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[3,4]

[1]*Department of Statistics, University of Auckland, Auckland 1142, New Zealand*
[2]*Institute for Gravitational Research, School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[3]*ARTEMIS, Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, CS 34229, F-06304 Nice Cedex 4, France*
[4]*Physics and Astronomy, Carleton College, Northfield, Minnesota 55057, USA*

(Received 11 October 2018; published 9 April 2019)

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 marginal likelihood of a model but can produce posterior samples as a byproduct. 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 fewer 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 signal model applied to simulated data from the Advanced LIGO and Advanced Virgo gravitational wave detectors. 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.

DOI: 10.1103/PhysRevD.99.084006

I. INTRODUCTION

It has now been two decades since Bayesian parameter estimation routines were first introduced for studies in astrophysics [1], gravitational waves (GWs) [2], and cosmology [3,4]. Bayesian parameter estimation routines have become extremely important for these disciplines, and their use is ubiquitous [5]. Recent dramatic observations with, for example, the cosmic microwave background and gravitational waves, have been used with Bayesian parameter estimation methods to significantly push our knowledge of the Universe and its history [6–12]. Advances in computer power, coupled with new and innovative Bayesian parameter estimation techniques, continue to push the applicability and importance of Bayesian methods [13].

The importance of accurate parameter estimation calculations was dramatically displayed with the observations of gravitational waves and gamma rays from the binary neutron star merger GW170817 and GRB 170817A [12,14]. Using the data from the two Advanced LIGO detectors [15] and the Advanced Virgo detector [16] an initial sky-map and distance estimate from the gravitational wave data was released five hours after the merger [17] using a specially designed method for sky position estimation [18]. A little over 11 hours after the gravitational wave—gamma ray event refined estimates were released based on the first comprehensive parameter estimation [19], giving a more accurate estimate of the sky position and distance to the source [17]. The parameter estimation calculations allowed astronomers to identify the source, providing for electromagnetic observations that yielded a plethora of astrophysical information, including the observation of a kilonova [17]. Parameter estimation of gravitational wave models will continue to be significant for multimessenger astronomy.

Also of critical importance is the ability to conduct model comparison and parameter estimation studies with the gravitational wave signals. For example, Bayesian parameter estimation methods were used to decipher the physical characteristics of the observed gravitational wave events, such as the first observed binary black hole merger GW150914 [9,10,20,21], and the binary neutron star merger GW170817 [12,22]. Similarly, model comparisons have been conducted in a number of ways using the data from the detected gravitational waves signals. This includes tests of general relativity [23], neutron star equation of state studies [12,24], constraining tidal instabilities in binary neutron star mergers [25], and the search for a stochastic gravitational wave background from binary black hole mergers over the history of the Universe [26]. When Advanced LIGO and Advanced Virgo made the first observation of a binary black merger using the data from three detectors it provided an opportunity to conduct a model comparison test as to whether the polarization of the gravitational waves was consistent with general relativity, or other theories of gravity; general relativity succeeded in this important model comparison [27]. Advanced LIGO data were also used to search for a stochastic gravitational wave background as described by general relativity or alternative theories of gravity, and model comparison was integral to this study [28,29]. Methods that improve the calculation of the evidence, the marginal likelihood of a model, would be well received in the gravitational wave community, and would certainly be of use in other areas of astrophysics and cosmology [5]. Equally important is an accurate estimation of the associated standard error. Here we introduce the moving block bootstrap (MBB) [30] that accounts for the autocorrelation between the samples, and provides a more accurate estimate than the standard bootstrap method.

This paper makes two main contributions to the existing literature on evidence calculation of GW models. First, this paper proposes the stepping-stone sampling (SS) algorithm [31] that provides an evidence estimator for Bayesian model selection and the MBB for computing its standard error. The main advantage of the SS algorithm is its reduced computational costs in comparison to the evidence estimate based on thermodynamic integration, a popular method used in the GW field. Furthermore, the evidence estimate based on the SS algorithm has lower standard error. The practicality of the SS algorithm for calculating the evidence of gravitational wave models is demonstrated. The SS algorithm could be a further advancement for model selection for gravitational wave data analysis, as well as for other applications in astrophysics and astronomy. Secondly, this paper introduces the MBB for a more accurate estimation of the standard error of the evidence estimate. The MBB provides a general method to compute the standard error of any quantity that is based on Markov chain Monte Carlo (MCMC) output.

