TOPICS IN ASTROSTATISTICS: STELLAR BINARY EVOLUTION, GRAVITATIONAL - WAVE SOURCE MODELLING AND STOCHASTIC PROCESSES

by

JAMES WILLIAM BARRETT

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF
BIRMINGHAM

School of Physics and Astronomy
College of Engineering and Physical Sciences
University of Birmingham
May, 2018

Abstract

The effective use of statistical techniques is one of the cornerstones of modern astrophysics. In this thesis we use sophisticated statistical methodology to expand our understanding of astrophysics. In particular, we focus on the physics of coalescing binary black holes, and the observation of these events using gravitational wave astronomy. We use Fisher matrices to explore how much we expect to learn from these observations, and then use cutting edge machine learning techniques, including random forests and Gaussian processes, to faciliate a Bayesian comparison of real observations to our model. We find that Gaussian processes perform poorly compared to random forests, and present a discussion of their performances. Finally, we develop a technique, based on Gaussian processes, for characterising stochastic variability in time series data.

Acknowledgements

Words can't express how grateful I am for all of the people in my life, but I'll give it a go.

Mum, thank you for the unwaivering support. I don't know where I'd be without you. I am proud to call you my mum.

Dad, thank you for always being there, for the frequent drives to Birmingham, for the snooker and for always being interested in what I do. You're a great man, and I'm proud to call you my dad.

Steph, who probably understands me better than anyone. You're the best sister a man could ask for. I'm proud of you.

Nana, Grandpa, Uncle, Julie, Granny, Auntie Ally, Uncle Neil, Auntie Debbie, my family is my strength. I hope you all understand how much I appreciate what you do for me.

To my Lowestoft friends; Robert Button, Tim Meadows, Lee Phillips, Phoebe Boatman, David Button, Andrew Button, Matt Rayner. For loving me for who I am. May there be many more years, and many more trips to the Triangle.

To the friends I met in Birmingham; Robin Smith, Daniel Töyra, Matt Hunt, Serena Vinciguerra, Hannah Middleton, Emily Takeva, Holly Ranger, Simon Killarney (Killa), Alan Woolaston (Singerman), Joe Kadar (Brush). My years in Birmingham have been the best of my life, and I have you to thank.

To team `COMPAS`, Alejandro Vigna-Gómez and Coen Neijssel, for your friendship and patience, and for all the pints, I say thank you.

To my supervisors; Ilya Mandel, thanks for taking me on as your student, and

for all of your guidance and support. Will Farr, I wouldn't be half the programmer or statistician I am today without your guidance, thank you.

David Stops, I hope you will enjoy reaching the end of a working day without me pestering you. Thank you for you friendship and unending patience for all my IT woes.

And to everyone else at the University of Birmingham. There are too many of you to mention, but you know who you are. Thank you.

I would also like to thank my thesis examiners, Peter Tino and Derek Bingham, for taking the time to read my thesis and providing useful and thorough feedback. I also thank them for the useful discussion during my viva.

Finally, of course, I thank the STFC for funding all of the work in this thesis.

Declaration

Chapter 1 introduces the astrophysical context for the chapters that follow, specifically binary evolution and black holes, and introduces the statistical tools used in later chapters. It also provides an introduction to stochastic variability and correlated noise processes.

The COMPAS software package, described in chapter 1 and used extensively for the work presented in chapters 2 and 3 was collaboratively developed by myself, Alejandro Vigna-Gómez, Simon Stevenson, Coenraad Neijssel and Ilya Mandel.

Chapter 2 reproduces the text of Barrett et al. [2017a], submitted to Monthly Notices of the Royal Astronomical Society. Some modifications have been made to reflect the wishes of the thesis examiners. This work was led by myself, in collaboration with Sebastian Gaebel, Coenraad Neijssel, Alejandro Vigna-Gómez, Simon Stevenson, Christopher Berry, Will Farr and Ilya Mandel. The code dealing with selection effects was written by Sebastian Gaebel, with modifications by myself. The code dealing with cosmic history was written by Coenraad Neijssel, with modifications by myself. All other analysis code was written by myself. The text of the paper was written by myself, with modifications from all authors.

Chapter 3 describes my work on emulators for population synthesis models. This work was conducted in collaboration with Coenraad Neijssel, Sebastian Gaebel, Alejandro Vigna-Gómez and Ilya Mandel. It makes use of the same selection effects and cosmic history codes attributed above to Sebastian Gaebel and Coenraad Neijssel respectively. All other code was written by myself, except for some libraries which are appropriately referenced in the text of this thesis. All text is my own.

Chapter 4 describes my work on inference using observations of stochastic phenomena and contains the modified text of an unpublished manuscript written by myself in collaboration with Will Farr. All code was written by myself.

Chapter 5 gives a summary of the thesis and my conclusions. The conclusions are my own.

Contents

Chapter 1

Introduction

1.1 Binary Stars

The methods in this thesis are all presented in the context of astrophysics, and in particular the physics of binary stars. Binary stars are defined to be two stars which are in a bound orbit about their mutual centre of mass. Observations suggest that a large fraction of stars are in binary systems, with almost all massive stars (whose mass is more than around 10 times the mass of our own sun) thought to be in binary (or higher multiple) systems [Sana et al., 2012a]. In some cases, when the stars within binaries come close enough to one another, they can interact. Binary evolution has a dramatic effect on the fate of the stars. [Tutukov and Yungelson, 1973]. There are many observable astrophysical phenomena which can be attributed to binary interactions, including X-Ray binaries [Podsiadlowski et al., 2002, Liu et al., 2006], short gamma ray bursts [Abbott et al., 2017a], type IA supernovae [Hillebrandt and Niemeyer, 2000], luminous red novae [Ivanova et al., 2013a] and gravitational-waves from compact object mergers [Abbott et al., 2016a, 2017b]. In this thesis we focus our attention on the observation of gravitational-waves from compact object mergers, and in particular from binary black holes.

In this section of the introduction, we give a brief overview of gravitational-waves, and how a stellar binary system evolves to become pair of black holes which will coalesce, producing observable gravitational-waves. We then introduce rapid population synthesis as a method of studying binary evolution through the characteristics of populations of binaries. We give some detail of methodology underpinning rapid population synthesis, paying particular attention to the physical processes we relate our statistical methodology to in the later chapters of this thesis. We also introduce some specifics of `COMPAS`, the rapid population synthesis code used to generate the data used in later chapters.

1.1.1 Gravitational-Waves

Gravitational-waves were first predicted in Einstein's seminal papers on the general theory of relativity [Einstein, 1916a,b]. They are perturbations to spacetime, radiating energy, angular momentum and information from gravitational interactions. They travel at the speed of light, and manifest as a planar wave "stretching" and "squashing" the spacetime they propagate through. gravitational-waves were recently detected for the first time from the merger of two black holes (GW150914, Abbott et al. [2016b]). Since then, five more gravitational wave detections have been made; four from binary black hole mergers (GW151226, GW170104, GW170608, GW170814, Abbott et al. [2016a, 2017c,d,e]) and one from a binary neutron star merger (GW170817,Abbott et al. [2017a]), together with one less statistically significant event (LVT151012,Abbott et al. [2016a]), which has an 86% probability of having astrophysical origin.

These events were detected by the Advanced Laser Interferometer Gravitational-wave Observatory (aLIGO), with the latter two also detected by Advanced VIRGO (AdV). These detectors use large scale Michelson interferometers to detect the perturbations to spacetime caused by gravitational-wave propagation. A schematic diagram of a simple Michelson interferometer can be seen in figure 1.1. The basic principle is that a laser beam is split and sent along two long, perpendicular "arms", reflected from mirrors (which act as test masses to be perturbed by gravitational-waves), returning along their paths to recombine, and then travel towards an instrument to measure the recombined beam. If there has been no change to the spacetime

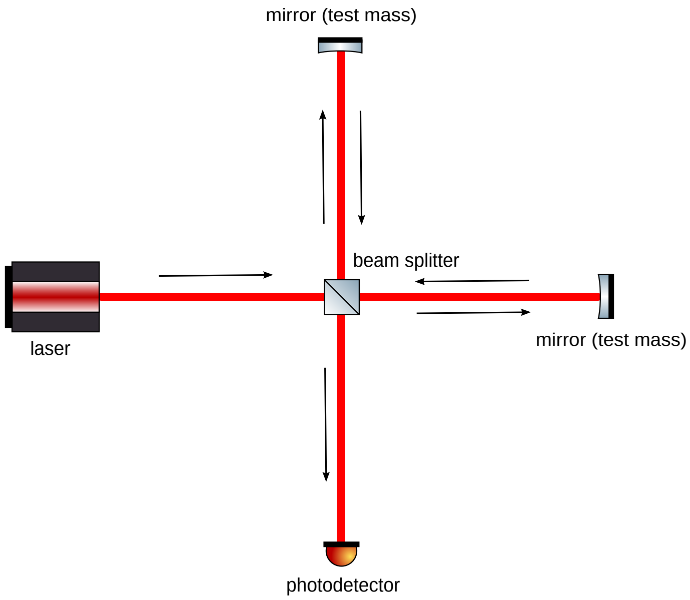


Figure 1.1: A cartoon of an interferometer. The gravitational-wave detectors of both aLIGO and AdV are based on this principle design, although with many more complicated optical components to increase their sensitivity.

of each beam's path, the beams are designed to recombine destructively. If, due to a passing gravitational wave the arm lengths have changed length, then there will be some detectable interference pattern [Abbott et al., 2004].

The gravitational wave signal from a compact object merger, detected by an interferometer such as this, will have a characteristic waveform, consisting of three distinct parts. The 'chirp'; a sinusoidal signal increasing in both frequency and amplitude over the last few orbits of the system, the merger; where the waveform decays rapidly and then the ringdown; where the signal goes to zero once the objects have merged. Figure 1.2 shows the measured characteristic strain from the first detection of gravitational-waves (GW150914) together with a theoretical waveform consistent with the data, calculated using numerical relativity Cardoso et al. [2015]. The characteristic strain (h) is a dimensionless quantity representing the amount that the path length of the laser (ΔL) has changed relative to its overall path length (L) so that $h = (\Delta L)/L$.

Gravitational-waves are a weak mechanism for dissipating orbital energy from a binary system. It can take a long time for systems to merge due to the emission of gravitational radiation. If we are to detect gravitational-waves from these systems

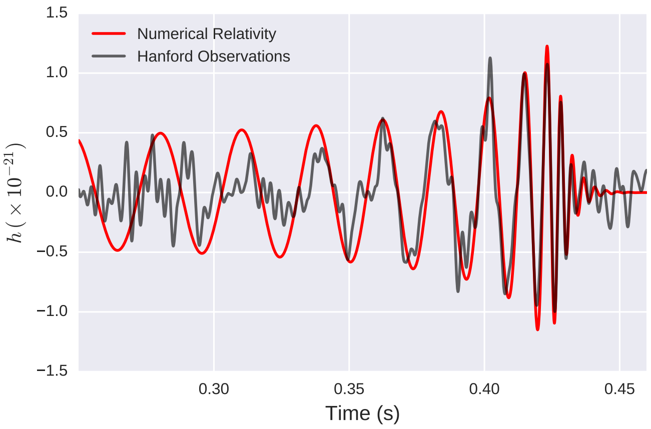


Figure 1.2: The characteristic strain h against time relative to September 14, 2015 at 09:50:45 UTC, which is the time of detection of GW150914 in the Hanford detector. The dark line shows the observed strain at Hanford, and the red curve shows a numerical relativity waveform consistent with the data. The red curve shows the full bandwidth of the gravitational wave (i.e., not just the frequencies the detector is sensitive to). Both waveforms are plotted from data available at `https://losc.ligo.org/events/GW150914/`

on Earth, then they must have coalesced within the lifetime of the universe, which means that they must have had a very small orbital separation. However, it is rarely possible for stars to spend their whole lifetime this close together, since during the lifetime of a massive star, it will expand to many times the radii necessary for a prompt merger, and thus merge before a double compact object has formed. It is therefore necessary to consider mechanisms which can bring compact objects close together after one or both have formed.

There have been several suggested mechanisms, including chemically homogeneous evolution [Mandel and de Mink, 2016] and dynamical formation [Rodriguez et al., 2016], together with classical isolated binary evolution with a common envelope phase. In this thesis we focus only on the isolated binary evolution channel.

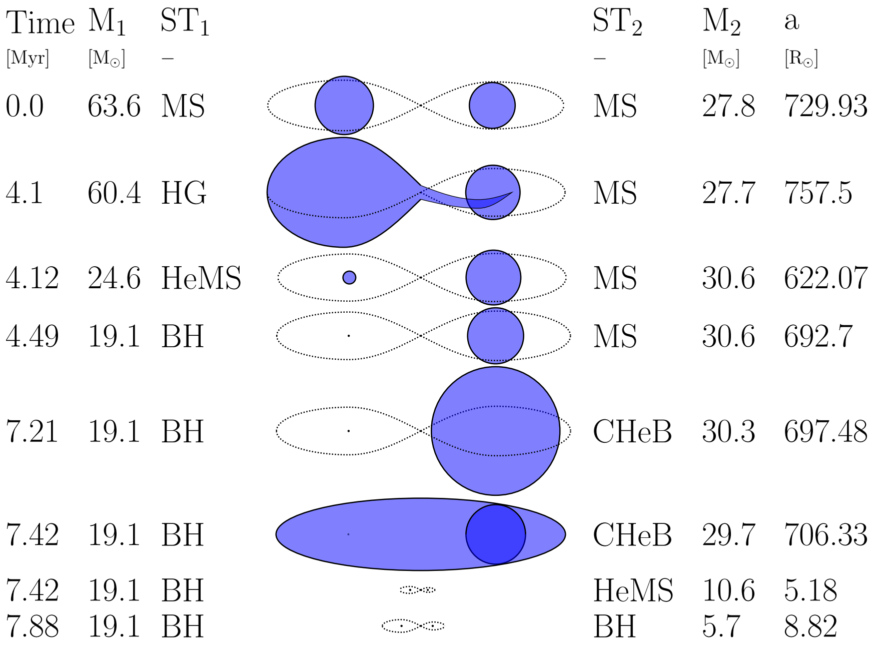
1.1.2 Isolated Binary Evolution

Isolated binary evolution is characterised by the fact that two stars are born together, and then evolve without interacting with anything but one another. Many different observable phenomena can be attributed to the interactions that take place in isolated binaries, and in particular it is one of the leading theories for the production of coalescing double compact objects (i.e., gravitational-wave sources) [De Marco and Izzard, 2017]. Indeed, the isolated binary evolution channel is capable of producing all of the gravitational wave events seen thus far by aLIGO and AdV [Stevenson et al., 2017a].

There are several different modes of interaction in isolated binary evolution, including mass transfer, common envelopes, supernovae and stellar winds. The former two involve the two stars interacting directly with one another, whilst the latter two affect the orbit as well as the amount of mass present, subsequently affecting later episodes of mass transfer and common envelopes Benacquista [2013].

The formation of binary black holes via isolated binary evolution can happen a number of different ways, although most commonly the following series of events occurs; each star starts life as a massive star, and evolves through its main sequence, slowly expanding. When the initially more massive star reaches the end of its main sequence it expands more rapidly, and its outer layers become more gravitationally attracted to the companion star, initiating an episode of mass transfer from the more massive to the less massive star. This mass transfer continues until the more massive star has shed its entire envelope, leaving behind a helium main sequence star. The less massive star may not be able to accept all of this transferred mass, depending on how similar the two stars are, so the total mass of the system may decrease.

The initially more massive star will continue to fuse helium into heavier elements, until it eventually runs out of fuel and collapses under its own gravity to form a black hole. Meanwhile the initially less massive star completes its main sequence evolution and expands, its hydrogen envelope entering the gravitational influence of the now black hole. This time, the mass transfer happens very quickly, and the outer layers of the star envelope both the black hole and helium core of the initially less massive star (a common envelope event). The Helium core and black hole continue to orbit one another within this dense envelope, experiencing significant frictional forces,



Time	M_1	ST_1	ST_2	M_2	a
[Myr]	[$M_\odot$]	–	–	[$M_\odot$]	[$R_\odot$]
0.0	63.6	MS	MS	27.8	729.93
4.1	60.4	HG	MS	27.7	757.5
4.12	24.6	HeMS	MS	30.6	622.07
4.49	19.1	BH	MS	30.6	692.7
7.21	19.1	BH	CHeB	30.3	697.48
7.42	19.1	BH	CHeB	29.7	706.33
7.42	19.1	BH	HeMS	10.6	5.18
7.88	19.1	BH	BH	5.7	8.82

Figure 1.3: A typical evolutionary pathway for a GW151226-like event formed through the classical isolated binary evolution channel. The quantities shown are time, the component masses $M_{1,2}$, the stellar types $ST_{1,2}$ and the semi major axis a. This figure is reproduced from [Stevenson et al., 2017a].

shrinking the orbit dramatically and imparting enough energy into the enveloping gas to allow it to become unbound from the system.

Finally, the now stripped helium core of the initially less massive star runs out of fuel and also collapses to a black hole, leaving a binary black hole system. The significant shrinking of the orbit during the common envelope phase will have brought the two objects close enough together that their emission of gravitational-waves and resultant orbital decay will happen sufficiently rapidly that the two black holes will merge, producing a burst of gravitational-waves potentially detectable using ground based detectors on earth [Stevenson et al., 2017a].

Figure 1.3 shows this sequence of events diagrammatically for a system whose final properties are consistent with GW151226, the second gravitational wave event to be detected. The details of the evolution of the system were computed using the code `COMPAS` (Compact Object Mergers: Population Astrophysics and Statistics), which we describe in the next section.

1.1.3 Rapid Population Synthesis and COMPAS

There are a few different approaches to studying binary evolution using simulations. Many studies use detailed hydrodynamical simulations [Mohamed et al., 2013] or complex physical models [Stancliffe and Eldridge, 2009] to produce a detailed study of a few systems. These approaches might take hours or days of computational time per system, and so it is unfeasible to study large populations in this way. Large populations are necessary, since the uncertainty in population synthesis predictions typically scales with the number of systems simulated. The volume of the input parameter space is also high, meaning a large number of simulations is required to effectively explore it.

Rapid population synthesis takes fits to detailed stellar models together with simplified physical prescriptions, designed to give a good approximation to the physics at the minimum possible computational expense, so that large numbers of systems can be simulated. The properties of populations of binaries can then be studied. `COMPAS` is one such code. On average, `COMPAS` is able to simulate the evolution of a binary from birth to double compact object coalescense in ~ 0.3 seconds, so that it takes ~ 40 hours to simulate a population of 500000 binaries on a typical processor [1].

There are two different classes of parameters assigned to a binary before their evolution is simulated using `COMPAS`. There are 'initial parameters', which are intrinsic to each binary, which include things like initial masses and separations. There are then 'population parameters', which are shared between all binaries in a population. These include the discrete choices of model assumptions used in `COMPAS` (e.g., pessimistic vs optimistic model of Dominik et al. [2012]) as well as continuous variables which parametrise uncertainties in certain model prescriptions. The continuous population parameters and the effect they have on populations of merging binary black holes is the focus of large parts of this thesis.

[1] These times were measured on a laptop with a 2.2GHz Intel Core i5-5200U processor

Initial Parameter Distributions

Being able to simulate large numbers of binaries means the plausible space of initial parameters can be thoroughly explored for a given set of population parameters. These initial parameters are drawn from empirical distributions derived from astrophysical observations. The primary mass (the mass of the initially more massive star) follows a broken power law distribution, known as the initial mass function. We use the unnormalised initial mass function of Kroupa [2001]

$$m_1 \propto \begin{cases} m_1^{-0.3} & \text{for } (0.01M_\odot < m_1 < 0.08M_\odot) \\ m_1^{-1.3} & \text{for } (0.08M_\odot < m_1 < 0.5M_\odot) \\ m_1^{-2.3} & \text{for } (0.5M_\odot < m_1) \end{cases} . \tag{1.1}$$

Whilst for all of the work presented here the primary mass is greater than $0.5\,M_\odot$, the full range of the distribution is important for normalisation purposes, to account for all star forming mass in the universe. Instead of drawing from this distribution again for the secondary star's mass m_2, we instead draw a mass ratio for the system q, which is drawn from a simple uniform distribution, as motivated by observations [Sana et al., 2012b]

$$m_2 = qm_1 \text{ where } q \propto \mathcal{U}[0,1). \tag{1.2}$$

Finally, the initial separation of the two stars a is drawn from a flat-in-the-log distribution [Öpik, 1924, Abt, 1983]

$$a \propto a^{-1}. \tag{1.3}$$

We assume that all binaries are initially circular throughout this thesis, although their eccentricity may change during their evolution.

In Stevenson et al. [2017a] we introduced the COMPAS fiducial model. The fiducial model is the collection of population parameters which represent the most accurate representation of the physics, within the framework of rapid population synthesis. The COMPAS fiducial model is able to reproduce all of the gravitational-wave events observed so far.

Supernova

As stars evolve, they fuse heavier and heavier elements until they reach Iron, which does not release energy under fusion. There is therefore a buildup of Iron in the core, which eventually becomes sufficiently heavy that it can not support itself against gravitational pressure, then the core collapses further, releasing an extremely large amount of energy into the outer layers of the star. Depending on its mass and metallicity, the core may collapse to either a neutron star or a black hole, and the rest of the material is either blown away, or some fraction of it may fall back again under gravity. This process is known as a core collapse supernova [Weiler and Sramek, 1988].

In COMPAS, the stellar evolution tracks evolve the stars until they have a core consisting of carbon and oxygen, which are then directly related to the remnant masses according to the prescriptions of Fryer et al. [2012].

The ejection of the outer material does not happen spherically symmetrically; some more material may get ejected in one direction than another [Janka and Mueller, 1994, Woosley, 1987]. This leads to a ‘natal kick’ of the compact object left behind after the supernova, which can be tens to hundreds of kms^{-1}. The strength and direction of these kicks are drawn randomly from distributions. The distribution from which the strength of the kick is drawn is of particular interest in this thesis. Observations suggest that supernova kick strength is well matched to a 3-Dimensional Maxwellian distribution [Hobbs et al., 2005a], which has a distribution parameter σ_{kick}. Figure 1.4 shows the shape of this distribution when the σ_{kick}

parameter is 250kms^{-1}, chosen to match the observations of Hobbs et al. [2005a], which is the fiducial value of this parameter within `COMPAS`. The distribution function for the Maxwellian distribution is

$$P(v_{\rm kick}) = \sqrt{\frac{2}{\pi}} \frac{v_{\rm kick}^2}{\sigma_{\rm kick}^3} \exp\left(\frac{-v_{\rm kick}^2}{2\sigma_{\rm kick}^2}\right). \tag{1.4}$$

The 3-D Maxwellian distribution represents the distribution of the magnitude of a velocity 3-vector, each component of which is drawn from a Normal distribution with zero mean and variance $\sigma^2_{mathrmkick}$. Kicks are important to compact binary evolution, since they can cause the binary to become unbound. Smaller values of $\sigma^2_{mathrmkick}$ mean that supernova kicks are lower on average, and so binary systems are less likely to be disrupted. Large values of $\sigma^2_{mathrmkick}$ would mean that systems are more likely to be disrupted.

