Costless correction of chain based nested sampling parameter estimation in gravitational wave data and beyond

Metha Prathaban[1,2,3]★ and Will Handley[1,2,4]†
[1]Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK
[2]Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge CB3 0HE, UK
[3]Pembroke College, Trumpington Street, Cambridge CB2 1RF, UK
[4]Gonville & Caius College, Trinity Street, Cambridge CB2 1TA, UK

arXiv:2404.16428v1 [astro-ph.IM] 25 Apr 2024

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Nested sampling parameter estimation differs from evidence estimation, in that it incurs an additional source of error. This error affects estimates of parameter means and credible intervals in gravitational wave analyses and beyond, and yet, it is typically not accounted for in standard error estimation methods. In this paper, we present two novel methods to quantify this error more accurately for any chain based nested sampler, using the additional likelihood calls made at runtime in producing independent samples. Using injected signals of black hole binary coalescences as an example, we first show concretely that the usual error estimation method is insufficient to capture the true error bar on parameter estimates. We then demonstrate how the extra points in the chains of chain based samplers may be carefully utilised to estimate this error correctly, and provide a way to check the accuracy of the resulting error bars. Finally, we discuss how this error affects p-p plots and coverage assessments.

Key words: nested sampling, parameter estimation, gravitational waves

1 INTRODUCTION

Nested Sampling (NS) (Skilling 2006) is a Bayesian inference tool widely used in the field of gravitational wave astronomy, both for parameter estimation and model comparison (Ashton et al. 2022; Buchner 2023; Veitch & Vecchio 2010; Thrane & Talbot 2019; Williams et al. 2021; Gair et al. 2010; Veitch et al. 2015a; Ashton et al. 2019). It enables us to perform posterior and evidence estimation on observed gravitational wave events, in turn enabling breakthrough science, such as an independent measurement of the Hubble constant (Schutz 1986; Abbott & et al. 2017b), reconstructed sky maps of binary neutron star mergers for electromagnetic follow-up (Abbott & et al. 2017a), and population inference to understand formation mechanisms of binaries (Talbot & Thrane 2017; Gerosa & Berti 2017; Mapelli 2021).

Nested sampling distinguishes itself from Markov Chain Monte Carlo (MCMC) algorithms (MacKay 2003; Foreman-Mackey et al. 2013; Ashton & Talbot 2021) in its ability to easily calculate evidences, and the errors associated with this evidence estimation are well understood (Skilling 2006). Whilst it is also popularly used for parameter inference in gravitational wave data analysis and beyond, nested sampling parameter estimation contains an additional source of error which affects its performance (Chopin & Robert 2010). Quantifying this error accurately is key, as it affects our estimates of parameter means and credible intervals on parameters. This error is essentially a result of the stochastic nature of the nested sampling algorithm, and so we are able to accurately quantify it by performing many nested sampling runs on the same data. However, this is computationally expensive and, for many gravitational wave problems, simply not viable. There is, as yet, little literature exploring accurately estimating this error from a single run. In response to this, (Higson et al. 2018a) proposed an innovative method of deconstructing nested sampling runs into single live point runs and using bootstrapping to determine a new error bar on parameter inferences.

Here, we present two novel approaches which utilise the extra likelihood evaluations already performed at runtime for any chain based nested sampler. Deemed to be too correlated to use in evidence estimation, the potential performance gains to parameter inference from this wealth of additional likelihood calls has been largely unexplored. As an example of this, we show that, by carefully harnessing these additional evaluations, we can correctly quantify parameter estimation errors and verify that these match the errors obtained from repeated nested sampling runs.

In the following section, we lay out necessary background on the nested sampling algorithm and identify the dominant sources of error for both evidence and parameter estimation. In Section 3, we introduce the two methods for estimating parameter inference error bars, in the context of a simulated black hole binary (BBH) signal, and demonstrate that both methods are able to accurately capture the error bar on the parameter means. We show that these methods also produce comparable results to the method presented in Higson et al. (2018a) in Section 4. How this error extends to estimates of credible intervals and coverage plots is discussed in Section 5 and, finally, our conclusions are presented in Section 6.

★ E-mail: myp23@cam.ac.uk
† E-mail: wh260@cam.ac.uk

© 2024 The Authors

2 BACKGROUND

2.1 Anatomy of a nested sampling run

A typical nested sampling run begins by populating the parameter space with a number of 'live points', drawn from the prior, and evaluating the corresponding likelihood values of these points. At each iteration of the algorithm, the live point with the lowest likelihood is deleted and a new point is drawn from the prior, with the constraint that it must have a likelihood higher than that of the deleted point. As this process continues, the collection of live points compresses exponentially in the parameter space towards the peak of the likelihood function (Hu et al. 2023) until some termination condition is satisfied.

At the end of the run, we are left with the series of deleted points, termed 'dead points', each associated with a set of parameter values, θ_i, and a likelihood, $\mathcal{L}_i$ (see Figure 1), and each defining an iso-likelihood contour in the parameter space. Every dead point is also assigned a value X_i, defined as the fractional prior volume contained within the iso-likelihood contour defined by the point:

$$X(\mathcal{L}_i) \equiv \int_{\mathcal{L}(\theta)>\mathcal{L}_i} \pi(\theta) d\theta. \tag{1}$$

By construction, X runs from 1 to 0. The exact prior volumes are unknown, but are modelled statistically using a set of shrinkage ratios, $\mathbf{t}$, where $X_i = t_i X_{i-1}$. t_i is defined as the largest of N random numbers drawn from a uniform distribution between zero and unity (Skilling 2006) and, thus,

$$P(t_i) = N t_i^{N-1}. \tag{2}$$

2.2 Evidence estimation

In the subsequent sections of the background, we will describe the different sources of error present in evidence estimation and parameter estimation from a nested sampling run. Much of this argument follows (Higson et al. 2018a).

Bayes' theorem tells us that the evidence may be computed by an integral over our parameters as

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta) d\theta. \tag{3}$$

Nested sampling enables the conversion of this many-dimensional integral into a one-dimensional integral, by changing the integration variable to the fractional prior volume within an iso-likelihood contour, X:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX. \tag{4}$$

In a typical nested sampling run, we only have access to information about a single point on each likelihood contour, not the entire contour itself. So, in practice, this integral is approximated by a sum over the dead points:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX \approx \sum_{i \in \text{dead points}} \mathcal{L}_i \Delta X_i. \tag{5}$$

Along a given contour, by construction, all points would have the same likelihood value, so using the likelihood value of a single dead point as a proxy for $\mathcal{L}(X)$ is an exact substitution that introduces no extra error. However, the exact values of the fractional prior volume 'shells', ΔX_i, are unknown, since the exact set of shrinkage ratios, $\mathbf{t}_i$, are unknown, and this is the dominant source of error in evidence estimation. Typically, the resulting error bar on the evidence is quantified by simulating sets of ΔX_i to use in the evidence calculation and quoting the error from this set of estimates.

