PHYSICAL REVIEW D **72,** 022001 (2005)

Bayesian modeling of source confusion in LISA data

Richard Umstätter,[1,*] Nelson Christensen,[2,†] Martin Hendry,[3,‡] Renate Meyer,[1,§] Vimal Simha,[3,‖] John Veitch,[3,¶] Sarah Vigeland,[2,**] and Graham Woan[3,††]

[1]*Department of Statistics, University of Auckland, Auckland, New Zealand*
[2]*Physics and Astronomy, Carleton College, Northfield, Minnesota 55057, USA*
[3]*Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*
(Received 9 June 2005; published 7 July 2005)

One of the greatest data analysis challenges for the Laser Interferometer Space Antenna (LISA) is the need to account for a large number of gravitational wave signals from compact binary systems expected to be present in the data. We introduce the basis of a Bayesian method that we believe can address this challenge and demonstrate its effectiveness on a simplified problem involving 100 synthetic sinusoidal signals in noise. We use a reversible jump Markov chain Monte Carlo technique to infer simultaneously the number of signals present, the parameters of each identified signal, and the noise level. Our approach therefore tackles the detection and parameter estimation problems simultaneously, without the need to evaluate formal model selection criteria, such as the Akaike Information Criterion or explicit Bayes factors. The method does not require a stopping criterion to determine the number of signals and produces results which compare very favorably with classical spectral techniques.

DOI: 10.1103/PhysRevD.72.022001 PACS numbers: 04.80.Nn, 02.70.Rr, 06.20.Dk

I. INTRODUCTION

The Laser Interferometer Space Antenna (LISA) is designed to detect gravitational radiation from astrophysical sources in the 10^{-2} mHz to 100 mHz band [1]. The sensitivity of LISA is such that very many such sources should be detectable. Paradoxically, this sensitivity will also make signal identification somewhat problematic: parts of the band will likely be swamped with tens of thousands of signals. One of the most abundant classes of source will be close-by white dwarf binaries, producing signals from 0.1 mHz to 3 mHz. There will be source confusion below 1 mHz and resolvable sources above 5 mHz; the 1 mHz to 5 mHz band for LISA therefore presents a tremendous data analysis challenge, potentially containing up to 10^5 sources [2–4]. LISA's ability to detect and characterize other astrophysical sources will be greatly helped if the thousands of background signals from binary systems can be identified. For a detailed look at the population of binary systems that produce signals in LISA's operating band, and how they affect LISA's performance, we direct the reader to Barack and Cutler [2] and to Nelemans *et al.* [5].

We approach this problem from a new direction. Markov chain Monte Carlo (MCMC) techniques have been demonstrated to be especially suited to parameter estimation problems involving numerous parameters [6]. We previously have used the Metropolis-Hastings (MH) algorithm [7,8] in other gravitational radiation problems, such as estimating astrophysical parameters for gravitational wave signals from coalescing compact binary systems [9] or pulsars [10,11]. It is our belief that MCMC methods could provide an effective means for identifying and characterizing the thousands of background binary signals to be found within the LISA data [12].

The method that we present in this paper is not a *source subtraction* approach [13], but one that identifies and characterizes binary produced periodic signals in the data. Signals that are sufficiently large in amplitude will have their parameters estimated. Sources that are weak will contribute to the noise; our method also produces an estimate of the overall level of the noise. We show that the noise level estimate from our method depends on the inherent detector noise level, and also the presence of unidentified signals. MCMC methods are robust and dynamic, and we believe that ultimately it will be possible to use them with LISA data to estimate simultaneously the parameters associated with a wide range of source types occurring in the presence of many thousands of white dwarf binaries.

In this paper we present the results of a simulation study, comprising a data stream of m sinusoidal signals embedded in Gaussian noise. Although our simulation study is simple, it does highlight a number of issues relevant to the real LISA data analysis problem. For example, the number of signals present in the data, m, will be unknown.

Our Bayesian approach does not need to fit each model with m signals, for $m = 1, \ldots M$, and then select the best fitting model via the evaluation of Bayes factors. The evaluation of Bayes factors [14,15] requires computation of the marginal likelihoods and thus marginalizations over the parameter vectors of each model. This is a formidable computational problem when the dimension of the parame-

*Electronic address: richard@stat.auckland.ac.nz
†Electronic address: nchriste@carleton.edu
‡Electronic address: martin@astro.gla.ac.uk
§Electronic address: meyer@stat.auckland.ac.nz
‖Electronic address: vimal_simha@hotmail.com
¶Electronic address: jveitch@astro.gla.ac.uk
**Electronic address: vigelans@carleton.edu
††Electronic address: graham@astro.gla.ac.uk

ter space is large. A shortcut to the calculation of Bayes factors, the harmonic mean of the likelihood values [16], is known to be unstable because the inverse likelihood does not possess a finite variance. Other large sample approximations to the Bayes factors such as the Bayesian Information Criterion, also referred to as *Schwarz Criterion*, and the related penalized likelihood ratio model choice criterion, Akaike Information Criterion, have been shown to be inconsistent when the dimension of the parameter space goes to infinity [17]. The newly developed deviance information criterion [18] is known to be controversial in mixture models. While MCMC algorithms like the Gibbs sampler and MH algorithm yield posterior distributions of the parameters, they do not provide marginal likelihoods. An indirect method of estimating marginal likelihoods from Gibbs sampling output has been developed [19] and extended to output from the MH algorithm [20]. These, however, are impractical when the number of candidate models is very large, as is the case for LISA data. Therefore, we use another strategy, the reversible jump algorithm [21], that samples over the model *and* parameter space in order to estimate posterior model probabilities/marginal likelihoods. We consider the number of sinusoids as an additional parameter and determine its marginal posterior distribution. Its modal value will give us the model with the most probable number of sinusoids. As illustrated in Sec. II, a Bayesian analysis naturally encompasses Occam's Razor [22] and a preference for a simpler (smaller m) model. We thus address and solve the detection and estimation of sinusoids simultaneously. In addition, our MCMC method is better than a classical periodogram at resolving signals that are very close in frequency, and we provide a detailed discussion of how to identify these signals. The natural ability of Bayesian methods to super-resolve signals (to better than the reciprocal of the observing time) depends on signal-to-noise ratio, just as the position of a bright star can be determined to much better than the diffraction limit of a telescope. This interesting topic is explored in detail in [23]. Finally, our method infers the noise level in the data together with the parameters of each of the m sinusoids.

The problem of identifying an unknown number of sinusoids is neither new nor simple [24,25]. Whereas previous studies have looked for a handful of unknown signals, here we show results for 100 signals. Another benefit of using MCMC methods is that computation time scales roughly linearly as the number of parameters increases and does not show an exponential increase in time [6]. In the future we will make the character of the signals more realistic, taking into account the orbit of the LISA satellites and the nature of the inspiral of binaries. With our study of sinusoids we hope to inform LISA data analysis researchers of another possible avenue for characterizing a large number of background signals.

The rest of the paper is organized as follows: Sec. II illustrates that Bayesian inference automatically implements the principle of Occam's Razor. Our Bayesian method and posterior computational algorithms are described in Sec. III. In Sec. IV we present the results of this study using synthesized data. We believe that this method offers great hope for signal extraction in real LISA data, and this point is discussed in Sec. V.

II. OCCAM'S RAZOR

The notion that Bayesian inference naturally implements a quantitative version of Occam's Razor [22] is well known, but this property is sufficiently central to our approach to LISA analysis to warrant a very simple demonstration:

We imagine two physical models, $\mathcal{M}_1$ and $\mathcal{M}_2$, constrained by a single datum, d. $\mathcal{M}_1$ has one parameter, a_1, to describe the datum. $\mathcal{M}_2$ uses the sum of two parameters, $s = a_1 + a_2$ to describe the same datum. Clearly, all other things being equal, if the datum is equally consistent with both models, we would prefer $\mathcal{M}_1$ over $\mathcal{M}_2$ on the grounds that $\mathcal{M}_1$ is "simpler." To see how this naturally occurs within the Bayesian framework we proceed by considering the ratio of probabilities (the *odds ratio*) of the two models:

$$\frac{p(\mathcal{M}_1|d)}{p(\mathcal{M}_2|d)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{p(d|\mathcal{M}_1)}{p(d|\mathcal{M}_2)}. \tag{1}$$

If we set the prior odds ratio to unity, reflecting no prior preference for either model, the right-hand side of the above equation is simply the ratio of marginal likelihoods (sometimes called the *evidences*) of the two models. This ratio is conventionally known as the *Bayes factor* for the comparison. We will take the priors for our parameters a_1 and a_2 to be each uniform in the range $0 \rightarrow R$. Under $\mathcal{M}_1$ we therefore have $p(a_1|\mathcal{M}_1) = 1/R$. The likelihood under $\mathcal{M}_1$ is assumed to be $p(d|a_1, \mathcal{M}_1) = \delta(d - a_1)$, where δ denotes the Dirac function. This yields

$$p(d|\mathcal{M}_1) = \int_0^R \frac{1}{R}\delta(d - a_1)\mathrm{d}a_1 = \begin{cases} 1/R & 0 < d < R \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Under $\mathcal{M}_2$ with parameter $s = a_1 + a_2$, the prior probability density function (pdf) for s is the convolution of the two uniform prior pdfs for a_1 and a_2 and therefore has the "triangle" form

$$p(s|\mathcal{M}_2) = \begin{cases} s/R^2 & 0 < s < R \\ 2/R - s/R^2 & R < s < 2R, \end{cases} \tag{3}$$

so the evidence for $\mathcal{M}_2$ is

$$\begin{aligned} p(d|\mathcal{M}_2) &= \int_0^{2R} p(s|\mathcal{M}_2)\delta(d - s)\mathrm{d}s \\ &= \begin{cases} d/R^2 & 0 < d < R \\ 2/R - d/R^2 & R < d < 2R \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4}$$

(see Fig. 1). If the datum lies in the range $0 < d < R$, the

Bayes factor is $(1/R)/(d/R^2) = R/d$, and $\mathcal{M}_1$ is always favored over $\mathcal{M}_2$. If $d = R$ neither model is preferred, and if $R < d < 2R$ then $\mathcal{M}_2$ (as the only viable model) is a relative certainty. We now see why $\mathcal{M}_1$ is more probable than $\mathcal{M}_2$ when the datum is equally consistent with both (i.e., $d < R$): $\mathcal{M}_2$ has more flexibility than is necessary to explain the datum and so penalizes itself by spreading its evidence more thinly. We can also see that the effect is not simply dependent on the *number* of parameters. If $\mathcal{M}_2$ depended on just a single parameter but with a uniform prior on this parameter over $0 \rightarrow 2R$, then it too would be discriminated against (in the case $0 < d < R$) for using more parameter space to do the same job as $\mathcal{M}_1$.

