An alternative marginal likelihood estimator for phylogenetic models

Serena Arima [*1] **and Luca Tardella**[2]

[1]Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale, Sapienza Università di Roma
via del Castro Laurenziano 9, 00161 Roma,
E-mail: serena.arima@uniroma1.it

[2]Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza Università di Roma
p.le Aldo Moro, 5, 00161 Roma,
E-mail: luca.tardella@uniroma1.it

June 22, 2010

Abstract

Bayesian phylogenetic methods are generating noticeable enthusiasm in the field of molecular systematics. Many phylogenetic models are often at stake and different approaches are used to compare them within a Bayesian framework. The Bayes factor, defined as the ratio of the marginal likelihoods of two competing models, plays a key role in Bayesian model selection. We focus on an alternative estimator of the marginal likelihood whose computation is still a challenging problem. Several computational solutions have been

[*]Dipartimento di studi geoeconomici, linguistici, statistici e storici per l'analisi regionale, Sapienza Università di Roma, via del Castro Laurenziano 9, 00161 Roma,
E-mail: serena.arima@uniroma1.it

proposed none of which can be considered outperforming the others simultaneously in terms of simplicity of implementation, computational burden and precision of the estimates. Practitioners and researchers, often led by available software, have privileged so far the simplicity of the harmonic mean estimator (HM) and the arithmetic mean estimator (AM). However it is known that the resulting estimates of the Bayesian evidence in favor of one model are biased and often inaccurate up to having an infinite variance so that the reliability of the corresponding conclusions is doubtful. Our new implementation of the generalized harmonic mean (GHM) idea recycles MCMC simulations from the posterior, shares the computational simplicity of the original HM estimator, but, unlike it, overcomes the infinite variance issue. The alternative estimator is applied to simulated phylogenetic data and produces fully satisfactory results outperforming those simple estimators currently provided by most of the publicly available software.

keywords: Bayes factor, harmonic mean, importance sampling, marginal likelihood, phylogenetic models.

1 Introduction

The theory of evolution states that all organisms are related through a history of common ancestor and that life on Earth diversified in a tree-like pattern connecting all living species. Phylogenetics aims at inferring the tree that better represents the evolutionary relationships among species studying differences and similarities in their genomic sequences. Alternative tree estimation methods such as parsimony methods (Felsenstein (2004), chapter 7) and distance methods (Fitch and Margoliash, 1967; Cavalli-Sforza and Edwards, 1967) have been proposed. In this paper, we will focus on stochastic models for substitution rates and we address the model choice issue within a fully Bayesian framework proposing an alternative model evidence estimation procedure. The paper is organized as follows: in Section 2 we briefly review basic phylogenetic concepts and the Bayesian inference for substitution models. In Section 3 we focus on the model selection issue for substitution models and discuss some available computational tools for Bayesian model evidence. One of the most popular tool for computing model evidence in phylogenetics is the

Harmonic Mean (HM) estimator proposed by Newton and Raftery (1994) as an easy-to-apply instance of a more general class of estimators called Generalized Harmonic Mean (GHM). An alternative version of GHM is considered in Section 4. It has been introduced in Petris and Tardella (2007) under the name of Inflated Density Ratio (IDR) and its implementation for substitution models is described in Section 5. Numerical examples and comparative performance are given in Section 6. We conclude with a brief discussion in Section 7.

2 Substitution models: a brief overview

Phylogenetic data consists of homologous DNA strands or protein sequences of related species. Observed data consists of a nucleotide matrix X with n rows representing species and k columns representing sites. Comparing DNA sequences of two related species, we define *substitution* the replacement in the same situs of one nucleotide in one species by another one in the other species. The stochastic models describing this replacement process are called *substitution models*. A phylogeny or a phylogenetic tree is a representation of the genealogical relationships among species, also called *taxa* or *taxonomies*. Tips (leaves or external nodes) represent the present-day species, while internal nodes usually represent extinct ancestors for which genomic sequences are no longer available. The ancestor of all sequences is the root of the tree. The branching pattern of a tree is called *topology*, and is denoted with τ, while the lengths ν_τ of the branches of the tree τ represent the time periods elapsed until a new substitution occurs.

DNA substitution models are probabilistic models which aim at modeling changes between nucleotides in homologous DNA strands. Changes at each site occur at random times. Nucleotides at different sites are usually assumed to evolve independently each other. For a fixed site, nucleotide replacements over time are modeled by a 4-state Markov process, in which each state represents a nucleotide. The Markov process indexed with time t is completely specified by a substitution rate matrix $Q(t) = r_{ij}(t)$: each element $r_{ij}(t)$, $i \neq j$, represents the instantaneous rate of substitution from nucleotide i to nucleotide j. The diagonal elements of the rate matrix are defined as $r_{ii}(t) = \sum_{j \neq i} r_{ij}(t)$ so that $\sum_{j=1}^{4} r_{ij}(t) = 0, \forall i$. The transition probability matrix $P(t) = \{p_{ij}(t)\}$, defines the probability of changing from state i to state j.

The substitution process is assumed homogeneous over time so that $Q(t) = Q$ and $P(t) = P$. It is also commonly assumed that the substitution process at each site is stationary with equilibrium distribution $\Pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ and time-reversible, that is

$$\pi_i r_{ij} = \pi_j r_{ji} \tag{2.1}$$

where π_i is the proportion of time the Markov chain spends in state i and $\pi_i r_{ij}$ is the amount of flow from state i to j. Equation (2.1) is known as *detailed-balance condition* and means that flow between any two states in the opposite direction is the same. Following the notation in Hudelot *et al.* (2003), we define $r_{ij}(t) = r_{ij} = \rho_{ij}\pi_i$, $\forall i \neq j$, where ρ_{ij} is the transition rate from nucleotide i to nucleotide j. This reparameterization is particularly useful for the specification of substitution models, since it makes clear the distinction between the nucleotide frequencies $\pi_A, \pi_G, \pi_C, \pi_T$ and substitution rates ρ_{ij}, allowing to spell out different assumptions on evolutionary patterns. The most general time-reversible nucleotide substitution model is the so-called **GTR** defined by the following rate matrix

$$Q = \begin{pmatrix} - & \rho_{AC}\pi_C & \rho_{AG}\pi_G & \rho_{AT}\pi_T \\ \rho_{AC}\pi_A & - & \rho_{CG}\pi_G & \rho_{CT}\pi_T \\ \rho_{AG}\pi_A & \rho_{CG}\pi_C & - & \rho_{GT}\pi_T \\ \rho_{AT}\pi_A & \rho_{CT}\pi_C & \rho_{GT}\pi_G & - \end{pmatrix} \tag{2.2}$$

and more thoroughly illustrated in Lanave *et al.* (1984). Several substitution models can be obtained simplifying the Q matrix reflecting specific biological assumptions: the simplest one is the **JC69** model, originally proposed in Jukes and Cantor (1969), which assumes that all nucleotides are interchangeable and have the same rate of change, that is $\rho_{ij} = \rho \ \ \forall i, j$ and $\pi_A = \pi_C = \pi_G = \pi_T$.
In this paper, for illustrative purposes we will consider only instances of GTR and JC69 models. One can look at Felsenstein (2004) and Yang (2006) for a wider range of alternative substitution models.

2.1 Bayesian inference for substitution models

The parameter space of a phylogenetic model can be represented as

$$\Omega = \{\tau, \nu_\tau, \theta\}$$

where $\tau \in \mathcal{T}$ is the tree topology, ν_τ the set of branch lengths of topology τ, and $\theta = (\rho, \pi)$ the parameters of the rate matrix. We denote N_T the cardinality of $\mathcal{T}$. Notice that N_T is a huge number even for few species. For instance with $n = 10$ species there are about $N_T \approx 2 \cdot 10^6$ different trees.