The paper is structured as follows. In Sec. II we review nested sampling, thermodynamic integration and introduce the stepping-stone algorithm for computing the evidence of Bayesian model selection. In Sec. III we introduce the moving block bootstrap for calculating the Monte Carlo standard error of the evidence estimates. In Sec. IV we show the enhanced performance of the SS algorithm over thermodynamic integration in a simulation study. The different algorithms are then applied to simulated LIGO-Virgo gravitational wave data in Sec. V. Their results are contrasted and the benefits of the SS algorithm and the moving black bootstrap for standard error estimation become evident. A summary discussion is given in Sec. VI.

II. COMPUTATION OF MARGINAL LIKELIHOOD

The evidence or marginal likelihood of a model M is a multidimensional 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 data set, $L(\boldsymbol{X}|\boldsymbol{\theta}, M)$ 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 [32]. Let $\boldsymbol{\theta}_i, i = 1, \ldots, n$ be samples from the prior; the AM estimator is an average of corresponding n likelihood values,

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

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,

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

This is the harmonic mean of likelihood values.

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 [31,33–35]. 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 the inverse temperature $0 \leq \beta \leq 1$, 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 [31–33], particularly in high dimensional problems. Among these methods, thermodynamic integration (TI) [33] is a popular method to estimate the evidence of GW models, showing in general good performance. Another method, widely applied in other fields such as phylogenetics is the SS algorithm [31]. 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 parallel tempering is commonly used in GW parameter estimation, the samples at different temperatures are available and can be recycled in order to use these methods.

However, as has been noticed in [19], 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, the SS algorithm seems very promising since it requires fewer temperature steps than TI to provide accurate evidence estimates as we show in Sec. IV.

Another method to estimate the evidence, not based on power posteriors, is nested sampling (NS) [36,37]. This Bayesian algorithm has been successfully applied in diverse fields, such as astronomy [38], cosmology [39], engineering [40] and phylogenetics [41,42]. To estimate the evidence of GW models, NS has been implemented in the software package LALInference [19]. The method has the unique property of yielding an estimation of the uncertainty associated with the evidence estimate in a single run (however, only for independent samples).

Alternatively, instead of estimating the evidence for each model being tested, a transdimensional reversible jump Markov chain Monte Carlo [43,44] 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 [45].

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

A. Thermodynamic integration

Thermodynamic integration or the more general path sampling [46] makes 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

$$z(\boldsymbol{X}|\beta) = \int_\Theta L(\boldsymbol{X}|\boldsymbol{\theta}, M)^\beta \pi(\boldsymbol{\theta}|M) \mathrm{d}\boldsymbol{\theta}, \tag{5}$$

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

$$\log(z) = \log\left(\frac{z(\boldsymbol{X}|\beta = 1)}{z(\boldsymbol{X}|\beta = 0)}\right) = \int_0^1 E_\beta[\log(p(\boldsymbol{X}|\boldsymbol{\theta}, M)] \mathrm{d}\beta. \tag{6}$$

Representation (6) follows by integration from

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

The samples from the parallel tempered chains for different values of β provide samples from the power posteriors and the expectation $E_\beta[\log(L(\boldsymbol{X}|\boldsymbol{\theta}, M)]$ is then estimated by the sample average. The integral in Eq. (6) 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 [31]. 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 Sec. 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 [47], 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 $\beta_0 = 0 < \beta_1 < \ldots < \beta_{K-2} < \beta_{K-1} = 1$, 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 Monte Carlo estimator

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

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

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

with log version

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

Although $\hat{z}$ is unbiased, the log transformation introduces a bias which can be alleviated by increasing K [31].

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, but not as strong as in TI (see [31] and our simulation study below). Along these lines, [31] 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 [35]. 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 these values are averaged [48]. 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 [41] 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 nonincreasing 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

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

where $0 < \xi_{m+1} < \xi_m < \cdots < \xi_1 < \xi_0 = 1$.

