The stepping-stone sampling algorithm for calculating the evidence of gravitational wave models

Patricio Maturana Russel[1], Renate Meyer[1], John Veitch[2] and Nelson Christensen[3]
[1] *Department of Statistics, University of Auckland, Auckland 1142, New Zealand*
[2] *Physics and Astronomy, University of Glasgow, United Kingdom*
[3] *ARTEMIS, Laboratoire de la Côte d'Azur, Nice, France*

Bayesian statistical inference has become increasingly important for the analysis of observations from the Advanced LIGO and Advanced Virgo gravitational-wave detectors. To this end, iterative simulation techniques, in particular nested sampling and parallel tempering, have been implemented in the software library LALInference to sample from the posterior distribution of waveform parameters of compact binary coalescence events. Nested sampling was mainly developed to calculate the evidence of a model but can produce posterior samples as a by-product. Thermodynamic integration is employed to calculate the evidence using samples generated by parallel tempering but has been found to be computationally demanding. Here we propose the stepping-stone sampling algorithm, originally proposed by Xie et al. (2011) in phylogenetics and a special case of path sampling, as an alternative to thermodynamic integration. The stepping-stone sampling algorithm is also based on samples from the power posteriors of parallel tempering but has superior performance as less temperature steps and thus computational resources are needed to achieve the same accuracy. We demonstrate its performance and computational costs in comparison to thermodynamic integration and nested sampling in a simulation study and a case study of computing the marginal likelihood of a binary black hole model applied to simulated LIGO data John: please correct, not sure which data and model you have been using. To deal with the inadequate methods currently employed to estimate the standard errors of evidence estimates based on power posterior techniques, we propose a novel block bootstrap approach and show its potential in our simulation study and LIGO application.

PACS numbers: 04.30.-w, 02.50.-r, 05.45.Tp, 97.60.Bw
Keywords: thermodynamic integration, stepping-stone sampling, nested sampling, model selection, gravitational waves

I. INTRODUCTION

general review of evidence estimates that are used in the gravitational wave literature

- Discuss: importance of "z" - how the uncertainty is estimated in power posterior methods

The paper is structured as follows: Section II . . . .

II. COMPUTATION OF MARGINAL LIKELIHOOD

The *evidence* or *marginal likelihood* of a model M is a multi-dimensional integral defined as

$$z = \int_{\Theta} L(\boldsymbol{X}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\theta} \in \Theta$ denotes the parameter vector, $\boldsymbol{X}$ the dataset, $L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ the likelihood function, and $\pi(\boldsymbol{\theta}|M)$ the prior density, assumed to be proper, i.e. $\int_{\Theta} \pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} = 1$.

In general, this integral (1) has no analytical solution and must be estimated using numerical methods. Importance sampling techniques, in particular the arithmetic mean (AM) and harmonic mean (HM) methods, provide the simplest way of estimating it [1]. Let $\boldsymbol{\theta}_i, i = 1, \ldots, n$ be samples from the prior, the AM estimator is an average of corresponding n likelihood values:

$$\widehat{z}_{AM} = \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{X}|\boldsymbol{\theta}_i, M). \tag{2}$$

In general, high-likelihood areas are very small and constitute a small fraction of the prior. Therefore, unless n is very large, the sample will not adequately represent these areas and yield a poor estimate. The HM estimator is based on samples $\boldsymbol{\theta}_i, i = 1, \ldots, n$ drawn from the posterior:

$$\widehat{z}_{HM} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{L(\boldsymbol{X}|\boldsymbol{\theta}_i, M)}\right)^{-1} \tag{3}$$

i.e. the inverse of an average of inverse likelihood values or *harmonic mean* of likelihood values.

Therefore, the AM and HM estimators are not recommended because they produce unreliable estimates of the evidence even though they are easily calculated. In this context, more complex approaches have been proposed, such as power posterior methods [2–5]. These methods rely on a set of transitional distributions which connect the prior and the posterior, reminiscent of simulated annealing. The geometric path is the most popular scheme used to connect these distributions and defines the *power* *posterior density* as

$$p_\beta(\boldsymbol{\theta}|\boldsymbol{X}, M) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta}, \tag{4}$$

for $0 \leq \beta \leq 1$, the inverse temperature, where z_β is the normalizing constant, which is defined as $\int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M) \mathrm{d}\boldsymbol{\theta}$. Note that the power posterior density turns into the prior and posterior for $\beta = 0$ and $\beta = 1$, respectively.

