Bayesian inference on astrophysical binary inspirals based on gravitational-wave measurements

Christian Röver

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy in Statistics,
The University of Auckland, 2007.

"One should always be a little improbable."

Oscar Wilde

Abstract

Gravitational waves are predicted by general relativity theory. Their existence could be confirmed by astronomical observations, but until today they have not yet been measured directly. A measurement would not only confirm general relativity, but also allow for interesting astronomical observations. Great effort is currently being expended to facilitate gravitational radiation measurement, most notably through earth-bound interferometers (such as LIGO and Virgo), and the planned space-based LISA interferometer. Earth-bound interferometers have recently taken up operation, so that a detection might be made at any time, while the space-borne LISA interferometer is scheduled to be launched within the next decade. Among the most promising signals for a detection are the waves emitted by the inspiral of a binary system of stars or black holes. The observable gravitational-wave signature of such an event is determined by properties of the inspiralling system, which may in turn be inferred from the observed data.

A Bayesian inference framework for the estimation of parameters of binary inspiral events as measured by ground- and space-based interferometers is described here. Furthermore, appropriate computational methods are developed that are necessary for its application in practice. Starting with a simplified model considering only 5 parameters and data from a single earth-bound interferometer, the model is subsequently refined by extending it to 9 parameters, measurements from several interferometers, and more accurate signal waveform approximations. A realistic joint prior density for the 9 parameters is set up. For the LISA application the model is generalised so that the noise spectrum is treated as unknown as well

and can be inferred along with the signal parameters. Inference through the posterior distribution is facilitated by the implementation of Markov chain Monte Carlo (MCMC) methods. The posterior distribution exhibits many local modes, and there is only a small "attraction region" around the global mode(s), making it hard, if not impossible, for basic MCMC algorithms to find the relevant region in parameter space. This problem is solved by introducing a parallel tempering algorithm. Closer investigation of its internal functionality yields some insight into a proper setup of this algorithm, which in turn also enables the efficient implementation for the LISA problem with its vastly enlarged parameter space. Parallel programming was used to implement this computationally expensive MCMC algorithm, so that the code can be run efficiently on a computer cluster. In this thesis, a Bayesian approach to gravitational wave astronomy is shown to be feasible and promising.

Acknowledgements

My thanks go to my PhD supervisors Renate Meyer and Nelson Christensen, for having just the right thesis topic at the right time for me, for introducing me to the "wunderbare Welt der Schwerkraft", and for managing to keep a good balance between devoted supervision and freedom. I would also like to thank everyone joining forces with us, that is, Gianluca M. Guidi, Andrea Viceré, Ed Bloomer, James Clark, Ian Harry, Martin Hendry, Chris Messenger, Matthew Pitkin, Emma L. Robinson, B. S. Sathyaprakash, Alexander Stroeer, Jennifer Toher, Alberto Vecchio, John Veitch, Graham Woan, Vicky Kalogera and Marc van der Sluys. Special thanks go to Richard Umstätter and Jennifer Wilcock for many fertile discussions, in and out of office, and partly even related to this thesis. I am grateful for financial support by the Royal Society of New Zealand Marsden Fund. Finally, another big thanks goes out to the many people not named here, who also had their share in making the last three years possible and enjoyable.

Contents

Chapter 1

Introduction

The disciplines of astronomy and statistics have always been fields of fertile interaction. Astronomical research on the one hand usually covers both the gathering and interpretation of numerical data. On the other hand, important contributions to statistical methodology originated from astronomical applications, the most famous one probably being the *theory of least squares* that was developed by C. F. Gauß as a 'spin-off' in the context of the determination of an asteroid's orbit based on observations that are subject to measurement errors [1]. Much of early astronomical statistics was actually done in a 'Bayesian spirit', although the distinction between the 'Bayesian' and the 'long-run frequency' approaches was usually not made in those days, until the advent of statistical testing and the accompanying theory around the late 19th century [2]. Some issues around Bayesian theory that formerly were unresolved have been clarified in the meantime; especially the alleged arbitrariness or subjectivity in the definition of prior distributions has been ameliorated by relating it to information theory and the concept of entropy. In particular, (Bayesian) probability theory is meanwhile established as the *extension of logic* when faced with *incomplete information*. In this sense, logic covers the theory of concluding from *certain* facts, while probability theory constitutes the extension that is able to deal with (and return) information that comes with *degrees of certainty* attached, and at the heart of which *Bayes' theorem*

gives the technical description of how to update a given state of information when provided with additional information (or data) [3].

As in any science, for about the last century the so-called 'orthodox' or 'frequentist' statistical theory has had the 'home field advantage' in astronomy, being taught and practiced as the standard procedure for any data analysis. The lack of popularity of Bayesian methods may be explained by the (naturally) limited interest in the underlying statistical theory on the part of the standard user, and by communication problems on the part of the experts [4].

Bayesian methods have a lot to offer in astronomical applications, in particular because astronomical problems are often well-posed, in the sense that the involved parameters are usually related to 'physical' counterparts, for which states of (especially prior-) information are easily formulated or interpreted [5]. In the context of gravitational wave measurements, where any analysis is essentially a time series analysis (see also the following chapter 2), Bayesian methods show great promise, as here these have already proven extremely useful and often superior to 'frequentist' practices [6]. The somewhat odd reasoning inherent in frequentist methods (where e.g. confidence levels refer to the average performance of a test procedure, and—strictly speaking—only indirectly refer to the investigated parameter itself) is once again exposed when viewed in the context of gravitational waves, as in many other astronomical applications. A Bayesian approach will be more suitable when faced with the question of what can be concluded from the one observation at hand, which is in particular not to be viewed as one in a potentially infinite series of observations [7].

The Bayesian approach has already proven useful for estimating signal parameters in gravitational wave measurements, for example for the parameters of spinning neutron stars, in resolving the number and parameters of superimposed sinusoidal signals [8], or for binary inspiral signals [9, 10, 11]. The aim of this thesis is to extend the work on the latter type of 'chirp' signal (see also the following chapter 2). In previous studies, simplified models were used, especially fixing certain parameters at their (known) true values. These will need to be taken into account, which es-

pecially means that simultaneous measurements from several instruments will also need to be considered in the analysis, as otherwise certain parameters cannot be estimated. Along with the extension of the model, appropriate computational methods need to be developed in order to facilitate inference within the model framework. In particular, Markov chain Monte Carlo (MCMC) methods are required to perform integration of the resulting posterior distributions of the parameters. Testing of the developed procedures is done using simulated data, partly because there is no 'real' data available yet.

When starting this work in mid-2004, there was initially a lot of background knowledge to acquire, including the physical background, C programming, signal processing techniques, and the theory of Fourier transformations. A first version of an analysis framework considered 5 parameters, data from a single (earth-bound) interferometer, and was based on the '2.0 PN stationary-phase approximation' of the gravitational wave signal, that is given in the frequency domain. First results were presented at the *2nd ASBA Bayesian retreat "Bayesian Topics in the Tropics"* (Brisbane, Australia, September 28–30, 2005), and were eventually published in [12]. This approach was then extended to incorporate 9 parameters, and to consider data from several separate interferometers, in particular allowing the determination of the source direction of the passing gravitational wave signal. This also required the implementation of advanced (parallel tempering) MCMC methods in order to reliably find the relevant region in the enlarged parameter space. The accuracy of the modelled signal was increased by using the '2.5 PN phase / 2.0 PN amplitude approximation' (given in the time domain), and a realistic and proper prior was specified for all parameters. These results were presented at the *ISBA Eighth World Meeting on Bayesian Statistics* (Valencia, Spain, June 1–6, 2006), as well as *Statistical Challenges in Modern Astronomy IV* (State College, PA, USA, June 12–15, 2006), and published in [13]. The approach was later upgraded further by introducing the '3.5 PN phase / 2.5 PN amplitude approximation' for the signal, and by further refinement of the prior definition; these results were presented at the *11th Gravitational wave data analysis work-*

shop (GWDAW-11) (Potsdam, Germany, December 18–21, 2006), as well as the *2007 Joint Statistical Meetings* (Salt Lake City, UT, USA, July 29–August 2, 2007), and are described in [14].

Beginning in mid-2006, a similar approach was developed for signals as observed by the planned space-borne Laser Interferometer Space Antenna (LISA) [15]; this work was done within the 'Global LISA Inference Group (GLIG)' that was set up in order to jointly work on analysis methods for data obtained by LISA. A major difference is in the size and mode of operation of this instrument, which will be sensitive to signals in a different frequency range, and which (in contrast to the modeling of earth-bound signals) requires the response to a binary inspiral's 'chirp' signal to be numerically derived (at least for now). A basic algorithm for inference on inspiral signals modeled by 9 parameters and the 'restricted PN approximation' was implemented by the end of the year, and first results were presented at the *11th Gravitational wave data analysis workshop (GWDAW-11)* (Potsdam, Germany, December 18–21, 2006) [16]. In order to be applicable in a more realistic setting, the model was lacking particular features, which were developed and implemented by mid-2007, and the resulting inference framework was presented at the *7th Edoardo Amaldi conference on gravitational waves* (Sydney, Australia, July 8–14, 2007).

Both applications to ground-based as well as space-based gravitational wave measurements benefited from the implementation of advanced MCMC methods taylored to the particular issues in each case. Along the way, some insight into the background of some methods was gained, which lead to the derivation of some universal algorithm characteristics that might also allow for more efficient implementations in any general applications. The joint prior for the signal parameters was properly defined in a general way that should also be transferable to similar applications. The application to LISA measurements, with its particularly complex expected background noise, led to the development of a very general Bayesian formulation of a noise model that might also find applications in other contexts.

The organisation of this thesis is as follows. Chapter 2 introduces some

general background on the physics of gravitational waves and gravitational wave astronomy. Chapter 3 introduces the methods that are used later, and also contains some recommendations on their effective implementations. In chapter 4, details of the applied statistical models are described. In chapter 5, the previous elements are assembled to inference frameworks for binary inspiral signals measured by earth-bound and space-bound interferometers, and example applications are illustrated for both cases. Chapter 6 finally gives some conclusions. The appendix contains some more details that are too lengthy or not necessarily vital to be included in the main part. In general, the descriptions of methods or background given here are intended to be sufficiently detailed to point out the relevant concepts (and further references), and to introduce consistent notation conventions that can later be referred back to; they are in particular *not* supposed to constitute an 'ultimate' reference. For example, all defintions related to Fourier transforms are restricted to real-valued inputs, which are all that is relevant here. Also, the exact definitions of the modeled waveforms are mostly skipped, since they would be extremely lengthy, are probably not enlightening beyond what can already be seen from the given simplified form, and in any case anyone trying to reproduce these would probably be better advised to consult the original references.

Chapter 2

Gravitational wave astronomy

2.1 Gravitational radiation

General relativity theory introduced the idea of a space-time that is curved by the presence of mass, energy and momentum within it. It implies the existence of gravitational waves, which can be thought of as distortions or 'ripples' in space-time, caused especially by rapidly moving heavy objects, and propagating at the speed of light [17]. Until now this effect has only been observed *indirectly*, when the observed slight change in the orbital period of binary neutron star system PSR 1913+16 matched exactly with the predicted amount due to the loss of energy through gravitational radiation [18]. The *direct* measurement of gravitational radiation would not only confirm General Relativity Theory, but would also complement 'traditional' observations in the electromagnetic spectrum by opening up an additional completely new 'window' for a wide range of astronomical observations. Gravitational waves will encode interesting information about the processes causing them, and they are emitted by events that are otherwise 'invisible', as they may include black holes, and may also be measurable to great distances [19, 20].

A gravitational wave's effect acts in directions orthogonal to its direction of travel. It is a *quadrupole* wave, which implies that the waveform can be decomposed into two orthogonal components that are offset by a

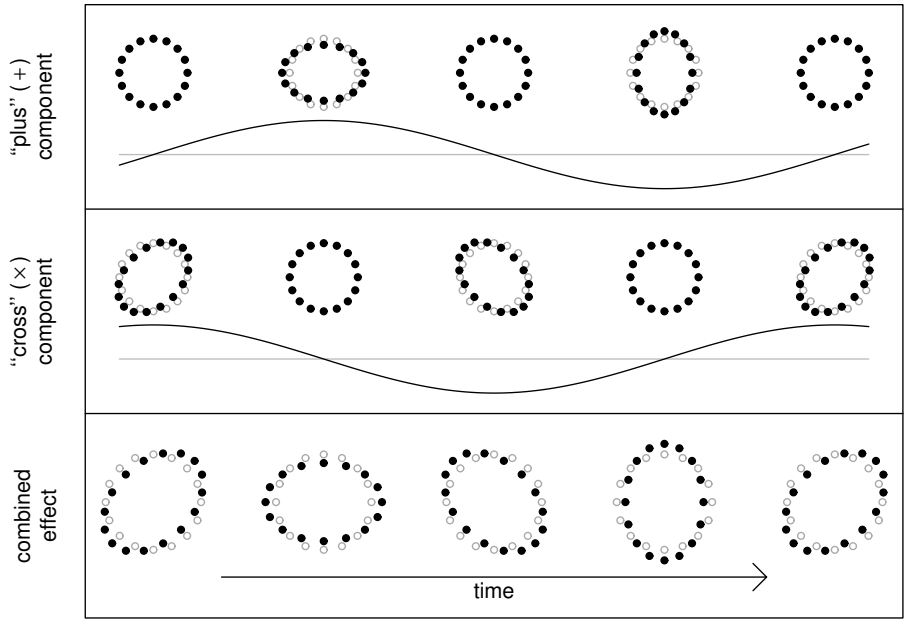

Figure 2.1: Illustration of the effect that a gravitational wave would have on a set of free falling test masses arranged in a circle. The wave's direction of travel here is orthogonal to the plane of the test masses. The wave's two orthogonal plus- ('+') and cross- ('×') components each cause an alternating 'squeezing' and 'stretching' into orthogonal directions, which are offset by 45° between plus- and cross-polarisation. The overall effect is shown in the bottom panel, and it would have different appearances for different plus- and cross-amplitudes.

rotation of 45° (along the direction of travel). This is illustrated in figure 2.1: imagine a set of free falling masses that are arranged in a circle. A gravitational wave travelling orthogonal to the circle's plane would then cause the ring of masses to be alternatingly stretched and squeezed. The waveform can be decomposed into its "plus" (+) and "cross" (×) components, and the overall effect gives the ring of masses a periodically 'wobbling' appearance. A rotation by 90° inverts the waveform, and a 180° rotation again yields the original wave.

2.2 Measuring gravitational waves

The effect of gravitational radiation is very weak, and so it takes very sensitive instruments to detect and measure it. In order to measure gravitational waves, one needs to measure the tiny relative motions of masses (as illustrated in figure 2.1) as a wave passes by. The most prominent approach towards the measurement of gravitational waves is currently by means of *laser interferometers*.

In earth-bound interferometers, the free falling masses are replaced by carefully suspended pendulums that are placed in the corners of a right-angled triangle, and a gravitational wave's squeezing/stretching effect in orthogonal directions is then measured by monitoring their motion over time using *laser interferometry*. Two test masses are placed in the opposite corners of a right-angled triangle. Originating from the third corner, laser beams are passed along the two right-angled sides, reflected at the masses, and matched as they return to the corner station. Monitoring the two lasers' interference then allows the detection of movements of the test masses in the direction of the laser beams that are well below the laser's wavelength. Several such interferometers have already been built, the two LIGO instruments in Hanford and Livingston (USA) [21], Virgo near Pisa (Italy) [22], Tama in Tokyo (Japan) [23], and GEO600 near Hanover (Germany) [24]. Another interferometer, AIGO, is planned to be built in Australia.

A variation of the same principle is going to be implemented in the *Laser Interferometer Space Antenna (LISA)*, which is a joint NASA/ESA project for a space-borne interferometer. This instrument is planned to be launched in about 10 years time, and will consist of three satellites forming an equilateral triangle of several million kilometres is size. Here, six laser beams will be passed between the satellites, again allowing the monitoring of the relative motions of test masses within each satellite [15].

The functional principle of these two kinds of laser interferometers is illustrated in figure 2.2. In the top panel, the black dots represent the test masses (pendulums), and the black square is the corner station. Laser

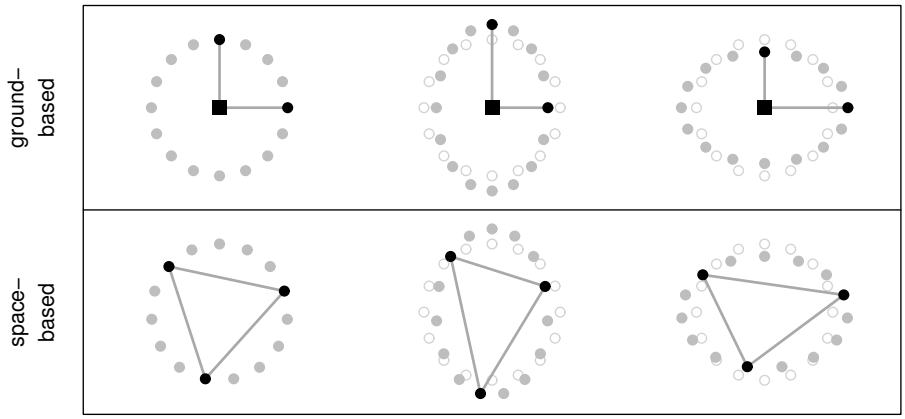

Figure 2.2: A sketch of how earth- and space-based laser interferometers 'perceive' a passing gravitational wave (in analogy to figure 2.1). In both cases the geometry of the test masses (shown in black) changes over time, which is monitored via the laser beams (shown as grey lines).

beams are directed from the corner station to the test masses, where they are reflected and return to the corner station. The passing gravitational wave changes the distances the beams have to cover, which is picked up through the resulting interference of the beams as they return to the corner station. The space-based interferometer works similarly, but is has three test masses (satellites) that emit and receive two laser beams each. The distortion of the triangle again is measured by matching the lasers' phases. The size of an earth-bound interferometer is of the order of kilometres, while that of the space-based interferometer is of the order of millions of kilometres. Pictures of existing interferometers and of the planned LISA instrument are shown in figure 2.3.

In any case, the data output of such measurements is one or more time series. A laser interferometer is not 'pointed' in any direction like a telescope, but will measure waves coming from (more or less) any direction, although with varying sensitivity. It will be most sensitive to gravitational waves passing the interferometer's plane orthogonally (as in figure 2.1), and less sensitive otherwise. Consequently, one might rather think of it as an 'antenna' that is 'listening' for passing gravitational waves [19].

Figure 2.3: Pictures of actual interferometers. The left one is the LIGO interferometer in Livingston (LA, USA) with 4-km-arms. The interferometer in the middle is the Virgo instrument near Pisa (Italy) that has 3-km-arms. The picture on the right is a sketch of the planned LISA interferometer consisting of three satellites at distances of 5 million kilometres [25].

2.3 Binary inspirals

Gravitational radiation is emitted by many kinds of objects and processes, but one of these that is best understood and is also expected to be detected first is the *binary inspiral*. A binary inspiral event develops as a pair of heavy objects (neutron stars or black holes) rapidly orbit around their centre of mass. Due to the radiation of energy in the form of gravitational waves, the orbital distance between the objects decreases while the orbital frequency increases, until the system eventually collapses and the two companions merge.

The gravitational wave signal emitted by such an event is a '*chirp*', an oscillation of increasing frequency and amplitude, whose evolution is primarily determined by the masses of the two involved objects. The signal in general is shorter and more violent for greater masses (i.e. follows a steeper frequency and amplitude increase), and is of longer duration for lower masses. The waveform that is measured at an interferometer is then further affected by the relative orientations of interferometer and inspiral event with respect to each other. If the two inspiralling companions have roughly similar masses, it is predicted that their orbits will have circularised over time, which simplifies the waveform that needs to be modeled. Otherwise, further parameters that might need to be taken into

account are eccentricities of the orbit, or spins that the companions may have. Estimates of rates by which events happen within the sensitivity range of today's earth-bound interferometers are of the order of several per year, but vary by orders of magnitude. With the upgrades to 'second generation' detectors that are planned to be installed in the near future, the sensitivity will be significantly increased. Also, the sensitivity of current detectors already is steadily being improved, and is expected to be significantly greater within a few years. The planned space-based interferometer LISA will be sensitive enough to be certain to measure numerous signals up to the limit of confusion [26, 27]. Due to their different layout, ground-based and space-based interferometers are sensitive to signals in different frequency bands, and will each only perceive a certain selection of all signals present [20, 28, 29].

2.4 Inference on gravitational waves

2.4.1 The statistical problem

Besides binary inspirals, there are more sources of gravitational waves that are expected to be detected by gravitational wave measurements. These include *bursts*, which are signals of short duration that are expected e.g. from supernovae, in conjunction with gamma ray bursts, or at the end of a binary inspiral when the two companions merge. *Spinning neutron stars* and *binary systems* (that are yet far from their eventual inspiral) are expected to emit periodic, sinusoidal signals. The merger of black holes is expected to be followed by a *ringdown* signal, a damped pulsation emitted by the resulting newly formed black hole. Eventually, studying properties of the *stochastic background noise* may be of interest, as it may shed light on properties of the early universe [20].

When studying these different kinds of sources, the aims may be very different, ranging from parameter estimation for known waveforms to exploration or classification of unknown waveforms, or the characterisation of the noise. The statistical problem will—at least initially—be a time se-

ries analysis problem. In this respect, gravitational wave astronomy is probably primarily comparable to radioastronomy, in particular in the sense that there are no 'pictures' being shot, but one is rather 'listening' for signals.

The data one is dealing with are a single time series for a single interferometer, or more for several interferometers or for the planned space-based LISA instrument. In any case, the data will need to be seen in conjunction with the instrument's location and earth's orbital parameters, in order to be able to account for orientation or Doppler effects.

The kind of noise one is faced with depends on the problem at hand. There will always be instrument noise, but depending on what particular 'signal' one is studying (which might be the background noise itself), all other 'signals' will need to be considered as 'noise'.

2.4.2 Some common approaches

With respect to problems of estimating parameters of signals where the waveform is (assumed) known, there are several different general approaches being followed. A *template bank search* (e.g. [30]) is, roughly speaking, a 'brute force' or 'grid' search, where the data are matched against a bank of signal templates and the optimal match with respect to a certain match criterion is sought. The match between data and a signal waveform (that corresponds to a certain parameter setting) is usually evaluated by the likelihood, and such methods are then, statistically speaking, maximum-likelihood (ML) methods. *Bayesian* estimation methods (in conjuction with MCMC methods) have been implemented as well (e.g. [9]). Other parameter estimation techniques that are being used include the Hough transform (e.g. [31]), the Radon transform, or the Hilbert-Huang transform (e.g. [32]).

When searching for signals in the data, and data from more than one instrument are involved, methods can usually be categorised as either *coincidence methods* or *coherent methods*. In a coincidence search, the data sets are searched (e.g. for inspiral event signals) individually, and the re-

sulting lists of event candidates are matched for entries that are close in time for further investigation. Coherent methods on the other hand consider all the data simultaneously. Coincidence searches are in general computationally less expensive, while coherent methods are in general more sensitive.

However, up to now, parameter estimation techniques have only been tested on simulated data (or at least simulated *signals*), since a 'real' signal has not been detected yet. Applications to real data have by now mostly resulted in upper limits on event rates and/or signal amplitudes, based on the observation that *no* signal was detected [33].

2.4.3 Dealing with noise

When measuring gravitational waves, the measurement errors are a composition of *instrumental noise* and *background noise*. Instrumental noise here refers to what the instrument would still measure if there were no gravitational waves, which would for example be due to vibrations within the instrument, or random fluctuations in the laser beam. If the instrument was able to 'perfectly' measure gravitational radiation, then it would still measure many signals that are simultaneously present, but that are not the signal(s) actually aimed for. The resulting noise is a non-white background that is made up of 'random' and 'deterministic' (but unaccounted for) contributions [19, 26, 27, 34].

In the context of ground-based measurements of inspiral signals, the assumption of Gaussian and stationary noise with a fixed spectrum might be justified, since the relevant observation times are relatively short (of the order of seconds to minutes). In particular, the noise spectrum may be assumed roughly constant over such short times, and may be estimated from immediately preceding or following measurements; also, the signals concerned are so rare that the probability of one inspiral signal 'contaminating' another's measurement is negligible. However, when dealing with space-based measurements, this approach is less appropriate. The necessary observation periods are significantly longer (of the order of months

or years, if not the mission's complete lifetime), and the spectrum cannot be assumed constant, since the background signals may evolve over time, and in any case are modulated over time by the instrument's orbital motion.

An ad hoc "solution" to this problem that has been proposed, is to derive estimates for the parameters of the unaccounted for signals, and then 'subtract' these from the data one by one (e.g. [35, 36]). However, there are some concerns about this approach. Firstly, the parameter estimation will always result in a mismatch between the true parameters and their estimates, plus there might be a mismatch between the simplified model and the actual signal, and hence each subtraction will in turn *add* a 'residual signal' to the noise (although this residual will probably be of far less magnitude than the original signal). One question is whether it is more harmful to have background signals in the data, which one then can account for, or to have those residuals in the data, which are probably harder to account for, if they are not supposed to be ignored in the following. Due to the large number of background signals, the 'residual signals' might accumulate to be quite large, especially considering that (due to the law of large numbers) there will be cases of bad mismatch among these. Another question is whether the background signals actually interfere with the signal of primary interest, i.e. whether there is some similarity between background and foreground signals, or whether the background signals' presence affects the foreground signal's parameter estimates. If it does not, one may as well leave them in the data. But if it does, their removal might actually remove parts of the sought for signal as well—which might be worse than leaving them where they are. In any event, as soon as signals get too close to each other in frequency (which they will in certain frequency bands), their individual parameters will not be resolvable any more [37], and so the subtraction will have to stop at some point anyway. Also, if in the signal subtraction phase the algorithm is not faced with the proper alternative of the signal actually sought for, it might consequently (also due to their relatively general form) attribute parts of it to be due to background noise. Since the signal subtraction is effectively a maximum-

likelihood (or maximum-a-posteriori) approach that does not account for uncertainty in parameter estimates, it would strictly speaking prevent an end-to-end Bayesian analysis of the resulting data. In the end, the noise was not Gaussian before subtraction, but whether the actual signal is 'unharmed', and whether the noise is easier to handle afterwards is not certain.

Note that the associated background signal parameter estimation algorithms that have been developed are very sophisticated and effective (e.g. [38]), and that a Bayesian approach to the same problem requires complex and time-consuming MCMC methods. Due to the nature of the problem, the derivation of Bayesian parameter estimates also is far from trivial [37]. It is just that there are reasonable concerns with the practice of the point estimation and 'subtraction' of the background noise signals that should be given serious consideration. It might actually turn out to be an effective and harmless way of dealing with background noise, but it might also turn out to be unnecessary.

Alternatively, considering that the background noise's parameters are not of primary interest, and in appreciation of the fact that there will be unexpected and unmodeled signals within the background noise anyway, one can try to set up a robust model that introduces minimum assumptions and still accounts for noise with unknown spectrum, including unaccounted for 'noise signals'. This is attempted later in section 4.5. The model is supposed to reflect the *randomness* in the noise, as well as *ignorance* about deterministic, but unaccounted for signals. The only explicit assumption being made is that the noise spectrum is finite, but otherwise it is inferred along with the signal parameters based on the data and prior information.

Chapter 3

Methods

3.1 Bayesian modeling

The main characteristic of Bayesian procedures, that distinguishes them from 'frequentist' or 'orthodox' methods, is the concept of *probability*. Here, probability is understood in a general sense. Probabilities are associated with any kind of event (and in particular not restricted to 'repeated trials'), and probability calculus is applied as a generalisation of logic, allowing one to process and infer *states of incomplete information* [3]. Bayes' theorem comes into play as it then turns out to be the natural way to update a state of information given some observation, or data. In the end, this leads to unique ways to approach and solve statistical problems.

A Bayesian data analysis starts with the specification of a model for the observables y, depending on some model parameters θ. This is done by defining the *sampling distribution* that describes how observations come about given certain parameter values, which is expressed through the probability density $p(y|\theta)$. The introduction of θ then entails the necessity to formulate the *a priori* information about the parameters, again, in terms of a probability distribution, the *prior distribution* $p(\theta)$. The 'probability distribution' here is to be understood as a distribution *of probability* across the parameter space, associating parameter values with probabili-

ties. In particular it does *not* denote expected frequencies in some sort of 'repeated draws' of parameter values. Application of Bayes' theorem then yields the *posterior distribution* $p(\theta|y)$, which expresses the 'updated' information (in terms of probability) about the parameters given the observations y:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{3.1}$$

[39]. How exactly to proceed from here and how to make use of $p(\theta|y)$ depends on the particular problem at hand. One can just illustrate $p(\theta|y)$ graphically, or record some characterising key figures. If one is interested in parameter estimates, then the problem is of a decision-theoretic nature. Decision theory will require the specification of a loss function, and the determination of an optimal estimate $\hat{\theta}$ then follows via optimisation over the parameter space. For example, in the special (but common) case of a quadratic loss, the optimal estimator of a parameter is its posterior expected value (where 'optimality' refers to a minimal expected loss) [40].

Technically, inference usually requires integration of the posterior distribution over the parameter space (or parts thereof). In general, such integration is hard to do analytically, and includes functions of the whole data y. For this reason, Monte Carlo integration is often used to approximate the desired integrals.

The specification of the parameters' prior distribution is also crucial for the eventual analysis. The prior distribution needs to reflect the prior information about the parameters *before* the data are taken into account, but, unfortunately, "the problem of translating prior information uniquely into a prior probability assignment represents the as yet unfinished half of probability theory" [3]. There are many approches to the setup of the prior, and it is hard to give universally applicable advice. However, especially in problems like the present one where all parameters are of a physical nature, the range of *reasonable* choices is often very limited, as is the relevance of differences among these. In general, if there is sufficient evidence about the parameter values in the data, the prior distribution will be outweighed by the likelihood in the resulting posterior distribution. On the

other hand, if there is little or nothing to learn from the data about the parameters, the posterior state of information will (almost) equal the prior information, and consequently the prior specification matters very well.

3.2 Monte Carlo integration

3.2.1 General case

Bayesian analyses generally require the solution of integrals involving the posterior distribution, which is typically specified in terms of its density function $p(\theta|y)$ (see also equation (3.1)). The posterior density is usually known only *up to a normalising constant*, since the value of the denominator $p(y) = \int p(y|\theta)p(\theta)\mathrm{d}\theta$ (see equation (3.1)), which is independent of θ and constant for given y, is another unknown integral. One way to approach the problem is to generate (or rather, simulate) a sample from the posterior distribution and then approximate the desired integrals by sample averages—so-called *Monte Carlo integration* [41]. For example, consider the case where one is interested in computing the expectation of a function h of a random variable X that has the probability density function $f_X(x)$:

$$\mathrm{E}[h(X)] = \int h(x)\,\mathrm{d}f_X = \int h(x)\,f_X(x)\,\mathrm{d}x. \tag{3.2}$$

Instead of doing the integration analytically, an approximation can be computed by obtaining a random sample $x_1, \ldots, x_N$ from the distribution f_X, and estimating the figure by the sample average

$$\overline{h(x)} = \frac{1}{N}\sum_{i=1}^{N} h(x_i). \tag{3.3}$$

Analogously, marginal densities, quantiles, etc. can be estimated from samples from the distribution of interest.

There are many ways to simulate or approximate such draws from a given probability distribution using computers and implementations of

(pseudo-) random number generators [41, 42]. In the following sections some *Markov chain Monte Carlo (MCMC)* methods are described in detail.

3.2.2 Notation

In the following, a general '*target*' distribution of interest is denoted by its density function $f(\theta)$. In the special case of posterior simulation, where the distribution $f(\theta) = p(\theta|y)$ is proportional to the product of prior and likelihood (see equation (3.1)), the prior density is denoted by $\pi(\theta)$, and the likelihood by $\mathcal{L}(\theta) = p(y|\theta)$ (so that $f(\theta) = p(\theta|y) \propto \pi(\theta)\mathcal{L}(\theta)$). The methods described here rely on the availability of basic random number generators.

3.2.3 MCMC simulation

A *Markov chain* is a random process assuming different states over a discretely proceeding time, in which each random step in the sequence only depends on the previous state of the chain (the *Markov property*). *Markov chain Monte Carlo (MCMC)* methods make use of Markov chains by setting up a random walk through an arbitrary state space, whose stationary distribution is set to be a certain distribution of interest (and whose state space is the domain of that distribution). This allows the generation of random sequences of numbers from an arbitrary target distribution which can then be used for Monte Carlo integration. In general, subsequent samples will be more or less correlated, which might turn out to be a problem. Another problem is that MCMC algorithms often need a 'burn-in' time of indefinite length before they properly sample from their stationary distribution. The most prominent examples of MCMC algorithms are probably the Metropolis, Metropolis-Hastings, and the Gibbs algorithm [41, 42].

3.2.4 The Metropolis algorithm

The Metropolis algorithm can be applied when the target distribution is only known up to a normalising constant. It proceeds as follows:

0. Given: a target distribution $f(\theta)$, a starting point θ^0 for which $f(\theta^0) > 0$, and a proposal (or jumping-) distribution with density $J(\theta^*|\theta^t)$ that is *symmetric* in the sense that $J(\theta^a|\theta^b) = J(\theta^b|\theta^a)$ for all θ^a, θ^b. Set $t = 0$.

1. Increase t by 1.
   Propose: draw a *candidate point* θ^* from $J(\,\cdot\,|\theta^{t-1})$.

2. Calculate the ratio of densities (acceptance probability):

$$r \;=\; \frac{f(\theta^*)}{f(\theta^{t-1})} \tag{3.4}$$

3. Set

$$\theta^t \;=\; \begin{cases} \theta^* & \text{with probability } \min(1, r) \\ \theta^{t-1} & \text{otherwise} \end{cases} \tag{3.5}$$

   and continue from step 1.

So in each step the current state θ^{t-1} of the chain is randomly manipulated using the proposal distribution J, and the proposed new state is either *accepted* or *rejected* according to the ratio of the target densities at current and proposed state. If $f(\theta^*) \geq f(\theta^{t-1})$, the proposal is always accepted, otherwise acceptance is a matter of chance. A random walk set up in this way then has the specified target distribution as its *stationary distribution* [42, 43].

Note that in equation (3.4) any constant multiplier to the target density cancels out. Note also that the generated samples are *not* independent from each other, and that the efficiency of the sampler heavily depends on sensible choices of starting point θ^0 and proposal distribution J.

3.2.5 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a generalisation of the Metropolis algorithm which does not require the proposal distribution $J(\theta^*|\theta^t)$ to satisfy the symmetry condition. This is taken into account when computing the acceptance probability r (see equation (3.4)), which becomes

$$r = \frac{f(\theta^*)\ J(\theta^{t-1}|\theta^*)}{f(\theta^{t-1})\ J(\theta^*|\theta^{t-1})} \tag{3.6}$$

instead [42, 44]. You can see that *if* the proposal distribution satisfies the symmetry condition, expression (3.6) again simplifies to (3.4). Note also that the proposal distribution's density J only needs to be specified up to a normalising constant as well.

