The University of Auckland

Digital thesis consent form

Author of thesis	Richard Umstaetter
Title of thesis	Bayesian Strategies for Gravitational Radiation Data Analysis
Name of degree	PhD

I agree that the University of Auckland Library may make a digital copy of my thesis available for consultation for the purposes of research and private study. It is understood that any person wishing to publish or otherwise use material from my thesis will require my permission before any material from the thesis is published or used.

This consent is in addition to any permissions I have already given for supply to the collection of another prescribed library, or for supply to any person in accordance with the provisions of Section 56 of the Copyright Act 1994.

I certify that the digital copy deposited with the University is the same as the final officially approved version of my thesis. Except in the circumstances set out below no emendation of content has occurred and if there are any minor variations in formatting they are the result of the conversion to digital format.

☒ I confirm that my thesis does not contain material, the copyright for which belongs to a third party

or

☐ I confirm that I have obtained written permission to use all third party copyright material in my thesis and attach copies of each of the permissions required.

or

☐ I confirm that where third party copyright material was included in my thesis and I have not been able to obtain written permission to include this material in the digital version of the thesis I have:
a) removed the third party copyright material from the thesis; and
b) fully referenced the deleted materials; and
c) where possible, provided links to electronic sources of the material

(Please tick the box which applies.)

Signed: .... (author of thesis)

Date: 8.1.2007

Bayesian Strategies for Gravitational Radiation Data Analysis

Richard Umstätter

A thesis
submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy in Statistics,

The University of Auckland, 2006

Abstract

This work addresses the exploration of Bayesian MCMC methods applied to problems in gravitational wave physics. The thesis consists of two parts. In the first part a Bayesian Markov chain Monte Carlo technique is presented for estimating the astrophysical parameters of gravitational radiation signals from a neutron star in laser interferometer data. This computational algorithm can estimate up to six unknown parameters of the target, including the rotation frequency and frequency derivative, using reparametrization, delayed rejection and Metropolis-Coupled Markov Chain Monte Carlo. Results will be given for different synthesized data sets in order to demonstrate the algorithm's behaviour for different observation lengths and signal-to-noise ratios. The probability of detecting weak signals is assessed by a model comparison, based on the BIC, between a model that postulates a signal and one that postulates solely noise within the data.

The second part of the thesis adresses the tremendous data analysis challenges for the Laser Interferometer Space Antenna (LISA) with the need to account for a large number of gravitational wave signals from compact binary systems expected to be present in the data. The basis of a Bayesian method is introduced that can address this challenge, and its effectiveness is demonstrated on a simplified problem involving one hundred synthetic sinusoidal signals in noise. The reversible jump Markov chain Monte Carlo technique is deployed to infer simultaneously the number of signals present, the parameters of each identified signal, and the noise level. This approach is specifically focused on the detection of a large number of sinusoids with separation of sinusoids that are close in frequency. A robust post-processing technique handles the label switching problem by a frequency interval separation technique with a subsequent classification according to a mixed model approximation. The algorithm 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 produces results which compare very favorably with classical spectral techniques.

Acknowledgements

A key feature of modern science is its synergy between traditional disciplines. I would therefore like to thank first of all my supervisors Dr. Renate Meyer and Dr. Nelson Christensen who brought this exciting interdisciplinary project into being and for guiding me through the different worlds of Bayesian statistics and gravitational wave physics. I would also like to extend my thanks to the research group at the University of Glasgow and in particular Dr. Graham Woan for his very helpful inputs.

Special thanks go to my colleagues at the Department of Statistics and in particular to Christian Röver and Dr. Andreas Berg for many fertile discussions.

This work was supported by the Royal Society of New Zealand Marsden Fund Grant UOA204.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, Bayesian methods have become increasingly popular in many different scientific fields. First devised by Bayes [1], the idea of inverse probability was then brought forward in greater generality by Laplace [2] in the beginning of the 19th century and much more clearly by Sir Harold Jeffreys [3] in the first half of the 20th century with a justification by Cox [4]. Despite the fact that the three latter protagonists were physicists applying Bayesian methods to problems in their field, the acceptance of Bayesian methods within the astrophysical community is still rather hesitant [5]. This is remarkable as inference in a pure scientific scope has always been an intuitive version of Bayes' theorem where the state of knowledge is represented by both the prior and the posterior distributions [6]. Scientists are naturally drawn to Bayesian thinking. However, the step to data processing requires complex numerical techniques that are based on Monte Carlo methods as the analytical integration of the posterior distribution is generally not possible. In the outgoing 20th century, Bayesian MCMC methods have received increasing attention from applied statisticians and scientists of fields like biometrics and econometrics and have had a profound effect on Bayesian statistics. Recent research has considerably broadened the diversity of methods and substantiated its theory. Bayesian inference involving MCMC techniques have demonstrated to be especially suited to parameter estimation problems involving numerous parameters [7]. The existence of the thesis at hand reveals the demand on interdisciplinary knowledge of Bayesian statistics and physics to approach the data analysis challenges in modern astronomy and addresses in particular the exploration of Bayesian MCMC methods applied to problems of gravitational wave physics.

Any sufficiently asymmetric distribution of matter that is accelerated emits gravitational radiation. A gravitational wave consists of energy transmitted in the form of a wave through the gravitational field of spacetime and will be emitted by physical objects with a pulsating shape, specifically objects with a changing quadrupole moment like binary systems or a non-precessing triaxial neutron star, called a pulsar. The detection of gravitational radiation is accomplished by laser interferometry. Ground-based interferometers are currently operating in the USA (LIGO), Italy (VIRGO), Germany/UK (GEO600), and Japan (TAMA300). LIGO as the biggest observatory operates two main 4 km gravitational wave observatories in Livingston, Louisiana and on the Hanford Nuclear Reservation near Richland, Washington. A third half-length interferometer operates in parallel also on the Hanford Nuclear Reservation. Although LIGO has recently reached its target sensitivity, the other observatories have not yet reached a level of sensitivity sufficient enough for the actual detection of gravitational radiation.

The data analysis challenges of ground-based gravitational wave detectors led to a fertile collaboration between statisticians and physicists. This gave rise to the development of MCMC techniques specifically designed for ground-based interferometric gravitational radiation problems, such as estimating astrophysical parameters for gravitational wave signals from coalescing compact binary systems [8, 9] or pulsars [10]. This path is pursued by this thesis, aiming for the expansion and development of efficient search algorithms in this field based on the Bayesian principle. The following Chapter 2 addresses the estimation of parameters of a model that postulates a gravitational wave signal from a sufficiently asymmetric pulsar as would be seen from a ground-based interferometer, like LIGO.

However, along with the operation of ground-based interferometers, a space based version is scheduled to be launched in 2015 as a joint project of the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA). The much higher sensitivity of LISA (Laser Interferometer Space Antenna) in the low frequency band will entail a tremendous data analysis challenge. Its sensitivity allows for the detection of 100 000 signals simultaneously, and Chapter 3 aims to explore the capability of Bayesian methods to address the problem of simultaneously estimating noise level, individual signal parameters of an unknown number of signals present within data expected from LISA. This requires the need of a statistical tool for model selection on a very large set of models that considers the countless gravitational wave signals of unknown quantity.

A unique and very important feature of Bayesian theory is model comparison. Bayesian model comparison justifies Occam's Razor which limits the complexity of a model to that amount that best explains the data without redundant degrees of freedom. For model comparison of the complexity described above, the Bayesian framework provides a numerical tool, the reversible jump MCMC (RJMCMC) technique [11, 12]. RJMCMC simultaneously addresses model selection and parameter estimation. This is especially true in situations where the number of models is vast, and where traditional methods that involve the evaluation of formal model selection criteria, such as the Akaike Information Criterion or explicit Bayes factors, are inapplicable.

The RJMCMC technique has demonstrated great potential for applications in Bayesian spectrum analysis by simultaneously estimating the number, noise level and parameters of a few sinusoids [13]. Chapter 3 elaborates an adaptive MCMC algorithm for estimating a large number of sinusoids [14, 15] in order to demonstrate the applicability on LISA data, including post processing tools to solve the label switching problem during the sampling procedure. With regard to the resolving power of the signals, the Bayesian approach is capable of surpassing the Rayleigh limit by an order of magnitude as already shown by E. T. Jaynes [16] and G. L. Bretthorst [17]. The Bayesian method allows for probabilistic statements of the number of signals and the noise level in the data and of the parameters that describe the individual signals.

Chapter 2

Parameter estimation for gravitational wave signals from neutron stars using an adaptive MCMC method

2.1 Introduction

Rapidly rotating neutron stars could be an important source of gravitational wave signals. Several mechanisms have been proposed that would cause them to emit quasi-periodic gravitational waves [18, 19].

Interferometric gravitational wave detectors are now operating in numerous locations around the world [20, 21, 22, 23], and much work has gone into the development of dedicated search algorithms for these signals. Radio observations can provide the sky location, rotation frequency and spin-down rate of known pulsars, and this knowledge simplifies the analysis. This was the case for the recent search for a signal from PSR J1939+2134 [24]. When the position and phase evolution of a source are not known, all-sky hierarchical strategies are required, and these have huge computational requirements [25, 26].

This chapter concentrates on the search for a gravitational wave signal where spin parameters of the rotating neutron star are not well known (but within a narrow band). SN1987A is a good example of a poorly parameterised potential source. If the marginal detection is assumed to be true and SN1987A indeed produced a putative neutron star, then there are large uncertainties in the frequency and spin-down parameters [27]. In particular, this text considers a search with six unknown parameters: the gravitational wave amplitude h_0, the polarization angle ψ (which depends on the position angle of the spin axis in the plane of the sky), the phase of the signal at a fiducial time ϕ_0, the inclination of the spin axis with respect to the line-of-sight ι and the deviations (from reference values) of the signal frequency Δf, and of the frequency derivative $\Delta \dot{f}$.

This problem is approached by a Bayesian Markov chain Monte Carlo (MCMC) technique as MCMC methods have been applied successfully to similar problems involving large numbers of parameters [7]. For a similar search, but with only five parameters (absence of frequency derivative $\Delta \dot{f}$) a Metropolis-Hastings (MH) algorithm [28, 29] had been applied [10]. When the frequency derivative $\Delta \dot{f}$ is included in the basic random walk Metropolis method of [10] the large correlation between Δf and $\Delta \dot{f}$ makes the parameter search difficult, and the basic MH algorithm becomes inefficient. In order to adequately sample the parameter space, a combination of three different strategies have been implemented for accelerating convergence of Markov chains: re-parameterisation, the delayed rejection method of Tierney and Mira [30] (which is an adaptive version of the MH algorithm), and exchange Monte Carlo technique (Metropolis-Coupled Chain, Parallel Tempering) [31] (which is a Monte Carlo approach suited for multimodal posterior sampling). At a previous stage of this research [32, 33] a simulated annealing [34] technique was used. This approach is faster but convergence can be critical if after the lapse of the annealing schedule no convergence could be achieved. Furthermore, the Metropolis-Coupled chain performs better in case of weak signals, a situation which this thesis specifically focus on.

The parameter Δf is highly correlated with $\Delta \dot{f}$, and a strong correlation also exists between h_0 and $\cos \iota$. An initial transformation of these variables to near orthogonality yields a more tractable parameter space that is more effectively sampled. The heterodyne manipulation of the data used in this study is identical to that presented in an end-to-end robust Bayesian method of searching for periodic signals in gravitational wave interferometer data [35], and is also described in [24]. A brief summary of this heterodyne technique is given in Sec. 2.2. Sec. 2.3 and 2.4 describe the Bayesian model and the posterior computational algorithms. In Sec. 2.5 the results of this study are presented, using synthesized signals, for this six parameter problem. The weak signal detection is elaborated in Sec. 2.6. A brief discussion and the long term goals for this work are presented in Sec. 2.7 and 2.8.

2.2 The gravitational wave signal

Gravitational waves from spinning neutron stars are expected to be weak at the Earth, so long integration periods are necessary to extract the signal. It is therefore important to take proper account of the antenna patterns of the detectors and the Doppler shift due to the motion of the Earth.

As in previous studies [10, 24, 35] the signal under consideration is one that is expected from a non-precessing triaxial neutron star. The gravitational wave signal from such an object is at twice its rotation frequency, $f_\mathrm{s} = 2f_\mathrm{r}$, and we characterise the amplitudes of each polarisation with overall strain factor, h_0. The measured gravitational wave signal will also depend on the antenna patterns of the detector for the 'cross' and 'plus' polarisations, $F_{\times,+}$, giving a signal

$$\begin{aligned} s(t) = \; & \tfrac{1}{2} \quad F_+(t;\psi,\alpha,\delta) h_0 (1+\cos^2\iota)\cos\Psi(t;\mathbf{n},f_s,\dot{f}_s) \\ & + \quad F_\times(t;\psi,\alpha,\delta) h_0 \cos\iota \sin\Psi(t;\mathbf{n},f_s,\dot{f}_s). \end{aligned} \tag{2.1}$$

The antenna pattern of the detector depends on time, the polarisation angle ψ of the pulsar and its location determined by right ascension α and declination angle δ. The location is assumed to be known from, for example, radio observations. A simple slowdown model provides the phase evolution of the signal as

$$\Psi(t;\mathbf{n},f_s,\dot{f}_s)=\phi_0+2\pi\left[f_\mathrm{s}(T_{(\alpha,\delta)}-T_0)+\frac{1}{2}\dot{f}_\mathrm{s}(T_{(\alpha,\delta)}-T_0)^2\right], \tag{2.2}$$

where

$$T_{(\alpha,\delta)}=t+\delta t=t+\frac{\mathbf{r}\cdot\mathbf{n}}{c}+\Delta T. \tag{2.3}$$

Here, $T_{(\alpha,\delta)}$ is the time of arrival of the signal at the solar system barycenter, ϕ_0 is the phase of the signal at a fiducial time T_0, $\mathbf{r}$ is the position of the detector with respect to the solar system barycenter that varies with time, $\mathbf{n}$ is a unit vector in the direction of the neutron star (depending on α and δ), c is the speed of light and ΔT contains the relativistic corrections to the arrival time [36] and also is a function of time.

If f_s, $\dot{f}_\mathrm{s}$, and $\mathbf{n}$ are known from (for example) radio observations, the signal can be *heterodyned* by multiplying the data by $\exp[-i\Psi(t;\mathbf{n},f_s,\dot{f}_s)]$, low-pass filtered and resampled, so that the only time varying quantity remaining is the antenna pattern of the interferometer. The reference sky location is needed for the heterodyning process prior to the MCMC simulation.

We are left with a simple model with four unknown parameters h_0, ψ, ϕ_0 and ι. If there is an uncertainty in the frequency and frequency derivative two additional parameters come into play, the differences between the signal and heterodyne frequency and frequency derivatives, Δf and $\Delta\dot{f}$. The unit vector in the direction of the neutron star is supposed to be known by the sky position that is determined by right ascension α and declination δ of the sky position of the pulsar.

A detailed description of the heterodyning procedure is presented elsewhere [24, 35]. An outline of this standard technique is as follows: The raw signal, $s(t)$, is centered at twice the rotation frequency of the neutron star, but is Doppler modulated due to the motion of the Earth and the orbit of the neutron star if it is in a binary system. The modulation bandwidth is typically 10^4 times less than the detector bandwidth, so one can greatly reduce the effective data rate by extracting this band and shifting it to zero frequency. In its standard form the result is one binned data point, B_k, every minute, containing all the relevant information from the original time series but at only 2×10^{-6} the original data rate. If the phase evolution has been correctly accounted for at this heterodyning stage then the only time-varying component left in the signal will be the effect of the antenna pattern of the interferometer, as its geometry with respect to the neutron star varies with Earth rotation. Any small error, Δf, in the heterodyne frequency will cause the signal to oscillate at Δf (plus the residual Doppler shift).

The data points, B_k are assumed to be uncorrelated. Note, that even in case of correlations, the resulting inferences would just yield more conservative results than inferences that account for correlation. In this paper the B_k's are generated using white Gaussian noise with $\mu = 0$ and $\sigma = 10^{-22}$. The variance, σ_k, associated with each bin is therefore known be be unity *a priori* and the noise is uncorrelated between bins [24, 37]. For real data, this assumption may not hold. However, practice has shown that when using sufficiently small bandwidths with GEO and LIGO data the noise is not significantly correlated between bins [24, 37]. It is also assumed that the noise is stationary over the 60 s of data contributing to each bin. This is also consistent with current instrumental performances.

2.3 The Bayesian full probability model

After heterodyning, the signal on which the MCMC analysis is intended to be carried out has the form [35]

$$\begin{aligned} y(t_k;\boldsymbol{a}) = & \ \tfrac{1}{4} \quad F_+(t_k;\psi,\alpha,\delta)h_0(1+\cos^2\iota)e^{i\Delta\Psi(t_k;\alpha,\delta,\Delta f,\Delta\dot{f})} \\ & - \frac{i}{2}F_\times(t_k;\psi,\alpha,\delta)h_0\cos\iota e^{i\Delta\Psi(t_k;\alpha,\delta,\Delta f,\Delta\dot{f})}, \end{aligned} \tag{2.4}$$

where t_k is the time of the k^{th} bin and $\boldsymbol{a} = (h_0,\ \cos\iota,\ \phi_0,\ \psi,\ \Delta f,\ \Delta\dot{f})$ is a vector of the unknown parameters. $\Delta\Psi(t;\alpha,\delta,\Delta f,\Delta\dot{f})$ represents the residual phase evolution of the signal, equalling $\phi_0 + 2\pi[\Delta f(T_{(\alpha,\delta)} - T_0) + \Delta\dot{f}(T_{(\alpha,\delta)} - T_0)^2/2]$, where $T_{(\alpha,\delta)}$ (Eq. 2.3) depends on the known sky location of the pulsar. The objective is to fit this model to the data

$$B_k = y(t_k; \boldsymbol{a}) + \epsilon_k, \tag{2.5}$$

where ϵ_k is assumed to be normally distributed noise with a mean of zero and known variance σ_k^2. Assuming statistical independence of the binned data points, B_k, the joint likelihood that these data $\boldsymbol{d} = \{B_k\}$ arise from a model with a certain parameter vector $\boldsymbol{a}$ is [35]

$$p(\boldsymbol{d}|\boldsymbol{a}) \propto \prod_k \exp\left[-\frac{1}{2}\left|\frac{B_k - y(t_k; \boldsymbol{a})}{\sigma_k}\right|^2\right] = \exp\left[\frac{-\chi^2(\boldsymbol{a})}{2}\right], \tag{2.6}$$

where

$$\chi^2(\boldsymbol{a}) = \sum_k \left|\frac{B_k - y(t_k; \boldsymbol{a})}{\sigma_k}\right|^2. \tag{2.7}$$

In order to draw any inference on the unknown parameter vector $\boldsymbol{a}$ the (posterior) probability is needed of $\boldsymbol{a}$ given $\boldsymbol{d}$, which can be obtained from the likelihood via an application of Bayes' theorem. The unnormalised posterior density

$$p(\boldsymbol{a}|\boldsymbol{d}) \propto p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a}) \tag{2.8}$$

is the product of the prior density of $\boldsymbol{a}$, $p(\boldsymbol{a})$, and the joint likelihood. Accordingly, appropriate priors have to be chosen for the particular parameters. In this study uniform priors are used with prior ranges $[0, 2\pi]$, $[-\pi/4, \pi/4]$ and $[-1, 1]$ for the angle parameters ϕ_0, ψ and $\cos\iota$ respectively.

For h_0, a uniform prior is specified with boundary $[0, 1\,000]$ in units of the rms noise [35]. For the frequency and spindown uncertainty, suitable uniform priors are used with ranges of $[-\frac{1}{120}, \frac{1}{120}]\,\mathrm{Hz}$ and $[-10^{-9}, 10^{-9}]\,\mathrm{Hz\,s^{-1}}$ for Δf and $\Delta\dot{f}$, respectively.

The normalized posterior density $p(\boldsymbol{a}|\boldsymbol{d}) = p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})/p(\boldsymbol{d})$ cannot be evaluated analytically, due to the difficult to compute normalisation constant. Therefore Monte Carlo methods are used here to explore $p(\boldsymbol{a}|\boldsymbol{d})$. If samples can be drawn from $p(\boldsymbol{a}|\boldsymbol{d})$, all interesting quantities can be estimated, including the posterior means of all parameters from the corresponding sample means, to any desired accuracy by increasing the sample size.

2.4 The adaptive Metropolis-Hastings algorithm

However, drawing independent samples in a six-dimensional parameter space is not feasible. It has already been shown that MCMC methods can be used to parameterise gravitational wave signals of low signal-to-noise ratio [10] with four unknown parameters. These simulate a Markov chain, constructed so that its stationary distribution coincides with the posterior distribution and the sample path averages converge to the expectations. A minimal requirement for this is the irreducibility of the chain and hence the ability of the chain to reach all parts of the state space [7]. A technique producing a large class of MCMC algorithms is the MH algorithm [28, 29] which does not require the normalization constant, only the unnormalised posterior density of Eq. 2.8. The MH algorithm has been employed for the four and five parameter pulsar detection problems [10]. The efficiency of the MH algorithm depends heavily on the choice of the proposal density. Intuition suggests that the closer the proposal distribution is to the target, the faster convergence to stationarity is achieved. Default choices such as a Gaussian proposal or a random walk result in very slow mixing for this 6-parameter problem. To increase the speed of convergence, an *adaptive* technique has been employed, adaptive in the sense that it allows the choice of proposal distribution to depend upon information gained from the already sampled states as well as the proposed but rejected states. This method does not break the Markov property that guarantees correctness of MCMC.

2.4.1 The delayed rejection method

The idea behind the delayed rejection algorithm specified by [30] is that persistent rejection, perhaps in particular parts of the state space, may indicate that locally the proposal distribution is badly calibrated to the target. Therefore, the MH algorithm is modified so that on rejection, a second attempt to move is made with a proposal distribution that depends on the previously rejected state. This adaptive Monte-Carlo method [30] was generalized for the variable dimension case [38] and renamed the ‘delayed rejection method’. The original version [30], and also the generalized version [38] have been implemented. It turned out that the generalised delayed rejection method was not that beneficial for this particular problem and thus the original delayed rejection algorithm [30] is explained here.

For the Metropolis-Hastings algorithm, a new state in a Markov chain is chosen first by sampling a candidate $\boldsymbol{a}'$ from a certain proposal distribution $q_1(\boldsymbol{a}'|\boldsymbol{a}_n)$ usually depending on the current state $\boldsymbol{a}_n$ and then accepting or rejecting it with a probability $\alpha_1(\boldsymbol{a}'|\boldsymbol{a}_n)$ depending on the distribution of interest. This rejection is essential for the convergence of the chain to the intended target distribution. The choice of a good proposal distribution is important to avoid persistent rejections in order to achieve good convergence of a chain. However in different parts of the state space, different proposals are required. When a proposed MH move is rejected, a second candidate $\boldsymbol{a}''$ can be sampled with a different proposal distribution $q_2(\boldsymbol{a}''|\boldsymbol{a}', \boldsymbol{a}_n)$ that can depend on the previously rejected proposal. Since a rejection suggests a bad fit of the first proposal, a different form of proposal can be advantageous in the second stage. 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 [39]

$$\boldsymbol{\alpha}_1(\boldsymbol{a}'|\boldsymbol{a}_n) = \min\left(1, \frac{p(\boldsymbol{a}')p(\boldsymbol{d}|\boldsymbol{a}')q_1(\boldsymbol{a}_n|\boldsymbol{a}')}{p(\boldsymbol{a}_n)p(\boldsymbol{d}|\boldsymbol{a}_n)q_1(\boldsymbol{a}'|\boldsymbol{a}_n)}\right) \tag{2.9}$$

and

$$\boldsymbol{\alpha}_2(\boldsymbol{a}''|\boldsymbol{a}_n) = \min\left(1, \frac{p(\boldsymbol{a}'')p(\boldsymbol{d}|\boldsymbol{a}'')q_1(\boldsymbol{a}'|\boldsymbol{a}'')q_2(\boldsymbol{a}_n|\boldsymbol{a}', \boldsymbol{a}'')[1-\boldsymbol{\alpha}_1(\boldsymbol{a}'|\boldsymbol{a}'')]}{p(\boldsymbol{a}_n)p(\boldsymbol{d}|\boldsymbol{a}_n)q_1(\boldsymbol{a}'|\boldsymbol{a}_n)q_2(\boldsymbol{a}''|\boldsymbol{a}_n, \boldsymbol{a}')[1-\boldsymbol{\alpha}_1(\boldsymbol{a}'|\boldsymbol{a}_n)]}\right) \tag{2.10}$$

respectively. Fig. 2.1 illustrates the idea of delayed rejection. When the second stage proposal step is applied due to rejection of the first, the chain has, in order to preserve the reversibility, to imply a return path which comprises a fictive stationary Markov chain consisting of a fictive stage 1 proposal step from $\boldsymbol{a}''$ to $\boldsymbol{a}'$ which is rejected followed by an accepted fictive second stage move to $\boldsymbol{a}_n$ [38]. Although the delayed rejection method provides better acceptance rates over the two stages, cross-correlations between the parameters still impede convergence of the Markov chain. Preliminary runs reveal that especially the parameters Δf and $\Delta\dot{f}$ are highly correlated. To a certain extent h_0 and $\cos\iota$ are also correlated after the Markov chain has found a potential mode. The consequence of which is poor mixing of the chain and therefore a re-parameterisation is required.

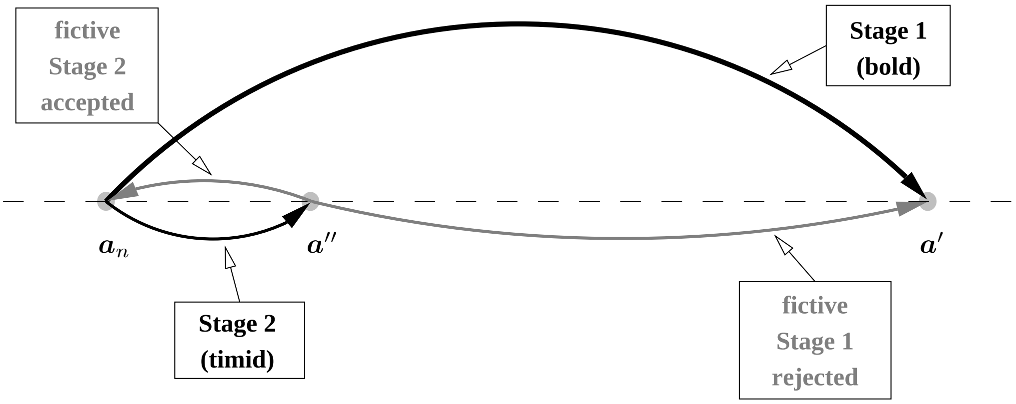


Figure 2.1: Diagram of the delayed rejection method. In case of rejection of the first, bold step a second, more timid move is proposed. In order to maintain the reversibility of the Markov Chain the acceptance probability has to consider a fictive return path.

2.4.2 Re-parameterisation

The coherence between Δf and $\Delta\dot{f}$ is obvious since the data is sampled from time t_{start} to t_{end}, where the heterodyned signal traverses a frequency from $f_{\text{start}} = \Delta f + \frac{1}{2}\Delta\dot{f}\cdot t_{\text{start}}$ to $f_{\text{end}} = \Delta f + \frac{1}{2}\Delta\dot{f}\cdot t_{\text{end}}$; time $t = 0$ is an epoch time when $f = \Delta f$.

Herein lies the major problem. When proposing new samples of $\Delta\dot{f}$, the frequency at time $t = 0$ is not affected at all whereas the impact on the frequency values at time points that lie far ahead or after the epoch time $t = 0$ is huge, especially when the epoch time is badly chosen outside the actual observation period. This is an issue that should be strictly avoided. But even for an epoch time that lies ideally in the middle of the observation period, implicating $t_{start} = -t_{end}$, we encounter this problem.

A re-parameterisation is required that controls both Δf and $\Delta\dot{f}$ in a way that it affects the frequencies at all time points in an equal measure. Hence it is much more natural to propose values for f_{start} and f_{end} instead and transform those parameters back into Δf and $\Delta\dot{f}$.

The original parameters are then obtained by the simple linear transformation

$$\Delta\dot{f} = 2 \cdot \frac{f_{\text{end}} - f_{\text{start}}}{t_{\text{end}} - t_{\text{start}}}. \tag{2.11}$$

and

$$\Delta f = f_{\text{start}} - \frac{1}{2}\Delta\dot{f} \cdot t_{\text{start}} = f_{\text{start}} - \frac{f_{\text{end}} - f_{\text{start}}}{t_{\text{end}} - t_{\text{start}}} \cdot t_{\text{start}} \tag{2.12}$$

Since the Jacobian of this transformation is constant, the prior distributions for the new parameters f_{start} and f_{end} are flat as well. It is notable that the new parameter space changes from a rectangle shape into a parallelogram that satisfies $f_{\text{start}} < f_{\text{end}}$. In practice, one is working in the $(f_{\text{start}}, f_{\text{end}})$ space and the boundaries are verified by back transformation into the $(\Delta f, \Delta\dot{f})$ space. It is important that a proposed transition outside the boundaries is counted as an iteration with a regular rejected proposal where the prior probability and hence the posterior probability is zero.

Another cross-correlation can be observed between the parameters h_0 and $\cos\iota$ that arises from the fact that h_0 can be seen as a scaling factor and $\cos\iota$ as a non-linear weighting between the plus and cross polarisation part of the model. As seen in Eq. 2.4, the plus part is multiplied by the factor $a_1 = \frac{1}{4}h_0(1+\cos^2\iota)$ while the cross part encloses the term $a_2 = \frac{1}{2}h_0\cos\iota$. The original parameters can be derived from

$$h_0 = 2\left(a_1 + \sqrt{a_1^2 - a_2^2}\right), \tag{2.13}$$

and

$$\cos\iota = \frac{2a_2}{h_0}. \tag{2.14}$$

As mentioned above, the prior of the parameters h_0 and $\cos\iota$ are chosen uniform with joint probability density function

$$f(h_0, \cos\iota) = \begin{cases} (2l_{h_0})^{-1}, & \text{if } 0 \leq h_0 < l_{h_0},\ -1 \leq \cos\iota \leq 1, \\ 0, & \text{otherwise,} \end{cases} \tag{2.15}$$

where for this study $l_{h_0} = 1\,000$ is in units of the rms noise. This implies a joint prior distribution for the parameters a_1 and a_2 of the form

$$g(a_1, a_2) = \left\{ \begin{array}{ll} (2l_{h_0})^{-1}, & \text{if } |a_2| \leq a_1 < \frac{4a_2^2 + l_{h_0}^2}{4l_{h_0}} \leq \frac{l_{h_0}}{2} \\ 0, & \text{otherwise} \end{array} \right\} |\det J| \tag{2.16}$$

with Jacobian

$$\begin{aligned} \det J &= \begin{pmatrix} \frac{\partial h_0}{\partial a_1} & \frac{\partial h_0}{\partial a_2} \\ \frac{\partial(\cos\iota)}{\partial a_1} & \frac{\partial(\cos\iota)}{\partial a_2} \end{pmatrix} \\ &= \begin{pmatrix} 2\left(1 + \frac{a_1}{\sqrt{a_1^2 - a_2^2}}\right) & -2\frac{a_2}{\sqrt{a_1^2 - a_2^2}} \\ -\frac{a_2\left(1 + \frac{a_1}{\sqrt{a_1^2 - a_2^2}}\right)}{\left(a_1 + \sqrt{a_1^2 - a_2^2}\right)^2} & \frac{a_2^2}{\sqrt{a_1^2 - a_2^2}\left(a_1 + \sqrt{a_1^2 - a_2^2}\right)^2} + \frac{1}{a_1 + \sqrt{a_1^2 - a_2^2}} \end{pmatrix} \\ &= \frac{2}{a_1 + \sqrt{a_1^2 - a_2^2}} + \frac{2a_1}{\sqrt{a_1^2 - a_2^2}(a_1 + \sqrt{a_1^2 - a_2^2})} \\ &= \frac{2}{\sqrt{a_1^2 - a_2^2}}. \end{aligned} \tag{2.17}$$

Since the Jacobian is positive for the above restrictions we can write

$$g(a_1, a_2) = \begin{cases} \frac{1}{l_{h_0}\sqrt{a_1^2 - a_2^2}}, & \text{if } |a_2| \leq\ a_1 < \frac{4a_2^2 + l_{h_0}^2}{4l_{h_0}} \leq \frac{l_{h_0}}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.18}$$

This joint prior density has the shape shown in Fig. 2.2. These re-parameterisations result in a faster mixing Markov chain. In fact, especially the re-parameterisation of the frequency parameters is the key of obtaining mixing at all, due to the extreme accuracy they have to be matched with many narrow modes in the posterior distribution. Hence, the choice of suitable proposal distributions is

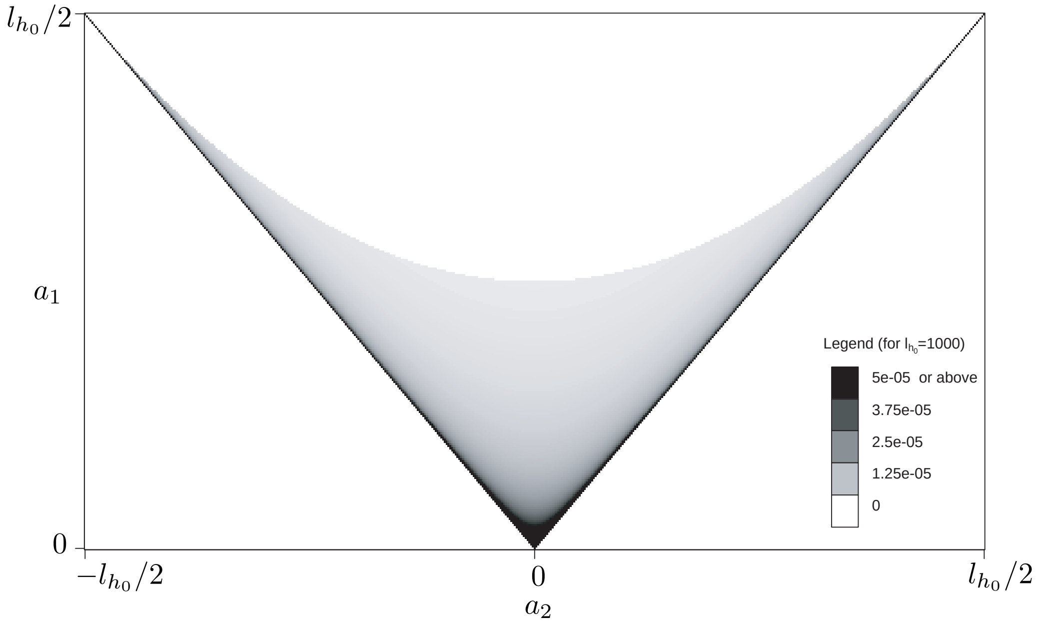


Figure 2.2: Joint prior density of a_1 and a_2 for a given boundary l_{h_0} for the parameter h_0.

essential, in particular for the frequency parameters.

2.4.3 The choice of proposal distributions

Therefore the parameter set is divided into three blocks to account for the different coherences between the parameters. Only parameters within a block are proposed at a time. The blocks under consideration are I : $\{a_1, a_2\}$, II : $\{f_{\text{start}}, f_{\text{end}}\}$, and III : $\{\psi, \phi_0\}$. One positive side effect here is that the antenna pattern that is only dependent on the parameter ψ is not needed to be recalculated when parameters of the first or second block are being proposed. This speeds up the likelihood calculation significantly.

Considering the three different blocks $b \in \{\text{I}, \text{II}, \text{III}\}$, usually, a multivariate Normal distribution is utilized for the proposal distributions $q_1(\boldsymbol{a}^{(b)'}|\boldsymbol{a}_n^{(b)})$ and $q_2(\boldsymbol{a}^{(b)''}|\boldsymbol{a}^{(b)'}, \boldsymbol{a}_n^{(b)})$, with means equal to the current state and different variances depending on the stage. Following Fig. 2.1, larger variances are chosen for the 'bold' first stage steps, while smaller variances are more beneficial for the 'timid' second stage candidates.

For parameter block I the proposals for the parameters a_1 and a_2 are sampled independently since they represent scaling factors for the plus and cross polarisation part, respectively. A bivariate Normal distribution with diagonal matrix is applied here with large standard deviations at stage 1 and a 100 times smaller standard deviation for stage 2.

The covariance matrix in block II that is determined by the proposal variances of f_{start} and f_{end} and their correlation ρ_f, indirectly controls the parameter $\Delta\dot{f}$. The logical choice is to use equal variances σ_f^2 for proposals for f_{start} and f_{end} with a correlation ρ_f. Then the variance by which $\Delta\dot{f}$ is affected is easily derived and given by $\sigma^2_{\Delta\dot{f}} = \frac{8\sigma_f^2}{(t_{\text{end}}-t_{\text{start}})^2} \cdot (1-\rho_f)$. This signifies that a correlation of $\rho_f = 1$ implies proposals for f_{start} and f_{end} that do not effect $\Delta\dot{f}$. On the other hand a correlation of $\rho_f = 0$ controls f_{start} and f_{end} in such a way that the interaction of Δf and $\Delta\dot{f}$ have a moderate effect on the frequency course during the observation time, decoupled from the arbitrary chosen epoch time.

Therefore, a correlation of $\rho_f = 1$ is applied at the first stage of the delayed rejection that just affects and proposes a change in the value of Δf. This stage is targeted for finding the frequency of the signal with large proposal variances σ_f^2. At the second stage of the delayed rejection, the choice of $\rho_f = 0$ will take care of the interaction between Δf and $\Delta\dot{f}$ resulting in a moderate and sensible frequency course over the observation period, affecting the frequencies values at all time points in equal measure. The proposal variance σ_f^2 at this stage has to be chosen rather small.

