

Gravitational-wave astronomy with an uncertain noise power spectral density

Colm Talbot^{1,2,3,*} and Eric Thrane^{2,3}

¹*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

²*Monash Centre for Astrophysics, School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia*

³*OzGrav: ARC Centre of Excellence for Gravitational-Wave Discovery, Clayton, VIC 3800, Australia*



(Received 16 June 2020; revised 18 September 2020; accepted 3 November 2020; published 1 December 2020)

In order to extract information about the properties of compact binaries, we must estimate the noise power spectral density of gravitational-wave data, which depends on the properties of the gravitational-wave detector. In practice, it is not possible to know this perfectly, only to estimate it from the data. Multiple estimation methods are commonly used, and each has a corresponding statistical uncertainty. However, this uncertainty is widely ignored when measuring the physical parameters describing compact binary coalescences, and the appropriate likelihoods which account for the uncertainty are not well known. In order to perform increasingly precise astrophysical inference and model selection, it will be essential to account for this uncertainty. In this work, we derive the correct likelihood for one of the most widely used estimation methods in gravitational-wave transient analysis, the median average. We demonstrate that simulated Gaussian noise follows the predicted distributions. We then examine real gravitational-wave data at and around the time of GW151012, a relatively low-significance binary black hole merger event. We find that the difference in our inference when using different PSD estimation techniques is larger than the predicted statistical uncertainty.

DOI: [10.1103/PhysRevResearch.2.043298](https://doi.org/10.1103/PhysRevResearch.2.043298)

I. INTRODUCTION

The astrophysical parameters of compact binaries are inferred from gravitational-wave data using Bayesian inference. A crucial first step for Bayesian inference is to choose the appropriate likelihood for our data. In gravitational-wave transient data analysis this, typically, hinges on assumptions that the noise is Gaussian and stationary over the period being analyzed [1]. If these conditions are met, and if the noise power spectral density (PSD) were known exactly, the appropriate likelihood would be the Whittle likelihood [2]:

$$\mathcal{L}(\tilde{d}|\theta, P) = \frac{2}{\pi TP} \exp\left[-\frac{2|\tilde{d} - \tilde{\mu}(\theta)|^2}{TP}\right]. \quad (1)$$

Here P is the PSD, \tilde{d} is the frequency domain interferometer data, T is the duration of the data being analyzed, and $\tilde{\mu}$ is our model for the expected signal.

However, in practice, we do not have access to the true power spectral density of gravitational-wave detectors, and so we have to rely on an empirical estimate. There are two commonly used methods to compute these estimates. The simplest method is to average over the power in neighboring stretches of data. This method assumes that the PSD does not vary over the duration being averaged and that there are no

non-Gaussian features in the data. The other commonly used method is to simultaneously fit a parameterized model to any non-Gaussian features and the power spectral density (e.g., Ref. [3]). While this method does not require analyzing as much data and hence is less effected by nonstationarity and non-Gaussianity, it is far more computationally expensive. In this work, we are going to focus on the former.

To generate an average PSD we typically either compute the mean or median of neighboring segments. Taking the mean of neighboring segments is a commonly used method (sometimes referred to as the Welch or Blackwell method [4]) in gravitational-wave data analysis and many other signal processing applications. However, it is not widely used in gravitational-wave transient data analysis due to its sensitivity to non-Gaussian transients, “glitches,” in the detector noise. To mitigate the effect of these glitches a median average is instead used to compute the PSD as the median is more robust to the presence of large outliers. However, there may be effective methods to either remove or exclude these non-Gaussian features [5–11].

The other assumption underlying Eq. (1) is that the PSD does not change over time; in other words, the data are stationary. In practice, the PSD of real interferometers varies over the timescale of minutes, and so care must be taken when estimating the PSD using longer stretches of time [1,3,12]. Methods for mitigating this nonstationarity have also been considered previously [11]. In this paper, we ignore these possible effects and assume the data we look at are both Gaussian and stationary.

In Chatziioannou *et al.* [13], the authors compare Advanced LIGO data whitened with a median average PSD with parametric PSD estimation methods. They show that the data

*ctalbot@caltech.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

whitened using the median PSD does not follow a unit normal distribution. They argue that this difference is due to non-Gaussianity and nonstationarity in the data. However, data whitened using an average of a finite number of segments are known to follow a non-Normal distribution *even for Gaussian noise*. For a mean average, the whitened Gaussian-noise data follow a Student's t distribution, and the correct likelihood to use is the Student-Rayleigh distribution [14–17]. In this work, we demonstrate that data whitened with a median PSD estimate follows a different distribution, and we show how to marginalize over the uncertainty in this estimated PSD to obtain the correct likelihood for stationary, Gaussian noise.

Parametric PSD estimation using the `BayesLine` algorithm [3] marginalizes over a prior which models the PSD as a combination of a slowly varying spline and Lorentzians to fit sharp spectral features. This algorithm is generally combined with the `BayesWave` algorithm [18] to fit astrophysical and terrestrial transients simultaneously with the PSD. However, this does not allow direct inference of physical parameters describing compact binary coalescences, e.g., the masses and spins of merging black holes.

A common approach in compact binary data analysis is to take a median average of the posterior distribution for the PSD obtained using `BayesLine` rather than an averaged PSD (e.g., Ref. [19]). Recently, Biscoveanu *et al.* [20] introduced a method to marginalize over the uncertainty in these parametric PSDs estimates. However, this is done at significant computational cost requiring $\sim 200\times$ the computational resources as a standard analysis. Additionally, under the formalism presented there, it is not possible to compute the Bayesian evidences necessary to perform model comparison.

The remainder of this paper is structured as follows. In Sec. II we derive the appropriate distributions for the likelihood and whitened data after marginalizing over the uncertainty in a median (and/or mean) PSD estimate. We provide a brief introduction to Bayesian inference in the context of gravitational-wave astronomy in Sec. III. We then demonstrate the efficacy of our formalism by applying it to simulated Gaussian data in Sec. IV. Following this, in Sec. V we consider a case study using real Advanced LIGO data. We analyze the marginal gravitational-wave candidate GW151012 with both mean and median PSD estimates to understand the effect of marginalizing over the statistical uncertainty and of using the different estimation techniques. This event is convenient for our present purposes since the effects we seek to study are most prominent for marginal signals like GW151012. Some closing comments are then provided in Sec. VI.

II. FORMALISM

A. Gaussian noise

For stationary Gaussian noise $n(t)$, if we do not manipulate the data in any way before performing a Fourier transform, the noise covariance can be written in the frequency domain as

$$C(f, f') = \langle \tilde{n}(f)\tilde{n}^*(f') \rangle = P(f)\delta(f - f'). \quad (2)$$

The angle braces denote an ensemble average over realizations. In practice, we work with discrete Fourier transforms

and noise covariance matrices

$$C_{ij} = \langle \tilde{n}_i\tilde{n}_j^* \rangle = \frac{T}{4}P_i\delta_{ij} = \frac{T}{4}A_i\delta_{ij}A_j. \quad (3)$$

Here T is the duration over which the discrete Fourier transforms is performed, i, j index frequency bins, and $A_i = P_i^{1/2}$ is the noise amplitude spectral density (ASD).

For real data, a number of manipulations are performed before the data are Fourier transformed, which makes things more complicated. The data are band-passed and windowed in the time domain to prevent aliasing and spectral leakage [1]. As long as the frequency limits of the band-pass filters do not overlap with the frequency range of interest, the band-passing can be ignored. However, the window applied to the data must be considered. Since the window is multiplicative in the time domain, there is a corresponding convolution in the frequency domain,

$$C_{ij}^w = \langle (\tilde{n} * \tilde{W})_i^* (\tilde{n} * \tilde{W})_j \rangle = \frac{T}{4}A_i\mathcal{T}_{ij}A_j, \quad (4)$$

$$\mathcal{T}_{ij} = \begin{cases} \tilde{W}_{i-j} & i \geq j \\ \tilde{W}_{j-i}^* & i < j \end{cases}, \quad (5)$$

where \mathcal{T} is a Hermitian Toeplitz matrix and there is no implied summation over i or j . For a rectangular window $\mathcal{T}_{ij} = \delta_{ij}$ and the covariance matrix is diagonal. For generic windows, there is a regular, predictable, off-diagonal power. In reality, the effect of this is much smaller than the effects considered here, and inverting the covariance matrix is a significant computational challenge. We leave a detailed analysis of the effect of nonrectangular windows on parameter estimation and model selection to a future study.

The real and imaginary components of the frequency-domain noise follow a normal distribution with variance matrix P ,

$$p(\tilde{n}_i|P_i) = \sqrt{\frac{2}{\pi T P_i}} \exp\left(-\frac{2\tilde{n}_i^2}{T P_i}\right). \quad (6)$$

This is *not* the likelihood which we use when analyzing gravitational-wave transients as we need to simultaneously consider the real and imaginary components of the noise. The likelihood is given by

$$\mathcal{L}(\tilde{n}_i, \tilde{n}_j|C_{ij}) = \frac{2}{\pi T \det(C_{ij})} \exp\left(-\frac{1}{2}\tilde{n}_i C_{ij}^{-1} \tilde{n}_j^*\right). \quad (7)$$

Since we assume the covariance matrix is diagonal this is often written in the simplified form known as the Whittle likelihood,

$$\mathcal{L}(\tilde{n}_i|C_i) = \frac{2}{\pi T P_i} \exp\left(-\frac{2|\tilde{n}_i|^2}{T P_i}\right). \quad (8)$$

We note that this likelihood is normalized over the complex plane. It is convenient to reduce to one dimension for visualization purposes, so we note that the power of the noise $P_i = |\tilde{n}_i|^2$ follows an exponential distribution,

$$p(P_i|P_i) = \frac{2}{T P_i} \exp\left(-\frac{2P_i}{T P_i}\right). \quad (9)$$

All of the expressions above assume that there are no non-Gaussian signals in the data. In order to include signals we

simply make the substitution $n = d - \mu$ where d is the data and μ is the signal.

