots \quad , \;\; \xi_k = \prod_{i=1}^{k} t_i,$$

where $t_i = \xi_i / \xi_{i-1} \in [0, 1]$, for $i = 1, \ldots, k$. This variable follows a Beta$(N, 1)$ distribution. This is because at the i^{th} iteration, NS takes N x_i points which follows a Uniform$(0, \xi_{i-1})$, with $i = 1, \ldots, N$. These values are cumulative probabilities and consequently have a uniform distribution. Their maximum value is ξ_i which is related to the minimum likelihood value (note that $L(\xi)$ is a non-increasing function). Since the distribution of x_i/ξ_{i-1} is a Uniform$(0, 1)$, their maximum value ξ_i/ξ_{i-1} follows a Beta$(N, 1)$ distribution.

Skilling (2006) defined two schemes for estimating the ξ-values: *stochastic* and *deterministic*.

- *Stochastic:* the t_i values are generated randomly from the Beta$(N, 1)$ distribution, for $i = 1, \ldots, k$.
- *Deterministic:* the t_i values are fixed by using their expectations as follows:
  - Considering its *arithmetic mean*, $t_i = N/(N + 1)$, approximate ξ-values would be given by

$$\xi_i = \left(\frac{N}{N+1} \right)^i.$$

– Considering its *geometric mean*, $t_i = e^{-1/N}$, the estimated prior mass would be

$$\xi_i = e^{-i/N}.$$

Thus, a sequence of ξ values can be generated and used in (4). 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 alternate version of NS with unbiased estimates of $\mathcal{Z}$. On the other hand, the use of the stochastic approach has the potential of estimating more accurately the uncertainty by replicating the estimates for different ξ-sequences.

3.3 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 fulfil 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 (Mukherjee et al., 2006; Feroz et al., 2009; Brewer et al., 2011). We use Skilling's method to generate the restricted prior samples in our application.

Unlike the proposal mechanisms used in standard MCMC methods to sample the posterior, a static distribution, in NS such mechanisms have to deal with a variable target distribution over time. In particular, this is a new scenario for tree proposals. Nested sampling compresses the prior at each iteration making it vary at a constant rate. The proposals have to explore a quite wide area at the beginning which becomes constrained over time. Tree proposal mechanisms should be able to adapt to this sampling characteristic. Frequently, a uniform prior distribution is assigned over the tree parameter space which is quite huge even for few taxa. Initially, bold moves would allow a good exploration using less steps than conservative ones. However, the acceptance probability would decrease drastically over time due to the fact that the target distribution gets 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 bold moves. Ideally, the proposal mechanism should take into account this dynamical behaviour of the target distribution over time. Brewer and Foreman-Mackey (2016) stated that heavy-tailed proposals are as efficient as slice sampling (Neal, 2003), at least in simple experiments. In this work we use the operators implemented in BEAST2 (Bouckaert et al., 2014) and described in Drummond and Bouckaert (2015), but switched off auto-optimisation features, since the target distribution changes through time.

3.4 Information

The idea of how much we have learned from the data is quantified through the notion of entropy. The measure of information (Sivia and Skilling, 2006; Knuth and Skilling, 2012) 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 data. By definition, it can be seen as the expected value $H = \mathbb{E}_P[\log(P(\boldsymbol{\theta})/\pi(\boldsymbol{\theta}))]$. It can be approximated by

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

with $w_i = \xi_{i-1} - \xi_i$ (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

$$\log \mathrm{G}_Y = \mathbb{E}[\log Y] \Leftrightarrow \mathrm{G}_Y = e^{\mathbb{E}[\log Y]}. \tag{5}$$

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 one that supports the same parameter values, has a lower information than one which likelihood function 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.

3.5 Uncertainty

The numerical uncertainty associated to the NS estimation of $\mathcal{Z}$ comes from two sources: i) approximating the prior volume ($w_i = \xi_{i-1} - \xi_i$), 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).

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

$$\mathrm{dev}\big[\log \mathcal{Z}\big] = \sqrt{\frac{H}{N}}. \tag{6}$$

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 ξ-sequences, i.e., using the stochastic approach, but keeping the same likelihood sequence. Thus, a distribution of $\log \widehat{\mathcal{Z}}$ 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.

3.6 Algorithm

The algorithm iterates between the following steps:

1. Sample N points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from the prior $\pi(\boldsymbol{\theta})$;
2. Initialize $\mathcal{Z} = 0$ and $\xi_0 = 1$;
3. Repeat for $i = 1, \ldots, k$;
   - **i)** out of the N live points, take the one with the lowest likelihood which we call $\boldsymbol{\theta}_l$ with corresponding likelihood $L_i = L(\boldsymbol{\theta}_l)$, where $l \in \{1, \ldots, N\}$;
   - **ii)** set $\xi_i = \exp(-i/N)$;
   - **iii)** set $w_i = \xi_{i-1} - \xi_i$ (or $w_i = (\xi_{i-1} - \xi_{i+1})/2$ for the trapezoidal rule);
   - **iv)** update $\mathcal{Z} = w_i L_i + \mathcal{Z}$; and

v) update the set of active points $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ replacing $\boldsymbol{\theta}_l$ by drawing a new point $\boldsymbol{\theta}$ from the prior distribution restricted to $L(\boldsymbol{\theta}) > L(\boldsymbol{\theta}_l)$.

Repeat the routine until a given stopping criterion is satisfied. However, there is not a rigorous criterion that guarantees that we have found most of the bulk of $\mathcal{Z}$. Nevertheless, some termination conditions have been proposed (Skilling, 2006), which are discussed below.