The proposal variances σ_f^2 at both stages have to be treated wisely. As already mentioned at the end of Sec. 2.4.2, the posterior distribution features very narrow modes caused by the frequency parameters Δf and $\Delta\dot{f}$ in a large parameter space that has to be scanned. Thus a simple Normal distribution is not suitable for a proposal distribution as pilot runs have revealed. Also the t-distribution is not sufficiently long-tailed for the present problem. Instead, a proposal distribution with long tails and strong narrow mode is required. This can easily be achieved by generating a random sample between two boundaries b_l and b_h for the standard deviation of the proposal by generating a random weight for the weighted geometric mean of these two boundaries. Hence the standard deviations are sampled according to $\sigma = b_h^w b_l^{1-w}$ where $w \sim \beta(a, b)$ is Beta-distributed with parameters a and b. The resulting proposal distribution is symmetric with very long tails and a strong narrow mode. In order to obtain higher standard deviations for the first stage the choice of $w \sim \beta(2, 1)$ (with mean $\mathrm{E}(w) = 2/3$) is adequate while for the second stage $w \sim \beta(1, 2)$ (with mean $\mathrm{E}(w) = 1/3$) samples smaller standard deviations. These final choices are the result of many test runs and result in tails that reach out over the entire prior range of the frequency and still yield proposal modes that are as narrow as the posterior modes that test runs have revealed.

Finally the proposals for parameter block III have to be considered that encompass the two parameters $\{\psi, \phi_0\}$. In the first stage of the delayed rejection method uncorrelated proposals are chosen for the search through the parameter space. When sampling from the actual posterior mode, pilot runs show that ψ and ϕ_0 are negatively correlated by the influence of the polarisation angle ψ on the antenna pattern and the phase ϕ_0 on the phase evolution given in Eq. 2.2. This can be accounted for in the second stage of the delayed rejection method by proposals with adapted correlations.

The implementation of the ideas outlined above leads to reasonably good acceptance rates and hence to a better convergence of the Markov chain. While during the burn-in period it is mainly the stage 1 candidates that are accepted, the Markov chain is driven mainly by stage 2 candidates after the burn-in. But still, the stationary distribution features many distinct modes that carry the risk of trapping the Markov chain.

Different approaches to encounter this problem have been introduced in the 1990s. Many approaches are related to the simulated annealing method [34] which flattens the posterior by raising it to a fractional power. A more advanced technique is the related Metropolis-Coupled MCMC technique introduced by Geyer [31].

2.4.4 The Metropolis-Coupled MCMC

Initially [32], a simulated annealing temperature scheme has been applied by raising the posterior distribution to a power of $\beta = 1/T$ in order to flatten out the posterior modes. The degree of flattening is determined by a temperature T. During a burn-in that usually starts at a higher temperature, say T_0, the temperature T was slowly lowered until the actual posterior distribution was reached, hoping that the sampler had enough time to find the posterior mode of highest evidence and interest.

Although longer temperature schemes procure better success the difficulty is still to find an optimal balance between length of temperature scheme and success probability for the sampler to find the global posterior maximum to sample from. The sampler is prone to get trapped in one of the local maxima of the posterior distribution. Once the scheme has reached the inverse destination temperature $\beta = 1$ that determines the actual posterior distribution without having found the global posterior maximum it is unlikely for the sampler to find it subsequently within a finite time period while abiding at inverse temperature $\beta = 1$.

In a different approach, Simulated Tempering [40], the temperature becomes a dynamic variable on which a random walk is conducted during the entire sampling process. The joint distribution of temperature and remaining parameters, however, requires the normalization constants of the distributions given the temperature. Other approaches like the Tempered Transition method [41] or the Metropolis-Coupled chain (a.k.a. parallel tempering algorithm) [31] do not need normalisation constants. The latter approach has been advocated in the astrophysical literature [42, 43] and was implemented here, as it yielded the best results.

In a Metropolis-Coupled chain [31], sampling is done from k different distributions $p_i(\boldsymbol{a}_i|\boldsymbol{d}), i \in \{1, \ldots, k\}$ in parallel. The real posterior distribution of interest is denoted by $p_1(\boldsymbol{a}_1|\boldsymbol{d})$ with parameter vector $\boldsymbol{a}_1$, whereas the distributions of higher orders $i > 1$ are chosen in such way that the sampling process is facilitated. Usually, different temperature coefficients are applied [44] that flatten out the posterior modes. During the sampling from the

k distributions, from time to time, attempts are made to swap the states of a randomly chosen pair of distributions. The acceptance probability for swapping the parameters of two different posterior distributions $p_i(\boldsymbol{a}_i|\boldsymbol{d})$ and $p_j(\boldsymbol{a}_j|\boldsymbol{d})$, $i \neq j$ is given by [31]

$$\boldsymbol{\alpha}_{i\leftrightarrow j} = \min\left(1, \frac{p_i(\boldsymbol{a}_j|\boldsymbol{d})p_j(\boldsymbol{a}_i|\boldsymbol{d})}{p_i(\boldsymbol{a}_i|\boldsymbol{d})p_j(\boldsymbol{a}_j|\boldsymbol{d})}\right). \tag{2.19}$$

The posterior in the present context can be regarded as a canonical distribution

$$p(\boldsymbol{a}|\boldsymbol{d}) \propto p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a}) = p(\boldsymbol{a})\exp\left[-\beta\chi^2(\boldsymbol{a})\right] \tag{2.20}$$

with inverse temperature $\beta = 0.5$. For lower values of β, the posterior modes are flattened out and the sampling process is eased. A temperature scheme for a Metropolis-Coupled chain uses k different values β_i, $i \in \{1,\ldots,k\}$, while the actual posterior distribution is represented by $p(\boldsymbol{a}_1|\boldsymbol{d})$ with inverse temperature $\beta_1 = 0.5$. The prior distribution is deliberately not involved in the temperature scheme as the prior information at high temperatures is preserved. The transformed posterior in Eq. 2.20 converges to the prior distribution for $\beta \longrightarrow 0$ whereas a temperature scheme that would be applied to the entire posterior distribution would merely yield a uniform distribution.

Using Eq. 2.19, the acceptance probability for swapping the parameters of two distribution $p_i(\boldsymbol{a}_i|\boldsymbol{d})$ and $p_j(\boldsymbol{a}_j|\boldsymbol{d})$, $i \neq j$ is given by [31]

$$\begin{aligned}\boldsymbol{\alpha}_{i\leftrightarrow j}(\beta_i,\beta_j,\boldsymbol{a}_i,\boldsymbol{a}_j) = {} & \min\left(1, \exp\left[(\beta_j-\beta_i)\cdots\right.\right.\\ & \left.\left.\cdots(\chi^2(\boldsymbol{a}_j)-\chi^2(\boldsymbol{a}_i))\right]\right)\end{aligned} \tag{2.21}$$

where the prior distributions cancel out as they are not influenced by the temperature.

The inverse temperatures $\beta_i, i \in 1,\ldots,k$, including their number k, are among the unknown parameters that have to be determined prior to each simulation. It is notable that there are other parameters. The choice of parameters that determine the proposal distributions for the original posterior distribution was addressed in Sec. 2.4.3. The adaptation to the annealed posteriors is accomplished automatically by the use of the delayed rejection method that was explained in Sec. 2.4.1. Simulations from posterior distributions at higher annealing temperatures have more stage 1 acceptance rates than those at cooler temperatures.

The determination of the inverse temperatures requires the assessment of the (unnormalised) posterior distribution $p(\boldsymbol{a}|\boldsymbol{d}) \propto p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})$ and the range of values it can attain. A coarse assessment for the purpose of deriving an appropriate temperature range is described below by considering that the critical part of the posterior distribution that influences the convergence is the likelihood.

There are basically two boundaries of interest for the log-likelihood and therefore $\chi^2(\boldsymbol{a})$ which allow for assessing its range. The expected lower boundary that can be reached arises purely from the complex noise ϵ_k of the signal, hence, from a perfectly fitted model and is estimated by two sums (over real and imaginary part) $\chi^2(\boldsymbol{a}) = \sum_k \left|\frac{\epsilon_k}{\sigma_k}\right|^2 = \sum_k \left(\frac{\epsilon_{k,\mathrm{re}}}{\sigma_k}\right)^2 + \sum_k \left(\frac{\epsilon_{k,\mathrm{im}}}{\sigma_k}\right)^2$. It has the expected value

$$\hat{b}_{\chi^2,\mathrm{fit}} = \mathrm{E}\left(\chi^2(\boldsymbol{a})\right) = 2n. \tag{2.22}$$

An approximation of the 95%-confidence interval of the chi-square level for a perfectly fitted model can be derived using a χ^2-distribution. The Wilson-Hilferty approximation [45] of a 95%-confidence interval with $2n$ degrees of freedom is given by

$$\mathrm{CI}_{\mathrm{fit},95\%} = \left[2n\left(1 - \frac{1}{9n} - 1.96\sqrt{\frac{1}{9n}}\right)^3, 2n\left(1 - \frac{1}{9n} + 1.96\sqrt{\frac{1}{9n}}\right)^3\right] \tag{2.23}$$

as an approximate 95%-confidence range of the lower chi-square level.

The second boundary of interest is approached during the burn-in when the frequency parameters have not yet been found. The only way to maximise the likelihood (and minimise $\chi^2(\boldsymbol{a})$) at this stage is by reducing the value of h_0. This is intuitively clear because when we assume the frequency parameters to be fixed at an incorrect frequency, then the conditional posterior for the amplitude of a signal is zero as there is no signal at the wrong frequency. This is the case even for small aberrations from the true frequency values if the likelihood is based on large data samples. In Sec. 2.5.2 we will see that the assumption of a zero-amplitude signal or equivalently of absence of a signal will help in deriving a limit for a signal's amplitude required for its detection.

The above mentioned phenomenon forces the algorithm to keep h_0 small while $\chi^2(\boldsymbol{a})$ stays on a certain level until the frequency parameters converge to the *true* values, facilitating the chance that the remaining parameters converge. This lower boundary of $\chi^2(\boldsymbol{a})$ for a non-fitted model is obtained by setting the model parameter h_0 to zero and has the form

$$\hat{b}_{\chi^2,\text{not-fit}} = \sum_k \left| \frac{B_k}{\sigma_k} \right|^2 . \tag{2.24}$$

The difference of the two levels derived above is roughly the barrier to cross when stepping from a point in parameter space that maximises the likelihood to a point in parameter space where no signal is postulated ($h_0 = 0$). This, by experience, seems to be the best natural guesswork of the MCMC sampler as long as the frequency parameters are not consistent with the true values. This assessment is essential for obtaining a sensible temperature pool for the parallel tempering scheme. An appropriate highest temperature level is therefore assessed by regarding the acceptance probability of a fictive step from χ^2-level $\hat{b}_{\chi^2,\text{fit}}$ to $\hat{b}_{\chi^2,\text{not-fit}}$ given by

$$\alpha_{\text{fit}\rightarrow\text{not-fit}} = \min\left(1, \exp[-\beta \cdot (\hat{b}_{\chi^2,\text{not-fit}} - \hat{b}_{\chi^2,\text{fit}})]\right) . \tag{2.25}$$

A suitable acceptance probability has to be chosen at a given temperature β that guarantees an occasional acceptance to jump the difference of the two levels. For an acceptance probability that yields an expected acceptance every nth step it is $\alpha_{\text{fit}\rightarrow\text{not-fit}} = 1/n$ which yields an inverse hot temperature of

$$\beta_k = 0.5 \log n / \left(\hat{b}_{\chi^2,\text{not-fit}} - \hat{b}_{\chi^2,\text{fit}} \right). \tag{2.26}$$

A sensible choice for n is 10 000, as it guarantees an occasional escape from an alleged mode during a typical run of hundreds of thousands of iterations for the chain that runs at the highest temperature β_k. The remaining unknown left to be determined is k, the number of chains needed to cover the inverse temperature range from $\beta_1 = 0.5$ to β_k. It depends on the uncertainty in the noise level of the signal which is represented by the confidence interval range $\Delta\text{CI}_{\text{fit},95\%}$ of Eq. 2.23 as it reflects the contribution of the noise to potentially jump a certain difference in the chisquare levels. The ratio between $\hat{b}_{\chi^2,\text{non-fit}} - \hat{b}_{\chi^2,\text{fit}}$ and $\Delta\text{CI}_{\text{fit},95\%}$ therefore affects the number of chains needed to cover the inverse temperature range and determines

$$k \approx \frac{\hat{b}_{\chi^2,\text{non-fit}} - \hat{b}_{\chi^2,\text{fit}}}{\Delta\text{CI}_{\text{fit},95\%}}. \tag{2.27}$$

The inverse temperature scheme to cover the range of $[\beta_1, \beta_k]$ is given by [46] $\beta_i = (1 + \lambda \cdot (i-1))/2$, $i \in \{1, \ldots, k\}$ and $\lambda = (2\beta_k - 1)/(k-1)$. The motivation of using this linearly spaced grid arose from the observations made in pilot runs that the sampler is prone to get stuck at log-likelihood levels that are rather linearly spaced. The inverse temperature is in turn a multiplicative factor to the log-likelihood (Eq. 2.20). The linear spacing can be seen later in Fig. 2.3 of Sec. 2.5.1.

Nonetheless, care must be taken in the presence of weak signals. In this case, the problem emerges that $\hat{b}_{\chi^2,\text{fit}}$ can fall into the confidence interval $\text{CI}_{\text{fit},95\%}$. This indicates an eventual futile retrieval of the parameters from the signal and requires special consideration. A discussion of the issue of weak signals will follow in Sec. 2.5.2.

2.5 Results with simulated signals

2.5.1 Simulation results for strong signals

Synthesised fictitious data has been used and passed through the six parameter MCMC routine. The artificial signals were embedded within white and normally distributed noise. The ability of the MCMC algorithm to successfully find the signal and estimate the six parameters was demonstrated, and is presented below. The artificial signals $s(t)$ were synthesised assuming a source at right ascension $5^{\mathrm{h}}\,35^{\mathrm{m}}\,28.03^{\mathrm{s}}$ and declination $-69^\circ\,16'\,11.79''$, as would be seen by the LIGO-Hanford interferometer. The signals were then added to noise; a signal at 300 Hz has been assumed and a corresponding noise spectral density of at that frequency of $h(f) = 1 \times 10^{-22}\,\mathrm{Hz}^{-1/2}$. The amplitude of the signal used in the test runs was $h_0 = 2.0 \times 10^{-22}$. The two data sets correspond to observation periods $\mathrm{OP}_1 : [2003-04-03\ 04:19:50\mathrm{UTC} - 2003-04-13\ 04:18:50\mathrm{UTC}]$ and $\mathrm{OP}_2 : [2003-04-03\ 04:19:50\mathrm{UTC} - 2003-05-14\ 20:18:50\mathrm{UTC}]$ over 10 days and 41.7 days respectively. At a sample rate of one sample per minute (which was the rate used for the LIGO/GEO S1 analysis described in [24]) this entails 14 400 and 60 000 samples, respectively.

Markov chains have been simulated for example signals with the *true* six parameters: $h_0 = 2.0 \times 10^{-22}$, $\psi = 0.35$, $\phi_0 = 0.22$ (both in radians), $\cos\iota = 0.5$, $\Delta f = 2.0 \times 10^{-3}\,\mathrm{Hz}$ and $\Delta\dot{f} = -1.2 \times 10^{-10}\,\mathrm{Hz\,s}^{-1}$. Following Eq. 2.27, the data set over observation period OP_1 required three Metropolis-Coupled Markov chains whereas for observation period OP_2, six chains were needed. For a better understanding, the log-posterior values of the (unnormalised) posterior distribution are drawn in Fig. 2.3 for all Metropolis-Coupled chains. This should demonstrate the ability of the parallel chains to improve mixing. The chains are coloured by their intrinsic temperature. The colour scheme goes from red to blue. The colder a chain, the more close it is to the real posterior. Black indicates the true posterior distribution at inverse temperature $\beta = 0.5$. This underpins the idea of the Metropolis-Coupled MCMC as by means of the hotter chains the entire ensemble can

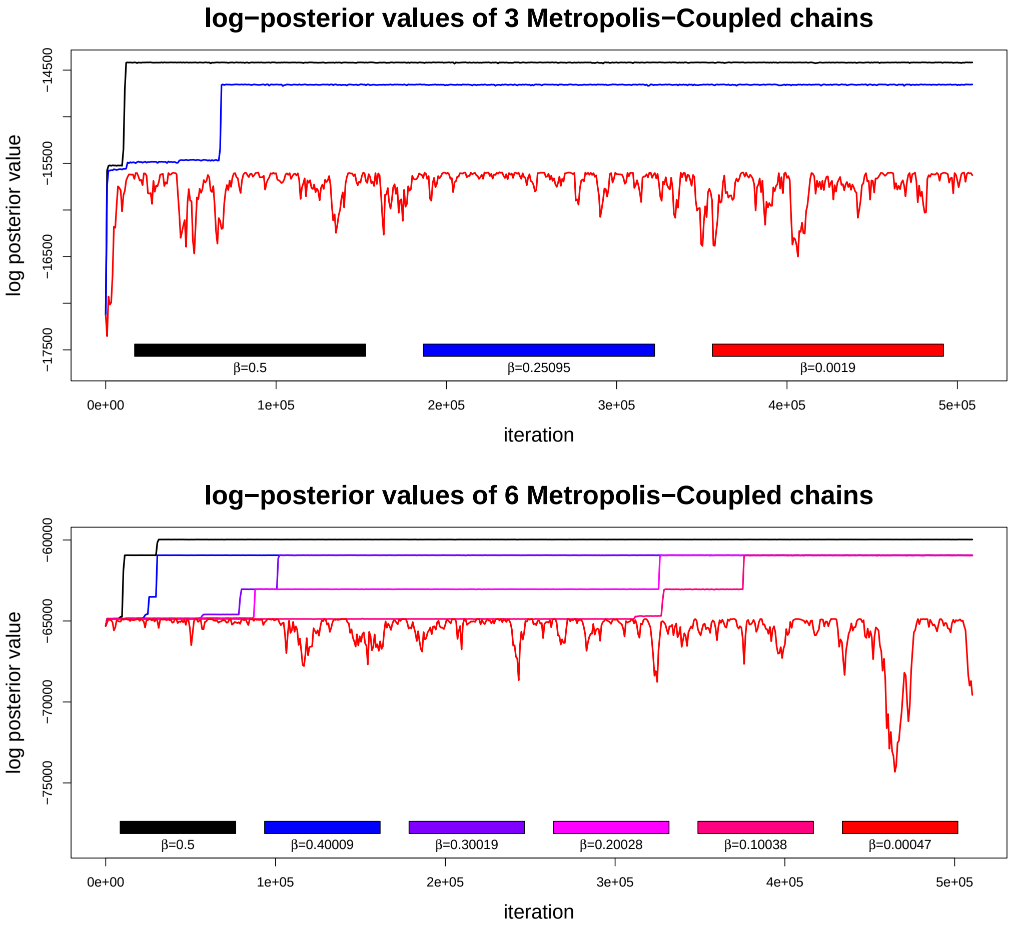


Figure 2.3: Log-posterior values for all Metropolis-Coupled chains. The upper and lower plot show simulations over observation periods OP_1 and OP_2 respectively. These are the true log posterior values and not the tempered values in order to maintain comparability.

hardly get trapped in a local maxima when the temperature scheme is chosen adequately. The different lower boundaries for the fitted model and the non-fitted model that were point of discussion in Sec. 2.4.4 are reflected in the hottest and coldest chain.

Eq. 2.27 specifies a vague lower limit for the number of chains that are required in order to prevent the sampler from getting trapped in local posterior modes. The interactions between the chains are rather small but the hotter chains fulfil the purpose of passing over information of other parts of the parameter space to the colder chains during the burn-in. Please note that some interactions can not be seen in Fig. 2.3 due to thinning out that will be discussed below.

An unfortunate feature of the Metropolis Coupled MCMC is that only the samples of the actual posterior distribution at inverse temperature $\beta = 0.5$ can be utilised. A larger number of chains improves the interactions but produces more wasted samples after the Markov chain of the true posterior distribution has converged. One way out of this could be to sample more often from the actual posterior distribution but this in turn discriminates the hotter chains that are important during the burn-in.

Fig. 2.4 and Fig. 2.5 display the MCMC generated posterior probability distribution functions (pdfs) for those example signals. The first 33 000 iterations of the chain were discarded as burn-in. Short-term correlations in the chain were eliminated by 'thinning' the remaining terms; every 100^{th} item was kept in the chain. This choice is rather arbitrary but is based on the fact that the sampler yields acceptance rates of around 20%. Therefore, there are, on average, 5 acceptances within 100 iterations. The autocorrelation of lag 100, after the burn-in, was about $1/4$ in all runs. The fact that the sampler only outputs every 100 iteration helps reducing the memory requirement of simulations over hundreds of thousands of iterations with multiple chains.

The bandwidth of the Gaussian kernel density estimator was chosen according to Silverman [47] as 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power.

The following tables Tab. 2.1 and Tab. 2.2 show the median values, marginal 95% posterior probability intervals and MCMC standard errors for the two data sets. The MCMC standard error gives a measure of how much the sample mean, as a point estimate of the true posterior mean, changes over repeated MCMC simulations. This precision depends on the number of iterations and the degree of autocorrelation within the sample. Here, Geweke's [48] method has been applied, often referred to as "time-series standard error" which is based on estimating the spectral density. The

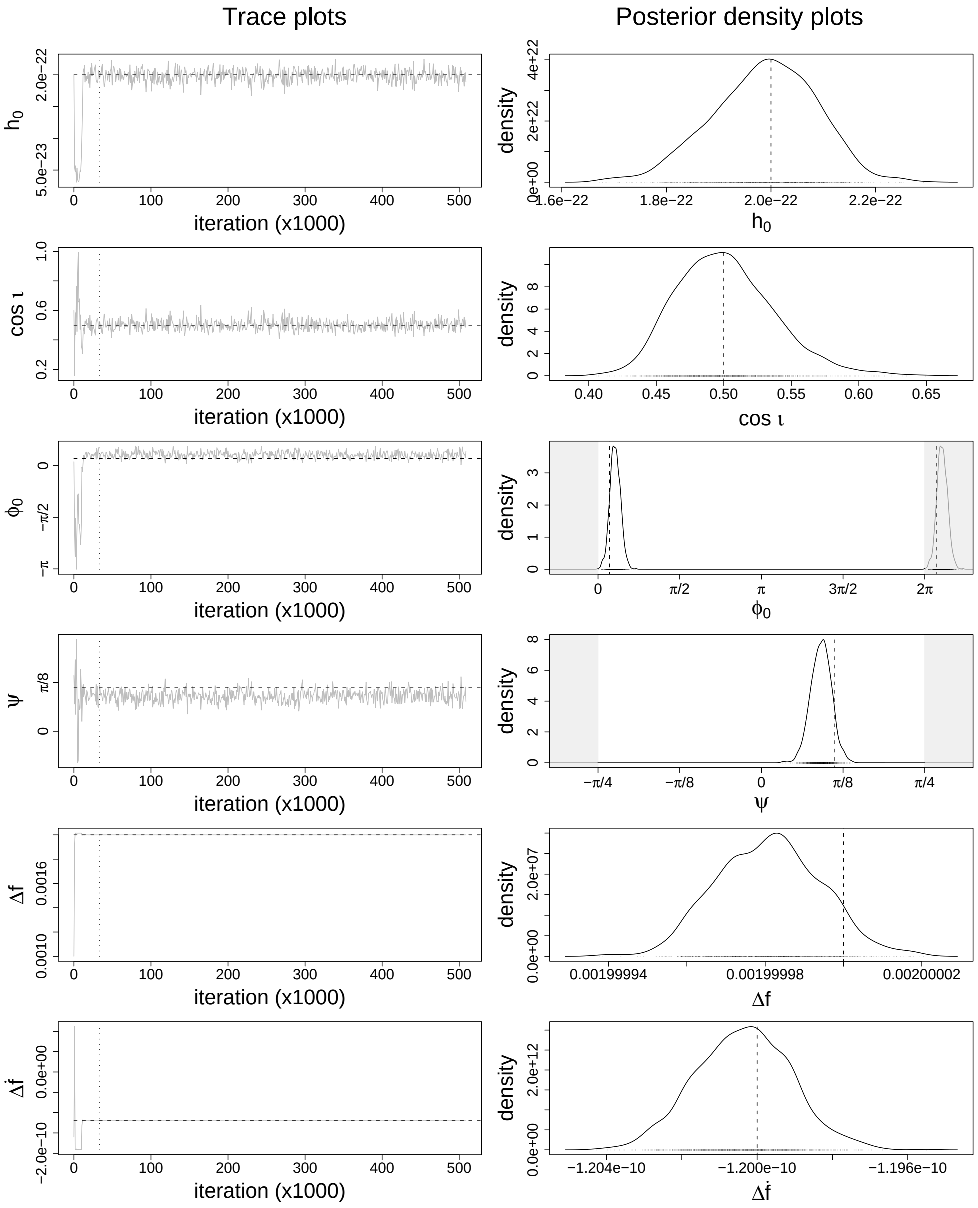


Figure 2.4: MCMC estimates of the posterior pdf (kernel density) for the six parameters h_0, ψ, ϕ_0, $\cos\iota$, Δf and $\Delta\dot{f}$ for the data set with 14 400 samples. This synthesised signal had *true* parameters of: $h_0 = 2.0 \times 10^{-22}$, $\psi = 0.35$, $\phi_0 = 0.22$ (both in radians), $\cos\iota = 0.5$, $\Delta f = 2.0 \times 10^{-3}\,\mathrm{Hz}$ and $\Delta\dot{f} = -1.2 \times 10^{-10}\,\mathrm{Hz\,s^{-1}}$. The dashed lines show the *true* parameter values and the dotted lines in the trace plots indicate the end of the burn-in.

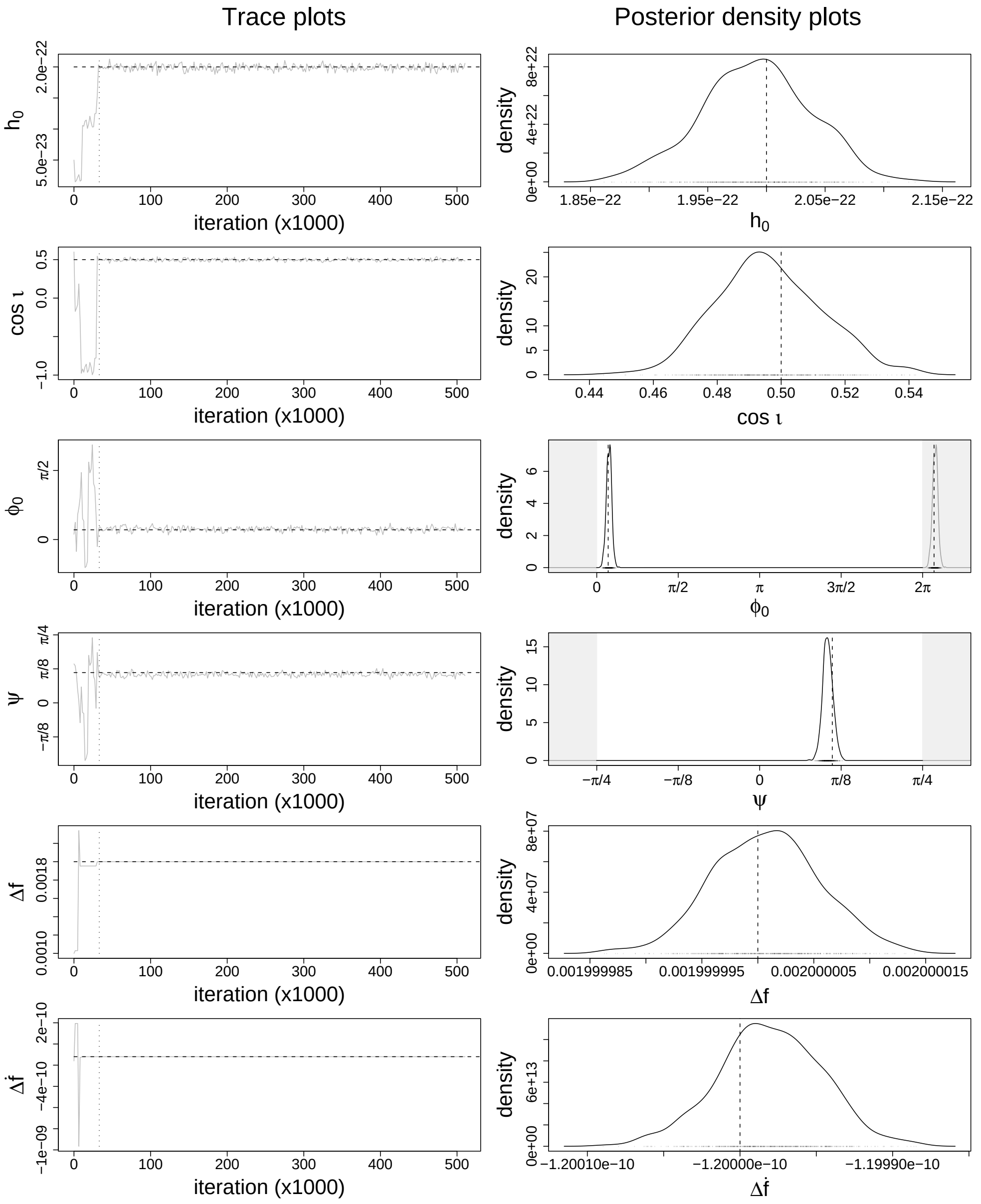


Figure 2.5: MCMC estimates of the posterior pdf (kernel density) for the six parameters h_0, ψ, ϕ_0, $\cos\iota$, Δf and $\Delta\dot{f}$ for the data set with 60 000 samples. This synthesised signal had *true* parameters of: $h_0 = 2.0 \times 10^{-22}$, $\psi = 0.35$, $\phi_0 = 0.22$ (both in radians), $\cos\iota = 0.5$, $\Delta f = 2.0 \times 10^{-3}\,\mathrm{Hz}$ and $\Delta\dot{f} = -1.2 \times 10^{-10}\,\mathrm{Hz\,s^{-1}}$. The dashed lines show the *true* parameter values and the dotted lines in the trace plots indicate the end of the burn-in.

interested reader should note the paper by Geyer [49] who develops improved window-estimates for the MCMC standard error by calculating the ‘optimal’ bandwidth using specific properties of the autocovariances of a Markov chain.

Table 2.1: Median values obtained by MCMC, 95% posterior probability intervals (p.c.i.) and MCMC standard errors for the data set of length 14 400 bins.

Parameter	True value	Posterior median	95% posterior probability intervals	MCMC standard error
h_0	2×10^{-22}	1.990×10^{-22}	$[1.784 \times 10^{-22}, 2.171 \times 10^{-22}]$	2.3355×10^{-25}
$\cos\iota$	0.5	0.499	$[0.440, 0.582]$	8.53×10^{-4}
ϕ_0	0.22	0.329	$[0.130, 0.535]$	2.7×10^{-3}
ψ	0.35	0.288	$[0.190, 0.383]$	1.04×10^{-3}
Δf	0.002	1.9999814×10^{-3}	$[1.9999556 \times 10^{-3}, 2.0000074 \times 10^{-3}]$	2.57×10^{-10}
$\Delta\dot{f}$	-1.2×10^{-10}	$-1.2003623 \times 10^{-10}$	$[-1.2028017 \times 10^{-10}, -1.1979087 \times 10^{-10}]$	5.4394×10^{-15}

Table 2.2: Median values obtained by MCMC, 95% posterior probability intervals and MCMC standard errors for the data set of length 60 000 bins.

Parameter	True value	Posterior median	95% posterior probability intervals	MCMC standard error
h_0	2×10^{-22}	1.993×10^{-22}	$[1.897 \times 10^{-22}, 2.080 \times 10^{-22}]$	1.1504×10^{-25}
$\cos\iota$	0.5	0.495	$[0.465, 0.530]$	4.51×10^{-4}
ϕ_0	0.22	0.240	$[0.138, 0.344]$	1.44×10^{-3}
ψ	0.35	0.327	$[0.278, 0.374]$	0.554×10^{-3}
Δf	0.002	2.0000008×10^{-3}	$[1.9999909 \times 10^{-3}, 2.0000104 \times 10^{-3}]$	2.6774×10^{-10}
$\Delta\dot{f}$	-1.2×10^{-10}	$-1.199981 \times 10^{-10}$	$[-1.200046 \times 10^{-10}, -1.1999129 \times 10^{-10}]$	1.8096×10^{-16}

As mentioned in Sec. 2.4.3 correlations occur between the parameters of block III : $\{\psi, \phi_0\}$. Two dimensional plots have been compiled to gain insight into the correlation of the parameter pairs $\{h_0, \cos\iota\}$ and $\{\psi, \phi_0\}$. This is shown in Fig. 2.6. The negative correlation $\rho_{\psi,\phi_0} = -0.69$ is revealed between the polarisation angle ψ and the phase ϕ_0 as well as the expectedly negative

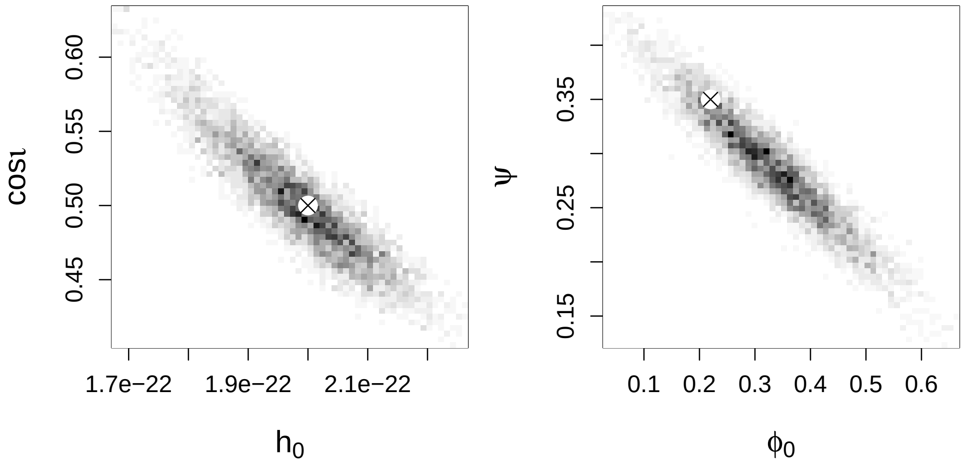


Figure 2.6: Joint 2-D posterior plots for two different parameter pairs $\{h_0, \cos\iota\}$ and $\{\phi_0, \psi\}$ of the simulation over observation period OP_1. The crosses indicate the *true* parameter values.

correlation $\rho_{h_0,\cos\iota} = -0.92$ between h_0 and $\cos\iota$. The conjunct influence of the two frequency parameters Δf and $\Delta\dot{f}$ on the phase evolution does not lead to a correlation of the samples of the target posterior mode. The many weaker posterior modes in its vicinity that attract to the target posterior mode are however better accounted for when considering the physical coherence that lead to the re-parametrisation in Sec. 2.4.2.

The two runs in tables Tab. 2.1 and Tab. 2.2 and also in Fig. 2.4 and Fig. 2.5 only differ in the length of the data set that was given. They both reveal that the frequency Δf and especially its first derivative $\Delta\dot{f}$ have to be found very accurately before convergence is achieved. However, the most striking feature between those two results is the accuracy improvement between runs with 14 400 and 60 000 bins (4.167× the data) based on the same parameters. The following table Tab. 2.3 reveals the ratios between the individual ranges of the 95% posterior probability intervals of the two runs.

While the accuracy of the estimates of the first four parameters improve by roughly 2.0 when involving 4.167 times the data, the accuracy of the estimate for parameters Δf and $\Delta\dot{f}$ yielded around 2.7 and 37 times better results, respectively. The reason for this is found in the fact that $\Delta\dot{f}$ is linked

Table 2.3: Ratios for all six parameters of the 95% posterior probability interval ranges between Tab. 2.1 and Tab. 2.2. Both runs are based on the same signals but the second (Tab. 2.2) involves 4.167× the data.

Parameter	Ratio of 95% posterior probability interval ranges
h_0	2.12
$\cos\iota$	2.17
ϕ_0	1.96
ψ	2.02
Δf	2.66
$\Delta\dot{f}$	36.77

to the phase evolution Eq. 2.2 by the squared time and any minor aberration in the frequency derivative is highly penalised by the likelihood.

The simulations made above have shown that for the constant noise level used here ($\sigma = 1.0 \times 10^{-22}$) the successful detection of signals with an amplitude $h_0 \geq 2.0 \times 10^{-22}$ and 10 days of data are straightforward. The next subsection therefore deals with the issue of weak signal detection with respect to the noise level and the amount of data collected. Especially the behaviour of the MCMC sampler in cases of a small signal-to-noise ratio will be examined and the question discussed for what signal-to-noise ratios a detection is still possible.

2.5.2 Simulation results for weak signals

MCMC sampling in presence of a low signal-to-noise ratio

In the work by Christensen et al. [10] with just four parameters (h_0, ψ, ϕ_0, and $\cos\iota$), the confident detection of signals with $h_0/\sigma \gtrapprox 0.1$ could be achieved with 14 400 bins of data. A look at the posterior distribution of h_0 reveals whether a signal is likely to be present in the data or not. In particular the comprehension of the new frequency parameters complicates the search due to the corresponding increase in the size of the parameter space. In the previous work [32] of this research a simple simulated annealing schedule was used and the detection of signals $h_0/\sigma \lessapprox 0.5$ with 14 400 data points was not possible. Detecting a signal by the use of an MCMC sampler implies discovering the posterior mode of the frequency parameters. No inference on any parameter can be drawn unless the chain is consistent enough with the true frequency parameters, and only then allowing the convergence of the remaining parameters.