Time-domain windows affect the noise and signal components differently. We assume that the window is always applied such that the window does not cause any loss of signal power in the observing frequency band. In addition to the correlation between different frequency bins induced by the window, there is a net power loss in the Gaussian noise given by the mean square value of the window function. Care must be taken to consistently correct for this power loss to avoid biasing our inference (e.g., Ref. [21]).

Now that we have established which distributions we want to use when the PSD is known, we can address the distributions we want to use when the PSD is uncertain.

B. Median PSD estimate

The generic expression for the likelihood marginalized over uncertainty in an estimated PSD, \hat{P} , is

$$\mathcal{L}_P(\tilde{d}|\theta, \hat{P}) = \int_0^\infty dP \mathcal{L}(\tilde{d}|\theta, P)\pi(P|\hat{P}), \quad (10)$$

where $\mathcal{L}(\tilde{d}|\theta, P)$ is the likelihood of obtaining the data given model parameters θ and the true PSD P , as defined in Eq. (1), and $\pi(P|\hat{P})$ is our prior on the true PSD given the estimated PSD. Similarly, using Eq. (6), we can write down an expression for the expected distribution of whitened strain residuals, $\tilde{v} = \tilde{n}/\hat{P}^{1/2}$,

$$p_P(\tilde{v}|\hat{P}) = \int_0^\infty dP p(\tilde{n}|P)\pi(P|\hat{P}). \quad (11)$$

Here \tilde{n} is the frequency-domain data after removing any signals present.

First we need to define the estimated PSD

$$\hat{P} = \frac{\text{median}(P_\ell)}{\alpha}, \quad (12)$$

where

$$\alpha = \sum_{\ell=1}^N \frac{(-1)^\ell}{\ell} \quad (13)$$

is a factor to account for the median being a biased estimator of the mean (see, e.g., Appendix B of Ref. [22]), and ℓ indexes the segments being averaged over. For simplicity, we assume that we are computing the median of an odd number, N , of stretches ensuring $\alpha > 0$. We use nonoverlapping segments to ensure that our average is over independent noise realizations, although we note that overlapping segments are widely used.

It is convenient to work with a regularized version of the PSD,

$$Q = 2\hat{P}/P. \quad (14)$$

Since the data are assumed to follow a zero-mean Gaussian distribution with variance P , the Q_i are drawn from a χ^2 distribution of order 2,

$$p(Q) = \chi_2^2(Q) = \frac{1}{2} \exp\left(-\frac{Q}{2}\right). \quad (15)$$

Additionally, we define the usual cumulative distribution function, Φ , and survival function, S , for this quantity:

$$\Phi(Q) = \int_0^Q dQ' p(Q') = 1 - \exp\left(-\frac{Q}{2}\right), \quad (16)$$

$$S(Q) = \int_Q^\infty dQ' p(Q') = \exp\left(-\frac{Q}{2}\right). \quad (17)$$

The probability of the median of an odd number of segments follows the median order statistic. This is the probability of the getting the median value from the distribution multiplied by the probability of having $m = (N - 1)/2$ measurements less than \hat{Q} and m measurements larger than \hat{Q} . Symbolically, this is

$$\pi(Q|\hat{P}) = \frac{p(Q)}{2\hat{P}} \frac{\Phi(Q)^m S(Q)^m}{\text{B}(m+1, m+1)} \quad (18)$$

$$= \frac{(1 - e^{-\frac{1}{2}Q})^m e^{-\frac{(m+1)}{2}Q}}{4\hat{P}\text{B}(m+1, m+1)} \quad (19)$$

$$= \sum_{k=0}^m \binom{m}{k} \frac{(-1)^k e^{-\frac{(m+k+1)}{2}Q}}{4\hat{P}\text{B}(m+1, m+1)}, \quad (20)$$

where B is the beta function and in the last line we perform a binomial expansion. The final piece we need is to relate our prior on Q to our prior on P ,

$$\pi(P|\hat{P})dP = \pi(Q|\hat{P})dQ. \quad (21)$$

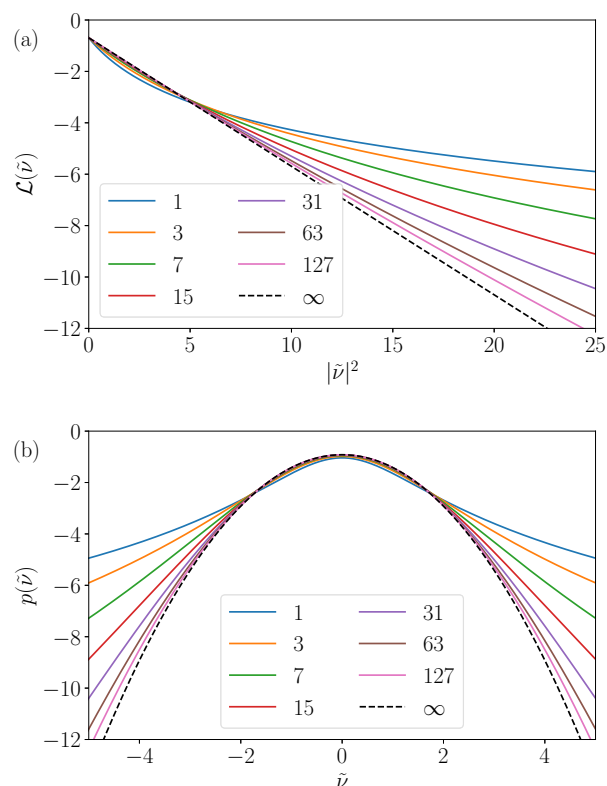


FIG. 1. The distributions of the whitened power (a) and whitened data (b) when using a median average power spectral density estimate for a range of numbers of segments used in the average. As the number of averages approaches infinity, the distributions converge to the unmarginalized case.

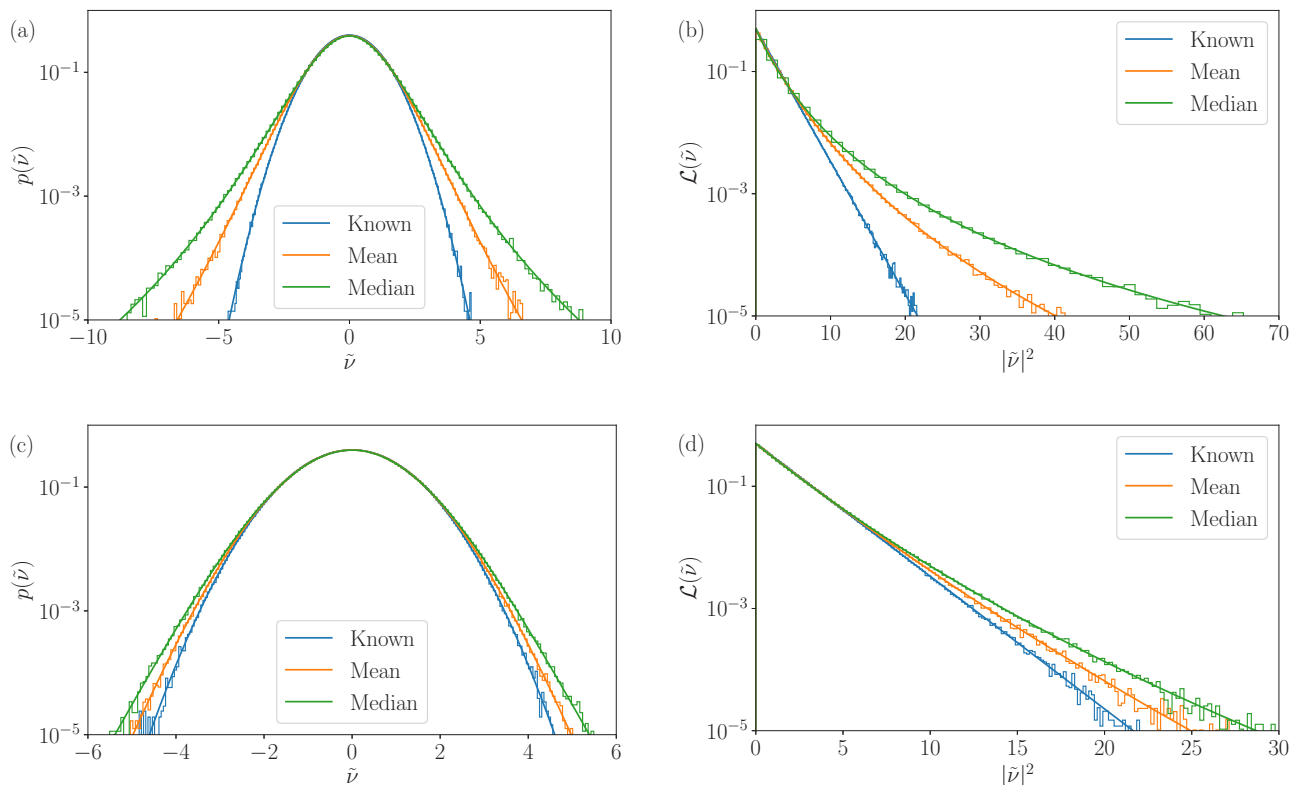


FIG. 2. Comparison of the expected and empirical distribution of whitened frequency-domain strain (left) and whitened noise power (right) when using three different PSDs for whitening and simulated Gaussian noise colored to the known PSD. The three PSDs used are the true “known” PSD (blue/darkest gray), a mean estimate (orange/lightest gray), and a median estimate (green/mid-gray). The number of averages used to generate the mean and median estimates are seven (top) and 31 (bottom). In all cases the data follow the predicted distributions.