3.6.1 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

$$\max\big(L(\boldsymbol{\theta}_1), L(\boldsymbol{\theta}_2), \ldots, L(\boldsymbol{\theta}_N)\big) w_i < \gamma \mathcal{Z}.$$

The algorithm can be stopped 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 $\xi \approx e^{-H}$, when most of the posterior mass in the prior has been found. Given that $\xi_i \approx e^{-i/N}$, a natural termination condition to estimate the log-evidence would be stopping the loop when i/N significantly exceeds H, i.e., when the posterior mass has been explored completely.

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 won't happen. In this work, we use the relative error as termination criterion.

3.7 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 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k$. These discarded points have contributed to estimate the marginal likelihood with their respective weights wL which are proportional to the posterior distribution, in other words, 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 (ESS) is related to the entropy of the posterior weights (Skilling, 2006) as

$$\text{ESS} = \exp\left(-\sum_{i=1}^{k} p_i \log p_i\right), \quad \text{where} \quad p_i = \frac{w_i L_i}{\mathcal{Z}}.$$

4 Application

The NS algorithm is assessed under different phylogenetic scenarios in order to show its performance for marginal likelihood estimation and parameter inference. First, a reasonably big dataset (Horn et al., 2014) where the alignment has been split up into several partitions, each with their own site model, is used to carry out model selection via SS and NS. In addition, it is used to assess the information provided by NS about the posterior distribution in comparison to a standard MCMC method. Then, two datasets (Tetrapod and Chloridoideae) consisting of sequences of eukaryote species are analysed. These datasets form part of a group of standard datasets for evaluating MCMC methods (Lakner et al., 2008; Höhna and Drummond, 2012; Larget, 2013; Whidden and Matsen, 2015) and possess interesting characteristics, challenging standard MCMC methods. NS is performed using the deterministic approach to generate the ξ-sequence and the trapezoidal

Model	N	H	SD	Iterations	$\log \widehat{\mathcal{Z}}$	lower	upper	SS
Strict clock	1	1356.10	36.83	1,427	-69611.60	-69683.78	-69539.42	-69603.88
Relaxed clock	1	1604.64	40.06	1,689	-69100.21	-69178.72	-69021.70	-69054.25

Table 1: NS and SS marginal likelihood estimates for the Euphorbia dataset. NS includes its 95% confidence intervals, which contain the SS estimates and make evident its potential for model selection even under the simplest specifications (using a single active point).

rule as method for the numerical integration. We use the relative error as termination criterion with a error tolerance 1e-13. All the analyses are performed in BEAST2 (Bouckaert et al., 2014) and the plots produced in R (R Core Team, 2015).

4.1 Euphorbia

This dataset (Horn et al., 2014) contains 197 taxa with 6,328 nucleotides. Only chloroplast sites were used, divided into 15 partitions. This represents a challenging case since involves many parameters and consequently a huge parameter space. We evaluate NS performance for model selection and compare it to SS results. Also, we assess the ability of NS in sampling the posterior distribution and compare it to a standard MCMC method.

4.1.1 Model selection

Two clock models are compared: an *strict clock model* and a *relaxed clock model* (Drummond et al., 2006). These two clock models are compared through their marginal likelihoods. First, we estimate these values by using SS with 400 steps and 500 samples per each transitional distribution. These specifications have been tested and yield reliable estimates (results not shown). The analysis is replicated using NS with a single active point and 30,000 MCMC steps per NS iteration, to generate the samples required.

The results are displayed in Table 1. SS estimates show the better fit of the relaxed clock model in comparison to the strict clock model; the former has a marginal likelihood substantially higher than the latter. In terms of the Bayes factor, there is strong evidence in favour of the relaxed clock model (Kass and Raftery, 1995).

NS is consistent with SS taking into account the uncertainty associated with the estimates. Actually, the 95% confidence intervals contain the SS estimates in both cases. The model selection can be made based on these intervals since they do not overlap.

This example shows the effectiveness of NS to carry out model selection. In the hypothetical case in which the intervals would have overlapped, the analyses should have been redone but increasing the number of active points in order to decrease the uncertainty and consequently the width of the intervals.

4.1.2 Parameter inference

Recycling the NS run executed in the previous analysis, it is possible to carry out parameter inference. For this, we recalculate the posterior weights for a new sequence of ξ-values (obtained from a Beta distribution), and generate a new posterior sample as described in Section 3.7. The mean and the standard deviation are calculated from this sample. This procedure was replicated 1,000 times registering these statistics. Note that this procedure is not computationally expensive since it does not require likelihood evaluations. In addition, an MCMC analysis is performed and its statistics registered. The chain length is 50,000,000 with a burn-in period of 10% and thinning factor of 10,000.

The results are displayed in Table 2. The NS posterior sample size fluctuated between 9 and 47 points, with a mean of around 15 and standard deviation of 7.3. NS point estimates also include their corresponding 95% confidence intervals. It is apparent that all the intervals contain the MCMC estimates for the means as well as