3.2.6 The Gibbs sampler

The Gibbs sampler works by first dividing the set of parameters into parameter subsets. Sampling is then done by alternately holding all parameters except a certain subset constant, and then drawing from the conditional distribution, conditional on the 'constant' parameters. Consider sampling from a posterior $p(\theta|y)$ where the parameter vector $\theta = \begin{pmatrix}\vec{\alpha}\\ \vec{\beta}\end{pmatrix}$ consists of sub-vectors $\vec{\alpha}$ and $\vec{\beta}$.

1. Current state is: $\theta^t = \begin{pmatrix}\vec{\alpha}^t\\ \vec{\beta}^t\end{pmatrix}$.
2. Draw α^{t+1} from the conditional distribution $p(\vec{\alpha}|\vec{\beta}^t, y)$.
3. Draw β^{t+1} from the conditional distribution $p(\vec{\beta}|\vec{\alpha}^{t+1}, y)$.
4. New state is: $\theta^{t+1} = \begin{pmatrix}\vec{\alpha}^{t+1}\\ \vec{\beta}^{t+1}\end{pmatrix}$. Repeat from step 1.

[42]. The basic algorithm can be varied in many ways. For example, one can use an arbitrary number of subvectors. Sometimes it is easy to sample from some of the conditional distributions when these are known, parametric distributions. Otherwise one can for example also use a Metropolis-step to implement sampling from the conditional distribution.

Note that the Gibbs sampler (as opposed to a Metropolis sampler) in general does not have an *acceptance probability*, and it does not require the explicit computation of likelihoods. In each step, the individual parameters, or subsets of parameters, are drawn (in an unspecified manner) from their conditional distributions, conditioning on the values of the remaining parameters. If some of these conditional draws happen to be implemented as Metropolis (-Hastings) steps, they will of course (internally) have an acceptance probability, but in general the sampler moves through parameter space along those sub-spaces and following the corresponding conditional distributions.

3.2.7 Enhancing and diagnosing MCMC performance

Convergence and mixing

As already noted above, a Metropolis (-Hastings) algorithm's performance depends on the choice of starting point θ^0 and proposal distribution J. Starting points should *at best* be drawn from the target distribution—whatever exactly qualifies a given point as such, besides the minimum requirement that $f(\theta) > 0$. The optimal proposal would be the target distribution itself (leading to acceptance rate of $\equiv 1$ and independent steps).

Metropolis- (and related) MCMC algorithms also possess optimization properties; in fact they happen to behave in a very similar way to e.g. a Nelder-Mead algorithm which is extended to a simulated annealing algorithm: on its random walk through parameter space it will always accept an 'uphill' step, and sometimes (randomly) a 'downhill' step as well [45]. This property often comes in handy since the problem usually is not only to *sample* from the posterior, but also to first *find* the global posterior mode(s) within a complex posterior surface, and among numerous minor modes. These convergence properties can also be enhanced through the implementation of the sampler, while care must be taken to maintain its ergodicity properties. The phase that the sampler spends 'converging towards its stationary distribution' is often called the *burn-in phase*.

Once the MCMC sampler has converged and is sampling from the tar-

get distribution, a good '*mixing*' of the sampler is desirable, which refers to a desired independence between subsequent samples. If for example the variance of the proposal distribution is chosen too small, then subsequent samples θ^t, θ^{t+1} are correlated because $\theta^t \approx \theta^{t+1}$, and the chain only moves very little in each step. If on the other hand the variance is chosen to be too large they are *also* highly correlated, because $\theta^t = \theta^{t+1}$ for most iterations, since most proposals get rejected. The trick is then to find a good balance between the two extremes.

Practical implementation

In practice, a good starting point is one for which $f(\theta^0)$ is large, the distribution's mode for example, if that can easily be determined. On the other hand, '*overdispersed*' starting values are desirable as well, i.e. points further away from the mode—starting a sampler repeatedly from such points, one can then see whether chains 'robustly' converge towards the same mode.

When using a proposal distribution centered around the current state, different variance choices were explored for the case of a Normal distribution for both target and proposal distribution, and were found to be optimal at a size of $\approx \frac{2.4}{\sqrt{d}}$ times the true covariance, where d is the distribution's dimension [42]. Looking at the proportion of accepted steps among all proposed steps in such a setup, an optimal proposal distribution results in an acceptance rate of $\approx$ 23% [42]. These figures must be treated with caution though; not every sampler with 23% acceptance rate is good, and not every good sampler has 23% acceptance. There may be good reasons for deviations from that rule-of-thumb.

One can run parallel chains and compute an individual convergence measure $\widehat{R}$ ('potential scale reduction factor') for each parameter, in order to assess the convergence of the sampler to the same stationary distribution when starting from different points θ^0 [46, 42]. $\widehat{R}$ has also been extended to the multivariate case $\widehat{R}^p$, which also serves as an upper bound to the individual $\widehat{R}$'s [47].

The MCMC output is often 'thinned out', by keeping only every kth sample and discarding the remaining ones. The advantage is that it reduces the correlation between subsequent samples, and reduces the amount of data that needs to be stored and handled. If data storage is not a problem, there would be no advantage in skipping iterations, however [42].

3.2.8 Metropolis-coupled MCMC

Metropolis-coupled MCMC ('MCMCMC') is a variation of the Metropolis (-Hastings) sampler in which k MCMC chains with *differing* stationary distributions f_i are run in parallel, where $f_1 = f$, so that only one chain samples from the distribution of actual interest. Then additional steps are introduced, proposing 'swaps' between two chains i and j currently being in states θ_i and θ_j. A swap means that the parameter sets θ_i and θ_j are exchanged between chains i and j, or equivalently, that the stationary distributions f_i and f_j are switched. These swaps are accepted with probability $r_s = \min(1, \omega)$, where

$$\omega = \frac{f_i(\theta_j)\, f_j(\theta_i)}{f_i(\theta_i)\, f_j(\theta_j)}. \tag{3.7}$$

The parallel chains are supposed to improve convergence and mixing, and eventually only the draws from the first chain are used for inference while the others are in general discarded [41, 48].

The different stationary distributions f_i can be different modifications of the actual target distribution, e.g. using tempering (see following sections) or cheaper approximations. Note that the individual chains are not Markov chains any more, but instead the whole set of k chains now forms a Markov process on the k-fold cartesian product of the original parameter space. Also, the swap acceptance probability ω is computed assuming that the involved density expressions actually reflect the frequencies with which states θ are assumed by the chains—in other words, the *whole set of k chains* needs to have converged and completed their burn-in in order for the first chain's samples to be valid.

When using Metropolis-coupled MCMC, another advantage is that to some extent it already works in the spirit of the 'parallel chains' approach described in the previous section 3.2.7 (or [42, 46]). When interpreting the 'swapping' operations as swaps of the stationary distributions (and not as swaps of parameter sets), then the individual chains still move independently of one another. Running several Metropolis-coupled MCMC algorithms in parallel would not make much sense.

3.2.9 Tempering methods

Simulated annealing and *parallel tempering* methods both utilise a 'tempered' version of the objective function. Tempering is motivated by the observation that e.g. chemical crystallisation processes behave differently at different temperatures, and in particular that more regular crystals (corresponding to a globally optimal alignment of molecules) are formed when a solution is slowly *annealed*, in contrast to irregular crystals (corresponding to locally optimal molecule alignments) that tend to form when they have to do so more rapidly. In analogy, an optimisation algorithm treats the target function differently under different 'temperatures' (specified by a temperature parameter T), the intention being to make it less likely to get caught in local optima when temperatures are higher.

With a probability distribution as objective function, the tempering can be defined by replacing the density function $f(\theta)$ by its 'tempered version'

$$f_{(T)}(\theta) \;=\; c_T\, f(\theta)^{\frac{1}{T}} \;\propto\; f(\theta)^{\frac{1}{T}} \tag{3.8}$$

where $T \geq 1$ is the temperature, and c_T is the corresponding 'new' normalising constant. $T = 1$ yields the initial distribution, and greater values correspond to tempered distributions. When applied in the context of a Metropolis sampler, there are (at least) two motivations for this manipulation:

Variance inflation / flattening out: Consider sampling from a univariate Normal distribution with density $f(\theta) \propto \exp\left(-\frac{1}{2}(\frac{\theta-\mu}{\sigma})^2\right)$. Then a chain

at temperature T will instead be sampling from density

$$f_{(T)}(\theta) \;\propto\; \exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2\right)^{\frac{1}{T}} \;=\; \exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\sqrt{T}\,\sigma}\right)^2\right), \quad (3.9)$$

so effectively the standard deviation σ is inflated by a factor of $\sqrt{T}$. Analogous effects apply in higher-dimensional cases, and, hopefully, for other, more interesting distributions.

Relaxed acceptance / jumping: In the case of a Metropolis algorithm, the 'regular' acceptance probability r (cp. equation (3.4)) changes for the tempered distribution to

$$r_{(T)} \;=\; \frac{f(\theta^*)^{\frac{1}{T}}}{f(\theta^{t-1})^{\frac{1}{T}}} \;=\; r^{\frac{1}{T}} \;=\; \sqrt[T]{r} \;\geq\; r, \quad (3.10)$$

so the tempering may also be perceived as a modification of the *MCMC algorithm* rather than of the *target distribution*. The MCMC random walk can thus also be considered to be exploring exactly the same ('uninflated') distribution f, but with 'extended leeway'.

Implementing tempering as above allows one to control the shape of the tempered distribution, which is equal to the target distribution for $T = 1$ and gets increasingly 'flat' for greater T.

In the context of posterior simulation where the target distribution $f(\theta) = \pi(\theta)\mathcal{L}(\theta)$ is the product of prior π and likelihood $\mathcal{L}$, it may make sense to apply the tempering only to the likelihood part, so that

$$f_{(T)}(\theta) \;=\; c_T\,\pi(\theta)\,\mathcal{L}(\theta)^{\frac{1}{T}} \;\propto\; \pi(\theta)\,\mathcal{L}(\theta)^{\frac{1}{T}}. \quad (3.11)$$

For the two extreme cases $T = 1$ and $T \to \infty$ the tempered distribution $f_{(T)}$ then is equal to posterior and prior respectively, so manipulating T allows to adjust between these two. This usually makes more sense than instead having a uniform distribution in the extreme case of $T \to \infty$, which would often be improper and would also lead to different be-

haviours for different parametrisations of the same problem. Another advantage is that this way the tempered distribution will always be proper, as long as the prior is proper (for a proof see appendix A.1). Note that if the likelihood is of a form as in section 4.5.2, then the tempered likelihood may also be seen as a likelihood that assumes the noise (spectrum) to be inflated by a factor of T instead. The Metropolis acceptance probability (cp. equations (3.10) and (3.4)) in this case becomes:

$$r_{(T)} = \frac{\pi(\theta^*)}{\pi(\theta^{t-1})} \left(\frac{\mathcal{L}(\theta^*)}{\mathcal{L}(\theta^{t-1})} \right)^{\frac{1}{T}}. \tag{3.12}$$

3.2.10 Simulated annealing

Simulated annealing is often applied in the context of optimisation. Tempering of the objective function is utilised by starting the optimisation at a high temperature and then lowering the temperature over time corresponding to some *annealing scheme* until optimisation is eventually carried out on the actual objective function [45]. The idea is that the algorithm is able to find the vincinity of the global mode at high temperature, and then narrows down to the optimum as the temperature is lowered. Here, sensible choices of starting temperature and annealing scheme are crucial for an effective implementation. In particular, if the temperature is lowered too quickly, the main mode may not yet have been found.

3.2.11 Parallel tempering

Parallel tempering is a special case of Metropolis-coupled MCMC sampling (see section 3.2.8), where the stationary distributions of the MCMC chains running in parallel are defined by 'tempering' the objective density. Each chain i runs at a different temperature T_i, where $T_1 = 1 < T_2 < \ldots < T_k$, so the first chain runs at temperature 1 and thus samples from the actual distribution of interest. If dealing with a posterior distribution as the target density, and the tempering is implemented as in equation (3.11), then a swap between chains i and j that are in states θ_i and θ_j (with $i < j$

and so $T_i < T_j$) is accepted with probability $r_s = \min(1, \omega)$, where the acceptance probability (see (3.7))

$$\begin{aligned}
\omega &= \frac{c_{T_i}\pi(\theta_j)\mathcal{L}(\theta_j)^{\frac{1}{T_i}}\, c_{T_j}\pi(\theta_i)\mathcal{L}(\theta_i)^{\frac{1}{T_j}}}{c_{T_i}\pi(\theta_i)\mathcal{L}(\theta_i)^{\frac{1}{T_i}}\, c_{T_j}\pi(\theta_j)\mathcal{L}(\theta_j)^{\frac{1}{T_j}}} \\
&= \frac{\mathcal{L}(\theta_j)^{\frac{1}{T_i}}}{\mathcal{L}(\theta_j)^{\frac{1}{T_j}}} \cdot \frac{\mathcal{L}(\theta_i)^{\frac{1}{T_j}}}{\mathcal{L}(\theta_i)^{\frac{1}{T_i}}} \\
&= \mathcal{L}(\theta_j)^{\frac{1}{T_i}-\frac{1}{T_j}} \cdot \mathcal{L}(\theta_i)^{\frac{1}{T_j}-\frac{1}{T_i}} \\
&= \left(\frac{\mathcal{L}(\theta_j)}{\mathcal{L}(\theta_i)}\right)^{\frac{1}{T_i}-\frac{1}{T_j}}
\end{aligned} \tag{3.13}$$

only depends on *likelihood* values, not on the prior. A swap between chains means that the states θ_i and θ_j are exchanged between chains, or equivalently, that temperatures are mutually substituted [41, 49, 50]. From (3.13) you can see that if the 'hotter' chain j comes across greater likelihoods than chain i, a swap is always accepted; otherwise the acceptance probability ω depends on the ratio of likelihoods and the difference in (inverse) temperatures. So, while high-temperature chains are allowed to move more freely, any 'promising' parameter sets (with greater likelihoods) are preferably 'handed down' to chains with lower temperatures.

Note that from a computational point of view, the swapping steps come (almost) 'for free' (as opposed to the regular proposals), in the sense that they do not require the computation of new likelihood values, but only the comparison of previously computed ones. An advantage over simulated annealing is that the global optimum no longer needs to be found while the temperature is high, but instead the tempered distributions are sampled from all the time. The tradeoff of course is that several chains are run simultaneously, and only one of them samples from the actual target distribution. Also, instead of an annealing schedule the '*temperature ladder*' (number and levels of temperatures) needs to be specified (see also following section).

3.2.12 Implementing parallel tempering

Proposal distribution

Due to the 'variance inflation' effect of a tempering factor T_i, as sketched in equation (3.9), it makes sense to choose the proposal variance of each chain i proportional to $\sqrt{T_i}$, the square root of its temperature. This works well in practice, leading to roughly the same acceptance rate for all parallel chains.

Temperature ladders

In the following some advice will be given on choice of 'temperature ladders', i.e. how to choose the levels T_i and the total number of temperatures k. In [41], a temperature ladder $T_i = 1 + \lambda(i-1)$ for $i = 1, \ldots, k$ and $\lambda > 0$ is mentioned, but in the following a different spacing is proposed, as well as hints on choice of an appropriate number of steps k.

Step widths

From the motivation given in section 3.2.9 (especially equation (3.9)), one might expect that the parallel chains' temperatures T_i should be chosen as

$$T_i = q^{(i-1)} \qquad (q > 1,\ i = 1, \ldots, k), \tag{3.14}$$

so that the stationary distributions of all neighbouring chains (i and $i+1$) differ by the same 'inflation factor' $\sqrt{q}$.

In fact, this strategy has been justified in the case of parallel tempering applications for molecular simulations [51]. There it was argued that the expected acceptance probability of a swap proposed between neighbouring chains only depends on the ratio of corresponding temperatures. In the same context it was shown that the optimal expected acceptance probability for swaps between neighbouring chains is at about 23% [52].

The same property can be motivated for the case of sampling from an annealed posterior distribution as defined in (3.11) using some asymptotic

arguments. In the following, an approximate expression for the expected acceptance probability is derived, which turns out to only depend on the ratio of the involved chains' temperatures, and which is also useful to estimate the necessary temperature ratio to yield a given swap acceptance rate.

Consider a swap between two chains i and j, without loss of generality assuming that $i < j$, and consequently $T_i < T_j$. Let d be the dimension of the parameter space that is sampled from. Then the (marginal) posterior distribution of the sampled loglikelihood values for chain i is asymptotically given by

$$\log(\mathcal{L}(\theta_i)) \;=\; L_{\max} - T_i X \tag{3.15}$$

where $L_{\max}$ is the maximum achievable loglikelihood, and X is a random variable following a Gamma($\frac{d}{2}$, 1)-distribution [53, 54]. In particular, this implies that

$$\begin{aligned} \mathrm{E}\left[\log(\mathcal{L}(\theta_i))\right] &= L_{\max} - T_i \tfrac{d}{2} \qquad \text{and} && (3.16)\\ \mathrm{Var}\left(\log(\mathcal{L}(\theta_i))\right) &= T_i^2 \tfrac{d}{2}. && (3.17) \end{aligned}$$

The acceptance probability of a swap between chains i and j was given in equation (3.13) as $r_s = \min(1, \omega)$ where

$$\begin{aligned} \omega &= \left(\frac{\mathcal{L}(\theta_j)}{\mathcal{L}(\theta_i)}\right)^{\frac{1}{T_i}-\frac{1}{T_j}} \\ &= \exp\Bigg(\underbrace{\left[\tfrac{1}{T_i}-\tfrac{1}{T_j}\right]\left[\log(\mathcal{L}(\theta_j)) - \log(\mathcal{L}(\theta_i))\right]}_{=:\,Z}\Bigg). \end{aligned} \tag{3.18}$$

Following from the previous asymptotic expression (3.15), the first two moments of Z are then given by

$$\mathrm{E}[Z] \;=\; d\left(1 - \frac{1}{2}\left(\frac{T_j}{T_i} + \frac{T_i}{T_j}\right)\right), \tag{3.19}$$

$$\mathrm{Var}(Z) \;=\; d\left(1 - \left(\frac{T_i}{T_j} + \frac{T_j}{T_i}\right) + \frac{1}{2}\left(\left(\frac{T_i}{T_j}\right)^2 + \left(\frac{T_j}{T_i}\right)^2\right)\right), \tag{3.20}$$

and, as conjectured, these only depend on the temperature ratio of the two chains involved. Assuming a Normal distribution for Z, the resulting distribution of ω would be a Log-normal distribution with parameters $\mu = \mathrm{E}[Z]$ and $\sigma^2 = \mathrm{Var}(Z)$ as given above. Under that assumption, which should be a good approximation especially for larger values of d, the expected swap acceptance probability $\mathrm{E}[r_\mathrm{s}]$ is given by:

$$\begin{aligned}
\mathrm{E}[r_\mathrm{s}] &= \mathrm{E}[\min(1,\omega)] \;=\; \mathrm{E}[\min(1,\exp(Z))] \\
&= \mathrm{E}[\omega \,|\, \omega \leq 1] \;+\; \mathrm{P}(\omega > 1) \\
&= \mathrm{E}[\exp(Z) \,|\, Z \leq 0] \;+\; \mathrm{P}(Z > 0) \\
&= \frac{1}{\sqrt{2\pi}} \int_0^1 \exp\left(-\frac{(\log(w)-\mu)^2}{2\sigma^2}\right) \mathrm{d}w \\
&\quad + \frac{1}{\sqrt{2\pi}} \int_1^\infty \frac{1}{w} \exp\left(-\frac{(\log(w)-\mu)^2}{2\sigma^2}\right) \mathrm{d}w \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \exp(z) \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \mathrm{d}z \\
&\quad + \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \mathrm{d}z
\end{aligned} \tag{3.21}$$

which is relatively easy to compute numerically. Note that the above expected probability is a *long-term* average. Due to the correlation of subsequent samples within the actual MCMC there may be longer phases where every other proposed swap is accepted, and other phases without any swapping taking place.

This expression now allows one to approximate the expected acceptance rate of swaps between chains with a given ratio of temperatures, or between all neighbouring chains if the temperature ladder is set up with fixed ratio intervals as defined in equation (3.14). On the other hand, it also allows the determination of the temperature ratio to be used when aiming for a given acceptance rate, by (again, numerically) inverting the function. An example coded in R [55] is shown in Appendix A.2. Table 3.1 shows such approximate temperature ratios q that would lead to an expected swap acceptance of 25% for a posterior of dimension d. Note that

Table 3.1: Approximate temperature ratios q leading to a swap acceptance rate of 25% for a posterior of dimension d. The figure in brackets ($\sqrt{q}-1$) gives the corresponding 'variance inflation' (see (3.9)).

d	q	($\sqrt{q}-1$)	d	q	($\sqrt{q}-1$)	d	q	($\sqrt{q}-1$)
2	6.98	(164%)	16	1.80	(34%)	100	1.26	(12%)
3	4.41	(110%)	18	1.74	(32%)	200	1.18	(8.5%)
4	3.48	(87%)	20	1.69	(30%)	500	1.11	(5.3%)
5	2.99	(73%)	25	1.59	(26%)	1000	1.075	(3.7%)
6	2.69	(64%)	30	1.53	(24%)	2000	1.053	(2.6%)
7	2.48	(58%)	45	1.41	(19%)	5000	1.033	(1.6%)
8	2.33	(53%)	50	1.39	(18%)	10 000	1.023	(1.2%)
9	2.21	(49%)	60	1.35	(16%)	20 000	1.016	(0.82%)
10	2.12	(46%)	70	1.32	(15%)	50 000	1.010	(0.52%)
12	1.98	(41%)	80	1.29	(14%)	100 000	1.0073	(0.36%)
14	1.88	(37%)	90	1.27	(13%)	200 000	1.0051	(0.26%)

the swap acceptance rate may be seen as a measure of the similarity between the two corresponding distributions. The acceptance rate is equal to 1 if the two distributions are equal, and gets lower for differing distributions.

Ladder height

Given that one has decided to use a temperature ladder $T_i = q^{(i-1)}$ and chosen a temperature ratio $q > 1$, the only detail left to define is k, the total number of temperature levels or chains, and with that the level of the highest temperature T_k. Two questions may give some guidance on the choice of k:

Multiple modes: Parallel tempering is supposed to prevent the sampler from getting stuck in local modes. How 'deep' does one expect the 'valleys' to be that need to be crossed between local modes?

Prior sampling: The coolest chain (sampling at temperature $T_1 = 1$) is sampling from the posterior, while the hottest chain is supposed to ap-

proximately sample from the prior (cp. equation (3.11)). How high does T_k need to be in order to leave the sampler 'sufficiently unimpressed' by the likelihood contribution to the (tempered) stationary distribution of chain k?

In both cases one needs to ensure that the sampler, when sampling at temperature T_k, is able to advance below a certain level of posterior density, either to traverse to another mode, or to descend from the posterior mode down to the remainder of the prior domain. Having one chain basically sampling from the prior is supposed to ensure a realistic best possible chance of any point in parameter space being reached, regardless of the posterior's shape or the starting values.

Neglecting the influence of the prior (or assuming the prior density to be constant), and again resorting to the approximation (3.15), this is equivalent to deciding on how far below $L_{\max}$ the (log-) likelihoods from which the 'hottest' chain k is sampling are supposed to be. This range depends on the shape of the posterior distribution itself, and in particular on the value of $L_{\max}$, which in general is not known beforehand. Assuming one knew the value of the maximum achievable loglikelihood $L_{\max}$, the problem remains to either define a 'minimum likelihood' value L_{low} that should be within the range of the 'hottest' chain at temperature T_k, or to define parameters $\theta^{\emptyset}$, for which $\mathcal{L}(\theta^{\emptyset}) = L_{\text{low}}$. The most sensible definition of such parameters depends very much on the statistical problem at hand; in the case of (gravitational wave) signal detection it may make sense to consider the loglikelihood that is achieved for a parameter value $\theta^{\emptyset}$, indicating the presence of *no signal*. In practice, this is also the level around which the sampler keeps sampling when it has not yet started to converge, or the level that most arbitary parameter sets yield which do not imply a very large signal-to-noise ratio. In the same context, a value of $L_{\text{low}} = L_{\max} - 4(L_{\max} - \mathcal{L}(\theta^{\emptyset}))$ might be sensible, since this would be about the level that a signal with an inverted sign but otherwise correct parameters should yield. For other contexts, other values might make sense, e.g. in a regression problem the likelihood one gets when only fitting a constant term.

Given that one has chosen a $\theta^{\emptyset}$, one can then try to aim chain k's stationary distribution at least down at the level for $\theta^{\emptyset}$. Constraining its expected sampled likelihood (and with that its median, which is similar) leads to the following requirement depending on $\check{\theta}$, the true parameter value:

$$\begin{aligned} \mathrm{E}_{\mathrm{P}_{T_k}}\left[\log(\mathcal{L}(\theta))\right] &= \mathrm{E}[L_{\max} - T_k\tfrac{d}{2}] && (3.22)\\ &= \mathrm{E}[L_{\max}] - T_k\tfrac{d}{2} && (3.23)\\ &= \log(\mathcal{L}(\check{\theta})) + \tfrac{d}{2} - T_k\tfrac{d}{2} && (3.24)\\ &= \log(\mathcal{L}(\check{\theta})) - (T_k - 1)\tfrac{d}{2} \overset{!}{\leq} \log(\mathcal{L}(\theta^{\emptyset})) && (3.25) \end{aligned}$$

so that

$$T_k = q^{k-1} \geq \tfrac{2}{d}\Big(\log(\mathcal{L}(\check{\theta})) - \log(\mathcal{L}(\theta^{\emptyset}))\Big) + 1 \qquad (3.26)$$

is a function of the (logarithmic) likelihood ratio $\log\left(\frac{\mathcal{L}(\check{\theta})}{\mathcal{L}(\theta^{\emptyset})}\right)$ for true and 'null' parameters. This likelihood ratio again is a random variable that depends on the actual noise realisation in the data, but is constant for a given data set. The figure $D = 2\log\left(\frac{\mathcal{L}(\check{\theta})}{\mathcal{L}(\theta^{\emptyset})}\right)$ is also known as the *deviance* with respect to the 'null' and 'alternative' point hypotheses $\theta^{\emptyset}$ and $\check{\theta}$, and is related to the evidence in favour of $\check{\theta}$ (against $\theta^{\emptyset}$) in the data [56, 57].

Transforming that (log-) likelihood ratio allows for some insight into its nature:

$$\log\left(\frac{\mathcal{L}(\check{\theta})}{\mathcal{L}(\theta^{\emptyset})}\right) \tag{3.27}$$

$$= \log\left(\frac{p(y|\check{\theta})}{p(y|\theta^{\emptyset})}\right) \tag{3.28}$$

$$= \log\left(\frac{\frac{p(\check{\theta}|y)}{p(\check{\theta})}p(y)}{\frac{p(\theta^{\emptyset}|y)}{p(\theta^{\emptyset})}p(y)}\right) \tag{3.29}$$

$$= \log\left(\frac{p(\check{\theta}|y)}{p(\theta^{\emptyset}|y)}\right) - \log\left(\frac{p(\check{\theta})}{p(\theta^{\emptyset})}\right) \tag{3.30}$$

$$= \log\left(\frac{p(\check{\theta}|y)}{p(\check{\theta})}\right) - \log\left(\frac{p(\theta^{\emptyset}|y)}{p(\theta^{\emptyset})}\right) \tag{3.31}$$

$$\approx \underbrace{\mathrm{E}_{\mathrm{P}(\theta|y)}\left[\log\left(\frac{p(\theta|y)}{p(\theta)}\right)\right]}_{=\mathcal{D}(p(\theta|y)\|p(\theta))} - \log\left(\frac{p(\theta^{\emptyset}|y)}{p(\theta^{\emptyset})}\right). \tag{3.32}$$

Thus the required ladder height depends on the ratios in (3.30) and (3.31), which are related to the *relative entropy* (or *Kullback-Leibler distance*) $\mathcal{D}(p(\theta|y)\|p(\theta))$ between prior and posterior [42, 58].

In the special (but common) case where the data are assumed to be the sum of a model term s and noise

$$y_i = s_i(\theta) + \varepsilon_i \tag{3.33}$$

where $i = 1, \ldots, N$, and ε_i i.i.d. with $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$, and the likelihood is of the form

$$\mathcal{L}(\theta) = \exp\left(-\sum_{i=1}^{N}\frac{(y_i - s_i(\theta))^2}{2\sigma_i^2}\right), \tag{3.34}$$

the above likelihood ratio term can be decomposed further:

$$\begin{aligned}
& \log(\mathcal{L}(\check{\theta})) - \log(\mathcal{L}(\theta^{\emptyset})) && (3.35)\\
= & -\sum_{i=1}^{N} \frac{(y_i - s_i(\check{\theta}))^2}{2\sigma_i^2} + \sum_{i=1}^{N} \frac{(y_i - s_i(\theta^{\emptyset}))^2}{2\sigma_i^2} && (3.36)\\
= & -\sum_{i=1}^{N} \frac{(s_i(\check{\theta}) + \varepsilon_i - s_i(\check{\theta}))^2}{2\sigma_i^2} + \sum_{i=1}^{N} \frac{(s_i(\check{\theta}) + \varepsilon_i - s_i(\theta^{\emptyset}))^2}{2\sigma_i^2} && (3.37)\\
= & -\sum_{i=1}^{N} \frac{\varepsilon_i^2}{2\sigma_i^2} + \sum_{i=1}^{N} \frac{(s_i(\check{\theta}) - s_i(\theta^{\emptyset}) + \varepsilon_i)^2}{2\sigma_i^2} && (3.38)\\
= & -\sum_{i=1}^{N} \frac{\varepsilon_i^2}{2\sigma_i^2} + \sum_{i=1}^{N} \frac{(s_i(\check{\theta}) - s_i(\theta^{\emptyset}))^2 + 2\varepsilon_i(s_i(\check{\theta}) - s_i(\theta^{\emptyset})) + \varepsilon_i^2}{2\sigma_i^2} && (3.39)\\
= & \underbrace{\sum_{i=1}^{N} \frac{(s_i(\check{\theta}) - s_i(\theta^{\emptyset}))^2}{2\sigma_i^2}}_{=:A} - \underbrace{2\sum_{i=1}^{N} \frac{\varepsilon_i(s_i(\check{\theta}) - s_i(\theta^{\emptyset}))}{2\sigma_i^2}}_{=:B} && (3.40)
\end{aligned}$$

where A is constant (for given $\check{\theta}$), and B is a random variable (depending on the noise realisation) with

$$\begin{aligned}
\mathrm{E}[B] &= 0 \qquad \text{and} && (3.41)\\
\mathrm{Var}(B) &= 2\sum_{i=1}^{N} \frac{\left(s_i(\check{\theta}) - s_i(\theta^{\emptyset})\right)^2}{2\sigma_i^2} \;=\; 2\,A. && (3.42)
\end{aligned}$$

So the likelihood ratio depends completely on the noise, which splits up into 'deterministic' and 'random' contributions A and B: A depends on the noise parameters, and B depends on the actual noise realisation. The required 'ladder height' $A + B$ then is a random variable with

$$\mathrm{E}[A + B] \;=\; A \qquad \text{and} \qquad \mathrm{Var}(A + B) \;=\; 2\,A. \tag{3.43}$$

In anticipation of section 4.6.3, one can see that A actually is closely related to the signal-to-noise ratio (SNR). If the likelihood is of the form as in (3.34) (or (4.28)), and $\theta^{\emptyset}$ was defined so that $s_i(\theta^{\emptyset}) \equiv 0$ for all i, as suggested

previously, then

$$A \propto \rho(\check{\theta})^2, \tag{3.44}$$

that is, the expectation and variance of the required ladder height are proportional to the (squared) SNR of the signal present in the data.

Conclusion

In conclusion, since the appropriate choice of k depends on properties of the data (especially the unknown true parameters) that usually are not known beforehand, one can choose k

- based on worst-case considerations,
- based on integration over the prior, or
- adaptively, i.e. add higher temperature chains as the MCMC keeps coming across greater likelihood values.

As long as there is no universally practical value for k available, the latter method should be a good choice, because once the sampler has converged, k should not need to be increased any more, and so the sampler's ergodicity properties would not be affected for the eventual posterior sample.

The step widths were derived based on the approximation in equation (3.15), which does not consider the prior's contribution to the posterior. However, for sufficiently large values of T, any neighbouring chains should both be roughly sampling from the prior, and so the acceptance probability should then be close to 1 as the distributions get increasingly similar. The approximate acceptance rate associated with a ladder setup as in (3.14) may consequently serve as a lower bound, and one should expect the actual acceptance rates to be larger as T increases.

Parallel tempering is supposed to improve both *mixing* and *convergence*. Suppose one has decided to use a temperature laddder as in (3.14). If one does not worry about convergence, but is only interested in improved mixing, the best choice of k might not depend on the likelihood ratio as argued in the previous section, as there might not be a need for

the sampler to reach down to certain likelihood levels, and consequently a lower value of k could be chosen. If on the other hand convergence is the primary concern, then the exact value of the temperature ratio might be less relevant. If, due to limited resources, one is only able to run a certain number of parallel chains, it might make more sense to 'stretch' the ladder (so that it still reaches down to the same level, but with larger steps sizes), than to use a ladder that has 'optimal' step sizes, but is too short. A swap acceptance rate that is constant across different neighbouring pairs of chains, and significantly $\gg 0$ should probably still be beneficial.

3.2.13 Evolutionary MCMC

The evolutionary MCMC algorithm [59] is an extension of parallel tempering, implementing additional proposals that are motivated by *Genetic Algorithms* [60]. 'Recombinations' of parameter samples from different MCMC chains are used as proposals in order to improve convergence and mixing.

Genetic algorithms are motivated by evolutionary principles, and in this context such terms as 'population', 'parents', 'crossover' and 'offspring' are frequently used. Applying this terminology to the parallel tempering algorithm, the set of MCMC chains running at different temperatures constitute the *population*, within which the individuals differ by their *genomes* (parameter values), and *evolve* over time t. Randomly, *mutations* (proposals) are formed, that are subject to the principle of *survival of the fittest* (acceptance/rejection). Evolutionary MCMC then introduces additional proposals in analogy to the *recombination* common in evolution. Two *parental* individuals are selected to 'mate', i.e. have their *genes* (parameters) recombined in order to form new *offsprings*.

Recombination of two parameter sets is implemented in two ways: the '*real crossover*', in which single elements of the two parental parameter sets are swapped in order to form two new offsprings, and the '*snooker crossover*', in which a single 'offspring' parameter set is proposed that lies on the line passing through the two 'parental' points in the parameter

space (see also [61]).

Note that with the introduction of crossovers the 'parallel sampling' property that was outlined in the last paragraph of section 3.2.8 is compromised.