Substituting this expression into Eq. (10), the PSD-marginalized likelihood is

$$\mathcal{L}_P = \int_0^\infty dQ \frac{Q}{8\pi\hat{P}} \frac{(1 - e^{-\frac{Q}{2}})^m e^{-\frac{Q}{2}(\frac{1}{2}|\tilde{v}|^2 + m + 1)}}{\mathbf{B}(m+1, m+1)} \quad (22)$$

$$= \sum_{k=0}^m \binom{m}{k} \frac{(-1)^k (m+k+1 + \frac{|\tilde{v}|^2}{2})^{-2}}{2\pi\hat{P} \mathbf{B}(m+1, m+1)}, \quad (23)$$

and using Eq. (11), the distribution of whitened residuals is

$$p(\tilde{n}) = \int_0^\infty dQ \frac{1}{4\sqrt{2\pi\hat{P}}} \frac{Q}{2\pi\hat{P}} \frac{(1 - e^{-\frac{Q}{2}})^m e^{-\frac{Q}{2}(\frac{1}{2}|\tilde{v}|^2 + m + 1)}}{\mathbf{B}(m+1, m+1)} \quad (24)$$

$$= \sum_{k=0}^m \binom{m}{k} \frac{(-1)^k (m+k+1 + \frac{|\tilde{v}|^2}{2})^{-3/2}}{\sqrt{2\pi\hat{P}} \mathbf{B}(m+1, m+1)}. \quad (25)$$

While the final expressions Eqs. (23) and (25) are exact closed-form solutions, they are numerically unstable and cannot be safely computed for $m \gtrsim 15$. We therefore simply construct an interpolant over numerically computed values of the integrals in Eqs. (22) and (24), which can be rapidly evaluated at run time. In Fig. 1 we show the distributions derived above for a range of values of N .

C. Mean PSD estimate

The appropriate distribution to use for a mean averaged PSD has been discussed and independently derived multiple

times in the literature (e.g., Refs. [15,17]); in this work, we just quote the relevant results. For a mean estimate:

$$\hat{P} = \frac{1}{N} \sum_{i=1}^N P_i, \quad (26)$$

$$\pi(P|\hat{P})dP = \pi(Q|\hat{P})dQ = \chi_{2N}^2(Q)dQ, \quad (27)$$

$$\mathcal{L}_P = \frac{1}{2\pi\hat{P}} \left(1 + \frac{|\tilde{v}|^2}{2N}\right)^{-(N+1)}, \quad (28)$$

$$p(\tilde{v}) = \frac{\Gamma(N+1/2)}{\sqrt{2\pi N\hat{P}}\Gamma(N)} \left(1 + \frac{|\tilde{v}|^2}{2N}\right)^{-(N+1/2)}. \quad (29)$$

Equation (28) is the F distribution with $d_1 = 2$, $d_2 = 2N$ [23] and Eq. (29) is the Student's t distribution with $2N$ degrees of freedom, two degrees for each segment being averaged over, coming from the real and imaginary components of the frequency domain strain.

D. Limiting cases

When $N = 1$ the mean and median are the same and so Eq. (25) should reduce to a Student's t distribution with two degrees of freedom. As expected, we find

$$p(\tilde{n}) = \frac{1}{\sqrt{2\pi\hat{P}}} \left(1 + \frac{|\tilde{v}|^2}{2}\right)^{-3/2}. \quad (30)$$

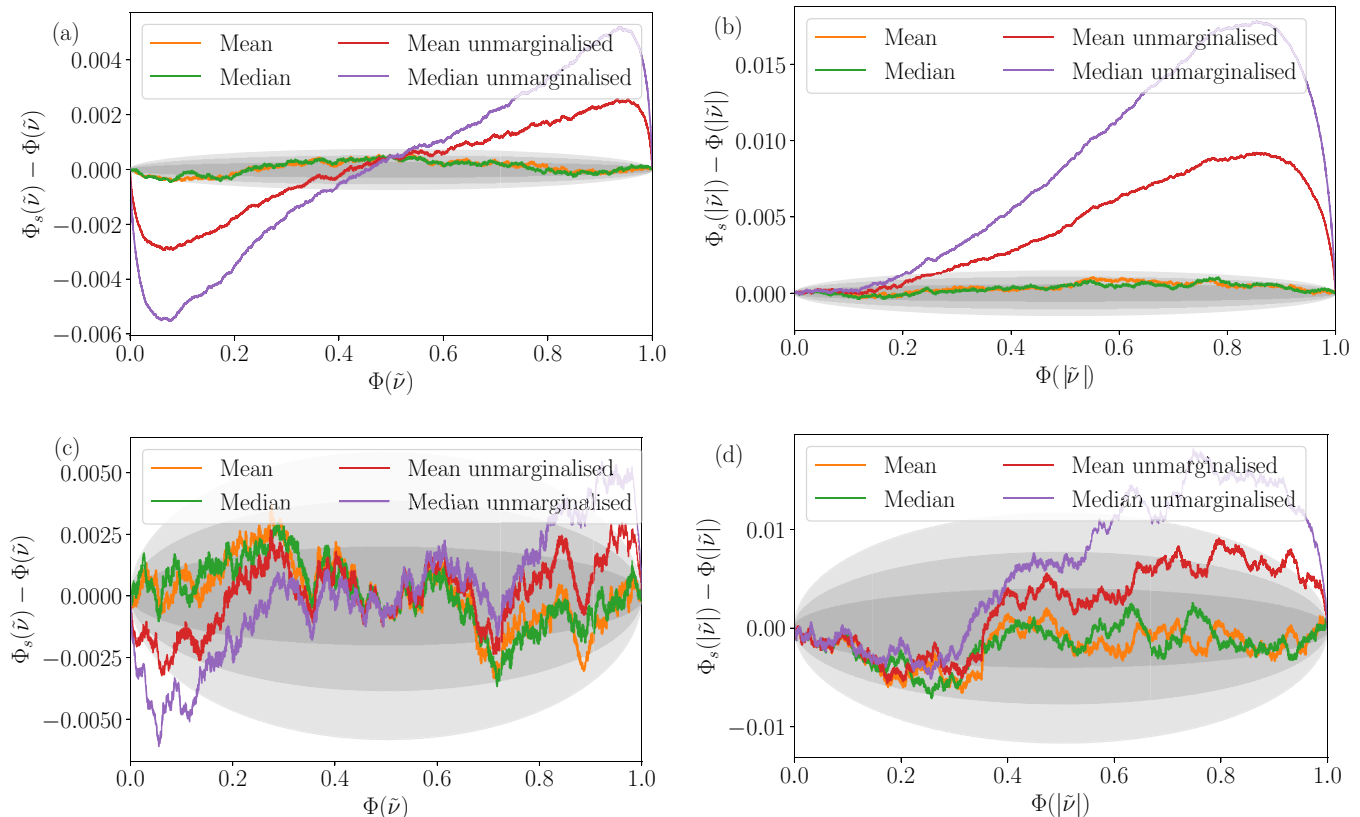


FIG. 3. Difference between the empirical and expected cumulative distributions of whitened frequency-domain strain (left) and whitened noise power (right) when using two different PSDs for whitening and simulated Gaussian noise colored to a known PSD. In the top panels we consider 512 s of simulated data and in the bottom panels we consider 8 s of simulated data. The two PSDs used are a mean estimate (orange/red, lightest/darkest gray) and a median estimate (green/purple, lighter/darker gray). The gray-shaded regions show the expected 1σ , 2σ , and 3σ uncertainties. We average over 31 realizations to generate the PSDs. There are significant deviations when not marginalizing over the uncertainty in the PSD once enough data are considered.

Analogously, the PSD uncertainty marginalized likelihoods also match and are both

$$\mathcal{L}_P = \frac{1}{2\pi\hat{P}} \left(1 + \frac{|\tilde{v}|^2}{2}\right)^{-2}. \quad (31)$$

The other important limiting case is when $N \rightarrow \infty$. It is a well-known result that the Student's t distribution converges to a Gaussian in this limit; this follows from Taylor-expanding the distribution. Performing a similar expansion, it is possible to demonstrate that the Student-Rayleigh distribution converges to the Whittle likelihood. We numerically confirm that (25) and (23) also converge to a Gaussian distribution and Whittle likelihood respectively although more slowly than (29) and (28).