Methods that make use of samples from the power posteriors are much more accurate than HM as has been widely documented [1–3], particularly in high dimensional problems. Among these methods, thermodynamic integration (TI) [3] is popular method to estimate the evidence of gravitational wave (GW) models, showing in general good performance. Another method, widely applied in other fields such as phylogenetics is stepping-stone sampling (SS) algorithm [2]. As this method can provide many advantages over the TI estimate, it is important to explore the performance of the SS estimator for GW models as to the best of our knowledge, the SS algorithm has not been used for evidence calculation in this context.

One of the drawbacks of power posterior methods is the significant computational cost required to produce a single evidence estimate as multiple Markov chains have to be run, one for each temperature. Fortunately, since the usual practice to carry out parameter inference on GW models is parallel tempering, the samples at different temperatures are available and can be recycled in order to use these methods. This is what software packages such as LALInference [6] do in order to estimate the evidence.

However, as has been noticed in [6], TI might require a larger number of temperatures than the one needed for parameter estimation in order to achieve accurate estimates. Note that the samples of chains at temperatures $T > 1$ ($\beta < 1$) are only used to aid the mixing of the chain at $T = \beta = 1$ whose stationary distribution is the posterior, and are therefore discarded from the inference process. In this context, SS algorithm seems very promising since it requires fewer temperature steps than TI to provide accurate evidence estimates as we will show in section IV.

Another method to estimate the evidence, not based on power posteriors, is nested sampling (NS) [7]. This Bayesian algorithm has been successfully applied in diverse fields, such as astronomy [8] John: please add references, cosmology [9], engineering [10] and phylogenetics [11, 12]. To estimate the evidence of GW models, NS has been implemented in the software package LALInference [6]. The method has the unique property of yielding an estimation of the uncertainty associated to the evidence estimate in a single run (subject to the independence of the samples), unlike power posterior methods. But this only holds for independent samples, not when running MCMC to sample from the restricted prior as is done in practice.

Alternatively, instead of estimating the evidence for each model being tested, a trans-dimensional Reversible Jump Markov chain Monte Carlo (RJMCMC) method can be used in order to explore the joint space of all models. Then the probability for each model can be calculated simply by calculating the relative frequency of visits to each model by the Markov chain. However, this exploration depends on tuning parameters which can be difficult to specify, leading to poor mixing of the Markov chain and subsequently to large statistical errors associated with the evidence estimates [13].

Below we describe TI, SS and NS in more detail before comparing their performance in sections IV and V.

A. Thermodynamic Integration

Thermodynamic integration or the more general path sampling [14] make use of an auxiliary variable β, $0 \leq \beta \leq 1$, to define transitional distributions, namely the power posterior distributions defined in (4) in the case of TI, that provide a path from the prior ($\beta = 0$) to the posterior distribution ($\beta = 1$). By explicitly denoting the evidence z_β as a function of β by

$$z(\boldsymbol{X}|\beta) = \int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M) \mathrm{d}\boldsymbol{\theta},$$

the log marginal likelihood has the representation as the integral over the 1-dimensional parameter β of half the mean deviance where the expectation is taken with respect to the power posterior:

$$\log(z) = \log\left(\frac{z(\boldsymbol{X}|\beta = 1)}{z(\boldsymbol{X}|\beta = 0)}\right) = \int_0^1 E_\beta \left[\log(p(\boldsymbol{X}|\boldsymbol{\theta}, M)\right] \mathrm{d}\beta. \tag{5}$$

Representation (5) follows by integration from

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\log(z(\boldsymbol{X}|\beta)) = \frac{1}{z(\boldsymbol{X}|\beta)}\frac{\partial}{\partial\beta}z(\boldsymbol{X}|\beta) \\
&= \frac{1}{z(\boldsymbol{X}|\beta)}\frac{\partial}{\partial\beta}\int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{z(\boldsymbol{X}|\beta)}\int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \log(L(\boldsymbol{X}|\boldsymbol{\theta}, M))\pi(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M)}{z_\beta}\log(L(\boldsymbol{X}|\boldsymbol{\theta}, M))\mathrm{d}\boldsymbol{\theta} \\
&= E_\beta\left[\log(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\right].
\end{aligned}$$

The samples from the parallel tempered chains for different values of β provide samples from the power posteriors and the expectation $E_\beta\left[\log(L(\boldsymbol{X}|\boldsymbol{\theta}, M)\right]$ is then estimated by the sample average. The integral in equation (5) is then approximated by numerical integration, e.g. using the trapezoidal or Simpson's rule.

