PHYSICAL REVIEW D **87,** 104021 (2013)

Hyper-efficient model-independent Bayesian method for the analysis of pulsar timing data

Lindley Lentati,* P. Alexander, and M. P. Hobson
Astrophysics Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, United Kingdom

S. Taylor and J. Gair
Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom

S. T. Balan
Astrophysics Group, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, United Kingdom and Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, United Kingdom

R. van Haasteren
Max-Planck-Institut fur Gravitationsphysik (Albert-Einstein-Institut), D-30167 Hannover, Germany
(Received 22 December 2012; published 20 May 2013)

A new model-independent method is presented for the analysis of pulsar timing data and the estimation of the spectral properties of an isotropic gravitational wave background (GWB). Taking a Bayesian approach, we show that by rephrasing the likelihood we are able to eliminate the most costly aspects of computation normally associated with this type of data analysis. When applied to the International Pulsar Timing Array Mock Data Challenge data sets this results in speedups of approximately 2–3 orders of magnitude compared to established methods, in the most extreme cases reducing the run time from several hours on the high performance computer ''DARWIN'' to less than a minute on a normal work station. Because of the versatility of this approach, we present three applications of the new likelihood. In the low signal-to-noise regime we sample directly from the power spectrum coefficients of the GWB signal realization. In the high signal-to-noise regime, where the data can support a large number of coefficients, we sample from the joint probability density of the power spectrum coefficients for the individual pulsars and the GWB signal realization using a ''guided Hamiltonian sampler'' to sample efficiently from this high-dimensional ($\sim$1000) space. Critically in both these cases we need make no assumptions about the form of the power spectrum of the GWB, or the individual pulsars. Finally, we show that, if desired, a power-law model can still be fitted during sampling. We then apply this method to a more complex data set designed to represent better a future International Pulsar Timing Array or European Pulsar Timing Array data release. We show that even in challenging cases where the data features large jumps of the order 5 years, with observations spanning between 4 and 18 years for different pulsars and including steep red noise processes we are able to parametrize the underlying GWB signal correctly. Finally we present a method for characterizing the spatial correlation between pulsars on the sky, making no assumptions about the form of that correlation, and therefore providing the only truly general Bayesian method of confirming a GWB detection from pulsar timing data.

DOI: 10.1103/PhysRevD.87.104021 PACS numbers: 04.30.−w, 95.30.Sf, 04.80.Nn, 04.80.Cc

I. INTRODUCTION

Millisecond pulsars (MSPs) have for some time been known to exhibit exceptional rotational stability, with decades-long observations providing timing measurements with accuracies similar to atomic clocks (e.g. [1,2]). Such stability lends itself well to the pursuit of a wide range of scientific goals, e.g. observations of the pulsar PSR B1913 + 16 showed a loss of energy at a rate consistent with that predicted for gravitational waves [3], while the double pulsar system PSR J0737 − 3039A/B has provided precise measurements of several ''post-Keplerian'' parameters allowing for additional stringent tests of general relativity [4].

By measuring the arrival times (TOAs) of the radio pulses to high precision it is possible to construct a timing model: a deterministic model that describes the physical properties of the pulsar e.g. its binary period and spin evolution, its trajectory, post-Keplerian terms, and so on. A detailed description of this process is available in the Tempo2 series of papers [5–7]. The timing model can then be subtracted from the TOAs resulting in a set of residuals that contain within them any physical effects not correctly accounted for by the timing model.

In this paper we will be concerned with extracting information from these residuals that results from time-correlated stochastic signals. These can include additional

*ltl21@cam.ac.uk

red noise terms due to rotational irregularities in the neutron star [8] or correlated noise between the pulsars due to a stochastic gravitational wave background (GWB) generated by, for example, coalescing black holes (e.g. [9,10]) or cosmic strings (e.g. [11–13]). These could be detected using a pulsar timing array (PTA), a collection of Galactic millisecond pulsars from which the cross-correlated signal induced by a GWB could be extracted. Current methods for the analysis of PTA data are for the most part extremely computationally expensive. This is particularly true for existing Bayesian methods ([14,15], henceforth vHL2013 and vH2009) with large dense matrix inversions resulting in a scaling with the number of data points of approximately $O(n^3)$. Recently new methods have been proposed to speed up this analysis. In [16] (henceforth vH2013), lossy data compression is used to reduce the time these matrix inversions require, resulting in a speedup of ~3–6 orders of magnitude over previous methods, while the authors of Ref. [17] make an approximation to the likelihood function that allows speedups proportional to the square of the number of pulsars in the array. As with other existing Bayesian techniques, however, these methods still assume specific models for the properties of both the GWB and the intrinsic pulsar noise, a statement of prior knowledge whose validity is unknown, since as yet any GWB remains undetected.

In this paper we present an alternative, model-independent approach to performing a Bayesian analysis of PTA data that results in a speedup of between 2 and 3 orders of magnitude when compared to vHL2013, is not limited by the number of free parameters fitted or system memory, using <1 GB of system memory for the analysis of the International Pulsar Timing Array (IPTA) data sets, and critically at no stage requires the specification of any prior form for the shape of the correlated power spectrum induced by a GWB, or the red noise present in a particular pulsar at the point of sampling. This represents a true model-independent means of performing inference on the shape of the power spectrum of a gravitational wave background, where we do not know the form that background will take. We accomplish this in two ways. In the low signal-to-noise regime (Sec. III) we sample directly from the power spectrum coefficients of the GWB signal realization. We show that for the IPTA data challenges, the number of coefficients required to describe the signal is roughly an order of magnitude less than the number of data points in the time domain, and so correspondingly the matrix inversions required in the likelihood are $\sim 10^3$ times faster to compute.

In the high signal-to-noise regime, when the number of coefficients to be sampled is larger, these matrix inversions once again become untenable, and so we sample from the joint probability density of the power spectrum coefficients for the individual pulsars and the GWB signal realization. This allows us to eliminate all matrix-matrix multiplications and costly matrix inversions from the likelihood calculation entirely, replacing them with matrix-vector operations and sparse, banded matrix inversions, so that this new likelihood scales as $O(n \times n_p^3)$ with the number of frequencies sampled n, and number of pulsars n_p while still retaining the ability to make robust statistical inferences about the white and red noise present in the PTA data with the same precision as in vH2009/vHL2013. We perform the sampling process in this case using a guided Hamiltonian sampler (GHS) (Balan, Ashdown, and Hobson [18], henceforth B13), which provides an efficient means of sampling in large numbers of dimensions (potentially $> 10^6$). This method of sampling, in combination with the new, simpler likelihood function, allows us to greatly extend what is computationally feasible from a Bayesian analysis of pulsar timing data. This includes the ability to parametrize the spatial correlations between pulsars directly, without having to assume anything about the form it might take. This spatial correlation is the ''smoking gun'' of a signal from a gravitational wave background, and so the ability to extract it directly from the data is crucial for the credibility of any future detections from pulsar timing data.

Finally, due to the versatility of this approach we show that where desired, models for the power spectrum of the GWB and additional red noise processes such as a single power law can still be applied at the point of sampling.

In Secs. II and III we derive the new likelihood functions. In Sec. IV we describe the guided Hamiltonian sampler and how it can be applied to PTA data analysis. In Sec. V we provide a way of estimating the number of coefficients that are supported by the data in both the low and high signal-to-noise cases. In Sec. VI we apply the three different methods described thus far to the first IPTA data challenge and compare the results with both the established method described in vHL2013 and the updated method described in vH2013. In Sec. VII we then describe and analyze a set of more challenging simulated data sets designed to represent better a future IPTA data release. Finally in Sec. VIII we describe our method of parametrizing the spatial correlation between pulsars.

This research is the result of the common effort to directly detect gravitational waves using pulsar timing, known as the European Pulsar Timing Array, Janssen *et al.* [19,20].

II. ESTIMATING THE POWER SPECTRUM

For any pulsar we can write the TOAs for the pulses as a sum of both a deterministic and a stochastic component:

$$\mathbf{t}_{\rm tot} = \mathbf{t}_{\rm det} + \mathbf{t}_{\rm sto}, \tag{1}$$

where $\mathbf{t}_{\rm tot}$ represents the n TOAs for a single pulsar, with $\mathbf{t}_{\rm det}$ and $\mathbf{t}_{\rm sto}$ the deterministic and stochastic contributions to the total, respectively, where any contributions to the latter will be modeled as random Gaussian processes. In estimating the timing model parameters for the pulsar,

a standard weighted least-squares fit, as performed in packages such as Tempo2, will model the stochastic contributions purely as white noise characterized by the TOA uncertainties. In doing so, a set of prefit timing residuals $\boldsymbol{\delta t}_{\rm pre}$ are produced using an initial estimate of the m timing model parameters β_{0i} such that

$$\boldsymbol{\delta t}_{\rm pre} = \mathbf{t}_{\rm tot} - \mathbf{t}_{\rm det}(\boldsymbol{\beta}_0). \tag{2}$$

From here a linear approximation of the timing model can be used such that any deviations from the initial guess of the timing model parameters are encapsulated using the m parameters ϵ_i such that

$$\epsilon_i = \beta_i - \beta_{0i}. \tag{3}$$

We can therefore write the set of postfit residuals $\boldsymbol{\delta t}$ that arise from this fitting process as

$$\boldsymbol{\delta t} = \boldsymbol{\delta t}_{\rm pre} + \mathbf{M}\boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{M}$ is the $n \times m$ "design matrix" which describes the dependence of the timing residuals on the model parameters. Thus any contribution to $\mathbf{t}_{\rm sto}$ not described by the TOA uncertainties, such as the signal from a GWB, will be absorbed by the timing model fit and so when the timing model is subtracted from the data, any attempt to characterize the power spectrum of the resulting postfit residuals will be incorrect. While some methods exist to model the intrinsic red noise at the point of fitting the timing model (e.g. [21]), and indeed, one can use Tempo2 in conjunction with the methods described in this paper to simultaneously fit for the red noise and the nonlinear timing model, this is not an approach we pursue in the following work.

In order to account for this, we instead begin by following the approach of vHL2013 that we describe in brief here so as to aid subsequent discussion. We begin by assuming that the effect of the additional noise processes beyond the TOA uncertainties on the timing model fit will be small, so that the linear approximation will still hold even in their presence. By refitting for the set of parameters $\boldsymbol{\epsilon}$ we can therefore write the stochastic component of the residuals as

$$\boldsymbol{\delta t}_{\rm sto} = \boldsymbol{\delta t} - \mathbf{M}\boldsymbol{\epsilon}. \tag{5}$$

We can then write the likelihood for the timing residuals as (vH2009)

$$\Pr(\boldsymbol{\delta t}|\boldsymbol{\epsilon}, \boldsymbol{\phi}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \times \exp\left(-\frac{1}{2}(\boldsymbol{\delta t} - \mathbf{M}\boldsymbol{\epsilon})^T \mathbf{C}^{-1} (\boldsymbol{\delta t} - \mathbf{M}\boldsymbol{\epsilon})\right), \tag{6}$$

where the $n \times n$ covariance matrix $\mathbf{C}$ describes the stochastic contributions to the timing residuals such that

$$\langle \delta t_{{\rm sto}_i} \delta t_{{\rm sto}_j} \rangle = C_{ij}, \tag{7}$$

and is described by a set of parameters $\boldsymbol{\phi}$.

We can then marginalize over all variables $\boldsymbol{\epsilon}$ in order to calculate the likelihood of a particular set of parameters $\boldsymbol{\phi}$ for the stochastic contributions to the residuals, i.e.

$$\Pr(\boldsymbol{\delta t}|\boldsymbol{\phi}) = \int {\rm d}^m \boldsymbol{\epsilon} \Pr(\boldsymbol{\epsilon}) \Pr(\boldsymbol{\delta t}|\boldsymbol{\epsilon}, \boldsymbol{\phi}). \tag{8}$$

In vHL2013 this marginalization is performed analytically assuming a uniform prior on $\boldsymbol{\epsilon}$ to give

$$\Pr(\boldsymbol{\delta t}|\boldsymbol{\phi}) = \frac{1}{\sqrt{(2\pi)^{(n-m)} \det(\mathbf{G}^T \mathbf{C} \mathbf{G})}} \times \exp\left(-\frac{1}{2}\boldsymbol{\delta t}^T \mathbf{G} (\mathbf{G}^T \mathbf{C} \mathbf{G})^{-1} \mathbf{G}^T \boldsymbol{\delta t}\right), \tag{9}$$

where $\mathbf{G}$ is a positive-definite symmetric $n \times (n - m)$ matrix, the derivation of which will not be described here.

For the IPTA data challenge, data sets consisted of 130 residuals for 36 pulsars such that $n = 4680$. $\mathbf{G}$ therefore is $\sim 4500 \times 4500$, and so the bottleneck in this calculation comes from the matrix inversion that must occur for every likelihood calculation, along with the set of matrix-matrix multiplications required to calculate $\mathbf{G}^T \mathbf{C} \mathbf{G}$.

Our goal is to remove this obstacle by rephrasing the likelihood such that its evaluation requires no matrix-matrix multiplications and to either eliminate the need to perform computationally intensive [i.e. $O(n^3)$] dense matrix inversions, or to reduce the size of these matrices sufficiently such that their inversion no longer dominates the evaluation time of the likelihood function, while retaining the ability to determine the power spectrum of the stochastic contributions to the residuals.

We do this by first writing our timing residuals $\boldsymbol{\delta t}$ as the sum of a signal $\mathbf{s}$ that we are interested in parametrizing, which will include contributions from both intrinsic red noise and the GWB signal, and some additional white noise $\mathbf{n}$ so that we have

$$\boldsymbol{\delta t} = \mathbf{s} + \mathbf{n}. \tag{10}$$

We can expand $\mathbf{s}$ in terms of its Fourier coefficients $\mathbf{a}$ so that $\mathbf{s} = \mathbf{F}\mathbf{a}$ where $\mathbf{F}$ denotes the Fourier transform such that for frequency ν and time t we will have both

$$F_{\nu,t} = \sin\left(\frac{2\pi}{T}\nu t\right), \tag{11}$$

and an equivalent cosine term. For a single pulsar the covariance matrix $\boldsymbol{\varphi}$ of the Fourier coefficients $\mathbf{a}$ will be diagonal, with components

$$\varphi_{ij} = \langle a_i a_j^* \rangle = \varphi_i \delta_{ij}, \tag{12}$$

where there is no sum over i, and the set of coefficients $\{\varphi_i\}$ represent the theoretical power spectrum for the residuals.

Note that, while this equation states that the Fourier modes are orthogonal to one another, this does not mean that we assume they are orthogonal in the time domain where they are sampled, and we will show explicitly later

that this nonorthogonality is accounted for within the likelihood. Instead, in Bayesian terms, Eq. (12) represents our prior knowledge of the power spectrum coefficients within the data. We are therefore stating that, while we do not know the form the power spectrum will take, we know that the underlying Fourier modes are still orthogonal by definition, regardless of how they are sampled in the time domain. It is here then that, should one wish to fit a specific model to the power spectrum coefficients at the point of sampling, such as a broken, or single power law, the set of coefficients $\{\varphi_i\}$ should be given by some function $f(\Theta)$, where we sample from the parameters Θ from which the power spectrum coefficients $\{\varphi_i\}$ can then be derived.

When dealing with a signal from a stochastic gravitational wave background, however, it is crucial to include the cross-correlated signal between the pulsars on the sky. We do this by using the Hellings-Downs relation [22]:

$$\alpha_{mn} = \frac{3}{2}\frac{1-\cos(\theta_{mn})}{2}\ln\left(\frac{1-\cos(\theta_{mn})}{2}\right) - \frac{1}{4}\frac{1-\cos(\theta_{mn})}{2} + \frac{1}{2} + \frac{1}{2}\delta_{mn}, \tag{13}$$

where θ_{mn} is the angle between the pulsars m and n on the sky and α_{mn} represents the expected correlation between the TOAs given an isotropic background. With this addition our covariance matrix for the Fourier coefficients becomes

$$\varphi_{mi,nj} = \langle a_{mi} a^*_{nj} \rangle = \alpha_{mn}\varphi_i\delta_{ij}, \tag{14}$$

where there is no sum over i, which results in a band diagonal matrix for which calculating the inverse is extremely computationally efficient.