This quantitative ability to identify the least flexible model consistent with the data makes Bayesian inference the natural methodology in problems where the number of parameters is unknown. Indeed we do not need to consider explicitly Bayes factors at all if we consider the number of parameters itself to be a parameter, and determine its marginal posterior probability in the usual way. This is the approach we will apply to the LISA white dwarf binary problem, as described in the next section.

III. EXTRACTION OF SIGNAL PARAMETERS

We consider a signal consisting of m superimposed sinusoidal signals where m is an unknown parameter. Therefore we confine our attention to a set of models $\mathcal{M}_m$:$m \in \{0, \cdots, M\}$ with M being the maximum number of sinusoidal signals we allow. Let $\boldsymbol{d} = [d_1, \cdots, d_N]$ be a vector of N samples recorded at times $\boldsymbol{t} = [t_1, \cdots, t_N]$. Model $\mathcal{M}_m$ takes the observed data to comprise a signal, $s^{(m)}$ plus noise, $\boldsymbol{\epsilon}^{(m)}[\epsilon_1^{(m)}, \cdots, \epsilon_N^{(m)}]$:

$$d_j = s^{(m)}(t_j, \boldsymbol{a}_m) + \epsilon_j^{(m)}, \quad \text{for } j = 1, \ldots, N, \tag{5}$$

where the noise terms ϵ_j are assumed to be independently and identically distributed $N(0, \sigma_m^2)$ random variables. The signal of model $\mathcal{M}_m$ has the form

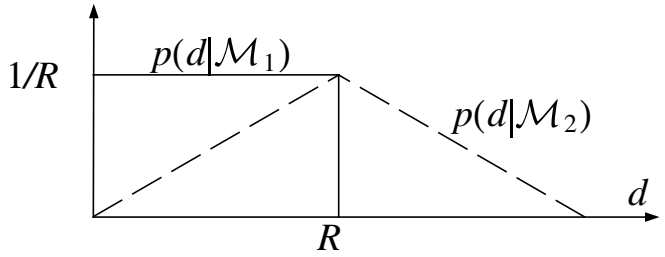

FIG. 1. The evidences for models $\mathcal{M}_1$ (solid line, one parameter fit) and $\mathcal{M}_2$ (dashed line, two parameter fit) as a function of the datum d. Note that the evidence ratio always favors the simpler model $\mathcal{M}_1$ when the datum is equally consistent with either.

$$s^{(m)}(t_j, \boldsymbol{a}_m) = \sum_{i=1}^{m} [A_i^{(m)} \cos(2\pi f_i^{(m)} t_j) + B_i^{(m)} \sin(2\pi f_i^{(m)} t_j)], \tag{6}$$

so that each sinusoid component is characterized by one frequency and two amplitudes. Model $\mathcal{M}_m$ is therefore characterized by a vector of $3m + 1$ unknown parameters which we denote $\boldsymbol{a}_m = [A_1^{(m)}, B_1^{(m)}, f_1^{(m)}, \cdots, A_m^{(m)}, B_m^{(m)}, f_m^{(m)}, \sigma_m^2]$. The objective is to find the model $\mathcal{M}_m$ that best fits the data and to estimate its parameters. We use a Bayesian approach as in [23], but instead of fitting each model separately and selecting the best using Bayes factors, we treat the number m of unknown sinusoids as an additional unknown parameter and determine the mode of its marginal posterior distribution (together with the posterior distribution of all the other parameters). The joint probability that these data $\boldsymbol{d}$ arise from the parameter vector $\boldsymbol{a}_m$ and model $\mathcal{M}_m$ is given by

$$p(\boldsymbol{d}|m, \boldsymbol{a}_m) \propto \frac{1}{\sigma_m^N} \exp\left\{ -\frac{1}{2\sigma_m^2} \sum_{j=1}^{N} [d_j - s^{(m)}(t_j, \boldsymbol{a}_m)]^2 \right\}. \tag{7}$$

For simplicity we take the prior distribution of the model dimension parameter m as uniform over $\{0, \ldots, M\}$. The variances σ_m^2 are given noninformative inverse gamma priors, discussed in Sec. III C.

Again, for simplicity our calculations use dimensionless frequencies with a Nyquist frequency of 0.5, and we use uniform priors for the component frequencies $f_i^{(m)}$ over $[0, 0.5]$.

Given m, σ_m^2, and the frequency vector $\boldsymbol{f}^{(m)}$, our model (5) is a linear regression model which can be written in matrix form as

$$\boldsymbol{d} = \boldsymbol{D}^{(m)} \boldsymbol{b}^{(m)} + \boldsymbol{\epsilon}^{(m)}, \tag{8}$$

where $\boldsymbol{b}^{(m)} = [A_1^{(m)}, B_1^{(m)}, A_2^{(m)}, B_2^{(m)}, \ldots, A_m^{(m)}, B_m^{(m)}]$ is the vector of $2m$ amplitudes and the $N \times 2m$ matrix $\boldsymbol{D}$ contains the entry $\cos(2\pi f_j^{(m)} t_i)$ in row i and column $2j - 1$, and $\sin(2\pi f_j^{(m)} t_i)$ in row i and column $2j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, m$. Thus, an obvious choice for the prior distribution of the amplitudes would be a g-prior [26], a multivariate Normal distribution with mean zero and covariance matrix equal to $g \times (\boldsymbol{D}'\boldsymbol{D})^{-1}$. This g-prior was used in [25] for situations with $m \leq 4$. However, this choice becomes impractical for a large number m of signals since each iteration of the MCMC algorithm would require the calculation and inversion of the $2m \times 2m$ covariance matrix of basis functions. Therefore, for our synthesized data we chose uniform priors for the amplitudes $A_i^{(m)}$, $B_i^{(m)}$ with ranges $A_i^{(m)} \in [-A_{\max}, A_{\max}]$ and $B_i^{(m)} \in [-B_{\max}, B_{\max}]$ and with $A_{\max} = B_{\max} = 5$.

Note that these priors are not uninformative, and we express the parameter space in Cartesian coordinates simply because this is convenient for the implementation of the MCMC algorithms. A model expressed in polar coordinates with uniform priors on amplitude and phase would correspond to a different prior distribution for the transformed amplitudes in Cartesian coordinates. The effect of these different priors on the frequency estimates is generally negligible when the data is informative and is discussed in [23]. The posterior pdf of the frequency reaches its maximum at the same value for both prior choices but with different curvatures at this maximum.

By applying Bayes's theorem, we obtain the posterior pdf for our model parameters of

$$p(m, \boldsymbol{a}_m|\boldsymbol{d}) = \frac{p(m, \boldsymbol{a}_m)p(\boldsymbol{d}|m, \boldsymbol{a}_m)}{p(\boldsymbol{d})}, \tag{9}$$

where $p(\boldsymbol{d}) = \sum_{i=0}^{M} \int p(m, \boldsymbol{a}_m)p(\boldsymbol{d}|m, \boldsymbol{a}_m)\mathrm{d}\boldsymbol{a}_m$. The direct evaluation of the normalization constant $p(\boldsymbol{d})$ is difficult due to the $(3m + 1)$-dimensional integration involved. Moreover, the computation of marginal posterior pdfs would require subsequent $3m$-dimensional integration. To overcome this problem, we use sampling-based MCMC techniques to carry out posterior inference (see [6] for an introduction and overview). These only require the unnormalized posterior $p(m, \boldsymbol{a}_m|\boldsymbol{d}) \propto p(m, \boldsymbol{a}_m)p(\boldsymbol{d}|m, \boldsymbol{a}_m)$ to sample from Eq. (9) and to estimate the quantities of interest. However, in the present context the overall model does not have fixed dimension, and classical MH techniques [7,8] cannot be used to propose trans-dimensional moves. We therefore use the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm for our model determination [21,27]. Additionally we use the delayed rejection (DR) method [28,29] for transitions within the same model. This allows better adaptation of the proposals in different parts of the state space by allowing the choice of the proposal distribution to depend on the proposed but rejected state as well as the current state.

A. The RJMCMC for model determination

To sample from the joint posterior $p(m, \boldsymbol{a}_m|\boldsymbol{d})$ we construct a Markov chain simulation with state space $\cup_{m=1}^{M}(m \times \mathbb{R}^{3m+1})$ where m is the current number of signals. When a new model is proposed we attempt a step between state spaces of different dimensionality. Suppose that at the nth iteration of the Markov chain we are in state $(k, \boldsymbol{a}_k)$. If model $\mathcal{M}_{k'}$ with parameter vector $\boldsymbol{a}'_{k'}$ is proposed, a reversible move has to be considered in order to preserve the detailed balance equations of the Markov chain. Therefore the dimensions of the models have to be matched by involving a random vector $\boldsymbol{r}$ sampled from a proposal distribution with pdf $q(\boldsymbol{r})$, say, for proposing the new parameters $\boldsymbol{a}'_{k'} = \mathrm{t}(\boldsymbol{a}_k, \boldsymbol{r})$ where t is a suitable deterministic transformation function of the current state and $\boldsymbol{r}$. Here we focus on transitions that either decrease or increase models by one signal, i.e. $k' \in \{k - 1, k + 1\}$. We use equal probabilities $p_{k\mapsto k'} = p_{k'\mapsto k}$ to either move up or down in dimensionality, and without loss of generality we consider the upward move $k' = k + 1$.

If the transformation $\mathrm{t}_{k\mapsto k'}$ from $(\boldsymbol{a}_k, \boldsymbol{r})$ to $\boldsymbol{a}'_{k'}$ and its inverse $\mathrm{t}^{-1}_{k\mapsto k'} = \mathrm{t}_{k'\mapsto k}$ are both differentiable, then reversibility is guaranteed if we define the acceptance probability for increasing a model by one signal as

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(\boldsymbol{a}'_{k'}, k')p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')p_{k\mapsto k'}}{p(\boldsymbol{a}_k, k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{r})p_{k'\mapsto k}}|J_{k\mapsto k'}|\right\}, \tag{10}$$

where $|J_{k\mapsto k'}| = \left|\frac{\partial \mathrm{t}(\boldsymbol{a}_{k'}, \boldsymbol{r})}{\partial(\boldsymbol{a}_k, \boldsymbol{r})}\right|$ is the Jacobian determinant of this transformation [21]. In this context, we suggest two types of transformations, ''split-and-merge'' and ''birth-and-death.''