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, is 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 $L_1, \ldots, L_m$ is generated.

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

$$\xi_1 = u_1, \xi_2 = u_2\xi_1, \ldots, \xi_m = u_m\xi_{m-1},$$

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

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

The nature of the NS algorithm also allows one to estimate the standard error of the $\log z$ estimate in a single run as

$$\widehat{\mathrm{s.e.}}_{\mathrm{NS}}(\log z) = \sqrt{\frac{H}{N}}, \tag{7}$$

where H is the negative entropy. However, this NS standard error estimate is only valid if the samples are drawn independently. In practice though, the samples are often serially dependent because Metropolis-Hastings algorithms are used for their generation. As an alternative to (7), for a fixed sequence of likelihood values and multiple sequence of ξ-values, generated from different $u \sim \mathrm{Beta}(N, 1)$ values, a distribution of marginal likelihood estimates can be generated and subsequently the uncertainty can be estimated.

III. ESTIMATION OF THE MONTE CARLO STANDARD ERROR OF THE EVIDENCE

The point estimate of the evidence is subject to random errors and therefore we need to have a measure of the Monte Carlo standard error of the evidence estimates. This is also important if we want to compare the performance of different types of evidence estimates. In the NS case, the algorithm provides direct ways of calculating its standard error from a single run as given in (7). However, power posterior methods lack a reliable direct way of calculating the standard error of the evidence. In [33] and [31], 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 resampling 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.

Bootstrap is a resampling procedure proposed by [49], initially for independent variables and later generalized by several authors. An extension for the case of time series was proposed in [30], which differs from the original algorithm by allowing the sampling in blocks. The method is known as MBB. This allows one 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 data set.

Variants of this method can be found in [50], such as stationary bootstrap, where the block length follows a geometric distribution, nonoverlapping block bootstrap, which as its name suggests, considers nonoverlapping blocks, and circular block bootstrap, which increases the original data set 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 need to 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) is 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 [33] to test TI and compare it to the harmonic mean method. Here, it is used to compare SS to TI. We also assess the error estimate via the MBB method and compare it to the empirical calculation of the error. In addition, we study NS performance for different sampling specifications.

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

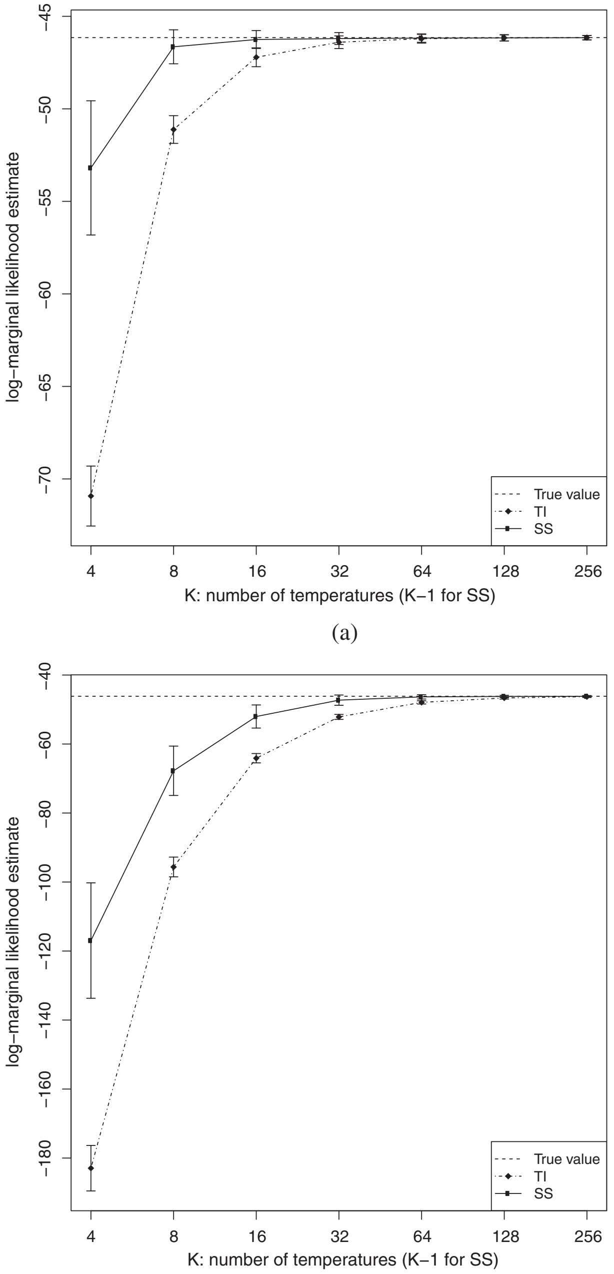
$$L(\boldsymbol{x}) = \prod_{i=1}^{d} e^{-\frac{x_i^2}{2v}},$$