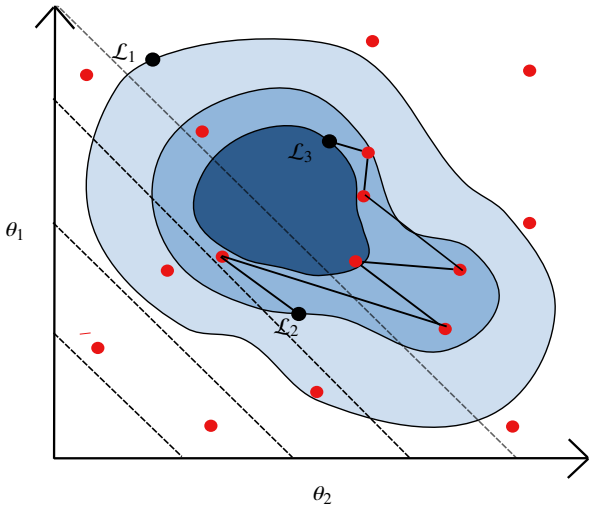

Figure 1. Schematic of a typical nested sampling run with a chain based sampler. At the end of the run, we are left with a set of dead points (black), which define a series of nested iso-likelihood contours. To generate a new live point from a given point, chain based samplers use a Markov-Chain based procedure to continually generate points within its likelihood contour, until the new point is deemed uncorrelated enough with the original point from which it was seeded. Thus, at the end of the run we are also left a set of 'phantom points' (red), the exact number of which depends on the chain lengths. An example chain is shown in black between dead points 2 and 3. In parameter estimation, we typically only use the dead points and, therefore, must use the parameter values of each dead point as a proxy for the average parameter value along the entire contour. For an example two-parameter case, where the parameter being estimated is the sum $f(\theta) = \theta_1 + \theta_2$, the contours of constant of constant $f(\theta)$ are shown (dashed). In this case, the parameter value of the dead points is not necessarily representative of the average parameter value over the contours, and this will be the dominant source of error in our $f(\theta)$ estimate. However, the phantom points can provide a better understanding of the variation of this parameter along the contours, enabling a more accurate quantification of this error.

2.3 Parameter estimation

To calculate the expected value of some function, $f(\theta)$, of the parameters, we must integrate the function over the posterior (Chopin & Robert 2010):

$$E[f(\theta)] = \int f(\theta) \frac{\mathcal{L}(\theta)\pi(\theta)}{\mathcal{Z}} d\theta = \frac{1}{\mathcal{Z}} \int \tilde{f}(X)\mathcal{L}(X) dX. \tag{6}$$

$\tilde{f}(X)$ represents the prior expectation of $f(\theta)$ given than $\mathcal{L}(\theta) = \mathcal{L}(X)$:

$$\tilde{f}(X) = E^{\pi}[f(\theta)|\mathcal{L}(\theta) = \mathcal{L}(X)]. \tag{7}$$

In order words, it can be seen as the average value of $f(\theta)$ over the given iso-likelihood contour. Discretising 6, as before, gives:

$$\frac{1}{\mathcal{Z}} \int \tilde{f}(X)\mathcal{L}(X) dX \approx \frac{1}{\mathcal{Z}} \sum_i \tilde{f}(X_i)\mathcal{L}_i \Delta X_i. \tag{8}$$

Again, the exact prior volumes are unknown, but there is an additional source of error here which was not present before: given that we only have information about a single $f(\theta)$ value on each contour, we are required to use $f(\theta_i)$ as a proxy for $\tilde{f}(X_i)$. Unlike with the likelihood, it is no longer true that this is an exact substitution, and in parameter estimation this becomes the dominant source of error. The stochasticity of nested sampling means that over a single run we will not necessarily be able to get representative values of $f(\theta)$ along every contour, and thus won't be able to capture our true error bar over multiple runs, even when simulating sets of prior volumes. This is demonstrated in Figure 1, where we show the example of trying to estimate the sum, $f(\theta) = \theta_1 + \theta_2$, of parameters in a simple two-parameter case; here, the values of $f(\theta)$ at each of the dead points on the contours do not reflect the average value over the contours. The key to capturing this error is better understanding the variation of $f(\theta)$ along the contours.

3 CORRECTING PARAMETER ESTIMATION ERRORS WITH PHANTOM POINTS

In order to study the nested sampling parameter estimation error, we consider the example of a simulated black hole binary (BBH) signal with parameters similar to those of GW150914 and with similar noise. The exact injected parameter values are shown in Figure 2. Since studying the error bars carefully requires hundreds of nested sampling runs to be performed, we work within a simplified framework in which only four of the parameters are sampled: the chirp mass ($\mathcal{M}$), mass ratio (q), luminosity distance (d_L), and zenith angle between the total angular momentum and line of sight (θ_{jn}). For all of the below results, we perform nested sampling runs with 500 live points per run and the default PolyChord chain length of $5 * \text{ndims} = 20$. The runs are performed using `bilby` (Ashton et al. 2019), with a modified version of the in-built PolyChord sampler option. All waveforms are both injected and sampled with the `IMRPhenomPv2` waveform model (Hannam et al. 2014) and we use two interferometers, LIGO-Hanford (H1) and LIGO-Livingston (L1).

3.1 Evidence error estimation method is insufficient for parameter estimation

The stochasticity of the fractional prior volumes in nested sampling means that the estimated mean evidence will vary from run to run. In order to accurately quantify the error on a single run, therefore, the error bar must reflect this run-to-run variation. For evidence estimation, this is typically achieved by sampling many sets of the shrinkage ratios, $\mathbf{t}$, from a known distribution and using these to calculate the evidences, quoting the error bar from the spread of these estimates. Higson et al. (2018a) term this the 'simulated weights method' and we shall refer to it henceforth as such too.

In the context of our simulated BBH signal, we can carefully verify that this method does indeed quantify the evidence error correctly. To do this, we first perform 100 nested sampling runs on our simulated signal. For each run, we apply the simulated weights method using `anesthetic` (Handley 2019) to calculate a set of evidence estimates. The resulting distributions of evidences are plotted in Figure 3. The dashed line represents the overall mean evidence calculated from all 100 runs.