We then write the joint probability density of the power spectrum coefficients and the signal realization $\Pr(\{\varphi_i\}, \mathbf{a}|\boldsymbol{\delta t})$, where here $\mathbf{a}$ refers to the concatenated vector of all coefficients a_i for all pulsars, as

$$\Pr(\{\varphi_i\}, \mathbf{a}|\boldsymbol{\delta t}) \propto \Pr(\boldsymbol{\delta t}|\mathbf{a})\Pr(\mathbf{a}|\{\varphi_i\})\Pr(\{\varphi_i\}) \tag{15}$$

and then marginalize over all $\mathbf{a}$ in order to find the posterior for the parameters $\{\varphi_i\}$ alone. For our choice of $\Pr(\{\varphi_i\})$ we use a uniform prior in $\log_{10}$ space, as the scale of the coefficients is largely unknown below some upper limit, and draw our samples from the parameter $\rho_i = \log_{10}(\varphi_i)$ instead of φ_i, which has the added advantage that we avoid unnecessary rejections due to samples that have negative coefficients in the sampling process. Given this choice of prior the conditional distributions that make up Eq. (15) can be written

$$\Pr(\boldsymbol{\delta t}|\mathbf{a}) \propto \frac{1}{\sqrt{\det(\mathbf{G}^T\mathbf{N}\mathbf{G})}}\exp\left[-\frac{1}{2}(\boldsymbol{\delta t}-\mathbf{F}\mathbf{a})^T \times \mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T(\boldsymbol{\delta t}-\mathbf{F}\mathbf{a})\right], \tag{16}$$

where $\mathbf{N} = \langle \mathbf{n}\mathbf{n}^T \rangle$ and represents the white noise errors in the residuals, which follows from Eq. (9) with $\mathbf{N}$ replacing $\mathbf{C}$, and substituting $\boldsymbol{\delta t} - \mathbf{F}\mathbf{a}$ for $\boldsymbol{\delta t}$, and

$$\Pr(\mathbf{a}|\{\rho_i\}) \propto \frac{1}{\sqrt{\det\boldsymbol{\varphi}}}\exp\left[-\frac{1}{2}\mathbf{a}^{*T}\boldsymbol{\varphi}^{-1}\mathbf{a}\right]. \tag{17}$$

Note that we can calculate $\mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T$ before the sampling starts and store it in memory, which eliminates the need for any dense matrix inversions, or matrix multiplications within the likelihood calculation.

A. Estimating the white noise properties

When dealing with realistic pulsar timing data, the properties of the white noise can be split into two components.

(1) For a given pulsar, each TOA has an associated error bar, the size of which will vary across a set of observations. We can therefore introduce an extra free parameter, an EFAC value, to account for possible miscalibration of this radiometer noise [7]. The EFAC parameter therefore acts as a multiplier for all the TOA error bars for a given pulsar, observed with a particular system.

(2) A second white noise component, independent of the size of the error bars, is also used to represent some additional source of time-independent noise. We call this parameter EQUAD.

In both the IPTA data challenges, and the simulations in Sec. VII, the TOAs for a given pulsar are all assigned a single value for the size of their error bars and so there is no need to include both an EFAC and EQUAD in their analysis, requiring only a single EFAC value per pulsar. Using the likelihood in Eq. (16), despite precalculating the product $\mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T$, we are still able to make inferences about the properties of this scaling factor. Denoting the EFAC parameter for each pulsar p as w_p, we can define a diagonal matrix $\mathbf{W}$ such that, if pulsar p has a set of o_p residuals, and a timing model described by m_p model fit parameters, the first o_1 diagonal elements of $\mathbf{W}$ will equal w_1, the next o_2 diagonal elements will equal w_2, and so on, we can rewrite the product $\mathbf{G}(\mathbf{G}^T\mathbf{W}\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T$. Exploiting the fact that the $\mathbf{G}$ are block diagonal, we can then rewrite this as

$$\begin{aligned}\mathbf{G}(\mathbf{G}^T\mathbf{W}\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T &= \mathbf{G}(\mathbf{W}'\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T \\ &= \mathbf{G}\mathbf{W}'^{-1}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T \\ &= \mathbf{W}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T,\end{aligned} \tag{18}$$

where $\mathbf{W}'$ will be a diagonal matrix where the first $(o_1 - m_1)$ entries are equal to w_1, the next $(o_2 - m_2)$ entries will be equal to w_2, and so on. The determinant of the inverted matrix is then given by

$$\det(\mathbf{W}'\mathbf{G}^T\mathbf{N}\mathbf{G}) = \prod_{p=1}^{N_p} w_p^{(o_p - m_p)} \det(\mathbf{G}^T\mathbf{N}\mathbf{G}), \quad (19)$$

where N_p is the total number of pulsars in the data set. Thus we can store $\mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T$ and the determinant $\det(\mathbf{G}^T\mathbf{N}\mathbf{G})$ in memory and the only additional overhead in the likelihood calculation is the calculation of $\det(\mathbf{W}')$, which is negligible.

For the sake of simplifying our notation we now redefine

$$\tilde{\mathbf{N}}^{-1} = \mathbf{W}^{-1}\mathbf{G}(\mathbf{G}^T\mathbf{N}\mathbf{G})^{-1}\mathbf{G}^T. \quad (20)$$

For more realistic data, where the size of the TOA error bars vary across an observation, and different observing systems are used such that multiple EQUAD and EFAC parameters are desired for the analysis, a slightly different approach is required. Rather than marginalizing over the timing model parameters for each pulsar analytically as in Eq. (16), we can simply perform that marginalization process numerically and so write

$$\Pr(\boldsymbol{\delta t}|\mathbf{a}, \boldsymbol{\epsilon}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{N}}} \exp\Big(-\frac{1}{2}(\boldsymbol{\delta t} - \mathbf{M}\boldsymbol{\epsilon} - \mathbf{F}\mathbf{a})^T \times \mathbf{N}^{-1}(\boldsymbol{\delta t} - \mathbf{M}\boldsymbol{\epsilon} - \mathbf{F}\mathbf{a})\Big). \quad (21)$$

In this way as many white noise parameters can be included as needed; however, this approach will not be pursued further in this paper given, as mentioned previously, the data sets under consideration can be analyzed fully using Eq. (16).

B. Including additional red noise

In order to account for uncorrelated red noise in the pulsar timing residuals we need only modify the covariance matrix $\boldsymbol{\varphi}$ in Eq. (14) by introducing an additional set of parameters $\kappa_{p\nu}$ along the diagonal such that

$$\varphi_{mi,nj} = \alpha_{mn} 10^{\rho_i} \delta_{ij} + 10^{\kappa_{mi}} \delta_{mn} \delta_{ij}, \quad (22)$$

where we then marginalize over all $\kappa_{p\nu}$.

C. Performing the sampling

How we now perform the sampling depends entirely on the number of Fourier coefficients we will be using to describe the stochastic signal in the timing residuals. As we shall see in Secs. VI and VII, even in data sets that exhibit an extremely high signal-to-noise ratio, the number of coefficients required to adequately describe the system is much less than the number of data points in the time domain, often by more than an order of magnitude. This is because practically all the power in the data sets analyzed in these sections comes from only a few low frequency modes that are heavily oversampled in the time domain. In this situation we can marginalize over the Fourier coefficients $\mathbf{a}$ analytically and sample directly from the power spectrum coefficients $\{\boldsymbol{\rho}, \boldsymbol{\kappa}\}$, a process we describe in Sec. III. While this marginalized likelihood function will still include the inversion of a dense matrix, if the number of coefficients sampled is an order of magnitude less than the number of time series data points, then the matrix to be inverted will be an order of magnitude smaller than that in Eq. (9) and will thus take a factor 1000 less time to be inverted.

If, however, we wish to sample over a larger number of Fourier coefficients, to include, for example, higher frequencies where we might expect to observe gravitational wave signals from bright individual sources, then in the limit that we wish to extend our analysis to all frequencies that are Nyquist sampled in the data, the matrix to be inverted when performing the marginalization analytically will be of the same size as that in Eq. (9), and we will have the same computational burden as when performing the analysis in the time domain. In this situation we can perform the marginalization numerically, sampling directly from the high dimension, joint probability distribution described in Eq. (15), a process made possible through the use of a GHS (B13), which we describe in the Sec. IV.

III. THE LOW SIGNAL-TO-NOISE REGIME: ANALYTICAL MARGINALIZATION OVER THE FOURIER COEFFICIENTS

In order to perform the marginalization over the Fourier coefficients $\mathbf{a}$, we first write the log of the likelihood in Eq. (15), which denoting $(\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\mathbf{F} + \boldsymbol{\varphi}^{-1})$ as $\boldsymbol{\Sigma}$ and $\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t}$ as $\mathbf{d}$ is given by

$$\log \mathrm{L} = -\frac{1}{2}\boldsymbol{\delta t}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t} - \frac{1}{2}\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} + \mathbf{d}^T\mathbf{a}. \quad (23)$$

Taking the derivative of $\log \mathrm{L}$ with respect to $\mathbf{a}$ gives us

$$\frac{\partial \log \mathrm{L}}{\partial \mathbf{a}} = -\boldsymbol{\Sigma}\mathbf{a} + \mathbf{d}^T, \quad (24)$$

which can be solved to give us the maximum likelihood vector of coefficients $\hat{\mathbf{a}}$:

$$\hat{\mathbf{a}} = \boldsymbol{\Sigma}^{-1}\mathbf{d}^{\mathbf{T}}. \quad (25)$$

Reexpressing Eq. (23) in terms of $\hat{\mathbf{a}}$,

$$\log \mathrm{L} = -\frac{1}{2}\boldsymbol{\delta t}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t} + \frac{1}{2}\hat{\mathbf{a}}^T\boldsymbol{\Sigma}\hat{\mathbf{a}} - \frac{1}{2}(\mathbf{a} - \hat{\mathbf{a}})^T\boldsymbol{\Sigma}(\mathbf{a} - \hat{\mathbf{a}}), \quad (26)$$

the third term in this expression can then be integrated with respect to the m elements in $\mathbf{a}$ to give

$$I = \int_{-\infty}^{+\infty} \mathrm{d}\mathbf{a} \exp\Big[-\frac{1}{2}(\mathbf{a} - \hat{\mathbf{a}})^T\boldsymbol{\Sigma}(\mathbf{a} - \hat{\mathbf{a}})\Big] = (2\pi)^m \det \boldsymbol{\Sigma}^{-\frac{1}{2}}. \quad (27)$$

Our marginalized probability distribution for a set of GWB coefficients is then given as

$$\Pr(\{\varphi_i\}|\boldsymbol{\delta t}) \propto \frac{\det(\boldsymbol{\Sigma})^{-\frac{1}{2}}}{\sqrt{\det(\boldsymbol{\varphi})\det(\tilde{\mathbf{N}})}} \times \exp\left[-\frac{1}{2}(\boldsymbol{\delta t}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t} - \mathbf{d}^T\boldsymbol{\Sigma}^{-1}\mathbf{d})\right], \quad (28)$$

where we can still precalculate both $\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\mathbf{F}$ and $\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t}$.

Equation (28) shows that the covariance matrix $\boldsymbol{\Sigma}$ both acts to whiten residuals and fully describes the nonorthogonality in the Fourier modes due to uneven sampling in the time domain. This, in combination with the marginalization over the timing model parameters included in $\tilde{\mathbf{N}}$, which includes a quadratic in t that describes the pulsar spin-down, and acts to project out any contribution from those frequencies lower than we can properly sample in the data means that no additional prewhitening steps are required by this method. Demonstrably this will be shown to have the desired effect; even for the data sets described in the Sec. VII, where we have large gaps in the data (~5-yr gaps in a 20-yr data set) we extract the correct power spectrum.

To perform the parameter estimation with this method we will then use the MULTINEST algorithm [23,24], which will simultaneously allow us to calculate the evidence for increasing numbers of Fourier modes until a maximum is reached, and to test whether or not the data supports the inclusion of additional red noise parameters.

For large numbers of Fourier modes, however, performing this marginalization analytically and sampling using MULTINEST no longer remains a viable option due to both the scaling of the matrix inversions required and the performance scaling of MULTINEST with dimensionality. In the following section we therefore describe a method for performing this marginalization numerically using a GHS, while in Sec. V we describe two possible options for estimating the evidence for different numbers of Fourier modes in order to find the optimal set.

We note that, in principle, one could also use the GHS when marginalizing analytically, where the superior scaling of the GHS with dimensionality when compared to MULTINEST could allow for the inclusion of greater numbers of power spectrum coefficients. Ultimately, however, this approach is still limited by the scaling of the matrix inversions and so we do not pursue this idea further.

IV. GUIDED HAMILTONIAN SAMPLING

For a detailed account of both Hamiltonian Monte Carlo (HMC) and GHS refer to (B13); or Appendix A, here we will describe only the key aspects of each. HMC sampling [25] has been widely applied in Bayesian computation [26] and has been successfully applied to problems with extremely large numbers of dimensions (~10^6 see e.g. [27]). Where conventional Markov Chain Monte Carlo methods move through the parameter space by a random walk and therefore require a prohibitive number of samples to explore high-dimensional spaces, HMC draws parallels between sampling and classical dynamics. By exploiting techniques developed for describing the motion of particles in potentials, it is possible to suppress random walk behavior. Introducing persistent motion of the chain through the parameter space allows HMC to maintain a reasonable efficiency even for high-dimensional problems.

We define a "potential energy" Ψ that is related to our posterior distribution $\Pr(\mathbf{x})$ by

$$\Psi(\mathbf{x}) = -\ln(\Pr(\mathbf{x})), \quad (29)$$

where $\mathbf{x}$ is the N-dimensional vector of parameters to be sampled. Each parameter x_i must be assigned a mass m_i and a momentum p_i so that we can write our Hamiltonian as

$$H = \sum_i \frac{p_i^2}{2m_i} + \Psi(\mathbf{x}). \quad (30)$$

The sampler is given a start point $\mathbf{x}$ and a set of initial momenta $\mathbf{p}$, which are drawn from a set of N-uncorrelated Gaussian distributions of width m_i in dimension i. The system can then evolve deterministically from then for some length of time τ using Hamilton's equations.

After it has reached its new position $(\mathbf{x}', \mathbf{p}')$, that point will be accepted with a probability

$$p = \min[1, \exp(-\delta H)], \quad (31)$$

where $\delta H = H(\mathbf{x}', \mathbf{p}') - H(\mathbf{x}, \mathbf{p})$. A new set of momenta can then be drawn and the process repeats. This implies that if we are able to integrate Hamilton's equations exactly then, as energy is conserved along such a trajectory, the probability of acceptance is unity. In practice, however, numerical inaccuracies mean that this is not the case.

In order to perform the integration along the systems trajectory at each state we use a "leapfrog" method as is common practice. Here n_s steps are taken of size λ such that $n_s\lambda = \tau$ such that

$$p_i\left(t + \frac{\lambda}{2}\right) = p_i(t) - \frac{\lambda}{2}\frac{\partial\Psi(\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}(\mathbf{t})}, \quad (32)$$

$$x_i(t+\lambda) = x_i(t) + \frac{\lambda}{m_i}p_i\left(t + \frac{\lambda}{2}\right), \quad (33)$$

$$p_i(t+\lambda) = p_i\left(t + \frac{\lambda}{2}\right) - \frac{\lambda}{2}\frac{\partial\Psi(\mathbf{x})}{\partial x_i}\bigg|_{\mathbf{x}(t+\lambda)}, \quad (34)$$

until $t = \tau$ where τ is varied to avoid resonant trajectories. HMC thus requires a large number of adjustable parameters, the mass m_i, step size λ_i, and the number of steps n_s in the trajectory. Adjusting the step size or the mass produces similar effects [28] and so one is usually fixed and the other tuned during sampling.

GHS is designed to eliminate much of the remaining tuning aspect by using the Hessian $\hat{\mathbf{H}}$ of the joint probability distribution calculated at its peak to set the step size λ for each parameter. The masses m_i are then set to unity and the only tunable parameter that remains is a global scaling parameter for the step size η that is chosen such that the acceptance rate for the GHS is $\sim$68%.

Therefore in order to perform sampling we need the following:

(i) the gradient of Ψ for each parameter x_i
(ii) the peak of the joint distribution, and
(iii) the Hessian at that peak.

The gradients of our parameters are given by the following:

$$\frac{\partial \Psi}{\partial \mathbf{a}} = -(\boldsymbol{\delta t} - \mathbf{F}\mathbf{a})^T \tilde{\mathbf{N}}^{-1}\mathbf{F} + \mathbf{a}^T \boldsymbol{\varphi}^{-1}, \tag{35}$$

$$\frac{\partial \Psi}{\partial w_i} = \frac{1}{2w_i}(o_i - m_i) - \frac{1}{w_i}(\boldsymbol{\delta t}_i - \mathbf{F}_i\mathbf{a}_i)^T \tilde{\mathbf{N}}^{-1}(\boldsymbol{\delta t}_i - \mathbf{F}_i\mathbf{a}_i), \tag{36}$$

$$\frac{\partial \Psi}{\partial \rho_i} = \frac{1}{2}\,\mathrm{Tr}\left(\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\right) - \frac{1}{2}\mathbf{a}^T \boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\mathbf{a}, \tag{37}$$

and the components of the Hessian are

$$\frac{\partial^2 \Psi}{\partial \mathbf{a}^2} = \mathbf{F}^T \tilde{\mathbf{N}}^{-1}\mathbf{F} + \boldsymbol{\varphi}^{-1}, \tag{38}$$

$$\frac{\partial^2 \Psi}{\partial w_i^2} = \frac{1}{w_i^2}(o_i - m_i) + \frac{2}{w_i^2}(\boldsymbol{\delta t}_i - \mathbf{F}_i\mathbf{a}_i)^T \tilde{\mathbf{N}}^{-1}(\boldsymbol{\delta t}_i - \mathbf{F}_i\mathbf{a}_i), \tag{39}$$

$$\frac{\partial^2 \Psi}{\partial \rho_i^2} = \mathbf{a}^T \boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\mathbf{a} - \frac{1}{2}\mathbf{a}^T \boldsymbol{\varphi}^{-1}\frac{\partial^2 \boldsymbol{\varphi}}{\partial \rho_i^2}\boldsymbol{\varphi}^{-1}\mathbf{a}, \tag{40}$$

$$\frac{\partial^2 \Psi}{\partial \rho_i \partial \mathbf{a}} = -\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\mathbf{a}. \tag{41}$$

For a set of power spectrum coefficients $\{\rho_i, \kappa_i\}$ and white noise coefficients $\{\Sigma_i\}$ we can solve for the maximum set of Fourier coefficients $\mathbf{a}_{\max}$ analytically using Eq. (25) so when searching for the global maximum we need only search over the subset of parameters $\{\rho_i, \Sigma_i, \kappa_i\}$. This is achieved by using either a particle swarm algorithm ([29,30] and for uses in cosmological parameter estimation see e.g. [31]; for a description of the particle swarm method applied to PTA data in this context see [32]) or using a gradient search optimization [33]. In the work to follow we use the former method and take an iterative approach, passing the maximum likelihood value at the end of a search to one of the particles as a starting point for the next iteration, enabling us to find the maximum using only 1 core per $\sim$10 free parameters.

