Bayesian Methods for Gravitational Waves and Neural Networks

Philip Graff

Cavendish Astrophysics and Queens' College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

Submitted 06 June 2012
In final form 25 June 2012

Declaration

This dissertation is the result of work carried out in the Astrophysics Group of the Cavendish Laboratory, Cambridge, between October 2008 and June 2012. Except where explicit reference is made to the work of others, the work contained in this dissertation is my own, and is not the outcome of work done in collaboration. No part of this dissertation has been submitted for a degree, diploma or other qualification at this or any other university. The total length of this dissertation does not exceed sixty thousand words.

Philip Graff

25 June 2012

Cambridge was the place for someone from the Colonies or the Dominions to go on to, and it was to the Cavendish Laboratory that one went to do physics.

Sir Aaron Klug

I would like to dedicate this thesis to my loving parents, who have supported and encouraged me through the years.

Acknowledgements

I acknowledge financial support from the Gates Cambridge Trust, founded by the Bill and Melinda Gates Foundation, which provided me with the opportunity to study in Cambridge. The Cambridge Philosophical Society, Queens' College, and the Astrophysics Group of the Cavendish Laboratory have also provided financial support for my studies and presentations at conferences.

I am very grateful to Anthony Lasenby and Mike Hobson for their continued encouragement and help over the past three and a half years, both in my projects and in how to perform proper scientific research. I also thank them for proof-reading this thesis and providing useful feedback. I would also like to thank Farhan Feroz for all of the help and advice he has given me. Without him, much of this work would not have been able to be completed in the time it was. I thank David Titterington for help with local computing issues and Stuart Rankin and Andrey Kaliazin for technical support with the Darwin and COSMOS supercomputing facilities.

I am very happy to have spent the last three and a half years in the Astrophysics Group, whose members past and present have been great friends. Whether in the office, in the tea room, at the pub after a hard day's work, at a party, or elsewhere, they have been wonderful compatriots in sharing complaints, jokes, and life as an astrophysicist. The members of Queens' College MCR, and especially my flatmate for two years Robert Lowe, also deserve great thanks for making my time in Cambridge as fantastic as it has been. My other friends in and around Cambridge – fellow Gates Cambridge Scholars, the Newnham College MCR matriculation class of 2009, the Gaelic Athletics Club, the Cambridge Royals baseball team, and many

more – have also been great sources of fun and support without whom my Cambridge experience would have been much more boring.

Lastly, I would like to thank my parents, siblings, and friends back in the USA for their continued support throughout my PhD.

LIGO Scientific Collaboration Acknowledgement

The second chapter of this thesis contains analysis of LIGO data performed by the author as a member of the LIGO Scientific Collaboration. The author gratefully acknowledges the support of the United States National Science Foundation for the construction and operation of the LIGO Laboratory and the Science and Technology Facilities Council of the United Kingdom, the Max-Planck-Society, and the State of Niedersachsen/Germany for support of the construction and operation of the GEO600 detector. The authors also gratefully acknowledge the support of the research by these agencies and by the Australian Research Council, the Council of Scientific and Industrial Research of India, the Istituto Nazionale di Fisica Nucleare of Italy, the Spanish Ministerio de Educación y Ciencia, the Conselleria d'Economia, Hisenda i Innovacio of the Govern de les Illes Balears, the Scottish Funding Council, the Scottish Universities Physics Alliance, The National Aeronautics and Space Administration, the Carnegie Trust, the Leverhulme Trust, the David and Lucile Packard Foundation, the Research Corporation, and the Alfred P. Sloan Foundation.

Abstract

Einstein's general theory of relativity has withstood 100 years of testing and will soon be facing one of its toughest challenges. In a few years we expect to be entering the era of the first direct observations of *gravitational waves*. These are tiny perturbations of space-time that are generated by accelerating matter and affect the measured distances between two points. Observations of these using the laser interferometers, which are the most sensitive length-measuring devices in the world, will allow us to test models of interactions in the strong field regime of gravity and eventually general relativity itself.

I apply the tools of Bayesian inference for the examination of gravitational wave data from the LIGO and Virgo detectors. This is used for signal detection and estimation of the source parameters. I quantify the ability of a network of ground-based detectors to localise a source position on the sky for electromagnetic follow-up. Bayesian criteria are also applied to separating real signals from glitches in the detectors. These same tools and lessons can also be applied to the type of data expected from planned space-based detectors. Using simulations from the Mock LISA Data Challenges, I analyse our ability to detect and characterise both burst and continuous signals. The two seemingly different signal types will be overlapping and confused with one another for a space-based detector; my analysis shows that we will be able to separate and identify many signals present.

Data sets and astrophysical models are continuously increasing in complexity. This will create an additional computational burden for performing Bayesian inference and other types of data analysis. I investigate the application of the MOPED algorithm for faster parameter estimation and

data compression. I find that its shortcomings make it a less favourable candidate for further implementation.

The framework of an *artificial neural network* is a simple model for the structure of a brain which can “learn” functional relationships between sets of inputs and outputs. I describe an algorithm developed for the training of feed-forward networks on pre-calculated data sets. The trained networks can then be used for fast prediction of outputs for new sets of inputs. After demonstrating capabilities on toy data sets, I apply the ability of the network to classifying handwritten digits from the MNIST database and measuring ellipticities of galaxies in the Mapping Dark Matter challenge.

The power of neural networks for learning and rapid prediction is also useful in Bayesian inference where the likelihood function is computationally expensive. The new BAMBI algorithm is detailed, in which our network training algorithm is combined with the nested sampling algorithm MultiNest to provide rapid Bayesian inference. Using samples from the normal inference, a network is trained on the likelihood function and eventually used in its place. This is able to provide significant increase in the speed of Bayesian inference while returning identical results. The trained networks can then be used for extremely rapid follow-up analyses with different priors, obtaining orders of magnitude of speed increase.

Learning how to apply the tools of Bayesian inference for the optimal recovery of gravitational wave signals will provide the most scientific information when the first detections are made. Complementary to this, the improvement of our analysis algorithms to provide the best results in less time will make analysis of larger and more complicated models and data sets practical.

Contents

Chapter 1

Introduction

> I was born not knowing and have had only a little time to change that here and there.
>
> Richard P. Feynman

1.1 Gravitational Waves

Einstein's general theory of relativity [1] (GR) predicts the existence of gravitational waves (GWs), perturbations to the curvature of spacetime that are generated in regions of strong gravity and will allow scientists to observe the universe through an entirely new window. The direct detection of gravitational waves will usher in a new era of astronomy. The GW spectrum in many ways independent of and complementary to electromagnetic radiation. GWs can be used to probe astrophysical events otherwise unobservable, such as the merger of two black holes, the inner processes of supernovae, and the equation of state of neutron star matter. Projects are currently underway across the world to detect these very weak signals. Indirect evidence for the existence of gravitational waves exists from observations of binary pulsars spinning in towards each other as the system loses energy due to the emission of GWs [2].

GR has withstood almost 100 years of testing and verification; however, it is now facing one of its toughest challenges in gravitational wave searches. Direct observations may allow for the testing of alternate theories of gravity as a test of the strong field predictions of general relativity.

1.1.1 General Relativity Theory

In the linearised regime of general relativity, applicable for weak gravitational fields, we can assume a flat Minkowski background metric for the spacetime, given by $\eta_{ab} = \text{diag}(-1,1,1,1)$. The true metric can then be defined as

$$g_{ab} = \eta_{ab} + h_{ab}, \quad \| h_{ab} \| \ll 1. \tag{1.1}$$

This constrains us to a weak gravitational field and we will refer to h_{ab} as the metric perturbation. We can now derive Einstein's equations in the linearised regime, where terms of order h_{ab}^2 and above will be ignored as they are significantlly small relative to the first-order perturbation. I will only present the main results of this procedure, full steps can be found in [3].

We calculate the Einstein tensor with this new metric, discarding small terms of $O(h_{ab}^2)$ and higher, and obtain

$$G_{ab} = \frac{1}{2}(\partial_c \partial_b h^c_{\ a} + \partial^c \partial_a h_{bc} - \Box h_{ab} - \partial_a \partial_b h - \eta_{ab} \partial_c \partial^d h^c_{\ d} + \eta_{ab} \Box h). \tag{1.2}$$

We substitute in for the trace-reversed perturbation, $\bar{h}_{ab} = h_{ab} - \frac{1}{2}\eta_{ab}h$ (where $h = h^a_{\ a}$) and choose to satisfy the Lorenz gauge condition that $\partial^a \bar{h}_{ab} = 0$. Doing so is allowed because there is freedom to choose the gauge (coordinate system) we want to work in without affecting the physical observables of the system. Therefore, the Einstein tensor simplifies to $G_{ab} = -\frac{1}{2}\Box \bar{h}_{ab}$, where $\Box = \partial_c \partial^c = \nabla^2 - \partial_t^2$. The linearised Einstein equation is thus

$$\Box \bar{h}_{ab} = -16\pi T_{ab}. \tag{1.3}$$

In a vacuum $T_{ab} = 0$ and this equation describes the propagation of plane waves traveling at the speed of light. These are gravitational waves.

Furthermore, in addition to the Lorenz gauge we can specify the gauge to be purely spatial and traceless,

$$h_{tt} = h_{ti} = 0 \quad \text{and} \quad h = 0. \tag{1.4}$$

This now fully specifies our coordinate system. The combination of these choices determines that the perturbation is transverse, meaning that $\partial_i h_{ij} = 0$. Therefore, we call this the transverse-traceless gauge and write it as h_{ij}^{TT}. Being traceless, $h_{ij}^{TT} = \bar{h}_{ij}^{TT}$. This gauge choice is very useful and leaves only two independent components of the

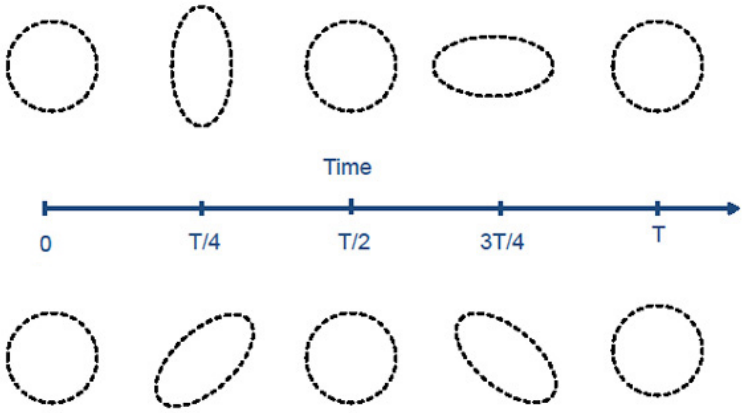


Figure 1.1: The effect of a passing sinusoidal gravitational wave with period T on a ring of particles. Image from `www.learner.org`.

metric perturbation. When orienting our coordinate system to describe waves traveling along the z axis, these degrees of freedom are given by

$$h_{xx}^{TT} = -h_{yy}^{TT} \equiv h_{+}(t-z), \tag{1.5a}$$

$$h_{xy}^{TT} = h_{yx}^{TT} \equiv h_{\times}(t-z). \tag{1.5b}$$

The two components can now be seen as two independent polarisations, plus and cross, whose effect on a ring of particles (for a sinusoidal GW source) can be seen in Figure 1.1. The matter source of a gravitational wave will specify the strengths of its two polarisations as a function of time.

We are able to detect gravitational waves because they affect the geodesics that free-falling bodies follow in spacetime. In the transverse-traceless gauge the spacetime coordinates of a body will remain the same, as they move with the waves, but the proper separation of two bodies will change as the metric varies with the passing wave. A straightforward analysis [3] shows that the proper separation of two bodies on the x axis will oscillate with a fractional length change of

$$\frac{\delta L}{L} \simeq \frac{1}{2} h_{xx}^{TT}(t, z=0), \tag{1.6}$$

from a wave travelling along the z axis. This formula can be generalized for any wave propagation direction and any coordinate separation. Because of this effect, the amplitude of the GW signal is known as the GW strain.

A complete mathematical background and derivation of general relativity, as well as its various applications and extensions, can be found in [4].

1.1.2 Gravitational Wave Sources

If we consider a gravitational wave source that is near the origin of our coordinate system, measure the field at a distance r that is large compared to the spatial extent of the source, and assume that the constituents of the source are moving slowly compared to c, then the solution to the linearised Einstein equation can be given by the compact-source approximation (a full derivation and further details on the following analysis can be found in [5]),

$$\bar{h}^{\mu\nu}(ct,\vec{x}) = -\frac{4G}{c^4 r}\int T^{\mu\nu}(ct-r,\vec{y})d^3\vec{y}. \tag{1.7}$$

Taking the different components of $T^{\mu\nu}$ separately, $\int T^{00}d^3\vec{y} = Mc^2$ and $\int T^{0i}d^3\vec{y} = \int T^{i0}d^3\vec{y} = P_i c$. These components describe the integrated energy and momentum of the system, respectively, and are therefore constant for a closed system and will not contribute to the varying gravitational wave strain. Additionally, we are able to transform our coordinates to the source's centre-of-momentum frame so that

$$\bar{h}^{00} = -\frac{4GM}{c^2 r}, \quad \bar{h}^{0i} = \bar{h}^{i0} = 0. \tag{1.8}$$

Using the transverse property of $T^{\mu\nu}$ and Gauss's law, we can re-write the integral over the remaining components as

$$\int T^{ij}d^3\vec{y} = \frac{1}{2c^2}\frac{d^2}{dt'^2}\int T^{00}y^i y^j d^3\vec{y}, \tag{1.9}$$

where $ct' = ct - r$. Therefore we obtain the *quadrupole* formula

$$\bar{h}^{ij}(ct,\vec{x}) = -\frac{2G}{c^6 r}\frac{d^2 I^{ij}(ct')}{dt'^2}, \tag{1.10}$$

where we define the quadrupole moment tensor of the source as

$$I^{ij}(ct) = \int T^{00}(ct,\vec{y})y^i y^j d^3\vec{y}. \tag{1.11}$$

For a given source, this can be used to determine the far-field gravitational radiation.

As an illustration, let us consider two particles of equal mass M moving non-relativistically in circular orbits of radius a in the $z = 0$ plane about their common centre of mass with angular frequency Ω. Since the particles are moving slowly, we may use the approximation that $T^{00} \approx c^2\rho$, where ρ is the density of the source in its own reference frame. Therefore,

$$I^{ij}(ct) = c^2 \int \rho(ct, \vec{y}) y^i y^j d^3\vec{y}. \tag{1.12}$$

It is straightforward to show that for this source

$$I^{ij}(ct) = Mc^2a^2 \begin{pmatrix} 1+\cos(2\Omega t) & \sin(2\Omega t) & 0 \\ \sin(2\Omega t) & 1-\cos(2\Omega t) & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{1.13}$$

Substituting this into Equation (1.10) yields

$$\bar{h}^{ij}(ct, \vec{x}) = \frac{8GMa^2\Omega^2}{c^4 r} \begin{pmatrix} \cos(2\Omega t') & \sin(2\Omega t') & 0 \\ \sin(2\Omega t') & -\cos(2\Omega t') & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{1.14}$$

Considering only the radiative part of $h^{\mu\nu}$ gives

$$\bar{h}^{\mu\nu}_{\text{rad}}(ct, \vec{x}) = \frac{8GMa^2\Omega^2}{c^4 r} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \cos(2\Omega t') & \sin(2\Omega t') & 0 \\ 0 & \sin(2\Omega t') & -\cos(2\Omega t') & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1.15}$$

We should note at this point two important aspects of our solution. First, the amplitude scales as $1/r$, which means that the GW is a spherical wave. However, at large radii it can be approximated as a plane wave. Additionally, the $1/r$ scaling implies that a factor of X improvement in detector sensitivity will yield the same factor of X in observable distance and thus X^3 in observable volume. Our second comment on the solution is that the frequency of the GW is twice the orbital frequency of the binary; this is due to the fact that the radiation is quadrupolar. Monopole and dipole radiation are zero due to the conservation of mass and momentum, respectively.

We now consider the polarisation of the GW received by an observer. For an observer on the z axis, we begin by noting that Equation (1.15) is already transverse-traceless so $\bar{h}^{\mu\nu}_{TT} = h^{\mu\nu}_{TT}$. $h_+(t) \propto \cos(2\Omega t)$ and $h_\times(t) \propto \sin(2\Omega t)$, corresponding to right-handed circularly polarised radiation. For an observer on the x axis, however, we must first convert Equation (1.15) to the transverse-traceless gauge. We zero the non-transverse components and then subtract one-half of the trace from the remaining diagonal components to obtain

$$(\bar{h}^{TT}_{\text{rad}})^{\mu\nu}(ct,\vec{x}) = \frac{4GMa^2\Omega^2}{c^4 r}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\cos(2\Omega t') & 0 \\ 0 & 0 & 0 & \cos(2\Omega t') \end{pmatrix}. \tag{1.16}$$

We observe that the radiation is in the plus polarisation entirely. It is also important to note that the coefficient in front is a factor of 2 smaller. This combined with the fact that there is radiation in only one polarisation state implies that the GW energy emitted in the z direction will be eight times larger than that in the x or y directions, making it highly anisotropic.

Integrating the energy flux over all angles and averaging over multiple GW periods yields that the energy lost by the system due to the emission of gravitational waves is

$$L_{GW} = \frac{128GM^2a^4\Omega^6}{5c^2}. \tag{1.17}$$

This loss of energy is not considered in the initial analysis where we maintained conservation of mass-energy. Waveforms defined in this manner must use a quasi-static approximation – the evolution of the orbital radius is defined by the internal energy and this further defines the orbital period through Kepler's third law. The loss of energy means that system will evolve inwards (decreasing a) with increasing Ω until collision/merger. The strong dependence of the emitted energy on frequency results in this occurring at an increasing rate and creates a burst of radiation at the end. This is the characteristic "chirp" signal that data analysis pipelines are searching for. For sources far from merger, the energy lost in gravitational wave emission is low enough that the frequency evolution can be modeled as an initial frequency f and a frequency derivative $\dot{f}$ over a period of observation that is short compared to the time to reach merger.

In general, this approximation holds for binaries with large separations. As the binary spirals inwards, further *post-Newtonian* [6–12] corrections must be made to the waveform models. These are corrections that expand in powers of v/c, where v is the transverse velocity of a binary member, and provide more accurate modeling

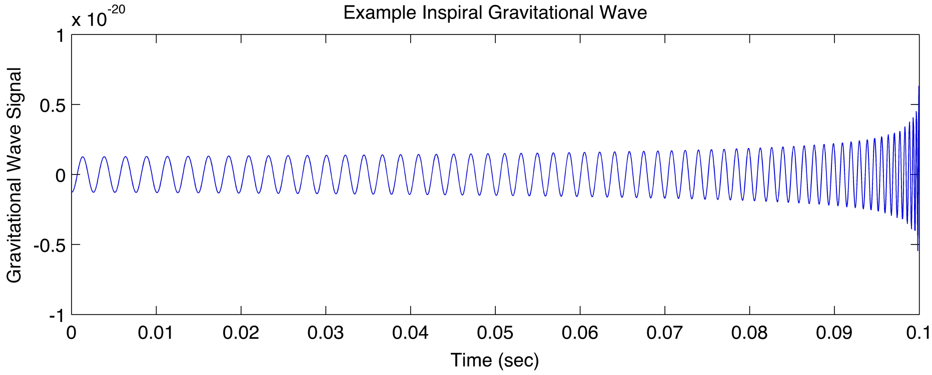


Figure 1.2: A sample inspiral GW signal that LIGO might observe. Image from `www.ligo.org`.

of the phase and amplitude of the GW signal. At higher post-Newtonian orders the spin of the binary members becomes important and spin-orbit ($\mathbf{L}\cdot\mathbf{S}_1$ and $\mathbf{L}\cdot\mathbf{S}_2$ terms) and spin-spin ($\mathbf{S}_1\cdot\mathbf{S}_2$ terms) interactions can have significant effects on the waveform (amplitude and frequency modulation and orbit precession) [13–18]. A sample inspiral GW signal is shown in Figure 1.2.

Near to the point of merger, the effective-one-body (EOB) approach to waveform modeling can provide a better template [19–23]. This method reconstitutes the problem as that of a test particle with some spin spiraling in to a Kerr black hole. Since accurate numerical simulations of Einstein's full equations have been performed [24–29], this waveform can then be attached to a fit of the signal coming from the merger of the two objects (this has been done for black hole binaries only). A *ringdown* waveform derived from black hole perturbation theory and numerical simulations is then used to complete the signal. This is characterised by constant modes of oscillation exponentially decreasing in amplitude over time. Combining post-Newtonian and/or EOB theory with numerical relativity simulations and black hole perturbation theory allows for the construction of complete inspiral-merger-ringdown (IMR) [30–36] hybrid waveform families that are analytic fits over specific parameter ranges. By modeling the entire signal, these allow for the extraction of the most information and signal-to-noise ratio. However, they are limited by our ability to perform numerical simulations with enough orbits prior to merger and covering a large enough area of the possible parameter space.

Similar brief derivations of the quadrupole formula for gravitational wave radiation from compact binaries can be found in [37, 38]. A more complete derivation and analysis of the gravitational waves from compact binaries is available in [39] and a review of the post-Newtonian formalism can be found in [40]. Gravitational waves are created by a variety of astrophysical sources, such as supernova explosions [41] and spinning neutron stars [42–47], but compact binaries are the most well-studied. A review of sources that may be found in high and low frequency ranges can be found in [48].

Follow-up searches to find GWs associated with observed gamma-ray bursts have been performed [49] and studies are underway to cross-correlate GW and neutrino observations [50]. Gravitational wave detections can be used to test general relativity and measure any deviations from Einstein's theory [51, 52], such as the existence of a massive graviton [53] (as GR predicts a massless graviton). Observations can also be used to test cosmological models by using binary mergers as a standard siren for luminosity distance and redshift measurements [54–58].

1.1.3 Gravitational Wave Detectors

The Laser Interferometer Gravitational-wave Observatory (LIGO) [59] is a major US project funded by the National Science Foundation for the detection of GWs. It consists of a pair of observatories located in Livingston, Louisiana and Hanford, Washington in the United States. The detectors are sensitive in the audio band (~ 30 Hz to several kHz) with strain amplitudes as low as 10^{-21}. The main research centres are located at the California Institute of Technology in Pasadena, California and the Massachusetts Institute of Technology in Cambridge, Massachusetts.

The basic geometry of each observatory is that of a 90° Michelson interferometer. However, some changes have been made so as to increase their sensitivity. Each arm consists of a resonant Fabry-Perot cavity, with the two mirrors also acting as the gravitational test masses. The resonance amplifies the effect of a passing gravitational wave on the phase difference between the two arms. Another mirror is placed between the laser source and the beamsplitter so as to create a resonant cavity that increases the total power in the interferometer by not letting light escape back towards the laser—this is called the power recycling mirror. An input mode cleaner before the PRM removes unwanted spatial modes from the laser source, leaving only the $(l,m) = (0,0)$ mode, a single central Gaussian peak. The Livingston observatory has an interferometer with 4km arm lengths, denoted L1. The Hanford observatory has two interferometers in the same vacuum system; H1 has 4km arms and H2 has 2km arms. In the final Initial LIGO science run, S5, all detectors operated at design strain sensitivity. The status of compact binary coalescence searches at that time is given in [60] and descriptions of the standard analysis pipeline are found in [61, 62].

Enhanced LIGO (eLIGO) [63] was an improvement upon the sensitivity of LIGO by a factor of approximately two that did not require opening the vacuum chambers. A plot of the sensitivity measured during eLIGO's science run (S6) as a function of GW frequency can be seen in Figure 1.3. At the low frequency end, the sensitivity is limited by seismic noise. At the high frequency end, the limiting factor is quantum shot noise from the error in counting the photons that reach the detector; the fractional error in photon counts on the shorter integration timescales approaches the same fractional distance to be measured. For eLIGO, a more powerful 35 W laser was installed, the GW signal was switched to a DC readout that operates slightly off the dark fringe, and an output mode cleaner was installed to remove unwanted spatial modes that add shot noise but do not contribute to the GW signal. The S6 science run was performed over 2009–2010 in this configuration. Despite the many controls, LIGO experiences transient noise sources that detection pipelines have to rule out as possible signals [64–67].

The next stage of development is Advanced LIGO (aLIGO) [69–71], which aims to increase the sensitivity by a further factor of five. aLIGO's upgrades include adding a signal recycling mirror that both increases the broadband sensitivity and allows the sensitivity of the interferometer to be "tuned". Additionally, the test masses will be hung from triple pendulums to decrease their sensitivity to seismic noise and reduce the lower limit on observations to $\sim$ 10Hz. A 180 W laser will be installed and active modulators and isolators will be used on the test masses to compensate for thermal deformation. The test masses are also larger in size at 40 kg, up from 10 kg in Initial and Enhanced LIGO. The H2 interferometer will be removed and the mirrors sent to India in support of the building of a detector there by the IndIGO project [72]. A schematic of the aLIGO setup can be seen in Figure 1.4.

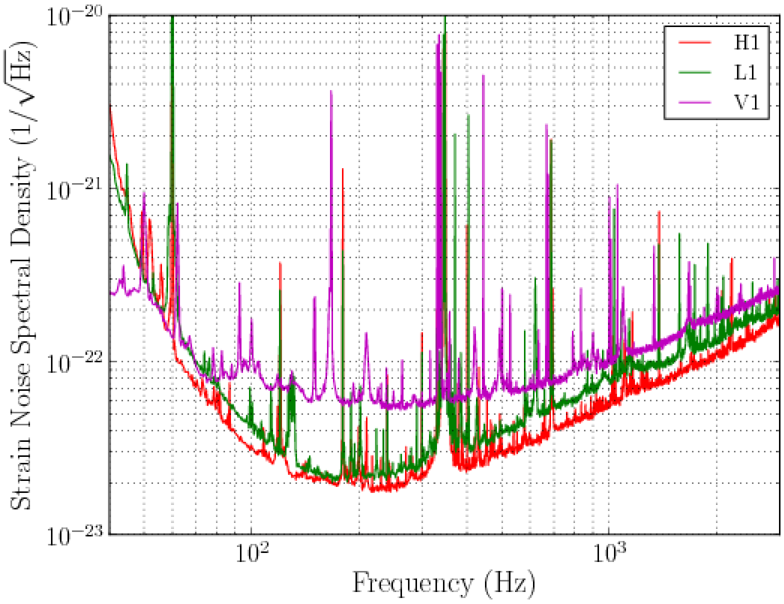


Figure 1.3: A plot of the strain sensitivities of the LIGO and Virgo interferometers from the S6 and VSR2/3 science runs [68].

In addition to the LIGO project in the United States, there is a joint Italian-French observatory, Virgo [73], located in Cascina, Italy. Virgo has a similar setup to LIGO but with 3km arms and active seismic isolation. Virgo's instrumental noise is similar to LIGO's as seen in Figure 1.3, but with a higher noise floor and better sensitivity at frequencies below $\sim$ 50 Hz. Virgo's second and third science runs were performed in coincidence with LIGO's S6. There is also a joint German-British observatory, GEO600 [74], near Sarstedt, Germany that is part of the LIGO Scientific Collaboration. GEO600 is primarily used for testing technology that will be used in upcoming LIGO upgrades, but is currently on "astrowatch" while LIGO and Virgo are offline being upgraded. Due to its shorter arm lengths and lack of Fabry-Perot cavities in the arms, GEO's sensitivity is much lower than LIGO or Virgo. In Japan the TAMA300 [75] observatory is operating while the KAGRA detector (the Large Cryogenic Gravitational-wave Telescope) is under construction [76, 77]. Together, LIGO, Virgo, GEO, LCGT, and IndIGO will form a worldwide network of detectors that will allow for simultaneous detection of GWs and improve our ability to determine the parameters of a signal.

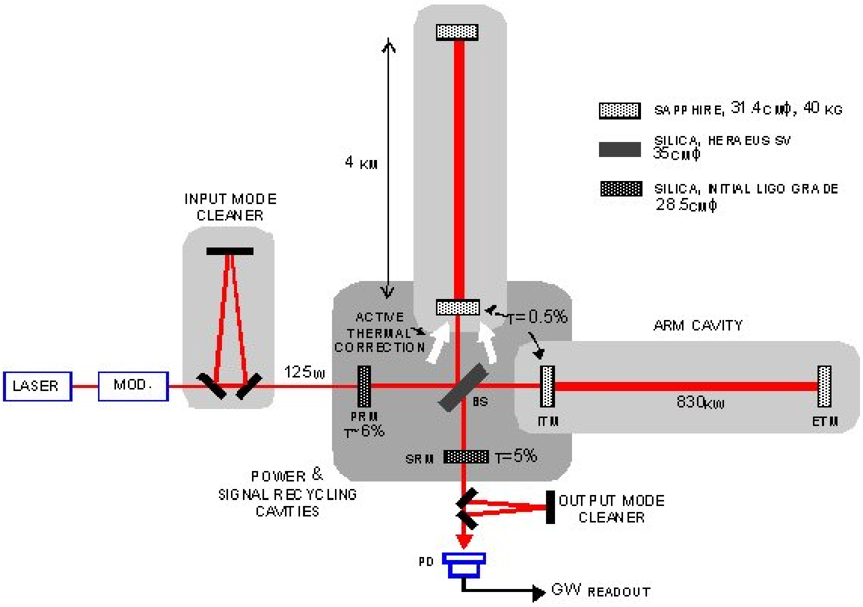

4 KM
▨ SAPPHIRE, 31.4CMϕ, 40 KG
■ SILICA, HERAEUS SV 35CMϕ
▩ SILICA, INITIAL LIGO GRADE 28.5CMϕ
INPUT MODE CLEANER
ACTIVE THERMAL CORRECTION
T=0.5%
ARM CAVITY
LASER
MOD.
125W
BS
ITM
830KW
ETM
PRM
T~6%
SRM
T=5%
POWER & SIGNAL RECYCLING CAVITIES
OUTPUT MODE CLEANER
PD
GW READOUT

Figure 1.4: A diagram of the aLIGO interferometer. ETM = end test mass, ITM = initial test mass, BS = 50/50 beam splitter, PD = photodetector, PRM = power recycling mirror, SRM = signal recycling mirror, and MOD = phase modulation. Image from `www.ligo.caltech.edu`.

Studies are also underway for the construction of a third-generation ground-based interferometric detector, the Einstein Telescope [78, 79].

The next generation GW detector under development is ESA's New Gravitational-wave Observatory (NGO) [80, 81], a re-scoped version of the Laser Interferometer Space Antenna project (LISA) [82]. NGO will consist of three satellites (one 'mother' and two 'daughters') that will be the vertices of an equilateral triangle one million km on a side. The NGO constellation is planned to orbit the Sun in the same orbit as the Earth, but behind and at an inclination of 60° to the ecliptic. NGO will form an interferometer that can observe GWs from $\sim$0.1 mHz up to $\sim$0.1 Hz. This will allow NGO to detect the mergers of supermassive black holes [83, 84], quasi-stationary signals from smaller compact binaries in the galaxy, extreme-mass-ratio inspirals, bursts from cusps in cosmic strings, and a stochastic GW background [85, 86].

In preparation for the NGO mission, LISA Pathfinder [87] will be launched by ESA to test new technology. LPF is an advanced geodesics explorer and will be a proof of principle for NGO. Pathfinder is scheduled for launch in 2014, with NGO launch currently unsure.

A summary of the use of interferometry for detection of gravitational waves both on the ground and in space is given by [88].

The near-perfect clock nature of pulsars also allows for the detection of very long wavelength gravitational waves. As a GW passes between a source pulsar and the Earth, it periodically distorts the spacetime that the pulsar signal must pass through. This results in periodic variations in the timing of the pulsar signal at Earth. By measuring many pulsars' timings to high precision over several years, these variations may be cross-correlated to indicate the passing of a gravitational wave or the presence of a GW background [89–91].

The Gravitational Wave International Committee's roadmap [92] outlines plans for GW detection in the coming decades.

1.2 Bayesian Inference

Gravitational wave detection poses a major data analysis problem. Most signals in the collected data streams are expected to be of low SNR. Additionally, there may be many overlapping signals, not all of the same type, and a non-stationary noise background.

Therefore, we need an efficient and robust method for identifying signals in the data and determining their physical parameters, both intrinsic and extrinsic. To do this, we turn to the methods of Bayesian statistics.

1.2.1 Bayes' Theorem

In our analysis, we wish to estimate the parameters Θ in a model or hypothesis H for given data D. Bayes' theorem states that

$$\Pr(\Theta|D,H) = \frac{\Pr(D|\Theta,H)\Pr(\Theta|H)}{\Pr(D|H)}. \tag{1.18}$$

$\Pr(\Theta|D,H) \equiv P(\Theta)$ is the posterior probability distribution, $\Pr(D|\Theta,H) \equiv \mathcal{L}(\Theta)$ is the likelihood function, $\Pr(\Theta|H) \equiv \pi(\Theta)$ is the prior distribution of the parameters, and $\Pr(D|H) \equiv \mathcal{Z}$ is the Bayesian evidence. The evidence is the factor required to normalise the posterior over Θ, so we can ignore it for parameter estimation and analyse just the un-normalised posterior. It can be found, however, with the following integral (where N is the dimension of the parameter space):

$$\mathcal{Z} = \int \mathcal{L}(\Theta)\pi(\Theta)d^N(\Theta). \tag{1.19}$$

To compare two models, H_0 and H_1, we compare their relative probabilities with respect to given data. This is known as the odds ratio and is given by

$$\frac{\Pr(H_1|D)}{\Pr(H_0|D)} = \frac{\Pr(D|H_1)\Pr(H_1)}{\Pr(D|H_0)\Pr(H_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{\Pr(H_1)}{\Pr(H_0)}. \tag{1.20}$$

The prior probability ratio between the two models, $\Pr(H_1)/\Pr(H_0)$, is taken to be unity when no further information is available as to which model is more probable. Thus, it can be seen from Equation (1.20) that the evidence plays a critical role in model selection as it determines the relative probabilities; $\mathcal{Z}_1/\mathcal{Z}_0$ is known as the evidence ratio. However, calculating the evidence requires the evaluation of a multi-dimensional integral which is far from trivial, especially in the multimodal and degenerate parameter spaces that we will be exploring.

In cases of both model selection and parameter estimation, we require a sampling of the parameter space that provides us with information about the shape of the likelihood function. From this information, we can numerically integrate and compare the evidence of different models or compare the relative likelihoods (and, as we will see, local evidences) of different parameter choices. A more detailed review of Bayesian statistics and its particular applications to cosmology can be found in [93].

1.2.2 MCMC and Nested Sampling

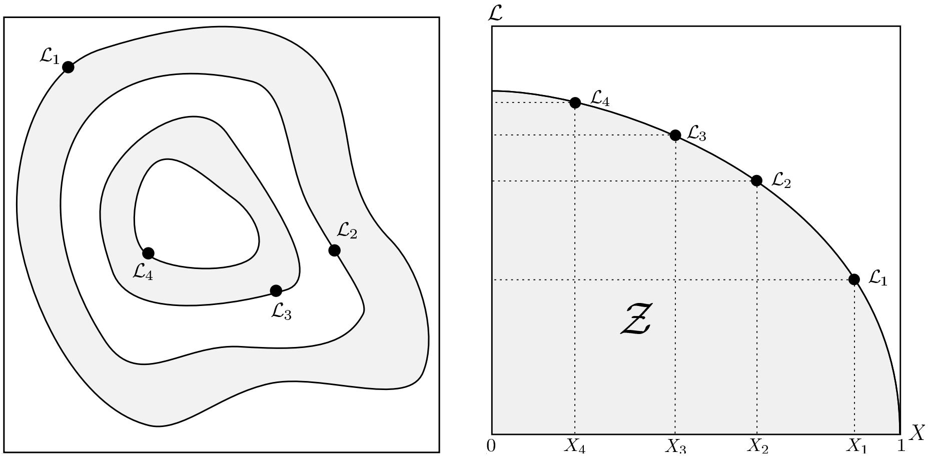
Due to the difficulty in integrating Equation (1.19) for complicated likelihood functions in many dimensions and the computational impracticality of grid approximations, Markov chain Monte Carlo (MCMC) methods have long been used for computing the integral and obtaining samples from the posterior distribution. One of the most popular implementations is the Metropolis-Hastings algorithm [94, 95]. In this algorithm, an initial point is chosen and steps to new points are chosen according to proposal distributions. New points are accepted with the following probability:

$$\Pr(x_{t+1}|x_t) = \min\left(1, \frac{P(x_{t+1})Q(x_{t+1},x_t)}{P(x_t)Q(x_t,x_{t+1})}\right), \tag{1.21}$$

where x_t is the initial point, x_{t+1} is the new point, and the proposal distribution $Q(x_{t+1},x_t)$ is the probability of jumping to point x_{t+1} from point x_t. If the point is not accepted then $x_{t+1} = x_t$. After an initial "burn-in" series of steps, the resulting distribution of samples will be equivalent to sampling from $P(x)$. We therefore set $P(x_t) = \mathcal{L}(x_t)\pi(x_t)$ to obtain samples from the posterior probability distribution. However, this methodology will require many samples and likelihood evaluations, which can be computationally expensive. Additionally, in order to obtain a calculation of the Bayesian evidence, additional techniques such as simulated annealing (a.k.a. parallel tempering) [96, 97] are typically used, further increasing the computational cost. Model selection with MCMC may also be performed using a technique known as "reversible jump" MCMC (RJMCMC), which allows jumps between different model parameter spaces [98].

Nested sampling [99, 100] is a Monte Carlo method targeted at the efficient calculation of the evidence, but also produces posterior inferences as a by-product. It calculates the evidence by transforming the multi-dimensional evidence integral into a one-dimensional integral that is easy to evaluate numerically. This is accomplished by defining the prior volume X as $dX = \pi(\Theta)d^N\Theta$, so that

$$X(\lambda) = \int_{\mathcal{L}(\Theta)>\lambda} \pi(\Theta)d^N\Theta, \tag{1.22}$$



(a)

(b)

Figure 1.5: Cartoon illustrating (a) the posterior of a two dimensional problem; and (b) the transformed $\mathcal{L}(X)$ function where the prior volumes X_i are associated with each likelihood $\mathcal{L}_i$. Images from [101].

where the integral extends over the region(s) of parameter space contained within the iso-likelihood contour $\mathcal{L}(\Theta) = \lambda$. The evidence integral, Equation (1.19), can then be written as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX, \tag{1.23}$$

where $\mathcal{L}(X)$, the inverse of Equation (1.22), is a monotonically decreasing function of X. Thus, if one can evaluate the likelihoods $\mathcal{L}_i = \mathcal{L}(X_i)$, where X_i is a sequence of decreasing values,

$$0 < X_M < \cdots < X_2 < X_1 < X_0 = 1, \tag{1.24}$$

as shown schematically in Figure 1.5, the evidence can be approximated numerically using standard quadrature methods as a weighted sum

$$\mathcal{Z} = \sum_{i=1}^{M} \mathcal{L}_i w_i, \tag{1.25}$$

where the weights w_i for the simple trapezium rule are given by $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$. An example of a posterior in two dimensions and its associated function $\mathcal{L}(X)$ is shown in Figure 1.5.

The summation in Equation (1.25) is performed as follows. The iteration counter is first set to $i = 0$ and N "active" (or "live") samples are drawn from the full prior $\pi(\Theta)$, so the initial prior volume is $X_0 = 1$. The samples are then sorted in order of their likelihood and the smallest (with likelihood $\mathcal{L}_0$) is removed from the active set and replaced by a point drawn from the prior subject to the constraint that the point has a likelihood $\mathcal{L} > \mathcal{L}_0$. The corresponding prior volume contained within the iso-likelihood contour associated with the new live point will be a random variable given by $X_1 = t_1 X_0$, where t_1 follows the distribution $\Pr(t) = Nt^{N-1}$ (i.e., the probability distribution for the largest of N samples drawn uniformly from the interval $[0,1]$). At each subsequent iteration i, the removal of the lowest likelihood point $\mathcal{L}_i$ in the active set, the drawing of a replacement with $\mathcal{L} > \mathcal{L}_i$ and the reduction of the corresponding prior volume $X_i = t_i X_{i-1}$ are repeated, until the entire prior volume has been traversed. The algorithm thus travels through nested shells of likelihood as the prior volume is reduced. The mean and standard deviation of $\log(t)$, which dominates the geometrical exploration, are:

$$E[\log t] = -1/N, \quad \sigma[\log t] = 1/N. \tag{1.26}$$

Since each value of $\log(t)$ is independent, after i iterations the prior volume will shrink down such that $\log X_i \approx -(i \pm \sqrt{i})/N$. Thus, one takes $X_i = \exp(-i/N)$.

The nested sampling algorithm is terminated when the evidence has been computed to a pre-specified precision. The evidence that could be contributed by the remaining live points is estimated as $\Delta\mathcal{Z}_{\mathrm{i}} = \mathcal{L}_{\mathrm{max}} X_{\mathrm{i}}$, where $\mathcal{L}_{\mathrm{max}}$ is the maximum-likelihood value of the remaining live points, and X_i is the remaining prior volume. The algorithm terminates when $\Delta\mathcal{Z}_{\mathrm{i}}$ is less than a user-defnied value (we use 0.5 in log-evidence).

Once the evidence $\mathcal{Z}$ is found, posterior inferences can be easily generated using the final live points and the full sequence of discarded points from the nested sampling process, i.e., the points with the lowest likelihood value at each iteration i of the algorithm. Each such point is simply assigned the probability weight

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}. \tag{1.27}$$

These samples can then be used to calculate inferences of posterior parameters such as means, standard deviations, covariances and so on, or to construct marginalised posterior distributions.

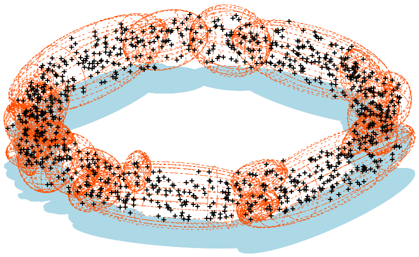
By providing both posterior samples and the value of the evidence, nested sampling can greatly reduce the number of points at which we need to evaluate the likelihood. This will provide us with what is necessary for both parameter estimation and model selection in less time than traditional MCMC methods.

1.2.2.1 The MultiNest Algorithm

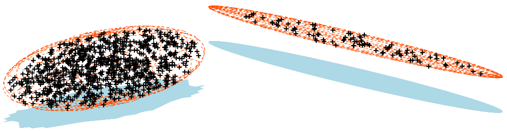
The most challenging task in implementing nested sampling is to draw samples from the prior within the hard constraint $\mathcal{L} > \mathcal{L}_i$ at each iteration i. The MultiNest algorithm [101, 102] tackles this problem through an ellipsoidal rejection sampling scheme. The live point set is enclosed within a set of (possibly overlapping) ellipsoids and a new point is then drawn uniformly from the region enclosed by these ellipsoids. The ellipsoidal decomposition of the live point set is chosen to minimize the sum of volumes of the ellipsoids. The ellipsoidal decomposition is well suited to dealing with posteriors that have curving degeneracies, and allows mode identification in multi-modal posteriors, as shown in Figure 1.6. If there are subsets of the ellipsoid set that do not overlap with the remaining ellipsoids, these are identified as a distinct mode and subsequently evolved independently. The MultiNest algorithm has proven very useful for tackling inference problems in cosmology and particle physics [103–106], typically showing two orders of magnitude improvement in efficiency over conventional techniques. More recently, it has been shown to perform well as a search tool for gravitational wave data analysis [107]. A different implementation of nested sampling for LIGO analysis is described in [108].

1.3 Thesis Outline

In Chapter 2 I describe applications of Bayesian inference to LIGO observations of binary inspirals. I begin with a simple example and proof-of-principle, followed by a systematic analysis of the ability of a LIGO-Virgo network to localise the position of a GW source on the sky. These are both performed with MultiNest for Bayesian inference. I then consider using coherent versus incoherent model selection to distinguish a real GW signal from glitches in actual LIGO data.



(a)



(b)

Figure 1.6: Examples of MULTINEST enclosing live points in sets of ellipsoids. 1000 points have been randomly sampled from (a) a torus and (b) two non-intersecting ellipsoids. Images from [102].

Chapter 3 describes Bayesian inference for LISA/NGO sources using MLDC simulations. The first section covers parameter estimation and model selection for burst sources from cusps on cosmic strings. I then analyse continuous gravitational wave sources for LISA, where there are many overlapping signals present in the data to the point of presenting foreground confusion noise at lower frequencies. Finally, I consider detecting burst sources in the presence of the foreground of continuous GW sources.

At this point I turn to looking at data analysis techniques in a more general sense. Chapter 4 analyses the benefits and drawbacks of the Multiple Optimised Parameter Estimation and Data compression (MOPED) algorithm. I consider cases where it works and where it fails and attempt to set a prescription for determining the viability of a MOPED analysis.

Artificial neural networks are introduced in Chapter 5 as a tool for learning and performing regression and classification functions. I detail the structure of neural networks and the algorithm used for training them on prepared data sets. Several examples are then provided to illustrate the capabilities of the networks and the training algorithm. These include the learning of analytic functions, classification of complicated data, and reduction of the dimensionality of data.

In order to further improve tools for general Bayesian inference, neural networks are integrated into the MULTINEST algorithm to learn the structure of the likelihood function. They can then be used to make future predictions of the likelihood at new points. This is especially beneficial for problems where likelihood evaluations are particularly time-consuming. In addition, the networks may be used for follow-up analyses that can now be performed up to a hundred thousand times faster. The new Blind Accelerated Multimodal Bayesian Inference (BAMBI) algorithm is demonstrated on a few toy problems and then applied to the task of cosmological parameter estimation, where significant increases in speed are demonstrated for initial analyses and exceptional speed-ups are achieved in follow-up.

Chapter 2

Analysis of LIGO Data

> In theory, there is no difference between theory and practice. But, in practice, there is.
>
> Jan L. A. van de Snepscheut

In the next few years the Advanced LIGO and Advanced Virgo detectors will be coming online and collecting science data. During their observation period, one or more compact binary coalescence events are expected to be detected [71, 109]. These signals are well-modeled by post-Newtonian theory [6–8, 40], the effective-one-body approach [19–22], and numerical relativity [24–27], giving us a variety of waveform models to use to search the data. Additionally, the sensitivity of the LIGO and Virgo detectors is designed to be optimal for these sources. In this chapter I begin by introducing a simple proof-of-principle for detecting the gravitational waveform from one of these events using Bayesian inference techniques. I then present work done as part of the LIGO Scientific Collaboration (LSC) to quantify the ability of Bayesian inference to determine the sky location of a detected event for use in electromagnetic follow-up searches. This was done in part with the Parameter Estimation sub-group of the Compact Binary Coalescence group of the LSC and will form part of a publication currently in preparation; code from the LSC Algorithm Library (LAL [110]) was used in generating and analysing this data. I conclude with my contribution to the analysis of a hardware blind injection into LIGO's sixth and Virgo's second science runs [68]. The analysis in this section was presented as a poster at the Ninth Edoardo Amaldi Conference on gravitational waves [111].

2.1 Simple Inspiral Model

Many stars form as part of a binary system [112]. Over their lifetimes, these stars will slowly spiral into each other as they lose energy from the emission of gravitational waves. These predictions have been verified by the discovery of binary pulsar PSR B1913+16 in 1974 by Hulse and Taylor and subsequent observations [2]. As the two stars spiral in on each other, their GW signal produces a "chirp", characterised by increasing amplitude and frequency until a specific frequency of the last stable orbit, after which they go through merger and ringdown phases.