The errors should be normally distributed, meaning that about 68% of the time our 'true evidence' (estimated from the mean of 100 runs) should lie within the error bars, and 95% of the time it should lie within two times the extent of the error bars. This condition can also be viewed in terms of the distribution of the percentiles of the 'true evidence'. If the errors are indeed correctly distributed, a histogram of these percentiles will yield a uniform distribution from 0 to 100. For our simulated BBH example, this histogram is plotted in Figure 3, along with the Kolmogorov-Smirnov (K-S) test p-value for a uniform distribution. We can conclude from this that the simulated weights method leads to correct evidence error estimates from a single run.

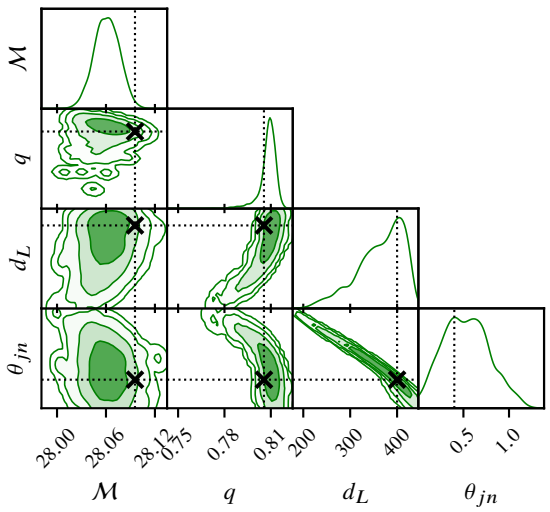

Figure 2. Posteriors recovered from the injected signal using PolyChord, with the injected parameter values indicated. Since parameter estimation studies require large numbers of runs, we only sample these 4 parameters, with the other parameters simply set to their injected values.

Skilling (2006) suggests using the same method to estimate parameter estimation errors. As an example, we shall attempt to estimate the chirp mass of our signal. As before, for each of the 100 runs, we can simulate sets of shrinkage ratios to obtains sets of ΔX and substitute these into equation 8 to calculate a set of chirp mass estimates. These are plotted in Figure 3, along with the overall mean chirp mass estimated from all 100 runs.

This time, even by eye it seems that the resulting distributions of chirp masses per run are not quite wide enough to capture the run-to-run variation. Performing a similar analysis of the percentiles as for the evidences yields the histogram in Figure 3. It is unmistakable from both the plot and the K-S test p-value that the simulated weights method alone does not lead to correct error estimates on the chirp mass. Far more often than ought to be the case, the mean chirp mass over many runs lies deep in the tails of the estimates from a single run. This is an empirical verification, in the context of gravitational waves, of results already established theoretically by Chopin & Robert (2010) and experimentally by Higson et al. (2018a).

3.2 Phantom points

In nested sampling, there are many ways to generate a new live point subject to the hard likelihood constraint $\mathcal{L} > \mathcal{L}_i$, and this is a key point of difference between existing nested sampling implementations. However, many implementations use a Markov-Chain based procedure, where new points are continually generated within the likelihood contour until we are satisfied that the new point is independent from the deleted point from which it was seeded. This point is then assigned as the new live point, and the points generated in the