3.2.14 Importance sampling

Importance sampling is another Monte Carlo method to approximate integrals. Consider the case where one wants to compute an expectation like the one in equation (3.2). Instead of obtaining a sample from density $f(\theta)$ as in the general case of Monte Carlo integration, one can also use a sample $\theta_1, \ldots, \theta_N$ from a distribution $g \approx f$ similar to the one actually sought. An estimate of the desired integral is then obtained by computing the *weighted* average

$$\overline{h(\theta)} \;=\; \frac{1}{\sum_{j=1}^{N} w_j} \sum_{i=1}^{N} w_i \, h(\theta_i), \tag{3.45}$$

where

$$w_i \;=\; \frac{f(\theta_i)}{g(\theta_i)} \tag{3.46}$$

are the *importance ratios* or *importance weights*. Both densities f and g do not need to be normalised here. The accuracy of this approximation depends heavily on the similarity of g and f. At best, all importance ratios would be of roughly the same size, at worst there are few w_i concentrating most of the total weight, or there are rare but important cases with $f(\theta) \gg g(\theta)$ completely missed in the sample. Note that if the densities f and g are given in normalised form, the estimate in (3.45) can be simplified to $\frac{1}{N} \sum_{i=1}^{N} w_i \, h(\theta_i)$ [42].

3.2.15 Importance sampling and parallel tempering

When using parallel tempering, the 'hot' chains are also sampling from a distribution similar to the actual target distribution, so instead of dispos-

ing of these samples, one can also use them for importance sampling. The resulting importance weights then are:

$$w_i \;=\; \frac{\pi(\theta_i)\mathcal{L}(\theta_i)}{\pi(\theta_i)\mathcal{L}(\theta_i)^{\frac{1}{T}}} \;=\; \mathcal{L}(\theta)^{(1-\frac{1}{T})}. \tag{3.47}$$

The magnitudes of these weights are ideally close to one, which they are especially for $T \to 1$. From equation (3.15) one can see that for greater values of T, $\mathrm{Var}(\log(w_i)) \approx \frac{d}{2}(1-\frac{1}{T})^2 \propto T$. Note that since the normalising constants to the tempered distributions are (in general) unknown and depend on the temperature T, the average needs to be normalised by the sum of all weights as in (3.45) and does not simplify.

3.2.16 Importance resampling

Importance resampling is closely related to importance sampling. It is a method for generating a sample that approximately follows a given distribution $f(\theta)$ when sampling from f directly is not possible, but an approximation g to f is available. One starts off by drawing a (comparatively large) sample $\theta_1, \ldots, \theta_N$ from the approximate target distribution $g(\theta)$. Then for each element θ_i the *importance ratio* $w_i = \frac{f(\theta_i)}{g(\theta_i)}$ is computed and the eventual sample of size $n \ll N$ is drawn out of $\{\theta_1, \ldots, \theta_N\}$ without replacement and with probabilities proportional to the weights w_i [42].

3.2.17 Implementing importance resampling

Sampling distributions

When importance resampling is used to generate starting points for MCMC chains, it might be desirable for these to be rather slightly 'overdispersed'. This can be procured by chosing an overdispersed approximation $g(\theta)$ [42, 46].

The best available approximation to the posterior will in general be the prior distribution (as long as one does not yet consider the data). However,

it might not necessarily be the best choice from which to draw starting values. When using importance resampling for MCMC starting values, the reasoning behind the method is not necessarily to get that initial guess *as close to the true values as possible*, but rather to draw starting values *such that the probability of convergence to the correct mode is highest*. There may be parts of the parameter space where only a very rough guess is sufficient for reliable convergence, while in other parts the posterior mode must be targetted very accurately in order to be eventually found. Consequently, the sampling should be denser in the latter regions and may be coarser in the former regions in order to have uniformly good chances of convergence—which may well lead to a sampling distribution different from the prior distribution.

Minimising memory usage

If the approximation g has a much wider mode than the target distribution f, the initial sample $\{\theta_1, \ldots, \theta_N\}$ will be made up of a large majority of elements that have a very low importance ratio w_i (since $f(\theta_i) \ll g(\theta_i)$), while only a small minority actually have a significant chance of eventually being drawn. In order to save memory and simplify management of the initial sample, it is of interest to sort out such samples and keep only those that have a reasonable chance of getting into the eventual sample.

One idea to achieve this is to restrict to those samples θ_j for which:

$$f(\theta_j) \geq \left(\max_i f(\theta_i)\right) \exp(-\delta) \tag{3.48}$$

$$\Leftrightarrow \quad \log(f(\theta_j)) \geq \left(\max_i \log(f(\theta_i))\right) - \delta \tag{3.49}$$

for some $\delta > 0$. That is, one only considers samples whose density differs by less than a factor of $\exp(\delta)$ from the maximum density so far, or equivalently, whose log-density is within a certain range δ of the greatest log-density reached so far. Since the maximum (log-) density in the sample can only increase while sampling, this allows to sort out samples imme-

diately while sampling and manage only a handful of samples while the effective sample size N is much larger.

The question of course is how to choose such a threshold δ. The approximate posterior distribution of log-likelihood values (equation (3.15)) provides a clue to what range of values below the maximum log-likelihood $L_{\max}$ one should expect for a distribution over a d-dimensional parameter space. Assuming that the posterior distribution of $\log(f(\theta))$ behaves similarly to that of $\log(\mathcal{L}(\theta))$, one can then e.g. set δ to at least 3 standard deviations above the mean difference from $L_{\max}$:

$$\delta \gtrsim \frac{d}{2} + 3\sqrt{\frac{d}{2}}, \tag{3.50}$$

which in practice yields the desired effect of identifying a small number of more or less reasonable samples from the vast majority of virtually impossible draws.

Technically, what happens when setting a limit δ is that (assuming continuity of the density $f(\theta)$) parameter values θ that concentrate less than $\exp(-\delta)$ times as much probability within the immediate neighbourhood around themselves, when compared to the 'best' so far, are identified.

3.3 Reparametrisation: transformation of random variables

In order to improve the efficiency of an MCMC sampler it is often sensible to use reparametrisations. Parameters may be highly correlated in their original form, and re-expressing them in different scales may yield a posterior distribution that is then easier to sample from [42]. When moving over from one domain to another, the transformation's effect on the (prior/posterior) density function needs to be accounted for as follows.

Let X be a continuous random variable with density $f_X(x)$ and domain $\mathcal{A} = \{x : f_X(x) > 0\}$. Let $y = g(x)$ be a one-to-one mapping from $\mathcal{A}$ to $\mathcal{B}$. If the derivative of $x = g^{-1}(y)$ is continuous and non-zero for all

$y \in \mathcal{B}$, then the distribution of $Y = g(X)$ has the density

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)) \tag{3.51}$$

More generally, if X and Y are d-dimensional, the density of Y is given by

$$f_Y(\vec{y}) = |\det(J(\vec{y}))| f_X(g^{-1}(\vec{y})). \tag{3.52}$$

where $|\det(J(\vec{y}))|$ is the absolute determinant of $J(\vec{y})$, the Jacobian of the inverse transformation $\vec{x} = g^{-1}(\vec{y})$ as a function of $\vec{y}$:

$$J(\vec{y}) = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial y_1} & \cdots & \frac{\partial x_d}{\partial y_d} \end{pmatrix} \tag{3.53}$$

[40, 42].
Note that for the special case of a 2×2-matrix the determinant is:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc. \tag{3.54}$$

The corresponding terms required for transformations used here are derived in appendix A.3.

In some cases assuming certain priors for parameters is equivalent to assuming a certain transformation of the parameter as the actual parameter which then has a uniform prior. Consider for example a parameter $\theta \in [0, \pi]$ that has a prior that is proportional to $\sin(\theta)$. When instead one uses its cosine, $\cos(\theta)$, as the actual parameter, then the necessary transformation coefficient (see appendix A.3) cancels with the prior density; so in this case using θ in conjunction with the sine prior is equivalent to using $\cos(\theta)$ with a uniform prior.

3.4 Fourier transformation in theory and application

3.4.1 Fourier transform

The Fourier transform (FT) allows the transformation of a function back and forth between *time* and *frequency* domains. Let h be a real-valued function $h : \mathbb{R} \to \mathbb{R}$; then its Fourier transform $\tilde{h} : \mathbb{R} \to \mathbb{C}$ is given by:

$$\tilde{h}(f) \;=\; \int_{-\infty}^{\infty} h(t)\exp(-2\pi \mathrm{i} f t)\mathrm{d}t \tag{3.55}$$

where h is considered a function of time while $\tilde{h}$ is a function of frequency. Since h is a real-valued function, $\tilde{h}$ is symmetric ($\tilde{h}(-f) = \overline{\tilde{h}(f)}$), allowing attention to be restricted to (non-redundant) positive frequencies and to consider $\tilde{h}$ as $\tilde{h} : \mathbb{R}^+ \to \mathbb{C}$. The back-transformation from $\tilde{h}$ to h is given by:

$$h(t) \;=\; \int_{-\infty}^{\infty} \tilde{h}(f)\exp(2\pi \mathrm{i} f t)\mathrm{d}f. \tag{3.56}$$

In view of applications in later sections, the relationships decribed in the following are of interest. Let '$\rightleftharpoons$' denote a transform pair, let

$$h(t) \;\rightleftharpoons\; \tilde{h}(f) \qquad \text{and} \qquad g(t) \;\rightleftharpoons\; \tilde{g}(f) \tag{3.57}$$

and let

$$(g * h)(t) \;=\; \int_{-\infty}^{\infty} g(t-\tau)\, h(\tau)\mathrm{d}\tau \tag{3.58}$$

define the *convolution* of g and h, where $g * h = h * g$. Then the following properties apply:

$$a g(t) + h(t) \;\rightleftharpoons\; a\tilde{g}(f) + \tilde{h}(f) \quad \text{(linearity)} \tag{3.59}$$

$$(g * h)(t) \;\rightleftharpoons\; \tilde{g}(f)\,\tilde{h}(f) \quad \text{(convolution theorem)} \tag{3.60}$$

for $a \in \mathbb{R}$ [45, 62]. Convolutions will be important especially in the following section. Note that there are several different definitions of the Fourier

transform, varying in signs or normalising constants.

3.4.2 Discrete Fourier transform

The analogue to the Fourier transform for functions that are sampled at discrete time points is the *Discrete Fourier Transform (DFT)*, which replaces the integral (3.55) by a sum. The input is a time series of length N (even) and sampling rate $\frac{1}{\Delta_t}$ (or resolution / sampling interval Δ_t):

$$\{h(t) \in \mathbb{R} : \; t = 0, \Delta_t, 2\Delta_t, \ldots, (N-1)\Delta_t\} \tag{3.61}$$

which the transform then maps to

$$\{\tilde{h}(f) \in \mathbb{C} : \; f = 0, \Delta_f, 2\Delta_f, \ldots, (N-1)\Delta_f\}, \tag{3.62}$$

where $\Delta_f = \frac{1}{N\Delta_t}$ and

$$\tilde{h}(f) \;=\; \Delta_t \sum_{j=0}^{N-1} h(j\Delta_t) \exp(-2\pi \mathrm{i} j f). \tag{3.63}$$

Since h is real-valued, for $j = 1, \ldots, \frac{N}{2} - 1$ the elements of $\tilde{h}$ are symmetric (and with that redundant) in the sense that $\tilde{h}(i\Delta_f) \;=\; \overline{\tilde{h}((N-i)\Delta_f)}$. The elements $\tilde{h}(0)$ and $\tilde{h}(\frac{N}{2}\Delta_f)$ of the DFT are always purely real, i.e. $\mathrm{Im}\big(\tilde{h}(0)\big) \;=\; \mathrm{Im}\big(\tilde{h}(\frac{N}{2}\Delta_f)\big) \;=\; 0$. The back-transformation to the time domain, the *inverse DFT*, is given by

$$h(t) \;=\; \Delta_f \sum_{j=0}^{N-1} \tilde{h}(j\Delta_f) \exp(2\pi \mathrm{i} j t) \tag{3.64}$$

[63]. The actual transforms are most commonly carried out using the *Fast Fourier Transform (FFT)* algorithm [45, 62]. Note that the '*FFTW*' implementation [64] returns the unnormalised transforms (not including the factors of Δ_t or Δ_f, i.e., one gets $\frac{1}{\Delta_t}\tilde{h}(f)$ instead of $\tilde{h}(f)$ as defined in (3.63)).

3.4.3 Windowing and convolution

Several problems arise when approximating the FT of a continuous function by the DFT of samples from that function. The DFT can only resolve frequencies below the *Nyquist (critical) frequency* $f_c = \frac{1}{2\Delta_t}$. If the FT of the actual function is non-zero for frequencies above f_c, this will still have an effect on the DFT. Also, any contribution of frequencies that do not exactly coincide with one of the f_n will result in *leakage* into neighbouring frequency bins [45].

The first problem may not appear if the examined function is known to be *bandwidth-limited* to frequencies below the Nyquist frequency, which might e.g. be the case when dealing with a signal that has passed a low-pass filter. For the second problem there are (at least) two motivations:

Discontinuity: When applying the DFT, it is implicitly assumed that the transformed points constitute a periodic function with period $N\Delta_t$. Any frequency contribution that does not coincide with one of the f_n (i.e. whose period is not a divisor of $N\Delta_t$) contributes to a discontinuity between the 'endpoints' of the function; this discontinuity then is the cause of the leakage [65].

Convolution: Transforming only a finite stretch of a function with infinite support is equivalent to transforming the product of the actual function and a 'rectangular' *windowing function* that is equal to 1 for the analysed interval and 0 otherwise. This 'windowing' in the time domain is equivalent to a convolution in the frequency domain, causing the spectral leakage [62].

The leakage effect can be reduced by clever choice of a (non-rectangular) windowing function that—following the former motivation—'*matches as many orders of derivative (of the weighted data) as possible at the boundary*', or—following the latter motivation—a window whose Fourier Transform behaves (in some sense) favourably, so that the convolution has the smallest possible undesirable effect [65].

Much of the literature that deals with optimal choice of windowing functions is primarily concerned with its effects when estimating spectral densities, or when trying to detect/resolve isolated sinusoidal signals (of constant amplitude) within noise. Directions given there may not necessarily be applicable for the case of the Fourier transformation of a GW signal, which evolves over time and for which time and phase information (real *and* imaginary components of the transformed signal) matter. This becomes especially important when one is matching *numerically* FT'd data with *analytically* FT'd waveform templates (and less important when both FTs are performed numerically).

The *rectangular window*, which is implicitly applied when 'no' window is used, applies the following weights

$$w_N(i) \;=\; \begin{cases} 1 & \text{for} \quad 0 \leq i \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{3.65}$$

to the individual observations $h(i)$ in the time domain. This is also known as the 'square' or 'boxcar' window.

For spectral density estimation purposes the *Hann window* is used, which is defined as:

$$w_N(i) \;=\; \begin{cases} \frac{1}{2}(1-\cos(\frac{i}{N}2\pi)) & \text{for} \quad 0 \leq i \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{3.66}$$

[65].

When Fourier-transforming the actual GW data (or signal templates), a window is used which is called the *Tukey window* (or *cosine-tapered window* or *split cosine bell window*). It possesses an additional parameter $\alpha \in [0,1]$ and is defined as:

$$w_N(i) \;=\; \begin{cases} \frac{1}{2}\left(1-\cos(\pi\frac{i}{\frac{\alpha}{2}N})\right) & \text{for} \quad 0 \leq i \leq \frac{\alpha}{2}N \\ 1 & \text{for} \quad \frac{\alpha}{2}N \leq i \leq (1-\frac{\alpha}{2})N \\ \frac{1}{2}\left(1-\cos(\pi\frac{N-i}{\frac{\alpha}{2}N})\right) & \text{for} \quad (1-\frac{\alpha}{2})N \leq i \leq N-1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.67}$$

The parameter α denotes the fraction of the window in which it behaves sinusoidally; for $\alpha = 0$ it is equal to the rectangular window, and for $\alpha = 1$ it equals the Hann window [65, 66, 67]. Its specific advantage is that it leaves the middle $(1 - \alpha)$ fraction of the data unaltered while 'mending' the discontinuity at the ends. (Note that there is another—different—window which is sometimes also referred to as the Tukey window.)

3.4.4 Power spectral density

The (one-sided) power spectral density $S_h(f)$ of a (real-valued) function $h(t)$ is defined as

$$S_h(f) \;=\; 2\,|\tilde{h}(f)|^2 \qquad \text{for } f \geq 0. \tag{3.68}$$

Note that there are a range of differing definitions of the spectral density, most notably, the *two-sided* spectrum is usually defined as above, but without the factor of 2 (which is due to the fact that more generally one can compute the spectrum by summing over positive *and* negative frequencies, which happen to be symmetric for real-valued $h(t)$, and hence is equivalent to summing over positive frequencies only and then multiplying by 2) [45].

3.4.5 Power spectral density estimation via the DFT

The power spectral density (PSD) of a time series can be estimated from its empirical discrete Fourier transform. The discrete analogue of the PSD for a discrete, real-valued function h as in (3.61) is

$$S_h(f) \;=\; \begin{cases} \frac{2}{N}\,|\tilde{h}(f)|^2 & \text{for } f = \Delta_f, 2\Delta_f, \ldots, (\frac{N}{2} - 1)\Delta_f \\ \frac{1}{N}|\tilde{h}(f)|^2 & \text{for } f = 0 \text{ or } f = \frac{N}{2}\Delta_f. \end{cases} \tag{3.69}$$

When dealing with larger amounts of data, it makes sense not to transform the complete data at once but to divide it into smaller segments, transform these, and average over the individual results. This may speed up computations, save memory, and result in less variance in the resulting estimate. The tradeoff on the other hand is a lower frequency resolution,

although this should not be a problem if the spectrum is assumed to be rather smooth.

The data segments can be chosen so that these overlap, with each data segment individually windowed before transformation. Figure 3.1 illus-

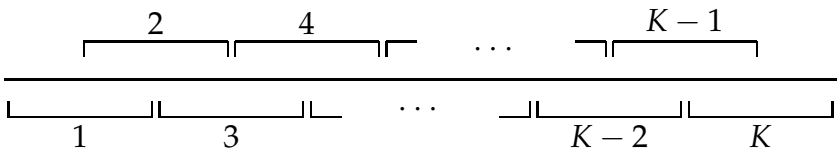

Figure 3.1: Dividing a time series into K overlapping segments.

trates such a segmentation. Let $\{x(t) : t = 0, \ldots, N-1\}$ be the original time series, which is subdivided in K segments of length L (L even), such that $(K+1)\frac{L}{2} = N$. Then $\{y_j(t) : t = 0, \ldots, L-1\}$, where $y_j(t) = w_L(t)x((j-1)L+t)$, is the windowed jth segment ($j = 1, \ldots, K$). Define $w_{ss} = \sum_{i=0}^{L-1}(w_L(i))^2$ as the sum of squared windowing coefficients (Note: $w_{ss} = L$ for 'no' windowing, i.e. the rectangular window). Then

$$\widehat{S}_n(f) = \frac{2}{K\,w_{ss}} \sum_{j=1}^{K} |\tilde{y}_j(f)|^2 \tag{3.70}$$

is an estimator for the (one-sided) Power Spectral Density at the dicscrete set of frequencies $\{f = \Delta'_f, 2\Delta'_f, \ldots, \frac{L-1}{2}\Delta'_f\}$, where $\Delta'_f = \frac{1}{L\Delta_t}$. In analogy to (3.69), the factor of 2 again drops out for $f = 0$ and $f = \frac{L}{2}\Delta'_f$ [68, 69, 45].

3.5 Downsampling and filtering

When working with (noiseless) discretely sampled data containing a band-width-limited signal, where the upper frequency limit is below the Nyquist frequency f_c, the data can be downsampled without loss of information. If the data is noisy, downsampling would move noise in the part of the spectrum between 'old' and 'new' Nyquist frequency below the new Nyquist frequency. In such a case, it makes sense to low-pass filter the data before

downsampling, because it will reduce noise in the downsampled data, and will not affect the signal. A general algorithm to do this is given in [70]. Procedures to design filters satisfying certain optimality criteria are described in [71].

3.6 Density estimation and confidence regions

The estimated densities shown in later sections (e.g. figure 5.10) are *kernel density estimates* [72]. Kernel density estimates are similar to histograms, but they return a continuous estimate of the density function (and not a step function). As with histograms, these can also be generalised to the 2-dimensional case.

Kernel density estimates are also used to derive 2-dimensional confidence regions (e.g. figure 5.11). In order to construct these, first a 2D histogram based on $k \times k$ bins was constructed, with $n_{i,j}$ giving the number of samples in the bin indexed by $i, j \in \{1, \ldots, k\}$. In order to construct a $1 - \alpha$ confidence region (where α denotes the confidence level; e.g. $\alpha = 0.05$ for a 95% confidence region), one *could* then use a contour line at a level n_α, enclosing the highest histogram bars that accumulate $(1 - \alpha)$ of observations. This would be defined as:

$$n_\alpha \;=\; \max\{n \in \mathbb{R}^+ : \tfrac{1}{N} \sum_{i,j:n_{i,j}>n} n_{i,j} > 1 - \alpha\}. \tag{3.71}$$

Since the heights of the histogram bars have a large variance, the resulting confidence region would have a very frayed appearance. Hence, in order to have a smoother contour, it is better to use a kernel density estimate as a smoothed version of the histogram. The kernel density estimate is computed and evaluated at each histogram bin's mid-point, with $f_{i,j}$ giving the estimated density at the (i, j)th bin. The confidence region is then constructed by determining a threshold density value f_α as

$$f_\alpha \;=\; \max\{f \in \mathbb{R}^+ : \tfrac{1}{N} \sum_{i,j:f_{i,j}>f} n_{i,j} > 1 - \alpha\} \tag{3.72}$$

and drawing the corresponding contour line at f_α into the kernel density plot. The *shape* of the contour line is then based on the kernel density estimate, while its *level* is based on the histogram. So its shape is smoothed, but still it always includes the desired fraction of samples. The resulting credibility region is then (an estimate of) a *$(1-\alpha)$ highest posterior density region* [42].

3.7 Recursive mean and covariance estimation

When trying to estimate variances or covariances of MCMC samples, these samples will in general not be available all at once in the computer memory. Furthermore, it might be desirable to have an intermediate estimate of variances/covariances available at any time while sampling is still going on. Both these motivations suggest not to use the straightforward formulas (the so-called 'textbook algorithm'), but to rather compute these 'on-the-fly', updating the estimate recursively with every new sample coming in. Formulas for estimating variances and covariances in a recursive manner are given in [73, 74].

3.8 Spherical statistics

When trying to characterise data that can be interpreted as locations on spheres or circles, the usual descriptive measures like mean or variance do not in general make sense, or only make sense if the distribution is confined to a very small region on the sphere so that the exact topology does not significantly differ from a linear manifold. Generalisations of these measures for 'spherical data' are derived in [75]; the analogues to mean and variance (*mean direction* and *spherical variance*) are reproduced in appendix A.4.

3.9 Parallel programming

Parallel programming means the simultaneous use of several processors (or processes) for computation, as opposed to a strictly sequential computation. A general introduction to parallel programming is given in [76]. The implemention of parallel tempering in a parallel fashion, so that different chains are running as different processes, can be done using *message passing* methods. This means that every single process manages a single chain, just as in a simple Metropolis-Hastings implementation, but that every iteration 'messages' are exchanged between processes to carry out the exchange of parameters or temperatures. A common protocol for message passing is the *Message Passing Interface (MPI)* [77].

Chapter 4

Model components

4.1 Data

In both cases of earth-bound or space-bound GW measurements, the retrieved data are first of all time series. In a network of earth-bound interferometers, each individual interferometer produces a data stream indicating the measured phase difference of the laser beams that went along its two arms. The planned space-based LISA interferometer will produce several data streams for each of its 3 satellites, which will probably usually be combined into 3 time series, the so-called *Time Delay Interferometry (TDI)* variables [78]. Sampling frequencies are of the order of 16 384–20 000 Hz for ground-based interferometers, and expected to be approximately $\frac{1}{15}$ Hz for LISA.

Since the sensitivity and the way different signals are modulated varies with the relative orientation of the interferometer and the passing gravitational wave, information about the exact time of measurement is vital. Together with the geographical location of earth-bound interferometers, or the orbital parameters for space-bound observations, this determines the instruments' locations and orientations during measurement.

A ground-based interferometer measures a passing gravitational wave 'instantaneously', and its output is basically a linear combination of the passing wave's plus- and cross-polarisations, depending on the interfer-

ometer's orientation with respect to the wave. The space-based LISA interferometer on the other hand is so large that it will take a photon roughly 15 seconds to travel from one satellite to another, and so the output will not (except in the limit of very low frequencies) be such a simple '1:1' mapping of the waveform. In fact, the TDI variables mentioned above will have an *eight-pulse-response* to a passing delta function shaped wave [78]. More details on TDI are given in section 4.4 below.

4.2 Parameters and parametrisations

4.2.1 General

The gravitational waves that are *emitted* from an inspiralling binary system depend on certain properties of the system, like masses of involved objects etc. The signal that is *measured* at a given point in space then also depends on the distance and orientation of inspiral and interferometer with respect to each other. Waves originating from the same event will result in different signals when observed from different directions. Due to the great distances in which such events generally happen, the signals measurable at different point within the solar system will not differ significantly, except for delays in the arrival time (or Doppler effects, if the receiver is in motion).

In the context of the models referred to in the following, the nine parameters determining the form of the wave originating from a binary inspiral event that passes a point in the solar system are

- the individual masses ($m_1, m_2 \in \mathbb{R}^+;\ m_1 \leq m_2$),
- luminosity distance ($d_L \in \mathbb{R}^+$),
- inclination angle ($\iota \in [0, \pi]$),
- coalescence phase ($\phi_0 \in [0, 2\pi]$),
- coalescence time at geocentre ($t_c^{\oplus} \in \mathbb{R}$)
  or coalescence time at solar system barycentre ($t_c^{\odot} \in \mathbb{R}$),

- declination ($\delta^{\oplus} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$)
  or ecliptic latitude ($\beta^{\odot} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$),
- right ascension ($\alpha^{\oplus} \in [0, 2\pi]$))
  or ecliptic longitude ($\lambda^{\odot} \in [-\pi, \pi]$) and
- polarisation angle w.r.t. earth frame ($\psi^{\oplus} \in [0, \pi]$)
  or w.r.t. ecliptic frame ($\psi^{\odot} \in [0, \pi]$).

The evolution of the inspiral event over time (orbits, velocities, etc.) is determined by the masses m_1 and m_2 of the two inspiralling compact objects. The event's gravitational wave 'signature' then is different when perceived from different directions, and in particular it makes a difference whether it is viewed from above, within, or below its orbital plane. The direction of the line-of-sight with respect to the event is denoted by the inclination angle ι (its angle relative to the orbital plane) and the coalescence phase ϕ_0 (its angle relative to the two objects *within* the orbital plane). The luminosity distance d_L only affects the signal's overall amplitude. The interferometric measurement of the signal depends on the orientation of the instrument relative to the passing wave. The direction towards the event's sky location (relative to the instrument) is given either in terms of declination and right ascension ($\delta^{\oplus}$, $\alpha^{\oplus}$), or in terms of ecliptic latitude and longitude ($\beta^{\odot}$, $\lambda^{\odot}$). The polarisation angle ψ indicates the orientation of the event's orbital plane relative to the sky coordinate system, and the coalescence time t_c defines the signal's arrival time.

Note that the coalescence time t_c here does *not* denote the instant of coalescence, but it is rather an 'un-physical' or 'virtual' figure. The waveform models used here are only valid until shortly before coalescence, and do not account for what exactly happens when or immediately before the two companion masses merge (see also section 4.3 later in this chapter). The instant t_c denotes the point in the signal's evolution where the expression for its frequency would become infinite. For more detailed parameter definitions and conventions see also [79].

For ground-based interferometry, the observed signal at a particular detector depends on its location and orientation. The response of a cer-

tain detector I depends on the above ('global') parameters via the 'local parameters'

- local coalescence time ($t_c^{(I)} \in \mathbb{R}$),
- altitude ($\vartheta^{(I)} \in [0, \pi]$),
- azimuth ($\varphi^{(I)} \in [0, 2\pi]$) and
- polarisation ($\psi^{(I)} \in [0, \frac{\pi}{2}]$).

How exactly to derive these from the 'global' parameters for a given detector is described in section 4.2.3 later in this chapter. Figure 5.5 in section 5.1.4 shows an example of a chirp signal that is received with differing amplitudes and time delays at different interferometer sites. For applications to space-based measurements, the response to a given gravitational wave signal is more complex and is here derived numerically, using the *LISA Simulator* [80, 81].

4.2.2 Reparametrisations

The mass parameters (m_1, m_2) may be re-expressed in terms of

- total mass ($m_t = m_1 + m_2$),
- reduced mass ($\mu = \frac{m_1 m_2}{m_t}$),
- chirp mass ($m_c = \frac{(m_1 m_2)^{3/5}}{m_t^{1/5}}$) or
- mass ratio ($\eta = \frac{m_1 m_2}{m_t^2}$).

Noteworthy relationships between these are $m_c = \mu^{\frac{3}{5}} m_t^{\frac{2}{5}}$, $\eta = \frac{\mu}{m_t} = \frac{m_1}{m_t}\frac{m_2}{m_t}$, and $\mu = m_t \eta$. Note also that the mass ratio ($0 < \eta \leq \frac{1}{4}$) is dimensionless and that the chirp mass is positively homogeneous in the sense that $f(m_1, m_2) = m_c \;\Leftrightarrow\; f(am_1, am_2) = am_c$ for $a \geq 0$. The mapping from (m_1,m_2) to (m_c,η) is illustrated in figure 4.1. Formulas necessary for reparameterising between some of these expressions (see section 3.3) are given in appendix A.3.

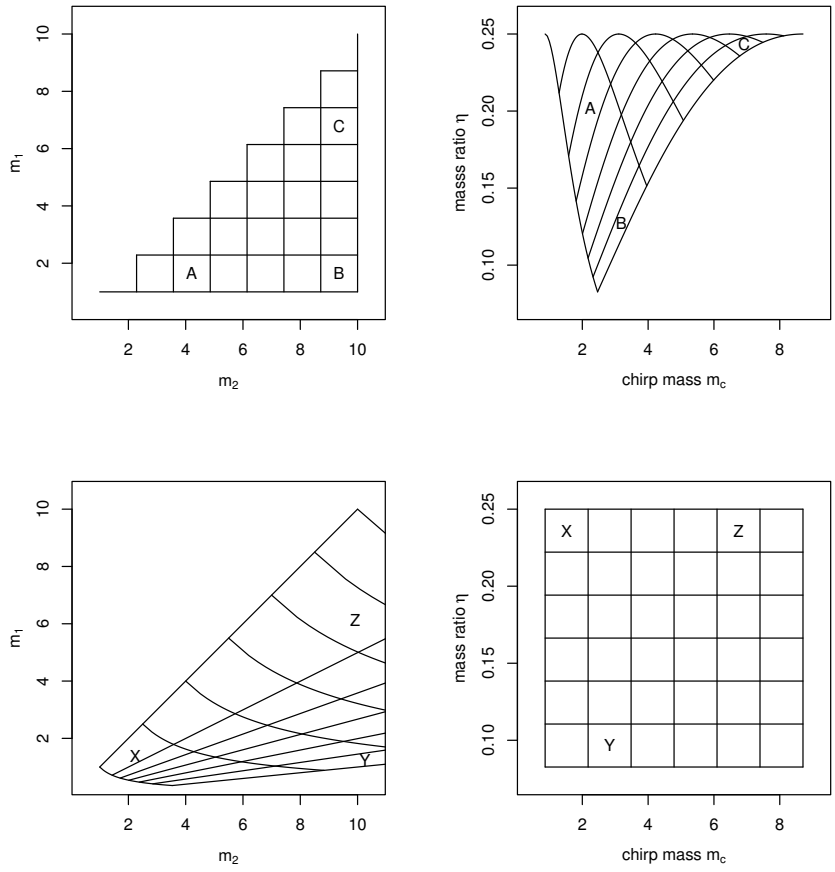

Figure 4.1: Illustration of the mapping between the two parametrisations in terms of individual masses (m_1, m_2) or chirp mass and mass ratio (m_c, η).

4.2.3 Deriving 'local' parameters

Preliminary definitions

The following section is only relevant to ground-based interferometry, since for LISA data the interferometer response to a given signal here is derived numerically, resorting to 'black box' code that only needs to be supplied with the signal waveform as well as location and orientation parameters $\beta^{\odot}$, $\lambda^{\odot}$ and $\psi^{\odot}$.

Locations and orientations of ground-based interferometers are given

in [82]. In order to derive the 'local' coordinates (with respect to a certain interferometer I), it is convenient to transform the polar coordinates (latitude, longitude) into a 3-dimensional cartesian coordinate system. The mapping from the celestial coordinates (declination, right ascension) to geographical coordinates (latitude, longitude) reduces to a shift in longitude / right ascension depending on the current Greenwich mean sidereal time (GMST) [83]. Vectors with respect to the (3-dimensional) cartesian earth frame are, in the following, defined by

- the 'Greenwich' vector $(1, 0, 0)^T$ pointing from the geocentre towards the intersection of the 0° 'Greenwich' meridian with the equator plane,
- the 'Ganges' vector $(0, 1, 0)^T$ pointing towards the intersection of the 90° E 'Ganges' meridian with the equator plane, and
- the 'North Pole' vector $(0, 0, 1)^T$ pointing towards the North Pole.

Note that earth coordinates (latitude, longitude) need to take into account the ellipsoidal earth model [84], while celestial coordinates (declination, right ascension) do not.

In the following, several more or less basic vector operations will be used, namely dot product, cross product, scalar triple product, orthogonal projections, rotations etc, which are explicitly defined in appendix A.6. '$\text{ANGLE}(\vec{x}, \vec{y})$' denotes the angle between two vectors $\vec{x}$ and $\vec{y}$, '$\text{RH}(\vec{x}, \vec{y}, \vec{z})$' means that the three vectors $\vec{x}$, $\vec{y}$ and $\vec{z}$ constitute a right-handed system, '$\text{OP}(\vec{a}, \vec{x}, \vec{y})$' denotes the orthogonal projection of $\vec{a}$ into the plane spanned by $\vec{x}$ and $\vec{y}$, and $R^\alpha_{\vec{a}}$ is the matrix that, when multiplied to a vector, rotates it around $\vec{a}$ by an angle of α. The following conventions are used (roughly following [79] and [85]):

- $\vec{a}^{(I)}$ is the vector pointing from the geocentre to interferometer I's corner station, defined in units of metres.
- $\vec{n}$ is the unit vector pointing from the geocentre along the line-of-sight to the event's sky location.

- $\vec{x}^{(I)}$, $\vec{y}^{(I)}$ and $\vec{z}^{(I)}$ are the unit vectors that are parallel to interferometer I's right arm, left arm and the normal vector of the latter two, respectively, pointing into the interferometer's zenithal direction. $\vec{x}^{(I)}$, $\vec{y}^{(I)}$ and $\vec{z}^{(I)}$ are orthonormal and right-handed.

Coalescence time

The local coalescence time $t_c^{(I)}$ is derived from the global coalescence time at the geocentre $t_c^{\oplus}$ and the event's sky location $(\delta^{\oplus}, \alpha^{\oplus})$ by projecting the line-of-sight $\vec{n}$ onto the vector $\vec{a}^{(I)}$ that points from the geocentre to the interferometer location. The time delay $\Delta_t^{(I)}$ (in seconds) with which the signal arrives at the interferometer relative to the geocentre is given by

$$\Delta_t^{(I)} \;=\; -\frac{\vec{a}^{(I)} \cdot \vec{n}}{\mathrm{c}}, \tag{4.1}$$

and the resulting local coalescence time is: $t_c^{(I)} = t_c^{\oplus} + \Delta_t^{(I)}$.