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 assess 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 it 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 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 1(a) and 1(b) display the results. It becomes clear in both cases that the SS algorithm requires fewer temperatures than TI to produce estimates around the true value. When the β values are calculated according to a Uniform (0,1), Fig. 1(b), the TI estimates are seriously biased for low number of temperatures, whereas the SS estimates, even though biased too, are closer to the true value. For equally spaced β values and $K = 4$ Fig. 1(b), TI is more than 130 units away from the true value compared to the around 25 units for β values spread according to quantiles of the Beta(0.3,1) distribution in Fig. 1(a). This shows that TI is more sensitive to the distribution of the temperatures as was similarly shown by [31].



(a)

(b)

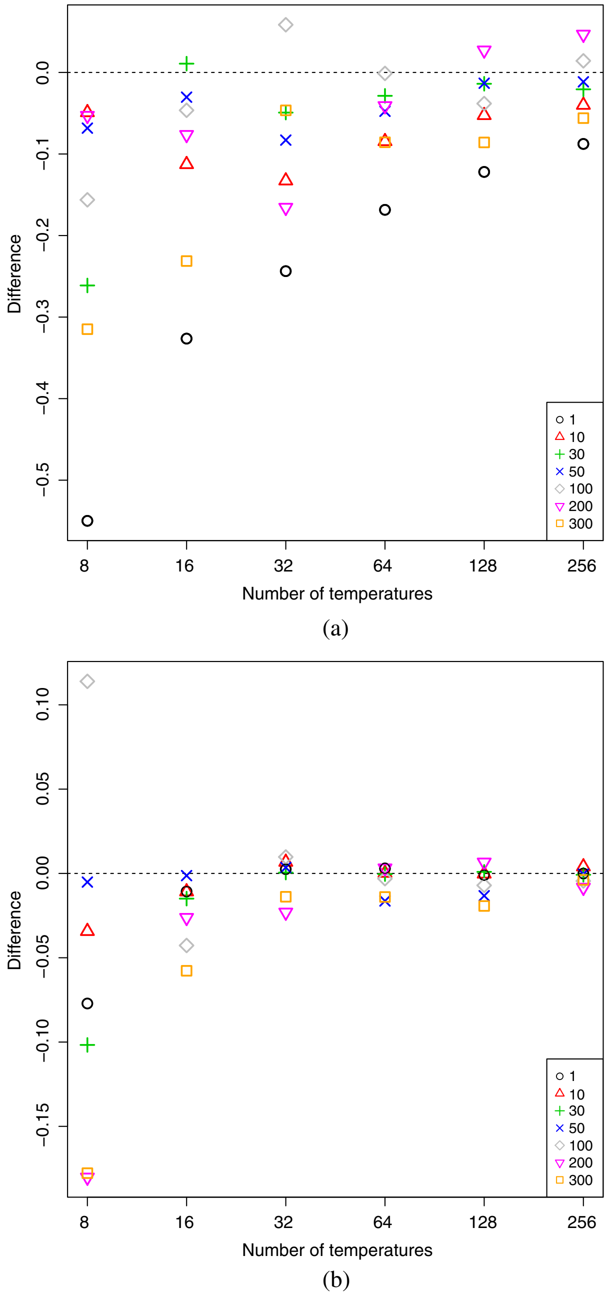
FIG. 1. Log-marginal likelihood estimates as a function of the number of temperatures K for the Gaussian model. Error bars depict ±1 standard error based on 1000 independent MCMC analyses. (a) β values spread according to evenly spaced quantiles of a Beta(0.3, 1) distribution. (b) β values equally spaced between 0 and 1.