2.1.1 Generating the Signal

The sensitivity of the LIGO observatory is best for the inspiral phase of a compact binary coalescence for objects with sizes on the order of solar masses, so for this reason and for computational simplicity we consider only the inspiral. We assume the two black holes are nonspinning and have masses m_1 and m_2. We therefore define the total mass, $m = m_1 + m_2$, symmetric mass ratio, $\eta = (m_1 m_2)/(m_1 + m_2)^2$, and chirp mass, $\mathcal{M} = (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5} = m\eta^{3/5}$. Following the approach of Veitch and Vecchio [113], we take the leading Newtonian quadrupole order of the strain in the frequency domain. The GW signal can thus be expressed as

$$\tilde{h}(f;\vec{\theta}) = \begin{cases} A\mathcal{M}^{5/6} f^{-7/6} e^{\imath\psi(f;\vec{\theta})} & f \leq f_{LSO} \\ 0 & f > f_{LSO} \end{cases}, \tag{2.1}$$

where

$$\psi(f;\vec{\theta}) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{4}(8\pi\mathcal{M}f)^{-5/3} \tag{2.2}$$

is the signal phase and f_{LSO} is the frequency of the last stable orbit in the Schwarzschild spacetime. For a general system this is given by

$$f_{LSO} = (6^{3/2} 2^{-1} \pi m)^{-1}, \tag{2.3}$$

but for simplicity we assume $m_1 = m_2$ so we can re-write f_{LSO} as

$$f_{LSO} = (6^{3/2} 2^{1/5} \pi \mathcal{M})^{-1}. \tag{2.4}$$

In equations (2.1) and (2.2), $\vec{\theta}$ refers to the 4-dimensional parameter vector, where our parameters are

$$\vec{\theta} = \{A, \mathcal{M}, t_c, \phi_c\}. \tag{2.5}$$

A is the overall signal amplitude and depends on the source masses, distance, sky location, inclination angle, and polarisation; t_c and ϕ_c are the time and phase at coalescence, respectively. In this analysis, geometrical units are being used, so $c = G = 1$ and therefore $M_\odot = 4.926 \times 10^{-6}$ s and A has the units of s^{-2}. It should be noted that only the chirp mass, $\mathcal{M}$, is being used in this formula. Therefore, a degeneracy exists between the two black hole masses—any two masses that produce the same chirp mass will give the same results with this formula. A more accurate model of the signal that also uses the symmetric mass ratio, or any other function of the masses, would be able to break this degeneracy.

The detector noise profile in the frequency domain is modelled as a Gaussian random process in the real and imaginary parts, with zero mean and a variance at frequency f_k ($k = 1, \ldots, K$, where K is the number of frequency bins) given by

$$\sigma_k^2 = \left(\frac{f_k}{f_0}\right)^{-4} + 1 + \left(\frac{f_k}{f_0}\right)^2. \tag{2.6}$$

This is representative, within a multiplicative constant, of the LIGO interferometers with $f_0 = 150$ Hz setting the scale for the frequency of maximum sensitivity.

2.1.2 Calculating the Likelihood

We define our likelihood function to be a multi-variate Gaussian, centred on the true parameter values:

$$\mathcal{L}(\vec{\theta}) = \prod_{k=1}^{K} \left[\frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{|\tilde{h}_k(\vec{\theta}) - \tilde{d}_k|^2}{2\sigma_k^2} \right) \right]. \tag{2.7}$$

To simplify this, we compute the log-likelihood and subtract out any constants. This will give us an evidence value that is off by a constant factor, but all local evidences will have this factor which will be cancelled out in the evidence ratio. The log-likelihood is therefore

$$\ln \mathcal{L}(\vec{\theta}) = \sum_{k=1}^{K} \left(-\frac{|\tilde{h}_k(\vec{\theta}) - \tilde{d}_k|^2}{2\sigma_k^2} \right). \tag{2.8}$$

Parameter	Measured	Actual
$A\,(\mathrm{s}^{-1}/M_\odot)$	4.809 ± 0.328	5.00
$\mathcal{M}\,(M_\odot)$	8.000 ± 0.015	8.009
$t_c\,(\mathrm{s})$	$20.0 \pm 0.21\ \times 10^{-3}$	20.0
$\phi_c\,(\mathrm{rad})$	1.262 ± 0.280	1.00

Table 2.1: Results from the MULTINEST parameter inference on a simple inspiral chirp signal.

In order to run the nested sampling algorithm MULTINEST [101, 102] on each data set we need to define the prior volume for it to explore. In order to have the most generality, we will use a uniform prior range defined over the following intervals:

$$A \in [0,1000]\ \mathrm{s}^{-1}/M_\odot, \tag{2.9a}$$

$$\mathcal{M} \in [1,20]M_\odot, \tag{2.9b}$$

$$t_c \in [19.5,20.5]\ \mathrm{s}, \tag{2.9c}$$

$$\phi_c \in [0,2\pi)\ \mathrm{rad}. \tag{2.9d}$$

This is a significantly wider prior range than that used by Veitch and Vecchio. Their prior ranges were $A \in [0,10]\,\mathrm{s}^{-1}/M_\odot$, $\mathcal{M} \in [7.7,8.3]M_\odot$, $t_c \in [19.9,20.1]\,\mathrm{s}$, and $\phi_c \in [0,2\pi)\,\mathrm{rad}$.

2.1.3 MULTINEST Results

MULTINEST was used to analyse the simulated data produced by the model described in Section 2.1.1 with parameters of $\vec{\theta} = \{5\mathrm{s}^{-1}/M_\odot, 8.009M_\odot, 20, 1\}$. The results of the run are given in Table 2.1 and the final marginalized posterior distributions can be seen in Figure 2.1. These results show that MULTINEST was able to estimate all four parameter values very closely.

In Figure 2.2 is plotted the log-likelihood function as it varies over the amplitude and chirp mass for the correct values of t_c and ϕ_c. We can clearly see a ridge at constant $\mathcal{M}$ and a background drop-off for increasing A and $\mathcal{M}$. In order to ensure sampling from this ridge, whose peak coincides with the true parameters, an increased number of live points for MULTINEST need to be used. This had the expected side effect of increasing

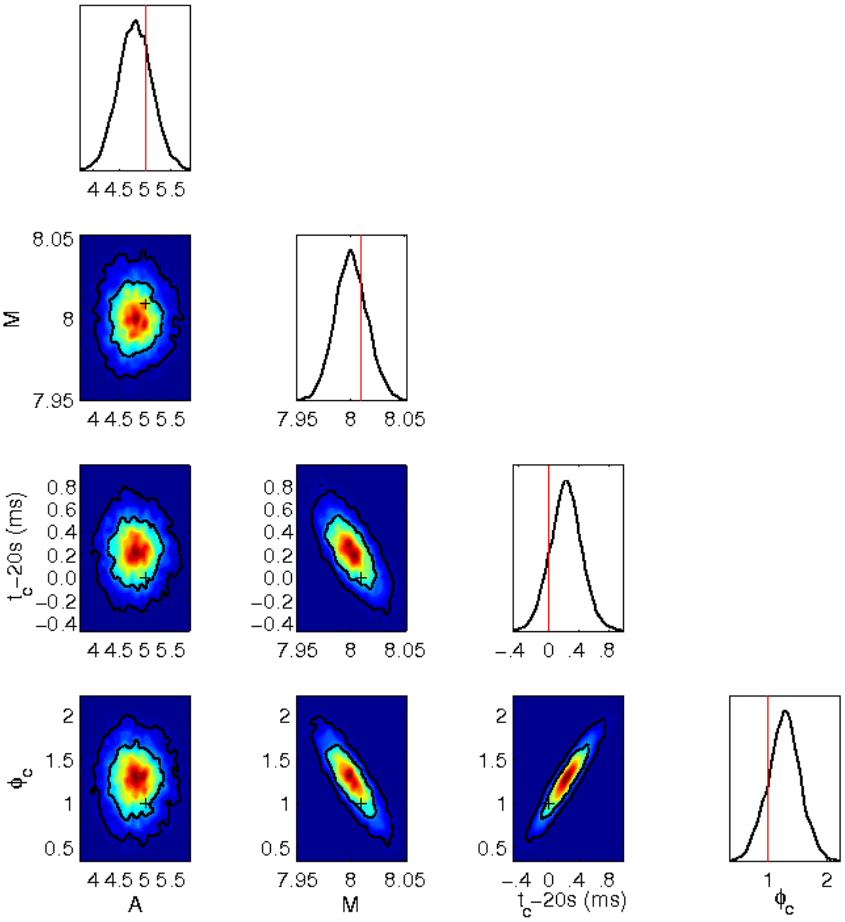


Figure 2.1: Plots of the one-and two-dimensional marginalized posterior distributions obtained by MULTINEST. Veritcal red lines and plus signs indicate true input values. $[A] = \mathrm{s}^{-1}/M_{\odot}$, $[\mathcal{M}] = M_{\odot}$ and $[\phi_c]$ =rad.

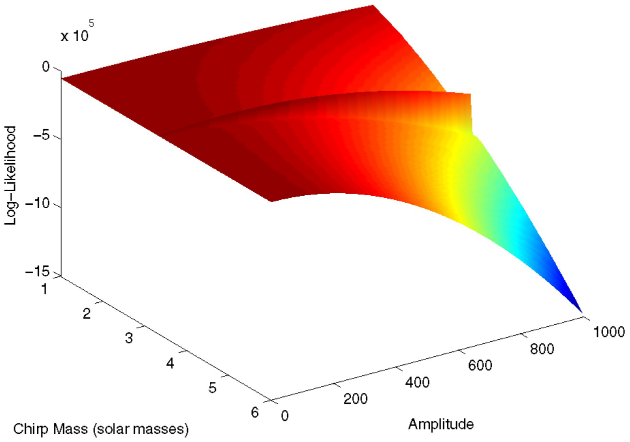


Figure 2.2: The log-likelihood function for a given range of amplitude and chirp mass values. The actual values of signal are $A = 300\mathrm{s}^{-1}/M_{\odot}$, $\mathcal{M} = 3.52M_{\odot}$, $t_c = 20\,\mathrm{s}$, and $\phi_c = 0\,\mathrm{rad}$. t_c and ϕ_c have been set to their actual values for this calculation.

the computation time. On a single Linux machine, MULTINEST still took less than an hour to run, even for signals with low SNRs.

If we were to look at the log-likelihood for the prior ranges of [113], the large background feature in A and $\mathcal{M}$ is not significant, as can be seen in Figure 2.3. The range of chirp masses they explore limits the effect of this background effect and also causes the ridge to occupy a larger fraction of the sampled space. This makes the likelihood easier to sample, but limits the detectability to signals from binaries whose parameters fall within a narrow range.

We can now measure how well MULTINEST can find the true signal embedded in noise. A natural measure of this is the Bayes factor, the ratio of the posterior odds of the signal and noise-only models. We assume a prior odds ratio of one so the Bayes factor is calculated simply by dividing the evidence of the signal model as measured by MULTINEST and the evidence of the noise-only model. A Bayes factor greater than one indicates that the signal model is more probable than the noise model; we ask for a slightly larger Bayes factor to declare a detection. As the Bayes factor increases, the confidence of our detection grows as well. To measure how detection is affected by weaker signals, I computed the Bayes factor for several signal-to-noise ratios (SNRs).

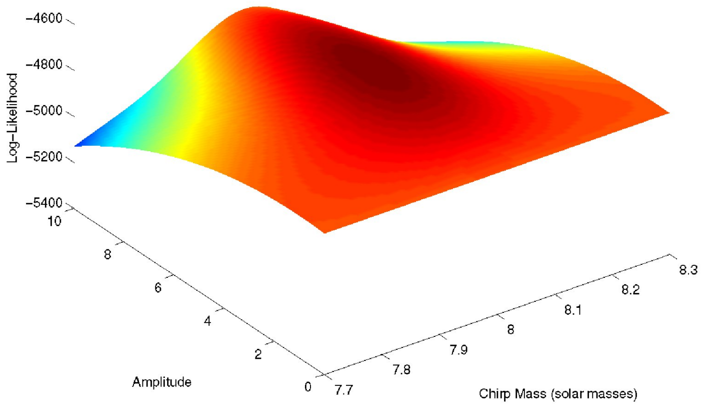


Figure 2.3: The log-likelihood function for data with an input signal of $\vec{\theta} = \{5/M_\odot, 8M_\odot, 20, 0\}$, where $t_c = 20$ s and $\phi_c = 0$ rad have been set.

The SNR, specified by ρ, provides a measure of the relative strength of the signal to the background noise and is calculated using the following formula:

$$\rho = \sqrt{\sum_{k=1}^{N} \frac{|\tilde{h}_k(\vec{\theta})|^2}{\sigma_k^2}}. \tag{2.10}$$

Varying values of the SNR were obtained by changing the amplitude of the signal. The chirp mass (total mass and symmetric mass ratio), time of coalescence, and phase of coalescence were all kept equal so as to minimize any effects other than varying signal strength. The global evidence of the signal model was given by MULTINEST at the end of each run, which were performed using the original log-likelihood function and 1000 live points to ensure a thorough sampling of the prior volume. The null evidence (that of the noise model) was computed by fixing the amplitude to zero. Figure 2.4 shows the log Bayes factor as a function of the measured SNR over a range of SNRs. We can see that once a certain threshold has been reached by an SNR of 10, we can be very confident in our detection of a signal. Below an SNR of ~ 6, however, we cannot be sure of a signal being present.

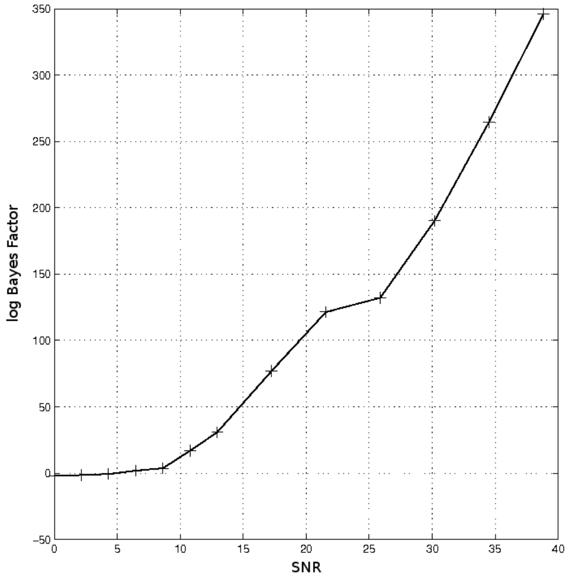


Figure 2.4: The log Bayes factor as a function of SNR. Past an SNR of ~ 10, our confidence of detection increases exponentially.

This clearly shows that Bayesian inference as performed by MULTINEST is a strong tool for detecting gravitational wave signals in instrument noise. Being able to make detections at SNRs below 10 is important as the current instruments do not expect to have many SNRs above that if a signal is in fact present. For comparison, see Veitch and Vecchio's original work on this same problem in [113, 114], which shows similar results for Bayesian evidences to those obtained here, and their use of delayed rejection in [115, 116]. MULTINEST has also been used to detect the fully parameterised waveform from non-spinning supermassive black holes in [107]. The usefulness of inspiral-only templates for detection of full inspiral-merger-ringdown waveforms is examined in [117]. The use of MULTINEST for the detection of galaxy clusters with a fully Bayesian analysis can be found in [118].

2.2 LIGO Sky Localisation

In order to confirm the first detections of gravitational waves and extract the most science from an observed event, we will want to obtain follow-up observations from electromagnetic telescopes [119–126]. For certain events, these will be able to confirm the event occurring as well as provide additional information. For binary black hole mergers, the lack of an observed signal will confirm theory and the detection of a signal could indicate the presence of an accretion disk around one or both of the black holes. To this end, a project was begun in the Bayesian parameter estimation subgroup of the CBC group in the LSC to determine our ability to localise the position of a source on the sky and measure how fast we would be able to accomplish this. As this was also a validation test of the Bayesian codes, we elected to use a simpler model for the inspiral signals and analytically generate the noise according to a known power spectral density (PSD).

Studies have already been performed to analyse the ability of GW detector networks to localise a source on the sky [127–134], but this is the first systematic Bayesian study that utilises coherent detection and compares many inference algorithms to confirm results and measure speed. The full paper presenting results from the Parameter Estimation group is in preparation. Other studies, such as [135], have measured statistical uncertainties in measured parameters with the advanced detector network, but use the Fisher matrix approximation that cannot capture the full nature of posterior degeneracies and does not apply at low SNRs. There have also been efforts to characterise the ability of a detector network to measure other source parameters [136].

2.2.1 The Injections

The injections were performed with the restricted frequency-domain post-Newtonian waveform model, TaylorF2 [10], at second order in phase (corrections of $O(v^4)$) and with zeroth order (Newtonian) amplitude. No spin was included in the models in order to simplify the analysis so there were 9 total parameters: chirp mass, symmetric mass ratio, luminosity distance, inclination of the orbital plane (angular momentum) to the line of sight, polarisation, phase at coalescence, time of coalescence at the geocentre, longitude (related to right ascension), and colatitude (related to declination). The last two of these are the position on the sky of the source. The parameter vector is thus given by $\vec{\theta} = \{\mathcal{M}, \eta, d_L, \iota, \psi, \phi_0, t_c, \phi_s, \theta_s\}$. ϕ_s and θ_s are related to the right ascension α and declination δ by $\phi_s = \alpha - \mathrm{GAST}$ (GAST is the Greenwich Apparent Sidereal

Time) and $\theta_s = \pi/2 - \delta$. The waveform is given in the source frame by [10]:

$$\tilde{h}_+(f) = \frac{\mathcal{M}^{5/6}}{d_L} f^{-7/6}(1+\cos^2(\iota))\cos(\Phi(f)) \tag{2.11a}$$

$$\tilde{h}_\times(f) = -\frac{\mathcal{M}^{5/6}}{d_L} f^{-7/6}\cos(\iota)\sin(\Phi(f)) \tag{2.11b}$$

$$\Phi(f) = 2\pi f t_c - \phi_0 - \frac{\pi}{4} + \frac{3}{128\eta v^5} \times \left[1 + \frac{20}{9}\left(\frac{743}{336} + \frac{11}{4}\eta\right)v^2 - 16\pi v^3 + 10\left(\frac{3058673}{1016064} + \frac{5429}{1008}\eta + \frac{617}{144}\eta^2\right)v^4\right] \tag{2.11c}$$

where $v = (\pi \mathcal{M} f)^{1/3}$. The waveform is evaluated for frequencies up to the last stable orbit given by Equation (2.3).

This is measured by each of the LIGO and Virgo detectors with sensitivity patterns given by:

$$F_+(\psi, \phi_s, \theta_s) = -\frac{1}{2}(1+\cos^2(\theta_s))\cos(2\phi_s)\cos(2\psi) - \cos(\theta_s)\sin(2\phi_s)\sin(2\psi), \tag{2.12}$$

$$F_\times(\psi, \phi_s, \theta_s) = \frac{1}{2}(1+\cos^2(\theta_s))\cos(2\phi_s)\sin(2\psi) - \cos(\theta_s)\sin(2\phi_s)\cos(2\psi), \tag{2.13}$$

such that the signal measured is given by

$$\tilde{h} = \tilde{h}_+ F_+ + \tilde{h}_\times F_\times . \tag{2.14}$$

We analyse 200 injections chosen uniformly from the distributions given in Table 2.2. Notice that the masses are sampled by m_1 and m_2 instead of $\mathcal{M}$ and η – this is to allow equal-mass binaries more easily. The additional constraints of $m_1 > m_2$ and $m_1 + m_2 \leq 20M_\odot$ are also applied. t_T is the pre-determined and supplied trigger time of the signal that we use as the basis for a search interval. These same ranges are also used as the priors for the subsequent data analysis.

The noise PSD was generated in the frequency domain according to the analytic formula given below with $f_0 = 150$Hz.

$$S_h^2(f) = 9 \times 10^{-46}\left(\left(4.49\frac{f}{f_0}\right)^{-56} + 0.16\left(\frac{f}{f_0}\right)^{-4.52} + 0.52 + 0.32\left(\frac{f}{f_0}\right)^2\right). \tag{2.15}$$

Parameter	Units	Minimum	Maximum
m_1	$M_\odot$	1	15
m_2	$M_\odot$	1	15
$\log(d_L)$	Mpc	10	40
$\cos(\iota)$	–	-1	1
ψ	rad	0	π
ϕ_0	rad	0	2π
t_c	sec	$t_T - 0.1$	$t_T + 0.1$
ϕ_s	rad	0	2π
$\cos(\theta_s)$	–	-1	1

Table 2.2: Ranges for the sampling distributions of signal parameters. These are used for both signal generation and prior probabilities.

This same function was used in the analysis for the PSD. Although this does not mirror the non-stationarity and non-Gaussianities in real LIGO noise, it does allow us to confirm the results of our analysis using statistical measures.

2.2.2 Analysis and Results

The 200 injections were analysed with MULTINEST on LIGO's Caltech cluster, from which we were able to obtain converged results on 187 (the remaining 13 were excluded due to computational issues that kept preventing analysis jobs from finishing). 5000 live points were used to ensure complete sampling of the prior. Run times on a single processor varied from approximately 2 days to 2 weeks depending on the number of likelihood evaluations necessary. The loudest and quietest signals had the lowest sampling efficiency due to sharp peaks or broad degenerate modes in the likelihood function, respectively, and so took the longest to converge.

A typical posterior distribution is shown in Figure 2.5 for an event with a total network SNR of 19.6. Strong correlations can be seen between certain parameters, particularly between distance and inclination and between different pairs of mass parameters. We can also see that the phase at coalescence is not well constrained and that there is a multimodal degeneracy in the polarisation of the GW signal. It is important to fully explore these modes and degeneracies as the true signal parameters can be located anywhere within them.

Focusing on the intrinsic parameters of the system, which for this simple model is limited to $\mathcal{M}$ (m_c) and η, we see in Figure 2.6 a degeneracy but accurate recovery of the true parameters within the 68% confidence interval. For the sky location, which is what we need to measure for electromagnetic follow-up, there is again a degeneracy but the location has been measured to within a fraction of a steradian and the true value recovered just outside the 68% confidence interval. The 68% interval contains less than 50 deg^2 and the 95% interval less than 100 deg^2; the posterior as mapped on the entire sky can be seen in Figure 2.7.

For our first test that posteriors were recovered correctly, we look at the fraction of injections recovered at a given level of each parameter's integrated one-dimensional posterior probability. We integrate the posteriors from the minimum to the maximum of each parameter and record the cumulative distribution function (CDF) value at the point of the true injected value. We expect the distribution of these values to map directly to the posterior distribution. If we therefore plot the fraction of injections recovered below a given CDF value against CDF values, the distribution should follow a line corresponding to $y = x$ from 0 to 1. We are essentially comparing the CDF of the injections with the CDF of the measured posteriors to confirm they are equal. In Figure 2.8 we show this for all 9 model parameters as well as m_1 and m_2, which our prior is defined on. We find that the 1D CDFs all match the expected distributions very closely, usually to within 2σ error. Some deviations are seen, but these are most exaggerated for poorly constrained parameters and correlations between parameters can cause these errors to affect other measurements. To quantitatively measure the differences in experimental and theoretical CDFs, we performed a Kolmogorov-Smirnov (K-S) test on each. This test measures the maximum difference in CDFs to decide the probability that the data supports the null hypothesis that both CDFs are from the same distribution. A p-value less than the threshold of 0.05 indicates that we reject the null hypothesis at 95% confidence. All parameters return p-values above 0.05 as indicated in the subplot titles. This is a good indication that the recovered posterior distributions do accurately represent our state of knowledge.

A more thorough check involves measuring the posterior confidence level at which the injection was recovered in sky position. This interval is is calculated by measuring

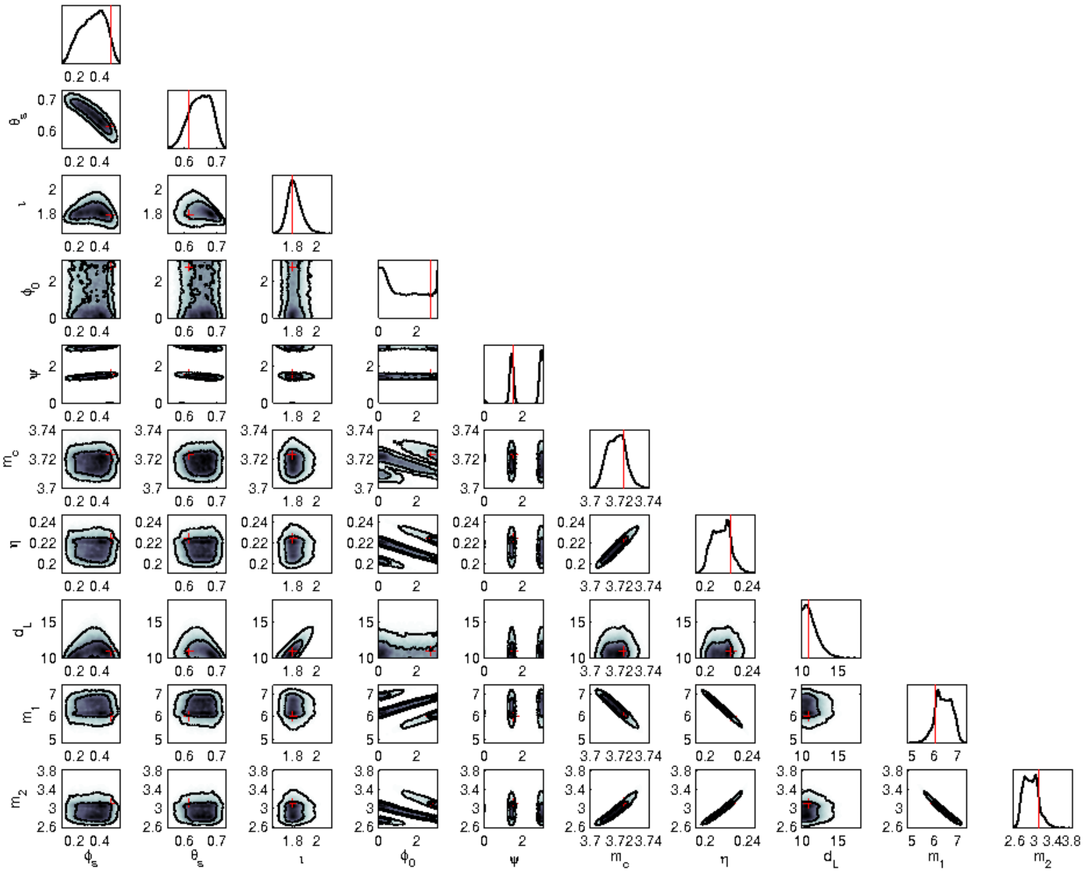
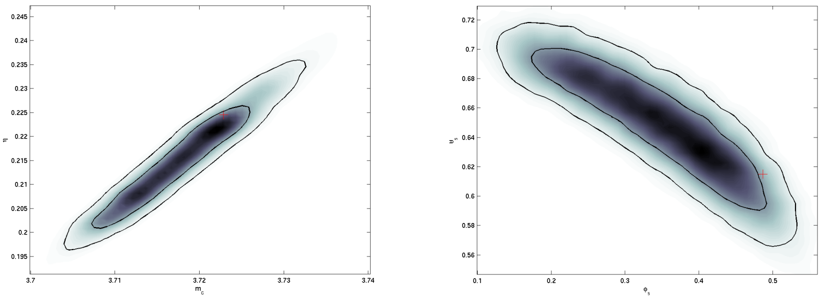


Figure 2.5: Marginalised 2D and 1D posterior distributions for event 22 (all parameters except t_c). This event has an SNR of 9.0 in H1, 13.4 in L1, and 11.1 in V1 for a total network SNR of 19.6. Red vertical lines and crosses indicate the true values. Contours on 2D plots indicate 68% and 95% confidence intervals.



(a) (b)

Figure 2.6: Marginalised 2D posterior distributions for event 22 in (a) chirp mass and symmetric mass ratio and (b) sky position. The red cross indicates the true value. Contours indicate 68% and 95% confidence intervals.

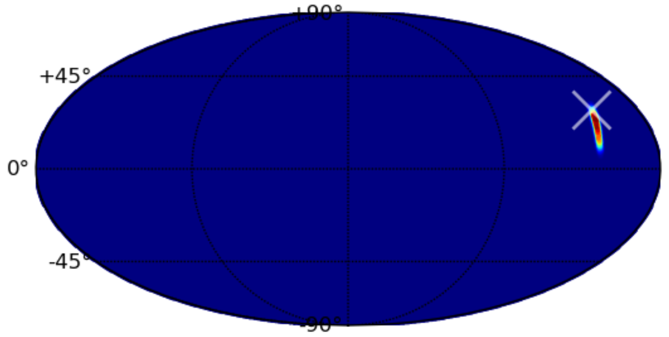


Figure 2.7: The marginalised posterior probability distribution for the sky position of event 22 as shown across the entire sky. The white × marks the true location.

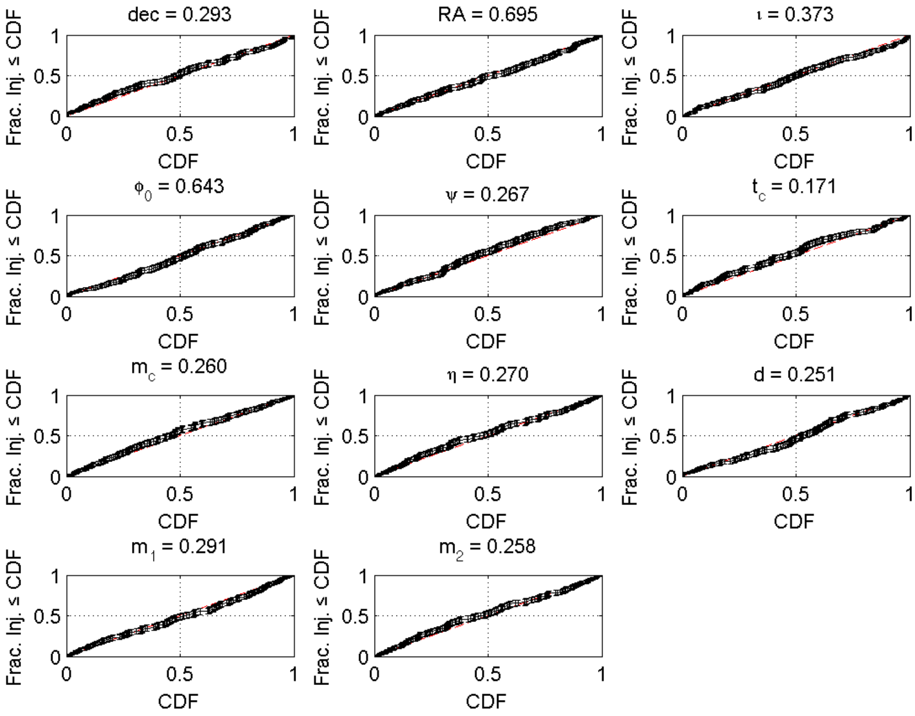


Figure 2.8: The fraction of injections recovered at or below 1D posterior CDF values. Error bars are given by $\sqrt{p(1-p)/N}$, where N is the number of injections. K-S test p-values are given in the titles for each subplot. All parameters pass the test for a threshold of 0.05.

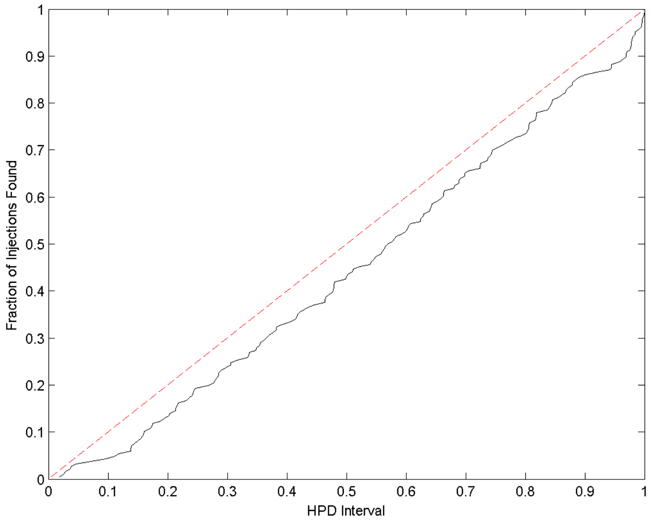


Figure 2.9: The fraction of injections recovered within high probability density intervals of the posterior as computed over θ_s and ϕ_s. The sag is due to posterior biases and K-D tree binning.

the accumulated posterior probability in regions with higher probability density than where the injection is located. We perform a K-D tree [137] binning on the two sky location variables and sum the posterior density for 'leaves' with higher density than where the injection is found. As in the 1D case, we expect to find $X\%$ of injections within the $X\%$ high probability density (HPD) region of the posterior. Figure 2.9 displays our measured recovery compared to the theoretical line. There is a degree of 'sag' below the line, but this is an expected artifact from the process of K-D tree binning and Poisson error on the number of samples in a certain region.

A final validation of the recovered posterior values involves direct measurement of the full nine-dimensional HPD interval at which each injection is recovered. We do this by calculating the log-prior at each posterior sample to add to the log-likelihood and obtain an non-normalised log-posterior density value. The same value is computed for the true parameters of the injection. We then count the fraction of posterior samples that have a higher density than the true parameters and this equals the confidence interval at which the injection was recovered. We expect the fraction of injections re-

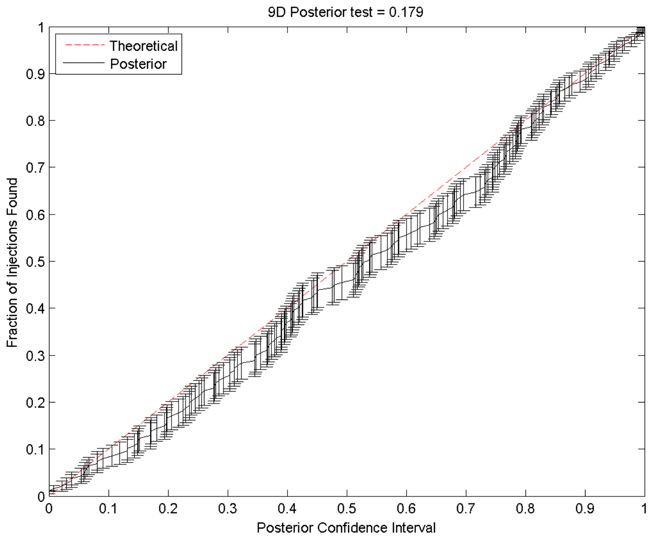


Figure 2.10: The fraction of injections recovered within high probability density intervals of the posterior as computed from posterior probability density values directly. Error bars are given by $\sqrt{p(1-p)/N}$, where N is the number of injections. The sag is due to accumulation of biases seen in the 1D cases (Figure 2.8). The K-S test p-value is 0.179.

covered within a given HPD to follow $y=x$ as in Figure 2.8 if the recovered posteriors are, in fact, representative of the correct priors and likelihood. This comparison is shown in Figure 2.10 along with 1σ error bars. The recovery is consistent to within 2σ of the theoretical value for all confidence intervals, but there is a systematic bias towards lower values. This is likely due to the biases seen in the one-dimensional CDFs in Figure 2.8 accumulating to push the peak away from the true value beyond the expected statistical variation. The K-S test results in a p-value of 0.179, so the null hypothesis that the CDFs are equal is not rejected.

Having passed these tests of the posterior distribution, we now look at the statistics for the areas within certain intervals. Our first consideration is the area contained within a particular posterior high probability density interval. Using the K-D tree binning, we calculate the sky area within 68%, 90%, and 95% confidence intervals. In Figure 2.11 we plot the fraction of injections whose confidence intervals contain a

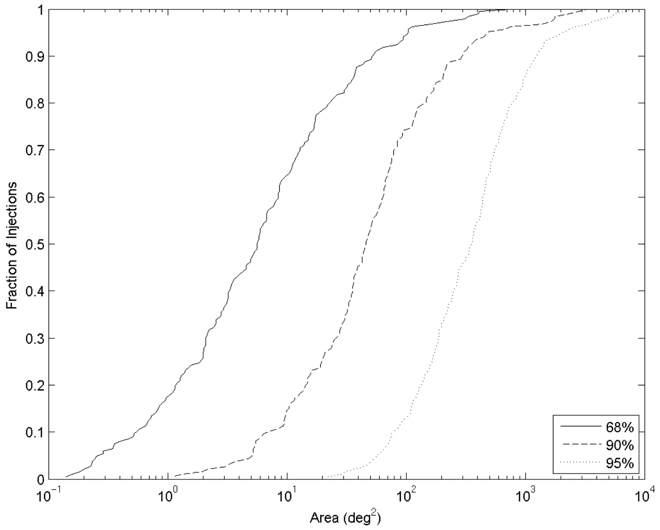


Figure 2.11: The fraction of injections whose HPD posterior confidence intervals contain the amount of sky area or less. Areas are calculated using K-D tree binning.

certain amout of sky area or less. This allows us to say that if we can search over a certain amount of sky area, say 100 deg^2, then $\sim$ 80% of all signals will have their 90% confidence intervals contained within that search area. We therefore expect to place constraints on $\sim$ 72% of signals. We can also determine that our confidence regions increase by approximately an order of magnitude in sky area in going from 68% to 90% and again from 90% to 95%. Therefore, using larger confidence intervals may make the search regions impractically large for electromagnetic follow-up. Posterior distributions tend to be ellipses in right ascension and declination, making them slightly curved on the sky when extending further from the peak.

The next consideration is the fraction of injections that were recovered within a given sky area. For this, we use the K-D tree binning to determine the sky area contained within bins with posterior density equal to or greater than that for the bin in which the true point is located. This comparison is shown in Figure 2.12 and allows us to determine the amount of sky area we need to be able to search to find counterparts for a given fraction of signals or conversely, the fraction of signals we can expect to find counterparts for in a given search area. For example, if we can only search over an

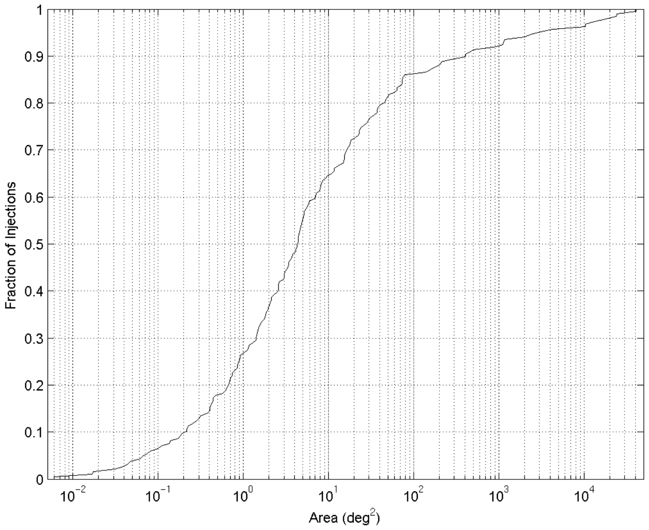


Figure 2.12: The fraction of injections that were recovered within the amount of sky area or less. Areas are calculated using K-D tree binning.

area of 4 deg^2, then we can expect to locate counterparts (where they exist) for 50% of signals; however, if we wish to identify counterparts for 90% of signals then we will need to search 300 deg^2 of sky. With this information we can optimise our follow-up search strategies based on either the sky area they can map or the fraction of signals we wish to recover.

The first two analyses were averaged over all injections. However, we expect to be able to localise signals with higher SNR to smaller sky areas. Therefore, the final comparison considers the sky area as a function of signal-to-noise ratio. Figure 2.13 shows the sky area contained within the 90% posterior confidence interval as a function of SNR. It also shows the area within the interval at which each signal was recovered, but this measurement is subject to statistical variations in the relative positions of the true location and the posterior peak due to the particular noise realisation. Overall, we can see in both sets of points that increased SNR does result in smaller sky areas, with the 90% interval area scaling roughly $\propto 1/\mathrm{SNR}^2$. For SNRs above 25, all 90% confidence regions contain 100 deg^2 or less, but for all signals with SNR below 20 this same interval contains at least 20 deg^2. We see then that in order to find electro-

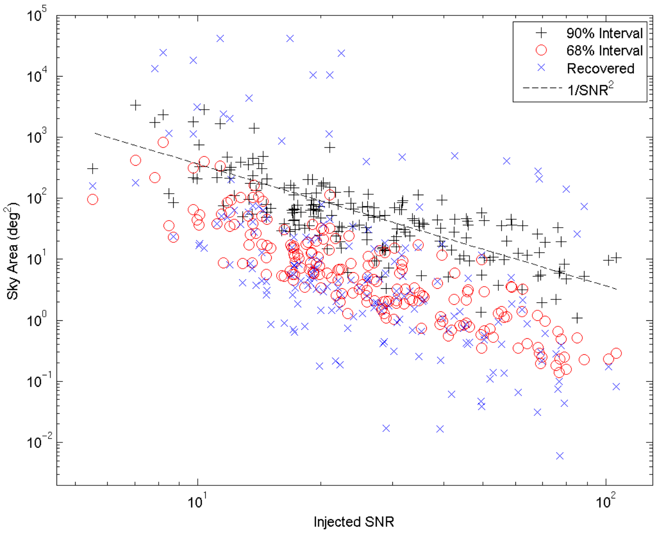


Figure 2.13: The sky area within 90% posterior confidence intervals and the posterior interval at which an injected signal was recovered as a function of signal-to-noise ratio. Areas are calculated using K-D tree binning. A line showing $A \propto 1/\mathrm{SNR}^2$ is plotted for comparison.

magnetic counterparts to weaker signals, we will have to be able to search larger areas of the sky. Even with very loud signals, 90% confidence intervals still contain over a square degree, which is a large area for electromagnetic telescopes.

Since our ability to localise sources is fundamentally limited to areas of $O(10)$ deg^2 for 90% confidence intervals, optimised methods for electromagnetic follow-up, such as those explored in [138], are necessary to increase the chances of observing a counterpart. Combined detections of both GWs and electromagnetic signals (and possibly neutrinos) will further confirm source models for GW observations and lead to improved science. It may be possible to discover more about the internal physics of supernovae, neutron star mergers that form black holes, and much more.

2.3 The “Big Dog” Event and Coherence Tests

Throughout 2009 and 2010, LIGO and Virgo were collecting data in their sixth and second/third science runs, respectively (called S6 and VSR2/3). One of the primary sources that these detectors could observe and that was searched for in the data is low-mass compact binary coalescences. These involve binaries with a total mass less than $25M_{\odot}$ and a minimum component mass of $1M_{\odot}$. The CBC group of the LVC has published a report on the results of this search [68, 139], which made no detections of gravitational waves. A complementary high-mass search was also performed [140]. A similar analysis was previously performed on the joint observations during S5 and VSR1 [141–144]. A review of knowledge of event source rates following S5/VSR1 is presented in [145].

During S6 and VSR2/3, a hardware injection was made into the detectors (by actuating test mass mirror controls) without the knowledge of the data analysis teams as part of a “blind injection challenge”. The task was for the detection pipelines to find the event and then for the analysis groups to correctly identify and characterise the signal that was injected. The event came to be known as the “big dog” since initial sky location estimates placed the source in the Canis Major constellation. The analysis in this section was presented as a poster at the Ninth Edoardo Amaldi Conference on gravitational waves [111] and used data from the S6 and VSR2 science runs.

One of the major considerations for determining the existence of a real signal is the possibility of false coincidences. Each of the detectors registered an event trigger at a similar time, but this could also potentially happen if they each glitched at approximately the same time. We are able to rule out coincident glitches by performing a coherent search and comparing the evidence to that of an incoherent search. In the case of just the LIGO detectors, this can be summarised as:

$$\begin{array}{rcl} \text{coherent} & \Longrightarrow & \int \mathcal{L}_H \mathcal{L}_L d\theta = \mathcal{Z}_{HL} \\ \text{incoherent} & \Longrightarrow & \int \mathcal{L}_H d\theta_1 \times \int \mathcal{L}_L d\theta_2 = \mathcal{Z}_H \mathcal{Z}_L \end{array}$$

If each detector did indeed observe the same signal, then the combined likelihood function will provide a larger evidence $\mathcal{Z}_{HL}$ than the product of the evidences from the separate likelihoods $\mathcal{Z}_H \mathcal{Z}_L$. If in fact this was not a real GW signal, then the two detectors will favour different parameters and the coherent evidence will be severely penalised. We can consider these as two separate models and compare the Bayesian evidences, such that if $\mathcal{Z}_{HL} > \mathcal{Z}_H \mathcal{Z}_L$ we will favour the coherent signal model and if $\mathcal{Z}_{HL} < \mathcal{Z}_H \mathcal{Z}_L$ we favour the incoherent model, which implies no GW signal. A toy example of this application can be found in [146].

A large number of time-slides were analysed in the calculation of the false alarm rate for the "big dog" event. These involve taking data from one of the two LIGO detectors and translating it in time such that there can be no coherent GW signal present. This data is then analysed for potential triggers. Data from Virgo was not used in the analysis that generated the time-slides and the signal in Virgo was of low SNR so V1 data is not used here, either. In total, 14 triggers were found with SNR or "newSNR" (SNR modified by χ^2 of data minus best-fit template) greater than that of the original event; all of these contained the injected signal in H1 and a glitch in L1. Details of the time-slides and newSNR calculation can be found in [68]. These time-slides provide us with an opportunity to test the Bayesian criterion for comparing coherent and incoherent signal models. The 14 time-slide triggers and the "big dog" were each analysed with the TaylorF2 waveform model at 2, 2.5, and 3.5 post-Newtonian order in phase (0pN amplitude). Table 2.3 reports the Bayesian evidence ratios calculated for each trigger using each waveform model.

The significantly negative log-evidence ratios clearly eliminate all time-slide triggers from being coherent signals with probabilities much less than 1%. The most significant time-slide, TS5, contained a weak signal in L1, so was thus able to disagree with the signal in H1 the least. L1 on its own returned a very low evidence that would not likely have passed as a trigger on its own. The cause for low evidence ratios for the "big dog" event is parameter biases from the fact that we are not using the correct waveform model. After the true signal was revealed, we found that the injection was calculated with a time-domain waveform and contained significant spin in the larger component, something the TaylorF2 waveform was not modeling. Even so, the lowest coherent probability was still 37.5%, more than enough to prevent us from eliminating the model as a possibility without further investigation.

This analysis demonstrated the usefulness of Bayesian evidence criteria for model selection and detection for LIGO. When the noise is difficult to model but uncorrelated between separate detectors, comparing the Bayesian evidence from coherent and incoherent signal models can be used as a further detection threshold to eliminate coincident glitches.

Event	$\log(R)$ 2pN	$\log(R)$ 2.5pN	$\log(R)$ 3.5 pN
BD	1.39	-0.509	1.90
TS1	-17.8	-15.4	-15.1
TS2	-11.5	-10.2	-10.1
TS3	-12.7	-14.3	-13.4
TS4	-11.1	-12.5	-10.8
TS5	-2.96	-3.88	-2.65
TS6	-14.2	-14.2	-14.1
TS7	-14.2	-12.0	-11.7
TS8	-8.38	-8.52	-6.59
TS9	-7.74	-9.21	-9.19
TS10	-14.2	-14.7	-14.0
TS11	-15.2	-13.1	-14.2
TS12	-7.39	-8.06	-8.07
TS13	-10.4	-7.53	-9.00
TS14	-10.6	-11.8	-12.3

Table 2.3: Logs of the Bayesian evidence ratios ($R = \mathcal{Z}_{HL}/(\mathcal{Z}_H \mathcal{Z}_L)$) comparing coherent and incoherent signal models for the time-slide triggers (TS1-TS14) and “big dog” event (BD). This analysis clearly rules out all time-slide triggers as signals but is not definitive with regards to the “big dog”.

2.4 Conclusion

Throughout this chapter I have shown the useufulness of Bayesian inference methods for the analysis of ground-based gravitational wave detector data. A first toy example demonstrated the ability of Bayesian techniques to provide accurate parameter estimation as well as a threshold for detection.

An analysis of many simulated signals injected into simulated noise representative of the LIGO-Virgo network further illustrates this applicability to ground-based detector data. We are able to obtain accurate posterior inferences about the source parameters, allowing us to measure intrinsic parameters of the system as well as determine the location of a source on the sky. This latter capability was quantified as it is important for obtaining electromagnetic follow-up observations to supplement the information about the source event. We found that the LIGO-Virgo network will be able to identify 50% of signals to within 4 deg^2 and 90% of signals to within 300 deg^2. Louder signals will be more precisely measured on the sky, with the area in a given posterior confidence interval scaling roughly $\propto 1/\mathrm{SNR}^2$.

Lastly, through analysis of the "big dog" blind injection event in S6 and VSR2, we were able to test a Bayesian criterion for the veracity of a detected signal. By requiring that the signal in distant detectors be from the same source, we compared coherent and incoherent signal models using the Bayesian evidence. Coincident but incoherent signals were simulated by using time-slides of detector data and the evidence ratio clearly indicated that these were all incoherent detections, while supporting the hyposthesis that the injection itself was a real signal.