B. Stepping-stone Sampling Algorithm

Stepping-stone sampling is another method to estimate the marginal likelihood. It has been widely used by the phylogenetic community where it was proposed by [2]. SS works basically by mixing elements from importance sampling and simulated annealing methods. This method relies on the same sampling scheme required by TI. Therefore, its implementation in any software package where TI or parallel tempering has already been implemented should be straightforward. SS has the advantage of requiring fewer path steps than TI to accurately estimate the marginal likelihood and yielding a less-biased estimator as demonstrated in section IV.

The marginal likelihood can be seen as the ratio $z = z_1/z_0$, where $z_0 = 1$ since the prior is assumed to be proper. The direct calculation of this ratio via importance sampling is not reliable because the distributions involved in the numerator and denominator (posterior and prior, respectively) are, in general, quite different. To solve this problem, SS expands this ratio in a telescope product of K ratios of normalizing constants of the transitional distributions [15], that is

$$z = \frac{z_1}{z_0} = \frac{z_{\beta_1}}{z_{\beta_0}}\frac{z_{\beta_2}}{z_{\beta_1}}\cdots\frac{z_{\beta_{K-2}}}{z_{\beta_{K-3}}}\frac{z_{\beta_{K-1}}}{z_{\beta_{K-2}}} = \prod_{k=1}^{K-1}\frac{z_{\beta_k}}{z_{\beta_{k-1}}} = \prod_{k=1}^{K-1} r_k,$$

for $\beta_0 = 0 < \beta_1 < \cdots < \beta_{K-2} < \beta_{K-1} = 1$, being the sequence of inverse temperatures, where $r_k = z_{\beta_k}/z_{\beta_{k-1}}$. These individual intermittent ratios can be estimated with higher accuracy than $\frac{z_1}{z_0}$ because the distributions in the numerator and denominator are generally quite similar when using a reasonable number of temperatures K. In this situation the importance sampling method works well.

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

$$r_k = \frac{z_{\beta_k}}{z_{\beta_{k-1}}} = \int_\Theta \frac{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^{\beta_k}}{L(\boldsymbol{X}|\boldsymbol{\theta}, M)^{\beta_{k-1}}}\, p_{\beta_{k-1}}(\boldsymbol{\theta}|\boldsymbol{X}, M)\mathrm{d}\boldsymbol{\theta},$$

which is estimated by its unbiased MC estimator

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

where $\boldsymbol{\theta}^1_{\beta_{k-1}}, \ldots, \boldsymbol{\theta}^n_{\beta_{k-1}}$ are drawn from $p_{\beta_{k-1}}$ with $k = 1, \ldots, K-1$.

Therefore, the SS estimate of the marginal likelihood is defined as

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

with log-version

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

Although $\widehat{z}$ is an unbiased, the log transformation introduces a bias which can be alleviated by increasing K [2].

The performance of this method depends naturally on its specifications such as the number of transitional distributions and number of samples from each of them (K and n, respectively). The dispersal of the β values has also a strong influence, even more so in TI (see [2] and our simulation study below). Along these lines, [2] proposed to spread the β values according to the evenly spaced quantiles of a Beta(0.3, 1) distribution. This distribution is right skewed, thereby putting half of the β values below 0.1 where most of the variability is found.

SS is closely related to annealed importance sampling [5]. The latter utilizes the same product of ratios, but instead of estimating each ratio separately, it estimates the entire product via importance sampling, that is the whole telescope product is evaluated multiple times and then the these values are averaged [16]. For the particular case of $K = 2$, that is considering only the prior, both methods reduce to the arithmetic mean, and for $n = 1$, they are equivalent.