1. Split-and-merge transitions

For a ''split'' transition we randomly choose one of our signals with parameter subvector $\boldsymbol{a}_{(i)} = (A_i^{(k)}, B_i^{(k)}, f_i^{(k)})$ from $\boldsymbol{a}_k$. This signal is chosen by sampling i uniformly from $\{1, \ldots, k\}$. The proposed parameter vector $\boldsymbol{a}'_{k'}$ comprises all the other $(k - 1)$ subvectors of $\boldsymbol{a}_k$ and two additional three-dimensional subvectors, say $\boldsymbol{a}'_{(i_1)} = (A_{i_1}^{(k')}, B_{i_1}^{(k')}, f_{i_1}^{(k')})$ and $\boldsymbol{a}'_{(i_2)} = (A_{i_2}^{(k')}, B_{i_2}^{(k')}, f_{i_2}^{(k')})$ each with half the amplitude of $\boldsymbol{a}_{(i)}$, but same frequency, to replace $\boldsymbol{a}_{(i)}$. A three-dimensional Gaussian random vector (with mean zero) $\boldsymbol{r} = (r_A, r_B, r_f)$ changes the current state $\boldsymbol{a}_{(i)}$ to the two resulting states $\boldsymbol{a}'_{(1i)}, \boldsymbol{a}'_{(2i)}$ through a linear transformation

$$\mathrm{t}_{k\mapsto k'}(\boldsymbol{a}_{(i)}, \boldsymbol{r}) = \begin{pmatrix} \frac{1}{2}A_i^{(k)} + r_A \\ \frac{1}{2}B_i^{(k)} + r_B \\ f_i^{(k)} + r_f \\ \frac{1}{2}A_i^{(k)} - r_A \\ \frac{1}{2}B_i^{(k)} - r_B \\ f_i^{(k)} - r_f \end{pmatrix} = \begin{pmatrix} A_{i_1}^{(k')} \\ B_{i_1}^{(k')} \\ f_{i_1}^{(k')} \\ A_{i_2}^{(k')} \\ B_{i_2}^{(k')} \\ f_{i_2}^{(k')} \end{pmatrix}.$$

The inverse transformation $\mathrm{t}^{-1}_{k\mapsto k'} := \mathrm{t}_{k'\mapsto k}$ accounts for the merger of two signals and can be written as

$$\mathrm{t}_{k'\mapsto k}(\boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)}) = \begin{pmatrix} A_{i_1}^{(k')} + A_{i_2}^{(k')} \\ B_{i_1}^{(k')} + B_{i_2}^{(k')} \\ \frac{1}{2}f_{i_1}^{(k')} + \frac{1}{2}f_{i_2}^{(k')} \\ \frac{1}{2}(A_{i_1}^{(k')} - A_{i_2}^{(k')}) \\ \frac{1}{2}(B_{i_1}^{(k')} - B_{i_2}^{(k')}) \\ \frac{1}{2}(f_{i_1}^{(k')} - f_{i_2}^{(k')}) \end{pmatrix} = \begin{pmatrix} A_i^{(k)} \\ B_i^{(k)} \\ f_i^{(k)} \\ r_A^{(k)} \\ r_B^{(k)} \\ r_f^{(k)} \end{pmatrix}.$$

Note that the determinant of the Jacobian of the transformation $\mathrm{t}_{k\mapsto k'}$, (i.e. the determinant of the above 6-by-6

matrix) is $|J_{k\mapsto k'}| = 2$, and that of its inverse is $1/2$. Thus, the acceptance probability for increasing a model by one signal is given by

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(k', \boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)})p(\boldsymbol{d}|\boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)}, k')}{p(k, \boldsymbol{a}_{(i)})p(\boldsymbol{d}|\boldsymbol{a}_{(i)}, k)q(\boldsymbol{r})}|J_{k\mapsto k'}|\right\}. \quad (11)$$

By analogy, the acceptance probability for the reversible move of a fusion of two signals is given by

$$\alpha_{k'\mapsto k}(\boldsymbol{a}_k|\boldsymbol{a}'_{k'}) = \min\left\{1, \frac{p(k, \boldsymbol{a}_{(i)})p(\boldsymbol{d}|\boldsymbol{a}_{(i)}, k)q(\boldsymbol{r})}{p(k', \boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)})p(\boldsymbol{d}|\boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)}k')}|J_{k'\mapsto k}|\right\}. \quad (12)$$

where $|J_{k'\mapsto k}| = |\frac{\partial(\boldsymbol{a}_{(i)}, \boldsymbol{r})}{\partial(\boldsymbol{a}'_{(1i)}, \boldsymbol{a}'_{(2i)})}|$.

We use a multivariate normal proposal distribution $N[\mathbf{0}, \mathrm{diag}(\sigma_A^2, \sigma_B^2, \sigma_f^2)]$ for $q(\boldsymbol{r})$. Suitable values for the variances of this multivariate normal distribution can be chosen by considering their effect on the acceptance probabilities. First, we consider the effect of a small proposal variance for a splitting transition. In this case the proposal has an insignificant effect on the likelihood since the two new signals in the model function are almost linearly dependent. On the other hand, the resulting large value for $q(\boldsymbol{r})$ considerably decreases the acceptance probability when proposing a split and increases it when a fusion of two signals is proposed. Now we consider the effect of a large proposal variance: the value of $q(\boldsymbol{r})$, and therefore its influence on the acceptance probability, is moderate. However it causes the likelihood to change considerably, resulting in a small acceptance probability. The choice of σ_A^2, σ_B^2, and σ_f^2 is therefore an important consideration for improving mixing. In each iteration we set σ_A^2, σ_B^2 equal to the noise level σ_m^2 of the current model m.

The posterior precision of the frequency in a single-frequency model depends on the signal-to-noise ratio $\gamma = \sqrt{(A^2 + B^2)/\sigma^2}$ and the number of samples N of the data set [30]. Using a Gaussian approximation to the posterior pdf of the frequency [30], its standard deviation is given by $\sigma_f'' = (2\pi\gamma)^{-1}\sqrt{48/N^3}$. This yields a distance in frequency for which two sinusoids can still be identified as distinct, neglecting the interference of other sinusoids. We therefore use σ_f'' as a frequency perturbation when splitting two sinusoids and use a normal distribution with this particular standard deviation when merging.

2. Birth-and-death transitions

A "birth" transformational step simply creates a new signal with parameter triple $\boldsymbol{a}'_{(i)}$ independent of other existing signals in the current model $\mathcal{M}_k$. The one-to-one transformation in this case is very simply given by $\mathrm{t}_{k\mapsto k'}(\boldsymbol{r}) = \boldsymbol{r} = \boldsymbol{a}'_{(i)}$. The inverse ("death") transformation that annihilates signal i', $\mathrm{t}^{-1}_{k\mapsto k'} := \mathrm{t}_{k'\mapsto k}$, has form $\mathrm{t}_{k'\mapsto k}(\boldsymbol{a}'_{(i)}) = \boldsymbol{a}'_{(i)} = \boldsymbol{r}$. The Jacobian for both of these is 1. The acceptance probability for the creation process is therefore

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(k')p(\boldsymbol{a}'_{(i)})p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')}{p(k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{r})}\right\}, \quad (13)$$

and that for the annihilation process is

$$\alpha_{k'\mapsto k}(\boldsymbol{a}_k|\boldsymbol{a}'_{k'}) = \min\left\{1, \frac{p(k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{a}'_{(i)})}{p(k')p(\boldsymbol{a}'_{(i)})p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')}\right\}. \quad (14)$$

As for split-and-merge transitions, a bold proposal distribution q results in a small acceptance probability due to the strong effect on the likelihood, whereas timid proposals have minor effects on the likelihood but are often rejected due the higher values of the proposal distribution. An effective way to make a proposal for the frequencies is to base it on the Schuster periodogram of the data [31], given by

$$C(f) = \frac{1}{N}[R(f)^2 + I(f)^2], \quad (15)$$

where $R(f) = \sum_{j=1}^{N} d_j \cos(2\pi f t_j)$ and $I(f) = \sum_{j=1}^{N} d_j \sin(2\pi f t_j)$ are the real and imaginary parts from the sums of the discrete Fourier transformation of the data. This technique already has been applied successfully by [25]. We use proposals from a normal distribution for the amplitudes with zero mean and a standard deviation derived from the average amplitude of all remaining amplitudes of the current state of the Markov chain.

Classical MCMC methods could be used for transitions within a particular model $\mathcal{M}_m$, however we use an adaptive MCMC technique here. The DR method has been introduced by [28,29,32] and several of us have successfully applied it to estimate the frequency and frequency derivative of potential gravitational radiation signals produced by a triaxial neutron star [11].

B. The delayed rejection method for parameter estimation

As sampling progresses, suppose that at the nth iteration the state of the Markov chain is $\boldsymbol{a} = \boldsymbol{a}_m$ from model $\mathcal{M}_m$. We can choose a new state within the same model by first sampling a candidate state $\boldsymbol{a}'$ from a proposal distribution $q_1(\boldsymbol{a}'|\boldsymbol{a})$ and then accepting or rejecting it with an MH-probability $\alpha_1(\boldsymbol{a}'|\boldsymbol{a})$ depending on the distribution of interest. When a proposed MH move is rejected, a second candidate $\boldsymbol{a}''$ can be sampled with a different proposal distribution $q_2(\boldsymbol{a}''|\boldsymbol{a}', \boldsymbol{a})$ that may depend on the previously rejected proposal. To preserve reversibility of the Markov chain and thus to comply with the detailed balance condition, the acceptance probabilities for both the first and the

second stage are given by [29]

$$\alpha_1(\boldsymbol{a}'|\boldsymbol{a}) = \min\left[1, \frac{p(\boldsymbol{a}')p(\boldsymbol{d}|\boldsymbol{a}')q_1(\boldsymbol{a}|\boldsymbol{a}')}{p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})q_1(\boldsymbol{a}'|\boldsymbol{a})}\right], \tag{16}$$

and

$$\alpha_2(\boldsymbol{a}''|\boldsymbol{a}', \boldsymbol{a}) = \min\left\{1, \frac{p(\boldsymbol{a}'')p(\boldsymbol{d}|\boldsymbol{a}'')q_1(\boldsymbol{a}'|\boldsymbol{a}'')q_2(\boldsymbol{a}|\boldsymbol{a},' \boldsymbol{a}'')[1 - \alpha_1(\boldsymbol{a}'|\boldsymbol{a}'')]}{p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})q_1(\boldsymbol{a}'|\boldsymbol{a})q_2(\boldsymbol{a}''|\boldsymbol{a}, \boldsymbol{a}')[1 - \alpha_1(\boldsymbol{a}'|\boldsymbol{a})]}\right\}. \tag{17}$$

We therefore apply distinct types of DR transition for the amplitudes and the frequency of a sinusoid, and these are considered below. The transitions are performed randomly and with equal probability for a randomly chosen sinusoid i.