If there is fallback on to the compact object after the supernova, the supernova kick can be diminished, especially for very massive stars [Fryer et al., 2012]. This means that supernova have a lesser effect on more massive systems, so that their evolution will be less sensitive to the value of $\sigma_{\rm kick}$. The direction of the kick given to a star when it experiences a supernova is drawn isotropically.

The kick strength and its direction are both random quantities in `COMPAS`, meaning that the same set of initial parameters doesn't necessarily lead to the same outcome. This means that `COMPAS`, and more generally rapid population synthesis models are not a deterministic mapping from initial to final conditions.

Common Envelope

When the two stars get close to one another, mass can be transferred from one star to the other. Sometimes this process is unstable, meaning that it is a runaway process where the mass transfer causes more mass transfer. This can lead to the donor and its companion becoming engulfed in a 'common envelope'. The core of the donor star and the companion star continue to orbit one another within this

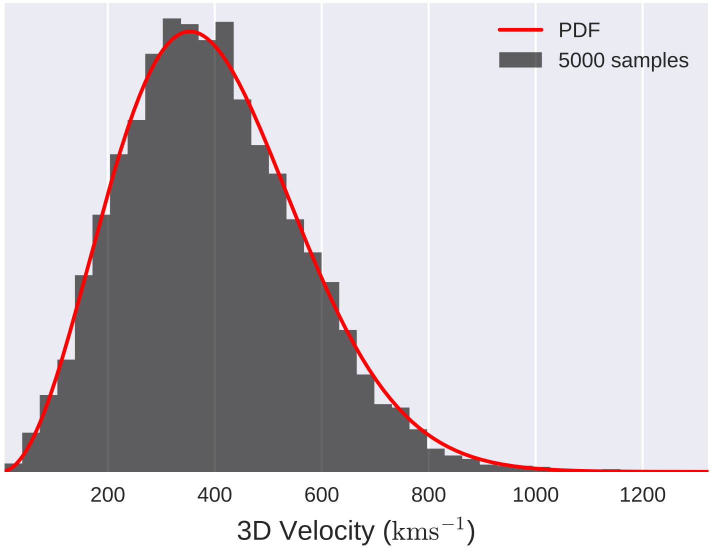


Figure 1.4: A 3 dimensional Maxwellian distribution following the probability density function in 1.4 with $\sigma_{\mathrm{kick}} = 250\,\mathrm{km s}^{-1}$, plotted over 5000 samples from that distribution in 40 bins.

common envelope, and are subject to strong frictional forces. The binary expends its orbital energy to these frictional forces, until either the binary merges within the envelope, or enough energy has been transferred into the envelope for the envelope to escape the binary system entirely, leaving behind the stripped core of the donor and its companion [Paczynski, 1976, Ivanova et al., 2013b].

The common envelope phase is especially important to the production of gravitational wave sources, since it is one of the most efficient mechanisms for tightening a binary. Tightening is important to ensure that coalescence happens promptly, so that the resultant gravitational-waves might be detected on Earth. In `COMPAS`, the energetics of common envelopes are parametrised using the classical energy formalism, which is a two parameter model [Webbink, 1984]. The first parameter, α_{CE}, represents the efficiency with which energy is transferred from the binary orbit into the envelope

$$\alpha_{\mathrm{CE}} = \frac{E_{\mathrm{bind}}}{\Delta E_{\mathrm{orbit}}}, \tag{1.5}$$

where ΔE_{orbit} is the change in the orbital energy of the binary, and E_{bind} is the change in the binding energy of the envelope, from when the common envelope is first formed until it is either ejected or the system merges. ΔE_{orbit} can equivalently be thought of as the energy deposited into the envelope [Hurley et al., 2002]. The binding energy is parametrised by looking at the gravitational force between the donor's envelope and its core

$$E_{\text{bind}} = -\frac{GMM_{\text{env}}}{\lambda_{\text{CE}}R}, \tag{1.6}$$

where M is the mass of the donor star before the common envelope event, M_{env} is the mass of its envelope and R is the radius of the donor star. λ_{CE} is the second parameter in the classical energy formalism, representing the structure of the donor star. In some studies, λ_{CE} is further parametrised to depend on properties of the system, however throughout this thesis its value is kept constant, so that the classical becomes a single parameter model depending on the product $\alpha_{\text{CE}}\lambda_{\text{CE}}$, although by convention we continue to talk in terms of adjustments to only α_{CE}. A constant λ_{CE} is unlikely to accurately represent reality, but we choose it to be constant for simplicity.

Stellar Winds

All stars lose some mass through stellar winds. In certain evolutionary phases these winds can be more or less important, or more or less well constrained by astrophysical observations. The stellar winds during two phases in particular are important to the work in this thesis: (1) the Wolf-Rayet phase, where a star has completed its main sequence evolution has become stripped of its Hydrogen envelope (2) the luminous blue variable phase, a short period of very rapid mass loss in massive stars.

The rate at which stars lose mass through winds during these phases is uncertain, so in `COMPAS`these rates are parametrised by a multiplicative factor which modifies their overall mass loss rate. For the luminous blue variable phase [Humphreys and

Davidson, 1994], this parameter is denoted $f_{\rm LBV}$, so that the mass loss rate becomes

$$\dot{M}_{\rm LBV} = f_{\rm LBV} \times 10^{-4}\ M_\odot\, {\rm yr}^{-1}, \tag{1.7}$$

as described in Belczynski et al. [2010]. During the Wolf-Rayet phase, there is a stronger dependence on the physical characteristics of the star in question, in particular its luminosity L and its metallicity Z [Belczynski et al., 2010, Hamann and Koesterke, 1998]. The mass loss rate is parametrised by the multiplicative factor $f_{\rm WR}$, and has the form

$$\dot{M}_{\rm WR} = f_{\rm WR} \left(\frac{L}{L_\odot}\right)^{1.5} \left(\frac{Z}{Z_\odot}\right)^{m} \times 10^{-13}\ M_\odot\ {\rm yr}^{-1}, \tag{1.8}$$

where $L_\odot$ and $Z_\odot$ are the luminosity and metallicity of our sun, respectively, and $m \approx 0.86$ is an empirically determined power law index [Vink and de Koter, 2005].

Mass loss due to winds can have a profound effect on the ultimate fate of the binaries, both in the obvious sense that the final masses are affected, and also the mass loss can widen the binary.

Summary of `COMPAS`

In summary, `COMPAS` takes a large set of binaries spanning the initial parameter space and evolves each of them using simplified physical prescriptions, in part governed by a set a population parameters, which are shared between all binaries. At the end of a `COMPAS` simulation, each set of initial parameters has either failed to become a system of interest, or has a set of 'system characteristics', which are either astrophysical observables (e.g., the masses or spins of the components), or details of the systems evolution which help to contextualise them within the population (e.g., the time from formation to coalescense). Due to the randomness inherent in supernova kicks, `COMPAS`is a non-deterministic model.

Parameter	**Symbol (Units)**	**Type**	**Distribution/Expression**
Primary Mass	$m_1(M_\odot)$	Initial Parameter	$m_1 \sim m_1^{-2.3} \ (m_1 > 0.5)$
Secondary Mass	$m_2(M_\odot)$	Initial Parameter	$m_2 \sim \mathcal{U}\,[0, m_1)$
Initial Separation	$a(AU)$	Initial Parameter	$a \sim a^{-1}$
Metallicity	Z (n/a)	Population Parameter	
Kick Velocity	v_{kick} (kms^{-1})	System Characteristic	$v_{\text{kick}} \sim \sqrt{\frac{2}{\pi}}\,\frac{v_{\text{kick}}^2}{\sigma_{\text{kick}}^3} \exp\left(\frac{-v_{\text{kick}}^2}{2\sigma_{\text{kick}}^2}\right)$
Kick Velocity Dispersion	σ_{kick} (kms^{-1})	Population Parameter	
Common Envelope Efficiency	α_{CE} (n/a)	Population Parameter	$\alpha_{\text{CE}} = \frac{\Delta E_{\text{bind}}}{\Delta E_{\text{orbit}}}$
Wolf–Rayet Multiplier	f_{WR} (n/a)	Population Parameter	
LBV Multiplier	f_{LBV} (n/a)	Population Parameter	
Final Primary Mass	M_1 ($M_\odot$)	System Characteristic	
Final Secondary Mass	M_2 ($M_\odot$)	System Characteristic	
Final Mass Ratio	q (n/a)	System Characteristic	$q = M_1/M_2$
Chirp Mass	$\mathcal{M}$ ($M_\odot$)	System Characteristic	$\mathcal{M} = \frac{(M_1 M_2)^{\frac{3}{5}}}{(M_1+M_2)^{\frac{1}{5}}}$
Delay Time	τ_{delay} (Myr)	System Characteristic	$\tau_{\text{delay}} = t_{\text{merge}} - t_{\text{form}}$

Table 1.1: A summary of the parameters in `COMPAS` central to the work described in this thesis.

A system characteristic of particular importance to binary black hole mergers is chirp mass $\mathcal{M}$, which is a particular combination of the component masses of a double compact object system M_1 and M_2 which is very well measured using observations of gravitational-waves. This is because it is a leading order coefficient in the frequency evolution of the chirp signal (as shown in figure 1.2). The chirp mass is given by

$$\mathcal{M} = \frac{(M_1 M_2)^{\frac{3}{5}}}{(M_1 + M_2)^{\frac{1}{5}}}. \tag{1.9}$$

In table 1.1, a list of the initial, population and system characteristics used throughout this thesis is given with their relevant statistical distribtuion or definition. This is by no means an exhaustive list for rapid population synthesis, but contains each of the parameters commonly considered to be the most important for the production of merging binary black holes.

1.2 Bayesian Methods

There are two main paradigms in statistics; frequentist statistics and Bayesian statistics. Frequentist statistics relies on repeated experiments to build a picture of a probabilistic distribution. The frequentist approach to statistics dominated the statistical literature and much of the statistical methodology in physics for the majority of the 20th century. Indeed, many fields, such as experimental particle physics, continue to principally use frequentist methods. In section 1.3, we will discuss some techniques which would fall under this category.

However the alternative view, Bayesian statistics, has experienced a boom in popularity in recent years, especially in astronomy and astrophysics. [Jaynes and Bretthorst, 2003]. In this section we will present some of the key results from Bayesian statistics and probability theory, and link them to the topic of scientific inference. We will also discuss the more practical aspects of scientific inference, in particular how to efficiently explore an unknown probability distribution and how to compare different models.

1.2.1 Probability Theory

A random variable is a quantity whose value is the outcome of a random process. The most that can be known about a random variable is values that it could possibly take, and how likely it is to take each of these values. This is known as its probability distribution.

From the Bayesian perspective, if A is the set of all possible outcomes, the probability that a random variable takes some value $a \in A$ is defined as a number $P(a) \in (0, 1)$ which represents our degree of belief that the outcome is a. A probability of 0 indicates that the outcome will never occur, a probability of 1 means that the outcome is certain to occur, and all values in between are ordered such that if $0 \leq P(a) < P(b) \leq 1$ then we believe that a is a more likely outcome than b.

Random outcomes can relate to other random outcomes. In order to talk about the probability of multiple random variables, we need to introduce the concepts of joint and conditional probabilities. If we have two random variables, each with its own set of possible outcomes A and B, the joint probability of two outcomes $P(a, b)$ is the probability that both outcomes $a \in A$ and $b \in B$ occur. A conditional probability $P(a|b)$, is the probability that outcome a occurs given that it is known that b occurs.

Outcome a is said to be independent of outcome b if its probability does not change when conditional on b, i.e., $P(a|b) = P(a)$. In the special case that both outcomes are mutually independent, then their joint probability is simply $P(a, b) = P(a)P(b)$. [Jaynes and Bretthorst, 2003]

Two key results in probability theory are the sum and product rules. Here we state the discrete results, however both are trivially generalisable to the continuous case. The sum rule states that the combined probability of all possible outcomes for a random variable must equal unity.

$$\sum_{a \in A} P(a) = 1 \tag{1.10}$$

The product rule states that the probability of two outcomes a and b both happening (i.e., their joint probability), is the combination of the probability that one outcome occured, $P(a)$, with the probability that the other occurred, given that the first one occurred $P(b|a)$. So

$$P(a, b) = P(b|a)P(a). \tag{1.11}$$

For a continuous random variable $a \in A$, there are an uncountable number of possible outcomes in A. In this case, we define probability over an interval of possible values by integrating the probability density function over that range. The probability density function is the function which represents the density of probability. This implies that the probability of an outcome taking a value in an interval over a region of high probability density is higher than it taking a value from an interval of the same size in a region of low probability density.

1.2.2 Bayes' Theorem

Bayes' theorem is a natural consequence of the symmetry of the product rule, $P(a,b) = P(a)P(b|a) = P(b)P(a|b)$, so that

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}. \tag{1.12}$$

In the context of data analysis, Bayes' theorem is invaluable for determining the probability distributions for model parameters conditioned on some observations. Say we have some observational data $\mathcal{D}$ and a parametrised model $\mathcal{M}$ which depends on some parameter vector $\vec{\theta}$. We wish to determine which values of the parameters best represent the data for that model;

$$P(\vec{\theta}|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|\vec{\theta},\mathcal{M})\ P(\vec{\theta}|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})}. \tag{1.13}$$

$P(\vec{\theta}|\mathcal{D},\mathcal{M})$ is known as the posterior distribution, and represents the joint probability distribution of the values of the parameters $\vec{\theta}$, conditional on the data. $P(\mathcal{D}|\vec{\theta},\mathcal{M})$ is the likelihood distribution, representing the probability of seeing the observed data, given the model and its parameters. $P(\vec{\theta}|\mathcal{M})$ is known as the prior distribution, and is used to represent any prior knowledge about the values of the parameters. Throughout this thesis, we shorten the names of these distributions to simply the 'posterior', 'likelihood' and 'prior' respectively. $P(\mathcal{D}|\mathcal{M})$ is called the evidence (or alternatively the margnalised likelihood or normalising constant) which combines all possible values that the parameters could take to normalise the posterior distribution, making it a probability

$$P(\mathcal{D}|\mathcal{M}) = \int d\vec{\theta} P(\mathcal{D}|\vec{\theta},\mathcal{M})\ P(\vec{\theta}|\mathcal{M}). \tag{1.14}$$

For a fixed model and data, the evidence term does not change, and so often it is sufficient to only consider the numerator of Bayes' theorem. However, when comparing models, the evidence becomes important. We will discuss this in more detail in section 1.2.4.

1.2.3 Stochastic Sampling

It is rare that the posterior probability can be computed analytically. In many cases, it is necessary to numerically explore the posterior probability surface. When the dimensionality of the posterior is low, the posterior values can be efficiently computed on a grid. However for problems of more than a couple of dimensions, this becomes highly inefficient. There are a number of more sophisticated methods for the efficient exploration of higher dimensional parameter spaces. These are known as stochastic sampling methods, and use fewer evaluations of the posterior probability to gain a good approximation of its distribution. The most commonly used stochastic sampling methods fall into two main families; Markov Chain Monte Carlo (MCMC) methods and nested sampling methods.[MacKay, 2003, Skilling et al., 2006].

Markov Chain Monte Carlo (MCMC)

A Markov chain is an ordered series of states of some process, where each state can be described entirely as a transformation of the previous state. In MCMC sampling, a Markov chain of evaluations of the posterior probability distribution, or something proportional to it, is constructed in such a way that the distribution of states in the Markov chain matches the posterior probability distribution. The simplest algorithm which achieves this is the Metropolis-Hastings algorithm [Metropolis et al., 1953].

In the Metropolis-Hastings algorithm, samples are generated by choosing a new point from a proposal distribution, which is typically centred around the previous sample in the chain. The posterior function is evaluated there and its value compared to its value at the previous point in the chain. If the posterior is greater, that proposal is accepted, whereas if it is smaller, then the proposal is probabilistically accepted or rejected, so that the proposal is accepted with probability $P_{\text{accept}} = P_{\text{new}}/P_{\text{old}}$.

Ensemble sampling is an extension to the Metropolis-Hastings algorithm, where instead of a single Markov chain there is an ensemble of such chains whose proposal distributions are dictated by the position of the rest of the ensemble. The most widely used type of ensemble sampling is Goodman and Weare's affine invariant ensemble sampler [Goodman and Weare, 2010, Foreman-Mackey et al., 2013].

In this algorithm, each chain's proposal distribution has support on the line intersecting its current state in parameter space X_i and the current state of another, randomly chosen chain X_{other}. The new proposed state X_{i+1} is then somewhere on this line, so that

$$X_{i+1} = X_i + S(X_{\text{other}} - X_i), \tag{1.15}$$

where S is a random variable dictating how far along the connecting line the new point is placed. The authors of Goodman and Weare [2010] recommend a sampling distribution $g(S)$ for S

$$S \propto \begin{cases} \frac{1}{\sqrt{S}} & \text{if } \; S \in \left[\frac{1}{2}, 2\right] \\ 0 & \text{otherwise} \end{cases} . \tag{1.16}$$

This new point is then either accepted or rejected in the same way as a standard Metropolis-Hastings.

The benefit of using an ensemble sampler over a standard Metropolis-Hastings sampler is that the ensemble can easily adapt itself to an appropriate scale, so that if the prior range is particularly broad when the true posterior is relatively narrow, then the ensemble will tend to spread out until it finds a peak. Then, when one chain finds a peak, since it is unlikely to leave, the other chains will, on average, tend to be attracted towards this peak. Their scale will adjust to the scale of the peak, so they can effectively explore it.

For example, when one chain finds an area of high posterior density, then it is unlikely to accept any proposals to jump towards the other chains in the ensemble. The other chains, however, will eventually choose their proposed jump towards the chain with high posterior density, and are likely to accept proposals in this direction. The chains will therefore tend to cluster around regions of high posterior density.

Ensemble sampling is a simple but effective method of exploring many posterior distributions. However, it has two major disadvantages. First, it is unable to cope well with multimodal distributions, since when the ensemble settles on one mode, it becomes highly unlikely for it to explore the prior volume to find other modes. One can make further modifications to the method to deal with this shortcoming, however these come with their own associated disadvantages Vousden et al. [2016].

Second, by itself ensemble sampling can only explore the unnormalised shape of the posterior distribution, since there is no way to compute the evidence integral of equation 1.14 using just this algorithm. Again, modifications can be made to allow estimates of this integral (e.g., parallel tempering with thermodynamic integration [Calderhead and Girolami, 2009]).

Nested Sampling

Nested sampling is an entirely different approach to sampling when compared to MCMC methods [Skilling et al., 2006], which deals directly with the denominator of equation 1.13; the evidence. The evidence is defined as an integral over all possible values of the parameters in the likelihood, as defined in equation 1.14

If we denote the prior volume element $dX = P(\vec{\theta}|\mathcal{M})d\vec{\theta}$, then we can write the integral in a form that we can approximate by quadrature

$$\begin{aligned} P(\mathcal{D}|\mathcal{M}) &= \int dX\ P(\mathcal{D}|\vec{\theta},\mathcal{M}) && (1.17) \\ &\approx \sum \Delta X\ P(\mathcal{D}|\vec{\theta},\mathcal{M}). && (1.18) \end{aligned}$$

The challenge is in computing an ordered set of prior volume elements in order to effectively approximate the integral. The nested sampling procedure achieves this by finding a series of isolikelihood contours (surfaces of equal likelihood in the prior volume) of decreasing volume, centred upon the region of high likelihood. In many practical cases, under the assumption that the likelihood surface is unimodal, the change in volume between contours can be reasonably approximated as decreasing exponentially, although it is also possible to approximate the change in volume by sampling.

In practise, these contours are constructed by drawing a number of 'live' points from the prior, computing and sorting their likelihood, then replacing the lowest likelihood point with a new point with the constraint that it must have a higher likelihood than the discarded point. Once this process has been repeated until a suitable precision in the value of the evidence is achieved, the discarded points can be assigned appropriate weights to turn them into samples from the posterior.

The most difficult aspect of the nested sampling algorithm is generating new points to satisfy the increasing likelihood condition. Nested sampling typically achieves this by drawing samples from an approximated likelihood contour, usually a D dimensional ellipse.

The benefit of nested sampling is that computes both the evidence and posterior samples as a bi-product, however, as with the basic MCMC algorithms, it doesn't handle multimodal distributions particularly well.

Improvements to the way nested sampling handles the issue of multi-modality have been explored in the literature. [Feroz et al., 2009, Feroz and Hobson, 2008, Feroz et al., 2013] Instead of drawing a single approximation to the likelihood contour, if clusters in the existing 'live' points exist, then multiple D dimensional ellipses are drawn to approximate the likelihood contour. Identifying these clusters is a highly non-trivial problem, and is dealt with in detail in the relevant papers.

1.2.4 Model Comparison

The key benefit of computing the evidence using methods such as nested sampling is that it allows for two models to be compared. We wish to compute the probability that a given model $\mathcal{M}$ is correct given some data $\mathcal{D}$ (i.e., $P(\mathcal{M}|\mathcal{D})$). We can once again use Bayes' theorem

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}. \tag{1.19}$$

We can recognise $P(\mathcal{D}|\mathcal{M})$ as the evidence for a parametrised model from equation 1.14. The $P(\mathcal{M})$ is simply our prior belief that the model is correct. $P(\mathcal{D})$ is the difficult term, since we can't know the absolute probability of obtaining some data without conditioning on a model. However, we can compare two models by computing the ratio of their probabilities so that this term cancels. This is known as the posterior odds ratio $\mathcal{R}$. If we have two models $\mathcal{M}_1$ and $\mathcal{M}_2$ then

$$\mathcal{R} = \frac{P(\mathcal{M}_1|\mathcal{D})}{P(\mathcal{M}_2|\mathcal{D})} = \frac{P(\mathcal{D}|\mathcal{M}_1)P(\mathcal{M}_1)}{P(\mathcal{D}|\mathcal{M}_2)P(\mathcal{M}_2)}. \tag{1.20}$$

It is frequently true that if we are comparing two models, we have no prior preference for one model or the other, so that $P(\mathcal{M}_1) = P(\mathcal{M}_2)$ and the posterior odds is simply a ratio of the evidences, also known as the Bayes' factor [Gelman et al., 2013].