These two tools of being able to distinguish signals from coincident glitches and measure the probability distribution for source parameters are vital to future LIGO and Virgo observations. Once detection claims are being made it is even more important to be sure that these are in fact real astrophysical signals and not conspiring glitches in the detectors. Additionally, measuring source parameters will allow for tests of general relativity and astrophysical source and population models. Determining sky locations facilitates electromagnetic follow-up observations to obtain even more information about a detection and perform multi-messenger astronomy.

Chapter 3

The Mock LISA Data Challenges

> Measure what can be measured, and make measureable what cannot be measured.
>
> Galileo Galilei

In addition to ground-based gravitational wave searches, there are long-term plans for a space-based detector. Originally the joint NASA and ESA project, the Laser Interferometer Space Antenna (LISA) [82], it is now an ESA-only venture, the New Gravitational-wave Observatory (NGO) [80, 81]. Preparation for this has involved both technical development of the satellites as well as exploration of the science case. This has involved modeling the populations of expected sources in the sensitivity range for the detector and then preparing and analysing expected observational data [147, 148]. In this chapter I present the analysis of data from a few of the mock data challenges that were run in order to spur development of data analysis code and demonstrate our ability to obtain scientific information from future observations. The work in Section 3.1 was done in collaboration with Jonathan Gair and Farhan Feroz and published in [149]; Section 3.2 was done in collaboration with F. Feroz and Stanislav Babak of the Max Plank Institute for Gravitational Physics (Albert Einstein Institute, AEI) in Potsdam, Germany; Section 3.3 was performed in collaboration with F. Feroz, S. Babak, and Antoine Petiteau also of the AEI.

3.1 Cosmic String Cusps

In preparation for the future launch of a space-based gravitational wave detector, such as LISA or NGO, development of data analysis algorithms has been a field of active research. To this end, LISA algorithm development has been encouraged by the Mock LISA Data Challenges (MLDCs) [150]. Round number 3, completed in April 2009, included data sets containing galactic white dwarf binaries (see Section 3.2), super-massive black hole binary mergers, extreme-mass-ratio inspirals, cosmic string cusps, or a stochastic background added to instrumental noise. In [149], we applied MULTINEST to the problem of detecting and characterising cosmic string cusps (MLDC 3.4). A summary of all submissions for MLDC 3 as well as plans for MLDC 4 can be found in [151]. A separate analysis of MLDC 3.4 is described in [152], where the authors focus on the problems of the degeneracies inherent in this signal type's posterior distribution.

Cosmic string cusps in the observable frequency range of LISA are a burst source with durations of hundreds of seconds or less, which is short on the timescale of a LISA mission lasting up to two years. However, they are just one type of potential source for burst signals. Since they can be physically modelled we can use a matched filtering search with expected waveforms. There may be bursts of gravitational waves in the LISA data stream from other sources and the question arises as to whether we would be able to detect them and if we can distinguish cosmic string bursts from bursts due to these other sources. The Bayesian evidence that MULTINEST computes is one tool that can be used to address model selection. Evidence has been used for gravitational wave model selection to test the theory of relativity in ground based observations [114], and, in a LISA context, to distinguish between an empty data set and one containing a signal [153]. Calculation of an evidence ratio requires a second model for the burst as an alternative to compare against. We chose to use a sine-Gaussian waveform as a generic alternative burst model, as this is one of the burst models commonly used in LIGO data analysis (see for instance [154]). We find that the evidence is a powerful tool for characterising bursts — for the majority of detectable bursts, the evidence ratio strongly favours the true model over the alternative. While the sine-Gaussian model does not necessarily well describe all possible un-modelled bursts, these results suggest that the evidence can be used to correctly identify any cosmic string bursts that are present in the LISA data stream.

3.1.1 Burst waveform models

3.1.1.1 Cosmic Strings

Gravitational waves can be generated by cosmic strings through the formation of cusps, where portions of the string are traveling at nearly the speed of light [155, 156]. Such radiation is highly beamed and when viewed along the emission axis is linearly polarised and takes a simple power-law form, $h(t) \propto |t - t_c|^{1/3}$ [155]. When viewed slightly off-axis, the waveform is still approximately linearly polarised but the cusp spectrum is rounded off and decays rapidly for frequencies above $f_{\mathrm{max}} \sim 2/(\alpha^3 L)$, where α is the viewing angle and L is the dimension of the feature generating the cusp [155, 157]. The particular model for the frequency domain waveform adopted in the MLDC is given by [150]

$$|h(f)| = \begin{cases} \mathcal{A} f^{-4/3} & f < f_{\mathrm{max}} \\ \mathcal{A} f^{-4/3} \exp\left(1 - \frac{f}{f_{max}}\right) & f > f_{\mathrm{max}} \end{cases} . \tag{3.1}$$

In addition, the MLDC waveforms include a fourth-order Butterworth filter to mitigate dynamic-range issues associated with inverse Fourier transforms. We adopt the same ansatz as the MLDC, namely that the Fourier domain waveform amplitude is given by

$$\begin{aligned} |h_+| &= \begin{cases} \mathcal{A} f^{-4/3} \left(1 + \left(\frac{f_{low}}{f}\right)^2\right)^{-4} & f < f_{\mathrm{max}} \\ \mathcal{A} f^{-4/3} \left(1 + \left(\frac{f_{low}}{f}\right)^2\right)^{-4} \exp\left(1 - \frac{f}{f_{max}}\right) & f > f_{\mathrm{max}} \end{cases} \\ |h_\times| &= 0 \end{aligned} \tag{3.2}$$

and the phase by $\exp(2\pi \imath f t_c)$, where t_c is the burst time.

3.1.1.2 Sine Gaussians

A sine-Gaussian waveform is centred on a particular frequency, and has exponentially suppressed power at nearby frequencies. We choose to consider a linearly polarised sine-Gaussian, for which the waveform magnitudes in the frequency domain are given by

$$|h_+| = \frac{\mathcal{A}}{2}\sqrt{\frac{Q^2}{2\pi f_c^2}}\exp\left(-\frac{Q^2}{2}\left(\frac{f-f_c}{f_c}\right)^2\right), \quad |h_\times| = 0 \tag{3.3}$$

where $\mathcal{A}$ is the dimensionless amplitude, f_c is the frequency of the sinusoidal oscillation, and Q is the width of the Gaussian envelope. The phase of the wave is again $\exp(2\pi\imath f t_c)$, where t_c is the central burst time. In the time domain, the sine-Gaussian is a small burst "packet" of particular frequency, f_c, with the number of cycles in the burst determined by Q.

3.1.1.3 Detector model

To include the LISA response we made use of the static LISA model as described in [158]. This approximation is valid for burst sources, as LISA does not move significantly over the typically short duration of the bursts, which is of the order of 1000s or less. The static LISA response model was also adopted for cosmic string bursts in [157].

Three optimal detection time-delay interferometry (TDI) channels, A, E and T, can be constructed from the LISA data stream [159]. The noise is uncorrelated within and across these channels. For the search of the MLDC data, the A, E and T channels were constructed as linear combinations of the three Michelson channels, X, Y and Z, that were provided in the data release.

$$A = \frac{1}{\sqrt{2}}(Z - X), \tag{3.4a}$$

$$E = \frac{1}{\sqrt{6}}(X - 2Y + Z), \tag{3.4b}$$

$$T = \frac{1}{\sqrt{3}}(X + Y + Z). \tag{3.4c}$$

The power spectral densities of the noise in the three channels are given by

$$\begin{aligned} S_A(f) = S_E(f) &= 16\sin^2(2\pi f t_L)\left(2\left(1+\cos(2\pi f t_L)+\cos^2(2\pi f t_L)\right)S_{\rm pm}(f)\right. \\ &\quad \left.+\left(1+\cos(2\pi f t_L)/2\right)S_{\rm sn}f^2\right) \qquad (3.5) \\ S_T(f) &= 16\sin^2(2\pi f t_L)\left(2\left(1-2\cos(2\pi f t_L)+\cos^2(2\pi f t_L)\right)S_{\rm pm}(f)\right. \\ &\quad \left.+\left(1-\cos(2\pi f t_L)\right)S_{\rm sn}f^2\right) \qquad (3.6) \\ S_{\rm pm}(f) &= \left(1+\left(\frac{10^{-4}\text{Hz}}{f}\right)^2\right)\frac{S_{\rm acc}}{f^2} \\ &\qquad (3.7) \end{aligned}$$

where $t_L = 16.678$s is the light travel time along one arm of the LISA constellation, $S_{\rm acc} = 2.5 \times 10^{-48}\text{Hz}^{-1}$ is the proof mass acceleration noise and $S_{\rm sn} = 1.8 \times 10^{-37}\text{Hz}^{-1}$ is the shot noise.

The LISA data stream will also contain a confusion noise foreground from unresolved white dwarf binaries in our galaxy. These were not included in the noise model for the MLDC round 3.4, and so we have ignored the confusion noise contribution in the current work. In practice, we found that the T channel was too noisy to be used in the MLDC search and did not contribute any significant SNR, so we used the A and E channels only for data analysis.

3.1.1.4 Likelihood evaluation

The space of gravitational waveform signals possesses a natural scalar product [160, 161],

$$\langle h|s\rangle = 2\int_0^\infty \frac{df}{S_n(f)}\left[\tilde{h}(f)\tilde{s}^*(f)+\tilde{h}^*(f)\tilde{s}(f)\right], \qquad (3.8)$$

where

$$\tilde{h}(f) = \int_{-\infty}^{\infty} dt\, h(t)e^{2\pi\imath ft} \qquad (3.9)$$

is the Fourier transform of the time domain waveform $h(t)$. The quantity $S_n(f)$ is the one-sided noise spectral density of the detector. For a family of sources with waveforms $h(t;\vec{\lambda})$ that depend on parameters $\vec{\lambda}$, the output of the detector, $s(t) = h(t;\vec{\lambda}_0)+n(t)$, consists of the true signal $h(t;\vec{\lambda}_0)$ and a particular realisation of the

Parameter	Minimum	Maximum
$\log(\mathcal{A})$	-23.3	-21
$f_{\max}$ (Hz)	0.001	0.5
t_c (s)	T_0	$T_0+\Delta T$
$\sin(\theta)$	-1	1
ϕ (rad)	0	2π
ψ (rad)	0	2π

Table 3.1: Prior probability distributions for the cosmic string cusp model.

noise, $n(t)$. Assuming that the noise is stationary and Gaussian, the logarithm of the likelihood that the parameter values are given by $\vec{\lambda}$ is

$$\log\mathcal{L}\left(\vec{\lambda}\right)=C-\frac{1}{2}\left\langle s-h\left(\vec{\lambda}\right)\middle|s-h\left(\vec{\lambda}\right)\right\rangle, \tag{3.10}$$

and it is this log-likelihood that is evaluated at each point in the search. The constant, C, depends on the dimensionality of the search space, but its value is not important as we are only interested in the relative likelihoods of different points. As mentioned earlier, the LISA data stream has several independent data channels. The total multi-channel likelihood is obtained by summing the scalar product in Equation (3.8) over the A and E channels that are used.

3.1.1.5 Priors

The parameter space over which to search for signals must also be specified. For the searches with the cosmic string model, we used uniform priors for each parameter that covered the signal space from which the MLDC sources were drawn. These correspond to the ranges in Table 3.1. Here, θ is the sky colatitude, ϕ is the sky azimuthal angle, and ψ is the polarization. The data stream was divided into segments for the search and T_0 is the start time of the particular data segment being searched and ΔT is that segment's length.

For the sine-Gaussian model, we maintained the approach of using uniform prior ranges while making sure that all signals would fall within our specified bounds. The priors used are given in Table 3.2. T_0, ΔT, θ, ϕ, and ψ are the same physical quantities

Parameter	Minimum	Maximum
$\log(\mathcal{A})$	-22.3	-18
f_c (Hz)	0.001	0.5
Q	1	10
t_c (s)	T_0	$T_0 + \Delta T$
$\sin(\theta)$	-1	1
ϕ (rad)	0	2π
ψ (rad)	0	2π

Table 3.2: Prior probability distributions for the sine-Gaussian model.

as in the cosmic string model; the prior on $\log(\mathcal{A})$ was modified in order to cover the same range of SNRs for the signals.

3.1.2 MLDC challenge 3.4 search

Mock LISA Data Challenge 3.4 [150] consisted of a data set of 2^{21} samples at a cadence of 1s, containing GW burst signals from cosmic string cusps [155, 156]. The signals occurred as a Poissonian random process throughout the data set, with an expectation value of five events. To analyse the data, we divided the data stream into segments of 32768s in length, with 2048s of overlap between neighbouring segments. The overlap was chosen to ensure that each of the bursts, which were of maximum duration of $\sim$ 1000s, would be entirely contained in at least one segment. To avoid artefacts in the Fourier domain, we used Welch windowing before performing the Fast Fourier Transform (FFT) of each segment. We also analysed data segments of the same length, offset from the first set by 16384s, to check no signals had been missed and to verify that the detected signals were not edge artefacts. We also repeated the search with segments of 16384s and 8192s in length to verify our results. MULTINEST was run with 4000 live points on each segment, using 8 3GHz Intel Woodcrest processors in parallel; the algorithm took approximately 5 minutes to run on a single segment when no signal was present, and twice this when one was present. The search of the offset and shorter segments returned all of the same detections. We ran the search on the MLDC training data set to test the algorithm, and then on the challenge data set.

To assess the quality of our solutions we will not only use the recovered parameters for a source, but also the signal-to-noise ratio (SNR) of the recovered solution. The degeneracies in the parameter space mean that a waveform can match the true waveform very well with very different parameter values. The recovered SNR is a measure of how much of a residual would be left in the data stream if the recovered parameters were used to subtract a given source from the data set. If the recovered SNR is close to the true SNR we can say that we have correctly identified the waveform of the source even if the parameters are quite different.

3.1.2.1 Challenge data

Three signals were found in the challenge data set at $t_c \sim 6 \times 10^5$s, $t_c \sim 1.07 \times 10^6$s and $t_c \sim 1.6 \times 10^6$s and this was the correct number of signals. We label these as source candidates 3, 2 and 1 respectively for consistency with the MLDC key file. For source candidates 1 and 3, MULTINEST returned a flat distribution in $f_{\max}$, thereby leading us to conclude that the actual maximum was above the Nyquist frequency of 0.5 Hz. We identified several modes in the posterior in each case, and used the local Bayesian evidence to characterise these. We decided to submit the two modes of highest evidence; these generally corresponded to two antipodal sky solutions. However, for source candidate 3 there was a third mode of almost equal local evidence, and so we submitted that mode as well. The middle source, 2, was found to have a break frequency of 0.0011 Hz, very close to the minimum value in the prior range of 0.001 Hz. At this low frequency, LISA is not able to resolve the sky position for a burst source. As expected, we found very broad posteriors for all parameters other than the maximum frequency. As the posterior was not well separated into modes, we chose to submit only one set of parameters in this case, which we took to be the maximum a-posteriori parameters.

In Table 3.3 we list the true parameters for the three sources in the challenge data set (available at [162]), and the parameters recovered by MULTINEST. For simplicity we only include the parameters for the mode of highest local evidence, not all of the modes submitted. The agreement between the true and recovered parameters is broadly consistent with our expectations, and the true parameters lie within a few standard deviations of the means in most cases. For source 1, t_c is quite far from the true value by this measure. However, this is due to the correlated degeneracies between t_c, θ and ϕ. The true time of coalesence was consistent with the full recovered posterior distribution. Additionally, for source 1 the true value of $f_{\max}$ is considerably below Nyquist, while our analysis recovered a flat distribution indicative of a value above Nyquist. We carried out further checks, but there is no indication that the data favours a particular value of $f_{\max}$ in this case.

In Figure 3.1 we show the true waveforms, in the A channel, for the three sources present in the data set, and overlay them with the waveforms corresponding to each of the modes that we included in our submission. We see clearly that, in all three cases, the recovered waveforms reproduce the true waveforms well, with small differences that arise from noise fluctuations in the detector. We expect the recovered SNR to be a better measure of the quality of the solution than the values of the recovered parameters. In Table 3.4 we list the true SNRs in the A and E channels, and the SNRs recovered by each of the solutions included in the submission. We have clearly recovered all of the SNR of the true signal, and the residual after subtraction would be at the level of instrumental noise. We thus regard the search as a success.

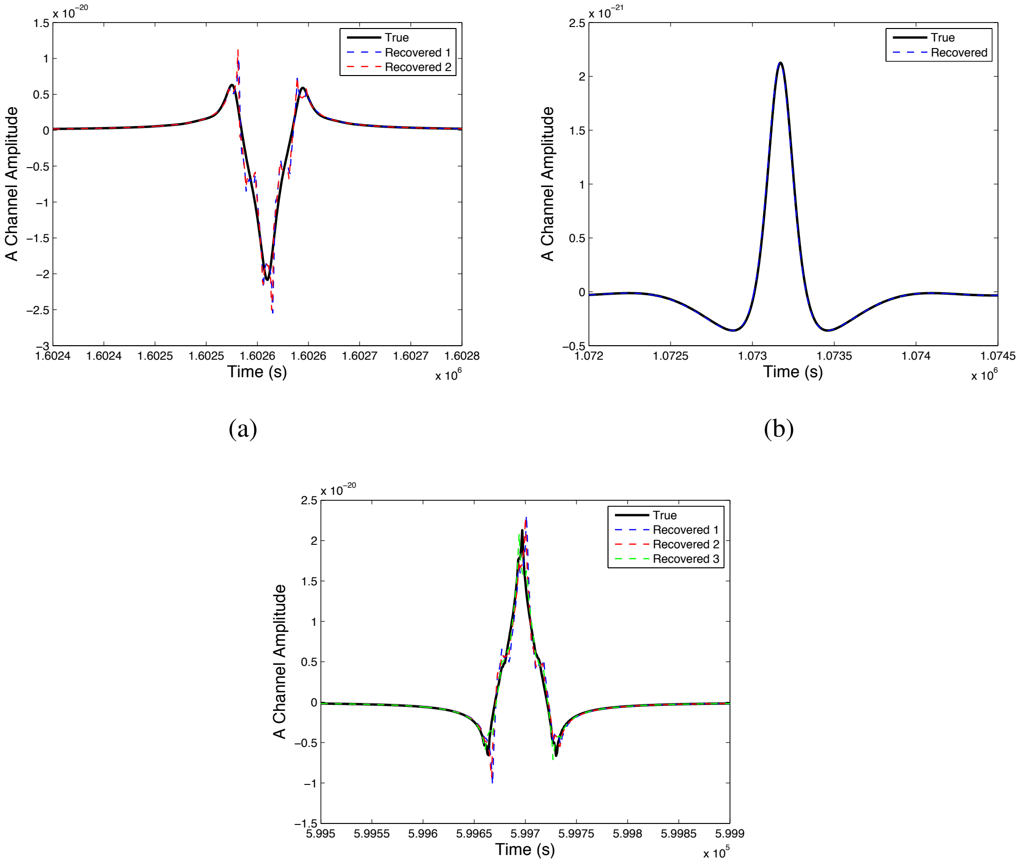
In Figure 3.2 we show 1D and 2D marginalised posterior probability distributions for the second of the sources present in the challenge data set. We see that in this case $f_{\max}$ can be measured very well, since it is far below Nyquist, but degeneracies in sky position and time of coalesence are clearly evident. The sky position degeneracies are particularly severe since the source is at very low frequency.

3.1.3 Model selection using Bayesian evidence

The MLDC search demonstrated that MULTINEST is able to correctly identify and characterize bursts from cosmic string cusps in LISA data. However, cosmic string cusps produce a very particular waveform signature. LISA might also detect bursts of gravitational radiation from other sources, such as supernovae of hypermassive objects, or even unmodelled sources. A search of the LISA data using a cosmic string cusp model could conceivably detect these other types of burst. In order to make scientific statements it is important to be able to say what the most likely origin for an observed burst could be. The Bayesian evidence provides a tool for carrying out model

Source		$\mathcal{A}$ ($\times 10^{-21}$)	ψ (rad)	θ (rad)	ϕ (rad)	$f_{\max}$ (Hz)	t_c ($\times 10^6$s)
1	True	0.866	3.32	0.556	3.71	0.0296	1.6022
	Recovered	1.65	2.82	0.349	6.01	0.5*	1.6030
	Mean	1.13	2.70	−0.141	4.38	0.27	1.6030
	Std. Dev.	±0.22	±0.14	±0.30	±2.6	±0.14	±0.000071
2	True	2.79	5.12	−0.444	3.17	0.00108	1.0727
	Recovered	2.97	0.271	−0.893	5.12	0.00108	1.0732
	Mean	2.75	0.91	−0.00804	5.31	0.00108	1.0733
	Std. Dev.	±0.63	±0.28	±0.52	±0.46	±0.000048	±0.00019
3	True	0.854	4.66	−0.800	0.217	6.15	0.60000
	Recovered	1.14	0.927	−0.562	5.45	0.5*	0.59991
	Mean	1.16	0.962	−0.488	5.48	0.24	0.59992
	Std. Dev.	±0.18	±0.16	±0.16	±0.18	±0.14	±0.000076

Table 3.3: Parameters of the three signals present in the challenge data set, and the parameters recovered by MULTINEST for the mode of highest local evidence identified by the search. The parameters are amplitude, $\mathcal{A}$, polarisation, ψ, sky colatitude, θ, and longitude, ϕ, break frequency, $f_{\max}$, and time of coalesence at the Solar System barycentre, t_c. In the row labelled "Recovered" we quote the maximum likelihood parameters of the mode with highest evidence identified by MULTINEST. In the rows labelled "Mean" and "Std. Dev." we quote the mean and standard deviation in that parameter, as computed from the recovered posterior for the mode of highest evidence. Where the value of $f_{\max}$ is marked by a *, the distribution in this parameter was very flat, indicating that $f_{\max}$ was probably above Nyquist. In these cases, we set $f_{\max}$ to the Nyquist frequency of 0.5Hz for the maximum likelihood parameters, but still quote the actual recovered mean and standard deviation.



(a)

(b)

(c)

Figure 3.1: Plots comparing the true waveforms with the waveforms for the various modes recovered by MULTINEST and included in our submission for the analysis of the challenge data set. (a) Source 1, (b) Source 2, and (c) Source 3.

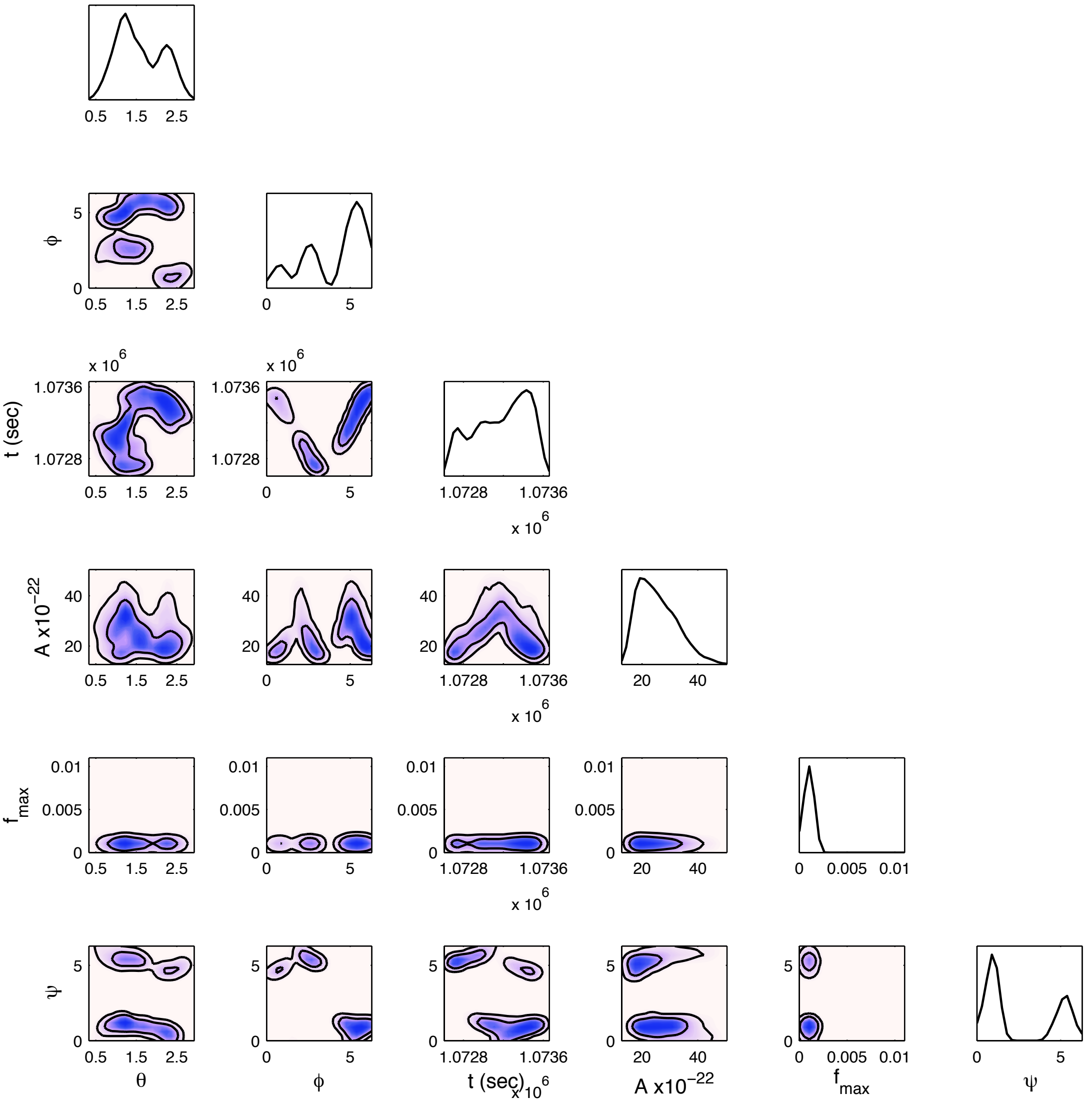


Figure 3.2: Two-dimensional marginalised posteriors as recovered by MULTINEST in the search for the second of the cosmic string bursts in the MLDC challenge data set. The parameters, from top-to-bottom and left-to-right, are colatitude, longitude, burst time, burst amplitude, burst break frequency and waveform polarization. At the top of each column we also show the one-dimensional posterior for the column parameter.

Source	t_c(s)	Channel	True SNR	Recovered SNR Mode 1	 Mode 2	 Mode 3
1	1.6×10^6	A	41.0	41.2	41.0	N/A
		E	14.5	14.5	14.6	
2	1.1×10^6	A	30.7	30.7	N/A	N/A
		E	13.9	13.9		
3	6×10^5	A	18.8	18.9	18.5	18.4
		E	36.9	36.7	37.1	36.8

Table 3.4: Signal-to-noise ratios for the three sources in the challenge data set. We show the true SNRs in the A and E channels, plus the SNRs for all of the modes we submitted for analysis.

selection of this type, although it relies on having two alternative hypotheses to compare. As an alternative to a cosmic string model, we consider a sine-Gaussian burst, which is a generic burst waveform commonly used in burst analysis for ground-based gravitational wave detectors [154]. We can use the evidence ratio to compare these two models when analysing a data stream containing a cosmic string burst or containing a sine-Gaussian burst. There are two stages to this analysis — 1) determining the signal-to-noise ratio at which we begin to be able to detect a signal of each type in a LISA data set; 2) determining the SNR at which the Bayesian evidence begins to favour the true model over the alternative.

3.1.3.1 Detection threshold

To determine the SNR needed for detection of a burst source with the MULTINEST algorithm, we generated a sequence of data sets containing a particular source and Gaussian instrumental noise generated using the theoretical noise spectral densities given in Equation (3.7), but varying the burst amplitude between the data sets to give a range of SNRs between 5 and 15. Note that here and in the following we quote total SNRs for the A and E channels combined, $\rho^2_{\text{Total}} = \rho^2_A + \rho^2_E$. Each data set was searched with MULTINEST and the global log-evidence of the data set, as computed by MULTINEST, was recorded. This was repeated for 8 different cosmic string sources and 9 different sine-Gaussian sources. The parameters for the cosmic string sources were taken from the parameters of the five training and three blind sources injected in the MLDC data sets, as these covered the parameter space of possible signals nicely. The important parameters of the sine-Gaussian model are the frequency and the burst width, so we considered three different values of each, and constructed waveforms for the nine possible combinations, while choosing a random value for the sky position and waveform polarisation in each case.

In Figure 3.3 we show the global log-evidence of the data as a function of the SNR of the injected signal for the cosmic string burst sources. Figure 3.4 shows the corresponding results for the sine-Gaussian bursts. The detection threshold for cosmic string bursts is at an SNR of 5–8 depending on the source parameters. Most of the sources have a significant log-evidence (greater than 3) at an SNR of 7, but the SNR required for detection is higher for the first training source and the second blind source. These are the two sources with the lowest values of the break frequency f_{max}, which makes the waveforms much smoother and simpler in the time domain. This may explain why it is more difficult to distinguish them from noise. The MLDC data sets had a prior for the source SNR, in a *single* Michelson channel, of $\rho \in [10, 100]$, so these results suggest we would be able to detect any source drawn from the MLDC prior.

The sine-Gaussian bursts require a slightly higher SNR for detection, of 6–9, but this increase in threshold is only of the order of 1 in SNR. The sine-Gaussian signals are in general much simpler in form than the cosmic strings and therefore it is perhaps unsurprising that it requires a higher SNR to distinguish them from noise. In this case, it was the sources with highest frequency, $f_c = 0.49$Hz, that were most difficult to detect. However, since the Nyquist frequency of the data sets was 0.5Hz, the difficulty of detection may have arisen because the signal was partially out of band.

3.1.3.2 Model selection

To explore model selection, we used MULTINEST to search the same data sets described above, but now using the alternative model, i.e., we searched the cosmic string data sets using the sine-Gaussian likelihood and vice versa. For sufficiently high SNR, in both cases the alternative model was able to successfully detect a signal in the data set. Typically, the best-fit sine-Gaussian signal to a cosmic string source has low Q and a frequency that matches the two lobes of the cosmic string burst. The parameter Q sets

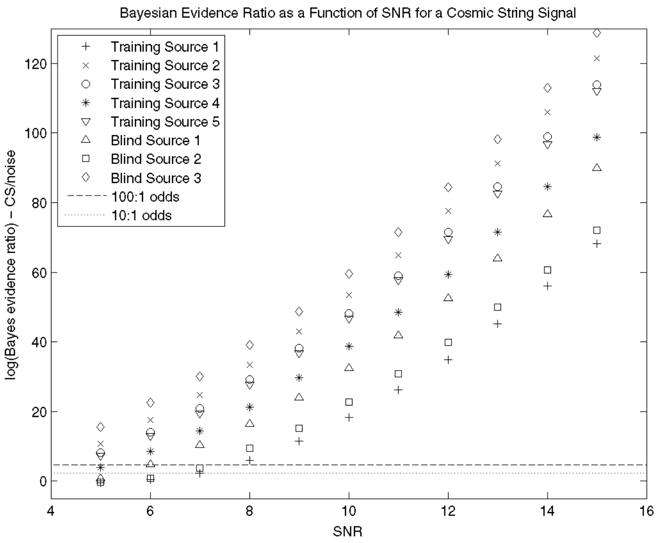


Figure 3.3: The Bayesian log-evidence for the data set as a function of the SNR for eight different cosmic string cusp signals. These signals had the same parameters, other than amplitude, as the various sources used in the training and blind data sets of the MLDC. The labels in the key are consistent with labels in Table 3.3 and the solutions at [162].

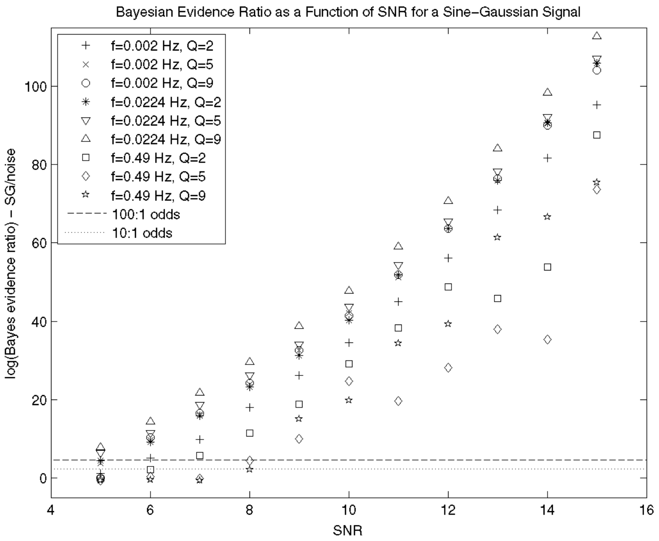


Figure 3.4: As Fig. 3.3, but now showing results for nine different sine-Gaussian signals, with frequency, f, and width, Q, as shown.

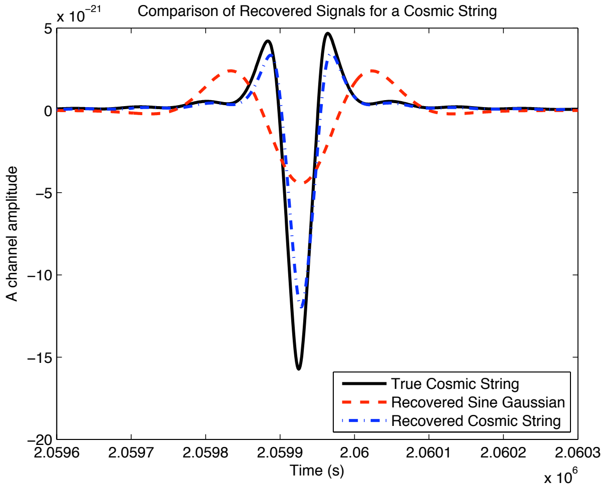


Figure 3.5: Typical example of confusion when searching a cosmic string data set with the wrong model. The plot shows a comparison of the injected cosmic string signal to the best-fit signals found by MULTINEST using the cosmic string model as templates and using the sine-Gaussian model as templates.

the number of cycles in the sine-Gaussian wave packet, and so a sine-Gaussain with $Q \sim 2$ most closely resembles a cosmic string event, which typically has two cycles. This is illustrated for a typical case in Figure 3.5. Similarly, the best-fit cosmic string source to a sine-Gaussian signal matches the central two peaks of the sine-Gaussian waveform as well as possible. A typical case is shown in Figure 3.6.

From the results of these searches it is possible to construct the evidence ratio of the two models for each of the data sets. In Figure 3.7 we show the ratio of the Bayesian evidence for the cosmic string model to that of the sine-Gaussian model when searching the cosmic string data sets. We see that for some signals the evidence ratio starts to significantly favour the true model, i.e., the cosmic string, at an injected SNR of ~ 5, which is the point at which we first start to be able to detect the cosmic string burst at all. However, for the two low frequency sources, training source 1 and blind source 2, the evidence ratio only starts to favour the true model at SNR$\sim$ 11 or higher. This partially reflects the higher SNR required to detect the source, but also includes the confusion caused by the signal very closely resembling a sine-Gaussian. We conclude that when a cosmic string burst is loud enough to be detected, then the evidence favours the interpretation of the event as a cosmic string burst, with

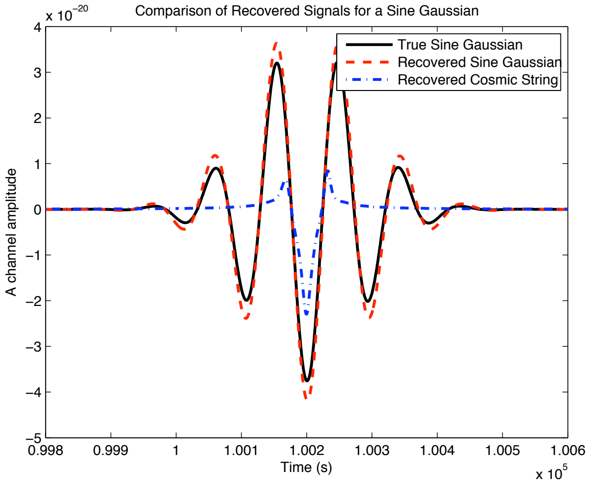


Figure 3.6: As Figure 3.5, but we now show a typical confusion example for searches of the sine-Gaussian data sets. We compare the injected sine-Gaussian signal to the best-fit signals recovered by MULTINEST using both the cosmic string and the sine Gaussian models.

lower frequency sources requiring additional SNR. Since MLDC sources have an SNR greater than 10 in each Michelson channel, they should be clearly distinguishable from sine-Gaussian bursts.

In Figure 3.8 we show the ratio of the evidence of the sine-Gaussian model to that of the cosmic string model when searching the data sets containing sine-Gaussian signals. These are distinguishable from cosmic strings at SNRs of $\sim$ 6–9. This slightly higher SNR in order to correctly choose the sine-Gaussian model reflects the fact that we need a somewhat higher SNR to detect the sine-Gaussians in the first place. The only case for which the evidence ratio does not begin to favour the sine-Gaussian model at the point where the source becomes detectable is the case with $f = 0.002$ Hz and $Q = 2$. This is a sine-Gaussian signal with only two smooth oscillations, and so it does look rather like a low frequency cosmic string event. Even in that case, the evidence begins clearly to favour the correct model for SNRs of $\sim$ 12 and higher.

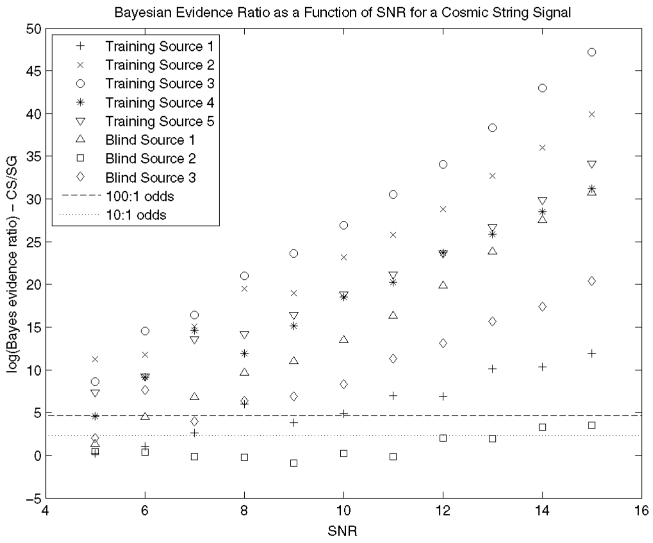


Figure 3.7: The ratio of the Bayesian evidence for the cosmic string model to that of the sine-Gaussian model when searching a data set containing a cosmic string burst source. We show the Bayesian evidence ratio as a function of signal SNR for a variety of different cosmic string sources.

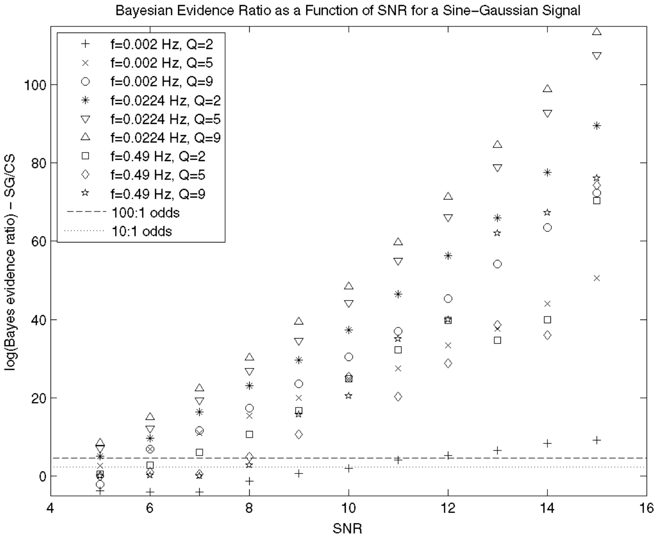


Figure 3.8: As Fig. 3.7, but now showing the ratio of the Bayesian evidence for the sine-Gaussian model to that of the cosmic string model when searching a data set containing a sine-Gaussian burst source.

3.1.4 Discussion

We have considered the use of the multi-modal nested sampling algorithm MULTINEST for detection and characterisation of cosmic string burst sources in LISA data. As a search tool, the algorithm was able successfully to find the three cosmic string bursts that were present in the MLDC challenge data set. These sources were correctly identified in the sense that the full signal-to-noise ratio of the injected source was recovered, and a posterior distribution for the parameters obtained. The maximum likelihood and maximum a-posteriori parameters were not particularly close to the true parameters of the injected signals, but this was a consequence of the intrinsic degeneracies in the cosmic string model parameter space and in all cases the true parameters were consistent with the recovered posterior distributions.

In controlled studies, we found that the SNR threshold required for detection of the cosmic string bursts was $\sim$ 5–8, depending on the burst parameters. Bursts with a low break-frequency require a higher SNR to detect than those with high break frequencies. We also explored the detection of sine-Gaussian bursts and in that case the SNR required for detection was slightly higher, being typically 6–9, with sources having frequency close to Nyquist being more difficult to detect.

MULTINEST is designed to evaluate the evidence of the data under a certain hypothesis, and this can be used to compare possible models for the burst sources. LISA may detect bursts from several different sources, and it is important for scientific interpretation that the nature of the burst be correctly identified. We used the Bayesian evidence as a tool to choose between two different models for a LISA burst source — the cosmic string model and the sine-Gaussian model, which was chosen to represent a generic burst. The Bayesian evidence works very well as a discriminator between these two models. The evidence ratio begins to clearly favour the correct model over the alternative at the same SNR that the sources become loud enough to detect in the first place.

The usefulness of MULTINEST as a search tool in this problem is a further illustration of the potential utility of this algorithm for LISA data analysis, as previously demonstrated in a search for non-spinning SMBH binaries [107]. Other algorithms based on Markov Chain Monte Carlo techniques have also been applied to the search for cosmic strings [157]. Both approaches performed equally well as search tools in the last round of the MLDC. MultiNest was not designed primarily as a search algorithm, but as a tool for evidence evaluation, and this work has demonstrated the utility of the Bayesian evidence as a tool for model selection in a LISA context. Other problems where the evidence ratio approach could be applied include choosing between relativity and alternative theories of gravity as explanations for the gravitational waves observed by LISA, or choosing between different models for a gravitational wave background present in the LISA data set. The Bayesian evidence was previously used in a LIGO context as a tool to choose between alternative theories of gravity [114] and in a LISA context to distinguish a data set containing a source from one containing purely instrumental noise [153].

In the context of interpretation of LISA and NGO burst events, what we have considered here is only part of the picture. We have shown that we are able to correctly choose between two particular models for a burst, and this can easily be extended to include other burst models. However, LISA/NGO might also detect bursts from unmodelled sources. In that case, algorithms such as MultiNest which rely on matched filtering would find the best fit parameters within the model space, but a higher intrinsic SNR of the source would be required for detection. In such a situation, we would like to be able to say that the source was probably not from a model of particular type, e.g., not a cosmic string burst. There are several clues which would provide an indication that this was the case. The sine-Gaussian model is sufficiently generic that we would expect it, in general, to provide a better match to unmodelled bursts than the cosmic string model, which has a very specific form. Therefore, we could say that if the evidence ratio favoured the cosmic string model over the sine-Gaussian model it was highly likely that the burst was in fact a cosmic string and not something else. Similarly, if we found that several of the alternative models had almost equal evidence, but the SNR was quite high, it would be indicative that the burst was not described by any of the models. We have seen that at relatively moderate SNRs, when the signal is described by one of the models, the evidence clearly favours the true model over an alternative. If we found that two models gave almost equally good descriptions of the source, it would suggest that the burst was not fully described by either of them. A third clue would come from the shape of the posterior for the source parameters. The cosmic string waveform space contains many degeneracies, but these can be characterised theoretically for a given choice of source parameters. If the signal was not from a cosmic string, we might find that the structure of the posterior was modified. Finally, some techniques have been developed for the Bayesian reconstruction of generic bursts [163] which could also be applied in a LISA/NGO context. The usefulness of these various approaches can be explored further by analysing data sets into which numerical supernovae burst waveforms have been injected. While the necessary mass of the progenitor is probably unphysically high for a supernova to produce a burst in the LISA/NGO frequency band, such waveforms provide examples of unmodelled burst signals on which to test analysis techniques. The final LISA/NGO analysis will employ a family of burst models to characterize any detected events. The work described here demonstrates that the Bayesian evidence will be a useful tool for choosing between such models, and MULTINEST is a useful tool for computing those evidences.

3.2 Galactic White Dwarf Binaries

The Milky Way Galaxy has a population of order tens of millions of objects that can be classified as compact binaries. These include separated binaries of two objects that interact only gravitationally and interacting binaries which have mass transfer from one object to the other. Separated binaries may consist of two white dwarfs, two neutron stars, or one of each. Types of interacting binaries include AM CVn stars (white dwarf accreting mass from a companion) and ultra-compact X-ray binaries (neutron star accreting mass from a companion). Mock LISA Data Challenge 3.1 [150] uses a population model which is described by [164]. It consists of $\sim$ 26 million detached binaries and $\sim$ 34 million interacting binaries, each modeled as two point masses in circular orbit with increasing or decreasing orbital frequency depending on the dominant process (gravitational radiation or mass transfer, respectively).

3.2.1 Data Model

The plus and cross polarizations of the binary signal are given by [150]

$$h_+(t) = A(1+\cos^2\iota)\cos\left[2\pi(ft+\dot{f}t^2/2)+\phi_0\right], \tag{3.11}$$

$$h_\times(t) = -2A(\cos\iota)\sin\left[2\pi(ft+\dot{f}t^2/2)+\phi_0\right], \tag{3.12}$$

$$A = \frac{4(G\mathcal{M})^{5/3}}{c^4 D_L}(\pi f)^{2/3}, \tag{3.13}$$

where $\mathcal{M}$ is the chirp mass as defined before, $\dot{f}$ is the constant frequency derivative, ϕ_0 is the phase at $t = 0$, and ι is the inclination of the binary to our line of sight. A fast-slow decomposition of the LISA orbital motion can also be used to simulate the detector phase measurements in the frequency domain, as described in [165].

This work was performed as an extension to that done in [166]. In that paper, the authors use an $\mathcal{F}$-statistic to analytically maximise over most amplitude and phase factors and then perform a template bank search followed by parameter refinement with the Nelder-Mead optimisation algorithm [167]. The search is done in the four-dimensional space defined by the initial frequency, f, frequency drift, $\dot{f}$, and sky latitude and longitude, β and λ.

The challenge data set was broken into 0.1 mHz frequency bands to simplify the analysis; each band was separated with Butterworth bandpass filters as described in Section V of [166]. In the initial analysis for each frequency band, the loudest signal is subtracted and then the search is performed again until no signal can be identified. This process is able to handle the large number of signals present in the data, especially below $\sim$ 7mHz, but can be very time-consuming at higher frequencies where larger template banks are needed to ensure a minimum correlation with any signals present. We therefore aimed to use the ability of MultiNest to identify multiple signals present in each band and evaluate local Bayesian evidences and a posterior maximum for each.

Each bandpassed frequency range was analysed for signals with frequencies in the centre of the band. The prior on f was $\mathcal{U}[f_c - 0.06, f_c + 0.06]$mHz. Since the bandpasses were performed at 0.1 mHz intervals, this allowed for 0.01 mHz of overlap on either side while not searching too close to the edges of the filter. The frequency drift, $\dot{f}$ for detached binaries evolving via gravitaitonal radiation is given by

$$\dot{f} = \frac{24}{5\pi} \left(\frac{G\mathcal{M}}{2c^3} \right)^{5/3} (2\pi f)^{11/3}. \tag{3.14}$$

Therefore, in order to set prior limits on this parameter, we assumed a uniform prior over $[(-0.25)\dot{f}_{\max}, \dot{f}_{\max}]$. $\dot{f}_{\max}$ is calculated with a maximum chirp mass of $1.5M_\odot$ at the highest frequency in the prior. A minimum $\dot{f}$ of $-1/4$ times the maximum is used because as [165, 168] describe, contact binaries with mass transfer are driven to longer periods with similar frequency evolution to Equation (3.14), but with opposite sign and slightly lower magnitude. We have the benefit of knowing the true signals present in the data and these prior bounds were confirmed to include them all. The sky position prior is taken to be uniform over the sphere: uniform in $\lambda \in [0,2\pi]$ and uniform in $\cos\beta \in [-1,1]$.

The noise in each TDI channel can be taken as approximately constant over the frequency band and is calculated at the central frequency using the formulas provided in [166].