Herein lies the problem as the simulations above reveal that just a slight inconsistency from the true frequency parameters yields a posterior probability that suggests no signal at all. In Bayesian terms, the range of values for Δf and especially $\Delta \dot{f}$ that are consistent with the posterior probability is extremely narrow and in order to draw conclusions from the output of the MCMC sampler it needs to identify that narrow range until we can speak of *convergence*. If the signal is strong, the MCMC sampler is forced to stay in a particular posterior mode due to its overwhelming evidence.

Things change when the evidence for a signal in the data fades. With diminishing signal-to-noise ratio in the data, the influence of the likelihood in the posterior fades and the prior gains more importance. However, the range of values for the frequency parameters that are consistent with the posterior probability ascribed to the likelihood are still narrow but with its vanishing probability mass the prior probability gains dominance. This increasing dominance of the prior binds the MCMC sampler to sample from a more spread out posterior distribution and not just from a narrow mode influenced by the likelihood that contains the actual information about a potential signal in the data. Rather, the MCMC sampler is bound to sample from posterior modes that are influenced by a faint likelihood as well as other parts of the posterior influenced by the prior.

This is a problem as it usually requires thousands of iterations to target the narrow range of frequency values Δf and $\Delta \dot{f}$ that are consistent with the evidence of a signal. It is clear that a straightforward detection by a single burn-in manner is impeded when the acceptance probability of leaving the posterior mode ascribed to the evidence of a signal does not allow the sampler to dedicate a viable sampling period. In addition, the MCMC sampler can not exploit the situation to find the strongest posterior mode influenced by the likelihood by approaching it in a stepwise manner over the weaker posterior modes in its vicinity. Those provide a rough but yet somewhat helpful attraction area towards the strongest posterior mode and are even more faint in the presence of a weak signal.

The idea of multiple Metropolis-Coupled chains provides a remedy here. As already stated in Sec. 2.4.4, an inverse temperature scheme is chosen to cover the range of $[\beta_1, \beta_k]$ determined by Eq. 2.26 and Eq. 2.27. In the presence of weak signals, however, $\hat{b}_{\chi^2,\text{not-fit}}$ given by Eq. 2.24 can fall into the confidence interval $\text{CI}_{\text{fit},95\%}$. If this happens it serves as a reliable indicator for a weak signal requiring an exceptional tempering scheme to account for the issues mentioned above.

Eq. 2.27 would normally suggest to just run a single chain due to the strong noise level, intrinsically preventing the Markov chain from being trapped. But in this case, a larger number of chains would indeed make more sense as multiple parallel chains sustain a memory about the posterior with its numerous modes. This helps to improve the mixing by helping to repeatedly retrieve those posterior modes which provide insufficient evidence to be sampled from continuously. Since this situation provides no helpful information that would allow to assess the number of chains required, a fixed number of chains must be chosen. A number $k = 10$ has proven to be quite efficient but can be increased if enough computing power is available. Following Eq. 2.26, the inverse temperature β_k is determined by $\Delta\text{CI}_{\text{fit},95\%}$ and has the form

$$\beta_k = 0.5 \log n / \Delta\text{CI}_{\text{fit},95\%}. \tag{2.28}$$

Simulation results of signals with weak signal-to-noise ratio

A run was conducted on a data set containing the synthesised signal of a pulsar as would be observed during observation period OP_1 at the LIGO interferometer, with all true parameters (except h_0) having the same values as in 2.1. The data set was created assuming a pulsar with amplitude $h_0 = 0.3 \times 10^{-22}$ embedded within noise $\sigma = 10^{-22}$ which yields a signal-to-noise ratio of $h_0/\sigma = 0.3$ which, from experience, exhibits a real challenge for the

MCMC sampler.

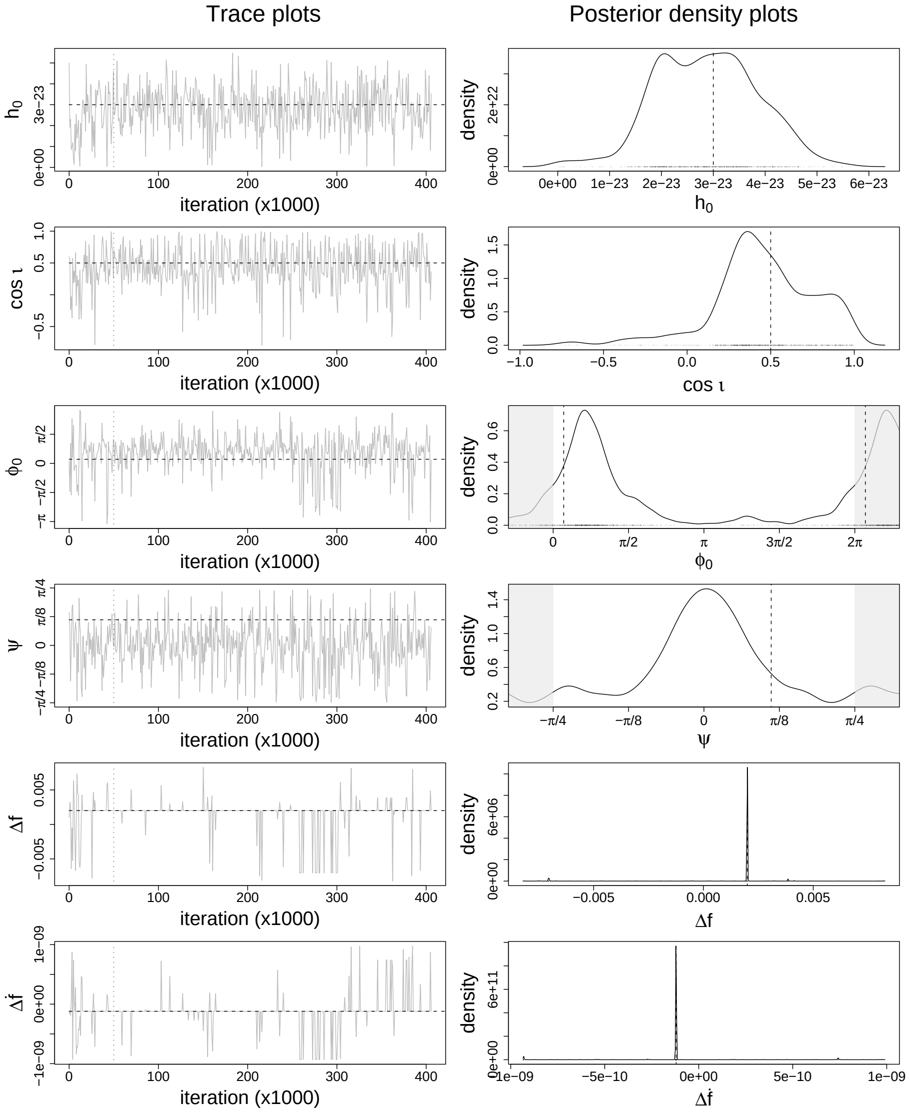


Figure 2.7: MCMC estimates of the posterior pdf (kernel density) for the six parameters h_0, ψ, ϕ_0, $\cos\iota$, Δf and $\Delta\dot{f}$ of a weak signal for a data set with 14 400 samples. This synthesised signal had *true* parameters of: $h_0 = 0.30 \times 10^{-22}$, $\psi = 0.35$, $\phi_0 = 0.22$ (both in radians), $\cos\iota = 0.5$, $\Delta f = 2.0 \times 10^{-3}\,\text{Hz}$ and $\Delta\dot{f} = -1.2 \times 10^{-10}\,\text{Hz}\,\text{s}^{-1}$. The dashed lines show the *true* parameter values and the dotted lines in the trace plots indicate the end of the burn-in.

Table 2.4: MCMC yielded median values, 95% posterior probability intervals and MCMC standard errors for the data set of length 14 400 bins.

Parameter	True value	Posterior median	95% posterior probability intervals	MCMC standard error
h_0	0.3×10^{-22}	0.289×10^{-22}	$[0.081 \times 10^{-22}$, $0.476 \times 10^{-22}]$	2.3305×10^{-25}
$\cos \iota$	0.5	0.431	$[-0.404, 0.961]$	77.9×10^{-4}
ϕ_0	0.22	0.687	$[-2.356, 2.229]$	41.8×10^{-3}
ψ	0.35	0.006	$[-0.731, 0.667]$	14.4×10^{-3}
Δf	0.002	1.9998411×10^{-3}	$[-7.0408822 \times 10^{-3}$, $3.8576193 \times 10^{-3}]$	1.122100×10^{-4}
$\Delta \dot{f}$	-1.2×10^{-10}	$-1.2026692 \times 10^{-10}$	$[-9.3092961 \times 10^{-10}$, $7.4088244 \times 10^{-10}]$	1.24780×10^{-11}

The six parameters h_0, $\cos \iota$, ψ, ϕ_0, Δf, and $\Delta \dot{f}$ are displayed in Fig. 2.7 and Tab. 2.4. It is demonstrated that the new MCMC sampler obtains sensible estimates and that it succeeded in retrieving the frequency parameters. However, it becomes apparent that it is forced to sample occasionally from other areas of the prior range due to the weak attraction of the posterior mode influenced by the likelihood, i.e. data.

For this simulation 10 Metropolis-Coupled Markov chains were maintained which are shown in Fig. 2.8. On the lines of Fig. 2.3, colour gradations between red and blue indicate the different temperatures of the Markov chains whereas black is the choice for the Markov chain that simulates from the actual posterior distribution. In Fig. 2.8 the brisk interactions between the 10 chains become clear due to the fact that the target posterior mode is indistinct. The two distinct boundaries for fitted and non-fitted model as discussed in Sec.2.4.4 are still apparent throughout this simulation but are very close due to the low evidence of a signal in the data.

For an injected signal of $h_0/\sigma = 0.25$, with the otherwise same true parameter values as used in the examples above, the parameter values could not be retrieved within 750 000 iterations. Fig. 2.9 shows the very low evidence over the course of the log-posterior values for all 10 Metropolis-Coupled chains. The difference between the two levels for a fitted and non-fitted model

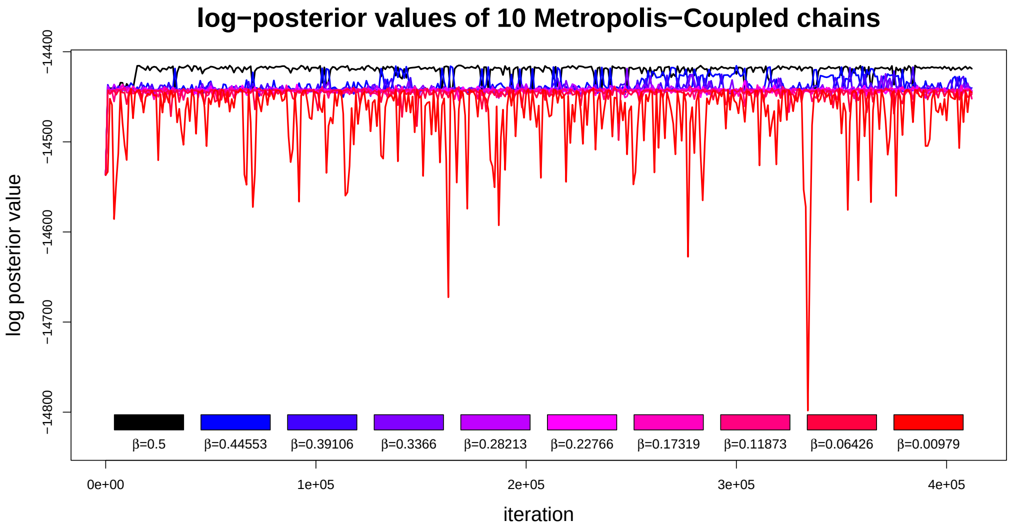


Figure 2.8: Log-posterior values for all Metropolis-Coupled chain for the simulation results displayed in Fig. 2.7. These are the true log posterior values and not the tempered values in order to maintain comparability.

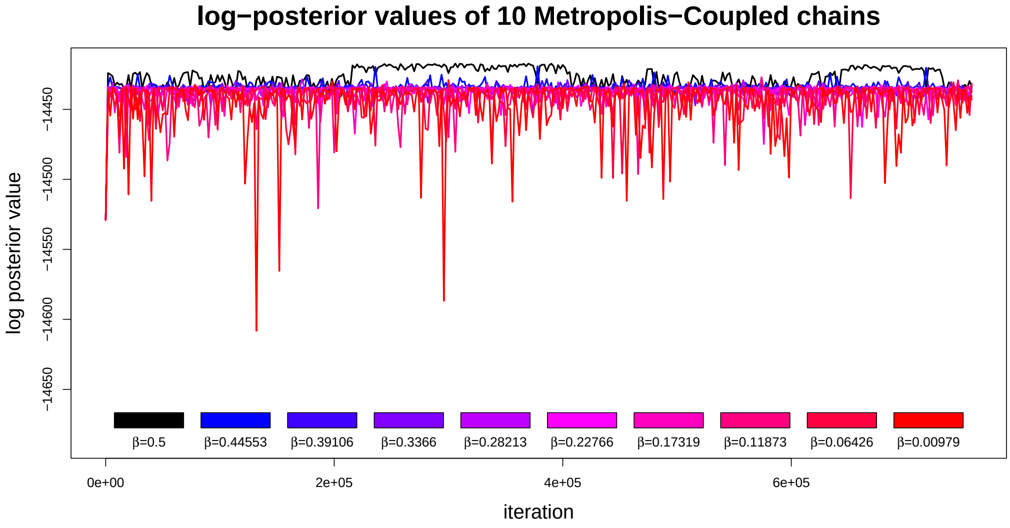


Figure 2.9: Log-posterior values for all Metropolis-Coupled chain for the simulation results displayed in Fig. 2.10. These are the true log posterior values and not the tempered values in order to maintain comparability.

is too small to be maintained throughout the course of the Markov chain. The obvious question is how can the correctness of estimates be assessed when a signal in that data has such a faint evidence and how can we assess from the posterior MCMC output whether a signal is faint.

In Fig. 2.10 trace plots are compiled for the interesting four of the six parameters h_0, $\cos\iota$, Δf, and $\Delta\dot{f}$. It becomes evident that no convergence could be achieved within 750 000 iterations. This demonstrates the difficulty of estimating parameters of signals with very low signal-to-noise ratios. The Markov chain jumps between modes and, in the last part, even comes very close to the true frequency parameter but narrowly misses its first derivative. The dilemma is that if the chain had stopped at the 400 000th iteration and

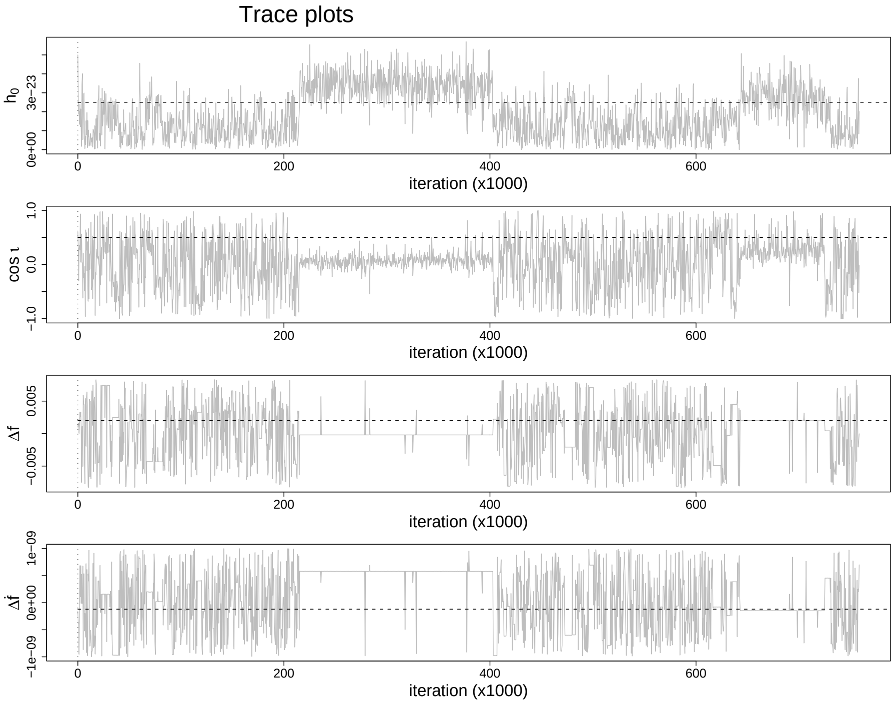


Figure 2.10: Trace plots for the four parameters of a pulsar which are of main interest (h_0, $\cos\iota$, Δf and $\Delta\dot{f}$). This simulation is based on the data set with 14 400 samples (OP$_1$). This synthesised signal had *true* parameters of: $h_0 = 0.25 \times 10^{-22}$, $\psi = 0.35$, $\phi_0 = 0.22$ (both in radians), $\cos\iota = 0.5$, $\Delta f = 2.0 \times 10^{-3}\,\mathrm{Hz}$ and $\Delta\dot{f} = -1.2 \times 10^{-10}\,\mathrm{Hz\,s^{-1}}$. The dashed lines show the *true* parameter values.

if a burn-in of 220 000 had been assumed, the chain could have passed as to have converged to an alleged signal. The problem that arises in the presence of weak signals is that the Markov chain has to be run for a much longer time in order to be sure whether an alleged signal is an actual signal.

2.6 The detection of weak signals

2.6.1 The problem with assessing the signal detection from MCMC outputs

The approach for a signal detection involves the idea of model comparison. All the calculations made above are based on a model that postulates a pulsar signal within the data with its six parameters, expressed in Eq. 2.5. In the following it is referred to as $\mathcal{M}_1$. However, we also consider model $\mathcal{M}_0$, in which we postulate $y(t_k; \boldsymbol{a}) = 0$ in Eq. 2.5. This is equivalent to postulating mere noise within the data and the model comprises no parameters at all, leading to a uniform posterior distribution consisting of a constant likelihood

$$p(\boldsymbol{d}|\boldsymbol{a}) \propto \exp\left[\frac{-\chi^2(\boldsymbol{a})}{2}\right] \text{ with } \chi^2(\boldsymbol{a}) = \sum_k \left|\frac{B_k}{\sigma_k}\right|^2. \tag{2.29}$$

only depending on the observed data, as $\boldsymbol{a}$ is just an empty set within the scope of $\mathcal{M}_0$.

Conditioned on a data set $\boldsymbol{d}$, the Bayesian approach allows for comparing the two models $\mathcal{M}_0$ and $\mathcal{M}_1$ by estimating $p(\mathcal{M}_1|\boldsymbol{d})$, the posterior probability of model $\mathcal{M}_1$, under consideration of $\mathcal{M}_0$ and $\mathcal{M}_1$. Following Bayes theorem, this probability is given by [50]

$$p(\mathcal{M}_1|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\mathcal{M}_1)p(\mathcal{M}_1)}{p(\boldsymbol{d}|\mathcal{M}_0)p(\mathcal{M}_0) + p(\boldsymbol{d}|\mathcal{M}_1)p(\mathcal{M}_1)} \tag{2.30}$$

with $p(\mathcal{M}_0)$ and $p(\mathcal{M}_1)$ being the prior probabilities for each model. The prior probabilities for the different models will be discussed later and for now $\mathcal{M}_0$ and $\mathcal{M}_1$ are assumed to be equally likely $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 0.5$. This leads to

$$p(\mathcal{M}_1|\boldsymbol{d}) = \frac{1}{1 + p(\boldsymbol{d}|\mathcal{M}_0)/p(\boldsymbol{d}|\mathcal{M}_1)} = \frac{1}{1 + B_{0,1}} \tag{2.31}$$

where $B_{0,1} = p(\boldsymbol{d}|\mathcal{M}_0)/p(\boldsymbol{d}|\mathcal{M}_1)$ is the so called Bayes factor [50].

The computation of Bayes factors from the output of an MCMC sampler requires the estimation of the marginal likelihood. Unfortunately, it is obtained by integrating the likelihood with respect to the prior distribution of the parameters. Therefore the MCMC output of the posterior distribution can not serve directly to estimate the marginal likelihood. Several suggestions have been made to estimate the marginal likelihood from the posterior MCMC output one of which is the harmonic mean estimator of the likelihood values obtained from the MCMC output [50] and that is based on the importance sampling approximation [51]. This, however, is known to be unstable. Another possibility for estimating marginal likelihoods from Gibbs sampling output has been developed by Chib et. al. [52] and extended to output from the MH algorithm [53] where a candidate point in a region of high posterior density, which is usually the posterior mean or median, is substituted into the so called basic marginal likelihood identity.

However, we will consider the newly developed deviance information criterion (DIC) [54] that can also be derived directly from the MCMC output. By deploying the deviance $D(\boldsymbol{a}_\mathcal{M}) = -2\log(p(\boldsymbol{d}|\boldsymbol{a}_\mathcal{M})) + 2\log(p(\boldsymbol{d}))$, the expected goodness of fit of the model $\mathcal{M}$ to the data $\boldsymbol{d}$ is given by $\bar{D}_\mathcal{M} = \mathrm{E}[D(\boldsymbol{a}_\mathcal{M})]$. The penalty term $p_\mathcal{M} = \bar{D}_\mathcal{M} - D(\mathrm{E}[\boldsymbol{a}_\mathcal{M}])$ encompasses the complexity of the model $\mathcal{M}$ that accounts for the principle of Occam's Razor and indicates the effective number of parameters by the difference between the average of the likelihood values and the likelihood value of the averaged parameter values. The DIC criterion is given by [54]

$$\mathrm{DIC}_\mathcal{M} = \bar{D}_\mathcal{M} + p_\mathcal{M}. \tag{2.32}$$

However, the long tails of the posterior distribution within this scope of low signal evidence suggest the median as an estimate for the average rather than the posterior mean.

Nonetheless, all methods that involve the posterior MCMC output for the assessment of a signal present within the data require the Markov chain to have converged. Herein lies a truculent problem as we are facing two scenarios when considering the posterior MCMC output for $\mathcal{M}_1$.

1. The posterior estimate of the MCMC output reveals a distinct peak for the frequency parameters over a longer course of the Markov chain, indicating a tangible signal within the data. Under these circumstances a sensible DIC value can be obtained for model $\mathcal{M}_1$. The results will always favor $\mathcal{M}_1$ as the evidence is high because otherwise no distinct posterior peak would be present. This can be true even if the sampler found a local maximum of the posterior mode that is not consistent with the true values. If we had stopped the Markov chain at the 400 000th iteration, discarding the first 220 000 iterations as a burn-in, then the DIC values had revealed the existence of a signal although the estimations are in fact inconsistent. In this example, for model $\mathcal{M}_1$ we obtain $\text{DIC}_{\mathcal{M}_1} = 28870.28$ (comprising the penalty value $p_{\mathcal{M}_1} = 3.63$), model $\mathcal{M}_0$ yields $\text{DIC}_{\mathcal{M}_0} = 28898.14$ which reveals a higher evidence for model $\mathcal{M}_1$. This is deceiving as we can see in the further progress of the Markov chain that convergence obviously has not been attained. This is merely a question of convergence and the DIC does not help to tackle that problem because of the second scenario given below.

2. The posterior estimate of the MCMC output reveals that the frequency parameters Δf and $\Delta \dot{f}$ are spread over the whole prior with many distinct modes. The calculation of a candidate for the basic marginal likelihood identity is futile and alternatively the penalty term of the DIC value of model $\mathcal{M}_1$ attains negative (!) values indicating that the chain has not converged and its derivation is not legitimate. The model comparison based on the posterior MCMC output fails and the only conclusions we can draw is that either the burn-in is still too short or that there is really no signal present in the data.

It is clear that this is a matter of convergence assessment. The diagnostic tools for convergence from Geweke [48] based on a single chain or from Gelman et. al. involving multiple runs [55] also provide no satisfactory remedy. With fading evidence for a signal, during the burn-in, the sampler merely samples from the prior pretending that the chain has converged at first glance. This, however, does not answer the question whether we would detect a signal if the chain would run for longer.

In Fig. 2.7 after several hundreds of thousands of iterations it becomes clear that there might be a hint of a signal because the Markov chain remains mostly within the same mode over a longer course. Convergence assessment can then provide information about the authenticity of the estimated parameters when an alleged posterior mode has been found but it can not negate the question whether there is a faint signal if we would simulate a longer Markov chain. In the example shown in Fig. 2.7, no convergence is achieved before 750 000 iterations and therefore no sensible parameter estimation can be made based on this output. However, if the chain would run for longer, we might be able to detect the signal. All we know is, the longer we run the chain the more we know. For a signal with $h_0/\sigma = 0.25$ the burn-in increases to an amount that leads to the conclusion that the true parameter values are actually not retrievable. Nonetheless, the question whether there is a faint signal within the data is not answered satisfactorily.

The purpose of above comments is to point out that Bayes factors or information criteria like the DIC that require MCMC outputs can only be applied when the sampler efficiently samples from the posterior distribution and converges properly. The dilemma is that MCMC sampling only works efficiently when the data set contains a sufficiently strong signal, in which case a model comparison is senseless. On the other hand, when the computation of a Bayes factor or DIC does make sense, precisely because we are in doubt about the existence of a signal within the data, then this is when their application is jeopardised, because the sampler is inefficiently sampling from the posterior distribution. In this matter it is unprofitable to argue about the advantages and disadvantages of Bayes factors or the use of the DIC which is, for example, controversial in mixture models [56].

A theoretical assessment of the required burn-in length is impossible. It depends on the sampler's performance as well as the data set. This problem needs to be approached by a theoretical assessment of the evidence for model $\mathcal{M}_1$, given an arbitrary data set. The following derivations therefore do not depend on any MCMC output, which renders the use of Bayes factors and DIC impossible.

2.6.2 Derivation of a theoretical detection probability

One possibility to compare models involves the calculation of the AICs (Akaike Information Criterion) [57]. The AIC is defined [57] as

$$\text{AIC} = -2\log(\text{maximum likelihood}) + 2d \tag{2.33}$$

where the penalty term $2d$ brings in the number $d = 6$ of parameters in the model. The penalty term penalises the number of parameters in a model in order to concede preference towards simpler models and to meet the principle of Occam's Razor. The AIC as an estimator of the relative expected value of Kullback-Leibler information loss, however, does not consider the uncertainty about parameter values [50, 58] and overestimates the required number of parameters [50, 59, 60] in which case it would overestimate the probability for a model that considers a signal present within the data. We also encounter large data sets and in the case where the prior information is small compared to the information provided by the data, the Schwarz criterion, also called Bayesian Information Criterion (BIC), yields a better approximation [50] and has the form

$$\text{BIC} = -2\log(\text{maximum likelihood}) + d\log n, \tag{2.34}$$

where the penalty term $d\log n$ comprises the number of $d = 6$ independent parameters that describe a model but also the number n of data samples. All derivations and results below are therefore based on the BIC, although it has the reputation of privileging smaller models, i.e. the model that postulates no signal. All the following calculations can also be applied to the AIC, the only difference being the penalty term. Therefore, results based on the AIC will also be given, providing a better comparison and insight into the problem.

The convenience of the AIC and BIC to merely require maximum likelihood estimates (MLEs) and easy to derive penalty terms make them eligible for the technique described within this section. The concern is to derive a theoretical limit for the detection of a signal within a data set observed during a determined observation period at a certain noise level. The observation period is a vector $\mathrm{OP} = (t_1, \ldots, t_n)'$ of n time points t_k with $k \in \{1, \ldots, n\}$ during which the data has been collected starting from t_{start} and ending at t_{end}. The noise vector is a vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)'$ for the n data bins. Given the *true* parameter vector of the pulsar from which the signal arises, the full information needed for a detection is determined by the vector $\boldsymbol{a}_* = (h_0^*, \cos\iota^*, \psi^*, \alpha^*, \delta^*, \Delta f^*, \Delta\dot{f}^*, \boldsymbol{\sigma}, \mathrm{OP})'$. Although some parameters like sky location are expected to be known, they are essential factors for the detection probability in connection with the observation period and the noise. These are essential parts of the parameter vector as the detection depends significantly on them.

A signal *detection* depends on the actual evidence of the model that postulates a signal from a pulsar within an arbitrary data set, under consideration of two models, postulating either mere noise or a signal with noise. By marginalising over all parameters except the amplitude h_0^*, which determines the signal within the data on which the model comparison is based, we aim for deriving a probability statement for a signal detection given a certain signal amplitude h_0^*. This is made possible by the fact that each potential data set under consideration is based on the *true* parameters of a potential pulsar. Therefore each model comparison is conditioned on a data set $\boldsymbol{d}_*$ that is conditioned on the parameter vector $\boldsymbol{a}_*$. Under regularity conditions this fact can be used to obtain an approximation for the maximum likelihood value since the MLE is sufficient and asymptotically consistent for large sample. Thus the estimates converge to the true values for large samples sizes. The samples sizes that we expect are in fact in the range of tens of thousands.

A potential data set $\boldsymbol{d}_*$ from a pulsar, based on a *true* parameter vector $\boldsymbol{a}_*$ is modelled according to Eq. 2.5 by $\mathcal{M}_* : \boldsymbol{d}_*^{(k)} = y(t_k; \boldsymbol{a}_*) + \epsilon_k$ with noise vector ϵ_k. Due to the fact that $\boldsymbol{d}_*$ is conditioned on $\boldsymbol{a}_*$, an approximate maximum log-likelihood under model $\mathcal{M}_1$ is given by

$$\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} \approx -\chi^2_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1}(\boldsymbol{a}_*)/2 = -\sum_k \left|\frac{\epsilon_k}{\sigma_k}\right|^2/2, \tag{2.35}$$

using Eq. 2.5 and Eq. 2.7. This term comprises the sum of the residuals as the model is fitted by the *true* parameter vector. It is clear that this asymptotical approximation underestimates the MLE. However, this entire approach is based on the approximation by involving the AIC and BIC and is a coarse assessment for a signal detection.

Under model $\mathcal{M}_0$ that encompasses no parameters, the log-likelihood has a constant value and therefore its maximum is given by

$$\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0} = -\chi^2_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0}/2 = -\sum_k \left|\frac{y_k(t_k;\boldsymbol{a}_*)+\epsilon_k}{\sigma_k}\right|^2/2, \tag{2.36}$$

where the summation term comprises the true and given parameter vector of the signal. It is clear that $\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} \leq \log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0} \forall \boldsymbol{a}_*$. As a result of this, naturally model $\mathcal{M}_1$ has to be preferred at all times. This, however, does not take into account the penalty term that comes into play due to the principle of Occam's razor. Equality of expressions 2.36 and 2.35 can only be achieved for a zero amplitude h_0^* in parameter vector $\boldsymbol{a}_*$. But how large do we have to choose this amplitude, also considering other influential parameters, in order to justify model $\mathcal{M}_1$ with its many more parameters? This is the essential idea behind this model comparison approach and the penalty terms play a key role in it.

We aim to compare model $\mathcal{M}_0$ and $\mathcal{M}_1$ conditioned on the data set $\boldsymbol{d}_*$, conditioned on a potential pulsar characterised by the *true* parameter vector $\boldsymbol{a}_*$. By substituting Eq. 2.35 and Eq. 2.36 into Eq. 2.34, we obtain

$$\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0} = -2\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0} \tag{2.37}$$

and

$$\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} = -2\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} + 6\log n, \tag{2.38}$$

with $d = 6$ parameters under model $\mathcal{M}_1$ and $d = 0$ parameters under model $\mathcal{M}_0$. With respect to the two models $\mathcal{M}_0$ and $\mathcal{M}_1$, a probability for model $\mathcal{M}_1$ can be derived by

$$\begin{aligned}
p(\mathcal{M}_1|\boldsymbol{d}_*, \boldsymbol{a}_*) &= \frac{p(\mathcal{M}_1)e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1}/2}}{p(\mathcal{M}_0)e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0}/2} + p(\mathcal{M}_1)e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1}/2}} \\
&= \frac{e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1}/2+\log p(\mathcal{M}_1)}}{e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0}/2+\log p(\mathcal{M}_0)} + e^{-\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1}/2+\log p(\mathcal{M}_1)}} \\
&= \frac{1}{1 + e^{(\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} - \mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0})/2 - \log p(\mathcal{M}_1) + \log p(\mathcal{M}_0)}} \\
&= \frac{1}{1 + e^{\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}/2 - \log p(\mathcal{M}_1) + \log p(\mathcal{M}_0)}}
\end{aligned} \tag{2.39}$$

Here, $p(\mathcal{M}_0)$ and $p(\mathcal{M}_1)$ are prior probabilities for $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. In practice, the prior probability whether a signal detection can be expected within the narrow frequency band of 1/60Hz under consideration depends on the location on which a particular search is focused on. Derivations with different prior scenarios could be made but here, equal probabilities for the models will be assumed as a natural choice when the interest is drawn to a model detection when there is no prior information about the possible existence of a signal. Substituting the prior probabilities $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 0.5$ yields

$$p(\mathcal{M}_1|\boldsymbol{d}_*, \boldsymbol{a}_*) = \frac{1}{1 + e^{\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}/2}} \tag{2.40}$$

where

$$\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*} := \mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} - \mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0}. \tag{2.41}$$

Eq. 2.40 presents the probability that the data $\boldsymbol{d}_*$ from a potential pulsar with given parameter vector $\boldsymbol{a}_*$ is better modelled by $\mathcal{M}_1$ (a signal) rather than $\mathcal{M}_0$ (no signal). In other words it is the probability for the existence of a signal in the data that is emitted by a pulsar with parameter vector $\boldsymbol{a}_*$. It is merely the difference of the two BIC values under consideration that is responsible for a signal detection. A difference of zero for example would yield a 50% probability for both models. Since the model comparison is conditioned on the data $\boldsymbol{d}_*$ from the parameter vector $\boldsymbol{a}_*$, the next step is to derive

$$\begin{aligned} p(\mathcal{M}_1|\boldsymbol{a}_*) &= \mathrm{E}\left[p(\mathcal{M}_1|\boldsymbol{d}_*, \boldsymbol{a}_*)|\boldsymbol{a}_*\right] \\ &= \mathrm{E}\left(\frac{1}{1+e^{\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}/2}}|\boldsymbol{a}_*\right). \end{aligned} \tag{2.42}$$

Although it is possible to analytically derive $\mathrm{E}(\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}|\boldsymbol{a}_*)$ it is more complex to derive Eq. 2.42, as it comprehends the sigmoid function of $\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}$. Thus, in the sequel, the characteristics of $\Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*}$ will be derived. By using equations Eq. 2.35, Eq. 2.36, Eq. 2.37, Eq. 2.38, and Eq. 2.41 we obtain

$$\begin{aligned} \Delta\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*} &= (\mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} - \mathrm{BIC}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0}) \\ &= -2\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_1} + 6\log n + 2\log \mathrm{ML}_{\boldsymbol{d}_*,\boldsymbol{a}_*,\mathcal{M}_0} \\ &\approx \sum_k \left|\frac{\epsilon_k}{\sigma_k}\right|^2 + 6\log n - \sum_k \left|\frac{y_k(\boldsymbol{a}_*)+\epsilon_k}{\sigma_k}\right|^2 \end{aligned} \tag{2.43}$$