$\mathcal{R}$ can be interpreted as the odds that $\mathcal{M}_1$ is correct compared to $\mathcal{M}_2$, so that an ratio $\mathcal{R} = 10$ would indicate that $\mathcal{M}_1$ is 10 times more likely to represent the data as $\mathcal{M}_2$.

1.3 Machine Learning

Machine learning, which is also commonly known as statistical learning, is the field of study concerned with building models and algorithms which have the freedom to learn from data. Within this broad definition, machine learning can be roughly divided into two sub disciplines: supervised learning and unsupervised learning.

Supervised learning, which includes all of the methods discussed in this thesis, concerns algorithms which make use of labeled data to model the behaviour of an unknown function. Labelled data is the a set of inputs to the unknown function $\{x_i\}$ for which we know the respective outputs from the function $\{y_i\}$. Modelling the function is then an exercise of predicting the unobserved outputs to the function $\{y_i^*\}$ for a set of inputs $\{x_i^*\}$. In the case that the output is made up of a finite set of 'classes', this is known as a classification problem, and when the output is continuous this is known as regression. Unsupervised learning is the study of algorithms which deal with unlabelled data. We do not make use of unsupervised learning methods in this thesis.

There are many different methods in supervised learning. In the preliminary work leading up to the work presented in this thesis, we explored a number of different methods, including nearest neighbour regression [Bentley, 1975] and neural network regression [McCulloch and Pitts, 1943, Rumelhart et al., 1988].

1.3.1 Gaussian Process Regression

Gaussian process regression is a flexible regression technique, which is particularly effective when (i) the size of the training set is modest, (ii) when the response surface is smooth, meaning that there are no discontinuous jumps in the response surface, (iii) stationary, meaning that the length scales over which changes in the inputs affect the outputs doesn't change with respect to the inputs, (iv) when it is important to quantify the uncertainty in the model predictions [Rasmussen, 2004, Ambikasaran et al., 2015].

The underlying assumption for Gaussian process regression is that the set of functions which can explain the observed data are jointly distributed with Gaussian statistics, meaning that we can write mean and covariance functions, $\mu(x)$ and $K(x, x')$ respectively, for the underlying 'true' function f(x)

$$\begin{aligned} \mu(x) &= \langle f(x) \rangle & (1.21) \\ K(x, x') &= \langle (f(x) - \mu(x))(f(x') - \mu(x') \rangle \, . & (1.22) \end{aligned}$$

It is often easier to understand this property with the derivative assumption that any subset of outputs in the training set are jointly Gaussian distributed, meaning that there exists a covariance matrix describing how they relate to one another. If we then learn how the elements of this covariance matrix depend on the model inputs, then we can use this to write down a conditional distribution for the unseen outputs, given the training set.

This is achieved by parametrising a covariance function, which is also known in this context as a kernel function. We train this covariance function $K(x_i, x_j)$, which gives the value of the covariance between y_i and y_j. K is only constrained to be symmetric with respect to x and positive semidefinite, however in practise there are a handful of functions which are commonly used. Probably the most common is the squared exponential function

$$K(x_i, x_j) = \exp\left(-\frac{1}{2}(x_j - x_i)^T M (x_j - x_i) \right) , \tag{1.23}$$

where M is a metric for the input parameter space. The elements of this matrix are free parameters, how changes in the input x influence the observed output y. It is possible to use any positive definite function of the input vectors as a kernel function. Another common choice is the Matern family of covariance functions, whose sample paths are less smooth (less differentiable) than the squared exponential kernel

$$K_{3/2}(x_i, x_j) = \left(1 + \sqrt{3(x_i - x_j)^T M(x_i - x_j)}\right) \exp\left(-\sqrt{3(x_i - x_j)^T M(x_i - x_j)}\right). \tag{1.24}$$

The training process involves finding the elements of M such that the covariance function $K(x_i, x_j)$ optimally represents the training data. This can be achieved by maximising the multivariate Gaussian likelihood $\mathcal{L}$ for the problem (which is, by definition, the probability of observing the observed data, given the parameters in the covariance function)

$$\log(\mathcal{L}) = N_{tr} \log(2\pi) - \frac{1}{2} \log |\mathbf{K}| + \mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y}, \tag{1.25}$$

where N_{tr} is the number of training examples, and $\mathbf{K}$ is the matrix whose elements are determined by the function K. The shape of the likelihood can be inferred using stochastic sampling methods (e.g MCMC), or it can be maximised using any of a number of point estimate optimisers (e.g LBFGS). Once the parameters in the covariance function have been optimised, the goal is to compute a conditional distribution for the values of the response surface at unseen points, given the observed 'training' input/output pairs. This is the distribution $P(y^*|y, x, x^*)$. Under the Gaussian assumption, we can write this joint distribution for all of the y^* and y (which, via the covariance function, are only functions of x^* and x). It can then be shown, for centred observations, that the conditional distribution is itself a multivariate Gaussian;

$$P(y^*|y, x, x^*) \sim N(\mu, \Sigma) \tag{1.26}$$

$$\mu = K(y, y^*)^T K(y, y)^{-1} y \tag{1.27}$$

$$\Sigma = K(y^*, y^*) - K(y, y^*)^T K(y, y)^{-1} K(y, y^*). \tag{1.28}$$

See appendix A.1 for a detailed derivation. These are known as the the Gaussian process update equations.

One of the principal benefits of using Gaussian process regression compared to most other regression methods is that the predictions come with a distribution of possible outputs, rather than just point estimates (as with, for example, neural networks). Moreover, if the training observations have Gaussian uncertainty, these can also be trivially taken into account with the Gaussian process model, by inflating the diagonal elements of $K(y, y)$ to account for this extra variance.

1.3.2 Decision Trees

Decision trees are an intuitively simple machine learning method which can be applied to both classification and regression problems, based on recursively making binary partitions of the input parameter space, parallel to one of the input directions [Hastie et al., 2013, Breiman et al., 1984]. Figure 1.5 provides a sketch of how the partitioning of the input space might look for a regression problem in one dimension.

Partitions are made to the input parameter space, both for classification and regression, with the object of minimising the predictive error across both sides of the partition. First, for a regression problem, if there are a set of N input variables, x_{ij} with their corresponding responses continuous y_i, where the i index corresponds to each element in the training set, and the j index spans the dimensionality D of

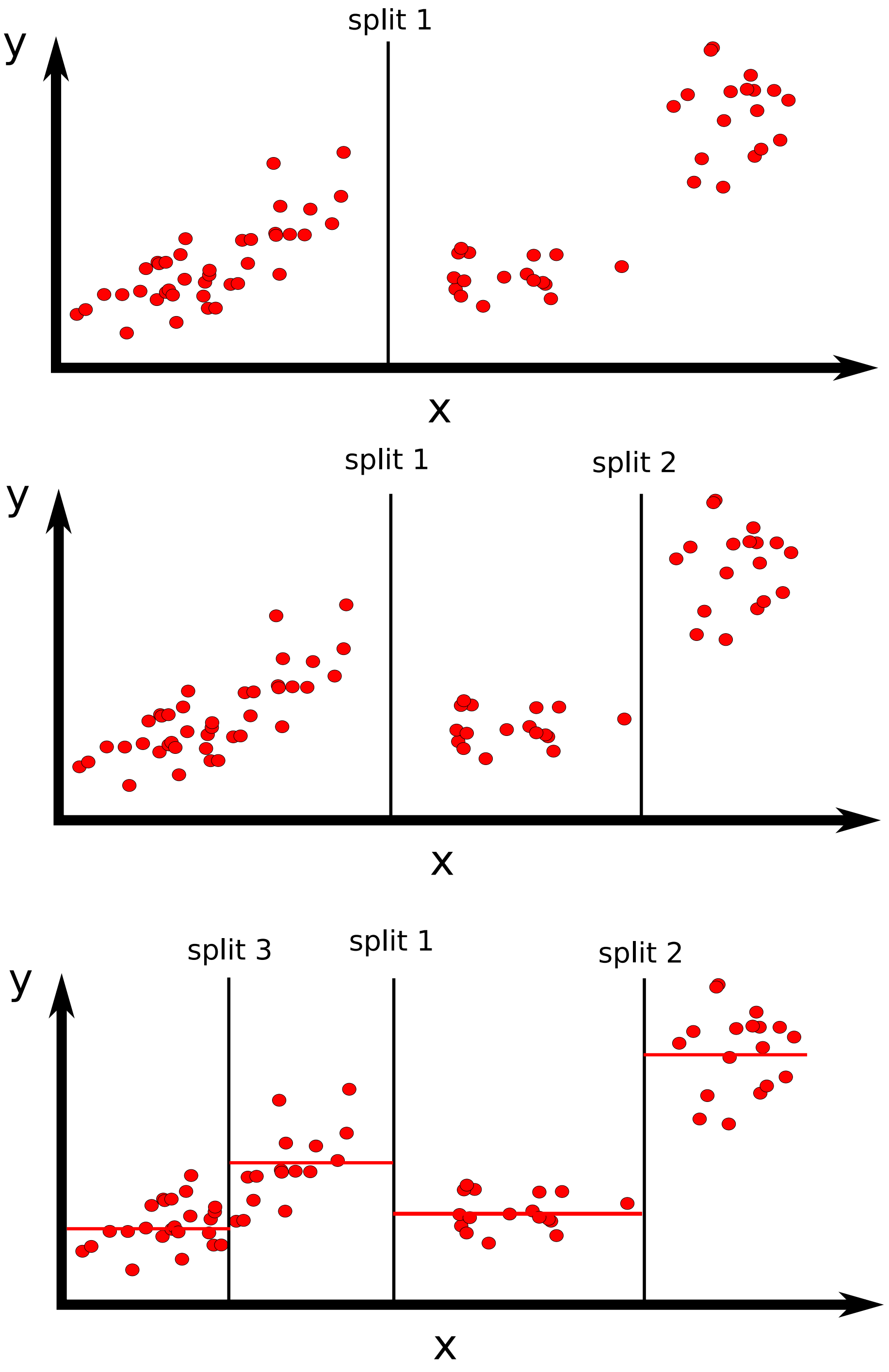


Figure 1.5: A cartoon example of the construction of a regression in one dimension, where x represents the input parameter, and y the response. Each subsequent split is made to make a binary partition of the dataset. Once all of the splits have been made, the predictions are averages of the responses in that region of parameter space, here represented by horizontal red lines.

the input. The predictive error is defined in the usual least squares sense

$$\varepsilon^2 = \sum_i \left(y_i - f(x_{ij})\right)^2, \tag{1.29}$$

where $f(x_{ij})$ is the decision tree's prediction. The optimum split in the input parameter space is chosen such that this predictive error is minimised with respect to both the input dimension the split is made with respect to, the index j, and the location of the split in that dimension, S. The problem then becomes

$$j, S = \underset{j,S}{\operatorname{argmin}} \left[\min_y \sum_{i:x_{ij}<S} \left(y_i - f(x_{ij})\right)^2 + \min_y \sum_{i:x_{ij}\geq S} \left(y_i - f(x_{ij})\right)^2 \right]. \tag{1.30}$$

Whilst this looks like a difficult optimisation problem, since the amount of training data is finite, there are only $D(N-1)$ non-degenerate splits that can be made. These can usually be explored exhaustively. This process is then repeated to partition each side of the split further, continuing recursively until some stopping condition is reached. Commonly, the stopping condition is that each partition or 'leaf' contains some minimum number of training examples.

For classification problems, the method proceeds in much the same way, except that the predictive error is defined differently. If we have K different classes into which an observation can be classified, and we define a probability p_k, which is the fraction of observations which are correctly classified into class k, then the splitting is carried out by optimising the Gini index G, which is defined as

$$G = \sum_{k\in K} p_k(1-p_k). \tag{1.31}$$

There are alternatives to using the Gini index, however this is the method we used in this thesis.

1.3.3 Bootstrapping

Bootstrapping is a technique for both reducing the variance in a statistical estimator based on a dataset, and for providing an estimate on the 'realisation uncertainty' of some feature derived from a dataset. In both cases, bootstrapping involves resampling the dataset; making a new dataset of the same size by uniformly randomly choosing datapoints from the original dataset with replacement [Hastie et al., 2013].

If the task is to derive a better estimator for the variance of a model prediction using some dataset, then it can be shown that the average of the results of the same unbiased estimator applied to each resampled dataset supplies an estimator for the same quantity, but with an estimate of the variance modified improved by a factor of $\sim 1/m$ where m is the number of new resampled datasets.

When bootstrapping for the purpose of uncertainty quantification for some derived feature of the dataset, that feature is simply computed for each of the resampled datasets. The distribution of values of this feature can then be used as a representation of the uncertainty of that feature, coming from the finite size of the dataset.

1.3.4 Random Forests

Random forests are an ensemble machine learning algorithm which can be used for either regression or classification, utilising both the decision tree algorithm and bootstrapping. The random forest algorithm involves growing a number of decision trees. Each decision tree is built with a different resampled (bootstrap) dataset, and each time a new split of the input space is considered, it is made using a random subset of the input dimensions [Hastie et al., 2013, Breiman, 2001].

Decision trees by themselves tend to overfit data, especially when grown to their maximal extent, however by using an ensemble of trees this effect is averaged, leading to a powerful machine learning algorithm for either regression or classification, depending on the type of constituent decision tree.

1.3.5 Parameter Optimisation

There are several approaches to optimising the parameters in a model, depending on what is known about the target distribution and the acceptable computational expense. In almost all cases, a Bayesian approach is optimal, writing down a likelihood function which represents how well a given set of parameters in a model allows that model to match the data, together with an appropriate prior, and then using one of the stochastic sampling techniques described in section 1.2.3 to explore the parameter space infer the shape of the posterior.

However, whilst the Bayesian approach is usually optimal in the sense that it provides the multivariate distribution of plausible parameter values, it is often not pragmatically sensible to use this approach, especially in cases where the posterior function is expensive to compute, or when a reasonable point estimate of the maximum of the posterior is good enough. In these cases, it is often sufficient to turn to maximum likelihood methods.

We use two such methods in this thesis. The first is the Normal equation for the solution to an overconstrained system of simultaneous linear equations. Overconstrained means that we have more observations of a linear system than we do parameters. If we have a system of linear equations of the form

$$y = \mathbf{X}\beta \tag{1.32}$$

Where y is a vector of responses, $\mathbf{X}$ is the design matrix of the inputs corresponding to the responses in y and β is the vector of parameters, in this case coefficients to the inputs in a linear model. Where the number of responses is greater than the number of parameters in the model, the optimal parameter vector $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y. \tag{1.33}$$

A derivation of this expression is given in appendix A.2. In the case where the responses are independently and identically distributed about the truth, this expression gives the maximum likelihood estimation for $\hat{\beta}$.

The second non-Bayesian parameter optimisation method we use is a variable metric method, the LBFGS algorithm, for the optimisation of a function [Nocedal and Wright, 1999, Press, 2007]. These methods rely on having gradient information of the function to be optimised, and can be powerful alternatives to a fully Bayesian treatment when that is prohibitively expensive, since they take relatively few function evaluations to find an extrema. However, they must be used with care, since they only give point estimates for the extrema, rather than a full posterior distribution, and so in cases where the function is not convex, these algorithms may settle on local extrema.

1.4 Stochastic Variability

Stochastic variability is when systems change their state unpredictably. Many physical systems demonstrate stochastic variability, including the pulsation of stars [Farr et al., 2018] and x-ray emission from x-ray binaries [Barnard et al., 2015]. It is therefore important to have methods to characterise and quantify the properties of stochastic processes. In this section we introduce some of the key concepts of stochastic variability, in particular correlations and stochastic periodicity, and describe continuous autoregressive moving average (CARMA) models. We also describe methods of visualising the characteristics of stochastic processes.

1.4.1 Correlation

Correlation in the context of stochastic variability is the relation of a system's current state on its previous states. The state of a system is the values of its observable characteristics at a fixed time. If a process were completely uncorrelated, then every measurement of it would be an independent draw from some distribution. This

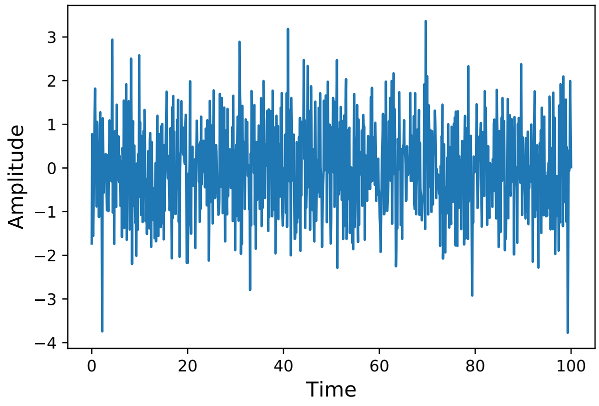


Figure 1.6: An example of uncorrelated $\sim \mathcal{N}(0,1)$ White Noise

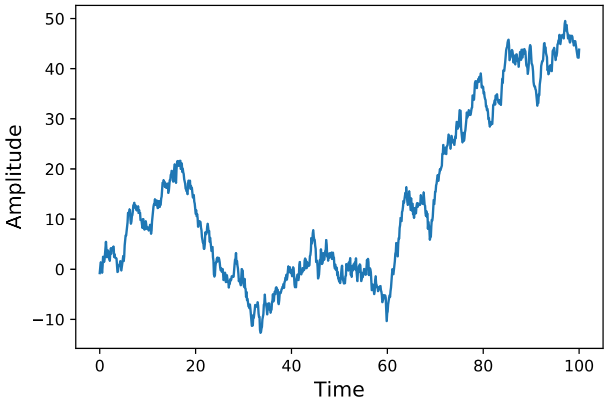


Figure 1.7: An example of a Brownian noise process, with each sample drawn from a Normal distribution with variance 1 about the previous sample.

kind of process is often referred to as a 'white noise' process. Physically interesting systems are seldom well described by white noise. We show an example of normally distributed white noise in figure 1.6.

One of the simplest examples of correlated noise is Brownian noise. Mathematically, this can be simply expressed as an integral of a Gaussian white noise process $dW(t) \sim N(0, \sigma^2)$

$$y_{Br}(t) = \int_0^t dW(t). \tag{1.34}$$

Equivalently, one could generate discrete samples of the same correlated process by moving the centre of the distribution with each sample

$$y_{Br}(t_i) \sim \mathcal{N}(y_{Br}(t_{i-1}), \sigma^2). \tag{1.35}$$

When discretised in this way, Brownian processes are Markov processes, so that each of their states is described by a probabilistic transformation of the previous state. Brownian processes can be a useful approximation to some physical processes, such as motion of particles in a gas, however it has the disadvantage that the properties of the noise depend on how often it is observed. We show an example of a Brownian process in figure 1.7.

Probably the simplest extension to the Brownian noise process which is useful for describing physical systems is a first order continuous auto-regressive (CAR) process. CAR processes are intuitively Brownian noise with friction, which means that the mean of the underlying sampling distribution decays, as a function of time, back to the mean of the overall series. The discrete representation of a (zero mean) CAR process (i.e an AR process) is

$$y_{AR}(t_i) \sim N\left(y_{AR}(t_{i-1}) \cdot \exp\left(-\frac{t_i - t_{i-1}}{\tau}\right), \sigma^2\right). \tag{1.36}$$

AR(1) noise is dependent on a single timescale τ, which dictates how quickly the instantaneous centre of the noise decays back towards the true centre of the series. By considering the case of $t_i - t_{i-1} = \tau$, it is clear that τ is in fact the e-folding timescale for the magnitude of the deviation of the process from the true mean. A centred autoregressive process can be expressed as a linear combination of a noise contribution and a term proportional to the previous observation

$$y_{AR}(t_i) = \exp\left(-\frac{\Delta t}{\tau}\right) y_{AR}(t_{i-1}) + \beta\varepsilon, \tag{1.37}$$

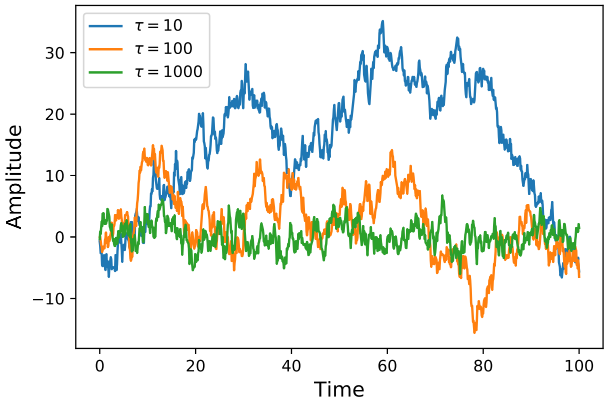


Figure 1.8: Examples of three different CAR(1) processes with different decay times τ, and variance $\sigma^2 = 1$.

where $\varepsilon \sim N(0,1)$ and $\beta = \sigma\sqrt{1 - \exp\left(-\frac{\Delta t}{\tau}\right)}$ is a constant that modulates the variance of the noise contributions according to the same decay timescale τ.

AR(1) models can be sufficient to explain many examples of stochastic variability, however in some cases it is necessary to extend to a model which accounts for multiple timescales, or periodicities, in the correlations. These higher order autoregressive models satisfy the following differential equation

$$\left[\prod_{j=0}^{p-1}\left(\frac{d}{dt} - r_j\right)\right](y(t) - \mu) = \eta(t). \tag{1.38}$$

Where p is the order of the autoregressive model, the r_j represent the (inverse) decay timescales and periodicities at play in the model, μ is the mean of the process and $\eta(t)$ is a normally distributed white noise process. A further generalisation to the model is to also include a moving average (MA) contribution to the process. A moving average model is a simple model representing a system whose mean is perturbed by random shocks of varying strengths. When coupled to the autoregressive models, these are known as continuous autoregressive moving average (CARMA) models, which satisfy

$$\left[\prod_{j=0}^{p-1}\left(\frac{d}{dt}-r_j\right)\right](y(t)-\mu)=\left[\prod_{i=1}^{q}\left(\frac{d}{dt}-b_i\right)\right]\eta(t). \tag{1.39}$$

where the q is the order of the moving average side of the model and the b_i represent the magnitude and times of the moving average perturbations to the mean of the process. In chapter 4 we present a thorough exploration of how CARMA models can be used to explore physical phenomena.