3.2.2 Results

We searched 71 frequency ranges, spanning 4.99–12.11mHz. The density of signals is much greater at lower frequencies so we expect more modes of the posterior to be identified in those bands. Modes were clustered by starting frequency only. Once the analyses were complete, the posteriors were evaluated to determine how many signals were found and with what parameter values. This was done by collecting all posterior modes with local log-evidences above 100. If any two were found to have frequencies within 0.16μHz and frequency derivative within 2.6×10^{-14}Hz/s, the one with larger evidence was considered. These constraints on saying two signals are potentially the same are determined by $10/T_{\text{obs}}$ and $100/T_{\text{obs}}^2$, where T_{obs} is the total two year observation time. This served to eliminate potential false detections and secondary maxima. Maximum likelihood parameter values were saved for each detected signal; very narrow modes meant that these did not differ significantly from maximum posterior values. In total, 754 signals were identified in the 4.99–12.11mHz range this way.

Using the criteria from [166] to determine the true/false positive rate for these detections we find that the majority of signals found are true. The frequency resolution is given by the width of bins in Fourier space, which is given by $\Delta f = 1/T_{\text{obs}} \simeq 16$nHz. The resolution in the frequency drift is the difference at which over the total observation time the frequency will drift by the resolution in frequency; this is given by $\Delta \dot{f} = 1/T_{\text{obs}}^2 \simeq 2.5 \times 10^{-16}$Hz/s. We therefore require our reported solutions to be correct within the resolution possible. However, in this frequency range we find only about one quarter of the 2924 total signals present. At higher frequency bands we find all or nearly all of the signals but the recovery decreases as more signals add to the confusion. Figure 3.9 shows the number of true signals in the MLDC3.1 key (bright

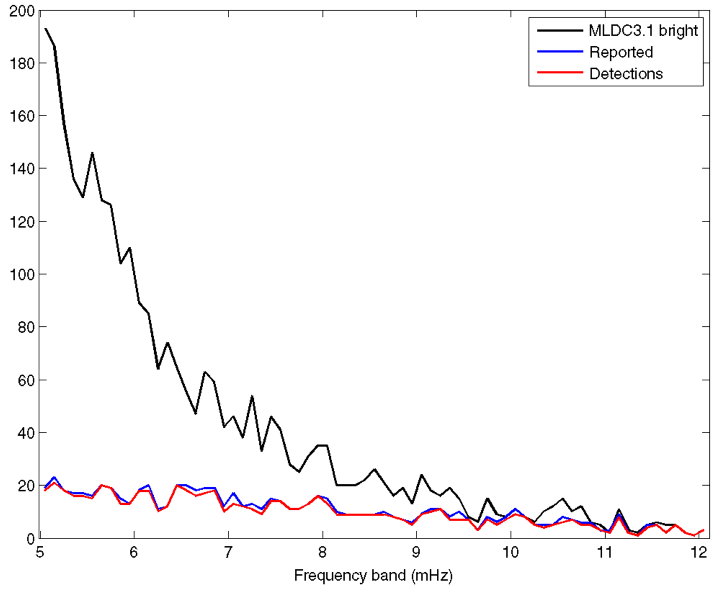


Figure 3.9: Number of true, recovered, and correctly detected signals in each frequency band analysed.

signals that can be measured above the background) as well as the number of reported and correct detections in each frequency band. Complementary to this, Figure 3.10 shows the purity and completeness of the reported detections from the analysis of each frequency band. The purity is the percent of reported detections that are correct and completeness is the percentage of all true signals that are correctly detected. The purity is roughly consistent across all frequency bands, staying above 80% most of the time and only once dipping just below 70%. Completeness is very poor at low frequencies, with values as low as 10%. However, at frequencies above 10mHz most signals present are recovered.

Values of f and $\dot{f}$ for true signals present in the data versus those recovered are shown in Figure 3.11. The first plot shows all signals within the 4.99–12.11mHz band, while the second and third plots show the upper and lower 0.5mHz ranges, respectively. We can compare in these last two the relative density of signals and how that has an effect on the completeness of the detections and the rate of false positives (indicated by a red ×). In the range of less dense signals we not only recover a larger fraction of

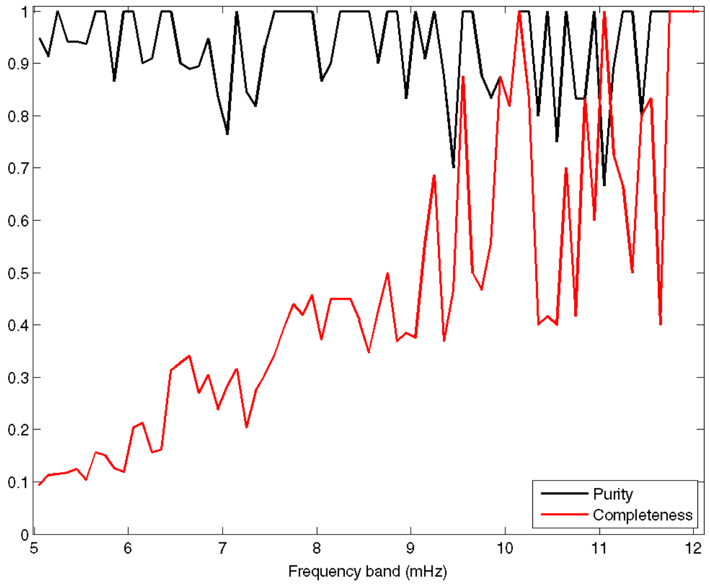


Figure 3.10: The purity and completeness of detections for all frequency bands analysed. The purity is the percent of reported detections that are correct and completeness is the percentage of all true signals that are correctly detected.

signals present, but also with smaller errors and, in this case, no false positives.

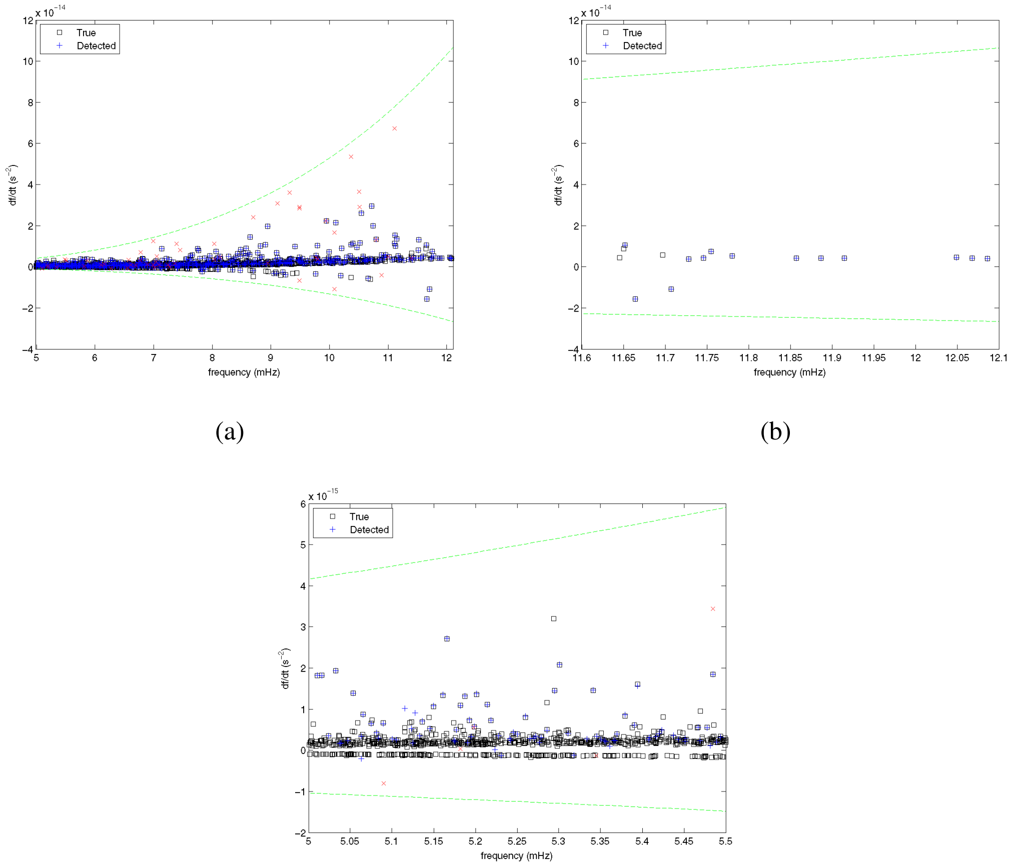
For all correctly recovered signals we can consider the recovery errors. Figure 3.12 plots histograms of the difference in the reported values of the measured parameters versus their true values. f and $\dot{f}$ are reported in units of the resolution size. We find that f and $\dot{f}$ are found to high precision, with most values being correct to within a fraction of the resolution. This is possible as the signals will not be entirely in one frequency bin over the entire observation period and the relative magnitudes can provide this extra resolution. In addition, the sky location is found to good precision with most errors much less than a radian.

3.2.3 Discussion

This analysis has shown that MULTINEST, and nested sampling in general, is a useful tool for detecting multiple signals present in a single data set. Even in the presence of a much larger number of signals causing confusion, the purity of recovered signals is not adversely affected. This will be very useful for analysis of eLISA/NGO data that will have many white dwarf binary signals present. In order to improve the purity and completeness of these detections, it will likely be necessary to implement a multi-stage analysis where the loudest detected signals are subtracted and the residual data re-analysed for more signals. This type of pipeline has been tested for white dwarf binary sources for LISA [169]. At lower frequencies it may also be necessary to use narrower frequency bands for analysis to lower the number of signals present. The status of detection at this time is very promising as this analysis may be performed very rapidly and in parallel to detect multiple signals at a time.

3.3 Cosmic Strings in the Presence of a Galactic Foreground

When a space-based GW detector is eventually launched, it will observe many signals from various sources all at the same time. As considered in the previous section, 3.2, there may be a foreground of galactic white dwarf binary signals. The impact of this will depend on the final sensitivity of the eLISA/NGO detector. Here we consider the data from the fourth round of the Mock LISA Data Challenge in an attempt to recover



(a)

(b)

(c)

Figure 3.11: Plots comparing the values of f and $\dot{f}$ for true signals present in MLDC 3.1 challenge data in the 4.99–12.11mHz range. True positive detections are indicated by a blue $+$, false positives by a red $\times$. (a) Displays the entire frequency range, (b) the upper 0.5mHz, and (c) the lower 0.5mHz. Dashed green lines indicate the maximum and minimum for the $\dot{f}$ prior.

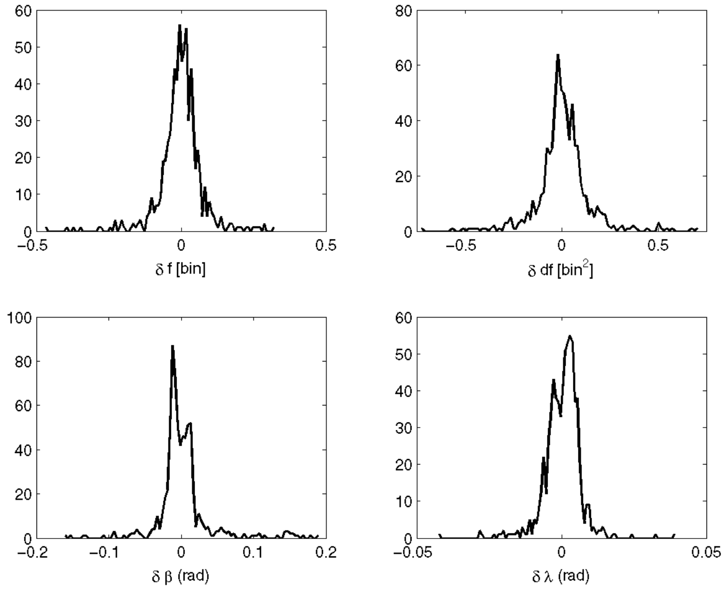


Figure 3.12: The error in recovered parameters for correctly detected signals. f and $\dot{f}$ are reported in terms of the resolution size.

cosmic string bursts in the presence of no galaxy, a reduced galactic component, and the full galaxy foreground.

The data generation procedure is described in [151]. With the training data key file for MLDC 4, we generated data sets with just the cosmic strings, instrumental noise, and one of three options for the level of inclusion for galactic white dwarf (WD) binaries. At the lowest level, no WDs are included. In the middle level, a confusion foreground of millions of WD binaries that cannot be separated from each other is added. These have an approximate power spectral density given by Equation (3.15), which was calculated in [170] (Equation 14) for the specific model of the galactic stellar population used in MLDC 4.

$$S_{\mathrm{conf}}(f) = \begin{cases} 10^{-44.62} f^{-2.3} & 10^{-4.0} < f \leq 10^{-3.0} \\ 10^{-50.92} f^{-4.4} & 10^{-3.0} < f \leq 10^{-2.7} \\ 10^{-62.80} f^{-8.8} & 10^{-2.7} < f \leq 10^{-2.4} \\ 10^{-89.68} f^{-20.0} & 10^{-2.4} < f \leq 10^{-2.0} \\ 0 & \text{else} \end{cases} \mathrm{m^2 Hz^{-1}} \tag{3.15}$$

Finally, the full WD galaxy population, including the confusion foreground and thousands of louder, separable binaries is added.

3.3.1 Data and Signal Models

The full MLDC 4 data spanned approximately 2 years and was sampled at a 1.875s cadence, yielding 2^{25} total samples. Cosmic string signals were injected at times randomly sampled from a uniform distribution over the entire span, with a number that was sampled from a Poisson distribution with a mean of 20 signals (22 were actually injected). In order to efficiently search the data for these sparse signals, we segmented the data into stretches that were 61440s long with 15360s of overlap (each start was 46080s after previous), to avoid missing signals potentially at the edges. Additionally, we applied a Tukey window to each segment to avoid edge effects in the Fourier transform; the Tukey window was used in place of the Welch window that was implemented in Section 3.1.2 as it ensures that no data point was windowed by more than 50% with the given overlap (the minimum overlap from the Welch window was $\sim 25\%$). There were 1366 segments in total for each data set.

The instrumental noise model was the same as in MLDC3, so Equations (3.5), (3.6), and (3.7) were used. Additionally, for the data sets that included either the reduced or full galaxy population we added the galactic foreground confusion noise given by Equation (3.15). Values from Equation (3.15) were further multiplied by a factor to bring them into the same units as the instrumental noise and our data values.

$$S'_{\mathrm{conf}}(f) = 64\pi^2 f^2 t_L^2 \sin^2(2\pi f t_L) \times S_{\mathrm{conf}}(f) \tag{3.16}$$

3.3.2 Comparison of Results

3.3.2.1 No Galaxy

In this data set, only instrumental noise and cosmic strings were present. MULTINEST was run on each segment of data and modes found with a local log-evidence of more than 7 were recorded as potential detections. This limit is chosen due to the analysis in Section 3.1.3.1. Additionally, the global log-evidence from each segment was recorded. Of the 1366 segments, 1339 contained no GW signals. The 22 signals were present in 27 of the segments due to overlap. When compiling the list of detected signals, any two within 998s of each other were considered to be the same signal; this time was chosen as it is approximately equal to the light travel time across the diameter of the Earth's orbit, along which LISA would travel. If a signal was found twice, the mode with higher log-evidence was chosen to give the evidence and parameters of detection. This always selected the segment in which the signal was further from the edge and thus suffered less from windowing. Each signal was recovered to high accuracy with log-evidences of 70 or greater. In Figure 3.13 we plot the receiver-operator curve (ROC) for using the log-evidence as a threshold for detection and the true and false positive rates as a function of this threshold. The ROC is perfect, which indicates that there is a clear detection threshold for these signals in the presence of only instrumental noise and that all signals pass. Figure 3.14 shows a histogram of the global log-evidences from noise-only segments and a plot of the same log-evidences for each segment versus time. The frequency distribution of log-evidences follows an inverse-χ^2 distribution and there is no time dependency, as expected. All segments that contained only noise had log-evidences of 4 or less, with 98.5% having $\ln(\mathcal{Z}) < 0$ and thus supporting the noise-only model.

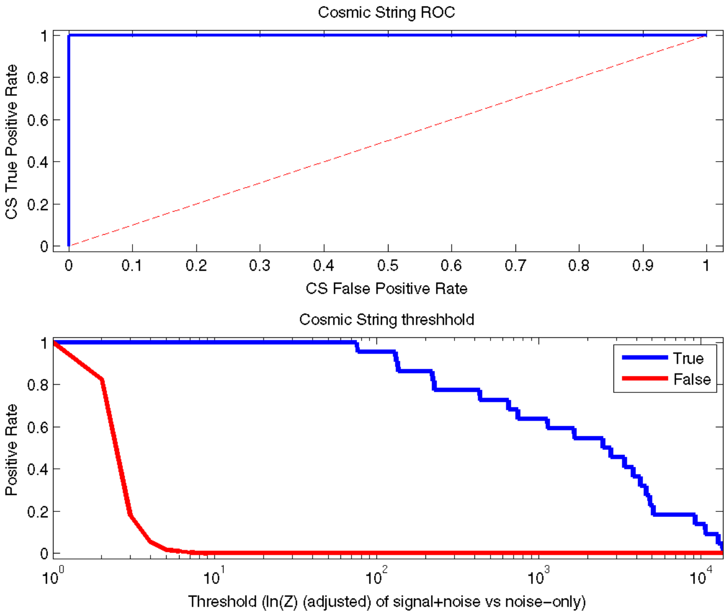


Figure 3.13: (Top) ROC for cosmic strings in instrumental noise, showing true positive rate vs. false positive rate as the threshold varies. (Bottom) The true and false positive detection rates as a function of the log-evidence. $\ln(\mathcal{Z})$ is adjusted by adding a constant to all values so that the minimum is greater than 1.

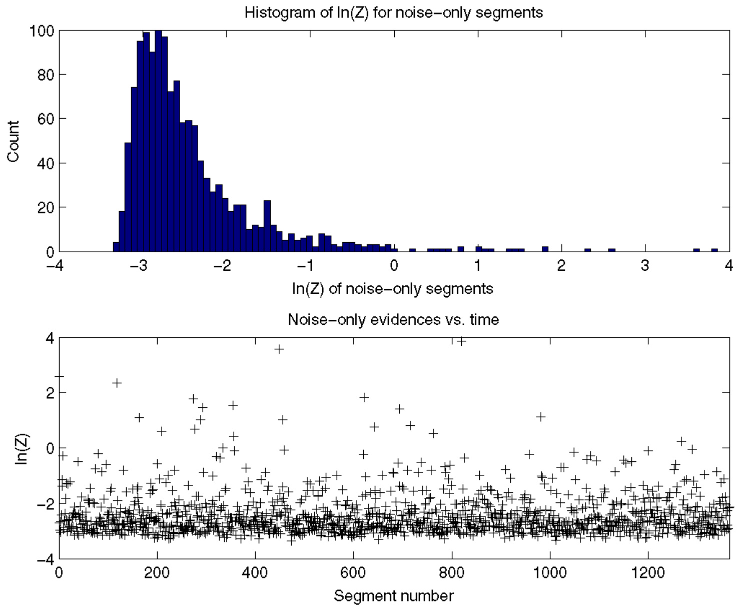


Figure 3.14: (Top) Histogram of global log-evidences for the segments that do not contain a cosmic string GW signal. These fit an inverse-χ^2 distribution and 98.5% have $\ln(\mathcal{Z}) < 0$. (Bottom) The log-evidence values of noise-only segments over time (segment number). No time dependence is present.

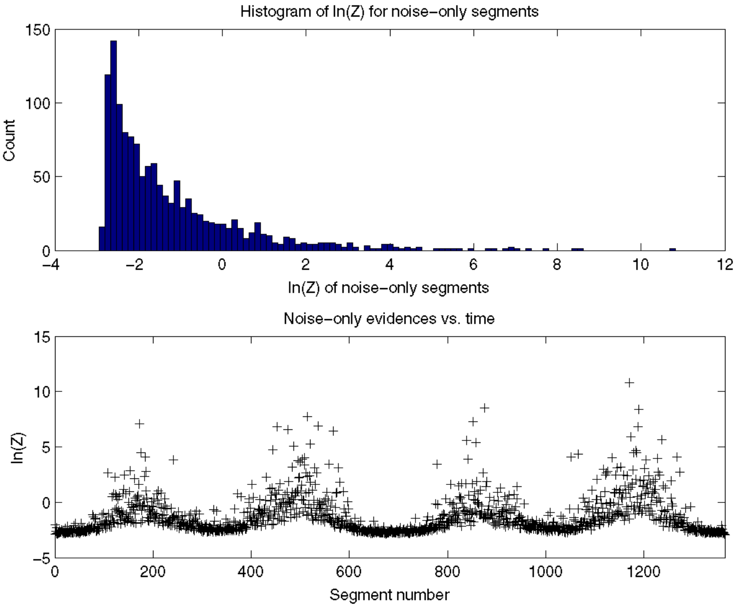


Figure 3.15: (Top) Histogram of global log-evidences for the segments that do not contain a cosmic string GW signal with the reduced galaxy population. (Bottom) The log-evidence values of noise-only segments over time (segment number). There is a clear time-dependence as the detector's sensitivity sweeps over the galaxy.

3.3.2.2 Reduced Galaxy

The second data set analysed contained the reduced population of galactic binaries in addition to instrumental noise. MULTINEST was run and we recorded detections and evidences as before. Figure 3.15 shows the histogram and time-dependent plot of log-evidences from segments that were known not to contain signals. The histogram does not follow the inverse-χ^2 distribution expected if Equation (3.15) correctly accounts for the presence of the galactic foreground confusion noise. We can also see in the time-dependent plot that the distribution of $\ln(\mathcal{Z})$ is not time-independent.

This time-dependence of the magnitude of the foreground confusion noise is due to the varying sensitivity of the LISA detector to signals coming from the galaxy over the course of the simulated mission. To account for this, the foreground confusion noise was multipled by a time-varying factor that applied the necessary modulation and

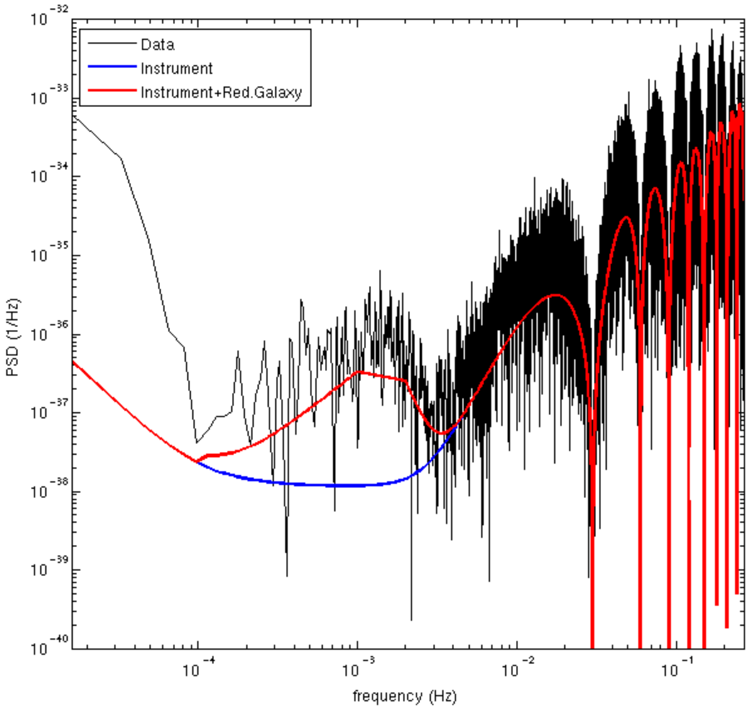


Figure 3.16: Power spectral density of a data segment containing only instrumental and reduced galaxy foreground confusion noise. On top are plotted the instrumental and instrumental plus reduced galaxy noise models.

increase in PSD when the detector is most sensitive to the galaxy. The new confusion noise is given, up to an unimportant normalisation factor, by

$$S''_{\rm conf} = \left(1 + \sin^2\left(2\pi \frac{t_m}{1\ \text{year}}\right)\right) \times S'_{\rm conf}, \tag{3.17}$$

where t_m is the time of the mid-point of a data segment. Figure 3.16 shows the power spectral density from a data segment containing no cosmic string signal, only instrumental and reduced galactic foreground confusion noise. The instrumental and instrumental plus reduced galaxy confusion noise models have been plotted on top for comparison.

The analysis was re-run with this new model for the galactic foreground and the resulting histogram and time-dependent plot of $\ln(\mathcal{Z})$ from segments without a cos-

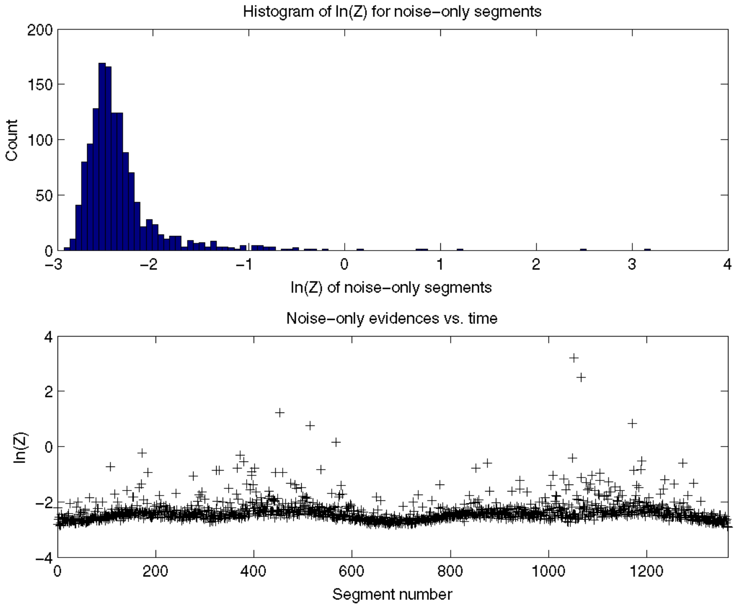


Figure 3.17: (Top) Histogram of global log-evidences for the segments that do not contain a cosmic string GW signal with the reduced galaxy population and modulated foreground noise model. These fit the correct distribution much more closely and 99.6% have $\ln(\mathcal{Z}) < 0$. (Bottom) The log-evidence values of noise-only segments over time (segment number). The time-dependence is much less significant.

mic string signal now have the features we expect, as seen in Figure 3.17. The time-dependence is nearly gone, but what remains does not significantly affect further results. The ROC along with true and false positive rates as a function of detection threshold are shown in Figure 3.18. Again the ROC is perfect and the detection threshold set is consistent with the results from Section 3.1.3.1.

Figures 3.19 and 3.20 show the true and recovered waveforms in the A and E channels for a selection of four of the 22 signals (true in solid black, recovered in dashed red). Differences from the true signal are a result of the addition of instrumental and galactic foreground noise as well as windowing for signals that were nearer to the edge of a segment. Low-frequency signals, such as signals 2 and 6 in Figure 3.19, were often recovered with higher precision due to their simpler structure and the fact that the entire signal is within the frequency sensitivity of the detector. Signals with higher

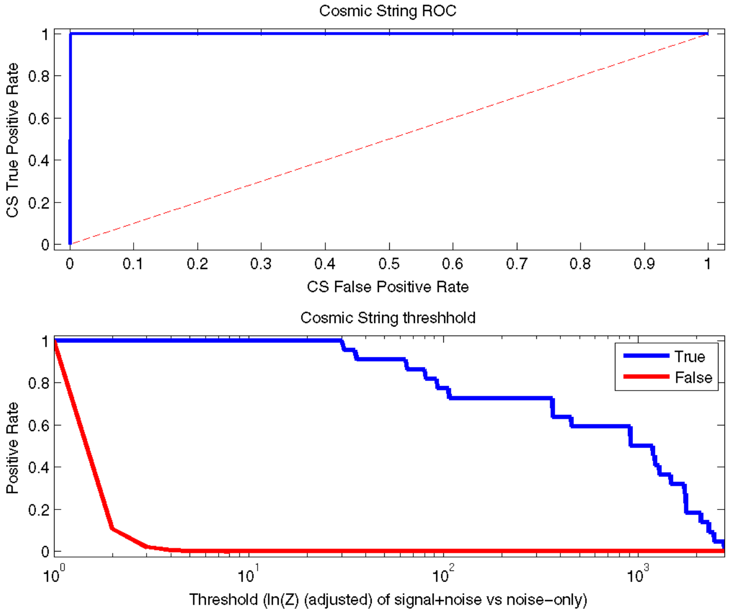


Figure 3.18: (Top) ROC for cosmic strings in instrumental and galactic foreground confusion noise, showing true positive rate vs. false positive rate as the threshold varies. (Bottom) The true and false positive detection rates as a function of the log-evidence. $\ln(\mathcal{Z})$ is adjusted by adding a constant to all values so that the minimum is greater than 1.

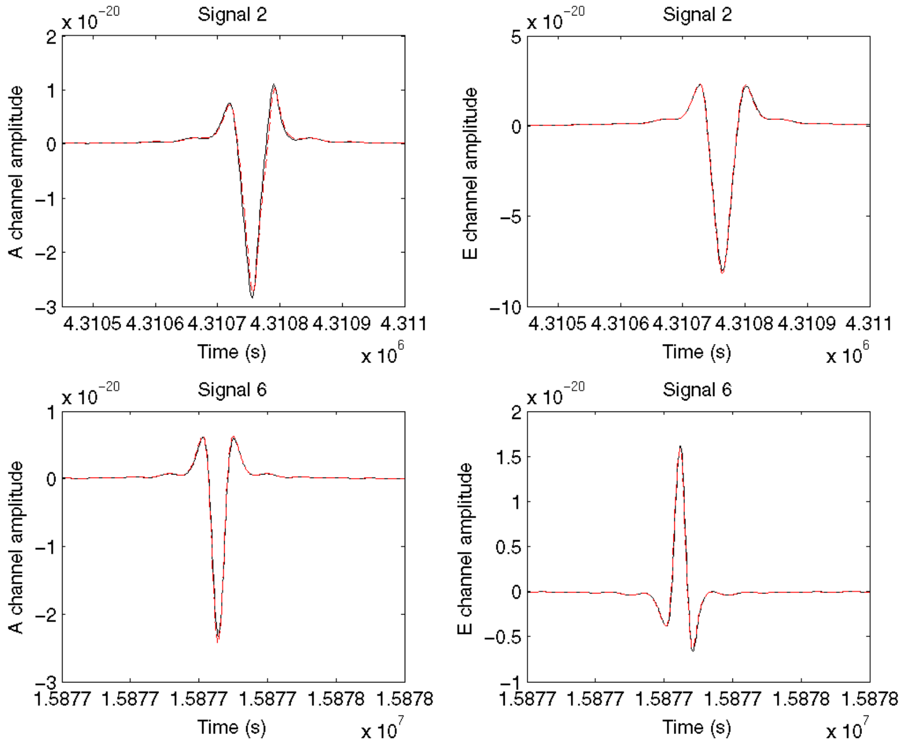


Figure 3.19: A comparison of the true and recovered waveforms in the A and E channels for two of the 22 signals. The true signal is in solid black and the recovered is in dashed red. These signals have a low value of f_{max}.

f_{max} show more discrepancies, as can be seen for signals 3 and 14 in Figure 3.20.

3.3.2.3 Full Galaxy

The final data set analysed consisted of instrumental noise, the full galaxy population of white dwarf binaries, and the cosmic strings we were looking for. The modulated foreground confusion noise model given in Equation (3.17) was used in this analysis, which was performed in the same way as for the previous two data sets. In this scenario, we do not expect the noise model to account for all contaminating sources and thus we will be presented with many false detections.

Due to the presence of loud signals across all frequencies in each data segment,

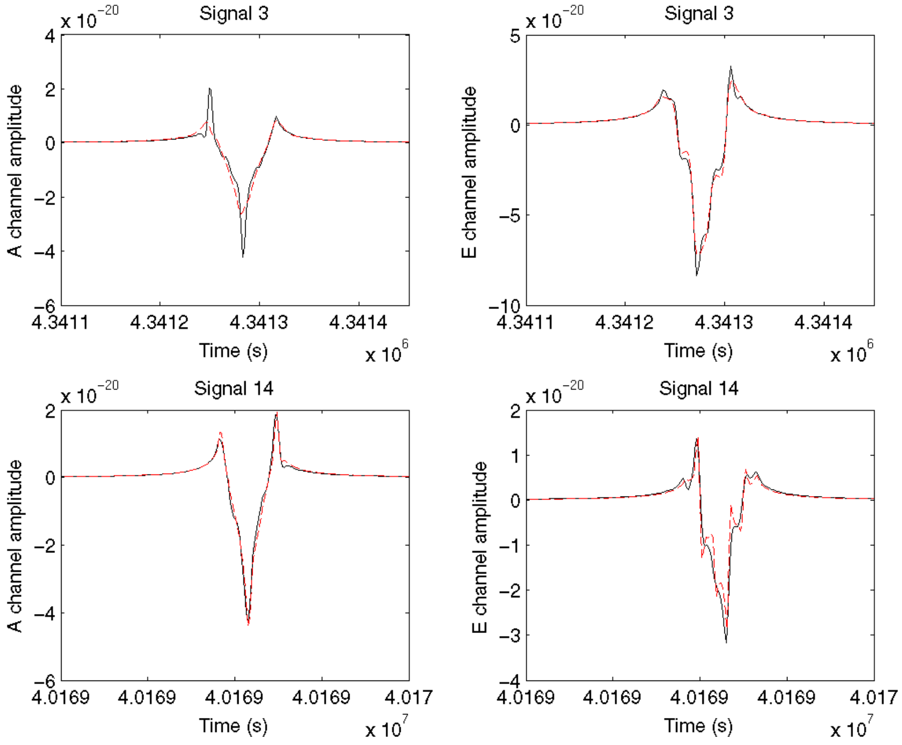


Figure 3.20: A comparison of the true and recovered waveforms in the A and E channels for two of the 22 signals. The true signal is in solid black and the recovered is in dashed red. These signals have a high value of f_{max}.

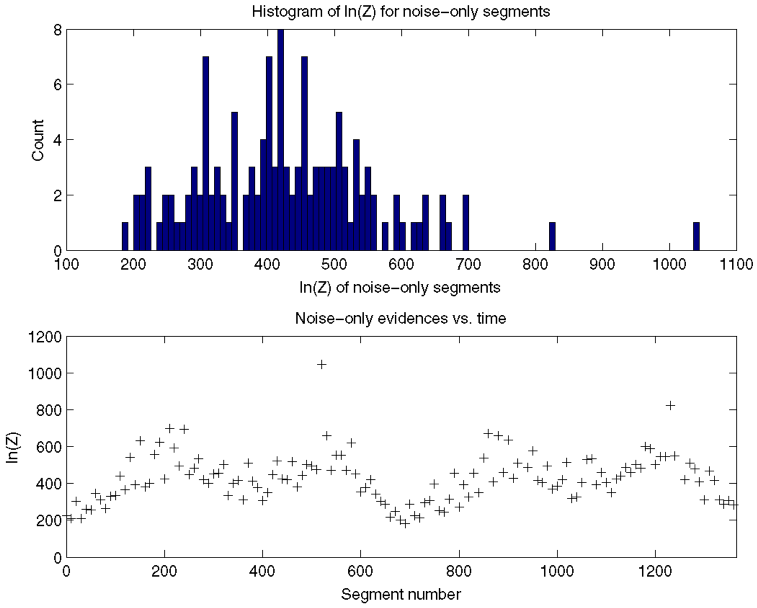


Figure 3.21: (Top) Histogram of global log-evidences for the segments that do not contain a cosmic string GW signal with the full galaxy population and modulated foreground noise model. (Bottom) The log-evidence values of noise-only segments over time (segment number).

evidences for all segments were significantly increased. In Figure 3.21 we show the histrogram and time scatter plot of log-evidences of segments that do not contain a cosmic string GW signal. There is clearly a much larger base level that will obscure finding true signals. The ROC along with true and false positive rates for detecting segments with signals present are shown in Figure 3.22. It is clear that some level of contamination will be present if we wish to set a threshold that does not exclude true cosmic string signals.

3.3.2.4 Summary

The results presented in this section clearly show that perfect identification and excellent recovery of cosmic string burst signals is possible in the presence of instrumental noise and reduced galactic foreground. The population of galactic WD binaries with

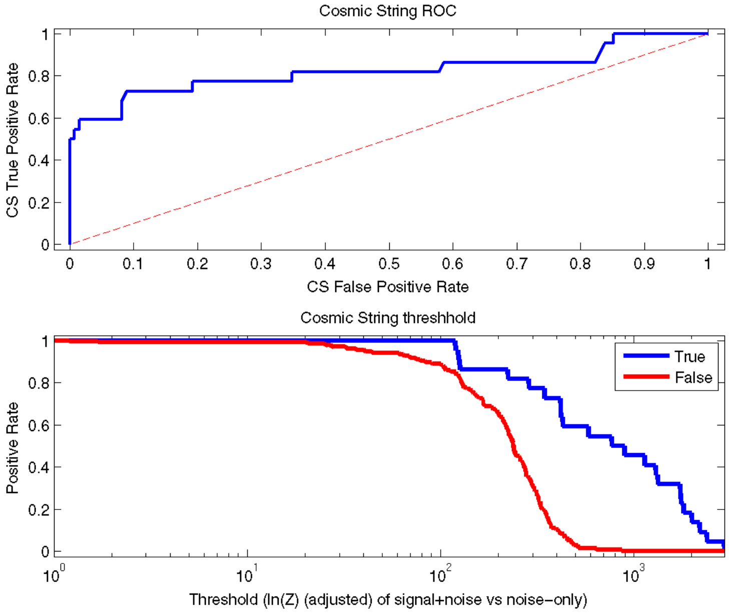


Figure 3.22: (Top) ROC for cosmic strings in instrumental noise with the full galactic population, showing true positive rate vs. false positive rate as the threshold varies. (Bottom) The true and false positive detection rates as a function of the log-evidence. $\ln(\mathcal{Z})$ is adjusted by adding a constant to all values so that the minimum is greater than 1.

the loudest signals removed forms a confusion foreground that, when appropriately modeled, can be accounted for and will not affect detection of cosmic string bursts. However, when the full WD population is left in the data, there is significant contamination and keeping this to a manageable amount will result in potentially missing some cosmic string signals. Spurious signals may be identified by comparing to either a sine-Gaussian or constant frequency model, using the Bayesian evidence to determine if one of these is preferred over the cosmic string burst model.

This demonstrates the need to identify and subtract out of the data the loudest WD binaries detected by LISA, or any similar detector, in order to be able to detect other signals underneath. Doing so should not be affected by cosmic strings still present in the data since the latter are transient while the former are present over the entire two year data span. A main caveat is that this analysis was performed without the presence of SMBH or EMRI signals, which are transient but significantly longer (typically of duration three months or longer) and potentially much louder than any cosmic string bursts. SMBH signals can have SNRs of thousands when coalescing during the observation period; EMRIs, however, build up an SNR of similar magnitude to cosmic strings (typically 10 to 100) but over a duration of months. Detection and removal of these sources in the presence of all others will need to be performed as well. We have shown, however, that cosmic string bursts can be detected even in the presence of the full galactic WD population and that the Bayesian evidence can be used as a detection threshold. Like any detection statistic, some false positives may be included and will require further analysis to be filtered out.

3.4 Discussion

The eLISA/NGO detector that is eventually launched will be able to observe multiple signals that evolve on very different time scales. From bursts that last only a few minutes to supermassive black hole binaries that coalesce over months to galactic white dwarf binaries that are essentially monochromatic over a two year observation, these signals will need to be detected and characterised from a single data stream. Data analysis methods investigated here that utilise nested sampling will be essential in obtaining the most information from the data as possible.

Segmenting data to look for transients will enable us to separate these short-lived signals from the more constant sources that require time to build up a significant SNR. Modeling of other sources present in the data in the noise estimation will further improve our ability to recover the signals we are looking for. In searching for cosmic string burst signals with a reduced WD binary foreground, the quality of parameter estimation was not adversely affected as the additional sources were well-modeled by the noise PSD and the signals of interest were of sufficiently high SNR. Analysing frequency bands for near-monochromatic signals such as white dwarf binaries can yield many detections that will not be influenced by noise at other frequencies. Subtracting the loudest signals and re-analysing should return a high rate of detection. For data with other signals present, analysing the two years of observation separately and requiring coherent detection should reduce the effect of transient sources passing through the frequency bands of interest.

The source environment for eLISA/NGO will require many diverse techniques for separating out the different signal types which can also vary over orders of magnitude in SNR. Some of these methods have been explored here and present promising results for future data analysis pipeline development.

Chapter 4

An Investigation into the MOPED Algorithm

> Science is what we understand well enough to explain to a computer.
>
> Donald Knuth

The analysis of gravitational wave data involves measuring a signal that is given by thousands of data points but is only defined by 17 or fewer individual parameters. This drastic increase in the number of values to be measured slows down likelihood calculations and requires large amounts of computer memory. Additionally, estimation and inversion of noise covariance matrices needs increased computational resources. Simplifying assumptions are used to reduce these requirements, but removing the need analytically without losing information would be preferable. With this goal in mind, I investigated implementing the Multiple Optimised Parameter Estimation and Data compression (MOPED) algorithm for use in gravitational wave burst searches for LISA.

I begin this chapter by introducing the general likelihood function and a simple method of decreasing its complexity in Section 4.1. I then introduce MOPED in Section 4.2 and define the MOPED likelihood function along with comments on the potential speed benefits of MOPED. In Section 4.3 I introduce my astrophysical scenario of gravitational wave burst signals where we found that MOPED did not accurately portray the true likelihood function. In Section 4.4 I expand upon this scenario to another where MOPED is found to work and to two other scenarios where it does not. I present a discussion of the criteria under which I believe MOPED will accurately represent the likelihood in Section 4.5, as well as a discussion of an implementation of a solution provided by [171]. Work presented in this chapter has been published in [172].

4.1 The Likelihood Function

We begin by defining our data as a vector, $\mathbf{x}$. Our model describes $\mathbf{x}$ by a signal plus random noise,

$$\mathbf{x} = \mathbf{u}(\boldsymbol{\theta}_T) + \mathbf{n}(\boldsymbol{\theta}_T), \tag{4.1}$$

where the signal is given by a vector $\mathbf{u}(\boldsymbol{\theta})$ that is a function of the set of parameters $\boldsymbol{\theta} = \{\theta_i\}$ defining our model, and the true parameters are given by $\boldsymbol{\theta}_T$. The noise is assumed to be Gaussian with zero mean and noise covariance matrix $\mathcal{N}_{jk} = \langle n_j n_k \rangle$, where the angle brackets indicate an ensemble average over noise realisations. In general this matrix may also be a function of the parameters $\boldsymbol{\theta}$, as in the case of galaxy spectra explored in [173] where noise in a frequency bin depends on the amplitude of the signal in the bin; however, in the examples explored in this chapter we will not be considering this fully general case in a GW context. The full likelihood for N data points in $\mathbf{x}$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}\sqrt{|\mathcal{N}(\boldsymbol{\theta})|}} \exp\left\{ -\frac{1}{2}[\mathbf{x} - \mathbf{u}(\boldsymbol{\theta})]^{\mathrm{T}} \mathcal{N}^{-1}(\boldsymbol{\theta}) [\mathbf{x} - \mathbf{u}(\boldsymbol{\theta})] \right\}. \tag{4.2}$$

At each point, then, this requires the calculation of the determinant and inverse of an $N \times N$ matrix. Both scale as N^3, so even for smaller datasets this can become cumbersome.

4.1.1 A Simple Iterative Approach

One can perform a simple approximation to avoid performing the $O(N^3)$ matrix operations repeatedly by following the steps below.

1. Choose a fiducial point from the prior, $\boldsymbol{\theta}_F$.

2. Calculate $\mathcal{N}^{-1}(\boldsymbol{\theta}_F)$ and $|\mathcal{N}(\boldsymbol{\theta}_F)|$.

3. Locate the peak of the posterior probability distribution of $\boldsymbol{\theta}$ assuming constant $\mathcal{N}^{-1}(\boldsymbol{\theta}_F)$ and $|\mathcal{N}(\boldsymbol{\theta}_F)|$ instead of evaluating these at each point $\boldsymbol{\theta}$.

4. Take the posterior maximum as the new fiducial point and repeat steps 2 and 3. Continue doing so until convergence is reached; the final posterior distribution is the solution.

This method will provide a quick approximation of the true likelihood function and only requires one matrix inversion and determinant calculation per iteration.

4.2 The MOPED Algorithm

Multiple Optimised Parameter Estimation and Data compression (MOPED; [173]) is a patented algorithm for the compression of data and the speeding up of the evaluation of likelihood functions in astronomical data analysis and beyond. It becomes particularly useful when the noise covariance matrix is dependent upon the parameters of the model and so must be calculated and inverted at each likelihood evaluation. However, such benefits come with limitations. Since MOPED only guarantees maintaining the Fisher matrix of the likelihood at the chosen fiducial point, multimodal and some degenerate distributions will present a problem. In this chapter I report on some of the limitations of the application of the MOPED algorithm. In the cases where MOPED does accurately represent the likelihood function, however, its compression of the data and consequent much faster likelihood evaluation does provide orders of magnitude improvement in runtime. In [173], the authors demonstrate the method by analysing the spectra of galaxies and in [174] they illustrate the benefits of MOPED for estimation of the CMB power spectrum. The problem of “badly” behaved likelihoods was found by [171] for the problem of light transit analysis; nonetheless, the authors present a solution that still allows MOPED to provide a large speed increase.

4.2.1 Data Compression with MOPED

Full details of the MOPED method are given in [173], here we will only present a limited introduction.

MOPED allows one to eliminate the need for the full matrix inversion by compressing the N data points in $\mathbf{x}$ into M data values, one for each parameter of the model. Additionally, MOPED creates the compressed data values such that they are independent and have unit variance, further simplifying the subsequent likelihood calculation to an $O(M)$ operation. Typically, $M \ll N$ so this gives us a significant increase in speed. A single compression is done on the data, $\mathbf{x}$, and then again for each point in parameter space where we wish to compute the likelihood. The compression is done by generating a set of weighting vectors, $\mathbf{b}_i(\boldsymbol{\theta}_F)$ $(i = 1 \ldots M)$, from which we can generate a set of MOPED components from the theoretical model and data,

$$y_i(\boldsymbol{\theta}_F) \equiv \mathbf{b}_i(\boldsymbol{\theta}_F) \cdot \mathbf{x} = \mathbf{b}_i^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathbf{x}. \tag{4.3}$$

Note that the weighting vectors must be computed at some assumed fiducial set of parameter values, $\boldsymbol{\theta}_F$. The only choice that will truly maintain the likelihood peak is when the fiducial parameters are the true parameters, but obviously we will not know these in advance for real analysis situations. Thus, we can choose our fiducial model to be anywhere and iterate the procedure, taking our likelihood peak in one iteration as the fiducial model for the next iteration. This process will converge very quickly, and may not even be necessary in some instances. For our later examples, since we do know the true parameters we will use these as the fiducial ($\boldsymbol{\theta}_F = \boldsymbol{\theta}_T$) in order to remove this as a source of confusion (all equations, however, are written for the more general case). Note that the true parameters, $\boldsymbol{\theta}_T$, will not necessarily coincide with the peak $\hat{\boldsymbol{\theta}}_O$ of the original likelihood or the peak $\hat{\boldsymbol{\theta}}_M$ of the MOPED likelihood (see below).