We assume white Gaussian noise here and therefore the distribution of the noise component ϵ_k of bin k is $\epsilon_k \propto N(0,\sigma_k^2)$ with known variances σ_k^2. By substituting

$$\begin{aligned} y(t_k;\boldsymbol{a}_*) &= \frac{1}{4}h_0^*(1+\cos^2\iota^*)F_+(t_k;\psi^*,\alpha^*,\delta^*)e^{i\Delta\Psi(t_k;\alpha^*,\delta^*,\Delta f^*,\Delta\dot{f}^*)} \\ &- \frac{i}{2}h_0^*\cos\iota^* F_\times(t_k;\psi^*,\alpha^*,\delta^*)e^{i\Delta\Psi(t_k;\alpha^*,\delta^*,\Delta f^*,\Delta\dot{f}^*)}. \end{aligned} \tag{2.44}$$

and defining some abbreviations for an ease of use, $F_+(t_k;\psi^*,\alpha^*,\delta^*) := F_k^+$, $F_\times(t_k;\psi^*,\alpha^*,\delta^*) = F_k^\times$, $e^{i\Delta\Psi(t_k;\alpha^*,\delta^*,\Delta f^*,\Delta\dot{f}^*)} := e^{i\Delta\Psi_k} = \cos(\Delta\Psi_k) + i\sin(\Delta\Psi_k)$, $\frac{1}{4}h_0^*(1+\cos^2\iota^*) =: A^+$, and $\frac{1}{2}h_0^*\cos\iota^* =: A^\times$ we can reformulate Eq 2.43. For clarity, the following calculations will concentrate on the difference of the two sums in Eq 2.43 and leave out the penalty term $6\log n$. For the two summation terms in Eq. 2.43, we obtain

$$\sum_k \frac{|\epsilon_k|^2}{\sigma_k^2} - \sum_k \frac{|y_k(\boldsymbol{a}_*)+\epsilon_k|^2}{\sigma_k^2}$$

$$\begin{aligned}
&= \sum_k \frac{\left|\epsilon_k\right|^2}{\sigma_k^2} - \sum_k \frac{\left|A^+F_k^+ e^{i\Delta\Phi_k} + \epsilon_{k,\mathrm{re}} - iA^\times F_k^\times e^{i\Delta\Phi_k} + i\epsilon_{k,\mathrm{im}}\right|^2}{\sigma_k^2} \\
&= \sum_k \left[\frac{\epsilon_{k,\mathrm{re}}^2}{\sigma_k^2} + \frac{\epsilon_{k,\mathrm{im}}^2}{\sigma_k^2}\right] - \sum_k \left[\left(\frac{A^+F_k^+\cos(\Delta\Psi_k) + A^\times F_k^\times \sin(\Delta\Psi_k) + \epsilon_{k,\mathrm{re}}}{\sigma_k}\right)^2\right. \\
&\quad \left. + \left(\frac{-A^\times F_k^\times \cos(\Delta\Psi_k) + A^+F_k^+\sin(\Delta\Psi_k) + \epsilon_{k,\mathrm{im}}}{\sigma_k}\right)^2\right] \\
&= \sum_k \left[\frac{\epsilon_{k,\mathrm{re}}^2}{\sigma_k^2} + \frac{\epsilon_{k,\mathrm{im}}^2}{\sigma_k^2}\right] - \sum_k \left[\left(\frac{A^+F_k^+\cos(\Delta\Psi_k)}{\sigma_k}\right)^2 + \left(\frac{A^\times F_k^\times \sin(\Delta\Psi_k)}{\sigma_k}\right)^2\right. \\
&\quad + \left(\frac{A^\times F_k^\times \cos(\Delta\Psi_k)}{\sigma_k}\right)^2 + \left(\frac{A^+F_k^+\sin(\Delta\Psi_k)}{\sigma_k}\right)^2 \\
&\quad \underbrace{+\,2\frac{A^+F_k^+\cos(\Delta\Psi_k)A^\times F_k^\times\sin(\Delta\Psi_k)}{\sigma_k^2} - 2\frac{A^\times F_k^\times\cos(\Delta\Psi_k)A^+F_k^+\sin(\Delta\Psi_k)}{\sigma_k^2}}_{=0} \\
&\quad +2\frac{A^+F_k^+\cos(\Delta\Psi_k)\epsilon_{k,\mathrm{re}}}{\sigma_k^2} + 2\frac{A^\times F_k^\times\sin(\Delta\Psi_k)\epsilon_{k,\mathrm{re}}}{\sigma_k^2} \\
&\quad \left. -2\frac{A^\times F_k^\times\cos(\Delta\Psi_k)\epsilon_{k,\mathrm{im}}}{\sigma_k^2} + 2\frac{A^+F_k^+\sin(\Delta\Psi_k)\epsilon_{k,\mathrm{im}}}{\sigma_k^2} + \frac{\epsilon_{k,\mathrm{re}}^2}{\sigma_k^2} + \frac{\epsilon_{k,\mathrm{im}}^2}{\sigma_k^2}\right] \\
&= -\sum_k \left[\left(\frac{A^+F_k^+\cos(\Delta\Psi_k)}{\sigma_k}\right)^2 + \left(\frac{A^\times F_k^\times\sin(\Delta\Psi_k)}{\sigma_k}\right)^2\right. \\
&\quad + \left(\frac{A^\times F_k^\times\cos(\Delta\Psi_k)}{\sigma_k}\right)^2 + \left(\frac{A^+F_k^+\sin(\Delta\Psi_k)}{\sigma_k}\right)^2 \\
&\quad +2\frac{A^+F_k^+\cos(\Delta\Psi_k)}{\sigma_k}\frac{\epsilon_{k,\mathrm{re}}}{\sigma_k} + 2\frac{A^\times F_k^\times\sin(\Delta\Psi_k)}{\sigma_k}\frac{\epsilon_{k,\mathrm{re}}}{\sigma_k} \\
&\quad \left. -2\frac{A^\times F_k^\times\cos(\Delta\Psi_k)}{\sigma_k}\frac{\epsilon_{k,\mathrm{im}}}{\sigma_k} + 2\frac{A^+F_k^+\sin(\Delta\Psi_k)}{\sigma_k}\frac{\epsilon_{k,\mathrm{im}}}{\sigma_k}\right]
\end{aligned} \tag{2.45}$$

Since the quadratic noise terms all cancel out, we are left with sums of constants plus sums of normally distributed terms. Therefore, for arbitrary data sets $\boldsymbol{d}_*$, given their true parameter vector $\boldsymbol{a}_*$, the term Eq. 2.45 is normally distributed with mean

$$
\begin{aligned}
\mu_{\boldsymbol{a}_*} &= -\Bigg((A^+)^2 \sum_k \left(\frac{F_k^+ \cos(\Delta\Psi_k)}{\sigma_k} \right)^2 + (A^\times)^2 \sum_k \left(\frac{F_k^\times \sin(\Delta\Psi_k)}{\sigma_k} \right)^2 \\
&\quad + (A^\times)^2 \sum_k \left(\frac{F_k^\times \cos(\Delta\Psi_k)}{\sigma_k} \right)^2 + (A^+)^2 \sum_k \left(\frac{F_k^+ \sin(\Delta\Psi_k)}{\sigma_k} \right)^2 \\
&\quad + \sum_k \left(\frac{2A^+ F_k^+ \cos(\Delta\Psi_k)}{\sigma_k^2} \underbrace{\mathrm{E}(\epsilon_{k,\mathrm{re}})}_{=0} \right) + \sum_k \left(\frac{2A^\times F_k^\times \sin(\Delta\Psi_k)}{\sigma_k^2} \underbrace{\mathrm{E}(\epsilon_{k,\mathrm{re}})}_{=0} \right) \\
&\quad - \sum_k \left(\frac{2A^\times F_k^\times \cos(\Delta\Psi_k)}{\sigma_k^2} \underbrace{\mathrm{E}(\epsilon_{k,\mathrm{im}})}_{=0} \right) + \sum_k \left(\frac{2A^+ F_k^+ \sin(\Delta\Psi_k)}{\sigma_k^2} \underbrace{\mathrm{E}(\epsilon_{k,\mathrm{im}})}_{=0} \right) \Bigg) \\
&= -\Bigg((A^+)^2 \sum_k \left(\frac{F_k^+ \cos(\Delta\Psi_k)}{\sigma_k} \right)^2 + (A^\times)^2 \sum_k \left(\frac{F_k^\times \sin(\Delta\Psi_k)}{\sigma_k} \right)^2 \\
&\quad + (A^\times)^2 \sum_k \left(\frac{F_k^\times \cos(\Delta\Psi_k)}{\sigma_k} \right)^2 + (A^+)^2 \sum_k \left(\frac{F_k^+ \sin(\Delta\Psi_k)}{\sigma_k} \right)^2 \Bigg) \\
&= -\Bigg((A^+)^2 \sum_k \left(\frac{F_k^+}{\sigma_k} \right)^2 \underbrace{\left(\cos^2(\Delta\Psi_k) + \sin^2(\Delta\Psi_k) \right)}_{=1} \\
&\quad + (A^\times)^2 \sum_k \left(\frac{F_k^\times}{\sigma_k} \right)^2 \underbrace{\left(\sin^2(\Delta\Psi_k) + \cos^2(\Delta\Psi_k) \right)}_{=1} \Bigg) \\
&= -(\frac{1}{4} h_0^* (1 + \cos^2 \iota^*))^2 \sum_k \left(\frac{F_+(t_k; \psi^*, \alpha^*, \delta^*)}{\sigma_k} \right)^2 \\
&\quad - (\frac{1}{2} h_0^* \cos \iota^*)^2 \sum_k \left(\frac{F_\times(t_k; \psi^*, \alpha^*, \delta^*)}{\sigma_k} \right)^2 .
\end{aligned}
\tag{2.46}
$$

and variance

$$
\sigma_{\boldsymbol{a}_*}^2 = \mathrm{Var}\left(\sum_k \left[\underbrace{2 \left(\frac{A^+ F_k^+ \cos(\Delta\Psi_k)}{\sigma_k} + \frac{A^\times F_k^\times \sin(\Delta\Psi_k)}{\sigma_k} \right)}_{:=C_k^+} \frac{\epsilon_{k,\mathrm{re}}}{\sigma_k} \right.\right.
$$

$$
\begin{aligned}
& \qquad + \underbrace{2\left(-\frac{A^{\times}F_k^{\times}\cos(\Delta\Psi_k)}{\sigma_k} + \frac{A^{+}F_k^{+}\sin(\Delta\Psi_k)}{\sigma_k}\right)}_{:=C_k^{\times}} \frac{\epsilon_{k,\mathrm{im}}}{\sigma_k} \Bigg]\Bigg) \\
\overset{\epsilon_{k,\mathrm{re}},\epsilon_{k,\mathrm{im}}\text{i.i.d.}}{=} & \quad \sum_k \left(\mathrm{Var}\left(C_k^{+}\frac{\epsilon_{k,\mathrm{re}}}{\sigma_k}\right) + \mathrm{Var}\left(C_k^{\times}\frac{\epsilon_{k,\mathrm{im}}}{\sigma_k}\right)\right) \\
= & \quad \sum_k \left(\left(C_k^{+}\right)^2 \underbrace{\mathrm{Var}\left(\frac{\epsilon_{k,\mathrm{re}}}{\sigma_k}\right)}_{=1} + \left(C_k^{\times}\right)^2 \underbrace{\mathrm{Var}\left(\frac{\epsilon_{k,\mathrm{im}}}{\sigma_k}\right)}_{=1} \right) \\
= & \quad 4\sum_k \left(\left(\frac{A^{+}F_k^{+}\cos(\Delta\Psi_k)}{\sigma_k} + \frac{A^{\times}F_k^{\times}\sin(\Delta\Psi_k)}{\sigma_k} \right)^2 \right. \\
& \qquad \left. + \left(\frac{A^{+}F_k^{+}\sin(\Delta\Psi_k)}{\sigma_k} - \frac{A^{\times}F_k^{\times}\cos(\Delta\Psi_k)}{\sigma_k} \right)^2 \right) \\
= & \quad 4\sum_k \left(\left(\frac{A^{+}F_k^{+}}{\sigma_k}\right)^2 + \left(\frac{A^{\times}F_k^{\times}}{\sigma_k}\right)^2 \right) \\
= & \quad \left(\frac{1}{2}h_0^{*}(1+\cos^2\iota^{*})\right)^2 \sum_k \left(\frac{F_{+}(t_k;\psi^{*},\alpha^{*},\delta^{*})}{\sigma_k}\right)^2 \\
& \qquad + \left(h_0^{*}\cos\iota^{*}\right)^2 \sum_k \left(\frac{F_{\times}(t_k;\psi^{*},\alpha^{*},\delta^{*})}{\sigma_k}\right)^2
\end{aligned}
\tag{2.47}
$$

Note that both expressions Eq. 2.46 and Eq. 2.47 are independent of the frequency parameters Δf^{*}, $\Delta\dot{f}^{*}$ and phase ϕ_0^{*} and they only depend on the remaining five parameters h_0^{*}, $\cos\iota^{*}$, ψ^{*}, α^{*}, and δ^{*}. From a physical point of view this is what we had to expect, as only the strength of the signal is decisive for a detection with those parameters modulating the amplitude of the signal. The parameters ψ^{*}, α^{*}, and δ^{*} only enter in the plus and cross polarisation terms F_k^{+} and $F_k^{\times}$ of the antenna pattern. The antenna pattern depends on the varying orientation of the interferometer towards the pulsar and the polarisation angle of the gravitational wave that it emits. The gravitational wave itself depends on the position angle of the spin axis. Hence, the summation terms in Eq. 2.46 containing the antenna pattern depend on that orientation sweep during the observation time and the noise.

Now that we know that the difference of the summation terms is normally distributed with mean $\mu_{\boldsymbol{a}_*}$ and variance $\sigma^2_{\boldsymbol{a}_*}$, for arbitrary data vectors $\boldsymbol{d}_*$, given the parameter vector $\boldsymbol{a}_*$, it follows that

$$\Delta\mathrm{BIC}|\boldsymbol{a}_* \sim \mathrm{N}(\mu_{\boldsymbol{a}_*} + 6\log n, \sigma^2_{\boldsymbol{a}_*}). \tag{2.48}$$

In order to be able to speak in terms of a simple signal-to-noise ratio we will expect a signal with constant noise $\sigma = \sigma_k$. In fact, the noise is fairly stable in reality. The constant noise assumption is not a necessary restriction, however it is much easier to handle a single signal-to-noise ratio h_0^*/σ. Eq. 2.48 depends on five parameters of the pulsar but also on the noise σ and the observation period OP due to the sums of the antenna pattern over that time. As already mentioned, it is sensible to combine the signal's amplitude h_0^* and the noise σ to a signal-to-noise ratio h_0^*/σ represented as one parameter. We can therefore define a new vector $\boldsymbol{a}_\bullet$ given by $\boldsymbol{a}_\bullet = (h_0^*/\sigma, \cos\iota^*, \psi^*, \alpha^*, \delta^*, \mathrm{OP})'$ with observation period $\mathrm{OP} = (t_1, \ldots, t_n)'$. Explicitly, the difference in the BIC values with respect to models $\mathcal{M}_0$ and $\mathcal{M}_1$, for arbitrary data sets, conditioned on $\boldsymbol{a}_\bullet$ follow the distribution

$$\Delta\mathrm{BIC}|\boldsymbol{a}_\bullet \sim \mathrm{N}(\mu_{\mathrm{BIC},\boldsymbol{a}_\bullet}, \sigma^2_{\mathrm{BIC},\boldsymbol{a}_\bullet}) \tag{2.49}$$

with

$$\begin{aligned}\mu_{\mathrm{BIC},\boldsymbol{a}_\bullet} \;=\;& 6\log n - \left(\frac{h_0^*}{\sigma}\right)^2 \left(\left[\frac{1}{4}(1+\cos^2\iota^*)\right]^2 \sum_k (F_+(t_k;\psi^*,\alpha^*,\delta^*))^2 \right.\\ & \left. + \left[\frac{1}{2}\cos\iota^*\right]^2 \sum_k (F_\times(t_k;\psi^*,\alpha^*,\delta^*))^2 \right)\end{aligned} \tag{2.50}$$

and

$$\sigma^2_{\mathrm{BIC},\boldsymbol{a}_\bullet} \;=\; \left(\frac{h_0^*}{\sigma}\right)^2 \left(\left[\frac{1}{2}(1+\cos^2\iota^*)\right]^2 \sum_k (F_+(t_k;\psi^*,\alpha^*,\delta^*))^2\right.$$

$$
\begin{aligned}
&\quad + [\cos\iota^*]^2 \sum_k \left(F_\times(t_k;\psi^*,\alpha^*,\delta^*)\right)^2 \Bigg) \\
&= \quad -4\left(\mu_{\mathrm{BIC},\boldsymbol{a}_\bullet} - 6\log n\right).
\end{aligned}
\tag{2.51}
$$

The use of Monte Carlo methods allow for simulating from Eq. 2.42 directly by creating l samples $s_i \sim \mathrm{N}(\mu_{\mathrm{BIC},\boldsymbol{a}_\bullet}, \sigma^2_{\mathrm{BIC},\boldsymbol{a}_\bullet})$, $i = 1, \ldots, l$ to obtain the estimate

$$
\hat{p}(\mathcal{M}_1|\boldsymbol{a}_\bullet) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{(1+e^{s_i/2})}. \tag{2.52}
$$

The following section deals with the signal detection for different scenarios.

2.6.3 Signal detection results for different scenarios

Following the examples from above, Fig. 2.11 below therefore displays the expected probabilities for a signal detection from a potential pulsar as would be seen by the LIGO-Hanford interferometer with respect to the two observation periods OP_1 and OP_2 and its sky position with full parameter vectors

$$
\boldsymbol{a}_\bullet^{(1)} = \begin{pmatrix} h_0^*/\sigma \\ \cos\iota^* \\ \psi^* = 0.35 \\ \alpha^* = 5^{\mathrm{h}}\,35^{\mathrm{m}}\,28.03^{\mathrm{s}} \\ \delta^* = -69^\circ\,16'\,11.79'' \\ \mathrm{OP}_1 \end{pmatrix}, \boldsymbol{a}_\bullet^{(2)} = \begin{pmatrix} h_0^*/\sigma \\ \cos\iota^* \\ \psi^* = 0.35 \\ \alpha^* = 5^{\mathrm{h}}\,35^{\mathrm{m}}\,28.03^{\mathrm{s}} \\ \delta^* = -69^\circ\,16'\,11.79'' \\ \mathrm{OP}_2 \end{pmatrix}
$$

in which the values of h_0^*/σ and $\cos\iota^*$ are shown in the 2-D-plane.

From Fig. 2.11 we can now read off the theoretical probabilities for model $\mathcal{M}_1$ for an arbitrary data set obtained by a pulsar with the true values as used in the examples displayed in Fig. 2.7 and Fig. 2.9. For the observation period OP_1, we obtain $\hat{p}(\mathcal{M}_1|\boldsymbol{a}_\bullet^{(1)}, h_0^*/\sigma = 0.3) = 0.405$ and $\hat{p}(\mathcal{M}_1|\boldsymbol{a}_\bullet^{(1)}, h_0^*/\sigma = 0.25) = 0.057$. Note, that in cases of low signal evidence the distribution of the probability for model $\mathcal{M}_1$ is bimodal due to the sigmoid function. For a given data set obtained over OP_1 it is therefore either

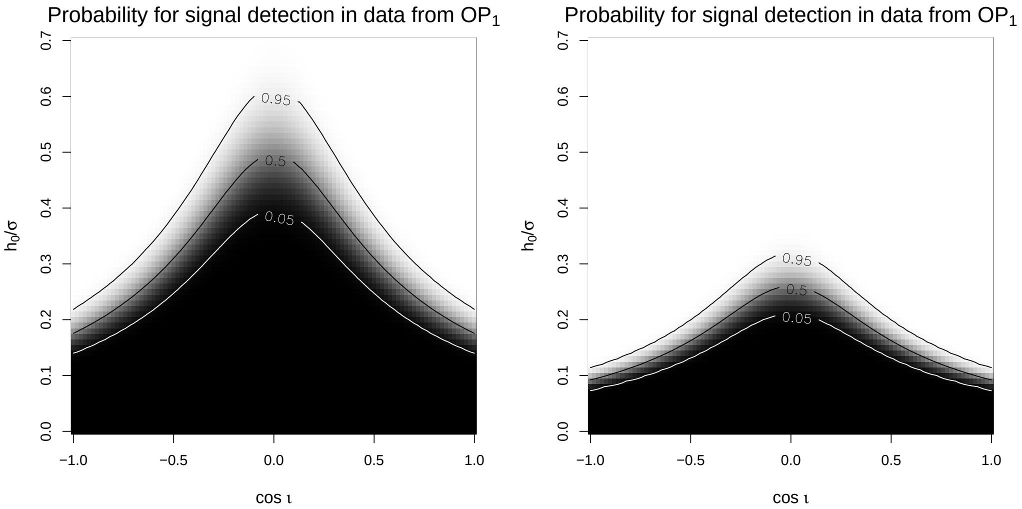


Figure 2.11: Expected probabilities $p(\mathcal{M}_1|\boldsymbol{a}_\bullet^{(1)})$ and $p(\mathcal{M}_1|\boldsymbol{a}_\bullet^{(2)})$ of a model that suggests a signal within the data that originate from a potential pulsar with parameter vector $\boldsymbol{a}_\bullet$. The probability is displayed as a function of the signal-to-noise ratio h_0^*/σ and the inclination $\cos\iota$. Both plots are conditioned on the same fixed observation periods (OP_1 with 14 400 bins and OP_2 with 60 000 bins), polarisation angle, and sky location as in the examples used above.

very likely to detect a signal or very unlikely and the mean expresses the general probability for an arbitrary data set. From the results shown in Fig. 2.7 we can anticipate that the detection probability would be rather low as it already reveals tendencies of sampling occasionally from other parts of the posterior distribution dominated by the prior. We also know that no convergence could be achieved in Fig. 2.9. A signal evidence of around 0.405 when $h_0^*/\sigma = 0.3$ seems to be a bit low but it underpins theoretically that a signal this weak is located very close the detection limit. The computation of an AIC based evidence for model $\mathcal{M}_1$ reveals the values 0.997 and 0.976 for $h_0^*/\sigma = 0.3$ and $h_0^*/\sigma = 0.25$ respectively and confirms that the AIC tends to overestimate the evidence of $\mathcal{M}_1$.

Eager to pursue the goal of estimating the probability of detecting an arbitrary pulsar, given only a signal-to-noise ratio h_0^*/σ, we start by marginalising over the remaining parameters by considering their individual priors. A

Monte Carlo method is used here, that samples from the prior distribution and calculates an estimation of $\hat{p}(\mathcal{M}_1|\boldsymbol{a}_\bullet)$. Since it is easy to sample from the prior distributions, a sufficient number of samples can be generated in order to obtain good estimates.

The prior of $\cos\iota$ and ψ are uniform $p_{\cos\iota}(\cos\iota^*) = 1_{[-1,1]}(\cos\iota^*)/2$ and $p_\psi(\psi^*) = 1_{[-\pi/4,\pi/4]}(\psi^*)/(\pi/2)$, respectively. The sky location parameters α^* and δ^* are not actual parameters but the detection probability is conditioned on them. Therefore, a prior distribution is required in order to be able to properly marginalise over them. A sensible distribution is uniform over the entire firmament. In order to account for the fact that lines of constant latitude are more dense in the polar region than near the equator, the proper distribution is given by $p_\alpha(\alpha^*) = 1_{[-\pi,\pi]}(\alpha^*)/(2\pi)$ and $p_\delta(\delta^*) = 1_{[-\pi/2,\pi/2]}(\delta^*)\cos(\delta^*)/2$.

The marginalisation over $\cos\iota^*$, ψ^*, α^*, and δ^* allows find the signal-to-noise ratio that is required in order to detect an arbitrary pulsar. However, when it comes to the actual process of observing data, the observation period is intrinsically given as well as the sky location on which the search is focused, due to the approach of this problem, by estimating the variation at a known sky location. Therefore, in the case of an observed data set we do not need to marginalise over those parameters if we intend to find out about the probability of detecting a signal within that very data set. In an intermediate step before marginalising over all parameters, we are first of all interested in

$$p(\mathcal{M}_1|h_0^*/\sigma, \alpha^*, \delta^*, \mathrm{OP}) = \int p(\mathcal{M}_1|\boldsymbol{a}_\bullet)p_{\cos\iota}(\cos\iota^*)p_\psi(\psi^*)d(\cos\iota^*)d\psi^* \tag{2.53}$$

with $\cos\iota^*$ and ψ^* being the only parameters that need to be integrated out. By using the parameters in the examples used above, Fig. 2.12 shows that probability of their detection subject to the signal-to-noise ratio.

If conclusions are to be drawn for arbitrary data obtained with respect to the specific sky location that was used in the example above, Fig. 2.12 provides the necessary information at what signal-to-noise ratio we are able to detect any signal. The shape of the 99% credibility region, indicated by the gray area, arises from the fact that the distribution of the probability

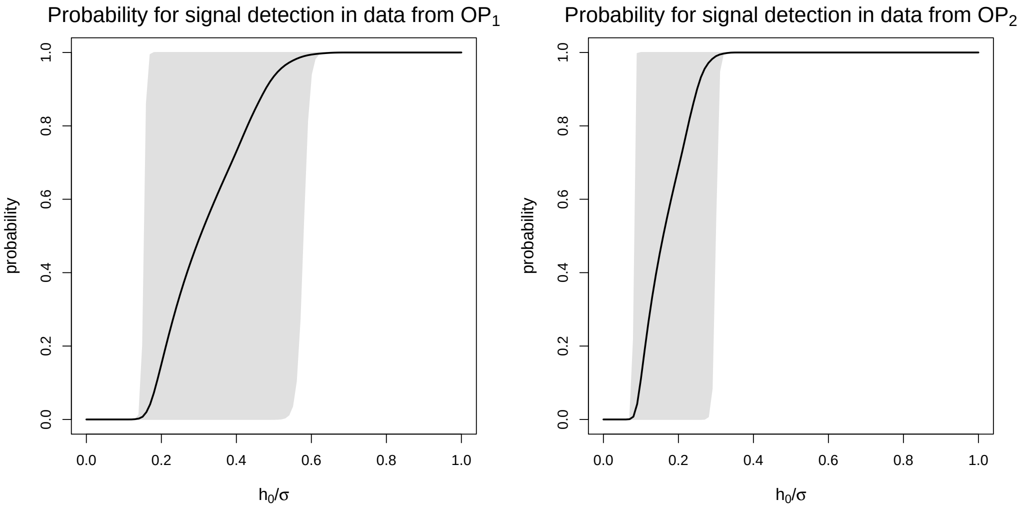


Figure 2.12: Marginal probability over the inclination $\cos\iota^*$ and polarisation angle ψ^* as a function of the signal-to-noise ratio h_0^*/σ. This is the probability of a model that suggests a signal within the data. The data from a potential pulsar are conditioned on the same fixed observation periods (OP_1 with 14 400 bins and OP_2 with 60 000 bins) and a sky location as in the examples used above. The gray area signifies the 99% credibility region. Note that the distribution of the probability for a given h_0^*/σ is bi-modal within the transition region with evidence of both models.

of a signal detection at a given signal-to-noise ratio is strongly bi-modal in the transition area where both models have evidence. This is due to the fact that the probability BIC is a function of the sigmoid function of the BIC difference.

In Fig. 2.11 and Fig. 2.12 the impact of the observation length becomes quite apparent. Naturally, with increasing observation time the detection becomes more likely. If the observation starts at the point in time at which both OP_1 and OP_2 started, then the detection probability as presented in Fig. 2.12 can be derived with respect to increasing observation lengths. Therefore Fig. 2.13 displays the probability of detection in a two dimensional plot with signal-to-noise ratio h_0/σ and different observation lengths.

We proceed with marginalising over the remaining parameters in order

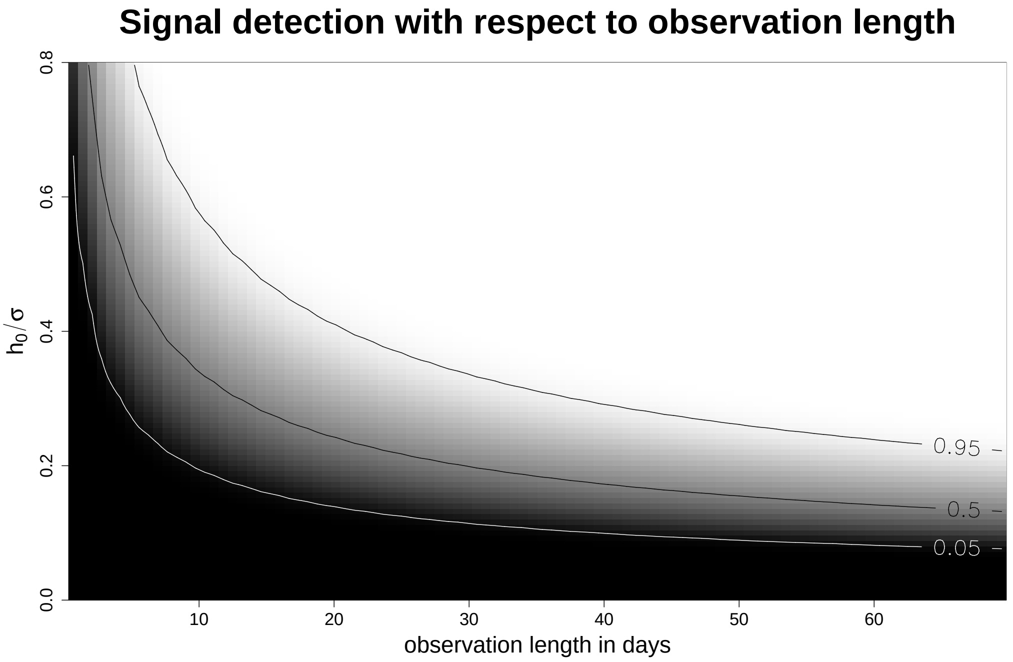


Figure 2.13: Marginal probability over the inclination $\cos\iota^*$ and polarisation angle ψ^* as a function of the signal-to-noise ratio h_0^*/σ. This is the probability of a model that suggests a signal within the data of different observation length, starting at the same time as OP_1 and OP_2. The sky location has the same value as in the examples used above.

to derive probabilities for arbitrary pulsars and obtain the probability

$$\begin{aligned} & p(\mathcal{M}_1|h_0^*/\sigma, \mathrm{OP}) \\ = & \int p(\mathcal{M}_1|\boldsymbol{a}_\bullet)p_{\cos\iota}(\cos\iota^*)p_\psi(\psi^*)p_\alpha(\alpha^*)p_\delta(\delta^*)d(\cos\iota^*)d\psi^*d\alpha^*d\delta^* \end{aligned} \tag{2.54}$$

as a function merely of the signal-to-noise ratio and the observation period. Still, $p(\mathcal{M}_1|h_0^*/\sigma, \mathrm{OP})$ depends on a fixed observation period OP, affecting the sum over the antenna pattern. Nevertheless, the marginalisation over all possible observation periods is hardly feasible but the fact that we integrate over all possible sky locations renders this step irrelevant as the trail of LIGO is of no importance against the background of all possible sky locations. The observation length with its n data samples, however, is a decisive factor as can be seen between OP_1 or OP_2 with OP_2 being $4.167\times$ longer than OP_1. Hence, $n = |\mathrm{OP}|$ remains the important parameter to condition $p(\mathcal{M}_1|h_0^*/\sigma, n)$ on. Thus, a detection probability merely depending on h_0^*/σ and the size n of the data set is displayed in the following Fig. 2.14.

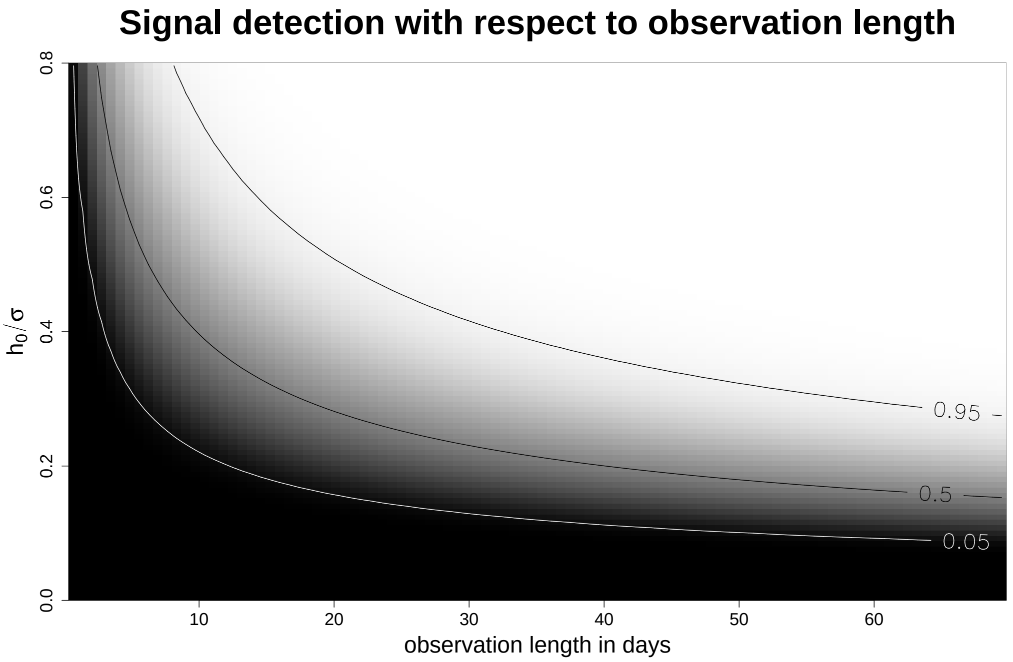


Figure 2.14: Marginal probability over the sky location, inclination $\cos\iota^*$ and polarisation angle ψ^* as a function of the signal-to-noise ratio h_0^*/σ. This is the probability of a model that suggests a signal within the data of different observation length.

The derivations made above help us to evaluate the limit for the MCMC sampler to provide us with sensible estimations of pulsar parameters. It provides a remedy for answering the question about the existence of a signal in those cases where no distinct posterior mode could be found during a sufficiently long run. From the numerous runs that have been conducted, confident statements can be made if the sampler reveals no sensible estimates for the frequency parameters after 1 000 000 iterations. Then, it can be assumed that the signal must lie below the detection limit that can be derived by using Eq. 2.53 with the known sky location parameters and observation time OP. The estimations based on the BIC will slightly overestimate values for h_0 whereas calculations based on the AIC would give smaller estimates.

The model comparison made above was based on the assumption that the model that postulates a signal in the data is parameterised by $d = 6$ parameters. But this context also allows for compiling plots that consider a model with exact known frequency parameters Δf and $\Delta \dot{f}$ ($d = 4$ parameters) [10]. The detection probability for different dimensions of the parameter space is characterised by the penalty term $6 \log n$ in Eq. 2.34. This term has a major impact as can be seen in Fig. 2.15 below.

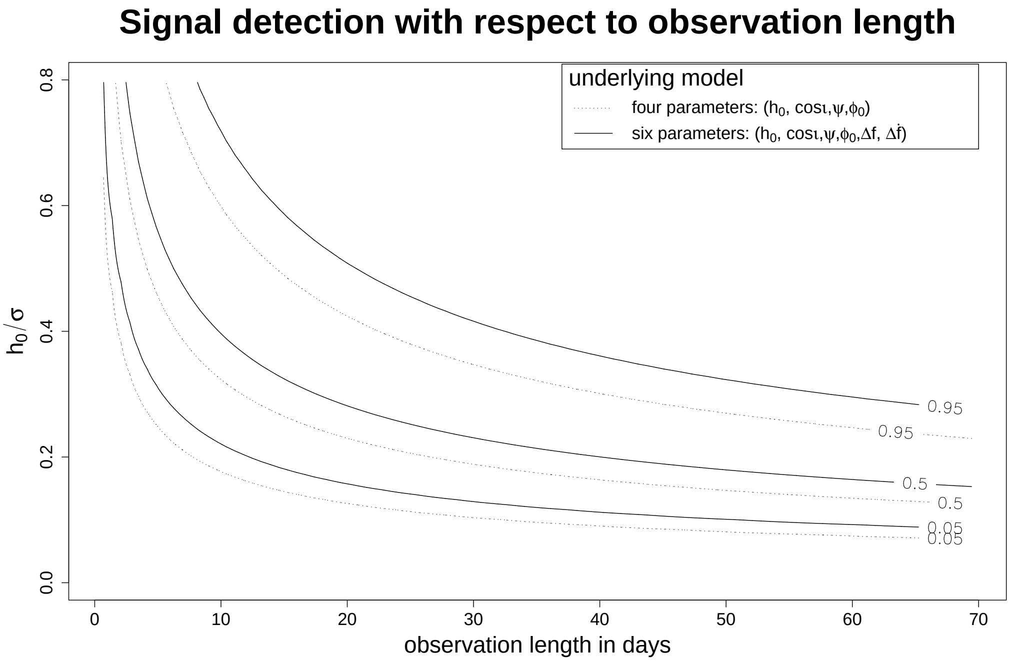


Figure 2.15: Marginal probability over the sky location, inclination $\cos \iota^*$ and polarisation angle ψ^* as a function of the signal-to-noise ratio h_0^*/σ. This is the comparison of the probabilities of two models of different complexity that suggest a signal within the data of different observation length.

2.7 Discussion

The code has also been successfully tested on real interferometer data, where the noise spectral density is coloured, containing artificially injected signals. The heterodyning process and the noise estimation procedure [35] were still successful. The MCMC routine was also successful and robust with this data although the results with coloured noise are likely to yield more conservative results. In summary, these MCMC methods potentially offer great benefits for gravitational radiation searches where the signals depend on a large number of parameters.

In the simplest application, the method demonstrated here could complement searches for signals from known pulsars [24, 35]; it could be used to verify the frequency and frequency derivative values. The real advantage of the technique would come about in a search for a signal at a known location, but where the frequency information pertaining to the neutron star is not well known; a search for a signal from SN1987A [27] would be a possible application.

The heterodyning process requires the sky location to be known from radio observations. Although the antenna pattern changes only on a small scale within the vicinity of a reference sky location, for a long observation period the motion of the earth induces a Doppler modulation into the signal which depends on the orientation between interferometer and sky location. If the heterodyning process was conducted to a vaguely known reference sky location the aberration between reference and actual sky location of the pulsar would induce a Doppler modulation into the heterodyne frequency. This would allow for a very exact determination of the sky location as the modulation pattern is characterised by the daily rotation of the earth and the yearly modulation around the sun and needs to match accurately over the entire observation time. In order to demonstrate the significant impact of small location offsets, following the examples in Sec. 2.5.1, Fig. 2.16 displays a Doppler modulation scenario for the observation period OP_2 if the signal had been heterodyned with respect to $\alpha = 5^{\mathrm{h}}\,35^{\mathrm{m}}$, $\delta = -70°$. The frequency modulation for two different sky locations ($\alpha = 5^{\mathrm{h}}\,32^{\mathrm{m}}\,18.93^{\mathrm{s}}$,

$\delta = -68^\circ\, 28'\, 6.44'')$ and $(\alpha = 5^{\mathrm{h}}\, 36^{\mathrm{m}}$, $\delta = -70^\circ\, 10')$ are shown if the pulsar is assumed to have 300 Hz.