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 high 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 Fig. 1(a). Even though TI improves its performance considerably, it cannot outperform SS, which still needs fewer step temperatures to produce estimates around the true value.

B. Standard error estimate

Based on 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, call this $\widehat{\text{s.e.}}_{\text{ind}}$ and compare it to the standard error estimates calculated using MBB ($\widehat{\text{s.e.}}_{\text{MBB}}$ for different block lengths, $\ell = 1$, 10, 30, 50, 100, 200, 300, via their differences, $\widehat{\text{s.e.}}_{\text{MBB}} - \widehat{\text{s.e.}}_{\text{ind}}$).



(a)

(b)

FIG. 2. Difference between the standard error calculated via MBB and the one from independent evidence estimates. The legend shows the different block lengths used in MBB. (a) The Markov chains contain a degree of autocorrelation. (b) The samples in the Markov chain are completely independent.

The results are shown in Fig. 2(a). 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 results are displayed in Fig. 2(b). In this case, the standard bootstrap method, that is $\ell = 1$, is sufficient to estimate the standard error reasonably well. Large block lengths cause, in general, a slight underestimation but only in the case of a low number of temperatures. As the number of temperatures increases, the estimates are located around the empirical error estimates, i.e., around 0, and less dispersed.

We caution against the use of the theoretical standard error estimate of NS in Eq. (7) when the Metropolis-Hastings algorithm is used to generate the samples rather than sampling independently, as the validity of this theoretical standard error estimate is based on the independence assumption. To this end, we include a comparison of this theoretical NS standard error estimate with the empirical standard error obtained from 100 independent runs in Table I. We observe a decrease in bias with increasing number of MCMC steps. However, the NS standard error estimates are far too small and thus underestimate the uncertainty even for a large number of MCMC steps of

TABLE I. Nested sampling results based on 100 independent NS runs. N is the number of live points, “steps” the number of MCMC steps used to generate the points at each iteration, Ave($\widehat{\text{s.e.}}_{\text{NS}}$) the average of the theoretical standard error estimate defined in (7), SD($\widehat{\text{s.e.}}_{\text{NS}}$) the standard deviation of the theoretical standard error estimates, $\widehat{\text{s.e.}}_{\text{NS,ind}}$ the standard error estimate based on the independent marginal likelihood estimates, and “bias” the difference between the true value and the mean of the NS marginal likelihood estimates.

N	Steps	Ave($\widehat{\text{s.e.}}_{\text{NS}}$)	SD($\widehat{\text{s.e.}}_{\text{NS}}$)	$\widehat{\text{s.e.}}_{\text{NS,ind}}$	Bias
10	10	1.9922	0.0921	3.6244	4.7724
10	100	1.8918	0.0562	2.3078	−0.3310
10	1000	1.8959	0.0542	2.1083	−0.1580
10	5000	1.8939	0.0562	2.3454	−0.3233

5000. This is a well-known shortcoming of the NS standard error estimate; e.g., a more detailed examination of this issue can also be found in Fig. 4 of [37].

V. APPLICATION WITH SIMULATED LIGO-VIRGO DATA

We apply the SS algorithm to an example analysis of a simulated binary black hole coalescence signal in the Advanced LIGO [15] and Advanced Virgo [16] gravitational wave detectors, operating at design sensitivity. The data contained 4 s of simulated Gaussian noise, generated using the design sensitivity curves of two Advanced LIGO detectors (Hanford, Livingston) and the Advanced Virgo detector, plus the GW signal. The simulated black hole binary had component masses 25 and 13 $M_\odot$, and lay at a luminosity distance of 614 Mpc, with a combined signal-to-noise ratio of 17.9 in the three-detector network. The analysis was performed in the frequency range 40–512 Hz using the IMRPhenomPv2 waveform approximant [51]. The system's total angular momentum was inclined at 95° to the line of sight to the binary, and the primary and secondary black holes had dimensionless spin magnitudes of 0.67 and 0.12, tilted at 45° and 90° to the orbital angular momentum, respectively. This configuration produces a precession of the orbital plane which results in a waveform that is not well approximated by a nonspinning signal. The analysis was performed using the 15-dimensional parametrized model for a quasicircular black hole binary commonly used in LIGO-Virgo analyses (e.g., [10,21]), implemented in the LALInference package [19].