Altitude, azimuth

The altitude is the angle between line-of-sight and the interferometer's normal vector:

$$\theta^{(I)} \;=\; \mathrm{ANGLE}(\vec{z}^{(I)}, \vec{n}). \tag{4.2}$$

The azimuth is the (directed) angle between the right interferometer arm and the vertical plane (as seen from the interferometer) that contains the source. It may be derived by first determining the orthogonal projection of the line-of-sight into the interferometer plane:

$$\vec{n}^{\perp} = \mathrm{OP}(\vec{n}, \vec{x}^{(I)}, \vec{y}^{(I)}) \tag{4.3}$$

and then determining the angle between right interferometer arm and the projection $\vec{n}^{\perp}$:

$$\varphi^{(I)} = \begin{cases} \mathrm{ANGLE}(\vec{x}^{(I)}, \vec{n}^{\perp}) & \text{if } \mathrm{RH}(\vec{x}^{(I)}, \vec{n}^{\perp}, \vec{z}^{(I)}) \\ 2\pi - \mathrm{ANGLE}(\vec{x}^{(I)}, \vec{n}^{\perp}) & \text{otherwise.} \end{cases} \tag{4.4}$$

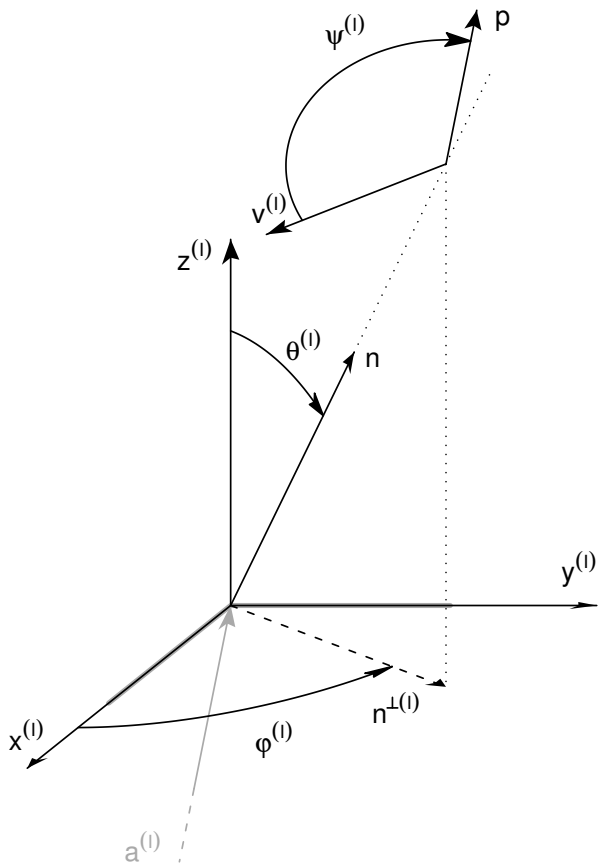

Figure 4.2: Illustration of some of the 'local' parameters, with respect to a certain interferometer I. The (orthonormal) vectors $\vec{x}^{(I)}$, $\vec{y}^{(I)}$ and $\vec{z}^{(I)}$ point along the interferometer's arms and to its zenith, and $\vec{n}$ points towards the source's sky location. $\vec{v}^{(I)}$ is orthogonal to $\vec{n}$ and $\vec{z}^{(I)}$.

Polarisation

The local polarisation angle $\psi^{(I)}$ is defined by the angle between the 'polarisation plane' and the vertical plane (as seen from the interferometer) that contains the source. The global polarisation $\psi^{\oplus}$ defines the angle between the polarisation plane and the vertical plane (relative to the earth

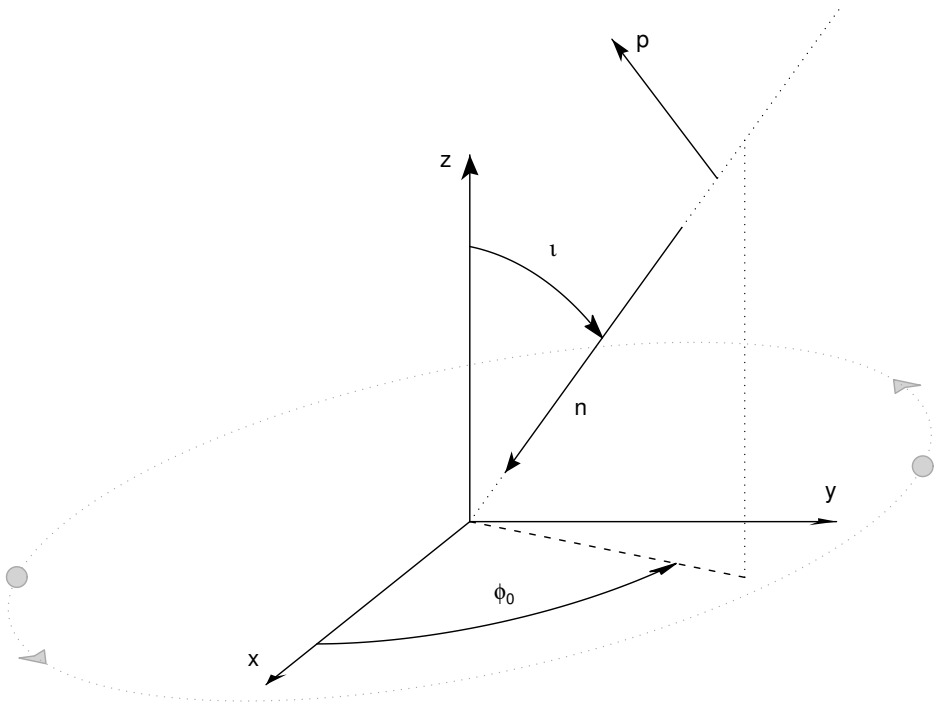

Figure 4.3: Illustration of some of the 'global' parameters; note the common vectors $\vec{n}$ and $\vec{p}$ also appearing in figure 4.2. The (orthonormal) vectors $\vec{x}$, $\vec{y}$ and $\vec{z}$ here define the binary system's orbital plane and its normal direction. The 'polarisation vector' $\vec{p}$ lies within the plane spanned by $\vec{n}$ and $\vec{z}$.

frame) containing the sky location's meridian. An angle of $\psi^{\oplus} = 0$ means that both planes coincide, and increasing it turns the polarisation plane counterclockwise (looking towards the event's sky location).

The local polarisation with respect to an interferometer I may be derived via the following steps. First, determine a (global) 'polarisation vector' $\vec{p}$ that is the normal vector to the polarisation plane:

$$\vec{p} \;=\; R_{\vec{n}}^{(\psi^{\oplus}-\frac{\pi}{2})}\Big((0,0,1)^T - \big(\vec{n}\cdot(0,0,1)^T\big)\vec{n}\Big) \tag{4.5}$$

Then determine the local polarisation as the angle between the above polarisation plane and the interferometer's vertical 'constant azimuth' plane

containing the event's sky location. The latter is given by its normal vector

$$\vec{v}^{(I)} \;=\; R^{\frac{\pi}{2}}_{\vec{z}^{(I)}}\,\mathrm{OP}(\vec{n},\vec{x}^{(I)},\vec{y}^{(I)}) \tag{4.6}$$

which is the line-of-sight, projected into the interferometer's plane and rotated around $\vec{z}^{(I)}$ by 90°. $\vec{n}$, $\vec{z}^{(I)}$ and $\vec{v}^{(I)}$ now constitute a right-handed triplet. The local polarisation then is the angle between the two planes' normal vectors:

$$\psi^{(I)} \;=\; \begin{cases} \mathrm{ANGLE}(\vec{p},\vec{v}^{(I)}) & \text{if } \mathrm{RH}(\vec{n},\vec{v}^{(I)},\vec{p}) \\ \pi - \mathrm{ANGLE}(\vec{p},\vec{v}^{(I)}) & \text{otherwise.} \end{cases} \tag{4.7}$$

4.3 Signal waveform templates

4.3.1 The quadrupole wave

As a gravitational wave propagates through space at light speed, its effect on masses is in directions orthogonal to its direction of travel (see also chapter 2). A binary inspiral event's gravitational wave passing a given point in space is defined by its two $+/\times$ polarisation waveforms, the associated polarisation angle, and the direction to the source (line-of-sight). A gravitational wave's effect on the distance between two (free falling) test masses depends on the orientation of their connecting line with respect to the line of sight, and may be modulated by their relative motion with respect to the source (Doppler effect).

4.3.2 The general binary inspiral 'chirp' signal

In a very simple approximation the $+/\times$ polarisation waveforms for a binary inspiral's GW signal are given by:

$$h_+ \;\propto\; -\frac{2\mu}{d_L}(m_t\omega)^{\frac{2}{3}}\left(1+\cos^2(\iota)\right)\,\cos(2\Phi) \tag{4.8}$$

$$h_\times \;\propto\; -\frac{2\mu}{d_L}(m_t\omega)^{\frac{2}{3}}\left(2\cos(\iota)\right)\,\sin(2\Phi) \tag{4.9}$$

where μ and m_t are the reduced mass and the total mass (see section 4.2.2), and $\Phi = \Phi(t - t_c)$, is the phase evolution as a function of time, giving the *instantaneous phase* for any $t < t_c$. Its time derivative, the *instantaneous frequency* $\frac{d\Phi}{dt} = \omega(t - t_c)$ describes the frequency evolution, in terms of *angular* frequency [79]. Phase evolution Φ and frequency ω are both monotonically increasing, and together with the frequency the signal's amplitude also increases over time, making it a so-called *chirp* signal. The signal's (*dominant*) frequency is twice the orbital frequency, so a complete orbit of the binary system corresponds to two oscillations of the resulting signal.

The chirp waveform (h_+ and $h_\times$) is not given in terms of an exact analytic expression, but it is only approximated to a certain accuracy, based on general relativity theory. In situations where gravity is weak, interactions of masses are usually sufficiently accurately described by Newton's theory of gravity. But as soon as higher masses or velocities are involved, relativistic effects become more and more relevant. Starting off from Newtonian theory, one can then subsequently add higher order corrections to the model and increase its precision. A common approach to this is via the *post-Newtonian formalism*, where the precision of the approximation is then denoted by its *post-Newtonian (PN) order*.

The formulas given for $h_+/h_\times$ in general are thus only valid until shortly before the binary's coalescence, and especially do not consider the effects of coalescence itself. In practice, the $h_+/h_\times$ expressions are used up to the point where the *(Schwarzschild) innermost stable circular orbit (ISCO)* is reached, i.e. when the orbital frequency reaches

$$\omega_{\mathrm{ISCO}} = \frac{c^3}{6^{\frac{3}{2}} G\, m_t}, \tag{4.10}$$

or when the approximation obviously fails, e.g. when the approximated orbital frequency starts decreasing [86].

What accuracy (in terms of PN order) is necessary, depends on the massed involved; for lower masses (neutron star inspirals), an order of 3.0 or 3.5 might be required [86]. The uncertainty in the underlying physical

constants might also need to be considered; for example, the current value of the gravitational constant G is given with an uncertainty of 0.015% [88], which is close to the order of magnitude of the accuracy by which the chirp mass m_c is expected to be determined by ground-based measurements of binary inspiral signals [29, 13]. Since the estimate of m_c is closely related to G, this extra uncertainty might be relevant, and it would be rather straightforward to consider it in a Bayesian setup [89].

In the following sections, some specific waveform approximations are pointed out. Formulas are only given for the two simplest cases, since these are very complex and not of primary interest here beyond what can already be seen from the above outline. For some of the inspiral signal approximations, it is possible to restrict the parameter space. For example, in the 'restricted PN approximation', the signal waveform's dependence on the phase parameter ϕ_0 is such that a certain value ϕ_0 leads to exactly the same waveform as $\phi_0 \pm \pi$. Here the parameter space can be narrowed down to the range of $[0, \pi]$ instead of $[0, 2\pi]$, while keeping in mind that any statement about a value ϕ_0 applies to $\phi_0 + \pi$ as well.

4.3.3 The restricted PN approximation

This approximation gives the $h_+/h_\times$ waveforms in the time domain and was used for mock data generation for the first and second round of the Mock LISA Data Challenges (MLDC) [90]. The formulas are given in [91, 92] and are reproduced in appendix A.7. This aproximation was used in the context of inference on space-based LISA measurements [16]. In contrast to the definitions in 4.2, the domain of the coalescence phase may be narrowed down to $\phi_0 \in [0, \pi]$ here.

4.3.4 The 2.0 PN stationary phase approximation

This approximation gives the Fourier transforms $\tilde{h}_+/\tilde{h}_\times$ of the chirp waveform, i.e. the signal in the frequency domain. The formulas are given in [10, 93] and are reproduced in appendix A.8. These were used in the

first 5-parameter version of the MCMC, where the likelihood computation took place completely in the frequency domain [12]. In contrast to the definitions in 4.2, the domains of coalescence phase and polarisation angle may be narrowed down to $\phi_0 \in [0, \pi]$ and $\psi \in [0, \frac{\pi}{2}]$ here.

4.3.5 The 2.5 PN phase / 2.0 PN amplitude approximation

This approximation gives templates in the time domain, with 2.5 PN accuracy in phase, and 2.0 PN in amplitude. Formulas are given in [79]. These were used in the coherent version of the MCMC algorithms, where templates were generated in the time domain, and then numerically Fourier-transformed to the frequency domain [13].

4.3.6 The 3.5 PN phase / 2.5 PN amplitude approximation

This is a more accurate version of the above templates. The expressions for the 3.5 PN phase are given in [86] (requiring terms from [94]), and the formulas for the 2.5 PN amplitude in [95]. For parametrisation in terms of coalescence phase ϕ_0 instead of 'a constant phase' τ_0, see appendix A.9. These templates were used in [14].

4.4 Detector response

4.4.1 Ground-based interferometry

For ground-based interferometry, the detector is assumed to have essentially zero extent and so the interferometer output h simplifies to the sum of the projections of the passing $+/\times$ polarisation waveforms h_+ and $h_\times$ along the directions of the interferometer arms. Depending on the interferometer's location, the wave will pass the interferometer with a slight time difference $\Delta_t^{(I)}$ with respect to the geocentre (within ± 0.022 seconds). The signal $h^{(I)}(t, \theta)$ that is eventually measured at a detector I depends on the

antennae pattern functions $F_+^{(I)}$ and $F_\times^{(I)}$:

$$h^{(I)}(t,\theta) \;=\; F_+^{(I)}\, h_+^{(I)}\big(t,\theta\big) \;+\; F_\times^{(I)}\, h_\times^{(I)}\big(t,\theta\big) \tag{4.11}$$

where

$$\begin{aligned} F_+^{(I)} \;&=\; +\tfrac{1}{2}\big(1+\cos^2(\vartheta^{(I)})\big)\cos(2\varphi^{(I)})\cos(2\psi^{(I)}) \\ &\qquad -\cos(\vartheta^{(I)})\sin(2\varphi^{(I)})\sin(2\psi^{(I)}) \end{aligned} \tag{4.12}$$

$$\begin{aligned} F_\times^{(I)} \;&=\; -\tfrac{1}{2}\big(1+\cos^2(\vartheta^{(I)})\big)\cos(2\varphi^{(I)})\sin(2\psi^{(I)}) \\ &\qquad -\cos(\vartheta^{(I)})\sin(2\varphi^{(I)})\cos(2\psi^{(I)}) \end{aligned} \tag{4.13}$$

[79], and the 'local' parameters $\vartheta^{(I)}$, $\varphi^{(I)}$, $\psi^{(I)}$ and $\Delta_t^{(I)}$ are defined as in section 4.2. For binary inspiral signal analysis the antennae pattern functions as well as the time shift $\Delta_t^{(I)}$ are assumed to be constant over the short time intervals of concern.

4.4.2 Space-based interferometry

The planned space-based LISA observatory will be of a size that is of the order of the wavelengths it is going to be able to detect. It will be sensitive to much lower frequencies, and its sampling frequencies will be lower as well, so any measurement will be significantly affected by the orbital motion and its change over time. The mapping from the passing gravitational wave to the output is not a simple '1:1' mapping as in the simplified model for earth-bound measurements, especially when the signal wavelength is of the order of LISA's armlength. The interferometer response to a given GW signal (specified in terms of its $+/\times$ polarisation waveforms h_+ and $h_\times$, polarisation angle and source direction) can in general be approximated numerically [81, 97, 98], or, in the special case of sinusoidal signals at low wavelengths, analytically [35, 99].

Due to LISA's layout, the 'raw' detector output variables will always be interspersed with highly correlated noise, originating e.g. from random motions of the individual spacecraft that will affect all in- and out-

going measurements. By clever combination of the different observables, many such unwanted effects can be made to cancel out again. The data produced by the spacecraft trio is typically combined to form three time-delay-interferometry (TDI) variables, denoted by X, Y and Z [78]. These can be linearly recombined into three stochastically independent components (A, E and T), two of which are sensitive to gravitational waves (A and E) and one component which is a 'null stream' that is only noise (T) (see also appendix A.10) [100]. The derivation of TDI varibles can be motivated as the signal's principal components or sufficient statistics, which concentrate the 'astronomical' information in the data [101]. In the following we will only be concerned with the former two variables, A and E.

4.5 Model

4.5.1 The data

In both cases of ground-based and space-based interferometers, the data are time series, in general several of them, denoting the measured gravitational wave, and also noise affecting the measurement for different interferometers or TDI variables. The noise may have individual characteristics for the different data streams, and in any case the streams are assumed to be stochastically independent. In the following two subsections, two models for these individual time series (corresponding to different interferometers or TDI variables) are introduced.

4.5.2 Known noise spectrum

The detector output $\{z(t_i)\}_{i=1,\ldots,N}$ is a time series, a data set of size N corresponding to a set of equidistant time points t_i. It is assumed to be the sum of the response to a gravity wave signal $s(t,\theta)$ depending on an (unknown) parameter vector θ, and noise $n(t)$, which is assumed to be

Normal and stationary with given power spectral density $S_n(f)$:

$$z(t_i) \;=\; s(t_i, \theta) \;+\; n(t_i) \qquad \text{for } i = 1, \ldots, N. \tag{4.14}$$

Data from different detectors or TDI variables then are assumed to be stochastically independent. The noise's Normal distribution is not necessarily only a property of the actual noise, but may also express uncertainty in the noise level (and sources) or its distribution itself; the Normal distribution is the maximum entropy distribution for a given second (and first) moment of the noise [102, 103, 3].

The signal waveforms $s(t, \theta)$ describing the detector response to a passing gravitational wave with parameters θ were defined in sections 4.3 and 4.4. For ground-based interferometers these are of a functional form, while for space-based applications these are derived numerically from the given $+/\times$ polarisation waveforms.

4.5.3 Unknown noise spectrum

Model specification

When the noise spectrum is not known in advance, it can also be incorporated into the inference in terms of a set of unknown parameters. The model setup is then similar to the above, the difference being the additional parameter vector $\sigma = (\sigma_0, \ldots, \sigma_{N/2})$ in the noise term:

$$z(t_i) \;=\; s(t_i, \theta) \;+\; n(t_i, \sigma) \qquad \text{for } i = 1, \ldots, N, \tag{4.15}$$

where the noise (-residuals) $n(t_i, \sigma)$ are modeled as:

$$\begin{aligned} n(t_i, \sigma) &= \sum_{j=0}^{N/2} a_j \cos(2\pi f_j t_i) + b_j \sin(2\pi f_j t_i) && (4.16) \\ &= \sum_{j=0}^{N/2} \sqrt{a_j^2 + b_j^2}\, \sin(2\pi f_j t_i + \varphi_j), && (4.17) \end{aligned}$$

$$\text{where } \varphi_j = \begin{cases} \arctan\left(\frac{b_j}{a_j}\right) & \text{if } a_j > 0 \\ \arctan\left(\frac{b_j}{a_j} \pm \pi\right) & \text{if } a_j < 0, \end{cases}$$

the f_j are the Fourier frequencies $f_j = j\Delta_f = \frac{j}{N\Delta_t}$, and the a_j and b_j corresponding to frequency f_j are random variables following a Normal distribution $\mathrm{N}(0, \sigma_j^2)$ with zero mean and variance σ_j^2 (except for $b_0 = b_{N/2} = 0$). This means that the resulting noise $n(t_i, \sigma)$ is Normal with zero mean (being a linear combination of Normally distributed random variables for each t_i). For given variance parameters $\sigma_0, \ldots, \sigma_{N/2}$ the noise's (one-sided) spectral density at a frequency f_j is

$$S_n(f_j) \;=\; N\,\Delta_t^2\,\sigma_j^2 \;=\; \frac{1}{N\,\Delta_f^2}\,\sigma_j^2 \qquad \text{for } j = 1, \ldots, \tfrac{N}{2} - 1, \qquad (4.18)$$

and half as much for $j = 0$ and $j = \frac{N}{2}$. If all the σ_j were known, this model would be exactly equivalent to the model from the previous section 4.5.2. Treating the variance parameters σ_j as unknown adds a set of $\frac{N}{2} + 1$ additional parameters to the model. The noise realisations $a_0, \ldots, a_{N/2}$ and $b_0, \ldots, b_{N/2}$ can be derived from the time-domain noise $n(t)$ by a discrete Fourier transform (for details see appendix A.11.1), and every σ_j can be estimated based on a_j and b_j (and prior information, of course). If the data consists of more than one time series and a common spectrum is assumed for these, the estimation of the σ_j can be based on more than two samples each.

Some remarks

At first glance, the above model might seem a bit odd, seemingly implying that the noise came about by randomly drawing amplitudes for each frequency, and then adding up the resulting sinusoids. But if one assumes all the variance parameters to be known, then the model turns out to be exactly equivalent to the more common model from the previous section. Models for non-white (or 'coloured') noise are commonly characterised either by their spectrum or (equivalently) by their autocorrelation function. Following the above reasoning, a definition in terms of an autocorrelation function might seem just as odd, since it would imply that the noise samples at each instance came about as linear combinations of all previous samples, plus an added error term. Still, these models are common and useful.

More insight into the model introduced above may be gained by considering the following motivations. Firstly, the set of trigonometric functions in (4.16) constitute an orthonormal basis of $\mathbb{R}^N$. This implies that there is a unique one-to-one mapping (via Fourier transformation) between the time-domain and frequency-domain representations of the noise. Looking at the relation between noise spectrum and noise parameters in equation (4.18), one can see that (for given noise parameters σ) this noise parametrisation is equivalent to the model for given noise spectrum, and that allowing σ to be unknown leads to a somewhat 'natural' generalisation. Implementation of the model is simplified by the availability of a conjugate prior distribution for σ_j (see the following section). Using this prior, all σ_j are independent in their conditional posterior distribution (conditional on a given vector of residual noise), and only depend on the corresponding a_j and b_j, which is very convenient for practical implementation. In fact, if one used the noise model for spectrum estimation of a given 'fixed' noise sample (instead of looking at *conditional* spectra, conditional on the current signal parameters within a Metropolis-algorithm, as in the following), the complete posterior distribution of the noise parameters σ_j, and with that the posterior distribution of the spec-

trum, would be given in a rather simple, closed form.

As stated earlier, the resulting noise is Normal with zero mean. Probably most importantly, the use of the Normal distribution does not necessarily mean that the noise actually *was sampled from a Normal distribution*, but it is also the maximum entropy distribution for given first and second moments (mean and variance) of the noise. This means that when specifying the noise in terms of a Normal distribution, the only assumptions entering the model are about its mean (which is fixed at zero), and its variance parameters (σ_j), for which prior information needs to be supplied, but which are otherwise inferred from the data. In particular, with this specification the model is also able to handle 'deterministic' but unaccounted for signals that may be present within the noise. This property is exposed in far more detail in [103]. This will be particularly useful in the application to gravitational-wave measurements where the 'noise' is assumed to be in significant part due to unaccounted for signals.

The great number of parameters in the noise model make it very general, nonrestrictive and flexible. This is a particularly desirable property when dealing with space-based 'LISA' data, where the noise spectrum is known to exhibit many narrow 'emission lines'. Restrictions may be introduced through the prior, or, if for example the spectrum is assumed to be somewhat smooth, via a hierarchical model and hyperparameters for the σ_j.

Implementation implications

The variance parameters σ_j are estimated from the observed values of a_j and b_j. Specifying the prior for each σ_j ($j = 0, \ldots, \frac{N}{2}$) in terms of their conjugate prior distribution makes inference for these very straightforward. For a given, fixed noise vector $n(\cdot)$, the posterior distribution of the noise parameters $\sigma_0, \ldots, \sigma_{N/2}$ then has a rather simple, closed form. When using MCMC methods for inference using the models introduced above, the additional noise parameters that appear in the 'unknown spectrum' model (4.15) can be drawn in a simple Gibbs step, conditional on the remaining

signal parameters θ, on which the noise parameters depend via the implied vector of residual noise. The conjugate prior distribution for σ_j is the *scaled inverse χ^2-distribution*:

$$p(\sigma_j^2) \;=\; \text{Inv-}\chi^2(\nu_j, s_j^2) \tag{4.19}$$

with degrees-of-freedom parameter ν_j and scale parameter s_j^2 [42]. Specifying ν_j and s_j^2 as independent from j would lead to a priori white noise. Varying s_j^2 with j on the other hand leads to a priori coloured noise, while a specification of ν_j dependent on j would indicate varying prior certainty over different parts of the spectrum.

For simplicity and clarity, in the following the running index j of the noise parameters σ_j is mostly assumed to be $j > 0$ and $j < \frac{N}{2}$. This excludes the two extreme cases $j = 0$ and $j = \frac{N}{2}$ which need to be treated slightly differently but analogously. For a given sample $(n(t_0), \ldots, n(t_{N-1}))$, a discrete Fourier transformation $(\tilde{n}(f_0), \ldots, \tilde{n}(f_{N-1}))$ yields the two realisations a_j and b_j for each $j = 0, \ldots, \frac{N}{2}$ from which the value of σ_j^2 (corresponding to frequency f_j) can be inferred; see appendix A.11.1 for an exact derivation. The (conditional) posterior distribution of σ_j^2 (conditional on the remaining parameters θ, on which it depends via the implied vector of residual noise) is then again a scaled inverse χ^2-distribution:

$$\sigma_j^2 \big| \{n(t_1), \ldots, n(t_N)\} \;\sim\; \text{Inv-}\chi^2\left(\nu_j + 2,\, \frac{\nu_j s_j^2 \,+\, a_j^2 + b_j^2}{\nu_j + 2}\right) \tag{4.20}$$

for $j = 1, \ldots, N-1$ [42]. The sum of the squared 'empirical' amplitudes $a_j^2 + b_j^2 = 4\Delta_f^2 |\tilde{n}(f_j)|^2$ here is the sufficient statistic for σ_j. More generally, if one has several independent time series with a *common* unknown spectrum that is to be estimated, then the corresponding conditional posterior distribution is

$$\sigma_j^2 \big| n_1(\cdot), \ldots, n_k(\cdot) \;\sim\; \text{Inv-}\chi^2\left(\nu_j + 2k,\, \frac{\nu_j s_j^2 \,+\, v_j}{\nu_j + 2k}\right) \tag{4.21}$$

for $j = 1, \ldots, N-1$, where k is the number of time series, and $v_j = 4\Delta_f^2 \sum_{i=1}^{k} |\tilde{n}_i(f_j)|^2$ is the total sum of squared amplitudes corresponding to frequency f_j.

When setting up an MCMC sampler for this model that uses parallel tempering, one also needs to be able to sample from the 'tempered' conditional posterior. If the tempering is only applied to the likelihood part of the posterior (as in (3.11)), the resulting *tempered* distribution is again an Inv-χ^2-distribution:

$$\text{Inv-}\chi^2\left(\nu_0 + \tfrac{2k}{T},\ \frac{\nu_j s_j^2 + \frac{1}{T} v_j}{\nu_j + \frac{2k}{T}}\right) \tag{4.22}$$

for $j = 1, \ldots, N-1$ (see also appendix A.12).

Given the sum of squared amplitudes v_j, one can also easily derive the (conditional) expected spectrum, via the expected values of the σ_j's (4.21) and their (linear) relationship to the spectrum (4.18). This is useful, because this way one can estimate the posterior expected spectrum within an MCMC algorithm as $\mathrm{E}[S_n(f)|y] = \mathrm{E}\big[\mathrm{E}[S_n(f)|\theta]|y\big]$ without having to integrate over the noise parameters (by sampling) as well. First of all, the expectation of σ_j^2 is:

$$\mathrm{E}[\sigma_j^2 | n_1(\cdot), \ldots, n_k(\cdot)] \;=\; \mathrm{E}[\sigma_j^2 | v_j^2] \;=\; \frac{\nu_j s_j^2 + v_j}{\nu_j + 2k - 2} \tag{4.23}$$

[42], and consequently the expected spectrum is

$$\mathrm{E}[S_n(f_j) | n_1(\cdot), \ldots, n_k(\cdot)] \;=\; N\Delta_t^2 \frac{\nu_j s_j^2 + v_j}{\nu_j + 2k - 2}. \tag{4.24}$$

Note that when doing the likelihood computation as described in the following section, the computation of the $v_j = 4\Delta_f^2 \sum_{i=1}^{k} |\tilde{n}_i(f_j)|^2$ do not require any *additional* Fourier transformations to those that are already done in the 'known spectrum' case; due to the linearity of the Fourier transform (3.59), the transform of the difference of data and signal tem-

plate is identical to the difference in the transforms of data and signal template: $\widetilde{z-s} = \widetilde{z} - \widetilde{s}$.

4.6 Likelihood

4.6.1 Overall likelihood

When processing data from different detectors or TDI variables, in both cases their noises are assumed to be detector/variable-specific and independent. Thus, the joint likelihood $\mathcal{L}(\theta) = p(\theta|y)$ is a product of the individual likelihoods, indexed by I:

$$\mathcal{L}(\theta) = \prod_I \mathcal{L}^{(I)}(\theta) \tag{4.25}$$

$$\Leftrightarrow \quad \log(\mathcal{L}(\theta)) = \sum_I \log(\mathcal{L}^{(I)}(\theta)). \tag{4.26}$$

The exact forms of the individual likelihoods are given in the following section.

4.6.2 Individual likelihood

'Known spectrum' model

Since the noise is defined as coloured and specified in terms of its spectral density, it is most convenient to perform likelihood computations in the frequency domain, based on Fourier transforms of data ($\tilde{z}$) and modeled waveform ($\tilde{s}(\theta)$). The likelihood for the I-th detector/variable is given by the following expression

$$\mathcal{L}^{(I)}(\theta) = p^{(I)}(z|\theta) = K \times \exp\left(-2\int_0^\infty \frac{|\tilde{z}(f) - \tilde{s}(f,\theta)|^2}{S_n(f)}\mathrm{d}f\right) \tag{4.27}$$

[104], which for discretised data corresponds to the sum of the squared differences between (discrete) Fourier transforms of observed signal and

signal template over the discrete set of Fourier frequencies $\{f_i = i\Delta_f : i_L \leq i \leq i_U\}$:

$$\mathcal{L}^{(I)}(\theta) \;=\; K \times \exp\left(-\frac{2}{N}\sum_{i=i_L}^{i_U}\frac{|\tilde{z}(f_i) - \tilde{s}(f_i,\theta)|^2}{S_n(f_i)}\right) \tag{4.28}$$

where f_{i_L} and f_{i_U} are the lower and upper bounds of the examined frequency range, Δ_f is the resolution of the (discrete) Fourier transformed data, $|\cdot|$ denotes the absolute value of the (complex-valued) difference, $S_n(\cdot)$ is the (one-sided) noise power spectral density, and K is a normalising constant. Note that (although not labeled as such) z, s, S, i_L, i_U, δ_t and Δ_f are specific for the Ith set of data (detector or TDI variable). For simplicity, it is here assumed that $i_L > 0$ and $i_U < \frac{N}{2}$, since this means the special cases of $i = 0$ and $i = \frac{N}{2}$ drop out (see section 3.4.2, (4.18), and appendix A.11.1).

The restriction to a limited frequency range in (4.28) may be interpreted as the assumption that the noise spectrum was infinite outside that range, or that $\tilde{s}$ was bandwidth-limited to within that range, or at least that a change in θ does not change $\tilde{s}$ outside the range. In either case the dropped terms are considered part of the normalising constant K.

'Unknown spectrum' model

The likelihood function for the model that does not assume the spectrum to be known beforehand is very similar to the above (4.28), but takes into account the noise parameters $(\sigma_1^2, \ldots, \sigma_{N/2}^2)$ as unknowns:

$$\mathcal{L}^{(I)}(\theta) = K \times \exp\left(\sum_{i=i_L}^{i_U}\left[-\frac{2}{N}\frac{|\tilde{z}(f_i) - \tilde{s}(f_i,\theta)|^2}{S_n(f_i)} - \log\big(S_n(f_i)\big)\right]\right) \tag{4.29}$$

where the relation between the idividual parameters σ_i^2 $(i = 0, \ldots, \frac{N}{2})$ and the implied (one-sided) spectrum $S_n(f_i)$ was given in equation (4.18). A more detailed derivation of the likelihood from the noise model is shown in appendix A.11.2.

4.6.3 Signal-to-noise ratio

An expression closely related to the likelihood is the *signal-to-noise ratio* (SNR). The SNR of a certain signal $s(\theta)$ received at interferometer I and embedded in noise with spectral density S_n is defined as:

$$\rho^{(I)}(s(\theta)) \;=\; \sqrt{4\int_0^\infty \frac{|\tilde{s}(f,\theta^{(I)})|^2}{S_n(f)}\mathrm{d}f} \tag{4.30}$$

[11]. In analogy to equation (4.28) it is in practice computed over the same frequency range that is relevant for the likelihood:

$$\rho^{(I)}(s(\theta)) \;=\; \sqrt{\frac{4}{N}\sum_{i=i_L}^{i_U} \frac{|\tilde{s}(i\times\Delta_f,\theta^{(I)})|^2}{S_n(i\times\Delta_f)}}. \tag{4.31}$$

The overall network SNR then is defined as

$$\rho(s(\theta)) \;=\; \sqrt{\sum_I \left(\rho^{(I)}(s(\theta))\right)^2} \tag{4.32}$$

[29].

Note that $\rho(s) \geq 0$, and $\rho(a\cdot s) = |a|\cdot\rho(s)$ for $a \in \mathbb{R}$ (i.e. the SNR is *positively homogeneous*).

4.7 Prior definition

4.7.1 A priori information

The prior distribution of the parameters expresses the information about the signal's parameter values *before considering the data*. In the following, one obvious prerequisite will be taken advantage of, namely that *the signal present in the data is strong enough to be detected*. The algorithm developed here is intended for use after the data has been preprocessed by a signal detection algorithm, and not for detection itself. So, once the signal has been picked up by the detection algorithm, it clearly must have

a certain intensity. Ruling out those parts of parameter space that imply signals that could not be detected anyway allows the definition of a proper and otherwise uninformative prior distribution for all parameters. First an (improper) distribution describing the occurrence of inspiral events will be derived, which is then constrained by considering the detectability for any given parameter combination.