	MCMC		NS					
	Mean	SD	Mean	lower	upper	SD	lower	upper
Posterior	-66831.62	17.99	-66838.90	-66904.90	-66772.91	14.36	0	35.51
Likelihood	-67524.00	16.32	-67531.22	-67602.26	-67460.19	13.57	0	37.48
Prior	692.39	7.99	692.32	683.45	701.18	5.29	0.99	9.59
TreeHeight	0.04	0.00	0.04	0.04	0.05	0.00	0.00	0.00
YuleModel	728.35	7.89	728.33	719.28	737.37	5.39	1.04	9.74
birthRate	112.63	9.20	113.22	105.35	121.09	7.06	1.34	12.77
kappa.rbcL_pos3	5.04	0.45	5.09	4.75	5.43	0.36	0.14	0.59
kappa.rpl16_pos1	1.34	0.47	1.47	1.07	1.87	0.42	0.16	0.67
kappa.rpl16_pos3	3.85	1.21	4.15	3.07	5.23	1.04	0.16	1.93
kappa.rpl16ex_pos2	7.79	3.67	7.33	4.52	10.14	2.55	0.16	4.95
ucldStdev	0.57	0.03	0.56	0.52	0.60	0.02	0.01	0.04
rate.mean	0.93	0.04	0.93	0.89	0.98	0.02	0.00	0.04
rate.variance	0.38	0.06	0.38	0.32	0.43	0.04	0.01	0.07
rate.coefficientOfVariation	0.62	0.04	0.61	0.57	0.65	0.03	0.01	0.05

Table 2: MCMC and NS posterior means and standard deviations of estimated features for the Euphorbia dataset – see Figure 2 for remaining kappa and relative substitution rate estimates. NS estimates also include their corresponding 95% confidence intervals, which in each case contain the MCMC point estimates.

for the standard deviations of the posterior distribution. These include posterior, likelihood, and prior values as well as some parameters. Figure 2 shows the 95% confidence intervals for the NS estimates for some other parameters which are in similar scales. The points stand for the MCMC estimates, which are all within the NS intervals.

4.2 Tetrapod

The DS1 alignment from Höhna and Drummond (2012) consists of 27 ribosomal RNA sequences of tetrapod with 1,949 nucleotides (Hedges et al., 1990). Its most remarkable feature is its tree space, which contains separate regions which form "islands" with high posterior probabilities (Höhna and Drummond, 2012). Whidden and Matsen (2015) showed that two of these islands are separated by only 2 SPR operations, but that the intermediate topology is so unlikely that it was never visited in their MCMC analysis. In general, the MCMC chains tend to get stuck in one of these tree islands, a common problem for standard MCMC methods.

These characteristics make DS1 a good case to assess the performance of NS and compare it to standard methods. We evaluate its attributes for parameter inference and marginal likelihood estimation. It will be assumed a GTR and a relaxed clock models are suitable. As prior for the tree topologies, a Yule model is assumed with a Uniform(0, 150) for its birth rate; for the clock rate a lognormal distribution; for the relative rates, the default Gamma priors implemented in BEAST2.

4.2.1 Parameter inference

Two independent MCMC analyses are performed in order to sample from the posterior distribution. Each chain has a length of 30 million with a burn-in period of 33%. The samples are taken every 10,000 steps. Furthermore, two independent NS are performed in order to sample the posterior. For each run, 100 active points are considered with 20,000 steps in order to generate the independent points required by NS at each iteration.

The posterior values for the 2 MCMC chains are displayed in Figure 3. It is apparent that they converge to different distributions. It is highly probable that the Markov chains got stuck on one of the tree islands, being unable to escape. Figure 4 shows their posterior clade probabilities which reflect how different the samples are. Note that there are clades with 100% support in one sample, whereas in the other sample, the same

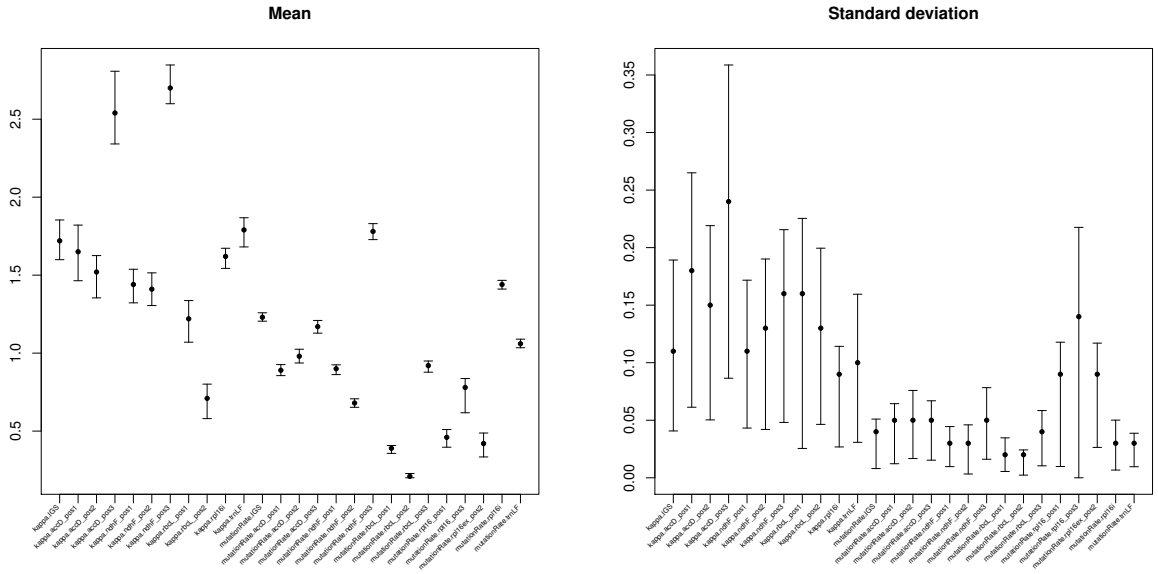

Figure 2: NS and MCMC estimates of the mean and standard deviation for some parameters in the Euphorbia dataset. The intervals are inferred by using NS and the dots stand for the point estimates obtained from the MCMC analysis. All the NS intervals contain the MCMC point estimates, suggesting its effectiveness in parameter inference.

clades have 0% support. On the other hand, NS yields samples with consistent clade probabilities (Figure 4). The low values are notably in total agreement. They tend to differ in case of problems in the sampling. NS posterior sample sizes in this example were of around 2,300. The analysis was replicated multiple times, obtaining similar posterior clade probabilities from the NS analyses (not shown).