C. Nested Sampling

NS transforms the multidimensional integral defined in (1), by making use of a property of positive random variables (see [11] for more details), into a one-dimensional one that utilizes a function that relates the prior with the likelihood as

$$z = \int_0^1 L(\xi)\mathrm{d}\xi,$$

where L is the likelihood as a function of the prior volume ξ. This function can be read as the proportion of prior volume ξ with likelihood values greater than $L(\xi)$.

This likelihood is a non-increasing function over the unit range. For a given decreasing sequence of ξ-values and an increasing sequence of L-values, the marginal likelihood can be estimated using, for instance, the trapezium rule

$$\widehat{z}_{NS} = \sum_{i=1}^{K}\frac{1}{2}(\xi_{i-1} - \xi_{i+1})L_i,$$

where $0 < \xi_{K+1} < \xi_K < \cdots < \xi_1 < \xi_0 = 1$.

NS explores the parameter space from the prior toward those areas of high likelihood values over time. For this, a set of N points, called *live* points, are drawn independently from the prior. The point $\boldsymbol{\theta}_1$ with the lowest likelihood associated to these points is detected and the latter is registered as L_1. Then, this point $\boldsymbol{\theta}_1$ is replaced by a new one $\boldsymbol{\theta}^*$ drawn from the prior but restricted to have a greater likelihood, that is $L(\boldsymbol{\theta}^*) > L(\boldsymbol{\theta}_1)$. This procedure is repeated until a given stopping criterion is satisfied. Thus, an increasing sequence of likelihood values $L_1, \ldots, L_K$ is generated.

Even though the ξ-values cannot be measured precisely, the nature of this algorithm allows to estimate them. The ξ-sequence can be defined as

$$\xi_1 = u_1,\ \xi_2 = u_2\xi_1, \ldots,\ \xi_K = u_K\xi_{K-1},$$

where $u_i \sim \text{Beta}(N, 1)$. The geometric mean is the most common method to estimate the u-values, which yields

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

The nature of NS algorithm also allows to estimate the uncertainty in a single run as

$$\text{SD}(\log z) = \sqrt{\frac{H}{N}}, \tag{6}$$

where H is the negative entropy. Alternatively, for a fixed sequence of likelihood values and multiple sequence of ξ-values, generated from different $u \sim \text{Beta}(N, 1)$ values, a distribution of marginal likelihood estimates can be generated and subsequently the uncertainty can be estimated. NS results are valid if and only if the samples drawn at each iteration are independent.

III. ESTIMATION OF THE MC STANDARD ERROR OF THE EVIDENCE

The point estimate of the evidence is not sufficient if we want to compare the performance of different types of evidence estimates. We need to have a measure of the Monte Carlo standard error of the evidence estimates. In the NS case, the algorithm provides direct ways of calculating its standard error from a single run as given in (6). However, power posterior methods lack a reliable direct way of calculating the standard error of the evidence. In [3] and [2], the authors proposed estimates which rely on the independence of the samples in the Markov chains at different temperatures, an assumption that is not met in general. Practitioners opt for the standard procedure of repeating the analysis multiple times and then calculating the standard error. This brute force technique can be very costly and is in some cases computationally not viable. Alternatively, some estimate the error internally in a single run, that is by re-sampling independently the Markov chains in order to generate multiple evidence estimates. However, this approach does not consider the potential autocorrelation in the samples, leading to wrong estimates. Here, we propose the use of a block bootstrap method for multivariate time series, which accounts for the autocorrelation between the samples within a Markov chain at a fixed temperature and the cross-correlation between parallel chains at different temperatures.

Cornish and Littenberg (2015) [13] estimate TI error based on RJMCMC results. Maybe we should mention it.

Bootstrap is a resampling procedure proposed by [17], initially for independent variables and later generalized by several authors. [18] proposed an extension for the case of time series, which differs from the original algorithm by allowing the sampling in blocks. The method is known as *moving block bootstrap*, in short MBB. This allows to take into account the presence of dependence in the data.