1. Proposing new amplitudes

Our amplitude proposal distribution is multivariate normal, with a covariance matrix that depends on the basis functions of the sinusoids. The closer two sinusoids are in frequency the more correlation there is in their recovered amplitudes.

In [25] the amplitudes were regarded as nuisance parameters and integrated out by treating them as parameters of a multiple regression model with a conditional posterior that is normally distributed with known mean and covariance matrix according to the g-prior. If deemed necessary the amplitudes can be updated by Gibbs sampling. However, the computation of the covariance matrix involves determining the inverse of $\boldsymbol{D}^{(m)}(\boldsymbol{f})^T\boldsymbol{D}^{(m)}(\boldsymbol{f})$ for each iteration, which is impractical considering the number of sinusoids expected for LISA. Mindful of this we consider the interference between the sinusoids by proposing those pairs of sinusoids that have the strongest correlation, namely, sinusoid pairs which are closest in frequency. Therefore for the randomly chosen sinusoid i we also involve the sinusoid i_c which is the closest in frequency. The matrix for this subset of basis vectors reduces to

$$\boldsymbol{D}^{(m)}_{(i,i_c)}(\boldsymbol{f}) = \begin{pmatrix} \cos(2\pi f_i^{(m)} t_1) & \sin(2\pi f_i^{(m)} t_1) & \cos(2\pi f_{i_c}^{(m)} t_1) & \sin(2\pi f_{i_c}^{(m)} t_1) \\ \cos(2\pi f_i^{(m)} t_2) & \sin(2\pi f_i^{(m)} t_2) & \cos(2\pi f_{i_c}^{(m)} t_2) & \sin(2\pi f_{i_c}^{(m)} t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(2\pi f_i^{(m)} t_N) & \sin(2\pi f_i^{(m)} t_N) & \cos(2\pi f_{i_c}^{(m)} t_N) & \sin(2\pi f_{i_c}^{(m)} t_N) \end{pmatrix}, \tag{18}$$

which results in a covariance matrix merely of type 4×4 for computing proposals for a sinusoid pair (i, i_c). Now suppose that $\boldsymbol{a}_{(i,i_c)} = (A_i^{(m)}, B_i^{(m)}, A_{i_c}^{(m)}, B_{i_c}^{(m)})^T$ is a vector containing the amplitudes of the sinusoid pair (but not their frequencies). Then

$$\Sigma_{(i,i_c)} = r[\boldsymbol{a}^T_{(i,i_c)}(\boldsymbol{D}^{(m)}_{(i,i_c)}(\boldsymbol{f}))^T(\boldsymbol{D}^{(m)}_{(i,i_c)}(\boldsymbol{f}))\boldsymbol{a}_{(i,i_c)}] \times (\boldsymbol{D}^{(m)}_{(i,i_c)}(\boldsymbol{f})^T\boldsymbol{D}^{(m)}_{(i,i_c)}(\boldsymbol{f}))^{-1} \tag{19}$$

is the covariance matrix of the proposal distribution $q_1(\boldsymbol{a}'_{(i,i_c)}|\boldsymbol{a}_{(i,i_c)}) = N(\boldsymbol{a}_{(i,i_c)}, \Sigma_{(i,i_c)})$, where r is a factor controlling the step size of the proposal and $\times$ represents a normal matrix product.

If the first proposal is rejected, another attempt is made by proposing the amplitudes of a single sinusoid i without considering correlations. The variance of the proposal is assessed by the mean amplitude of all signals which is given by $\bar{a}^{(m)}(\boldsymbol{a}) = (2m)^{-1}\sum_{i=1}^{m}[(A_i^{(m)})^2 + (B_i^{(m)})^2]$. This yields the second proposal $q_2(\boldsymbol{a}''_{(i)}|\boldsymbol{a}) = N[\boldsymbol{a}_{(i)}, r \cdot \mathrm{diag}(\bar{a}^{(m)}(\boldsymbol{a}), \bar{a}^{(m)}(\boldsymbol{a}))]$ where $\boldsymbol{a}_{(i)} = (A_i^{(m)}, B_i^{(m)})^T$ is the subvector of $\boldsymbol{a}$ containing the amplitudes of sinusoid i.

2. Proposing a new frequency

A new frequency is proposed as follows: In the first stage a new frequency for sinusoid i is sampled from a proposal density that is proportional to the periodogram $q_1(\boldsymbol{a}'_{(i)}) \propto (0, 0, C(f))^T$ and independent of the actual stage. This is similar to the sinusoid proposal scheme for RJMCMC. The main objective of this stage is to coarsely scan the whole parameter space for frequencies.

The event of a rejection suggests to sample from the local frequency mode, and a proposal is made conditional on the actual state of the frequency by slightly perturbing the state. In the same manner as in the split-and-merge transition the perturbation is oriented on the possible achievable accuracy $\sigma_{f_i^{(m)}} = (2\pi)^{-1} \times \sqrt{48\sigma_m^2[(A_i^{(m)})^2 + (B_i^{(m)})^2]^{-1}N^{-3}}$ of a frequency by [23]. This yields the proposal $q_2(\boldsymbol{a}''_{(i)}|\boldsymbol{a}_{(i)}) = N(\boldsymbol{a}_{(i)}, \mathrm{diag}(0, 0, \sigma_{f_i^{(m)}}))$ and aims to draw representative samples from the local mode in the second stage.

C. Updating the noise parameter

The sum of the squared residuals between the model and the data, taken from the likelihood in Eq. (7), is

$$S^2 = \sum_{j=1}^{N} [d_j - s_m(t_j, \boldsymbol{a}_m)]^2. \tag{20}$$

Using this, we choose a vague prior for the noise parameter σ_m^2, defined by $IG(\alpha, \beta) = IG(N_p/2, N_p \cdot S_p^2/2)$ with a shape parameter $\alpha = N_p/2 = 0.001$ and a scale parameter $\beta = N_p S_p^2/2 = 0.001$. This yields $N_p = 0.002$ and $S_p^2 = 1$ for the parameters of the vague prior. The full conditional distribution

$$p(\sigma_m^2 | m, \boldsymbol{a}_m, \boldsymbol{d}) \propto IG\left(\frac{N_p + N}{2}, \frac{N_p S_p^2 + S^2}{2}\right) \tag{21}$$

is used for drawing samples for σ_m^2 in a Gibbs update.

D. Initial values

The initial values of a Markov chain are crucial for the length of the burn-in period needed to converge to the real posterior distribution. We could start with an empty model $\mathcal{M}_0$, but it is obvious that it would then take the sampler many steps to find all the signals. Instead, we perform a fast fourier transformation of the data and use this to generate our initial values. We use the frequencies corresponding to the local maxima in the periodogram, $f_{0,i}$, as starting values for $f_{0,i}^{m_0}$ and $A_{0,i} = 2R(f_{\max_i})/N$, $B_{0,i} = 2I(f_{\max_i})/N$ as starting values for $A_i^{m_0}$ and $B_i^{m_0}$, respectively. Theoretically we could use all the local maxima as initial values, but as most of them are due to noise we select only those that exceed a certain threshold. We set this threshold low, as it is easier to delete nonrelevant sinusoids than create good ones.

The frequency resolution depends on sample size N, suggesting that the convergence of the Markov chain is also dependent on N. As we will see, spectral estimates based on the fast fourier transformation are significantly worse than those we obtain by our MCMC method, but they are sufficient to serve as initial values.

E. Identifying the sinusoids

Although the RJMCMC method enables us to select the most probable model, we still encounter the *label-switching problem*. This is a general problem caused by the invariance of the likelihood under relabelling of the sinusoidal components and has been extensively discussed in the context of mixture models [33]. During the MCMC simulation, parameter triples are constantly changing their affiliation to individual sinusoids, either due to the creation and annihilation of sinusoids or following transitions within the same model. We therefore need an additional step in our analysis if we are to break this symmetry and talk meaningfully about individual sinusoid components. This step involves associating the samples in the final Markov chain with particular sinusoids, which we know neither by number nor by location.

The sinusoids that are contained in the model $\hat{\mathcal{M}}$, with the highest posterior probability of m, are permutations of $\hat{m}$ coexistent sinusoids from a list of unknown size. We will assume therefore that this list is the same size as the upper limit $m_{\max}$ of the marginal posterior of m. However, in practice the coexistent sinusoids in $\hat{\mathcal{M}}$ are usually clear, and it is rather unlikely that others present in higher order models could replace them. The parameter vector sampled in each iteration of the Markov chain (corresponding to model $\hat{\mathcal{M}}$) is a permutation of $\hat{m}$ parameter triples determining $\hat{m}$ out of $m_{\max}$ sinusoids. The problem is to determine which parameter triple belongs to which sinusoid. We find that the parameter that contributes significantly to identifying a sinusoid is its frequency. Thus in order to obtain the dominant sinusoids that characterize model $\hat{\mathcal{M}}$ we calculate the marginal posterior of the frequency and obtain the $\hat{m}$ strongest peaks together with their frequency ranges by finding the threshold that separates those peaks.