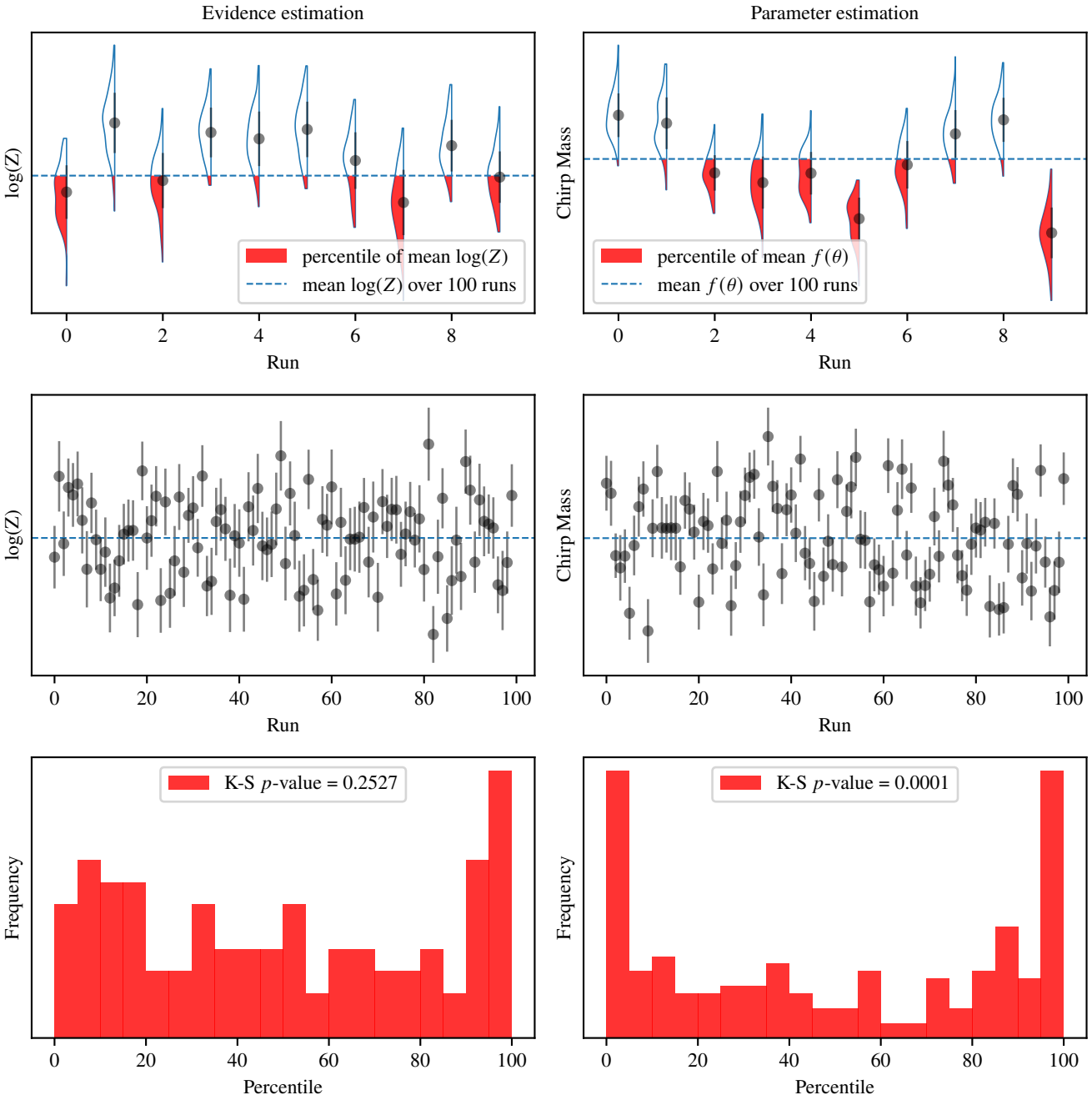

Figure 3. Typically, a distribution of evidence estimates is computed from a single run using the 'simulated weights' method (top left). The mean and standard deviation of this distribution are then quoted as the evidence value and its corresponding error (black). If the error bars from single runs accurately quantify the error on the evidence due to unknown prior volumes, the overall mean evidence, computed over many runs, should lie in the 1st percentile of estimates 1% of the time, in the 50th percentile 50% of the time, and so on. Hence, plotting a histogram of the percentiles from each run in which the overall mean (blue dashed line) lies should give a uniform distribution from 0 to 100 (bottom left). We see here that this is indeed the case for the evidences, with the Kolmogorov-Smirnov p-value being 0.25, demonstrating that resampling the prior volumes is sufficient for estimating the error bar on the evidence for this example. For the chirp mass, even by eye the estimates per run are not wide enough to capture the variation in estimates between runs (top and middle right). This indicates that the variation in chirp mass along a given contour is not being properly accounted for. The percentiles plot (bottom right) confirms that the error estimate on a single nested sampling run is too optimistic. More often than would be the case for correctly distributed errors, the overall mean estimate for the chirp mass over 100 runs (blue dashed line) lies deep in the tails of the estimated chirp mass from a single run. The simulated weights method is not sufficient for calculating the parameter estimation error.

chain between the deleted and new live point are typically discarded. These points are shown in red in Figure 1, with an example chain between a deleted and new live point illustrated by the black lines.

Though deemed too correlated to the deleted point to use as our new live point, these discarded points still have the potential to provide useful information about the parameter space, though this potential has been largely unexplored. For the remainder of the paper we shall refer to these 'intra-chain' points as 'phantom points', following the terminology of PolyChord (Handley et al. 2015b,a), the nested sampling implementation we use to obtain the results for this paper; tailored for high-dimensional parameter spaces and parallelised using openMPI, it is particularly suitable for gravitational wave parameter inference studies. However, we emphasise that the existence of these phantom points is not unique to PolyChord and can be found in any chain based sampler.

The key piece lacking from the simulated weights method is that it uses the single value of $f(\theta)$ available for each iso-likelihood contour as a proxy for the average $f(\theta)$ over the contour, since there is no extra information in the live points to do otherwise. However, phantom points, drawn from the likelihood-constrained prior and each with a corresponding likelihood evaluation, have the potential to provide this additional information about the variation of $f(\theta)$ along contours. Below, we present two novel error estimation methods which incorporate this, and demonstrate that they result in the correct error distributions.

3.3 Likelihood binning method

In this first approach, we bin the phantom points from the run by their likelihood values, such that each phantom point is assigned to the dead point to which it is closest in log-likelihood. In this way, each dead point is now associated with a set of points which sit very close to the contour defined by it; we make the assumption that, though the phantom points do not lie exactly on the dead point's iso-likelihood contour, they are still representative of the distribution of the $f(\theta)$ values along the contour. Then, for each dead point in our sum in equation 8, we resample a new $f(\theta)$ value from the associated bin, which includes the original dead point itself, as well as sampling a new ΔX. The error estimate is then obtained by repeating this process many times and taking the standard deviation of the resulting distribution of estimates.

In reality, the phantom points are slightly correlated, which can lead to biased results due to the points not independently populating the prior. However, the large majority of the correlation can be removed by only using every, for example, 5^{th} point in the chains, as the correlation in the chirp mass becomes negligible after a few steps. The results for the chirp mass are shown in Figure 4. As demonstrated by the K-S test p-value and the histogram, the distribution of percentiles is now much more consistent with a uniform distribution. This shows that the error bars now accurately capture the variation in the chirp mass over the contours, and reflect the run-to-run differences from nested sampling due to this additional stochasticity.

3.4 Reconstructed runs method

In order to explore the second method, we first note that there is nothing wrong with any of the phantom points in terms of their suitability for use in a nested sampling run, except that we have already chosen to use another point (the associated dead point) which may be too correlated with the phantom point to use both. Hence, we may take the 1^{st} phantom point in every chain in the run and combine

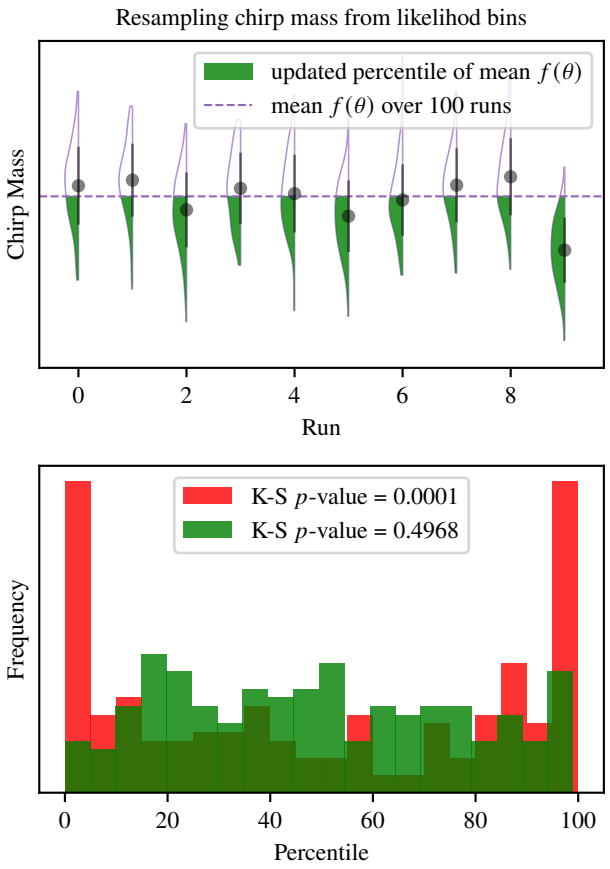

Figure 4. For each of the first 10 nested sampling runs performed on our simulated dataset, the new distribution of chirp mass estimates are plotted, obtained from resampling both shrinkage ratios and chirp mass values from likelihood bins around each contour (top). Resampling both the weights and the chirp mass gives wider error bars, as expected, and even by eye these seem more consistent with the spread of estimates across runs. Testing the likelihood binning method more rigorously, we see that the percentiles of the overall mean chirp mass are now consistent with being uniformly distributed (bottom, green). This indicates that the true variation of the chirp mass along the contours is now being correctly accounted for, and thus the stochasticity of nested sampling parameter estimation is properly captured.

these carefully to form an equally valid nested sampling run to the original. Thus, from every run we are able to reconstruct multiple equally valid 'phantom runs'. It is true that some of these runs will be correlated to each other, but, as before, this correlation can be largely removed by only using a subset of the phantom points.