Let $X_1, \ldots, X_n$ be the observed values from a sequence of stationary random variables, in our case, a Markov chain. Define the overlapping blocks $B_i = (X_i, , \ldots, X_{i+\ell-1})$ of length ℓ, for $1 \le i \le n - \ell + 1$ and $1 \le \ell \le n$, that is

$$\begin{aligned} B_1 &= (X_1, X_2, X_3, \ldots, X_\ell) \\ B_2 &= (X_2, X_3, X_4, \ldots, X_{\ell+1}) \\ &\vdots \\ B_m &= (X_{n-\ell+1}, \ldots, X_n), \end{aligned}$$

where $m = n-\ell+1$. MBB works by resampling randomly b blocks (for didactic reasons, suppose that $b = n/\ell$) and concatenating them in order to form a set of bootstrap observations $X_1^*, \ldots, X_n^*$. For $\ell = 1$, the original bootstrap method for i.i.d. data is recovered. This procedure is repeated as usual, generating the distribution of the statistic of interest, in our case the marginal likelihood. In the general case that n is not a multiple of ℓ, we can concatenate the random sample of b block bootstraps, where b is n/ℓ rounded up, and discard the leftover points $X_{n+1}^*, \ldots, X_{b\ell}^*$, such that the bootstrap observation set has length n, as the original dataset.

Variants of this method can be found in [19], such as *stationary bootstrap*, where the block length follows a geometric distribution; *nonoverlapping block bootstrap*, which as its name say, considers nonoverlapping blocks; and *circular block bootstrap*, which increases the original dataset with the first $\ell - 1$ observations in order to give equal weights to all of them.

In the context of parallel tempering, in which case there are multiple Markov chains, we can generate the bootstrap observations using the same scheme for all the chains. For instance, assuming equal chain lengths, a bootstrap observation set for a Markov chain consisting in (B_6, B_4, B_2), can be replicated across the other chains. This procedure takes into account the potential autocorrelation within the chains and the cross-correlation between the chains due to the swaps in parallel tempering sampling. This is the approach applied in our examples.

IV. SIMULATION STUDY

We consider a simple Gaussian model used by [3] to test TI and compare it to the harmonic mean method.

Here, it is used to compare SS to TI. In addition, we assess the error estimate via the moving block bootstrap method and compare it to the empirical calculation of the error.

The model is parametrized by a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ of dimension d. The prior on $\boldsymbol{x}$ is a product of independent standard normal distributions on each x_i, for $i = 1, \ldots, d$. The likelihood is

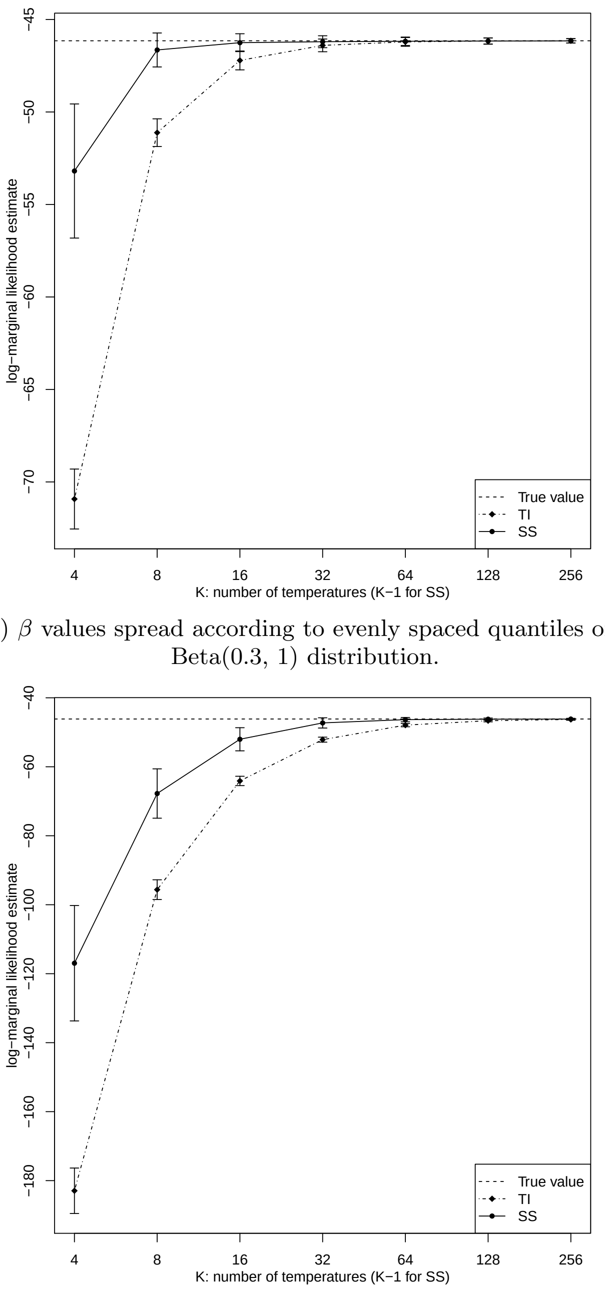
$$L(\boldsymbol{x}) = \prod_{i=1}^{d} e^{-\frac{x_i^2}{2v}},$$