III. BAYESIAN INFERENCE FOR GRAVITATIONAL-WAVE TRANSIENTS

In the previous section, we derived likelihood functions, marginalized over the statistical uncertainty in an estimate of the PSD. These likelihood functions $\mathcal{L}(\tilde{d}|\theta, \hat{P}, \mathcal{M})$ are the probability of obtaining data \tilde{d} given a signal model \mathcal{M} described by parameters θ and a PSD estimate \hat{P} . However, we are generally interested in measuring the source-model parameters and performing model comparison. Using Bayes'

theorem we get

$$p(\theta|\tilde{d}, \hat{P}, \mathcal{M}) = \frac{\mathcal{L}(\tilde{d}|\theta, \hat{P}, \mathcal{M})\pi(\theta, \hat{P}, \mathcal{M})}{\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M})}. \quad (32)$$

The term on the LHS, $p(\theta|\tilde{d}, \hat{P}, \mathcal{M})$, is the posterior probability distribution, the probability of the parameters describing the model given the data. The term $\pi(\theta, \hat{P}, \mathcal{M})$ is our prior distribution which is based on our expectation before analyzing the data. The term $\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M})$ is the evidence for the data given the model \mathcal{M} .

The evidence is used for model comparison by computing Bayes factors for two models,

$$BF_0^1 = \frac{\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M}_1)}{\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M}_0)}. \quad (33)$$

While the Bayes factor is often used for model selection, strictly speaking, we should compare the probability of the model given the data, rather than the probability of the data given the model. This is given by the odds

$$\mathcal{O}_0^1 = \frac{\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M}_1) \pi(\mathcal{M}_1|\hat{P})}{\mathcal{Z}(\tilde{d}|\hat{P}, \mathcal{M}_0) \pi(\mathcal{M}_0|\hat{P})}, \quad (34)$$

which is the Bayes factor comparing the two models multiplied by the prior odds. Throughout this work, we will assume

all models have equal prior odds and so the odds reduces to the Bayes factor.

Finally, we define the coherent versus incoherent Bayes factor [24], BCI, as a measure of the relative probability that the data contain a coherent signal or incoherent signals in different detectors,

$$\text{BCI} = \frac{\mathcal{Z}(\{\tilde{d}_k\}|\{\hat{P}_k\}, \mathcal{M})}{\Pi_k \mathcal{Z}(\tilde{d}_k|\hat{P}_k, \mathcal{M})}. \quad (35)$$

Here the k index multiple independent interferometers, e.g., LIGO Hanford and LIGO Livingston. As in Ref. [25] we assume that any incoherent signals are described by the same model as the coherent signals; however, this is not necessarily the case [26]. We note that the BCI is not used as the final discriminator between the coherent and incoherent models as it is missing a prior for the relative rates of coherent and incoherent signals. In Refs. [25,26] the priors on rate are empirically calibrated delta functions; however, in Ref. [27] the authors fit the rates of coherent and incoherent signals.

In general relativity, noneccentric binary black hole coalescences are fully described by fifteen parameters. Eight parameters which describe the “intrinsic” properties of the binary (two masses and two three-dimensional angular momentum vectors), and seven “extrinsic” parameters to specify the position, orientation, and coalescence time of the binary relative to Earth. This parameter space is typically explored using stochastic samplers using either Markov-chain Monte Carlo [28] or nested sampling [29].

In order to improve the convergence of the sampling and accelerate our inference, it is possible to use a modification of the Whittle likelihood which is marginalized over the coalescence time, orbital phase, and distance of the source [24,30]. It is not possible to perform these marginalizations as easily while also marginalizing over uncertainty in the PSD. Therefore, in this work, we perform our inference in two stages following Ref. [31]:

(1) First, we analyze the data using the Whittle likelihood marginalized over coalescence time, binary orbital phase, and distance to obtain samples from the posterior distribution and an estimate of the signal evidence and Gaussian noise evidence. Posterior distributions for these marginalized parameters are then recomputed in postprocessing. We use *dynesty* [32], an implementation of the nested sampling algorithm, as implemented in *Billby* [33] to sample the space.

(2) After this, we importance resample the posterior obtained in the previous step by the ratio of the PSD-marginalized likelihood to the Whittle likelihood to obtain posterior samples and an evidence which include the marginalization over the statistical uncertainty.

We note that the importance sampling in step 2 works only when resampling to a distribution which is similar to the original posterior distribution. We quantify this by evaluating the efficiency of the resampling and the number of effective samples from the PSD-marginalized posterior. Since the marginalized likelihoods converge to the nonmarginalized likelihood when averaging many segments, we expect the resampling to be efficient. This method also generically gives a much smaller uncertainty on the Bayes factor comparing the two models than would be obtained by performing two

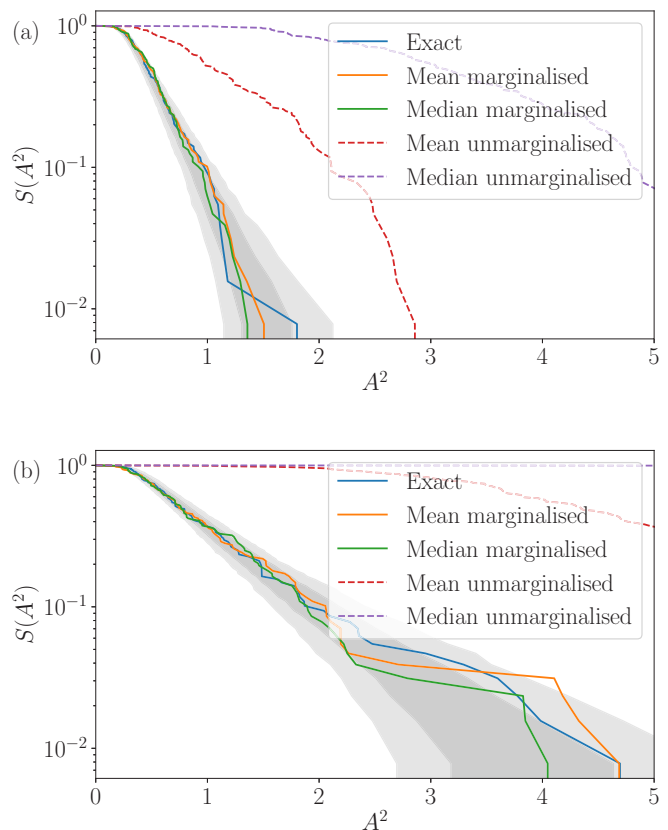


FIG. 4. The survival function of the Anderson-Darling statistic comparing Gaussian noise whitened with three PSDs with the distributions which do/do not marginalize over the statistical uncertainty in the PSD estimate. In panel (a), we analyze the real and imaginary components of the whitened strain. In panel (b), we analyze whitened power. The blue (dark solid) curve is generated using data whitened by the exact known PSD. The solid curves are generated by comparing data whitened by the mean/median (orange/green, lightest/mid-gray) PSD estimates with the distributions which marginalize over the statistical uncertainty. The dashed curves are generated by comparing data whitened by the mean/median (red/purple, darkest/less dark gray) PSD estimates with the distributions which do not marginalize over the statistical uncertainty. The grey shaded regions show the expected 1σ and 2σ uncertainties. We average over 31 realizations to generate the PSDs. The Anderson-Darling statistic does not follow the expected distribution when not marginalizing over the uncertainty in the PSD.

independent sampling runs [34]. A similar method has previously been employed for cosmological inference in [35] to marginalize over uncertainty in an estimated covariance matrix. We find that including the marginalization over extrinsic parameters reduces the run time by up to approximately an order of magnitude, therefore sampling directly using the PSD-marginalized likelihood would increase the computational cost by an equivalent factor compared to the method used here.

IV. DEMONSTRATION WITH GAUSSIAN NOISE

To demonstrate the accuracy of the methods described in Sec. II we analyze simulated Gaussian noise colored by

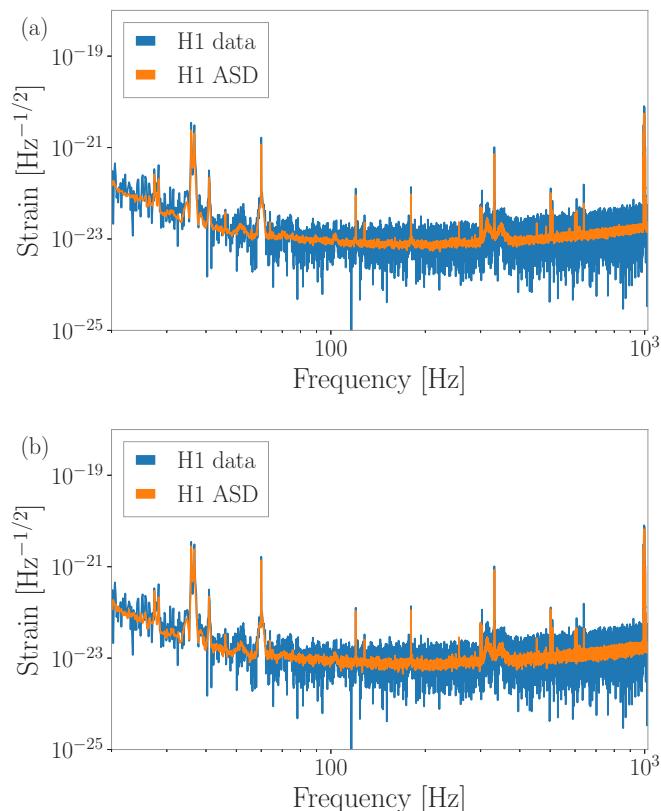


FIG. 5. Frequency domain strain power (blue/dark gray) and mean (a) and median (b) estimated power spectral densities (orange/light gray) for LIGO Hanford at the time of GW151012.

the Advanced LIGO design sensitivity PSD [36]. Following Ref. [13], we perform three tests on data whitened using median and mean PSD estimates for verification. As an extension to the analysis presented in Ref. [13], we consider the whitened power, $|\tilde{v}|^2$ in addition to the real and imaginary components of the whitened strain \tilde{v} . For all estimated PSDs, we average nonoverlapping segments with the same duration as the analysis segment.