These extra phantom runs are akin to performing multiple nested sampling runs on the same dataset, but, crucially for gravitational wave inference, at no extra computational cost. The parameter estimates from each of these reconstructed runs, as well as the original run, can then be combined and the corresponding new error bar may be computed from this. The resulting percentiles plot is shown in Figure 5, where it can be seen that, as with the first method, the new error bars are now much more consistent with the spread of estimates across multiple runs.

3.5 Verifying accuracy of error bar

For certain parameters, the chain length of the sampler may not be long enough to accumulate enough uncorrelated phantom points that

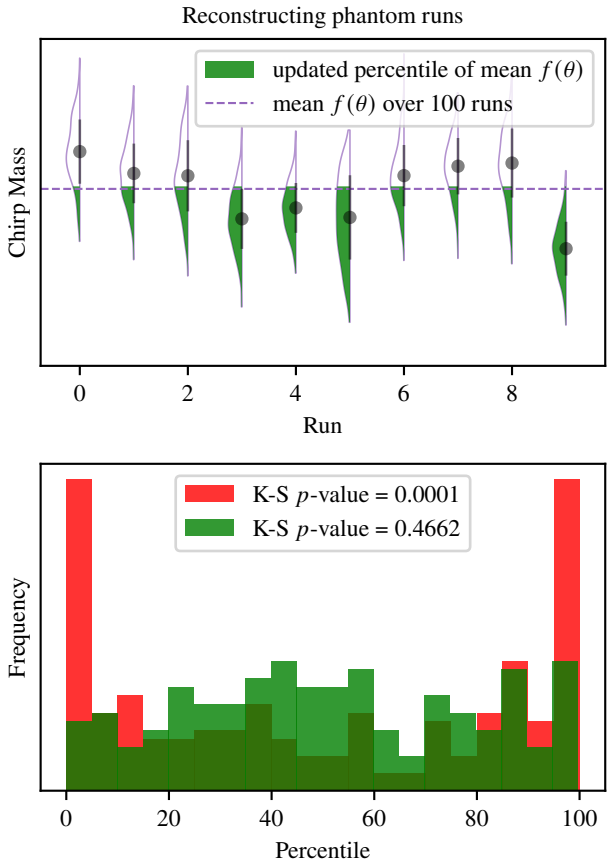

Figure 5. As with the likelihood binning method, computing parameter estimates from the reconstructed phantom runs gives correctly distributed percentiles. Again, we capture the stochasticity of nested sampling parameter estimation without additional computational cost.

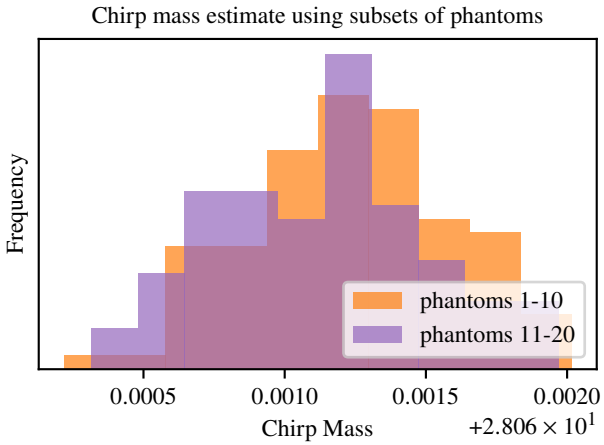

Figure 6. The chirp mass was estimated from a single run using the likelihood binning method described in Section 3.3. Two sets of estimates were produced, one using only the first ten phantom points in all the chains and the other using only the last ten. The resulting distributions are in agreement with each other, with the K-S 2-sample test statistic being 0.14, with a p-value of 0.28. This indicates that the phantom points were well distributed in the chirp mass and give an accurate estimate of the true error bar.

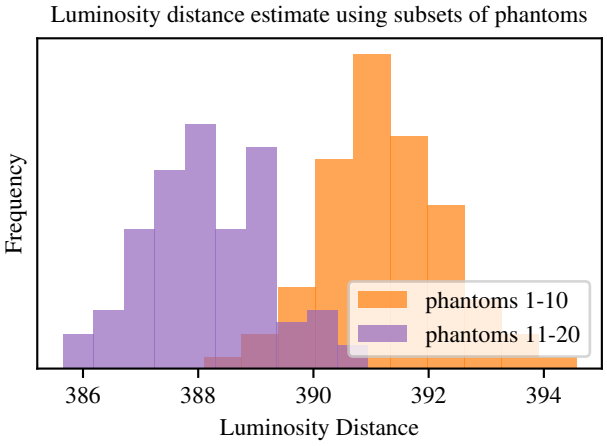

Figure 7. The same analysis was repeated for the luminosity distance, a parameter on which our methods do not perform well with the default PolyChord chain length. This time the K-S 2-sample test statistic was 0.82, with a p-value of 4.11×10^{-34}. It is likely that the phantom points are still a little too correlated to their associated dead point in luminosity distance, and a longer chain length is needed as more of the points in the chain must be discarded.

populate the prior in the correct way. In this case, the new error bars from the above methods, though certainly wider than those from the usual method, may still not be wide enough to capture the full variation of the parameter along the contours. For gravitational waves, parameters like the mass ratio and the luminosity distance may suffer from this. If one is able to perform many nested sampling runs on the dataset, the error bars can be checked rigorously in a manner similar to the percentiles analysis above. However, this is inefficient and in many cases simply not viable. We need to be able to know from a single run whether the new error bars are now fully correct for a given parameter. Below, we present a method to check for convergent results of the parameter estimates, indicating whether the error bars have been quantified correctly.

For the default chain length in PolyChord the phantom points are very well distributed in their chirp mass values in order to be able to obtain an accurate error bar for this parameter. But they are more correlated in their luminosity distance values, meaning that the error bars obtained for this parameter will still be an underestimate of the true error. We need to employ a longer chain length in the run in order to obtain sufficient usable phantom points for the two methods described above. We can check whether there are enough uncorrelated phantom points for a given parameter by dividing the points into two halves, according to their position within the chains. We can then take each half separately and use either the likelihood binning method or the reconstructed runs method to obtain a set of parameter estimates, checking whether the two sets of estimates are in agreement using the Kolmogorov-Smirnov 2-sample test.