Observed data consists of a nucleotide matrix X once specified the substitution model M, the likelihood $p(X|\tau, \nu_\tau, \theta, M)$ can be computed using the pruning algorithm, a practical and efficient recursive algorithm proposed in Felsenstein (1981). One can then make inference on the unknown model parameters looking for the values which maximize the likelihood. Alternatively one can adopt a Bayesian approach where the parameter space is endowed with a joint prior distribution $\pi(\tau, \nu_\tau, \theta)$ on the unknowkn parameters and the likelihood is used to update the prior uncertainty about (τ, ν_τ, θ) following the Bayes' rule:

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

where

$$m(X|M) = \sum_{\tau \in \mathcal{T}} \int_{\nu_\tau} \int_{\theta} p(X|\tau, \nu_\tau, \theta, M)\pi(\tau, \nu_\tau, \theta) d\nu_\tau d\theta$$

The resulting distribution $p(\tau, \nu_\tau, \theta|X, M)$ is the posterior distribution which coherently combines prior believes and data information. Prior believes are usually conveyed as $\pi(\tau, \nu_\tau, \theta) = \pi(\tau)\pi(\nu_\tau)\pi(\theta)$. The denominator of the Bayes' rule $m(X|M)$ is the marginal likelihood of model M and it plays a key role in discriminating among alternative models.

More precisely, suppose, we aim at comparing two competing substitution models M_0 and M_1. The Bayes Factor is defined as the ratio of the marginal likelihoods as follows

$$BF_{10} = \frac{m(X|M_1)}{m(X|M_0)}$$

where, for $i = 0, 1$

$$c^{(i)} = m(X|M_i) = \sum_{\tau \in \mathcal{T}} \int_{\nu_\tau^{(i)}} \int_{\theta^{(i)}} p(X|\theta, \tau, M_i)\pi(\theta^{(i)}) d\theta^{(i)} \pi(\nu_\tau^{(i)}|\tau) d\nu_\tau^{(i)} \pi(\tau) \qquad (2.3)$$

Numerical guidelines for interpreting the evidence scale are given in Kass and Raftery (1995). Values of $BF_{10} > 1$ ($log(BF_{10}) > 0$) can be considered as evidence in favor of M_1 but only a value of $BF_{10} > 100$ ($log(BF_{10} > 4.6)$ can be really considered as decisive.

Most of the times the posterior distributions $p(\tau, \nu_\tau, \theta|X, M_i)$ and marginal likelihoods are not analytically computable but can be approached through appropriate approximations. Indeed, over the last ten years, powerful numerical methods based on Markov Chain Monte Carlo (MCMC) have been developed, allowing one to carry out Bayesian inference under a large category of probabilistic models, even when dimension of the parameter space is very large. Indeed, several ad-hoc MCMC algorithms have been tailored for phylogenetic models (Larget and Simon, 1999; Li *et al.*, 2000) and are currently implemented in publicly available software such as in MrBayes (Ronquist and Huelsenbeck, 2003) and PHASE (Gowri-Shankar and Jow, 2006).

3 Model selection for substitution models

Given the variety of possible stochastic substitution mechanisms, an important issue of any model-based phylogenetic analysis is to select the model which is most supported by the data. Several model selection procedures have been proposed depending also on the inferential approach. A classical approach to model selection for choosing between alternative nested models is to perform the hierarchical likelihood ratio test (LRT) (Posada and Crandall, 2001). A number of popular programs allow users to compare pairs of models using this test such as PAUP (Swafford, 2003), PAML (Yang, 2007) and the R package APE (R Development Core Team, 2008). However, Posada and Buckely (2004) have shown some drawbacks of performing systematic LRT for model selection in phylogenetics. This is because the model that is finally selected can depend on the order in which the pairwise comparisons are performed (Pol, 2004). Moreover, it is well-known that LRT tends to favor parameter rich models.

The Akaike Information Criterion (AIC) is another model-selection criterion commonly used also in phylogenetics (Posada and Buckely, 2004): one of the advantages of the AIC is that it allows to compare nested as well as non nested models and it can be easily implemented. However, also the AIC tends to favor parameter-rich models. To overcome this selection bias one can use the Bayesian Information Criterion (BIC) (Schwartz, 1978) which better penalizes parameter-rich models.

Sometimes these criteria applied to the same data can end up selecting very different substitution models, as shown in Abdo *et al.* (2005). Indeed they compare ratios of likelihood values penalized for an increase in the dimension of one of the models, without directly accounting for uncertainty in the estimates of model parameters. The latter aspect is addressed within a fully Bayesian framework through the use of the Bayes Factor. Bayes Factor directly incorporates this uncertainty and its meaning is more intuitive than other methods since it can be directly used to assess the comparative evidence provided by the data in terms of the most probable model under equal prior model probabilities.

Bayes Factor for comparing phylogenetic models was first introduced in Sinsheimer *et al.* (1996) and Suchard *et al.* (2001). Since then its popularity in phylogenetics has grown so that some publicly available software provide in their standard output approximations of marginal likelihoods for model evidence and Bayes Factor evaluation. Indeed the complexity of phylogenetic models and the computational burden in the light of high-dimensional parameter space make the problem of finding alternative and more efficient computational strategy for computing Bayes Factor still open and in continuous development (Lartillot *et al.*, 2007; Ronquist and Deans, 2010).

3.1 Available computational tools for Bayesian model evidence

The computation of the marginal likelihood $m(X|M)$ of a phylogenetic model M is not straightforward. It involves integrating the likelihood over k-dimensional subspaces for the branch length parameters ν_τ and the substitution rate matrix $\theta = (\rho, \pi)$ and eventually summing over all possible topologies.

Most of the marginal likelihood estimation methods proposed in the literature have been applied extensively also in molecular phylogenetics (Minin *et al.*, 2003; Lartillot *et al.*, 2007; Suchard *et al.*, 2001). Among these methods, many of them are valid only under very specific conditions. For instance, the Dickey-Savage ratio (Verdinelli and Wasserman, 1995) applied in phylogenetics in Suchard *et al.* (2001), assumes nested models. Laplace approximation (Kass and Raftery, 1995) and BIC

(Schwartz, 1978), applied in phylogenetics firstly in Minin *et al.* (2003), require large sample approximations around the maximum likelihood, which can be sometimes difficult to compute or approximate for very complex models. A recent appealing variation of the Laplace approximation has been proposed in Rodrigue *et al.* (2007): however, its applicability and performance are endangered when the posterior distribution deviates from normality and the maximization of the likelihood can be neither straightforward nor accurate.

The reversible jump approach (Green, 1995; Bartolucci *et al.*, 2006) is another MCMC option applied to phylogenetic model selection in Huelsenbeck *et al.* (2004). Unfortunately the implementation of this algorithm is not straightforward for the end user and it often requires appropriate delicate tuning of the Metropolis Hastings proposal. Moreover its implementation suffers extra difficulties when comparing models based on an entirely different parametric rationale (Lartillot *et al.*, 2007).

As recently pointed out in Ronquist and Deans (2010) among the most widely used methods for estimating the marginal likelihood of phylogenetic models are the thermodynamic integration, also known as path sampling, and the harmonic mean approach. The thermodynamic integration reviewed in Gelman and Meng (1998) and first applied in a phylogenetic context in Lartillot and Philippe (2006) produces reliable estimates of Bayes Factors of phylogenetic models in a large varieties of models. Although this method has the advantage of general applicability, it can incur high computational costs and may require specific adjustments. For certain model comparisons, a full thermodynamic integration may take weeks on a modern desktop computer, even under a fixed tree topology for small single protein data sets (Rodrigue *et al.*, 2007; Ronquist and Deans, 2010). On the other hand, the HM estimator can be easily computed and it does not demand further computational efforts other than those already made to draw inference on model parameters, since it only needs simulations from the posterior distributions. However, it is well known that the HM estimator is unstable since it can end up with an infinite variance. As highlighted by Ronquist and Deans (2010), thermodynamic integration and reversible jump are, until now, the most accurate tools for computing the marginal likelihood. However, until these methods become more user-friendly and more widely available, simple tools for exploring in a quicker way the more interesting models are useful. For this reason in the next sections we focus on an alternative generalized harmonic

mean estimator, the IDR estimator, which shares the computational simplicity of HM estimator but, unlike it, better copes with the infinite variance issue. Its simple implementation makes the IDR estimator a useful and more reliable method for easily comparing competing substitution models. It can be used also as a confirmatory tool even in those models for which more complex estimation methods, such as the path sampling, can be applied.