First, we perform a visual test of the whitened data. In Fig. 2 we show the distribution of the real and imaginary components of the whitened strain (left) and whitened power (right) along with the theoretical expectations. In the top (bottom) pair of panels, we average over seven (31) independent noise realizations. We see that the data whitened using the averaged PSD estimates follow the expected distributions in each case. We emphasize that throughout we consider only whitened frequency-domain data and so use the ASD as an acausal whitening filter.

In Fig. 3 we show the difference between the empirical and expected cumulative distribution functions, $\Phi_s - \Phi$, plotted against the expected cumulative distribution function for the same data as in Fig. 2. In orange and red we compare the data whitened with the mean PSD estimate with the expected distributions with and without marginalizing over the uncertainty in the PSD respectively. In green and purple we compare the data whitened with the median PSD estimate with the expected distributions with and without marginalizing over the uncertainty in the PSD respectively. The gray regions

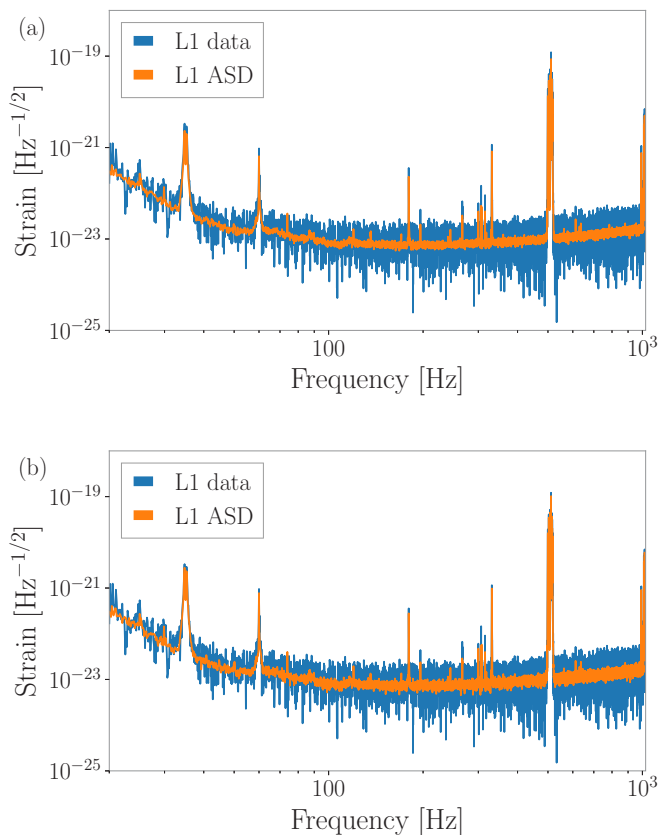


FIG. 6. Frequency domain strain power (blue/dark gray) and mean (a) and median (b) estimated power spectral densities (orange/light gray) for LIGO Livingston at the time of GW151012.

indicate the expected 1σ , 2σ , and 3σ fluctuations. For both PSD estimation methods, we see that the data agree better with the distributions which marginalize over the uncertainty in the PSD.

When comparing the marginalized distributions to the nonmarginalized distributions we see two clear deviations from the expected behavior. The whitened strain uncertainty-marginalized distribution has wider, symmetric, tails than a normal distribution leading to the negative $\Phi_s - \Phi$ for small \tilde{v} and positive $\Phi_s - \Phi$ for large \tilde{v} . The distribution of the whitened power, however, has a wide tail only out to large $|\tilde{v}|$, leading to the positive $\Phi_s - \Phi$ for large \tilde{v} .

To quantify the similarity of the data to the expected distributions we compute the Anderson-Darling statistic

$$A^2 = N \int_{-\infty}^{\infty} d\Phi \frac{(\Phi_s - \Phi)^2}{\Phi(1 - \Phi)}. \quad (36)$$

Here N is the number of samples, in this case, the number of frequency bins. The numerator is the square of the quantity on the vertical axis of Fig. 3, and the integral is over the horizontal axis.

In Fig. 4 we show the survival function of the distribution of the Anderson-Darling statistic for four cases: for both the mean and median PSD estimation methods we compare the distribution of the whitened strain to a unit normal distribution and the expected distribution as described in the previous section for the whitened strain (a) and whitened power (b). We

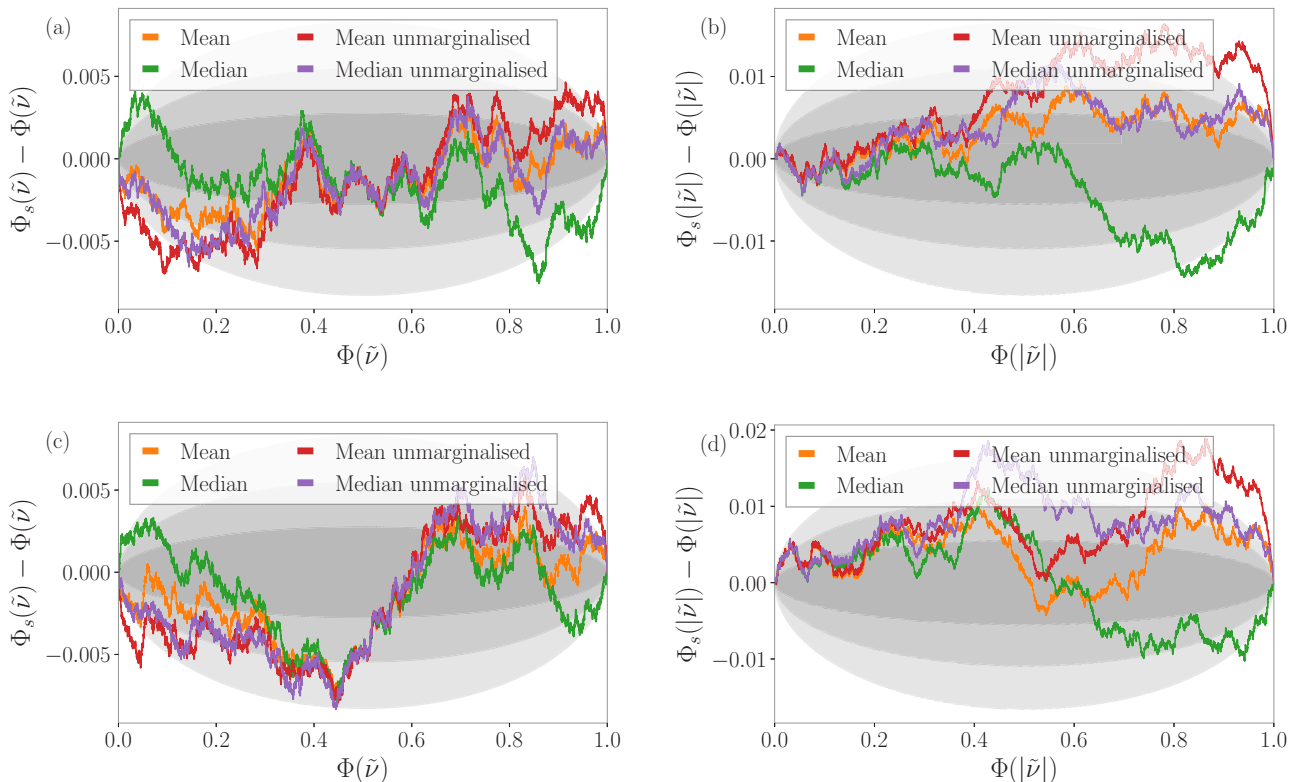


FIG. 7. The difference between the empirical and analytic cumulative distribution functions of whitened strain \tilde{v} and whitened power $|\tilde{v}|^2$ in the LIGO Hanford (top) and Livingston (bottom) interferometers with two different power spectral density estimation methods at the time of GW151012. The gray-shaded regions show the expected 1σ , 2σ , and 3σ fluctuations. For the orange (lightest gray) and green (lighter gray) curves, the data are compared with the distributions which marginalize over uncertainty in the power spectral density estimate. For the red (darkest gray) and purple (darker gray) curves, the data are compared with the distributions which do not marginalize over uncertainty in the power spectral density estimate. We note that the latter pair of curves deviate from the 3σ region, while the former does not. The data are well described by a stationary Gaussian process when marginalizing over uncertainty in the power spectral density.

also show the expected distribution if the two distributions are the distributions when the PSD is exactly known.