Since we have to deal with a vast number of output samples grouped within very small regions we cannot apply kernel density estimates or histograms, as the required fixed bin size for a histogram would be too small to be feasible. Instead, we use a variable bin size and calculate densities for a fixed number of samples per bin. We initially sort all individual parameter triples for all MCMC samples of the considered model $\hat{\mathcal{M}}$ by frequency. If we have generated n MCMC samples of model $\hat{\mathcal{M}}$ during a run then we have to deal with $\hat{m}n$ parameter triples and hence $f_1 < \ldots < f_{n\hat{m}}$ ordered frequency samples. The density can be assessed by calculating the frequency range spanned by a fixed number of sorted frequencies. The $\hat{m}n$ parameter triples are from $\hat{m}$ sinusoids and therefore we can expect n frequency samples per peak since we assume the number of peaks to be similar to the number of sinusoids. Hence the fixed number of sorted frequencies that spans the frequency ranges must be some fraction, r, of the number of MCMC samples, n, and is the counterpart to the required bin width in a histogram or the bandwidth h of a kernel density estimate $\mathrm{f}(f) = (2hn\hat{m})^{-1} \sum_{i=1}^{n\hat{m}} \mathrm{I}_{|f-f_i|<h}$ with uniform kernel, where I is the indicator function. The advantage of this approach is the automatic adaptation of the bandwidth to the situation by involving the information of the expected parameter triples per peak. We found $r = 0.05$ (5%) to be a good value, and hence rn serves as an estimate of the number of samples needed for assessing the spans of the frequency ranges. In analogy to the kernel density estimate with uniform kernel, where the number of samples are counted that fall into a range of length $2h$, the density value ρ_j corresponds to each frequency sample j and its $rn - 1$ subsequent samples that fall into a frequency range of length $(f_{j+rn-1} - f_j)$. Hence

the density is given by $\rho_j = rn/[n\hat{m}(f_{j+rn-1} - f_j)] = r/[\hat{m}(f_{j+rn-1} - f_j)]$ where $j = 1, \cdots, n(\hat{m} - r)$. Since each ρ_j comprises the samples $j, \cdots, j + rn - 1$ this has to be considered later when deriving spans for the peaks. We find the smallest density threshold l that separates $\hat{m}$ distinct peaks with respect to the values of ρ_j. The frequency range for each peak $k \in \{1, \cdots, \hat{m}\}$ is $[f_{j_{k,\text{start}}}, f_{j_{k,\text{end}}+rn-1}]$ where $j_{k,\text{start}}$ and $j_{k,\text{end}}$ are the indices of the first and the last member of the set of ρ_j's in peak k, respectively. Because of the fact that we always focus on frequency ranges that contain a fixed amount of frequency samples we efficiently deal with large frequency ranges of low density. This technique is fast and requires a minimum of memory. The greatest computational cost is in sorting the frequencies, although this can be carried out fairly quickly using a heap sort.

It is still possible that individual peaks contain more than one sinusoid or even none. This can be assessed by the histogram of the number of sinusoids in the MCMC samples for the restricted frequency range under consideration. To separate more than one present sinusoid, we then consider the two amplitudes and apply an agglomerative hierarchical cluster analysis that involves all three parameters. We use a modified Ward technique [34] that minimizes the within-cluster variance using a normalized Euclidean distance between the parameters by adjusting the frequency range to the much larger range of the amplitudes. We use the software package R [35] for this task. The Ward technique starts with each parameter triple belonging to a singleton cluster. Iteratively, cluster pairs are joined that produce the smallest possible increase in within-cluster sum of squares. In this particular case we have no difficulty in detecting when to stop the agglomeration, which is an important issue in cluster analysis, as we know the expected number of sinusoids in a peak and hence the number of clusters. However, due to the vast number of parameter triples involved here it is not possible to carry out a cluster analysis simultaneously on all of the data. Instead we have to divide the set of samples into randomly chosen subsets of equal size and perform separate cluster analyses (with R) for each of these subsets. Finally, we perform a cluster analysis on the median points of the subset clusters to allocate each of those clusters to a supercluster. Each single parameter triple is then allocated to a supercluster and hence to a presumed sinusoid. Those parameters which fall into small peaks below the threshold can be allocated to an additional noise contribution.

IV. SIMULATIONS

We created an artificial data set of 1000 samples containing 100 random sinusoids in Gaussian noise. The amplitude coefficients were chosen randomly in the range $[-1 \ldots 1]$ and the noise standard deviation was $\sigma = 1$, making our maximum signal-to-noise ratio $\gamma_i = \sqrt{(A_i^2 + B_i^2)/\sigma^2}$ about 1.4. We will present our results with dimensionless units. We chose a uniform prior for m on $\{0, 1, 2, \ldots, M = 60\,000\}$, and set $A_{\max} = B_{\max} = 5$. Our choice of M is sufficiently high that we would not expect there to be significant evidence for a greater number of sinusoids in the data. In practice, if the marginal posterior for m were to peak at (or close to) M, then this would provide some evidence that more than M sinusoids were present, and the reasons for choosing the value of M should be reexamined. The Markov chain ran for 3×10^8 iterations, the first 5×10^6 of which were considered as burn-in and discarded. The chain was then thinned by storing every 2000th iteration. The MCMC simulation was implemented in C on a 2.8 GHz Intel P4 PC and took about 78 hours to run. Figure 2(a) gives the histogram of the marginal posterior model probabilities obtained by the reversible jump algorithm. As each model $\mathcal{M}_m$ is characterized by a different noise level σ_m, we also have plotted the marginal posterior distributions of the noise standard deviations in Fig. 2(b). Note that σ_m decreases with higher model order m since a model comprising more sinusoids accounts for more of the available power.

All subsequent results presented here are based solely on MCMC samples corresponding to model $\mathcal{M}_{95}$, so we will omit the superscripts and denote the parameter vector of

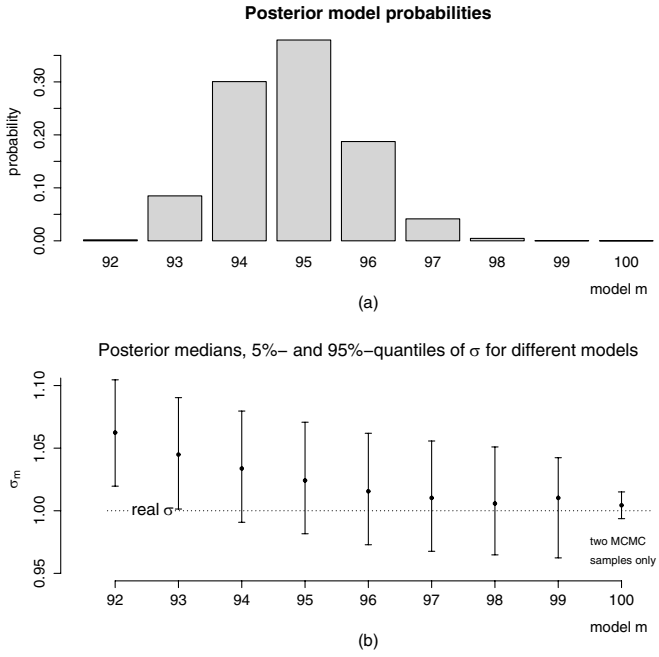

FIG. 2. (a) shows the model probabilities deduced from our analysis. The model corresponding to 95 sinusoids has highest probability. Each model has a different noise level, and (b) shows the estimated noise standard deviation and their confidence intervals for the different models. The dotted line indicates the real noise level. As model 100 has a very low probability only two MCMC were obtained during the run of 3×10^8 iterations, making it poorly determined.

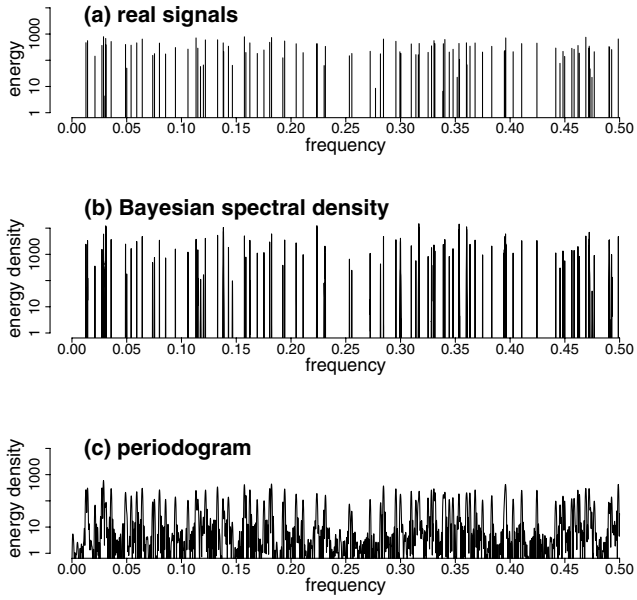

FIG. 3. Comparison of (a) the real signal energies, (b) the Bayesian spectral density, and (c) the classical periodogram of the test data. Note that (a) shows component energies, whereas (b) and (c) show energy densities (inferred energy per unit frequency). The peaks generated by the Bayesian method are higher than those of the periodogram, reflecting its greater ability to constrain the frequency.

model $\mathcal{M}_{95}$ by $(A_1, B_1, f_1, \ldots, A_{95}, B_{95}, f_{95}, \sigma^2_{95})$. The power spectral density of the signal can be estimated from the product of the conditional expectation of the energy of each sinusoid i given its frequency f_i, $E(A_i^2 + B_i^2|\boldsymbol{d}, m, f_i)$, and the posterior pdf of f_i given the data, $p(f_i|\boldsymbol{d})$ as described in [30]. An advantage of using a Bayesian spectral analysis is the subsequent ability to calculate confidence areas for the spectrum by grouping our MCMC samples and calculating posterior confidence intervals for frequency bins. An appropriate width for the bins can be estimated using the frequency resolution of the spectrum, $\sigma_f = (2\pi\gamma)^{-1}(48/N^3)^{1/2}$, given by [23]. In this example the choice of 20 000 bins is sufficient to resolve sinusoids of $\gamma \leq 1.4$.

Figure 3 compares the known real sinusoids from which the data set had been created, the Bayesian spectral density estimate as described above, and the classical periodogram derived from Eq. (15). We note here that Fig. 3(b) and Fig. 3(c) are spectral densities, whereas Fig. 3(a) is a plot of the component energies, defined as $N(A_i^2 + B_i^2)/2$. The theoretical spectral density of these components would consist of delta functions of formally zero width and hence infinite density. Furthermore the accuracy of the Bayesian spectral estimate is better than the one obtained from the periodogram, which also leads to different values on the ordinate. The Bayesian spectrum estimate usually reveals much higher values in the density since the same energy contribution from a sinusoid is distributed over a smaller frequency range. However the locations of the peaks can be compared directly, which is the main purpose of Fig. 3.

We have picked out three different frequency ranges for special attention. Each contain two close sinusoid pairs and we will refer to these as Region-1:$f \in [0.0486, 0.0506]$, Region-2:$f \in [0.1375, 0.1395]$, and Region-3:$f \in [0.3594, 0.3614]$. Joint posteriors for these regions are presented in Fig. 4.