3.2 Harmonic Mean estimators

We introduce the basic ideas and formulas for the class of estimators known as Generalized Harmonic Mean (GHM). Since the marginal likelihood is nothing but the normalizing constant of the unnormalized posterior density, we illustrate the GHM estimator as a general solution for estimating the normalizing constant of a non-negative, integrable density g defined as

$$c = \int_{\Omega} g(\theta) d\theta \tag{3.1}$$

where $\theta \in \Omega \subset \Re^k$ and $g(\theta)$ is the unnormalized version of the probability distribution $\tilde{g}(\theta)$. The GHM estimator of c is based on the following identity

$$c = \frac{1}{E_{\tilde{g}}\left[\left(\frac{g(\theta)}{f(\theta)}\right)^{-1}\right]} \tag{3.2}$$

where f is a convenient instrumental Lebesgue integrable function which is only required to have a support which is contained in that of g and to satisfy

$$\int_{\Omega} f(\theta) d\theta = 1. \tag{3.3}$$

The GHM estimator, denoted as $\hat{c}_{GHM}$ is the empirical counterpart of (3.2), namely

$$\hat{c}_{GHM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T} \frac{f(\theta_t)}{g(\theta_t)}}. \tag{3.4}$$

where $\theta_1, \theta_2, ..., \theta_T$ are sampled from $\tilde{g}$. In Bayesian inference the very first instance of such GHM estimator was introduced in Gelfand and Dey (1994) to estimate the marginal likelihood considered as the normalizing constant of the unnormalized

posterior density $g(\theta) = \pi(\theta)L(\theta)$. Hence, taking $f(\theta) = \pi(\theta)$ one obtains as special case of (3.4) the Harmonic Mean estimator

$$\hat{c}_{HM} = \frac{1}{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L(\theta_t)}} \tag{3.5}$$

which can be easily computed by recycling simulations $\theta_1, ..., \theta_T$ from the target posterior distribution $\tilde{g}(\theta)$ available from MC or MCMC sampling scheme. This probably explains the original enthusiasm in favor of $\hat{c}_{HM}$ which indeed was considered a potential convenient competitor of the standard Monte Carlo Importance Sampling estimate given by the (Prior) Arithmetic Mean (AM) estimator

$$\hat{c}_{AM} = \frac{1}{T}\sum_{t=1}^{T} L(\theta_t) \tag{3.6}$$

where $\theta_1, ..., \theta_T$ are sampled from the prior π.

The implementation of and ,more generally, (3.4) requires a relatively light computational burden hence reducing computing time with respect to thermodynamic integration. The simplicity of the computation has then favored the widespread use of the Harmonic Mean estimator with respect to more complex methods. In fact, the Harmonic Mean estimator is implemented in several Bayesian phylogenetic software as shown in Table 1 and recent biological papers (Yamanoue *et al.*, 2008; Wang *et al.*, 2009; Normana *et al.*, 2009) report the HM as a routinely used model selection tool.

Table 1 about here

However, both $\hat{c}_{AM}$ and $\hat{c}_{HM}$ can end up with a very large variance and unstable behavior. This fact cannot be considered as an unusual exception but it often occurs and the reason for that can be argued on a theoretical ground.

For $\hat{c}_{AM}$ the erratic behavior is simply explained by the fact that the likelihood usually gives support to a region with low prior weight hence sampling from the prior yields low chance to hit high likelihood region and large chance to hit much lower likelihood region ending up in a large variance of the estimate $\hat{c}_{AM}$. Indeed, starting from the original paper Newton and Raftery (1994), (see in particular R. Neal's discussion) it has been shown that even in very simple and standard gaussian models also the HM estimator can end up having an infinite variance hence yielding

unreliable approximations. This fact raises sometimes the question whether they are reliable tools and certainly has encouraged researchers to look for alternative solutions. Several generalizations and improved alternatives have been proposed and recently reviewed in Raftery *et al.* (2007).

In the following sections we will consider a new marginal likelihood estimator, the Inflated Density Ratio (IDR) estimator, proposed in Petris and Tardella (2007), which is a particular instance of the Generalized Harmonic Mean (GHM) approach. This new estimator basically shares the original simplicity and the computation feasibility of the HM estimator but, unlike it, it can guarantee important theoretical properties, such as a bounded variance.

4 IDR: Inflated Density Ratio estimator

The inflated density ratio estimator is a different formulation of the GHM estimator, based on a particular choice of the instrumental density $f(\theta)$ as originally proposed in Petris and Tardella (2007). The instrumental $f(\theta)$ is obtained through a perturbation of the original target function g. The perturbed density, denoted with g_{P_k}, is defined so that its total mass has some known functional relation to the total mass c of the target density g as in (3.1). In particular, g_{P_k} is obtained as a parametric inflation of g so that

$$\int_{\Omega} g_{P_k}(\theta) = c + k \tag{4.1}$$

where k is a known inflation constant which can be arbitrarily fixed. The perturbation device comes from an original idea in Petris and Tardella (2003) and is deteiled in Petris and Tardella (2007) for unidimensional and multidimensional densities. In the unidimensional case the perturbed density is

$$g_{P_k}(\theta) = \begin{cases} g(\theta + r_k) & \text{if } \theta < -r_k \\ g(0) & \text{if } -r_k \leq \theta \leq r_k \\ g(\theta - r_k) & \text{if } \theta > r_k \end{cases} \tag{4.2}$$

with $2r_k = \frac{k}{g(0)}$ corresponding to the length of the interval centered around the origin where the density is kept constant. In Figure 1 one can visualize how the perturbation acts. The perturbed density allows one to define an instrumental density $f_k(\theta) = \frac{g_{P_k}(\theta) - g(\theta)}{k}$ which satisfies the requirement (3.3) needed to define the GHM estimator as in (3.4). The Inflated Density Ratio estimator $\hat{c}_{IDR}$ for c is then obtained as follows

$$\hat{c}_{IDR} = \frac{k}{\frac{1}{T}\sum_{t=1}^{T} \frac{g_{P_k}(\theta_t)}{g(\theta_t)} - 1} \tag{4.3}$$

where $\theta_1, ..., \theta_T$ is a sample from the normalized target density $\tilde{g}$. The use of the perturbed density as importance function leads to some advantages with respect to the other instances of c_{GHM} proposed in the literature. In fact $\hat{c}_{IDR}$ defined as in (4) yields a finite-variance estimator under mild sufficient conditions and a wide range of g densities Petris and Tardella (2007) (Lemma 1, 2 and 3). Notice that in order for the perturbed density g_{P_k} to be defined it is required that the original density g has full support in $\Re^k$. Moreover, the use of a parametric perturbation makes the method more flexible and efficient with a moderate extra computational effort.