We note that the gradient of the expected distribution of the Anderson-Darling statistic is steeper for the whitened strain than for the whitened power. When applying a window the data before performing the discrete Fourier transforms, the real and imaginary components of the frequency domain strain are no longer independent, reducing the appropriate value of N by a factor of two (see Appendix A of Ref. [21]). We, therefore, avoid the case identified in [13] where the distribution of v appeared to match the correct distribution better than possible.

V. A CASE STUDY: GW151012

To examine the effect of non-Gaussianity and nonstationarity on the noise properties of real gravitational-wave detectors, we analyze the data in the two Advanced LIGO interferometers [37] at and around the time of GW151012 [38], the lowest significance binary black hole merger included in the first gravitational-wave transient catalog [19]. We analyze 256 s of data ending 2 s after the merger time. We subdivide the data into 32 8 s chunks, the first 31 chunks are used to compute the PSD and the final 8 s are the analyzed data [39].

We apply a Tukey window with a roll off of 0.2 s to each of the chunks to suppress spectral leakage. We then fast-Fourier

transform the windowed time-domain strain before averaging the PSD chunks. We do not apply the conventional window amplitude correction factor to either the PSD or the data. After applying the fast-Fourier transform, we remove all data below 20 Hz and above 1024 Hz.

The resulting PSDs and the power in the analyzed data are shown in Figs. 5 and 6. Figure 5 shows data from the LIGO Hanford interferometer and Fig. 6 data from the LIGO Livingston interferometer. The orange curves show the mean estimated PSDs, and the green show the median estimated PSDs. All the PSD estimates are at the center of the scatter in the analyzed data, as expected. We note that the width of the scatter on the mean PSDs is slightly smaller than for the median due to the slower convergence of the median estimate.

A. Data quality tests

The main reason for using a median estimate over a mean estimate is to mitigate the effect of large non-Gaussian transients. However, the formalism derived above is incorrect if there is a large outlier in the data being averaged over; see Appendix A for a more detailed discussion of the issue and a potential resolution. Therefore, we try to identify if any of the segments are clear outliers. We compute the power per segment divided by the mean power in all the other segments.

TABLE I. Values of the Anderson-Darling statistic for the whitened strain for the mean and median marginalized likelihood at the time of GW151012. Larger values of the Anderson-Darling statistic indicate comparatively worse agreement. The marginalized distributions match the data better.

	Mean vs marginalized	Median vs marginalized	Mean vs not marginalized	Median vs not marginalized
Livingston strain	0.53	0.46	1.62	2.66
Hanford strain	0.56	0.42	0.48	2.21
Livingston power	1.48	1.03	4.22	6.71
Hanford power	0.79	0.48	0.84	5.55

This is essentially testing how well the data in each of the segments is whitened by the data in the other segments. We apply an empirically tuned threshold of 1.5 for the mean whitened power per segment. Any segment with a mean power above this value we discard and repeat the test. We identify that one segment of the Hanford data which fails this test with a mean whitened power of 2.66. Visual inspection reveals that this segment has a larger amplitude than all the others below ~ 100 Hz. No significant outliers are present in the selected Livingston data.

Additional tests of the quality are possible and performed routinely during gravitational-wave data analysis. For example, researchers often remove specific frequency bins if the noise at that frequency is known to be non-Gaussian, e.g., around the frequency (and higher harmonics) of mains electricity [1]. A possible extension would be to use the normalized average power used above to track nonstationarity in the data, a similar method is used in [11]. Implementing further data quality cuts and vetoes will improve the quality of our PSD estimates and is an interesting avenue for further study.

B. Data whitening

We repeat the tests performed in Sec. IV on the data. In Fig. 7 we show the deviations from the expected cumulative distribution functions for the data from the Hanford (top) and Livingston (bottom) interferometers. On the left, we show the real and imaginary components of the whitened strain and on the right the whitened power. In orange we show the difference between the empirical mean-estimated PSD whitened data and expected mean-marginalized distributions (29, 28), in green the difference between the empirical median-estimated PSD whitened data and expected median-marginalized distributions (25, 23). In red (purple) we compare the data whitened using the mean- (median-) estimated PSDs with the distributions which do not account for the uncertainty. In gray we show the 1σ , 2σ , and 3σ expected deviations.

In Table I we quote the corresponding values of the Anderson-Darling statistic for each of these lines. We see that the largest deviations are observed when using data whitened with a median estimated PSD and compared to the nonmarginalized distributions. We also see that, with the exception of the strain components in Hanford, the Anderson-Darling statistic is always smaller when using the appropriate marginalized distributions.

C. Impact on inference

We analyze the data using Bayesian inference as described in Sec. III twice, once each with the mean-averaged and

median-averaged PSDs to obtain samples from the posterior distribution and Bayesian evidences under four sets of assumptions:

- (1) The data are well described by the mean-estimated PSD and the Whittle likelihood.
- (2) The data are well described by the mean-estimated PSD and the Student-Rayleigh likelihood.
- (3) The data are well described by the median-estimated PSD and the Whittle likelihood.
- (4) The data are well described by the median-estimated PSD and the median marginalized likelihood, Eq. (23).

In Table II we show the natural logarithm of the BCI under these four set of assumptions. We find that both PSD estimation methods have $\ln \text{BCI} \approx 10$ which is a moderately strong preference for the coherent hypothesis, although we note that a full treatment requires careful consideration of prior odds. For both PSD estimates the BCI decreases slightly when marginalizing over the uncertainty. The increase in the BCI when using the median estimated PSD is mirrored in the increased signal-to-noise ratio ρ in the lower panel of Fig. 8. This is likely due to the different handling of non-Gaussian features in the mean and median PSD estimation methods.

In Table III we show the natural log Bayes factors comparing the marginalized to unmarginalized likelihoods for both PSD estimation methods. In both cases, we see a strong preference for the model which marginalizes over the uncertainty. This preference is much larger for the median PSD estimate. This can be understood by the fact that the large $|\nu|$ tail of the median marginalized likelihood is broader than the mean marginalized likelihood; cf. Fig. 2, lower-right panel.

In Figs. 8 and 9 we show selected posterior distribution under our four sets of assumptions. In Fig. 8(a) we show the posterior distribution for the best measured combination of the component masses, the chirp mass

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}. \quad (37)$$

TABLE II. Values of the (natural) log coherent vs incoherent Bayes factor ($\ln \text{BCI}$) for different PSD estimates and likelihoods. In both cases the difference between the $\ln \text{BCI}$ with and without marginalizing over the PSD uncertainty is within typical uncertainties due to finite sampling.

PSD	No marg.	Marg.
Mean	9.98	9.89
Median	10.39	10.16

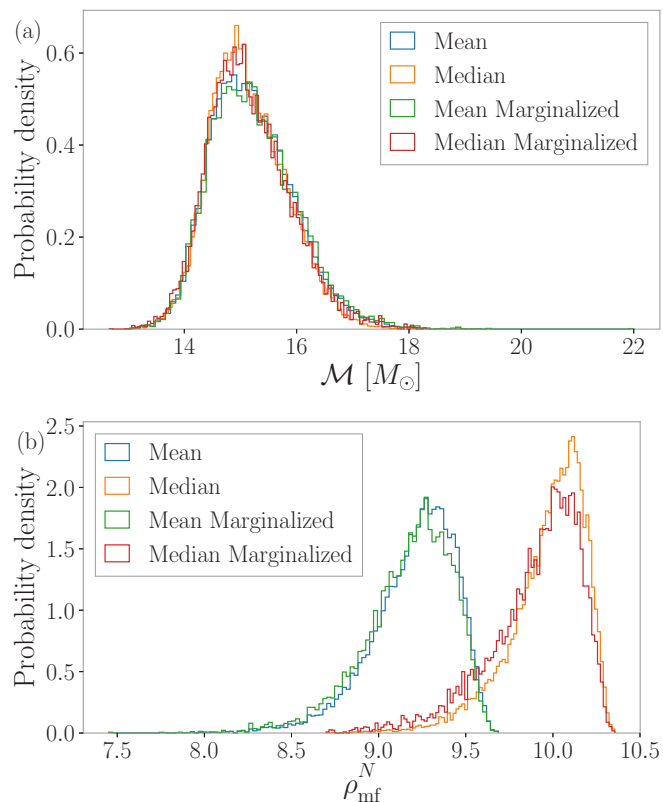


FIG. 8. The posterior distribution for chirp mass (a) and matched filter SNR (b) for GW151012 for four different models. In blue (darker gray) and green (lighter gray) we use the mean estimated PSD while in orange (lightest gray) and red (darker gray) we use the median estimate. In blue and orange we neglect the uncertainty in the PSD estimate and in green and red we marginalize over the appropriate statistical uncertainty. We note that in both cases, marginalizing over the uncertainty increases the width of the chirp mass posterior and decreases the average SNR. The mean estimated PSD gives a wider chirp mass posterior than the median PSD estimate and we see a corresponding decrease in the recovered matched filter SNR.

The parameters m_i are the masses of the two component black holes. In Fig. 8(b) we show the network matched filter signal-to-noise ratio (SNR). We find that the recovered SNR is larger when using the median PSD estimate and the marginalizing over the uncertainty in the PSD increases the posterior support at SNR less than the maximum found SNR but does not decrease the maximum SNR. Correspondingly, we see that the posterior for chirp mass is slightly less strongly peaked when marginalizing over uncertainty in the PSD, and when using the median PSD estimate.