V. DETERMINING THE OPTIMAL NUMBER OF FOURIER MODES

In the low signal-to-noise regime, where we need only sample small numbers of Fourier coefficients, we are able to use MULTINEST to calculate the evidence directly and thus determine the optimal number of frequencies to describe the data by choosing the set for which the evidence is maximized. When we wish to sample greater numbers of Fourier coefficients, so the dimensionality of the problem is large, this approach is no longer computationally practical. While in principle we could ensure that we always include a sufficient number of coefficients so that our model is able to correctly describe the data simply by including all possible Fourier coefficients, this will in most cases be suboptimal. Therefore we would like to perform model selection between models where we include different sets of frequencies $\{\mathbf{w}\}$ prior to sampling by maximizing an approximation to the evidence with respect to the set $\{\mathbf{w}\}$, and use that set for the analysis that follows.

We do this in two ways: first by considering the Laplace approximation (e.g. [34]) of the marginalized posterior given by Eq. (28), and second by considering the analytical evaluation of the evidence for an approximate likelihood function. We then compare the results of applying these two approaches to the result calculated using MULTINEST for each of the IPTA data challenges in Sec. VI.

A. Laplace approximation

Given a model with a set of m maximum likelihood parameters $\hat{\boldsymbol{\rho}}$ we can approximate the likelihood around the peak using a Gaussian such that given a different set of parameters $\boldsymbol{\rho}$ we can write

$$\Pr(\boldsymbol{\delta t}|\boldsymbol{\rho}, m)\Pr(\boldsymbol{\rho}, m) \approx \mathrm{P}(\hat{\boldsymbol{\rho}})\Pr(\hat{\boldsymbol{\rho}}, m)\exp\left[-\frac{1}{2}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}})^T \hat{\mathbf{H}}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}})\right], \tag{42}$$

where $\hat{\mathbf{H}}$ is the Hessian of the negative log likelihood evaluated at the peak as before. This can be integrated with respect to $\boldsymbol{\rho}$ to give the Laplace approximation to the evidence given the set of model parameters m:

$$\Pr(\boldsymbol{\delta t}|m) \propto (2\pi)^{m/2}\det \hat{\mathbf{H}}^{-1/2}\mathrm{P}(\hat{\boldsymbol{\rho}})\Pr(\hat{\boldsymbol{\rho}}, m). \tag{43}$$

Denoting $(\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\mathbf{F} + \boldsymbol{\varphi}^{-1})^{-1}$ as $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{F}^T\tilde{\mathbf{N}}^{-1}\boldsymbol{\delta t}$ as $\mathbf{d}$ as before, we can write the first derivative of $\Psi = -\log\Pr(\{\rho_i\}|\boldsymbol{\delta t})$ as

$$\frac{\partial \Psi}{\partial \rho_i} = \frac{1}{2}\,\mathrm{Tr}\left(\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\right) - \frac{1}{2}\mathbf{d}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial \boldsymbol{\varphi}}{\partial \rho_i}\boldsymbol{\varphi}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{d}. \tag{44}$$

In order to estimate the number of coefficients $\boldsymbol{\rho}$ to be used, we then assume that all the signal in the data for the set of

N_p pulsars is the result of a GWB so that this simplifies slightly to

$$\frac{\partial\Psi}{\partial\rho_i} = \frac{1}{2}\log(10)N_p - \frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\right) - \frac{1}{2}\mathbf{d}^T\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\mathbf{\Sigma}^{-1}\mathbf{d}. \quad (45)$$

Writing $\bar{\mathbf{d}}^T = \mathbf{d}^T\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}$ our Hessian is therefore given by

$$\frac{\partial^2\Psi}{\partial\rho_i^2} = \frac{1}{2}\mathrm{Tr}\left(-\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1} + \mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial^2\boldsymbol{\varphi}}{\partial\rho_i^2}\boldsymbol{\varphi}^{-1}\right) - \bar{\mathbf{d}}^T\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\bar{\mathbf{d}} + \bar{\mathbf{d}}^T\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\bar{\mathbf{d}} - \frac{1}{2}\bar{\mathbf{d}}^T\frac{\partial^2\boldsymbol{\varphi}}{\partial\rho_i^2}\bar{\mathbf{d}}, \quad (46)$$

$$\frac{\partial^2\Psi}{\partial\rho_i\partial\rho_j} = \frac{1}{2}\mathrm{Tr}\left(-\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_j}\boldsymbol{\varphi}^{-1}\right) - \bar{\mathbf{d}}^T\frac{\partial\boldsymbol{\varphi}}{\partial\rho_i}\boldsymbol{\varphi}^{-1}\mathbf{\Sigma}^{-1}\boldsymbol{\varphi}^{-1}\frac{\partial\boldsymbol{\varphi}}{\partial\rho_j}\bar{\mathbf{d}}. \quad (47)$$

We can thus use Eqs. (46) and (47) to evaluate expression (43) and approximate the evidence. While this calculation requires that we calculate the maximum likelihood values for incremental numbers of parameters m, we believe that in any practical data set this will still prove less costly than performing the analysis using the full set of Fourier coefficients present in the data.

B. Approximating the likelihood

We now take a second alternate approach to the subject of model selection, by considering a simpler problem for which we can calculate the evidence directly. We begin with a simple example where for some time series data $\mathbf{d}$ of length N with uniform white noise we would like to determine the number of basis functions that the data can support as derived in [35]. We include the complete derivation of the results given in this section in Appendix B; however, below we include only a brief outline.

1. Uniform white noise

Suppose we have a single realization of some time series data $\mathbf{d}$ of length N. We then define a set of hypotheses $\{H\}$ such that each H_i purports that our data $\mathbf{d}$ is described by some function f_i where

$$f_i(t) = \sum_{k=1}^{m} b_k M_k(t, \mathbf{w}), \quad (48)$$

with M_k a set of general basis functions. The number of functions m, the parameters that describe them (e.g. their frequencies) $\mathbf{w}$, and the model coefficients b_k are allowed to vary for each f_i. We then transform this set of basis functions into an orthonormal set F_k through the transformation

$$F_k(t) = \frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{m} e_{kj}M_j(t), \quad (49)$$

where e_{kj} is the kth element of the jth eigenvector and λ_k is the kth eigenvalue of the covariance matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$. Our function f_i can now be written in terms of these new basis vectors,

$$f_i(t) = \sum_{k=1}^{m} a_k F_k(t, \mathbf{w}), \quad (50)$$

where the coefficients a in the orthonormal basis are related to the coefficients b in the original basis through

$$b_k = \sum_{j=1}^{m}\frac{a_k e_{jk}}{\sqrt{\lambda_j}}. \quad (51)$$

The probability of the data given a model f_i, assuming that the noise is described by a zero mean random Gaussian process with variance σ is given by

$$\Pr(\mathbf{d}|\mathbf{a}, \mathbf{w}, \sigma, f_i) = (2\pi\sigma^2)^{-N/2}\exp\left[\frac{1}{2\sigma^2}\sum_{k=1}^{N}[d_k - f_i(t_k)]^2\right]. \quad (52)$$

We begin by integrating over both the set of coefficients $\mathbf{a}$ and frequencies $\mathbf{w}$. We assume that the two parameters are logically independent, in so far as we can write the priors:

$$\Pr(\mathbf{a}, \mathbf{w}) = \Pr(\mathbf{a})\Pr(\mathbf{w}). \quad (53)$$

For the amplitude coefficients, we choose an uninformative Gaussian prior given by

$$\Pr(\mathbf{a}|\delta) = (2\pi\delta^2)^{-m/2}\exp\left[-\sum_{k=1}^{m}\frac{a_k^2}{2\delta^2}\right], \quad (54)$$

with $\delta \gg \sigma$. For our frequencies, we consider that for any given model f_i we are selecting a set of frequencies chosen from an evenly spaced grid. Therefore we will have a delta function prior for each frequency w_j in the set $\mathbf{w}$ and thus arrive at the expression

$$\Pr(\mathbf{d}|\delta, \sigma, f_i) = (2\pi\delta^2)^{-m/2}(2\pi\sigma^2)^{-(N-m)/2} \times \exp\left[\frac{\mathbf{d}^2 - \mathbf{h}(\mathbf{w_i})^2}{2\sigma^2}\right]\exp\left[\frac{\mathbf{h}(\mathbf{w_i})^2}{2\delta^2}\right]. \quad (55)$$

We are now in a position to integrate over our unknown variances σ and δ. As in [35], we set an upper bound H and lower bound L to this integral, which will therefore be of the form

$$\frac{1}{\log(H/L)}\int_L^H \mathrm{d}s\,\frac{s^{-a}\exp\left[-\frac{Q}{s^2}\right]}{s}. \quad (56)$$

Making a substitution $u = Q/s^2$, this becomes

$$\frac{Q^{-a/2}}{2\log(H/L)}\int_{Q/H^2}^{Q/L^2} \mathrm{d}u u^{a/2-1}\exp[-u]. \quad (57)$$

If we assume that H is sufficiently large, and L is sufficiently small that we may write $Q/H^2 \ll 1$ and $a/2-1 \ll Q/L^2$, then the integral will evaluate to approximately $\Gamma(a/2)$. Therefore we can finally write the probability of the data D given a model f_i as

$$\Pr(\mathbf{d}|f_i) = \frac{\Gamma(m/2)}{2\log(R_\delta)}\left[\frac{\mathbf{h}(\mathbf{w})^2}{2}\right]^{-m/2}\frac{\Gamma((N-m)/2)}{2\log(R_\sigma)} \times\left[\frac{\mathbf{d}^2-\mathbf{h}(\mathbf{w})^2}{2}\right]^{-(N-m)/2}. \quad (58)$$

2. Nonuniform white noise

In general when dealing with pulsar residuals the white noise level across a data set for a single pulsar will vary with time, where for example different instruments have been used to collect data for the same pulsar. In this case the expansion of our likelihood function is not so simple, because the covariance matrix $\mathbf{G}^T\mathbf{N}\mathbf{G}$ will no longer reduce to a diagonal matrix. If we define $\mathbf{C} = \mathbf{G}^T\mathbf{N}\mathbf{G}$ where we consider $\mathbf{C}$ to be a general dense covariance matrix, Eq. (52) will take the form

$$\Pr(\mathbf{d}|\mathbf{a},\mathbf{w},f_i) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}\exp\left[\frac{-1}{2}(\mathbf{d}-\mathbf{F}\mathbf{a})^T\mathbf{C}^{-1}(\mathbf{d}-\mathbf{F}\mathbf{a})\right]. \quad (59)$$

As in Sec. II A, we would like to fit for a global scaling factor that modifies the overall noise level in the data set, i.e. we would like to write $\mathbf{C}' = \mathbf{G}^T(\alpha^2\mathbf{N})\mathbf{G}$ where α is a constant to be determined. Taking the same priors as the uniform noise case described previously, and following a similar process to integrate over the Fourier coefficients $\boldsymbol{a}$, frequencies $\boldsymbol{w}$, and variances α and δ, we arrive at the final probability for a set of m functions f_i:

$$\Pr(\mathbf{d}|f_i) = \frac{\Gamma(m/2)}{2\log(R_\delta)}\left[\frac{1}{2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}'^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}'^{-1}\mathbf{F_i}}\right)^2\right]^{-m/2} \times\frac{\Gamma((N-m)/2)}{2\log(R_\alpha)}\left[-\frac{1}{2}(\mathbf{d}^T\bar{\mathbf{C}}^{-1}\mathbf{d})\right]^{-(N-m)/2}, \quad (60)$$

where we have defined

$$\bar{\mathbf{C}}^{-1} = \mathbf{C}'^{-1} - \mathbf{C}'^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}'^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}'^{-1}. \quad (61)$$

VI. THE IPTA DATA CHALLENGE

We will now apply the three methods discussed thus far to the first IPTA data challenge. Henceforth we will refer to the numerical marginalization using the GHS as method (A), the analytical marginalization using MULTINEST as method (B), and the approach of fitting directly for a model power spectrum, where we use a power law model of the form $\mathrm{P}(f) = Af^{-\gamma}$, as method (C). Each of these methods will therefore be sampling a different number of parameters, which for clarity we outline explicitly below:

Method (A): With the exception of closed data set 3, we are simultaneously parametrizing the white noise for each pulsar (N_p dimensions), a set of n GWB coefficients, and ($N_p \times n \times 2$) Fourier coefficients. For closed data set 3 we also include an additional set of ($N_p \times n$) coefficients to allow for red noise parametrization such that we allow different pulsars to have different red noise spectra.

Method (B): For method (B) we are parametrizing a set of n GWB coefficients only, with the exception of closed data set 3 where we include an additional n parameters to describe the average red noise across the pulsars, where we assume the data set has used a single power spectrum model for all pulsar realizations as in the open 3 data set. In all cases we assume the level of the white noise in the data set is consistent with that given for the TOAs in the data files.

Method (C): For method (C) we directly parametrize the slope and amplitude of the gravitational wave signal in the data using a power law model of the form $\mathrm{P}(f) = Af^{-\gamma}$, resulting in only two dimensions per data set, with the exception of closed data set 3 where we include an additional two parameters to describe the average amplitude and slope of the red noise properties in the data. In all cases we assume the level of the white noise in the data set is consistent with that given for the TOAs in the data files.

In total there are six data sets in the IPTA data challenge, three of which comprise the "open" challenge, where the properties of the injected signals are known prior to analysis, and three which make up the "closed" challenge, where at the time of analysis the details were unknown. We will outline the properties of these data sets below.

Where present in the data, the injected GWB power spectrum has a characteristic strain spectrum given by

$$h_c(f) = A_g\left(\frac{f}{1\ \mathrm{yr}^{-1}}\right)^{\alpha}, \quad (62)$$

with A_g a dimensionless amplitude at a frequency of (yr^{-1}) and α a power law index. Parametrizing the spectral density as in vHL2013,

$$S(f) = A^2\left(\frac{1}{1\ \mathrm{yr}^{-1}}\right)\left(\frac{f}{1\ \mathrm{yr}^{-1}}\right)^{-\gamma}, \quad (63)$$

the strain spectrum will result in an observed spectral density within the residuals of

$$S(f) = \frac{A_g^2}{12\pi^2}1\ \mathrm{yr}^3\left(\frac{f}{1\ \mathrm{yr}^{-1}}\right)^{-\gamma}, \quad (64)$$

where in both instances $\gamma = 2\alpha - 3$. The parameters of the open and closed data sets are listed below.

Open challenge 1: 36 pulsars with 130 observations each evenly sampled in time. Each data set has white noise with an amplitude of 10^{-7} s and an injected GWB signal with $A_g = 5 \times 10^{-14}$ and $\gamma = 13/3$.

Open challenge 2: As open challenge 1, but the sampling in the time domain is no longer even, and the amplitude of the white noise varies between different pulsars in the range $\sim 10^{-8} \rightarrow 10^{-6}$ s.

Open challenge 3: As open challenge 2, but now $A_g = 10^{-14}$, and there is an additional red noise signal present in each data set of the form $\mathrm{P}(f) = Af^{-\gamma_{\mathrm{red}}}$ where $A = 5.77 \times 10^{-22}$ $\mathrm{s}^{1.3}$ and $\gamma_{\mathrm{red}} = 1.7$.

Closed challenge 1: As open challenge 1, with the injected GWB signal parameters changed to $A_g = 1 \times 10^{-14}$ and $\gamma = 13/3$.

Closed challenge 2: As open challenge 2, with the injected GWB signal parameters changed to $A_g = 6 \times 10^{-14}$ and $\gamma = 13/3$.

Closed challenge 3: As open challenge 3, but now $A_g = 5 \times 10^{-15}$, and the red noise signal present in each data set is given by $A = 3.66 \times 10^{-18}$ $\mathrm{s}^{1.8}$ and $\gamma_{\mathrm{red}} = 1.2$.

In analyzing the data, we chose a fundamental frequency f_0 to be equal to $1/T_{\mathrm{max}}$, where T_{max} represents the greatest observing time span for any of the pulsars in the data set. Defining $f_n = n/T_{\mathrm{max}}$, we then fit for the coefficients corresponding to some set of $\{n\}$ Fourier modes.