This is demonstrated in Figures 6 and 7. For the chirp mass, it can be seen that the estimates obtained from using only the first ten phantom points in the chain agree well with the estimates from the last ten phantom points in the chain. The same cannot be said for the luminosity distance, indicating that we may need a chain length longer than 20 in order to obtain enough uncorrelated phantom points to estimate the error on this parameter. It is important to note here that this does not mean the chain length of the run wasn't long enough to produce uncorrelated posterior samples in the luminosity distance, only that it was not long enough to produce sufficient suitable phantom points.

To examine this further, we performed a run with a chain length of

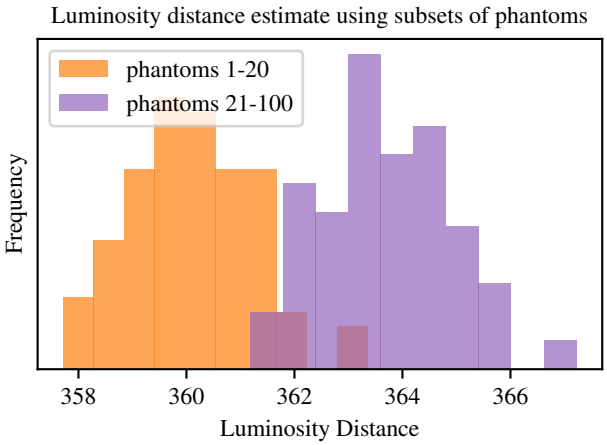

Figure 8. For a run with a chain length of 100, the luminosity distance estimates from the first 20 points in the chain are compared to those from the rest of the phantom points. The K-S 2-sample test statistic for the resulting set of estimates is 0.79, with an accompanying p-value of 3.05×10^{-31}. This indicates that the first 20 phantom points are not well distributed enough to make an accurate estimate of the luminosity distance error bars and more points are needed.

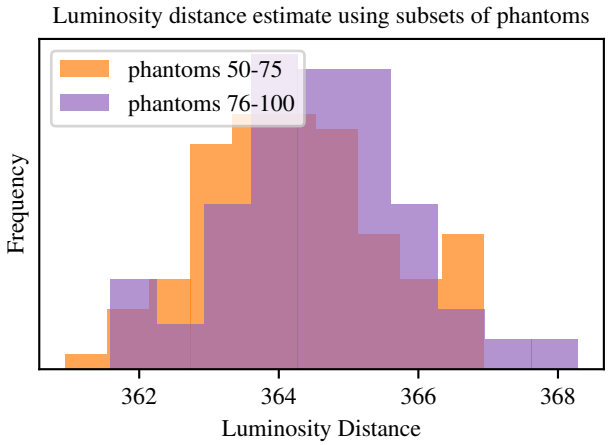

Figure 9. The same analysis is repeated, but comparing estimates obtained from the phantom points in the chains between positions 50 and 75 and the estimates from the just the last 25 points in each chain. These two sets of estimates now show good agreement, with a test statistic of 0.19 and a p-value of 0.05. Using the last 50 phantom points in the chains for either method described in this paper would therefore give an accurate estimate of the true luminosity distance error bar.

100, to see how many phantom points must be discarded before we have enough uncorrelated points to provide convergent estimates of the luminosity distance. First, the estimates obtained from the first 20 phantom points in the chain are compared to the those from the last 80 points, shown in Figure 8. The two sets of estimates are very different and this is a clear indication that the original chain length of 20 was insufficient to accurately estimate the luminosity distance error bars. Next, the estimates obtained from phantoms points at positions in the chains between 50 and 75 were compared to estimates from the last 25 points in the chains. The two sets of estimates in Figure 9 are now in agreement and show that the last 50 phantom points in the chain may be used in the likelihood binning or reconstructed runs method to provide an accurate estimate of the error bars.

4 COMPARISON TO HIGSON METHOD

In order to compare the methods presented above to those obtained using the bootstraps method presented in Higson et al. (2018a), we use the `nestcheck` (Higson et al. 2018c) package to apply the Higson method on the same set of simulated BBH signals. In this method, a nested sampling run is split up into single live point runs, and then n_{live} runs are sampled with replacement from these. This is repeated many times to build up a distribution of parameter estimates from which the error bar is estimated.

Figure 10 shows the resulting percentiles plot for the chirp mass; the Higson method performs similarly to the phantom methods and provides an independent way to correctly estimate the error bars. On parameters where the phantom methods do not work as well for the default settings, such as the luminosity distance, the Higson method also does not produce wide enough error bars. The advantage of using phantom points is that they can confirm from a single run whether or not the error bars are correct for a given parameter, as described in the previous section.

The phantom methods are distinct from the Higson method, and so they can be used in combination. Using phantom points and single live point run bootstrapping together may also help to obtain the correct error bars for parameters like the luminosity distance without having to run a longer chain.

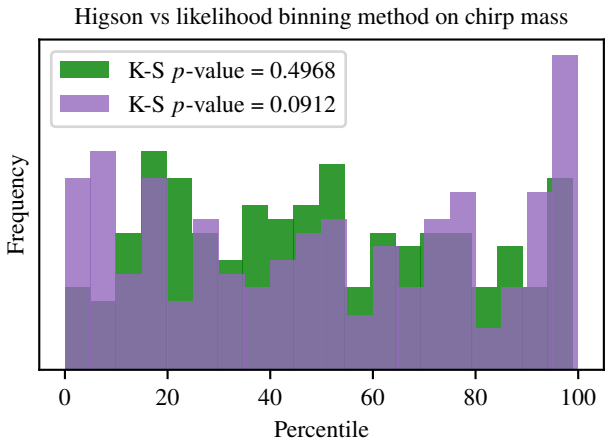

Figure 10. The Higson method (purple) and the methods presented in this paper (green) perform similarly well on the simulated BBH signals, producing percentile plots that are consistent with a uniform distribution. Both are a significant improvement on the usual 'simulated weights' method, and are distinct in their approaches so may be used in conjunction with each other.

5 COVERAGE AND CREDIBLE INTERVALS

The additional error in nested sampling parameter estimation described in Section 3 affects not only our estimates of parameter means, but also the credible intervals. If a parameter inference procedure is accurate, we expect that the percentiles of the true parameters across many analysed datasets should be uniformly distributed (Talts et al. 2020; Samantha R Cook & Rubin 2006). The tool typically used to check this is a p-p plot (Morisaki & Raymond 2020; Veitch et al. 2015b; Pankow et al. 2015; Smith et al. 2020; Berry et al. 2015;

Del Pozzo et al. 2018; Romero-Shaw et al. 2020; Biwer et al. 2019; Sidery et al. 2014; Green & Gair 2020), constructed from N runs on injected signals where the true parameters are known.