4.7.2 Occurrence

Assuming that an inspiral event is equally likely to happen in any direction and orientation (or there is none a priori 'preferred') leads to independent prior distributions that are uniform across their domains for coalescence phase ϕ_0, right ascension $\alpha^{\oplus}$, and polarisation $\psi^{\oplus}$. The prior density for the declination $\delta^{\oplus}$ is

$$f(\delta^{\oplus}) \;=\; \tfrac{1}{2}\cos(\delta^{\oplus}), \tag{4.33}$$

proportional to the circumference of the corresponding parallel (circle of latitude). Analogously, the prior density for the inclination ι is

$$f(\iota) \;=\; \tfrac{1}{2}\sin(\iota). \tag{4.34}$$

For parametrisation in terms of (ecliptic) latitude $\beta^{\odot}$, longitude $\lambda^{\odot}$ and polarisation $\psi^{\odot}$ the prior is defined accordingly.

The coalescence time (t_c) is assumed to be known in advance with a certain precision through preprocessing of the data. For ground-based applications this is implemented as a uniform distribution across $\pm 10\,\mathrm{ms}$ around the true value (which of course is known for simulated data sets). For space-based (LISA) applications an (improper) unbounded uniform prior distribution is used. In realistic applications, a uniform prior, either across a conservatively wide range, or an unbounded (improper) one, might be appropriate as well.

The prior for the masses (m_1, m_2), reflecting the distribution of the masses among binary inspirals, could be based on observational evidence

[105, 106] as well as theoretical considerations [107, 108]. For testing purposes with simulated data in ground-based applications (as well as for the plots shown later in this section) a uniform prior across 1–10 $M_\odot$ (solar masses: $M_\odot \approx 2\times10^{30}$ kg) was used. For the LISA application, the prior also was defined to be uniform over the parameter range specified for the corresponding application [92], which was defined by $m_1 \in [10^6 M_\odot, 5\times 10^6 M_\odot]$ and $m_2/m_1 \in [1,4]$.

Assuming that inspirals happen uniformly across space leads to a prior for the luminosity distance d_L with

$$P(d_L \leq y) \propto y^3. \tag{4.35}$$

Without explicitly specifying any upper bound for d_L, so far this is an improper prior (that has an infinite integral), seemingly implying that there was an 'infinite' probability for 'infinitely remote' inspiral events. For some useful formulas for a bounded, proper version of this distance 'occurrence' prior see appendix A.14.

4.7.3 Detectability

The above definitions alone lead to a prior that is not only improper, but also unrealistic—obviously, signals cannot originate from arbitrarily great distances, since beyond a certain point they would be too faint to be noticeable at all. Rather, they need to happen within a certain range in order to be detected. This restriction is incorporated into the prior definition by considering the *detection probability* for any point in parameter space. The detection probability $D(\theta)$ of a signal with parameters θ is defined to depend on the signal's signal-to-noise ratio (SNR) $\rho(\theta)$:

$$D(\theta) \propto \rho(\theta). \tag{4.36}$$

The SNR is computationally expensive to determine, so for the prior definition a 'cheap' approximation to the SNR is sought. Due to the SNR's homogeneity property (see section 4.6.3) one can simplify and only con-

sider those parameters that affect the wave's amplitude $\mathcal{A}$, assuming that

$$\rho(\theta) \propto \mathcal{A}(\theta), \tag{4.37}$$

i.e. the SNR $\rho(\theta)$ is approximated using the overall signal amplitude $\mathcal{A}(\theta)$, which is then defined as

$$\begin{aligned} \mathcal{A}(m_1, m_2, d_L, \iota) &= \log\left(\frac{\sqrt{\eta}\, m_t^{\frac{5}{6}}}{d_L} \underbrace{\sqrt{\left(1+\cos^2(\iota)\right)^2 + \left(2\cos(\iota)\right)^2}}_{\geq 1 \text{ and } \leq \sqrt{8} \approx 2.8} \right) && (4.38) \\ &= \tfrac{1}{2}\big(\log(m_1) + \log(m_2)\big) - \tfrac{1}{6}\log(m_1 + m_2) \\ &\quad - \log(d_L) + \tfrac{1}{2}\log\big(1 + 6\cos(\iota)^2 + \cos(\iota)^4\big) && (4.39) \end{aligned}$$

[12, 14]. $\mathcal{A}$ here actually denotes the *logarithmic* amplitude (see (A.28), (4.8), (4.9), [12]). This simplification only considers the properties of the arriving wave (at a given point in space), neglecting any detector properties like its relative orientation or noise. From equation (4.38) one can see how the amplitude is affected by different parameters: it grows with the mass ratio (more similar masses mean a greater amplitude), it grows with the total mass (twice the total mass m_t yields $2^{\frac{5}{6}} \approx 1.78$ times the amplitude), it decreases with the distance (twice the distance d_L—half the amplitude), and depending on the inclination angle ι, it may vary by a factor of almost 3 (and is lowest at the most likely value $\iota = \frac{\pi}{2}$).

The detection probability for certain parameter values θ then is modeled in dependence on the logarithmic SNR (approximated by $\mathcal{A}$):

$$D(\theta) \;=\; D\big(\log(\rho(\theta))\big) \;\approx\; D\big(\mathcal{A}(\theta)\big). \tag{4.40}$$

Due to the random noise contribution to the measured signal, detectability of a given signal in general is a matter of chance. While detectability is certain for strong signals and impossible for weak signals, there is a transition region in between; the same effect was found to be present in a similar context (long-term observations of pulsar's gravitational wave signals) [8]. The dependence between a given amplitude x and detection probability D

is modeled using a (sigmoidal) logistic function of the form:

$$D_{a,b}(x) = \frac{1}{1 + \exp(-\frac{x-a}{b})} \tag{4.41}$$

where a and b are set so that $D_{a,b}(x_L) = p$ and $D_{a,b}(x_U) = 1 - p$ for some $0 < p < 0.5$ and lower and upper thresholds x_L and x_U. So x_L denotes the amplitude at which the detection probability reaches p, and x_U is the amplitude where the probability falls below $(1 - p)$. In order to fit d through these points, its parameters are then set to:

$$a = \frac{x_L + x_U}{2} \qquad \text{and} \qquad b = \frac{x_L - x_U}{2 \log(\frac{p}{1-p})}. \tag{4.42}$$

For example, one could set $p = 0.1$, $x_U = \mathcal{A}(2\mathrm{M}_\odot, 2\mathrm{M}_\odot, 50\mathrm{Mpc}, \frac{\pi}{2})$ and $x_L = \mathcal{A}(2\mathrm{M}_\odot, 2\mathrm{M}_\odot, 60\mathrm{Mpc}, \frac{\pi}{2})$, assuming that a 2-2-$\mathrm{M}_\odot$ inspiral with an inclination of $\frac{\pi}{2}$ is detectable out to 50 and 60 Mpc with probabilities of 90% and 10%, respectively.

Fitting the logistic function (4.41) also yields the interpretation that there is a linear relationship between logarithmic detection-*odds* and the (logarithmic) signal amplitude $\mathcal{A}$. Odds are derived from probabilities as:

$$\mathrm{odds}(p) \;=\; \frac{p}{1-p}. \tag{4.43}$$

The parameter a then defines the amplitude value $\mathcal{A}$ where detection chances are "50:50" (i.e. the probability is $p = \frac{1}{2}$), and b determines how much the (logarithmic) odds change with an increasing (logarithmic) amplitude. A value of $b = 1$ would mean that an amplitude change by a certain factor changes the detection odds by the same factor.

Instead of approximating the SNR $\rho(\theta)$ by the amplitude $\exp(\mathcal{A}(\theta))$, one could directly use the SNR to model the detection probability. Computing an SNR is in general about as computationally expensive as a likelihood computation. But within an Metropolis algorithm, where one would compute the likelihood for each proposed step anyway (except if the prior is already zero), it would simplify to some extent. Due to their similar-

ity, SNR and likelihood computation could share common intermediate stages, and an additional SNR computation might come at little additional computational cost. The shape of the prior distribution would then depend on properties of the instrument(s) used to obtain the measurement, e.g. instrument noise, or its orientation at the time of observation.

4.7.4 Prior

Given the above *occurrence* and *detection* probabilities, the joint prior for parameters θ is:

$$p(\theta) \;=\; p(\text{occurrence} \,|\, \theta) \;\times\; p(\text{detection} \,|\, \theta, \text{occurrence}), \qquad (4.44)$$

so the detection probability enters the prior definition as an extra factor in addition to the independent components of the 'occurrence' prior definition. The resulting prior distribution is proper, i.e. it has a finite integral (see appendix A.15 for a proof).

The dependence on the signal's amplitude implies that greater masses are (a priori) more likely even if *initially* any mass is assumed to be equally likely to occur: signals involving greater masses may originate from greater distances, while low-mass inspirals need to be close to be noticeable at all. Masses, distance and inclination angle are not stochastically independent any more. Figure 4.4 shows some marginal prior densities resulting from the above definitions (and a uniform prior across 1–10 $\mathrm{M}_\odot$ for the individual masses m_1, m_2). A similar, analogous effect is known in astronomy as the *Malmquist effect*. Considering it in the prior definition will compensate for selection bias that would otherwise also affect parameter estimates and other inference [109, 110].

4.7.5 Noise prior

When using the model that does not assume the noise spectrum to be known beforehand (see section 4.5.3), the prior for the additional noise parameters $\sigma_0, \ldots, \sigma_{N/2}$ needs to be specified as well. If available, one can

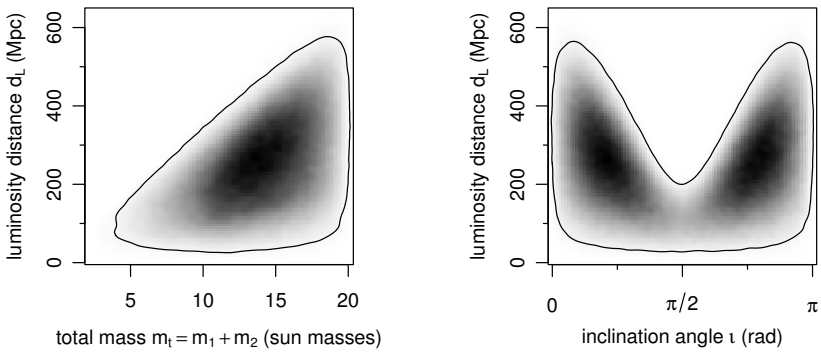

Figure 4.4: Marginal prior densities for two pairs of parameters.

derive a rough spectrum estimate from an (independent) sample of the noise, by first determining its empirical spectrum $\widehat{S}_n$ (as described in section 3.4.5 for example) and then defining the prior scale s_j^2 for each of the parameters σ_j by inverting equation (4.18) and setting

$$s_j^2 = \frac{\widehat{S}_n(f_j)}{N\Delta_t^2}. \tag{4.45}$$

Alternatively, theoretical predictions of the spectrum (e.g. [34, 26, 27]) could be used instead. The certainty in this estimate then also needs to be specified via the prior degrees-of-freedom ν_0. Setting ν_0 to zero yields a non-informative, but also improper prior for the σ_j that does not depend on the prior scale s_j^2 (see also appendix A.12).

Chapter 5

Application

5.1 Inference on inspiral signals using ground-based detectors

5.1.1 Introduction

In the following, applications of a Bayesian inference framework to the analysis of inspiral signals are illustrated. The (simulated) data come from earth-bound interferometers, and the parameters determining the measured waveform are inferred.

The model and methods used here originate from a longer evolution process (see also chapter 1). A first version of the approach used a simplified model considering only 5 parameters and data from a single interferometer. The 4 neglected parameters were altitude, azimuth, polarisation and inclination angle; it was then assumed that the inspiral event was 'optimally oriented', having zero inclination and being located in the interferometer's zenith. The resulting posterior distributions exhibit multiple modes, so that a basic MCMC algorithm usually would end up and be stuck in one of the many local modes. Reliable convergence was then ensured by using importance resampling to generate some approximate posterior draws from a large sample covering the whole parameter space. This approach was no longer feasible with the enlarged parameter space

(9 parameters) when going over to analyse data from several interferometers. Starting values were still generated the same way, but convergence and mixing of the MCMC chains are ensured by the use of parallel tempering (and its extension, evolutionary MCMC). In the simplified problem, the Metropolis-algorithm's proposal distribution was adaptively fitted to the distribution of some initial few thousand samples (see also section 3.7). This was possible because the sampler was certain to have correctly converged after a few thousand iterations. Since the sampler spends an indefinite amount of time searching for and converging towards the main posterior mode in the enlarged, 9-dimensional parameter space, it is now working with fixed covariance settings. While initially the 2.0 PN stationary phase model was used to derive inspiral waveforms directly in the frequency domain (see section 4.3.4), these were meanwhile replaced by higher PN time-domain models, where the resulting waveforms are then numerically Fourier-transformed. When working with signal templates that are *analytically* Fourier-transformed waveforms, these are of course still matched with *numerically* Fourier-transformed data. This may actually introduce some discrepancies, because the analytically Fourier-transformed templates may consider features of the waveform that actually fall outside the observed time stretch of data, while the numerically Fourier-transformed data will be affected by leakage effects that are not reflected in the template. Having both data and matched waveform in the time domain, and only doing the matching (likelihood computation) in the frequency domain, based on numerical transforms, should circumvent these potential pitfalls.

In the following sections, examples are shown for an application of the 5-parameter model using data from a single interferometer, as well as applications of the code developed for the 9-parameter problem, considering data from a network of detectors.

5.1.2 Model and code details

The MCMC sampler was coded in C, and any post-processing, plots etc. of the MCMC output were undertaken in R [55]. The data used was given in the *Frame format* and ported into C using the freely available *Frame library* [111]. This 'data' here was either a file containing separate noise and signal channels which needed to be combined internally, or only the noise, if the simulated signal was generated and injected by the MCMC code. The inspiral signal waveform was modeled using either the 2.0 PN stationary-phase approximation (see section 4.3.4), the 2.5 PN phase / 2.0 PN amplitude approximation (see section 4.3.5), or the 3.5 PN phase / 2.5 PN amplitude approximation (see section 4.3.6). Because the signal is known to be bandwidth-limited well below the data's sampling rate, it was possible to filter and downsample the original data by a factor of 4 (for details see section 3.5), which was done *after* the signal injection. The necessary low-pass filter was designed using a freely available implementation of the 'Parks-McClellan' (or 'Remez exchange') algorithm [112]. For all numerical Fourier transforms the *FFTW* library was used [64]. The noise was modeled as stationary and Gaussian, with a known spectral density (see the description in section 4.5.2). The noise spectrum that goes into the likelihood computations was estimated from a separate section of data that is disjoint from the actually analysed data set, as described in section 3.4.5. In order to reduce leakage effects, windowing was applied to the data; a 'Hann' window was used for spectrum estimation, and a 'Tukey' window was used for the actual data (see section 3.4.3). The Tukey window's parameter was set to $\alpha = 5\%$, and the data segment to be analysed was selected such that the windowing only interfered with the 'left' end (the beginning) of the signal, while the endpoint was set so that the downweighing of data points only set in well after coalescence time. Some of the parameters were reparameterised to allow for easier sampling from the posterior distribution (see also section 3.3). Instead of the individual masses m_1 and m_2, chirp mass m_c and mass ratio η were used (see section 4.2.2), greatly reducing correlations between these pa-

rameters. Instead of the luminosity distance d_L, its logarithm was used, which implicitly leads to an unbounded parameter space, and proposal step widths that are proportional to the distance itself. Inclination angle ι and declination δ were tranformed to $\cos(\iota)$ and $\sin(\delta)$. The 'local' parameters determining the signal observed at a specific interferometer were derived as described in section 4.2.3. The necessary specifications of the interferometers' locations and orientations are given in [82]. In order to determine the eventually relevant coordinates, the Greenwich mean sidereal time (GMST) needs to be derived [83], and the geographical coordinate system needs to be considered [84]. Random number generation within the MCMC sampler was implemented using *Randlib* [113]. The proposal distribution used in the sampler was a multivariate t-distribution with 3 degrees of freedom [42]. This distribution has 'heavier tails' than a Normal distribution, which means that extreme values are more likely to occur; this property makes it a more robust choice as a proposal distribution [42]. In addition to the 'regular' proposals, sometimes (randomly) draws from the prior or moves to 'related' parts of the parameter space were proposed for some parameters, in order to improve convergence and mixing. 'Related' parts of the parameter space could e.g. be a move from inclination ι to $\pi - \iota$, or from phase ϕ_0 to $\phi_0 \pm \pi$, both leading to similar waveforms. If care is taken that the proposal distribution's symmetry property is maintained, the resulting sampler is still a simple Metropolis (not a Metropolis-Hastings) sampler, so that the proposal distribution's density does not need to be considered. For the case of 9 parameters and data from a network of interferometers, the sampler was extended to a parallel tempering algorithm, and then further to an evolutionary MCMC (see section 3.2.9 and following). The evolutionary MCMC algorithm was set up such that 75% of all proposals still were 'regular' Metropolis ('mutation') steps; half of the remaining ('recombination') proposals then were 'real' crossovers, while the other half were 'snooker' crossovers. Due to the high computational costs and the great variability in the output, tuning and evaluation of algorithm parameters is hard or would require a greater systematical effort, and so the setup remained at these ad-hoc settings.

The tempering was implemented so that it only affects the likelihood part of the posterior (and not the prior), as defined in equation (3.11). The temperature ladder was defined through a constant temperature ratio, as described in section 3.2.12. The sampler eventually was set to log every 25th or 50th sample to a text file that was then read into R for eventual analysis.

5.1.3 Single interferometer example

Example setup

The 5 parameters considered here are the two companions' masses m_1 and m_2, coalescence time t_c, coalescence phase ϕ_0 and distance d_E. In this example the distance parameter is denoted by d_E, the *effectice distance*, instead of the luminosity distance d_L. This is owed to the fact that certain parameters are assumed known or fixed here, which otherwise would also affect the signal's amplitude. Since in a realistic setting those parameters (sky location etc.) cannot be known, they would also affect the overall signal amplitude, and with that the distance estimate, which would then not give the "physical" distance, but rather the "apparent" distance. The signal waveform is modeled using the 2.0 PN stationary-phase approximation (see section 4.3.4).

The signal analysed had an effective distance of 25 Mpc, and was embedded in Gaussian and stationary noise that had its noise power spectral density match that of LIGO's target sensitivity [21]. The embedded signal had a signal-to-noise ratio of 10. The true parameter values are given in table 5.1. Approximate posterior samples as starting values for the MCMC chains are generated using importance resampling (see section 3.2.16). The frequency range over which the likelihood is computed was set to 40–1800 Hz. The prior for the 5 parameters was defined as described in section 4.7, the prior range for the masses was defined to be between $0.6\,\mathrm{M}_\odot$ and $3\,\mathrm{M}_\odot$, and the coalescence time's range was within ± 5 ms of the true value.

Six parallel chains were run; the starting points of the chains were gen-

erated by importance resampling of 100 000 draws, a number that proved to yield enough eventual draws that were sufficiently close to the main posterior mode to ensure reliable and fast convergence of the Metropolis algorithm. The first 30 000 iterations of each chain were considered the burn-in-phase, during which the iterations 15 000–30 000 were used to tune the proposal covariance. The code was then run for 2 million iterations in total, which after thinning out of the samples and discarding the burn-in yielded a sample of size 236 400 from the posterior. The 'multivariate potential scale reduction factor' $\widehat{R}^p$ was close to 1 ($\widehat{R}^p = 1.0034$), indicating convergence of all chains [47].

Posterior inference

Figures 5.1–5.4 illustrate estimates of marginal posterior densities. Firstly, figure 5.1 shows posterior densities for the five individual parameters. Most of them exhibit a mode near the true parameter value (indicated by

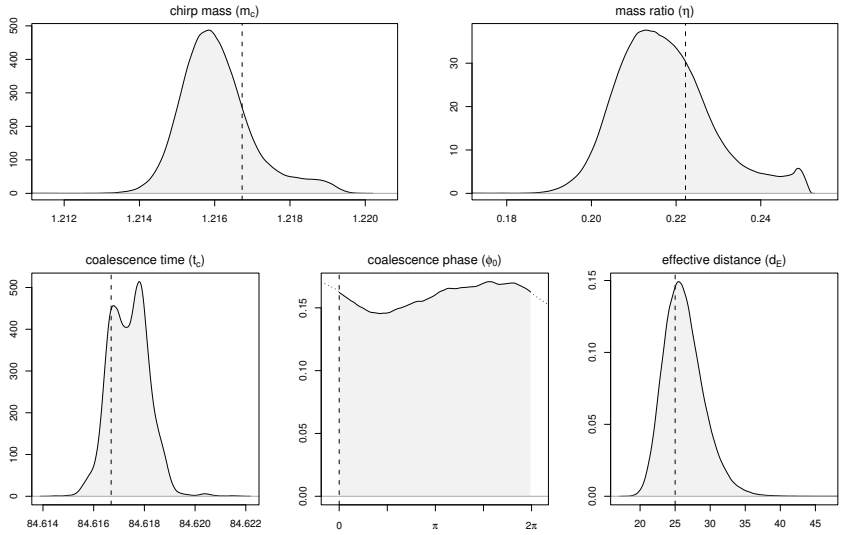

Figure 5.1: Marginal posterior densities of the five parameters. Dashed lines indicate the true values.

dashed lines). One can see that the relative precision of parameter estimation varies significantly between different parameters. For example, the

posterior of the chirp mass covers a range of about 0.006 $M_\odot$, while the prior range initially was some 2.1 $M_\odot$. The coalescence phase's posterior, on the other hand, still covers the complete prior domain.

Figure 5.2 allows for some insight into joint distributions of some of the parameters. The joint density of chirp mass and mass ratio (5.2a) shows a positive correlation between the two parameters. Figure 5.2b shows interaction between two parameters (ϕ_0 and η), and in particular demonstrates that although the marginal density of ϕ_0 alone is almost uniform (see figure 5.1), this does not imply that its effect on the posterior was negligible.

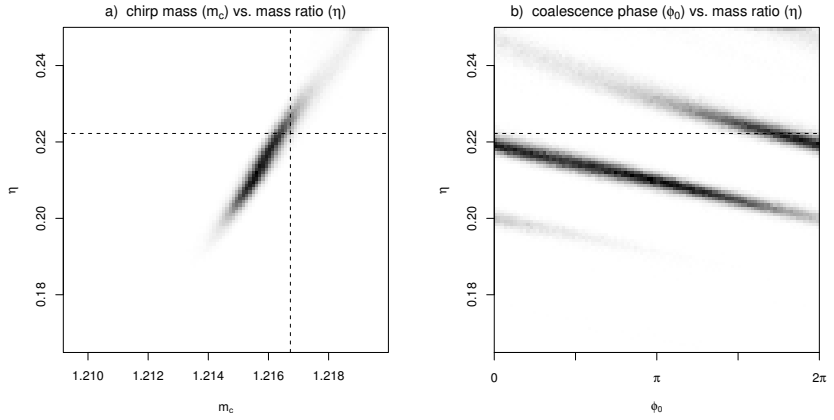

Figure 5.2: Bivariate marginal posterior densities for two pairs of parameters. Dashed lines indicate the true values, the true coalescence phase is $\phi_0 = 0$ (Histograms, the greyscale plots show relative densities normalised to the mode).

The MCMC sampler internally works with chirp mass (m_c) and mass ratio (η) instead of individual masses (m_1, m_2). A posterior sample of the individual masses still can easily be obtained by back-transforming each pair of (m_c, η) samples. Figure 5.3 shows these two marginal densities combined into one plot.

Analogously, other functions of the parameters can be derived and distributional features investigated; if e.g. one was interested in whether the masses differ 'significantly' or are 'almost equal', we can estimate:

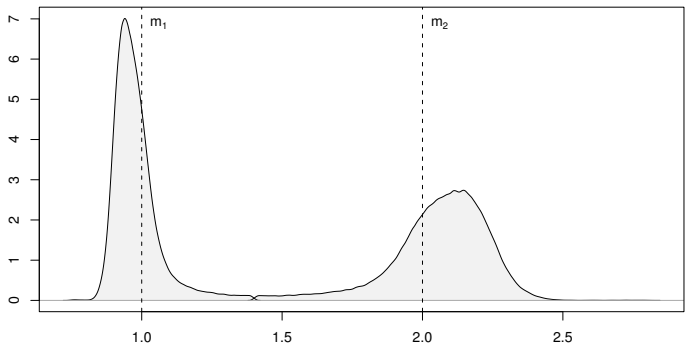

Figure 5.3: Combined plot of marginal posterior densities of the two companions' individual masses. Dashed lines indicate true values.

$\mathrm{P}(m_2 > 3m_1) = 0.11\%$ or $\mathrm{P}(m_2 < 1.5m_1) = 4.84\%$. Figure 5.4 shows the posterior density of the logarithmic amplitude $\mathcal{A}(m_1, m_2, d_E)$ (4.38). Comparing it to the prior density you can see that, since it is significantly above the reference points x_U and x_L, the particular specification of the lower bound of the parameter space does not affect our conclusions. Table 5.1 shows summary statistics of the posterior distributions of the inspiral's parameters.

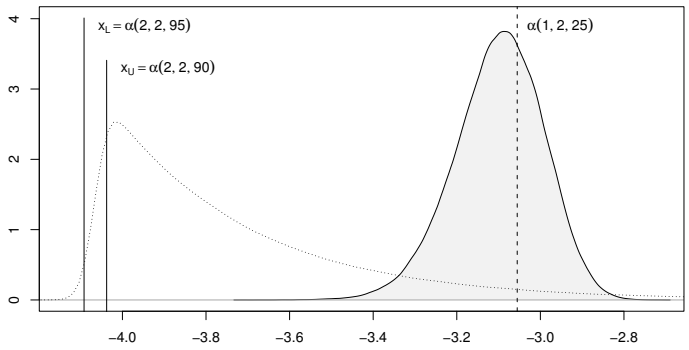

Figure 5.4: Posterior density of the signal's logarithmic amplitude $\mathcal{A}(m_1, m_2, d_E)$; the dashed line indicates the true value. The prior density (dotted line) and x_U and x_L are shown as well.

Table 5.1: Posterior estimates: Means, medians and 95% central credible intervals for several parameters.

parameter		mean	median	95% c.c.i.	true	unit
chirp mass	(m_c)	1.2161	1.2159	$[1.2145, 1.2186]$	1.2167	$M_\odot$
mass ratio	(η)	0.2174	0.2162	$[0.1987, 0.2457]$	0.2222	
coalescence time	(t_c)	84.6174	84.6174	$[84.6160, 84.6189]$	84.6167	s
coalescence phase	(ϕ_0)		*— not meaningful —*		0.0	radian
effective distance	(d_E)	26.28	25.99	$[21.55, 32.68]$	25.00	Mpc
mass 1	(m_1)	0.980	0.964	$[0.876, 1.229]$	1.0	$M_\odot$
mass 2	(m_2)	2.062	2.085	$[1.600, 2.327]$	2.0	$M_\odot$

5.1.4 Coherent network inference example

Example setup

The simulated data in this example represent an inspiral event that is measured at three interferometers, namely the two LIGO sites Hanford (WA, USA) and Livingston (LA, USA), and the Virgo detector near Pisa (Italy). The simulated inspiral event is an inspiral of two companions with masses of $2\,M_\odot$ and $5\,M_\odot$, taking place at a distance of 30 Mpc from Earth; the values of the remaining parameters are given in table 5.2. Figure 5.5 shows the different waveforms that are measured at each site, without noise and for the very last orbits before coalescence.

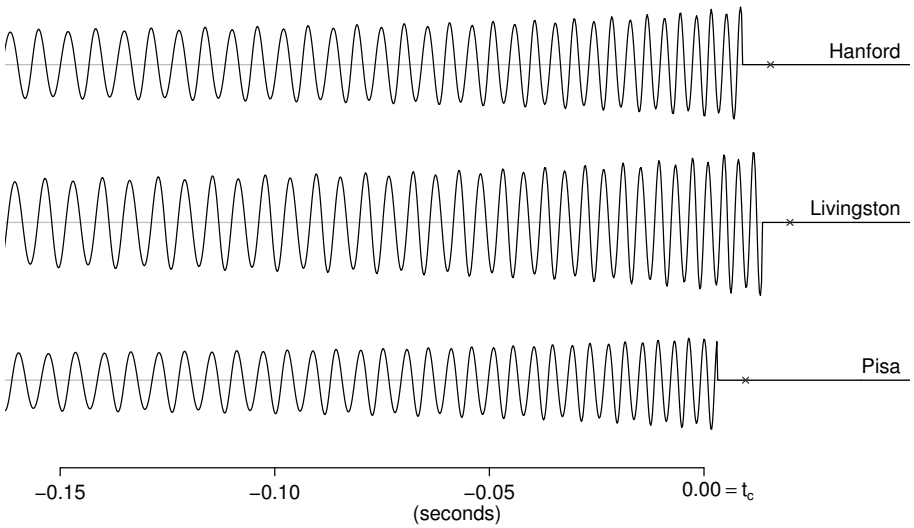

Figure 5.5: The signal waveforms (without noise) measured at the 3 different interferometer sites and modeled with 3.5PN phase and 2.5PN amplitude accuracy. Note especially the slightly differing arrival times and amplitudes due to the interferometers' locations and orientations. The time axis refers to the coalescence time *at the geocentre* ($t_c^{\oplus}$); the local coalescence times are marked by crosses.

The simulated signals were embedded in (synthetic, Gaussian) coloured noise with noise spectral densities matching those expected for the instruments at their target sensitivities [21, 22]. The noise curves are shown in

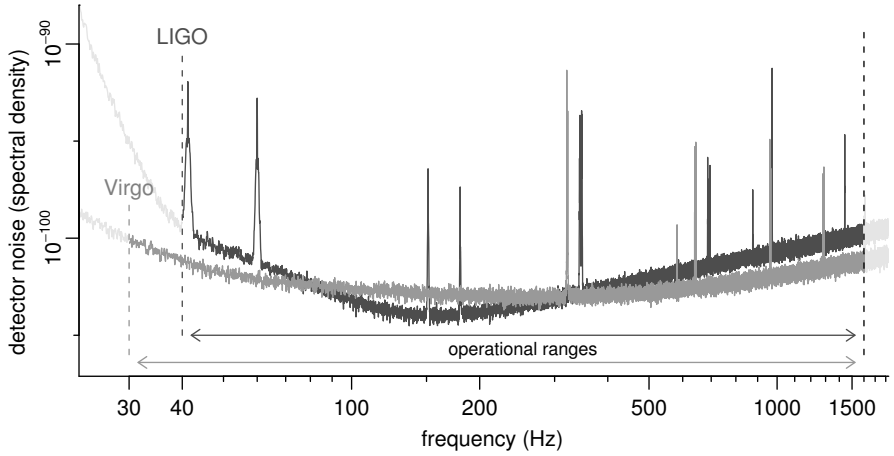

Figure 5.6: Noise spectra for the different interferometers. The spectra are (virtually) the same for the two LIGO interferometers (Hanford, Livingston), but different for the Virgo interferometer (Pisa). The operational ranges (frequency ranges considered for inference) are indicated as well.

figure 5.6. The resulting signal-to-noise ratios at the individual interferometers were: Hanford 7.05, Livingston 9.04, Pisa 5.51, and in total 12.72.

The original time resolution of the data (before downsampling) was 16 384 Hz sampling rate for LIGO data, and 20 000 Hz for Virgo measurements. The frequency range to be considered for likelihood computations was set to 40–1600 Hz for LIGO and 30–1600 Hz for Virgo. The amount of data to be considered for inference was 12 s for LIGO, and 23 s for Virgo. This reflects the time an inspiral of this kind would spend radiating in the corresponding interferometers' sensitivity bands, and could in a realistic setting be defined either with respect to rough estimates from the detection pipeline, or to worst-case considerations.

The prior was defined as described in section 4.7, and in particular the prior distribution for the masses was set to be uniform across 1–10 $M_\odot$, and the coalescence time prior was defined as uniform across ± 10 ms around the true value. The settings for an inspiral event's detectability were (the same as in the example in section 4.7.3) such that an inspiral of two masses of 2 $M_\odot$ each is assumed to be detectable out to 50 Mpc and

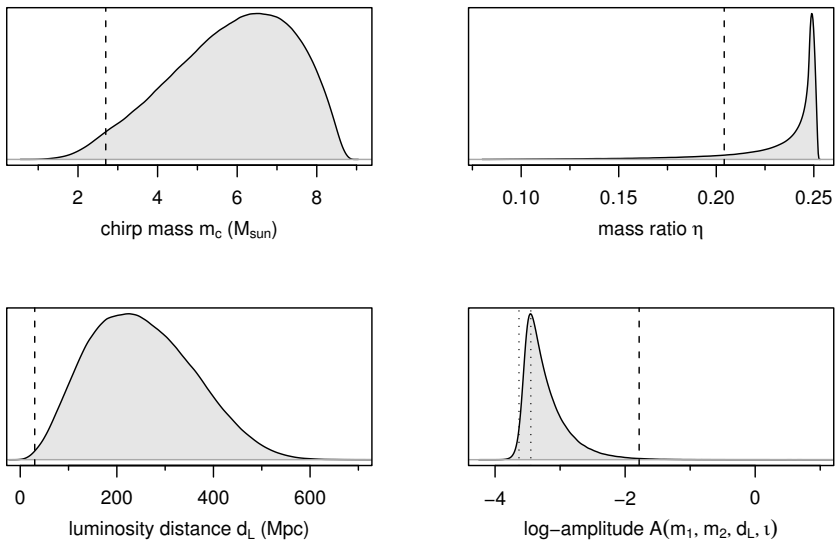

Figure 5.7: Kernel density estimates of some marginal prior densities. Dashed lines indicate the parameter values of the injected signal. The two dotted lines in the lower right plot indicate the values of x_L and x_U.

60 Mpc distance with 90% and 10% probabiliy, respectively (that is, $x_U = \mathcal{A}(2\mathrm{M}_\odot, 2\mathrm{M}_\odot, 50\mathrm{Mpc}, \frac{\pi}{2})$, $x_L = \mathcal{A}(2\mathrm{M}_\odot, 2\mathrm{M}_\odot, 60\mathrm{Mpc}, \frac{\pi}{2})$, and $p = 0.1$).