NS has the potential of exploring the parameter space in different areas at each iteration, granted by the active points. Even in the case of using a single active point, the particle explores the parameter space according to the prior, with the likelihood restriction, but not according to the posterior, which embodies the difficulties.

4.2.2 Marginal likelihood estimation

The marginal likelihood is estimated by means of SS and NS. For the SS algorithm, the estimation is carried out by using 384 and 2,560 steps. In general, these specifications should be enough to yield reliable estimates. For NS, 1 and 100 active points are considered, respectively.

Table 3 displays the results. For the two different specifications in SS, the estimates are quite similar which could lead one to trust the outcome. However, these values are not around the NS estimates. Presumably, SS has some problems sampling near the posterior distribution as was observed in the MCMC analyses carried out before, which could lead to the underestimation. This is also a situation in which GSS would fail in estimating the marginal likelihood, due to the incapacity of the MCMC methods in approximating the posterior and consequently, in generating the reference distribution (see Appendix 6.1 for an example which illustrates this situation).

4.3 Chloridoideae

The DS9 alignment (Höhna and Drummond, 2012) consists of 67 sequences of Chloridoideae with 955 nucleotides (Ingram and Doyle, 2004). Its tree space is rather flat which requires a big sample size in order to

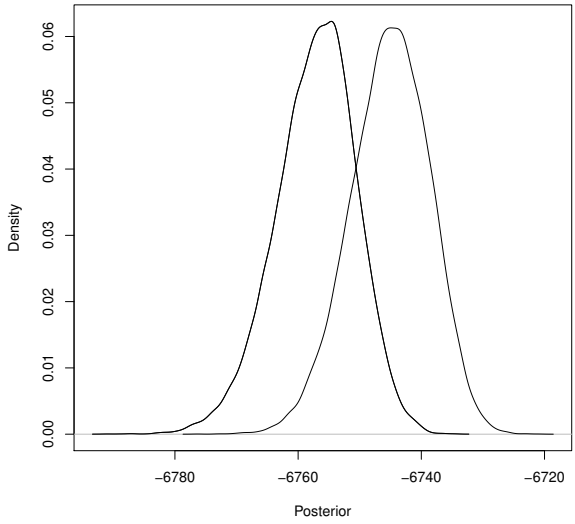

Figure 3: Posterior densities obtained from two independent MCMC chains for the Tetrapod dataset, converging to different distributions due to the tree islands in the parameter space.

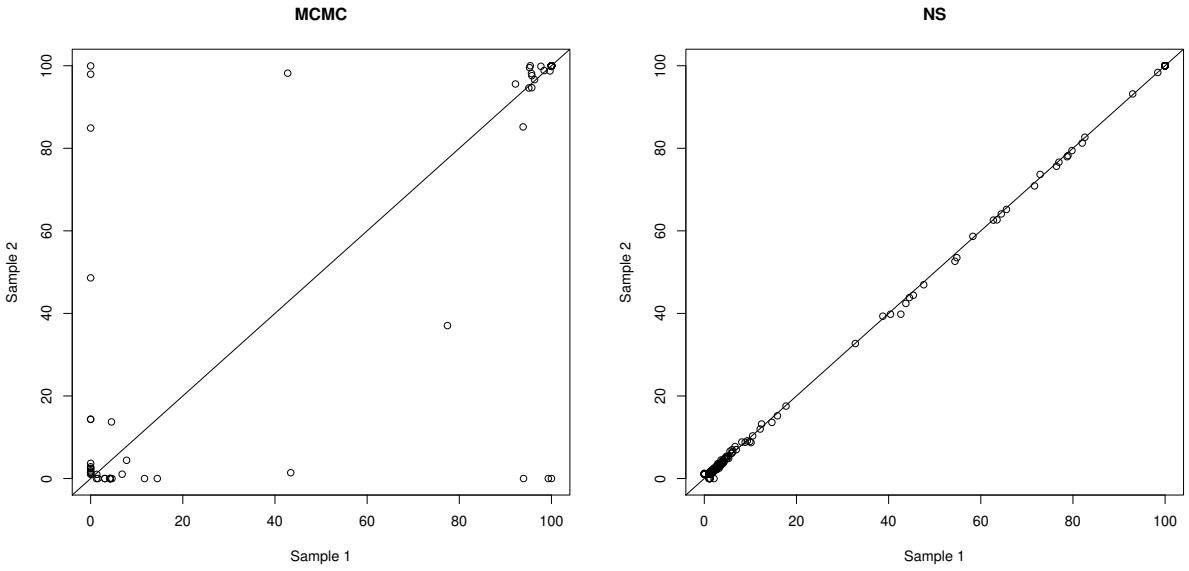

Figure 4: Comparison of 2 MCMC and NS posterior clade probabilities for the Tetrapod dataset. The graph on the left shows the difficulty of sampling a parameter space which contain tree islands, where the MCMC chains tend to get stuck and consequently converge to different distributions. The graph on the right shows consistency of NS yielding posterior samples in independent runs, unlike standard MCMC methods.

Method	Steps	Samples		$\log \widehat{\mathcal{Z}}$	
SS	384	500	-6956.49	-6959.06	-6955.41
SS	2560	500	-6956.99	-6958.36	-6953.69
	N	H	SD	Iterations	$\log \widehat{\mathcal{Z}}$
NS	1	149.36	12.22	183	-6952.69
NS	100	146.90	1.21	17,754	-6950.93

Table 3: Marginal likelihood estimates obtained through SS and NS under different specifications for the Tetrapod dataset. SS does not yield estimates around the NS ones due to its difficulty in sampling near the posterior distribution.