A p-p plot shows the fraction of events for which the true parameter values lie within a given credible interval as a function of the credible interval. In the ideal case where the parameter inference procedure leads to the correct coverage, this plot should be a diagonal line. There is a statistical error that arises from the finite number, N, of datasets analysed, often indicated by a gray region around the ideal diagonal line (as in Figures 11 and 12), and this is usually the dominant source of error in p-p plot analyses. However, as N increases, the nested sampling error becomes important. This error can be evaluated using either the likelihood binning or reconstructed runs method, and we can place a 'nested sampling confidence interval' on our p-p plot to account for the variation of the calculated percentiles from run to run on the same set of events.

Figures 11 and 12 were produced using the likelihood binning method. For each posterior sample from a given event, the individual parameter values were resampled from bins of neighbouring phantom points. These resampled parameter values were then used to calculate a set of credible intervals for that event, from which the error on the credible interval was estimated. In Figure 11, a total of 200 injected signals were analysed, with the injection parameters drawn from the prior, and the gray region shows the 3-σ (99.7%) confidence interval for $N = 200$ due to a finite event sample size. The purple region shows the corresponding 3-σ confidence interval on the p-p curve due to the nested sampling parameter estimation error. In scenarios where the p-p plot lies outside of the gray region, using the nested sampling error bars may help to assess how much of the deviation is simply due to the stochastic nature of nested sampling. Figure 12 is produced from 1000 injected signals, where now the statistical error for $N = 1000$ and the nested sampling parameter estimation error are of similar sizes. Here, it is especially useful to consider the latter error in assessing to what extent the p-p plot displays the expected coverage. Ideally, the gray confidence intervals should include both sources of error.

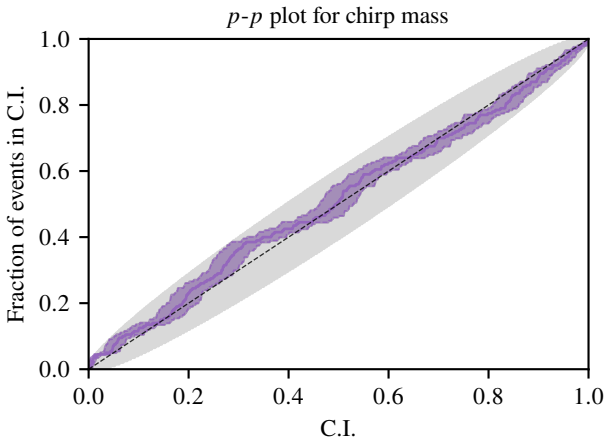

Figure 11. The p-p plot obtained for the chirp mass using `bilby` and PolyChord is shown. As with the parameter mean estimates, we only sample over four parameters, simply setting the rest to the injected value. The plot shows the cumulative distribution function (CDF) of the percentiles of the true chirp mass for 200 events, calculated in the usual way, with the associated 3-σ confidence interval due to the statistical error from the finite number of events (gray). The purple region shows the 3-σ confidence interval on the calculated CDF of the percentiles due to the stochastic nature of nested sampling. The binomial error still dominates, but the nested sampling parameter estimation error can be useful in assessing the source of deviations from the expected distribution.

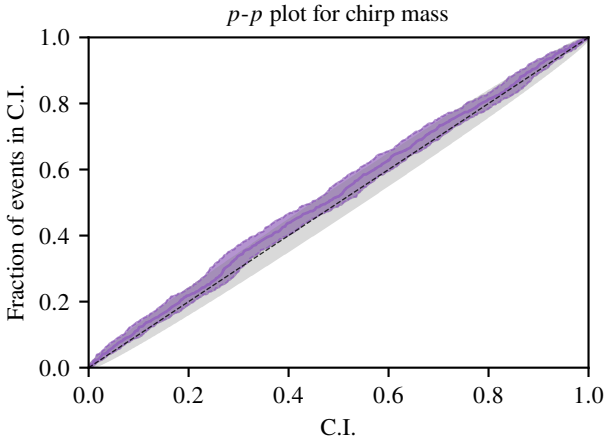

Figure 12. The CDF of the percentiles of the true chirp mass is shown for 1000 events, along with the 3-σ confidence interval due to a finite event sample size (gray). The 3-σ confidence interval due to the nested sampling parameter estimation error is also shown (purple). These are now of a comparable size, and in order to make an accurate assessment of the coverage, it is necessary to account for both sources of error.

6 CONCLUSIONS

In nested sampling, the evidence error bar is dominated by the unknown exact prior volumes of the 'shells' between iso-likelihood contours. This error affects parameter estimation too, but here there is an additional source of error, from using the parameter value of a single sample, $f(\theta_i)$, as a proxy for the average parameter value over the entire iso-likelihood contour defined by that sample, $\tilde{f}(X_i)$ (Chopin & Robert 2010). Though the prior volume error is well understood and accurately estimated from the 'simulated weights method', the added stochasticity due to parameter variation over a given contour is typically ignored, despite the fact that this can be a significant source of error in the analysis of any dataset, including gravitational wave signals (Thrane & Talbot 2019).

Here we have proposed two novel approaches to account for this additional error using the extra likelihood calls made at run-time by any chain based sampler. Though these 'phantom points' are not suitable for use in evidence estimation, they are valuable in parameter inference for exploring the variation in a parameter value over individual contours. This is the first demonstration of the use of nested sampling phantom points for inference. We have also shown how to use phantom points to verify the accuracy of the error bars on a given parameter, by splitting the set of the phantom points in two and checking for convergence in the resulting estimates. This could be a useful approach at runtime in tuning the chain length of chain based samplers to speed up nested sampling whilst still ensuring independent samples. In other MCMC procedures, such as Hamiltonian Monte Carlo (Carpenter et al. 2017; Cabezas et al. 2024), the emphasis is placed on tracking convergence on specific parameters of interest, and dynamic nested sampling (Higson et al. 2018b; Speagle 2020) provides a way to apply this in nested sampling too.

Furthermore, we have shown how this error impacts estimates of

credible intervals and coverage plots, such as the p-p plot commonly used in gravitational wave analyses. Though this source of error is often secondary to the statistical error due to the finite number of events sampled for a typical coverage plot, the two become comparable in size for larger numbers of analysed events. This highlights the importance of accounting for the additional nested sampling parameter estimation error when interpreting coverage plots.

The methods described in this paper differ from the `nestcheck` methods proposed in Higson et al. (2018a) in that we incorporate information from the chain based phantom points and do not make use of bootstraps, but they produce comparable results and should be viewed as complementary. They can be combined and used together. This may be particularly helpful for parameters where the original chain length was not long enough to produce sufficiently many uncorrelated phantom points for use with the methods in this paper on their own, and may avoid the need to repeat the run with a longer chain. The advantage of incorporating phantom points is that they can be used to assess whether the true run-to-run variation has been captured from a single run, in contrast with bootstrapping.