We estimate the marginal likelihood via TI and SS. Because the true evidence is unknown, we include NS estimates for comparison purposes using 256, 512, 1024, 1536 and 2048 live points. For TI and SS, we considered 8, 12, 17, 31, 46 and 61 temperatures, evenly spaced on a logarithmic scale, with 4 million samples from each with a burn-in period of 1 million and a thinning factor of 2 000. To compute the standard error of the TI and SS evidence estimates, we applied the MBB method for different block lengths, including the standard bootstrap ($\ell = 1$), and took the one that yielded the maximum standard deviation as a conservative way of estimation. The standard error for the NS estimates was obtained from 32 independent runs. The results are displayed in Table II and visualized in Fig. 3.

The NS evidence estimates have an increasing trend, and their standard error values decrease as the number of live points increase. This is a well-known behavior of this method. Taking the NS estimate with the largest number of live points, that is $N = 2048$, we can see that all SS estimates are closer than the TI ones to this value. The TI estimate which is the closest to the NS estimate is 1.14 units away, that is 3.16 NS SDs, whereas the SS estimate, which is utmost, is 0.82 units away, that is 2.28 NS SDs. In other words, the best TI estimate is farther away than the worst SS estimate from the NS estimate.

TABLE II. Evidence estimates and their corresponding standard errors from the NS method for different number of live points N, and TI and SS methods for different number of temperatures K.

Method	N	K	$\log z$	SD
NS	256	⋯	−5732.39	2.10
	512	⋯	−5731.21	0.92
	1024	⋯	−5730.83	0.60
	1536	⋯	−5730.72	0.40
	2048	⋯	−5730.82	0.36
TI	⋯	8	−5731.96	0.32
	⋯	12	−5732.14	0.34
	⋯	17	−5732.29	0.36
	⋯	31	−5732.12	0.33
	⋯	46	−5732.13	0.37
	⋯	61	−5732.09	0.21
SS	⋯	8	−5730.06	0.46
	⋯	12	−5730.00	0.21
	⋯	17	−5730.10	0.20
	⋯	31	−5730.35	0.10
	⋯	46	−5730.07	0.11
	⋯	61	−5730.19	0.07

Interestingly, as the number of temperature increases, the TI estimates do not seem to converge to the NS estimate, which could be due to the number of samples considered per temperature. This is not the case for SS, which always yields results closer to the NS estimate with the largest number of active points.

The lowest TI standard error of $\mathrm{SD} = 0.21$ in this example, which is reached for $K = 61$, required 2.44×10^8 (61×2000^2) likelihood evaluations. On the other hand, the same SS SD is obtained for $K = 12$, which only required 4.4×10^7 (11×2000^2) likelihood evaluations.

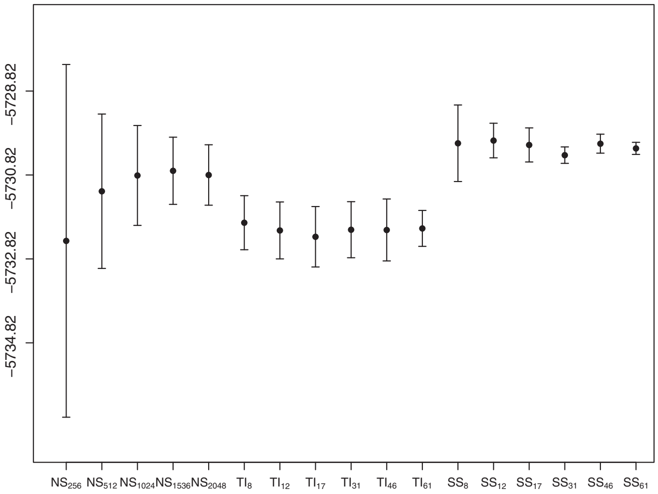


FIG. 3. Evidence estimates ±2 standard errors. Subscripts in NS stand for the number of live points N, and in TI and SS stand for the number of temperatures K.

Thus, SS is roughly 5.5 times more efficient than TI at achieving the same level of precision in this example.