Like all methods based on importance sampling strategies, the properties of the estimator $\hat{c}_{IDR}$ strongly depend on the ratio $\frac{g_{P_k}(\theta)}{g(\theta)}$. To evaluate its performance one can use an asymptotic approximation (via standard delta-method) of the Relative Mean Square Error of the estimator

$$RMSE_{\hat{c}_{IDR}} = \sqrt{E_{\tilde{g}}\left[\left(\frac{\hat{c}_{IDR} - c}{c}\right)^2\right]} \approx \frac{c}{k}\sqrt{Var\left[\frac{g_{P_k}(\theta)}{g(\theta)}\right]} = RMSE_{\hat{c}_{IDR,Delta}} \tag{4.4}$$

which can be ultimately estimated as follows:

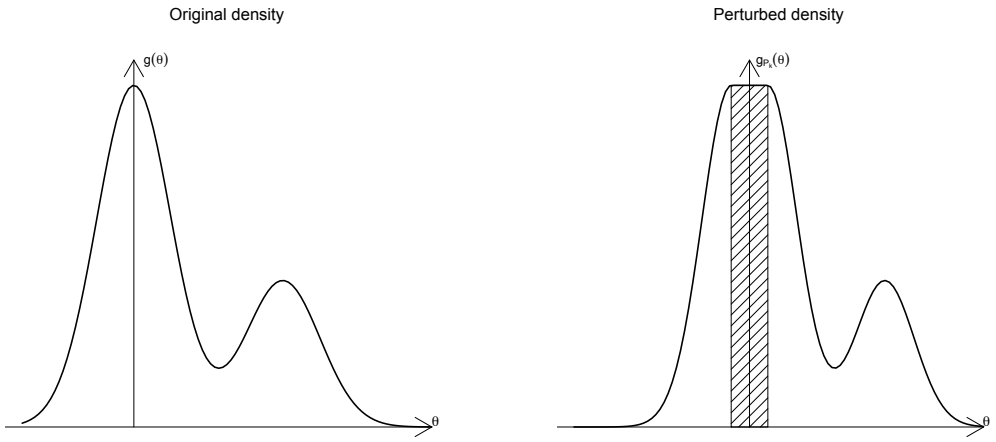

Figure 1: *Left Panel: original density g with total mass c. Right Panel: perturbed density* g_{P_k} *defined as in (4.2). The total mass of the perturbed density is then* $c+k$. *The shaded area correspond to the inflated mass* k *with* $k = 2 \cdot r_k \cdot g(0)$ *as in (4.2).*

$$\widehat{RMSE}_{\hat{c}_{IDR,Delta}} = \frac{\hat{c}_{IDR}}{k} \sqrt{\widehat{Var}_{\tilde{g}} \left[\frac{g_{P_k}(\theta)}{g(\theta)} \right]} \tag{4.5}$$

where $\widehat{Var}_{\tilde{g}}$ is the observed sample variance of the ratio $g_{P_k}(\theta)/g(\theta)$.
The expression in Equation (4.5) clarifies the key role of the choice of k with respect to the error of the estimator: for $k \to 0$, the variance of the ratio $\frac{g_{P_k}(\theta)}{g(\theta)}$ tends to 0, since g_{P_k} is very close to g, but $\frac{c}{k}$ tends to infinity: in other words, if $Var_{\tilde{g}} \left[\frac{g_{P_k}(\theta)}{g(\theta)} \right]$ would favor as little values of k as possible, $\frac{1}{k}$ acts in the opposite direction. In order to address the choice of k, Petris and Tardella (2007) suggested to choose the perturbation k which minimizes the estimation error (4.5). In practice, one can calculate the values of the estimator for a grid of different perturbation values, $\hat{c}_{IDR}(k)$, $k = 1, ..., K$ and choose the optimal k^{opt} as the k for which $\widehat{RMSE}_{\hat{c}_{IDR}(k)}$ is minimum. This procedure for the calibration of k requires iterative evaluation of $\hat{c}_{IDR}(k)$ hence is relatively heavier than the HM estimator, but it does not require extra simulations which in the phylogenetic context is often the main time-consuming part. Hence, the computational cost is alleviated by the fact that one uses the same sample from $\tilde{g}$ and the only quantity to be evaluated K times is the inflated density g_{P_k}.

Once obtained the ratio of the perturbed and the original density, the computation of $\hat{c}_{IDR}(k)$ is straightforward.

For practical purposes, the computation of the inflated density when the support of g is the whole $\Re^k$ can be easily implemented in R using a function available at `http://sites.google.com/site/idrharmonicmean/home`. In Petris and Tardella (2007) and Arima (2009) it has been shown that in order to improve the precision of $\hat{c}_{IDR}$ it is recommended to standardize the simulated data $\theta_1, ..., \theta_T$ with respect to a (local) mode and the sample variance-covariance matrix so that the corresponding density has a local mode at the origin and approximately standard variance-covariance matrix. This is automatically implemented in the publicly available `R` code.

In order to assess its effectiveness, the IDR method has been applied to simulated data from differently shaped distributions for which the normalizing constant is known. As shown in Petris and Tardella (2007), the estimator produces fully convincing results with simulated data from several known distributions, even for a 100-dimensional multivariate Gaussian distribution. In terms of estimator precision, these results are comparable with those in Lartillot and Philippe (2006) obtained with the thermodynamic integration. In Arima (2009) simple antithetic variate tricks allow the IDR estimator to perform well even for those distributions with severe variations from the symmetric Gaussian case such as asymmetric and even some multimodal distributions. Table 2 shows the estimates obtained by applying the IDR method in several controlled scenarios: the method correctly reproduces the true value of the normalizing constant for different shape and dimension of the target function. Some real data implementation with standard generalized linear models have been also reported in Petris and Tardella (2007). In the next Section, we extend the IDR method in order to use $\hat{c}_{IDR}$ in more complex settings such as phylogenetic models.

Table 2 about here

5 Implementing IDR for substitution models with fixed topology

We extend the Inflated Density Ratio approach in order to compute the marginal likelihood of phylogenetic models. In this section we show how to compute the marginal likelihood when it involves integration of substitution model parameters θ and the branch lengths ν_τ which are both defined in continuous subspaces. Indeed the approach can be used as a building block to integrate also over the tree topology τ.

For a fixed topology τ and a sequence alignment X, the parameters of a phylogenetic model M_τ are denoted as $\omega = (\theta, \nu_\tau) \subset \Omega_\tau$. The joint posterior distribution on ω is given by

$$p(\theta, \nu_\tau | X, M_\tau) = \frac{p(X|\theta, \nu_\tau, M_\tau)\pi(\theta)\pi(\nu_\tau)}{m(X|M_\tau)} \tag{5.1}$$

where

$$m(X|M_\tau) = \int_\theta \int_{\nu_\tau} p(X|\theta, \nu_\tau, M_\tau)\pi(\theta)\pi(\nu_\tau) d\theta d\nu_\tau \tag{5.2}$$

is the marginal likelihood we aim at estimating.

When the topology τ is fixed, the parameter space Ω_τ is continuous. Hence, in order to apply the IDR method we only need the following two ingredients:

- a sample $(\theta^{(1)}, \nu_\tau^{(1)}), ..., (\theta^{(T)}, \nu_\tau^{(T)})$ from the posterior distribution, $p(\theta, \nu_\tau | X, M_\tau)$
- the likelihood and the prior distribution evaluated at each posterior sampled value $(\theta^{(k)}, \nu_\tau^{(k)})$, that is $p(X|\theta^{(k)}, \nu_\tau^{(k)}, M_\tau)$ and $\pi(\theta^{(k)}, \nu_\tau^{(k)}) = \pi(\theta^{(k)})\pi(\nu_\tau^{(k)})$

The first ingredient is just the usual output of the Monte Carlo Markov Chain simulations derived from model M and data X. The computation of the likelihood and the joint prior is usually already coded within available software. The first one is accomplished through the pruning algorithm while computing the prior is straightforward. Indeed a necessary condition for the inflation idea to be implemented as prescribed in Petris and Tardella (2007) is that the posterior density must have full support on the whole real k-dimensional space. In our phylogenetic models this is not always the case hence we explain simple and fully automatic remedies to overcome this kind of obstacle.