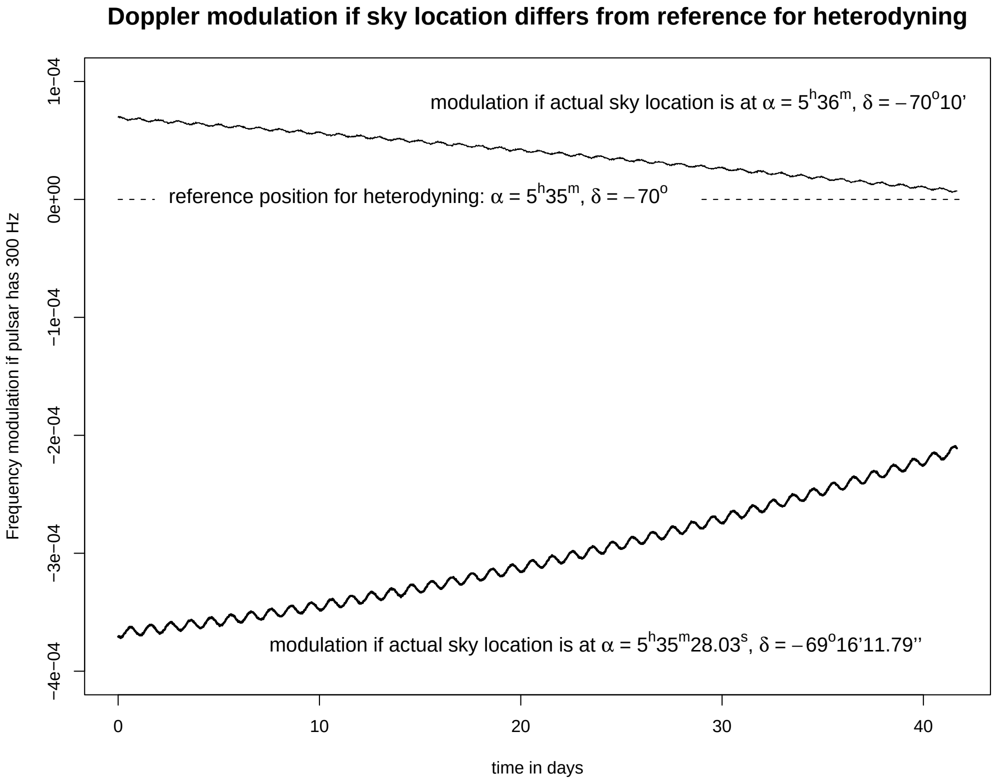


Figure 2.16: Doppler modulation of the heterodyne frequency in case of an offset between reference sky location for heterodyning and actual sky location of the pulsar. Two possible sky locations are shown in order to exhibit the strong influence of the sky location offset on the Doppler modulation.

This most welcome fact of being able to accurately determine the actual sky location of the pulsar by the Doppler shift modulation, however, exhibits a challenge to the MCMC sampler. This research has demonstrated the difficulty in finding the frequency discrepancy and its derivative. The two additional parameters of a sky location aberration from the reference location would introduce a phase evolution change in a highly nonlinear manner and complicates the construction of an efficient MCMC sampler that can tackle an eight parameter problem with vaguely known sky location.

2.8 Outlook

A future goal is to expand the model to eight parameters in order to estimate the exact sky location of a pulsar in cases where it is known only vaguely. It also has to take into account that the frequency shift might suppress a signal if the frequency modulation exceeds the 1/60 Hz band under consideration.

In the demonstration here the heterodyning process provides a band of 1/60 Hz. It would be straightforward to expand this search to a bandwidth of 5 Hz by running the code on 300 processors, a task easily accomplished on a cluster of computers. For 10 days of data it takes a single 2.8 GHz personal computer approximately an hour to conduct about $3 - 5 \times 10^4$ iterations per chain of the MCMC depending on the stage of the Markov chain. Due to the delayed rejection method, at the beginning of a run more iterations are performed per time interval because at that time more stage-1 steps are accepted.

Chapter 3

Bayesian estimation of confusion noise expected from LISA capture sources

3.1 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 [61]. It will be the first space-based gravitational wave observatory and is scheduled to be launched in 2015 as a NASA and ESA joint project. Since LISA is an all-sky monitor the information of all sources of gravitational waves has to be extracted simultaneously from the observed data stream. Unlike the ground-based interferometers, LISA will be sensitive in a much lower frequency band covering the detection of signals from close by White Dwarf Binaries, captures of stellar-mass objects by massive black holes (capture sources), and black hole binaries. Therefore especially in the frequency band below 10^{-3} Hz, LISA encounters confusion noise [62, 63] from an expected number of $100,000$ signals. In particular the frequency band between 10^{-3} Hz to 5×10^{-3} Hz presents a great data analysis challenge.

LISA will be placed in a solar orbit trailing the earth by about 20 degrees. Its three identical spacecraft form an almost equilateral triangle with an arm length of 5×10^6 km. The time dependent directional properties of the LISA response due to its motion around the sun and the nonisotropic and inhomogeneous distribution of the white-dwarf-white-dwarf binary systems within the galactic disk as seen by LISA, imply that the magnitude of the background noise observed by LISA will not be stationary [64]. Hence, for a time period of about two months within the yearly orbit the background noise level is expected to be only a factor less than two larger than the level of instrumental noise [64].

The complete removal of background noise in the LISA data is expected to be feasible only above a cutoff frequency. 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 [63, 65, 66]. LISA's ability to detect and characterise 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, the reader might be interested in Barack and Cutler [63] and in Nelemans et al. [67].

This problem is approached from a new direction. Markov chain Monte Carlo (MCMC) techniques have been demonstrated to be especially suited to parameter estimation problems involving numerous parameters [7]. In Chapter 2 the idea of the Metropolis-Hastings (MH) algorithm [28, 29] has been applied to a gravitational radiation problem, such as estimating astrophysical parameters for gravitational wave signals from pulsars but also in the literature for coalescing compact binary systems [8]. MCMC methods promise an effective means for identifying and characterizing the thousands of background binary signals to be found within the LISA data.

The method that is presented in this chapter is not a *source subtraction* approach [68] or a grid based template search using optimal filtering [69], but one that identifies and characterises 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; the method described here also produces an estimate of the overall level of the noise. It will be shown that the noise level estimate from this method depends on the inherent detector noise level, and also the presence of unidentified signals.

MCMC methods are robust and dynamic, and ultimately offer the possibility to be utilised 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. Monte Carlo methods and more recently genetic algorithms have been applied to LISA data [70, 71] using a generalised $\mathcal{F}$-statistic to extract the likelihood of a fixed dimension model of galactic binary signals within a small frequency band. Bayes factors serve for comparison of competing models in order to estimate the number of signals.

The Bayesian approach presented in this chapter does not need to fit each model with m signals, for $m = 1, \dots M$, and then select the best fitting model via the evaluation of Bayes factors. The evaluation of Bayes factors [50, 72] requires computation of the marginal likelihoods and thus marginalisations over the parameter vectors of each model. This is a formidable computational problem when the dimension of the parameter space is large. A shortcut to the calculation of Bayes factors, the harmonic mean of the likelihood values [51], 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 (BIC), also referred to as *Schwarz Criterion*, and the related penalised likelihood ratio model choice criterion, Akaike Information Criterion (AIC), have been shown to be inconsistent when the dimension of the parameter space goes to infinity [73]. The newly developed deviance information criterion (DIC) [54] is known to be controversial in mixture models [56]. 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 [52] and extended to output from the MH algorithm [53]. These, however, are impractical when the number of candidate models is very large, as is the case for LISA data. Therefore, another strategy is used, the reversible jump algorithm [11], that samples over the model *and* parameter space in order to estimate posterior model probabilities/marginal likelihoods. The number of sinusoids is considered as an additional parameter with its marginal posterior distribution. Its modal value will give us the model with the most probable number of sinusoids.

A Bayesian analysis naturally justifies Occam's Razor [74, 75] due to the penalisation of unreasonably complex models by the integration over the parameter space resulting in the preference for a simpler (smaller m) model. In this chapter, the simultaneous detection and estimation of sinusoids is addressed and solved. In addition, the MCMC method developed in this context is better than a classical periodogram at resolving signals that are very close in frequency, and a detailed discussion is provided of how to identify these signals. Finally, the method infers the noise level in the data together with the parameters of each of the m sinusoids.

This chapter presents the results of a simulation study, comprising a data stream of m sinusoidal signals embedded in Gaussian noise, where m and noise variance are among the parameters to be estimated. Although this simulation study is simple, it does highlight a number of issues relevant to the real LISA data analysis problem.

The problem of identifying an unknown number of sinusoids is neither new nor simple [13, 76]. Whereas previous studies have looked for a handful of unknown signals, here results are shown for 100 signals. Another benefit of using MCMC methods is that the number of required iterations scale roughly linearly as the number of parameters increases, and does not show an exponential increase in time [7]. In view of the present problem, in order to maintain mixing of the Markov chain, only the parameters of one randomly chosen sinusoid are proposed in each iteration. It is therefore plausible that with increasing number of sinusoids the short-term correlations in the chain increase in the same way and the resulting Markov chain needs to be thinned out proportional to the number of sinusoids.

In the future the signals have to be more realistic, taking into account the orbit of the LISA satellites and the nature of the inspiral of binaries. The intention of this work is to inform LISA data analysis researchers of another possible avenue for characterizing a large number of background signals.

The rest of the chapter is organised as follows: Section 3.2 describes the Bayesian model. The posterior computational algorithms are explained in Sections 3.3 and 3.4. In Section 3.5 the results of this study, using synthesised data, are presented. This method promises great hope for signal extraction in real LISA data, and this point is discussed in Section 3.6.

3.2 The Bayesian full probability model

We consider a signal consisting of m superimposed sinusoidal signals where m is an unknown parameter. Therefore the attention is confined to a set of models $\{\mathcal{M}_m : m \in \{0, \cdots, M\}\}$ with M being the maximum number of sinusoidal signals allowed. 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, $\boldsymbol{\epsilon}^{(m)} = [\epsilon_1^{(m)}, \cdots, \epsilon_N^{(m)}]$:

$$d_j = s^{(m)}(t_j, \boldsymbol{a}_m) + \epsilon_j^{(m)}, \qquad \text{for } j = 1, \ldots, N \tag{3.1}$$

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

$$s^{(m)}(t_j, \boldsymbol{a}_m) = \sum_{i=1}^{m} \Big[A_i^{(m)} \cos(2\pi f_i^{(m)} t_j) + B_i^{(m)} \sin(2\pi f_i^{(m)} t_j) \Big], \tag{3.2}$$

so that each sinusoid component is characterised by one frequency and two amplitudes. Model $\mathcal{M}_m$ is therefore characterised 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 [17], but instead of calculating Bayes factors for every model, a trans-dimensional MCMC algorithm makes transitions between high-probability models. The number m of unknown sinusoids becomes an additional unknown parameter and the algorithm permits the determination of 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

$$p(\boldsymbol{d}|m, \boldsymbol{a}_m) \propto \frac{1}{\sigma_m^N} \exp\left\{ -\frac{1}{2\sigma_m^2} \sum_{j=1}^{N} \left[d_j - s^{(m)}(t_j, \boldsymbol{a}_m) \right]^2 \right\}. \tag{3.3}$$

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 Section 3.3.3. For simplicity the calculations use dimensionless frequencies with a Nyquist frequency of 0.5, and uniform priors are used 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 (3.1) is a linear regression model which can be written in matrix form as

$$\boldsymbol{d} = \boldsymbol{D}^{(m)} \boldsymbol{b}^{(m)} + \boldsymbol{\epsilon}^{(m)}, \tag{3.4}$$

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 $\boldsymbol{D}^{(m)}$ 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 [77], a multivariate Normal distribution with mean zero and covariance matrix $\sigma_m^2 \Sigma$. In this context, the covariance matrix is scaled by a hyperparameter g^2 in $\Sigma = g^2 \times (\boldsymbol{D}^{(m)^T} \boldsymbol{D}^{(m)})^{-1}$. The g-prior was used in [13] 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. Another choice could be to use uniform priors for the amplitudes. However, these priors are not uninformative, and the parameter space is expressed 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. In fact, it has been shown in

[17] that independent normal distributions on the Cartesian amplitude coefficients correspond to a uniform prior on the phase when the sinusoids are expressed in polar coordinates. From this point of view, the g-prior on the Cartesian amplitude coefficient is the natural choice.

The effect of the different priors on the frequency estimates is generally negligible when the data is informative, and is discussed in [17]. The posterior pdf of the frequency reaches its maximum at the same value for both prior choices but with different curvatures at this maximum. However, uniform priors impose abrupt restrictions onto the values of the amplitudes by the predefined prior ranges. This problem is caused by the correlation of sinusoids that are close in frequency in which case the different linear combination of the sinusoids allow large amplitude values without significantly affecting the likelihood. This phenomenon was encountered in a previous stage of this research [14].

It is therefore advisable to choose a prior, similar to the g-prior that favors smaller amplitudes using a normal distribution. One possibility is to choose a simplified g-prior based on an identity matrix as covariance matrix $\sigma_m^2 \Sigma = \sigma_m^2 g^2 \cdot \mathrm{I}_{2m}$ [78] and hyperparameter g^2. Interestingly, this hyperparameter is related to the expected signal-to-noise ratio (of the signal containing all sinusoids) and $\sigma_m^2 g^2$ can be exploited later for finding an appropriate proposal distribution for the amplitudes.

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

$$p(m, \boldsymbol{a}_m | \boldsymbol{d}) = \frac{p(m, \boldsymbol{a}_m) p(\boldsymbol{d} | m, \boldsymbol{a}_m)}{p(\boldsymbol{d})}, \tag{3.5}$$

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, sampling-based MCMC techniques are used to carry out posterior inference (see [7] for an introduction and overview.) These only require the unnormalised 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. 3.5 and to estimate the quantities of interest. However, in the present context the overall model does not have fixed dimension and classical MH techniques [28, 29] cannot be used to propose trans-dimensional moves. We therefore use the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm for the model determination [11, 12]. Additionally we use the delayed rejection (DR) method [39, 30] 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.

3.3 Sampling from the posterior distribution

3.3.1 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 $\cup_{m=1}^{M}\left(m \times I\!R^{3m+1}\right)$ 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 $\boldsymbol{r}$. The transitions considered here imply either decreasing or increasing a model by one signal, i.e. $k' \in \{k-1, k+1\}$. Equal probabilities $p_{k\mapsto k'} = p_{k'\mapsto k}$ are used 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

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(\boldsymbol{a}'_{k'}, k')p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')p_{k\mapsto k'}}{p(\boldsymbol{a}_k, k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{r})p_{k'\mapsto k}}\right\}|J_{k\mapsto k'}| \tag{3.6}$$

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 [11]. In this context, we suggest two types of transformations, 'split-and-merge' and 'birth-and-death'.

Split-and-merge transitions

For a 'split' transition we randomly choose one of the available 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 3-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

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

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

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

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

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(k', \boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)})p(\boldsymbol{d}|\boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)}, k')}{p(k, \boldsymbol{a}_{(i)})p(\boldsymbol{d}|\boldsymbol{a}_{(i)}, k)q(\boldsymbol{r})}\right\} |J_{k\mapsto k'}|\,. \tag{3.7}$$

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

$$\alpha_{k'\mapsto k}(\boldsymbol{a}_k|\boldsymbol{a}'_{k'}) = \min\left\{1, \frac{p(k, \boldsymbol{a}_{(i)})p(\boldsymbol{d}|\boldsymbol{a}_{(i)}, k)q(\boldsymbol{r})}{p(k', \boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)})p(\boldsymbol{d}|\boldsymbol{a}'_{(i_1)}, \boldsymbol{a}'_{(i_2)}, k')}\right\} |J_{k'\mapsto k}|\,, \tag{3.8}$$

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

A multivariate normal proposal distribution $N[\mathbf{0}, \mathrm{diag}(\sigma_A^2, \sigma_B^2, \sigma_f^2)]$ is used 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 [16]. Using a Gaussian approximation to the posterior pdf of the frequency [16], 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.

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 $\mathrm{t}_{k'\mapsto k}\left(\boldsymbol{a}'_{(i)}\right) = \boldsymbol{a}'_{(i)} = \boldsymbol{r}$. The Jacobian for both of these is 1. The acceptance probability for the creation process is therefore

$$\alpha_{k\mapsto k'}(\boldsymbol{a}'_{k'}|\boldsymbol{a}_k) = \min\left\{1, \frac{p(k')p(\boldsymbol{a}'_{(i)})p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')}{p(k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{r})}\right\}, \tag{3.9}$$

and that for the annihilation process is

$$\alpha_{k'\mapsto k}(\boldsymbol{a}_k|\boldsymbol{a}'_{k'}) = \min\left\{1, \frac{p(k)p(\boldsymbol{d}|\boldsymbol{a}_k, k)q(\boldsymbol{a}'_{(i)})}{p(k')p(\boldsymbol{a}'_{(i)})p(\boldsymbol{d}|\boldsymbol{a}'_{k'}, k')}\right\}. \tag{3.10}$$

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 to 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 [79], given by

$$C(f) = \frac{1}{N}\left[R(f)^2 + I(f)^2\right], \tag{3.11}$$

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. In practice, samples are drawn from a density proportional to $C(f)$. This technique has already been applied successfully by [13].

With regard to the amplitudes we can exploit the fact that the hyperparameter g^2 is related to the expected signal-to-noise ratio, as mentioned previously. Therefore $g^2\sigma_m^2$ is a good scale for the expected variance to choose for the amplitudes. A suitable proposal is therefore a normal distribution with mean zero and the variance derived from the current state of the parameter $g^2\sigma_m^2$ in 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 delayed rejection (DR) method has been introduced by [30, 38, 39] and has been successfully applied in Chapter 2 to estimate the frequency and frequency derivative of potential gravitational radiation signals produced by a triaxial neutron star.

3.3.2 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 [39]

$$\alpha_1(\boldsymbol{a}'|\boldsymbol{a}) = \min\left[1, \frac{p(\boldsymbol{a}')p(\boldsymbol{d}|\boldsymbol{a}')q_1(\boldsymbol{a}|\boldsymbol{a}')}{p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})q_1(\boldsymbol{a}'|\boldsymbol{a})}\right] \tag{3.12}$$

and

$$\alpha_2(\boldsymbol{a}''|\boldsymbol{a}',\boldsymbol{a}) = \min\left\{1, \frac{p(\boldsymbol{a}'')p(\boldsymbol{d}|\boldsymbol{a}'')q_1(\boldsymbol{a}'|\boldsymbol{a}'')q_2(\boldsymbol{a}|\boldsymbol{a}',\boldsymbol{a}'')[1-\alpha_1(\boldsymbol{a}'|\boldsymbol{a}'')]}{p(\boldsymbol{a})p(\boldsymbol{d}|\boldsymbol{a})q_1(\boldsymbol{a}'|\boldsymbol{a})q_2(\boldsymbol{a}''|\boldsymbol{a},\boldsymbol{a}')[1-\alpha_1(\boldsymbol{a}'|\boldsymbol{a})]}\right\}. \tag{3.13}$$

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.

Proposing new amplitudes

In [13] 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. As already mentioned, the computation of the covariance matrix involves determining the inverse of $\boldsymbol{D}^{(m)^T}\boldsymbol{D}^{(m)}$ given the frequency vector $\boldsymbol{f}^{(m)}$ for each iteration, which is impractical against the background of the large number of signals expected for LISA, and in the LISA context, the amplitudes are parameters of interest. Therefore, a simplified g-prior distribution was chosen that uses an identity matrix for the covariance matrix scaled by the hyperparameter g^2. However, for the choice of an appropriate proposal distribution the covariance matrix is important in order to account for the covariance of the amplitudes between sinusoids that are close in frequency.

It is possible to consider the covariance matrix of single pairs of sinusoids that are close in frequency, limiting the size of the covariance matrix to 4×4. The delayed rejection method provides the possibility to either consider a single sinusoid in a first stage or a pair of sinusoids in a second stage in case of rejection in the first stage.

In a first stage, the change of the two amplitudes of a single, randomly chosen sinusoid i, is proposed. The variance of the proposal is derived from the hyperparameter g^2 that is the expected signal-to-noise ratio and the noise σ_m^2, scaled down by an additional parameter r for which the value

10 offers good results in this context. This yields the proposal distribution $q_2(\boldsymbol{a}'_{(i,i_c)}|\boldsymbol{a}_{(i,i_c)}) = N\left[\boldsymbol{a}_{(i)}, r \cdot g^2\sigma_m^2 \mathrm{I}_2\right]$ where $\boldsymbol{a}_{(i)} = \left(A_i^{(m)}, B_i^{(m)}\right)^T$ is the subvector of $\boldsymbol{a}$ containing the amplitudes of sinusoid i.

If the first proposal is rejected, another attempt is made that considers the interference between the chosen sinusoid and its closest neighbor on the frequency scale. The covariance matrix that depends on the basis functions of those sinusoids is used for the multivariate normal proposal distribution. The closer two sinusoids are in frequency the more correlation there is in their recovered amplitudes. In the following, the notation $\boldsymbol{D}^{(m)}(\boldsymbol{f})$ is used for the $N \times 2m$ matrix $\boldsymbol{D}^{(m)}$ given the frequency vector $\boldsymbol{f}^{(m)}$. For the randomly chosen sinusoid i and sinusoid i_c that is closest in frequency to sinusoid i, the covariance matrix derived from the subset of basis vectors has the form

$$\boldsymbol{D}_{(i,i_c)}^{(m)}(\boldsymbol{f}) = \begin{pmatrix} \cos(2\pi f_i^{(m)} t_1) & \sin(2\pi f_i^{(m)} t_1) & \cos(2\pi f_{i_c}^{(m)} t_1) & \sin(2\pi f_{i_c}^{(m)} t_1) \\ \cos(2\pi f_i^{(m)} t_2) & \sin(2\pi f_i^{(m)} t_2) & \cos(2\pi f_{i_c}^{(m)} t_2) & \sin(2\pi f_{i_c}^{(m)} t_2) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(2\pi f_i^{(m)} t_N) & \sin(2\pi f_i^{(m)} t_N) & \cos(2\pi f_{i_c}^{(m)} t_N) & \sin(2\pi f_{i_c}^{(m)} t_N) \end{pmatrix} \tag{3.14}$$

which results in a covariance matrix that is merely of type 4 × 4 for computing proposals for a sinusoid pair (i, i_c). Now suppose that $\boldsymbol{a}_{(i,i_c)} = \left(A_i^{(m)}, B_i^{(m)}, A_{i_c}^{(m)}, B_{i_c}^{(m)}\right)^T$ is a vector containing the amplitudes of the sinusoid pair (but not their frequencies). The covariance matrix is scaled by the current value of the variance parameter σ_m^2 of model m and normalised by the number of samples N yielding

$$\Sigma_{(i,i_c)} = \sigma_m^2 \cdot N/2 \left(\boldsymbol{D}_{(i,i_c)}^{(m)}(\boldsymbol{f})^T \boldsymbol{D}_{(i,i_c)}^{(m)}(\boldsymbol{f})\right)^{-1} \tag{3.15}$$

as a covariance matrix for the proposal distribution $q_2(\boldsymbol{a}''_{(i)}|\boldsymbol{a})=\mathrm{N}(\boldsymbol{a}_{(i,i_c)},\Sigma_{(i,i_c)})$ of the second stage.

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 incident 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 potentially achievable accuracy $\sigma_{f_i^{(m)}} = (2\pi)^{-1}\sqrt{48\sigma_m^2 \left[(A_i^{(m)})^2 + (B_i^{(m)})^2\right]^{-1} N^{-3}}$ of a frequency by [17]. This yields the proposal $q_2(\boldsymbol{a}''_{(i)}|\boldsymbol{a}_{(i)}) = N\left(\boldsymbol{a}_{(i)}, \mathrm{diag}\left(0, 0, \sigma_{f_i^{(m)}}\right)\right)$ and aims to draw representative samples from the local mode in the second stage.

3.3.3 Updating the noise parameter

The sum of the squared residuals between the model and the data, taken from the likelihood in Eq. 3.3, is

$$S^2 = \sum_{j=1}^{N} [d_j - s_m(t_j, \boldsymbol{a}_m)]^2. \tag{3.16}$$

Using this, we choose a vague prior for the noise parameter σ_m^2, defined by $\mathcal{IG}(\alpha, \beta) = \mathcal{IG}(N_p/2, N_p \cdot S_p^2/2)$ 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

$$p(\sigma_m^2|m, \boldsymbol{a}_m, \boldsymbol{d}) \propto \mathcal{IG}\left(\frac{N_p + N}{2}, \frac{N_p S_p^2 + S^2}{2}\right) \tag{3.17}$$

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

3.3.4 Updating the hyperparameter g^2

Similar to Ref. [13], we consider g^2 as a scale parameter of an extended hierarchical model and ascribe a vague prior $g^2 \propto \mathcal{IG}(\alpha = 2, \beta = 1)$ to it.

The samples for g^2 are then drawn in a Gibbs update from the full conditional distribution

$$p(g^2|k, \boldsymbol{a}_k) \propto \mathcal{IG}\left(k+\alpha, \mathcal{A}^T\mathcal{A}/(2\sigma_k^2)+\beta\right) \tag{3.18}$$

where $\mathcal{A} = (A_1^{(k)}, \ldots, A_k^{(k)}, B_1^{(k)}, \ldots, B_k^{(k)})^T$ is the vector containing all amplitudes of $\boldsymbol{a}_k$.

3.3.5 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 (FFT) of the data and use this to generate our initial values. Those frequencies are used that correspond 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 non-relevant sinusoids than create good ones.

If we assume fixed and equidistant time intervals, then the frequency resolution depends 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 FFT are significantly worse than those obtained by the MCMC method developed here, but they are sufficient to serve as initial values.

3.4 Identifying the sinusoids

Although the RJMCMC method facilitates the selection of the most probable model, we still encounter the *label-switching problem*. This is a general problem caused by the invariance of the likelihood under relabeling of the sinusoidal components, and has been extensively discussed in the context of mixture models [76, 80, 81, 82, 83]. 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 the 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. There are general approaches to the label switching problem discussed in the literature:

- identifiability constraints
- relabelling algorithms
- label invariant loss functions

3.4.1 Identifiability constraints

The label switching problem is often addressed by imposing artificial identifiability constraints on the parameter space. This is equivalent to imposing constraints on the prior in order to break down its symmetry and thus the symmetry of the posterior distribution. Concerns have been expressed regarding the impact on the performance of the sampler when it is bound to obey such constraints [81]. It has been shown that the identifiability constraints can be applied in a post processing procedure [83, 84] by defining a prior with zero probability when the parameter vector violates the constraint. However, especially against the background of high model orders that we face here, such a constraint would be a burden to the entire sampling and inference process.

Another issue when applying artificial identifiability constraints is their appropriate choice. Particularly for multivariate parameter spaces this constitutes a problem [83] and for the situation at hand the only physically meaningful choice at first sight would be an ordering constraint on the frequency parameters. Although a sinusoid can be primarily identified by its frequency, there are two reasons why such a constraint should be avoided. First of all, the problem of confusion noise is addressed here. That is, we expect an overlap in the marginal posterior densities of the frequency among the sinusoids. Therefore, there is no physical justification to impose a prior that expects the frequencies to be in perfect order. The amplitudes play a role in the identification of overlapping frequencies but there is no sensible way to incorporate all parameters into a straightforward identifiability constraint.

Although ordering the frequencies [13] might be appropriate for a few well separated sinusoids, artificial identifiability constraints do not prevent multimodality in the components [83]. In view of the confusion noise problem with many sinusoidal components of unknown number, a prior with a simple ordering constraint on the frequency does not provide meaningful inference. The labels of the sinusoidal components of each MCMC sample that pertain to model $\hat{\mathcal{M}}$ are a permutation of the $\hat{m}$ coexistent sinusoids that determine the model. Unfortunately, this coexistence does not imply that, over the course of the Markov chain, there are exactly $\hat{m}$ sinusoids involved within model $\hat{\mathcal{M}}$. The occurrence of higher model orders within the multidimensional Markov chain suggests the existence of sinusoids that might turn up in lower model orders as well, albeit with smaller probability. Therefore, in general, the truncation of the parameter space is not necessarily conform with the unrestricted posterior [81].

3.4.2 Relabelling algorithms

Another approach to the label switching problem is based on k-means type clustering algorithms [80, 82, 83]. They are based on loss functions that depend on a classification action and the possible permutations. These loss functions are not label invariant and depend on a permutation, other then label invariant loss functions that will be described in Sec. 3.4.3. By turns, the loss function is minimised with respect to the classification action given a permutation and to the permutation given the classification action until convergence is achieved. Finding the minimum loss with respect to a permutation given the classification action for each output MCMC sample vector bears the most computational burden in the present problem.

An additional issue is that the k-means clustering type algorithm is based on the assumption that there are k existing components within the model under consideration. As mentioned in Sec. 3.4.1, this assumption is delicate as model $\hat{\mathcal{M}}$ comprises $\hat{m}$ coexistent sinusoids from a possibly larger set due to the existence of higher order models. An algorithm like the k-means type algorithm could therefore jeopardise a meaningful inference. This issue and the huge computational burden lead to a different cluster analysis approach and was used at a prior state of this research [14].

The predominant parameter that contributes significantly to identifying a sinusoid is its frequency. We can therefore divide the problem by considering smaller frequency intervals. The dominant sinusoids that characterise model $\hat{\mathcal{M}}$ are identified by the kernel density estimate of the marginal posterior of the frequency. The frequency intervals are obtained by finding a threshold that separates $\hat{m}$ strongest peaks together with their covering frequency intervals. This technique will be tersely outlined below. It will help later to better understand the motivation and need for elaborating the method that was finally applied and will be described in Sec. 3.4.3.

Since we have to deal with a vast number of output samples grouped within many very small regions it is neither possible to apply classical kernel density estimates nor histograms. The required fixed bin size for a histogram would be too small for a feasible implementation. Instead, a variable bin size is used and the densities calculated for fixed samples per bin. Initially, all individual sinusoidal components from all MCMC samples of the considered model $\hat{\mathcal{M}}$ are sorted by their frequency. After having generated n MCMC samples of model $\hat{\mathcal{M}}$ during a run, there are $\hat{m}\cdot n$ parameter triples and hence $f_1 < \ldots < f_{n\cdot\hat{m}}$ ordered frequency samples. The density can be assessed by calculating the frequency range spanned by a fixed number of sorted frequencies. The $\hat{m} \cdot n$ sinusoidal components determined by their 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 $\mathrm{f}(f) = (2hn\hat{m})^{-1} \cdot \sum_{i=1}^{n \cdot \hat{m}} \mathrm{I}_{|f-f_i|<h}$ with uniform kernel. 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. The choice $r = 0.05$ (5%) was found to be a good value, and hence $r \cdot n$ serves as an estimate of the number of members 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 $2 \cdot h$, the density value ρ_j corresponds to each frequency sample j and its $r \cdot n - 1$ subsequent samples that fall into a frequency range of length $(f_{j+r \cdot n-1} - f_j)$. Hence, the density is given by $\rho_j = rn / \left(n\hat{m}\left(f_{j+r \cdot n-1} - f_j\right)\right) = r / \left(\hat{m}(f_{j+r \cdot n-1} - f_j)\right)$ where $j = 1, \cdots, n(\hat{m} - r)$. Since each ρ_j comprises the samples $j, \cdots, j + r \cdot n - 1$ 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 $[f_{j_{k,\mathrm{start}}}, f_{j_{k,\mathrm{end}}+r \cdot n-1}]$ where $j_{k,\mathrm{start}}$ and $j_{k,\mathrm{end}}$ are the indices of the first and the last member of the set of ρ_j's in peak k, respectively. Due to the fact that we always focus on frequency ranges that contain a fixed number 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.

There are two issues to consider that arise when applying a peak analysis: Firstly, due to the possible frequency overlaps between sinusoids, not all peaks can be clearly separated and a single peak can contain more than one sinusoid. Secondly, we can expect the existence of more then $\hat{m}$ sinusoids over the course of the Markov chain even when restricting to model $\hat{\mathcal{M}}$. Dividing the entire frequency band into intervals of high posterior density does not necessarily yield a fixed number of sinusoids per MCMC sample within each interval. However, for a multidimensional subspace the posterior probabilities for the number of sinusoids within a frequency interval can be estimated, based on the MCMC samples. A histogram can be compiled of the number of samples that fall into the restricted frequency range under consideration.

A cluster analysis serves to separate multiple sinusoids. Therefore, we include the two amplitudes and apply an agglomerative hierarchical cluster analysis that involves all three parameters. For this purpose a modified Ward technique [85] is used that minimises the within-cluster variance using a normalised Euclidean distance between the parameters by adjusting the frequency range to the much larger range of the amplitudes. The software package R [86] is deployed 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 posterior 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 super-cluster. Each single parameter triple is then allocated to a super-cluster and hence to a presumed sinusoid.

This method has proven to work well [14] but problems occur with large numbers of sinusoids. The fact that $\hat{m}$ coexistent sinusoids in model $\hat{\mathcal{M}}$ do not necessarily mean that there are exactly $\hat{m}$ sinusoids involved within this model leads to minor peaks and spread out single sinusoidal component samples in model $\hat{\mathcal{M}}$. Individually, they have a small contribution to the entire probability mass but, in sum, gain significance and require proper identification. Moreover, the correct identification within major peaks is difficult as their intervals do not necessarily cover a fixed number of sinusoids per MCMC sample. The described method can not adhere to the permutational restriction that all sinusoidal components within an MCMC sample must belong to different sinusoids. Even though, naturally, samples of individual sinusoidal components belong to different peaks and clusters, this method is not bound to it.

The considerations made above lead to the conclusion that the posterior distribution of sinusoids can be multimodal. A peak and cluster analysis based approach is therefore to be handled with care. The key problem we face is the possible existence of sinusoids that embody multiple sinusoids, interchanged in higher model orders, appearing as a single sinusoid within the model under consideration. Those entities are therefore a blend of several sinusoids and in the following context those sinusoids are referred to as *blendoids*. They are hard to identify and estimates are difficult due to their multimodal features. However, the $\hat{m}$ sinusoidal components of each MCMC sample determine the likelihood under model $\hat{\mathcal{M}}$. It is therefore imperative to regard the $\hat{m}$ sinusoids as entities. Note, that the posterior distribution of one and the same sinusoid can alter among different models due to the changing interaction between sinusoids.

Although most sinusoids have a unimodal appearance, the appearance of blendoids disorder the ranking of the sinusoidal components within each MCMC sample. Those entities are the reason for the existence of other countless minor peaks and spread out samples that can not be allocated to unimodal posterior distributions. It is essential to trace back the origin and to identify possible blendoids as entities.

3.4.3 Interval separation of sinusoids by their frequency using a label invariant loss function

Label invariant loss functions have been applied to mixture distributions [81] and mixture regressions [87]. The idea is to define a label invariant loss function of a classification action and to compute the posterior expected loss based on the MCMC output. The minimum of the posterior expected loss with respect to the classification actions is usually derived by Monte Carlo methods [83].

For this process, an expedient loss function needs to be specified that is not too computationally expensive for the minimisation procedure based on all MCMC samples. Again, we can exploit the fact that the frequency is the predominant parameter for identifying a sinusoid. Therefore, within model $\hat{\mathcal{M}}$, the entire frequency band can be divided into $\hat{m}$ disjoint and adjacent intervals I_k :$]b_k, b_{k+1}]$, $k \in \{1, \ldots, \hat{m}\}$ with their $\hat{m}+1$ boundaries b_k. The outer two b_1 and $b_{\hat{m}+1}$ are predetermined by the minimum and maximum over all frequency samples that appear in the Markov chain and it is $b_1 < b_2 < \ldots < b_{k+1}$. Let us consider n MCMC samples of model $\hat{\mathcal{M}}$ with vectors $\boldsymbol{a}_i$, $i \in \{1, \ldots, n\}$. The ith MCMC sample encompasses the sinusoidal components $\boldsymbol{a}_i^{(k)}$, $k \in \{1, \ldots, \hat{m}\}$ each with two amplitudes $A_i^{(k)}$, $B_i^{(k)}$, and frequency $f_i^{(k)}$. With this setup it is possible to count, for each MCMC sample i, the number of frequency samples $f_i^{(k)}$ that fall into the individual frequency intervals $I_1 : [b_1, b_2]$, I_2 :$]b_2, b_3]$, ... , $I_{\hat{m}}$:$]b_{\hat{m}}, b_{\hat{m}+1}]$. When this is done for all n MCMC samples, a histogram can be compiled for each interval I_k revealing how many occurrences there are in the n MCMC samples in which $l \in \{0, 1, \ldots, \hat{m}\}$ frequency samples fall into the frequency interval I_k.

Deploying the indicator function

$$\mathbf{1}_{\mathcal{A}}(x) = \begin{cases} 1 & \text{if } x \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$$

allows for defining

$$H_k^{(l)} = \sum_{i=1}^{n} \mathbf{1}_{\{l\}} \left(\sum_{h=1}^{\hat{m}} \mathbf{1}_{I_k}(f_i^{(h)}) \right) \tag{3.19}$$

as the number of occurrences in which $l \in \{1, \ldots, \hat{m}\}$ components of the individual MCMC samples fall into interval I_k. The upper bound for all $H_k^{(l)}$ is n and we therefore define the normalisation $\hat{H}_k^{(l)} := H_k^{(l)}/n$. The following graph in Fig. 3.1 exemplifies the label switching issue and the difficulty in finding optimal intervals, in case of presence of a blendoid. This example contains one blendoid that complicates the correct allocation. The described interval method, however, helps identifying those as entities. The histogram of interval I_4 in Fig. 3.1 features the lack of single occupancies. Yet, it covers the major proportion of the blendoid that can be ascribed to an actual sinusoid. The other corresponding samples are spread over all other intervals.