The weighting vectors must be generated in some order so that each subsequent vector (after the first) can be made orthogonal to all previous ones using Gram-Schmidt orthonormalisation. We begin by writing the derivative of the model with respect to the i^{th} parameter as $\left.\frac{\partial \mathbf{u}}{\partial \theta_i}\right|_{\boldsymbol{\theta}_F} = \mathbf{u}_{,i}(\boldsymbol{\theta}_F)$. This gives us a solution for the first weighting vector, properly normalised, of

$$\mathbf{b}_1(\boldsymbol{\theta}_F) = \frac{\mathcal{N}^{-1}(\boldsymbol{\theta}_F)\mathbf{u}_{,1}(\boldsymbol{\theta}_F)}{\sqrt{\mathbf{u}_{,1}^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathcal{N}^{-1}(\boldsymbol{\theta}_F)\mathbf{u}_{,1}(\boldsymbol{\theta}_F)}}. \tag{4.4}$$

The first compressed value is $y_1(\boldsymbol{\theta}_F) = \mathbf{b}_1^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathbf{x}$ and will weight up the data combination most sensitive to the first parameter. The subsequent weighting vectors are made orthogonal by subtracting out parts that are parallel to previous vectors and are then normalised. The resulting formula for the remaining weighting vectors is

$$\mathbf{b}_m = \frac{\mathcal{N}^{-1}\mathbf{u}_{,m} - \sum_{q=1}^{m-1}(\mathbf{u}_{,m}^{\mathrm{T}}\mathbf{b}_q)\mathbf{b}_q}{\sqrt{\mathbf{u}_{,m}^{\mathrm{T}}\mathcal{N}^{-1}\mathbf{u}_{,m} - \sum_{q=1}^{m-1}(\mathbf{u}_{,m}^{\mathrm{T}}\mathbf{b}_q)^2}}, \tag{4.5}$$

where $m = 2 \dots M$ and all values are evaluated at $\boldsymbol{\theta}_F$. Weighting vectors generated with Equations (4.4) and (4.5) form an orthonormal set with respect to the noise covariance matrix so that

$$\mathbf{b}_i^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathcal{N}(\boldsymbol{\theta}_F)\mathbf{b}_j(\boldsymbol{\theta}_F) = \delta_{ij}. \tag{4.6}$$

This means that the noise covariance matrix of the compressed values y_i is the identity, which significantly simplifies the likelihood calculation. The new likelihood function is given by

$$\mathcal{L}_{\mathrm{MOPED}}(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{M}\left(y_i(\boldsymbol{\theta}_F) - \langle y_i\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)\right)^2\right\}, \tag{4.7}$$

where $y_i(\boldsymbol{\theta}_F) = \mathbf{b}_i^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathbf{x}$ represents the i^{th} compressed data value and $\langle y_i\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F) = \mathbf{b}_i^{\mathrm{T}}(\boldsymbol{\theta}_F)\mathbf{u}(\boldsymbol{\theta})$ represents the i^{th} compressed signal value. This is a much easier likelihood to calculate and is time-limited by the generation of a new signal template instead of the inversion of the noise covariance matrix. The peak value of the MOPED likelihood function is not guaranteed to be unique as there may be other points in the original parameter space that map to the same point in the compressed parameter space; this is a characteristic that we will investigate.

The MOPED likelihood will closely approximate the true likelihood function. Most importantly, the peak value will remain the same when the true parameters (or very similar) are used for the fiducial. Additionally, in the limiting case where the noise covariance matrix does not depend on the parameters, the Fisher matrix describing the curvature of the likelihood surface at the peak is identical to the original. However, properties of the likelihood away from the peak value are not guaranteed to be so well-behaved.

4.2.2 Speed Gains with MOPED

MOPED implicity assumes that the covariance matrix, $\mathcal{N}$, is independent of the parameters $\boldsymbol{\theta}$. With this assumption it is appropriate to compare MOPED to the approximate method described in Section 4.1.1. Therefore, a full likelihood calculation with N data points would require an $O(N^2)$ operation at each point in parameter space (or $O(N)$ if $\mathcal{N}$ is diagonal). In MOPED, however, the compression of the theoretical data is an $O(NM)$ linear operation followed by an $O(M)$ misfit calculation, leading to an overall complexity of $O(NM)+O(M) \simeq O(NM)$ for each likelihood calculation. Thus, one loses on speed if $\mathcal{N}$ is diagonal but gains by a factor of N/M otherwise. For the data sets we will analyze, $N/M > 100$. We begin, though, by assuming a diagonal $\mathcal{N}$ for simplicity, recognizing that this will cause a speed reduction but that it is a necessary step before addressing a more complex noise model.

One can iterate the parameter estimation procedure if necessary, taking the maximum likelihood or posterior point found as the new fiducial and re-analyzing (perhaps with tighter prior constraints) as in the approximate method suggested in the previous section. This procedure is recommended for MOPED in [173] but is not always found to be necessary.

MOPED has the additional benefit that the weighting vectors, $\mathbf{b}_i$, need only be computed once provided the fiducial model parameters are kept constant over the analysis of different data sets. Computed compressed parameters, $\langle y_i \rangle$, can also be stored for reference and require significantly less memory than storing the entire theoretical data set.

4.3 Simple Example With One Parameter

In order to demonstrate some of the limitations of the applicability of the MOPED algorithm, we will consider a few test cases. These originate in the context of gravitational wave data analysis for LISA since it is in this scenario that we first discovered such cases of failure. The full problem is seven-dimensional parameter estimation, but we have fixed most of these variables to their known true values in the simulated data set in order to create a lower-dimensional problem that is simpler to analyse.

We consider the case of a sine-Gaussian burst signal present in the LISA detector. The short duration of the burst with respect to the motion of LISA allows us to use the static approximation to the response. As in Section 3.1.1.2, the frequency-space waveform is described by [149]

$$\tilde{h}(f) = A\tfrac{Q}{f}\exp\left\{-\tfrac{1}{2}Q^2(\tfrac{f-f_c}{f_c})^2\right\}\exp(2\pi \imath t_0 f). \tag{4.8}$$

Here A is the dimensionless amplitude factor; Q is the width of the Gaussian envelope of the burst measured in cycles; f_c is the central frequency of the oscillation being modulated by the Gaussian envelope; and t_0 is the central time of arrival of the burst. This waveform is further multiplied by a projection factor dependent on the sky position of the burst source, θ and ϕ, and the burst polarisation, ψ, with respect to the detector. The one-sided noise power spectral density of the LISA detector is given by [149]

$$\begin{aligned} S_h(f) &= 16\sin^2(2\pi f t_L)\times \\ &\quad \left(2\left(1+\cos(2\pi f t_L)+\cos^2(2\pi f t_L)\right)S_{\rm pm}(f)\right. \\ &\quad \left.+(1+\cos(2\pi f t_L)/2)\,S_{\rm sn}f^2\right), & (4.9) \\ S_{\rm pm}(f) &= \left(1+\left(\frac{10^{-4}\rm Hz}{f}\right)^2\right)\frac{S_{\rm acc}}{f^2}, & (4.10) \end{aligned}$$

where $t_L = 16.678$s is the light travel time along one arm of the LISA constellation, $S_{\rm acc} = 2.5\times 10^{-48}\rm Hz^{-1}$ is the proof mass acceleration noise and $S_{\rm sn} = 1.8\times 10^{-37}\rm Hz^{-1}$ is the shot noise. This is independent of the signal parameters and all frequencies are independent of each other, so the noise covariance matrix is constant and diagonal. This less computationally expensive example allows us to show some interesting properties.

We begin by taking the one-dimensional case where the only unknown parameter of the model is the central frequency of the oscillation, f_c. We set $Q = 5$ and $t_0 = 10^5$s; we then analyze a 2048s segment of LISA data, beginning at $t = 9.9\times 10^4$s, sampled at a 1s cadence. For this example, the data were generated with random noise (following the LISA noise power spectrum) at an SNR of ~ 34 with $f_{c,T} = 0.1$ Hz (thus $f_{c,F} = 0.1$ Hz for MOPED). The prior range on the central frequency is uniform from 10^{-3} Hz to 0.5 Hz. 10,000 samples uniformly spaced in f_c were taken and their likelihoods calculated using both the original and MOPED likelihood functions. The log-likelihoods are shown in Figure 4.1. Note that the absolute magnitudes are not important but the relative values within each plot are meaningful. Both the original and MOPED likelihoods have a peak close to the input value $f_{c,T}$.

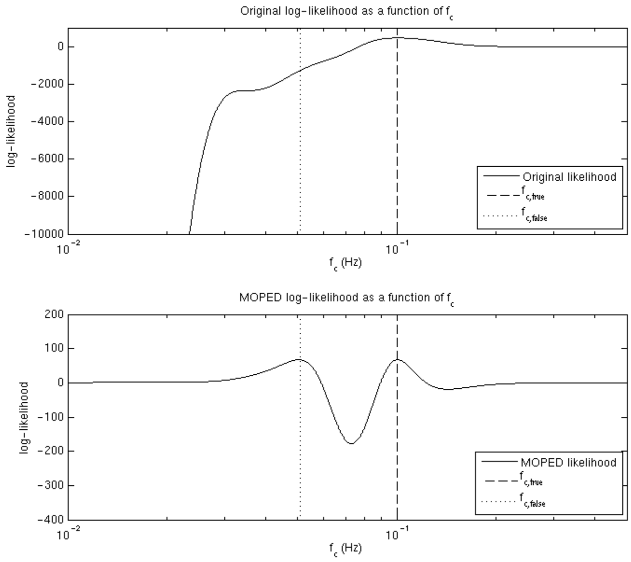


Figure 4.1: The original and MOPED log-likelihoods as a function of f_c for the chosen template.

We see, however, that in going from the original to MOPED log-likelihood, the latter also has a second peak of equal height at an incorrect f_c. To see where this peak comes from, we look at the values of the compressed parameter $\langle y_1 \rangle (f_c; f_{c,F})$ as it varies with respect to f_c as shown in Figure 4.2. The true compressed value peak occurs at $f_c \simeq 0.1$ Hz, where $y_1(f_{c,F}) = \langle y_1 \rangle (f_c; f_{c,F})$. However, we see that there is another frequency that yields this exact same value of $\langle y_1 \rangle (f_c; f_{c,F})$; it is at this frequency that the second, incorrect peak occurs. By creating a mapping from f_c to $\langle y_1 \rangle (f_c; f_{c,F})$ that is not one-to-one, MOPED has created the possibility for a second solution that is indistinguishable in likelihood from the correct one. This is a very serious problem for parameter estimation.

4.4 Recovery in a 2 Parameter Case

Interestingly, we find that even when MOPED fails in a one-parameter case, adding a second parameter may actually rectify the problem, although not necessarily. If we now allow the width of the burst, Q, to be a variable parameter, there are now two

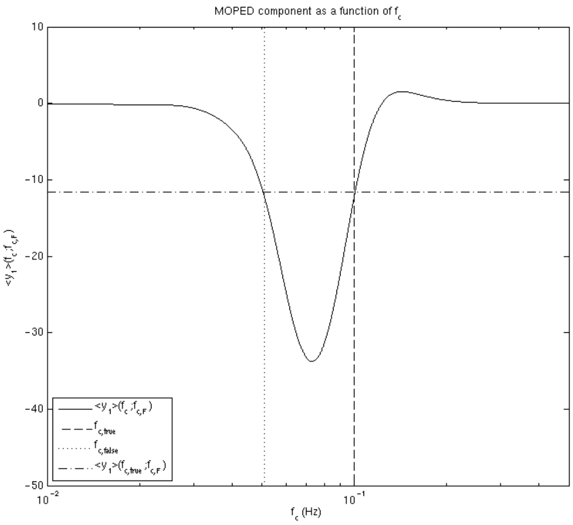


Figure 4.2: The value of the MOPED compressed parameter as a function of the original frequency parameter.

orthogonal MOPED weighting vectors that need to be calculated. This gives us two compressed parameters for each pair of f_c and Q. Each of these may have its own unphysical degeneracies, but in order to give an unphysical mode of equal likelihood to the true peak, these degeneracies will need to coincide. In Figure 4.3, we plot the loci in (f_c, Q) space where $\langle y_i \rangle (\boldsymbol{\theta}; \boldsymbol{\theta}_F) = \langle y_i \rangle (\hat{\boldsymbol{\theta}}_M; \boldsymbol{\theta}_F)$ as $\boldsymbol{\theta}$ ranges over f_c and Q values. We can clearly see the degeneracies present in either variable, but since these only overlap at the one location, near to where the true peak is, there is no unphysical second mode in the MOPED likelihood. Hence, when we plot the original and MOPED log-likelihoods in Figure 4.4, although the behaviour away from the peak has changed, the peak itself remains in the same location and there is no second mode.

Adding more parameters, however, does not always improve the situation as the signal varies in different ways for each. We now consider the case where Q is once again fixed to its true value and we instead make the polarisation of the burst, ψ, a variable parameter. There are degeneracies in both of these parameters and in Figure 4.5 we plot the loci in (f_c, ψ)-space where the compressed values are each equal to the value at the maximum MOPED likelihood point. These two will necessarily intersect at the maximum likelihood solution, near the true value ($f_c = 0.1$ Hz and $\psi = 1.3$ rad),

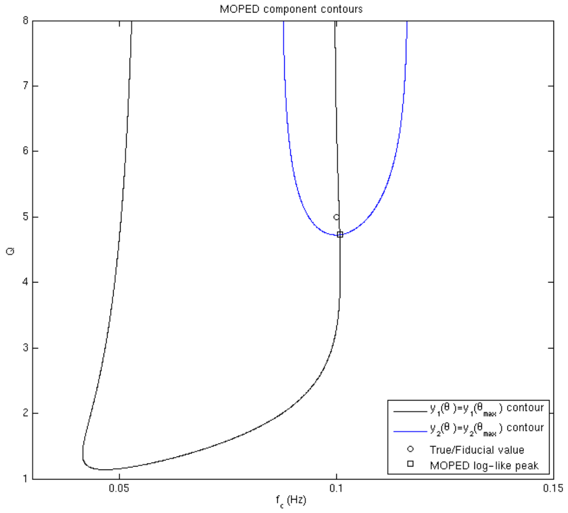


Figure 4.3: Loci of $\langle y_1 \rangle (\boldsymbol{\theta};\boldsymbol{\theta}_F) = \langle y_1 \rangle (\hat{\boldsymbol{\theta}}_M;\boldsymbol{\theta}_F)$ and $\langle y_2 \rangle (\boldsymbol{\theta};\boldsymbol{\theta}_F) = \langle y_2 \rangle (\hat{\boldsymbol{\theta}}_M;\boldsymbol{\theta}_F)$ in the parameter space $\boldsymbol{\theta} = \{f_c, Q\}$. The one intersection is the true maximum likelihood peak.

but a second intersection is also apparent. This second intersection will have the same likelihood as the maximum and be another mode of the solution. However, as we can see in the left plot of Figure 4.6, this is not a mode of the original likelihood function. MOPED has, in this case, created a second mode of equal likelihood to the true peak.

For an even more extreme scenario, we now fix to the true ψ and allow the time of arrival of the burst t_0 to vary (we also define $\Delta t_0 = t_0 - t_{0,T}$). In this scenario, the loci in $(f_c, \Delta t_0)$-space where $\langle y_i \rangle (\boldsymbol{\theta};\boldsymbol{\theta}_F) = \langle y_i \rangle (\hat{\boldsymbol{\theta}}_M;\boldsymbol{\theta}_F)$ are much more complicated. Thus, we have many more intersections of the two loci than just at the likelihood peak near the true values and MOPED creates many alternative modes of likelihood equal to that of the original peak. This is very problematic for parameter estimation. In Figure 4.7 we plot these loci so the multiple intersections are apparent. Figure 4.8 shows the original and MOPED log-likelihoods, where we can see the single peak becoming a set of peaks.

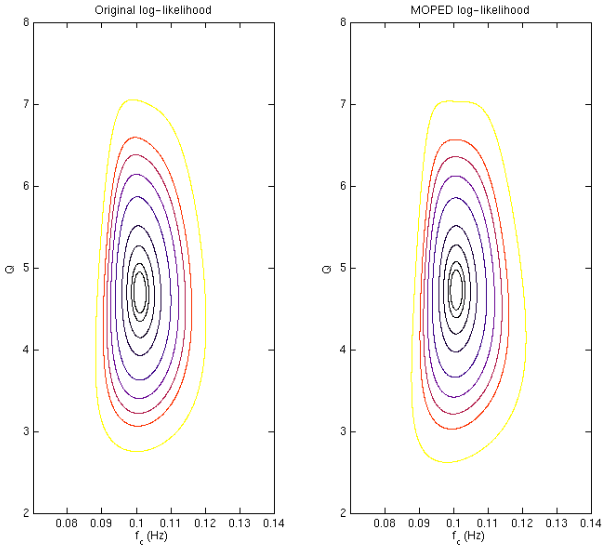


Figure 4.4: Contours of the original and MOPED log-likelihoods (left and right, respectively). The MOPED likelihood has been multiplied by a constant factor so that its peak value is equal to the peak of the original likelihood. Contours are at 1, 2, 5, 10, 20, 30, 40, 50, 75, and 100 log-units below the peak going from the inside to outside.

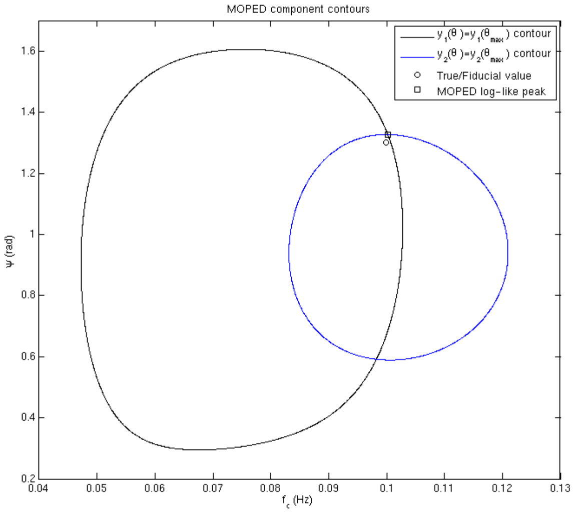


Figure 4.5: Loci of $\langle y_1\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)=\langle y_1\rangle(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}_F)$ and $\langle y_2\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)=\langle y_2\rangle(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}_F)$ in the parameter space $\boldsymbol{\theta}=\{f_c,\psi\}$.

4.5 Discussion

What we can determine from the previous two sections is a general rule for when MOPED will generate additional peaks in the likelihood function equal in magnitude to the true one. For an M-dimensional model, if we consider the $(M-1)$-dimensional hyper-surfaces where $\langle y_i\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)=\langle y_i\rangle(\hat{\boldsymbol{\theta}}_M;\boldsymbol{\theta}_F)$, then any point where these M hyper-surfaces intersect will yield a set of $\boldsymbol{\theta}$-parameter values with likelihood equal to that at the peak near the true values. We expect that there will be at least one intersection at the likelihood peak corresponding to approximately the true solution. However, we have shown that other peaks can exist as well. The set of intersections of contour surfaces will determine where these additional peaks are located. This degeneracy will interact with the model's intrinsic degeneracy, as any degenerate parameters will yield the same compressed values for different original parameter values.

Unfortunately, there is no simple way to find these contours other than by mapping out the $\langle y_i\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)$ values, which is equivalent in procedure to mapping the MOPED likelihood surface. The benefit comes when this procedure is significantly faster than mapping the original likelihood surface. The mapping of $\langle y_i\rangle(\boldsymbol{\theta};\boldsymbol{\theta}_F)$ can even be per-

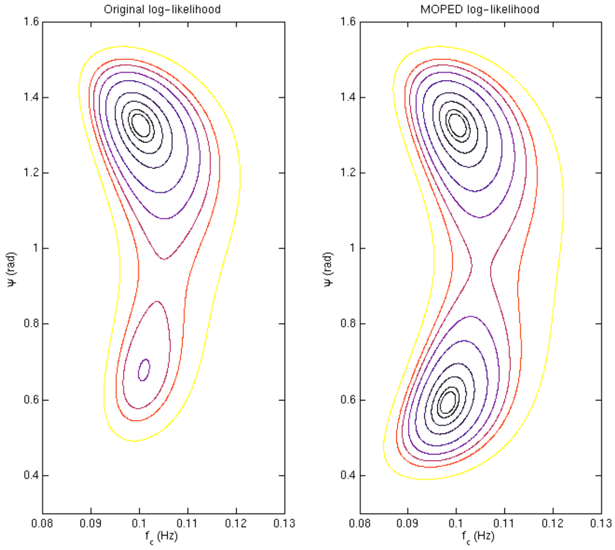


Figure 4.6: Contours of the original and MOPED log-likelihoods (left and right, respectively). The MOPED likelihood has been multiplied by a constant factor so that its peak value is equal to the peak of the original likelihood. Contours are at 1, 2, 5, 10, 20, 30, 40, 50, 75, and 100 log-units below the peak going from the inside to outside.

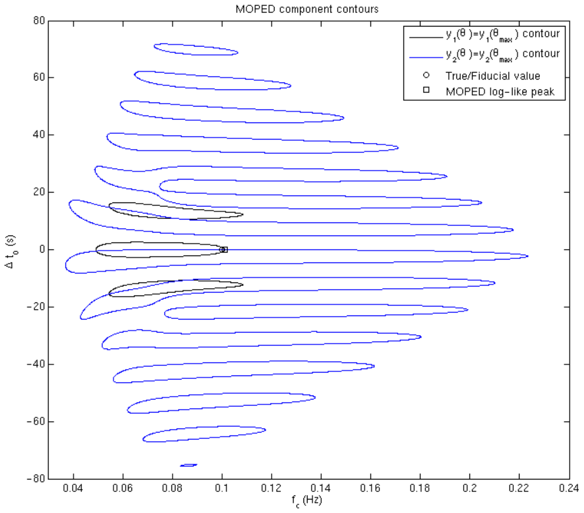


Figure 4.7: Loci of $\langle y_1 \rangle (\boldsymbol{\theta};\boldsymbol{\theta}_F) = \langle y_1 \rangle (\hat{\boldsymbol{\theta}};\boldsymbol{\theta}_F)$ and $\langle y_2 \rangle (\boldsymbol{\theta};\boldsymbol{\theta}_F) = \langle y_2 \rangle (\hat{\boldsymbol{\theta}};\boldsymbol{\theta}_F)$ in the parameter space $\boldsymbol{\theta} = \{f_c, \Delta t_0\}$. We can see many intersections here other than the true peak.

formed before data is obtained or used, if the fiducial model is chosen in advance; this allows us to analyse properties of the MOPED compression before applying it to data analysis. If the MOPED likelihood is mapped and there is only one contour intersection, then we can rely on the MOPED algorithm and will have saved a considerable amount of time, since MOPED has been demonstrated to provide speed-ups of a factor of up to 10^7 in [174]. However, if there are multiple intersections then it is necessary to map the original likelihood to know if they are due to degeneracy in the model or were created erroneously by MOPED. In this latter case, the time spent finding the MOPED likelihood surface can be much less than that which will be needed to map the original likelihood, so relatively little time will have been wasted. If any model degeneracies are known in advance, then we can expect to see them in the MOPED likelihood and will not need to find the original likelihood on their account.

One possible way of determining the validity of degenerate peaks in the MOPED likelihood function is to compare the original likelihoods of the peak parameter values with each other. By using the maximum MOPED likelihood point found in each mode and evaluating the original likelihood at this point, we can determine which one is

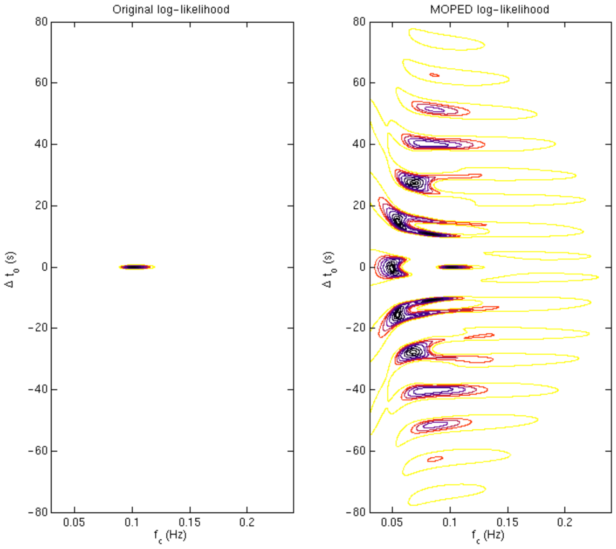


Figure 4.8: Contours of the original and MOPED log-likelihoods (left and right, respectively). The MOPED likelihood has been multiplied by a constant factor so that its peak value is equal to the peak of the original likelihood. Contours are at 1, 2, 5, 10, 20, 30, 40, 50, 75, and 100 log-units below the peak going from the inside to outside.

correct. The correct peak and any degeneracy in the original likelihood function will yield similar values to one another, but a false peak in the MOPED likelihood will have a much lower value in the original likelihood and can be ruled out. This means that a Bayesian evidence calculation cannot be obtained from using the MOPED likelihood; however, the algorithm was not designed to be able to provide this.

The solution for this problem presented in [171] is to use multiple fiducial models to create multiple sets of weighting vectors. The log-likelihood is then averaged across these choices. Each different fiducial will create a set of likelihood peaks that include the true peaks and any extraneous ones. However, the only peaks that will be consistent between fiducials are the correct ones. Therefore, the averaging maintains the true peak(s) but decreases the likelihood at incorrect values. This was tested with 20 random choices for $\boldsymbol{\theta}_F$ for the two-parameter models presented and was found to leave only the true peak at the maximum likelihood value. Other, incorrect, peaks are still present, but at log-likelihood values five or more units below the true peak. When applied to the full seven parameter model, however, the SNR threshold for signal recovery is increased significantly, from $\simeq 10$ to $\simeq 30$.

The MOPED algorithm for reducing the computational expense of likelihood functions can, in some examples, be extremely useful and provide orders of magnitude of improvement. However, as we have shown, this is not always the case and MOPED can produce erroneous peaks in the likelihood that impede parameter estimation. It is important to identify whether or not MOPED has accurately portrayed the likelihood function before using the results it provides. Some solutions to this problem have been presented and tested but further implementation was not pursued.

Chapter 5

Artificial Neural Networks

> The computer was born to solve problems that did not exist before.
>
> Bill Gates

Many calculations in gravitational wave physics, and astrophysics in general, are very computationally expensive. If a calculation needs to be performed a large number of times then it can slow down simulations and analysis significantly. One example in gravitational wave physics is the calculation of waveforms for binary black hole mergers where the full spin dynamics of both black holes are considered. A single waveform can take on the order of seconds to calculate, or even longer in the case of numerical relativity simulations. For this and other problems, we set out to develop a generic algorithm for the training of artificial neural networks in order to facilitate and expedite these repeated difficult computations.

Artificial neural networks (NNs) are a method of computation loosely based on the structure of a brain. They consist of a group of interconnected nodes, which process information that they receive and then pass this product along to other nodes via weighted connections. We will consider only feed-forward NNs, for which the structure is directed; an input layer of nodes passes information to an output layer via zero, one, or many "hidden" layers in between. A network is able "learn" a relationship between inputs and outputs given a set of training data and can then make predictions of the outputs for new input data. Further introduction can be found in [175].

In section 5.1 I describe the general structure of NNs and how they may be trained on a specific problem. Section 5.2 will then detail the procedure used to train NNs to solve a particular task. This is then applied to some toy examples in Section 5.3. Applying NNs to learning how to classify images of handwritten digits is described in Section 5.4 and to the problem of measuring the shape of sheared, blurred, and pixelated images of galaxies in Section 5.5. Work in this chapter was performed in collaboration with Farhan Feroz.

5.1 Network Structure

A multilayer perceptron artificial neural network (NN) is the simplest type of network and consists of ordered layers of perceptron nodes that pass scalar values from one layer to the next. The perceptron is the simplest kind of node, and maps an input vector $\mathbf{x} \in \Re^n$ to a scalar output $f(\mathbf{x};\mathbf{w},\theta)$ via

$$f(\mathbf{x};\mathbf{w},\theta) = \theta + \sum_{i=1}^{n} w_i x_i, \tag{5.1}$$

where $\mathbf{w} = \{w_i\}$ and θ are the parameters of the perceptron, called the “weights” and “bias”, respectively. For a 3-layer NN, which consists of an input layer, a hidden layer, and an output layer as shown in Figure 5.1, the outputs of the nodes in the hidden and output layers are given by the following equations:

$$\text{hidden layer: } h_j = g^{(1)}(f_j^{(1)});\ f_j^{(1)} = \theta_j^{(1)} + \sum_l w_{jl}^{(1)} x_l, \tag{5.2}$$

$$\text{output layer: } y_i = g^{(2)}(f_i^{(2)});\ f_i^{(2)} = \theta_i^{(2)} + \sum_j w_{ij}^{(2)} h_j, \tag{5.3}$$

where l runs over input nodes, j runs over hidden nodes, and i runs over output nodes. The functions $g^{(1)}$ and $g^{(2)}$ are called activation functions and must be bounded, smooth, and monotonic for our purposes. We use $g^{(1)}(x) = 1/(1+e^{-x}) = \text{sig}(x)$ (sigmoid) and $g^{(2)}(x) = x$; the non-linearity of $g^{(1)}$ is essential to allowing the network to model non-linear functions.

The weights and biases are the values we wish to determine in our training (described in Section 5.2). As they vary, a huge range of non-linear mappings from inputs to outputs is possible. In fact, a universal approximation theorem [176] states that a

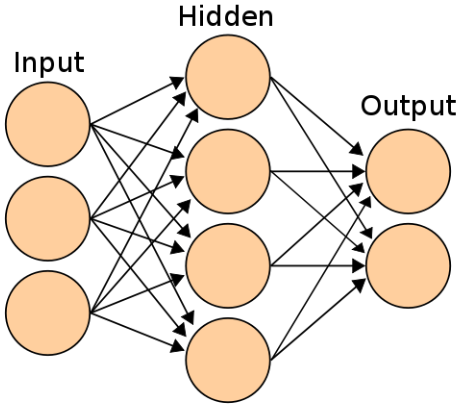


Figure 5.1: A 3-layer neural network with 3 inputs, 4 hidden nodes, and 2 outputs. Image courtesy of Wikimedia Commons.

NN with three or more layers can approximate any continuous function as long as the activation function is locally bounded, piecewise continuous, and not a polynomial (hence our use of sigmoid $g^{(1)}$, although other functions would work just as well, such as a tanh). By increasing the number of hidden nodes, we can achieve more accuracy at the risk of overfitting to our training data. To expand the NN to include more hidden layers, we mirror Equation (5.2) for each connection from one hidden layer to the next, each time using the same $g^{(1)}$ activation function. The final hidden layer will connect to the output layer with Equation (5.3).

5.1.1 Choosing the Number of Hidden Layer Nodes

An important choice when training a NN is the number of hidden layer nodes to use. The optimal number and organisation into one or more layers is a complex relationship between the number of training data points, the number of inputs and outputs, and the complexity of the function to be trained. Choosing too few nodes will mean that the NN is unable to learn the relationship to the highest possible accuracy; choosing too many will increase the risk of overfitting to the training data and will also slow down the training process. As a general rule, we find that a NN should not need more hidden nodes than the number of training data values used (each with a vector of inputs and outputs).

We choose to use 5–10 nodes in the hidden layer in our toy examples (see Section 5.3). These choices allow the network to model the complexity of the function without unnecessary work. In practice, it will be better to slightly over-estimate the number of hidden nodes required. There are checks built in to prevent over-fitting (described in Section 5.2) and the additional training time will not be a large penalty if an optimal network can be obtained in an early attempt. The optimal NN structure can be determined by comparing the Bayesian evidence of different trained NNs as calculated in Section 5.2.6.

5.2 Network Training

In training a NN, we wish to find the optimal set of network weights and biases that maximise the accuracy of predicted outputs. However, we must be careful to avoid overfitting to our training data at the expense of making predictions for input values the network has not been trained on. The NN can be trained from a random initial state or can be pre-trained, a procedure which will be discussed in Section 5.2.5. But first we will discuss the general process of NN training. The set of training data inputs and outputs, $\mathcal{D} = \{\mathbf{x}^{(k)}, \mathbf{t}^{(k)}\}$, is provided by the user. Approximately 75% should be used for actual NN training and the remainder provided as a validation set of data that will be used to determine convergence to avoid overfitting. This ratio of 3:1 gives plenty of information for training but still leaves a representative subset of the data for checks to be made.

5.2.1 Overview

The weights and biases we will collectively call the network parameter vector **a**. We can now consider the probability that a given set of network parameters is able to reproduce the known training data outputs – representing how well our NN model of the original function reproduces the true values. For problems of regression (fitting the model to a function), this gives us a log-likelihood function for **a**, depending on a standard χ^2 error function, given by

$$\log(\mathcal{L}(\mathbf{a};\boldsymbol{\sigma})) = -\frac{K\log(2\pi)}{2} - \sum_{i=1}^{N}\log(\sigma_i) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\left[\frac{t_i^{(k)} - y_i(\mathbf{x}^{(k)};\mathbf{a})}{\sigma_i}\right]^2, \tag{5.4}$$

where N is the number of outputs, K is the number of data points and $\mathbf{y}(\mathbf{x}^{(k)};\mathbf{a})$ is the NN's predicted output vector for the input vector $\mathbf{x}^{(k)}$ and network parameters $\mathbf{a}$. The values of $\boldsymbol{\sigma} = \{\sigma_i\}$ are hyper-parameters of the model that describe the standard deviation (error size) of each of the outputs.

For a classification network that aims to learn the probabilities that a set of inputs belongs to a set of output classes, the outputs are transformed according to the *softmax* procedure in order that they are all non-negative and sum to one.

$$y_i'(\mathbf{x}^{(k)};\mathbf{a}) = \frac{\exp\left(y_i(\mathbf{x}^{(k)};\mathbf{a})\right)}{\sum_{j=1}^{N}\exp\left(y_j(\mathbf{x}^{(k)};\mathbf{a})\right)} \tag{5.5}$$

The classification likelihood is then given by the *cross entropy* function [175],

$$\log(\mathcal{L}(\mathbf{a};\boldsymbol{\sigma})) = -\sum_{k=1}^{K}\sum_{i=1}^{N} t_i^{(k)} \log(y_i'(\mathbf{x}^{(k)};\mathbf{a})). \tag{5.6}$$

In this scenario, the true and predicted output values are probabilities (in $[0,1]$). In the true outputs, all are zero except for the correct output which has a value of one. For classification networks, the $\boldsymbol{\sigma}$ hyper-parameters do not factor into the log-likelihood.

Our algorithm considers the parameters $\mathbf{a}$ to be probabilistic with a log-prior distribution given by

$$\log(\mathcal{S}(\mathbf{a};\alpha)) = -\frac{\alpha}{2}\sum_i a_i^2. \tag{5.7}$$

α is a hyper-parameter of the model, called the “regularisation constant”, that gives the relative influence of the prior and the likelihood.

The posterior probability of a set of NN parameters is thus

$$\Pr(\mathbf{a};\alpha,\boldsymbol{\sigma}) \propto \mathcal{L}(\mathbf{a};\boldsymbol{\sigma}) \times \mathcal{S}(\mathbf{a};\alpha). \tag{5.8}$$

The network training begins by setting random values for the weights, sampled from a normal distribution with zero mean. The initial value of $\boldsymbol{\sigma}$ is set by the user and can be set on either the true log-likelihood values themselves or on their whitened values (as defined in Section 5.2.2). The only difference between these two settings is the magnitude of the error used. The algorithm then calculates a large initial estimate for α,

$$\alpha = \frac{|\nabla \log(\mathcal{L})|}{\sqrt{Mr}}, \tag{5.9}$$

where M is the total number of weights and biases (NN parameters) and r is a rate set by the user ($0 < r \leq 1$, default $r = 0.1$) that defines the size of the "confidence region" for the gradient. This formula for α sets larger regularisation (a.k.a. "damping") when the magnitude of the gradient of the likelihood is larger. This relates the amount of "smoothing" required to the steepness of the function being smoothed. The rate factor in the denominator allows us to increase the damping for smaller confidence regions on the value of the gradient. This results in smaller, more conservative steps that are more likely to result in an increase in the function value but results in more steps being required to reach the optimal weights.

The training then uses conjugate gradients to calculate a step, $\Delta\mathbf{a}$, that should be taken (see Section 5.2.3). Following a step, adjustments to α and $\boldsymbol{\sigma}$ may be made before another step is calculated. The methods for calculating the initial α value and then determining subsequent adjustments of α and/or $\boldsymbol{\sigma}$ are as developed for the MEMSYS software package, described in [177].

5.2.2 Data Whitening

Most of the time, it is prudent to "whiten" the data before training a network. Whitening normalises the input and/or output values according to a specified format, which makes it easier to train from the initial weights which are small and centred on zero. The network weights in the first and last layers can then be "unwhitened" after training so that the network will be able to process the original inputs and outputs.

Standard whitening will transform each input to a standard normal distribution by subtracting the mean and dividing by the standard deviation over all elements in the training data.

$$x_l^{(k)\prime} = \frac{x_l^{(k)} - \bar{x}_l}{\sigma_l} \tag{5.10a}$$

$$\bar{x}_l = \frac{1}{K} \sum_{k=1}^{K} x_l^{(k)} \tag{5.10b}$$

$$\sigma_l^2 = \frac{1}{K-1} \sum_{k=1}^{K} (x_l^{(k)} - \bar{x}_l)^2 \tag{5.10c}$$

Data can also be whitened by bringing all values into $[0,1]$.

$$x_l^{(k)\prime} = \frac{x_l^{(k)} - \min_k(x_l^{(k)})}{\max_k(x_l^{(k)}) - \min_k(x_l^{(k)})} \tag{5.11}$$

These two functions are normally performed separately on each input, but can be calculated across all inputs if the inputs are related. The mean, standard deviation, minimum, or maximum would then be computed over all inputs for all data values. The same whitening functions are also used for whitening the outputs. Since both functions consist of subtracting an offset and multiplying by a scale factor, they can easily be performed and reversed. To unwhiten network weights the inverse transform is applied, with the offset and scale determined by the source input node or target output node. Outputs for a classification network are not whitened since they are already simple probabilities.

5.2.3 Finding the next step

In order to find the most efficient path to an optimal set of parameters, we perform conjugate gradients using second-order derivative information. Newton's method gives the second-order approximation of a function,

$$f(\mathbf{a}+\Delta\mathbf{a}) \approx f(\mathbf{a}) + (\nabla f(\mathbf{a}))^{\mathrm{T}}\Delta\mathbf{a} + \frac{1}{2}(\Delta\mathbf{a})^{\mathrm{T}}\mathsf{B}\Delta\mathbf{a}, \tag{5.12}$$

where B is the Hessian matrix of second derivatives of f at $\mathbf{a}$. In this approximation, the maximum of f will occur when

$$\nabla f(\mathbf{a}+\Delta\mathbf{a}) \approx \nabla f(\mathbf{a}) + \mathsf{B}\Delta\mathbf{a} = 0. \tag{5.13}$$

Solving this for $\Delta\mathbf{a}$ gives us

$$\Delta\mathbf{a} = -\mathsf{B}^{-1}\nabla f(\mathbf{a}). \tag{5.14}$$

Iterating this stepping procedure will eventually bring us to a local maximum value of f. For our purposes, the function f is the log-posterior distribution of the NN parameters and hence Equation (5.12) is a Gaussian approximation to the posterior. The Hessian of the log-posterior is the regularised ("damped") Hessian of the log-likelihood function, where the prior – whose magnitude is set by α – provides the regularisation. If we define the Hessian matrix of the log-likelihood as $\mathbf{H}$, then $\mathbf{B} = \mathbf{H} + \alpha \mathbf{I}$ ($\mathbf{I}$ being the identity matrix). Regularisation increases the probability that the local maximum found is also the global maximum.

Using the second-order information provided by the Hessian allows for more efficient steps to be made, since curvature information can extend step sizes in directions where the gradient varies less and shorten where it is varying more. Additionally, using the Hessian of the log-posterior instead of the log-likelihood adds the regularisation of the prior. As mentioned before, this helps prevent the algorithm from getting stuck in a local maximum by smoothing out the function being explored. It also aids in reducing the region of confidence for the gradient information which will make it less likely that a step results in a worse set of parameters.

Given the form of the log-likelihood, Equation (5.4), is a sum of squares (plus a constant), we can also save computational expense by utilising the Gauss-Newton approximation of its Hessian, given by

$$\begin{aligned} \mathbf{H}_{ij} &= -\sum_{k=1}^{K} \left(\frac{\partial \mathbf{r_k}}{\partial a_i} \cdot \frac{\partial \mathbf{r_k}}{\partial a_j} + \mathbf{r_k} \cdot \frac{\partial^2 \mathbf{r_k}}{\partial \mathbf{a_i} \partial \mathbf{a_j}} \right) \\ &\approx -\sum_{k=1}^{K} \left(\frac{\partial \mathbf{r_k}}{\partial a_i} \cdot \frac{\partial \mathbf{r_k}}{\partial a_j} \right), \end{aligned} \tag{5.15}$$

where

$$\mathbf{r_k} = \frac{\mathbf{t^{(k)}} - \mathbf{y}(\mathbf{x^{(k)}}; \mathbf{a})}{\boldsymbol{\sigma}}. \tag{5.16}$$

The drawback of using second-order information is that calculation of the Hessian is computationally expensive and requires large storage, especially so in many dimensions as we will encounter for more complex networks. In general, the Hessian is not guaranteed to be positive semi-definite and so may not be invertible; however, the Gauss-Netwon approximation does have this guarantee and does not require calculating second derivatives. Inversion of the very large matrix will still be computationally expensive.

As noted in [178] however, we only need products of the Hessian with a vector to compute the solution, never actually the full Hessian itself. To calculate these approximate Hessian-vector products, we use a fast approximate method given in [179, 180]. This method takes advantage of the approximation made in Equation (5.13). If we choose $\Delta\mathbf{a} = r\mathbf{v}$ where $\mathbf{v}$ is any vector and r is a small number we can solve for the Hessian-vector product,

$$\mathsf{B}\mathbf{v} \approx \frac{\nabla f(\mathbf{a}+r\mathbf{v}) - \nabla f(\mathbf{a})}{r}. \tag{5.17}$$

This is a simple approximation for the product and can be computed in about the same time as required to find the gradient, which must be calculated anyway. In order to make this an exact method and less susceptible to numerical errors, we take the limit as $r \to 0$. Doing so gives us

$$\mathsf{B}\mathbf{v} = \frac{\partial}{\partial r}\nabla f(\mathbf{a}+r\mathbf{v})\bigg|_{r=0}. \tag{5.18}$$

We can therefore define the $\mathcal{R}_{\mathbf{v}}\{\cdot\}$ operator,

$$\mathcal{R}_{\mathbf{v}}\{f(\mathbf{a})\} \equiv \frac{\partial}{\partial r} f(\mathbf{a}+r\mathbf{v})\bigg|_{r=0}, \tag{5.19}$$

such that $\mathsf{B}\mathbf{v} = \mathcal{R}_{\mathbf{v}}\{\nabla f(\mathbf{a})\}$. This operator is then applied to all of the same equations and procedures used to calculate the gradient. These equivalent functions have the same cost as a gradient calculation, which is an $O(M)$ operation for M weights. This is a vast improvement over calculating and storing the entire Hessian matrix.

By combining all of these methods, second-order information is practical to use and significantly improves the rate of convergence of NN training.

5.2.4 Convergence

Following each step, the posterior, likelihood, correlation, and error squared values are calculated for the training data and the validation data that was not used in training (calculating the steps). Convergence to a best-fit set of parameters is determined by maximising the posterior, likelihood, correlation, or negative of the error squared of the validation data, as chosen by the user. This prevents overfitting as it provides a check that the network is still valid on points not in the training set. We use the error squared as the default function to maximise as it does not include the model hyper-parameters in its calculation and is less prone to problems with zeros than the correlation.

5.2.5 Autoencoders and Pre-Training

Autoencoders are a specific type of neural network wherein the inputs are mapped to themselves through one or more hidden layers. These typically have a symmetric setup of several hidden layers with a central layer containing fewer nodes than there are inputs. The NN can be split in half, with one part mapping the inputs to this central layer and the second part mapping the central layer weights to the outputs (same as original inputs). These two parts are called the “encoder” and “decoder” respectively and map either to or from a reduced basis set of “feature vectors” embodied in the central layer. This is akin to a non-linear principle component analysis (PCA). The non-linearity should allow for more complex relationships and more information to be contained in the same number of components/feature vectors. The individual feature vectors can be found by decoding $(1,0,0,...,0)$, $(0,1,0,...,0)$, and so on. A basic diagram of an autoencoder is shown in Figure 5.2. The three inputs $(x1,x2,x3)$ are mapped to themselves via three symmetric hidden layers, with 2 nodes in the central layer. The weights of the central layer $(z1,z2)$ are feature vector weights for the reduced non-linear basis set.

Autoencoders are notoriously difficult to train. A broad local maximum exists wherein all outputs are the average value of the inputs. Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh developed a method for “pre-training” networks to obtain a set of weights near the true global maximum [181]. This method was created with symmetric autoencoders in mind. The map from the inputs to the first hidden layer and then back is treated symmetrically and the weights are adjusted through a number of “epochs”, gradually reducing the reproduction error. This is repeated for the first to second hidden layer and so on until the central layer is reached. The network weights can then be “unfolded” by using the transpose for the symmetric connections in the decoding half to provide a decent starting point for the full training to begin. This is shown in Figure 5.2, where the $W1$ and $W2$ weights matrices are defined by pre-training. More details can be found in [181] and [182] has useful diagrams and explanations.

This pre-training can also be used for non-autoencoder networks. All layers of weights, except for the final one that connects the last hidden layer to the outputs, are pre-trained as if they were the first half of a symmetric autoencoder. However,

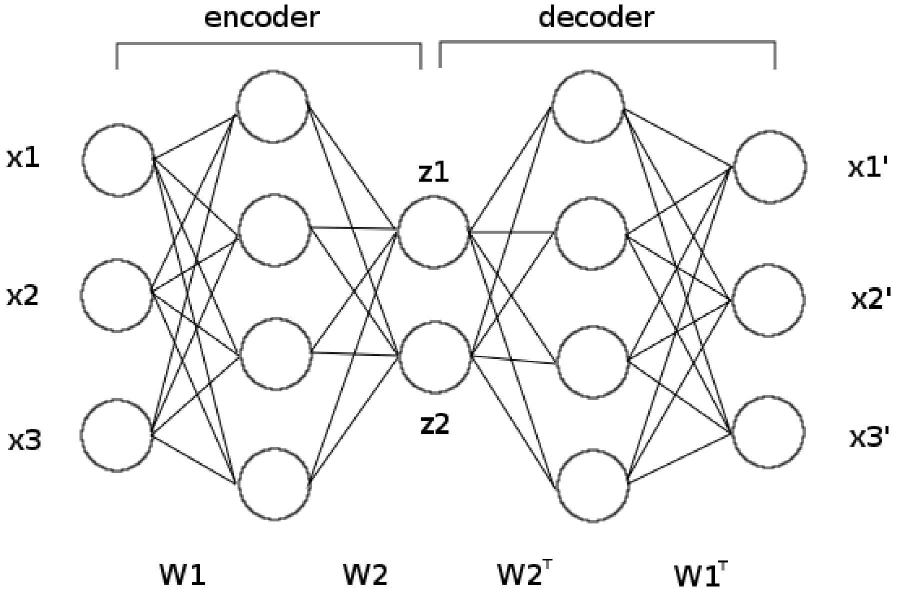


Figure 5.2: Schematic diagram of an autoencoder. The 3 input values are being encoded to 2 feature vectors. Pre-training defines the $W1$ and $W2$ weight matrices to provide a starting point for fine-tuning in training.

the network weights are not unfolded; instead the final layer of weights is initialised randomly as would have been done without pre-training. In this way, the network "learns the inputs" before mapping to a set of outputs.

When an autoencoder is pre-trained, the activation function to the central hidden layer is made linear and the activation function from the final hidden layer to the outputs is made sigmoidal. Regular networks that are pre-trained continue to use the original activation functions. Examples of the use of pre-training for autoencoders and other networks will be provided in the following sections of this chapter.

5.2.6 The Evidence of a Network

As we are determining the NN's optimal weights from a Bayesian formalism with likelihood and prior functions, we can also define the Bayesian evidence of the network model. This may be used to compare different types of networks and determine the optimal NN structure. The best NN model will have a larger evidence than other models and will, in general, be the smallest network that provides the best predictions possible for the given data. We approximate the evidence as deriving from a single Gaussian peak about the optimal weights found as in [183]. In practice, this approximation is close enough to the true value to provide a good guide. From [183] the evidence for an optimised network is given by

$$\begin{aligned}\log(\mathcal{Z}_{\text{network}}) = &\log(\mathcal{S}(\mathbf{a}_{\text{MP}};\alpha)) + \log(\mathcal{L}(\mathbf{a}_{\text{MP}});\boldsymbol{\sigma}) - \frac{1}{2}\log(|\mathbf{H}_{\text{MP}}|)\\ &+ \frac{M}{2}\log(\alpha) - \frac{1}{2}\sum_{i=1}^{N}\log(\sigma_i) - \frac{N}{2}\log(2\pi),\end{aligned} \tag{5.20}$$

where 'MP' indicates the value at the posterior maximum and M and N are defined as before (number of NN weights and outputs, respectively). There can be significant uncertainty in this measurement due to the statistical method of calculating $\log(|\mathbf{H}|)$ that avoids calculating $\mathbf{H}$ itself. However, later examples will display trends that indicate when an optimal network structure has been obtained.