1.4.2 Power Spectral Densities (PSDs)

It is often fruitful to examine a time series in the frequency domain. The translation between the time and frequency domain is most commonly achieved by performing a Fourier transform. [Arfken, 2013]

$$\tilde{h}(f)=\int_{-\infty}^{\infty} dt\; h(t)e^{2\pi i f t} \tag{1.40}$$

In experimental practise, one almost never deals with a continuous function, instead measurements are made as discrete samples of the process. The reciprocal of the time between subsequent discrete measurements is called the sampling frequency. The discrete time equivalent of the Fourier transform mapping N time domain samples to N points in the frequency domain can be written as

$$\tilde{h}=\sum_{k=0}^{N-1} h_k e^{-\frac{2\pi i k}{N}}. \tag{1.41}$$

A closely related concept to the frequency representation of a function is its power spectral density (PSD). This name is misleading, since the ‘power’ referred to doesn’t necessarily have to mean physical power. The power spectral density $P(f)$ is the absolute square of the amplitude of the process in the frequency domain [Press, 2007]. This is equivalent to the amount of power contained in a frequency interval $[f, f+df]$, so that the total power of a process is the sum of these contributions over all frequencies

$$P_{total} = \int_{-\infty}^{\infty} df \, \left|\tilde{h}(f)\right|^2 = \int_{-\infty}^{\infty} dt \, \left|h(t)\right|^2 , \tag{1.42}$$

where the final equality comes from invoking Parceval's theorem.

Another fundamental property of PSDs is that the PSD of a process is very closely related to the process' autocorrelation function, $\varrho(\tau)$ so that they form a Fourier transform pair. This relationship is called the Wiener-Khinchin theorem [Chatfield, 2013]

$$\varrho(\tau) = \int_{-\infty}^{\infty} df \; e^{2\pi i f \tau} P(f) \tag{1.43}$$

$$P(f) = \int_{-\infty}^{\infty} d\tau \; e^{-2\pi i f \tau} \varrho(\tau) \tag{1.44}$$

.

The autocorrelation itself is very closely related to the autocovariance, so that the two terms are often used synonymously in the literature. In fact they are simply proportional to one another, up to a factor of σ^2, the variance of the process.

Power spectral densities are a very useful visual tool for describing the properties of stochastic variability.

The simplest example of stochastic variability, white noise, is, by defintion, completely uncorrelated. Its correlation function must therefore be zero everywhere except at lag zero (since everything correlates perfectly with itself at zero lag), and so $\varrho(\tau) = \delta(\tau)$. Considering the Wiener-Khinchin theorem of equation 1.44, this means that the PSD of white noise is constant across all frequencies.

Extending the complexity to Brownian noise, which is defined in equation 1.34 as an integral of a white noise process, meaning that $W(t) = \frac{dy_{br}(t)}{dt}$. One can calculate the power spectral density straightforwardly, by Fourier transforming both sides of the above into the frequency domain and taking absolute squares

$$\begin{aligned} \left|\tilde{W}(f)\right|^2 &= \left|\mathcal{F}\left[W(t)\right]\right|^2 = \left|\mathcal{F}\left[\frac{dy_{br}(t)}{dt}\right]\right|^2 \\ &= (2\pi f)^2 \left|y_{br}(f)\right|^2 . \end{aligned} \tag{1.45}$$

Since it has already been deduced that the power spectral density of white noise $\left|\tilde{W}(f)\right|^2 = const$, it is clear that the power spectral density of brownian noise falls off like f^{-2}. The fall-off of the power spectral density with frequency is a sufficient indicator for correlation. The power spectral density of a CARMA process, derived in appendix A.3.1, is given by

$$P(f) = \sigma^2 \frac{\left|\prod_{i=1}^{q} (2\pi i f - b_i)\right|^2}{\left|\prod_{j=0}^{p-1} (2\pi i f - r_j)\right|^2}. \tag{1.46}$$

This expression shows that the PSD of a CARMA process has a frequency dependence of order $P(f) \sim f^{2(q-p)}$. For physical systems $p > q$, since otherwise the PSD would grow with frequency, allowing the total power to diverge. The CARMA PSD also therefore requires a slope of at least $\sim f^{-2}$ at high frequencies. The effect of the real r_j can then be intuitively seen to be an 'offset' for the influence of each f^{-2} falloff. If one considers a CAR(1) process described in equation 1.36 with $\tau = \frac{1}{r}$, then the PSD will take the form.

$$P(f) = \frac{\sigma^2}{4\pi^2 f^2 + r^2}. \tag{1.47}$$

At frequencies much lower than the auto-regressive root r, the PSD is dominated by the r^{-2} term, and is approximately constant. At higher frequencies the f^{-2}

term dominates and the PSD falls off. The transition between these two regimes is continuous and occurs at $r \approx f$, and forms what we will refer to as a 'knee' feature of a PSD. For higher order CARMA processes, with more than one real root, the 'knee' features combine independently, with each knee increasing the falloff rate of the PSD by order 2 at the corresponding frequency.

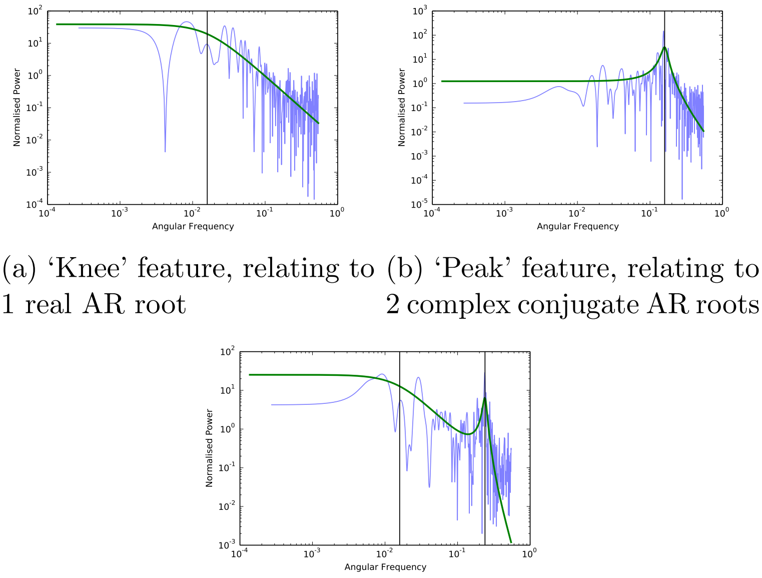
When the autoregressive roots are complex, they must come in complex conjugate pairs. The denominator of a PSD for a CAR(2) model with a pair of complex roots $r = a \pm ib$ contains a real 4^{th} order polynomial in f with all terms having non-zero coefficients. As with the real roots, this translates to a constant falloff for frequencies much lower than $f = a$ and a fall-off with frequency of f^{-4} for high frequencies. However, in the intermediate regime, the frequencies relate to a peak in the PSD, centred at $\frac{b}{2\pi}$ [Kelly et al., 2014]. The width of the peak is closely related to the quality factor of the oscillations in the autocorrelation, with a quality factor $Q = (a/b)$. We will refer to these as 'peak' features of the PSD.

In summary, a peak feature in the PSD, which corresponds to a periodicity to the stochastic variability, indicates a pair of complex conjugate auto-regressive roots. For a decay timescale, a single real auto-regressive root is needed, corresponding to a knee feature in the PSD. Figure 1.9 shows examples of PSDs exhibiting each of these features.

The moving average roots have a more subtle effect on the shape of the PSD. It is difficult to give a physically intuitive analogy to their effect. They ease the restriction that the PSD must fall off like f^{-2p} for purely auto-regressive models. This allows for more flexibility in the shape of the PSD, especially where there is a high auto-regressive order p.

1.4.3 Methods of Estimating PSDs

The most naive approach to computing PSDs is a direct computation, by Fourier transforming the measured process and taking its modulus squared at each frequency $P(f) \sim |\tilde{y}(f)|^2$. However, this is heavily dependent on the width of a the frequency



(a) ‘Knee’ feature, relating to 1 real AR root

(b) ‘Peak’ feature, relating to 2 complex conjugate AR roots

(c) ‘Peak’ and ‘Knee’ features, relating to 1 real and 2 complex AR roots.

Figure 1.9: PSDs of 3 different noise realisations with different features. The faint blue gives the Lomb-Scargle periodogram, the green gives the ‘true’ PSD and the black lines show the frequencies of the true correlation timescales.

bins used in the discrete Fourier transform. This dependence on the bin width manifests in ‘leakage’ between bins, where the convolution of the shape of the frequency bin can corrupt the estimate of the PSD at that frequency. [Press, 2007]

A partial solution to this problem is to select a different functional shape of the frequency bin, via a window function, so that leakage falls off as quickly as possible at the edges of the frequency bin. There are several practical choices of window function and all make valid estimates of the PSD. Howevever, this method still relies on the computational tractability of taking a Fourier transform of the data. In some situations this is not straightforward.

One of the biggest drawbacks to Fourier transform methods is that they are unable to deal with unevenly sampled data. In some cases, data can be approximated to being evenly sampled, which loses accuracy in the PSD estimate. In others, it is possible to use interpolation to make the data evenly sampled, which can either be computationally very expensive or, as before, costs accuracy in the PSD estimate.

Lomb-Scargle Periodogram

The Lomb-Scargle periodogram [Lomb, 1976] [Scargle, 1982] is a widely used method for approximating PSDs from time domain processes. It is derived by inferring the signal at each frequency as a weighted sum of sinusoids, such that the periodogram reduces to the Fourier result in the limit of evenly sampled data [VanderPlas, 2017]. The resultant estimator for the PSD can be computed analytically using observations and their mean and variance, and the cost of this computation scales modestly with the number of data points and with the number of sampled frequencies.

It can be shown that with the appropriate normalisation factor, this estimation of the PSD follows an exponential probability distribution. The probability distribution of the values of power being known mean that the statistical significance of features in a PSD can be computed. [Press, 2007]

According to the above distribution, it is clear that the frequency of points occuring above a value P falls off like e^{-P}, so that we expect a single frequency sample to appear below P with probability $1 - e^{-P}$. If N independent frequencies are sampled, then the probability of the distribution rising above P for at least one frequency is

$$Pr(P(f) > P) = 1 - (1 - e^{-P})^N. \tag{1.48}$$

This expression acts as a significance test against the null hypothesis for the power rising above a certain P. The Lomb-Scargle periodogram is thus a powerful tool for searching for periodicities in an unevenly sampled time series. The focus of chapter 4 is to introduce a new method for inferring the shape of PSDs.

Chapter 2

Fisher Matrices

2.1 Introduction

We currently have gravitational-wave observations of 5 merging black holes so far, and we may have to wait a number of years before we have an extensive catalogue of events. Until that time, we can still evaluate how much we are likely to be able to learn from observations once we make them. In this chapter we use `COMPAS` to compute the Fisher information matrix to quantify how much we expect to learn from observations in the context of rapid population synthesis.

This chapter reproduces the text of Barrett et al. [2017a], accepted for publication Monthly Notices of the Royal Astronomical Society. This work was led by myself, in collaboration with Sebastian Gaebel, Coenraad Neijssel, Alejandro Vigna-Gómez, Simon Stevenson, Christopher Berry, Will Farr and Ilya Mandel. The code dealing with selection effects was written by Sebastian Gaebel, with modifications by myself. The code dealing with cosmic history was written by Coenraad Neijssel, with modifications by myself. All other analysis code was written by myself. The text of the paper was written by myself, with modifications from all authors.

2.2 Abstract

The properties of the population of merging binary black holes encode some of the uncertain physics underlying the evolution of massive stars in binaries. The binary black hole merger rate and chirp-mass distribution are being measured by ground-based gravitational-wave detectors. We consider isolated binary evolution, and explore how accurately the physical model can be constrained with such observations by applying the Fisher information matrix to the merging black hole population simulated with the rapid binary-population synthesis code COMPAS. We investigate variations in four COMPAS parameters: common-envelope efficiency, kick-velocity dispersion, and mass-loss rates during the luminous blue variable and Wolf–Rayet stellar-evolutionary phases. We find that ~ 1000 observations would constrain these model parameters to a fractional accuracy of a few per cent. Given the empirically determined binary black hole merger rate, we can expect gravitational-wave observations alone to place strong constraints on the physics of stellar and binary evolution within a few years. Our approach can be extended to use other observational data sets; combining observations at different evolutionary stages will lead to a better understanding of stellar and binary physics.

2.3 Introduction

Gravitational waves from binary black hole coalescences [Abbott et al., 2016a, 2017c,d,e] have recently been observed by the ground-based gravitational-wave detectors of the Advanced Laser Interferometer Gravitational-Wave Observatory [aLIGO; Aasi et al., 2015] and Advanced Virgo [AdV; Acernese et al., 2015]. These observations provide a revolutionary insight into the properties of the population of binary black holes. The catalogue of detections will grow rapidly as the instruments continue to improve their sensitivity [Abbott et al., 2017f]. In this paper, we analyse how such a catalogue will make it possible to infer the physics of binary evolution by performing inference on parametrised population synthesis models.

A number of channels (sequences of physical phenomena) for the formation of binary black holes have been proposed [see, e.g., Abbott et al., 2016c, Miller, 2016, Mandel and Farmer, 2017, for reviews]. In this study, we assume that all merging binary black holes form through classical isolated binary evolution via a common-envelope phase [Postnov and Yungelson, 2014, Belczynski et al., 2016]. While all events observed to date are consistent with having formed through this channel [Stevenson et al., 2017a, Eldridge et al., 2017, Giacobbo et al., 2017], a future analysis would need to hierarchically include the possibility of contributions from multiple channels [e.g., Stevenson et al., 2017b, Zevin et al., 2017, Talbot and Thrane, 2017].

Previous efforts to explore how stellar and binary population synthesis models could be constrained with gravitational-wave observations [e.g., Bulik et al., 2004, Bulik and Belczynski, 2003, Mandel and O'Shaughnessy, 2010, Gerosa et al., 2014, Stevenson et al., 2015] have typically focused on a discrete set of models, usually obtained by varying one evolutionary parameter at a time [e.g., Voss and Tauris, 2003, Dominik et al., 2012, Mennekens and Vanbeveren, 2014]. In this paper, we consider the realistic scenario in which the astrophysical model is described by a multi-dimensional set of continuous parameters which may be strongly correlated. We ask how well we could constrain these parameters with a large observational data set.

The main tool we use to tackle this problem is the Fisher (information) matrix. Fundamentally, if we make an observation of a process, and we have a model for that process that depends on some parameters, then the Fisher matrix quantifies how much we can learn about the parameters in our model from the observation we made. It also captures how the information content of different parameters relate to one another. We derive an expression for the Fisher matrix for binary-population synthesis. We use this to quantify how much we can learn about the population parameters from observations of binary black holes using the current generation of ground-based gravitational-wave detectors. While we concentrate on gravitational-wave observations here, the method is applicable to other data sets, and the best constraints may come from combining multiple complementary observations.

We use Fisher matrices to demonstrate that it may be possible to precisely measure the population parameters in binary-population synthesis models with ~ 1000 observations of binary black hole mergers. At the expected rate of gravitational-wave detections [Abbott et al., 2017d], this could be within a few years of the detectors reaching design sensitivity ($\sim$ 2–3 yr at design sensitivity for our fiducial model); the observing schedule for gravitational-wave observatories is given in Abbott et al. [2017f].

We first give an introduction to our binary population synthesis model in section 2.4, together with a description of the model parameters we wish to infer using gravitational-wave observations. In section 2.5, we demonstrate how we transform the raw outputs of our binary population synthesis model by considering observational selection effects and redshift- and metallicity-dependent star formation rates. In section 2.6 we introduce the statistical tools used in this paper: (i) the likelihood function representing the probability of an observation given our model, (ii) a method for including measurement uncertainties in observations, and (iii) the Fisher matrix, which quantifies the sensitivity of our model to changes in its underlying parameters. The results of applying this methodology to binary population synthesis models are presented and discussed in section 2.7, and we discuss our conclusions in section 2.8.

2.4 Population synthesis of massive stellar binaries

Many of the details of binary evolution are currently uncertain [Postnov and Yungelson, 2014, De Marco and Izzard, 2017]. Population synthesis models efficiently, albeit approximately, simulate the interactions of a large number of binaries in order to capture population wide behaviour and thoroughly explore the space of initial conditions for binary evolution (i.e., initial masses and separations). Uncertainties in the physics underlying isolated binary evolution are captured within population synthesis models through tunable parameters, which we call population parameters. In this paper we focus on four population parameters which have an impact on binary black hole formation. We use the rapid population synthesis code COMPAS.[1] This uses the stellar evolutionary models of Hurley et al. [2000]. Final black hole masses are calculated using the delayed model of [Fryer et al., 2012]. With the exception of the variations to the four population parameters we describe in section 2.4.1, we employ the Stevenson et al. [2017a] fiducial model throughout this paper.

2.4.1 Population parameters

Supernova kick velocity

The asymmetric ejection of matter [Janka and Mueller, 1994, Burrows and Hayes, 1996, Janka, 2013] or emission of neutrinos [Woosley, 1987, Bisnovatyi-Kogan, 1993, Socrates et al., 2005] during a supernova can provide a kick to the stellar remnant. This birth kick is on the order of hundreds of $\mathrm{km\,s^{-1}}$ for neutron stars [Hobbs et al., 2005b]. The typical strength of supernova kicks imparted to black holes is not well constrained observationally [Wong et al., 2014, Mandel, 2016, Repetto et al., 2017], although they may be reduced relative to neutron star through the accretion of material falling back onto the stellar remnant [Fryer et al., 2012].

In COMPAS, the strength of supernova kicks is parametrised using the dispersion parameter for a 3-dimensional Maxwell–Boltzmann distribution σ_{kick}. A kick velocity v_{kick} is drawn from the distribution

$$P(v_{\mathrm{kick}}) = \sqrt{\frac{2}{\pi}}\,\frac{v_{\mathrm{kick}}^2}{\sigma_{\mathrm{kick}}^3}\exp\left(\frac{-v_{\mathrm{kick}}^2}{2\sigma_{\mathrm{kick}}^2}\right). \tag{2.1}$$

Alternative parametrisations for the supernova kick have been considered by Bray and Eldridge [2016], who did not find sufficient evidence to prefer them; here, we only consider continuous variations to model parameters, including the kick velocity

[1] Further details and sample COMPAS simulations are available at www.sr.bham.ac.uk/compas/.

in the Maxwell–Boltzmann distribution.

The kick is modified to account for mass fallback, so that the final kick imparted to the black hole is

$$v^*_{\text{kick}} = (1 - f_{\text{fb}})v_{\text{kick}}, \tag{2.2}$$

where f_{fb} is the fraction of matter that falls back on to the black hole, calculated according to the delayed model of Fryer et al. [2012]. For carbon–oxygen core masses greater than $11M_\odot$, $f_{\text{fb}} = 1$ and so many heavy black holes receive no natal kick in this model [Belczynski et al., 2016, Stevenson et al., 2017a]. Whilst observations of the proper motions of isolated Galactic pulsars [Hobbs et al., 2005a] suggest a value of $\sigma_{\text{kick}} = 265\ \text{km}\,\text{s}^{-1}$, we choose a fiducial $\sigma_{\text{kick}} = 250\ \text{km}\,\text{s}^{-1}$ to match Stevenson et al. [2017a].

Common-envelope efficiency

When mass transfer is dynamically unstable and initially proceeds on the very short dynamical timescale of the donor, a shared, non co-rotating common envelope is formed around the donor core and the companion [Paczynski, 1976]. The details of the common-envelope phase are amongst the least well understood across all phases of isolated binary evolution [for a review, see Ivanova et al., 2013b].

In COMPAS, the classical energy formalism [Webbink, 1984] is employed to parametrise uncertainty in the physics of the common envelope. When a binary begins a common-envelope phase, each envelope is bound to its core, with a total binding energy approximated by

$$E_{\text{bind}} = -G\left[\frac{M_1(M_1 - M_{\text{core},1})}{\lambda_{\text{CE},1}R_1} + \frac{M_2(M_2 - M_{\text{core},2})}{\lambda_{\text{CE},2}R_2}\right], \tag{2.3}$$

where G is Newton's constant, $M_{\text{core},(1,2)}$ are the core masses of the two stars, $M_{(1,2)}$ and $R_{(1,2)}$ are the stellar masses and radii, respectively, and $\lambda_{\text{CE}(1,2)}$ are the corresponding stellar-structure parameters introduced by de Kool [1990] and are functions of star's evolutionary state [e.g., Dewi and Tauris, 2000, Kruckow et al., 2016].

The loss of co-rotation between the orbit of the cores and the common envelope leads to energy dissipation which causes the cores to spiral in. Some of this lost orbital energy may be eventually used to eject the common envelope. The efficiency with which this transfer of energy occurs is uncertain, and is characterised by the free parameter α_{CE}. In order to determine the separation after the common-envelope phase, the classical energy formalism compares the binding energy of the envelope to the energy transferred from the orbit $\Delta E_{\mathrm{orbit}}$ so that

$$E_{\mathrm{bind}} = \alpha_{\mathrm{CE}} \Delta E_{\mathrm{orbit}} \,. \tag{2.4}$$

If the binary has sufficient orbital energy to completely expel the envelope, we consider this a successful common-envelope event. Unsuccessful ejections lead to a merger before a binary black hole system is formed. We follow Stevenson et al. [2017a] in assuming that common-envelope phases initiated by main sequence of Hertzsprung gap donors always lead to mergers [cf. the pessimistic model of Dominik et al., 2012].

The fiducial choices of the parameters in `COMPAS` are $\lambda_{\mathrm{CE}} = 0.1$ and $\alpha_{\mathrm{CE}} = 1.0$. We explicitly leave λ_{CE} fixed whilst making small perturbations to α_{CE}; however, this is an issue of labelling, since it is the product of these two free parameters which is ultimately of importance to the common-envelope physics [Dominik et al., 2012].

Mass-loss multipliers

Throughout their lives, stars lose mass through stellar winds. The wind mass-loss rate depends strongly on the star's luminosity, since brighter stars will tend to excite more material to leave the star. Mass loss rates are generally highest for high mass, high metallicity stars. The dearth of observations of low metallicity environments means wind mass-loss rates are poorly constrained at low metallicities, and at high masses where stars are intrinsically rare. These are precisely the regimes where the progenitors of gravitational-wave sources are likely to form [Belczynski et al., 2016, Eldridge and Stanway, 2016, Lamberts et al., 2016, Stevenson et al., 2017a, Giacobbo et al., 2017].