In Region-1 we see two sinusoids with a frequency gap of approximately $1/N = 0.001$ which, for a discrete Fourier transform, is the size of a single-frequency bin. Region-2 and Region-3 contain sinusoids with smaller frequency gaps of 0.000 413 and 0.000 517, respectively. The posteriors for Region-1 are strongly peaked at the true values. The same is true for Region-2 and Region-3, but we observe a larger uncertainty in their frequencies and even more in their amplitudes especially when the frequencies of the sinusoids are very close. The reason for this is clear: when the frequencies of the sinusoids are very close to-

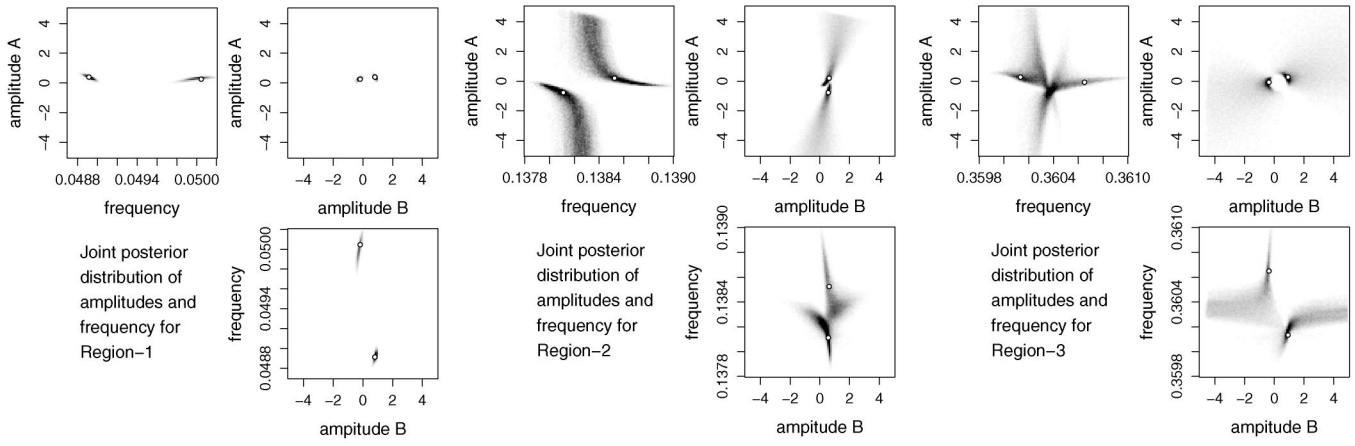

FIG. 4. Joint posteriors for the amplitudes and frequency in Regions-1, -2 and -3. The circles show the *true* values of the parameters comprising signal pairs.

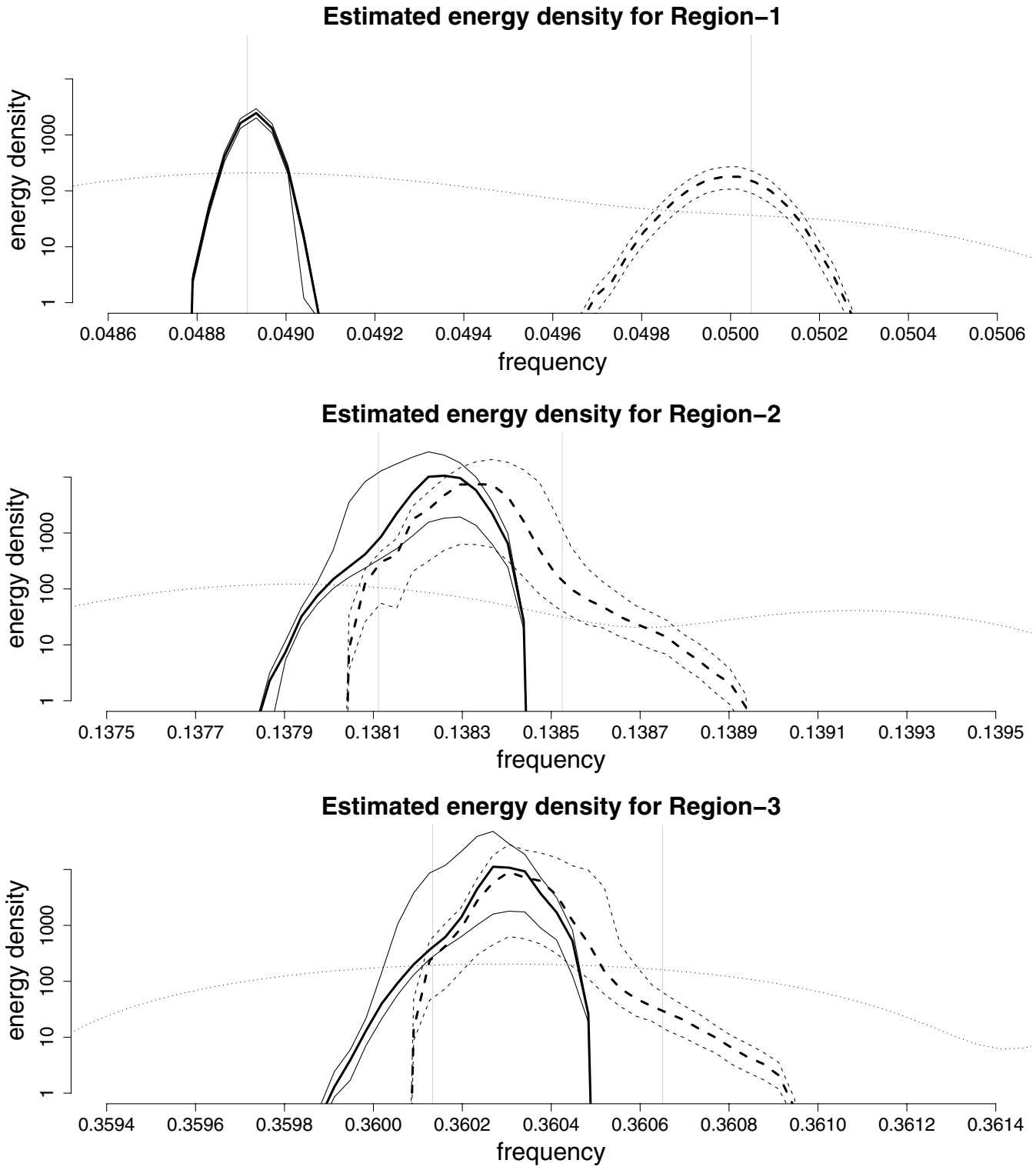

FIG. 5. Bayesian spectral estimates, showing 95% confidence bounds in spectral density, for the three regions considered. Each regions contain two distinct sinusoids, one shown with solid lines and the other with dashed lines. The three lines for each sinusoid show the median value (thick) and 95% bounds (thin) of the spectral estimate at each frequency. The vertical lines show the true frequency values of these components. The dotted line traversing the entire width of the plots is the corresponding classical periodogram.

gether they slowly beat against each other, and for these particular signals the beat period is greater than the observing time so that their total energy is difficult to assess.

The spectral estimates for these regions are shown in greater detail in Fig. 5. Rather than scale the frequency posterior by the expectation of the energy, as in Fig. 3, here we scale by the 2.5%, 50%, and 97.5% quantiles of the energy for each frequency bin to indicate the uncertainty in the value of the spectral density.

The gap between the two sinusoids in Region-1 of Fig. 5 is about $1/N$, and the sinusoid with the lower frequency has a significantly larger energy than its neighbor, so its frequency is inferred with greater accuracy. Region-2 and Region-3 however comprise pairs of considerably closer sinusoids, making it considerably more difficult to infer their separate frequencies. The corresponding uncertainties in their amplitudes can be seen in both Fig. 4 and Fig. 5, which shows large energy density values, especially in the area where the frequencies overlap.

Although spectral estimation is an important topic, for LISA analysis a more pertinent issue is the identification of orbital parameters for the compact binary systems generating the signals. In our example, we can therefore consider the joint posterior of the frequency f_i and the energy $E = N(A_i^2 + B_i^2)/2$ for the six sinusoids in our three regions (Fig. 6). Unlike in Fig. 5, the ordinate now displays an energy rather than an energy density, and the contour is at the 95% credible level.

To better point out the strength of this Bayesian approach, in Fig. 7 we display 95% confidence areas for a wider frequency range, including more sinusoids. As stated in [23], "If the signal one is analyzing is a simple harmonic frequency plus noise, then the maximum of the periodogram will be the best estimate of the frequency that we can make in the absence of additional prior information about it." This is demonstrated in Fig. 7 showing the concordance in frequency estimates for the well-separated sinusoids from both the joint posterior and periodogram plots. However, the periodogram peaks are significantly wider than the 95% areas and are clearly suboptimal for closely spaced sinusoids. In fact the posterior probability density for frequency is the exponent of the ratio of the periodogram $C(f)$ and the noise variance σ^2 [23]. The Bayesian approach therefore takes account of the noise