VI. DISCUSSION

SS is a method to estimate the marginal likelihood which has enjoyed great popularity in phylogenetics where it has been shown to work well. It requires less computational effort than TI to yield an accurate estimate of the evidence. In a simulation study with a simple Gaussian model, we have shown that it is less sensitive to the dispersal of the inverse temperature β values and achieves a higher accuracy with a smaller number of power posterior distributions. To the best of our knowledge, it has not been applied for calculating the evidence of 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 GW software packages such as LALInference.

The performance of SS depends on its specifications, such as the number of MCMC steps n in each parallel tempering chain, the number of temperatures K and the distribution of the inverse temperature β values. In addition, it depends on how different the prior and the posterior are. To mitigate the dependence on the prior distribution, we aim to explore a recent extension of SS known as generalized stepping-stone sampling [52]. 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 could be more accurate than its simple version and require fewer steps to yield the same accuracy. For this, the reference distribution needs to be a reasonable approximation of the posterior; otherwise it can dramatically fail [41].

One of the drawbacks of power posterior methods is the lack of a direct formula for the standard error of the evidence estimate. In practice, the methods are run multiple times in order to obtain an empirical standard error estimate. This brute force approach will prove too computationally expensive in most practical applications. Alternatively, the standard bootstrap has been applied. It is computationally much cheaper than the brute force approach, but it does not take the dependencies within and between the Markov chains into account. In this paper, we have proposed a moving block bootstrap method. This approach has the ability to allow for potential autocorrelation within the chains and cross-correlation between chains. We showed in example IV B of our simulation study that the standard bootstrap severely underestimates the standard error in the presence of autocorrelation in Markov chains but that the moving block bootstrap significantly improves the standard error estimates of the evidence.

The good performance of SS, shown in the simulation study presented in Sec. IV, was replicated in a simulated LIGO-Virgo GW data set in Sec. V. SS outperformed TI in terms of the number of temperatures and the number of likelihood evaluations required to yield accurate evidence estimates, using a nested sampling estimate as a reference. In addition, its estimates have lower standard errors, which were estimated via MBB using different block lengths and considering the largest standard error as a conservative estimate.

This LIGO-Virgo application showed the advantages of SS in the context of GW models. This method requires less computational effort than TI to yield accurate estimates of the evidence with lower standard errors. As [19] noticed, TI sometimes requires a larger number of temperatures than needed for parameter inference. This could add a significant computational cost to the GW data analysis. SS could potentially be based on the same number of temperatures used for parameter inference, or at least require generally less temperatures than TI. This makes SS a better alternative in GW data analysis than TI.

An implementation of the SS and TI algorithms and the MBB in R and Python is available on Github [53,54].

ACKNOWLEDGMENTS

We thank Claudia Kirch for helpful discussions on the block bootstrap for time series and John Siryj for the translation of R-code to Python. 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. The Observatoire de la Côte d'Azur also provided support for this research. P. M. and R. M.'s work is supported by Grant No. 3714568 from the University of Auckland Faculty Research Development Fund, DFG Grant No. KI 1443/3-1 and the James Cook Fellowship from Government funding, administered by the Royal Society Te Apārangi. J. V. is supported by STFC Grant No. ST/K005014/1. N. C.'s work is supported by NSF Grants No. PHY-1806990 and No. PHY-1505373. This paper has been given LIGO document number P1800299. All analysis was conducted in R, an open-source statistical software available on CRAN and LALInference.

[1] P. Saha and T. B. Williams, Astron. J. **107**, 1295 (1994).

[2] N. Christensen and R. Meyer, Phys. Rev. D **58**, 082001 (1998).

[3] N. Christensen, R. Meyer, L. Knox, and B. Luey, Classical Quantum Gravity **18**, 2677 (2001).

[4] L. Knox, N. Christensen, and C. Skordis, Astrophys. J. Lett. **563**, L95 (2001).

[5] S. Sharma, Annu. Rev. Astron. Astrophys. **55**, 213 (2017).

[6] G. Hinshaw *et al.*, Astrophys. J. Suppl. Ser. **208**, 19 (2013).