We start with branch length parameters which are constrained to lie in the positive

half-line. In that case the remedy is straightforward. One can reparameterize with a simple logarithmic transformation

$$\nu_{\tau}^{'} = log(\nu_{\tau}) \tag{5.3}$$

so that the support corresponding to the reparameterized density becomes unconstrained. Obviously the $log(\nu_{\tau})$ reparameterization calls for the appropriate Jacobian when evaluating the corresponding transformed density. For model parameters with linear constraints like the substitution $\theta = \{\rho, \pi\}$, a little less obvious transformation is needed. In this case $\theta = \{\rho, \pi\}$ are subject to the following set of constraints:

$$\begin{aligned} \sum_{i \in \{A,T,C,G\}} \pi_i &= 1 \\ \sum_{j \in \{A,T,C,G\}} \rho_{ij}\pi_j &= 0 \qquad \forall \;\; i \in \{A,T,C,G\} \end{aligned}$$

Similarly to the first simplex constraint the last set of constraints together with the reversibility can be rephrased (Gowri-Shankar, 2006) in terms of another simplex constraint concerning only the extra-diagonal entries of the substitution rate matrix (2.2) namely

$$\rho_{AC} + \rho_{AG} + \rho_{AT} + \rho_{CG} + \rho_{CT} + \rho_{GT} = 1.$$

In order to bypass the constrained nature of the parameter space we have relied on the so-called *additive logistic transformation* (Tiao and Cuttman, 1965; Aitchinson, 1986) which is a one-to-one transformation from $\mathbb{R}^{D-1}$ to the $(D-1)$-dimensional simplex

$$S^D = \{(x_1, ..., x_D) : x_1 > 0, ..., x_D > 0; x_1 + ...x_D = 1\}.$$

Hence we can use its inverse, called *additive log-ratio transformation*, which is defined as follows

$$y_i = log\left(\frac{x_i}{x_D}\right) \qquad i = 1, ..., D-1$$

for any $x = (x_1, ..., x_D) \in S^D$. Here the x_i's are the ρ's and $D = 6$. Applying these transformations to nucleotide frequencies π_i and to exchangeability parameters ρ's, the transformed parameters assume values in the entire real support and the IDR estimator can be applied. Again the reparameterization calls for the appropriate

change-of-measure Jacobian when evaluating the corresponding transformed density (see Aitchinson (1986) for details).

6 Numerical examples and comparative performance

In this section the successful implementation of the IDR estimator is illustrated with data simulated from some typical phylogenetic models.

Here IDR method has been applied using the MCMC output of the simulations from the posterior distribution obtained using the MrBayes software; the likelihood has been computed using the R package PML while the reparameterization on $\Re^k$ and IDR perturbation $g_{P_k}(\theta)$ have called for specifically developed R functions. The whole R code is available upon request from the first author.

Two of the simplest and most favorite model evidence output in the publicly available software are used as benchmarks: the Harmonic Mean estimator and the Arithmetic Mean estimator. Indeed, while the former is guaranteed to be a consistent estimate of the marginal likelihood, though possibly with infinite variance, the latter one is consistent only when formula (3.6) is applied when $\theta_1, ..., \theta_T$ are sampled from the prior. Since it is known such a prior AM turns out to be very unstable and unreliable it has often been replaced by a posterior AM where $\theta_1, ..., \theta_T$ are sampled from the posterior rather than from the prior. In that case one must be aware that the resulting quantity can be interpreted only as a surrogate evidence in favor of one model and it should by no means be confused with the rigorous concept of marginal likelihood and related to Bayes Factor. We now show the performance of IDR in two phylogenetic examples. Simulated data is used to have a better control of what one should expect from marginal likelihood and the corresponding comparative evidence of alternative models.

6.1 Hadamard 1: marginal likelihood computation

We use as a first benchmark the synthetic data set Hadamard 1 already employed in Felsenstein (2004). It consists of a sequence 1000 of amino acid alignments of six

species, A, B, C, D, E and F simulated from a $GTR + \Gamma$ model. The true tree is shown in the left Panel of Figure 2.

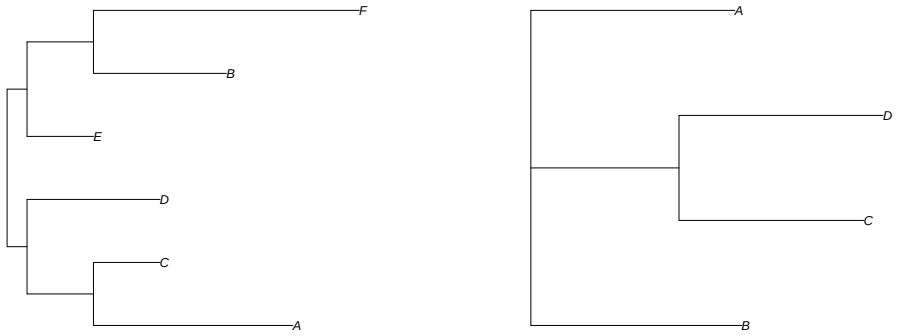

Figure 2: *True phylogenetic trees of the two synthetic data used as benchmarks:* Hadamard 1 *(left Panel)* Hadamard 2 *(right Panel).*

$GTR + \Gamma$ model is implemented in MrBayes which uses the Metropolis Coupling algorithm (Altekar *et al*, 2004) and provides as output the simulated Markov Chain and some evaluations of Bayesian model evidence in terms of posterior AM and marginal likelihood via HM.

The simulated Markov chain can be used as a sample from the posterior distribution in order to make inference on the model parameters. For this model the whole parameter space consists of 18 parameters. In order to reduce the autocorrelation and improve the convergence of the 1100000 sampled values, 100000 have been discarded with thinning rate equal to 10. We have recycled this MCMC output to estimate the marginal likelihood of the $GTR + \Gamma$ model for the known true topology through IDR method. In Table 3 we list the corresponding values of the IDR estimator on the log scale ($\log \hat{c}_{IDR}$), the Relative Mean Square Error estimate ($\widehat{RMSE}_{\hat{c}_{IDR}}$) as in Equation (4.5) and the confidence interval $\widehat{CI}$ for different perturbation masses k. In order to take into account the autocorrelation of the posterior simulated values, a correction has been applied to $\widehat{RMSE}_{\hat{c}_{IDR}}$ replacing n in (4.5)

with the effective sample size given by

$$n_{ESS} = n \times \frac{1}{1 + 2\sum_{i=1}^{I} \hat{\rho}_i} \tag{6.1}$$

Since the optimal corrected error $\widehat{RMSE}^*_{\hat{c}_{IDR}}$ corresponds to a perturbation value $k^{opt} = 10^{-7}$, the IDR estimator (on a logarithmic scale) for the $GTR + \Gamma$ model is $\log \hat{c}_{IDR} = -7258.200$.

Table 3 about here

We compare the results of the IDR method with those obtained with the Harmonic Mean (HM) and the posterior Arithmetic Mean (AM).

For $\hat{c}_{HM}$ and $\hat{c}_{AM}$, relative errors have been estimated respectively as

$$\widehat{RMSE}_{HM} = \widehat{c}_{HM} \sqrt{\widehat{Var}\left(\frac{1}{g(\theta)}\right)} \tag{6.2}$$

$$\widehat{RMSE}_{AM} = \frac{1}{\widehat{c}_{AM}} \sqrt{\frac{\widehat{Var}(g(\theta))}{n}} \tag{6.3}$$

Similarly to $\widehat{RMSE}^*_{\hat{c}_{IDR}}$ we have denoted with $\widehat{RMSE}^*_{\hat{c}_{HM}}$ and $\widehat{RMSE}^*_{\hat{c}_{AM}}$ the estimates of the relative errors adjusted with the effective sample size n_{ESS}.

The three methods produce somewhat different quantities although sometimes compatible once accounted for the estimation error. For each method, the Monte Carlo relative error of the estimate has been computed re-estimating the model 10 times ($\widehat{RMSE}_{\hat{c},MC}$) and recording the corresponding different values of $\hat{c}$.
Although it is known that under critical conditions such MC error is not sufficient to guarantee its precision it still remains a necessary premise for an accurate estimate. We have also looked at another precision measurement $\widehat{RMSE}_{\hat{c},Boot}$ based on bootstrap replications. $\widehat{RMSE}_{\hat{c},Boot}$ is defined as

$$\widehat{RMSE}_{\hat{c},Boot} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\frac{\hat{c}_b}{\hat{c}} - 1\right)^2}$$

where $\hat{c}_b$ denotes the bootstrap replicate of the generic marginal likelihood estimate with one of the three alternative formulae $\hat{c}_{AM}$, $\hat{c}_{HM}$ or $\hat{c}_{IDR}$ and $B = 1000$. In order to account for the autocorrelation, also $\widehat{RMSE}_{\hat{c},Boot}$ has been corrected using the effective sample size in (6.1).