Though phantom points are specific to chain based samplers, other samplers may generate discarded but valid points. These points also would not be suitable for evidence estimation but may be of use for other purposes such as parameter inference. Both importance sampling methods (Williams et al. 2021, 2023; Lange 2023) and methods using machine learning proxies to accelerate nested sampling (Graff et al. 2012; Lange 2023) generate samples that are not used, but may provide additional information about the parameter space. Likewise, MultiNest (Feroz et al. 2009) also produces discarded points. We hope that this work will renew interest within the community (Ashton et al. 2022) in methods for quantifying nested sampling parameter estimation.

ACKNOWLEDGEMENTS

MP was supported by the Harding Distinguished Postgraduate Scholars Programme (HDPSP). WH was supported by a Royal Society University Research Fellowship. This work was performed using the Cambridge Service for Data Driven Discovery (CSD3), part of which is operated by the University of Cambridge Research Computing on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The DiRAC component of CSD3 was funded by BEIS capital funding via STFC capital grants ST/P002307/1 and ST/R002452/1 and STFC operations grant ST/R00689X/1. DiRAC is part of the National e-Infrastructure.

DATA AVAILABILITY

All the data used in this analysis, including the relevant nested sampling dataframes, can be obtained from Prathaban & Handley (2024). We also include a notebook with all the code to reproduce the plots in this paper.

REFERENCES

Abbott B. P., et al. 2017a, Phys. Rev. Lett., 119, 161101
Abbott B. P., et al. 2017b, Nature, 551, 85–88
Ashton G., Talbot C., 2021, Monthly Notices of the Royal Astronomical Society, 507, 2037–2051
Ashton G., et al., 2019, The Astrophysical Journal Supplement Series, 241, 27
Ashton G., et al., 2022, Nature Reviews Methods Primers, 2
Berry C. P. L., et al., 2015, The Astrophysical Journal, 804, 114
Biwer C. M., Capano C. D., De S., Cabero M., Brown D. A., Nitz A. H., Raymond V., 2019, Publications of the Astronomical Society of the Pacific, 131, 024503
Buchner J., 2023, Statistics Surveys, 17
Cabezas A., Corenflos A., Lao J., Louf R., 2024, BlackJAX: Composable Bayesian inference in JAX (arXiv:2402.10797)
Carpenter B., et al., 2017, Journal of Statistical Software, 76, 1–32
Chopin N., Robert C. P., 2010, Biometrika, 97, 741
Del Pozzo W., Berry C. P. L., Ghosh A., Haines T. S. F., Singer L. P., Vecchio A., 2018, Monthly Notices of the Royal Astronomical Society
Feroz F., Hobson M. P., Bridges M., 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601–1614
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, Publications of the Astronomical Society of the Pacific, 125, 306–312
Gair J. R., Feroz F., Babak S., Graff P., Hobson M. P., Petiteau A., Porter E. K., 2010, J. Phys. Conf. Ser., 228, 012010
Gerosa D., Berti E., 2017, Phys. Rev. D, 95, 124046
Graff P., Feroz F., Hobson M. P., Lasenby A., 2012, Monthly Notices of the Royal Astronomical Society, pp no–no
Green S. R., Gair J., 2020, Complete parameter inference for GW150914 using deep learning (arXiv:2008.03312)
Handley W., 2019, Journal of Open Source Software, 4, 1414
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, Monthly Notices of the Royal Astronomical Society: Letters, 450, L61–L65
Handley W. J., Hobson M. P., Lasenby A. N., 2015b, Monthly Notices of the Royal Astronomical Society, 453, 4384
Hannam M., Schmidt P., Bohé A., Haegel L., Husa S., Ohme F., Pratten G., Pürrer M., 2014, Phys. Rev. Lett., 113, 151101
Higson E., Handley W., Hobson M., Lasenby A., 2018a, Bayesian Analysis, 13
Higson E., Handley W., Hobson M., Lasenby A., 2018b, Statistics and Computing, 29, 891–913
Higson E., Handley W., Hobson M., Lasenby A., 2018c, Monthly Notices of the Royal Astronomical Society, 483, 2044–2056
Hu Z., Baryshnikov A., Handley W., 2023, aeons: approximating the end of nested sampling (arXiv:2312.00294)
Lange J. U., 2023, NAUTILUS: boosting Bayesian importance nested sampling with deep learning (arXiv:2306.16923)
MacKay D. J. C., 2003, Information Theory, Inference, and Learning Algorithms. Copyright Cambridge University Press
Mapelli M., 2021, Formation Channels of Single and Binary Stellar-Mass Black Holes. Springer Singapore, p. 1–65, doi:10.1007/978-981-15-4702-7_16-1, http://dx.doi.org/10.1007/978-981-15-4702-7_16-1
Morisaki S., Raymond V., 2020, Physical Review D, 102
Pankow C., Brady P., Ochsner E., O'Shaughnessy R., 2015, Phys. Rev. D, 92, 023002
Prathaban M., Handley W., 2024, Costless correction of chain based nested sampling parameter estimation in gravitational wave data and beyond, doi:10.5281/zenodo.10911044, https://doi.org/10.5281/zenodo.10911044
Romero-Shaw I. M., et al., 2020, Monthly Notices of the Royal Astronomical Society, 499, 3295–3319
Samantha R Cook A. G., Rubin D. B., 2006, Journal of Computational and Graphical Statistics, 15, 675
Schutz B. F., 1986, Nature, 323, 310
Sidery T., et al., 2014, Physical Review D, 89
Skilling J., 2006, Bayesian Analysis, 1, 833
Smith R. J. E., Ashton G., Vajpeyi A., Talbot C., 2020, Monthly Notices of the Royal Astronomical Society, 498, 4492–4502
Speagle J. S., 2020, Monthly Notices of the Royal Astronomical Society, 493, 3132–3158
Talbot C., Thrane E., 2017, Physical Review D, 96
Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2020, Validating Bayesian Inference Algorithms with Simulation-Based Calibration (arXiv:1804.06788)

Thrane E., Talbot C., 2019, Publications of the Astronomical Society of Australia, 36
Veitch J., Vecchio A., 2010, Physical Review D, 81
Veitch J., et al., 2015a, Phys. Rev. D, 91, 042003
Veitch J., et al., 2015b, Phys. Rev. D, 91, 042003
Williams M. J., Veitch J., Messenger C., 2021, Phys. Rev. D, 103, 103006
Williams M. J., Veitch J., Messenger C., 2023, Machine Learning: Science and Technology, 4, 035011

This paper has been typeset from a TEX/LATEX file prepared by the author.