In order to determine the optimal set of Fourier modes to include for each data set for method (A) we use both the Laplace approximation and analytic approximation methods described in Secs. V A and V B, respectively. Figure 1 shows an example of the analytic approximation applied to one pulsar from each of the three open data sets. The red line shows how the evidence changes as the number of frequencies in the model increases, while the blue dotted and green dashed lines show the injected level and the best estimate of the rms amplitude for the white noise in the data for each model, where the latter is calculated using the expression in [35] as

TABLE I. Number of frequencies supported by the evidence for the IPTA data challenges.

	Optimal Number of Frequencies		
Data set	Laplace	Analytic	MULTINEST
Open 1	11	9	9
Open 2	15	12	11
Open 3	9	6	9
Closed 1	6	5	6
Closed 2	17	13	17
Closed 3	9	8	8 (r)

$$\langle \sigma^2 \rangle = \frac{1}{N - m - 2}(\mathbf{d}^2 - \mathbf{h}^2). \tag{65}$$

In all three cases the evidence can be seen to reach its maximum when the change in the estimated rms amplitude no longer justifies an increase in the number of model parameters. Since we wish to include all relevant frequencies, we therefore choose the maximum number of frequencies supported by any single pulsar as the set of frequencies to sample for the GWB.

The values for these approaches are given in Table I for the three open and three closed IPTA challenge data sets where those data sets for which the evidence supported the inclusion of additional red noise are marked with an (r).

A comparison of the three methods shows that while the analytical estimate performs well in four of the six data sets, for both closed 2 and open 3 there is a marked underestimate in the optimal number of coefficients suggested. The change in the log evidence calculated using MULTINEST going from 13 to 17 coefficients in closed data set 2 is $\Delta \log E = 13$ while going from 6 to 9 coefficients in open data set 3 resulted in an increase of $\Delta \log E = 7$, both representing significant losses of information for not including the additional coefficients. While the analytical approximation to the likelihood would likely hold in the case where the signal is dominated by uncorrelated red noise in the individual pulsars, here we see that the additional information gained through the coherence between

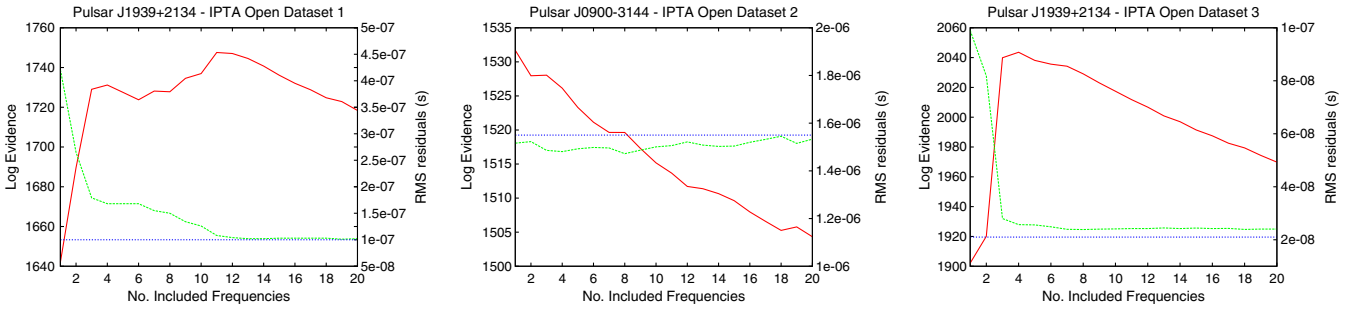


FIG. 1 (color online). Calculated using the analytical approximation to the likelihood described in Sec. V B we plot the evidence (red solid line) for models with different numbers of frequency modes, and the rms residuals (green dashed line) compared with the injected value (blue dotted line) for those models. Examples are given for open data set 1 (left), 2 (middle), and 3 (right) where the evidence is maximized for 11, 1, and 4 frequencies for each, respectively.

pulsars is enough to warrant additional Fourier coefficients in the analysis. In comparison, the Laplace approximation agrees well with the results found using MULTINEST in all six data sets. For the later simulations we will therefore take this approach; however, for the IPTA data sets all the results in the following section are derived using the number of Fourier modes found to be optimal via the numerical analysis using MULTINEST.

A. Results

Table II summarizes the results for the six IPTA data sets for methods (A), (B), and (C) described in this paper, and also the method described in vHL2013. For methods (A) and (B) we give the best fit values and errors for both the dimensionless amplitude A_g and the power law index γ that results from a weighted least squares fit to the 1D GWB power coefficients for each of the IPTA data sets, while for method (C) and that from vHL2013 we give the values of A_g and γ estimated directly from the data and the errors returned by MULTINEST. For comparison we also include the injected values of the GWB spectrum for each data set. Figures 2–4 then show a more detailed representation of the results from the open data challenges. In each figure the top left panel shows a log-log plot of the parametrized GWB power spectrum coefficients for that data set. The red and green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors for methods (A) and (B), respectively. For clarity we have offset the frequency position for method (B) but for the analysis both methods were evaluated for the same frequencies. The blue points represent the injected values for those coefficients, while the dashed blue and purple lines show the best fit power spectrum to the marginalized coefficients for methods (A) and (B), respectively. The top right panel then shows the parametrized values for the white noise in each pulsar in that data set. For open data set 2 and 3, where the pulsars each have a different white noise level, the injected value is indicated by the green crosses while the parametrized values are shown by the red points with their respective errors. The lower plot in each figure shows the one- and two-dimensional marginalized posteriors for the GWB power spectrum coefficients $\{\rho_i\}$ from method (B) fitted for that data set with vertical lines in the 1D distributions representing the power in the injected background at the frequency of that coefficient. Contours in the 2D plots represent 68% and 95% confidence levels. For the 3 closed data we show only the parametrized GWB power spectrum coefficients from methods (A) and (B) in red (solid) and green (dashed), respectively, for each data set in Fig. 5, and the injected values for each coefficient in blue.

The predominant message from these results is that for all the data sets methods (A)–(C) are all able to extract the correct power spectrum from the data with the same fidelity as the method in vHL2013. Comparing our results with those in [36], where the data compression method of vH2013 is applied to the IPTA closed data sets, we likewise see consistency between the values and precision of the inferred parameters. This is true despite the fact that methods (A) and (B) at no stage prescribe any form for the shape of the power spectrum, which we believe is the only correct way to perform an analysis of this kind where the true shape of the power spectrum is unknown.

B. Discussion

1. Run times

Table III shows a comparison of the run times for the three different sampling methods presented in this paper, and for the method described in vHL2013, when using a single 16 core Sandy Bridge node on the high performance computer "DARWIN". For our implementation of the method in vHL2013 we use the same number of free parameters as for method (C) described at the start of Sec. VI. In every case method (C) is 100–1000 times faster than the method described in vHL2013, precisely what we would expect given the order of magnitude decrease in the size of the covariance matrix that requires inverting when compared to the time domain analysis. Comparing the run times between methods (A) and (B) we can see at what point the numerical marginalization becomes favorable over the analytical form. Below ~15 coefficients performing the marginalization analytically is clearly the preferred choice, being a factor of a few faster than performing the process numerically; however, the increase in the number of calculations required for convergence, combined with the $O(n^3)$ scaling of the matrix inversion means that beyond this point it rapidly begins to lose out, ultimately

TABLE II. IPTA data challenge results.

	This paper (A)		This paper (B)		This paper (C)		vHL2013		Injected values	
Data set	$A_g \times 10^{-14}$	γ	$A_g \times 10^{-14}$	γ	$A_g \times 10^{-14}$	γ	$A_g \times 10^{-14}$	γ	$A_g \times 10^{-14}$	γ
Open 1	5.1 ± 0.2	4.34 ± 0.10	4.62 ± 0.19	4.30 ± 0.08	4.6 ± 0.2	4.32 ± 0.09	4.82 ± 0.18	4.4 ± 0.08	5	4.333
Open 2	5.2 ± 0.3	4.36 ± 0.12	5.1 ± 0.3	4.36 ± 0.11	5.4 ± 0.3	4.29 ± 0.12	5.5 ± 0.3	4.30 ± 0.09	5	4.333
Open 3	1.08 ± 0.12	4.2 ± 0.2	1.08 ± 0.12	4.17 ± 0.2	1.09 ± 0.13	4.13 ± 0.20	1.17 ± 0.13	4.13 ± 0.19	1	4.333
Closed 1	1.07 ± 0.05	4.2 ± 0.2	1.12 ± 0.13	4.36 ± 0.08	1.07 ± 0.11	4.25 ± 0.19	1.11 ± 0.09	4.31 ± 0.15	1	4.333
Closed 2	5.6 ± 0.3	4.40 ± 0.12	5.6 ± 0.3	4.36 ± 0.08	5.59 ± 0.28	4.4 ± 0.11	6.32 ± 0.15	4.27 ± 0.05	6	4.333
Closed 3	0.32 ± 0.09	4.5 ± 0.4	0.32 ± 0.09	4.2 ± 0.3	0.44 ± 0.08	4.0 ± 0.3	0.5 ± 0.16	4.2 ± 0.4	0.5	4.333

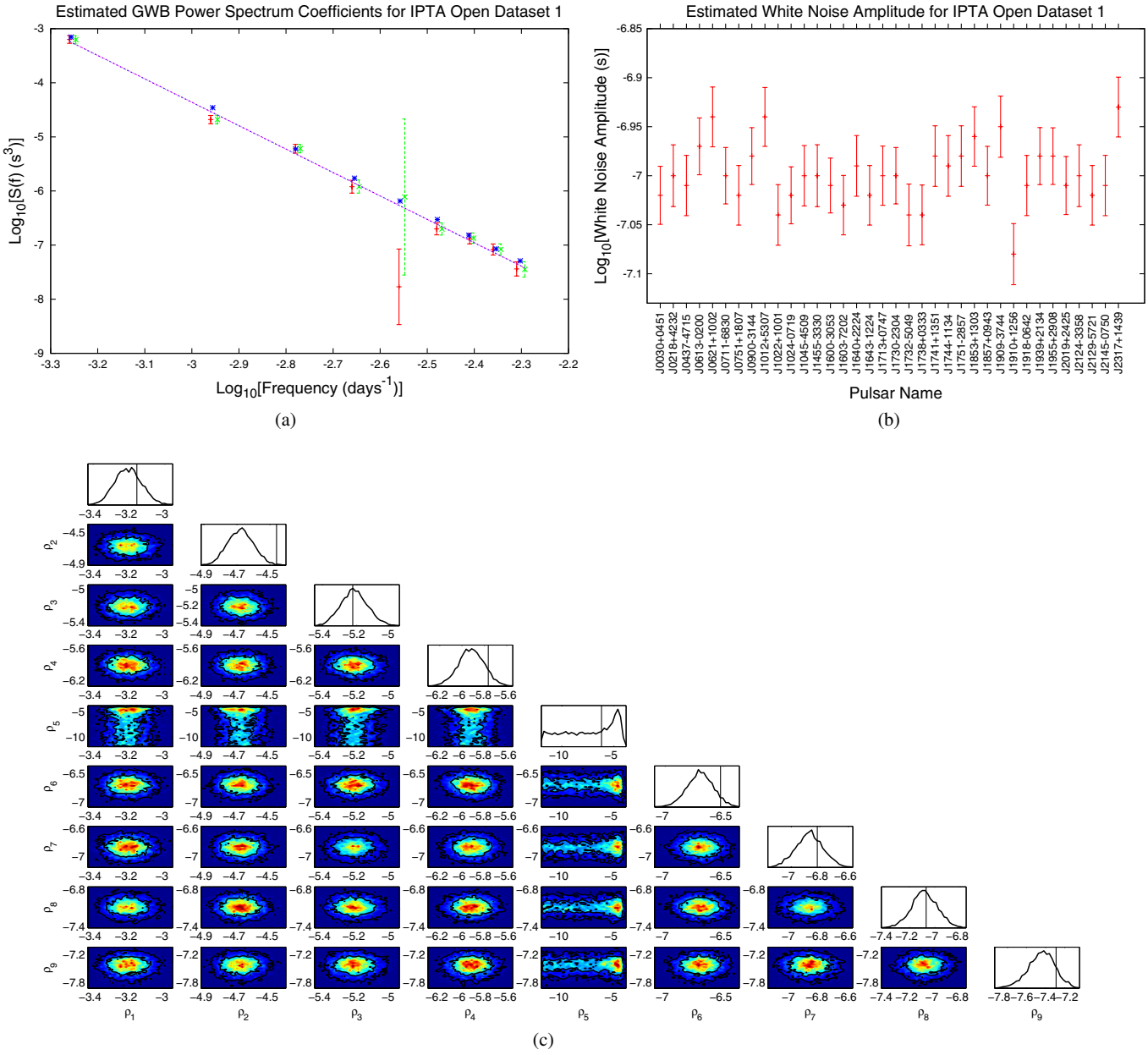


FIG. 2 (color online). (a) Log-log plot of the parametrized GWB power spectrum in open data set 1. The red and green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors for methods (A) and (B), respectively. For clarity we have offset the frequency position for method (B); however, for the analysis both methods were evaluated for the same frequencies. The blue points represent the power of the injected power spectrum at the sampled frequencies, while the dashed blue and purple lines show the best fit power spectrum to the marginalized coefficients for methods (A) and (B), respectively. (b) Parametrized values for the white noise in each pulsar in open data set 1 from the IPTA data challenge. Each Pulsar has a white noise component to their residuals with an amplitude of $\sigma_p = 10^{-7}$ s. Averaging across all pulsars we find an rms value for the white noise of $\bar{\Sigma}_{\rm avg} = -6.999 \pm 0.005$, which is thus consistent with the value in the data set to within 1σ errors. (c) 1D and 2D marginalized posteriors for the nine GWB power spectrum coefficients $\{\rho_i\}$ for method (B). The vertical line in the 1D distribution represents the power in the injected background at the frequency of that coefficient. Contours in the 2D plots represent 68% and 95% confidence levels.

degrading to become the slowest method with which to perform the analysis for the closed 2 data set.

While the comparisons in Table III have all been made with the method of vHL2013, it is of interest to see how the speedup compares with the data compression method presented in vH2013. We therefore used a dummy likelihood function that contained all the computational overhead associated with the data compression algorithm and set the number of pulsars, the number of observations, and the level of compression used to represent those values

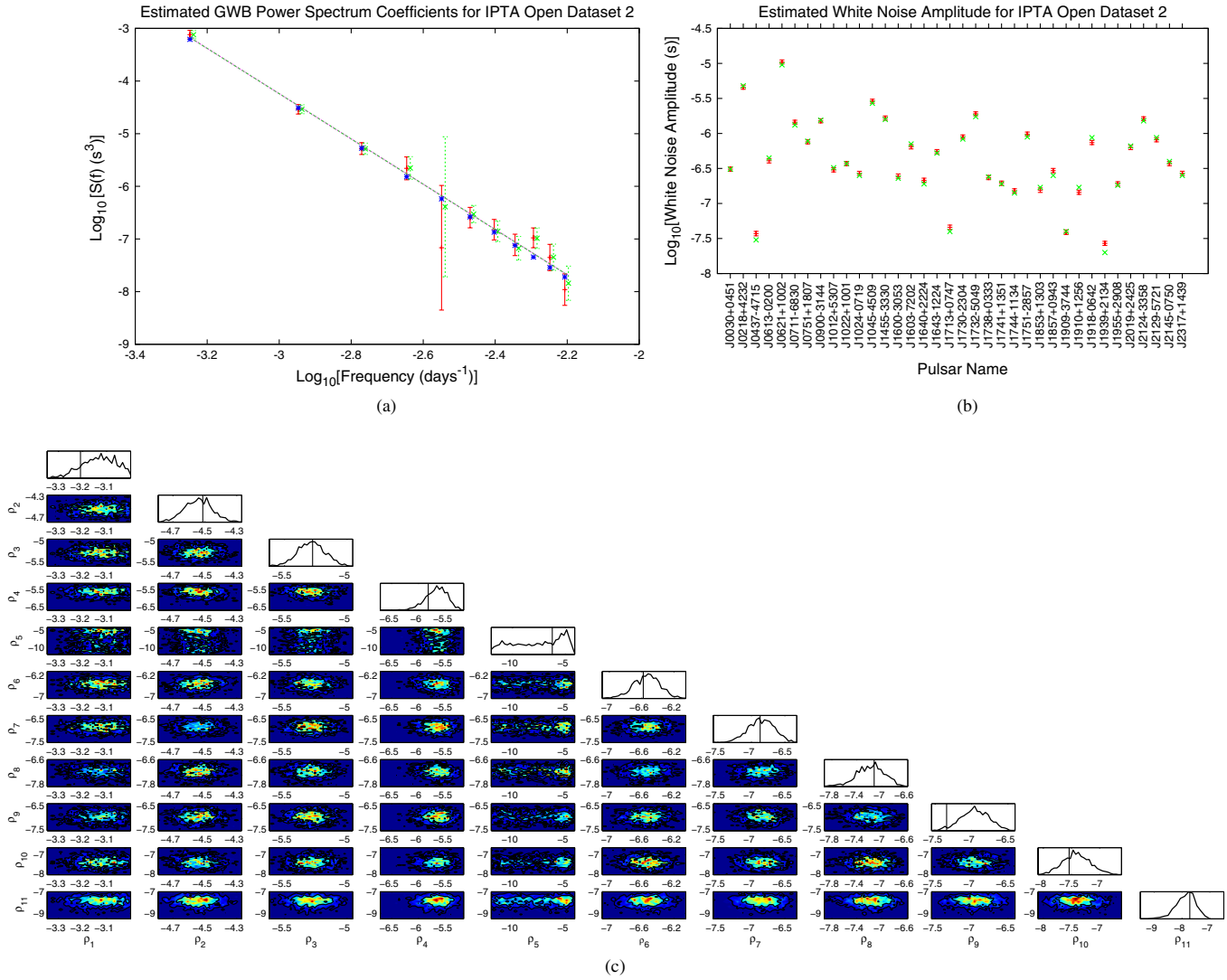
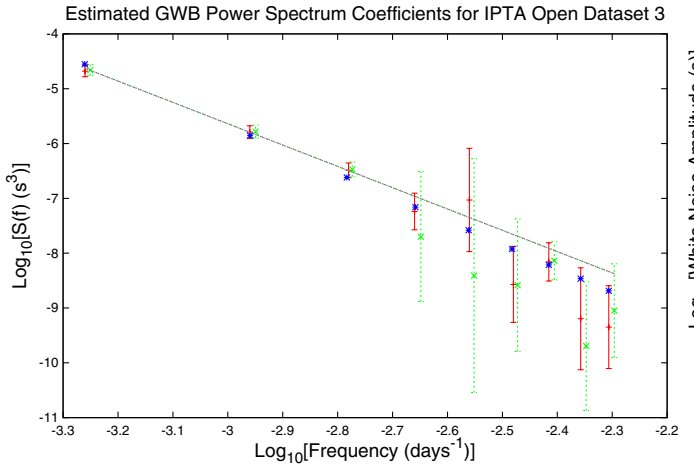