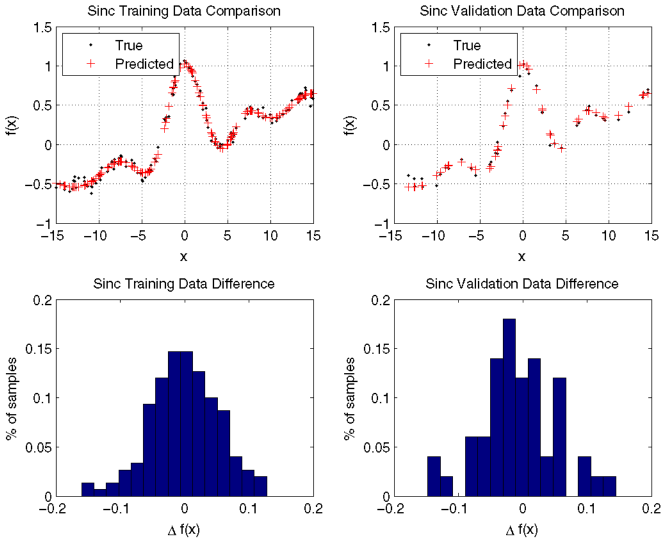


Figure 5.3: Comparisons of the true and predicted values for the sinc function on the training and validation data sets.

5.3 Toy Examples

5.3.1 Sinc Function with Noise

The first toy example we examined was a simple regression problem. In this, we generate 200 samples of $x \in \mathcal{U}[-5\pi, 5\pi]$ for which we evaluate a modified sinc function,

$$y(x) = \frac{\sin(x)}{x} + 0.04x, \tag{5.21}$$

and then add Gaussian noise with zero mean and a standard deviation of 0.05. The noise is added to make the learning of the function more difficult and prevent any exact solution being possible. The samples are split, using a random selection of 150 for training and 50 for validation. We use Equation (5.10) to whiten the inputs and outputs and train a network with 7 hidden layer nodes, obtaining correlations of greater than 99.3%. A comparison of the true and predicted outputs is shown in Figure 5.3.

The optimal number of hidden layer nodes can be determined by comparing the calculated evidence from networks with different numbers of hidden nodes. These results are shown in Table 5.1. The evidence clearly increases until we reach 6 hidden

N_{hid}	Avg. Err. Sqr.	Correlation %	$\ln(\mathcal{Z}_{\text{net}}) \pm 25$
3	7.08×10^{-3}	98.23	-279
4	6.44×10^{-3}	98.38	-256
5	3.20×10^{-3}	99.26	-161
6	2.93×10^{-3}	99.31	-142
7	2.76×10^{-3}	99.35	-143
8	2.86×10^{-3}	99.34	-144
9	2.87×10^{-3}	99.34	-143
10	2.73×10^{-3}	99.36	-137
11	2.75×10^{-3}	99.36	-134
12	2.92×10^{-3}	99.32	-154
13	2.77×10^{-3}	99.35	-134
14	2.72×10^{-3}	99.37	-170
15	2.82×10^{-3}	99.35	-183

Table 5.1: Results for training regression networks on a modified sinc function with noise.

nodes and then levels off to within its own measurement error. We can say then that any increase in accuracy for larger networks is offset by the increased complexity of the network. This additional complexity does not contribute enough additional accuracy beyond 13 nodes, when the evidence begins to drop significantly again.

5.3.2 Three-Way Classification

Radford Neal created a three-way classification data set [184] for testing his own algorithm for NN training. In this data set, four variables x_1, x_2, x_3, and x_4 are sampled from $\mathcal{U}[0,1]$ 1000 times each. If the two-dimensional Euclidean distance between (x_1,x_2) and $(0.4,0.5)$ is less than 0.35, the point is placed in class 0; otherwise, if $0.8x_1 + 1.8x_2 < 0.6$, the class was set to 1; and if neither of these conditions is true, the class was set to 2. Note that the values of x_3 and x_4 play no part in the classification. Gaussian noise with standard deviation 0.1 was then added to the input values. Figure 5.4 shows the data set with classifications. Approximately 75% of data was used for training and the remaining 25% for validation. A network was trained using 8

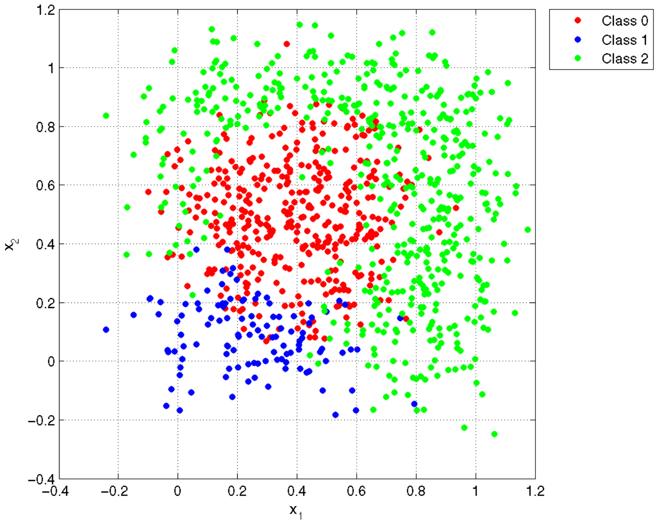


Figure 5.4: The classifications for all data points (training and validation).

True Class	Number	Predicted Class (%)		
		0	1	2
0	282	84.0	4.96	11.0
1	93	14.0	82.8	3.2
2	386	7.0	1.3	91.7

Table 5.2: Classifications for the toy training data set.

hidden layer nodes with the inputs whitened using Equation (5.10). In total, 87.8% of training data points and 85.4% of validation points were correctly classified. The summary of classifications is given in Tables 5.2 and 5.3. These results compare well with Neal's own original results [184] and are similar to classifications based on applying the original criteria to the new points that have noise added.

5.3.3 Autoencoders as Non-Linear PCA

As a first test, we trained an autoencoder network on the three-way classification inputs. For this we used hidden layers of 10 and then 4 units, so the final network

True Class	Number	Predicted Class (%)		
		0	1	2
0	99	75.7	6.1	18.2
1	19	21.1	78.9	0.0
2	121	5.0	0.8	94.2

Table 5.3: Classifications for the toy validation data set.

was 4+10+4+10+4. This network is not reducing the dimensionality, but does perform a non-linear transformation. Without pre-training, even this relatively simple autoencoder with 188 weights continuously achieves an average error squared value of 0.0950269, which corresponds to each output being equal to the average value of that particular input across all data points. However, with pre-training we can obtain an average error squared of 0.00352961, which corresponds to a correlation of 99.7%. After just one step of training the error squared is already below the best without pre-training. Reducing the central layer to 3 nodes does not affect the results without pre-training; with pre-training we are now only able to obtain an error squared of 0.00621163, which is a correlation of 92.6%. Clearly some significant information is lost, which agrees with the fact that there are four independent inputs.

To provide a quick comparison with traditional principle component analysis (PCA), we use the example of data points sampled from a multivariate Gaussian distribution. The eigenvalues and eigenvectors of this data represent the components that a PCA analysis would measure and use for data compression. In the first example, a 3D non-singular covariance matrix is used to generate samples from a multivariate Gaussian. The eigenvalues and eigenvectors of the covariance matrix calculated from these samples are then computed, to compare with the analytic values. As expected, these match very closely. Autoencoders with an increasing number of hidden layer nodes (in a single layer) are then trained. In the simplest case of a single hidden layer node, we expect that the one feature vector that will be represented will be the eigenvector with the largest eigenvalue, as this captures as much information as possible. For two hidden layer nodes, the two feature vectors will now be a linear combination of the two eigenvectors with the two largest eigenvalues, so that the same plane in the 3D space is

N_{hid}	Avg. Err. Sqr.	Correlation %	$\ln(\mathcal{Z}_{\text{net}})$
1	0.00191	94.06	5984
2	5.73×10^{-4}	98.08	7510
3	5.93×10^{-5}	99.82	10243

Table 5.4: Results for training autoencoder networks on a 3D multivariate Gaussian.

spanned. Finally, with three hidden layer nodes the three feature vectors should span the 3D space and be able to generate a nearly 100% correlation.

Pretraining was used as, even in these very small networks, it is easy to fall into the large local maximum of each output always containing the average value. Table 5.4 shows the results obtained when training autoencoders with increasing hidden layer nodes. As more nodes are added, there is a significant decrease in the average error squared as well as increasing correlation, up to 99.82%. The evidence value for the network also increases, giving another indication of a better fit. Furthermore, the feature vector from the network with a single hidden layer node was indeed an eignenvector with the highest eigenvalue and the feature vectors from the two hidden layer nodes network spanned approximately the same plane as the two eigenvectors with highest eigenvalues.

Additionally, we tested our ability to determine the optimal number of hidden layer nodes for an autoencoder when additional, redundant, information is provided. To accomplish this, a 3D multivariate Gaussian was rotated into 5D space. Therefore, there were only three independent combinations of the five provided values. The known optimal number of hidden nodes is thus also three and the autoencoders reflect that clearly in the evidences shown in Tabke 5.5. Once three hidden nodes are used, adding more neither decreases the error squared or increases the correlation; however, the evidence does decrease as Occam's razor applies a penalty for using more parameters without improving the fit to the data.

In both of these examples, it is found that with one hidden layer node the feature vector that is represented is equivalent to the primary eigenvector of the covariance matrix of samples, exactly consistent with a PCA analysis. Having shown that PCA-like results can be reproduced on a simple example, we can now apply autoencoders

N_{hid}	Avg. Err. Sqr.	Correlation %	$\ln(\mathcal{Z}_{\text{net}})$
1	0.00613	79.6	−6293
2	0.00127	96.0	12430
3	4.87×10^{-5}	99.86	17164
4	4.87×10^{-5}	99.86	16823
5	4.87×10^{-5}	99.86	16859

Table 5.5: Results for training autoencoder networks on a 3D multivariate Gaussian in 5D space.

for finding non-linear features in more complicated data sets by using larger network structures.

5.4 MNIST Handwriting Recognition

The MNIST database of handwritten digits is a subset of a larger set available from NIST (National Institute for Standards and Technology). It consists of $60,000$ training and $10,000$ validation samples of handwritten digits. The images have been size-normalised and centred in 28×28 pixel greyscale images. They are publicly available at [185], which also provides more information on the generation of the data set and results from previous analyses by other researchers. This data set has become a standard for testing of machine learning algorithms. The learning task is to correctly identify the digit written in each image, with the chosen digit being given by the output class (0 to 9) with the highest probability. Some sample digits are shown in Figure 5.5.

We have trained several different networks on this data set, using pre-training for any hidden layers. All networks whitened the inputs using Equation (5.11) across all inputs. Additionally, all networks had 784 inputs (each pixel of the image) and 10 outputs (one for each possible digit). Deeper and larger networks were able to obtain the best results. Results are summarised in Table 5.6, where the error rates are those calculated on the validation set of images. These can be compared with results referenced at [185], which obtain error rates as low as 0.35% [186] but more typically between 1% and 5% [187].

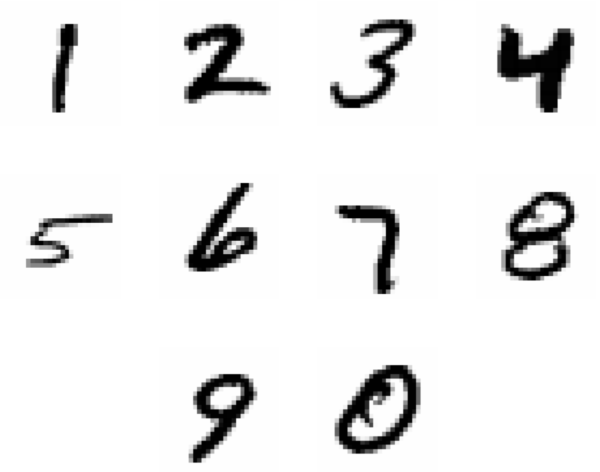


Figure 5.5: Sample handwritten digits from the MNIST database.

Hidden Layer Nodes	Error Rate (%)
0	8.08
100	4.15
250	3.00
1000	2.38
300+30	2.83
500+50	2.62
1000+300+30	2.31
500+500+2000	1.76

Table 5.6: Error rates for different networks trained to predict the MNIST data set. Training with larger networks was stopped before full convergence due to the large computational cost for marginal improvements.

True	Count	Predicted Digit (%)									
		0	1	2	3	4	5	6	7	8	9
0	980	99.1	0.0	0.2	0.1	0.0	0.3	0.2	0.1	0.0	0.0
1	1135	0.0	99.0	0.1	0.3	0.0	0.3	0.1	0.1	0.2	0.0
2	1032	0.2	0.1	97.6	0.9	0.1	0.1	0.2	0.4	0.5	0.0
3	1010	0.0	0.0	0.0	98.4	0.0	0.4	0.0	0.3	0.7	0.2
4	982	0.0	0.0	0.2	0.0	98.3	0.0	0.5	0.0	0.0	1.0
5	892	0.1	0.0	0.0	1.1	0.0	98.0	0.3	0.0	0.3	0.1
6	958	0.3	0.2	0.0	0.2	0.2	0.3	98.3	0.0	0.4	0.0
7	1028	0.0	0.1	0.5	0.5	0.1	0.0	0.0	98.0	0.2	0.6
8	974	0.2	0.0	0.1	0.8	0.1	0.4	0.0	0.1	97.7	0.5
9	1009	0.2	0.1	0.0	0.5	0.5	0.2	0.1	0.4	0.2	97.8

Table 5.7: Classifications for the MNIST blind data set. (Rows may not add up to exactly 100 due to rounding.)

In Table 5.7 we provide the classification rates made the best-performing (784 + 500 + 500 + 2000 + 10) network on the blind data set. From this we can see how the network is distributing its correct and incorrect predictions. To further illustrate, in Figure 5.6 we show a sample selection of the digits predicted incorrectly. Some are indeed hard even for a human to distinguish but some are unclear as to why they were mis-identified.

A pair of auto-encoder networks were also trained on the data set, one with hidden layers of $1000+300+30+300+1000$ (called AE-30) and another with hidden layers of $1000+500+50+500+1000$ (called AE-50). Both had 784 inputs/outputs to match the image size. The AE-30 network was able to obtain an average total error squared of only 4.64 on reproducing the digits and AE-50 obtained an average total error squared of 3.29. The networks encoded each image as 30 or 50 feature vector weights, respectively, and then decoded these weights to produce an image of the digit comparable to the original. The error squared values are comparable to those obtained by Hinton and Salakhutdinov in [182]. The feature vectors from the AE-30 network are shown in Figure 5.7 – each one being obtained by setting a single decoder input weight to one and all others to zero. Non-linear weighted combinations of these fea-

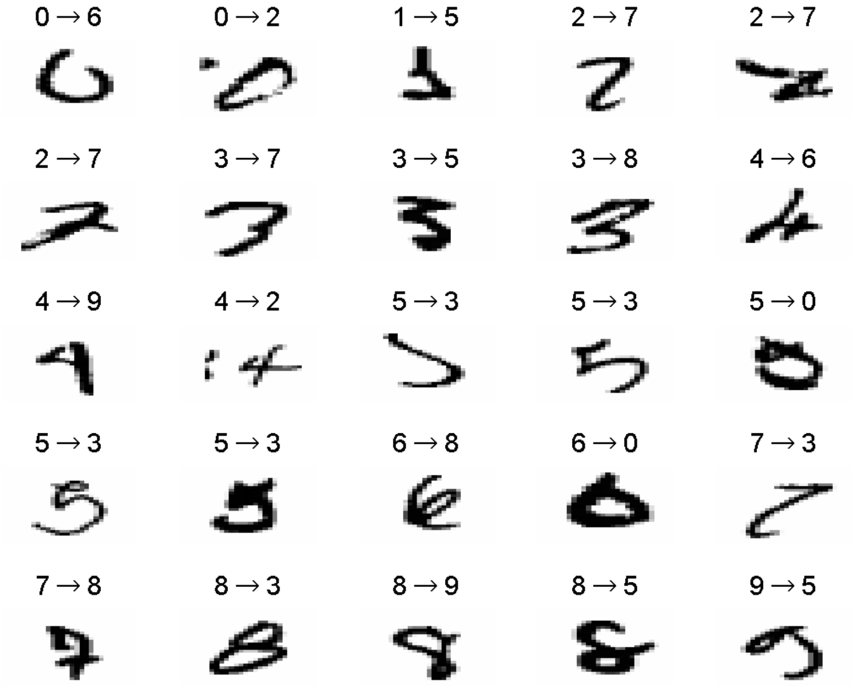


Figure 5.6: A sample of digits predicted incorrectly by the 784+500+500+2000+10 network. The true and predicted digit values are given above each image (true → predicted).

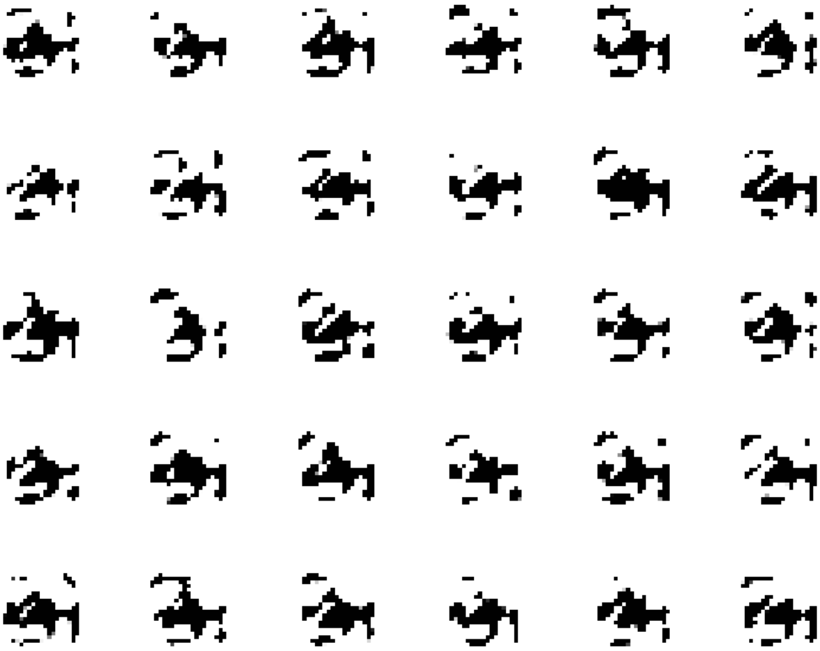


Figure 5.7: Feature vectors of the MNIST handwriting samples for the AE-30 network. The total error squared from these features is 4.64.

tures are used to reproduce each of the handwriting samples to within a small error.

We are able to use the feature vector weights of the encoded writing samples for classification. All of the training data images were passed through the two encoders to obtain the sets of 30 or 50 feature vector weights. Networks with 30 or 50 inputs and the ten classification outputs were then trained on this compressed data set. Results of the NN training are given in Table 5.8.

An autoencoder that calculated only two feature vectors using a deep network $(1000+500+250+2$ symmetric hidden layers) was also trained. Although this network was significantly less able to reproduce the original digits, having an average total error squared of 31.0, we can plot the values of the two feature vector weights for

Encoder	Hidden Layers	Error Rate (%)
AE-30	0	9.57
	10	6.39
	30	3.03
	100+50+10	2.55
AE-50	0	8.68
	10	6.61
	50	2.65
	100+50+10	2.71

Table 5.8: Error rates for different networks trained to predict the MNIST data set based on encoded feature vector weights.

each of the classes to illustrate the classification differences. We do so in Figure 5.8 for the $10,000$ blind data points. There is some clear overlap between digits with similar shapes, but some digits do occupy distinct regions of the parameter space (particularly 1 in the top right, some 0s in the bottom right, and many 2s in the middle right).

Through this example, we have shown the NN training algorithm to be able to handle the training of both deep classification and autoencoder networks. Using pre-training and second-order gradient descent we trained several networks to classify images of handwritten digits to a high degree of accuracy. Although this may be an easy task for a human brain, it is quite difficult for a computer to learn. Additionally, we were able to train autoencoder networks with millions of weights in order to reduce the dimensionality of the information we were attempting to classify from 784 pixels to 30 or 50 non-linear feature vector weights. These reduced basis sets retained enough information about the original images to reproduce them to within small errors, thus enabling us to perform classification using these weights. This classification was nearly as accurate as our best-performing network trained on the full images.

5.5 Mapping Dark Matter Challenge

The Mapping Dark Matter Challenge was presented on `www.kaggle.com` as a simplified version of the GREAT10 Challenge [188]. In this problem, each data value

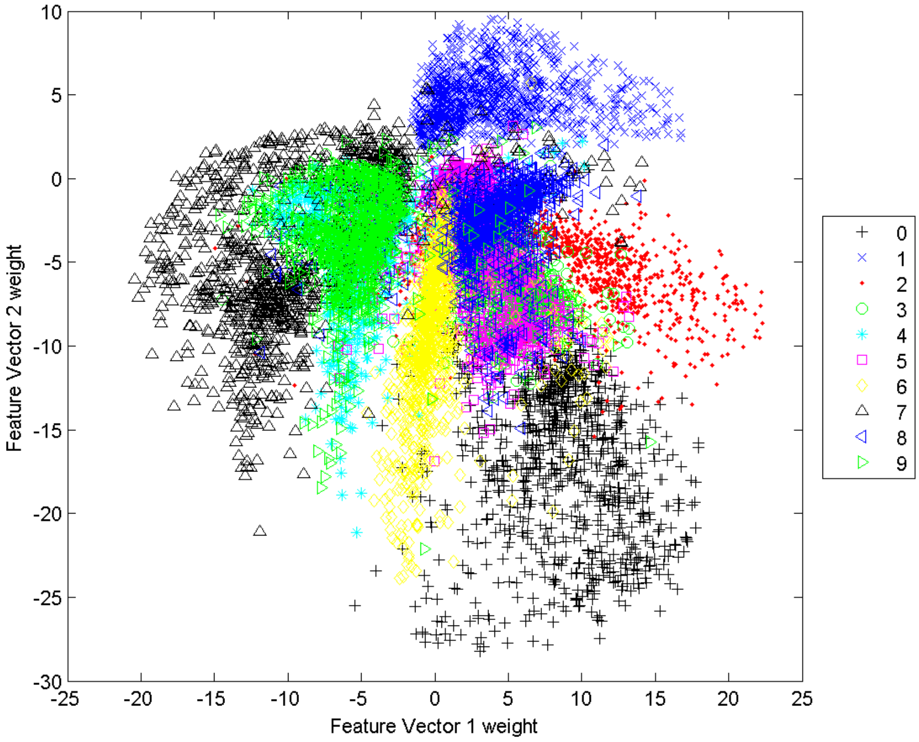


Figure 5.8: Scatterplot distribution of feature vector weights for blind data set digits as given from the encoding half of the 784+1000+500+250+2 symmetric autoencoder. This is comparable to Figure 3B in [182].

consists of an image of a galaxy and an image of a corresponding star. Each image is 48×48 pixels and greyscale. The galaxy has some ellipticity that has been obscured by a point-spread function and statistical noise. The star is a point source that has also been transformed by the same point-spread function and statistical noise. The trained network must take the galaxy and star images as inputs and predict the galaxy's ellipticity. The training data set contained $40,000$ image pairs and a challenge (validation) data set contained $60,000$ image pairs. During the challenge, the solutions for the validation data set were kept secret and participating teams submitted their predictions. Further details on the challenge and descriptions of the top results can be found in [189].

5.5.1 The Data and Model

Each galaxy image is an ellipse with a simple brightness profile. This is then convolved with a point-spread function which blurs the ellipse in a similar way to telescope observations. There is also Poisson noise degrading the image. Accompanying each galaxy image is a star image, which is a point source convolved with the same point-spread function and similar noise. A sample galaxy and star image pair is shown in Figure 5.9.

The ellipticity of a galaxy, which is essentially an ellipse, is given by two parameters, e_1 and e_2. These measure the relative lengths of the major and minor axes and the angle of the ellipse and may vary in $[-1,1]$. Equation (5.22) gives these definitions with a, b, and θ shown in Figure 5.10.

$$e_1 = \frac{a-b}{a+b}\cos(2\theta) \tag{5.22a}$$

$$e_2 = \frac{a-b}{a+b}\sin(2\theta) \tag{5.22b}$$

Further details about the data set can be found at the challenge's webpage [190]. The website also gives the unweighted quadrupole moments (UWQM) formula for calculating the ellipticity. However, as the competition organisers note, this formula will not provide very good measurements as it does not account for the point-spread function. Our aim was to use NNs to provide a better way of measuring the ellipticity.

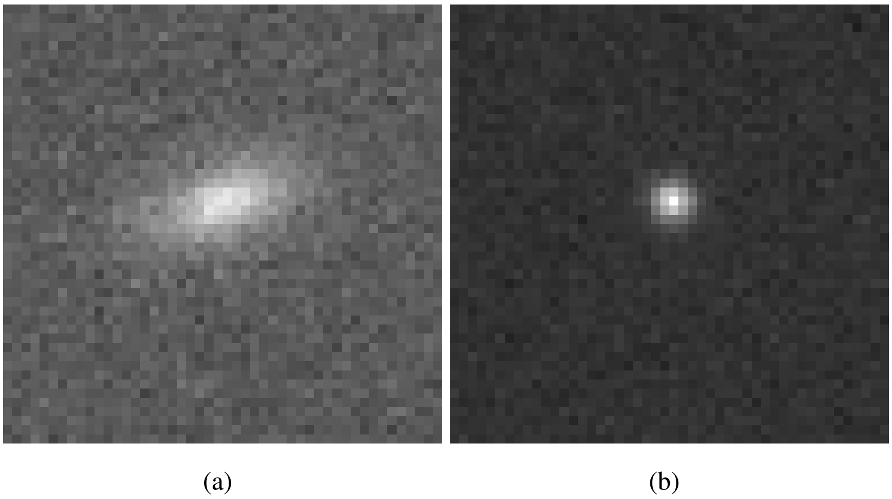


Figure 5.9: Example pair of (a) galaxy and (b) star images for the Mapping Dark Matter challenge. Each image is 48 × 48 greyscale pixels.

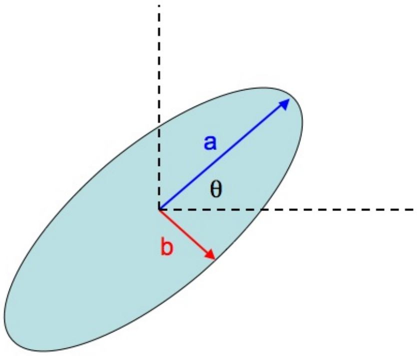


Figure 5.10: Definition of ellipse measurements for Equation (5.22). Image from [190].

Data Set	Hidden Layers	RMSE
Full Images	0	0.0224146
	2	0.0186944
	5	0.0184237
	10	0.0182661
Cropped Star	0	0.0175578
	2	0.0176236
	5	0.0175945
	10	0.0174997
	50	0.0172977
	50+10	0.0171719
Galaxy Only	0	0.0234740
	2	0.0234669
	5	0.0236508
	10	0.0226440

Table 5.9: Root mean square error rates on ellipticity predictions for different networks trained on the MDM data sets and evaluated on the $60,000$ challenge image pairs.

5.5.2 Results of Training

The quality of a network's predictions are measured by the root mean squared error of its predictions of the ellipticities for the $60,000$ challenge galaxies. Better predictions will result in lower values of the RMSE. The magnitude of the problem and the size of the dataset limited the ability to train large networks due to immense computational cost. Therefore, for this demonstration we only trained relatively smaller networks, but used three different data sets: (A) the full galaxy and star images, (B) the full galaxy image and a centrally cropped star image, and (C) just the full galaxy image. 75% of the provided training data was used for the training subset and the remaining 25% was used for validation. We originally trained on just 25% of the training subset and without using pre-training. Networks were then fine-tuned using the entire training subset. RMSE values for trained networks evaluated on the challenge set are given in Table 5.9.

These scores show naive first results that already perform very well; the standard software package SourceExtractor scores 0.0862191 on this test data and UWQM scores 0.1855600. The results could potentially be further improved by fitting profiles to the images and using these fits for training, which would reduce the number of inputs by about two orders of magnitude. Additionally, an autoencoder could be trained to allow us to decompose the images into feature vectors and use the corresponding weights for training, thereby reducing the dimensionality and the impact of noise in the images.

By reducing the number of inputs without affecting the information content – in this problem by cropping the star images to the central 24×24 pixels – we were able to improve the precision of predictions and lower our RMSE. Due to its better performance on the initial networks trained, the "CroppedStar" data set was also used to train two slightly larger networks with 50 and $50+10$ hidden layer nodes. These continued to show improving predictions, indicating that even more complex networks could further improve ellipticity measurements. The best results obtained, with an RMSE of 0.0171719, compare well with the competition winners who had an RMSE of 0.0150907 [189, 190] and used a mixture of methods that included NNs. This scored would have placed us in 32nd place out of 66 teams that entered.

The "GalaxyOnly" data set shows us that removing the star images does not allow the NN to account for the point-spread function and therefore gives a significant increase in the RMSE as might be expected.

The RMSE on the challenge data was slightly increased relative to that obtained on initial training data because it was generated with a non-zero mean ellipticity (actual mean of 0.01). Since the network is only able to predict what it has been trained on, the statistical differences between the two data sets led to a slightly increased prediction error on average.

5.6 Discussion

In this chapter I have introduced the framework of artificial neural networks used for machine learning. By using an efficient and robust algorithm, we are able to train large and deep networks. We use second-order information to allow convergence in fewer steps and prevent overfitting with prior information and checks on separate validation data. The algorithm is demonstrated on toy examples of regression, classification, and autoencoder networks. The framework was then applied to handwriting classification in determining digits from the MNIST database. Shallow and deep classification networks as well as autoencoders were used to accomplish this machine learning task. Lastly, NNs were applied to measuring the ellipticity of noisy and convolved galaxy images in the Mapping Dark Matter Challenge. NNs proved to perform well given the raw, un-treated data.

The learning power of NNs combined with the training algorithm described is a powerful tool for use in further astrophysics projects, such as the generation of gravitational wave signals, as well as computational tasks of other types that involve repeated and complex functions. This includes evaluations of likelihoods for Bayesian inference, an application that is investigated in the next chapter.

Chapter 6

The BAMBI Algorithm

> G-d does not care about our mathematical difficulties. He integrates empirically.
>
> Albert Einstein

Bayesian methods of inference are widely used in astronomy and cosmology and are gaining popularity in other fields, such as particle physics. Some uses have already been described in Chapters 2 and 3. At each point in parameter space, Bayesian methods require the evaluation of a likelihood function describing the probability of obtaining the data for a given set of model parameters. For some cosmological and particle physics problems each such function evaluation takes up to tens of seconds. MCMC applications may require millions of these evaluations, making them prohibitively costly. MultiNest is able to reduce the number of likelihood function calls by an order of magnitude or more, but further gains can be achieved if we are able to speed up the evaluation of the likelihood itself.

An artificial neural network is ideally suited for this task. A universal approximation theorem [176] assures us that we can accurately and precisely approximate the likelihood with a three-layer, feed-forward NN. The training of NNs is one of the most widely studied problems in machine learning, so techniques for learning the likelihood function are well established. In the Blind Accelerated Multimodal Bayesian Inference (BAMBI) algorithm [191], we incorporate NNs with MultiNest. Samples from MultiNest are used to train a NN or set of NNs on the likelihood function which can subsequently be used to predict new likelihood values in a tiny fraction of the time originally required. We implement the algorithm described in Chapter 5 for our NN training.

In this chapter, Section 6.1 will describe the structure of the new algorithm, BAMBI. Section 6.2 then demonstrates this algorithm performing on a few toy examples. The algorithm is applied to the computationally costly problem of cosmological parameter estimation in Section 6.3. For this same problem, we then demonstrate rapid follow-up analyses in Section 6.4. Work in this chapter was performed in collaboration with Farhan Feroz and large portions are published in [191].

6.1 The Structure of BAMBI

The Blind Accelerated Multimodal Bayesian Inference (BAMBI) algorithm combines nested sampling and neural networks. After a specified number of new samples from MultiNest have been obtained (specified by the `updInt` run parameter), BAMBI uses these to train a regression network on the log-likelihood function. Approximately 80% of the samples are used for training and the remaining 20% are used for the validation set. These values are slightly different from the previous chapter to give more information for training on the likelihood, but at the sacrifice of a smaller validation set. After convergence to the optimal NN weights, we test that the network is able to predict likelihood values to within a specified tolerance level. If not, sampling continues using the original log-likelihood until enough new samples have been made for training to be resumed. Once a network is trained that is sufficiently accurate, its predictions are used in place of the original log-likelihood function for future samples in MultiNest. Consistency checks are made to ensure the NN is not making predictions outside the range of data on which it was trained. Using the network reduces the log-likelihood evaluation time from seconds to microseconds, allowing MultiNest to complete analysis much more rapidly. As a bonus, the user also obtains a network or set of networks that are trained to easily and quickly provide more log-likelihood evaluations near the peak if needed, or in subsequent analyses.

6.1.1 When to Use the Trained Network

The optimal network possible with a given set of training data may not be able to predict log-likelihood values accurately enough, so an additional criterion is placed on when to use the trained network. This requirement is that the root-mean-square error of log-likelihoods evaluations by the network is less than a user-specified tolerance, `tol`. When the trained network does not pass this test, then BAMBI will continue using the original log-likelihood function to obtain `updInt`/2 new samples to generate a new training data set of the last `updInt` accepted samples. Network training will then resume, beginning with the weights that it had found as optimal for the previous data set. Since samples are generated from nested contours and each new data set contains half of the previous one, the saved network will already be able to produce reasonable predictions on this new data; resuming therefore enables us to save time as fewer steps will be required to reach the new optimum weights.

Once a NN is in use in place of the original log-likelihood function its evaluations are taken to be the actual log-likelihoods, but checks are made to ensure that the network is maintaining its accuracy. If the network makes a prediction outside of $[\min_{\rm training}(\log\mathcal{L}) - \texttt{tol}, \max_{\rm training}(\log\mathcal{L}) + \texttt{tol}]$, then that value is discarded and the original log-likelihood function is used for that point. Additionally, the central 95th percentile of the output log-likelihood values from the training data used is calculated and if the network is making $> 95\%$ of its predictions outside of this range then it will be re-trained. To re-train the network, BAMBI first substitutes the original log-likelihood function back in and gathers the required number of new samples from MultiNest. Training then commences, resuming from the previously saved network. These criteria ensure that the network is not trusted too much when making predictions beyond the limits of the data it was trained on, as we cannot be sure that such predictions are accurate.

The flow of sampling and training within BAMBI is demonstrated by the flowchart given in Figure 6.1. Red boxes indicate sampling with the original likelihood function and we want to minimise the time spent in these. The brown box is time spent training the NN and so does not advance the Bayesian inference; time here should be minimized as well. Lastly, the green box is sampling done with a trained NN and this is where we

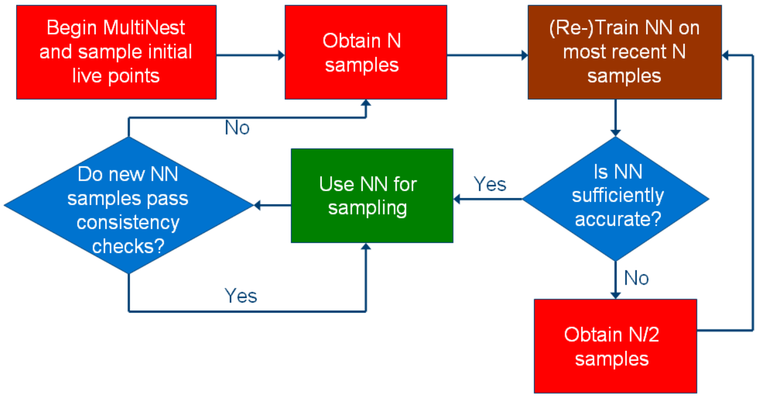

Begin MultiNest and sample initial live points
Obtain N samples
(Re-)Train NN on most recent N samples
No
Do new NN samples pass consistency checks?
Use NN for sampling
Yes
Is NN sufficiently accurate?
No
Yes
Obtain N/2 samples

Figure 6.1: A flowchart depicting the transitions between sampling and NN training within BAMBI. N is given by `updInt` from MULTINEST.

want to maximise usage, so that a larger percentage of the log-likelihood evaluations are done rapidly.

6.2 BAMBI Toy Examples

In order to demonstrate the ability of BAMBI to learn and accurately explore multimodal and degenerate likelihood surfaces, we first tested the algorithm on a few toy examples. The eggbox likelihood has many separate peaks of equal likelihood, meaning that the network must be able to make predictions across many different areas of the prior. The Gaussian shells likelihood presents the problem of making predictions in a very narrow and curving region. Lastly, the Rosenbrock function gives a long, curving degeneracy that can be extended to higher dimensions. They all require high accuracy and precision in order to reproduce the posterior truthfully and each presents unique challenges to the NN in learning the log-likelihood. It is important to note that running BAMBI on these problems required more time than the straightforward analysis; this was as expected since the actual likelihood functions are simple analytic expressions that do not require much computational expense.

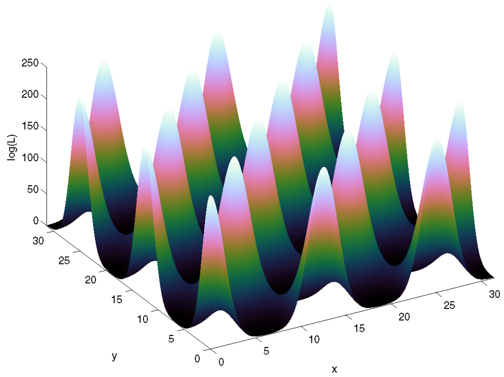


Figure 6.2: The eggbox log-likelihood surface, given by Equation (6.1).

6.2.1 Eggbox

This is a standard example of a very multimodal likelihood distribution in two dimensions. It has many peaks of equal value, so the network must be able to take samples from separated regions of the prior and make accurate predictions in all peaks. The eggbox log-likelihood [102] is given by

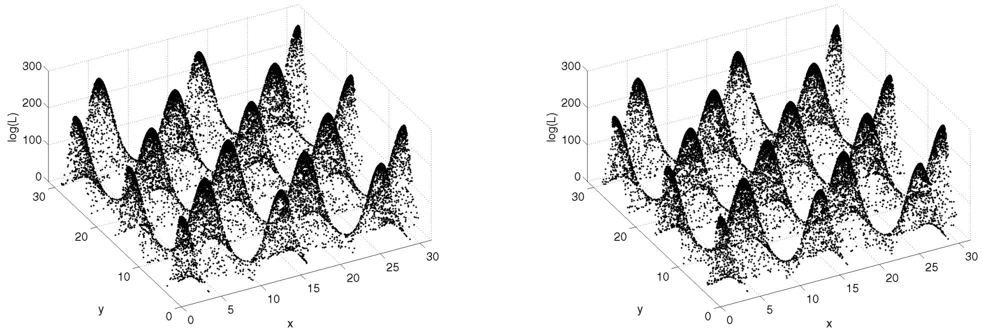
$$\log(\mathcal{L}(x,y)) = \left(2+\cos(\tfrac{x}{2})\cos(\tfrac{y}{2})\right)^5, \tag{6.1}$$

where we take a uniform prior $\mathcal{U}[0,10\pi]$ for both x and y. The structure of the surface can be seen in Figure 6.2.

We ran the eggbox example in both MULTINEST and BAMBI, both using 4000 live points. For BAMBI, we used 4000 samples for training a network with 50 hidden nodes. These values were chosen after some initial testing to give the NN sufficient complexity and data to learn the function. In Table 6.1 we report the evidences recovered by both methods as well as the true value obtained analytically from Equation (6.1). Both methods return evidences that agree with the analytically determined value to within the given error bounds. Figure 6.3 compares the posterior probability distributions returned by the two algorithms via the distribution of lowest-likelihood points removed at successive iterations by MULTINEST. We can see that they are identical distributions; therefore, we can say that the use of the NN did not reduce the quality of the results either for parameter estimation or model selection. During

Method	$\log(\mathcal{Z})$
Analytical	235.88
MULTINEST	235.859 ± 0.039
BAMBI	235.901 ± 0.039

Table 6.1: The log-evidence values of the eggbox likelihood as found analytically and with MULTINEST and BAMBI.



(a)

(b)

Figure 6.3: Points of lowest likelihood of the eggbox log-likelihood from successive iterations as given by (a) MULTINEST and (b) BAMBI.

the BAMBI analysis 51.3% of the log-likelihood function evaluations ($\sim 70,000$ total) were done using the NN; if this were a more computationally expensive function, significant speed gains would have been realised.

6.2.2 Gaussian Shells

The Gaussian shells likelihood function has low values over most of the prior, except for thin circular shells that have Gaussian cross-sections. We use two separate Gaussian shells of equal magnitude so that this is also a mutlimodal inference problem. Therefore, our Gaussian shells likelihood is

$$\mathcal{L}(\mathbf{x}) = \mathrm{circ}(\mathbf{x};\mathbf{c}_1,r_1,w_1) + \mathrm{circ}(\mathbf{x};\mathbf{c}_2,r_2,w_2), \tag{6.2}$$

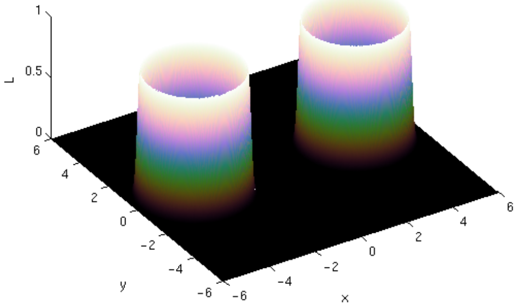


Figure 6.4: The Gaussian shell likelihood surface, given by Equations (6.2) and (6.3).

where each shell is defined by

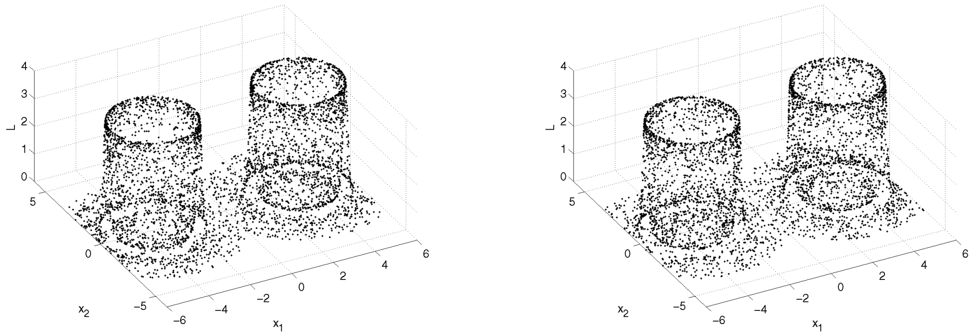
$$\mathrm{circ}(\mathbf{x};\mathbf{c},r,w) = \frac{1}{\sqrt{2\pi w^2}} \exp\left[-\frac{(|\mathbf{x}-\mathbf{c}|-r)^2}{2w^2}\right]. \tag{6.3}$$

We used values of $r_1 = r_2 = 2$, $w_1 = w_2 = 0.1$, $\mathbf{c}_1 = (-3.5,0)$, and $\mathbf{c}_2 = (3.5,0)$. This is shown in Figure 6.4.

As with the eggbox problem, we analysed the Gaussian shells likelihood with both MultiNest and BAMBI using uniform priors $\mathcal{U}[-6,6]$ on both dimensions of $\mathbf{x}$. MultiNest sampled with 1000 live points and BAMBI used 2000 samples for training a network with 100 hidden nodes (again, tests were performed to find suitable values for NN learning). In Table 6.2 we report the evidences recovered by both methods as well as the true value obtained analytically from Equations (6.2) and (6.3). The evidences are both consistent with the true value. Figure 6.5 compares the posterior probability distributions returned by the two algorithms (in the same manner as with the eggbox example). Again, we see that the distribution of returned values is nearly identical when using the NN, which BAMBI used for 18.2% of its log-likelihood function evaluations ($\sim 10,000$ total). This is a significant fraction, especially since they are all at the end of the analysis when exploring the peaks of the distribution.

Method	$\log(\mathcal{Z})$
Analytical	-1.75
MULTINEST	-1.768 ± 0.052
BAMBI	-1.757 ± 0.052

Table 6.2: The log-evidence values of the Gaussian shell likelihood as found analytically and with MULTINEST and BAMBI.



(a)

(b)

Figure 6.5: Points of lowest likelihood of the Gaussian shell likelihood from successive iterations as given by (a) MULTINEST and (b) BAMBI.

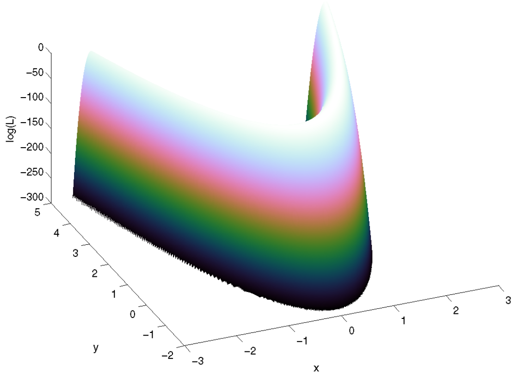


Figure 6.6: The Rosenbrock log-likelihood surface given by Equation (6.4) with $N = 2$.

6.2.3 Rosenbrock function

The Rosenbrock function is a standard example used for testing optimization as it presents a long, curved degeneracy through all dimensions. For our NN training, it presents the difficulty of learning the likelihood function over a large, curving region of the prior. We use the Rosenbrock function to define the negative log-likelihood, so the log-likelihood function is given in N dimensions by

$$\log(\mathcal{L}(\mathbf{x})) = -\sum_{i=1}^{N-1} \left[(1-x_i)^2 + 100(x_{i+1} - x_i^2)^2\right]. \tag{6.4}$$

Figure 6.6 shows how this appears for $N = 2$.

We set uniform priors of $\mathcal{U}[-5,5]$ in all dimensions and performed analysis with both MultiNest and BAMBI with $N = 2$ and $N = 10$. For $N = 2$, MultiNest sampled with 2000 live points and BAMBI used 2000 samples for training a NN with 50 hidden-layer nodes. With $N = 10$, we used 2000 live points, 6000 samples for network training, and 50 hidden nodes. Tests ensured that these settings allowed good sampling of the prior and ample information and NN structure to allow training to a high enough accuracy. Table 6.3 gives the calculated evidences returned by both algorithms as well as the analytically calculated values from Equation (6.4) (there does not exist an analytical solution for the 10D case so this is not included). Figure 6.7 compares the posterior probability distributions returned by the two algorithms for $N = 2$. For

Method	$\log(\mathcal{Z})$
Analytical 2D	-5.804
MULTINEST 2D	-5.799 ± 0.049
BAMBI 2D	-5.757 ± 0.049
MULTINEST 10D	-41.54 ± 0.13
BAMBI 10D	-41.53 ± 0.13

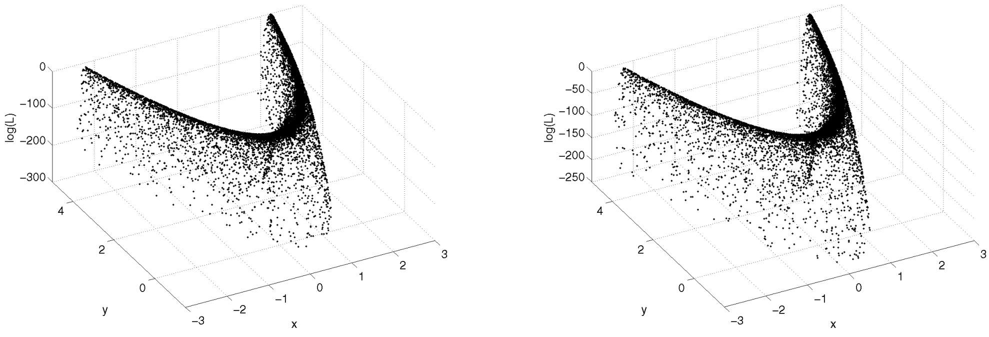
Table 6.3: The log-evidence values of the Rosenbrock likelihood as found analytically and with MULTINEST and BAMBI.