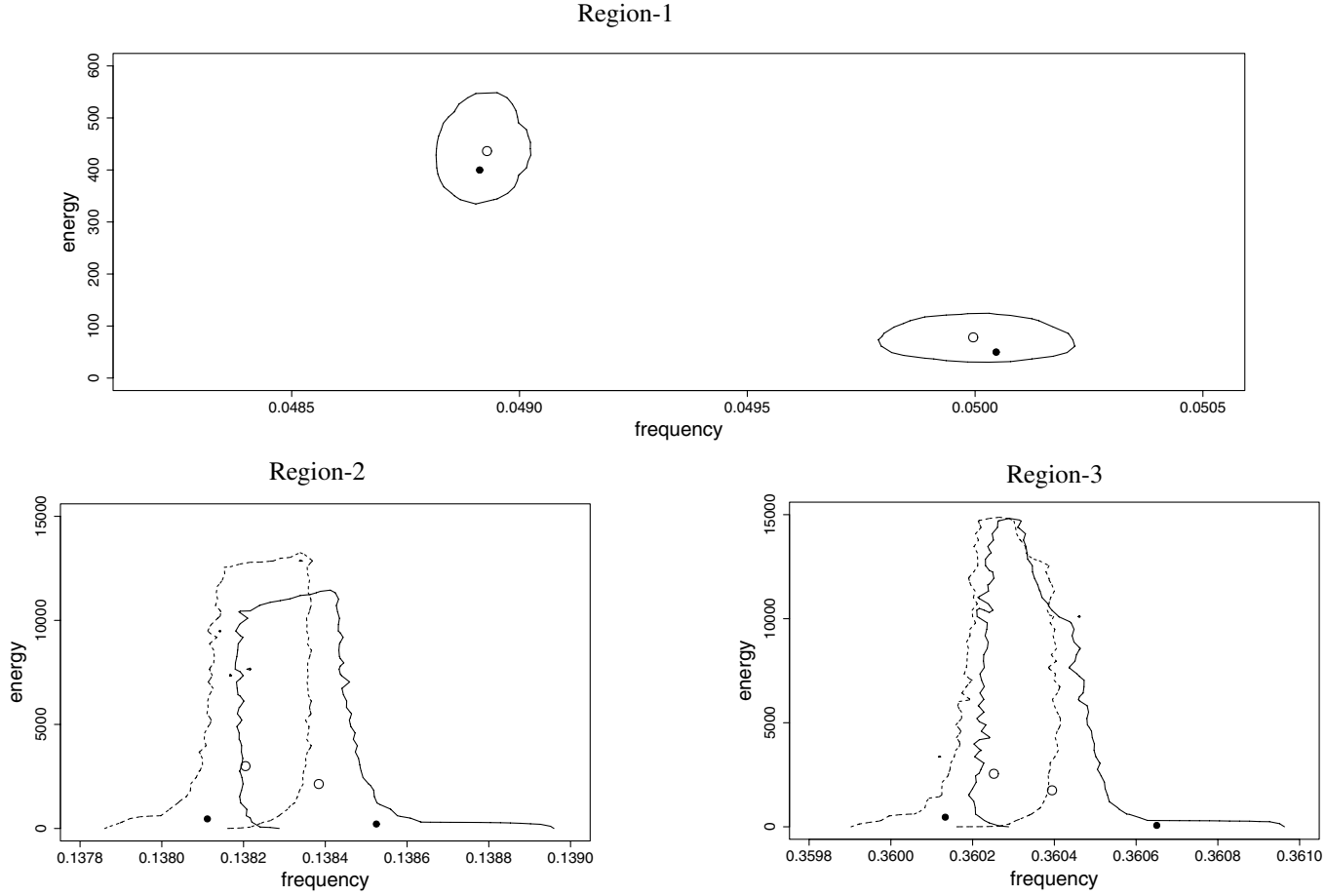

FIG. 6. 95% posterior confidence contours of energy and frequency for the six sinusoids considered. The empty circles show the median value of the posterior peak, and the filled circles show the true value.

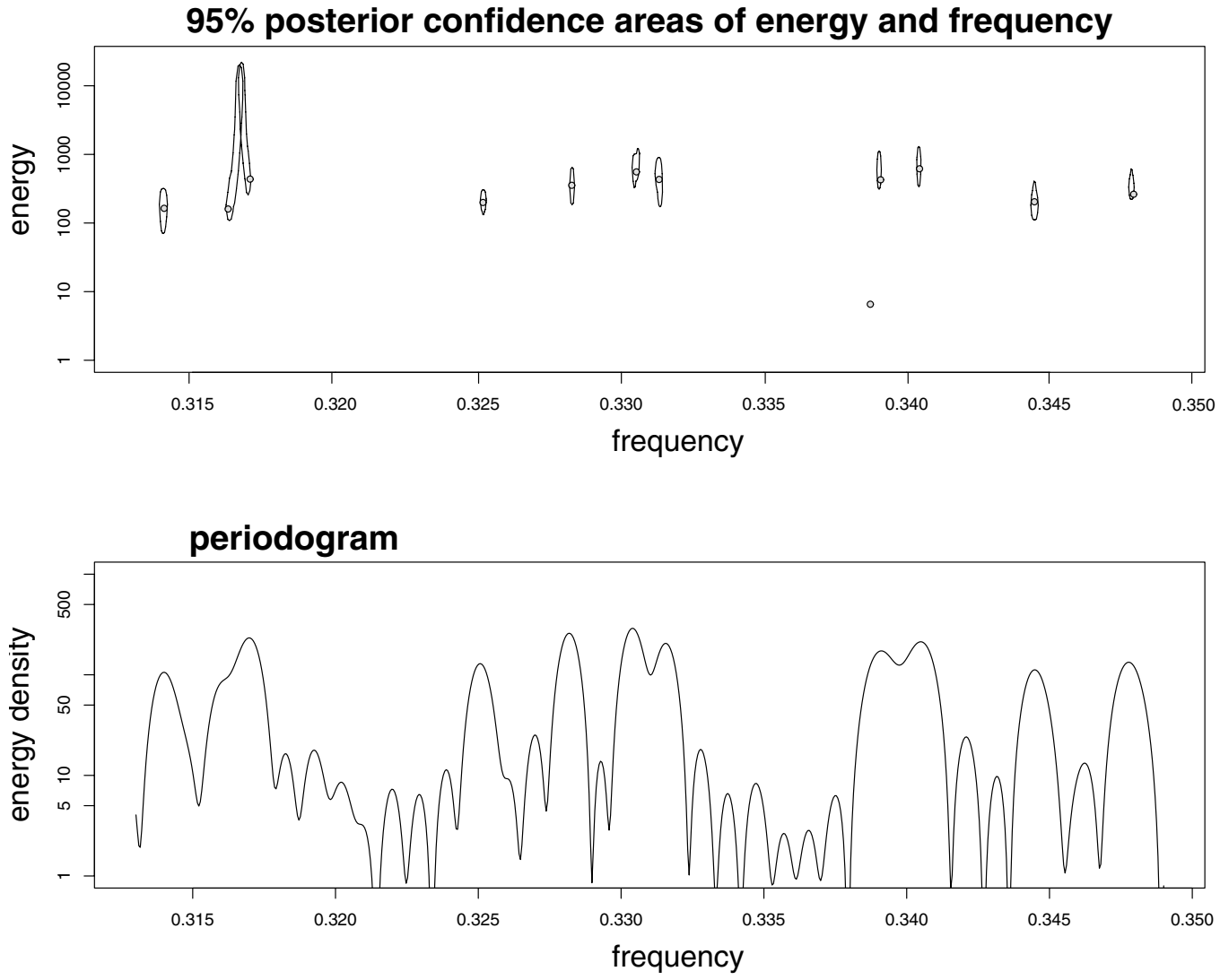

FIG. 7. Top: A section of the (energy, frequency) joint posterior probability density containing eleven 95% posterior confidence areas. The bottom plot shows the classical periodogram for the same frequency range. Twelve sinusoids are actually present in this region, indicated by gray dots and one sinusoid, at $f = 0.3\,387\,209$, was too weak to be identified by our method ($\gamma = 0.11$).

variance in the estimation process, which explains why the confidence regions are significantly narrower than the periodogram peaks.

This Bayesian method therefore shows great power when tackling strong signals close together in parameter space (in this case, frequency). In addition it delivers confidence intervals for the parameters (frequency and amplitude) and can take account of relevant prior information when applied to LISA data.

We have seen that Figs. 4–6 reveal strong interference between very close spaced signal pairs resulting in poor estimation of their parameters. Nevertheless, the Bayesian approach succeeds in revealing even these as separate sinusoids, at a level far beyond the ability of a classical periodogram. To investigate this further, we ran a series of simulations with two sinusoids gradually approaching each other in frequency. The results of these simulations are presented in Fig. 8 which shows how the ability of the method to separate the signals depends on their strengths, their relative phase, and on observing time. In our examples we use $t \in \{0, \ldots, N-1\}$. We ran 10 series of simulations for five different phase shifts $(0, \pi/4, \pi/2, 3\pi/4, \pi)$ between the two equal-amplitude sinusoids, and for two different signal strengths, with $\sigma = 1$. During each set of simulations the frequency gap between the two sinusoids was increased by 1% of a $(1/N)$ step. (Recall that $N = 1000$ so that $1/N = 0.001$). The left-hand column of Fig. 8 displays the results of runs in which the sinusoids have $A_1^2 + B_1^2 = A_2^2 + B_2^2 = 2$. The right-hand column displays results for a quarter of this energy, i.e., $A_1^2 + B_1^2 = A_2^2 + B_2^2 = 0.5$. A high probability for model 2 indicates that the two sinusoids could be