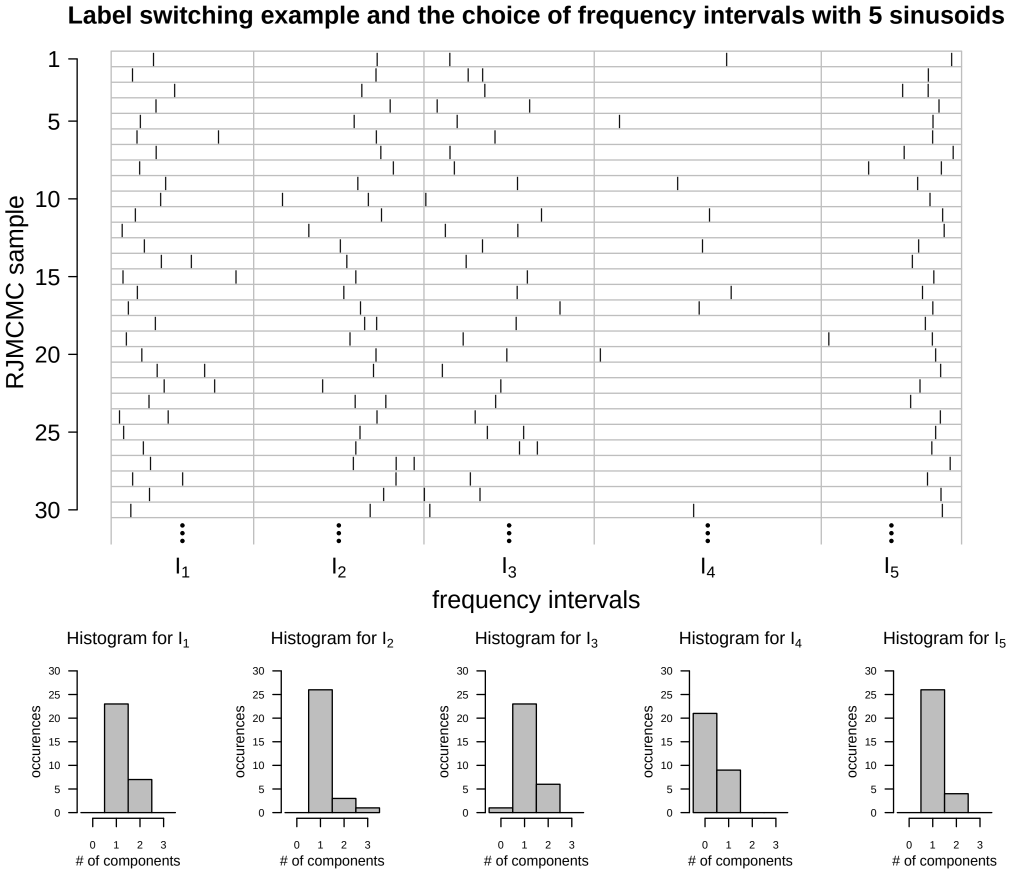


Figure 3.1: Example for the label switching problem with optimal chosen intervals to separate the sinusoidal components. This example displays the difficulty that arises in the presence of a blendoid that can be identified in interval I_4. The samples from a blendoid encompasses samples from several sinusoids of higher model order but appear as an entity in the likelihood of the model under consideration. The intervals are arranged such that all histograms yield in sum the most single occupancies. This yields the highest number of possible allocations and helps identify the core probability mass of blendoids.

The goal is to find intervals for which exactly one sinusoidal component of each single MCMC sample falls into one of the intervals, yielding $\hat{H}_k^{(1)} = 1$ for all $k \in \{1, \ldots, \hat{m}\}$. In this unlikely case, any further separation is needless.

As mentioned above, we encounter frequency samples intermingling between sinusoids that are close in frequency and furthermore the existence of blendoids. We therefore aim for maximising the term

$$H_\Sigma(b_1, \ldots, b_{\hat{m}+1}) = \sum_{k=1}^{\hat{m}} \hat{H}_k^{(1)} \tag{3.20}$$

with regard to the intervals I_k and therefore their boundaries $\{b_1, \ldots, b_{k+1}\}$. This yields the intervals that carry the most single occupancies or in other words the most unambiguous set of intervals with respect to the separation of the sinusoidal components of all MCMC samples. Eq. 3.20 has an upper bound of $\hat{m}$. Note, that b_1 and $b_{\hat{m}+1}$ can be determined by the lower and upper bound of the entire occupied frequency band whereas the inner $\hat{m}-1$ boundaries $b_2 < b_3 < \ldots < b_{\hat{m}}$ remain as parameters.

In view of the label invariant loss function, the inner boundaries represent the classification actions. The actual loss function is represented by

$$\mathcal{L}(b_1, \ldots, b_{\hat{m}+1}) := \hat{m} - H_\Sigma(b_1, \ldots, b_{\hat{m}+1}) \tag{3.21}$$

and we therefore seek to minimise Eq. 3.21 or maximise Eq. 3.20 with respect to those boundaries in order to obtain the posterior expected loss based on the MCMC output.

Finding frequency intervals that maximise the number of single occupancies

The maximisation of $H_\Sigma(b_1, \ldots, b_{\hat{m}+1})$ with respect to the boundaries of the intervals presents a global optimisation problem. A Monte Carlo technique has been deployed that tackles this challenge. It requires the sorting of all frequencies $f_i^{(1)} \leq f_i^{(2)} \leq \ldots \leq f_i^{(\hat{m})}$ within the ith MCMC sample, $\forall i \in \{1, \ldots, n\}$. The algorithm can be outlined as follows:

1. Determine the fixed outer bounds $b_1 = \min_{k,i} f_i^{(k)}$ and $b_{\hat{m}+1} = \max_{k,i} f_i^{(k)}$.

2. Determine initial values for $\{b_2, \ldots, b_{\hat{m}}\}$ by choosing randomly an MCMC sample i then calculating the in-between points $b_{k+1} = r_k \cdot (f_i^{(k)} + f_i^{(k+1)}), k \in \{1, \ldots, \hat{m} - 1\}$ of its frequency components with samples $r_k \sim \beta(5, 5)$ drawn from a beta distribution that yields values close to the center of both boundaries.

3. Compute $H_\Sigma(b_1, \ldots, b_{\hat{m}+1})$

4. Choose one of the following proposal techniques randomly:

    - <u>New:</u> Find a new set of boundaries as described in (2). This step prevents the maximisation procedure from getting trapped in a local maximum.

    - <u>Variation:</u> Choose randomly a boundary $b_k, k \in \{2, \ldots, \hat{m}\}$ with equal probability. Choose to either shift its position up or down by either setting $b'_k = r \cdot (b_{k+1} - b_k)$ or $b'_k = r \cdot (b_{k_1} - b_k)$, where r is drawn from a beta distribution $r \sim \beta(0.1, 1)$ which yields mostly timid but sufficiently often bold changes of b_k, always incorporating the distance to the neighbour boundaries.

    - <u>Exchange:</u> A particular interval can lack components $(\hat{H}_{k_1}^{(0)} \gg 0)$ while another reveals a surplus $(\sum_{l=2}^{\hat{m}} \hat{H}_{k_2}^{(l)} \gg 0)$. The boundary within the sparsely occupied interval is placed into the interval of abundant occupation. Pairs of intervals that show those conditions are preferably selected randomly. One of the framing boundaries of the sparse interval, say b_{k_1}, is chosen to be placed within the interval I_{k_2} with boundaries b_{k_2} and b_{k_2+1} that has abundant occupation. The new position is obtained by $b'_{k_1} = (1 - r) \cdot b_{k_2} + r \cdot b_{k_2+1}$ where r is drawn from $r \sim \beta(5, 5)$. This makes sure for b'_k to be placed somewhere close the center of the interval I_{k_2}. Subsequently the boundaries have to be sorted in order to maintain $b'_1 < \ldots < b'_{\hat{m}}$.

5. Compute $H_\Sigma(b'_1, \ldots, b'_{\hat{m}+1})$. If the new value is better or equal, accept $\{b'_1, \ldots, b'_{\hat{m}+1}\}$. Otherwise reject it.

6. Proceed with step (4) until no improvement could be made for 10 000 iterations.

The described algorithm repeatingly finds the same maximum value for $H_\Sigma(b_1, \ldots, b_{\hat{m}+1})$. One reason that prevents the algorithm from getting trapped in a local maximum is the fact that permanently an entire new set of boundaries is proposed by $b_{k+1} = r \cdot (f_i^{(k)} + f_i^{(k+1)}), k \in \{1, \ldots, \hat{m} - 1\}$ from $r \sim \beta(5, 5)$ for all neighboring frequency pairs f_i^k and $f_i^{(k+1)}$. Here, i is chosen randomly among all possible MCMC samples. This procedure has the following significance. One could argue that a good set of boundaries could be obtained by computing the mean or median of the individual frequencies over all MCMC samples that share the same rank of the sorted n sets $f_i^{(1)} \leq f_i^{(2)} \leq \ldots \leq f_i^{(\hat{m})}, \forall i \in \{1, \ldots, n\}$. The boundaries could then be derived in the same way as above. However, the presence of blendoids disorders the ranking and yields bad estimates. In contrast, among the many individual MCMC samples, it is very likely to find representatives that serve ideally for maximising $H_\Sigma(b_1, \ldots, b_{\hat{m}+1})$.

In order to speed up the computation one should start the algorithm with a subset of MCMC samples then subsequently accumulate more until all MCMC samples are incorporated. After having obtained the intervals

$$\{I_1 = [b_1, b_2], I_2 =]b_2, b_3], \ldots, I_{\hat{m}} =]b_{\hat{m}}, b_{\hat{m}+1}]\} : \max H_\Sigma(b_1, \ldots, b_{\hat{m}+1}),$$

that maximise Eq. 3.20, the number of samples possible to be allocated to sinusoids by means of the frequency intervals is at a maximum.

The allocation based on minimising the posterior expected loss with respect to Eq. 3.21 is merely based on the frequency parameters but in return is not computational expensive which makes its use feasible. This, however, is just the first step in the allocation procedure and further allocation steps are required that include the information of the amplitudes. We aim for determining allocations for all sinusoidal components based on the preallocations possible by the interval separation technique.

Preallocation scheme by means of the intervals $I_1, \ldots, I_{\hat{m}}$

If the allocation of the kth sinusoidal component $\boldsymbol{a}_i^{(k)}$ in the ith MCMC sample is non-ambiguous, it is allocated to the corresponding interval. In the ambiguous cases it is necessary to decide which of the candidates that could possibly occupy a particular interval is more likely to do so. From a physical point of view it is clear that a sinusoidal component $\boldsymbol{a}_i^{(k)}$ which carries the highest energy contribution $[(A_i^{(k)})^2 + (B_i^{(k)})^2]/2$ among those candidates is to be preferred. The higher the energy, the smaller is the expected discrepancy between frequency and corresponding interval. It is in the nature of blendoids to have a low energy contribution because otherwise they would maintain enough evidence to be fully present in model $\hat{\mathcal{M}}$. In contrast a sinusoid with high energy contribution and thus higher evidence is expected to reveal accurate estimates.

All remaining components within the ith MCMC sample get temporarily labeled zero to identify no sensible preallocation by the interval separation. The labeling of the individual sinusoidal components is structured in such a way that a sinusoid that is labeled k belongs to interval I_k that mainly covers it. Let us define $L(i,j)$ as the label for the jth sinusoidal component of the ith MCMC samples, $j \in \{1, \ldots, \hat{m}\}, i \in \{1, \ldots, n\}$ by

$$L(i,j) = \begin{cases} 1 & \text{if} f_i^{(j)} \in I_1 \text{ and } E_i^{(j)} > E_i^{(l)} \forall l \in \{1, \ldots, \hat{m}\} \setminus \{j\} : f_i^{(l)} \in I_1, \\ 2 & \text{if} f_i^{(j)} \in I_2 \text{ and } E_i^{(j)} > E_i^{(l)} \forall l \in \{1, \ldots, \hat{m}\} \setminus \{j\} : f_i^{(l)} \in I_2 \\ \vdots & \vdots \\ k & \text{if} f_i^{(j)} \in I_k \text{ and } E_i^{(j)} > E_i^{(l)} \forall l \in \{1, \ldots, \hat{m}\} \setminus \{j\} : f_i^{(l)} \in I_k \\ 0 & \text{otherwise, (identifies no preallocation)} \end{cases} \tag{3.22}$$

where $E_i^{(j)} = [(A_i^{(j)})^2 + (B_i^{(j)})^2]/2$ is the energy contribution of the jth sinusoidal component of the ith MCMC sample. The following Fig. 3.2 exemplifies the preallocation of Fig. 3.1.

By the procedure described above the allocation of the major part of the components can be achieved but there are cases in which more information from the amplitudes needs to be exploited. The interval separation is abrupt

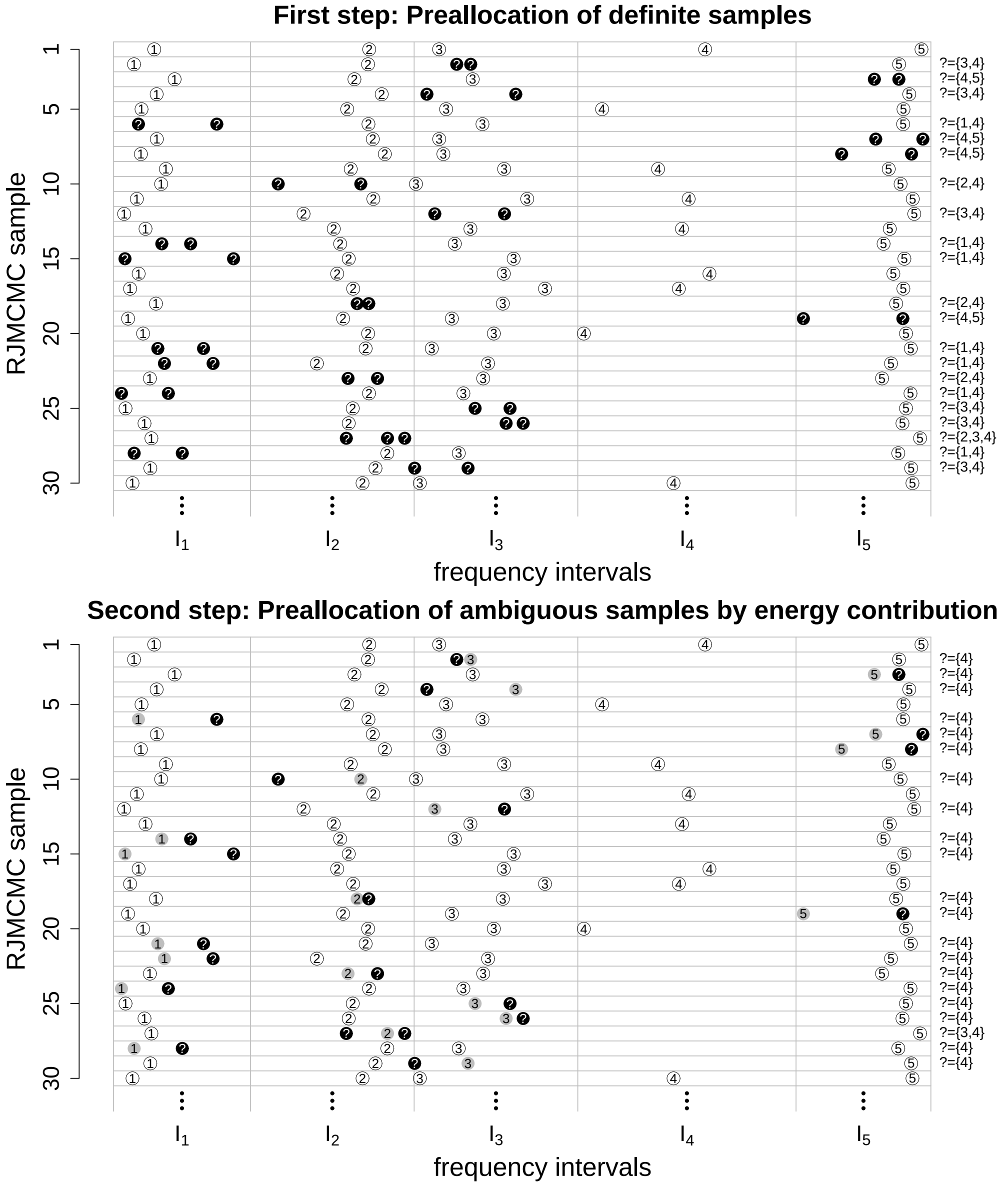


Figure 3.2: The two steps of preallocating the sinusoidal components. In the first step, the non-ambiguous samples are allocated to the corresponding intervals that maximise Eq. 3.20. In a second step, those candidates of the ambiguous samples that have the highest energy contribution are allocated to the corresponding intervals (gray discs). The others are not preallocated and receive an extra label (black discs).

and does not allow density overlaps. Neighboring intervals usually tend to share samples as there can be no clear frequency threshold drawn between the two distributions. For example in Fig. 3.2 in the 27th MCMC sample the sinusoids labelled $\{2, 3, 4\}$ fall into interval I_2. Blendoids operate within the entire frequency band with huge overlaps in frequency with other sinusoids.

With the preallocation information at hand and the fact that all sinusoidal components within an MCMC sample must belong to different sinusoids as they determine the likelihood of the model under consideration, a further classification step is possible. This problem can be approached by imputing that the posterior distribution is a mixture model, as described in the following section.

3.4.4 Further classification using a mixture model approximation

A preallocation by the interval technique delivers abundant information about the means and covariance matrix of the parameters of the individual sinusoids. The posterior distribution of the parameters of the isolated sinusoids present typical multivariate normal distributions of all parameters. Only the correlation between the amplitudes of sinusoids that are close in frequency results in a skewed posterior distributions in the marginal densities of the amplitudes. The marginal posterior densities of the frequencies, however, are approximately Gaussian [16]. The idea is to model the posterior distribution by a Gaussian mixture model which is assumed to be a sum of multivariate normal distributions. The parameters of the multivariate normal components can be estimated by the samples that could be successfully preallocated. The estimation of this Gaussian mixture model then allows for classifying the sinusoidal components by means of the multivariate distribution that incorporates all parameters.

The difficulty is to ascribe an appropriate distribution to a blendoid. A blendoid that appears within the model under consideration represents multiple sinusoids with low evidence. However, its existence must result from one sinusoid that has the highest evidence among those that justify its existence. The interest is drawn to that particular sinusoid that shares the major proportion of the probability mass which resides within its allocated frequency interval. The remaining probability mass must be ascribed to the weaker sinusoids that are expected to be found across the entire frequency band. A separation of a blendoid into all its components is hardly feasible due to the very low evidence of the remaining components and the best approximation to such a distribution is a uniform distribution on the entire frequency band and a normal distribution on the amplitudes. Estimates for mean and variance of the normal distribution can be obtained from the samples that could not be preallocated clearly by the interval technique. In the example that is displayed in Fig. 3.2 those samples are identified as solid black circles. They are almost solely from the blendoid and the amplitudes can be expected to be very small.

The weight between the distribution that embodies the major part of the sinusoid and the remaining blend can easily be estimated by the proportion between samples that could be preallocated and those that could not. Theoretically, every sinusoid can have a positive proportion of a blend of other sinusoids within a model but in most cases this proportion is almost zero when the evidence for the sinusoid is strong. All weights can be determined from the preallocation.

The parameter estimation of the mixture model components is done as follows. Let

$$n_k = \sum_{(i,j):L(i,j)=k} 1 \tag{3.23}$$

be the number of possible preallocations to interval I_k and $\boldsymbol{a}_i^{(j)}=(A_i^{(j)}, B_i^{(j)}, f_i^{(j)})^T$ the parameter vector of the jth sinusoidal component if the ith MCMC sample. In the following, the cth element of vector $\boldsymbol{a}_i^{(j)}$ will be referred to as $\boldsymbol{a}_i^{(j),c}$, $c \in \{1, 2, 3\}$ and indicates the amplitudes A, B, and the frequency f, respectively. This yields the mean vector

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{(i,j):L(i,j)=k} \boldsymbol{a}_i^{(j)} \tag{3.24}$$

and the empirical covariance matrix with elements

$$\hat{\Sigma}_k^{(c_1,c_2)} = \frac{1}{n_k - 1} \sum_{(i,j):L(i,j)=k} (\boldsymbol{a}_i^{(j),c_1} - \hat{\mu}_k^{(c_1)})(\boldsymbol{a}_i^{(j),c_2} - \hat{\mu}_k^{(c_2)}) \tag{3.25}$$

where c_1 and c_2 are indicating the different elements of the vectors. The sinusoidal component of sinusoid labeled k is expected to follow a multivariate normal distribution in the mixture model with

$$S_k(\boldsymbol{a}) = \frac{1}{(2\pi)^{3/2}\sqrt{\mathrm{Det}(\hat{\Sigma}_k)}} \exp\left(-\frac{1}{2}(\boldsymbol{a} - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (\boldsymbol{a} - \hat{\mu}_k)\right) \tag{3.26}$$

The remaining sinusoidal components without preallocation do therefore most likely belong to a blendoid. Estimates are needed for the distribution that can be expected from the remaining blend proportion. It is spread equally likely over the entire frequency band and therefore needs to be uniformly distributed in frequency. A covariance matrix could easily be derived but assuming uncorrelated amplitude samples is a more sensible choice. All required information can be derived from Eq. 3.24 and Eq. 3.25 for $k = 0$ indicating the label that indicates no successful preallocation. For the parameter vector $\boldsymbol{a} = (A, B, f)^T$ we obtain

$$P_k(\boldsymbol{a}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_0^{(A)}} e^{-\frac{1}{2}\left(\frac{A-\mu_0^{(A)}}{\hat{\sigma}_0^{(A)}}\right)^2} \frac{1}{\sqrt{2\pi}\hat{\sigma}_0^{(B)}} e^{-\frac{1}{2}\left(\frac{B-\mu_0^{(B)}}{\hat{\sigma}_0^{(B)}}\right)^2} \cdot 2 \tag{3.27}$$

as a suitable distribution we expect from the blend proportion of a sinusoid.

Finally the terms $S_k(\boldsymbol{a})$ (main sinusoid proportion) and $P_k(\boldsymbol{a})$ (blend proportion) need to be combined. The weights $w_k = n_k/n$ are easy to obtain from the n_k in Eq. 3.23. The $\hat{m}$ distribution components for the mixture model are hence

$$M_k(\boldsymbol{a}) = w_k P_k(\boldsymbol{a}) + (1 - w_k) S_k(\boldsymbol{a}) \tag{3.28}$$

with $k \in \{1, \ldots, \hat{m}\}$.

The final step remains to allocate the sinusoidal components based on the information of the approximated mixture model distribution conditioned on the permutational restrictions. The fact that the sinusoidal components within each MCMC sample are bound to be allocated to different components of the approximated mixture model distribution highly reduces the possible allocations to the different components within the mixture model. That makes the method very robust despite using a multivariate normal distribution for modelling the sinusoidal components and a coarse model for the blended components by Eq. 3.27. It is therefore important to adhere in the final classification to the permutational restrictions. From there, the classification procedure follows the scheme described below in order to circumvent double classifications of sinusoids within an MCMC sample to the same distribution component of the mixture model:

1. For the ith MCMC sample vector $\boldsymbol{a}_i$ with its $l \in \{1, \ldots, \hat{m}\}$ components $\boldsymbol{a}^{(l)}$, a $\hat{m} \times \hat{m}$-matrix is computed with elements $M_k(\boldsymbol{a}_i^{(l)}), k \in \{1, \ldots, \hat{m}\}$ (Eq. 3.28) in the kth row and lth column.
2. Find the largest element $(k, l) : \max(M_k(\boldsymbol{a}_i^{(l)}))$ in the matrix denoting the highest probability for the lth sinusoidal component of the ith MCMC sample to belong to the kth mixture model component.
3. Label the lth sinusoidal component with k.
4. Nullify all elements of the matrix that belong to row k or column l in order to invalidate them for further use.
5. Proceed with step (2) until all elements are nullified.
6. Proceed with step (1) until all MCMC samples are processed.

This algorithm allocates $\hat{m}$ unique labels to each of the sinusoidal component within each MCMC sample and we are left with n samples for each of the $\hat{m}$ sinusoids within model $\mathcal{M}_{\hat{m}}$. This allows for deriving estimates for the parameters of each sinusoid on the basis of the n samples. In presence of a blendoid, care has to be taken due to the multimodal structure. Since it represents multiple sinusoids from higher model orders which all have a share in the evidence of appearing in the model under consideration, it is necessary to identify the highest proportion that can be attributed to a single sinusoid. This proportion is obviously covered by its separation interval. A more sophisticated approach to identify that proportion by comparing the empirical quantiles of the marginal posterior distribution of the frequency to those of a normal distribution will be described in the next section Sec. 3.5 by means of a simulated example data set. We will see that any further separation of blendoids into its components is futile.

3.5 Simulation results

An artificial data set of $1,000$ samples at time points $t \in \{0, \ldots, 999\}$ containing 100 random sinusoids in Gaussian noise was created. The Cartesian amplitude coefficients were chosen randomly from a $\mathrm{N}(0,1)$ distribution and the frequency values from a uniform distribution with range $[0, 0.5]$. The results are presented with dimensionless units. The signal is embedded within white noise with standard deviation $\sigma = 0.6$. The following Fig. 3.3 displays the signal-to-noise ratios (SNRs) of all 100 sinusoids. The SNR with units in dB, as it is used here, is defined by $\mathrm{SNR} = 10 \log_{10}[(A_i^2 + B_i^2)/(2\sigma^2)]$ with A_i and B_i being the amplitude coefficients for sinusoid i.

The uniform prior range for m was set to $\{0, 1, 2, \ldots, M = 60\,000\}$. From the 2.05×10^8 iterations that the Markov chain has run for the first 5×10^6 were considered as burn-in and discarded. The chain was then thinned by storing every 1 000th iteration. The MCMC simulation was implemented in C on a 2.8 GHz Intel P4 PC and took about 50 hours to run. Fig. 3.4(a) gives the histogram of the marginal posterior model probabilities obtained by the reversible jump algorithm. As each model $\mathcal{M}_m$ is characterised by a different noise level σ_m, also the marginal posterior distributions of the noise standard deviations have been plotted in Fig. 3.4. Note that σ_m decreases with higher model order m since a model comprising more sinusoids accounts for more of the available power. The highest posterior model probability is observed for model $\mathcal{M}_{99}$ with 99 sinusoids and $\hat{\sigma}_{99} = 0.604$ ($p_{0.573<\hat{\sigma}_{99}<0.637} = 95\%$) since the weakest sinusoid (at $f = 0.4178$ with $\mathrm{SNR} = -22.5$ dB) could not

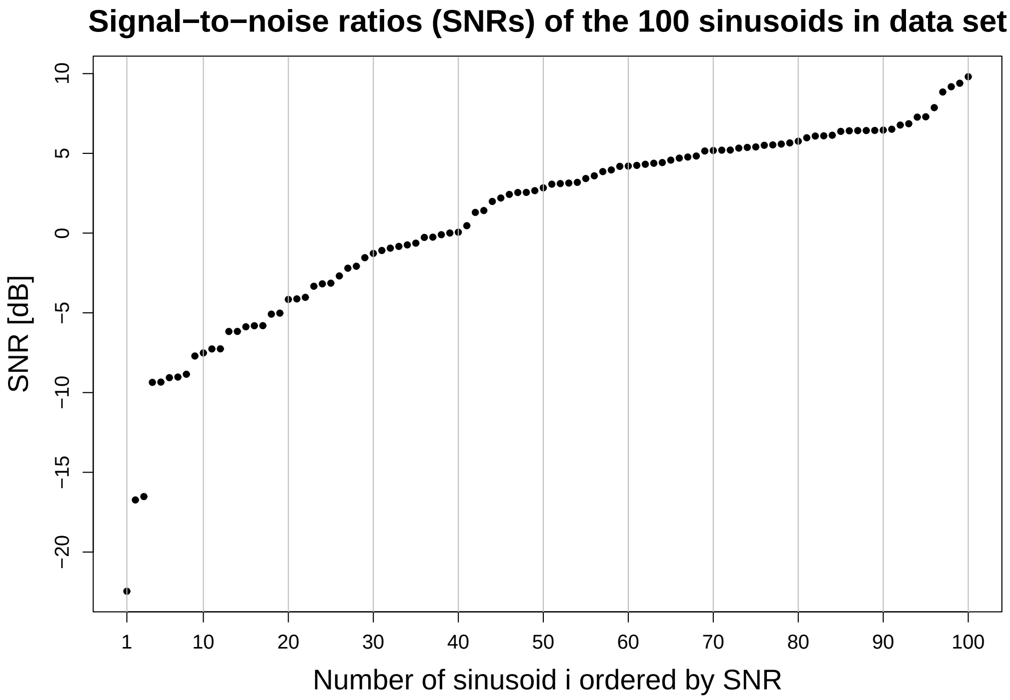


Figure 3.3: Signal-to-noise ratios (SNRs) for each individual sinusoid used in the artificial data set, ordered by SNR.

be retrieved and contributes to the overall noise level. All subsequent results presented here are based solely on MCMC samples corresponding to model $\mathcal{M}_{99}$. The superscripts will be omitted and denote the parameter vector of model $\mathcal{M}_{99}$ by $(A_1, B_1, f_1, \ldots, A_{99}, B_{99}, f_{99}, \sigma_{99}^2)$.

The initial step for post processing the MCMC samples is the separation into frequency intervals, as described in Sec. 3.4.3. The following table Tab. 3.1 shows the 99 intervals that maximise the number of single occupancies.

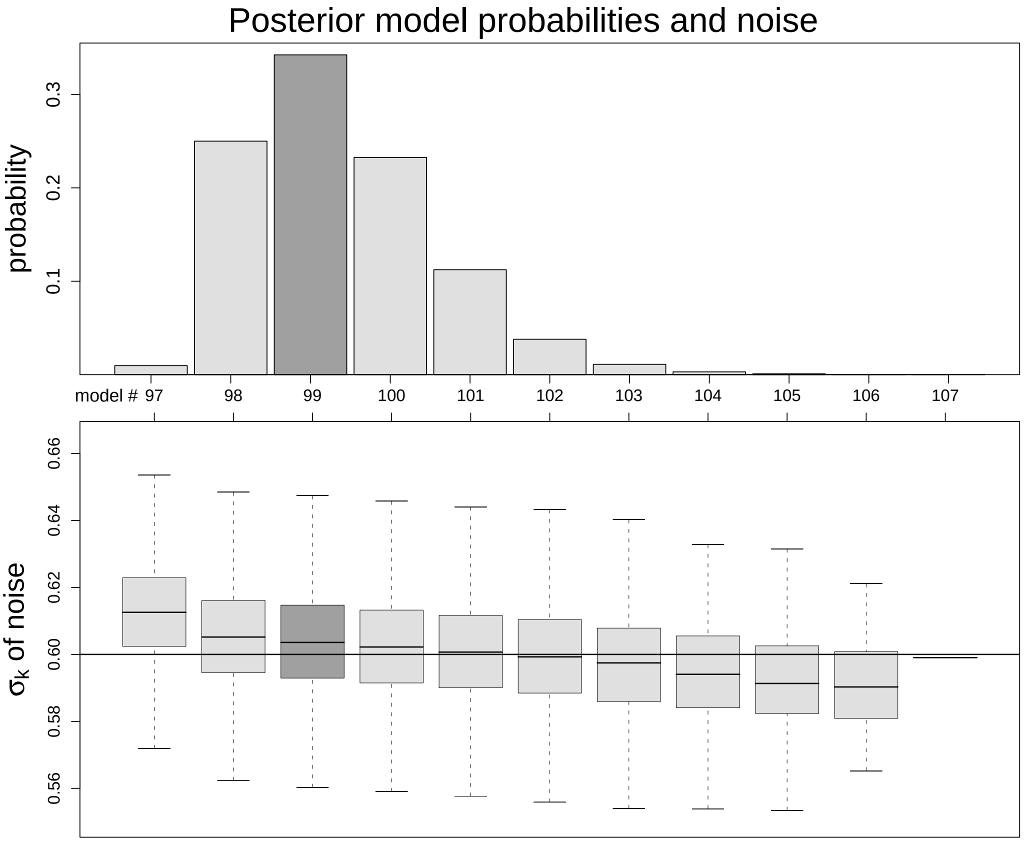


Figure 3.4: The upper plot shows the posterior model probabilities obtained from the analysis. The model corresponding to 99 sinusoids (dark shaded) has highest posterior probability. Each model has a different noise level. The bottom plot shows the corresponding estimated noise standard deviations and their posterior probability intervals. The vertical line indicates the real noise level.

Table 3.1: Frequency intervals that maximise the number of single occupancies in model $\mathcal{M}_{99}$ and percentages of individual occupancy numbers.

ID	frequency	no. of sinusoids			
k	range	$\hat{H}_k^{(0)}$	$\hat{H}_k^{(1)}$	$\hat{H}_k^{(2)}$	$\hat{H}_k^{(3)}$
1	[0.0000, 0.0027]		99.51%	0.49%	
2	] 0.0027, 0.0101]		97.91%	2.09%	
3	] 0.0101, 0.0111]		99.23%	0.77%	

continued...

Table 3.1 (continued)

k	frequency range	$\hat{H}_k^{(0)}$	$\hat{H}_k^{(1)}$	$\hat{H}_k^{(2)}$	$\hat{H}_k^{(3)}$
4	] 0.0111, 0.0122]		99.28%	0.72%	
5	] 0.0122, 0.0238]		99.07%	0.93%	
6	] 0.0238, 0.0251]		99.90%	0.10%	
7	] 0.0251, 0.0338]		99.39%	0.60%	0.01%
8	] 0.0338, 0.0374]		99.31%	0.69%	
9	] 0.0374, 0.0384]		99.92%	0.08%	
10	] 0.0384, 0.0525]		95.65%	4.34%	0.01%
11	] 0.0525, 0.0572]		99.51%	0.49%	
12	] 0.0572, 0.0589]		99.82%	0.18%	
13	] 0.0589, 0.0599]		99.91%	0.09%	
14	] 0.0599, 0.0690]		98.23%	1.77%	
15	] 0.0690, 0.0780]		99.33%	0.67%	
16	] 0.0780, 0.0875]		99.07%	0.93%	
17	] 0.0875, 0.0885]		99.72%	0.28%	
18	] 0.0885, 0.0901]		99.53%	0.47%	
19	] 0.0901, 0.0927]		99.83%	0.17%	
20	] 0.0927, 0.0999]		98.51%	1.49%	
21	] 0.0999, 0.1086]		99.39%	0.61%	
22	] 0.1086, 0.1103]		99.45%	0.55%	
23	] 0.1103, 0.1128]		99.83%	0.17%	
24	] 0.1128, 0.1155]		99.86%	0.14%	
25	] 0.1155, 0.1167]		99.92%	0.08%	
26	] 0.1167, 0.1318]		98.37%	1.62%	0.01%
27	] 0.1318, 0.1335]		99.67%	0.33%	
28	] 0.1335, 0.1381]		99.48%	0.52%	
29	] 0.1381, 0.1437]		99.50%	0.50%	
30	] 0.1437, 0.1581]		98.37%	1.63%	
31	] 0.1581, 0.1600]		99.77%	0.23%	
32	] 0.1600, 0.1610]		99.57%	0.43%	
33	] 0.1610, 0.1635]		99.59%	0.41%	
34	] 0.1635, 0.1643]		99.92%	0.08%	
35	] 0.1643, 0.1701]		99.12%	0.88%	
36	] 0.1701, 0.1736]		99.60%	0.40%	
37	] 0.1736, 0.1808]		99.68%	0.32%	
38	] 0.1808, 0.1845]		99.70%	0.30%	
39	] 0.1845, 0.1938]		99.30%	0.70%	
40	] 0.1938, 0.1980]		99.58%	0.42%	

continued...

Table 3.1 (continued)

k	frequency range	$\hat{H}_k^{(0)}$	$\hat{H}_k^{(1)}$	$\hat{H}_k^{(2)}$	$\hat{H}_k^{(3)}$
41	] 0.1980, 0.1993]		99.65%	0.35%	
42	] 0.1993, 0.2144]		94.25%	5.75%	
43	] 0.2144, 0.2267]		95.58%	4.42%	
44	] 0.2267, 0.2288]		99.89%	0.11%	
45	] 0.2288, 0.2345]		99.67%	0.33%	
46	] 0.2345, 0.2364]		99.72%	0.28%	
47	] 0.2364, 0.2425]		97.80%	2.20%	
48	] 0.2425, 0.2460]		99.12%	0.88%	
49	] 0.2460, 0.2488]		99.76%	0.24%	
50	] 0.2488, 0.2523]		99.30%	0.70%	
51	] 0.2523, 0.2537]		99.90%	0.10%	
52	] 0.2537, 0.2567]		99.78%	0.22%	
53	] 0.2567, 0.2639]	0.03%	97.72%	2.25%	
54	] 0.2639, 0.2648]	0.03%	99.31%	0.66%	
55	] 0.2648, 0.2669]		99.73%	0.27%	
56	] 0.2669, 0.2779]		99.36%	0.64%	
57	] 0.2779, 0.2816]		99.75%	0.25%	
58	] 0.2816, 0.2833]		99.82%	0.18%	
59	] 0.2833, 0.2887]		96.58%	3.42%	
60	] 0.2887, 0.2896]	0.04%	99.71%	0.24%	
61	] 0.2896, 0.2913]		98.26%	1.74%	0.00%
62	] 0.2913, 0.2951]		99.16%	0.84%	
63	] 0.2951, 0.3027]		99.64%	0.36%	
64	] 0.3027, 0.3214]		97.38%	2.58%	0.04%
65	] 0.3214, 0.3228]		99.61%	0.39%	
66	] 0.3228, 0.3247]		97.02%	2.90%	0.09%
67	] 0.3247, 0.3317]	4.62%	94.80%	0.58%	
68	] 0.3317, 0.3332]		99.80%	0.20%	
69	] 0.3332, 0.3362]	0.09%	99.63%	0.28%	
70	] 0.3362, 0.3408]	86.19%	13.72%	0.09%	
71	] 0.3408, 0.3510]		99.66%	0.34%	
72	] 0.3510, 0.3573]		99.10%	0.90%	
73	] 0.3573, 0.3615]		98.26%	1.74%	
74	] 0.3615, 0.3632]		98.91%	1.09%	
75	] 0.3632, 0.3661]		99.42%	0.58%	
76	] 0.3661, 0.3698]		99.64%	0.36%	
77	] 0.3698, 0.3717]		99.02%	0.98%	

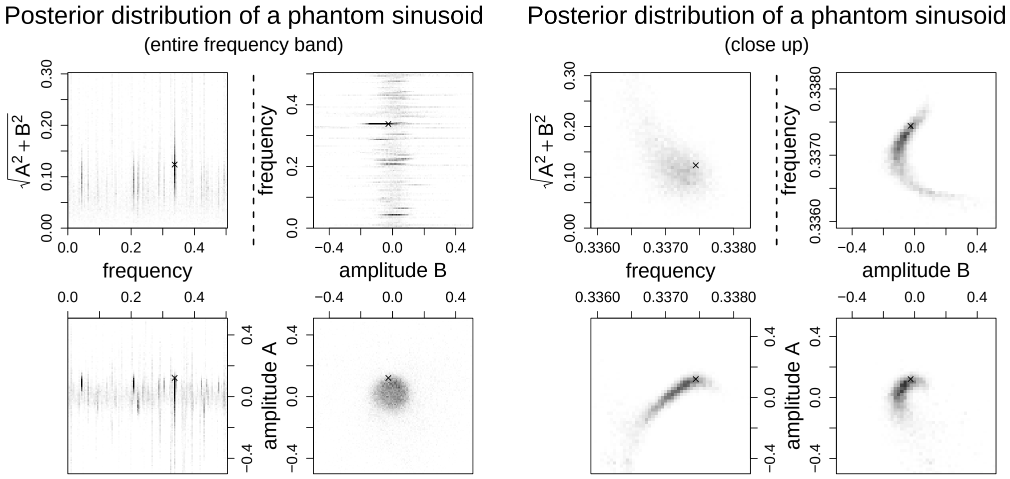
continued...