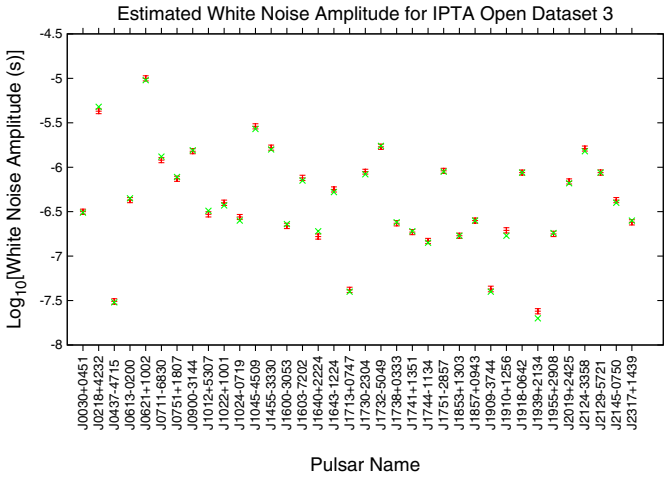
FIG. 3 (color online). (a) Log-log plot of the parametrized GWB power spectrum in open data set 2. The red and green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors for methods (A) and (B), respectively. For clarity we have offset the frequency position for method (B); however, for the analysis both methods were evaluated for the same frequencies. The blue points represent the power of the injected power spectrum at the sampled frequencies, while the dashed blue and purple lines show the best fit power spectrum to the marginalized coefficients for methods (A) and (B), respectively. (b) Parametrized values for the white noise in each pulsar in open data set 2 from the IPTA data challenge. Each pulsar has a different white noise component marked by the green crosses; red data points show the estimated white noise level from the analysis. (c) 1D and 2D marginalized posteriors for the 11 GWB power spectrum coefficients $\{\rho_i\}$ for method (B). The vertical line in the 1D distribution represents the power in the injected background at the frequency of that coefficient. Contours in the 2D plots represent 68% and 95% confidence levels.

that would be chosen for an analysis of the IPTA open 1 data set. This function was then compiled and linked to the same libraries used in the analysis of the previous section, at which point 10 sets of 1000 iterations each were performed and timed. We then used the likelihood function of methods (A) and (C), for which the latter provides the most direct comparison to the approach of vH2013, once again set the model parameters to be the same as those used in our analysis of open data set 1, and performed the same test. We found that the average computation time for 1000 evaluations of the three likelihood functions were approximately 45, 1.5, and 47 s for vH2013, method (A), and method (C), respectively. The consistency between vH2013 and method (C) is not surprising, the computational burden for each likelihood evaluation is still in the matrix inversions, which are of similar order, with the data compression method resulting in 10 data points per pulsar, and method (C) utilizing nine Fourier coefficients to describe the signal.

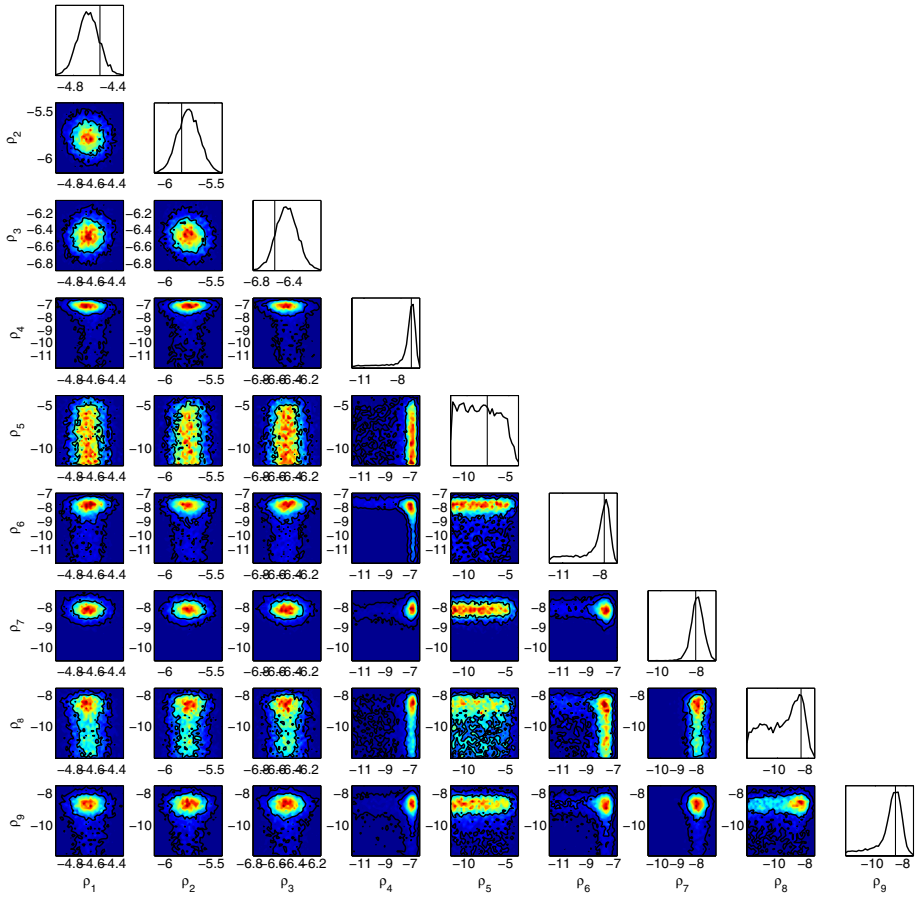
One important consideration when discussing the run times of these different methods is how well they scale with the inclusion of more parameters. While the method



(a)



(b)



(c)

FIG. 4 (color online). (a) Log-log plot of the parametrized GWB power spectrum in open data set 3. The red and green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors for methods (A) and (B), respectively. For clarity we have offset the frequency position for method (B); however, for the analysis both methods were evaluated for the same frequencies. The blue points represent the power of the injected power spectrum at the sampled frequencies, while the dashed blue and purple lines show the best fit power spectrum to the marginalized coefficients for methods (A) and (B), respectively. (b) Parametrized values for the white noise in each pulsar in open data set 3 from the IPTA data challenge. Each pulsar has a different white noise component marked by the green crosses; red data points show the estimated white noise level from the analysis. (c) 1D and 2D marginalized posteriors for the 9 GWB power spectrum coefficients $\{\rho_i\}$. The vertical line in the 1D distribution represents the power in the injected background at the frequency of that coefficient. Contours in the 2D plots represent 68% and 95% confidence levels.

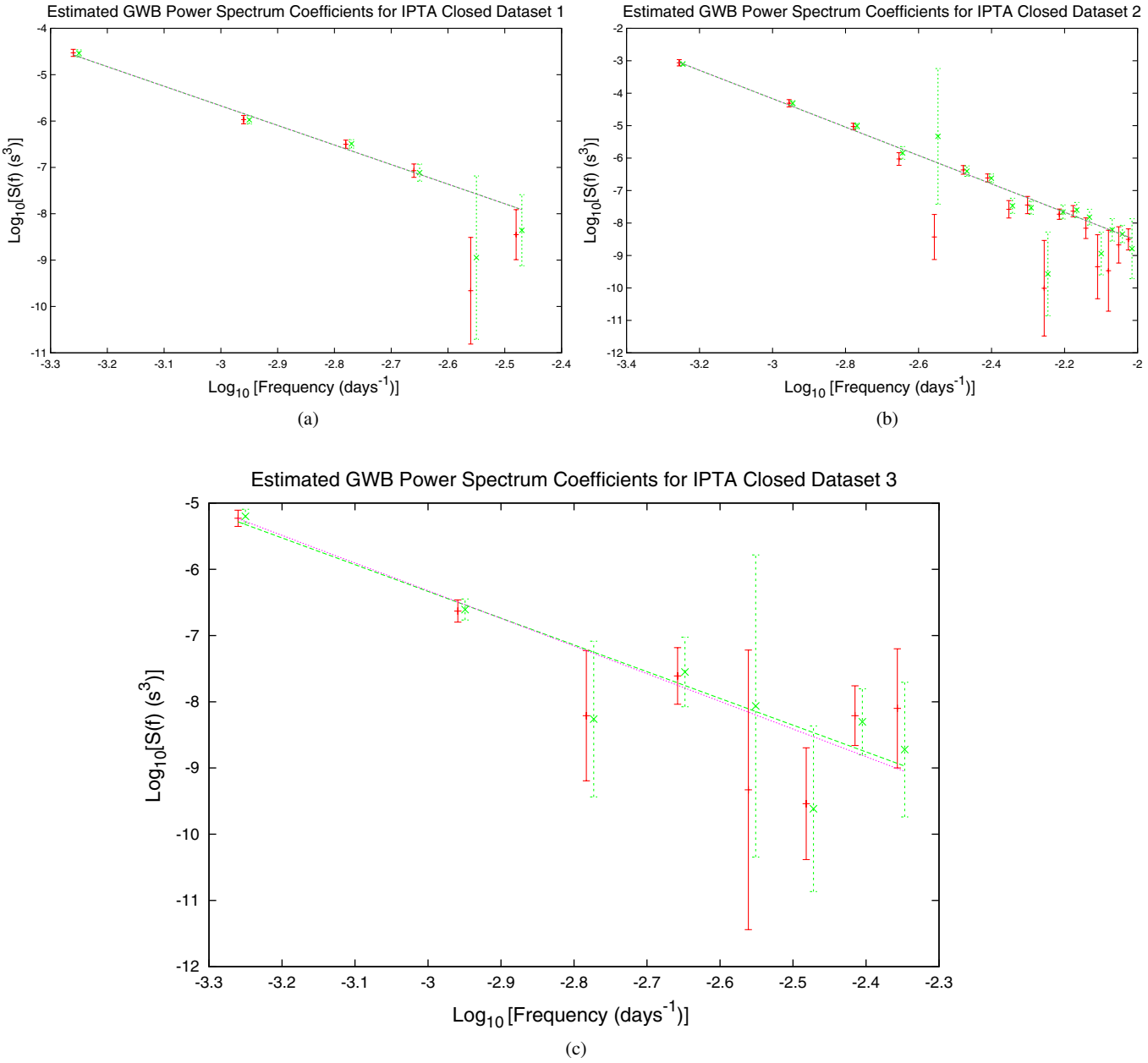


FIG. 5 (color online). Log-log plots of the parametrized GWB power spectrum in closed data sets 1 (a), 2 (b), and 3 (c). The red and green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors for methods (A) and (B), respectively. For clarity we have offset the frequency position for method (B); however, for the analysis both methods were evaluated for the same frequencies. The blue points represent the injected values for those coefficients, while the dashed blue and purple lines show the best fit power spectrum to the marginalized coefficients for methods (A) and (B), respectively.

described in vHL2013 is shown here to have comparable run times to method (A), we have only been using it to evaluate a two- or four-dimensional case. If for example one increases the dimensionality from 2 to 38 in order to include white noise estimation for each pulsar, the run time increases from 2 h, to over 100. Including white noise estimation in method (A), where the increase in dimensionality (36) is small compared to the total ($\sim$1000) results in a similarly small increase in the total run time of $\sim$15 min. This is one of the key advantages of the numerical marginalization coupled with the guided Hamiltonian sampler and one that we exploit in Sec. VIII as we introduce an additional 630 dimensions to parametrize the spatial correlations between pulsars. Even this though is still an extremely small parameter space compared to the greater than 10^6 dimensional problems that it has been used to solve in other work (B13). This therefore leaves a practically unlimited space in which to expand, with the inclusion of additional parameters such as simultaneous dispersion measure correction or even the full

TABLE III. Comparison of run times for different sampling methods.

	Method							
	This paper (A)		This paper (B)		This paper (C)		vHL2013	
Data set	Dimensionality	Run time (minutes)	Dimensionality	Run time (minutes)	Dimensionality	Run time (minutes)	Dimensionality	Run time (minutes)
Open 1	702	35	9	10	2	<1	2	145
Open 2	839	55	11	35	2	<1	2	130
Open 3	702	40	9	10	2	<1	2	140
Closed 1	474	30	9	2	2	<1	2	140
Closed 2	1277	110	17	180	2	4	2	160
Closed 3	908	130	16	145	4	3	4	235

nonlinear timing model that have previously been not thought feasible.

2. *Frequencies of* 1 $\mathbf{yr}^{-1}$

Perhaps one of the most striking features of the 1D and 2D confidence contours in Figs. 2–4 is that without exception the GWB coefficient ρ_5 is totally unconstrained. All of the data sets in the IPTA data challenge are approximately 1820 days in length, and so in every case ρ_5 corresponds to a frequency of $\sim 1\ \mathrm{yr}^{-1}$. That this should occur at such a distinct frequency is no coincidence; as part of the timing model fit performed by Tempo2 the pulsar's position and proper motion are all included as free parameters. Inaccuracies in the fitted values of these parameters can result in power being introduced to the residuals at frequencies of $1\ \mathrm{yr}^{-1}$ (see e.g. [2]). When we perform the analytic marginalization over all the model timing parameters, we therefore effectively project out contributions to the signal from components with these periods. The model Fourier coefficients corresponding to frequencies of $1\ \mathrm{yr}^{-1}$ therefore have no effect on the likelihood when the linear approximation to the timing model holds, and therefore the very way in which we account for the timing models for each pulsar results in us being able to make no inferences on the properties of the power spectrum at this frequency.

That this is so clear in the results is a testament to the success of the method; by not assuming any form for the power spectrum and simply asking in the most general way how the power is distributed in the signal, we are able to infer much more information than simply by fitting for a power law. In this instance that extra information is that we are unable to constrain anything about the spectrum at frequencies of $1\ \mathrm{yr}^{-1}$; however, where the true power spectrum is unknown this approach is the only way of ensuring an optimal estimate of that power spectrum and of extracting the maximal amount of information possible.

VII. A MORE REALISTIC SIMULATION

While the IPTA data challenges serve as a good introduction to analyzing PTA data, they still represent comparatively simplistic data sets when compared to genuine observations. For example, while some of the challenge data sets featured uneven sampling in the time domain, all pulsars within a data set shared the same TOAs, and thus also shared the same total time span. Similarly, when included, the properties of the red noise were the same for all the pulsars in the data sets. There were also no gaps in the data greater than a few weeks, whereas jumps of more than a year can be expected when analyzing real data. We have therefore constructed two simulations designed to represent better a potential future IPTA data release and thus provide a more difficult test for the analysis method presented in this paper.

A. Generating the residuals

The simulations are generated using the time domain covariance matrix $\mathbf{C}^{\mathrm{GW}}_{(ai)(bj)}$ between observations i and j and pulsars a and b for a GWB given in vH2009:

TABLE IV. Parameters for simulation one and two.