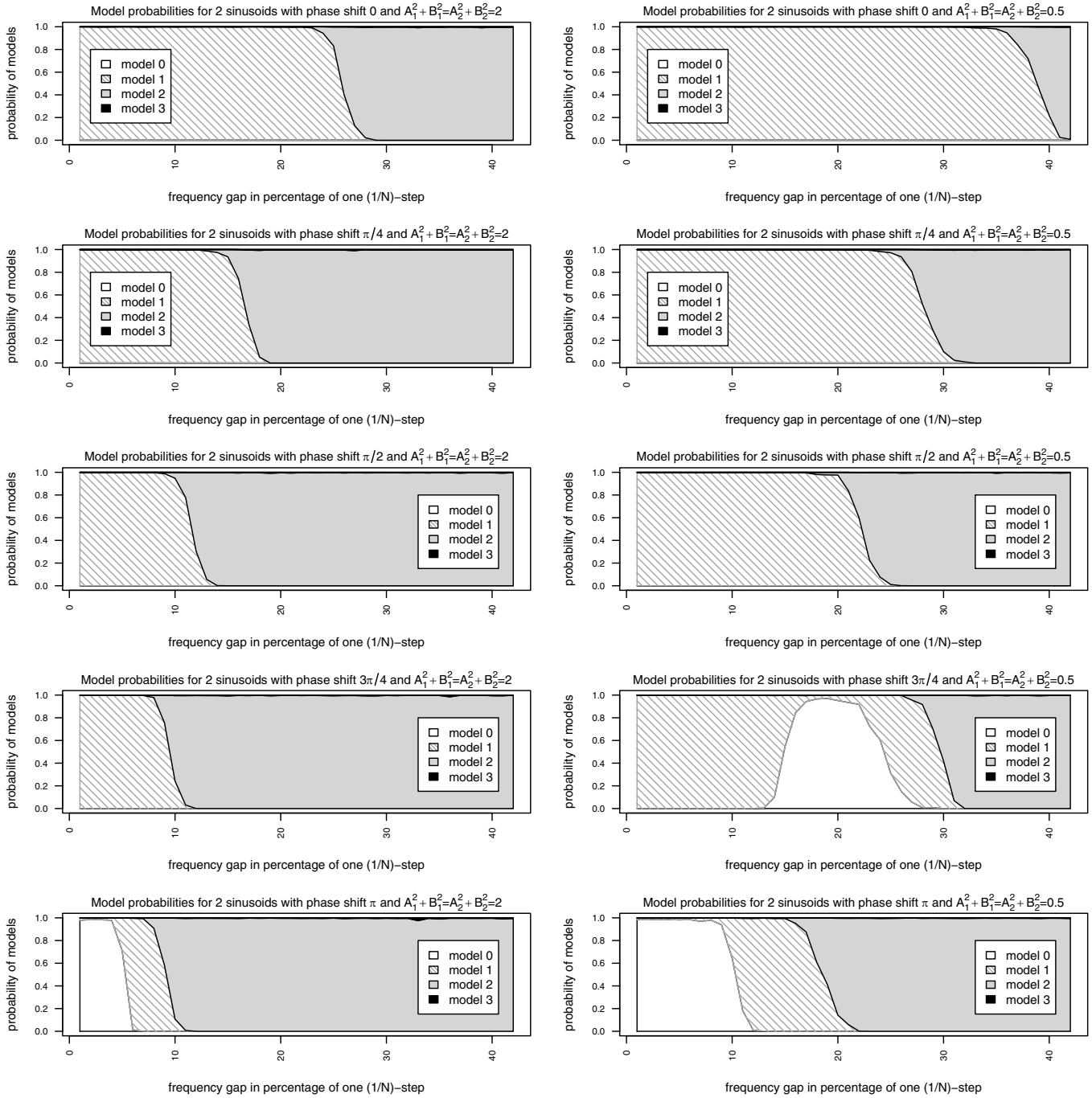

FIG. 8. Model probabilities for different pairs of equal-amplitude sinusoids in Gaussian noise ($\sigma = 1$). Each plot displays two sinusoids with varying frequency gaps expressed as a percentage of one $(1/N)$ step (with $N = 1000$). Each row corresponds to a different phase shift between the sinusoids: from top to bottom the phase shift is $0, \pi/4, \pi/2, 3\pi/4, \pi$. The two columns display results with different signal strengths, as given in the title of the plots.

separated well. If model 1 has the highest probability then the two sinusoids could only be identified as one signal, whereas model 0 indicates that the two sinusoids effectively annihilated each other with respect to the observation time and the resulting signal is too weak to be identified against the background noise. This last effect can only be observed for phase shifts close to π since the amplitudes then have opposite signs and for a frequency gap of around zero almost no resultant amplitude is developed within the observation time considered.

It is obvious that stronger sinusoids are easier to detect and separate, so model 2 is favored with much smaller frequency gaps in the left-hand column of Fig. 8 than in the right-hand column. Furthermore, in both the left- and right-

hand columns the separation ability is worse for small phase shifts and generally improves as the phase shift increases, but this again depends on the noise level. The most striking effect can be seen in the right column at a phase shift of $3\pi/4$. For a small frequency gap the sinusoids can be detected but not separated, then from 14% of a $(1/N)$ step model 0 starts to be favored. One might expect an increase in the probability of model 2 at this stage, but in fact their frequency and phase separation mean the pair almost annihilate each other over the time period considered, so that the resultant cannot be seen against the noisy background. With further separation the effect is reduced, and for larger frequency gaps [30% of a $(1/N)$ step] model 2 does indeed become dominant. This phenomenon also can be seen in the periodogram revealing a single peak slowly decreasing in height from 1% to about 19% of a $(1/N)$ step. Then the signal has the same height as the surrounding noise and is slowly increasing beyond 20% of a $(1/N)$ step. Note that for a different observing period, say $t \in \{t_{\text{start}}, \ldots, t_{\text{start}} + N - 1\}$, the results would have been different. However even in the worst case the detection can be achieved within 40% of a $(1/N)$ step for the weaker sinusoids and within 26% for the stronger sinusoids.

V. DISCUSSION

In this paper we have presented a Bayesian approach to identifying a large number of unknown periodic signals in a set of noisy data. Our reversible jump Markov chain Monte Carlo method can be used to estimate the number of signals present in the data, their parameters, and the noise level. This method compares favorably with classical spectral techniques. Our approach allows for simultaneous detection and parameter estimation and does not require a stopping criterion for determining the number of signals.

Although the parameters of even strong components are not well determined when they are sufficiently close together in frequency, we still obtain useful confidence intervals. Importantly, the noise level is itself a parameter in the overall fit so that the energy present in the data is automatically allocated to either signal or noise. The large uncertainties in amplitudes we obtained for the closely spaced components in our example of 100 sinusoids will, for a practical problem, be lessened by a sensible choice of priors. A prior similar to the g-prior, using a diagonal matrix and a hyper-parameter g as a multiplicative factor, might be a good choice. For the real LISA data we would use priors for the signal amplitudes that would reflect our astrophysical knowledge.

The motivation for our research is to address the difficulty that LISA will ultimately encounter in what is loosely called the *confusion problem*. LISA may see as many as 100 000 signals from binary systems in the 1 mHz to 5 mHz band. We therefore view our work as a powerful new technique for identifying and characterizing these signals in the LISA data stream.

The work presented here is of course a highly simplified toy problem: we are neglecting signal modulation due to LISA's orbit and beam pattern and are not considering an appropriate data model for LISA. In addition, the signals from compact binary systems differ from simple sinusoids. However, we believe that it demonstrates the applicability of the approach to LISA data analysis, and the next step is to deal with these more complicated signals and to develop a realistic strategy for applying our MCMC methods. We do not claim that this will be a trivial extension; in fact, we acknowledge the complexity of the situation. However, we believe that MCMC methods, like those presented here, do give a realistic strategy for identifying and characterizing the large number of signals, of all types, that will exist in LISA data.

Besides LISA, the methods discussed here are likely to be useful in other fields of study where the data contain an unknown number of periodic signals. One timely area of astrophysics is in the field of asteroseismology, where attempts are made to measure vibrational modes of a star; see, for example, Handler *et al.* [36]. We would expect that our method could be directly applied to asteroseismic data on stellar oscillations. Similarly, spectroscopic nuclear magnetic resonance studies are often concerned with spectral decomposition into an unknown number of components, dependent on the composition of the sample [37], and we believe the parameter estimation technique we present could be usefully applied here too.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation Grant Nos. PHY-0071327 and PHY-0244357, the Royal Society of New Zealand Marsden Grant No. UOA204, the Nuffield Foundation and the Particle Physics and Astronomy Research Council, United Kingdom.

[1] K. Danzmann and A. Rüdiger, Classical Quantum Gravity **20**, S1 (2003).
[2] L. Barak and C. Cutler, Phys. Rev. D **70**, 122002 (2004).
[3] M. J. Benacquista, J. DeGoes, and D. Lunder, Classical Quantum Gravity **21**, S509 (2004).
[4] J. Crowder and N. J. Cornish, Phys. Rev. D **70**, 082004 (2004).

[5] G. Nelemans, L. R. Yungleson, and S. F. Protegies Zwart, Astron. Astrophys. **375**, 890 (2001).
[6] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London, 1996).
[7] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[8] W. K. Hastings, Biometrika **57**, 97 (1970).
[9] N. Christensen, R. Meyer, and A. Libson, Classical Quantum Gravity **21**, 317 (2004).
[10] N. Christensen, R. J. Dupuis, G. Woan, and R. Meyer, Phys. Rev. D **70**, 022001 (2004).
[11] R. Umstätter, R. Meyer, R. J. Dupuis, J. Veitch, G. Woan, and N. Christensen, Classical Quantum Gravity **21**, S1655 (2004).
[12] R. Umstätter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigland, and G. Woan, gr-qc/0503121.
[13] N. J. Cornish and S. L. Larson, Phys. Rev. D **67**, 103001 (2003).
[14] R. E. Kass and A. E. Raftery, J. Am. Stat. Assoc. **90**, 773 (1995).
[15] C. Han and B. P. Carlin, J. Am. Stat. Assoc. **96**, 1122 (2001).
[16] M. A. Newton and A. E. Raftery, J. R. Stat. Soc. Ser. B. Methodol. **56**, 3 (1994).
[17] J. O. Berger, J. K. Ghosh, and N. Mukhopadhyay, Journal of Statistical Planning and Inference **112**, 241 (2003).
[18] D. J. Spiegelhalter, N. G. Best, B. R. Carlin, and A. van der Linde, J. R. Stat. Soc. Ser. B. Methodol. **64**, 583 (2002).
[19] S. Chib, J. Am. Stat. Assoc. **90**, 1313 (1995).
[20] S. Chib, and I. J. Jeliazkov, J. Am. Stat. Assoc., **96**, 270 (2001).
[21] P. J. Green, Biometrika **82**, 711 (1995).
[22] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, England, 2003).
[23] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Springer Lecture Notes in Statistics #48 (Springer, New York, 1988).
[24] S. Richardson and P. J. Green, J. R. Stat. Soc. Ser. B. Methodol. **59**, 731 (1997).
[25] C. Andrieu, A. Doucet, IEEE Transactions on Signal Processing **47**, 2667 (1999).
[26] A. Zellner, in *Bayesian Inference and Decision Techniques*, edited by P. Goel and A. Zellner, (Elsevier, New York, 1986), p. 233.
[27] P. J. Green, in *Highly Structured Stochastic Systems*, edited by P. J.Green, N. L. Hjort, and S. Richardson, (Oxford University Press, New York, 2003).
[28] L. Tierney and A. Mira, Statistics in Medicine **18**, 2507 (1999).
[29] A. Mira, Ph.D. thesis, University of Minnesota, 1998.
[30] E. T. Jaynes, in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, edited by C. Ray Smith, G. J. Erickson, and D. Reidel, (Dordrecht, Holland, 1987), p. 1.
[31] A. Schuster, Proc. R. Soc. A **77**, 136 (1905).
[32] P. J. Green and A. Mira, Biometrika **88**, 1035 (2001).
[33] G. Celeux, M. Hurn, and C. P. Robert, J. Am. Stat. Assoc. **95**, 957 (2000).
[34] J. H. Ward, J. Am. Stat. Assoc. **58**, 236 (1963).
[35] R. Gentleman, R. Ihaka, *et al.*, R Development Core Team, *R: A Language and Environment for Statistical Computing*, (R Foundation for Statistical Computing, Vienna, Austria, 2004).
[36] G. Handler *et al.*, Mon. Not. R. Astron. Soc. **347**, 454 (2004).
[37] M. H. Levitt, *Spin Dynamics: Basics of Nuclear Magnetic Resonance*, (Wiley, West Sussex, 2001).