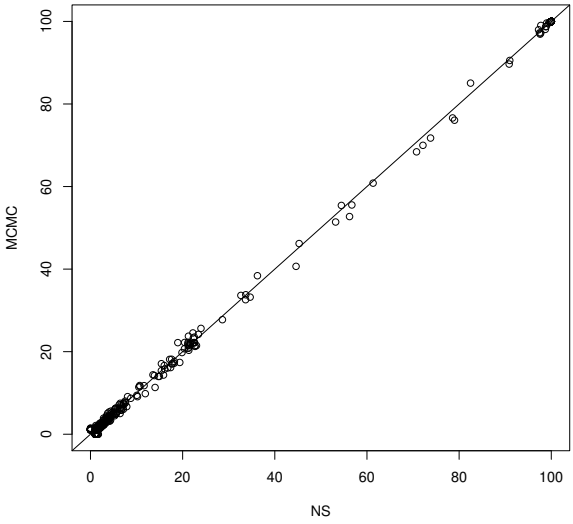

Figure 5: Posterior clade probabilities obtained via MCMC and NS for the Chloridoideae dataset. Both samples are in agreement, showing the proficiency of NS in sampling the tree space.

correctly infer the posterior probabilities (Höhna and Drummond, 2012). In this context, a Markov chain of length 10,000,000, with a burn-in period of 10% and thinning factor of 10,000, is compared to an NS posterior sample considering 100 active points. In particular, we are interested in how well NS performs for low probability clades.

The results are displayed in Figure 5. The MCMC analysis has effective sampling sizes of at least 400, indicating a reliable approximation of the posterior. The posterior clade probabilities obtained from NS are quite similar to those obtained from the MCMC analysis. These values are located around the straight line, reflecting the agreement between the tree posterior samples. Principally, the low probabilities are highly correlated, values which tend to differ in case of disagreement between the samples. NS posterior sample sizes were of around 3,500.

5 Conclusion

Nested sampling is a general Bayesian algorithm that provides the means to estimate marginal likelihoods and to carry out parameter inference. We have introduced it to phylogenetic inference under variable tree topology. Its performance has been compared to established methods available in many phylogenetic software packages.

NS performance has been assessed in different and challenging phylogenetic contexts. Firstly, we compared NS to SS using a dataset which contain 197 sequences and 15 partitions. In particular, we compared a strict clock model to a relaxed clock model, showing that NS with a single active point was enough to carry out model selection. In this analysis, the relaxed clock model was found to fit the data better. Then, we showed that the single NS run used to estimate the marginal likelihood for this model provides also the means to estimate the means and standard deviations of its posterior distribution. Secondly, we showed that NS yields consistent posterior clade probabilities in independent runs in the case that the tree topology space contain tree islands, unlike standard MCMC methods. In this context, we also showed that NS can differ from SS in marginal likelihood estimation, which can potentially have problems dealing in this scenario. Finally, we evaluated NS performance in a flat tree parameter space, showing that NS posterior clade probabilities are in agreement with those obtained from a standard MCMC method.

PS and SS have become popular because of their high accuracy estimating the marginal likelihood. Specifically, SS has become popular due to its implementation in widely-used phylogenetic software packages. However, they rely on several problem-specific tuning parameters which need to be specified by the user, namely number and distribution of the β values, number of samples from each transitional distribution and burn-in periods. GSS dispenses with the distribution of the β values, but it requires a number of posterior samples to calibrate the reference distribution. These calibrations are essential to get good estimates. On the other hand, NS only requires the number of active points and the number of MCMC steps used to generate the replacement points. Finding more efficient transition kernels remains an area open for further research. The length of the run is determined by the termination conditions described previously. NS is in practice more user-friendly than the methods presented.

One of the most important attributes of NS is that it not only yields a marginal likelihood estimate but also provides a measure of its uncertainty in a single run, unlike the other methods used in phylogenetics. This quantity is inversely proportional to the square root of the number of active points; the higher this number, the more accurate the estimate. Under similar computational conditions, NS can have a higher uncertainty than GSS, though this depends on the problem and whether the GSS parameters are well-tuned. However, its uncertainty can be calculated directly in a single run whereas GSS requires several replications to estimate its uncertainty, which can be carried out only after the reliability of the specifications for the tuning parameters have been tested. In practice, one could use NS to estimate the marginal likelihood intervals for the competitive models, and thus carry out model selection based on them. In the case that the intervals overlap, NS could be ran again for those specific models using more active points to increase the precision and narrow the intervals.

Nested sampling also provides the means to carry out parameter inference. This does not involve an extra cost since the points used to estimate the marginal likelihood are recycled. We have assessed the method to study clade probabilities and certain statistics of the posterior distribution. In particular, we showed that NS, even in the case of using a single active point, can be used to generate confidence intervals for the parameters. The method possesses interesting attributes. For instance, unlike conventional MCMC methods, NS does not require a burn-in period. In general, this period represents a high computational cost for MCMC methods. Furthermore, the method explores the parameter space in a quite different way, which allows it to deal well in complex scenarios, such as those parameter spaces composed of tree islands, a challenging scenario for standard MCMC methods.

NS possesses several positive characteristics which make it a very competitive algorithm in comparison to the established methods used currently in phylogenetics. It has been applied successfully to different fields and we believe this success can be replicated in phylogenetics as has been shown in this work.