TABLE III. Values of the (natural) log Bayes factor comparing the marginalized and unmarginalized likelihood hypotheses. In both cases there is a strong preference for the marginalized likelihood better describing the data than the unmarginalized likelihood.

PSD	Marg. vs no marg.
Mean	19.26
Median	91.67

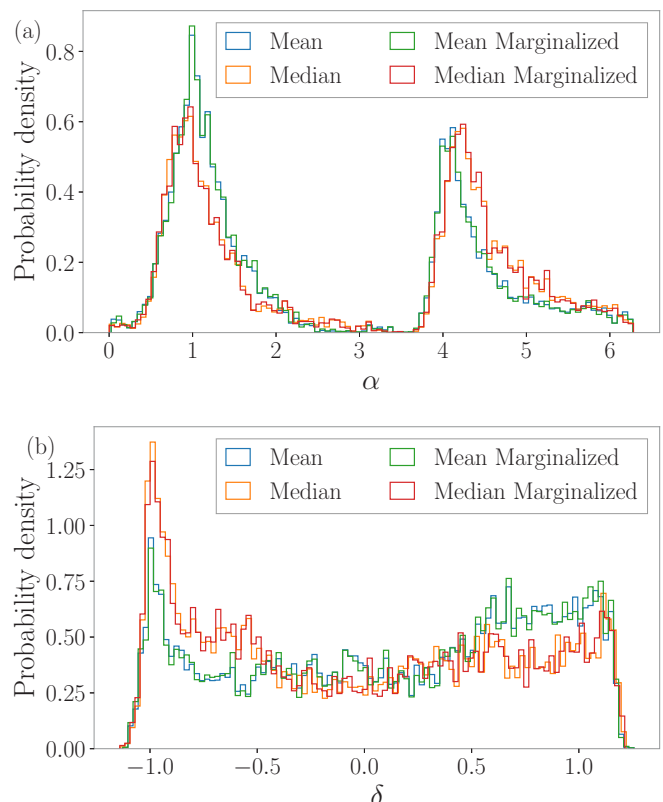


FIG. 9. The posterior distribution for right ascension (a) and declination (b) for GW151012 for four different models. In blue (darker gray) and green (lighter gray) we use the mean estimated PSD while in orange (lightest gray) and red (darkest gray) we use the median estimate. In blue and orange we neglect the uncertainty in the PSD estimate and in green and red we marginalize over the appropriate statistical uncertainty. In this case, marginalizing over the uncertainty does not make a large difference to the inferred posterior distributions. However, the different PSD estimation techniques while giving consistent posterior distributions give different posterior weights to different parts of the sky.

In Fig. 9 we show the posterior distribution for the parameters describing the position on the sky, right ascension α and declination δ . The impact of marginalizing over the

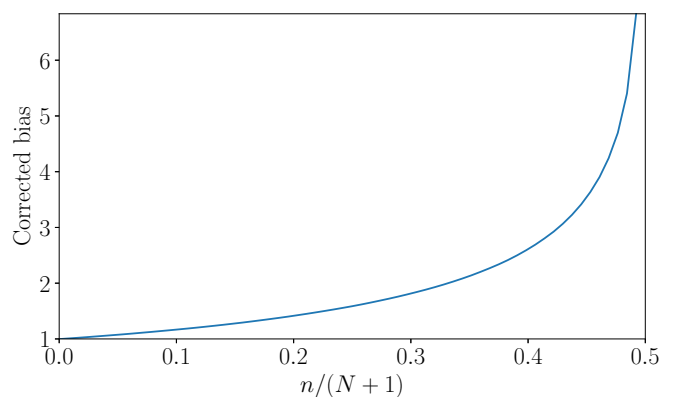


FIG. 10. The ratio of the true bias factor in the $(N+1)/2$ -th sample compared to the median bias as a function of the fraction of segments with large outliers.

uncertainty in the PSD does not significantly affect the inferred sky localization of the binary. However, the two PSD estimation methods recover different posterior distributions within the same region on the sky.

The fact that the differences in the posterior distributions and Bayes factors when using the different PSD estimation methods are larger than the corrections due to marginalizing over the statistical uncertainty mean that either one or both of the estimation methods are producing a biased estimate of the true PSDs. The source of this bias is presumably nonstationarity and/or non-Gaussianity in the data used to estimate the PSD. It is not possible to determine which estimate is less biased and so this must be considered as an extra source of systematic uncertainty in average PSD estimates.

VI. DISCUSSION

Performing astrophysical inference on gravitational-wave data requires an estimate of the noise power spectral density (PSD). In practice it is not possible to know this perfectly, only to estimate it from the data. Multiple methods of estimating the PSD are used, and each carries with it a different class of statistical uncertainty. In this work, we derived the relevant statistical uncertainty for an estimation method, which is widely used when analyzing gravitational-wave transients, the median average. We obtained a closed-form expression for the likelihood, which marginalizes over the statistical uncertainty, and demonstrated that simulated Gaussian data matches this distribution.

We then applied our new results to the lowest significance transient in the first LIGO/Virgo gravitational-wave transient catalog, GW151012. We analyzed this event using two different PSDs with likelihoods, which did and did not marginalize over the appropriate statistical uncertainty, one using a median average, and one using a mean average. We showed the PSD estimation method has a clear effect on the inferred posterior probability distribution and Bayesian evidence. The changes in the posterior distributions and Bayesian evidence when marginalizing over the statistical uncertainty is more subtle. However, for applications which require precise estimates of the evidence such as Refs. [25–27], these small differences will be crucial.

There are many interesting extensions to the work presented here, which are left to future work. These include implementing data quality tests when analyzing real gravitational-wave data, which are known to be non-Gaussian and nonstationary over timescales of minutes to hours. There are also other average PSD estimation methods employed. For example, the “median-mean” estimate [22], a hybrid method in which the mean of two median estimated PSDs is taken. The method presented here can be applied to that method, however, in practice the statistics describing this average are more complex.

Examples of handling non-Gaussianity and nonstationarity can be found in other areas of gravitational-wave data analysis. For example, searches for gravitational-wave transient signals implement methods to track and mitigate nonstationarity [11] and remove large non-Gaussian transients [10,11]. Searches for continuous gravitational-wave sources and the stochastic gravitational-wave background include algorithms

to detect and remove stretches of data where the PSD is rapidly fluctuating [5] or frequencies where the data are known to be non-Gaussian, e.g., around the frequency of AC mains electricity [1]. By combining these methods and the statistical models presented here, we can enable precision astrophysical inference for gravitational-wave transients without large computational overheads.

The code to generate the distributions in this work are available at Ref. [47] including implementations in `scipy` and `Bilby` formats. This work used the following software packages: `numpy` [40], `matplotlib` [41], `scipy` [42], `gwpy` [43], `lalsuite` [44], `Bilby` [33].

ACKNOWLEDGMENTS

We thank Sharan Banagiri, Sylvia Biscoveanu, Katerina Chatziioannou, Pat Meyers, Joe Romano, and Alan Weinstein for helpful discussions and comments on the manuscript. C.T. and E.T. are supported by the Australian Research Council (ARC) CE170100004. C.T. acknowledges the support of the National Science Foundation and the LIGO Laboratory. E.T. is supported by ARC FT150100281.

This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center [45,46,48], a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO is funded by the U.S. National Science Foundation. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN), and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. The authors are grateful for computational resources provided by the LIGO Lab and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459.

APPENDIX: ROBUSTNESS OF MEDIAN PSD ESTIMATE

The widespread use of median PSD estimates in the gravitational-wave literature is motivated by the additional robustness of the median estimator compared to a mean estimate. While it is true that the median estimate is less biased than the mean estimate, it is still biased by outliers.

The median of N samples is the value $(N + 1)/2$ -th sample (for odd N). If there are n outliers we have $(N - n)$ samples from the expected distribution and n samples from the distribution of outliers, which we assume will be the n largest values. When we compute a median, we are therefore taking the $(N + 1)/2$ -th of $(N - n)$ samples drawn from the expected distribution, and so rather than the median order statistic, we must modify Eq. (18) to use this new order statistic

$$\pi(Q|\hat{P}, n) = \frac{p(Q)}{2\hat{P}} \frac{\Phi(Q)^m S(Q)^{m-n}}{B(m+1, m-n+1)} \quad (\text{A1})$$

$$\frac{(1 - e^{-\frac{1}{2}Q})^{m-n} e^{-\frac{(m+1)Q}{2}}}{4\hat{P} B(m+1, m-n+1)} \quad (\text{A2})$$

$$\sum_{k=0}^{m-n} \binom{m-n}{k} \frac{(-1)^k e^{-\frac{(m+k+1)Q}{2}}}{4\hat{P} B(m+1, m-n+1)}. \quad (\text{A3})$$

If we expect that there is a Poisson process producing glitches with some rate R and we average T second segments to estimate the PSD, the expected number of glitches per segment follows a Poisson distribution with $\lambda = T/R$. The prior probability that n segments contain at least one glitch is then

$$\pi(n) = p^{N-n}(k=0, \lambda)[1 - p(k=0, \lambda)]^n \quad (\text{A4})$$

$$= e^{-\frac{(N-n)T}{R}} \left(1 - e^{-\frac{T}{R}}\right)^n. \quad (\text{A5})$$

The full expression for use in Eq. (18) is then

$$\pi(Q|\hat{P}) = \sum_{n=0}^{(N-1)/2} \pi(Q|\hat{P}, n)\pi(n). \quad (\text{A6})$$

Here, the upper limit for the sum is $(N-1)/2$ corresponding to the case where one in every two segments contains a non-Gaussian outlier. Therefore, if $R < 2T$ the median PSD method cannot produce correct (or even sensible) estimates.