Table 4 shows the obtained results.

Table 4 about here

Indeed the estimates of the model evidence as well as the corresponding estimates of their relative errors are very different. In fact, the smallest relative error is obtained with the arithmetic mean of the likelihood values simulated from the posterior. We have already pointed out the fact that this posterior AM does not really aim at estimating the marginal likelihood, but we have nonetheless considered it in our examples to verify how distant the corresponding values are and how different the conclusions and their strengths can be when comparing alternative models via the posterior AM estimator. For the Harmonic Mean method the estimated MC relative error is formidably high and unstable resulting in a serious warning on the reliability of the HM estimator in this context. On the other hand, the Inflated Density Ratio approach seems to be a good compromise in terms of order of magnitude of the error of the estimate $\hat{c}_{IDR}$ and robustness of its relative error estimation ranging from 0.15 with independent Monte Carlo re-estimation to 0.30 of the $\widehat{RMSE}^*_{\hat{c}}$.

6.2 Hadamard 2: Bayes Factor computation

We have also considered the Hadamard 2 data in (Felsenstein, 2004) as a second benchmark synthetic data set. It consists of 200 amino acids and four species, A, B, C, D. This dataset have been simulated from the Jukes-Cantor model (JC69) and the true tree is shown in the right Panel of Figure 2.

For the true topology, we compute the marginal likelihood for the JC69 model and for the GTR+Γ model: parameters lie in a 5-dimensional space for the JC69 model and in 14-dimensional space for the GTR+Γ model. The simulated values from the Metropolis-Coupled algorithm have been rearranged to evaluate the model evidence of both models using IDR, HM and the AM approach. As for the Hadamard 1 data, Monte Carlo RMSE have been also computed by repeating the estimates 10 times. Table 5 shows the results obtained respectively for the **GTR**+Γ and the **JC69** models.

Table 5 about here

All methods produce somewhat different results in terms of model evidence; as for the previous example, the smallest relative Monte Carlo error is associated once

again with the Arithmetic Mean method. Also in this case, estimates of the relative errors of the Harmonic Mean method are always larger than those produced by the Inflated Density Ratio method. The corresponding Bayes Factors (on logarithmic scale) for JC69 and GTR+Γ are shown in Table 6. Considering the reference values for the Bayes Factor defined in Kass and Raftery (1995), all methods consistently give support to the Jukes-Cantor model, which is known to be the true model. However, we highlight that the strongest evidence in favor of the correct model corresponds to the Bayes Factor as estimated by the Inflated Density Ratio .

Table 6 about here

6.3 Hadamard 2: tree selection

We now show how it is possible to extend the IDR approach for dealing with selecting competing trees when the topology is not fixed in advance. For a fixed substitution model, competing trees can be compared by considering the evidence of the data for a fixed tree topology. The evidence in support of each tree topology $\tau_i \in N_T$ can be evaluated in terms of its posterior probability $p(\tau_i|X)$ derived from the Bayes theorem as

$$p(\tau_i|X) = \frac{p(X|\tau_i)\pi(\tau_i)}{\sum_{\tau \in N_T} p(X|\tau)\pi(\tau)}$$

where the experimental evidence in favor of the model M_{τ_i} with fixed tree topology τ_i is contained in the marginal likelihood

$$m(X|M_{\tau_i}) = p(X|\tau_i) = \int_{\Omega_{\tau_i}} p(X|\omega_i, M_{\tau_i})\pi(\omega_i|\tau_i)d\omega_i$$

where the continuous parameters $\omega_i \in \Omega_{\tau_i}$ of the evolutionary process corresponding to M_{τ_i} are integrated out as nuisance parameters. Indeed when prior beliefs on trees are set equal $\pi(\tau_i) = \pi(\tau_j)$ comparative evidence discriminating τ_i against τ_j, is summarized in the Bayes Factor

$$BF_{ij} = \frac{m(X|M_{\tau_i})}{m(X|M_{\tau_j})}. \tag{6.4}$$

We have considered the problem of selecting competing trees of a substitution model for Hadamard 2 data. In the previous Subsection, we have verified the feasibility of the Inflated Density Ratio approach in comparing JC69 with the GTR+Γ

model. Under a fixed topology JC69 was favored. Indeed we know that Hadamard 2 data was simulated from JC69 model with true topology labeled as $\tau = 1$. Now we aim at comparing $N_T = 3$ alternative topologies within the JC69 model and we compute and compare the corresponding marginal likelihoods. Results are shown in Table 7.

Table 7 about here

Also in this case, the IDR exhibits the most convincing performance in terms of evidence in support of the true tree as well as precision and robustness of estimates.

7 Discussion

In this paper, we investigate the possibility of using simple effective recipes for evaluating model evidence of competing models of complex phylogenetic models. In a Bayesian framework, several methods have been proposed in order to approximate the marginal likelihood of a single model and then eventually estimate the Bayes Factor of two competing models.

Probably, the most widely used methods to date are the thermodynamic integration and the harmonic mean approach. The thermodynamic integration has been proved to produce more reliable estimates of Bayes Factors of competing phylogenetic models in a large varieties of contexts. Although this method has the advantage of general applicability, it is computationally demanding and may require fine tunings and adjustments. Indeed, the simplicity of implementation combined with a relatively light computational burden are two appealing features which explain why the HM is still currently one of the most favorite option for routine implementation (see von Reumont *et al.* (2009)). However, the simplicity of HM is often not matched with its accuracy and recent literature is highlighting unreliability of HM estimators in phylogenetic models (Lartillot and Philippe, 2006) as well as in more general biological applications (Calderhead and Girolami, 2009). In this paper, we have provided evidence of improved effectiveness of a simple alternative marginal likelihood estimator, the Inflated Density Ratio estimator (IDR), belonging to the class of generalized harmonic mean estimators. It shares the original simplicity and

computation feasibility of the HM estimator but, unlike it, it enjoys important theoretical properties, such as the finiteness of the variance. Moreover it allows one to recycle posterior simulations which is particularly appealing in those contexts – such as phylogenetic models – where the computational burden of the simulation is heavier than the evaluation of the likelihood, posterior densities and the like. Like all importance sampling techniques based on a single stream of simulations the computational burden can be shared in a parallel computing environment reducing the computing time. Also the grid search for optimizing the estimated RMSE can be speeded up with a parallel evaluation for each inflated density.

We have verified the effectiveness of the IDR estimator in some of the most common phylogenetic substitution models under different model complexity including mixed parameter space and evaluated the comparative performance with respect to HM and posterior AM estimators. In all circumstances the IDR estimator outperformed the HM estimator in terms of precision and robustness of the estimates and it is then an interesting candidate to be included in standard software as a simple and more reliable model evidence output. Its simple implementation makes the IDR estimator a useful tool to be possibly used as a simple confirmation/benchmark even in those models where fine-tuned approximation tools such as thermodynamic integration are available and, when appropriately fine-tuned, may yield more precise estimates.

References

Abdo, Z., Minin, V.N., Joyce, P., and Sullivan, J. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Molecular Biology and Evolution*, 22(3):691–703, 2005.

Aitichinson, J. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London, 1986.

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P. and Ronquist, F. Parallel metropolis-coupled Markov Chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20:407 – 415, 2004.

Arima, S. Bayesian tools for complex statistical models in genetics, (PhD Thesis). *DSPSA, Sapienza Universitá di Roma*, 2009.

Bartolucci, F., Scaccia, L. and Mira, A. Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, 93(1):41-52, 2006.