The nested sampling algorithm is implemented in the NS package for BEAST 2, available from https://

github.com/BEAST2-Dev/nested-sampling under the LGPL licence. A fully parallel version (Feroz et al., 2009) is available that runs K nested sampling analyses with N particles, but selects starting points from the pool of all available $K \times N$ active points (conditioned on having an appropriate likelihood to start with). The NS package allows for phylogenetic inference under any of the models available to BEAST.

References

Aitken, S. and O. Akman. 2013. Nested sampling for parameter inference in systems biology: application to an exemplar circadian model. BMC Syst. Biol. 7:72.

Arima, S. and L. Tardella. 2014. Inflated density ratio (IDR) method for estimating marginal likelihoods in Bayesian phylogenetics. chap. 3, Pages 25–58 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.

Baele, G. and P. Lemey. 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. chap. 4, Pages 59–94 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Baele, G., P. Lemey, and M. A. Suchard. 2016. Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. Syst. Biol. 65:250–264.

Baele, G., P. Lemey, and S. Vansteelandt. 2013a. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. BMC Bioinformatics 14:85.

Baele, G., W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. 2013b. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 30:239–243.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLOS Comput. Biol. 10:1–6.

Brewer, B. J. and C. P. Donovan. 2015. Fast Bayesian inference for exoplanet discovery in radial velocity data. Mon. Not. R. Astron. Soc. 448:3206–3214.

Brewer, B. J. and D. Foreman-Mackey. 2016. DNest4: Diffusive nested sampling in C++ and Python. ArXiv preprint arXiv:1606.03757.

Brewer, B. J., L. B. Pártay, and G. Csányi. 2011. Diffusive nested sampling. Stat. Comput. 21:649–656.

Chopin, N. and C. Robert. 2010. Properties of nested sampling. Biometrika 97:741–755.

Drummond, A. J. and R. Bouckaert. 2015. Bayesian evolutionary analysis with BEAST. Cambridge University Press.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88.

Drummond, A. J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Fan, Y., R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28:523–532.

Feroz, E., M. P. Hobson, and M. Bridges. 2009. Multinest: an efficient and robust bayesian inference tool for cosmology amd particle physics. Mon. Not. R. Astron. Soc. 398:1601–1614.

Friel, N. and A. N. Pettitt. 2008. Marginal likelihood estimation via power posteriors. J. Roy. Stat. Soc. B 70:589–607.

Handley, W. J., M. P. Hobson, and A. N. Lasenby. 2015. POLYCHORD: next-generation nested sampling. Mon. Not. R. Astron. Soc. 453:4384–4398.

Hedges, S. B., K. D. Moberg, and L. R. Maxson. 1990. Tetrapod phylogeny inferred from 18s and 28s ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

Höhna, S. and A. J. Drummond. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Syst. Biol. 61:1–11.

Holder, M., P. O. Lewis, D. L. Swofford, and D. Bryant. 2014. Variable tree topology stepping-stone marginal likelihood estimation. chap. 5, Pages 95–111 *in* Bayesian phylogenetics : methods, computational algorithms, and applications (M. Chen, L. Kuo, and P. O. Lewis, eds.). Chapman and Hall/CRC, New York.

Horn, J. W., Z. Xi, R. Riina, J. A. Peirson, Y. Yang, B. L. Dorsey, P. E. Berry, C. C. Davis, and K. J. Wurdack. 2014. Evolutionary bursts in Euphorbia (Euphorbiaceae) are linked with photosynthetic pathway. Evolution 68:3485–3504.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21:1123–1133.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Ingram, A. L. and J. J. Doyle. 2004. Is Eragrostis (Poaceae) monophyletic? insights from nuclear and plastid sequence data. Syst. Bot. 29:545–552.

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. J. Amer. Statist. Assoc. 90:773–795.

Knuth, K. H. and J. Skilling. 2012. Foundations of inference. Axioms 1:38–73.

Lakner, C., P. Van Der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. Syst. Biol. 57:86–103.

Larget, B. 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. Syst. Biol. 62:501–511.

Larget, B. and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot, N. and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24:2669–2680.

MacKay, D. J. C. 2002. Information theory, inference & learning algorithms. Cambridge University Press, New York, NY, USA.

Maturana R., P., B. Brewer, and S. Klaere. 2017. Model selection and parameter inference in phylogenetics using nested sampling. ArXiv preprint arXiv:1703.05471v1.

Mukherjee, P., D. Parkinson, and A. R. Liddle. 2006. A nested sampling algorithm for cosmological model selection. Astrophys. J. Lett. 638:L51–L54.

Neal, R. 2003. Slice sampling. Ann. Stat. 31:705–767.

Newton, M. A. and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Roy. Statist. Soc. Ser. B 56:3–48.

Pullen, N. and R. J. Morris. 2014. Bayesian model comparison and parameter inference in systems biology using Nested Sampling. PLoS ONE 9:e88419.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Sivia, D. S. and J. Skilling. 2006. Data analysis: a Bayesian tutorial. Oxford University Press, USA.

Skilling, J. 2006. Nested sampling for general Bayesian computation. Bayesian Analysis 1:833–860.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.

Verdinelli, I. and L. Wasserman. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Amer. Statist. Assoc. 90:614–618.

Walter, C. 2017. Point process-based Monte Carlo estimation. Stat. Comput. 27:219–236.