`COMPAS` employs the approximate wind mass-loss prescriptions detailed in Belczynski et al. [2010]. For hot O/B-stars, we employ the wind mass-loss prescription of Vink et al. [2001]. Our Wolf–Rayet wind mass-loss rates come from Hamann and Koesterke [1998]. For other phases the mass-loss prescriptions from Hurley et al. [2000] are used. Uncertainty in mass-loss rates can have a significant impact on stellar evolution; for example, Renzo et al. [2017] find that there is a ~ 50 per cent uncertainty in the mapping between initial and final masses when considering different mass-loss prescriptions when modelling solar-metallicity, non-rotating, single stars, with initial masses between $15M_{\odot}$ and $35M_{\odot}$. There are particular phases of stellar evolution where the mass-loss rates lack strong constraints by observations. We parametrise the mass-loss rates in two of these phases with tunable population parameters.

During the luminous blue variable (LBV) phase [Humphreys and Davidson, 1994], extremely massive stars undergo a relatively short episode of rapid mass loss which strongly impact the binary's future evolutionary trajectory [e.g., Mennekens and Vanbeveren, 2014]; observational constraints on the physics of LBV stars are currently uncertain [Smith, 2017].[2] Following Belczynski et al. [2010], we parametrise this rate in terms of a multiplicative factor f_{LBV} used to modify the basic prescription, so that the rate becomes

$$\dot{M}_{\mathrm{LBV}} = f_{\mathrm{LBV}} \times 10^{-4}\ M_{\odot}\,\mathrm{yr}^{-1}; \tag{2.5}$$

our fiducial value for this factor is $f_{\mathrm{LBV}} = 1.5$ [Belczynski et al., 2010].

During the Wolf–Rayet phase, stars have lost their hydrogen envelopes and have high but relatively poorly constrained mass-loss rates [Crowther, 2007]. We use a

[2] As in Hurley et al. [2000], we assume stars are in an LBV-like phase if their luminosity and radius satisfy $L > 6 \times 10^5 L_{\odot}$ and $(R/R_{\odot})(L/L_{\odot})^{1/2} > 10^5$.

multiplicative constant f_{WR} to modify the base rate:

$$\dot{M}_{\mathrm{WR}} = f_{\mathrm{WR}} \left(\frac{L}{L_\odot}\right)^{1.5} \left(\frac{Z}{Z_\odot}\right)^{m} \times 10^{-13}\, M_\odot\, \mathrm{yr}^{-1}, \tag{2.6}$$

where L is the stellar luminosity, Z is the metallicity, $Z_\odot = 0.02$ is approximately the bulk metallicity of our Sun, and $m = 0.86$ is an empirically determined scaling factor [Vink and de Koter, 2005, Belczynski et al., 2010]. The fiducial choice for this population parameter is $f_{\mathrm{WR}} = 1.0$. We use the same mass-loss prescription for all Wolf-Rayet subtypes [Belczynski et al., 2010], as the Hurley et al. [2000] evolutionary tracks do not distinguish between them. Recent investigations of mass loss for Wolf–Rayet stars of varying composition include McClelland and Eldridge [2016], Tramper et al. [2016], Yoon [2017].

2.5 Model predictions

In this paper we evaluate the impact of the tunable parameters described above on the rate of detections and the measured chirp-mass distribution of binary black holes. The chirp mass $\mathcal{M}$ is a particular combination of the component masses M_1, M_2 which is measured well from the gravitational-wave frequency evolution during the binary inspiral [Cutler and Flanagan, 1994, Abbott et al., 2016d],

$$\mathcal{M} = \frac{(M_1 M_2)^{3/5}}{(M_1 + M_2)^{1/5}}. \tag{2.7}$$

The chirp mass is just one of the parameters measurable through gravitational waves, other observables such as component masses, spins and the distance to the source can also be inferred [Abbott et al., 2016d]. For simplicity, we have chosen to focus on chirp mass since it is the best measured. This is a conservative approach, as we have neglected information about other parameters; however, the methods presented here are easily extendible to include other observables.

In order to represent the distribution of chirp masses produced by the population synthesis model, we chose to bin our systems by chirp mass. Throughout this paper, we use 30 bins of equal width, ranging from the lowest to the highest chirp masses present in our dataset. The number of bins is determined by the scale length of variability in the chirp-mass distribution and the chirp-mass measurement uncertainty discussed below; the results are insensitive to halving the number of bins.

The raw output of a population synthesis model is a list of the initial conditions and final outcomes of all the binaries simulated. In order to compare this output to astronomical observations, it is necessary to process the data further, in order to account for the relatively well known history of star formation in the Universe and the observational selection effects. We describe this processing below.

2.5.1 Cosmic history

In order to focus our computation on black hole progenitors, we only simulate systems with primary masses between $7M_\odot < M_1 < 100M_\odot$. We assume that all stars are in binaries with primary masses ranging between 0.01–$150M_\odot$ following the initial mass function of Kroupa [2001] with a flat mass-ratio distribution [Sana et al., 2012b]. At formation, binaries are assumed to have a uniform-in-the-logarithm distribution of orbital separations [Öpik, 1924, Abt, 1983] and zero orbital eccentricity; for more detailed studies of mass-ratio and orbital distributions, see Duchêne and Kraus [2013], Moe and Di Stefano [2017]. COMPAS simulations produce a rate of binary black hole formation per unit star formation mass M_form,

$$\mathcal{R}_\texttt{COMPAS} = \frac{\mathrm{d}^3 N_\mathrm{form}}{\mathrm{d}M_\mathrm{form}\,\mathrm{d}\tau_\mathrm{delay}\,\mathrm{d}\mathcal{M}}, \tag{2.8}$$

where τ_delay is the delay time, defined as the time from the birth of a binary to its coalescence [Peters, 1964]. To compute the total rate of binary black hole mergers per unit comoving volume per unit time we need to convolve the COMPAS formation rate with the amount of metallicity-specific star formation per unit volume per unit time at the birth of the binaries. Delay times can range from a few Myr to Gyr, and observations show that both the metallicity and star formation rates in galaxies evolve significantly over these timescales [Madau and Dickinson, 2014]. We use the star formation rate distribution of Madau and Dickinson [2014] and the metallicity distribution of Langer and Norman [2006]. Other distributions have been suggested [e.g., Savaglio et al., 2005, Ma et al., 2016, Vangioni et al., 2015], and the cosmic history of metallicity evolution adds an additional source of uncertainty to our model predictions, although we do not account for this extra uncertainty in this paper. Future studies could consider how metallicity evolution could be included with the other model parameters and inferred from binary observations. In figure 2.1 we provide an illustration of the metallicity-specific star formation rate at redshifts $z = 0.5$, 1, and 1.5, and also indicate metallicities at which we performed simulations for this study. We use these to translate the star formation rate into the merger rate at redshift z

$$\begin{aligned}\frac{\mathrm{d}^3 N_{\mathrm{merge}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}\mathcal{M}}(z) = \int \mathrm{d}Z \int \mathrm{d}\tau_{\mathrm{delay}} \Bigg[&\frac{\mathrm{d}^3 N_{\mathrm{form}}}{\mathrm{d}M_{\mathrm{form}}\,\mathrm{d}\tau_{\mathrm{delay}}\,\mathrm{d}\mathcal{M}}(Z) \\ &\times \frac{\mathrm{d}^3 M_{\mathrm{form}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}Z}(Z, t_{\mathrm{form}} = t_{\mathrm{merge}}(z) - \tau_{\mathrm{delay}}) \Bigg] ,\end{aligned} \tag{2.9}$$

where t_{s} is the time measured in the frame of reference of the merger, V_{c} is the comoving volume and we use cosmological parameters from Ade et al. [2016]. Figure 2.2 shows the local merger rate at three different redshifts after accounting for changes in star formation rate and cosmology.

2.5.2 Selection effects

Gravitational-wave detectors are not equally sensitive to every source. The distance to the source, its orientation and position relative to the detectors, as well as the physical characteristics of the source all affect how likely it is that the system would be detectable. The detectability of a signal depends upon its signal-to-noise ratio

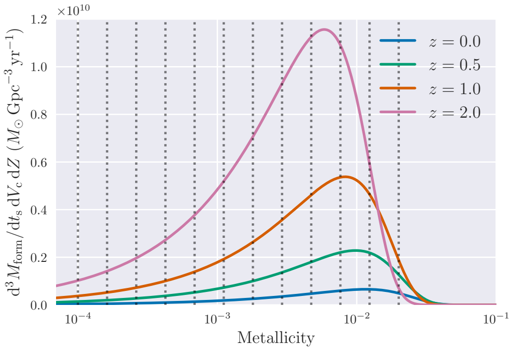


Figure 2.1: The metallicity-specific star formation rate as a function of metallicity at three different redshifts, using the star-formation-rate distribution of Madau and Dickinson [2014] and the metallicity distribution of Langer and Norman [2006]. The vertical dashed lines indicate the metallicities at which we undertook simulations for this study. Metallicities above $Z_\odot = 0.02$ contribute negligibly to the binary black hole merger rate.

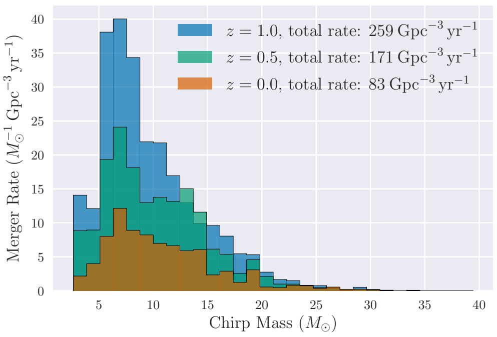


Figure 2.2: The binary black hole merger rate predicted by the `COMPAS` fiducial model at three different redshifts, taking into account the cosmic evolution of the metallicity-specific star formation rate. For comparison, the total inferred merger rate density from gravitational-wave observations is 12–213 Gpc^{-3} yr^{-1} [Abbott et al., 2017d].

(SNR). The SNR in a single detector is defined as [Finn, 1992]

$$\mathrm{SNR}^2 = \langle h|h\rangle = 4\Re\int_{f_{\min}}^{f_{\max}} \mathrm{d}f \; \frac{h^*(f)h(f)}{S(f)}, \tag{2.10}$$

where $h(f)$ is the waveform measured by the detector, $S(f)$ is the one-sided noise power spectral density, and $f_{\min}$ and $f_{\max}$ are the limits of the frequency range considered.

where $\Re$ means the real part. For simplicity, we assume that signals are detected if their single-detector SNR exceeds a threshold value of 8 [Abbott et al., 2017f]. To model the waveforms, we use the IMRPhenomPv2 [Hannam et al., 2014, Husa et al., 2016, Khan et al., 2016] and SEOBNRv3 [Pan et al., 2014, Babak et al., 2017] approximants;[3] these include the inspiral, merger and ringdown phases of a binary black hole coalescence, and allow for precession of the black hole spins. We incorporate the effects of cosmological redshift, which manifest as an apparent increase in the system masses, $M_{\mathrm{obs}} = (1+z)M_{\mathrm{s}}$ [Krolak and Schutz, 1987, Holz and Hughes, 2005]. We assume a detector sensitivity equal to aLIGO in its design configuration [Aasi et al., 2015, Abbott et al., 2017f].

We optimise our computations, reducing the number of waveform calculations required, by exploiting the fact that the parameters extrinsic to the gravitational-wave source, such as its position and orientation, only couple into the overall amplitude of the wave via

$$\begin{aligned}
\mathcal{A} &\propto \frac{1}{D_{\mathrm{L}}}\sqrt{F_+^2(1+\cos^2 i)^2 + 4F_\times^2\cos^2 i}, && (2.11)\\
F_+ &\equiv \frac{1}{2}\cos(2\psi)(1+\cos^2(\theta))\cos(2\phi), \\
&\quad - \sin(2\psi)\cos(\theta)\sin(2\phi), && (2.12)\\
F_\times &\equiv \frac{1}{2}\sin(2\psi)(1+\cos^2(\theta))\cos(2\phi), \\
&\quad + 2\cos(2\psi)\cos(\theta)\sin(2\phi), && (2.13)
\end{aligned}$$

[3] We use the implementations publicly available in the LAL suite software package wiki.ligo.org/DASWG/LALSuite.

where $\mathcal{A}$, D_L, i, ψ, θ and ϕ are the gravitational-wave amplitude, luminosity distance, inclination, polarization, and polar and azimuthal angles of the source location in the detector frame, respectively [Krolak and Schutz, 1987, Cutler and Flanagan, 1994]. Therefore, we need only compute the phase evolution for a given combination of intrinsic binary parameters, such as masses, once, and then marginalize over the extrinsic parameters (with the exception of D_L) as described in Finn and Chernoff [1993].

For a system with a given (M_1, M_2, D_L), we determine the fraction of extrinsic parameter realisations for which the observed SNR passes our threshold, and label this as our detection probability P_det.

We can use this detection probability to transform the merge rate given in Eq. (3.5) into a rate of detections. Integrating over the merger redshift gives the total detection rate

$$\frac{\mathrm{d}N_\mathrm{obs}}{\mathrm{d}t_\mathrm{obs}\,\mathrm{d}\mathcal{M}} = \int \mathrm{d}z \, \left[\frac{\mathrm{d}^3 N_\mathrm{merge}}{\mathrm{d}t_\mathrm{s}\,\mathrm{d}V_\mathrm{c}\,\mathrm{d}\mathcal{M}} \frac{\mathrm{d}V_\mathrm{c}}{\mathrm{d}z} \frac{\mathrm{d}t_\mathrm{s}}{\mathrm{d}t_\mathrm{obs}} P_\mathrm{det}\right], \tag{2.14}$$

where t_s is time in the source frame and $t_\mathrm{obs} = (1+z)t_\mathrm{s}$ is time in the observer's frame.

Figure 2.3 shows the rate and chirp-mass distribution of binary black hole mergers detected at aLIGO design sensitivity. The mass distribution is shifted to higher masses relative to the intrinsic merger rate plotted in figure 2.2 because selection effects favour heavier systems which emit louder gravitational-wave signals. Some of the sharp features in this plot are the consequence of simulating systems on a discrete grid of metallicities [cf. Dominik et al., 2013]. LBV winds tend to reduce high mass stars to a narrow, metallicity-dependent range of black hole masses. We discuss the impact of these features in section 2.8.

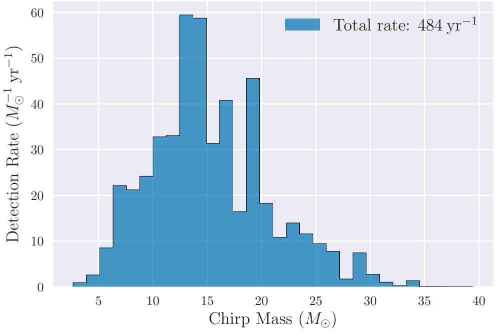


Figure 2.3: The rate and chirp-mass distribution of the binary black hole coalescences we expect aLIGO to observe at design sensitivity, taking into account cosmic history and selection effects, for the COMPAS fiducial model as described in Stevenson et al. [2017a]. The detection rate is per unit observing time.

2.6 The covariance matrix for population parameters

2.6.1 The Fisher information matrix

The Fisher matrix quantifies the amount of information that a set of observable random variables (in our case, the merger rate and chirp-mass distributions) carries about the parameters (in our case, the four tunable parameters described in section 2.4) of a distribution that models these observables. Getting more information from a random variable about a model parameter means that we are able to more accurately infer the value of that parameter.

Specifically, the Fisher matrix F for distribution over a set of observations of random variables $\mathcal{D}$ (the data) which are dependent on a set of parameters $\{\lambda\}$ is defined element-wise as

$$F_{ij} = -\left\langle \frac{\partial^2 \log\left[\mathcal{L}\left(\mathcal{D}|\{\lambda\}\right)\right]}{\partial\lambda_i\,\partial\lambda_j} \right\rangle, \tag{2.15}$$

where $\mathcal{L}$ is the likelihood function, defined as the probability of acquiring the observed data $\mathcal{D}$ given the model parameters, and the angle brackets indicate an expectation over the data realisation for a fixed value of the $\{\lambda\}$. We introduce the likelihood for our problem in the section below.

Under certain conditions, the inverse of the Fisher matrix gives a lower bound (the Crámer–Rao bound) on the covariance matrix for those dependent parameters [Vallisneri, 2008]; we discuss the regime of validity of the Fisher matrix inverse as an approximation to the covariance matrix in section 2.7.2. The covariance matrix tells us how sensitive our data is to a change in the model parameters, or equivalenty, which parameters, or combinations of parameters, would have the greatest effect on our observations if they changed. We can also examine which combinations of dependent parameters are degenerate and which combinations yield the greatest information gain.[4]

The Fisher matrix quantifies the sensitivity of predicted observations to model parameters, and provides a bound on the accuracy of parameter inference. This approach assumes that the model is correct. The correctness of the model can be evaluated through other means. For example, model selection can be used to compare distinct models, whether these are different formation channels or different prescriptions for describing the physical processes of binary evolution [e.g., Mandel and O'Shaughnessy, 2010, Vitale et al., 2017a, Stevenson et al., 2017b, Zevin et al., 2017, Talbot and Thrane, 2017], or model-independent clustering can be used without reference to particular models [e.g., Mandel et al., 2015, 2017].

2.6.2 The `COMPAS` likelihood function

For this study we assume that we have a gravitational-wave catalogue of merging binary black holes, formed via the isolated binary evolution channel, and we focus on two observable characteristics of such a dataset: the rate of detections and the distribution of chirp masses for the observed systems.

The likelihood function contains a term for each observational characteristic:

$$\log \mathcal{L}\left(\mathcal{D}|\{\lambda\}\right) = \log \mathcal{L}\left(N_{\mathrm{obs}}|\{\lambda\}, t_{\mathrm{obs}}\right) + \log \mathcal{L}\left(\{\mathcal{M}\}|\{\lambda\}\right). \tag{2.16}$$

[4] This is analogous to identifying the chirp mass as being the best measured combination of masses from gravitational-wave observations.

The first term is the likelihood of observing binary black holes at a given rate. We assume that observable binary black holes coalesce in the Universe as a Poisson process with rate parameter μ, which is predicted by our population synthesis model, and total number of observations N_{obs}, accumulated in a time t_{obs}. The Poisson likelihood is

$$\log \mathcal{L}\left(N_{\text{obs}}|\{\lambda\}, t_{\text{obs}}\right) = N_{\text{obs}} \log(\mu t_{\text{obs}}) - \mu t_{\text{obs}} - \log(N_{\text{obs}}!). \tag{2.17}$$

The second term is the likelihood of observing a given chirp-mass distribution. As described in section 2.5, we have chosen to represent our chirp-mass distribution in bins. This means that our data is a set of counts dsitributed over a set of N_{bins} categories. In this case the correct likelihood is a multinomial distribution [Stevenson et al., 2015]

$$\log \mathcal{L}\left(\{\mathcal{M}\}|\{\lambda\}\right) = \log(N_{\text{obs}}!) + \sum_{k}^{K} \left[c_k \log(p_k) - \log(c_k!)\right], \tag{2.18}$$

where K is the number of chirp-mass bins, c_k is the number of observed systems falling into the k-th bin with $\sum_k c_k = N_{\text{obs}}$, and p_k is the probability predicted by the model that a system falls into the k-th bin. Thus, μ and p_k are functions of the tunable model parameters λ, while c_k and N_{obs} are observables. Given the likelihood, we can now calculate the Fisher matrix.

2.6.3 Computing the Fisher matrix

In order to compute the Fisher matrix, we need to find the second derivatives of the likelihood with respect to the population parameters and average over the possible observations drawn according to the same likelihood distribution. First differentiating the total-rate log likelihood,

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}\left(N_{\text{obs}}|\{\lambda\}\right)}{\partial\lambda_i\,\partial\lambda_j} =&\frac{\partial}{\partial\lambda_j}\left[\left(\frac{N_{\text{obs}}}{\mu} - t_{\text{obs}}\right)\frac{\partial\mu}{\partial\lambda_i}\right]\\ =&-\frac{N_{\text{obs}}}{\mu^2}\frac{\partial\mu}{\partial\lambda_i}\frac{\partial\mu}{\partial\lambda_j}\\ &+\left(\frac{N_{\text{obs}}}{\mu} - t_{\text{obs}}\right)\frac{\partial^2\mu}{\partial\lambda_i\,\partial\lambda_j}.\end{aligned} \tag{2.19}$$

Meanwhile, differentiating the chirp-mass portion of the log likelihood yields

$$\begin{aligned}\frac{\partial^2 \log \mathcal{L}\left(\{\mathcal{M}\}|\{\lambda\}\right)}{\partial\lambda_i\,\partial\lambda_j} =&\frac{\partial}{\partial\lambda_j}\left(\sum_k^K \frac{c_k}{p_k}\frac{\partial p_k}{\partial\lambda_i}\right)\\ =&\sum_k^K\left(-\frac{c_k}{p_k^2}\frac{\partial p_k}{\partial\lambda_i}\frac{\partial p_k}{\partial\lambda_j} + \frac{c_k}{p_k}\frac{\partial^2 p_k}{\partial\lambda_i\,\partial\lambda_j}\right).\end{aligned} \tag{2.20}$$

The expectation value of N_{obs} over this Poisson likelihood with rate parameter μt_{obs} is just $\langle N_{\text{obs}}\rangle = \mu t_{\text{obs}}$; similarly, $\langle c_k\rangle = \mu t_{\text{obs}} p_k$. Therefore, the Fisher matrix is

$$F_{ij} = \mu t_{\text{obs}}\left[\frac{1}{\mu^2}\frac{\partial\mu}{\partial\lambda_i}\frac{\partial\mu}{\partial\lambda_j} + \sum_k^K \frac{1}{p_k}\frac{\partial p_k}{\partial\lambda_i}\frac{\partial p_k}{\partial\lambda_j}\right], \tag{2.21}$$

where we used $\sum_k p_k = 1$ to eliminate the second term from Eq. (2.20). Crucially, this expression contains only first-order derivatives of the observables with respect to the population parameters. These derivatives can be readily and reliably estimated using population synthesis models, as described below.

2.6.4 Evaluating the first derivatives

We have shown in Eq. (2.21) that the Fisher matrix can be computed using just the first derivatives of the binned rates with respect to the population parameters. To compute derivatives, we simulated binary populations using a suite of variations to the population parameters discussed in section 2.4.1. We used the same set of random seeds to the random number generator in COMPAS, so that for each variation the initial conditions (i.e. masses and separation) and random effects (i.e.

kick directions) remain fixed. This allows us to accurately measure the derivatives by estimating the differential rather than absolute rates, reducing the uncertainty associated with a limited number of simulations.