Calderhead, B. and Girolami, M. Estimating bayes Factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

Cavalli-Sforza, L.L. and Edwards, A.W.F. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21:550:570, 1967.

Chen, M.H. Importance-weighted marginal bayesian posterior density estimation. *Journal of the American Statistical Association*, **89**, 818–824, 1994.

Do, K.D., Muller, P. and Vannucci, M. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, New York, 2006.

Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

Felsenstein, J. *Inferring Phylogenies*. Sinauer, Sunderland, MA, 2004.

Fitch, W.M. and Margoliash, E. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

Gelfand, A.E. and Dey, D.K. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B: Methodological*, 56: 501–514, 1994.

Gelman, A. and Meng, X.L. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

Gowri-Shankar, V. and Jow, H. PHASE: a software package for Phylogenetics And Sequence Evolution. 2.0. *University of Manchester*, 2006.

Gowri-Shankar, V. RNA phylogenetic inference with heterogeneous substitution models. *PhD Thesis: University of Manchester*, 2006.

Green, P. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. and Higgs, P.G. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Molecolar Phylogenetic Evolution*, 28(2):241–252, 2003.

Huelsenbeck, J.P. and Ronquist, F. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, 17(8):754–755, August 2001.

Huelsenbeck, J.P., Larget, B. and Alfaro, E. Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo. *Molecular Biology and Evolution*, 21(6):1123–1133, 2004.

Larget, B., Miller, R.E., Huelsenbeck, J.P. and Ronquist, F. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

Jukes, T.H. and Cantor, C.R. Evolution of protein molecules. *In Mammalian protein metabolism*, 1969.

Kass, R.E. and Raftery, A. Bayes Factor. *Journal of American Statistical Association*, 90:773–795, 1995.

Lanave, C., Preparata, G. and Serio, G. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

Larget, B., and Simon, D.L. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16:750–759, 1999.

Lartillot, N., Brinkmann, H. and Philippe, H. Suppression of long branch attraction artefacts in the animal phylogeny using site-heterogeneous model. *BMC Evol. Biol.*, 7, 2007.

Lartillot, N. and Philippe, H. Computing Bayes Factors using Thermodynamic Integration. *Systematic Biology*, 55(2), 2006.

Li, S., Pearl, D. and Doss, H. Phylogenetic tree reconstruction using Markov Chain Monte Carlo. *Journal of American Statistical Association*, 95:493–508, 2000.

Minin, V.N, Abdo, Z., Joyce, P. and Sullivan, J. Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology*, 52(5):674–683, 2003.

Newton, M.A. and Raftery, A. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of Royal Statistical Society (Series B)*, 56:3–48 (with discussion), 1994.

Normana, J.A., Ericsonc, G.P., Jonssond, K. A., Fjeldsad, J. and Christidis, L. A multi-gene phylogeny reveals novel relationships for aberrant genera of AustraloPapuan core Corvoidea and polyphyly of the Pachycephalidae and Psophodidae (Aves: Passeriformes). *Molecular Phylogenetics and Evolution*, 52:488–497, 2009.

Petris, G. and Tardella, L. A geometric approach to trandimensional Markov Chain

Monte Carlo. *Canadian Journal of Statistics*, 2003.

Petris, G. and Tardella, L. New perspectives for Estimating Normalizing Constants via Posterior Simulation (Technical Report). *DSPSA, Sapienza Universitá di Roma*, 2007.

Pol, D. Empirical problems of the hierarchical likelihood ratio test for model selection. *Systematic Biology*, 53(6):949–962, 2004.

Posada, D. and Crandall, K.A. Selecting the best-fit model of nucleotide substitution. *Systematic Biology*, 50(4):580–601, 2001.

Posada, D. and Buckely, T.R. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approach over likelihood ratio test. *Systematic Biology*, 53(5):793–808, 2004.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

Raftery, M.A., Newton, A., Satagopan, J.M. and Krivitsky, P.N. Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. *Bayesian Statistics 8 - Proceedings of the Eighth Valencia International Meeting*, pages 1–45, 2007.

Robert, R.C. and Wraith, D. Computational methods for Bayesian model choice. *arXiv:0907.5123*, 2009.

Rodrigue, N., Philippe, H. and Lartillot, N. Exploring Fast Computational Strategies for Probabilistic Phylogenetis Analysis. *Systematic Biology*, 55:137–157, 2007.

Ronquist, F. and Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572:1574, 2003.

Ronquist, F. and Deans, A.R. Bayesian Phylogenetics and Its Influence on Insects Systematics. *Annu. Rev. Entomol*, 55: 189-206, 2010.

Schwartz, G. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

Sinsheimer, J.S., Lake, J.A.,ì and Little, J.A. Bayesian Hypothesis Testing of Four-Taxon Topologies Using Molecular Sequence Data. *Biometrics*, 52(1):193–210, 1996.

Weiss, R.E., Suchard, M.A. and Sinsheimer, J.S. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*,

18:1001–1013, 2001.

Swafford, D.L. *PAUP: phylogenetic analysis using parsimony (and other methods).* Sinauer Associates, Sunderland, MA, 2003.

Tiao, G.G. and Cuttman, I. The Inverted Dirichlet Distribution with Applications. *Journal of the American Statistical Association*, 311:793–805, 1965.

Verdinelli, I. and Wasserman, L. Computing Bayes Factor using a generalization of the Savage-Dickey density ratio. *Journal of American Statistical Association*, 90:614–618, 1995.

von Reumont, B., Meusemann, K., Szucsich, N.U., Dell'Ampio, E., Gowri-Shankar, V., Bartel, D., Simon, S., Letsch, H.O., Stocsits, R.R., Luan, Y., Wagele, J.W., Pass, G., Hadrys, H. and Misof, B. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evolutionary Biology*, 9:1–19, 2009.

Wang, Z., Johnston, P.R., Yang, Z.L. and Townsend, J.P. Evolution of Reproductive Morphology in Leaf Endophytes. *PlosOne*, 4, 2009.

Yamanoue, Y., Miya, M., Matsuura, K., Katoh, M., Sakai, H. and Nishida, M. A new perspective on phylogeny and evolution of tetraodontiform fishes (Pisces: Acanthopterygii) based on whole mitochondrial genome sequences: Basal ecological diversification? *SBMC Evolutionary Biology*, 8:212–226, 2008.

Yang, Z. *Computational Molecular Evolution.* OSEE, Oxford, UK, 2006.

Yang, Z. PAML 4: phylogentic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

Software	Marginal likelihood estimation method
BayesTraits	Harmonic Mean
BEAST	Harmonic Mean or bootstrapped Harmonic Mean
MrBayes	Harmonic Mean
PHASE	Harmonic Mean and Reversible Jump
PhyloBayes	Thermodynamic Integration under normal approximation

Table 1: *Availability of marginal likelihood estimation methods in Bayesian phylogenetics software: in spite of its inaccuracy, the harmonic mean estimator is still one of the most diffuse marginal likelihood estimation tool.*

Distribution	$\log c$	$\log \hat{c}_{IDR}$	$\widehat{RMSE}_{\hat{c}_{IDR}}$	k^{opt}
$N(\mu, \sigma)$	0	0	10^{-4}	10^{-4}
$N_{100}(\mu, \sigma)$	0	0	0.01	10^{23}
$SN(\mu, \sigma, \tau)$	0	10^{-4}	0.004	10^{-4}
$SN_5(\mu, \sigma, \tau)$	3,467	3.444	0.014	20
$SN_{30}(\mu, \sigma, \tau)$	2,302	2.403	0.047	10^4
Mix N_2	2.079	2.078	0.003	0.01
Mix N_3	2.772	2.766	0.002	1
Mix N_{10}	0	0.056	0.012	2