Whidden, C. and F. A. Matsen. 2015. Quantifying MCMC exploration of phylogenetic tree space. Syst. Biol. 64:472–491.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

6 Appendix

6.1 Statistical example

Skilling (2006) presented a statistical example in order to illustrate NS performance in the case that the likelihood (as a function of the cumulative prior probabilities) is partly convex. This function was defined as the sum of two Gaussians centred at zero. The first has a standard deviation of 0.1 and the second a standard deviation of 0.01. In addition, the second Gaussian has a factor of 100 which makes it contribute more to the shape of the posterior distribution. Thus, the likelihood is a relatively flat density with a spike in its centre. This model, which includes a phase transition, poses problems to power posterior methods in their simple form, i.e., when the transitional densities define the path between the prior and the posterior. On the other hand, they can perform well in their generalised forms, but if and only if an adequate reference distribution is used.

Maturana R. et al. (2017) showed that a uniform works well as reference distribution in this example. A tentative alternative would have been a normal distribution. However, this poses some issues. The standard MCMC methods used to draw from the posterior are highly dependent on the starting values. If these are around zero, the MCMC chain will get trapped in the spike area, being almost impossible to escape from it. As a result, the sample will fail in representing adequately the posterior and will lead to a poor reference distribution. Actually, this will be centred at zero with a standard deviation of 0.01, as the narrow Gaussian in the likelihood function. Thus, the reference distribution will only encapsulate the area where the spike is located and leave without consideration the rest of the parameter space, which will have a direct effect on the marginal likelihood estimation. In their particular case, the excluded areas do not contain a significant volume and therefore the bias in the estimate would be small.

That analysis prompted the study of the same model, but where the excluded areas contain a bigger volume. For this, we consider a different version of the statistical model analysed by Skilling (2006). In our model the likelihood is composed of the sum of the same two Gaussians, but with the factor which scales the spike reduced to 1, that is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{d} \frac{1}{v\sqrt{2\pi}} \exp\left(-\frac{\theta_i^2}{2v^2} \right) + \prod_{i=1}^{d} \frac{1}{u\sqrt{2\pi}} \exp\left(-\frac{(\theta_i - \mu)^2}{2u^2} \right), \tag{7}$$

where $\boldsymbol{\theta}$ is a d-dimensional parameter vector, $d = 20$, $\mu = 0$, $v = 0.1$ and $u = 0.01$. We consider a uniform prior in the unit cube $[-0.5, 0.5]^d$ for $\boldsymbol{\theta}$. Thus, the marginal likelihood is 2.

We assess the marginal likelihood estimation by using NS and GSS. For NS, we use 99 active points which makes it require around 10,000 iterations/samples. We also include the NS estimate with only a single active point to evaluate its performance in the simplest condition. To calibrate the reference distribution required by GSS, we consider two approaches when generating the posterior samples: i) starting the MCMC chain at zero and ii) starting at random values drawn from a Uniform(-0.5, 0.5). The estimates in both cases will be referred as GSS_0 and GSS_R, respectively. For both approaches, we use 1,000 posterior samples to parameterise 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. The β values are chosen according to evenly spaced quantiles of a Beta(0.3, 1.0) distribution, and following a "melting" scheme, that is, starting from the posterior and moving down to the reference distribution. The estimations are replicated 1,000 times and are displayed in Figure 6 in where the horizontal dotted line stands for the the true value $\log(2) = 0.693$.

GSS_0 underestimates slightly the true value. Its estimates are around 0 with a standard deviation of 0.001. Its reference distributions are centred 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 now, unlike in the original example, have a larger amount of probability mass. This is clearly illustrated in one dimension posterior sample in Figure 7. 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 original model. Even in the case of increasing significantly the number of steps, the GSS_0 estimates do not change (results not shown), with which one would be tempted to trust in the reliability of the estimate. On the other hand, GSS_R overestimates the true value due to the failing of its reference distribution on pondering the different areas of the parameter space. Its estimates have a mean of 17.56 and a standard deviation of 1.18. Even in the case of increasing significantly the number of steps, the estimates are far away from the true value (results not shown).

On the contrary, NS estimates are around the true value. Even in its simple case, with a single active point. The estimates have a standard deviation of 0.57 and 5.85 for the case of 1 a 99 active points, respectively. NS also provides accurate posterior samples which cannot be obtained by conventional MCMC methods (see Figure 7), such as Metropolis-Hastings or slice sampling.

Due to the likelihood shape, the approximation of the reference distribution for this model is determined by the starting point in the Markov chain (see Figure 7). If all the components of the starting point are around 0, the distribution will be approximately a N(0, 0.01) for each component of $\boldsymbol{\theta}$, but if there is at least one of them a bit far from its centre, let say outside the approximated interval $[-0.023, 0.023]$, the reference distribution