Pulsar No.	T_{span} years	N_{obs}	γ_{red}	$\log_{10}[A_g]$
1	3.18	22	3.3	14.3
2	14.86	1057	2.1	15.1
3	17.10	343	1.6	13.8
4	14.45	814	1.1	13.3
5	15.89	692	2.3	14.6
6	17.01	368	1.5	14.2
7	9.90	721	4.2	13.8
8	15.31	289	1.8	13.5
9	14.96	427	2.4	16.0
10	17.79	940	1.9	14.5
11	18.37	1291	1.6	14.0
12	17.80	422	2.2	14.2
13	8.04	153	5.1	15.0
14	16.96	728	3.4	14.6
15	5.75	164	2.6	13.9
16	4.75	35	3.5	14.0
17	9.02	728	1.5	13.4
18	10.46	284	2.3	14.4
19	15.42	293	2.8	14.1
20	17.54	914	1.2	13.7
21	14.95	402	3.4	14.0

TABLE V. Results from the two simulations in Sec. VII.

	Method (A)		Injected values	
Data set	$A_g \times 10^{-14}$	γ	$A_g \times 10^{-14}$	γ
Sim. 1	1.1 ± 0.2	4.2 ± 0.1	1	4.333
Sim. 2	0.61 ± 0.07	4.0 ± 0.2	0.5	4.333

$$\mathbf{C}^{\mathrm{GW}}_{(ai)(bj)} = \frac{\beta_{ab} A_g^2 \mathrm{yr}^{3-\gamma}}{12\pi^2 f_L^{\gamma-1}} \left\{ \Gamma(1-\gamma) \sin\left(\frac{\pi\gamma}{2}\right) (f_L \tau)^{\gamma-1} - \sum_{n=0}^{\infty} (-1)^n \frac{(f_l \tau)^{2n}}{(2n)!(2n+1-\gamma)} \right\}, \quad (66)$$

where β_{ab} is the Hellings-Downs coefficient between pulsars a and b, f_L is a low frequency cutoff, chosen only so that $1/f_L$ is much greater than the observing time span, and $\tau = 2\pi(t_{ai} - t_{bj})$ with t_{ai} the ith TOA for pulsar a. The covariance matrix for the included red noise $\mathbf{C}^{\mathrm{RN}}_{(ai)(bj)}$ is identical; however, the term β_{ab} is replaced with a delta function δ_{ab} as it will be uncorrelated between pulsars. Finally denoting the white noise covariance matrix $\mathbf{C}^{W}_{(ai)(bj)} = \sigma_w^2 \delta_{ab} \delta_{ij}$ we can write the total covariance matrix describing our simulated residuals $\mathbf{C}^{T}_{(ai)(bj)}$ as

$$\mathbf{C}^{T}_{(ai)(bj)} = \mathbf{C}^{\mathrm{GW}}_{(ai)(bj)} + \mathbf{C}^{\mathrm{RN}}_{(ai)(bj)} + \mathbf{C}^{W}_{(ai)(bj)}. \quad (67)$$

We then take the Cholesky decomposition of this matrix and use it to generate the residuals. A quadratic is then fitted to and subtracted from each of the pulsar residuals independently to mimic the effect of subtracting the timing model. The design matrix used to generate the matrix $\mathbf{G}$ in Eq. (9) and beyond will then simply be that of a quadratic polynomial.

B. The simulations

Both simulations use a set of 21 pulsars with observations spanning periods of between 4 and 18 years, with spacings between observations ranging from less than a day up to 5 yr. Simulation one then injects a gravitational wave background with parameters $\gamma = 4.33$ and dimensionless amplitude $A_g = 10^{-14}$ and white noise with an amplitude $\sigma_w = 10^{-7}$ s. The second simulation uses the same sampling times as the first; however, the background now has an amplitude of $A_g = 5 \times 10^{-15}$, and red noise is included for each pulsar, with γ_{red} covering a range from $1.1 \rightarrow 5.1$ and amplitudes extending from $A_g = 10^{-16} \rightarrow 5 \times 10^{-14}$. Table IV gives a more complete overview of the simulated data listing the total time span T_{span} for each pulsar, the number of observations N_{obs} in that observation window, and the red noise parameters γ_{red} and A_g present in simulation two.

In analyzing the data we chose a fundamental frequency f_0 to be equal to $1/T_{\mathrm{max}}$, where T_{max} represents the greatest time span for any of the pulsars in the data set, which for both simulations is $\sim$18.4 yr. We then use the Laplace approximation method described in Sec. V to determine the number of frequencies to be used in the analysis. We find that 21 coefficients should be sufficient to describe the first simulation while a maximum of 12 Fourier modes are required for the second. We then apply method (A) to the two data sets. In the first case we parametrize only the Fourier coefficients for the 21 pulsars, their white noise, and the set of 21 GWB power spectrum coefficients, while for the second data set we also include red noise parameters for each of the pulsars resulting in 264 and 903 dimensional spaces for each, respectively. The results are shown in Table V while we plot the GWB coefficients in both cases in Fig. 6 with the blue points representing the theoretical power at the sampled frequency given the injected spectrum. In the case of simulation 2 we plot only

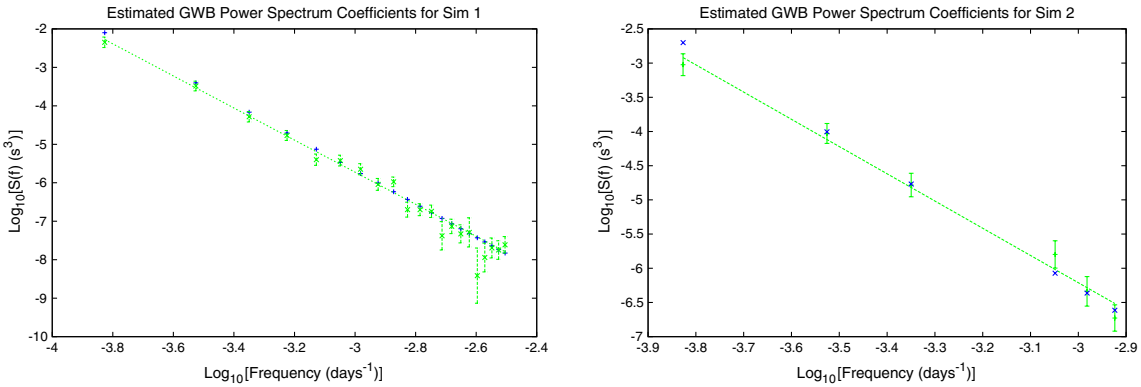


FIG. 6 (color online). Log-log plot of the parametrized GWB power spectrum in simulations one (left) and two (right). The green bars represent the marginalized values of the fitted GWB power coefficients $\{\rho_i\}$ and their errors derived using method (A) applied to that data set. The blue points represent the injected values for those coefficients, while the green line shows the best fit power law spectrum to the marginalized coefficients.

a subset of the frequency coefficients as only those corresponding to frequency modes 1–3 and 6–9 resulted in detections of a correlated signal within the data.

We see the results are once again consistent with the injected values, demonstrating that even in extremely challenging data where there is a great deal of additional red noise and highly irregular sampling we are able to correctly parametrize the GWB signal.

VIII. FITTING FOR THE CROSS CORRELATION

Thus far we have parametrized the angular correlations between different pairs of pulsars using the Hellings-Downs curve; the result derived assuming an isotropic background of gravitational waves when only those polarization states predicated by general relativity are considered. Different metric theories of gravity, however, predict different angular correlations, and anisotropies in the background due to bright individual sources can lead to deviations in this description [37]. Furthermore, terrestrial clock errors and inaccuracies in the solar system ephemeris can also generate spatial correlations within pulsar residuals, the latter for example would result in the residuals taking on a dipole signature [38]. As such, performing the analysis of PTA data assuming the Hellings-Downs curve explicitly could result in a false detection if there is a spatially correlated component, even if the form of that correlation is better described by something other than a GWB.

Methods for generalizing the Hellings-Downs curve at the point of sampling are relatively new, for example the authors of [39] present two possible approaches. First they fit for the angular correlation at a set of 5 angular separations and then use cubic splines to interpolate between those points in order to determine the angular correlations at intervening values, and second, they use a generalized Hellings-Downs model to parametrize the correlation. These methods were successfully able to extract the form of the Hellings-Downs curve in the case of the first IPTA open challenge; however, we would like to generalize this approach further and fit for the correlations between all pairs of pulsars directly. This therefore relieves us of the assumption that the background is isotropic, with pairs of pulsars at the same angular separation able to have different correlation coefficients, and still at no point assume any prescribed form of the correlation that might bias the end result in order to test whether or not the Hellings-Downs curve is distinguishable in simulated data from, for example, a dipole.

When fitting for the cross correlations between the pulsars, we must ensure that the covariance matrix describing those correlations remains positive definite. Many methods exist where the elements of the upper-triangular elements in the covariance matrix are reparametrized such that the resultant covariance matrix is ensured to be positive definite [40].

For any positive definite covariance matrix $\mathbf{\Sigma}$ we are able to take a Cholesky decomposition such that the matrix can be represented as the product $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. In general, however, such a decomposition is not unique. If $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{\Sigma}$ then so is any matrix obtained by multiplying a subset of the rows of $\mathbf{L}$ by -1. This can therefore give rise to multimodal distributions that will increase the complexity of the sampling process unnecessarily. This problem can be circumvented by ensuring that the diagonal elements of $\mathbf{L}$ are positive, in which case $\mathbf{L}$ is unique for a given $\mathbf{\Sigma}$, which can be achieved by fitting for the log of the diagonal elements. In this form, however, there is no straightforward way of fixing the elements of the matrix $\mathbf{\Sigma}$, such that the diagonal elements are equal to unity. We therefore use a spherical parametrization of the elements in $\mathbf{L}$ as in [40], which we describe below.

A. Spherical parameterization

If we denote the jth element of the ith column of the upper triangular matrix $\mathbf{L}$ as L_{ij}, and define a second upper triangular matrix $\mathbf{l}$ that contains the spherical parametrization of $\mathbf{L}$, we can write any element of $\mathbf{L}$ in the form

$$
\begin{aligned}
L_{i,1} &= l_{i,1}\cos(l_{i,2}),\\
L_{i,2} &= l_{i,1}\sin(l_{i,2})\cos(l_{i,3}),\\
L_{i,3} &= l_{i,1}\sin(l_{i,2})\sin(l_{i,3})\cos(l_{i,4}),\\
&\vdots\\
L_{i,i-1} &= l_{i,1}\sin(l_{i,2})\ldots\sin(l_{i,i-1})\cos(l_{i,i}),\\
L_{i,i} &= l_{i,1}\sin(l_{i,2})\ldots\sin(l_{i,i-1})\sin(l_{i,i}).
\end{aligned}
$$

The diagonal elements of the covariance matrix Σ_{ii} are then given by $\Sigma_{ii} = l_{i,1}^2$, and so we can trivially ensure a unit diagonal by setting all $l_{i,1} = 1$. Therefore for an $n \times n$ covariance matrix we need only fit for $n(n-1)/2$ elements, which for 36 pulsars, results in an increase of dimensionality of $N_{\mathrm{corr}} = 36 \times 35/2 = 630$.

The uniqueness of the spherical parametrization is then ensured by defining a new set of parameters $\mathbf{\Theta}$ such that

$$
l_{i,j} = \frac{\pi \exp(\Theta_{i,j})}{1 + \exp(\Theta_{i,j})}. \tag{68}
$$

While in principle this choice of parametrization should guarantee positive definiteness, in practice machine precision requires that we limit the values that $\Theta_{i,j}$ can take. Allowing $\mathbf{\Theta}$ to vary beyond ± 1.5 results in erroneous behavior due to this limitation, and so we require that $\mathbf{\Theta}$ lie within the range $\{-1, 1\}$, and therefore introduce a final set of parameters $\mathbf{X}$ such that

$$
\Theta_{i,j} = 2\left(\frac{\exp(X_{i,j})}{1 + \exp(X_{i,j})} - 0.5\right). \tag{69}
$$

Figure 7 shows the ability for this parametrization, with these limits in place, to reproduce the Hellings-Downs curve, zero correlation, and $\cos\theta/2$ between the pulsars.

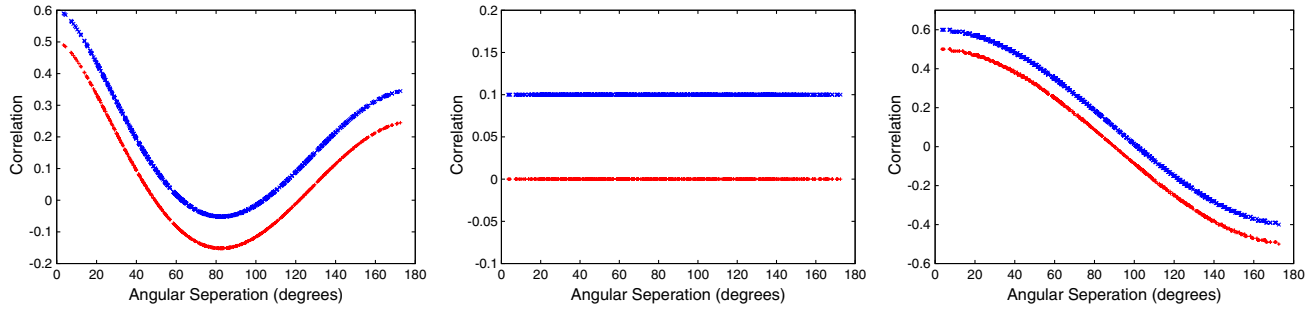


FIG. 7 (color online). Demonstration of the parametrization described in Sec. VIII A given the constraints on the parameter space imposed to ensure positive definiteness to reproduce the Hellings-Downs curve (left), no correlation (middle), and $\cos\theta/2$ (right). In each case the red line is the analytical evaluation, while the blue line is the best fit result. For clarity we have offset the blue line by 0.1 on the y axis.

We show the analytical expressions in red, while the best fit results are in blue. For clarity we have offset the two lines by 0.1 on the y axis, as the two forms are completely indistinguishable to within machine precision at all points.

B. Performing the sampling using the GHS

As before, in order to perform the sampling with the guided Hamiltonian sampler we will need both the gradients and the Hessian for our new likelihood function. By necessity we are sampling uniformly in the parameter $\mathbf{X}$; however, we would like to be sampling uniformly in the parameter space of the correlation coefficients $\mathbf{C}$. As such we must make a probability transformation so that the prior on our parameters $\mathbf{X}$ will be given by

$$\Pr(\mathbf{X}) = \Pr(\mathbf{C})|\mathbf{J}(\mathbf{X} \to \mathbf{C})|, \tag{70}$$

where writing the cross-correlation coefficient C_i in terms of its position in the cross-correlation matrix C_{mn}, the Jacobian can be written

$$J_{iq} = \frac{\partial C_i}{\partial X_q} = \left[\frac{\partial(\mathbf{L}\mathbf{L}^T)}{\partial l_q}\right]_{mn} \frac{\partial l_q}{\partial \Theta_q} \frac{\partial \Theta_q}{\partial X_q}. \tag{71}$$

This gives us our new log likelihood expression, which as in Sec. IV we write as the negative log, Ψ, so that ignoring constant terms

$$\Psi = \frac{1}{2}|\tilde{\mathbf{N}}| + \frac{1}{2}|\boldsymbol{\varphi}| + \frac{1}{2}(\boldsymbol{\delta t} - F\mathbf{a})^T\tilde{\mathbf{N}}^{-1}(\boldsymbol{\delta t} - F\mathbf{a}) + \frac{1}{2}\mathbf{a}^T\boldsymbol{\varphi}^{-1}\mathbf{a} - |J|. \tag{72}$$

At first sight calculating the gradient of such an expression with respect to the parameters $\mathbf{X}$ for every likelihood evaluation would seem a formidable computational task. However, because the $\partial\mathbf{L}/\partial l_q$ are all extremely sparse, featuring at most $N_{\rm corr}$ elements the scaling goes as $\sim O(N_{\rm corr}^2)$ and thus does not significantly impact the evaluation time. The gradient and second derivative of Ψ with respect to $\mathbf{X}$ are then of similar form to Eqs. (37) and (40) with extra terms corresponding to the derivatives of the Jacobian.

C. Results

We use this approach on the first open data challenge fitting for both the set of 630 cross-correlation coefficients between the 36 pulsars in the data set and 9 GWB coefficients. Figure 8 shows the cross-correlation coefficients and their associated errors as a function of the angular separation between pairs of pulsars in red, along with the analytical value for the Hellings-Downs curve at those values in blue. Fitting both the Hellings-Downs curve and no correlation as potential models results in χ^2 values of 630 and 1061, respectively, heavily favoring the presence of the Hellings-Downs curve, without having assumed its presence at the point of sampling.