where v is a parameter. Doing some calculations, it is easy to see that the posterior distribution is given by a product of independent $\mathrm{N}(0, v/(1+v))$ distributions, and therefore, its marginal likelihood has an analytical solution, which is $z = (v/(1+v))^{d/2}$. The power posterior or transitional distributions are given by a product of independent $\mathrm{N}(0, v/(v+\beta))$ distributions. All the involved distributions are Gaussians, so the sampling required to calculate TI and SS is straightforward. However, we use the Metropolis algorithm to sample these densities and thus allow a certain degree of autocorrelation in the samples, making the analysis more realistic in an evidence estimation context. The Markov chains have a lag of around 18 on average. In addition, we consider independent samples to asses mbb performance in the context of error estimation.

A. Evidence estimate

We consider the following model specifications: $v = 0.01$ and $d = 20$. This yields a log-marginal likelihood value -46.15. The analysis is performed for $n = 1000$ and $K = 4, 8, 16, 32, 64, 128, 256$. Strictly speaking, SS uses $K-1$ temperatures, since does not require samples from the posterior. For the arrangement of the β values, we test two approaches: evenly spaced values from 0 to 1, and values spread according to evenly spaced quantiles of a $\mathrm{Beta}(0.3, 1)$ distribution. The MCMC analysis is replicated 1000 times (with different random seeds) in order to calculate the error associated with the estimates. The same power posterior samples are used to estimate SS and TI.

Figures 1a and 1b display the results. It can be clearly noticed in both cases that SS requires less temperatures than TI to produce estimates around the true value. For the case the β values are calculated according to a Uniform(0,1), Figure 1b, TI fails dramatically for low number of temperatures, whereas SS, even though it fails too, its estimates are closer to the true value. For $K = 4$, TI is more than 130 units away from the true value. This shows that TI is more sensitive to the distribution of the temperatures as was similarly shown by [2].



(a) β values spread according to evenly spaced quantiles of a Beta(0.3, 1) distribution.

(b) β values equally spaced between 0 and 1.

FIG. 1: Log-marginal likelihood estimates as a function of the number of temperatures K for the Gaussian model. Error bars depict ±1 standard error based on 1000 independent MCMC analyses.

Both methods improve their performance when most of the computational effort is allocated in sampling in power posterior distributions near the prior, that is for hight temperatures. This is the effect of the Beta(0.3, 1) distribution, which allows that half of the β values are less than 0.1. The results for this case are displayed in Figure 1a. Even though TI improves its performance considerably, it can not outperform SS, which still needs fewer step temperatures to produce estimates around the true value.

B. Standard error estimate

Based in the case that the β values follow a Beta(0.3, 1) distribution, we study the performance of the MBB method for estimating the evidence error. For this, we calculate the standard error from the 1000 independent evidence estimates used in the previous analysis and compare it to the values calculated via MBB for different block lengths, $\ell = 1, 10, 30, 50, 100, 200, 300$.

The results are shown in Figure 2a. The case $\ell = 1$ is the original bootstrap method, which is used frequently for power posterior methods, but which ignores the dependence in the sampled values of the Markov chain. It is obvious that in the simple bootstrap with block length $\ell = 1$, the standard error is severely underestimated. On the other hand, the standard error estimates improved significantly using the MBB with larger block lengths, but still some underestimate the standard error. However, this example is an extreme case of highly correlated Markov chains.

We have also performed the analysis in the ideal case that the samples in the Markov chains are completely independent. The result are displayed in Figure 2b. In this case, the standard bootstrap method, that is $\ell = 1$, is sufficient to estimate the standard error reasonable. Large block lengths cause, in general, a slight underestimation in the case of low temperature numbers, but as the number of temperatures increases, the estimates are located around the empirical error estimates and less dispersed.