These settings also correspond to what was shown in figure 4.4 (section 4.7.3). The resulting 1-dimensional marginal prior densities for some of the parameters and for the (logarithmic) signal amplitude $\mathcal{A}$ are shown in figure 5.7. The values of x_L and x_U defining the lower bound for the amplitude, and consequently affecting the shape of the other involved parameters' distributions are shown as well. Larger values of $\mathcal{A}$ are less likely because they occur too rarely, and lower values of $\mathcal{A}$ are less likely because the resulting signals become too faint to be noticeable. The parameter values of the simulated signal are not very centrally located, compared to the prior densities as shown in figure 5.7. That should not be a problem though, because it is the 'local' shape of the prior density that is more important for the shape of the posterior (if the parameters are well determined by the data, i.e. if posterior distribution is narrow). It might however become a problem if the parameters are not well determined through

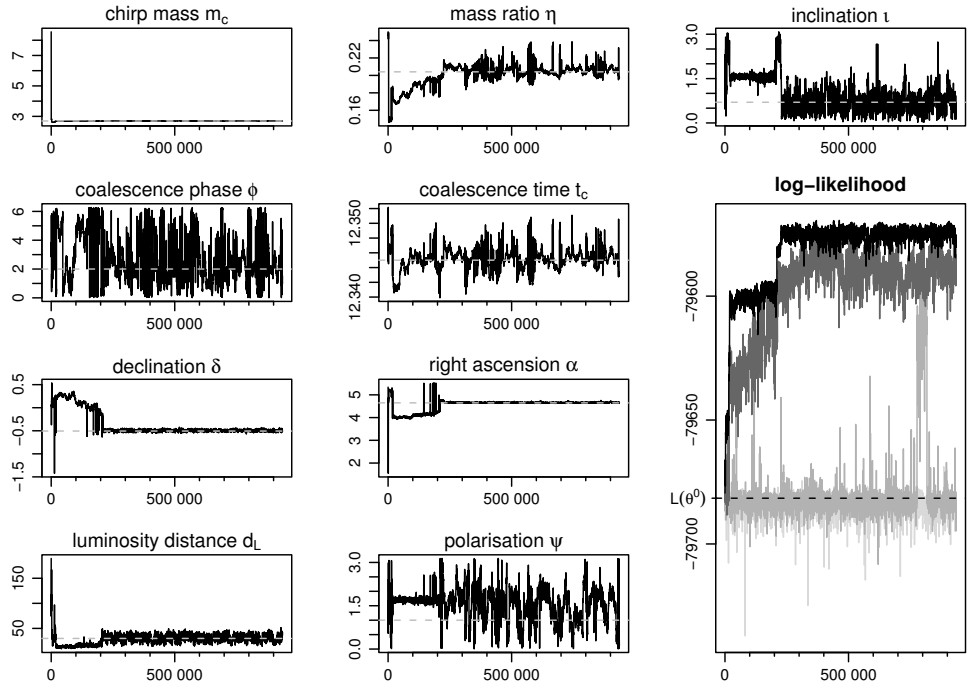

Figure 5.8: Trace plots of the 9 parameters of the first chain in a run with four chains, dashed lines indicate the true parameter values. The bottom right plot shows the corresponding (unnormalised) logarithmic likelihood values for *all four* chains, higher temperatures are indicated by ligher shades of grey. The dashed line here indicates the value of $\mathcal{L}(\theta^{\emptyset})$, the "null" likelihood for $\theta^{\emptyset}$ indicating *no signal*.

the data—if there is little to be learned from the data, then the prior knowledge becomes of greater relevance. What places the parameter values at the margin of the prior distribution (especially with respect to d_L and $\mathcal{A}$) is the conservative (low) choice of the prior's boundary (x_L and x_U).

Posterior inference

Different runs of the MCMC code on the simulated data were started, one starting from the true parameter values, and all eventually agreed and converged towards the same region of the parameter space. Figure 5.8 shows trace plots for one of the chains, together with a trace plot of the

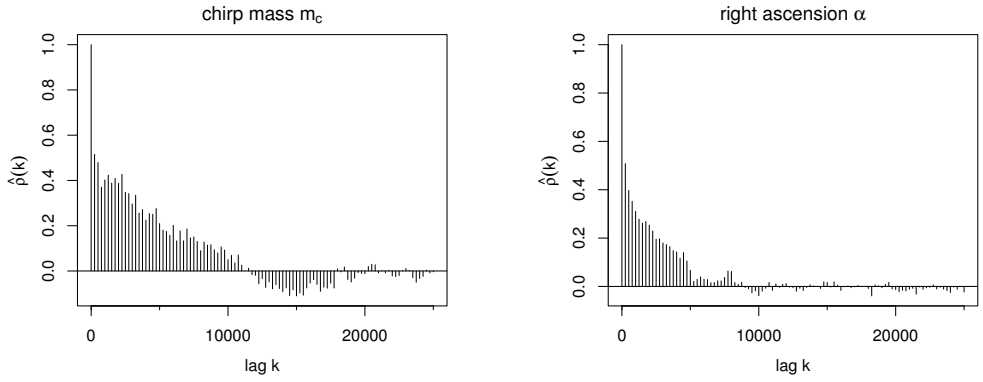

Figure 5.9: Autocorrelation functions for two of the parameters, corresponding to the MCMC run illustrated in figure 5.8.

corresponding (unnormalised) likelihood values. Here, the first chain (at temperature $T_1 = 1$) converged after about 250 000 iterations. In the bottom right plot you can see how good parameter sets are "handed down" to lower-temperature chains, e.g. at around the 200 000th iteration, from 2nd to 1st chain. The 3rd chain briefly visits a mode, but returns to regions of lower likelihood, while the 4th chain keeps sampling at about the prior. The dashed line in the likelihood plot indicates the value of $\mathcal{L}(\theta^{\emptyset})$, the "null" likelihood for $\theta^{\emptyset}$ indicating *no signal* (see also section 3.2.12). Chains at high temperatures obviously keep sampling at around this level of likelihood. Relative to $\mathcal{L}(\theta^{\emptyset})$, the (unnormalised) log-likelihood values are directly proportional to the *deviance* $D(\theta) = 2\log\left(\frac{\mathcal{L}(\theta)}{\mathcal{L}(\theta^{\emptyset})}\right)$ with respect to $\theta^{\emptyset}$, indicating the evidence against $\theta^{\emptyset}$ in the data (see also section 3.2.12). Note also the increasing variance and decreasing mean in the sampled posterior values for higher-temperature chains as predicted by (3.15). Acceptance rates for these four chains were at 26, 29, 60 and 63 percent, with greater acceptance rates at higher temperatures. The resulting MCMC chain exhibits a great amount of correlation between subsequent samples. Figure 5.9 shows estimated autocorrelation functions $\hat{\rho}(k)$ for two of the parameters: autocorrelations here decay to near zero only after a lag of some 10 000 iterations. The *integrated autocorrelation time*

(IACT) gives an idea of the time lag to be expected between two "effectively independent samples":

$$IACT \;=\; 1 + 2\sum_{k=1}^{\infty} \rho(k) \tag{5.1}$$

[114]. Estimated IACTs (derived by substituting the autocorrelation ρ by its estimate $\hat{\rho}$ in 5.1 and truncating the sum at $k = 20\,000$) corresponding to the trace plots in figure 5.8 are m_c : 4200, η : 5000, t_c : 7000, δ : 760, α : 2600, ψ : 10 000, ι : 3800, ϕ_0 : 6800 and d_L : 3500 (note that these figures, just as the axis labels in figures 5.8 and 5.9, refer to the individual MCMC iterations *before* any thinning out of samples).

Figure 5.10 shows kernel density estimates of the marginal posterior distributions of each of the 9 individual parameters. All the marginal distributions cover the corresponding true parameter values of the in-

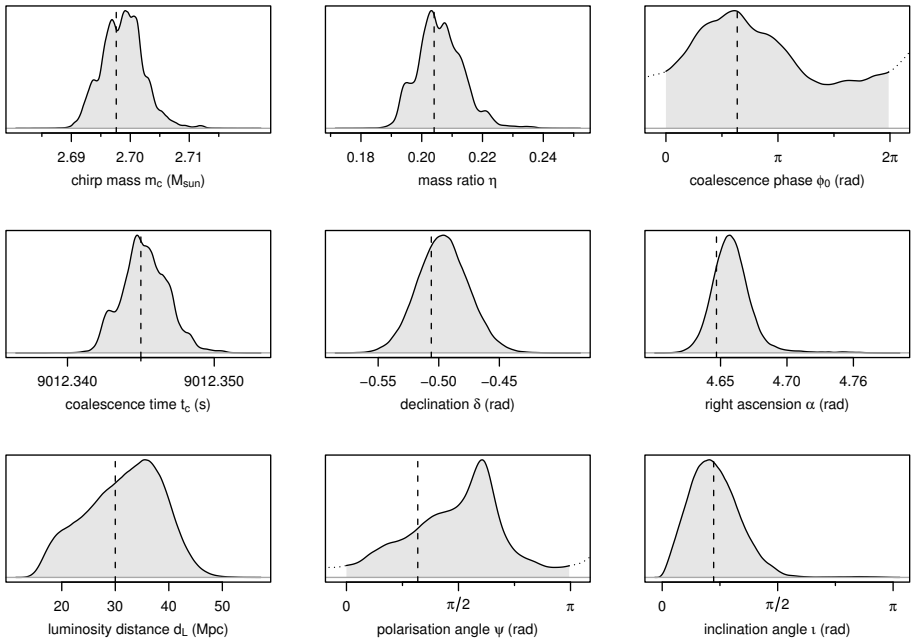

Figure 5.10: Kernel density estimates of the (marginal) posterior densities for each of the 9 parameters. Dashed lines indicate the true parameter values.

jected signal, but there are great differences in the accuracies with which these can be inferred from the data. The chirp mass m_c for example is determined with great accuracy (the posterior standard deviation is only 0.125% of the true value), while on the other hand the posterior distribution of the coalescence phase ϕ_0 still covers the complete prior range $[0, 2\pi]$.

More detailed insight into the posterior distribution is gained by looking at joint (marginal) distributions of pairs of parameters. Some estimates of such densities are shown in figure 5.11. Although correlation between the mass parameters was greatly reduced through the reparametrisation, some correlation still remains between the new parameters chirp mass m_c and mass ratio η. From the marginal density of phase ϕ_0 and coalescence

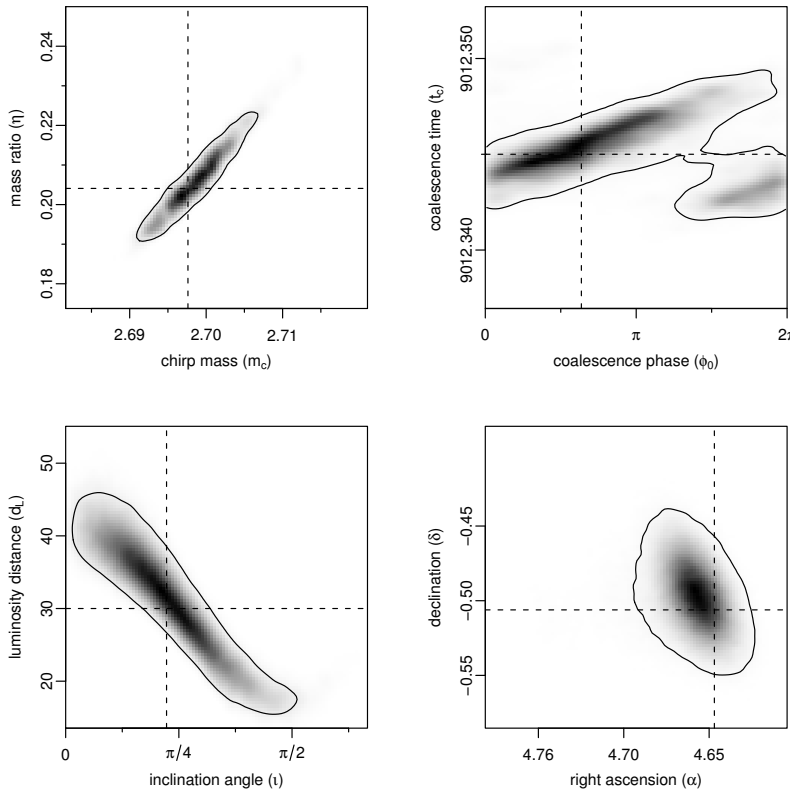

Figure 5.11: Estimates of 2-dimensional (marginal) joint posterior densities for 4 pairs of parameters, and 95% confidence regions. Dashed lines indicate the true parameter values.

time t_c one can see that while the marginal distribution of the phase alone appears rather undetermined when considered alone (see figure 5.10), the value of this parameter is important when seen in conjuction with other parameters. The lower left plot shows that the uncertainty in the distance d_L is tied to the uncertainty in the inclination ι. Both parameters affect the signal's amplitude (see (4.8) and (4.9) in section 4.3.2); the distance only affects the overall scale, and the inclination also shifts weight between '$+$' and '$\times$' terms. Looking at the posterior distribution of $\mathcal{A}$, the (approximate) logarithmic overall amplitude (see table 5.2), one can see that this is determined with greater accuracy than the distance—the posterior standard deviation of the amplitude is 7.7% of the true value, while the distance varies by 23% of the true value, three times as much. So the variation of ι and d_L that can be seen in figure 5.11 is mostly along values implying similar *total* amplitudes.

Table 5.2 finally lists some numerical estimates for the signal parameters or derived quantities. The posterior distribution of the (approximate) logarithmic amplitude $\mathcal{A}$, a function of masses, distance and inclination, is shown as well. Compared to its prior distribution (shown in figure 5.7) one can see that the posterior is far from the prior's boundary specified by x_L and x_U, suggesting that the exact settings for these do not affect the shape of the posterior distribution. Looking at the corresponding posterior means, chirp mass and luminosity distance are slightly overestimated, while the amplitude is slightly underestimated; this is probably due to the prior definition (i.e., the Malmquist effect). The true amplitude here is at about the posterior's 84% quantile.

Effects of varying signal characteristics

In order to check how the posterior distribution is affected by different signal properties, additional MCMC runs were performed with varying settings of the true parameter values of the simulated signal. The following simulations were undertaken using slightly different signal templates (2.5 PN phase, 2.0 PN amplitude approximation), and a prior that does not

Table 5.2: Some key figures summarising the marginal posterior distributions of individual parameters, where meaningful. Mean and standard deviation indicate location and spread, and the 95% central credible interval gives a range that contains the true parameter with 95% probability, given the data at hand.

	mean	st.dev.	95% c.c.i.	true	unit
chirp mass (m_c)	2.6987	0.0036	(2.6922, 2.7062)	2.6976	$M_\odot$
mass ratio (η)	0.2062	0.0076	(0.1931, 0.2224)	0.2041	
coalescence time (t_c)	12.3453	0.0016	(12.3424, 12.3485)	12.3450	s
luminosity distance (d_L)	31.6	7.0	(17.7, 43.6)	30.0	Mpc
inclination (ι)	0.722	0.356	(0.154, 1.465)	0.700	rad
declination (δ)	-0.496^a	0.025^a	$(-0.537, -0.455)$	-0.506	rad
right ascension (α)	4.659^a		(4.633, 4.692)	4.647	rad
coalescence phase (ϕ_0)	1.878^a	1.235^a		2.0	rad
polarisation (ψ)	1.602^a			1.0	rad
mass 1 (m_1)	2.028	0.089	(1.888, 2.227)	2.0	$M_\odot$
mass 2 (m_2)	4.936	0.232	(4.441, 5.334)	5.0	$M_\odot$
total mass (m_t)	6.964	0.144	(6.668, 7.222)	7.0	$M_\odot$
log-amplitude ($\mathcal{A}$)	-1.866	0.084	$(-2.050, -1.715)$	-1.785	

[a] *mean direction* and *spherical standard deviation*

yet take into account the inclination angle's effect on the signal amplitude [13].

As one would expect, the precision of parameter estimation is proportional to the signal's strength; table 5.3 shows the standard deviations of some of the parameters' posterior distributions. Here the mass and distance parameters were varied while the other parameters were held constant. The posterior is narrowest for a close-by inspiral of high masses, and gets wider for both lower mass or greater distance.

These results are in agreement with earlier estimates of the accuracy to be expected from such parameter estimates [29]. The great difference in relative accuracies of parameters related to phase evolution (like chirp mass m_c and reduced mass $\mu = \frac{m_1 m_2}{m_1+m_2} = m_t\eta$) versus those affecting the signal's amplitude (like distance d_L) is confirmed, and the correlation between m_c and μ is verified as well.

At decreasing SNRs, certain parameters cannot be determined unam-

Table 5.3: Individual and total SNRs for different signals, and some characteristics of the resulting posterior distributions. The accuracy of some of the parameters is illustrated by the posterior standard deviations for (δ, α), t_c, d_L, m_c and μ (percentages refer to the true value). The correlation coefficient for m_c and μ shows the (posterior) interdependence between the two parameters. These results are consistent with those presented in [29].

masses	distance	network	posterior standard deviations					
m_1-m_2	d_L	SNR	(δ, α)[a]	t_c	d_L	m_c	μ	$\mathrm{Cor}(m_c, \mu)$
1.5-2.0 $M_\odot$	10 Mpc	29.6	0.011 rad	0.26 ms	20 %	0.016 %	0.35 %	0.95
1.5-2.0 $M_\odot$	20 Mpc	14.8	0.030 rad	0.49 ms	25 %	0.031 %	0.69 %	0.94
1.5-2.0 $M_\odot$	30 Mpc	9.9	0.207 rad	1.04 ms	25 %	0.074 %	1.33 %	0.91
2.0-2.0 $M_\odot$	10 Mpc	33.3	0.008 rad	0.14 ms	14 %	0.009 %	0.14 %	0.80
2.0-2.0 $M_\odot$	20 Mpc	16.7	0.017 rad	0.28 ms	18 %	0.014 %	0.23 %	0.73
2.0-2.0 $M_\odot$	30 Mpc	11.1	0.026 rad	0.42 ms	21 %	0.021 %	0.37 %	0.78

[a] *spherical standard deviation*

biguously any more. One example is the inclination angle ι, which still has a 'well-behaving' posterior distribution at 10 Mpc distance (similar to that shown in figure 5.11). For a weaker signal originating from 30 Mpc distance, the distribution then turns bimodal (figure 5.12). The 'orientation' of the inclination angle is not clear any more, the result being two roughly equally likely 'mirror image' solutions with $\mathrm{P}(\iota < \frac{\pi}{2}) \approx \frac{1}{2} \approx \mathrm{P}(\iota > \frac{\pi}{2})$. Note that the two solutions ι and $\pi - \iota$ correspond to opposite orbital directions (clockwise/counterclockwise), as seen from Earth, which might be of minor interest anyway.

The sky location's posterior also exhibits multiple modes for this weaker signal (figure 5.12). This illustrates some potential pitfalls of maximum-likelihood (ML) or maximum-a-posteriori (MAP) methods; these would advise picking the highest of the several modes, which might just be the narrowest one, but not necessarily the most likely. If one then proceeded by extrapolating the curvature at that mode and deriving error bounds from the Fisher Information matrix, the resulting estimates might not only be far off, but also associated with overestimated accuracies.

MCMC runs with a modified prior setting were also tried; the prior for the coalescence time t_c was extended from its original range of $\pm$5ms

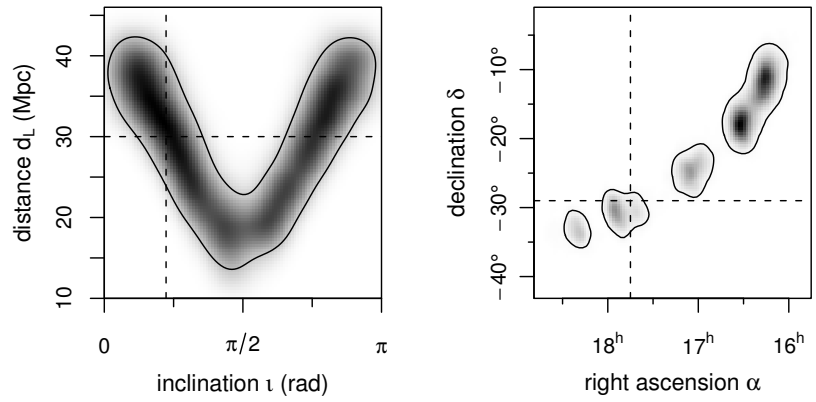

Figure 5.12: At greater distance the 'orientation' of the inclination angle ι cannot be resolved any more, both directions are roughly equally likely ($P(\iota < \frac{\pi}{2}) \approx \frac{1}{2} \approx P(\iota > \frac{\pi}{2})$). At the same time, with the lower SNR the sky location's posterior turns multimodal. (Dashed lines indicate the true values.)

around the true value to $\pm$27ms, allowing for an additional margin of 22ms, which is the time it takes a gravitational wave to travel from Earth's surface to its center. This setting reflects the case where the inspiral detection pipeline received triggers from less than three interferometer sites, so the signal's arrival time at the geocenter could not be estimated to greater accuracy in advance. The MCMC algorithm is still able to find the mode in the enlarged time parameter range, but takes more iterations to converge.

One scenario in which such an approach would be necessary is when the SNR for one of the interferometers is almost zero. In such a case the data from the interferometer under consideration also would not (directly) contribute to the estimation of phase- and frequency-related parameters, but would still carry information about amplitude-related parameters—by implicitly 'ruling out' those parameter combinations that *would* have resulted in a response at that interferometer. Figure 5.13 shows the sky location posteriors for such a signal, a 1.5-2.0 $M_\odot$ inspiral at 10 Mpc distance, where the SNRs at the three interferometer sites are: Hanford 9.6, Livingston 13.9, Pisa 0.18 (total 16.9). Including the data from the third

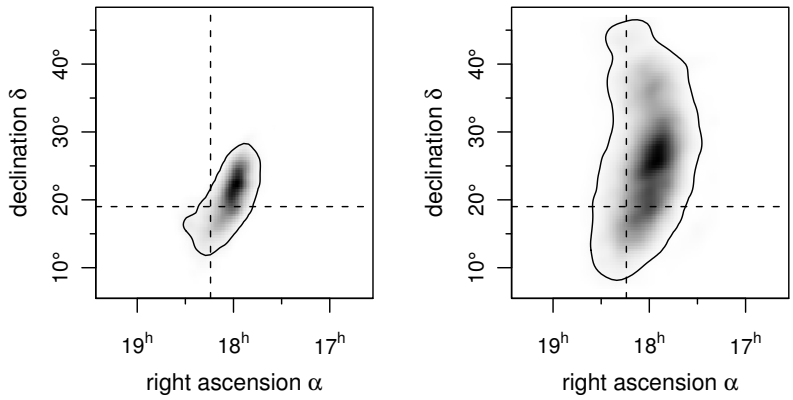

Figure 5.13: Even if the SNR at one of the interferometers is almost zero, it still contributes to estimates' accuracies—the posterior is much narrower if its data is included (left plot) than if it is omitted (right plot). (Dashed lines indicate the true values.)

interferometer (with almost zero SNR) into the analysis still yields a much more accurate estimate of the sky location. Table 5.4 compares the resulting parameter accuracies of these two settings. The posteriors for sky location (δ, α) and coalescence time t_c, which are closely related, are much narrower when the Virgo data is considered in the analysis, while estimates of the rather phase- and frequency-related parameters m_c and μ do not gain from the additional information.

Table 5.4: Relative accuracies of different parameters (in analogy to table 5.3) when considering / not considering the Virgo data (where the example situation is such that the SNR is nearly zero).

Virgo data...	(δ, α)	t_c	d_L	m_c	μ
...included	0.071 rad	0.81 ms	21 %	0.023 %	0.33 %
...excluded	0.150 rad	2.38 ms	23 %	0.022 %	0.31 %

On the one hand, not only a high (total) SNR is desirable but also one that is rather 'evenly spread' over different interferometers. On the other hand, even a near-zero SNR at one of the interferometers does not make its

measurement useless. Inference on different parameters will be affected to different degrees by such an unbalanced SNR arrangement.

5.2 Inference on inspiral signals using LISA data

5.2.1 Introduction

This section describes an example application of a Bayesian inference framework for the analysis of binary inspiral gravitational wave signals as observed through the Laser Interferometer Space Antenna (LISA). The data here are taken from the 2nd round of the *Mock LISA Data Challenges (MLDC)* [90]. The MLDC were initiated in order to foster the development of methods related to the analysis of data as produced by LISA. In each round, data sets are published in order to be analysed by different participating groups. At certain deadlines, results are submitted and compared to the true values that were to be inferred, and between groups [32]. The types of analysis problems posed so far cover continuous sources forming a galactic foreground, and different types of chirping sources. On the way towards more realistic scenarios, the 2nd round in particular included a single data set containing many different superimposed signals.

The results presented below were produced in collaboration within the *Global LISA Inference Group (GLIG)* that was formed in mid-2006 in order to work on different aspects of data analysis for LISA. Very important in this context, the *LISA Simulator* [80] was 'dissected' and adapted for our purpose (due in large measure to GLIG member Alexander Stroeer) in order to use it for numerical derivation of LISA's response to passing gravitational waves. This is necessary for likelihood computation, at least when there is no other approximation available. Resorting to existing code firstly spared us from having to "re-invent the wheel", and since the LISA simulator was also used for data generation, there were in principle no worries about matching of simulated detector responses. A disadvantage on the other hand is that the LISA Simulator code runs relatively slowly, since originally it was never intended to run repeatedly and fast. Parts of the developed code where then shared and also used in a different MLDC application [115].

A simple Metropolis sampler was set up in late 2006, and first results of an application to data from the 1st round of the MLDC were presented at the *11th Gravitational Wave Data Analysis Workshop (GWDAW-11)* in December that year [16]. At that point it was obvious that the algorithm was too slow and too inflexible. Sampling worked fine in principle, but the sampler was too slow to converge towards the global posterior mode in time, and it also could not move efficiently in parameter space, due to the posterior surface exhibiting multiple narrow modes, and due to the comparably large expected signal-to-noise ratios. Both problems were approached by extending the code to a parallel tempering algorithm, and implementing it in a parallel fashion. Due to the nature of the data, where the noise spectrum is not known beforehand, the model also needed to be generalised to incorporate the spectrum as an unknown. After these modifications were implemented, results were presented at the *7th Edoardo Amaldi conference on gravitational waves (Amaldi7)* in July 2007. It is these results that are illustrated in the following.

5.2.2 Model and code details

The (simulated) data analysed here is taken from the 2nd round of the Mock LISA Data Challenges (MLDC) [92]. The data sets are published in XML format (*LisaXML*, [91]), and they contain the three X, Y and Z observables recorded over time. The analysis code is written in C, where the data is imported and then the 'A' and 'E' TDI variables (see section 4.4.2) are constructed. The $+/\times$ polarisation waveforms are modeled following the *restricted PN* inspiral signal description (see section 4.3.3), and the corresponding detector response (in terms of A and E TDI variables) is then numerically derived using code that was extracted and adapted from the *LISA Simulator* [80]. Due to the nature of the noise, which is partly made up of many 'deterministic' but unknown signals, and whose spectrum is not known in advance, the noise spectrum was included as unknown into the model (see also section 4.5.3). Discrete Fourier transforms within the algorithm were performed using the *FFTW* library [64]. The data (as well

as the corresponding matched signal templates) were windowed using a Tukey window (with $\alpha = 2\%$; see section 3.4.3). The MCMC algorithm was then set up as a Metropolis sampler (for the 9 'signal' parameters), in combination with a Gibbs step (for the remaining 'noise' parameters). The proposal distribution used within the Metropolis sampler was a multivariate Student-t-distribution. Random number generation was done using the *Randlib* library [113]. Some of the signal parameters were reparameterised for easier sampling, in particular the masses were expressed in terms of chirp mass m_c and mass ratio η, instead of the inclination angle its cosine was used, and the luminosity distance was transformed to its logarithm (see sections 3.3 and 4.2). The 'basic' MCMC was extended to a parallel tempering algorithm, where the tempering is only applied to the likelihood part of the posterior (as described in section 3.2.9), and consequently included the conditional distributions of the noise parameters (see section 4.5.3). With the large number of parameters and the large SNR (and consequently the large number of parallel chains required) it was essential to have a guideline for the setup of the temperature ladder available, since this clearly could not be set up through trial-and-error. In order to gain speed, the parallel tempering was implemented in a parallel fashion [76]. Using the *Message Passing Interface (MPI)* [77], the algorithm can be set up so that each of the tempered MCMC chains is handled by a single process, and for each attempted 'swap' (of parameters or, equivalently, of temperatures), messages are passed between processes in order to compare and possibly swap their current states. Here, the *Open MPI* implementation was used [116]. When running such a code on a cluster, the different processes can then easily be distributed over multiple processors, where the number of processes may be larger than the number of processors.

5.2.3 Example setup

The sampling interval of the data is 15 seconds, and in the following, the sample size is always $N = 2^{17} = 131\,072$ samples (for each of the two A

and E variables), which corresponds to $\approx$ 23 days of measurements. The simulated data originates from *challenge 2.2* of the MLDC, which (besides the non-white instrumental noise) contains a superposition of signals from a galactic population of binary systems ($\approx$ 26 million signals at different frequencies and locations), and about 5 massive black hole inspiral signals (chirp signals) and extreme mass ratio inspiral signals (modulated chirp signals) each [92].

Since the two TDI variables can be considered as perceiving the same noise, a common noise spectrum was assumed for both. The prior for the noise spectrum was defined by using an empirical spectrum estimate from a disjoint stretch of data to set the prior scale parameters (see also section 3.4.5). The prior's degrees of freedom for each frequency bin were set to 2, and consequently each bin's (conditional) posterior distribution then has $2 + 4 = 6$ degrees of freedom. The prior for the remaining signal parameters was set as described in section 4.7, with its boundary for now specified by assuming an inspiral event of $2 \times 10^6\,\mathrm{M}_\odot$ and $\frac{\pi}{2}$ inclination to be detectable out to distances of 100 000 Mpc and 110 000 Mpc with 90% and 10% probability, respectively.

With 2^{17} samples (of two variables) in the data, there are $2^{17}/2 = 2^{16} \approx$ 65 000 noise parameters in the model. Since the signal is bandwidth-limited, the likelihood computation was simplified by restricting it to a limited frequency band of 0.000 05–0.01 Hz. This way only $\approx$ 20 000 noise parameters need to be considered for likelihood computations.

Rough estimates of chirp mass and coalescence time for the massive black hole inspiral signals were derived [117] using a *time-frequency analysis*, i.e. by investigating spectra of short data segments over time. This way, three inspiral signals were found in the data set, and in the following the analysis was aimed at the first of these three triggers. Runs of the MCMC code from these starting points did not converge in time for the submission deadline for results, and so the algorithm was then started from the true parameter values that were announced later.

Based on the number of parameters and the approximation given in section 3.2.12, a parallel tempering algorithm yielding a swap acceptance

rate of 25% would need a temperature ladder with a ratio of $q = 1.017$ between neighbouring temperatures. Due to the large number of parallel chains required, this ratio was increased to $q = 1.025$, which should still result in a swap acceptance rate of $\approx 8.4\%$. When running the algorithm starting from the rough parameter estimates, different numbers of parallel chains (up to 96) were tried, but due to the limited time and computing resources the chains would not converge to a unique mode, although there was some indication of convergence for some of the parameters. Once the true paramater values were published, the algorithm was then restarted from there, again, with varying numbers of chains, but since there was no indication of multiple (relevant) modes in the posterior, the results presented in the following were then produced using only four parallel chains.

Every 10th MCMC sample (of the 9 signal parameters) was eventually stored in a text file for subsequent posterior analysis. The posterior spectrum was logged by (internally) computing the conditionally expected spectrum (as in equation (4.24)) for chain 1 at temperature $T = 1$ every 10th iteration and then recursively updating an average spectrum (as in section 3.7) which was then recorded every 5000 iterations in a seperate file. This way only the posterior *mean* spectrum can be inferred (over arbitrary segments of the MCMC chain), but only comparably little storage space is required, although this alone still accounts for a significant share.

5.2.4 Posterior inference

The parallel tempering algorithm was run for 650 000 iterations, starting from the true parameter values. There was no value given for the coalescence phase ϕ_0, since the data were generated using a different parametrisation (*initial* phase) [92], but that one missing parameter could easily and unambiguously be inferred. The resulting marginal posterior distributions for individual parameters are shown in figure 5.14. Some of the parameters are more or less correlated, for example the mass parameters or the location parameters; the (marginal) bivariate distributions of these two

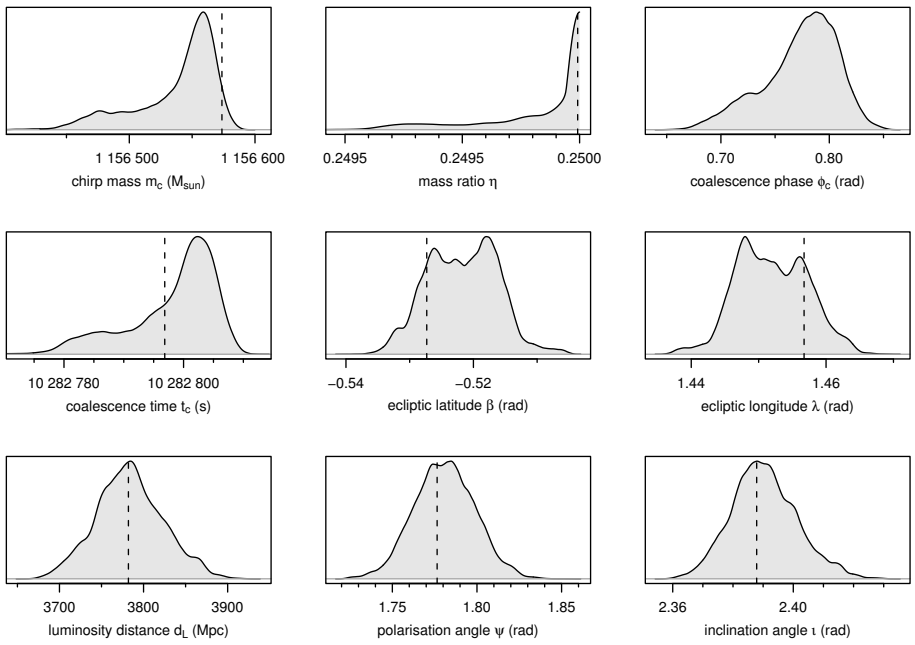

Figure 5.14: Kernel density estimates of the (marginal) posterior densities for each of the 9 parameters. Dashed lines indicate the true parameter values (except for ϕ_0, due to a different parametrisation).

pairs are shown in figure 5.15. Table 5.5 lists some numerical estimates of the parameters together with their true values; all these estimates were derived from all four (tempered) chains of the parallel tempering algorithm using importance sampling as described in section 3.2.15.

Along with the 9 signal parameters, the noise parameters (i.e., the spectrum) were also inferred within the MCMC algorithm. Figure 5.16 shows the spectrum's marginal posterior mean. In figure 5.17, the posterior is shown for a narrow frequency band; the top half of the plot illustrates the individual galactic binary background signals contributing to the background noise within that frequency range, and the bottom half shows how these are reflected in the posterior spectrum.