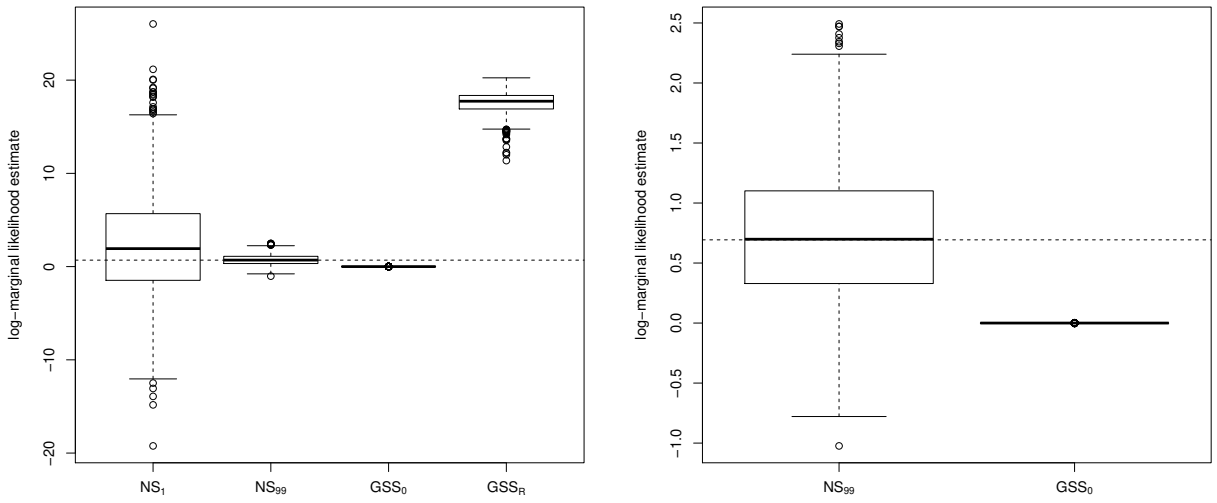

Figure 6: NS and GSS marginal likelihood estimates in the statistical example. The subscripts in the NS methods stand for the number of active points, whereas in the GSS method depict the starting value specifications to generate the posterior samples (see text for more details). The horizontal dotted line stands for the true log marginal likelihood value.

for each component will be approximately a N(0, 0.1). Actually, they are the plateau and the narrow normal distributions, respectively, which compose the likelihood. The consequences of failing to approximate the posterior distribution directly impacts the marginal likelihood estimates. Therefore, in situations where the posterior is not an easy distribution to sample from, (for instance, DS1), GSS should not be used.

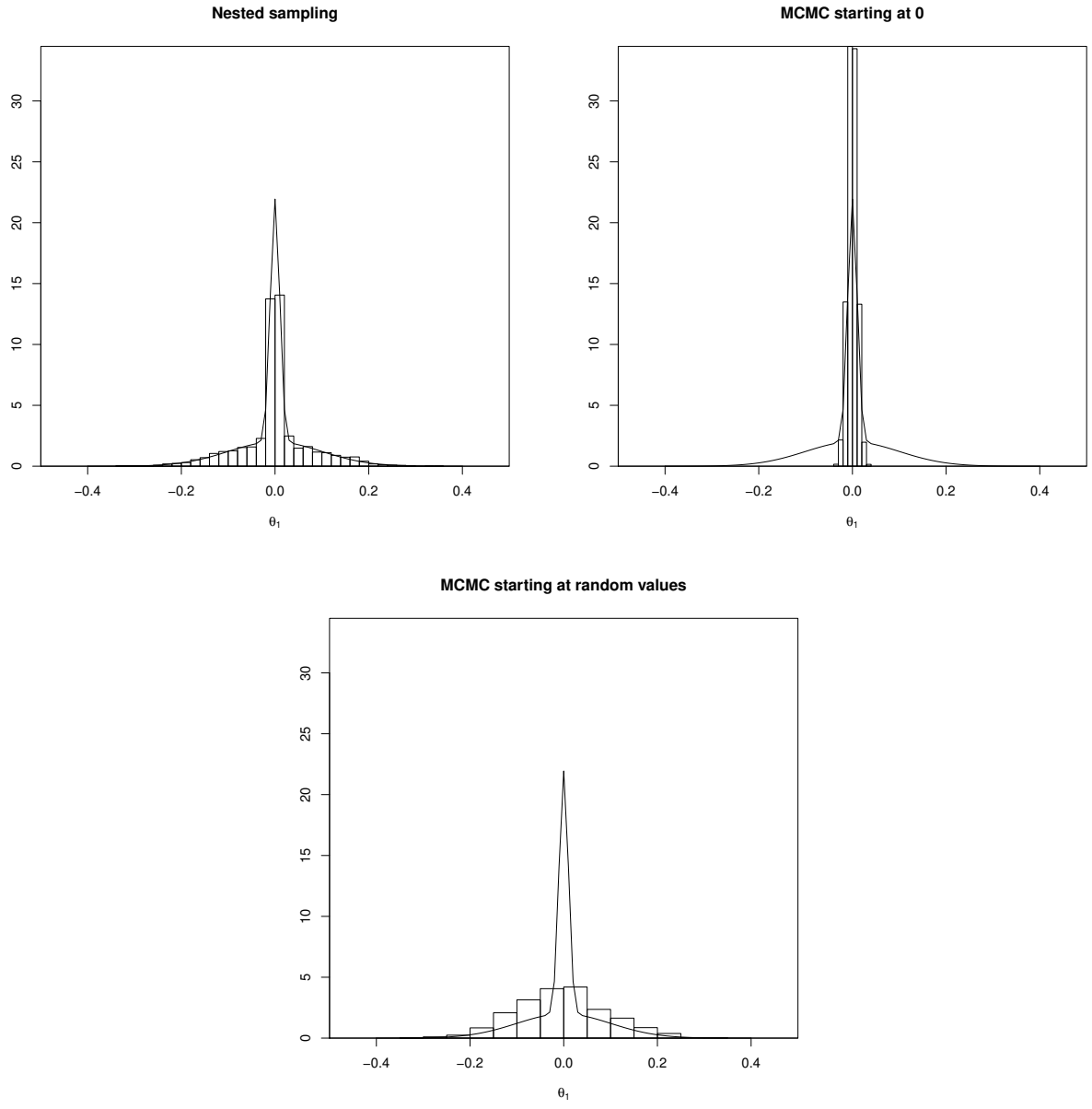

Figure 7: The histograms stand for 10,000 posterior samples for the first component of $\boldsymbol{\theta}$ by using nested sampling, and MCMC samples using two different starting values. This behaviour is similar in all the components of $\boldsymbol{\theta}$. The continuous lines depict the true marginal density.