V. APPLICATION TO LIGO DATA

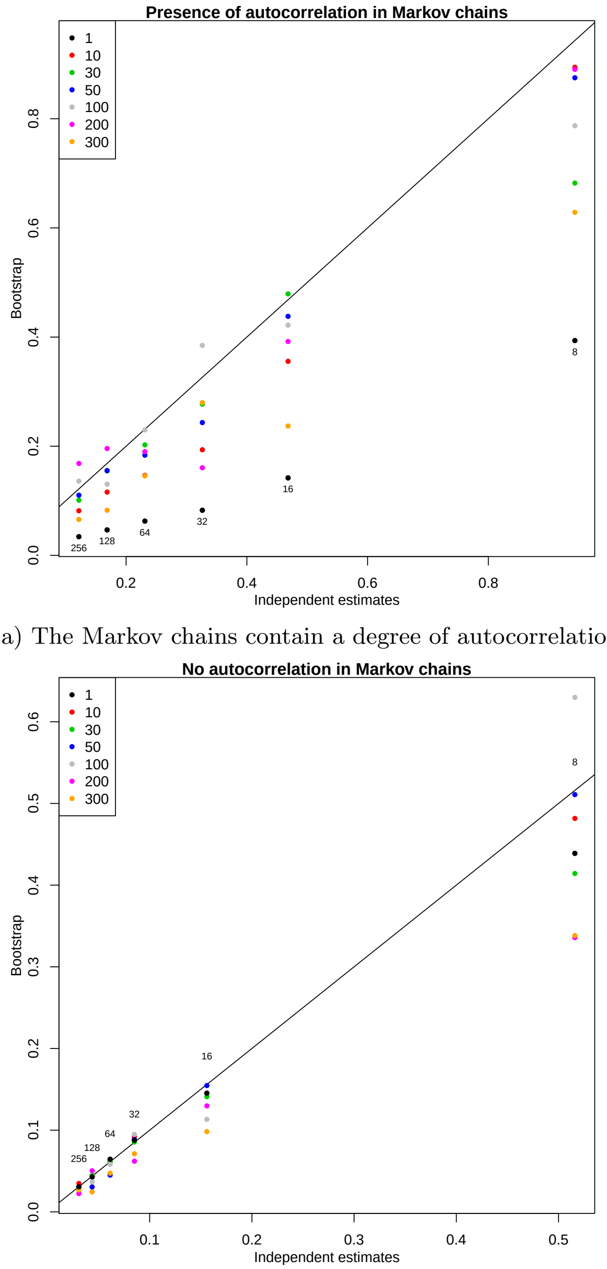
John: Some description of the data etc

Could/should we also calculate the marginal likelihood here for the same data but for a model without signal present and calculate the BF and compare?

Include a comparison of TI and SS to the nested sampling estimate

VI. DISCUSSION

SS is a method to estimate the marginal likelihood which has great popularity in phylogenetics where has shown to work well. It requires slightly less computational cost than TI to yield a less biased estimator. For our simple Gaussian model, we have shown that it is less sensitive to the disposal of the β values and the number of power posterior distributions. To the best of our knowledge, it has not been applied into gravitational wave models yet. Its implementation in this context should be straightforward since its main complexity lies with sampling from the power posterior, like TI. However, this can be done by using the parallel tempering method, which has been widely implemented in software packages in the gravitational wave context.



(a) The Markov chains contain a degree of autocorrelation.

(b) The samples in the Markov chain are completely independent.

FIG. 2: Standard error of independent evidence estimates versus the one calculated via MBB. The numbers inside the plot stand for K, number of temperatures. The legend describes the different block lengths used in MBB.

The performance of SS depends on its specifications, such as n, K and the distribution of the β values. In addition, it depends on how different the prior and the posterior are. To avoid the dependence on the prior distribution, an extension of SS has been proposed, which is known as generalized steppingstone sampling [GSS; 20].

This method makes use of a reference distribution which aims to shorten the distance between the prior and the posterior. Even though it requires posterior samples to construct the reference distribution, it is much more accurate than its simple version and requires less steps to yield the same accuracy. This happens only if the reference distribution is a reasonable approximation of the posterior, otherwise it can dramatically fail [11].