We made six perturbations to the fiducial model for each population parameter (three negative and three positive). The perturbations were chosen to be sufficiently small that we could reliably estimate first derivatives numerically. A full list of the variations we used can be found in table 2.1. For each of the quantities we are differentiating, we have a set of overconstrained simultaneous equations for the first and second derivatives according to the leading terms in the Taylor series, which we can write in matrix form

$$\begin{pmatrix} f(\lambda+\Delta_1)-f(\lambda) \\ \vdots \\ f(\lambda+\Delta_6)-f(\lambda) \end{pmatrix} = \begin{pmatrix} \Delta_1 & \frac{1}{2}\Delta_1^2 \\ \vdots & \vdots \\ \Delta_6 & \frac{1}{2}\Delta_6^2 \end{pmatrix} \begin{pmatrix} \dfrac{\partial f(\lambda)}{\partial \lambda} \\ \\ \dfrac{\partial^2 f(\lambda)}{\partial \lambda^2} \end{pmatrix}. \tag{2.22}$$

If we label the three terms in Eq. (2.22) as $\mathbf{y}$, $\mathbf{X}$ and β respectively, then the maximum-likelihood solution for the derivatives $\hat{\beta}$ can be computed directly as [Anton and Rorres, 2000, section 9.3]

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.23}$$

We use this approach to compute all of the derivatives in Eq. (2.21) and combine them into an estimate of the Fisher matrix. The Fisher matrix can then be inverted to provide the Crámer–Rao lower bound on the covariance matrix of the astrophysical parameters evaluated at the `COMPAS` fiducial model.

2.6.5 Measurement uncertainty

The measurements of chirp masses will be subject to a certain amount of measurement uncertainty. We use a simplified treatment of this measurement uncertainty based on the methodology of Gair et al. [2010], see their appendix A. We assume

σ_{kick} [$\mathrm{km\,s^{-1}}$]	α_{CE}	f_{WR}	f_{LBV}
250.0	1.00	1.00	1.50
240.0	1.00	1.00	1.50
244.0	1.00	1.00	1.50
247.0	1.00	1.00	1.50
253.0	1.00	1.00	1.50
256.0	1.00	1.00	1.50
260.0	1.00	1.00	1.50
250.0	0.95	1.00	1.50
250.0	0.97	1.00	1.50
250.0	0.99	1.00	1.50
250.0	1.01	1.00	1.50
250.0	1.03	1.00	1.50
250.0	1.05	1.00	1.50
250.0	1.00	0.90	1.50
250.0	1.00	0.94	1.50
250.0	1.00	0.97	1.50
250.0	1.00	1.03	1.50
250.0	1.00	1.06	1.50
250.0	1.00	1.10	1.50
250.0	1.00	1.00	1.45
250.0	1.00	1.00	1.47
250.0	1.00	1.00	1.49
250.0	1.00	1.00	1.51
250.0	1.00	1.00	1.53
250.0	1.00	1.00	1.55

Table 2.1: The 25 population-parameter variations used in this paper. The population parameters are described in section 2.4.1: σ_{kick} is the dispersion parameter for a Maxwellian used to draw the magniutde of natal kicks from Eq. (2.1); α_{CE} is the efficiency of common-envelope ejection from Eq. (2.4); f_{WR} is the multiplier for Wolf–Rayet wind mass loss from Eq. (2.6), and f_{LBV} is the multiplier for luminous blue variable mass loss described in Eq. (2.5). Our fiducial model appears in the top row. For each of these population parameter combinations we also varied metallicity. We used 12 different metallicities, which were evenly spaced in the log between $0.005Z_{\odot}$ and $Z_{\odot}$, where we use a solar metallicity $Z_{\odot} = 0.02$. We therefore had a total of 300 model variations. We simulated $1,197,989$ binaries for each of these variations.

that the probability of finding a system in an incorrect bin is given by a Gaussian distribution about the centre of the correct bin into which the system was placed in the simulation.

Let f_i be the fraction of system predicted by the simulation to lie in the i-th bin, which is centred on chirp mass μ_i and has left and right edges at chirp masses μ_i^- and μ_i^+, respectively. Then the probability p_i of observing a system in the i-th bin is

$$p_i = \sum_{j=1}^{K} \frac{f_j}{\sqrt{2\pi\sigma_j^2}} \int_{\mu_i^-}^{\mu_i^+} \mathrm{d}x \, \exp\left[\frac{-(x-\mu_j)^2}{2\sigma_j^2}\right], \tag{2.24}$$

where σ_i is the standard deviation of the measurement in the i-th bin. In the limit of σ_i tending to zero, we recover perfect measurement accuracy, $p_i = f_i$. An illustration of this treatment of the measurement errors is presented in figure 2.4.

The chirp-mass measurement uncertainty depends strongly on the total mass of the source, with the most massive sources spending the fewest inspiral cycles in band, leading to the largest measurement uncertainty [e.g., Abbott et al., 2016a]. It also scales inversely with the source SNR. Here, we crudely approximate this as a fixed fractional uncertainty on the chirp mass of 3 per cent [cf., Mandel et al., 2017, Vitale et al., 2017b]. We therefore modify the binned rates according to Eq. (2.24), using a standard deviation $\sigma_i = 0.03\mu_i$.

This method of incorporating measurement errors is a simplification. The formally correct approach would be to incorporate them on a per-system basis, which would involve a modification of the likelihood function. Performing the analysis in this way would correctly account for correlations between bins, whereas in the simplified approach bins are modified independently, losing information and slightly swelling the uncertainty.

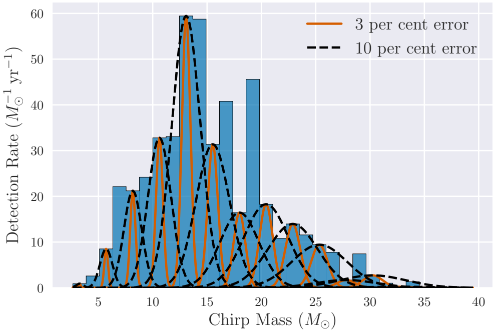


Figure 2.4: An illustration of how we include measurement errors in our analysis. A Gaussian is centred on each bin, with a standard deviation proportional to the value at the centre of that bin. That bin's counts are then distributed to other bins according to the fraction of that Gaussian falling in each bin.

2.6.6 Uncertainty quantification

The rate derivatives used to compute the Fisher matrix at the `COMPAS` fiducial model depend on the particular population realisation used in the calculation. We quantify the impact of simulation realisation noise, due to the finite number of simulated binaries, with bootstrapping. We recompute the Fisher matrix by re-creating data sets of the same size as the original simulated data set by drawing samples from it with replacement.

By repeating this process many times and observing the spread in the results, we can observe how much the point estimates change under different population realisations (different sets of binary initial conditions). Our full dataset consists of $359,396,700$ binary simulations, which consists of the same set of $1,197,989$ ZAMS binaries evolved under each of 300 different model variations (the 25 population parameter combinations listed in table 2.1, each simulated at the 12 different metallicities shown in figure 2.1). To generate one bootstrap sample Fisher matrix:

1. We randomly choose $1,197,989$ initial conditions, with replacement, from our original set of initial conditions.

2. For each of the 25 population parameter combinations in table 2.1, we find the systems from the bootstrap initial conditions which become merging binary

black holes, and calculate their total rate and chirp-mass distribution (taking into account cosmic history, selection effects and measurement uncertainty).

3. We use Eq. (2.22) and Eq. (2.23) to compute the derivatives of the total rate and chirp-mass distribution bin heights, with respect to each population parameter.

4. We use these derivatives to compute the Fisher matrix, using Eq. (2.21).

We repeat the above steps 1500 times in order to express the uncertainty coming from the realisation of the initial conditions, i.e. from the simulation statistical fluctuations. In principle, this model uncertainty could be overcome with more simulations, unlike the more fundamental uncertainties stemming from a finite number of observations and chirp-mass measurement uncertainty. We discuss the relative contributions of these sources of uncertainty in section 2.8.

2.7 Results and discussion

Using the method described in section 2.6 we computed the elements of the Crámer-Rao lower bound on the covariance matrix for the population parameters $\sigma_{\rm kick}$, $\alpha_{\rm CE}$, $f_{\rm LBV}$ and $f_{\rm WR}$. We computed simulation uncertainties on these elements by taking 1500 bootstrap samples from the $1,197,989$ sets of initial conditions simulated for the binaries, specifically varying the metallicities, initial masses and separations. Using these results we are able to explore what can be learned about these population parameters using gravitational-wave observations of binary black holes. Results are presented for $N_{\rm obs} = 1000$ observations, a sufficiently large number to ensure the validity of our results; we discuss the effect of changing the number of observations in section 2.7.2.

Figure 2.5 shows the distribution of standard deviations of each of the population parameters. We see that it will be possible to measure $\alpha_{\rm CE}$, $f_{\rm LBV}$ and $f_{\rm WR}$ with fractional accuracies of ~ 2 per cent after 1000 observations. We will be less sensitive to the value of $\sigma_{\rm kick}$. This is an expected result, since the natal kicks of black holes

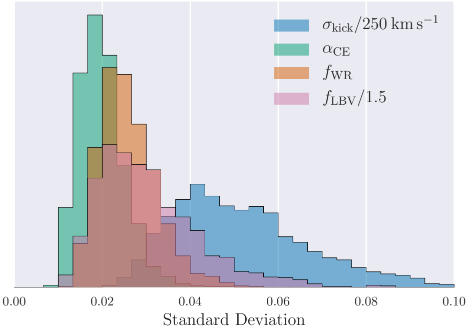


Figure 2.5: The inferred measurement accuracy for each of the four population parameters after observing 1000 systems, as estimated by taking the square root of the diagonal elements of the estimated covariance matrices for each of the 1500 bootstrapped sets. The histograms are normalised such that they all have the same area.

are reduced according to Eq. (2.2), and many of the more massive ones do not get a kick at all.

The fractional uncertainties on all of the parameters are quantities of order $N_{\rm obs}^{-1/2} \approx 0.03$ for $N_{\rm obs} = 1000$. Varying the parameters by their full dynamic range would change the rate by $\mathcal{O}(N_{\rm obs})$. For example, reducing $\alpha_{\rm CE}$ from 1 to 0 would make binary black hole formation through a common-envelope phase impossible, reducing the expected number of detections from $N_{\rm obs}$ to ~ 0.

The measurement accuracy with which the tunable population parameters can be inferred using 1000 gravitational-wave observations can be alternatively interpreted from the perspective of model selection.For example, the median of the distribution for the standard deviation of $\alpha_{\rm CE}$ is ~ 0.02. Therefore, if $\alpha_{\rm CE}$ different from the fiducial value by 6 per cent, the fiducial model could be ruled out with a confidence of $\sim 3\sigma \approx 99.7$ per cent.

We can examine the full multivariate normal behaviour of the population parameters. Figure 2.6 shows marginalised univariate distributions and bivariate projections of the 90 per cent confidence interval for each of the bootstrap samples. This plot shows that most pairwise correlations between most population parameters are negligible. Figure 2.7 shows the correlations between $\alpha_{\rm CE}$ and $f_{\rm WR}$, and between $\alpha_{\rm CE}$ and $f_{\rm LBV}$. Bootstrapping indicates an 88 per cent confidence that $\alpha_{\rm CE}$ and

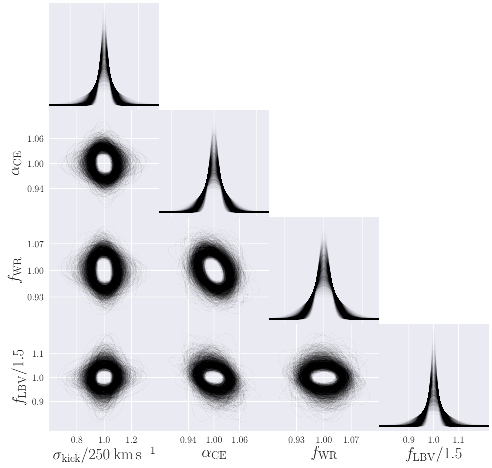


Figure 2.6: 1500 bootstrap samples of the marginalised univariate distributions and bivariate 90 per cent confidence intervals from the Crámer–Rao lower bound on the covariance matrix for the `COMPAS` population parameters. The univariate distributions are the Gaussian distributions corresponding to the standard deviations of figure 2.5, and have been normalised to have the same area.

$f_{\rm WR}$ are anti-correlated. Increasing $\alpha_{\rm CE}$ increases the efficiency with which orbital energy is transferred into the common envelope. An increased efficiency means that there will be less tightening of the binary, so fewer systems will come sufficiently close together to merge within a Hubble time. Losing mass through winds widens the orbit, meaning that increasing the Wolf–Rayet wind mass-loss rate creates more systems which are too wide to merge within a Hubble time. Increased mass loss also results in the black holes being less massive, therefore increasing the time required for them to merge through gravitational-wave emission from a given initial separation [Peters, 1964]. These correlations mean that increasing (or decreasing) both $\alpha_{\rm CE}$ and $f_{\rm WR}$ would compound the effect on the rates, so their bivariate distribution (in figure 2.6) is narrower in this direction. Conversely, the effects of increasing one whilst decreasing the other would partially cancel out, and thus the bivariate distribution is wider in that direction. The confidence in the anti-correlation between $\alpha_{\rm CE}$ and $f_{\rm LBV}$ is only 76 per cent, and there is insufficient evidence for correlation between other parameter pairs.

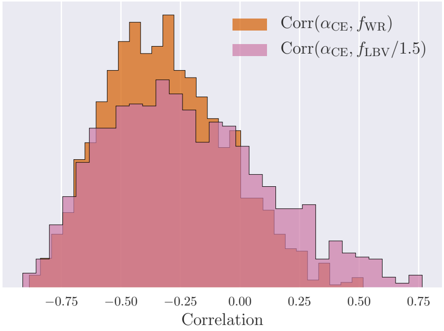


Figure 2.7: Distribution of correlations between α_{CE} and each of f_{LBV} and f_{WR}. The histograms have been normalised to have the same area.

2.7.1 Information from the total detection rate

To gain further insight into the correlations between the inferred parameters, we now consider what we could learn about the population parameters by considering only the total rate at which gravitational waves are observed. It is impossible to constrain the four-dimensional population parameter vector considered in this paper with a single observable, the binary black hole detection rate. In this case, all that can be learned about the population parameters is the value of some linear combination of them.

We construct a detection rate Fisher matrix, using only the total rate log likelihood of Eq. (3.16),

$$F_{ij}^{\mathrm{RO}} = \frac{t_{\mathrm{obs}}}{\mu} \frac{\partial \mu}{\partial \lambda_i} \frac{\partial \mu}{\partial \lambda_j}, \tag{2.25}$$

and perform an eigendecomposition. We expect to see that there is only one eigenvector whose eigenvalue is non-zero. We verified that this is true for all 1500 of our bootstrap samples, which provided a useful sanity check of our results.

Next, by examining the eigenvector whose eigenvalue is non-zero, we can find the linear combination of population parameters to which we are sensitive. Figure 2.8 shows a univariate representation of this direction (with its distribution from bootstrapping over simulations). The components of the vector parallel to f_{LBV} and σ_{kick}

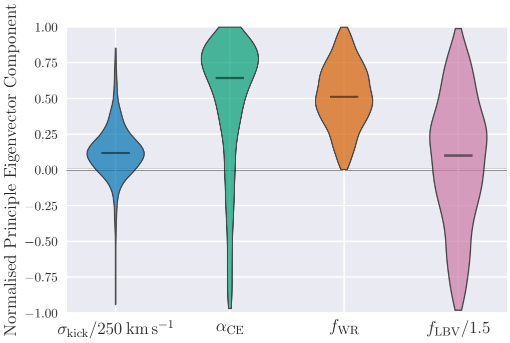


Figure 2.8: Violin plot showing components of the normalised principal eigenvector of the Fisher matrices calculated using only the total detection rate. The coloured regions give the bootstrapped distribution of the principle eigenvector direction, with medians marked in black.

axes are broadly consistent with zero. Most of the information learned solely from the total detection rate is in the α_{CE}–f_{WR} plane. The fact that both values are simultaneously positive implies that they are positively correlated; this is the same correlation as was discussed at the beginning of this section.

Whilst we can only measure this specific combination of population parameters using only the total detection rate, we can constrain parameter combinations in the $\sim \alpha_{\mathrm{CE}} + f_{\mathrm{WR}}$ direction to within a few per cent from the total rate. Figure 2.9 shows the standard deviation along the line defined by this combination of population parameters $a^{-1/2}$, where a is the principal eigenvalue. This can be interpreted in the same way as the standard deviations in figure 2.5, and matches the expected value of $\mathcal{O}(N_{\mathrm{obs}}^{-1/2})$. We see that if this combination of population parameters differed from our fiducial values by more than a few per cent, we would be able to confidently rule our model out after 1000 observations. However, we also see from figure 2.9 that including the chirp-mass distribution would significantly improve measurements of this parameter combination.

2.7.2 Number of observations

The expected number of observations only appears as a multiplicative term in Eq. (2.21), so that the standard deviations in figure 2.5 simply scale as $N_{\mathrm{obs}}^{-1/2}$.

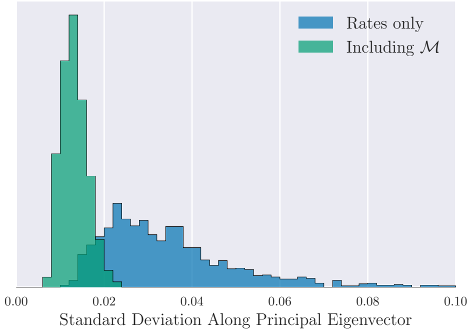


Figure 2.9: Distribution of the standard deviation of the particular linear combination of population parameters corresponding to the principal eigenvector of the total detection rate Fisher matrix. The measurement accuracy is computed using only information from the total rate (blue) and after including information from the chirp-mass distribution (green). The distributions come from considering all 1500 bootstrapped sets.

However, the results presented here are predicated on the assumption that the inverse of the Fisher information matrix is a good approximation to the covariance, and not just a lower bound. This in turn requires the likelihood to be approximately Gaussian, i.e. the linear single approximation [LSA; Vallisneri, 2008] should hold. Only if the predicted parameter uncertainties are smaller than the neighbourhood in which the LSA is valid does the Fisher matrix provide a self-consistent estimate of the accuracy of parameter inference. This effectively sets a minimal threshold on the number of observations required for self-consistency in our estimates.

When computing the derivatives, as described in section 2.6.4, we measure the terms in a Taylor expansion of an observable (binned) rate f as a function of the population parameter λ,

$$f(\lambda+\Delta)-f(\lambda)\approx\Delta f'(\lambda)+\frac{\Delta^2}{2}f''(\lambda). \tag{2.26}$$

In order to verify the validity of the LSA, we need to check that each f is indeed linear when Δ is of the same order as the computed standard deviations for the population parameters. We require that the linear term is dominant in the Taylor series, so that

$$f'(\lambda) \gg \frac{\Delta}{2} f''(\lambda). \tag{2.27}$$

We find $N_{\mathrm{obs}} = 1000$ to be a sufficient lower limit on the number of observations necessary to ensure the LSA is valid. At 1000 observations, the best measured combination of parameters is constrained at the per cent level, and this will continue to improve as we expand the catalogue of observations.

For smaller numbers of observations, the LSA will break down. The probability distribution for the model parameters may no longer be a multi-dimensional Gaussian so the Fisher matrix is likely to under-estimate the inference uncertainty.

2.8 Conclusions

We have, for the first time, quantitatively analysed how accurately gravitational-wave observations of binary black hole mergers will constrain binary population synthesis models described by a multi-dimensional parametrisation. When ground-based detectors have accumulated 1000 observations of merging binary black holes, we have shown that we will measure binary population synthesis model parameters with an accuracy of a few per cent. Equivalently, we will be able to distinguish models for which the population parameters only differ by a few per cent.

Our analysis accounts for three distinct sources of uncertainty in the inference of population parameters using gravitational-wave observations. The first is due to the finite number of observations. We show when the linear signal approximation holds (section 2.7.2), the accuracy with which population parameters can be inferred scales with the inverse square root of the number of observations. The second is the chirp-mass measurement uncertainty in individual observations. We only model this approximately (section 2.6.5) but find that it is unlikely to be limiting factor in inference. The third source of uncertainty is simulation uncertainty: the accuracy in predicted detection rate and chirp-mass distribution is limited by the finite number of COMPAS simulations. This uncertainty, which we quantify with bootstrapping (section 2.6.6), is only limited by computational cost, and be reduced indefinitely with more simulations or more efficient sampling [e.g., Andrews et al., 2017].

There is, of course, potential systematic uncertainty in the binary evolution models themselves: for example, it is probable that the α_{CE} parameter is not universal, as assumed here, but depends on the binary properties during the common-envelope phase. Model inference techniques such as those described here should be combined with model verification and with weakly modelled inference [e.g., Mandel et al., 2017].

We show the expected detection rate and chirp-mass distribution of merging binary black holes in figure 2.3. The sharp features in the chirp-mass distribution are due to only simulating systems at a small number (12) of metallicities, replacing the integral over metallicity in Eq. (3.5) with a discrete sum. Mass loss, particularly during the luminous blue variable phase, leads to a pile up of black hole masses from the most massive stars at particular metallicity-dependent values. The subsequent discrete sum over metallicities overpopulates some bins in the chirp-mass distribution relative to neighbouring bins [cf. Dominik et al., 2013]. This can impact our results, causing us to over-state the accuracy with which we will be able to measure population parameters. This issue can be addressed in the future by interpolating model predictions over metallicity [e.g., using Gaussian process emulators as described by Barrett et al., 2017b], producing a smooth set of predictions.

Our primary intention with this paper was to introduce a methodology for evaluating the accuracy with which astrophysical model parameters can be estimated based on the rates and properties of observed transients. We considered a four-dimensional parameter space, but the number of dimensions is limited only by computational cost. It is also straightforward to add more terms than just the chirp-mass distribution to Eq. (3.15) in order to investigate other observable characteristics of binary populations such as mass ratios and spins [e.g., Stevenson et al., 2017b, Talbot and Thrane, 2017, Zevin et al., 2017]. Furthermore, this analysis can be used for other populations than observations of binary black hole mergers via gravitational-waves in this paper. Other observed populations, such as Galactic binary pulsars, X-ray binaries, Wolf–Rayet stars, short gamma-ray bursts or luminous red novae [for a review, see De Marco and Izzard, 2017], can provide orthogonal constraints on the parameters governing binary evolution (cf. figure 2.9). Over the coming decade, such measurements will help us to accurately determine the physics of massive binary evolution.

Chapter 3

Emulators

3.1 Introduction

In the previous chapter, we described how much can be learned about population parameters in population synthesis models, given a number of observations of merging black holes. This is a fruitful exercise in the absence of data. However, we are now entering a regime where we do have data, and our model predicts, in figure 2.3, that in as few as 2-3 years we expect to enter a data rich regime, with 100s or even 1000s of detections. Once we have a catalogue of these detections, we can begin to directly compare them to our model.