Table 3.1 (continued)

k	frequency range	$\hat{H}_k^{(0)}$	$\hat{H}_k^{(1)}$	$\hat{H}_k^{(2)}$	$\hat{H}_k^{(3)}$
78	] 0.3717, 0.3762]		98.75%	1.25%	
79	] 0.3762, 0.3765]		99.98%	0.02%	
80	] 0.3765, 0.3811]		99.25%	0.75%	
81	] 0.3811, 0.3871]		99.29%	0.71%	
82	] 0.3871, 0.3898]		97.31%	2.69%	
83	] 0.3898, 0.3904]		99.87%	0.13%	
84	] 0.3904, 0.3979]		99.01%	0.99%	
85	] 0.3979, 0.4060]		99.53%	0.47%	
86	] 0.4060, 0.4216]		97.28%	2.72%	0.01%
87	] 0.4216, 0.4301]		96.94%	3.06%	
88	] 0.4301, 0.4326]		99.17%	0.82%	0.01%
89	] 0.4326, 0.4509]		97.65%	2.35%	
90	] 0.4509, 0.4580]		99.49%	0.51%	
91	] 0.4580, 0.4629]		99.77%	0.23%	
92	] 0.4629, 0.4812]		97.67%	2.33%	
93	] 0.4812, 0.4817]		99.93%	0.07%	
94	] 0.4817, 0.4902]	0.54%	96.68%	2.75%	0.03%
95	] 0.4902, 0.4912]	0.71%	98.45%	0.84%	
96	] 0.4912, 0.4922]		99.92%	0.08%	
97	] 0.4922, 0.4968]		98.63%	1.37%	
98	] 0.4968, 0.4971]	0.00%	99.95%	0.05%	
99	] 0.4971, 0.5000]		99.44%	0.56%	

The total number of single occupancies (Eq. 3.20) that could be achieved was $H_\Sigma(b_1, \ldots, b_{\hat{m}+1}) = 97.156$ with upper bound 99. This implies that dispersed sinusoid fractions totalling in 1.844 sinusoids can not be specified by the sole separation in frequency intervals. Almost all values of $\hat{H}_k^{(1)}$ are close to 100%. Sparse single occupancies $\hat{H}_k^{(1)}$ can either mean deficiency of occupancies ($\hat{H}_k^{(0)} > 0$) or multiple occupancies of the interval ($\hat{H}_k^{(l)} > 0, l \geq 2$). Overlaps in frequency of neighboring intervals usually cause marginal absence of single occupancies. However, interval I_{70} reveals explicit values for $\hat{H}_{70}^{(0)}$ and $\hat{H}_{70}^{(1)}$. Only 13.79% of the sinusoids probability mass is occupied by interval I_{70} with no compensation that could possibly be found in the neighboring intervals. It becomes clear that this candidate is highly likely to be a blendoid.

The further relabeling is accomplished by the method described in Sec. 3.4.4 and allows for taking a closer look at the marginal posterior distribution of the sinusoid that is expected to be a blendoid. Fig. 3.5 shows two views of this sinusoid. One that displays the entire frequency range and a second that focuses on that part of the frequency band that contains the core of the actual sinusoid that encompasses ≈ 13% of its probability mass. The features are quite striking. The samples occupy the entire frequency band with some accumulations, one of which lies within interval I_{70} which encompasses the most probability mass. This is displayed in the close-up. The distribution of the amplitudes show a very low energy contribution far below the noise level. This sinusoid is definitely very close to the limit of being eligible to exist within model $\mathcal{M}_{99}$.



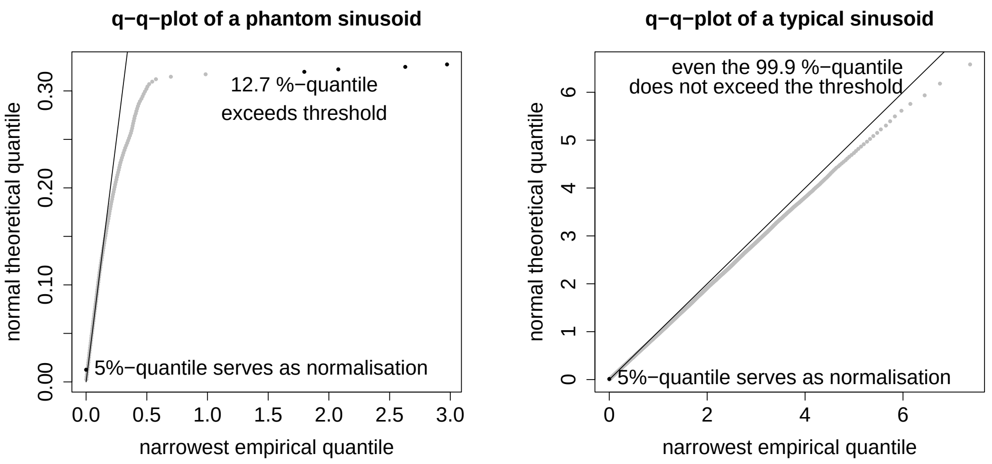
(a)

(b)

Figure 3.5: Marginal MCMC posterior distribution for different parameter pairs of the blendoid. The left plot covers the entire frequency band whereas the right plot focuses on the frequency band that covers the most probability mass that can be assigned to an actual sinusoid. The crosses indicate the corresponding true parameter values.

Estimations of such entities are difficult as they consist of mixtures. The main interest is to identify the sinusoidal component that appears most often in MCMC samples of the model. The technique that has been applied here in order to recover this main proportion continues to assume that the marginal posterior density of the frequency follows approximatively a normal distribution [16]. In order to identify the significant probability mass of a blendoid that can be attributed to a single sinusoid it is obvious to compare its empirical quantiles to those of a normal distribution, similar to a q-q-plot. Naturally, this technique requires the p-quantiles that span the narrowest frequency interval $[f_a, f_b]$ and therefore cover the frequency range with the highest density. This means, if $p(f|\boldsymbol{d})$ is the marginal posterior density of a sinusoid, then we are looking for the narrowest interval $[f_a, f_b]$ such that $p(f_a < f < f_b|\boldsymbol{d}) = p$. We aim for finding the largest p-quantile that follows a normal distribution. The normalisation is achieved by the $p = 0.05$ quantile that is chosen to intersect the bisection in a graph that displays empirical quantiles and quantiles of the normal distribution. The reason why such a small quantile has to be chosen for a normalisation is the possible small evidence. On the other hand it would not make sense to speak of evidence below such a limit of 5%. The following Fig. 3.6 demonstrates the technique by applying it to two different sinusoids, one of which is the blendoid that is covered by interval I_{70} in figure Fig. 3.5. The other sinusoid is the one covered by interval I_{98} and constitutes a typical sinusoid with strong evidence. Alternatively, a series of Kolmogorov-Smirnov tests or Shapiro-Wilk tests could be applied but the quantile comparison works superbly and enables a graphically verification.

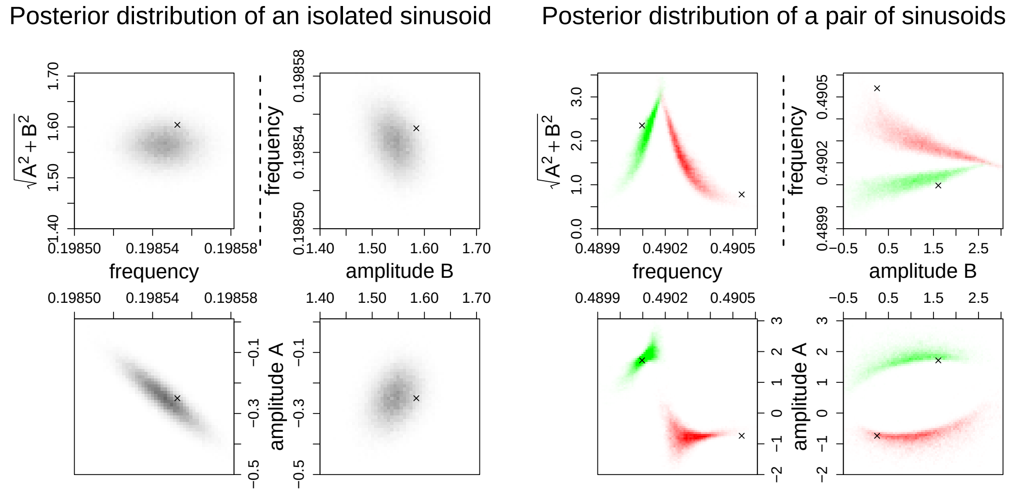
Any further separation of a blendoid is inane. If candidates of weaker sinusoids are to be examined, it is more appropriate to look at higher model orders but we are more and more likely to identify what would be considered as noise in models of lower order. In model $\mathcal{M}_{100}$, for example, the same sinusoid that corresponds to the one that is labeled $k = 70$ in $\mathcal{M}_{99}$ maintains more evidence by occupying 27.8% probability mass within its assigned interval but it still has to be considered as a blendoid. Instead another blendoid appears in $\mathcal{M}_{100}$ making the allocation even harder. This additional sinusoid in model $\mathcal{M}_{100}$ is not the weakest of the 100 in the data set (see Fig. 3.3) as we would expect. Care has to be taken when looking at models of higher order as we do not necessarily see more signals but also noise that is mistakenly identified as belonging to a signal. It is imperative



(a) (b)

Figure 3.6: Revelation of a blendoid (a) by comparing the smallest empirical quantiles to the corresponding quantiles of a normal distribution. Exploiting the condition that the marginal posterior density of the frequency follows very well a normal distribution it is possible to obtain the frequency range that encompasses the largest proportion of probability mass that can be allocated to an actual sinusoid. As a comparison, the q-q-plot for a typical sinusoid is opposed in (b). Its marginal posterior of the frequency naturally reveals almost 100% of its probability mass to accord to a normal distribution. The threshold is chosen one standard deviation from the actual normal quantile.

to mind the posterior model probabilities that favor model $\mathcal{M}_{99}$. These data only reveal the existence of one distinct blendoid and it is important to focus on the remaining other 98 sinusoids and the estimation of their parameters. Therefore, in Fig. 3.7, two plots have been compiled that show examples of ordinary sinusoids. This figure aims for comparing the posterior distribution of well separated sinusoids and those that have close neigbours in frequency. The marginal MCMC posterior distribution of a well separated sinusoid reveals a multivariate normal shape with little correlation in its recovered amplitudes whereas sinusoids that are close in frequency to others reveal stronger correlations yielding larger credibility areas. The red coloured sinusoid in Fig. 3.7(b) corresponds to the q-q-plot of Fig. 3.6(a). Note that despite the interactions between the sinusoids that are reflected in the am-



(a) (b)

Figure 3.7: Marginal MCMC posterior distribution of a single, isolated sinusoid (a) and of a pair of correctly separated sinusoids that are close in frequency (b). The closer the difference in frequency between two sinusoids the more correlation there is in their recovered amplitudes yielding larger credibility areas. The crosses indicate the corresponding true parameter values.

plitudes, the marginal posterior density of the frequency still conforms to a normal distribution.

The comparison of Bayesian methods to classical approaches like the Schuster periodogram turns out to be difficult. In a Bayesian analysis we obtain samples of the posterior distributions for each sinusoid whereas the Schuster periodogram merely yields an energy density over the entire frequency band including noise without any proper separation. As large number of sinusoids can not be all displayed in full detail, the following plots will focus on special sections of frequency bands.

The comparison of the classical Schuster periodogram is accomplished by contrasting it to the joint MCMC posterior distribution of the Cartesian amplitude $\hat{A}_i^{(99)} = [(A_i^{(99)})^2 + (B_i^{(99)})^2]^{1/2}$ and the frequency. The Cartesian way of expressing the sinusoids preserves the full information about the strength of a signal but neglects its phase that is of less interest here. The energy contribution of a sinusoid is proportional to the squared Cartesian amplitude and therefore allows for a better comparison with the periodogram that displays the energy density on the same axis in a log scale. The following plots display the 90%- and 50%- MCMC posterior contour areas of the joint MCMC posterior distribution of the Cartesian amplitude $\hat{A}_i^{(99)}$ and the frequency $f_i^{(99)}$.

In order to illustrate the versatile features of the entire posterior distribution, a gradual magnification in two steps has been compiled. The Schuster periodogram is contrasted on the same frequency scales. In Fig. 3.8, the attention is drawn to a pair of sinusoids that are separated by a frequency difference of 0.00047, less then one half of a Nyquist step of $1/N = 0.001$. It is the same sinusoid pair on which Fig. 3.7(b) was focused on. Further, the blendoid from Fig. 3.5 that is close-by is also included in this plot. Note, that due to the fact that the sensible probability mass that can be assigned to that sinusoid is $\approx 13\%$, only low credibility intervals can be given. A 90% credibility area would result in a huge interval that would stretch over the entire frequency band and would not make any sense.

Another approach to display the Bayesian results in a way that is more in the nature of a periodogram is to derive a Bayesian power spectral density according to Jaynes [16]. Fig. 3.8 indicates the uncertainty about both the true frequency f and the true power $A^2 + B^2$ of a signal. Jaynes derives a Bayesian power spectral density by combining the estimation of a power in the spectrum line and the estimation of the frequencies present [16]. For clarity we omit the index i indicating the ith sinusoid. For each sinusoid with parameters $\{A, B, f\}$, the power spectral density of the signal can be estimated from the product of the conditional expected power

$$(N/2)\mathrm{E}(A^2 + B^2|f, \boldsymbol{d}, m) = (N/2)\int (A^2 + B^2)p(A, B|f, \boldsymbol{d}, m)dAdB \quad (3.29)$$

of a line at frequency f and the posterior probability $p(f|\boldsymbol{d}, m)$ of the frequency [16]. In practice the Bayesian power spectral density

$$\hat{S}(f) = p(f|\boldsymbol{d}, m)(N/2)\mathrm{E}(A^2 + B^2|f, \boldsymbol{d}, m) \quad (3.30)$$

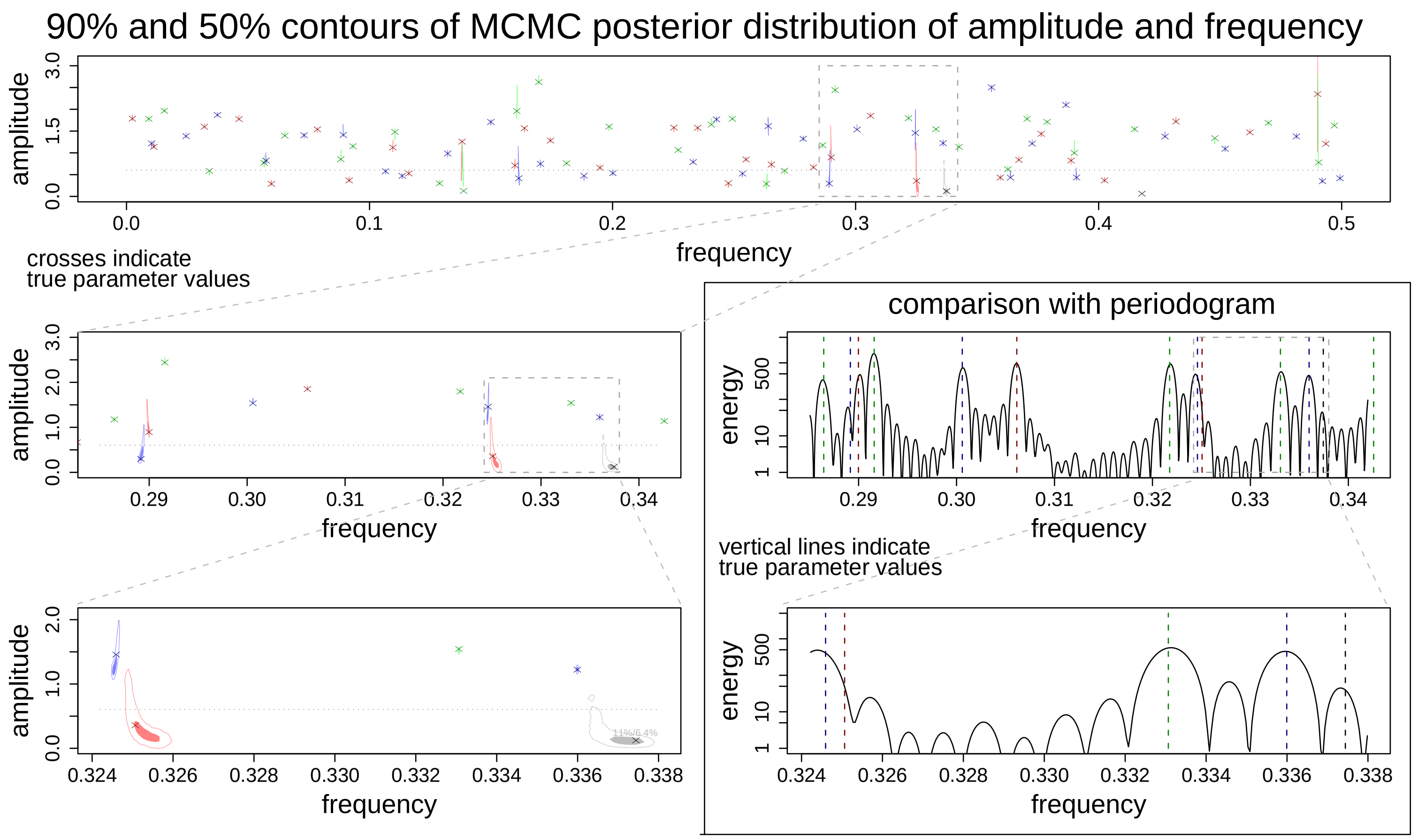


Figure 3.8: Gradually magnified area of a sinusoid pair with frequency separation 0.00047 and one that has a very low signal-to-noise ratio. Bayesian estimates and classical Schuster periodogram are opposed. The two sinusoids can clearly be separated. The closer the frequency gap the more correlation there is in their recovered amplitudes yielding larger confidence ranges for the estimates. The blendoid with the low signal-to-noise ratio only reveals 13% probability mass. This sinusoid vanishes in lower model orders due to insufficient evidence. The dotted horizontal lines indicate the noise level σ_{99}.

needs to be computed by binning the MCMC samples into frequency bins of width δf. We obtain $\hat{S}(f)\delta f$ as an estimate of the expected power within the frequency bin $[f, f+\delta f]$, multiplied by the probability that its frequency is in the interval $[f, f+\delta f]$. The Bayesian power spectral density is the power per unit frequency. Since the samples of the sinusoids are available in separated form, δf can be individually derived for each sinusoid. A number of 20 bins leads to the best results for covering the frequency range of a sinusoid.

In Fig. 3.9, this affords the compilation of $\hat{S}(f)$, the Schuster periodogram, and the theoretical spectral lines from the true values. Two different axes for the periodogram and the Bayesian power spectral density $\hat{S}(f)$ need to be chosen. The ordinate of the periodogram displays the energy and the ordinate of $\hat{S}(f)$ shows the energy per unit frequency. The theoretical spectral density would consist of delta functions of formally zero width and hence infinite density. Therefore the adjusted heights of the theoretical spectral lines in this context indicate the energy contribution $N(A^2 + B^2)/2$ (left axis) of a sinusoid with amplitudes A and B over the observation period N. The following Fig. 3.9 displays the same frequency section that was objective in Fig. 3.8.

The next two figures Fig. 3.10 and Fig. 3.11 focus on an area that contains three close sinusoids. The periodogram can hardly separate any of the three sinusoids due to their close proximity in frequency.

Finally, the spotlight is set to a region of three sinusoids displayed in Fig. 3.12. Here, the difference in frequency is 0.00044 between the left two sinusoids and is the closest within the set of 100 sinusoids. Again the periodogram fails to separate them. The interaction between those sinusoids is obvious in the MCMC posterior distribution.

Figures Fig. 3.8, 3.10, and 3.12 as well as Fig. 3.9, 3.11, and 3.13 demonstrate the superiority of the Bayesian approach to resolve signals and give credibility about the estimates. Pairs of sinusoids that are separated by a frequency difference less than one half of a Nyquist step $1/N$ can clearly be separated, while there is no evidence for distinct peaks in the periodogram that would allow for any separation. Naturally, the marginal posterior densities of different parameter pairs reveal the influence of small frequency

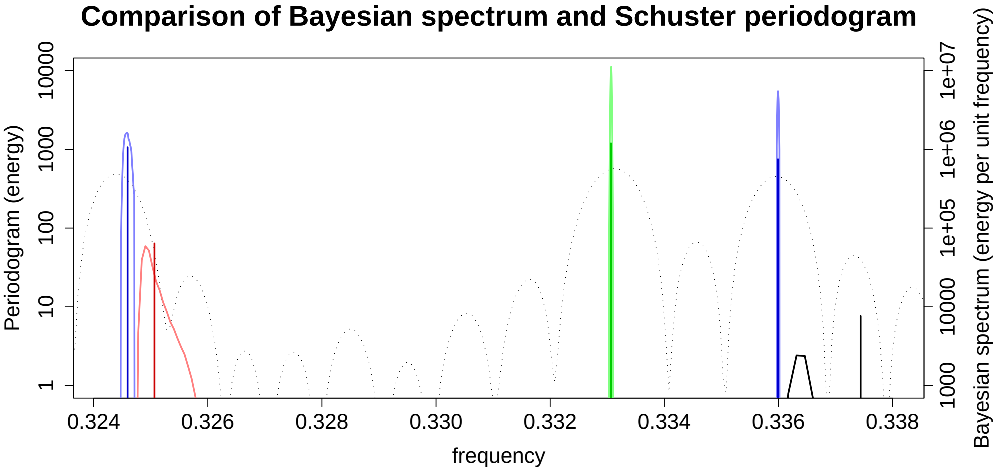


Figure 3.9: Comparison of true spectral lines (vertical lines), Bayesian spectral density estimate $\hat{S}(f)$ (coloured for each sinusoid), and classical Schuster periodogram (black, dotted). The periodogram uses the left ordinate (energy) whereas the Bayesian power spectrum density uses the right ordinate (energy per unit frequency). The different colours used for the Bayesian spectrum estimate underline the separability. Since the infinite small widths of the theoretical spectral lines would yield an infinite height, the height here corresponds to the left axis with energy contribution $N(A^2+B^2)/2$ of a particular sinusoid with its amplitudes A and B over the observation period N. The particular frequency band under consideration corresponds to the highest magnification in Fig. 3.8.

proximity on the accuracy of their parameter estimations. The posterior confidence areas are much larger for close sinusoids.

The concordance in Fig. 3.8 of the frequency estimates for the well-separated sinusoids, however, from both the Bayesian posterior estimates and the periodogram maxima, are quite striking. As stated in [17], “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”. However, the periodogram peaks are significantly wider than the Bayesian estimates, and are clearly sub-optimal for closely spaced sinusoids. In fact, the posterior probability density for the frequency is the exponent of the

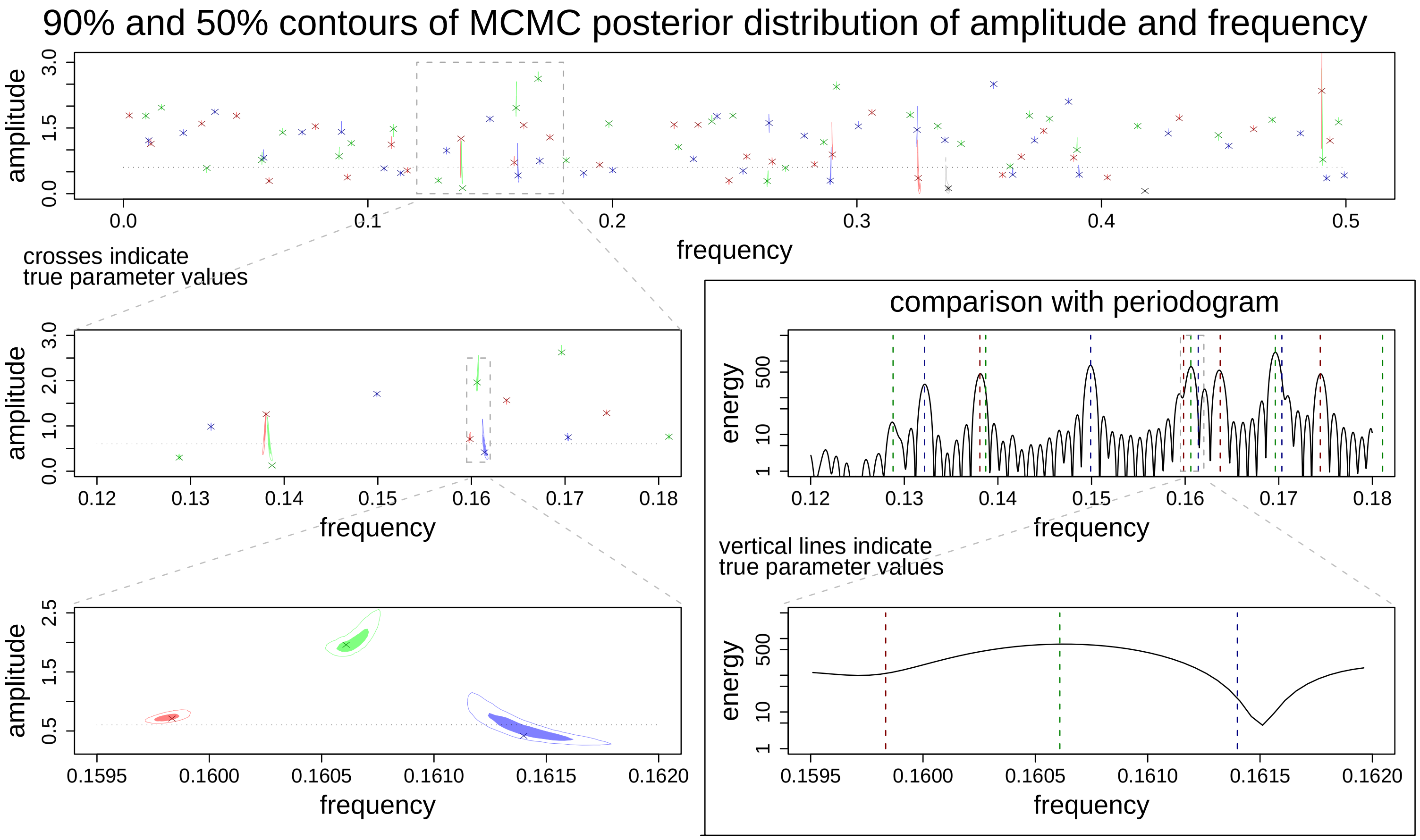


Figure 3.10: Gradually magnified area of a group of four sinusoids. All sinusoids can clearly be separated with the Bayesian technique whereas the Schuster periodogram does not reveal even a hint of separation.

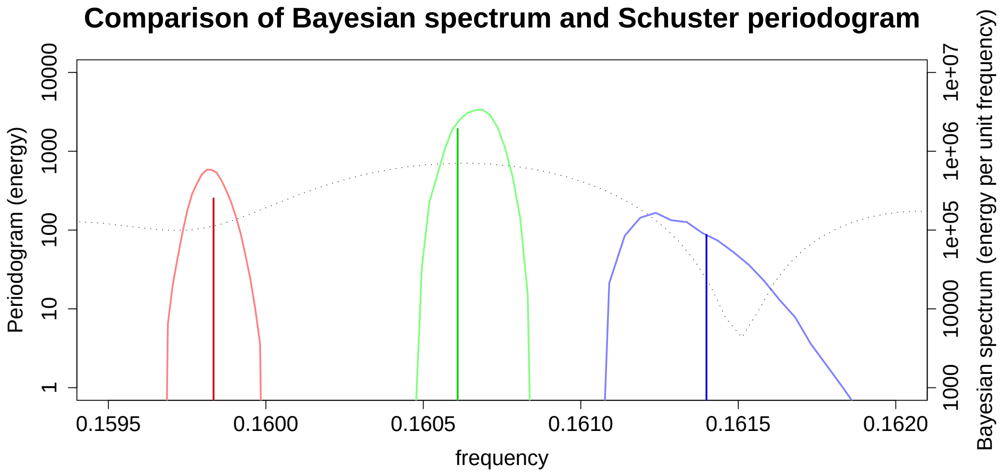


Figure 3.11: Comparison of true spectral lines (vertical lines), Bayesian spectral density estimate $\hat{S}(f)$ (coloured for each sinusoid), and classical Schuster periodogram (black, dotted). The periodogram uses the left ordinate (energy) whereas the Bayesian power spectrum density uses the right ordinate (energy per unit frequency). The different colours used for the Bayesian spectrum estimate underline the separability. Since the infinite small widths of the theoretical spectral lines would yield an infinite height, the height here corresponds to the left axis with energy contribution $N(A^2+B^2)/2$ of a particular sinusoid with its amplitudes A and B over the observation period N. The particular frequency band under consideration corresponds to the highest magnification in Fig. 3.10.

ratio of the periodogram $C(f)$ and the noise variance σ^2 [17]. The Bayesian approach therefore takes account of the noise variance in the estimation process, which explains why the confidence regions are significantly narrower than the periodogram peaks.

As to the energy contribution, the results of the periodogram are sensitive to the discreteness of the data and to the finite observation time. This, however, does not reflect the possible energy contribution of the real signals. By contrast, the Bayesian estimates of the amplitudes are honest by yielding large confidence intervals for sinusoids close in frequency, due to the diminishing impact on the possible linear combinations of the sinusoidal amplitudes on the likelihood when the frequencies of a sinusoid pair are getting close.

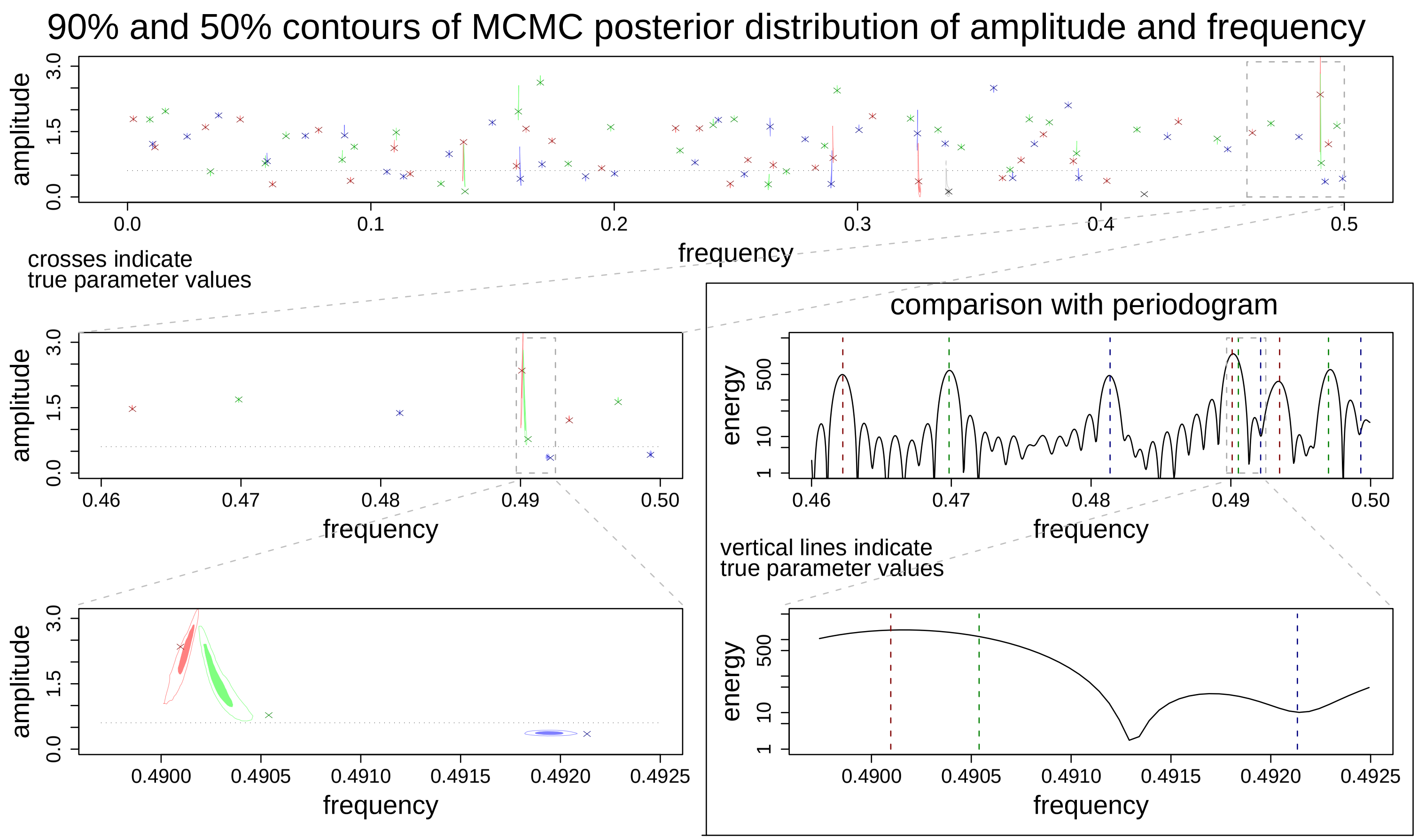


Figure 3.12: Gradually magnified area of a sinusoid pair with frequency separation 0.00044. Bayesian estimates and classical Schuster periodogram are opposed. The two sinusoids can clearly be separated with the Bayesian approach.

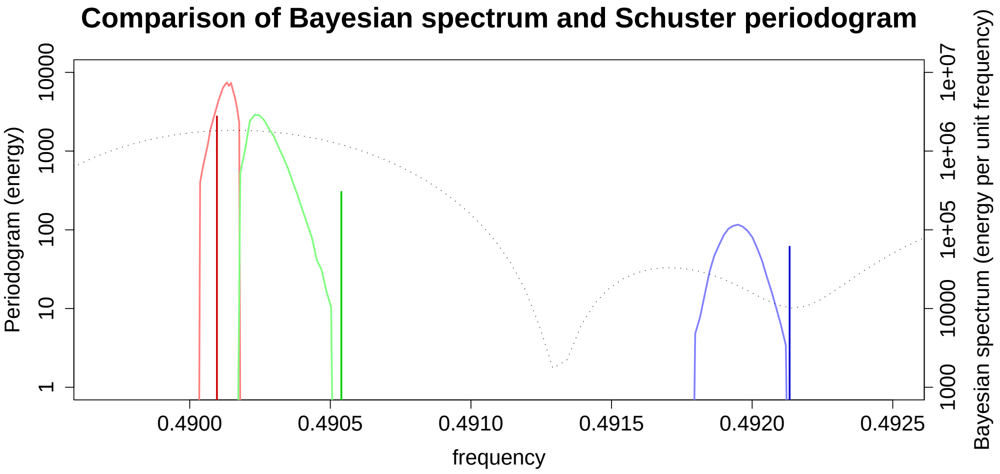


Figure 3.13: Comparison of true spectral lines (vertical lines), Bayesian spectral density estimate $\hat{S}(f)$ (coloured for each sinusoid), and classical Schuster periodogram (black, dotted). The periodogram uses the left ordinate (energy) whereas the Bayesian power spectrum density uses the right ordinate (energy per unit frequency). The different colours used for the Bayesian spectrum estimate underline the separability. Since the infinite small widths of the theoretical spectral lines would yield an infinite height, the height here corresponds to the left axis with energy contribution $N(A^2+B^2)/2$ of a particular sinusoid with its amplitudes A and B over the observation period N. The particular frequency band under consideration corresponds to the highest magnification in Fig. 3.12.

An energy could, in any case, also be derived by calculating amplitudes from the orthonormal model functions with respect to the discrete sampling times. The amplitudes of the orthonormal model are, however, linear combinations of the amplitudes of the original model. It would therefore not be possible to differentiate the energy contribution between the individual sinusoids but only for the entire pool of sinusoids.

After presenting the results in graphical form we will address the derivation of 95% credibility intervals and means of the marginal posterior distribution of the individual paramters of all sinusoids. As described above, in case of the blendoid, only those samples are involved in the estimation that contribute to the highest density area and follow a normal distribution according to the corresponding normal quantiles.