One of the drawbacks of power posterior methods is the lack of a direct procedure to estimate the error associated to the evidence estimation. In practice, the evidence is estimated multiple times in order to estimate empirically its standard error. This procedure might be highly expensive. Alternatively, some use the standard bootstrap method. It is much cheaper than the first approach, but it does not have the power of taking into account any kind of dependency within and between the Markov chains. We have proposed a moving block bootstrap method (MBB). This approach has the power of taking into account potential autocorrelation within the chains and cross-correlation between chains. We showed in Example IV B that standard bootstrap underestimates severally the empirical error in the presence of autocorrelation in the Markov chains. On the other hand, the MBB method improves significantly the estimates.

importance here demonstrated for compact binary inspirals but point out that algorithm is important as well for astrophysical and cosmological model selection.

ACKNOWLEDGMENTS

We thank Claudia Kirch for helpful discussions on the block-bootstrap for time series. We also thank the New Zealand eScience Infrastructure (NeSI) for their high performance computing facilities, and the Centre for eResearch at the University of Auckland for their technical support. PM's and RM's work is supported by Grant 3714568 from the University of Auckland Faculty Research Development Fund and the DFG Grant KI 1443/3-1, and NC's work by NSF grants PHY-1204371 and PHY-1505373. This paper has been given LIGO Document Number xxxxx. All analysis was conducted in R, an open-source statistical software available on CRAN (cran.r-project.org) and LALInference.

[1] M. A. Newton and A. E. Raftery, J. Roy. Statist. Soc. Ser. B **56**, 3 (1994).

[2] W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen, Syst. Biol. **60**, 150 (2011).

[3] N. Lartillot and H. Philippe, Syst. Biol. **55**, 195 (2006).

[4] N. Friel and A. N. Pettitt, J. Roy. Stat. Soc. B **70**, 589 (2008).

[5] R. M. Neal, Stat. Comput. **11**, 125 (2001).

[6] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin, W. Del Pozzo, F. Feroz, J. Gair, C.-J. Haster, V. Kalogera, T. Littenberg, I. Mandel, R. O'Shaughnessy, M. Pitkin, C. Rodriguez, C. Röver, T. Sidery, R. Smith, M. Van Der Sluys, A. Vecchio, W. Vousden, and L. Wade, Phys. Rev. D **91**, 042003 (2015).

[7] J. Skilling, Bayesian Analysis **1**, 833 (2006).

[8] B. J. Brewer and C. P. Donovan, Mon. Not. R. Astron. Soc. **448**, 3206 (2015).

[9] F. Feroz, M. Hobson, and M. Bridges, Monthly Notices of the Royal Astronomical Society **398**, 1601 (2009), cited By 723.

[10] R. Henderson, P. Goggans, and L. Cao, Digital Signal Processing: A Review Journal **70**, 84 (2017), cited By 1.

[11] P. Maturana R., B. J. Brewer, S. Klaere, and R. Bouckaert, "Model selection and parameter inference in phylogenetics using Nested Sampling," ArXiv preprint arXiv:1703.05471v2 (2017).

[12] P. Maturana Russel, in Bayesian Inference and Maximum Entropy Methods in Science and E edited by A. Polpo, J. Stern, F. Louzada, R. Izbicki, and H. Takada (Springer International Publishing, Cham, 2018) pp. 211–219.

[13] N. J. Cornish and T. B. Littenberg, Class. Quant. Grav. **32**, 135012 (2015).

[14] A. Gelman and X. Meng, Statistical Science **13**, 163 (1998), cited By 437.

[15] R. M. Neal, (1993).

[16] P. Maturana R., Bayesian inference in phylogenetics using Nested Sa Ph.D. thesis, The University of Auckland (2017).

[17] B. Efron, Ann. Statist. **7**, 1 (1979).

[18] H. R. Kunsch, Ann. Statist. **17**, 1217 (1989).

[19] S. N. Lahiri, Resampling Methods for Dependent Data, Springer series in statistics (Springer New York, New York, NY, 2003).

[20] Y. Fan, R. Wu, M.-H. Chen, L. Kuo, and P. O. Lewis, Mol. Biol. Evol. **28**, 523 (2011).