The correct approach to this problem would be to write down a likelihood function to compare the observed data to a prediction of the model, and then explore the population parameter space, using one of the stochastic sampling techniques described in 1.2.3, to find the combination of population parameters for which the model predictions best match the observations. 'Prediction', in this context, means the output of COMPAS after it is used to reproduce a potential observation. This approach would involve making a model prediction for each new set of population parameters proposed by the sampling algorithm. Even for rapid population synthesis, these model predictions can be expensive. For example, `COMPAS`, on average, takes ~ 0.3 seconds to simulate the outcome of a single binary. A model prediction involves simulating the outcome of $\sim 10^5 - 10^6$ binaries, taking tens of hours. Even with efficient stochastic sampling techniques, millions of model predictions may need to be made, making this problem computationally intractable.

This chapter describes our use of machine learning techniques to build and use an emulator; a computationally inexpensive approximation to the `COMPAS` model.

3.2 `COMPAS` Model Predictions

A `COMPAS` model prediction involves drawing a large number of initial conditions from the distributions described in section 1.1.3, and then simulating the outcome of each of these initial conditions. Each simulation depends on a set of 'population parameters' which are shared between all of the initial conditions. In this chapter, we again focus on a few key population parameters, namely the common envelope efficiency α_{CE}, supernova kick velocity dispersion σ_{kick} and mass loss rate during the luminous blue variable phase f_{LBV}, which are described in detail in section 2.4.1. We further include variations in metallicity, which is another population parameter representing the fraction of the stars (by mass) which are neither Helium nor Hydrogen.

Each of the initial conditions is simulated according to the physics prescribed by the population parameters. The binaries which are 'interesting' are then identified and their properties recorded. "Interesting", for this chapter, means binary black holes which will merge within the lifetime of the universe. These merging systems are binned by their chirp mass $\mathcal{M}$.

$$\mathcal{M} = \frac{(M_1 M_2)^{\frac{3}{5}}}{(M_1 + M_2)^{\frac{1}{5}}}. \tag{3.1}$$

The number of counts in each bin is expected to follow a Poisson distribution. We made the assumption that each bin is sufficiently populated that we were able to approximate the errors on the counts in each bin as being approximately Gaussian with standard deviation $\pm\sqrt{N}$, where N is the number of counts in the bin. In the cases where we had zero counts in a bin, we inflated the errors to ±1 count. We refer to these counts and their associated errors as the 'raw' chirp mass distribution.

The rates at which systems merge within COMPAS is not physically meaningful. In order to make this rate into a physical quantity, we must take into account factors such as the star formation rate at the time the system formed, and how much of all star forming mass is represented in our simulation. We begin by summing the total mass described by the intial conditions of the model prediction. This only represents a fraction of the total mass formed, since we typically only simulate massive systems (i.e., systems which stand a chance of becoming binary black holes). We therefore compute the fraction of star forming mass represented by our simulations by considering the initial mass function of Kroupa [2001], described in section 1.1.3.

$$m \propto \begin{cases} m^{-0.3} & \text{for } (0.01 < m < 0.08M_\odot) \\ m^{-1.3} & \text{for } (0.08M_\odot < m < 0.5M_\odot) \cdot \\ m^{-2.3} & \text{for } (0.5M_\odot < m) \end{cases} \tag{3.2}$$

This is also where we take into account the fraction of systems formed in binaries, although for massive stars we make the assumption that all stars are formed in binaries. Once we have computed the total mass represented by our simulation, we can write down a basic merger rate

$$\mathcal{R} = \frac{\mathrm{d}^3 N_{\mathrm{form}}}{\mathrm{d}M_{\mathrm{form}}\,\mathrm{d}\tau_{\mathrm{delay}}\,\mathrm{d}\mathcal{M}}, \tag{3.3}$$

where N_{form} is the number of merging binary black hole systems that form, M_{form} is the amount of mass formed in stars and $\mathcal{M}$ is the chirp mass. Each system takes a different amount of time to evolve and become a double compact object, and then depending on their mass and separation, each system will take a different amount of time to coalesce. The time from the birth of the binary to its coalescence is known as the delay time τ_{delay}. The rate of star formation in the universe changes appreciably over the typical delay timescale, so we must take this into account when computing a physical merger rate in the universe. Star formation rate also depends on the metallicity of the stars, as demonstrated in figure 2.1. We described in detail how this rate can be estimated in chapter 2.

In order to consider the total rate of mergers in the universe per unit comoving volume V_c, we must consider contributions from all possible formation times and from all possible delay times

$$\frac{\mathrm{d}^4 N_{\mathrm{merge}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}\mathcal{M}\,\mathrm{d}Z}(t_{\mathrm{merge}}, Z) = \int \mathrm{d}t_{\mathrm{form}} \int \mathrm{d}\tau_{\mathrm{delay}}\,\mathcal{R}(Z)\frac{\mathrm{d}^3 M_{\mathrm{form}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}Z}(Z, t_{\mathrm{form}}). \tag{3.4}$$

These quantities are related by the definition of the delay time $t_{\mathrm{merge}} = t_{\mathrm{form}} + \tau_{\mathrm{delay}}$, where t_{merge} is the time at which the system merges, in its own reference frame, due to the emission of gravitational-waves. We can therefore simplify the expression to a single integral

$$\frac{\mathrm{d}^4 N_{\mathrm{merge}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}\mathcal{M}\,\mathrm{d}Z}(t_{\mathrm{merge}}, Z) = \int \mathrm{d}\tau_{\mathrm{delay}}\,\mathcal{R}(Z)\frac{\mathrm{d}^3 M_{\mathrm{form}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}Z}(Z, t_{\mathrm{merge}} - \tau_{\mathrm{delay}}). \tag{3.5}$$

Finally, we must consider that not all systems in the universe are equally detectable, principally depending on how massive the systems are as well as how far away from the detector (Earth) they are, here quantified by their redshift z. The overall detection rate is then given by considering contributions from all metallicities and redshifts

$$\frac{\mathrm{d}^2 N_{\mathrm{obs}}}{\mathrm{d}t_{\mathrm{obs}}\,\mathrm{d}\mathcal{M}} = \int \mathrm{d}z \int \mathrm{d}Z \left[\frac{\mathrm{d}^4 N_{\mathrm{merge}}}{\mathrm{d}t_{\mathrm{s}}\,\mathrm{d}V_{\mathrm{c}}\,\mathrm{d}\mathcal{M}\,\mathrm{d}Z}\frac{\mathrm{d}V_{\mathrm{c}}}{\mathrm{d}z}\frac{\mathrm{d}t_{\mathrm{s}}}{\mathrm{d}t_{\mathrm{obs}}}P_{\mathrm{det}}\right]. \tag{3.6}$$

The detection probability P_{det} can be computed either at 'design sensitivity', representing the probability of making a detection on Earth when the aLIGO detectors are operating at their expected peak sensitivity, or at 'O1' sensitivity, to represent the sensitivity of the instruments during the first and second aLIGO observing runs (where all detections have been made so far). This 'model integrand' can be computed using a `COMPAS`. We use the same assumptions about cosmology and selection effects described in sections 2.5.1 and 2.5.2 for the work in this chapter. The integrand is similar to the raw chirp mass distribution described at the start of this section, but reweighted to take account of cosmological and selection effects. We denote the detection rate in the k^{th} bin of the i^{th} model prediction as μ_k^i, and the left and right edges of this k^{th} bin are $\mathcal{M} = \mu_{k-}^i$ and $\mathcal{M} = \mu_{k+}^i$ respectively, so that

$$\mu_k^i = \int_{\mu_{k-}^i}^{\mu_{k+}^i} \mathrm{d}\mathcal{M} \, \left[\frac{\mathrm{d}^4 N_{\mathrm{merge}}}{\mathrm{d}t_{\mathrm{s}} \, \mathrm{d}V_{\mathrm{c}} \, \mathrm{d}\mathcal{M} \, \mathrm{d}Z} \frac{\mathrm{d}t_{\mathrm{s}}}{\mathrm{d}t_{\mathrm{obs}}} P_{\mathrm{det}} \right] . \tag{3.7}$$

Each model integrand therefore consists of a set of μ_k^i, computed at a particular combination of population parameters. The errors on these integrands, $(\sigma\mu_k^i)$ are scaled by comparing the counts from the raw chirp mass distribution to the μ_k^i. When there are zero counts in a bin, we use the minimum non zero error for that integrand. The cases where there are no counts in any bin in an integrand are dealt with in section 3.5. It is these integrands which we seek to build an emulator for.

Model predictions are finally made by evaluating the integral in equation 3.6. In practise we do this by quadrature, so that

$$\frac{\mathrm{d}N_{k,\mathrm{obs}}}{\mathrm{d}t_{\mathrm{obs}}} = \int \mathrm{d}z \int \mathrm{d}Z \, \mu_k^* \frac{\mathrm{d}V_{\mathrm{c}}}{\mathrm{d}z} \approx \sum_z \sum_Z \mu_k^* \, \Delta z \, \Delta Z \, \frac{\Delta V_{\mathrm{c}}}{\Delta z}, \tag{3.8}$$

where μ_k^* represents the integrands as defined in 3.7 computed on a grid of metallicities and redshifts.

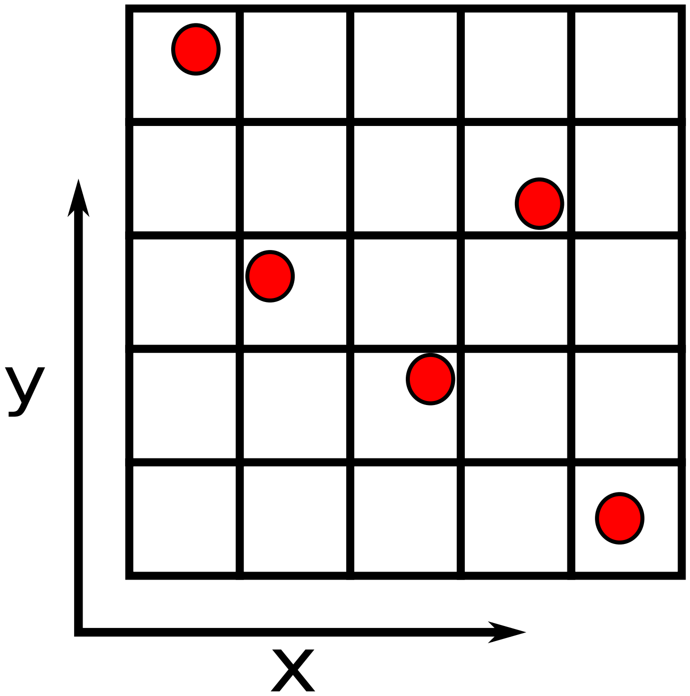


Figure 3.1: An example of a Latin Hypercube experimental design in a two dimensional parameter space. Each sample point (red) is placed such that the marginalised distribution in both x and y are uniform.

3.3 The Training Set

Emulators, also known as surrogate models, are built using supervised machine learning methods, and so require a training set [Conti and O'Hagan, 2010]. As described in the previous section, we have chosen to emulate the model integrands defined in equation 3.7, as a function of the population parameters α_{CE}, σ_{kick}, f_{LBV} and Z. In order to maximise the chance that the emulator performs well in all regions of population parameter space, it can be important to carefully design the the training set. It is not feasible to sample the space on a sufficiently dense regular grid, since the number of sample points required increases exponentially with dimensionality.

A common strategy to ensure that the parameter is explored well is to use a technique called Latin Hypercube Sampling (LHS), which is a space filling algorithm to place sample points in the parameter space such that the distribution of sample points is uniform, when looking at a one dimensional marginalisation of each of its dimensions [McKay et al., 1979]. Figure 3.1 gives a simple example of a LHS experimental design in two dimensions.

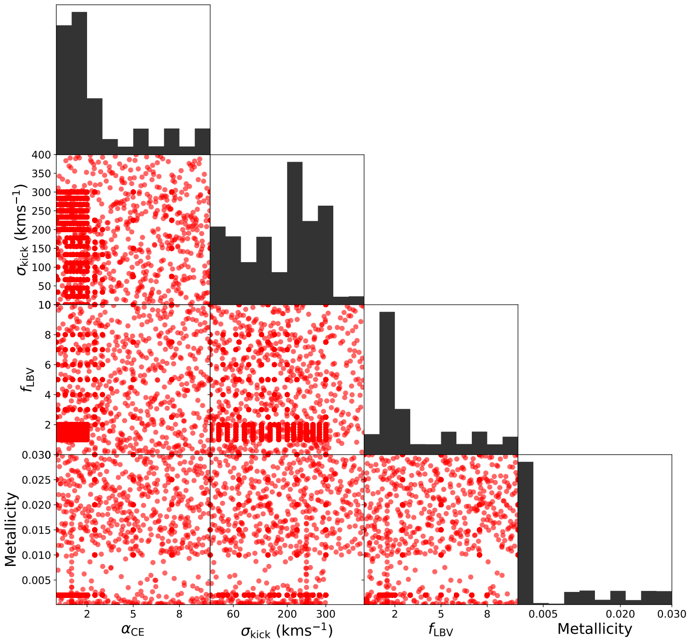


Figure 3.2: The location in population parameter space of all 2352 `COMPAS` populations used in construction of the emulator, with the marginalised distribution of each plotted on the diagonal, and pairwise scatter plots on the off-diagonal panels. The dense patches of training examples are typically where we have incorporated data from other projects.

Since we used a standard version of `COMPAS` for the work described in this chapter, we already had access to $\sim$ 1300 training populations from our preliminary investigations [Barrett et al., 2017b] and from other active projects within the research group. These were largely on regular grids. We subsequently generated a further $\sim$ 1000 training populations using latin hypercube sampling experimental designs, so that our entire training set consists of 2352 training populations, representing the evolution of a total of 634 980 019 binary simulations with `COMPAS`. The distribution of training examples in the population parameter space is plotted in figure 3.2.

For each of the training populations, we computed a set of training examples μ_k^i as defined in equation 3.7 for each of 100 different redshifts, evenly spaced in the log between $z \in [0, 2]$, to give a total of 235200 total sets of model integrands in the training set.

3.4 Performance Metric

In order to make a quantitative comparison between different emulators, we need to develop a metric which measures their performance. For each emulator, we hold back a 'validation set' of model integrands from the training process, which we can then agnostically predict. Since we have well defined Gaussian uncertainties for each of the true μ_k^i in our training set, we chose to quantify performance by measuring the significance of the deviations of emulator predictions from the truth.

For each training example in the validation set, we use the following procedure. For each bin, we compute a 2 sided p-value for the predicted rate coming from the distribution defined by the true rate and its uncertainty. For example, if the true rate was $0.5 \pm 0.1\,\mathrm{yr}^{-1}$, and the emulator predicts $0.3\,\mathrm{yr}^{-1}$, then this result has a p-value of $p = 0.046$. This is the probability of seeing an absolute deviation from the mean greater than or equal to $0.2\,\mathrm{yr}^{-1}$. This p-value represents the extent to which the emulator and `COMPAS` are statistically distinguishable. We are effectively performing a null hypothesis test, where the null hypothesis is that `COMPAS` and the emulator are the same model.

Combining the p-values across distributions is challenging, since measuring each bin does not provide an independent test of the null hypothesis, as the prediction success will be correlated across bins. This means that the p-values are not uniformly distributed, since if one bin is poorly emulated, it is likely other bins will also be poorly emulated. We therefore elect to judge a distribution on its most poorly predicted bin, by choosing the lowest p-value for each distribution.

3.5 Gaussian Process Emulators

We introduced Gaussian process regression in section 1.3.1. Gaussian process regression deals with problems with a scalar output, making the assumption that any subset of outputs share a joint, multivariate Gaussian distribution. A parametrised equation for the covariance between any two outputs is then written down and the parameters optimised with respect to the training data. This trained covariance function can then be used to compute the distribution of unseen outputs. Gaussian processes are a natural choice of method for this problem, since the amount of training samples is modest, and we expect the chirp mass rates to vary smoothly as a function of the input parameters. The other attractive property of Gaussian process emulators is that they give a full distribution for the outputs, giving a measure of the uncertainty of estimates, potentially informing future experimental designs.

Since the training examples we described in section 3.2 are multidimensional, we can not predict them all using a single Gaussian process. We therefore need to train an ensemble of Gaussian processes to predict the outputs. Since we expect the rates in neighbouring bins to be correlated, naively modelling them with indpendent Gaussian processes could lead to unrealistic predictions. We therefore employ a singular value decomposition (SVD) on the entire training set. A singular value decomposition takes the $N_{tr} \times N_{bin}$ (Number of training examples × number of chirp mass bins) matrix of transformed histograms M and finds a complete basis of orthogonal eigenvectors V with a diagonal matrix of their associated eigenvalues D, as well as the training set's representation in this basis U, so that

$$M = UDV. \tag{3.9}$$

Since the basis is now orthogonal, we no longer have to worry about correlations, and can train an independent Gaussian process for each component. Moreover, since the decomposition is linear (it's just a rotation of the 'bin space'), each training example in the new representation is just a linear combination of its components in bin space. It is therefore trivial to propagate errors through the decomposition. If we define the dense matrix (δM) as the matrix of errors on each element of M, and the corresponding matrix of errors (δU) in the rotated basis, then

$$\begin{aligned}(\delta M) &= (\delta U)DV && (3.10)\\ (\delta U) &= (\delta M)(DV)^{-1} && (3.11)\\ &= \left((V^T D^T)^{-1}(\delta M)^T\right)^T , && (3.12)\end{aligned}$$

where the last line is how the calculation is implemented for numerical stability.

Another benefit of using a singular value decomposition is that it allows us to make a dimensionality reduction to our training set. The eigenvalues returned by the SVD represent the relative importance of the corresponding eigenvector for describing the data. Typically, the vast majority of the data can be described with a small subset of the basis eigenvectors, so it's possible to simply ignore some of the less important directions without any significant loss of accuracy.

Next we transform the training set. We normalised the components of the training examples in U space so that they were within the range $[0, 1]$ by dividing by the range and subtracting the minimum value for each component. We also transformed the population paramters so that they were both log spaced and within the range $[0, 1]$. If the original population parameters are $\{\lambda_i\}$, then we instead emulate over t_i, where

$$t_i = \frac{g_i - \min(g_i)}{\max(g_i) - \min(g_i)} \qquad (3.13)$$

$$g_i = \log\left(\lambda_i + \min(\lambda_i)_{x_i>0}\right). \qquad (3.14)$$

Due to computational considerations, Gaussian processes can't be used on the full set of 235200 training examples spanning all redshifts. Our emulator therefore consists a small number of ensembles of GPs, each at a different redshift, and each with 2352 associated training populations (each at different redshifts). Again for computational reasons, and to limit the dimensionality of the input, we treated the populations coming from different redshifts as being independent. The emulator was therefore built to predict the rate integrands defined in equation 3.7 for fixed redshifts, but as a function of Z, α_{CE}, σ_{kick} and f_{LBV}.

Each ensemble of GPs had an independent GP for each component. We trained each GP individually, using the LBFGS gradient based optimiser [Press, 2007, Nocedal and Wright, 1999] to maximise the likelihood in equation 1.25. We ran the optimiser 5 times for each GP with randomised starting conditions, and kept the combination of hyperparameters with the highest overall likelihood. We use the `GPy` implementation of Gaussian Processes [1]. We used squared exponential covariance kernels for each GP, as defined in equation 1.23, with a diagonal metric. We also tried using a Matern-3/2 kernel, which made no significant difference to the success of our model.

In preliminary testing, we found that the emulator performed very badly with 'zero rate' training examples, where the rate in every bin was zero. We therefore also trained a random forest binary classifier to predict the 'zero rate' locations in the training set. Since random forests are capable of dealing with significantly more data than Gaussian process emulators, we were able to use the full 235200 example training set for the classifier. We used the `Scikit-Learn` implementation of random forest classifiers [2].

We trained 8 ensembles of GPs at redshifts spaced logarithmically between 0.23 and 1.24. We chose these redshifts since they give a good representation of cosmic history, and rates from more distant systems make insignificant contributions to the overall detection rate. We also trained a random forest classifier with an ensemble of 500 underlying decision trees in order to predict the locations of zero rate training examples. We kept back 5% of the training set for validation.

The classifier performed well, with a $< 2\%$ misclassification rate. On further investigation, the misclassified examples were exclusively cases where the validation point had a very small but non zero rate distribution, and so would have a negligible effect on the model prediction in equation 3.8. We then proceeded to use this classifier paired with the GP emulator to predict the shape of the model integrand.

[1] `https://github.com/SheffieldML/GPy`

[2] `https://github.com/scikit-learn/scikit-learn`

In figure 3.3 we show a sample of the validation set and their predictions.

By eye, the emulator performs reasonably well for the majority of validation points, and very badly at others. The emulator also seems to be confidently wrong in its predictions, with some errors on the predictions clearly not being consistent with the true distributions coming from `COMPAS`. We used the minimum p-value performance metric described in section 3.4 in order to evaluate the performance of the emulator at each of the validation points. In figure 3.4 we show the results of this analysis throughout the population parameter space, without making any modifications to the validation set. The results confirm the visually poor performance of the emulator seen in figure 3.3. However, the extremeness of many of the p-values are surprising. On further investigation, it was found that these very small values come from small deviations about zero for bins where the rate was identically zero, which in most cases had very small error bars on the true `COMPAS` output.

In order to see through this effect, we inflated the error bars on the true distribution for the validation set, so that the error on every bin matched the maximum error for that example. The result of running the minimum p-value analysis on this modified validation set are shown in figure 3.5. With these modifications, the emulator still doesn't perform well, but the results are commensurate with the intuition coming from figure 3.3. We also removed the 29 most poorly predicted examples from the validation set and repeated the analysis in figure 3.6, in order to facilitate an easier visual comparison with the random forest emulator we describe in the sections that follow.