$N = 10$, we show in Figure 6.8 comparisons of the marginalised two-dimensional posterior distributions for 12 variable pairs. We see that MULTINEST and BAMBI return nearly identical posterior distributions as well as consistent estimates of the evidence. For $N = 2$ and $N = 10$, BAMBI was able to use a NN for 64.7% and 30.5% of its log-likelihood evaluations respectively ($\sim 40{,}000$ total for $N = 2$ and $\sim 1.3 \times 10^6$ for $N = 10$). Even when factoring in time required to train the NN this would have resulted in large decreases in running time for a more computationally expensive likelihood function. In the $N = 2$ case the network was trained twice, with re-training required because the first training did not achieve sufficient accuracy. In the $N = 10$ case the network was trained 12 times since the first 10 times did not achieve enough accuracy and the first network used needed re-training once the sampling moved further up in log-likelihood contours and the distribution of predictions no longer resembled the data on which it was trained.

6.3 Cosmological Parameter Estimation with BAMBI

While likelihood functions resembling our previous toy examples do exist in real physical models, we would also like to demonstrate the usefulness of BAMBI on simpler likelihood surfaces where the time of evaluation is the critical limiting factor. One such example in astrophysics is that of cosmological parameter estimation and model selection.

We implement BAMBI within the standard COSMOMC code [192], which by default uses MCMC sampling. This allows us to compare the performance of BAMBI



(a)

(b)

Figure 6.7: Points of lowest likelihood of the Rosenbrock likelihood for $N = 2$ from successive iterations as given by (a) MULTINEST and (b) BAMBI.

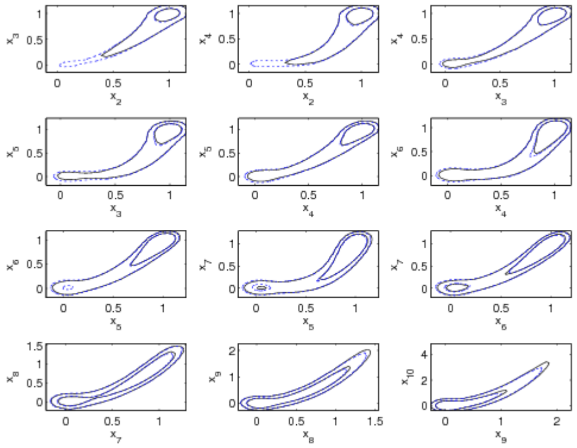


Figure 6.8: Marginalised 2D posteriors for the Rosenbrock function with $N = 10$. The 12 most correlated pairs are shown. MULTINEST is in solid black, BAMBI in dashed blue. Inner and outer contours represent 68% and 95% confidence levels, respectively.

with other methods, such as MULTINEST, COSMONET, INTERPMC, PICO, Particle Swarm Optimization, and others [102, 193–198]. In this work, we will only report performances of BAMBI and MULTINEST, but these can be compared with reported performance from the other methods.

Bayesian parameter estimation in cosmology requires evaluation of theoretical temperature and polarisation CMB power spectra (C_l values) using cosmological evolution codes such as CAMB [199]. These evaluations can take on the order of tens of seconds depending on the cosmological model. The C_l spectra are then compared to WMAP and other observations for the likelihood function. Considering that tens of thousands of these evaluations will be required, this is a computationally expensive step and a limiting factor in the speed of any Bayesian analysis. With BAMBI, we use samples to train a NN on the combined likelihood function which will allow us to avoid evaluating the full power spectra. This has the benefit of not requiring a pre-computed sample of points as in COSMONET and PICO, which is particularly important when including new parameters or new physics in the model. In these cases we will not know in advance where the peak of the likelihood will be and it is around this location that the most samples would be needed for accurate results.

The set of cosmological parameters that we use as variables and their prior ranges are given in Table 6.4; the parameters have their usual meanings in cosmology [200]. The prior probability distributions are uniform over the ranges given. A non-flat cosmological model incorporates all of these parameters, while we set $\Omega_K = 0$ for a flat model. We use $w = -1$ in both cases. The flat model thus represents the standard ΛCDM cosmology. We use two different data sets for analysis: (1) CMB observations alone and (2) CMB observations plus Hubble Space Telescope constraints on H_0, large-scale structure constraints from the luminous red galaxy subset of the SDSS and the 2dF survey, and supernovae Ia data. The CMB dataset consists of WMAP seven-year data [200] and higher resolution observations from the ACBAR, CBI, and BOOMERanG experiments. The COSMOMC website [192] provides full references for the most recent sources of these data.

Analyses with MULTINEST and BAMBI were run on all four combinations of models and data sets. MULTINEST sampled with 1000 live points and an efficiency of 0.5, both on its own and within BAMBI; BAMBI used 2000 samples for training

Parameter	Min	Max	Description
$\Omega_b h^2$	0.018	0.032	Physical baryon density
$\Omega_{DM} h^2$	0.04	0.16	Physical cold dark matter density
θ	0.98	1.1	100× Ratio of the sound horizon to angular diameter distance at last scattering
τ	0.01	0.5	Reionisation optical depth
Ω_K	-0.1	0.1	Spatial curvature
n_s	0.8	1.2	Spectral index of density perturbations
$\log[10^{10} A_s]$	2.7	4	Amplitude of the primordial spectrum of curvature perturbations
A_{SZ}	0	2	Amplitude of the Sunyaev-Zel'dovich spectrum

Table 6.4: The cosmological parameters and their minimum and maximum values. Uniform priors were used on all variables. Ω_K was set to 0 for the flat model.

a NN on the likelihood, with 50 hidden-layer nodes for both the flat model and non-flat model. These values were chosen based on the toy examples and analysis runs confirmed their viability. Table 6.5 reports the recovered evidences from the two algorithms for both models and both data sets. It can be seen that the two algorithms report equivalent values to within statistical error for all four combinations. In Figures 6.9 and 6.10 we show the recovered one- and two-dimensional marginalised posterior probability distributions for the non-flat model using the CMB-only data set. Figures 6.11 and 6.12 show the same for the non-flat model using the complete data set. We see very close agreement between MULTINEST (in solid black) and BAMBI (in dashed blue) across all parameters. The only exception is A_{SZ} since it is unconstrained by these models and data and is thus subject to large amounts of variation in sampling. The posterior probability distributions for the flat model with either data set are extremely similar to those of the non-flat flodel with setting $\Omega_K = 0$, as expected, so we do not show them here.

A by-product of running BAMBI is that we now have a network that is trained to predict likelihood values near the peak of the distribution. To see how accurate this network is, in Figure 6.13 we plot the error in the prediction ($\Delta \log(\mathcal{L}) = \log(\mathcal{L}_{\text{predicted}}) - \log(\mathcal{L}_{\text{true}})$) versus the true log-likelihood value for the different sets of training and val-

Algorithm	Model	Data Set	$\log(\mathcal{Z})$
MULTINEST	ΛCDM	CMB only	-3754.58 ± 0.12
BAMBI	ΛCDM	CMB only	-3754.57 ± 0.12
MULTINEST	ΛCDM	all	-4124.40 ± 0.12
BAMBI	ΛCDM	all	-4124.11 ± 0.12
MULTINEST	ΛCDM+Ω_K	CMB only	-3755.26 ± 0.12
BAMBI	ΛCDM+Ω_K	CMB only	-3755.57 ± 0.12
MULTINEST	ΛCDM+Ω_K	all	-4126.54 ± 0.13
BAMBI	ΛCDM+Ω_K	all	-4126.35 ± 0.13

Table 6.5: Evidences calculated by MULTINEST and BAMBI for the flat (ΛCDM) and non-flat (ΛCDM+Ω_K) models using the CMB-only and complete data sets. The two algorithms are in close agreement in all cases.

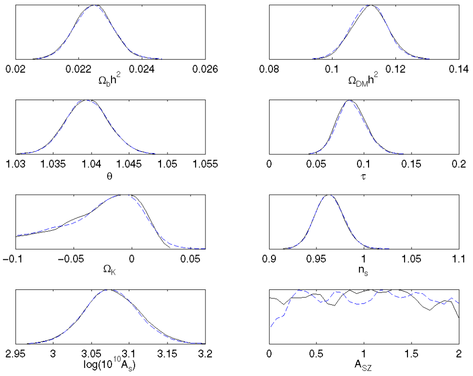


Figure 6.9: Marginalised 1D posteriors for the non-flat model (ΛCDM+Ω_K) using only CMB data. MULTINEST is in solid black, BAMBI in dashed blue.

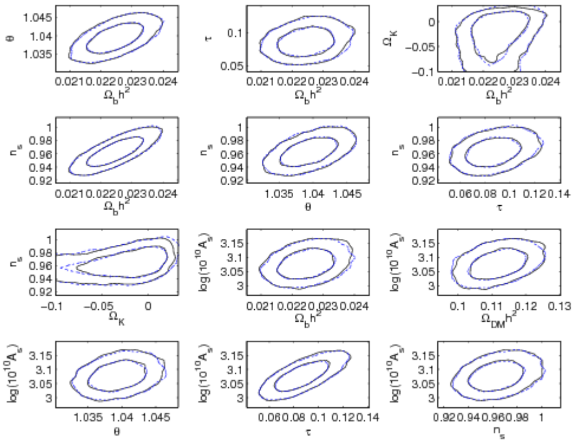


Figure 6.10: Marginalised 2D posteriors for the non-flat model (ΛCDM+Ω_K) using only CMB data. The 12 most correlated pairs are shown. MULTINEST is in solid black, BAMBI in dashed blue. Inner and outer contours represent 68% and 95% confidence levels, respectively.

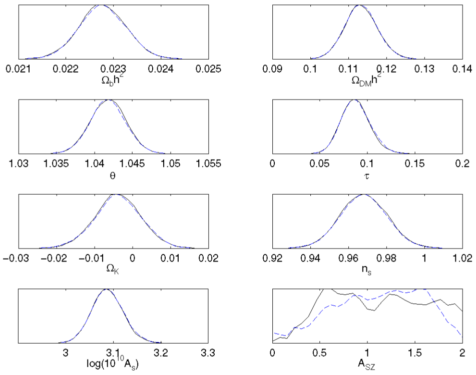


Figure 6.11: Marginalised 1D posteriors for the non-flat model (ΛCDM+Ω_K) using the complete data set. MULTINEST is in solid black, BAMBI in dashed blue.

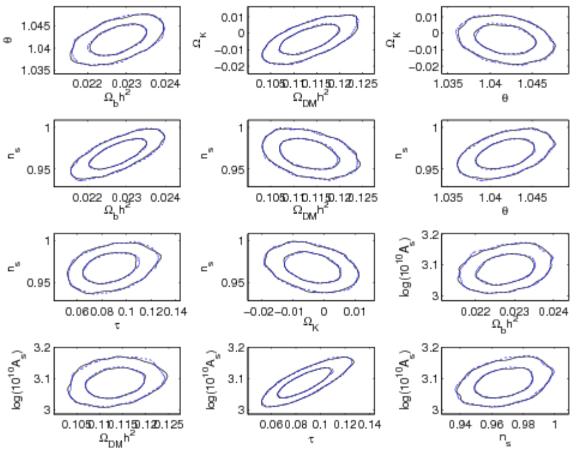


Figure 6.12: Marginalised 2D posteriors for the non-flat model (ΛCDM+Ω_K) using the complete data set. The 12 most correlated pairs are shown. MULTINEST is in solid black, BAMBI in dashed blue. Inner and outer contours represent 68% and 95% confidence levels, respectively.

idation points that were used. What we show are results for networks that were trained to sufficient accuracy in order to be used for making likelihood predictions; this results in two networks for each case shown. We can see that although the flat model used the same number of hidden-layer nodes, the simpler physical model allowed for smaller error in the likelihood predictions. Both final networks (one for each model) are significantly more accurate than the specified tolerance of a maximum average error of 0.5. In fact, for the flat model, all but one of the 2000 points have an error of less than 0.06 log-units. The two networks trained in each case overlap in the range of log-likelihood values on which they trained. The first network, although trained to lower accuracy, is valid over a much larger range of log-likelihoods. The accuracy of each network increases with increasing true log-likelihood and the second network, trained on higher log-likelihood values, is significantly more accurate than the first.

The final comparison, and perhaps the most important, is the running time required. The analyses were run using MPI parallelisation on 48 processors. We recorded the time required for the complete analysis, not including any data initialisation prior to initial sampling. We then divide this time by the number of likelihood evaluations performed to obtain an average time per likelihood ($t_{\text{wall clock, sec}} \times N_{\text{CPUs}}/N_{\log(\mathcal{L})\text{ evals}}$).

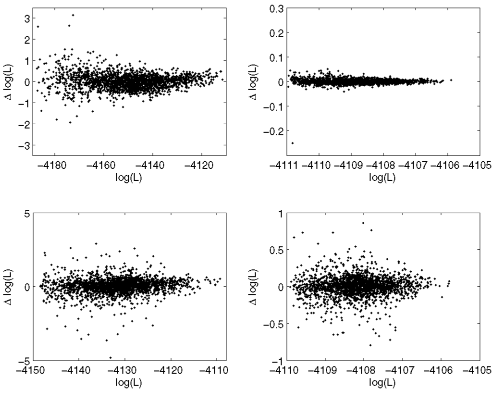


Figure 6.13: The error in the predicted likelihood ($\Delta\log(\mathcal{L}) = \log(\mathcal{L}_{\text{predicted}}) - \log(\mathcal{L}_{\text{true}})$) for the BAMBI networks trained on the flat (top row) and non-flat (bottom row) models using the complete data set. The left column represents predictions from the first NN trained to sufficient accuracy; the right column are results from the second, and final, NN trained in each case. The flat and non-flat models both used 50 hidden-layer nodes.

Therefore, time required to train the NN is still counted as a penalty factor. If a NN takes more time to train, this will hurt the average time, but obtaining a usable NN sooner and with fewer training calls will give a better time since more likelihoods will be evaluated by the NN. The resulting average times per likelihood and speed increases are given in Table 6.6. Although the speed increases appear modest, one must remember that these include time taken to train the NNs, during which no likelihoods were evaluated. This can be seen in that although 30–40% of likelihoods are evaluated with a NN, as reported in Table 6.7, we do not obtain the full equivalent speed increase. We are still able to obtain a significant decrease in running time while adding in the bonus of having a NN trained on the likelihood function.

6.3.1 Updated Timing Comparisons

The NN training algorithm code was updated to improve memory efficiency and to use single-precision variables by default. Following these updates, we decided to rerun the analyses on the cosmological parameter estimation to confirm that results were

Model	Data set	MULTINEST $t_{\mathcal{L}}$ (s)	BAMBI $t_{\mathcal{L}}$ (s)	Speed factor
ΛCDM	CMB only	2.394	1.902	1.26
ΛCDM	all	3.323	2.472	1.34
ΛCDM+Ω_K	CMB only	12.744	9.006	1.42
ΛCDM+Ω_K	all	12.629	10.651	1.19

Table 6.6: Time per likelihood evaluation, factor of speed increase from MULTINEST to BAMBI ($t_{\mathrm{MN}}/t_{\mathrm{BAMBI}}$).

Model	Data set	$\%\log(\mathcal{L})$ with NN	Equivalent speed factor	Actual speed factor
ΛCDM	CMB only	40.5	1.68	1.26
ΛCDM	all	40.2	1.67	1.34
ΛCDM+Ω_K	CMB only	34.2	1.52	1.42
ΛCDM+Ω_K	all	30.0	1.43	1.19

Table 6.7: Percentage of likelihood evaluations performed with a NN, equivalent speed factor, and actual factor of speed increase.

Model	Data set	MULTINEST	BAMBI	Speed
		$t_{\mathcal{L}}$ (s)	$t_{\mathcal{L}}$ (s)	factor
ΛCDM	CMB only	2.70	2.02	1.34
ΛCDM	all	3.69	2.46	1.50
ΛCDM+Ω_K	CMB only	13.56	9.33	1.45
ΛCDM+Ω_K	all	13.85	10.36	1.34

Table 6.8: Time per likelihood evaluation, factor of speed increase from MULTINEST to BAMBI ($t_{\mathrm{MN}}/t_{\mathrm{BAMBI}}$).

not adversely affected by use of single-precision. As before, a single hidden layer with 50 nodes was used for all four pairings of model and data and a tolerance of 0.5 was set for accepting a network as sufficiently accurate. MULTINEST used 1000 live points and 2000 points were used for network training. After completing, all calculated evidences were confirmed to agree between MULTINEST and BAMBI and previous results. Table 6.8 is like Table 6.6 but for the new runs. The reduction in time spent training a network is reflected in the speed factors which have in all cases increased from the initial comparisons.

6.4 Using Trained Networks for Follow-up in BAMBI

A major benefit of BAMBI is that following an initial run the user is provided with a trained NN, or set of NNs, that model the log-likelihood function. These can be used in a subsequent analysis with different priors to obtain much faster results. This is a comparable analysis to that of COSMONET [193, 194], except that the NNs here are a product of an initial Bayesian analysis where the peak of the distribution was *not* previously known. No prior knowledge of the structure of the likelihood surface was used to generate the networks that are now able to be re-used. In this section we will examine two methods for calculating the error in a network's prediction and provide results for performing a follow-up analysis on the cosmological parameter estimation problems.

6.4.1 Exact Error Calculation

When multiple NNs are trained and used in the initial BAMBI analysis, we must determine which network's prediction to use in the follow-up. The approximate uncertainty error in a NN's prediction of the value $y(\mathbf{x};\mathbf{a})$ ($\mathbf{x}$ denoting input parameters, $\mathbf{a}$ the NN weights and biases as before) that models the log-likelihood function is given by [201] as

$$\sigma^2 = \sigma^2_{\text{pred}} + \sigma^2_\nu, \tag{6.5}$$

where

$$\sigma^2_{\text{pred}} = \mathbf{g}^{\text{T}}\mathsf{B}^{-1}\mathbf{g} \tag{6.6}$$

and σ^2_ν is the variance of the noise on the output from the network training (the network hyper-parameter $\boldsymbol{\sigma}^2$ on the single output defined in Equation (5.4)). In Equation (6.6), B is the Hessian of the log-posterior as before, and $\mathbf{g}$ is the gradient of the NN's prediction with respect to the weights about their maximum posterior values,

$$\mathbf{g} = \left.\frac{\partial y(\mathbf{x};\mathbf{a})}{\partial \mathbf{a}}\right|_{\mathbf{x},\mathbf{a}_{\text{MP}}}. \tag{6.7}$$

This uncertainty of the prediction is calculated for each training and validation data point used in the initial training of the NN for each saved NN. The threshold for accepting a predicted log-likelihood from a network is then set to be 1.2 times the maximum uncertainty value found for points in its training and validation data sets. We can therefore ignore the value of σ^2_ν in Equation (6.5) as it is a constant value for all inputs and we only ever compare predictions of a network to previous predictions of that same network.

A log-likelihood is calculated by first making a prediction with the final NN to be trained and saved and then calculating the error for this prediction. If the error, σ_{pred}, is greater than that NN's threshold, then we consider the previous trained NN. We again calculate the predicted log-likelihood value and error to compare with its threshold. This continues to the first NN saved until a NN makes a sufficiently confident prediction. If no NNs can make a confident enough prediction, then we set $\log(\mathcal{L}) = -\infty$. This is justified because the NNs are trained to predict values in the highest likelihood regions of parameter space and if a set of parameters lies outside their collective region of validity, then it must not be within the region of highest likelihoods. This effectively limits the prior and so will bias evidences upwards but not significantly affect parameter estimation.

To demonstrate the speed-up potential of using the NNs, we first ran an analysis of the cosmological parameter estimation using both models and both data sets. This time, however, we set the tolerance of the NNs to 1.0 instead of 0.5, so that they would be valid over a larger range of log-likelihoods and pass the accuracy criterion sooner. Each analysis produced two trained NNs. We then repeated each of the four analyses, but set the prior ranges to be uniform over the region defined by $\mathbf{x}_{\max(\log(\mathcal{L}))} \pm 4\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the vector of standard deviations of the marginalised one-dimensional posterior probabilities.

In Figure 6.14 we show predictions from the two trained NNs on the two sets of validation data points in the case of the non-flat model using the complete data set. In the left-hand column, we can see that the first NN trained is able to make reasonable predictions on its own validation data as well as on the second set of points, in red crosses, from the second NN's validation data. The final NN is able to make more precise predictions on its own data set than the initial NN, but is unable to make accurate predictions on the first NN's data points. The right-hand column shows the error bar sizes for each of the points shown. For both NNs, the errors decrease with increasing log-likelihood. The final NN has significantly lower uncertainty on predictions for its own validation data, which enables us to set the threshold for when we can trust its prediction. The cases for the other three sets of cosmological models and data sets are very similar to this one. This demonstrates the need to use the uncertainty error measurement in determining which NN prediction to use, if any. Always using the final NN would produce poor predictions away from the peak and the initial NN does not have sufficient precision near the peak to properly measure the best-fit cosmological parameters. But by choosing which NN's prediction to accept, as we have shown, we can quickly and accurately reproduce the log-likelihood surface for sampling. Furthermore, if one were interested only in performing a re-analysis about the peak, then one could use just the final NN, thereby omitting the calculational overhead associated with choosing the appropriate network.

For this same case, we plot the one- and two-dimensional marginalised posterior probabilities in Figures 6.15 and 6.16, respectively. Although the priors do not cover exactly the same ranges, we expect very similar posterior distributions since the priors

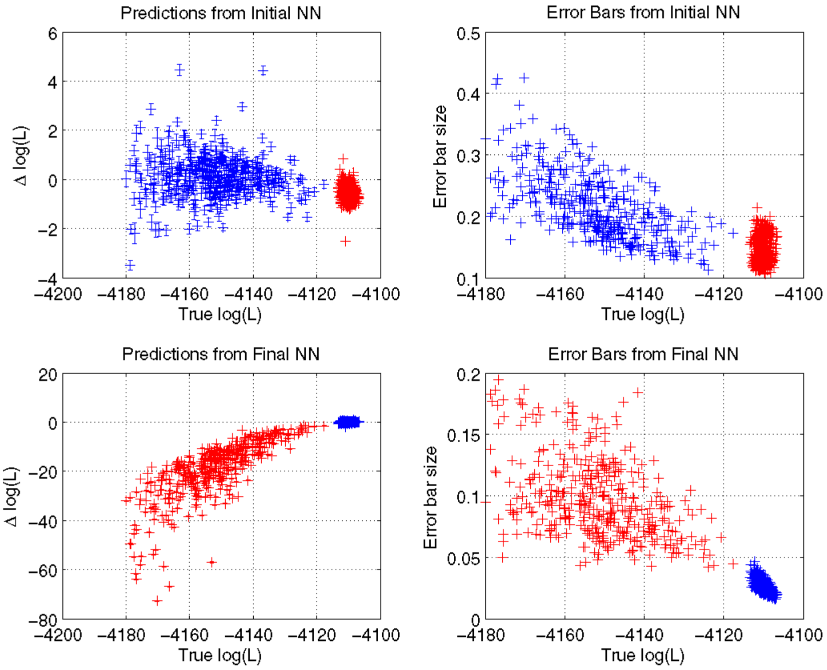


Figure 6.14: Predictions with uncertainty error bars for NNs saved by BAMBI when analysing the non-flat model using the complete data set. The left-hand side shows predictions with errors for the two NNs on their own and the other's validation data sets. Each network's own points are in blue plusses, the other NN's points are in red crosses. As many error bars are too small to be seen, the right-hand side uses the same colour and label scheme to show the magnitudes of the error bars from each NN on the predictions.

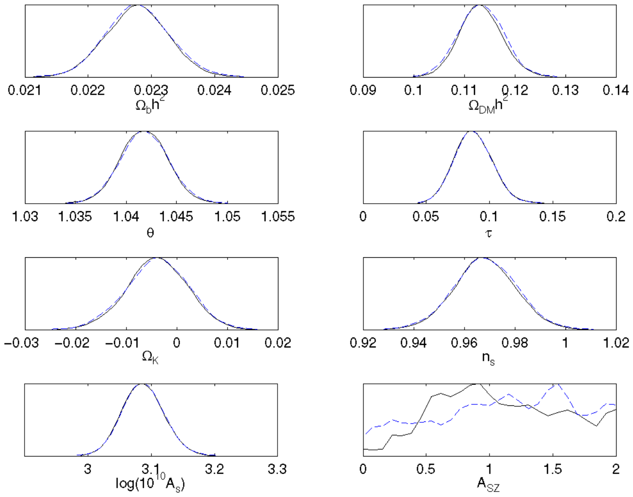


Figure 6.15: Marginalised 1D posteriors for the non-flat model (ΛCDM+Ω_K) using the complete data set. BAMBI's initial run is in solid black, the follow-up analysis in dashed blue.

are sufficiently wide as to encompass nearly all of the posterior probability. We see this very close agreement in all cases.

Calculating the uncertainty error requires calculating approximate inverse Hessian-vector products which slow down the process. We sacrifice a large factor of speed increase in order to maintain the robustness of our predictions. Using the same method as before, we computed the time per likelihood calculation for the initial BAMBI run as well as the follow up; these are compared in Table 6.9. We can see that in addition to the initial speed-up obtained with BAMBI, this follow-up analysis obtains an even larger speed-up in time per likelihood calculation. This speed-up is especially large for the non-flat model, where CAMB takes longer to compute the CMB spectra. The speed factor also increases when using the complete data set, as the original likelihood calculation takes longer than for the CMB-only data set; NN predictions take equal time regardless of the data set.

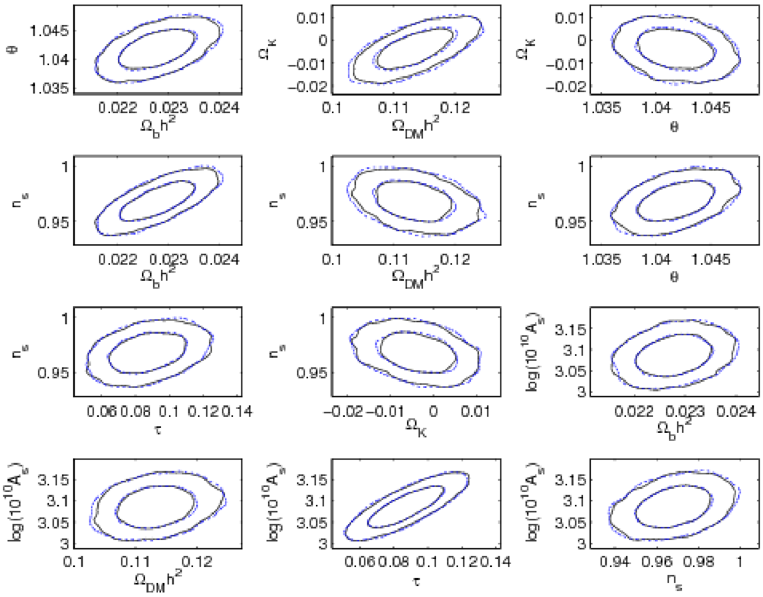


Figure 6.16: Marginalised 2D posteriors for the non-flat model (ΛCDM+Ω_K) using the complete data set. The 12 most correlated pairs are shown. BAMBI's initial run is in solid black, the follow-up analysis in dashed blue. Inner and outer contours represent 68% and 95% confidence levels, respectively.

Model	Data set	Initial $t_{\mathcal{L}}$ (s)	Follow-up $t_{\mathcal{L}}$ (s)	Speed factor
ΛCDM	CMB only	1.635	0.393	4.16
ΛCDM	all	2.356	0.449	5.25
ΛCDM+Ω_K	CMB only	9.520	0.341	27.9
ΛCDM+Ω_K	all	8.640	0.170	50.8

Table 6.9: Time per likelihood evaluation, factor of speed increase from BAMBI's initial run to a follow-up analysis.

6.4.2 Fast Error Calculation

One possible way to avoid the computational cost of computing error bars on the predictions is that suggested by [201] (see Section 9). One can take the NN training data and add Gaussian noise and train a new NN, using the old weights as a starting point. Performing many realisations of this will quickly provide multiple NNs whose average prediction will be a good fit to the original data and whose standard deviation from this mean will measure the error in the prediction. This will reduce the time needed to compute an error bar since multiple NN predictions are faster than a single inverse Hessian-vector product. The steps are given as:

1. Start with the converged network with weights $\mathbf{w}^*$ trained on true data set $D^* = \{\mathbf{x}^{(m)}, \mathbf{t}^{(m)}\}$. Estimate the Gaussian noise level of the residuals using $\sigma^2 = \sum_m (t^{(m)} - y(\mathbf{x}^{(m)}, \mathbf{w}^*))^2/N$.

2. Define a new data set D^1 by adding Gaussian noise of zero mean and variance σ^2 to the outputs in D^*.

3. Train a NN on D^1 using $\mathbf{w}^*$ as a starting point. Training should converge rapidly as the new data set is only slightly different from the original. Call the new network $\mathbf{w}_1$.

4. Repeat Steps 2 and 3 multiple times to find an ensemble of networks $\mathbf{w}_j$.

5. Use each of the networks $\mathbf{w}_j$ to make a prediction for a given set of inputs. The error on the predicted value can be estimated as the standard deviation of this set of values.

In addition to these steps, we include the option for the user to add a random Gaussian offset to the saved weights read in before training is performed on the new data set (Step 3). This offset will aid the training in moving the optimisation from a potential local maximum in the posterior distribution of the network weights, but its size must be chosen for each problem. We are thus adding noise to both the training data and the saved network weights before training a new network whose posterior maximum will be near to, but not exactly the same as, the original network's.

The training of the additional networks is performed by a separate program and the resulting error estimates can be compared with those from the exact, slow calculation

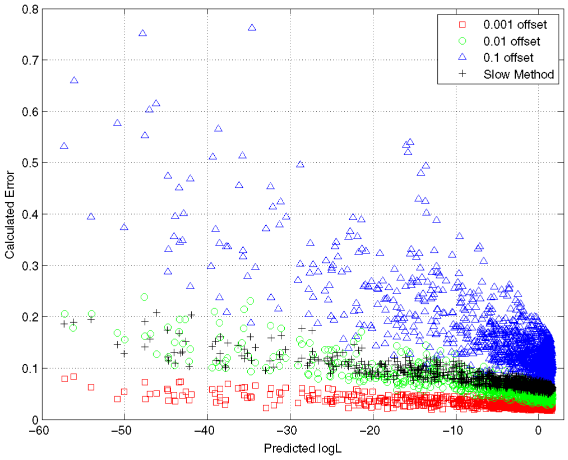


Figure 6.17: Comparison of NN prediction error calculations for the 2D Gaussian shells likelihood from the exact, slow method (Equation (6.6)) and the fast method with different weight offsets applied. Fast method calculations employ the original network and networks from 29 realisations of Gaussian noise added to the training data outputs for a total of 30 networks used in the error estimate.

(Equation (6.6)) described in the previous section (6.4). Time taken for this additional training of networks depends only on the size of the networks and the training data sets, not on the complexity of the likelihood function being modeled.

The method for estimating error was initially tested on two of the problems already analysed with BAMBI – the 2D Gaussian shells and the non-flat cosmology using the complete data set. Comparisons of the error estimates from different choices of the random offset applied to saved weights and from the exact, slow method are shown in Figures 6.17 and 6.18. The estimates from the fast method just described use the original network plus 29 different realisations of Gaussian noise added to the outputs to give a total of 30 networks, which results in total training time of less than one hour when run on 24 parallel CPUs.

From Figures 6.17 and 6.18 we can see that the value of the offset applied has a significant effect on the size of the resulting error estimates. In general, larger offsets result in larger error estimates. However, the general trend of increasing error for lower log-likelihoods (further from the peak) is followed in all cases. We can determine

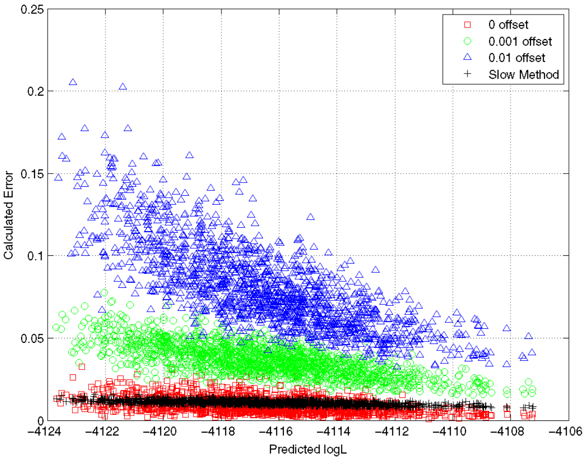


Figure 6.18: Comparison of NN prediction error calculations for the non-flat cosmology with complete data set likelihood from the exact, slow method (Equation (6.6)) and the fast method with different weight offsets applied (Gaussian random variable with mean 0 and standard deviation as noted). Fast method calculations employ the original network and networks from 29 realisations of Gaussian noise added to the training data outputs for a total of 30 networks used in the error estimate.

from Figure 6.17 that an offset with standard deviation $s \sim 0.01$ is best for the 2D Gaussian shells and from Figure 6.18 that an offset with $s \sim 0.0$ is best for the non-flat cosmology with the complete data set. Therefore, we find that the value of the best offset is roughly inversely proportional to the prior weight, $s \propto 1/\alpha$, from the best fit network $\mathbf{w}^*$. In the 2D Gaussian shells example $\alpha \simeq 5$ and in the non-flat cosmology example $\alpha \simeq 54000$. Therefore, the non-flat cosmology will require a significantly smaller offset to accurately reproduce the prediction error values. Since the absolute error values are not used for determining the quality of a fit, but rather the relative magnitudes, the fast calculations do not need to be too precise. We suggest using an offset with a standard deviation of $s \lesssim 1/\alpha$ as this will yield usable error estimates.

An additional option has been included at this stage to not discard points whose NN predictions fail error-checking; instead the original likelihood function will be evaluated at these points. Discarding the points (assigning them $\log(\mathcal{L}) = -\infty$) essentially removes them from the prior and will only affect points further from the peak where the networks are not all well trained. As such it will bias the evidence upwards but should not generally affect parameter estimation as the discarded points will be of significantly lower likelihood so as to be in the far tails of the distribution. These points will also be heavily weighted towards early in the analysis before the likelihood contours move well within the NNs' region of confident predictions.

To observe the potential speed increases with this method of error estimation, we re-analysed the cosmological models and data sets and then performed both the slow and fast methods of error calculation in separate follow-up analyses on restricted prior ranges. Our initial analyses were performed with the updated version of the NN training algorithm mentioned in Section 6.3.1. The follow-up analysis limited the prior to $\mathbf{x}_{\max(\log(\mathcal{L}))} \pm 4\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the vector of standard deviations of the marginalised one-dimensional posterior probabilities as in the previous section. When training networks for the fast error approximation, an offset with standard deviation $s \approx 1/\alpha$ was used for an ensemble of 20 total networks per saved network.

Table 6.10 shows the average time taken per log-likelihood function evaluation when this follow-up was performed with MULTINEST and with BAMBI using both the slow error calculation and the fast method. This timing includes any MULTINEST operations as well as loading the saved networks in the beginning and preparing to use them (it does not include COSMOMC data initialisation). In all cases where points

Method	Model	Data	$t_{\mathcal{L}}$ (ms)	$t_{\mathcal{L}}$ (ms)	factor	factor
				(discard)		(discard)
MultiNest	ΛCDM	CMB	2775	—	—	—
		all	3813	—	—	—
	ΛCDM+Ω_K	CMB	12830	—	—	—
		all	10980	—	—	—
BAMBI slow	ΛCDM	CMB	558.7	378.1	4.96	7.33
		all	639.5	393.4	5.96	9.68
	ΛCDM+Ω_K	CMB	1823.8	97.67	7.03	131.4
		all	1181.5	219.8	10.86	49.95
BAMBI fast	ΛCDM	CMB	3.883	0.2077	714.7	13360
		all	28.01	0.2146	136.1	17770
	ΛCDM+Ω_K	CMB	1105	0.08449	11.61	151900
		all	402.9	0.2032	27.25	54040

Table 6.10: Time per likelihood evaluation in a follow-up analysis and speed factors with respect to MultiNest. Average times without discarding bad predictions are limited by the number of calls to the original likelihood function.

were not discarded, the BAMBI evidences matched the MultiNest values to within statistical error. When discarding points, only the non-flat model with CMB-only data calculated an evidence that did not agree; this was due to this being the only case with only a single trained NN. In the other three cases, two trained networks were available to use. These three cases therefore required fewer points to be calculated with the original likelihood or discarded ($< 2\%$ for nonflat model and $< 0.2\%$ for flat model) but took additional time to load and find the best network prediction. The analysis using the non-flat model and CMB-only data, that had only a single network, runs more rapidly when points are discarded, but requires more likelihood samples and picks up a small error in the evidence calculation of approximately one log-unit (compared to an evidence uncertainty of approximately a tenth of a log-unit). When not discarding, this model and data combination required $\sim 10\%$ of points to have their log-likelihood computed with the original function.

Overall, the massive gains in speed from using the fast method along with discarding points (that the networks cannot predict precisely enough) of $O(10^4)$ to $O(10^5)$ make it worthwhile to use this as an initial follow-up procedure. Should there be too many discarded points, the analysis may be performed again without discarding to compute a more reliable evidence value but still obtaining a speed increase of $O(10)$ to $O(100)$. We can thus obtain accurate posterior distributions and evidence calculations orders of magnitude faster than originally possible. Follow-up without discarding is limited by the number of calls to the original likelihood function while follow-up with discarding is limited by the number of stored networks being used.

6.5 Summary

We have introduced and demonstrated a new algorithm for rapid Bayesian data analysis. The Blind Accelerated Multimodal Bayesian Inference algorithm combines the sampling efficiency of MultiNest with the predictive power of artificial neural networks to reduce significantly the running time for computationally expensive problems.

The first applications we demonstrated are toy examples that demonstrate the ability of the NN to learn complicated likelihood surfaces and produce accurate evidences and posterior probability distributions. The eggbox, Gaussian shells, and Rosenbrock functions each present difficulties for Monte Carlo sampling as well as for the training of a NN. With the use of enough hidden-layer nodes and training points, we have demonstrated that a NN can learn to accurately predict log-likelihood function values.

We then apply BAMBI to the problem of cosmological parameter estimation and model selection. We performed this using flat and non-flat cosmological models and incorporating only CMB data and using a more extensive data set. In all cases, the NN is able to learn the likelihood function to sufficient accuracy after training on early nested samples and then predict values thereafter. By calculating a significant fraction of the likelihood values with the NN instead of the full function, we are able to reduce the running time by a factor of up to 1.50 and potentially more. This is in comparison to use of MultiNest only, which already provides significant speed-ups in comparison to traditional MCMC methods [102].

Through all of these examples we have shown the capability of BAMBI to increase the speed at which Bayesian inference can be done. This is a fully general method and one need only change the settings for MULTINEST and the network training in order to apply it to different likelihood functions. For computationally expensive likelihood functions, the network training takes less time than is required to sample enough training points and sampling a point using the network is extremely rapid as it is a simple analytic function. Therefore, the main computational expense of BAMBI is calculating training points as the sampling evolves until the network is able to reproduce the likelihood accurately enough.

With the trained NNs, we can perform additional analyses using the same likelihood function but different priors and save large amounts of time in sampling points with the original likelihood and in training a NN. Follow-up analyses using already trained NNs provide much larger speed increases, with factors of 4 to 11 obtained for cosmological parameter estimation relative to BAMBI speeds when not discarding poorly estimated points, and 7 to 130 when discarding using the slow exact method. Having trained an ensemble of networks to provide error estimates, the speed-ups possible increase drastically to potentially $O(10^2)$ times faster when not discarding and $O(10^4)$ to $O(10^5)$ times faster when discarding. The limiting factor in these runs is the fraction of points early in the analysis that the NNs are unable to make sufficiently accurate predictions for and thus require the use of the original likelihood function or re-sampling. The calculation of the error of predictions is a flat cost based on the size of the NN and data set regardless of the original likelihood function, so the more computationally expensive the original likelihood function is the more benefit is gained from using NNs.

The NNs trained by BAMBI for cosmology cover a larger range of log-likelihoods than the one trained for COSMONET. This allows us to use a wider range of priors for subsequent analysis and not be limited to the four-sigma region around the maximum likelihood point. By setting the tolerance for BAMBI's NNs to a larger value, fewer NNs with larger likelihood ranges can be trained, albeit with larger errors on the predictions. Allowing for larger priors requires us to test the validity of our NNs' approximations, which ends up slowing the overall analysis when no networks can provide a precise enough prediction.

Since BAMBI uses a NN to calculate the likelihood at later times in the analysis where we typically also suffer from lower sampling efficiency (harder to find a new point with higher likelihood than most recent point removed), we are more easily able to implement Hamiltonian Monte Carlo [202, 203] for finding a proposed sample. This method uses gradient information to make better proposals for the next point that should be sampled. Calculating the gradient is usually a difficult task, but with the NN approximation they are very fast and simple. This improvement may be investigated in future work.

As larger data sets and more complicated models are used in cosmology, particle physics, and other fields, the computational cost of Bayesian inference will increase. The BAMBI algorithm can, without any pre-processing, significantly reduce the required running time for these inference problems. In addition to providing accurate posterior probability distributions and evidence calculations, the user also obtains NNs trained to produce likelihood values near the peak(s) of the distribution that can be used in extremely rapid follow-up analysis.

Chapter 7

Conclusions and Future Work

> Somehere, something incredible is waiting to be known.
>
> Carl Sagan

7.1 Summary of Results

Einstein's general theory of relativity predicts the existence of gravitational waves, which are perturbations to the curvature of spacetime caused by the acceleration of matter and propagate at the speed of light. The periodic strain signal they carry affects measured distances between points in space, enabling their passing to be detected by laser interferometers. These signals originate in the regimes of strongest gravity, allowing for one of the first tests of GR where the full non-linear equations must be used.

Binary star systems emit gravitational radiation and thus lose energy and fall in toward one another in a characteristic inspiral, generating a "chirp" signal observable by ground-based detectors such as LIGO and Virgo. By approximating the evolution of these systems we can obtain waveform functions that model the signal we expect to observe. Using the tools of Bayesian inference, we are then able to measure probability distributions for source parameters given these waveform models and detector data.

In Chapter 2 we consider these binary systems and test the posterior probability distributions given by the MULTINEST algorithm. We then use these posteriors to quantitatively analyse the ability of a network of ground-based detectors to localise the position of a source on the sky. We do this for a sample of injected waveforms into simulated noise from which we can infer the area of sky that will need to be observed in electromagnetic follow-ups in order to find a counterpart signal. We then use Bayesian model selection criteria for ruling out incoherent signals in LIGO data. This is performed in the context of the “big dog” blind hardware injection, where the coherent signal model was shown to be favoured over the incoherent; this same statistic ruled out manufactured incoherent signals, displaying its ability to differentiate between real gravitational waves and detector glitches.

Further advances in gravitational wave detection will be realised by the launch of a space-based laser interferometer (formerly LISA and now NGO). This will allow for the detection of GWs at much lower frequencies and with much higher SNRs. Simulated data was generated in the MLDCs in order to encourage the development of data analysis code for LISA and demonstrate the community’s ability to extract scientific information from the kind of data that was expected. In Chapter 3 I analyse some of this data, initially looking for burst signals in long stretches of instrument noise. We are able to use Bayesian criteria to detect and characterise signals and demonstrate the ability to perform model selection to determine the correct injected signal waveform. A similar analysis is then performed on LISA data containing a large number of continuous sources. Despite the confusion, we are able to accurately find many signals at once, with near-constant purity across a large range of frequencies and high completeness at higher frequencies where there are fewer signals present. Lastly we combine these two in order to attempt to detect burst signals in the presence of many continuous sources. Again Bayesian criteria prove to be powerful in separating true from false detections.

As more detailed signal models and larger data sets are analysed, performing Bayesian inference will become a more cumbersome task. To assist in simplifying our data analysis procedures, I investigated the application of the MOPED (Multiple Optimised Parameter Estimation and Data compression) algorithm. I found, as detailed in Chapter 4, that in some cases the algorithm will work, but as the analysis involves more parameters, there is the potential for additional modes in the likelihood to be falsely created, each with equal likelihood to that of the peak. A solution is tested, but the total overhead is found to not warrant further implementation.

An ever-increasingly popular tool for handling large computations is artificial neural networks. These mimic the structure of a brain by passing information through a set of interconnected nodes with the ability to "learn" a function by adjusting connection weights. In Chapter 5 I consider the application of feed-forward networks, where nodes are arranged in ordered layers. These can be trained on a pre-calculated set of data points from the function to be learned, which can either involve regression or classification. We implement a Bayesian optimisation algorithm that uses second-order information on the log-posterior function in order to find the optimal set of weights in as few steps as possible. We also prevent over-fitting to the precise training data set provided by using regularisation and a validation data set. The capabilities of the network are first demonstrated on a few toy problems. The network training algorithm is then applied to solve two larger and more complex examples. In the first, we identify handwritten digits from the MNIST database, which involves finding patterns in images to make ideal classifications. In the second, we measure the ellipticities of sheared and convolved galaxies, a difficult task addressed by the Mapping Dark Matter challenge. As its name suggests, if this procedure can be done accurately and rapidly it may be applied to large surveys of galaxies and can help in mapping dark matter distributions throughout the universe and testing cosmological models. Our neural network training algorithm is able to learn both of these data sets very well; prediction error on a further sample of handwritten digits produces an error rate of less than 2% and the error in galaxy ellipticity measurements in reduced by a factor of ~ 5 compared to previous standard methods.

In Chapter 6 I look to applying neural networks directly to Bayesian inference. The Blind Accelerated Multimodal Bayesian Inference (BAMBI) algorithm is described, wherein samples of the log-likelihood function from MULTINEST are used to train a neural network. This can then be used in place of the original log-likelihood once it is able to make sufficiently accurate predictions. I first test BAMBI on a few toy likelihoods to ensure that the neural network is able to learn the function from a reasonable number of samples before MULTINEST's sampling converges. We then address the task of performing cosmological parameter estimation. This is a problem with a well-defined likelihood function and common data sets through the CosmoMC software package. Likelihood function calls normally take on the order of seconds each so analysis may take many CPU days. With BAMBI, however, we are able to demonstrate speed increases of 50%, indicating that $\sim 33\%$ of likelihood samples were performed using a trained neural network. Even more impressive gains in speed come from follow-up analyses with different priors. We can re-use the trained NNs, thereby avoiding the original likelihood function entirely. Doing so with a rapid method of calculating the error in NN predictions yields posterior inferences $O(10^4)$ to $O(10^5)$ times faster than using only MULTINEST. BAMBI's ability to decrease the computational expense of initial Bayesian inference and perform incredibly rapid follow-up in a fully general way means it can be applied to all types of parameter estimation and model selection problems.

7.2 Avenues for Future Work

Analysis of data from ground-based GW detectors in the advanced era will need to utilise more advanced waveform families. Inspirals that include the full effects of spin in the phase and amplitude corrections will provide the best measurements of compact binary signals we hope to detect. Additionally, the inclusion of merger and ringdown stages to create a hybrid waveform should provide the most information about a GW source. Creating these waveform models and analysing their different parameter estimation biases is important before and during the advanced detector era where detections are expected. Model selection will prove to be important when determining which waveform type best fits an observed signal and for filtering out detector glitches.

As the LISA mission has been reconfigured to NGO, we must reconsider the scientific potential of the new detector. Data analysis pipelines will need to work even harder as there will be lower SNR for NGO-observed signals than for LISA.

Neural networks and the BAMBI algorithm can play a large role in these projects. NNs can be applied to computing very difficult gravitational waveforms in order to more rapidly model systems. BAMBI will greatly increase the speed at which Bayesian analysis may be performed on GW signals, providing sky location and other parameter estimates with significantly less lag time from the initial trigger. We are always looking to improve our NN training and BAMBI and will continue to develop these two algorithms to add more functionality and improve the quality and speed of results.

There is an optimistic future for the search for gravitational waves and the scientific potential they hold. With the right tools we can efficiently and rapidly analyse detector data to find and characterise any signals present. These tools are not limited to just gravitational wave astronomy, but may be applied to solve computational or data analysis problems in a wide number of fields.