Table 2: *Performance of IDR approach with $n = 10^5$ i.i.d. draws simulated from distributions with known normalizing constants: univariate and multivariate gaussian distributions (up to 100 dimension), univariate and multivariate skew normal distributions (SN) (30 dimension) and multivariate mixtures of two Normal components (Mix N) (in dimensions 2,3 and 10). The value* $\log c$ *represents the logarithm of the true value of the normalizing constant;* $\log \hat{c}_{IDR}$ *is the value of the estimated normalizing constant on a logarithmic scale.* k^{opt} *is the optimal inflation coefficient which minimizes the Relative Mean Square Error* $\widehat{RMSE}_{\hat{c}_{IDR}}$ *computed as in (4.5). For the SN case a small sensitivity study (not reported here) showed that the performance of the method is robust to different* (μ, σ, τ) *parameter choices.*

	k	$\log \hat{c}_{IDR}$	$\widehat{RMSE}_{\hat{c}_{IDR}}$	$\widehat{RMSE}^*_{\hat{c}_{IDR}}$	$\widehat{CI}$
	10^{-10}	-7264.438	0.1710	0.4515	$[-7264.852, -7263.718]$
	10^{-9}	-7262.150	0.1689	0.4514	$[-7262.560, -7261.443]$
	10^{-8}	-7259.939	0.1602	0.3664	$[-7260.332, -7259.284]$
$(*)$	10^{-7}	-7258.200	0.1178	0.3008	$[-7258.503, -7257.764]$
	10^{-6}	-7257.554	0.1407	0.3694	$[-7257.907, -7257.006]$

Table 3: *Inflated Density Ratio method applied to* Hadamard 1 *data with a GTR+*Γ *model with a 18-dimensional parameter space. IDR estimates on the log scale for a small regular grid of perturbation values. The relative mean square errors* $\widehat{RMSE}_{\hat{c}_{IDR}}$ *are computed as in (4.5) without accounting for autocorrelation.* $\widehat{RMSE}^*_{\hat{c}_{IDR}}$ *are the relative mean square errors corrected for the autocorrelation.* $\widehat{CI}$ *are confidence intervals on a log scale. Since the smallest error in the grid corresponds to a perturbation value* $k^{opt} = 10^{-7}$*, the IDR estimator for the* $GTR + \Gamma$ *model is taken to be* $\log \hat{c}_{IDR} = -7258.200$.

Method	$log(\widehat{c})$	$\widehat{RMSE}_{\hat{c}}$	$\widehat{RMSE}_{\hat{c},MC}$	$\widehat{RMSE}_{\hat{c},Boot}$	$\widehat{RMSE}^*_{\hat{c}}$
IDR	-7258.200	0.1178	0.1538	0.1698	0.3008
HM	-8365.509	173.2080	$> 10^{10}$	$> 10^{100}$	296.3475
AM	-7204.245	0.0197	0.0119	0.0202	0.065

Table 4: *Hadamard 1 data: marginal likelihood estimates obtained with the Inflated Density Ratio method* $\widehat{c_{IDR}}$*, with the Harmonic Mean approach* $\widehat{c_{HM}}$ *and with the Arithmetic Mean approach* $\widehat{c_{AM}}$*. The estimates are based on* $n = 10^6$ *from* 10^7 *simulated values. Three different RMSE estimates are provided:* $\widehat{RMSE}_{\hat{c}}$ *has been computed as in (4.5) for IDR, (6.3) for AM and (6.2) for HM;* $\widehat{RMSE}_{\hat{c},MC}$ *comes from 10 Monte Carlo independent replicates of the estimation;* $\widehat{RMSE}_{\hat{c},Boot}$ *is a boostrap approximation of RMSE (B=1000);* $\widehat{RMSE}^*_{\hat{c}}$ *is the estimated RMSE corrected for the autocorrelation as in (6.1).*

GTR+Γ					
Method	$log(\widehat{c})$	$\widehat{RMSE}_{\hat{c}}$	$\widehat{RMSE}_{\hat{c},MC}$	$\widehat{RMSE}_{\hat{c},Boot}$	$\widehat{RMSE}^*{}_{\hat{c}}$
IDR	-611.8571	0.1153	0.1087	0.1175	0.3608
HM	-594.648	31.5330	0.1329	0.3488	141.3285
AM	-588.286	0.0184	0.0863	0.0187	0.0826
JC69					
Method	$log(\widehat{c})$	$\widehat{RMSE}_{\widehat{c}}$	$\widehat{RMSE}_{\widehat{c},MC}$	$\widehat{RMSE}_{\widehat{c},Boot}$	$\widehat{RMSE}^*{}_{\widehat{c}}$
IDR	-595.5919	0.0068	0.0161	0.0081	0.0179
HM	-589.0289	34.6759	0.1415	0.6787	59.5918
AM	-589.4194	0.0057	0.0146	0.0056	0.0182

Table 5: *Hadamard 2 data: marginal likelihood estimates of the GTR+Γ model ($\omega = \Re^{14}$) obtained with the IDR method $\widehat{c_{IDR}}$, with the HM approach $\widehat{c_{HM}}$ and with the AM approach $\widehat{c_{AM}}$. The estimates are based on $n = 10^6$ from 10^7 simulated values. Three different RMSE estimates are provided: $\widehat{RMSE}_{\hat{c}}$ has been computed as in (4.5) for IDR, (6.3) for AM and (6.2) for HM; $\widehat{RMSE}_{\hat{c},MC}$ comes from 10 Monte Carlo independent replicates of the estimation; $\widehat{RMSE}_{\hat{c},Boot}$ is a bootstrap approximation of RMSE (B=1000); $\widehat{RMSE}^*{}_{\hat{c}}$ is the RMSE corrected for the auto-correlation as in (6.1).*

Method	$log(\widehat{BF}_{GTR+\Gamma-JC69})$	$\widehat{CI}^{MC}_{log(\widehat{BF})}$
IDR	16.2652	$[16.1726, 16.3578]$
HM	5.6241	$[5.0206, 5.1546]$
AM	1.1334	$[1.0617, 1.2051]$

Table 6: *Hadamard 2 data: Bayes Factors computed with IDR, HM and AM approach. The estimates are based on* $n = 10^6$ *from* 10^7 *simulated values.* $\widehat{CI}^{MC}_{log(\widehat{BF})}$ *is the confidence interval obtained as* $log(\hat{BF}) \pm 2 \cdot SD_{MC}(log(\hat{BF}))$. $log(\hat{BF})$ *is obtained by averaging the estimated Bayes Factors (on logarithmic scale) of* 10×10 *possible pairings of 10 MC replicates.* $SD_{MC}(log(\hat{BF}))$ *is computed as the standard error of the estimated Bayes Factors (on logarithmic scale) in* 10×10 *possible combinations of 10 MC replicates.*

τ	$Size$	$\widehat{CI}^{MC}_{log(\widehat{BF}_{IDR})}$	$\widehat{CI}^{MC}_{log(\widehat{BF}_{HM})})$	$\widehat{CI}^{MC}_{log(\widehat{BF}_{AM})}$
$log(BF_{12})$	10^4	$[3.511, 3.599]$	$[2.37, 4.066]$	$[2.929, 2.989]$
$log(BF_{13})$	10^4	$[3.817, 3.901]$	$[2.503, 3.163]$	$[2.053, 3.131]$

Table 7: *Hadamard 2 data: the Bayes Factor is computed in order to compare competing topologies The Bayes Factor is approximated with the IDR method (BF_{IDR}), the HM (BF_{HM}) and the AM (BF_{AM}) approach.* $\widehat{CI}^{MC}_{log(\widehat{BF})}$ *is the confidence interval obtained as* $log(\hat{BF}) \pm 2 \cdot SD_{MC}(log(\hat{BF}))$. $log(\hat{BF})$ *is obtained by averaging the estimated Bayes Factors (on logarithmic scale) of* 10×10 *possible combinations of 10 MC replicates.* $SD_{MC}(log(\hat{BF}))$ *is computed as the standard error of the estimated Bayes Factors (on logarithmic scale) in* 10×10 *possible combinations of 10 MC replicates.*