Table 3.2: Posterior means and 95% posterior credibility intervals (p.c.i.) of the frequency and the Cartesian amplitude $\hat{A} = \sqrt{A^2 + B^2}$

ID	frequency f			amplitude $\hat{A}$		
	post. mean	95% p.c.i.	true value	post. mean	95% p.c.i.	true value
1	0.002401	[0.002385 , 0.002417]	0.00241	1.792	[1.737 , 1.848]	1.785
2	0.009174	[0.009148 , 0.009203]	0.009174	1.758	[1.702 , 1.819]	1.778
3	0.010364	[0.010319 , 0.010415]	0.010351	1.181	[1.108 , 1.258]	1.218
4	0.011232	[0.011178 , 0.011287]	0.01125	1.172	[1.095 , 1.248]	1.138
5	0.015561	[0.015546 , 0.015577]	0.015566	1.978	[1.925 , 2.032]	1.965
6	0.024497	[0.024476 , 0.024519]	0.024493	1.398	[1.346 , 1.452]	1.384
7	0.032043	[0.032023 , 0.032062]	0.032037	1.609	[1.555 , 1.66]	1.599
8	0.034083	[0.034024 , 0.034143]	0.0341	0.532	[0.481 , 0.586]	0.588
9	0.037421	[0.037405 , 0.037437]	0.037428	1.869	[1.815 , 1.921]	1.869
10	0.046294	[0.046277 , 0.04631]	0.046301	1.794	[1.74 , 1.848]	1.776
11	0.056684	[0.056602 , 0.056766]	0.056637	0.796	[0.688 , 0.916]	0.771
12	0.057371	[0.057291 , 0.057446]	0.057375	0.858	[0.754 , 0.981]	0.822
13	0.059567	[0.059453 , 0.059668]	0.059614	0.302	[0.246 , 0.357]	0.289
14	0.065115	[0.065107 , 0.065151]	0.065137	1.431	[1.378 , 1.485]	1.395
15	0.073066	[0.073045 , 0.073087]	0.073068	1.391	[1.337 , 1.442]	1.405
16	0.078537	[0.078517 , 0.078557]	0.078534	1.51	[1.456 , 1.563]	1.54
17	0.088227	[0.088132 , 0.088323]	0.088196	0.899	[0.778 , 1.027]	0.854
18	0.089169	[0.089108 , 0.089228]	0.089206	1.484	[1.368 , 1.614]	1.412
19	0.091638	[0.091538 , 0.091741]	0.091624	0.386	[0.33 , 0.444]	0.368
20	0.093188	[0.093157 , 0.093219]	0.093182	1.143	[1.087 , 1.196]	1.153
21	0.10662	[0.106564 , 0.106676]	0.106575	0.567	[0.513 , 0.62]	0.578
22	0.109615	[0.109548 , 0.109682]	0.109597	1.152	[1.032 , 1.277]	1.122
23	0.110463	[0.11041 , 0.110517]	0.110435	1.437	[1.314 , 1.562]	1.48
24	0.113416	[0.113347 , 0.113486]	0.113425	0.458	[0.406 , 0.513]	0.473
25	0.116146	[0.11609 , 0.116204]	0.116151	0.536	[0.482 , 0.589]	0.528
26	0.128804	[0.128706 , 0.128893]	0.1288	0.315	[0.262 , 0.367]	0.3
27	0.132189	[0.132158 , 0.13222]	0.132173	0.978	[0.923 , 1.029]	0.985
28	0.13841	[0.137734 , 0.138057]	0.138077	0.827	[0.393 , 1.224]	1.26
29	0.139265	[0.138198 , 0.138655]	0.138702	0.63	[0.224 , 1.048]	0.127
30	0.149901	[0.149882 , 0.149919]	0.149908	1.685	[1.633 , 1.742]	1.71
31	0.159811	[0.159726 , 0.159903]	0.159834	0.729	[0.634 , 0.837]	0.71

continued...

Table 3.2 (continued)

ID	post. mean	95% p.c.i.	true value	post. mean	95% p.c.i.	true value
32	0.160639	[0.160539 , 0.160749]	0.160609	2.024	[1.769 , 2.408]	1.96
33	0.161454	[0.161129 , 0.161662]	0.161399	0.57	[0.244 , 0.982]	0.417
34	0.163758	[0.163735 , 0.16378]	0.163745	1.533	[1.476 , 1.589]	1.567
35	0.169631	[0.16961 , 0.169652]	0.169631	2.629	[2.557 , 2.709]	2.624
36	0.170299	[0.170221 , 0.170374]	0.170333	0.745	[0.667 , 0.82]	0.748
37	0.174449	[0.174426 , 0.174472]	0.174433	1.295	[1.243 , 1.349]	1.283
38	0.181103	[0.181064 , 0.181141]	0.181092	0.762	[0.706 , 0.813]	0.761
39	0.188245	[0.18817 , 0.188314]	0.188235	0.418	[0.365 , 0.471]	0.476
40	0.194852	[0.194808 , 0.194897]	0.194801	0.663	[0.611 , 0.717]	0.659
41	0.198545	[0.198523 , 0.198567]	0.198553	1.562	[1.509 , 1.619]	1.604
42	0.200118	[0.200055 , 0.200182]	0.200122	0.543	[0.485 , 0.598]	0.534
43	0.225233	[0.225212 , 0.225254]	0.225225	1.535	[1.48 , 1.589]	1.581
44	0.227021	[0.226992 , 0.227051]	0.227036	1.053	[1 , 1.107]	1.066
45	0.233209	[0.233169 , 0.233248]	0.233183	0.799	[0.744 , 0.855]	0.789
46	0.235017	[0.234997 , 0.235038]	0.235017	1.549	[1.494 , 1.605]	1.575
47	0.240659	[0.24064 , 0.240678]	0.240658	1.708	[1.654 , 1.764]	1.647
48	0.242772	[0.242754 , 0.242791]	0.242768	1.759	[1.706 , 1.816]	1.768
49	0.247676	[0.247535 , 0.247804]	0.247675	0.261	[0.206 , 0.315]	0.306
50	0.249276	[0.249257 , 0.249294]	0.249282	1.804	[1.748 , 1.858]	1.781
51	0.253415	[0.253334 , 0.253497]	0.253464	0.499	[0.444 , 0.553]	0.526
52	0.254885	[0.254838 , 0.254934]	0.25492	0.845	[0.789 , 0.897]	0.849
53	0.263388	[0.263168 , 0.263747]	0.263286	0.317	[0.176 , 0.518]	0.289
54	0.264067	[0.263976 , 0.264161]	0.264064	1.573	[1.346 , 1.755]	1.614
55	0.265334	[0.26524 , 0.265433]	0.265354	0.731	[0.652 , 0.811]	0.733
56	0.270753	[0.270699 , 0.270809]	0.270747	0.563	[0.509 , 0.616]	0.591
57	0.278465	[0.278442 , 0.278488]	0.278442	1.325	[1.272 , 1.379]	1.323
58	0.282634	[0.282587 , 0.282681]	0.282635	0.68	[0.626 , 0.732]	0.668
59	0.286422	[0.286393 , 0.286451]	0.286442	1.169	[1.115 , 1.224]	1.176
60	0.289235	[0.288972 , 0.289491]	0.289156	0.467	[0.207 , 0.913]	0.299
61	0.289874	[0.289759 , 0.290012]	0.289982	1.052	[0.807 , 1.503]	0.895
62	0.291595	[0.291579 , 0.291614]	0.29159	2.426	[2.373 , 2.492]	2.441
63	0.300583	[0.300563 , 0.300602]	0.300587	1.587	[1.534 , 1.639]	1.535
64	0.306152	[0.306136 , 0.306167]	0.306152	1.855	[1.804 , 1.91]	1.851
65	0.321766	[0.321749 , 0.321784]	0.321757	1.807	[1.752 , 1.862]	1.796
66	0.324486	[0.324487 , 0.324679]	0.324594	1.354	[1.119 , 1.885]	1.458
67	0.325252	[0.324839 , 0.325712]	0.325065	0.311	[0.078 , 0.783]	0.357

continued...

Table 3.2 (continued)

ID	post. mean	95% p.c.i.	true value	post. mean	95% p.c.i.	true value
68	0.333066	[0.333046 , 0.333086]	0.333066	1.515	[1.461 , 1.567]	1.545
69	0.335994	[0.335969 , 0.33602]	0.335993	1.224	[1.171 , 1.279]	1.224
70	*0.336835*	*[0.336137 , 0.337699]*	*0.337441*	*0.179*	*[0.013 , 0.612]*	*0.124*
71	0.342607	[0.342581 , 0.342633]	0.342583	1.155	[1.102 , 1.209]	1.137
72	0.355929	[0.35591 , 0.355936]	0.355927	2.442	[2.39 , 2.5]	2.503
73	0.359537	[0.359465 , 0.359614]	0.359536	0.43	[0.369 , 0.482]	0.435
74	0.362626	[0.362543 , 0.362705]	0.362654	0.627	[0.559 , 0.689]	0.623
75	0.36371	[0.363611 , 0.363805]	0.363729	0.495	[0.427 , 0.557]	0.431
76	0.367163	[0.367126 , 0.367201]	0.367197	0.859	[0.804 , 0.911]	0.839
77	0.370567	[0.370548 , 0.370584]	0.370551	1.801	[1.747 , 1.858]	1.779
78	0.372636	[0.372608 , 0.372661]	0.372636	1.238	[1.181 , 1.293]	1.213
79	0.376349	[0.376326 , 0.376373]	0.376368	1.416	[1.362 , 1.469]	1.437
80	0.37881	[0.378791 , 0.37883]	0.3788	1.683	[1.629 , 1.737]	1.713
81	0.386544	[0.386528 , 0.38656]	0.386543	2.1	[2.047 , 2.158]	2.099
82	0.388661	[0.388596 , 0.38872]	0.388649	0.81	[0.747 , 0.877]	0.824
83	0.390008	[0.389929 , 0.390094]	0.389981	1.082	[0.962 , 1.235]	0.999
84	0.390833	[0.39066 , 0.390986]	0.39092	0.484	[0.359 , 0.628]	0.435
85	0.402412	[0.402327 , 0.402492]	0.402414	0.377	[0.325 , 0.431]	0.368
86	0.414801	[0.414781 , 0.414821]	0.414792	1.543	[1.491 , 1.6]	1.544
-	not identified		0.417803	not identified		0.064
87	0.427305	[0.427282 , 0.427327]	0.427314	1.405	[1.351 , 1.463]	1.374
88	0.431817	[0.431795 , 0.431829]	0.431809	1.74	[1.683 , 1.792]	1.72
89	0.447796	[0.447773 , 0.447819]	0.447792	1.26	[1.207 , 1.312]	1.339
90	0.45205	[0.452022 , 0.452077]	0.452068	1.087	[1.033 , 1.139]	1.093
91	0.462216	[0.462196 , 0.462236]	0.462238	1.495	[1.443 , 1.548]	1.469
92	0.469839	[0.469822 , 0.469857]	0.46984	1.667	[1.613 , 1.719]	1.688
93	0.481361	[0.481339 , 0.481382]	0.48136	1.382	[1.329 , 1.436]	1.377
94	0.490106	[0.490024 , 0.490177]	0.490097	2.056	[1.05 , 2.939]	2.35
95	0.49023	[0.490195 , 0.490428]	0.490539	1.506	[0.615 , 2.489]	0.779
96	0.491939	[0.491827 , 0.492056]	0.492133	0.366	[0.31 , 0.424]	0.349
97	0.493483	[0.493455 , 0.493512]	0.493489	1.236	[1.181 , 1.293]	1.208
98	0.496985	[0.496966 , 0.497004]	0.496988	1.673	[1.62 , 1.727]	1.627
99	0.499312	[0.499243 , 0.499376]	0.499299	0.43	[0.372 , 0.487]	0.417

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

We have seen that Fig. 3.8, 3.10, and also 3.12, reveal strong interference between very closely 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 in further detail, a series of simulations have been conducted with two sinusoids gradually approaching each other in frequency. The results of these simulations are presented in Fig. 3.14 and Fig. 3.15. The ability of the method to separate the signals depends on the signal-to-noise ratio, their relative phase and on observing time. The prior that is chosen for the amplitudes does not impact the detectability but the the model probabilities show slight variations in the model probabilities.

In the examples at hand, an observation time of $t \in \{0, \ldots, N-1\}$ with $N = 1000$ is used and the two sinusoids have a signal-to-noise ratio of either $\hat{A}/\sigma = 2$ or $\hat{A}/\sigma = 4$. The noise vector was drawn once from a standard normal distribution. All data sets with the sinusoid pairs at different frequency gaps use the same noise vector in order to maintain the comparability. Due to the huge computational effort this series of runs can not be repeated easily for several noise vectors in order to obtain credibility intervals for each frequency gap. The single points within each plot were obtained from a lengthy RJMCMC run over 5.5×10^6 iterations with a burn-in of 5×10^5 iterations. These plots merely serve as an example for giving an insight into the separation ability of the Bayesian approach under certain conditions.

The g-prior that is also used in all examples in this context is opposed to the application of a uniform prior on the amplitudes with prior range $[-10, 10]$. In total, twelve series of simulations were conducted with two different priors on the amplitudes, three different phase shifts $(0, \pi/2, \pi)$, and two different signal-to-noise ratios. During each set of simulations the frequency gap between the two sinusoids was increased by 1% of a $(1/N)$-step. (Recall that $N = 1\,000$ so that $1/N = 0.001$). In Fig. 3.14 the results displayed are based on the uniform prior whereas Fig. 3.15 reveals results based on the g-prior used in recent examples. All signal pairs have coequal strength. The left columns exhibit a scenario with a signal-to-noise ratio of $\hat{A}/\sigma = 2$, while the right columns deal with a signal-to-noise ratio of $\hat{A}/\sigma = 4$. The most striking and also anticipated feature that is revealed by

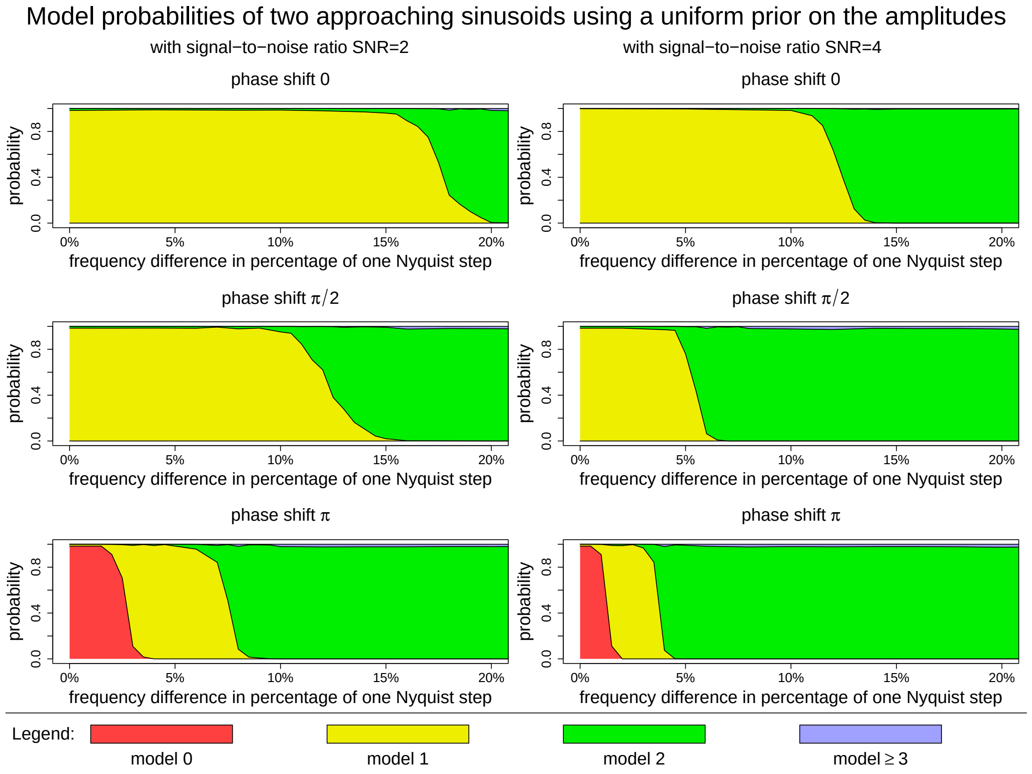


Figure 3.14: Model probabilities of a pair of sinusoids in Gaussian noise with varying difference in frequency and uniform prior on the amplitudes. The sinusoids on in the left and right column have a signal-noise-ratio of $\hat{A}/\sigma = 2$ and $\hat{A}/\sigma = 4$ respectively each. The three different rows show the different phase shifts $(0, \pi/2, \pi)$ between the sinusoid pairs. The uniform prior range for the amplitudes is $[-10, 10]$.

Fig. 3.14 and Fig. 3.15 is the fading separation ability with increasing phase shift. Best separation is achieved at a phase shift of π. For small frequency gaps this yields a cancellation of the sinusoids whereas for a phase shift of zero, a sinusoid can always be seen. Interestingly, the g-prior tends to prefer higher model orders. This is caused by the fact that the g-prior prefers

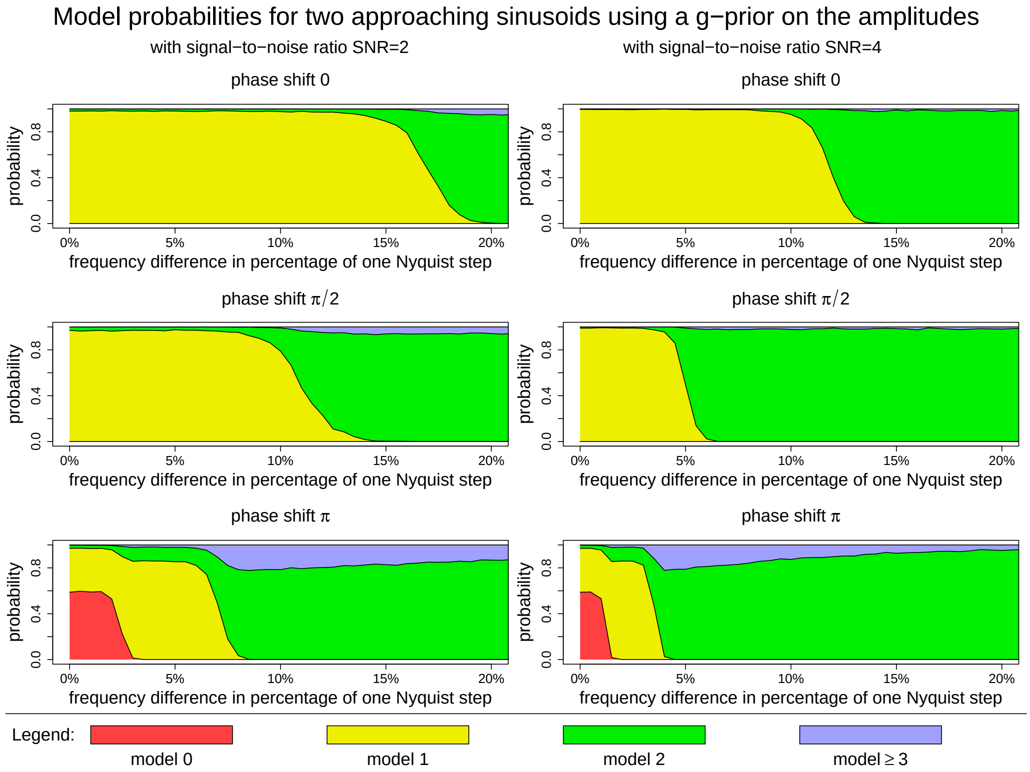


Figure 3.15: Model probabilities of a pair of sinusoids in Gaussian noise with varying difference in frequency and g-prior on the amplitudes. The sinusoids on in the left and right column have a signal-noise-ratio of $\hat{A}/\sigma = 2$ and $\hat{A}/\sigma = 4$ respectively each. The three different rows show the different phase shifts $(0, \pi/2, \pi)$ between the sinusoid pairs.

small amplitudes controlled by the variance determined by all sinusoidal amplitudes within the model. In this example the variance is only sustained by two sinusoids. When the two sinusoids are close in frequency and maintain a phase shift of π, then, for the limited observation time, those sinusoids annihilate each other and the variance of the g-prior obtains small values. In contrast a model with many sinusoids would yield more representative variances for the g-prior and this effect would diminish. Therefore, applying a g-prior results in a preference for a model that contains sinusoids with extreme small amplitudes. In model classes with many sinusoids, which this thesis is mainly focused on, this effect is not an issue.

What makes the g-prior so suitable is the fact that it prevents the amplitude values from obtaining huge values. When a pair of sinusoids come close in frequency, then, with respect to the limited observation time at discrete time points, more and more linear combinations of the amplitudes become likely that allow the amplitudes to obtain large values that are in fact inappropriate. This is a problem when applying a uniform prior to the amplitudes in which case the amplitudes have an abrupt cut-off at the end of their prior ranges. Intuitively it is clear that huge amplitudes are less likely and the information about the distribution of the amplitudes is provided by the other amplitudes. This information is exploited by using the g-prior. Actually, the more sinusoids there are present, the better the g-prior responds to the model class by its hyper parameter g.

3.6 Discussion

In this thesis a Bayesian approach has been presented to identifying a large number of unknown periodic signals in a set of noisy data. The reversible jump Markov chain Monte Carlo method developed here can be used to estimate the number of signals present in the data, their parameters, and the noise level. This method compares favourably with classical spectral techniques. This 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 motivation for this 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. This work can therefore be seen as a powerful new technique for identifying and characterising these signals in the LISA data stream.

It has been demonstrated that the simultaneous detection and estimation of sinusoids is possible also for a large unknown number of sinusoids. The label switching problem could be solved, by choosing frequency intervals that minimise the posterior expected loss, using a label-invariant loss function based on interval separation. It provides an excellent means to find the allocation of MCMC samples to individual sinusoids and allows us to draw meaningful inference of the MCMC output.

The work presented here is of course a highly simplified toy problem: The signal modulation is neglected due to LISA's orbit and beam pattern and there is no consideration of an appropriate data model for LISA. In addition, the signals from compact binary systems differ from simple sinusoids. Furthermore, with LISA we have to deal with six data streams from the laser beams between each of the three spacecraft that will deliver data which are expected to be dominated by frequency noise from its lasers. The six LISA data streams can be linearly combined to Time-Delay Interferometry (TDI) variables [88] in order to cancel out noise from the three lasers by appropriate time delays.

However, the purpose of this work is to 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 these MCMC methods. This will not be a trivial extension; in fact, the complexity of the situation has to be acknowledged. However, MCMC methods, like those presented here, have shown to 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. They are a general tool to perform a Bayesian spectrum analysis that has been optimised on working with a large number of signals.

Chapter 4

Conclusions

The Bayesian approach to gravitational radiation data analysis is a scientific basis for combining knowledge from different other sources like radio astronomy that can serve as prior information in the Bayesian analysis of gravitational radiation data. The possibility to make probability statements about parameters based on prior and sampling information are essential in this scientific context.

In addition, for high-dimensional problems the Bayesian approach provides the only viable solution particularly with regard to model selection on a large set of models. However, it has been shown that the Markov chain Monte Carlo framework provides the means of tackling such complex Bayesian analysis problems. The re-parametrisation in Chapter 2 is of vital importance and yields a more tractable parameter space that is more effectively sampled. The implementation of a Metropolis-Coupled Chain [31], although slowing down the computation, is necessary in order to avoid getting trapped in the countless local posterior modes. Different types of proposal distributions are required in order to be responsive to different areas of the posterior distribution. In this matter, the delayed rejection method [30] speeds up the burn-in process due to the fact that more stage 1 proposals are accepted leaving out unnecessary timid moves. The combination of these methods allow for a parameter estimation of a single signal described by a realistic model of a gravitational wave emitted by a pulsar. A modern standard desktop PC is sufficient to conduct the computation.

For the LISA data analysis challenge with its expected many thousands of signals, however, it appears that the demand on the hardware will be far higher, particularly for realistically modelled signals. The simplified problem that was discussed in Chapter 3 by using a single PC helps to explore on a small scale problems and potentials of Bayesian MCMC methods when facing many signals with signal confusion. It demonstrates that this enormous challenge must and can be countered by the Bayesian RJMCMC methodology, particularly against the background of the perpetual improvement in the computer hardware sector.

The RJMCMC method provides the framework that can handle the model selection between thousands of models. The label switching problem that is caused by the invariance of the likelihood under relabelling of the components exhibits a challenge in presence of thousands of signals. In order to draw inference from MCMC output using ergodic averaging, an appropriate relabelling technique is required. The fact that the parameters have a physical meaning in which a signal is identified by its frequency has been exploited here in order to circumvent the consideration of the possible permutations. The interval separation technique has been demonstrated to be highly efficient and able to cope with a large number of signals.

In a next step, the models need to be extended to data from the six LISA data streams which can be linearly combined to Time-Delay Interferometry (TDI) variables [88] in order to cancel out noise from the three lasers by appropriate time delays. In addition, the development of realistic models and suitable Bayesian MCMC methods will be an essential task for the next years before LISA will be launched.

Bibliography

[1] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. 1

[2] P. S. Laplace. *Theorie Analytiques des Probabilités.* Courcier, Paris, 1812. 1

[3] H. Jeffreys. *Theory of Probability.* Oxford University Press, Oxford, 3rd edition, 1961. (1st edition 1939). 1

[4] R. T. Cox. Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946. 1

[5] T. J. Loredo. The return of the prodigal: Bayesian inference for astrophysics. 1

[6] E. T. Jaynes. Bayesian methods: General background. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25. Cambridge University Press, 1986. 1

[7] W. R. Gilks, S. Richardson, and P. J. Spiegelhalter, D. J. Green, editors. *Markov chain Monte Carlo in practice.* Chapman and Hall, London, 1996. 1, 2.1, 2.4, 3.1, 3.2

[8] N. Christensen, R. Meyer, and A. Libson. A Metropolis-Hastings routine for estimating parameters from compact binary inspiral events with laser interferometric gravitational radiation data. *Classical and Quantum Gravity*, 21:317–330, 2004. 1, 3.1

[9] C. Röver, R. Meyer, and N. Christensen. Bayesian inference on compact binary inspiral gravitational radiation signals in interferometric data. *Classical and Quantum Gravity*, 23(15):4895–4906, August 2006. 1

[10] N. Christensen, R. J. Dupuis, G. Woan, and R. Meyer. Metropolis-Hastings algorithm for extracting periodic gravitational wave signals from laser interferometric detector data. *Physical Review D*, 70(2):022001–1–022001–7, 2004. 1, 2.1, 2.2, 2.4, 2.5.2, 2.6.3

[11] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995. 1, 3.1, 3.2, 3.3.1

[12] P. J. Green, N. L. Hjort, and S. Richardson, editors. *Highly Structured Stochastic Systems*. Oxford University Press, 2003. 1, 3.2

[13] C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Trans. on Signal Processing*, 47:2667, 1999. 1, 3.1, 3.2, 3.3.1, 3.3.2, 3.3.4, 3.4.1

[14] R. Umstätter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan. Bayesian modeling of source confusion in LISA data. *Physical Review D*, 72(2):022001, 2005. 1, 3.2, 3.4.2

[15] R. Umstätter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan. LISA source confusion: Identification and characterization of signals. *Classical and Quantum Gravity*, 22(18):S901–S911, 2005. 1

[16] E. T. Jaynes. *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*. D. Reidel, Dordrecht-Holland, 1987. Bayesian Spectrum and Chirp Analysis. 1, 3.3.1, 3.4.4, 3.5, 3.5, 3.5

[17] G. L. Bretthorst. *Lecture Notes in Statistics: Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, 1988. 1, 3.2, 3.2, 3.3.2, 3.5

[18] C. Cutler. Gravitational waves from neutron stars with large toroidal b fields. *Physical Review D*, 66(8):084025, 2002. 2.1

[19] L. Bildsten. Gravitational radiation and rotation of accreting neutron stars. *Astrophysical Journal*, 501(1):L89–L93, 1998. 2.1

[20] B. Willke et al. The GEO 600 gravitational wave detector. *Classical and Quantum Gravity*, 19(7):1377–1387, 2002. 2.1

[21] A. Abramovici, W. E. Althouse, R. W. P. Drever, Y. Gürsel, S. Kawamura, F. J. Raab, D. Shoemaker, L. Sievers, R. E. Spero, K. S. Thorne, R. E. Vogt, R. Weiss, S. E. Whitcomb, and M. E. Zucker. LIGO - The Laser-Interferometer-Gravitational-Wave-Observatory. *Science*, 256(5055):325–333, 1992. 2.1

[22] B. Caron et al. Status of the VIRGO experiment. *Nuclear Physics B - Proceedings Supplements*, 48(1):107–109, 1996. 2.1

[23] K. et al. Tsubono. Gravitational wave detection. In K. Tsubono, M.-K. Fujimoto, and K. Kurodo, editors, *Proceedings of TAMA Workshop, Saitama, Japan*, pages 183–191, Tokyo, 1997. Universal Academic Press. 2.1

[24] B. Abbott et al. Setting upper limits on the strength of periodic gravitational waves from PSR J1939+2134 using the first science data from the GEO 600 and LIGO detectors. *Physical Review D*, 69(8):082004–1–082004–16, 2004. 2.1, 2.2, 2.2, 2.5.1, 2.7

[25] P. Jaranowski, A. Krolak, and B. F. Schutz. Data analysis of gravitational-wave signals from spinning neutron stars: The signal and its detection. *Physical Review D*, 58(6):063001, 1998. 2.1

[26] P. Brady and Creighton T. Searching for periodic sources with LIGO. II. hierarchical searches. *Physical Review D*, 61(8):082001, 2000. 2.1

[27] J. Middleditch, J. A. Kristian, W. E. Kunkel, K.M. Hill, R. D. Watson, R. Lucinio, J. N. Imamura, T. Y. Steiman-Cameron, A. Shearer,

R. Butler, M. Redfern, and A.C. Danks. Rapid photometry of supernova 1987a: a 2.14 ms pulsar? *New Astronomy*, 5(5):243–283, 2000. 2.1, 2.7

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953. 2.1, 2.4, 3.1, 3.2

[29] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. 2.1, 2.4, 3.1, 3.2

[30] L. Tierney and A. Mira. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18:2507–2515, 1999. 2.1, 2.4.1, 3.2, 3.3.1, 4

[31] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas, editor, *Computing science and statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, VA, 1991. 2.1, 2.4.3, 2.4.4, 2.4.4, 4

[32] R. Umstätter, R. Meyer, R.J. Dupuis, J. Veitch, G. Woan, and N. Christensen. Estimating the parameters of gravitational waves from neutron stars using an adaptive MCMC method. *Classical and Quantum Gravity*, 21:S1655–S1665, 2004. 2.1, 2.4.4, 2.5.2

[33] R. Umstätter, R. Meyer, R.J. Dupuis, J. Veitch, G. Woan, and N. Christensen. Detecting gravitational radiation from neutron stars using a six-parameter adaptive MCMC method. In *AIP Conference Proceedings - Bayesian inference and Maximum Entropy Methods in Science and Engineering: 24th*, volume 735, pages 336–343. American Institute of Physics, 2004. International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. 2.1

[34] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. 2.1, 2.4.3

[35] R. J. Dupuis and G. Woan. A Bayesian method to search for periodic gravitational waves. preprint, 2003. 2.1, 2.2, 2.2, 2.3, 2.3, 2.3, 2.7

[36] J. H. Taylor. Binary pulsars and relativistic gravity. *Reviews of Modern Physics*, 66(3):711–719, 1994. 2.2

[37] B. Abbott et al. Limits on gravitational-waves emission from selected pulsars using LIGO data. *Physical Review Letters*, 94:181103, 2005. 2.2

[38] P. J. Green and A. Mira. Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, 88(4):1035–1053, 2001. 2.4.1, 2.4.1, 3.3.1

[39] A. Mira. *Ordering, Slicing and Splitting Monte Carlo Markov chain.* PhD thesis, University of Minnesota, 1998. 2.4.1, 3.2, 3.3.1, 3.3.2

[40] E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19:451–454, 1992. 2.4.4

[41] R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996. 2.4.4

[42] P. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences.* Cambridge University Press, New York, NY, USA, 2005. 2.4.4

[43] P. C. Gregory. A bayesian kepler periodogram detects a second planet in hd 208487, 2006. 2.4.4

[44] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281:140–150, 1997. 2.4.4

[45] E. B. Wilson and M. M. Hilferty. The distribution of Chi-square. In *Proceedings of the National Academy of Sciences*, volume 17, pages 684–688, 1931. 2.4.4

[46] W. R. Gilks, S. Richardson, and P. J. Spiegelhalter, D. J. Green, editors. *Markov chain Monte Carlo in practice*, chapter 11, page 103. Chapman and Hall, London, 1996. 2.4.4

[47] B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, London, 1986. 2.5.1

[48] J. Geweke. *Bayesian Statistics 4*, chapter Evaluating the accuracy of sampling-based approaches to calcualting posterior moments, pages 169–193. Clarendon Press, Oxford, UK, 1992. 2.5.1, 2.6.1

[49] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–482, 1992. 2.5.1

[50] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. 2.6.1, 2.6.1, 2.6.2, 3.1

[51] M. A. Newton and A. E. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994. 2.6.1, 3.1

[52] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. 2.6.1, 3.1

[53] S. Chib and I. Jeliazkov. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96(453):270–281, 2001. 2.6.1, 3.1

[54] D.J. Spiegelhalter, N.G. Best, B.R. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 64(4):583–616, 2002. 2.6.1, 3.1

[55] A. Gelman and D. B. Rubin. *A single series from the Gibbs sampler provides a false sense of security.*, volume Bayesian Statistics 4. Oxford University Press, Oxford, 1992. 2.6.1

[56] S. Richardson, M. DeIorio, and C. P. Robert. Discussions on “Bayesian measures of model complexity and fit” by Spiegelhalter, D.J., Best, N.G., Carlin, B.R., and van der Linde, A. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 64(4):616–639, 2002. 2.6.1, 3.1

[57] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csake, editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest: Akademiai Kiado, 1973. 2.6.2

[58] I. R. Aitchison and J. Dunsmore. *Statistical Prediction Analysis*. Cambridge University Press, Cambridge, 1975. 2.6.2

[59] R. Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117–126, 1976. 2.6.2

[60] R. W. Katz. On some criteria for estimating the order of a Markov chain. *Technometrics*, 23:243–249, 1981. 2.6.2

[61] K. Danzmann and A. Rüdiger. LISA technology - concept, status, prospects. *Classical and Quantum Gravity*, 20(10):S1–S9, 2003. 3.1

[62] J. Sylvestere and Tinto. M. Noise characterization for LISA. *Physical Review D*, 68(10):102002, 2003. 3.1

[63] L. Barack and C. Cutler. Confusion noise from LISA capture sources. *Physical Review D*, 70(12):122002–1–122002–21, 2004. 3.1

[64] J. A. Edlund, M. Tinto, A. Krolak, and G. Nelemans. White-dwarf-white-dwarf galactic background in the LISA data. *Physical Review D*, 71(12):122003, 2005. 3.1

[65] M. J. Benacquista, J. DeGoes, and D. Lunder. A simulation of the laser interferometer space antenna data stream from galactic white dwarf binaries. *Classical and Quantum Gravity*, 21(5):S509–514, 2004. 3.1

[66] J. Crowder and N. J. Cornish. LISA source confusion. *Physical Review D*, 70(8):082004, 2004. 3.1

[67] G. Nelemans, L. R. Yungelson, and S. F. P. Zwart. The gravitational wave signal from the galactic disk population of binaries containing two compact objects. *Astronomy and Astrophysics*, 375(3):890–898, 2001. 3.1

[68] N. J. Cornish and S. L. Larson. LISA data analysis: Source identification and subtraction. *Physical Review D*, 67(10):103001, 2003. 3.1

[69] J. R. Gair, L. Barack, T. Creighton, C. Cutler, S. L. Larson, E. S. Phinney, and M. Vallisneri. Event rate estimates for LISA extreme mass ratio capture sources. *Classical and Quantum Gravity*, 21(20):S1595–S1606, 2004. 3.1

[70] N. J. Cornish and J. Crowder. LISA data analysis using Markov chain Monte Carlo methods. *Physical Review D*, 72(4):043005–1–043005–15, 2005. 3.1

[71] J. Crowder, N. J. Cornish, and J. Reddinger. LISA data analysis using genetic algorithms. *Physical Review D*, 73(6):063011–1–063011–11, 2006. 3.1

[72] C. Han and B. P. Carlin. Markov chain Monte Carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132, 2001. 3.1

[73] J. O. Berger, J.K. Ghosh, and N. Mukhopadhyay. Approximations and consistency of bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, 112(1-2):241–258, Mar 2003. 3.1

[74] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. 3.1

[75] T. J. Loredo. *Statistical challenges in modern astronomy*, chapter The promise of Bayesian inference for astrophysics, pages 275–297. Springer-Verlag, New York, 1992. 3.1

[76] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B - Methodological*, 59(4):731–758, 1997. 3.1, 3.4

[77] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel P. and A. A. Zellner, editors,

Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, pages 233–243. Elsevier, New York, 1986. 3.2

[78] M. Davy and S. Godsill. Bayesian harmonic models for musical pitch estimation and analysis, 2002. 3.2

[79] A. Schuster. The periodogram and its optical analogy. *Proceedings of the Royal Society of London*, 77:136, 1905. 3.3.1

[80] G. Celeux. Bayesian inference for mixture: The label switching problem. In R. Payne and P. J. Green, editors, *COMPSTAT 98 - Proceedings in Computational Statistics*, pages 227–232. Physica, Heidelberg, 1998. 3.4, 3.4.2

[81] G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000. 3.4, 3.4.1, 3.4.3

[82] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 62(4):795–809, 2000. 3.4, 3.4.2

[83] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005. 3.4, 3.4.1, 3.4.2, 3.4.3

[84] M. Stephens. *Bayesian methods for mixtures of normal distributions.* PhD thesis, University of Oxford, 1997. 3.4.1

[85] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236, 1963. 3.4.2

[86] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0. 3.4.2

[87] M. Hurn, A. Justel, and C. P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003. 3.4.3

[88] M. Tinto and S. V. Dhurandhar. Time-delay interferometry. *Living Reviews*, in press (2005). 3.6, 4