[7] P. A. R. Ade *et al.* (Planck Collaboration), Astron. Astrophys. **571**, A16 (2014).

[8] P. A. R. Ade *et al.* (Planck Collaboration), Astron. Astrophys. **594**, A13 (2016).

[9] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).

[10] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 241102 (2016).

[11] R. Meyer and N. Christensen, Significance **13**, 20 (2016).

[12] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **119**, 161101 (2017).

[13] J. Zhu, J. Chen, W. Hu, and B. Zhang, Natl. Sci. Rev. **4**, 627 (2017).

[14] A. Goldstein *et al.*, Astrophys. J. Lett. **848**, L14 (2017).

[15] J. Aasi *et al.* (LIGO Scientific and Virgo Collaborations), Classical Quantum Gravity **32**, 115012 (2015).

[16] F. Acernese *et al.*, Classical Quantum Gravity **32**, 024001 (2015).

[17] B. P. Abbott *et al.*, Astrophys. J. Lett. **848**, L12 (2017).

[18] L. P. Singer and L. R. Price, Phys. Rev. D **93**, 024013 (2016).

[19] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).

[20] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **6**, 041014 (2016).

[21] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **6**, 041015 (2016).

[22] B. P. Abbott *et al.* (Virgo and LIGO Scientific Collaborations), Phys. Rev. X **9**, 011001 (2019).

[23] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 221101 (2016).

[24] B. P. Abbott *et al.* (Virgo and LIGO Scientific Collaborations), Phys. Rev. Lett. **121**, 161101 (2018).

[25] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **122**, 061104 (2019).

[26] R. Smith and E. Thrane, Phys. Rev. X **8**, 021019 (2018).

[27] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **119**, 141101 (2017).

[28] T. Callister, A. S. Biscoveanu, N. Christensen, M. Isi, A. Matas, O. Minazzoli, T. Regimbau, M. Sakellariadou, J. Tasson, and E. Thrane, Phys. Rev. X **7**, 041058 (2017).

[29] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **120**, 201102 (2018).

[30] H. R. Kunsch, Ann. Stat. **17**, 1217 (1989).

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

[32] M. A. Newton and A. E. Raftery, J. R. Stat. Soc. Ser. B **56**, 3 (1994).

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

[34] N. Friel and A. N. Pettitt, J. R. Stat. Soc. Ser. B **70**, 589 (2008).

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

[36] J. Skilling, Bayesian Anal. **1**, 833 (2006).

[37] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010).

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

[39] F. Feroz, M. Hobson, and M. Bridges, Mon. Not. R. Astron. Soc. **398**, 1601 (2009).

[40] R. Henderson, P. Goggans, and L. Cao, Digit. Signal Process. **70**, 84 (2017).

[41] P. Maturana Russel, B. J. Brewer, S. Klaere, and R. R. Bouckaert, Syst. Biol. **68**, 219 (2019).

[42] P. Maturana Russel, in Bayesian Inference and Maximum Entropy Methods in Science and Engineering, edited by A. Polpo, J. Stern, F. Louzada, R. Izbicki, and H. Takada (Springer International Publishing, Cham, 2018), pp. 211–219, DOI: 10.1007/978-3-319-91143-4_20.

[43] P. J. Green, Biometrika **82**, 711 (1995).

[44] R. Umstätter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan, Phys. Rev. D **72**, 022001 (2005).

[45] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).

[46] A. Gelman and X. Meng, Stat. Sci. **13**, 163 (1998).

[47] R. M. Neal, *Probabilistic inference using Markov chain Monte Carlo methods, Technical Report, Department of Computer Science* (University of Toronto, Toronto, Ontario, Canada, 1993).

[48] P. Maturana Russel, Ph.D. thesis, The University of Auckland, 2017, http://hdl.handle.net/2292/36784.

[49] B. Efron, Ann. Stat. **7**, 1 (1979).

[50] S. N. Lahiri, *Resampling Methods for Dependent Data*, Springer Series in Statistics (Springer, New York, NY, 2003).

[51] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).

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

[53] An implementation of the SS and TI algorithms and the MBB in R https://github.com/pmat747/powModSel.

[54] The implementation in Python https://github.com/pmat747/pythonPowModSel.