Clearly this represents the simplest possible case, with no red noise present in the data. Where red noise is present the ability to recover the Hellings-Downs curve in this manner will inevitably degrade, and it might not prove

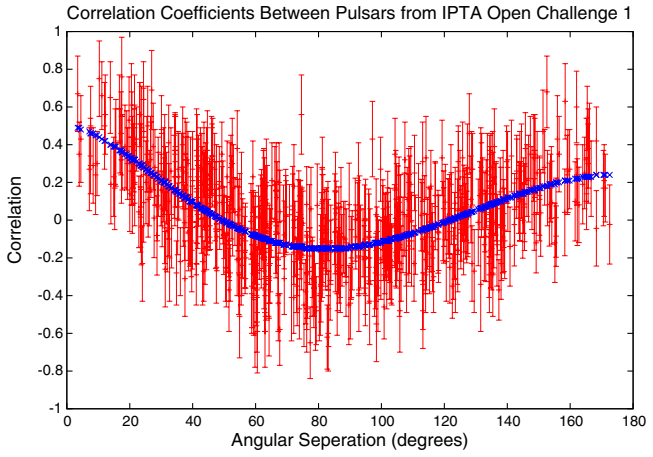


FIG. 8 (color online). Cross correlation coefficients between pairs of pulsars as a function of their angular separation parametrized using the approach in Sec. VIII. The blue points represent the analytical values that the Hellings-Downs curve takes for those angular separations. Fitting both the Hellings-Downs curve and no correlation as potential models results in χ^2 values of 630 and 1061, respectively, heavily favoring the presence of the Hellings-Downs curve.

possible to extract the cross correlation signal in such a completely general way. In such cases one might wish to reduce the number of free parameters by either assuming a model that has only an angular dependence and binning the coefficients up in angular separation as in [39] or by fitting some more general model that allows for spatial variation; in either case, the extrapolation of this method to these cases is straightforward.

IX. CONCLUSIONS

We have presented a new model-independent method for analyzing pulsar timing array data and estimating the spectral properties of a gravitational wave background.

We have shown that this method results in a speedup of approximately 2 orders of magnitude when compared to methods found in vHL2013 and, where the signal-to-noise ratio of the GWB is low, can reduce run times from several hours on a high performance computer to minutes on a regular workstation. We have accomplished this by sampling either directly from the power spectrum coefficients of the GWB where the number of coefficients to be sampled is small compared to the number of data points in the time domain, or, where the number of coefficients to be sampled increases, from the joint probability density of the power spectrum coefficients for the individual pulsars and the GWB signal realization, rephrasing the likelihood function to eliminate all matrix-matrix multiplications and costly dense matrix inversions. This latter approach therefore scales as $\mathrm{O}(n \times n_p^3)$ where n is the number of frequencies sampled, and n_p is the number of pulsars, as opposed to $\mathrm{O}(n_o^3)$ where n_o is the total number of observations in the data set across all pulsars.

We have shown this method requires no prior assumptions to be made regarding the shape of the power spectrum of the GWB. This is therefore currently the only method that provides a general approach to extracting a GWB signal from pulsar timing data, which we suggest is the only correct way of approaching the problem while we have no prior knowledge of the form of the power spectrum. We have also shown the ability for this method to parametrize correctly the correlation between pairs of pulsars. This correlation is the defining feature of a GWB signal, and extracting it from the data without first assuming that it is present will thus be a necessary step in any detection process.

Finally we have applied this method both to the first IPTA data challenge, as well as a more realistic pair of simulations and have shown that in all cases it correctly parametrizes the properties of the injected signals where they are known and is consistent with other established methods where they are not known.

ACKNOWLEDGMENTS

This work was performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (www.hpc.cam.ac.uk), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and funding from the Science and Technology Facilities Council.

APPENDIX A: GUIDED HAMILTONIAN SAMPLING

The following is a description of both Hamiltonian Monte Carlo and the Guided Hamiltonian Sampler as described in (B13).

1. Standard Hamiltonian Monte Carlo sampling

In HMC, one begins by defining the potential energy $\psi(\boldsymbol{x})$ of the target density $\Pr(\boldsymbol{x})$ as its negative logarithm, namely,

$$\psi(\boldsymbol{x}) = -\ln \Pr(\boldsymbol{x}). \tag{A1}$$

For each parameter, x_i we then introduce a "momentum" parameter p_i and a constant "mass" m_i and construct a kinetic energy term that, when added to the potential, leads to the Hamiltonian

$$\mathcal{H}(\boldsymbol{x}, \boldsymbol{p}) = \sum_i \frac{p_i^2}{2m_i} + \psi(\boldsymbol{x}). \tag{A2}$$

Our new objective is to draw samples from a distribution that is proportional to $\exp[-\mathcal{H}(\boldsymbol{x}, \boldsymbol{p})]$. The form of the Hamiltonian is such that this distribution is separable into a Gaussian in $\boldsymbol{p}$ and the target distribution, i.e.

$$\exp[-\mathcal{H}(\boldsymbol{x}, \boldsymbol{p})] = \Pr(\boldsymbol{x}) \prod_i \exp\left(-\frac{p_i^2}{2m_i}\right). \tag{A3}$$

We can then obtain samples from $\Pr(\boldsymbol{x})$ by marginalizing over $\boldsymbol{p}$.

To find a new sample we first draw a set of momenta from the distribution defined by our kinetic energy term, i.e. an N-dimensional uncorrelated Gaussian with a variance in dimension i of m_i. We then allow our system to evolve deterministically from our starting point $(\boldsymbol{x}, \boldsymbol{p})$ in the phase space for some fixed time τ according to Hamilton's equations,

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \nabla_p \mathcal{H}(\boldsymbol{x}, \boldsymbol{p}), \tag{A4}$$

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = -\nabla_x \mathcal{H}(\boldsymbol{x}, \boldsymbol{p}) = -\nabla_x \psi(\boldsymbol{x}). \tag{A5}$$

At the end of this trajectory we have reached the point $(\boldsymbol{x}', \boldsymbol{p}')$ and we accept this point with probability

$$p_A = \min[1, \exp(-\delta\mathcal{H})], \tag{A6}$$

where

$$\delta\mathcal{H} = \mathcal{H}(\boldsymbol{x}', \boldsymbol{p}') - \mathcal{H}(\boldsymbol{x}, \boldsymbol{p}). \tag{A7}$$

This implies that if we are able to integrate Hamilton's equations exactly then, as energy is conserved along such a trajectory, the probability of acceptance is unity. In practice, however, numerical inaccuracies mean that this is not the case. After a new proposed sample is generated the momentum variable is discarded and the process restarts by randomly drawing a new set of momenta as described above.

In fact the method is more general than outlined above since, provided one uses the Metropolis acceptance criterion [Eq. (A6)], it is permitted to follow any trajectory to generate a new candidate point. However, only trajectories that approximately conserve the value of the Hamiltonian [Eq. (A2)] will result in high acceptance rates. For some problems it may be advantageous to generate trajectories using an approximate Hamiltonian that can be computed rapidly, and bear the cost of lowering the acceptance probability.

To integrate the equations of motions it is common practice to use the leapfrog method [28]. This method has the property of exact reversibility that is required to ensure the chain satisfies detailed balance. It is also numerically robust and allows for the simple propagation of errors. We make n steps with a finite step size ϵ, such that $n\epsilon = \tau$, as follows:

$$\boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) = \boldsymbol{p}(t) + \frac{\epsilon}{2}\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\bigg|_{t}, \tag{A8}$$

$$\boldsymbol{x}(t + \epsilon) = \boldsymbol{x}(t) + \epsilon\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\bigg|_{t+\frac{\epsilon}{2}}, \tag{A9}$$

$$\boldsymbol{p}(t + \epsilon) = \boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) + \frac{\epsilon}{2}\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}\bigg|_{t+\epsilon}, \tag{A10}$$

until $t = \tau$. Substituting for the time derivatives using Hamilton's equations (A4), one thus obtains explicit relations for the leapfrog steps, which read

$$\boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) = \boldsymbol{p}(t) - \frac{\epsilon}{2}\nabla_{\boldsymbol{x}}\mathcal{H}|_{t}, \tag{A11}$$

$$\boldsymbol{x}(t + \epsilon) = \boldsymbol{x}(t) + \epsilon\nabla_{\boldsymbol{p}}\mathcal{H}|_{t+\frac{\epsilon}{2}}, \tag{A12}$$

$$\boldsymbol{p}(t + \epsilon) = \boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\nabla_{\boldsymbol{x}}\mathcal{H}|_{t+\epsilon}. \tag{A13}$$

The interval τ must be varied, usually by drawing n and ϵ randomly from uniform distributions, to avoid resonant trajectories; we therefore draw n and ϵ from $\mathrm{U}(1, n_{\max})$, $\mathrm{U}(0, \epsilon_{\max})$, respectively. The leapfrog method may be replaced by higher-order integration schemes provided exact reversibility is maintained; such methods yield greater accuracy, although generally incur significant additional computational costs.

2. Setting masses in HMC

HMC can be extremely sensitive to the choice of masses, in particular, when the marginal distributions of different parameters show considerable variation in the width of their posterior distributions. Reference [41] suggests that one should set the mass associated with each parameter to be approximately equal to the variance of that parameter in the target density. This is an attempt to circularize the trajectories in the $(\boldsymbol{x}, \boldsymbol{p})$ space. Interestingly, [28] suggests precisely the opposite approach, where the mass for a parameter is inversely proportional to the width of the distribution.

The authors of Ref. [27] follow the latter suggestion and justify it by generalizing the framework in [26] to describe the application of the leapfrog method. In particular, for the case where the N-dimensional target distribution $\Pr(\boldsymbol{x})$ is (well approximated by) a multivariate Gaussian with covariance matrix $\mathbf{C}$, they show that the leapfrog method is stable if $\mathbf{M} = \mathbf{C}^{-1}$ and $\epsilon \leq 2$, where $\mathbf{M}$ is the $N \times N$ "mass matrix" that appears in the generalized kinetic term $\frac{1}{2}\boldsymbol{p}^{\mathrm{t}}\mathbf{M}^{-1}\boldsymbol{p}$ of the Hamiltonian.

If the dimensionality of the problem is such that it is impractical to perform the required matrix inversion and decomposition of $\mathbf{M}$ (to compute the Hamiltonian and to draw new values for the momentum variables, respectively) then simple approximations must be employed. Typically one might construct a diagonal mass matrix with the mass associated with each parameter inversely proportional to the variance of that parameter.

Moreover, if the target distribution is not Gaussian, it seems reasonable to use some appropriate measure of the width of the distribution, such as the curvature at the peak [28], to set the masses.

3. Guided Hamiltonian sampling

Guided Hamiltonian sampling builds on the ideas explored in [27] to produce an HMC algorithm with just a single adjustable parameter, thereby eliminating the need for tuning masses. In particular, GHS takes advantage of, although does not rely on, the fact that one often wishes to sample from a target distribution that is unimodal, albeit, in general, non-Gaussian and high dimensional.

In GHS, one first sets the mass matrix in the kinetic term of the Hamiltonian to the identity, $\mathbf{M} = \mathbf{I}$. For the target distribution $\Pr(\boldsymbol{x})$, one then locates the peak $\hat{\boldsymbol{x}}$, typically using some iterative gradient-search optimization algorithm starting from, in general, some random initial point. One then calculates the Hessian (or curvature) matrix $\hat{\mathbf{H}}$ of $\ln\Pr(\boldsymbol{x}) = -\psi(\boldsymbol{x})$ (i.e. the negative of the potential energy, for convenience of sign conventions) at the maximum, either analytically or using numerical differentiation; this thereby defines a Gaussian approximation to $\Pr(\boldsymbol{x})$ in the neighborhood of the peak $\hat{\boldsymbol{x}}$.

Once the Hessian at the peak has been calculated, one then determines its N eigenvalues λ_i and N normalized

eigenvectors $\hat{\boldsymbol{e}}_i$. Denoting the matrix containing these normalized eigenvectors as its columns by $\boldsymbol{S}$, one first defines a new set of variables $\boldsymbol{x}' = \boldsymbol{S}^{\mathrm{t}}\boldsymbol{x}$ in which the Hessian becomes diagonal with the eigenvalues λ_i as its diagonal entries. One then rescales each x_i' to obtain a new set of variables $y_i = \sqrt{\lambda_i}x_i'/\eta$, where the scaling factor η is the single adjustable parameter in GHS, which we will discuss later. It is straightforward to show that the new variables are related to the original variables by

$$\boldsymbol{y} = \frac{1}{\eta}\hat{\mathbf{H}}^{1/2}\boldsymbol{x}. \tag{A14}$$

Consequently, in the new variables, the Hessian at the peak has the trivial form $\eta^2\mathbf{I}$. One then performs Hamiltonian sampling employing the standard leapfrog method [(A11)–(A13)] but in terms of the new variables $\boldsymbol{y}$, rather than $\boldsymbol{x}$. Thus, GHS may be considered simply as standard HMC but performed in a set of variables (or coordinates) that are tailored to the target distribution, namely, the scaled eigendirections of the Hessian at its peak. Consequently, although GHS may take advantage if $\Pr(\boldsymbol{x})$ possesses a single well-defined peak (with zero gradient), it does not rely on this, since it retains the generality of standard HMC.

Rather than working in terms of the new variables $\boldsymbol{y}$, one can, if desired, return to using the original variables $\boldsymbol{x}$, in which case the relation (A14) shows that the leapfrog steps take the modified form

$$\boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) = \boldsymbol{p}(t) - \frac{\epsilon}{2}\eta\hat{\mathbf{H}}^{-1/2}\nabla_x\mathcal{H}|_t, \tag{A15}$$

$$\boldsymbol{x}(t+\epsilon) = \boldsymbol{x}(t) + \epsilon\eta\hat{\mathbf{H}}^{-1/2}\nabla_p\mathcal{H}|_{t+\frac{\epsilon}{2}}, \tag{A16}$$

$$\boldsymbol{p}(t+\epsilon) = \boldsymbol{p}\left(t + \frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\eta\hat{\mathbf{H}}^{-1/2}\nabla_x\mathcal{H}|_{t+\epsilon}. \tag{A17}$$

Using the original variables $\boldsymbol{x}$ or the new variables $\boldsymbol{y}$, it is necessary to calculate either the (inverse) square root of the $N \times N$ Hessian matrix $\hat{\mathbf{H}}$ at the peak or (equivalently) its eigendecomposition (and, subsequently, the calculation of the square roots of its eigenvalues). Performing the above calculations can be computationally expensive, particularly for large N, although it should be noted that one need only perform these calculations once.

In summary, GHS aims to increase the efficiency of standard HMC, particularly for high-dimensional, unimodal target distributions, by performing the sampling in the principal coordinates defined by the Gaussian approximation at its peak. In this way, one may largely eliminate the tuning aspect of HMC: the single remaining adjustable parameter is the scaling η, the optimal value of which depends on the dimensionality of the parameter space, and should be chosen such that the acceptance rate is approximately 68%.

APPENDIX B: ANALYTICAL APPROXIMATION TO THE LIKELIHOOD

1. Uniform white noise

Suppose we have a single realization of some time series data $\mathbf{d}$ of length N. We then define a set of hypotheses $\{H\}$ such that each H_i purports that our data $\mathbf{d}$ is described by some function f_i where

$$f_i(t) = \sum_{k=1}^{m} b_k M_k(t, \mathbf{w}) \tag{B1}$$

with M_k a set of general basis functions. The number of functions m, the parameters that describe them (e.g. their frequencies) $\mathbf{w}$, and the model coefficients b_k are allowed to vary for each f_i. We then transform this set of basis functions into an orthonormal set F_k through the transformation

$$F_k(t) = \frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{m} e_{kj}M_j(t), \tag{B2}$$

where e_{kj} is the kth element of the jth eigenvector and λ_k is the kth eigenvalue of the covariance matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$. Our function f_i can now be written in terms of these new basis vectors,

$$f_i(t) = \sum_{k=1}^{m} a_k F_k(t, \mathbf{w}), \tag{B3}$$

where the coefficients a in the orthonormal basis are related to the coefficients b in the original basis through

$$b_k = \sum_{j=1}^{m}\frac{a_k e_{jk}}{\sqrt{\lambda_j}}. \tag{B4}$$

The probability of the data given a model f_i, assuming that the noise is described by a zero mean random Gaussian process with variance σ is given by

$$\Pr(\mathbf{d}|\mathbf{a}, \mathbf{w}, \sigma, f_i) = (2\pi\sigma^2)^{-N/2}\exp\left[\frac{1}{2\sigma^2}\sum_{k=1}^{N}[d_k - f_i(t_k)]^2\right]. \tag{B5}$$

Writing the projection of the data onto our basis functions as

$$h_i = \sum_{k=1}^{N} d_k F_i(t_k), \tag{B6}$$

and writing $\mathbf{d}^2 = \mathbf{d}^T\mathbf{d}$, Eq. (B5) can be written

$$\Pr(\mathbf{d}|\mathbf{a}, \mathbf{w}, \sigma, f_i) = (2\pi\sigma^2)^{-N/2}\exp\left[-\frac{1}{2\sigma^2}\left[\mathbf{d}^2 - \sum_{l=1}^{m} 2a_l h_l + a_l^2\right]\right]. \tag{B7}$$