The parallel tempering algorithm performed well, and the approximate results regarding the parallel tempering setup derived in section 3.2.12 in fact do provide a handle on an efficient setup of the algorithm. Fig-

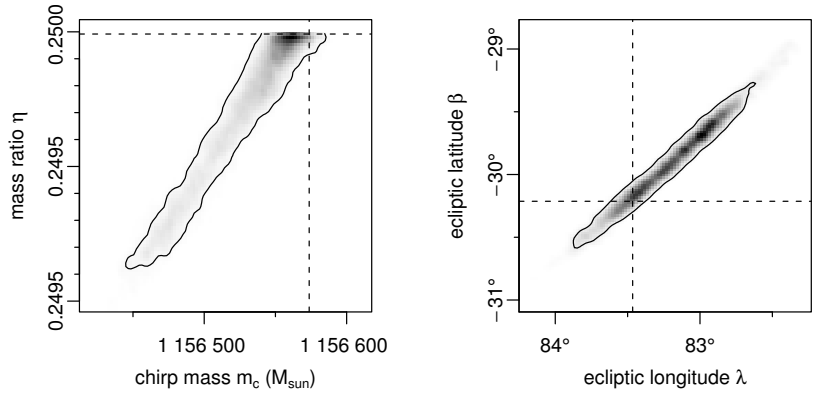

Figure 5.15: Estimates of 2-dimensional (marginal) joint posterior densities for 2 pairs of parameters, and 95% confidence regions. Dashed lines indicate the true parameter values.

ure 5.18 shows the acceptance rates of swaps between neighbouring chains within the parallel tempering scheme. At low temperatures, the algorithm behaves as predicted, and for higher temperatures the acceptance rates increase as expected, since, as the likelihood contribution to the posterior is "melted away", the tempered distributions become increasingly similar to the prior and to each other.

The speed of the code is rather slow. The time it takes for each iteration scales roughly linearly with the amount of data considered, and the majority ($\approx$ 95%) goes into the numerical derivation of the TDI response to given $+/\times$ polarisation waveforms. With 2^{17} samples (corresponding to 23 days of data) considered, it takes $\approx$ 3.3 seconds per iteration running on an AMD Opteron 1000 MHz dual-core processor. Again, the speed also scales roughly linearly with the number of chains and available processors: the code using 96 parallel chains took roughly 40 seconds per iteration on a machine with 8 of the above processors. The MCMC algorithm is working in principle, but it would gain from a faster TDI derivation (numerically, analytically or approximative) or greater computational resources, as this would then also allow for easier tuning and performance evaluation.

Table 5.5: Some key figures of the the marginal posterior distributions of individual parameters, derived by importance sampling from all tempered chains. Mean and standard deviation indicate location and spread, and the 95% central credible interval gives a range that contains the true parameter with 95% probability, given the data at hand. The interval stated for the mass ratio η is a one-sided interval. No true value was given for the phase ϕ_0, due to a differing parametrisation.

	mean	st.dev.	95% c.c.i.	true	unit
chirp mass (m_c)	1 156 536	33	(1 156 459, 1 156 575)	1 156 574	$M_\odot$
mass ratio (η)	0.249 78	0.000 26	(0.249 23, 0.250 00)	0.249 99	
coalescence time (t_c)	10 282 797.9	7.7	(10 282 780.3, 10 282 807.6)	10 282 796.9	s
luminosity distance (d_L)	3783	40	(3707, 3868)	3782	Mpc
inclination (ι)	2.390	0.011	(2.370, 2.415)	2.388	rad
ecliptic latitude (β)	-0.5214[a]	0.0071[a]	$(-0.5318, -0.5109)$	-0.5273	rad
ecliptic longitude (λ)	1.4517[a]		(1.4418, 1.4619)	1.4567	rad
coalescence phase (ϕ_0)	0.773	0.036	(0.693, 0.829)		rad
polarisation (ψ)	1.780	0.019	(1.744, 1.817)	1.776	rad
mass 1 (m_1)	1 298 000	23 000	(1 254 000, 1 328 000)	1 320 976	$M_\odot$
mass 2 (m_2)	1 361 000	25 000	(1 329 000, 1 408 000)	1 336 184	$M_\odot$

[a] *mean direction* and *spherical standard deviation*

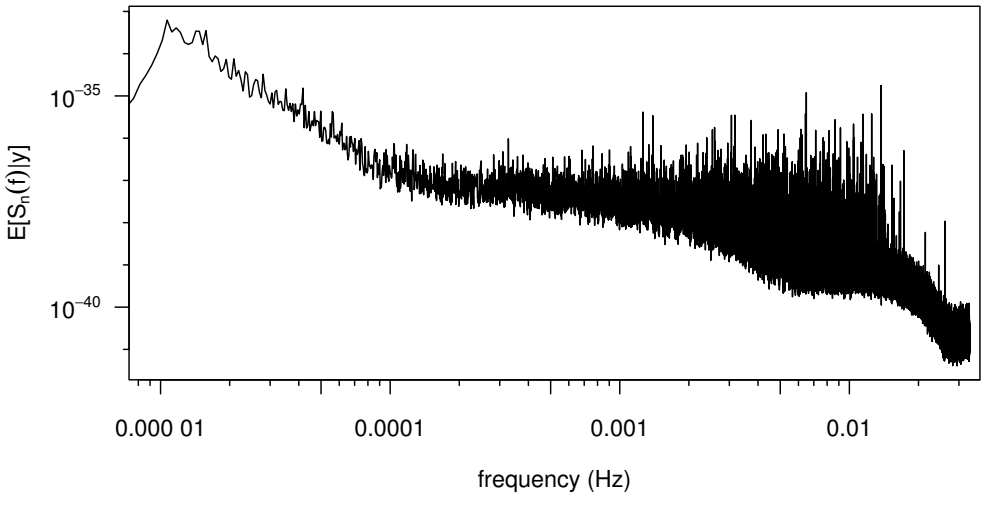

Figure 5.16: The marginal posterior mean spectrum $\mathrm{E}[S_n(f)|y]$.

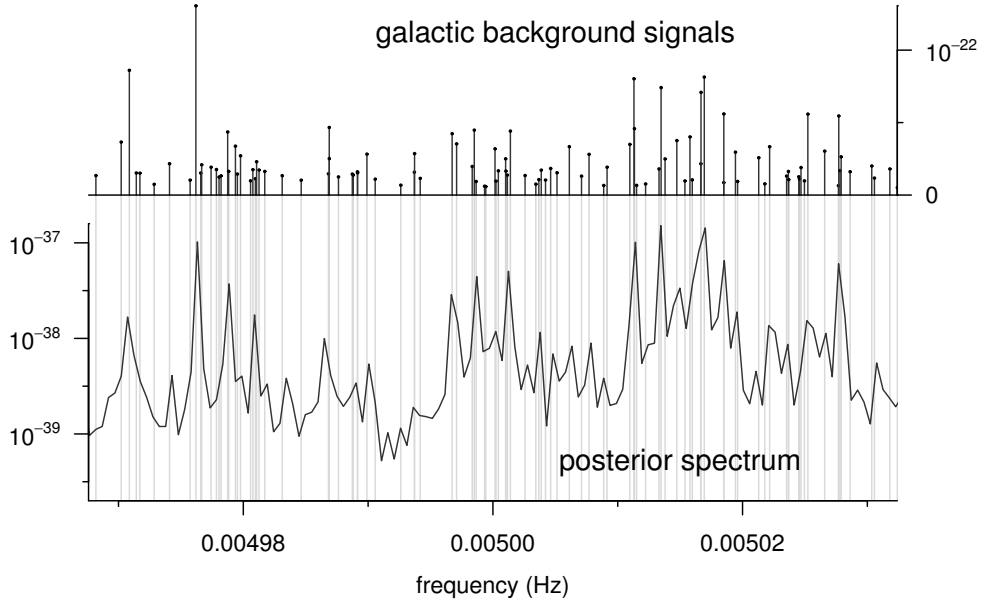

Figure 5.17: (True) frequencies and amplitudes of individual background signals within a narrow frequency band (top panel), and how these are reflected in the posterior spectrum (lower panel).

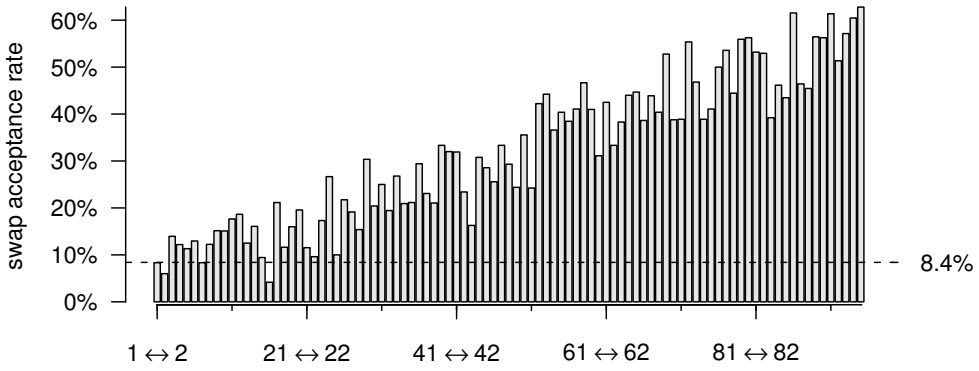

Figure 5.18: Acceptance rates of swaps for all neighbouring chains in a parallel tempering run with 96 chains (started from the true values). For the first few chains, at low temperatures, the acceptance rates are near the predicted level of 8.4%, and after that they are steadily increasing. The overall mean is 32%. The acceptance rates for regular proposals were roughly constant for all chains at 25 ($\pm$3) percent.

Chapter 6

Conclusions

A Bayesian model framework for the analysis of the gravitational-wave signals of binary inspiral events was developed, and applied in two different scenarios of ground-based and space-based laser-interferometric measurements. Computational methods (in particular MCMC methods) that are vital for practical application were adapted and successfully implemented. The standard model (assuming the noise to have a known spectrum) was generalised so that the noise parameters are treated as unknowns as well, and are inferred along with the signal parameters. This model extension also allows to account for a background noise that consists to significant parts of umodeled deterministic signals. In the course of implementing the parallel tempering MCMC algorithm, some insight was gained into its internal functionality, which lead to the development of hints for an efficient setup of such an algorithm.

The implementation developed for ground-based interferometric gravitational-wave measurements is running well and would be applicable as a tool for parameter estimation at the end of a signal-detection pipeline. In the future, this code could be extended to incorporate additional parameters, or to include the ring-down signal emitted by the newly formed black hole after the merger of the inspiralling companions. Figure 6.1 shows first results from the ongoing development of an MCMC algorithm for coherent inference on the 12 parameters that determine the signal of a *spinning*

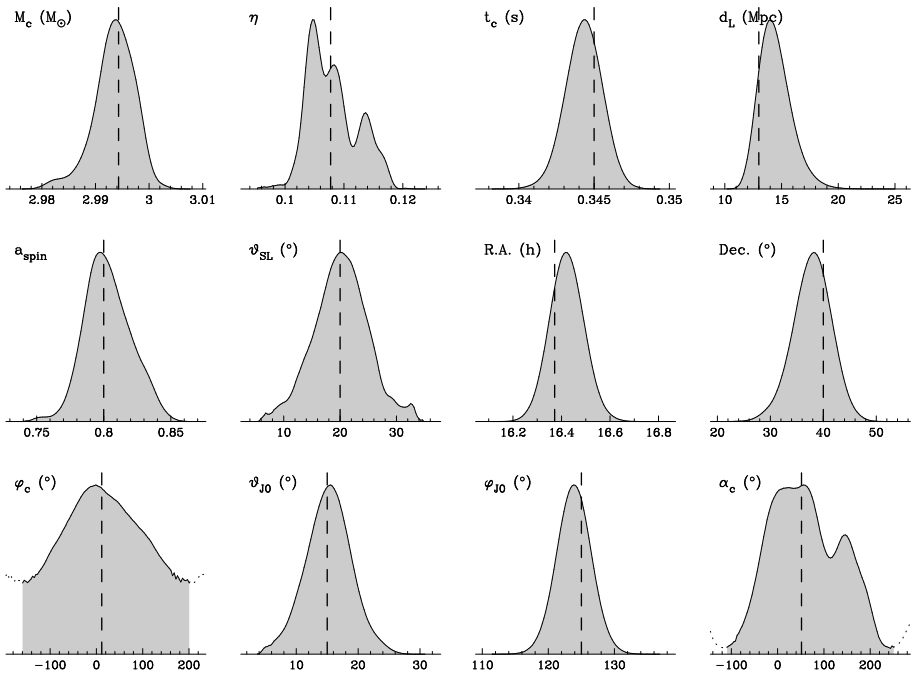

Figure 6.1: First results (marginal density estimates) from an MCMC code that infers the 12 parameters of a *spinning* binary inspiral from measurements of 2 interferometers [118]. The 3 additional parameters are: spin magnitude $a_{\rm spin}$, angle between spin and orbital angular momentum $\vartheta_{\rm SL}$, and precession phase α_c. The paramaters $\vartheta_{\rm J0}$ and $\varphi_{\rm J0}$ correspond to polarisation and inclination. Dashed lines indicate the true parameter values.

binary inspiral, based on measurements from several ground-based interferometers [118].

The implementation for space-based measurements is suffering from the computationally more expensive numerical derivation of the detector response to given gravitational-wave signals. Whilst it is working in principle, it would gain in applicability from a more efficient implementation, more computational resources, or the development of an analytical derivation (or approximation) of TDI responses.

The noise model that was developed for the latter application is a very general and also computationally convenient formulation, and it should be applicable in a wide range of contexts where time series are involved. In particular, it is not limited to signal processing applications, but may

be convenient for Bayesian spectrum estimation in general. A logical step from here might be to investigate its usefulness for Bayesian inference on autocorrelation functions, or to try to constrain the model in order to reduce its number of parameters. The results concerning the interior functional principles of the parallel tempering algorithm are also very generally applicable and should be helpful for setting up efficient implementations.

Appendix A

Appendix

A.1 Properness of tempered distributions

The properness of the resulting tempered distribution when using tempering as defined in (3.11) is shown by demonstrating that the integral is bounded above:

$$\begin{aligned}
\int f_{(T)}(\theta)\,\mathrm{d}\theta &\propto \int \pi(\theta)\,\mathcal{L}(\theta)^{\frac{1}{T}}\,\mathrm{d}\theta && \text{(A.1)}\\
&\leq \int \pi(\theta)\,\max\big(1,\mathcal{L}(\theta)\big)\,\mathrm{d}\theta && \text{(A.2)}\\
&\leq \int \pi(\theta)\,\big(1+\mathcal{L}(\theta)\big)\,\mathrm{d}\theta && \text{(A.3)}\\
&= \underbrace{\int \pi(\theta)\,\mathrm{d}\theta}_{<\infty} + \underbrace{\int \pi(\theta)\,\mathcal{L}(\theta)\,\mathrm{d}\theta}_{<\infty} && \text{(A.4)}
\end{aligned}$$

As long as the prior is proper, the posterior is proper [42], and consequently the tempered posterior is proper as well.

A.2 Parallel tempering setup

R code [55] to determine (approximate) acceptance rates for given temperature T or vice versa as derived in section 3.2.12.

```
expectedprob <- function(d=10, q=2)
{
  integrand <- function(x, mu=0, sigma=1)
  {
    return(1/(sqrt(2*pi) * sigma)
           * exp(-(log(x)-mu)^2/(2*sigma^2)))
  }
  Zmean <- d * (1 - 0.5*(q + 1/q))
  Zsd <- sqrt(d * (1 - (q+1/q) + 0.5*(q^2 + (1/q)^2)))
  E <- 1-pnorm(0,Zmean,Zsd)
  E <- E+integrate(integrand,0,1,mu=Zmean,sigma=Zsd)$value
  return(E)
}

# acceptance rate for 9 parameters and factor 3:
# expectedprob(9,3)

# factor for 9 parameters to yield 20% acceptance:
# uniroot(function(x){expectedprob(9,x)-0.2},c(1.0001,100))
```

A.3 Random variable transformations

General

In the following, the terms necessary for the variable transformations used here are given (see also section 3.3). The rather trivial cases are given in Table A.1, and the more complex formulas for mass reparametrisation are given in the following subsection.

Chirp mass & mass ratio

Transformation from masses $\{(m_1, m_2) : \ m_1, m_2 \in \mathbb{R}^+, \ m_1 \leq m_2\}$ to *chirp mass* and *mass ratio* $\{(m_c, \eta) : \ m_c \in \mathbb{R}^+, \ \eta \in]0, 0.25]\}$ (see also

Table A.1: Terms necessary for random variable transformations (see also section 3.3).

	domains ($x \to y$)	$f(x)$	$f^{-1}(y)$	$J(y)$	
logarithm	$\mathbb{R}^+ \to \mathbb{R}$	$\log(x)$	$\exp(y)$	$\exp(y)$	$(= x)$
sine	$[-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$	$\sin(x)$	$\arcsin(y)$	$\frac{1}{\sqrt{1-y^2}}$	$(= \frac{1}{\cos(x)})$
cosine	$[0, \pi] \to [-1, 1]$	$\cos(x)$	$\arccos(y)$	$-\frac{1}{\sqrt{1-y^2}}$	$(= -\frac{1}{\sin(x)})$
square root	$\mathbb{R}^+ \to \mathbb{R}^+$	$\sqrt{x}$	y^2	$2y$	

section 4.2.2).

$$f\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}\right) = \begin{pmatrix} \frac{(m_1 m_2)^{0.6}}{(m_1+m_2)^{0.2}} \\ \frac{m_1 m_2}{(m_1+m_2)^2} \end{pmatrix} = \begin{pmatrix} m_c \\ \eta \end{pmatrix} \tag{A.5}$$

$$f^{-1}\left(\begin{pmatrix} m_c \\ \eta \end{pmatrix}\right) = \begin{pmatrix} m_c \frac{\left(1+\frac{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}\right)^{0.2}}{\left(\frac{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}\right)^{0.6}} \\ m_c \frac{\left(1+\frac{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}\right)^{0.2}}{\left(\frac{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}\right)^{0.6}} \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \tag{A.6}$$

which, substituting $g(\eta) = \frac{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}$ and $g^*(\eta) = \frac{1}{g(\eta)} = \frac{\frac{1}{2}-\sqrt{\frac{1}{4}-\eta}}{\frac{1}{2}+\sqrt{\frac{1}{4}-\eta}}$, simplifies to:

$$f^{-1}\left(\begin{pmatrix} m_c \\ \eta \end{pmatrix}\right) = \begin{pmatrix} m_c \frac{(1+g(\eta))^{0.2}}{g(\eta)^{0.6}} \\ m_c \frac{(1+g^*(\eta))^{0.2}}{g^*(\eta)^{0.6}} \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \tag{A.7}$$

$$
\begin{aligned}
&\det(J(m_1, m_2)) = \\
&\quad \Bigg[\left(\tfrac{3}{5} m_1^{-\frac{2}{5}} m_2^{\frac{3}{5}} (m_1+m_2)^{-\frac{1}{5}} - \tfrac{1}{5}(m_1+m_2)^{-\frac{6}{5}} (m_1 m_2)^{\frac{3}{5}} \right) \\
&\quad \left(m_1 (m_1+m_2)^{-2} - 2 m_1 m_2 (m_1+m_2)^{-3} \right) \\
&\quad - \left(\tfrac{3}{5} m_1^{\frac{3}{5}} m_2^{-\frac{2}{5}} (m_1+m_2)^{-\frac{1}{5}} - \tfrac{1}{5}(m_1+m_2)^{-\frac{6}{5}} (m_1 m_2)^{\frac{3}{5}} \right) \\
&\quad \left(m_2 (m_1+m_2)^{-2} - 2 m_1 m_2 (m_1+m_2)^{-3} \right) \Bigg]^{-1}
\end{aligned}
\tag{A.8}
$$

Note that the determinant of J is expressed in terms of m_1 and m_2. In order to obtain $\det(J)$ in terms of m_c and η, use $J(f^{-1}(m_c, \eta))$.

A.4 Mean direction and spherical variance

The *mean direction* and the *spherical variance* are descriptive measures of location and spread for data representing locations on a sphere (see section 3.8). The data are represented as (p-dimensional) unit vectors in a sphere S^{p-1} in p-dimensional space $\mathbb{R}^p$. So for $p = 2$ one is dealing with a point on a circle S^1, for $p = 3$ with points on a sphere S^2, and so on. Let $x_1, \ldots, x_N$ be points on S^{p-1}. Then their *sample mean* $\bar{x}$ and their *mean resultant length* $\bar{R}$ are given by

$$
\bar{x} \;=\; \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \text{and} \qquad \bar{R} \;=\; \|\bar{x}\|, \tag{A.9}
$$

where $\bar{x} \in \mathbb{R}^p$ and $0 \leq \bar{R} \leq 1$. The unit vector

$$
\bar{x}_0 \;=\; \frac{1}{\bar{R}}\, \bar{x} \tag{A.10}
$$

then defines the *mean direction*, and

$$
s_0^2 \;=\; 2\,(1 - \bar{R}) \tag{A.11}
$$

defines the *spherical variance* [75].

If the data is *axial* instead of spherical, i.e. every data point maps to

a point on a hemisphere (or semicircle: $[0, \pi]$), then at least in the 2-dimensional case the mean is still defined and can be derived by doubling all the angles (mapping the data on the full circle), determining the mean direction, and again dividing the result by two [75].

A.5 Inverting a covariance matrix given its Cholesky decomposition

Deriving a covariance matrix' inverse from its Cholesky decomposition is useful for determining the proposal distribution's density within a Metropolis-Hastings algorithm when using the 'Randlib' library [113] to generate proposals. For multivariate Normal (or Student-t) proposals, computing the density requires the covariance matrix' inverse, and the Cholesky decomposition is available as a 'side effect' when initialising Randlib.

The *Cholesky decomposition* of a symmetric positive definite $(p \times p)$-matrix A is a (unique) factorisation $A = U^T U$ where U is an upper triangular matrix with positive entries on its main diagonal. Given this factorisation, in order to derive the inverse A^{-1} of A, such that $AA^{-1} = A^{-1}A = I(p)$ (where $I(p)$ is the $(p \times p)$ unity matrix) one can then substitute:

$$AA^{-1} = I(p) \tag{A.12}$$

$$\Leftrightarrow \quad U^T \underbrace{UA^{-1}}_{=:Y} = U^T Y = I(p) \tag{A.13}$$

which can be solved for Y straightforwardly, since U is upper diagonal. Now, given matrices U and Y and their relationship

$$UA^{-1} = Y \tag{A.14}$$

one can then solve for the inverse A^{-1} [119].

More specifically, equation (A.13) leads to the following elements of Y, that can be computed successively column-wise for a fixed i and for j

from 1 up to p:

$$y_{1,i} = \frac{I(p)_{1,i}}{u_{1,1}} \tag{A.15}$$

$$y_{j,i} = \frac{I(p)_{j,i} - \sum_{k=1}^{j-1} u_{k,j} y_{k,i}}{u_{j,j}} \tag{A.16}$$

Following equation (A.14), the elements $\alpha_{\cdot,\cdot}$ of A^{-1} can again be computed column-wise for fixed i, and for j from p down to 1:

$$\alpha_{p,i} = \frac{y_{p,i}}{u_{p,p}} \tag{A.17}$$

$$\alpha_{j,i} = \frac{y_{j,i} - \sum_{k=j+1}^{p} u_{j,k} \alpha_{k,i}}{u_{j,j}}. \tag{A.18}$$

A.6 Some vector operations

A.6.1 Vector products

The *dot product* of two vectors $\vec{x}$ and $\vec{y}$ of equal dimension d is defined by

$$\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \sum_{i=1}^{d} x_i y_i \in \mathbb{R}. \tag{A.19}$$

The dot product also defines a vector's *norm* through $\|\vec{x}\| = \sqrt{\vec{x} \cdot \vec{x}}$.

The *cross product* of two 3-dimensional vectors $\vec{x}$ and $\vec{y}$ is defined by

$$\vec{x} \times \vec{y} = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix} \in \mathbb{R}^3. \tag{A.20}$$

The *scalar triple product* of three 3-dimensional vectors $\vec{x}$, $\vec{y}$ and $\vec{z}$ is defined as $\vec{x} \cdot (\vec{y} \times \vec{z}) \in \mathbb{R}$. It is also equal to the determinant of the 3×3 ma-

trix having the three vectors as rows.

A triplet $(\vec{x}, \vec{y}, \vec{z})$ of vectors is referred to as *right-handed* if and only if its triple product is positive:

$$\text{RH}(\vec{x}, \vec{y}, \vec{z}) \quad \Leftrightarrow \quad \vec{x} \cdot (\vec{y} \times \vec{z}) \geq 0. \tag{A.21}$$

A.6.2 Angles between vectors

The angle between two vectors $\vec{x}$ and $\vec{y}$ is given by:

$$\text{ANGLE}(\vec{x}, \vec{y}) \;=\; \arccos\left(\frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \; \|\vec{y}\|}\right). \tag{A.22}$$

A.6.3 Orthogonal projection

The orthogonal projection of a vector $\vec{a}$ into the plane spanned by the two orthonormal vectors $\vec{x}$ and $\vec{y}$ is given by:

$$\text{OP}(\vec{a}, \vec{x}, \vec{y}) \;=\; (\vec{a} \cdot \vec{x})\vec{x} + (\vec{a} \cdot \vec{y})\vec{y}. \tag{A.23}$$

A.6.4 Vector rotations

A vector $\vec{x}$ is rotated around an 'axis' unit vector $\vec{n} = (n_1, n_2, n_3)^T$ by an angle of α by multiplying it with the rotation matrix

$$R^{\alpha}_{\vec{n}} = \begin{pmatrix} c + n_1^2(1-c) & n_1 n_2(1-c) + n_3 s & n_1 n_3(1-c) - n_2 s \\ n_2 n_1(1-c) - n_3 s & c + n_2^2(1-c) & n_2 n_3(1-c) + n_1 s \\ n_3 n_1(1-c) + n_2 s & n_3 n_2(1-c) - n_1 s & c + n_3^2(1-c) \end{pmatrix} \tag{A.24}$$

where $s = \sin\alpha$ and $c = \cos\alpha$. The rotation $\vec{y} = R^{\alpha}_{\vec{n}}\vec{x}$ is then clockwise when looking along $\vec{n}$ while it is pointing towards the observer [120].

A.7 The restricted PN approximation

The restricted PN approximation gives the inspiral gravitational wave signal in the time domain. The formulas given here are taken from [91, 92]. The general idea was sketched in section 4.3.2, but expressions for the instantaneous phase Φ and frequency ω were omitted. These depend on the time t (in seconds) via the dimensionless time variable

$$\tau \;=\; \frac{\eta}{5\,m_t}\,(t_c - t). \tag{A.25}$$

The instantaneous frequency ω (multiplied by the total mass) then is given by

$$\begin{aligned} m_t\omega \;&=\; \frac{1}{8}\tau^{-3/8}\left(1 + \left(\frac{11}{32}\eta + \frac{743}{2688}\right)\tau^{-1/4} - \frac{3}{10}\pi\tau^{-3/8}\right. \\ &\quad \left. + \left(\frac{1855\,099}{14\,450\,688} + \frac{56\,975}{258\,048}\eta + \frac{371}{2048}\eta^2\right)\tau^{-1/2}\right) \end{aligned} \tag{A.26}$$

and the instantaneous phase is given by

$$\begin{aligned} \Phi \;&=\; -\frac{1}{32\eta}(m_t\omega)^{-5/3}\left(1 + \left(\frac{3715}{1008} + \frac{55}{12}\eta\right)(m_t\omega)^{2/3} - 10\pi(m_t\omega)\right. \\ &\quad \left. + \left(\frac{15\,293\,365}{1016\,064} + \frac{27\,145}{1008}\eta + \frac{3085}{144}\eta^2\right)(m_t\omega)^{4/3}\right). \end{aligned} \tag{A.27}$$

A.8 The 2.5 PN (and 2.0 PN) stationary-phase waveform approximation

This approximation was introduced in section 4.3.4 and models the inspiral signal in the frequency domain. The Fourier-transformed signal $\tilde{s}_\theta(f)$ is a function of the frequency f and depends on the ('local') parameters as described in section 4.2. Depending on the orientation of binary and interferometer with respect to each other, it is a linear combination of *cosine chirp* $\tilde{h}_c(f,\theta)$ and *sine chirp* $\tilde{h}_s(f,\theta)$. The (Fourier-transformed) cosine

chirp is defined as:

$$\tilde{h}_c(f,\theta) = \frac{\sqrt{\eta}\, m_t^{\frac{5}{6}}}{d_L} \frac{\sqrt{5}\, \mathrm{G}^{\frac{5}{6}}}{2\sqrt{6}\, \pi^{\frac{2}{3}}\, \mathrm{c}^{\frac{3}{2}}} f^{-\frac{7}{6}} \exp\Big(-\mathrm{i}\big(\underbrace{\psi(f) + \phi_0 + 2\pi f t_c^{(I)}}_{\text{phase evolution}}\big)\Big) \tag{A.28}$$

where

$$\psi(f) \;=\; \sum_{i=1}^{5} a_i \zeta_i(f), \tag{A.29}$$

$$a_1 \;=\; \frac{3}{128\eta} q^{-\frac{5}{3}}, \tag{A.30}$$

$$a_2 \;=\; \frac{1}{384\eta}\left(\frac{3715}{84} + 55\eta\right) q^{-1}, \tag{A.31}$$

$$a_3 \;=\; -\frac{1}{128\eta} 48\pi q^{-\frac{2}{3}}, \tag{A.32}$$

$$a_4 \;=\; \frac{3}{128\eta}\left(\frac{15\,293\,365}{508\,032} + \frac{27\,145}{504}\eta + \frac{3085}{72}\eta^2\right) q^{-\frac{1}{3}}, \tag{A.33}$$

$$a_5 \;=\; \frac{\pi}{128\eta}\left(\frac{38\,645}{252} + 5\eta\right), \tag{A.34}$$

$\zeta_1(f) = f^{-\frac{5}{3}}$, $\zeta_2(f) = f^{-1}$, $\zeta_3(f) = f^{-\frac{2}{3}}$, $\zeta_4(f) = f^{-\frac{1}{3}}$, $\zeta_5(f) = \log(f)$, and $q = \pi \mathrm{G} m_t \mathrm{c}^{-3}$.

The sine chirp as the orthogonal waveform to the cosine chirp is:

$$\tilde{h}_s(f,\theta) \;=\; \mathrm{i}\, \tilde{h}_c(f,\theta). \tag{A.35}$$

From sine and cosine chirp the *plus*- and *cross*-waveforms are derived:

$$\tilde{h}_+(f,\theta) \;=\; -\tfrac{1}{2}(1+\cos^2(\iota))\, \tilde{h}_c(f,\theta) \tag{A.36}$$

$$\tilde{h}_\times(f,\theta) \;=\; -\cos(\iota)\, \tilde{h}_s(f,\theta) \tag{A.37}$$

And finally the actual chirp signal $\tilde{h}(f,\theta)$ measured at the detector depends on the *antennae pattern functions* F_+ and $F_\times$ that were defined in

equations (4.12) and (4.13) in section 4.4:

$$\tilde{h}(f,\theta) \;=\; F_+\tilde{h}_+(f,\theta) \;+\; F_\times\tilde{h}_\times(f,\theta) \tag{A.38}$$

[10, 93].

The 2.0 PN approximation can be derived from this 2.5 PN formula by simply leaving out the factors a_5 and ζ_5, i.e., by only summing from $i = 1$ to 4 in equation (A.29).

A.9 The 3.5 PN / 2.5 PN waveform parametrisation

In [79], the (2.5 PN) chirp waveform's phase is parameterised in terms of the coalescence phase ϕ_0, which is convenient for our purposes. In the more recent publications [86, 87], the (3.5 PN) phase is (equivalently) expressed in terms of 'a constant phase τ_0' instead. The following shows how these two expressions relate to each other, and how to re-express the 3.5 PN phase in terms of ϕ_0 instead of τ_0.

In [87], equation (13), the 3.5 PN instantaneous phase Φ is defined as:

$$\begin{aligned}
\Phi &= -\frac{1}{\eta}\left(f_1(\tau,\eta) + \left(-\frac{38\,645}{172\,032} + \frac{65}{2048}\eta\right)\pi\log\left(\frac{\tau}{\tau_0}\right) + f_2(\tau,\eta)\right) \\
&= -\frac{1}{\eta}\left(-\frac{38\,645}{172\,032} + \frac{65}{2048}\eta\right)\pi\log\left(\frac{\tau}{\tau_0}\right) \;-\; \frac{1}{\eta}\left(f_1(\tau,\eta) + f_2(\tau,\eta)\right) \\
&= -\frac{1}{\eta}\left(-\frac{38\,645}{172\,032} + \frac{65}{2048}\eta\right)\pi\left(\log(\tau) - \log(\tau_0)\right) \\
&\quad -\frac{1}{\eta}\left(f_1(\tau,\eta) + f_2(\tau,\eta)\right) \\
&= -\frac{1}{\eta}\left(-\frac{38\,645}{172\,032} + \frac{65}{2048}\eta\right)\pi\log(\tau) \\
&\quad +\frac{1}{\eta}\left(-\frac{38\,645}{172\,032} + \frac{65}{2048}\eta\right)\pi\log(\tau_0) - \frac{1}{\eta}\left(f_1(\tau,\eta) + f_2(\tau,\eta)\right)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}-\frac{15}{2048}\eta+\frac{80}{2048}\eta\right)\pi\log(\tau)\\
&\quad+\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}+\frac{65}{2048}\eta\right)\pi\log(\tau_0)-\frac{1}{\eta}\Big(f_1(\tau,\eta)+f_2(\tau,\eta)\Big)\\
&= -\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}-\frac{15}{2048}\eta\right)\pi\log(\tau)-\frac{1}{\eta}\left(\frac{80}{2048}\eta\right)\pi\log(\tau)\\
&\quad+\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}+\frac{65}{2048}\eta\right)\pi\log(\tau_0)-\frac{1}{\eta}\Big(f_1(\tau,\eta)+f_2(\tau,\eta)\Big)\\
&= \underbrace{\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}+\frac{65}{2048}\eta\right)\pi\log(\tau_0)}_{\text{`}\phi_0\text{ substitute' term}}\\
&\quad-\frac{1}{\eta}\Bigg(\underbrace{f_1(\tau,\eta)+\left(-\frac{38\,645}{172\,032}-\frac{15}{2048}\eta\right)\pi\log(\tau)}_{\text{2.5 PN terms as in [79]}}\\
&\qquad+\underbrace{\left(\frac{80}{2048}\eta\right)\pi\log(\tau)+f_2(\tau,\eta)}_{\text{extra 3.5 PN terms}}\Bigg)
\end{aligned}
\tag{A.39}
$$

(cp. equation (6.13) in [79]). So, by defining

$$
\phi_0 = \frac{1}{\eta}\left(-\frac{38\,645}{172\,032}+\frac{65}{2048}\eta\right)\pi\log(\tau_0) \tag{A.40}
$$

$$
\Leftrightarrow \quad \tau_0 = \exp\left(\frac{\phi_0}{\frac{1}{\eta}\left(-\frac{38\,645}{172\,032}+\frac{65}{2048}\eta\right)\pi}\right), \tag{A.41}
$$

the instantaneous phase Φ can be expressed in terms of ϕ_0 instead of τ_0 by substituting the corresponding term. In this way one can use the same parametrisation (and prior specification) as for the 2.5 PN model.