We also note that the bias factor α should be smaller for these not-quite-median PSD estimates. In Fig. 10 we show the ratio of the correct bias factor to the median bias as a function of the fraction of segments containing a glitch. As the number of averages increases, the bias decreases. As expected, the ratio diverges for $n = (N+1)/2$.

Another factor to consider is that many glitches do not affect the full frequency band being analyzed and the rate of glitches changes over time. We can account for this by using a time-and-frequency-dependent glitch rate.

-
- [1] B. P. Abbott *et al.*, A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals, *Class. Quantum Gravity* **37**, 055002 (2020).
- [2] J. D. Romano and N. J. Cornish, Detection methods for stochastic gravitational-wave backgrounds: A unified treatment, *Living Rev. Relativ.* **20**, 2 (2017).
- [3] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, *Phys. Rev. D* **91**, 084034 (2015).
- [4] P. Welch, The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.* **15**, 70 (1967).
- [5] B. P. Abbott *et al.*, Upper Limits on the Stochastic Gravitational-Wave Background from Advanced LIGO's First Observing Run, *Phys. Rev. Lett.* **118**, 121101 (2017).
- [6] S. A. Usman, A. H. Nitz, I. W. Harry, C. M. Biwer, D. A. Brown, M. Cabero, C. D. Capano, T. Dal Canton, T. Dent, S. Fairhurst *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, *Class. Quantum Gravity* **33**, 215004 (2016).
- [7] C. Pankow, K. Chatziioannou, E. A. Chase, T. B. Littenberg, M. Evans, J. McIver, N. J. Cornish, C.-J. Haster, J. Kanner, V. Raymond *et al.*, Mitigation of the instrumental noise transient in gravitational-wave data surrounding GW170817, *Phys. Rev. D* **98**, 084016 (2018).
- [8] J. C. Driggers, S. Vitale, A. P. Lundgren, M. Evans, K. Kawabe, S. E. Dwyer, K. Izumi, R. M. S. Schofield, A. Effler, D. Sigg *et al.*, Improving astrophysical parameter estimation via offline noise subtraction for Advanced LIGO, *Phys. Rev. D* **99**, 042001 (2019).
- [9] D. Davis, T. Massinger, A. Lundgren, J. C. Driggers, A. L. Urban, and L. Nuttall, Improving the sensitivity of Advanced LIGO using noise subtraction, *Class. Quantum Gravity* **36**, 055011 (2019).
- [10] S. Sachdev, S. Caudill, H. Fong, R. K. L. Lo, C. Messick, D. Mukherjee, R. Magee, L. Tsukada, K. Blackburn, P. Brady *et al.*, The GstLAL search analysis methods for compact binary mergers in Advanced LIGO's second and Advanced Virgo's first observing runs, [arXiv:1901.08580](https://arxiv.org/abs/1901.08580) (2019).
- [11] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO, *Phys. Rev. D* **100**, 023011 (2019).
- [12] G. Vajente, Y. Huang, M. Isi, J. C. Driggers, J. S. Kissel, M. J. Szczepańczyk, and S. Vitale, Machine-learning nonstationary noise out of gravitational-wave detectors, *Phys. Rev. D* **101**, 042003 (2020).
- [13] K. Chatziioannou, C.-J. Haster, T. B. Littenberg, W. M. Farr, S. Ghonge, M. Millhouse, J. A. Clark, and N. Cornish, Noise spectral estimation methods and their impact on gravitational wave measurement of compact binary mergers, *Phys. Rev. D* **100**, 104004 (2019).
- [14] C. Röver, Student-t based filter for robust signal detection, *Phys. Rev. D* **84**, 122004 (2011).
- [15] C. Röver, R. Meyer, and N. Christensen, Modelling coloured residual noise in gravitational-wave signal processing, *Class. Quantum Gravity* **28**, 015010 (2011).
- [16] T. Yamamoto, K. Hayama, S. Mano, Y. Itoh, and N. Kanda, Characterization of non-Gaussianity in gravitational wave detector noise, *Phys. Rev. D* **93**, 082005 (2016).
- [17] S. Banagiri, M. W. Coughlin, J. Clark, P. D. Lasky, M. A. Bizouard, C. Talbot, E. Thrane, and V. Mandic, Constraining the gravitational-wave afterglow from a binary neutron star coalescence, *Mon. Not. Roy. Astron. Soc.* **492**, 4945 (2020).
- [18] N. J. Cornish and T. B. Littenberg, BayesWave: Bayesian inference for gravitational wave bursts and instrument glitches, *Class. Quantum Gravity* **32**, 135012 (2015).
- [19] B. P. Abbott *et al.*, GWTC-1: A gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs, *Phys. Rev. X* **9**, 031040 (2019).
- [20] S. Biscoveanu, C.-J. Haster, S. Vitale, and J. Davies, Quantifying the effect of power spectral density uncertainty on gravitational-wave parameter estimation for compact binary sources, *Phys. Rev. D* **102**, 023008 (2020).
- [21] C. Talbot, Astrophysics of binary black holes at the dawn of gravitational-wave astronomy, Ph.D. thesis, Monash University, 2020.
- [22] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, *Phys. Rev. D* **85**, 122006 (2012).

- [23] This distribution is sometimes referred to as the Student-Rayleigh distribution in the gravitational-wave literature [15].
- [24] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, *Phys. Rev. D* **91**, 042003 (2015).
- [25] M. Isi, R. Smith, S. Vitale, T. J. Massinger, J. Kanner, and A. Vajpeyi, Enhancing confidence in the detection of gravitational waves from compact binaries using signal coherence, *Phys. Rev. D* **98**, 042007 (2018).
- [26] G. Ashton, E. Thrane, and R. J. E. Smith, Gravitational wave detection without boot straps: A Bayesian approach, *Phys. Rev. D* **100**, 123018 (2019).
- [27] R. Smith and E. Thrane, Optimal search for an astrophysical gravitational-wave background, *Phys. Rev. X* **8**, 021019 (2018).
- [28] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97 (1970).
- [29] J. Skilling, Nested sampling for general Bayesian computation, *Bayesian Anal.* **1**, 833 (2006).
- [30] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models, *Publ. Astron. Soc. Aust.* **36**, e010 (2019).
- [31] E. Payne, C. Talbot, and E. Thrane, Higher order gravitational-wave modes with likelihood reweighting, *Phys. Rev. D* **100**, 123017 (2019).
- [32] J. S. Speagle, DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences, *Mon. Not. Roy. Astron. Soc.* **493**, 3132 (2020).
- [33] G. Ashton, M. Hübner, P. D. Lasky, C. Talbot, K. Ackley, S. Biscoveanu, Q. Chu, A. Divakarla, P. J. Easter, B. Goncharov *et al.*, Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
- [34] M. Hübner, C. Talbot, P. D. Lasky, and E. Thrane, Measuring gravitational-wave memory in the first LIGO/Virgo gravitational-wave transient catalog, *Phys. Rev. D* **101**, 023011 (2020).
- [35] E. Sellentin and A. F. Heavens, Parameter inference with estimated covariance matrices, *Mon. Not. R. Astron. Soc. Lett.* **456**, L132 (2016).
- [36] B. P. Abbott *et al.*, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, *Living Rev. Relativity* **21**, 3 (2018).
- [37] J. Aasi *et al.*, Advanced LIGO, *Class. Quantum Gravity* **32**, 115012 (2015).
- [38] B. P. Abbott *et al.*, Binary Black Hole Mergers in the First Advanced LIGO Observing Run, *Phys. Rev. X* **6**, 041015 (2016).
- [39] We note that another common convention is to take data symmetrically around the analysis segment. This method can perform better in the presence of a long-term nonstationarity. We found qualitatively similar results when using a symmetric data stretch.
- [40] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, Array programming with NumPy, *Nature (London)* **585**, 357 (2020).
- [41] J. D. Hunter, Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.* **9**, 90 (2007).
- [42] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.* and SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python, *Nature Methods* **17**, 261 (2020).
- [43] D. Macleod, A. L. Urban, S. Coughlin, T. Massinger, Paulaltin, J. Areeda, E. Quintero, the Gitter Badger, and K. Leinweber, <https://zenodo.org/record/2580401#.X7vvnVNXhb8>.
- [44] LIGO Scientific Collaboration, *LIGO Algorithm Library—LALSuite, free software (GPL)* (2018).
- [45] M. Vallisneri, J. Kanner, R. Williams, A. Weinstein, and B. Stephens, The LIGO Open Science Center, *J. Phys. Conf. Ser.* **610**, 012021 (2015).
- [46] The LIGO Scientific Collaboration, the Virgo Collaboration, R. Abbott *et al.*, Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, [arXiv:1912.11716](https://arxiv.org/abs/1912.11716) [gr-qc] (2019).
- [47] <https://www.github.com/ColmTalbot/median-marginalised-distributions>.
- [48] <https://www.gw-openscience.org>.