We begin by integrating over both the set of coefficients $\mathbf{a}$ and frequencies $\mathbf{w}$. We assume that the two parameters are logically independent, insofar as we can write the priors

$$\Pr(\mathbf{a}, \mathbf{w}) = \Pr(\mathbf{a}) \Pr(\mathbf{w}). \tag{B8}$$

For the amplitude coefficients, we choose an uninformative Gaussian prior given by

$$\Pr(\mathbf{a}|\delta) = (2\pi\delta^2)^{-m/2} \exp\left[-\sum_{k=1}^{m} \frac{a_k^2}{2\delta^2}\right] \tag{B9}$$

with $\delta \gg \sigma$. Therefore, our probability, marginalized over $\mathbf{a}$ and $\mathbf{w}$ can be written

$$\begin{aligned}\Pr(\mathbf{d}|\delta, \sigma, f_i) = \int \mathrm{d}\mathbf{w} \Pr(\mathbf{w}) (2\pi\delta^2)^{-m/2} (2\pi\sigma^2)^{-N/2} \\ \times \int_{-\infty}^{+\infty} \mathrm{d}a_1 \ldots \mathrm{d}a_m \exp\left[-\sum_{k=1}^{m} \frac{a_k^2}{2\delta^2}\right] \\ \times \exp\left[-\frac{1}{2\sigma^2}\left[\mathbf{d}^2 - \sum_{l=1}^{m} 2a_l h_l + a_l^2\right]\right].\end{aligned} \tag{B10}$$

We have chosen δ such that the prior term $\exp[-\sum_{k=1}^{m} a_k^2/2\delta^2]$ is constant where the likelihood is large but goes to zero sufficiently quickly outside this region so as to be normalizable. Therefore, if we define $\hat{a}_i$ to be the maximum likelihood value for the parameter a_i, we can write our probability as

$$\begin{aligned}&\Pr(\mathbf{d}|\delta, \sigma, f_i) \\ &= \int \mathrm{d}\mathbf{w} \Pr(\mathbf{w}) (2\pi\delta^2)^{-m/2} (2\pi\sigma^2)^{-N/2} \exp\left[-\sum_{k=1}^{m} \frac{\hat{a}_k{}^2}{2\delta^2}\right] \\ &\quad \times \int_{-\infty}^{+\infty} \mathrm{d}\mathbf{a} \exp\left[-\frac{1}{2\sigma^2}\left[\mathbf{d}^2 - \sum_{l=1}^{m} 2a_l h_l + a_l^2\right]\right].\end{aligned} \tag{B11}$$

If we take the elements of $\mathbf{a}$ to be independent on our orthonormal basis, then we can write the expectation value of a single element a_i as

$$\langle a_i \rangle = \frac{\int_{-\infty}^{+\infty} \mathrm{d}a_i a_i \exp[\frac{-1}{2\sigma^2}[-2a_i h_i + a_i^2]]}{\int_{-\infty}^{+\infty} \mathrm{d}a_i \exp[\frac{-1}{2\sigma^2}[-2a_i h_i + a_i^2]]}, \tag{B12}$$

which evaluates to $\langle a_i \rangle = h_i$, i.e. the expectation value of the basis vector coefficient is just the projection of the data onto that basis. Substituting this into our equation for the probability in the place of $\hat{a}$ and performing the Gaussian integral over $\mathbf{a}$, we arrive at the expression

$$\begin{aligned}\Pr(\mathbf{d}|\delta, \sigma, f_i) = \int \mathrm{d}\mathbf{w} \Pr(\mathbf{w}) (2\pi\delta^2)^{-m/2} (2\pi\sigma^2)^{-(N-m)/2} \\ \times \exp\left[\frac{\mathbf{d}^2 - \mathbf{h}^2}{2\sigma^2}\right] \exp\left[\frac{\mathbf{h}^2}{2\delta^2}\right].\end{aligned} \tag{B13}$$

For our integral over our frequencies, we are for any given model f_i considering a set of frequencies chosen from an evenly spaced grid. Therefore we will have a set of delta function priors for each frequency w_j in the set $\mathbf{w}$ and the integral can be simply evaluated:

$$\begin{aligned}\Pr(\mathbf{d}|\delta, \sigma, f_i) = (2\pi\delta^2)^{-m/2} (2\pi\sigma^2)^{-(N-m)/2} \\ \times \exp\left[\frac{\mathbf{d}^2 - \mathbf{h}(\mathbf{w_i})^2}{2\sigma^2}\right] \exp\left[\frac{\mathbf{h}(\mathbf{w_i})^2}{2\delta^2}\right].\end{aligned} \tag{B14}$$

We are now in a position to integrate over our unknown variances σ and δ. As in [35] we set an upper bound H and lower bound L to this integral, which will therefore be of the form

$$\frac{1}{\log(H/L)} \int_L^H \mathrm{d}s \frac{s^{-a} \exp[-\frac{Q}{s^2}]}{s}. \tag{B15}$$

Making a substitution $u = Q/s^2$ this becomes

$$\frac{Q^{-a/2}}{2\log(H/L)} \int_{Q/H^2}^{Q/L^2} \mathrm{d}u u^{a/2-1} \exp[-u]. \tag{B16}$$

If we assume that H is sufficiently large, and L is sufficiently small that we may write $Q/H^2 \ll 1$ and $a/2 - 1 \ll Q/L^2$ then the integral will evaluate to approximately $\Gamma(a/2)$. Thus our integral over δ will become

$$\frac{1}{\log(H/L)} \int_L^H \mathrm{d}\delta \frac{\delta^{-m} \exp[-\frac{\mathbf{h}^2}{2\delta^2}]}{\delta} \approx \frac{\Gamma(m/2)}{2\log(R_\delta)} \left[\frac{\mathbf{h}(\mathbf{w})^2}{2}\right]^{-m/2}, \tag{B17}$$

and similarly for σ the integral evaluates to approximately

$$\frac{\Gamma((N-m)/2)}{2\log(R_\sigma)} \left[\frac{\mathbf{d}^2 - \mathbf{h}(\mathbf{w})^2}{2}\right]^{-(N-m)/2}. \tag{B18}$$

Therefore we can finally write the probability of the data D given a model f_i as

$$\begin{aligned}\Pr(\mathbf{d}|f_i) = \frac{\Gamma(m/2)}{2\log(R_\delta)} \left[\frac{\mathbf{h}(\mathbf{w})^2}{2}\right]^{-m/2} \frac{\Gamma((N-m)/2)}{2\log(R_\sigma)} \\ \times \left[\frac{\mathbf{d}^2 - \mathbf{h}(\mathbf{w})^2}{2}\right]^{-(N-m)/2}.\end{aligned} \tag{B19}$$

2. Nonuniform white noise

In general when dealing with pulsar residuals the white noise level across a data set for a single pulsar will vary with time, where for example different instruments have been used to collect data for the same pulsar. In this case the expansion of our likelihood function is not so simple, because the covariance matrix $\mathbf{G}^T\mathbf{N}\mathbf{G}$ will no longer reduce to a diagonal matrix. If we define $\mathbf{C} = \mathbf{G}^T\mathbf{N}\mathbf{G}$ where we consider $\mathbf{C}$ to be a general dense covariance matrix, Eq. (B5) will take the form

$$\Pr(\mathbf{d}|\mathbf{a},\mathbf{w},f_i) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2} \times \exp\left[\frac{-1}{2}(\mathbf{d}-\mathbf{Fa})^T\mathbf{C}^{-1}(\mathbf{d}-\mathbf{Fa})\right]. \quad \text{(B20)}$$

In this case, writing $\mathbf{F_i}^T\mathbf{C}^{-1}\mathbf{F_i} = \mathbf{C}_i^{-1}$ the maximum likelihood value of a particular coefficient a_i will be given by

$$\langle a_i\rangle = \frac{\int_{-\infty}^{+\infty} da_i a_i \exp[-\frac{1}{2}[a_i\mathbf{C}_i^{-1}a_i - 2\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}a_i]]}{\int_{-\infty}^{+\infty} da_i \exp[-\frac{1}{2}[a_i\mathbf{C}_i^{-1}a_i - 2\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}a_i]]} \quad \text{(B21)}$$

and evaluates to

$$\langle a_i\rangle = \frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}}{\mathbf{C}_i^{-1}}. \quad \text{(B22)}$$

In the case that $\mathbf{C}$ once again describes uniform white noise across the observation, this will reduce to $\langle a_i\rangle = \mathbf{d}^T\mathbf{F_i} = h_i$ as before. Using the same uninformative prior on our coefficients as in Eq. (B9), we can then write our integral over the basis coefficients as

$$\Pr(\mathbf{d}|\delta,f_i) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}(2\pi\delta^2)^{-m/2} \times \exp\left[-\frac{1}{2\delta^2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}^{-1}\mathbf{F_i}}\right)^2\right] \times \int_{-\infty}^{+\infty} d\mathbf{a}\exp\left[\frac{-1}{2}(\mathbf{d}-\mathbf{Fa})^T\mathbf{C}^{-1}(\mathbf{d}-\mathbf{Fa})\right]. \quad \text{(B23)}$$

If we define

$$\boldsymbol{\chi} = (\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}^{-1}\mathbf{d}, \quad \text{(B24)}$$

then we can reexpress this probability as

$$\Pr(\mathbf{d}|\delta,f_i) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2}(2\pi\delta^2)^{-m/2} \times \exp\left[-\frac{1}{2\delta^2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}^{-1}\mathbf{F_i}}\right)^2\right] \times \exp\left[-\frac{1}{2}\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}\right]\exp\left[\frac{1}{2}\boldsymbol{\chi}^T\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}\boldsymbol{\chi}\right] \times \int_{-\infty}^{+\infty} d\mathbf{a}\exp\left[-\frac{1}{2}(\mathbf{a}-\boldsymbol{\chi})^T\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}(\mathbf{a}-\boldsymbol{\chi})\right],$$

which evaluates to

$$\Pr(\mathbf{d}|\delta,f_i) = ((2\pi)^{N-m}|\mathbf{C}||\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}|)^{-1/2}(2\pi\delta^2)^{-m/2} \times \exp\left[-\frac{1}{2\delta^2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}^{-1}\mathbf{F_i}}\right)^2\right] \times \exp\left[-\frac{1}{2}\mathbf{d}^T\mathbf{C}^{-1}\mathbf{d}\right]\exp\left[\frac{1}{2}\boldsymbol{\chi}^T\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F}\boldsymbol{\chi}\right]. \quad \text{(B25)}$$

Thus far we have assumed that we know the level of the noise in $\mathbf{C}$ exactly; however, in general we would like to fit for a global scaling factor that modifies the overall noise level in the data set, i.e. we would like to write $\mathbf{C}' = \mathbf{G}^T(\alpha^2\mathbf{N})\mathbf{G}$ where α is a constant to be determined. Including this in our probability we can write

$$\Pr(\mathbf{d}|\alpha,\delta,f_i) = ((2\pi\alpha)^{(N-m)}|\mathbf{C}||\mathbf{F}^T\mathbf{C}'^{-1}\mathbf{F}|)^{-1/2}(2\pi\delta^2)^{-m/2} \times \exp\left[-\frac{1}{2\delta^2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}'^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}'^{-1}\mathbf{F_i}}\right)^2\right] \times \exp\left[-\frac{1}{2\alpha^2}(\mathbf{d}^T\mathbf{C}'^{-1}\mathbf{d} - \boldsymbol{\chi}^T\mathbf{F}^T\mathbf{C}'^{-1}\mathbf{F}\boldsymbol{\chi})\right]. \quad \text{(B26)}$$

We can then finally proceed as before integrating over both α and δ to arrive at the final probability

$$\Pr(\mathbf{d}|f_i) = \frac{\Gamma(m/2)}{2\log(R_\delta)}\left[\frac{1}{2}\sum_{k=1}^{m}\left(\frac{\mathbf{d}^T\mathbf{C}'^{-1}\mathbf{F_i}}{\mathbf{F_i}^T\mathbf{C}'^{-1}\mathbf{F_i}}\right)^2\right]^{-m/2} \times \frac{\Gamma((N-m)/2)}{2\log(R_\alpha)}\left[-\frac{1}{2}(\mathbf{d}^T\bar{\mathbf{C}}^{-1}\mathbf{d})\right]^{-(N-m)/2}, \quad \text{(B27)}$$

where we have defined

$$\bar{\mathbf{C}}^{-1} = \mathbf{C}'^{-1} - \mathbf{C}'^{-1}\mathbf{F}(\mathbf{F}^T\mathbf{C}'^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}'^{-1}. \quad \text{(B28)}$$

[1] V. M. Kaspi, J. H. Taylor, and M. F. Ryba, Astrophys. J. **428**, 713 (1994).
[2] D. N. Matsakis, J. H. Taylor, and T. M. Eubanks, Astron. Astrophys. **326**, 924 (1997).
[3] J. H. Taylor and J. M. Weisberg, Astrophys. J. **345**, 434 (1989).
[4] M. Kramer *et al.*, Science **314**, 97 (2006).
[5] R. T. Edwards, G. B. Hobbs, and R. N. Manchester, Mon. Not. R. Astron. Soc. **372**, 1549 (2006).
[6] G. Hobbs, F. Jenet, K. J. Lee, J. P. W. Verbiest, D. Yardley, R. Manchester, A. Lommen, W. Coles, R. Edwards, and C. Shettigara, Mon. Not. R. Astron. Soc. **394**, 1945 (2009).
[7] G. B. Hobbs, R. T. Edwards, and R. N. Manchester, Mon. Not. R. Astron. Soc. **369**, 655 (2006).

[8] R. M. Shannon and J. M. Cordes, Astrophys. J. **725**, 1607 (2010).
[9] A. H. Jaffe and D. C. Backer, Astrophys. J. **583**, 616 (2003).
[10] E. S. Phinney, arXiv:astro-ph/0108028.
[11] M. Kawasaki, K. Miyamoto, and K. Nakayama, Phys. Rev. D **81**, 103523 (2010).
[12] S. Ölmez, V. Mandic, and X. Siemens, Phys. Rev. D **81**, 104028 (2010).
[13] S. A. Sanidas, R. A. Battye, and B. W. Stappers, Phys. Rev. D **85**, 122003 (2012).
[14] R. van Haasteren and Y. Levin, arXiv:1202.5932.
[15] R. van Haasteren, Y. Levin, P. McDonald, and T. Lu, Mon. Not. R. Astron. Soc. **395**, 1005 (2009).
[16] R. van Haasteren, Mon. Not. R. Astron. Soc. **429**, 55 (2013).
[17] J. Ellis, X. Siemens, and R. van Haasteren, arXiv:1302.1903.
[18] S. Balan, M. A. J. Ashdown, and M. P. Hobson (unpublished).
[19] G. H. Janssen, B. W. Stappers, M. Kramer, M. Purver, A. Jessner, and I. Cognard, AIP Conf. Proc. **983**, 633 (2008).
[20] European Pulsar Timing Array Home Page, www.epta.eu.org.
[21] W. Coles, G. Hobbs, D. J. Champion, R. N. Manchester, and J. P. W. Verbiest, Mon. Not. R. Astron. Soc. **418**, 561 (2011).
[22] R. W. Hellings and G. S. Downs, Astrophys. J. **265**, L39 (1983).
[23] F. Feroz and M. P. Hobson, Mon. Not. R. Astron. Soc. **384**, 449 (2008).
[24] F. Feroz, M. P. Hobson, and M. Bridges, Mon. Not. R. Astron. Soc. **398**, 1601 (2009).
[25] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).
[26] R. M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report No. CRG-TR-93-1, Department of Computer Science, University of Toronto (1993).
[27] J. F. Taylor, M. A. J. Ashdown, and M. P. Hobson, Mon. Not. R. Astron. Soc. **389**, 1284 (2008).
[28] R. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag, New York, 1996).
[29] J. Kennedy and R. C. Eberhart, IEEE Int. Conf. Neural Networks **4**, 1942 (1995).
[30] J. Kennedy and R. C. Eberhart, *Swarm Intelligence* (Morgan Kufmann, San Francisco, 2001).
[31] J. Prasad and T. Souradeep, Phys. Rev. D **85**, 123008 (2012).
[32] S. R. Taylor, J. R. Gair, and L. Lentati, arXiv:1210.3489.
[33] J. Gilbert and C. Lemarchal, Math. Program. **45**, 407 (1989).
[34] A. Azevedo-filho (unpublished).
[35] G. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation* (Springer-Verlag, New York, 1988), p. 48.
[36] R. van Haasteren, C. M. F. Mingarelli, A. Vecchio, and A. Lassus, arXiv:1301.6673.
[37] S. J. Chamberlin and X. Siemens, Phys. Rev. D **85**, 082001 (2012).
[38] G. Hobbs, A. Lyne, and M. Kramer, Chin. J. Astron. Astrophys. **6**, 020000 (2006).
[39] S. R. Taylor, J. R. Gair, and L. Lentati, Phys. Rev. D **87**, 044035 (2013).
[40] J. Pinheiro and D. Bates, Stat. Comput. **6**, 289 (1996).
[41] K. M. Hanson, Proc. SPIE Int. Soc. Opt. Eng. **4322**, 456 (2001).
[42] M. M. Davis, J. H. Taylor, J. M. Weisberg, and D. C. Backer, Nature (London) **315**, 547 (1985).
[43] J. W. T. Hessels, S. M. Ransom, I. H. Stairs, P. C. C. Freire, V. M. Kaspi, and F. Camilo, Science **311**, 1901 (2006).
[44] G. Hobbs *et al.*, Classical Quantum Gravity **27**, 084013 (2010).
[45] G. Hobbs, A. G. Lyne, M. Kramer, C. E. Martin, and C. Jordan, Mon. Not. R. Astron. Soc. **353**, 1311 (2004).
[46] D. R. Madison, S. Chatterjee, and J. M. Cordes, arXiv:1210.2469.