References

[1] A. Einstein, *Relativity: The special and general theory*. New York, NY, USA: Henry Holt and Co., 1920. 1

[2] J. M. Weisberg, D. J. Nice, and J. H. Taylor, “Timing Measurements of the Relativistic Binary Pulsar PSR B1913+16,” *The Astrophysical Journal*, vol. 722, pp. 1030–1034, Oct. 2010, arXiv:1011.0718. 1, 22

[3] É. É. Flanagan and S. A. Hughes, “The basics of gravitational wave theory,” *New Journal of Physics*, vol. 7, p. 204, Sept. 2005, arXiv:gr-qc/0501041. 2, 3

[4] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*. San Francisco, CA, USA: W. H. Freeman and Company, 1973. 4

[5] M. P. Hobson, G. P. Estathiou, and A. N. Lasenby, *General Relativity: An Introduction for Physicists*. Cambridge, UK: Cambridge University Press, 2006. 4

[6] L. Blanchet, “Post-Newtonian theory and the two-body problem,” *ArXiv e-prints*, July 2009, arXiv:0907.3596. 6, 21

[7] M. Sasaki and H. Tagoshi, “Analytic black hole perturbation approach to gravitational radiation,” *Living Reviews in Relativity*, vol. 6, no. 6, 2003.

[8] T. Futamase and Y. Itoh, “The post-newtonian approximation for relativistic compact binaries,” *Living Reviews in Relativity*, vol. 10, no. 2, 2007. 21

[9] 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*, vol. 65, p. 061501, Mar. 2002, arXiv:gr-qc/0105099.

[10] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, “Comparison of post-newtonian templates for compact binary inspiral signals in gravitational-wave detectors,” *Phys. Rev. D*, vol. 80, p. 084043, Oct 2009. 29, 30

[11] L. E. Kidder, “Coalescing binary systems of compact objects to (post)$^{5/2}$-Newtonian order. V. Spin effects,” *Physical Review D*, vol. 52, pp. 821–847, July 1995, arXiv:gr-qc/9506022.

[12] C. M. Will and A. G. Wiseman, “Gravitational radiation from compact binary systems: Gravitational waveforms and energy loss to second post-newtonian order,” *Phys. Rev. D*, vol. 54, pp. 4813–4848, Oct 1996. 6

[13] C. Van Den Broeck and A. S. Sengupta, “Phenomenology of amplitude-corrected post-Newtonian gravitational waveforms for compact binary inspiral: I. Signal-to-noise ratios,” *Classical and Quantum Gravity*, vol. 24, pp. 155–176, Jan. 2007, arXiv:gr-qc/0607092. 7

[14] K. G. Arun, A. Buonanno, G. Faye, and E. Ochsner, “Higher-order spin effects in the amplitude and phase of gravitational waveforms emitted by inspiraling compact binaries: Ready-to-use gravitational waveforms,” *Physical Review D*, vol. 79, p. 104023, May 2009, arXiv:0810.5336.

[15] P. Ajith, “Addressing the spin question in gravitational-wave searches: Waveform templates for inspiralling compact binaries with nonprecessing spins,” *Physical Review D*, vol. 84, p. 084037, Oct. 2011, arXiv:1107.1267.

[16] D. A. Brown, A. Lundgren, and R. O’Shaughnessy, “Nonspinning searches for spinning binaries in ground-based detector data: Amplitude and mismatch predictions in the constant precession cone approximation,” *ArXiv e-prints*, Mar. 2012, arXiv:1203.6060.

[17] G. Faye, L. Blanchet, and A. Buonanno, “Higher-order spin effects in the dynamics of compact binaries. I. Equations of motion,” *Physical Review D*, vol. 74, p. 104033, Nov. 2006, arXiv:gr-qc/0605139.

[18] L. Blanchet, A. Buonanno, and G. Faye, “Higher-order spin effects in the dynamics of compact binaries. II. Radiation field,” *Physical Review D*, vol. 74, p. 104034, Nov. 2006, arXiv:gr-qc/0605140. 7

[19] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, “Inspiral-merger-ringdown multipolar waveforms of nonspinning black-hole binaries using the effective-one-body formalism,” *Phys. Rev. D*, vol. 84, p. 124052, Dec 2011. 7, 21

[20] A. Buonanno and T. Damour, “Transition from inspiral to plunge in binary black hole coalescences,” *Phys. Rev. D*, vol. 62, p. 064015, Aug 2000.

[21] A. Buonanno and T. Damour, “Effective one-body approach to general relativistic two-body dynamics,” *Phys. Rev. D*, vol. 59, p. 084006, Mar 1999.

[22] T. Damour, “Coalescence of two spinning black holes: An effective one-body approach,” *Phys. Rev. D*, vol. 64, p. 124013, Nov 2001. 21

[23] Y. Pan, A. Buonanno, L. T. Buchman, T. Chu, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, “Effective-one-body waveforms calibrated to numerical relativity simulations: Coalescence of nonprecessing, spinning, equal-mass black holes,” *Physical Review D*, vol. 81, p. 084041, Apr. 2010, arXiv:0912.3466. 7

[24] F. Pretorius, “Evolution of binary black-hole spacetimes,” *Phys. Rev. Lett.*, vol. 95, p. 121101, Sep 2005. 7, 21

[25] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, “Gravitational-wave extraction from an inspiraling configuration of merging black holes,” *Phys. Rev. Lett.*, vol. 96, p. 111102, Mar 2006.

[26] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, “Accurate evolutions of orbiting black-hole binaries without excision,” *Phys. Rev. Lett.*, vol. 96, p. 111101, Mar 2006.

[27] S. T. McWilliams, "The status of black-hole binary merger simulations with numerical relativity," *Classical and Quantum Gravity*, vol. 28, p. 134001, July 2011, arXiv:1012.2872. 21

[28] J. Centrella, J. G. Baker, B. J. Kelly, and J. R. van Meter, "The Final Merger of Black-Hole Binaries," *Annual Review of Nuclear and Particle Science*, vol. 60, pp. 75–100, Nov. 2010, arXiv:1010.2165.

[29] J. Centrella, J. G. Baker, B. J. Kelly, and J. R. van Meter, "Black-hole binaries, gravitational waves, and numerical relativity," *Reviews of Modern Physics*, vol. 82, pp. 3069–3119, Oct. 2010, arXiv:1010.5260. 7

[30] F. Ohme, "Analytical meets numerical relativity - status of complete gravitational waveform models for binary black holes," *ArXiv e-prints*, Nov. 2011, arXiv:1111.3737. 7

[31] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel, "A prototype effective-one-body model for non-precessing spinning inspiral-merger-ringdown waveforms," *ArXiv e-prints*, Feb. 2012, arXiv:1202.0790.

[32] P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan, J. T. Whelan, B. Brügmann, P. Diener, J. Gonzalez, M. Hannam, S. Husa, M. Koppitz, D. Pollney, L. Rezzolla, L. Santamaría, A. M. Sintes, U. Sperhake, and J. Thornburg, "A phenomenological template family for black-hole coalescence waveforms," *Classical and Quantum Gravity*, vol. 24, p. 689, Oct. 2007, arXiv:0704.3764.

[33] P. Ajith, M. Hannam, S. Husa, Y. Chen, B. Brügmann, N. Dorband, D. Müller, F. Ohme, D. Pollney, C. Reisswig, L. Santamaría, and J. Seiler, "Inspiral-Merger-Ringdown Waveforms for Black-Hole Binaries with Non-precessing Spins," *Physical Review Letters*, vol. 106, p. 241101, June 2011, arXiv:0909.2867.

[34] L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig, E. L. Robinson, J. Seiler, and B. Krishnan, "Matching post-Newtonian and numerical relativity waveforms: Sys-

tematic errors and a new phenomenological model for nonprecessing black hole binaries," *Physical Review D*, vol. 82, p. 064016, Sept. 2010, arXiv:1005.3306.

[35] I. MacDonald, S. Nissanke, and H. P. Pfeiffer, "Suitability of post-Newtonian/numerical-relativity hybrid waveforms for gravitational wave detectors," *Classical and Quantum Gravity*, vol. 28, p. 134002, July 2011, arXiv:1102.5128.

[36] R. Sturani, S. Fischetti, L. Cadonati, G. M. Guidi, J. Healy, D. Shoemaker, and A. Vicere', "Phenomenological gravitational waveforms from spinning coalescing binaries," *ArXiv e-prints*, Dec. 2010, arXiv:1012.5172. 7

[37] P. C. Peters and J. Mathews, "Gravitational radiation from point masses in a keplerian orbit," *Physical Review*, vol. 131, no. 1, pp. 435–440, 1963. 8

[38] B. F. Schutz, "Gravitational waves on the back of an envelope," *American Journal of Physics*, vol. 52, pp. 412–419, May 1984. 8

[39] S. A. Hughes, "Gravitational Waves from Merging Compact Binaries," *Annual Review of Astronomy & Astrophysics*, vol. 47, pp. 107–157, Sept. 2009, arXiv:0903.4877. 8

[40] L. Blanchet, "Gravitational radiation from post-newtonian sources and inspiralling compact binaries," *Living Reviews in Relativity*, vol. 9, no. 4, 2006. 8, 21

[41] K. N. Yakunin, P. Marronetti, A. Mezzacappa, S. W. Bruenn, C.-T. Lee, M. A. Chertkow, W. R. Hix, J. M. Blondin, E. J. Lentz, O. E. Bronson Messer, and S. Yoshida, "Gravitational waves from core collapse supernovae," *Classical and Quantum Gravity*, vol. 27, p. 194005, Oct. 2010, arXiv:1005.0779. 8

[42] P. Jaranowski, A. Królak, and B. F. Schutz, "Data analysis of gravitational-wave signals from spinning neutron stars: The signal and its detection," *Phys. Rev. D*, vol. 58, p. 063001, Aug 1998. 8

[43] E. Goetz and K. Riles, "An all-sky search algorithm for continuous gravitational waves from spinning neutron stars in binary systems," *Classical and Quantum Gravity*, vol. 28, p. 215006, Nov. 2011, arXiv:1103.1301.

[44] M. Pitkin, "Prospects of observing continuous gravitational waves from known pulsars," *Monthly Notices of the Royal Astronomical Society*, vol. 415, pp. 1849–1863, Aug. 2011, arXiv:1103.5867.

[45] R. Prix, S. Giampanis, and C. Messenger, "Search method for long-duration gravitational-wave transients from neutron stars," *Physical Review D*, vol. 84, p. 023007, July 2011, arXiv:1104.1704.

[46] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, C. Affeldt, B. Allen, and *et al.*, "Beating the Spin-down Limit on Gravitational Wave Emission from the Vela Pulsar," *The Astrophysical Journal*, vol. 737, p. 93, Aug. 2011, arXiv:1104.2712.

[47] C. Cutler, "An improved, "phase-relaxed" F-statistic for gravitational-wave data analysis," *ArXiv e-prints*, Apr. 2011, arXiv:1104.2938. 8

[48] C. Cutler and K. S. Thorne, "An overview of gravitational-wave sources," in *Proceedings of GR16* (N. T. Bishop and S. D. Maharaj, eds.), Singapore: World-Scientific, 2002, gr-qc/0204090. 8

[49] M. S. Briggs, V. Connaughton, K. C. Hurley, P. A. Jenke, A. von Kienlin, A. Rau, X.-L. Zhang, The LIGO Scientific Collaboration, Virgo Collaboration: J. Abadie, B. P. Abbott, and *et al.*, "Search for gravitational waves associated with gamma-ray bursts during LIGO science run 6 and Virgo science runs 2 and 3," *ArXiv e-prints*, May 2012, arXiv:1205.2216. 8

[50] S. Rosswog, T. Piran, and E. Nakar, "The multi-messenger picture of compact object encounters: binary mergers versus dynamical collisions," *ArXiv e-prints*, Apr. 2012, arXiv:1204.6240. 8

[51] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, "Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence," *ArXiv e-prints*, Oct. 2011, arXiv:1110.0530. 8

[52] S. Gossan, J. Veitch, and B. S. Sathyaprakash, “Bayesian model selection for testing the no-hair theorem with black hole ringdowns,” *ArXiv e-prints*, Nov. 2011, arXiv:1111.5819. 8

[53] W. Del Pozzo, J. Veitch, and A. Vecchio, “Testing general relativity using Bayesian model selection: Applications to observations of gravitational waves from compact binary systems,” *Physical Review D*, vol. 83, p. 082002, Apr. 2011, arXiv:1101.1391. 8

[54] S. R. Taylor and J. R. Gair, “Cosmology with the lights off: standard sirens in the Einstein Telescope era,” *ArXiv e-prints*, Apr. 2012, arXiv:1204.6739. 8

[55] S. R. Taylor, J. R. Gair, and I. Mandel, “Cosmology using advanced gravitational-wave detectors alone,” *Physical Review D*, vol. 85, p. 023535, Jan. 2012, arXiv:1108.5161.

[56] C. Messenger and J. Read, “Measuring a Cosmological Distance-Redshift Relationship Using Only Gravitational Wave Observations of Binary Neutron Star Coalescences,” *Physical Review Letters*, vol. 108, p. 091101, Mar. 2012, arXiv:1107.5725.

[57] W. Del Pozzo, “Cosmology with Gravitational Waves: statistical inference of the Hubble constant,” *ArXiv e-prints*, Aug. 2011, arXiv:1108.1317.

[58] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, “Exploring short gamma-ray bursts as gravitational-wave standard sirens,” *The Astrophysical Journal*, vol. 725, no. 1, p. 496, 2010. 8

[59] B. P. Abbott, R. Abbott, R. Adhikari, P. Ajith, B. Allen, G. Allen, R. S. Amin, S. B. Anderson, W. G. Anderson, M. A. Arain, and *et al.*, “LIGO: the Laser Interferometer Gravitational-Wave Observatory,” *Reports on Progress in Physics*, vol. 72, p. 076901, July 2009, arXiv:0711.3041. 8

[60] A. S. Sengupta, LIGO Scientific Collaboration, and Virgo Collaboration, “LIGO-Virgo searches for gravitational waves from coalescing binaries: A status update,” *Journal of Physics Conference Series*, vol. 228, p. 012002, May 2010, arXiv:0911.2738. 9

[61] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, "FINDCHIRP: an algorithm for detection of gravitational waves from inspiraling compact binaries," *ArXiv General Relativity and Quantum Cosmology e-prints*, Sept. 2005, arXiv:gr-qc/0509116. 9

[62] S. Bose, T. Dayanga, S. Ghosh, and D. Talukder, "A blind hierarchical coherent search for gravitational-wave signals from coalescing compact binaries in a network of interferometric detectors," *Classical and Quantum Gravity*, vol. 28, p. 134009, July 2011, arXiv:1104.2650. 9

[63] S. Ballmer and *et al.*, "Enhancements to the laser interferometer gravitational-wave observatory (ligo)," *Bulletin of the American Astronomical Society*, vol. 41, p. 443, 2009. 9

[64] L. Blackburn and *et al.*, "The lsc glitch group: Monitoring noise transients during the fifth ligo science run," *Classical and Quantum Gravity*, vol. 25, p. 184004, 2008. 9

[65] T. B. Littenberg and N. J. Cornish, "Separating gravitational wave signals from instrument artifacts," *Physical Review D*, vol. 82, p. 103007, Nov. 2010, arXiv:1008.1577.

[66] V. Raymond, M. V. van der Sluys, I. Mandel, V. Kalogera, C. Röver, and N. Christensen, "The effects of LIGO detector noise on a 15-dimensional Markov-chain Monte Carlo analysis of gravitational-wave signals," *Classical and Quantum Gravity*, vol. 27, p. 114009, June 2010, arXiv:0912.3746.

[67] T. Prestegard, E. Thrane, N. L. Christensen, M. W. Coughlin, B. Hubbert, S. Kandhasamy, E. MacAyeal, and V. Mandic, "Identification of noise artifacts in searches for long-duration gravitational-wave transients," *Classical and Quantum Gravity*, vol. 29, p. 095018, May 2012, arXiv:1111.1631. 9

[68] J. Abadie and *et al.*, "Search for gravitational waves from low mass compact binary coalescence in ligo's sixth science run and virgo's science runs 2 and 3," *Physical Review D*, vol. 85, p. 082002, Apr 2012. 10, 21, 41, 42

[69] G. M. Harry and LIGO Scientific Collaboration, “Advanced LIGO: the next generation of gravitational wave detectors,” *Classical and Quantum Gravity*, vol. 27, p. 084006, Apr. 2010. 9

[70] S. J. Waldman, “The Advanced LIGO Gravitational Wave Detector,” *ArXiv e-prints*, Mar. 2011, arXiv:1103.2728.

[71] A. J. Weinstein, for the LIGO Scientific Collaboration, and for the Virgo Collaboration, “Astronomy and astrophysics with gravitational waves in the Advanced Detector Era,” *ArXiv e-prints*, Dec. 2011, arXiv:1112.1057. 9, 21

[72] IndIGO, “Indian initiative in gravitational-wave observation,” May 2012. `http://www.gw-indigo.org`. 9

[73] F. Acernese, M. Alshourbagy, P. Amico, F. Antonucci, S. Aoudia, K. G. Arun, P. Astone, S. Avino, and *et al.*, “Virgo status,” *Classical and Quantum Gravity*, vol. 25, p. 184001, Sept. 2008. 10

[74] H. Grote and *et al.*, “The status of geo 600,” *Classical and Quantum Gravity*, vol. 25, no. 11, p. 114043, 2008. 10

[75] M. Ando, “Current status of the tama300 gravitational-wave detector,” *Classical and Quantum Gravity*, vol. 22, no. 18, pp. 881–889, 2005. 10

[76] K. Kuroda and *et al.*, “Status of lcgt,” *Classical and Quantum Gravity*, vol. 27, p. 084004, 2010. 10

[77] K. Somiya and for the KAGRA Collaboration, “Detector configuration of KAGRA - the Japanese cryogenic gravitational-wave detector,” *ArXiv e-prints*, Nov. 2011, arXiv:1111.7185. 10

[78] M. Punturo and *et al.*, “The Einstein Telescope: a third-generation gravitational wave observatory,” *Classical and Quantum Gravity*, vol. 27, p. 194002, Oct. 2010. 12

[79] B. Sathyaprakash and *et al.*, “Scientific Potential of Einstein Telescope,” *ArXiv e-prints*, Aug. 2011, arXiv:1108.1423. 12

[80] P. Amaro-Seoane, S. Aoudia, S. Babak, P. Binétruy, E. Berti, A. Bohé, C. Caprini, M. Colpi, N. J. Cornish, K. Danzmann, J.-F. Dufaux, J. Gair, O. Jennrich, P. Jetzer, A. Klein, R. N. Lang, A. Lobo, T. Littenberg, S. T. McWilliams, G. Nelemans, A. Petiteau, E. K. Porter, B. F. Schutz, A. Sesana, R. Stebbins, T. Sumner, M. Vallisneri, S. Vitale, M. Volonteri, and H. Ward, "eLISA: Astrophysics and cosmology in the millihertz regime," *ArXiv e-prints*, Jan. 2012, arXiv:1201.3621. 12, 45

[81] P. Amaro-Seoane, S. Aoudia, S. Babak, P. Binetruy, E. Berti, A. Bohe, C. Caprini, M. Colpi, N. J. Cornish, K. Danzmann, J.-F. Dufaux, J. Gair, O. Jennrich, P. Jetzer, A. Klein, R. N. Lang, A. Lobo, T. Littenberg, S. T. McWilliams, G. Nelemans, A. Petiteau, E. K. Porter, B. F. Schutz, A. Sesana, R. Stebbins, T. Sumner, M. Vallisneri, S. Vitale, M. Volonteri, and H. Ward, "Low-frequency gravitational-wave science with eLISA/NGO," *ArXiv e-prints*, Feb. 2012, arXiv:1202.0839. 12, 45

[82] O. Jennrich, "Lisa – a mission overview," *37th COSPAR Scientific Assembly*, vol. 37, p. 1367, 2008. 12, 45

[83] J. S. Key and N. J. Cornish, "Characterizing spinning black hole binaries in eccentric orbits with LISA," *Physical Review D*, vol. 83, p. 083001, Apr. 2011, arXiv:1006.3759. 12

[84] A. Petiteau, S. Babak, and A. Sesana, "Constraining the Dark Energy Equation of State Using LISA Observations of Spinning Massive Black Hole Binaries," *The Astrophysical Journal*, vol. 732, p. 82, May 2011, arXiv:1102.0769. 12

[85] T. Regimbau, "The astrophysical gravitational wave stochastic background," *Research in Astronomy and Astrophysics*, vol. 11, pp. 369–390, Apr. 2011, arXiv:1101.2762. 12

[86] P. A. Rosado, "Gravitational wave background from binary systems," *Physical Review D*, vol. 84, p. 084004, Oct. 2011, arXiv:1106.5795. 12

[87] M. Armano and *et al.*, "Lisa pathfinder: the experiment and the route to lisa," *Classical and Quantum Gravity*, vol. 26, p. 094001, 2009. 12

[88] M. Pitkin, S. Reid, S. Rowan, and J. Hough, “Gravitational Wave Detection by Interferometry (Ground and Space),” *Living Reviews in Relativity*, vol. 14, p. 5, July 2011, arXiv:1102.3355. 12

[89] K. J. Lee, N. Wex, M. Kramer, B. W. Stappers, C. G. Bassa, G. H. Janssen, R. Karuppusamy, and R. Smits, “Gravitational wave astronomy of single sources with a pulsar timing array,” *Monthly Notices of the Royal Astronomical Society*, vol. 414, pp. 3251–3264, July 2011, arXiv:1103.0115. 12

[90] A. Sesana and A. Vecchio, “Measuring the parameters of massive black hole binary systems with pulsar timing array observations of gravitational waves,” *Physical Review D*, vol. 81, p. 104008, May 2010, arXiv:1003.0677.

[91] S. J. Chamberlin and X. Siemens, “Stochastic backgrounds in alternative theories of gravity: Overlap reduction functions for pulsar timing arrays,” *Physical Review D*, vol. 85, p. 082001, Apr. 2012, arXiv:1111.5661. 12

[92] J. Marx, K. Danzmann, J. Hough, K. Kuroda, D. McClelland, B. Mours, S. Phinney, S. Rowan, B. Sathyaprakash, F. Vetrano, S. Vitale, S. Whitcomb, and C. Will, “The Gravitational Wave International Committee Roadmap: The future of gravitational wave astronomy,” *ArXiv e-prints*, Nov. 2011, arXiv:1111.5825. 12

[93] R. Trotta, “Bayes in the sky: Bayesian inference and model selection in cosmology,” *Contemporary Physics*, vol. 49, pp. 71–104, 2008, arXiv:0803.4089. 14

[94] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *J. Chem. Phys.*, vol. 21, no. 6, p. 1087, 1953. 14

[95] W. K. Hastings, “Monte carlo sampling methods using markov chains and their applications,” *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970. 14

[96] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, p. 671, 1983. 14

[97] D. J. Earl and M. W. Deem, “Parallel tempering: Theory, applications, and new perspectives,” *Physical Chemistry Chemical Physics (Incorporating Faraday Transactions)*, vol. 7, p. 3910, 2005, arXiv:physics/0508111. 14

[98] W. M. Farr and I. Mandel, “An Efficient Interpolation Technique for Jump Proposals in Reversible-Jump Markov Chain Monte Carlo Calculations,” *ArXiv e-prints*, Apr. 2011, arXiv:1104.0984. 14

[99] J. Skilling, “Nested sampling,” in *24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (R. Fischer, R. Preuss, & U. V. Toussaint, ed.), vol. 735 of *American Institue of Physics Conference Proceedings*, pp. 395–405, American Institute of Physics, 2004. 14

[100] J. Skilling, “Nested sampling for general bayesian computation,” *Bayesian Analysis*, vol. 1, no. 4, pp. 833–860, 2006. 14

[101] F. Feroz and M. P. Hobson, “Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses,” *Monthly Notices of the Royal Astronomical Society*, vol. 384, no. 2, pp. 449–463, 2008, arXiv:0704.3704. 15, 17, 24

[102] F. Feroz, M. P. Hobson, and M. Bridges, “Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics,” *Monthly Notices of the Royal Astronomical Society*, vol. 398, no. 4, pp. 1601–1614, 2009, arXiv:0809.3437. 17, 18, 24, 137, 144, 162

[103] F. Feroz, P. J. Marshall, and M. P. Hobson, “Cluster detection in weak lensing surveys,” *ArXiv e-prints*, Oct. 2008, arXiv:0810.0781. 17

[104] F. Feroz, M. P. Hobson, J. T. L. Zwart, R. D. E. Saunders, and K. J. B. Grainge, “Bayesian modelling of clusters of galaxies from multifrequency-pointed Sunyaev-Zel’dovich observations,” *Monthly Notices of the Royal Astronomical Society*, vol. 398, pp. 2049–2060, Oct. 2009, arXiv:0811.1199.

[105] F. Feroz, B. C. Allanach, M. Hobson, S. S. Abdus Salam, R. Trotta, and A. M. Weber, “Bayesian selection of sign μ within mSUGRA in global fits including

WMAP5 results," *Journal of High Energy Physics*, vol. 10, p. 64, Oct. 2008, arXiv:0807.4512.

[106] R. Trotta, F. Feroz, M. Hobson, L. Roszkowski, and R. Ruiz de Austri, "The impact of priors and observables on parameter inferences in the constrained MSSM," *Journal of High Energy Physics*, vol. 12, p. 24, Dec. 2008, arXiv:0809.3792. 17

[107] F. Feroz, J. Gair, M. P. Hobson, and E. K. Porter, "Use of the multinest algorithm for gravitational wave data analysis," *Classical and Quantum Gravity*, vol. 26, no. 21, p. 215003, 2009, arXiv:0904.1544. 17, 28, 63

[108] J. Veitch and A. Vecchio, "Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network," *Physical Review D*, vol. 81, p. 062003, Mar. 2010, arXiv:0911.3820. 17

[109] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, B. Allen, and *et al.*, "TOPICAL REVIEW: Predictions for the rates of compact binary coalescences observable by ground-based gravitational-wave detectors," *Classical and Quantum Gravity*, vol. 27, p. 173001, Sept. 2010, arXiv:1003.2480. 21

[110] L. S. Collaboration, "Lalsuite home page," May 2012. `https://www.lsc-group.phys.uwm.edu/daswg/projects/lalsuite.html`. 21

[111] P. Graff, "A coherent bayesian method tested on hardware injection time-slide triggers," 2011. LIGO-G1100599. 21, 41

[112] C. J. Lada, "Stellar Multiplicity and the Initial Mass Function: Most Stars Are Single," *The Astrophysical Journal*, vol. 640, pp. L63–L66, Mar. 2006, arXiv:astro-ph/0601375. 22

[113] J. Veitch and A. Vecchio, "Bayesian approach to the follow-up of candidate gravitational wave signals," *Physical Review D*, vol. 78, no. 2, p. 022001, 2008, arXiv:0801.4313. 22, 26, 28

[114] J. Veitch and A. Vecchio, "Assigning confidence to inspiral gravitational wave candidates with bayesian model selection," *Classical and Quantum Gravity*, vol. 25, no. 18, p. 184010, 2008, arXiv:0807.4483. 28, 46, 64

[115] M. Trias, A. Vecchio, and J. Veitch, "Delayed rejection schemes for efficient markov-chain monte-carlo sampling of multimodal distributions," *ArXiv e-prints*, 2009, arXiv:0904.2207. 28

[116] M. Trias, A. Vecchio, and J. Veitch, "Studying stellar binary systems with the laser interferometer space antenna using delayed rejection markov chain monte carlo methods," *Classical and Quantum Gravity*, vol. 26, no. 20, p. 204024, 2009, arXiv:0905.2976. 28

[117] B. Farr, S. Fairhurst, and B. S. Sathyaprakash, "Searching for binary coalescences with inspiral templates: Detection and parameter estimation," *Classical and Quantum Gravity*, vol. 26, no. 11, p. 114009, 2009, arXiv:0902.0307. 28

[118] F. Feroz, P. J. Marshall, and M. P. Hobson, "Cluster detection in weak lensing surveys," *ArXiv e-prints*, 2008, arXiv:0810.0781. 28

[119] J. S. Bloom, D. E. Holz, S. A. Hughes, K. Menou, A. Adams, S. F. Anderson, A. Becker, G. C. Bower, N. Brandt, B. Cobb, K. Cook, A. Corsi, S. Covino, D. Fox, A. Fruchter, C. Fryer, J. Grindlay, D. Hartmann, Z. Haiman, B. Kocsis, L. Jones, A. Loeb, S. Marka, B. Metzger, E. Nakar, S. Nissanke, D. A. Perley, T. Piran, D. Poznanski, T. Prince, J. Schnittman, A. Soderberg, M. Strauss, P. S. Shawhan, D. H. Shoemaker, J. Sievers, C. Stubbs, G. Tagliaferri, P. Ubertini, and P. Wozniak, "Astro2010 Decadal Survey Whitepaper: Coordinated Science in the Gravitational and Electromagnetic Skies," *ArXiv e-prints*, Feb. 2009, arXiv:0902.1527. 28

[120] The LIGO Scientific Collaboration, Virgo Collaboration: J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, and *et al.*, "Implementation and testing of the first prompt search for gravitational wave transients with electromagnetic counterparts," *ArXiv e-prints*, Sept. 2011, arXiv:1109.3498.

[121] J. Kanner, T. L. Huard, S. Márka, D. C. Murphy, J. Piscionere, M. Reed, and P. Shawhan, "LOOC UP: locating and observing optical counterparts to gravitational wave bursts," *Classical and Quantum Gravity*, vol. 25, p. 184034, Sept. 2008, arXiv:0803.0312.

[122] D. B. for the LVC, "Very low latency search pipeline for low mass compact binary coalescences in the ligo s6 and virgo vsr2 data," *Classical and Quantum Gravity*, vol. 27, no. 19, p. 194013, 2010.

[123] D. M. Coward, B. Gendre, P. J. Sutton, E. J. Howell, T. Regimbau, M. Laas-Bourez, A. Klotz, M. Boër, and M. Branchesi, "Towards an optimal search strategy of optical and gravitational wave emissions from binary neutron star coalescence," *Monthly Notices of the Royal Astronomical Society*, vol. 415, pp. L26–L30, July 2011, arXiv:1104.5552.

[124] P. A. Evans, J. K. Fridriksson, N. Gehrels, J. Homan, J. P. Osborne, M. Siegel, A. Beardmore, P. Handbauer, J. Gelbord, J. A. Kennea, and *et al.*, "Swift follow-up observations of candidate gravitational-wave transient events," *ArXiv e-prints*, May 2012, arXiv:1205.1124.

[125] T. Piran, E. Nakar, and S. Rosswog, "The Electromagnetic Signals of Compact Binary Mergers," *ArXiv e-prints*, Apr. 2012, arXiv:1204.6242.

[126] J. A. Rueda and R. Ruffini, "Gravitational Waves versus Electromagnetic Emission in Gamma-Ray Bursts," *ArXiv e-prints*, May 2012, arXiv:1205.6915. 28

[127] L. Wen and Y. Chen, "Geometrical expression for the angular resolution of a network of gravitational-wave detectors," *Physical Review D*, vol. 81, p. 082001, Apr. 2010, arXiv:1003.2504. 29

[128] S. Nissanke, J. Sievers, N. Dalal, and D. Holz, "Localizing Compact Binary Inspirals on the Sky Using Ground-based Gravitational Wave Interferometers," *The Astrophysical Journal*, vol. 739, p. 99, Oct. 2011, arXiv:1105.3184.

[129] S. Klimenko, G. Vedovato, M. Drago, G. Mazzolo, G. Mitselmakher, C. Pankow, G. Prodi, V. Re, F. Salemi, and I. Yakushin, "Localization of gravitational wave sources with networks of advanced detectors," *Physical Review D*, vol. 83, p. 102001, May 2011, arXiv:1101.5408.

[130] A. Manzotti and A. Dietz, "Prospects for early localization of gravitational-wave signals from compact binary coalescences with advanced detectors," *ArXiv e-prints*, Feb. 2012, arXiv:1202.4031.

[131] I. Mandel, L. Z. Kelley, and E. Ramirez-Ruiz, "Towards Improving the Prospects for Coordinated Gravitational-Wave and Electromagnetic Observations," in *IAU Symposium*, vol. 285 of *IAU Symposium*, pp. 358–360, Apr. 2012, arXiv:1111.0005.

[132] L. K. Nuttall and P. J. Sutton, "Identifying the host galaxy of gravitational wave signals," *Physical Review D*, vol. 82, p. 102002, Nov. 2010, arXiv:1009.1791.

[133] S. Fairhurst, "Source localization with an advanced gravitational wave detector network," *Classical and Quantum Gravity*, vol. 28, no. 10, p. 105021, 2011.

[134] S. Fairhurst, "Improved source localization with LIGO India," *ArXiv e-prints*, May 2012, arXiv:1205.6611. 29

[135] P. Ajith and S. Bose, "Estimating the parameters of nonspinning binary black holes using ground-based gravitational-wave detectors: Statistical errors," *Physical Review D*, vol. 79, p. 084032, Apr. 2009, arXiv:0901.4936. 29

[136] C. Cutler and É. E. Flanagan, "Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform?," *Physical Review D*, vol. 49, pp. 2658–2697, Mar. 1994, arXiv:gr-qc/9402014. 29

[137] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975. 36

[138] L. Singer, L. Price, and A. Speranza, "Optimizing optical follow-up of gravitational-wave candidates," *ArXiv e-prints*, Apr. 2012, arXiv:1204.4510. 40

[139] L. S. Collaboration, “The science of lsc research,” May 2012. `http://www.ligo.org/science/GW100916/`. 41

[140] D. Buskulic, Virgo Collaboration, and LIGO Scientific Collaboration, “Very low latency search pipeline for low mass compact binary coalescences in the LIGO S6 and Virgo VSR2 data,” *Classical and Quantum Gravity*, vol. 27, p. 194013, Oct. 2010. 41

[141] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, B. Allen, and *et al.*, “Search for gravitational waves from compact binary coalescence in LIGO and Virgo data from S5 and VSR1,” *Physical Review D*, vol. 82, p. 102001, Nov. 2010, arXiv:1005.4655. 41

[142] The LIGO Scientific Collaboration, the Virgo Collaboration: J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, and *et al.*, “Sensitivity to Gravitational Waves from Compact Binary Coalescences Achieved during LIGO’s Fifth and Virgo’s First Science Run,” *ArXiv e-prints*, Mar. 2010, arXiv:1003.2481.

[143] J. Abadie, B. P. Abbott, R. Abbott, T. Accadia, F. Acernese, R. Adhikari, P. Ajith, B. Allen, G. Allen, E. Amador Ceron, and *et al.*, “All-sky search for gravitational-wave bursts in the first joint LIGO-GEO-Virgo run,” *Physical Review D*, vol. 81, p. 102001, May 2010, arXiv:1002.1036.

[144] J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, C. Affeldt, and *et al.*, “All-sky search for periodic gravitational waves in the full S5 LIGO data,” *Physical Review D*, vol. 85, p. 022001, Jan. 2012, arXiv:1110.0208. 41

[145] I. Mandel and R. O’Shaughnessy, “Compact binary coalescences in the band of ground-based gravitational-wave detectors,” *Classical and Quantum Gravity*, vol. 27, p. 114007, June 2010, arXiv:0912.1074. 41

[146] F. Feroz, B. C. Allanach, M. P. Hobson, S. S. A. Salam, R. Trotta, and A. M. Weber, “Bayesian selection of sign μ within msugra in global fits including wmap5

results," *Journal of High Energy Physics*, vol. 10, p. 64, 2008, arXiv:0807.4512. 42

[147] L. Barack and C. Cutler, "LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy," *Physical Review D*, vol. 69, p. 082005, Apr. 2004, arXiv:gr-qc/0310125. 45

[148] J. R. Gair, A. Sesana, E. Berti, and M. Volonteri, "Constraining properties of the black hole population using LISA," *Classical and Quantum Gravity*, vol. 28, p. 094018, May 2011, arXiv:1009.6172. 45

[149] F. Feroz, J. Gair, P. Graff, M. P. Hobson, and A. Lasenby, "Classifying lisa gravitational wave burst signals using bayesian evidence," *Classical and Quantum Gravity*, vol. 27, no. 7, p. 075010, 2010, arXiv:0911.0288. 45, 46, 93

[150] S. Babak and *et al.*, "The mock lisa data challenges: from challenge 1b to challenge 3," *Classical and Quantum Gravity*, vol. 25, no. 18, p. 184026, 2008, arXiv:0806.2110. 46, 47, 51, 65

[151] S. Babak and *et al.*, "The mock lisa data challenges: from challenge 3 to challenge 4," *Classical and Quantum Gravity*, vol. 27, no. 8, p. 084009, 2010, arXiv:0912.0548. 46, 73

[152] M. I. Cohen, C. Cutler, and M. Vallisneri, "Searches for cosmic-string gravitational-wave bursts in mock lisa data," *Classical and Quantum Gravity*, vol. 27, no. 18, p. 185012, 2010, arXiv:1002.4153. 46

[153] T. Littenberg and N. Cornish, "Bayesian approach to the detection problem in gravitational wave astronomy," *Physical Review D*, vol. 80, no. 6, p. 063007, 2009, arXiv:0902.0368. 46, 64

[154] F. Beauville and *et al.*, "A first comparison of search methods for gravitational wave bursts using ligo and virgo simulated data," *Classical and Quantum Gravity*, vol. 22, no. 18, pp. S1293–S1301, 2005. 46, 57

[155] T. Damour and A. Vilenkin, "Gravitational wave bursts from cusps and kinks on cosmic strings," *Physical Review D*, vol. 64, no. 6, p. 064008, 2001, arXiv:gr-qc/0104026. 47, 51

[156] X. Siemens and K. D. Olum, “Cosmic string cusps with small-scale structure: Their forms and gravitational waveforms,” *Physical Review D*, vol. 68, no. 8, p. 085017, 2003, arXiv:gr-qc/0307113. 47, 51

[157] J. S. Key and N. J. Cornish, “Characterizing the gravitational wave signature from cosmic string cusps,” *Physical Review D*, vol. 79, no. 4, p. 043014, 2009, arXiv:0812.1590. 47, 48, 63

[158] L. J. Rubbo, N. J. Cornish, and O. Poujade, “Forward modeling of space-borne gravitational wave detectors,” *Physical Review D*, vol. 69, no. 8, p. 082003, 2004, arXiv:gr-qc/0311069. 48

[159] M. Tinto and S. V. Dhurandhar, “Time-delay interferometry,” *Living Reviews in Relativity*, vol. 8, no. 4, 2005. 48

[160] S. W. Helstrom, *Statistical Theory of Signal Detection*. London, UK: Pergamon, 1968. 49

[161] B. J. Owen, “Search templates for gravitational waves from inspiraling binaries: Choice of template spacing,” *Physical Review D*, vol. 53, no. 12, pp. 6749–6761, 1996, arXiv:gr-qc/9511032. 49

[162] “Mock lisa data challenge,” Jan. 2012. `http://astrogravs.nasa.gov/docs/mldc/`. 52, 59

[163] C. Röver, M.-A. Bizouard, N. Christensen, H. Dimmelmeier, I. S. Heng, and R. Meyer, “Bayesian reconstruction of gravitational wave burst signals from simulations of rotating stellar core collapse and bounce,” *Physical Review D*, vol. 80, no. 10, p. 102004, 2009, arXiv:0909.1093. 65

[164] G. Nelemans, “The galactic gravitational wave foreground,” *Classical and Quantum Gravity*, vol. 26, no. 9, p. 094030, 2009, arXiv:0901.1778. 65

[165] N. J. Cornish and T. Littenberg, “Tests of bayesian model selection techniques for gravitational wave astronomy,” *Physical Review D*, vol. 76, no. 8, p. 083006, 2007, arXiv:0704.1808. 66

[166] A. Błaut, S. Babak, and A. Królak, “Mock lisa data challenge for the galactic white dwarf binaries,” *Physical Review D*, vol. 81, p. 063008, 2010, arXiv:0911.3020. 66, 67

[167] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *Computer Journal*, vol. 7, pp. 308–313, 1965. 66

[168] G. Nelemans, L. R. Yungelson, and S. F. Portegies Zwart, “Short-period AM CVn systems as optical, X-ray and gravitational-wave sources,” *Monthly Notices of the Royal Astronomical Society*, vol. 349, pp. 181–192, Mar. 2004, arXiv:astro-ph/0312193. 66

[169] T. B. Littenberg, “Detection pipeline for Galactic binaries in LISA data,” *Physical Review D*, vol. 84, p. 063009, Sept. 2011, arXiv:1106.6355. 70

[170] S. E. Timpano, L. J. Rubbo, and N. J. Cornish, “Characterizing the galactic gravitational wave background with lisa,” *Physical Review D*, vol. 73, p. 122001, 2006, arXiv:gr-qc/0504071. 73

[171] P. Protopapas, R. Jimenez, and C. Alcock, “Fast identification of transits from light-curves,” *Monthly Notices of the Royal Astronomical Society*, vol. 362, no. 2, pp. 460–468, 2005, arXiv:astro-ph/0502301. 88, 89, 102

[172] P. Graff, M. P. Hobson, and A. Lasenby, “An investigation into the multiple optimised parameter estimation and data compression algorithm,” *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 413, no. 1, pp. L66–L70, 2011, arXiv:1010.5907. 88

[173] A. Heavens, R. Jimenez, and O. Lahav, “Massive lossless data compression and multiple parameter estimation from galaxy spectra,” *Monthly Notices of the Royal Astronomical Society*, vol. 317, no. 4, pp. 965–972, 2000, arXiv:astro-ph/9911102. 88, 89, 92

[174] S. Gupta and A. Heavens, “Fast parameter estimation from the cosmic microwave background power spectrum,” *Monthly Notices of the Royal Astronomical Society*, vol. 334, no. 1, pp. 167–172, 2002, arXiv:astro-ph/0108315. 89, 100

[175] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003. 103, 107

[176] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networksuniversal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, p. 359, 1990. 104, 133

[177] S. F. Gull and J. Skilling, *Quantified Maximum Entropy: MemSys 5 Users' Manual*. Bury St. Edmonds, UK: Maximum Entropy Data Consults, Ltd., 1999. 108

[178] J. Martens, "Deep learning via hessian-free optimization," in *Proceedings of the 27th International Conference on Machine Learning* (J. Fürnkranz and T. Joachims, eds.), (Haifa, Israel), pp. 735–742, Omnipress, 2010. 110

[179] N. N. Schraudolph, "Fast curvature matrix-vector products for second-order gradient descent," *Neural Computation*, vol. 14, no. 7, pp. 1723–1738, 2002. 110

[180] B. A. Pearlmutter, "Fast exact multiplication by the hessian," *Neural Computation*, vol. 6, no. 1, pp. 147–160, 1994. 110

[181] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006. 112

[182] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006. 112, 122, 126

[183] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992. 114

[184] R. Neal, "A three-way classification problem," Jan. 2012. `http://www.cs.toronto.edu/∼radford/fbm.2004-11-10.doc/Ex-netgp-c.html`. 116, 117

[185] Y. LeCun and C. Cortes, "Mnist handwritten digit database," Jan. 2012. `http://yann.lecun.com/exdb/mnist/`. 120

[186] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, “Deep big simple neural nets excel on handwritten digit recognition,” *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, 2010, arXiv:1003.0358. 120

[187] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 120

[188] T. Kitching, S. Balan, G. Bernstein, M. Bethge, S. Bridle, F. Courbin, M. Gentile, A. Heavens, M. Hirsch, R. Hosseini, A. Kiessling, A. Amara, D. Kirk, K. Kuijken, R. Mandelbaum, B. Moghaddam, G. Nurbaeva, S. Paulin-Henriksson, A. Rassat, J. Rhodes, B. Schölkopf, J. Shawe-Taylor, M. Gill, M. Shmakova, A. Taylor, M. Velander, L. van Waerbeke, D. Witherick, D. Wittman, S. Harmeling, C. Heymans, R. Massey, B. Rowe, T. Schrabback, and L. Voigt, “Gravitational Lensing Accuracy Testing 2010 (GREAT10) Challenge Handbook,” *ArXiv e-prints*, Sept. 2010, arXiv:1009.0779. 125

[189] T. D. Kitching, J. Rhodes, R. M. C. Heymans, Q. Liu, M. Cobzarenco, B. L. Cragin, A. Hassaine, D. Kirkby, E. J. Lok, D. Margala, J. Moser, M. O’Leary, A. M. Pires, and S. Yurgenson, “Image analysis for cosmology: Shape measurement challenge review and results from the mapping dark matter challenge,” *New Astronomy Reviews (submitted)*, 2012, arXiv:1204.4096. 127, 130

[190] K. Inc, “Mapping dark matter challenge,” Jan. 2012. `http://www.kaggle.com/c/mdm/`. 127, 128, 130

[191] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby, “BAMBI: blind accelerated multimodal Bayesian inference,” *Monthly Notices of the Royal Astronomical Society*, vol. 421, pp. 169–180, Mar. 2012, arXiv:1110.2997. 133, 134

[192] A. Lewis and S. Bridle, “Cosmological parameters from cmb and other data: A monte carlo approach,” *Physical Review D*, vol. 66, no. 10, p. 103511, 2002, arXiv:astro-ph/0205436. 142, 144

[193] T. Auld, M. Bridges, M. P. Hobson, and S. F. Gull, “Fast cosmological parameter estimation using neural networks,” *Monthly Notices of the Royal Astronomical*

Society: Letters, vol. 376, no. 1, pp. L11–L15, 2007, arXiv:astro-ph/0608174. 144, 151

[194] T. Auld, M. Bridges, and M. P. Hobson, “Cosmonet: fast cosmological parameter estimation in non-flat models using neural networks,” *Monthly Notices of the Royal Astronomical Society*, vol. 387, no. 4, pp. 1575–1582, 2008, arXiv:astro-ph/0703445. 151

[195] A. Bouland, R. Easther, and K. Rosenfeld, “Caching and interpolated likelihoods: accelerating cosmological monte carlo markov chains,” *Journal of Cosmology and Astroparticle Physics*, vol. 5, p. 16, 2011, arXiv:1012.5299.

[196] W. A. Fendt and B. D. Wandelt, “Pico: Parameters for the impatient cosmologist,” *The Astrophysical Journal*, vol. 654, no. 1, pp. 2–11, 2007, arXiv:astro-ph/0606709.

[197] J. Prasad and T. Souradeep, “Cosmological parameter estimation using particle swarm optimization (pso),” *ArXiv e-prints*, 2011, arXiv:1108.5600.

[198] S. F. Daniel, A. J. Connolly, and J. Schneider, “An Efficient Parameter Space Search as an Alternative to Markov Chain Monte Carlo,” *ArXiv e-prints*, May 2012, arXiv:1205.2708. 144

[199] A. Lewis, A. Challinor, and A. Lasenby, “Efficient computation of cosmic microwave background anisotropies in closed friedmann-robertson-walker models,” *The Astrophysical Journal*, vol. 538, no. 2, pp. 473–476, 2000, arXiv:astro-ph/9911177. 144

[200] D. Larson, J. Dunkley, G. Hinshaw, E. Komatsu, M. R. Nolta, C. L. Bennett, B. Gold, M. Halpern, R. S. Hill, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, N. Odegard, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, “Seven-year wilkinson microwave anisotropy probe (wmap) observations: Power spectra and wmap-derived parameters,” *The Astrophysical Journal Supplement*, vol. 192, no. 2, p. 16, 2011, arXiv:1001.4635. 144

[201] D. J. C. MacKay, “Probable networks and plausible predictions - a review of practical bayesian methods for supervised neural networks,” *Network: Computation in Neural Systems*, vol. 6, p. 469, 1995. 152, 157

[202] M. Betancourt, “Nested sampling with constrained hamiltonian monte carlo,” in *American Institute of Physics Conference Series* (A. Mohammad-Djafari, J.-F. Bercher, & P. Bessiére, ed.), vol. 1305 of *American Institute of Physics Conference Series*, (New York, New York), pp. 165–172, American Institute of Physics, 2011, arXiv:1005.0157. 164

[203] J. Sohl-Dickstein, “Hamiltonian Monte Carlo with Reduced Momentum Flips,” *ArXiv e-prints*, May 2012, arXiv:1205.1939. 164

The End.