A.10 TDI variables

The TDI variables (see section 4.4.2) are defined in terms of the observables X, Y and Z as [100]:

$$A = \frac{1}{\sqrt{2}}(Z - X) \tag{A.42}$$

$$\wedge \quad E = \frac{1}{\sqrt{6}}(X - 2Y + Z) \tag{A.43}$$

$$\wedge \quad T = \frac{1}{\sqrt{3}}(X + Y + Z), \tag{A.44}$$

and the back-transformation is then given by:

$$\Leftrightarrow \quad X = -\frac{\sqrt{2}}{2}A + \frac{\sqrt{6}}{6}E + \frac{\sqrt{3}}{3}T \tag{A.45}$$

$$\wedge \quad Y = -2\frac{\sqrt{6}}{6}E + \frac{\sqrt{3}}{3}T \tag{A.46}$$

$$\wedge \quad Z = +\frac{\sqrt{2}}{2}A + \frac{\sqrt{6}}{6}E + \frac{\sqrt{3}}{3}T. \tag{A.47}$$

A.11 The 'unknown spectrum' noise model

A.11.1 Noise model and DFT

Here the exact relationship between the noise model that includes the noise spectrum as an unknown (as described in section 4.5.3) and the output of a discrete Fourier transform (as described in section 3.4.2) is derived. Let

$$\alpha_j = \mathrm{Re}\big(\tilde{h}(f_j)\big) \quad \text{and} \quad \beta_j = \mathrm{Im}\big(\tilde{h}(f_j)\big), \tag{A.48}$$

i.e.: $\tilde{h}(f_j) = \alpha_j + \beta_j \mathrm{i}$. The inverse DFT was defined as (3.64):

$$
\begin{aligned}
h(t) &= \Delta_f \sum_{j=0}^{N-1} \tilde{h}(f_j) \exp(2\pi \mathrm{i} f_j t) && \text{(A.49)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \left[\tilde{h}(f_j) \exp(2\pi \mathrm{i} f_j t) + \overline{\tilde{h}(f_j)} \exp(2\pi \mathrm{i} f_{N-j} t) \right] \\
&\quad + \Delta_f \tilde{h}(f_0) \exp(2\pi \mathrm{i} f_0 t) \; + \; \Delta_f \tilde{h}(f_{N/2}) \exp(2\pi \mathrm{i} f_{N/2} t) && \text{(A.50)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \Big[(\alpha_j + \beta_j \mathrm{i}) \big(\cos(-2\pi f_j t) - \sin(-2\pi f_j t) \mathrm{i} \big) \\
&\quad + (\alpha_j - \beta_j \mathrm{i}) \big(\cos(-2\pi \, f_{N-j} t) - \sin(-2\pi f_{N-j} t) \mathrm{i} \big) \Big] \\
&\quad + \Delta_f \alpha_0 \cos(-2\pi \mathrm{i} f_0 t) \; + \; \Delta_f \alpha_{N/2} \cos(-2\pi \mathrm{i} f_{N/2} t) && \text{(A.51)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \Big[(\alpha_j + \beta_j \mathrm{i}) \big(\cos(-2\pi f_j t) - \sin(-2\pi f_j t) \mathrm{i} \big) \\
&\quad + (\alpha_j - \beta_j \mathrm{i}) \big(\cos(-2\pi f_j t) + \sin(-2\pi f_j t) \mathrm{i} \big) \Big] \\
&\quad + \Delta_f \alpha_0 \; + \; \Delta_f \alpha_{N/2} \cos(-2\pi \mathrm{i} f_{N/2} t) && \text{(A.52)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \Big[\big(\alpha_j \cos(\ldots) + \beta_j \sin(\ldots) \big) + \big(-\alpha_j \sin(\ldots) + \beta_j \cos(\ldots) \big) \mathrm{i} \\
&\quad + \big(\alpha_j \cos(\ldots) + \beta_j \sin(\ldots) \big) + \big(\alpha_j \sin(\ldots) - \beta_j \cos(\ldots) \big) \mathrm{i} \Big] \\
&\quad + \Delta_f \alpha_0 \; + \; \Delta_f \alpha_{N/2} \cos(-2\pi \mathrm{i} f_{N/2} t) && \text{(A.53)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \Big[\big(2\alpha_j \cos(-2\pi f_j t) + 2\beta_j \sin(-2\pi f_j t) \big) \Big] \\
&\quad + \Delta_f \alpha_0 \; + \; \Delta_f \alpha_{N/2} \cos(-2\pi \mathrm{i} f_{N/2} t) && \text{(A.54)}\\
&= \Delta_f \sum_{j=1}^{\frac{N}{2}-1} \Big[\big(2\alpha_j \cos(2\pi f_j t) + 2(-\beta_j) \sin(2\pi f_j t) \big) \Big] \\
&\quad + \Delta_f \alpha_0 \; + \; \Delta_f \alpha_{N/2} \cos(2\pi \mathrm{i} f_{N/2} t) && \text{(A.55)}
\end{aligned}
$$

where $t \in \{0, \Delta_t, 2\Delta_t, \ldots, (N-1)\Delta_t\}$, and $f_j = j\Delta_f = \frac{j}{N\Delta_t}$ are the Fourier frequencies. So, comparing to (4.16), one can see that the realisations of $a_0, \ldots, a_{N/2}$ and $b_0, \ldots, b_{N/2}$ are derived from a given noise vector (in the time domain) by Fourier-transforming and then setting

$$a_j \;=\; 2\Delta_f \alpha_j \quad \text{and} \quad b_j \;=\; -2\Delta_f \beta_j \quad \text{for } j = 1, \ldots, \tfrac{N}{2} - 1, \tag{A.56}$$

which especially implies that

$$a_j^2 + b_j^2 \;=\; 4\Delta_f^2 \left(\alpha_j^2 + \beta_j^2\right) \;=\; 4\Delta_f^2 \left|\tilde{h}(f_j)\right|^2 \quad \text{for } j = 1, \ldots, \tfrac{N}{2} - 1, \tag{A.57}$$

and analogously, but without the factor of 2, for $j = 0$ and $j = \frac{N}{2}$. Note that when working with unnormalised Fourier transforms, the factor of $2\Delta_f$ in (A.56) and (A.57) changes to $\frac{2}{N}$.

A.11.2 Likelihood

In the following, the likelihood function for the noise model that includes the noise spectrum as an unknown (as described in section 4.5.3) that was stated in section 4.6.2 is explicitly derived, up to a normalising constant:

$$\begin{aligned}
\mathcal{L}(\theta) \quad &\propto \quad \prod_{i=i_L}^{i_U} \left[\frac{1}{\sigma_i} \exp\left(-\frac{a_i^2}{2\sigma_i^2}\right) \; \frac{1}{\sigma_i} \exp\left(-\frac{b_i^2}{2\sigma_i^2}\right) \right] && \text{(A.58)} \\
&= \quad \exp\left(\sum_{i=i_L}^{i_U} \left[-\frac{a_i^2 + b_i^2}{2\sigma_i^2} - \log\left(\sigma_i^2\right) \right] \right) \\
&\overset{(4.18)}{=} \quad \exp\left(\sum_{i=i_L}^{i_U} \left[-\frac{a_i^2 + b_i^2}{2N\Delta_f^2 S_n(f_i)} - \log\left(N\Delta_f^2 S_n(f_i)\right) \right] \right) \\
&\overset{(A.57)}{=} \quad \exp\left(\sum_{i=i_L}^{i_U} \left[-\frac{4\Delta_f^2 \left|\tilde{h}(f_i)\right|^2}{2N\Delta_f^2 S_n(f_i)} - \log\left(N\Delta_f^2 S_n(f_i)\right) \right] \right) \\
&\propto \quad \exp\left(\sum_{i=i_L}^{i_U} \left[-\frac{2}{N} \frac{\left|\tilde{h}(f_i)\right|^2}{S_n(f_i)} - \log\left(S_n(f_i)\right) \right] \right) && \text{(A.59)}
\end{aligned}$$

Note that if the Fourier transform $\tilde{h}$ is not normalised, the factor of $\frac{2}{N}$ in front of the sum-of-squares term changes to $\frac{2\Delta_t^2}{N}$.

A.12 Properties of the Inv-$\chi^2(\nu, s^2)$ distribution

The scaled inverse-chi-square (Inv-$\chi^2(\nu, s^2)$) distribution is the conjugate prior distribution for the variance parameter(s) in section 4.5.3. In order to be able to judge the implications of parameter choices when defining a prior, some properties of the distribution are derived here. The density function for σ^2 is defined as:

$$f_{\nu,s}(\sigma^2) \;=\; \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\, s^{\nu}\, \left(\sigma^2\right)^{-(\nu/2+1)} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) \tag{A.60}$$

for $\nu \in \mathbb{R}^+$ and $s \in \mathbb{R}^+$ [42]. The density function for $\sigma = \sqrt{\sigma^2}$ can be derived by reparametrisation as described in section 3.3:

$$f_{\nu,s}(\sigma) \;=\; \frac{2(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\, s^{\nu}\, \sigma^{-(\nu+1)} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right), \tag{A.61}$$

where s now is a *scale parameter*, in the sense that $f_{\nu,s}(\sigma) = \frac{1}{s} f_{\nu,1}(\frac{\sigma}{s})$. So setting the prior distribution's parameter s defines the a priori 'order of magnitude' of σ (corresponding to the square root of the spectrum). The degrees-of-freedom parameter ν on the other hand defines the a priori certainty about σ, where greater values of ν indicate greater certainty. The variance of σ is only finite for $\nu > 2$. Table A.2 shows some quantiles of σ's distribution for varying values of ν. Due to the relationship to the χ^2-distribution [42], the Inv-$\chi^2(\nu, s^2)$ distribution's α-quantile is $s\sqrt{\nu/\chi^2_{\nu;\alpha}}$ where $\chi^2_{\nu;\alpha}$ is the α-quantile of the χ^2_ν-distribution. Note that due to the expression for the resulting posterior (equation (4.21)) one can also view the settings of ν and s as "providing the information equivalent to ν observations with average squared deviation s^2" [42]. Note that ν does not need to be an integer, and that setting ν to zero yields the standard non-informative (and improper) prior $f(\sigma^2) = \frac{1}{\sigma^2}$ for σ^2, independent from s^2.

Table A.2: Some quantiles of σ's distribution assuming a scaled inverse Chi-square distribution for σ^2 and varying the degrees-of-freedom parameter ν.

ν	1%	5%	25%	50%	75%	95%	99%
1	0.39s	0.51s	0.87s	1.48s	3.14s	15.9s	79.8s
2	0.47s	0.58s	0.85s	1.20s	1.86s	4.42s	9.97s
3	0.51s	0.62s	0.85s	1.13s	1.57s	2.92s	5.11s
4	0.55s	0.65s	0.86s	1.09s	1.44s	2.37s	3.67s
5	0.58s	0.67s	0.87s	1.07s	1.37s	2.09s	3.00s
6	0.60s	0.69s	0.87s	1.06s	1.32s	1.92s	2.62s
8	0.63s	0.72s	0.88s	1.04s	1.26s	1.71s	2.20s
10	0.66s	0.74s	0.89s	1.03s	1.22s	1.59s	1.98s
20	0.73s	0.80s	0.92s	1.02s	1.14s	1.36s	1.56s
50	0.81s	0.86s	0.94s	1.01s	1.08s	1.20s	1.30s
100	0.86s	0.90s	0.96s	1.00s	1.05s	1.13s	1.19s

When implementing a parallel tempering MCMC sampler for a model where some of the parameters can be drawn in a Gibbs step from an Inv-χ^2 (conditional) distribution, then one needs to be able to sample from the 'tempered' conditional distribution. If the tempering is applied in the general form, to the complete posterior distribution (as in (3.8)), then the tempered Inv-$\chi^2(\nu, s^2)$-distribution again is an Inv-χ^2-distribution:

$$\text{Inv-}\chi^2\left(\frac{\nu+2}{T}-2,\ \frac{\nu s^2}{2+\nu-2T}\right). \tag{A.62}$$

If the temperring is only applied to the likelihood part of the posterior (as in (3.11)), the resulting tempered distribution also again is an Inv-χ^2-distribution. If the prior was defined as Inv-$\chi^2(\nu_0, s_0^2)$, with prior degrees of freedom ν_0 and prior scale s_0^2, then the regular (un-tempered) posterior is an

$$\text{Inv-}\chi^2\left(\nu_0+n,\ \frac{\nu_0 s_0^2+nv}{\nu_0+n}\right) \tag{A.63}$$

distribution, where n is the sample size and v is the observed mean squared

deviation [42]. The tempered version of this is simply

$$\text{Inv-}\chi^2\left(\nu_0 + \tfrac{n}{T},\ \frac{\nu_0 s_0^2 + \frac{n}{T} v}{\nu_0 + \frac{n}{T}}\right), \tag{A.64}$$

so implicitly the 'weight' of the observed data in the posterior is down-weighed by a factor of $\frac{1}{T}$ from n to $\frac{n}{T}$.

A.13 Declination- / inclination prior

Density function f, (cumulative) distribution function F, and quantile function F^{-1} of the declination ($\delta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$):

$$f(\delta) = \begin{cases} \frac{1}{2}\cos(\delta) & \text{if } -\frac{\pi}{2} \leq \delta \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases} \tag{A.65}$$

$$F(\delta) = \begin{cases} 0 & \text{if } \delta < -\frac{\pi}{2} \\ \frac{1}{2}(\sin(\delta) + 1) & \text{if } -\frac{\pi}{2} \leq \delta \leq \frac{\pi}{2} \\ 1 & \text{if } \delta > \frac{\pi}{2} \end{cases} \tag{A.66}$$

$$F^{-1}(p) = \arcsin(2p - 1) \quad \text{for } 0 \leq p \leq 1 \tag{A.67}$$

The prior for the inclination angle ($\iota \in [0, \pi]$) is defined analogously.

A.14 Luminosity distance (d_L) prior

When restricting the 'occurrence' probability as defined in section 4.7.2 (particularly equation (4.35)) to a finite range $[\alpha, \beta]$ (where $0 \leq \alpha < \beta$), then the resulting distribution is proper, and its density, distribution func-

tion and quantile function are given by:

$$f(y) = \begin{cases} \frac{3y^2}{\beta^3-\alpha^3} & \text{if } \alpha \le y \le \beta \\ 0 & \text{otherwise} \end{cases} \tag{A.68}$$

$$\begin{aligned} F(y) &= \int_{-\infty}^{y} f(z)\,\mathrm{d}z \\ &= \begin{cases} 0 & \text{if} \quad y < \alpha \\ \frac{y^3-\alpha^3}{\beta^3-\alpha^3} & \text{if} \quad \alpha \le y \le \beta \\ 1 & \text{if} \quad y > \beta \end{cases} \end{aligned} \tag{A.69}$$

$$F^{-1}(p) = \sqrt[3]{p(\beta^3-\alpha^3)+\alpha^3} \qquad \text{for } 0 \le p \le 1. \tag{A.70}$$

A.15 Properness of the distance prior

In the following it is shown that the introduction of the *detection probability* into the prior (section 4.7.3) actually fixes the improperness of the distance prior as in section 4.7.2. As long as the prior for the all other parameters is proper, it is sufficient to show that the conditional prior of d_L (conditional on the remaining parameters) is proper, where

$$p(d_L \mid m_1, m_2, \iota, \phi_0, t_c, \delta, \alpha, \psi) = p(d_L \mid m_1, m_2, \iota), \tag{A.71}$$

i.e., the conditional prior depends only on masses and inclination angle. Since in this context normalising constant factors to the prior are not of interest, it only needs to be shown that the (conditional) density's integral is finite. The luminosity distance's domain is $\mathbb{R}^+$. The integral up to one $\int_0^1 p(d_L|m_1, m_2, \iota)\mathrm{d}d_L$ is obviously finite (see also previous section). For the remaining improper integral (from 1 to ∞) one can can find an upper

bound for the density, which then has a finite integral:

$$
\begin{aligned}
p(d_L|m_1,m_2,\iota) &\propto d_L^2 \times D_{a,b}(\mathcal{A}(m_1,m_2,d_L,\iota)) && \text{(A.72)}\\
&= \frac{d_L^2}{1+\exp\left(\frac{(c-\log(d_L))-a}{b}\right)} && \text{(A.73)}\\
&= \frac{d_L^2}{1+\exp\left(\frac{-\log(d_L)}{b}\right)\exp\left(\frac{c-a}{b}\right)} && \text{(A.74)}\\
&< \frac{d_L^2}{\exp\left(\frac{c-a}{b}\right)\, d_L^{-\frac{1}{b}}} && \text{(A.75)}\\
&= \frac{1}{\exp\left(\frac{c-a}{b}\right)}\, d_L^{2+\frac{1}{b}} && \text{(A.76)}
\end{aligned}
$$

where

$$
\begin{aligned}
c &= \mathcal{A}(m_1,m_2,d_L,\iota) + \log(d_L)\\
&= \tfrac{1}{2}\big(\log(m_1)+\log(m_2)\big) - \tfrac{1}{6}\log(m_1+m_2)\\
&\quad + \tfrac{1}{2}\log\big(1+6\cos(\iota)^2+\cos(\iota)^4\big)\\
&\quad \in \mathbb{R}. && \text{(A.77)}
\end{aligned}
$$

The integral $\int_1^\infty d_L^{2+\frac{1}{b}}\mathrm{d}d_L$ then is finite for

$$
\begin{aligned}
& 2+\frac{1}{b} < -1 && \text{(A.78)}\\
\Leftrightarrow \quad & b > -\frac{1}{3} && \text{(A.79)}\\
\Leftrightarrow \quad & \frac{\overbrace{x_U - x_L}^{>0}}{\underbrace{2\log\left(\frac{p}{1-p}\right)}_{<0}} < \frac{1}{3} && \text{(A.80)}\\
\Leftrightarrow \quad & \underbrace{x_U - x_L}_{>0} > \underbrace{\frac{2}{3}\log\left(\frac{p}{1-p}\right)}_{<0} && \text{(A.81)}
\end{aligned}
$$

(b was defined in (4.42)). So the prior is proper as long as $x_U > x_L$ and $p < \frac{1}{2}$, which they are by definition.

Index

Bibliography

[1] E. D. Feigelson. Statistics in astronomy. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of statistical sciences*. Wiley & Sons, New York, 1988.

[2] T. J. Loredo. From Laplace to supernova SN 1987A: Bayesian inference in astrophysics. In Fougère P. F., editor, *Maximum entropy and Bayesian methods*, pages 81–142. Kluver Academic Publishers, Dordrecht, The Netherlands, 1990.

[3] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge University Press, Cambridge, 2003.

[4] E. T. Jaynes. Bayesian methods: General background. In J. H. Justice, editor, *Maximum-Entropy and Bayesian methods in applied statistics*. Cambridge University Press, Cambridge, 1986.

[5] T. J. Loredo. The promise of Bayesian inference for astrophysics. In E. D. Feigelson and G. J. Babu, editors, *Statistical challenges in modern astronomy*, chapter 12, pages 275–297. Springer-Verlag, New York, 1992.

[6] P. C. Gregory. A Bayesian revolution in spectral analysis. In A. Mohammad-Djafari, editor, *Bayesian inference and Maximum Entropy methods in science and engineering*, volume 568 of *AIP Conference Proceedings*, pages 557–568. American Institute of Physics Proceedings, May 2001.

[7] L. S. Finn. Issues in gravitational wave data analysis. *Arxiv preprint gr-qc/9709077*, September 1997.

[8] R. Umstätter. *Bayesian strategies for gravitational radiation data analysis*. PhD thesis, The University of Auckland, 2006. URL `http://hdl.handle.net/2292/377`.

[9] N. Christensen and R. Meyer. Markov chain Monte Carlo methods for Bayesian gravitational radiation data analysis. *Physical Review D*, 58(8), October 1998.

[10] N. Christensen and R. Meyer. Using Markov chain Monte Carlo methods for estimating parameters with gravitational radiation data. *Physical Review D*, 64(2):022001, July 2001.

[11] 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(1):317–330, January 2004.

[12] 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.

[13] C. Röver, R. Meyer, and N. Christensen. Coherent Bayesian inference on compact binary inspirals using a network of interferometric gravitational wave detectors. *Physical Review D*, 75(6):062004, March 2007.

[14] C. Röver, R. Meyer, G. M. Guidi, A. Viceré, and N. Christensen. Coherent Bayesian analysis of inspiral signals. *Classical and Quantum Gravity*, 24(19):S607–S615, October 2007.

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

[16] C. Röver, A. Stroeer, E. Bloomer, N. Christensen, J. Clark, M. Hendry, C. Messenger, R. Meyer, M. Pitkin, J. Toher, R. Umstätter, A. Vecchio, J. Veitch, and G. Woan. Inference on inspiral signals using LISA MLDC data. *Classical and Quantum Gravity*, 24(19):S521–S527, October 2007.

[17] A. Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. *Sitzungsberichte der Preußischen Akademie der Wissenschaften*, pages 688–696, June 1916.

[18] J. H. Taylor and J. M. Weisberg. Further experimental tests of relativistic gravity using the binary pulsar PSR1913+16. *The Astrophysical Journal*, 345:434–450, October 1989.

[19] B. F. Schutz. Gravitational wave astronomy. *Classical and Quantum Gravity*, 16(12A):A131–A156, December 1999.

[20] C. Cutler and K. S. Thorne. An overview of gravitational-wave sources. *Arxiv preprint gr-qc/0204090*, April 2002.

[21] D. Sigg. Commissioning of LIGO detectors. *Classical and Quantum Gravity*, 21(5):S409–S415, March 2004.

[22] F. Acernese et al. The status of VIRGO. *Classical and Quantum Gravity*, 23(8):S63–S70, April 2006.

[23] R. Takahashi et al. Operational status of TAMA300. *Classical and Quantum Gravity*, 20(7):S593–S598, September 2003.

[24] H. Lück et al. Status of the GEO600 detector. *Classical and Quantum Gravity*, 23(8):S71–S78, April 2006.

[25] LIGO Livingston site aerial photo by Aero Data, Baton Rouge, Louisiana (`http://www.ligo-la.caltech.edu`). Virgo aerial photo courtesy of European Gravitational Observatory, Cascina, Italy (`http://www.virgo.infn.it`). LISA illustration courtesy of NASA/JPL-Caltech (`http://lisa.jpl.nasa.gov`).

[26] L. Barack and C. Cutler. Confusion noise from LISA capture sources. *Physical Review D*, 70(12):122002, December 2004.

[27] 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–S514, March 2004.

[28] M. Vallisneri. What can we learn about neutron stars from gravity-wave observations? *Arxiv preprint gr-qc/0202037*, February 2002.

[29] C. Cutler and É. É. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform? *Physical Review D*, 49(6):2658–2697, March 1994.

[30] A. Pai, S. Dhurandhar, and S. Bose. Data-analysis strategy for detecting gravitational-wave signals from inspiraling compact binaries with a network of laser-interferometric detectors. *Physical Review D*, 64(4):042004, August 2001.

[31] B. Krishnan et al. Hough transform search for continuous gravitational waves. *Physical Review D*, 70(8):082001, October 2004.

[32] K. A. Arnaud et al. Report on the first round of the Mock LISA Data Challenges. *Classical and Quantum Gravity*, 24(19):S529–S539, October 2007.

[33] H. M. Collins. Lead into gold: the science of finding nothing. *Studies in History and Philosophy of Science*, 34(4):661–691, December 2003.

[34] S. L. Larson, W. A. Hiscock, and R. W. Hellings. Sensitivity curves for spaceborne gravitational wave detectors. *Physical Review D*, 62(6):052001, August 2000.

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

[36] J. Crowder and N. J. Cornish. Solution to the galactic foreground problem for LISA. *Physical Review D*, 75(4):043008, February 2007.

[37] 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, July 2005.

[38] J. Crowder and N. J. Cornish. Extracting galactic binary signals from the first round of Mock LISA Data Challenges. *Classical and Quantum Gravity*, 24(19):S575–S585, October 2007.

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

[40] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the theory of statistics*. McGraw-Hill, New York, 3rd edition, 1974.

[41] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall / CRC, Boca Raton, 1996.

[42] A. Gelman, J. B. Carlin, H. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall / CRC, Boca Raton, 1997.

[43] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, September 1949.

[44] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[45] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C: The art of scientific computing*. Cambridge University Press, Cambridge, 2nd edition, 1992.

[46] A. Gelman and D. B. Rubin. Inference from interative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, March 1992.

[47] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, December 1998.

[48] 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, 1991.

[49] K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to Spin Glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, June 1996.

[50] U. H. E. Hansmann. Parallel Tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1–3):140–150, December 1997.

[51] D. A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *Journal of Chemical Physics*, 117(15):6911–6914, October 2002.

[52] A. Kone and D. A. Kofke. Selection of temperature intervals for parallel-tempering simulations. *Journal of Chemical Physics*, 122(20):206101, May 2005.

[53] A. E. Raftery, M. A. Newton, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Working paper 60R, Center for Statistics and the Social Sciences, University of Washington, Seattle, June 2006.

[54] P. J. Bickel and J. K. Ghosh. A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. *The Annals of Statistics*, 18(3):1070–1090, September 1990.

[55] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`.

[56] M. Aitkin, R. J. Boys, and T. Chadwick. Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, 15(3):217–230, July 2005.

[57] A. P. Dempster. The direct use of likelihood for significance testing. *Statistics and Computing*, 7(4):246–252, December 1997.

[58] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley & Sons, New York, 1991.

[59] F. Liang and H. W. Wong. Real-parameter Evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association*, 96(454):653–666, June 2001.

[60] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading, Mass., 1989.

[61] W. R. Gilks, G. O. Roberts, and E. I. George. Adaptive direction sampling. *The Statistician*, 43(1):179–189, 1994.

[62] C. Gasquet and P. Witomski. *Fourier analysis and applications*. Springer-Verlag, New York, 1998.

[63] P. C. Gregory. *Bayesian logical data analysis for the physical sciences*. Cambridge University Press, Cambridge, 2005.

[64] M. Frigo and S. G. Johnson. FFTW 3.0.1, a C subroutine library for computing the discrete Fourier Transform (DFT), 2003. URL `http://www.fftw.org`.

[65] F. J. Harris. On the use of windows for harmonic analysis with the Discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.

[66] C. Bingham, M. D. Godfrey, and J. W. Tukey. Modern techniques of power spectrum estimation. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2):56–66, June 1967.

[67] P. Bloomfield. *Fourier analysis of time series: An introduction*. Wiley & Sons, New York, 1976.

[68] G. M. Jenkins. General considerations in the analysis of spectra. *Technometrics*, 3(2):167–190, May 1961.

[69] P. D. Welch. The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2):70–73, June 1967.

[70] R. E. Crochiere. A general program to perform sampling rate conversion of data by rational ratios. In A. C. Schell et al., editors, *Programs for digital signal processing*, chapter 8.2. IEEE Press, New York, 1979.

[71] E. C. Ifeachor and B. W. Jervis. *Digital signal processing: A practical approach*, chapter 6, pages 285–309. Addison-Wesley, Wokingham, England, 1993.

[72] D. W. Scott. *Multivariate density estimation: Theory, practice and visualization*. Wiley & Sons, New York, 1992.

[73] T. F. Chan, G. H. Golub, and R. J. Le Veque. Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247, August 1983.

[74] B. P. Welford. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420, August 1962.

[75] K. V. Mardia and P. E. Jupp. *Directional statistics*. Wiley & Sons, Chichester, 2000.

[76] M. Paprzycki and P. Stpiczyński. A brief introduction to parallel computing. In E. J. Kontoghiorghes, editor, *Handbook of parallel computing and statistics*, chapter 1, pages 3–41. Chapman & Hall / CRC, Boca Raton, 2006.

[77] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI: portable parallel programming with the message-passing interface*. MIT Press, Cambridge, Mass., 1994.

[78] J. W. Armstrong, F. B. Estabrook, and M. Tinto. Time-delay interferometry for space-based gravitational wave searches. *The Astrophysical Journal*, 527(2):814–826, December 1999.

[79] L. Blanchet. Post-Newtonian computation of binary inspiral waveforms. In I. Ciufolini, V. Gorini, U. Moschella, and P. Fré, editors,

Gravitational waves: Proceedings of the Como school on gravitational waves in astrophysics. Institute of Physics Publishing, Bristol, 2001. See also Arxiv preprint gr-qc/0104084.

[80] N. J. Cornish, L.J. Rubbo, and O. Poujade. The LISA Simulator, version 2.1.1, June 2006. URL `http://www.physics.montana.edu/LISA`.

[81] J. Cornish, N and L. J. Rubbo. LISA response function. *Physical Review D*, 67(2):022001, January 2003.

[82] B. Allen. Gravitational wave detector sites. *Arxiv preprint gr-qc/9607075*, July 1996.

[83] K. R. Lang. *Astrophysical formulae*, volume II. Springer-Verlag, Berlin, 3rd edition, 1999.

[84] European Organization for the Safety of Air Navigation (EUROCONTROL), Institute of Geodesy and Navigation (IfEN), Brussels/Munich. *WGS 84 implementation manual*, 1998. URL `http://www.wgs84.com`.

[85] D. Brown. *Searching for gravitational radiation from black hole machos in the galactic halo*. PhD thesis, The University of Wisconsin-Milwaukee, 2004.

[86] L. Blanchet, G. Faye, B. R. Iyer, and B. Joguet. Gravitational-wave inspiral of compact binary systems to 7/2 post-Newtonian order. *Physical Review D*, 65(6):061501, March 2002. Note the erratum [87].

[87] L. Blanchet, G. Faye, B. R. Iyer, and B. Joguet. Erratum: Gravitational-wave inspiral of compact binary systems to 7/2 post-Newtonian order. *Physical Review D*, 71(12):129902, June 2005. (See also [86]).

[88] P. J. Mohr and B. N. Taylor. CODATA recommended values of the fundamental physical constants: 2002. *Review of Modern Physics*, 77(1):1–107, January 2005.

[89] R. Kacker and A. Jones. On use of Bayesian statistics to make the 'Guide to the expression of uncertainty in measurement' consistent. *Metrologia*, 40(5):235–248, October 2003.

[90] K. A. Arnaud et al. An overview of the Mock LISA Data Challenges. In S. M. Merkowitz and J. C. Livas, editors, *Laser Interferometer Space Antenna: 6th International LISA Symposium*, volume 873 of *AIP conference proceedings*, pages 619–624, November 2006.

[91] MLDC taskforce. Document for challenge 1, draft v1.0, August 2006. URL `http://svn.sourceforge.net/viewvc/*checkout*/lisatools/Docs/challenge1.pdf`.

[92] K. A. Arnaud et al. An overview of the second round of the Mock LISA Data Challenges. *Classical and Quantum Gravity*, 24(19):S551–S564, October 2007.

[93] T. Tanaka and H. Tagoshi. Use of new coordinates for the template space in a hierarchical search for gravitational waves from inspiraling binaries. *Physical Review D*, 62(8):082001, October 2000.

[94] L. Blanchet, T. Damour, G. Esposito-Farèse, and B. R. Iyer. Gravitational radiation from inspiralling compact binaries completed at the third post-Newtonian order. *Physical Review Letters*, 93(9):091101, August 2004.

[95] K. G. Arun, L. Blanchet, B. R. Iyer, and M. S. S. Qusailah. The 2.5 PN gravitational wave polarizations from inspiralling compact binaries in circular orbits. *Classical and Quantum Gravity*, 21(15):3771–3801, August 2004. Note the erratum [96].

[96] K. G. Arun, L. Blanchet, B. R. Iyer, and M. S. S. Qusailah. Corrigendum: The 2.5PN gravitational wave polarizations from inspiralling compact binaries in circular orbits. *Classical and Quantum Gravity*, 22(14):3115–3117, July 2005. (See also [95]).

[97] L. J. Rubbo, N. J. Cornish, and O. Poujade. Forward modeling of space-borne gravitational wave detectors. *Physical Review D*, 69(8):082003, April 2004.

[98] M. Vallisneri. Synthetic LISA: Simulating time delay interferometry in a model LISA. *Physical Review D*, 72(2):022001, January 2005.

[99] S. E. Timpano, L. J. Rubbo, and N. J. Cornish. Characterizing the galactic gravitational wave background with LISA. *Physical Review D*, 73(12):122001, June 2006.

[100] T. A. Prince, M. Tinto, S. L. Larson, and J. W. Armstrong. LISA optimal sensitivity. *Physical Review D*, 66(12):122002, December 2002.

[101] J. D. Romano and G. Woan. Principal component analysis for LISA: The time delay interferometry connection. *Physical Review D*, 73(10):102001, May 2006.

[102] E. T. Jaynes. Prior probabilities. *IEEE transactions on systems science and cybernetics*, SEC-4(3):227–241, September 1968.

[103] G. L. Bretthorst. The near-irrelevance of sampling frequency distributions. In W. v. d. Linden et al., editors, *Maximum Entropy and Bayesian Methods*, pages 21–46. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[104] L. S. Finn and D. F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Physical Review D*, 47(6):2198–2219, March 1993.

[105] L. S. Finn. Observational constraints on the neutron star mass distribution. *Physical Review Letters*, 73(14):1878–1881, October 1994.

[106] M. H. van Kerkwijk, J. van Paradijs, and E. J. Zuiderwijk. On the masses of neutron stars. *Astronomy and Astrophysics*, 303:497–501, 1995.

[107] K. Belczynski, V. Kalogera, and T. Bulik. A comprehensive study of binary compact objects as gravitational wave sources: evolutionary channels, rates, and physical properties. *The Astrophysical Journal*, 572(1):407–431, June 2002.

[108] H.-T. Janka. Neutron star formation and birth properties. *Arxiv preprint astro-ph/0402200*, February 2004.

[109] P. Teerikorpi. Observational selection bias affecting the determination of the extragalactic distance scale. *Annual Review of Astronomy and Astrophysics*, 35:101–136, September 1997.

[110] M. A. Hendry and J. F. L. Simmons. Distance estimation in cosmology. *Vistas in Astronomy*, 39(3):297–314, 1995.

[111] Frame library (Fr), version v6r19, May 2005. URL `http://lappweb.in2p3.fr/virgo/FrameL`.

[112] J. Janovetz. Parks-McClellan algorithm for FIR filter design (C version), 1998. URL `http://www.janovetz.com/jake`.

[113] B. W. Brown, J. Lovato, K. Russell, and J. Venier. Randlib - Library of C routines for random number generation, 1997. URL `http://biostatistics.mdanderson.org/SoftwareDownload`.

[114] A. D. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lecture notes at the Cargèse summer school on "Functional integration: basics and applications", September 1996.

[115] A. Stroeer, J. Veitch, C. Röver, E. Bloomer, N. Christensen, J. Clark, M. Hendry, C. Messenger, R. Meyer, M. Pitkin, J. Toher, R. Umstätter, A. Vecchio, and G. Woan. Inference on white dwarf binary systems using the first round Mock LISA Data Challenges data sets. *Classical and Quantum Gravity*, 24(19):S541–S549, October 2007.

[116] Open MPI, version 1.1.1, February 2007. URL `http://www.open-mpi.org`.

[117] E. Bloomer and J. Clark. Personal communication, April 2006.

[118] M. van der Sluys, C. Röver, A. Stroeer, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio. Parameter estimation of spinning binaries using MCMC. Poster presentation at the 7th Edoardo Amaldi conference on gravitational waves, July 2007.

[119] D. S. Watkins. *Fundamentals of matrix computations*, chapter 1.4, pages 32–49. Wiley & Sons, New York, 2nd edition, 2002.

[120] R. Levy and W. R. Spillers. *Analysis of geometrically nonlinear structures*. Chapman & Hall / CRC, New York, 1995.