We finally tested the Gaussian process emulator's success on making a model prediction, by using it to predict a grid of μ^* to complete the sums in equation 3.8. We used a grid of 100 metallicities together with the 8 redshifts used to create each ensemble of Gaussian processes. It is not possible to make a validation point for this prediction using the simulations described in this chapter, and creating model prediction using `COMPAS` requires hundreds of hours of computation. However figure 2.3 from the previous chapter is a full model prediction, and is made using

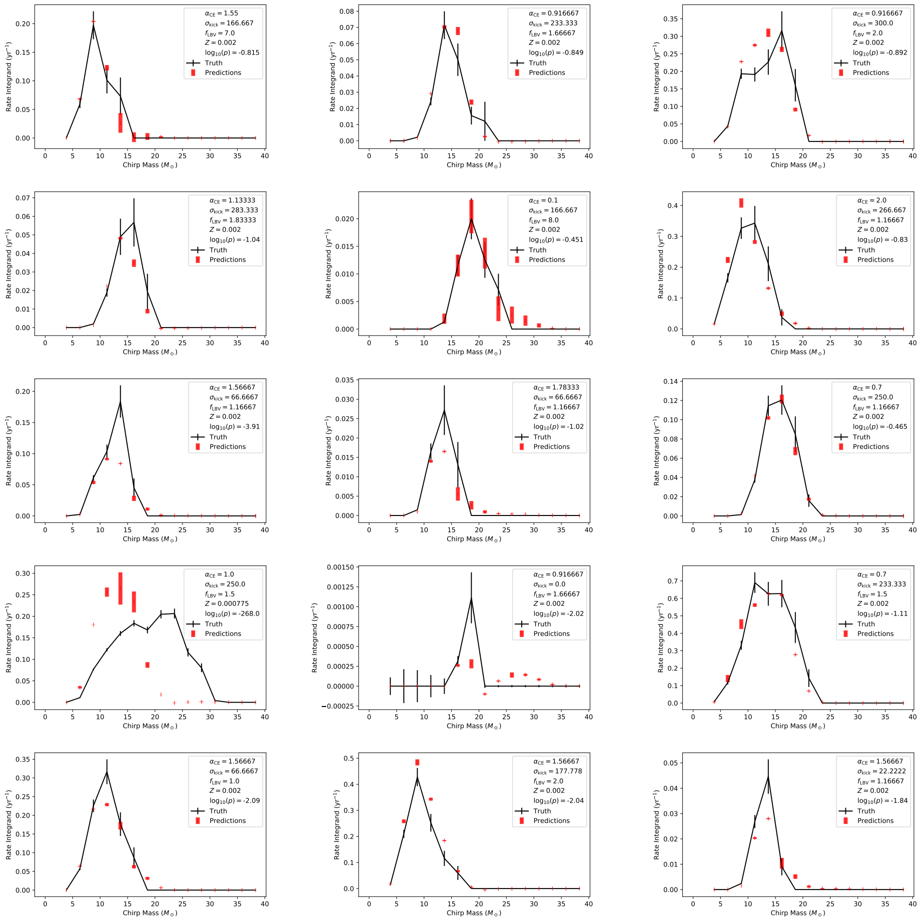


Figure 3.3: 15 random examples from the validation set with the Gaussian process emulator predictions, with the validation set examples in black and the Gaussian process predictions in red. In the legend, we include the p-values calculated using the inflated errors as described in the caption of figure 3.5, although the uncertainties are displayed in their non-inflated form.

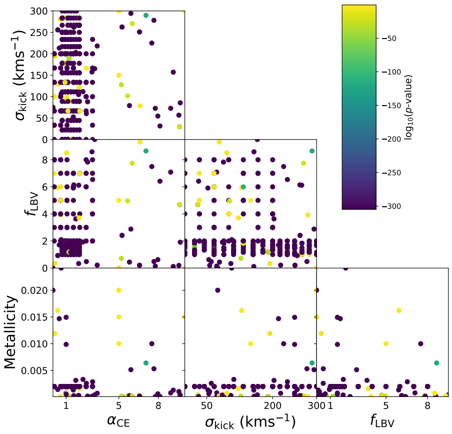


Figure 3.4: The p-value for the most poorly predicted bin at each point in population parameter space by the Gaussian process emulator.

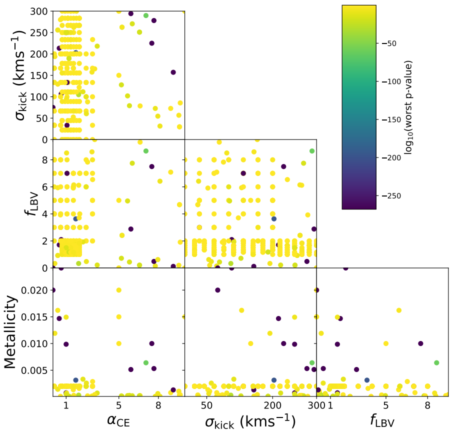


Figure 3.5: The p-value for the most poorly predicted bin at each point in population parameter space by the Gaussian process emulator. Here we have inflated the errors on each bin in the validation set, so that the error on every binned rate matches the error on the highest binned rate for that validation example. This penalises fluctuations about ~ 0 rate bins less severely.

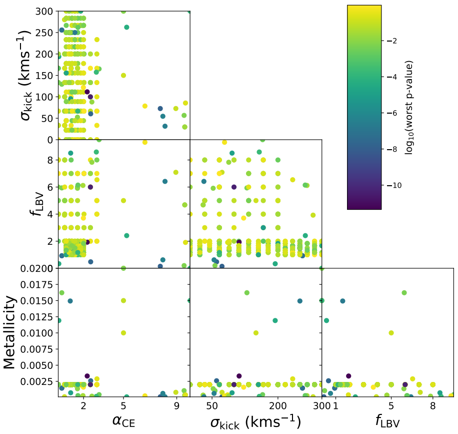


Figure 3.6: The p-value for the most poorly predicted bin at each point in population parameter space by the Gaussian process emulator. This is identical to figure 3.5, except that the 29 most poorly predicted validation points have been excluded.

a completely independent dataset from the one presented here. In figure 3.7 we present a number of draws from the full integral prediction using the Gaussian process emulator over a realisation of this independent model prediction.

Our implementation of a Gaussian process emulator fails to adequately predict the fiducial model prediction, but perhaps more worryingly it is extremely confident in its prediction. It is clear from this plot that the Gaussian process emulator is not a sufficiently good emulator for use in inferring population parameters. We defer a thorough discussion of reasons why the emulator may be performing poorly, and how it might be improved, to section 3.10.

3.6 Random Forest Emulator

Since the random forest classifier had performed so well in predicting the location of zero rate distributions, we decided to try a full random forest regression on the whole dataset. In contrast to the Gaussian process emulator, we were able to use the full 235200 example training set, allowing us to also use redshift as an input

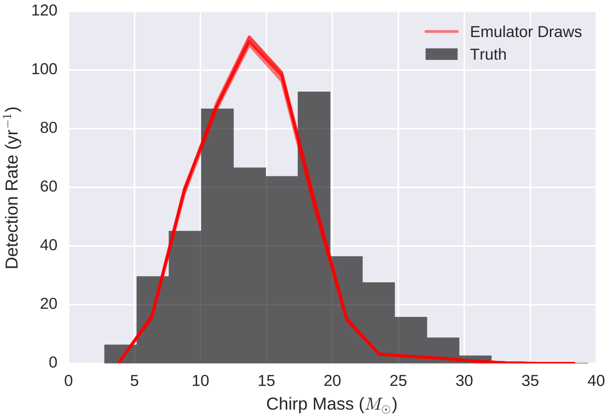


Figure 3.7: The result of using the Gaussian process emulator to predict a grid of integrands in metallicity and redshift to compute the integral in equation3.8. We plot 50 draws from the Gaussian processes, predicting the detection rate for the fiducial model, which we compare agnostically to a prediction made using the data in figure 2.3

parameter, so that we now have a 5 dimensional input space (α_{CE}, σ_{kick}, f_{LBV}, Z, z). We trained the model using a forest of 200 underlying decision trees, growing each tree to its maximal extent (1 training example per leaf node).

Initially, in order to make a fair comparison to the Gaussian process emulator, we built the emulator at a single redshift, so that the model had just 2352 training examples. Figure 3.8 shows a random selection of validation points and their associated random forest emulator predictions. It is optically clear that the random forest emulator outperforms the Gaussian process emulator.

We applied the minimum p-value analysis described in section 3.4, and the results are plotted in figure 3.9. It is again clear that the random forest emulator thoroughly outperforms the Gaussian process emulator, with a median p-value across all predictions of $p \approx 0.1$. It is also important to note that no modifications had to be made to the validation set to achieve this performance.

Next, we built two emulators, one using a training set built such that P_{det} in equation 3.7 corresponds to the detectors at design sensitivity, and the other using the a training set calibrated to O1 sensitivity. In each case we used the full 235200 example training set across all redshifts. We used these to predict the μ^* necessary to compute the integral in equation 3.8. We predicted a grid of 25 redshifts and

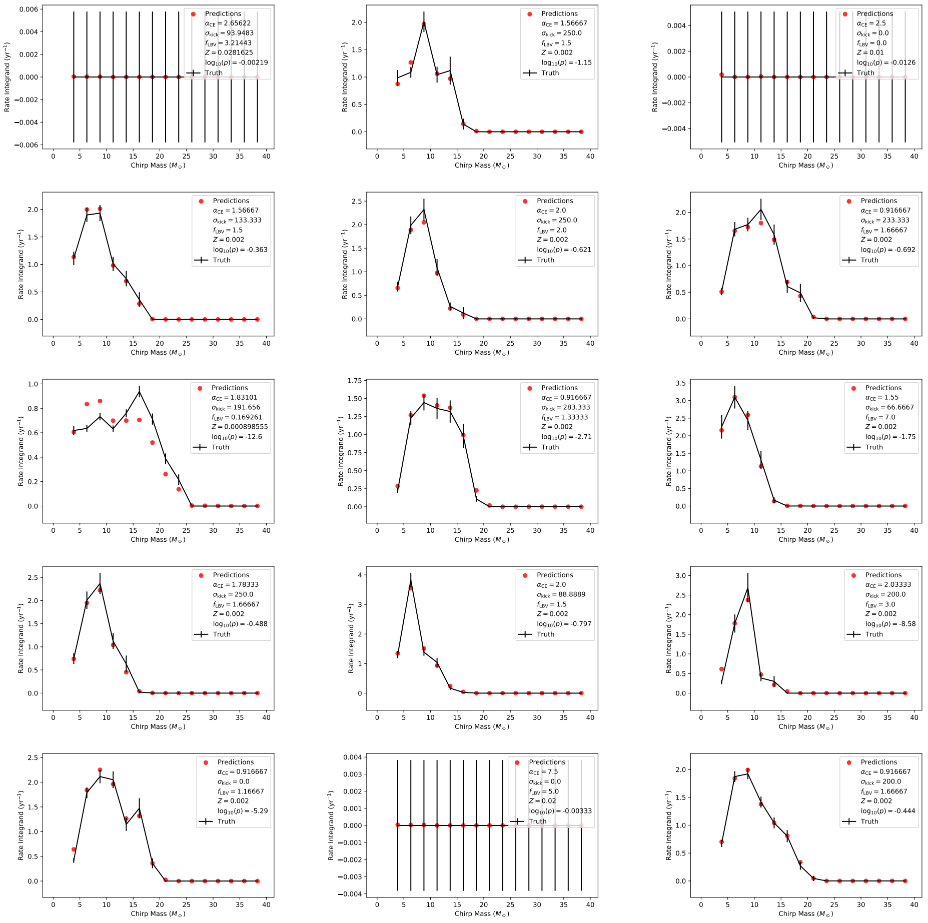


Figure 3.8: 15 random samples from the validation set built with the detection probability $P_{\rm det}$ calibrated to design sensitivity and only considering systems from a single redshift bin (to compare to figure 3.3, plotted in black. The random forest emulator predictions are plotted in red.

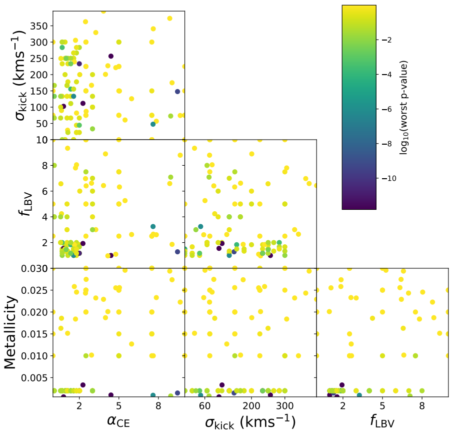


Figure 3.9: The p-value for the most poorly predicted bin at each point in population parameter space by the random forest process emulator

100 metallicities, and again used the dataset used for figure 2.3 to compare the true output of `COMPAS` to the predictions of the emulator.

We show the design sensitivity plot in figure 3.10 and for O1 sensitivity in figure 3.11. In both cases the emulator performs well. We also plot the distribution of expected observed chirp mass distributions, by computing the 90% confidence interval for Poisson draws about the emulator model prediction.

Since the random forest emulator performs so well, we chose to use this emulator instead of the Gaussian process emulator, to continue the work and make inferences about the population parameters. We leave a more detailed discussion of why the random forest emulator performs well to section 3.10.

3.7 Inference

The overall goal of building the emulator is to be able to efficiently compare the chirp masses of observed binary black hole coalescences to the predictions of `COMPAS`. In order to do this we must write down a posterior distribution function. We use

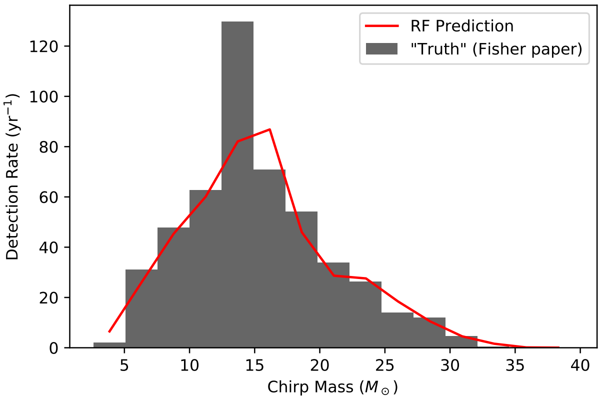


Figure 3.10: The result of using the random forest emulator, calibrated to design sensitivity, to predict a grid of integrands in metallicity and redshift to compute the integral in equation3.8. The prediction is in solid red, and is compared to the data used for figure 2.3 in grey. the 95% confidence interval on the expected Poisson fluctuations in the detected rate are in pale red

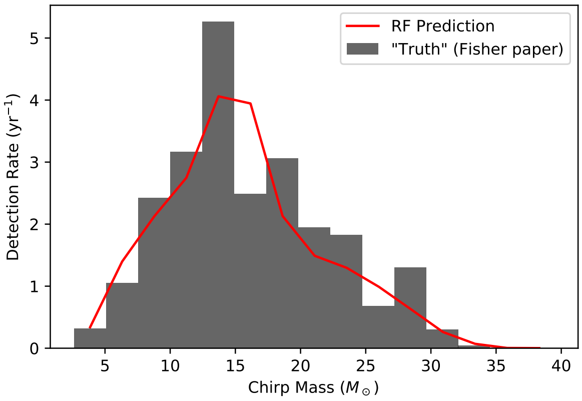


Figure 3.11: The result of using the random forest emulator, calibrated to O1 sensitivity, to predict a grid of integrands in metallicity and redshift to compute the integral in equation3.8. The prediction is in solid red, and is compared to the data used for figure 2.3 in grey. the 95% confidence interval on the expected Poisson fluctuations in the detected rate are in pale red

the same likelihood that was introduced in depth in section 2.6.2, which we will summarise again here. The likelihood function is made up of two terms.

$$\begin{aligned}\log(\mathcal{L}\left(\mathcal{D}|\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}}\right)) &= \log\mathcal{L}\left(N_{\mathrm{obs}}|\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}},t_{\mathrm{obs}}\right) \\ &+ \log\mathcal{L}\left(\{\mathcal{M}\}|\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}}\right).\end{aligned} \tag{3.15}$$

We model the rate at which binary black holes arrive in the detector as a Poisson process with model parameter ν, so that the likelihood of observing N_{obs} events during t_{obs} time spent observing is given by

$$\begin{aligned}\log\mathcal{L}\left(N_{\mathrm{obs}}|\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}},t_{\mathrm{obs}}\right) &= N_{\mathrm{obs}}\log(\nu(\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}})t_{\mathrm{obs}}) \\ &- \nu(\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}})t_{\mathrm{obs}} - \log(N_{\mathrm{obs}}!).\end{aligned} \tag{3.16}$$

For the binned rate distributions, we use the multinomial likelihood function. If we have a set of observed chirp masses $\{\mathcal{M}\}$, then the multinomial likelihood is given by

$$\log\mathcal{L}\left(\{\mathcal{M}\}|\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}}\right) = \log(N_{\mathrm{obs}}!)+\sum_{k}^{K}\left(c_k\log(p_k(\alpha_{\mathrm{CE}},\sigma_{\mathrm{kick}},f_{\mathrm{LBV}})) - \log(c_k!)\right), \tag{3.17}$$

where K is the number of chirp mass bins, c_k is the number of observed systems falling into the k-th bin with $\sum_k c_k = N_{\mathrm{obs}}$, and p_k is the probability predicted by the emulator that a system falls into the k^{th} bin.

We choose a uniform prior spanning the whole volume covered by the training examples, so that the prior $\pi(\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}})$ is

$$\log\left(\pi(\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}})\right) = \begin{cases} -\infty & \text{if } 0 > \alpha_{\text{CE}} > 10 \\ -\infty & \text{if } 0 > \sigma_{\text{kick}} > 400 \\ -\infty & \text{if } 0 > f_{\text{LBV}} > 10 \\ 0 & \text{otherwise} \end{cases} . \tag{3.18}$$

The `COMPAS` posterior $P(\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}}|\mathcal{D})$ can then be computed from the product of the prior and likelihood. Since for a fixed set of observations, the evidence term doesn't change, we do not need to compute it in this context. We can therefore simply use the emulator to compute

$$P(\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}|\mathcal{D}}) \propto \pi(\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}}) \cdot \mathcal{L}\left(\mathcal{D}|\alpha_{\text{CE}}, \sigma_{\text{kick}}, f_{\text{LBV}}\right). \tag{3.19}$$

3.8 Posterior Predictive Checking

In order to verify that the emulator is working correctly in the context of inference on the population parameters, we created mock datasets using known combinations of population parameters, in order to see if we could correctly infer them again using the emulator.

We chose three random points in population parameter space, and used the design sensitivity emulator to generate a 'true' detection rate distribution at that point. We computed the total detection rate at that point by adding up the heights of all bins, and then used this total rate as the rate parameter to draw a poisson distributed random number, representing the number of detections to be made in a year. We then drew this many chirp masses, such that they would be distributed according to the shape of the chirp mass distribution for those parameters. These were then our mock dataset $\mathcal{D}$.

We then used the `emcee`[3] implementation of the affine invariant ensemble sampling algorithm described in section 1.2.3 to sample from the posterior in equation 3.19, using the mock observations. We ran the sampler for 7000 steps, after which we visually inspected the evolution of each chain's likelihood and the mean path of the ensemble through population parameters in order to verify that the sampler had 'burned-in'.

We took the final 3000 iterations for each chain in the ensemble and computed its autocorrelation length in log-likelihood using the `acor` package [4], and thinned the chain by this length (i.e., only keeping samples which were at least one autocorrelation length apart) to obtain a set of independent samples from the posterior distribution.

For each of figures 3.12, 3.13 and 3.14 we plot both the univariate and bivariate projections of the sampled posterior distributions, with the 'true' population parameter values from that mock dataset overplotted. We also fairly drew 100 samples from this posterior distribution and used the emulator to compute the shape of the chirp mass distribution, allowing us to produce credible intervals. In all three cases, the true parameter values are well recovered, although the resultant projections of the posterior distribution show some multimodality in some dimensions.

It is possible that this multimodality comes from insufficient sampling of the posterior distribution, although every effort was made to ensure that the sampling was successful. The chirp mass distribution posteriors indicate that this multimodality may indicate a degeneracy between the shape of the chirp mass distribution across different regions of population parameter space. We will discuss this degeneracy in more detail in section 3.10.

[3] https://github.com/dfm/emcee

[4] https://github.com/dfm/acor

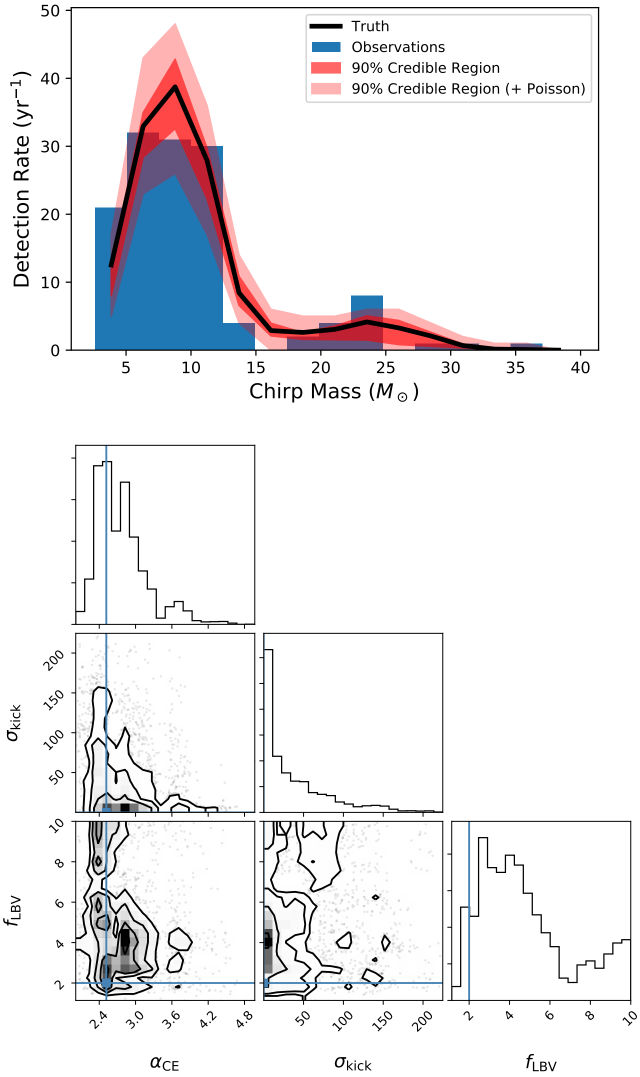


Figure 3.12: The results of agnostically inferring the shape of the chirp mass detection rate distribution for $\alpha_{\rm CE} = 2.52$, $\sigma_{\rm kick} = 0$ kms^{-1} and $f_{\rm LBV} = 2$. The mock dataset represented a hypothetical one year of observations, with 135 detections. The top panel shows this dataset, with the true distribution they were drawn from and the inferred 90% credible region in dark red. The lighter red distribution incorporates the region that could be expected to be obsereved, by adding Poisson uncertainty to the 90% region. The bottom panel shows univariate and bivariate marginalised projections of the posterior distribution, with the true population parameter values indicated in blue.

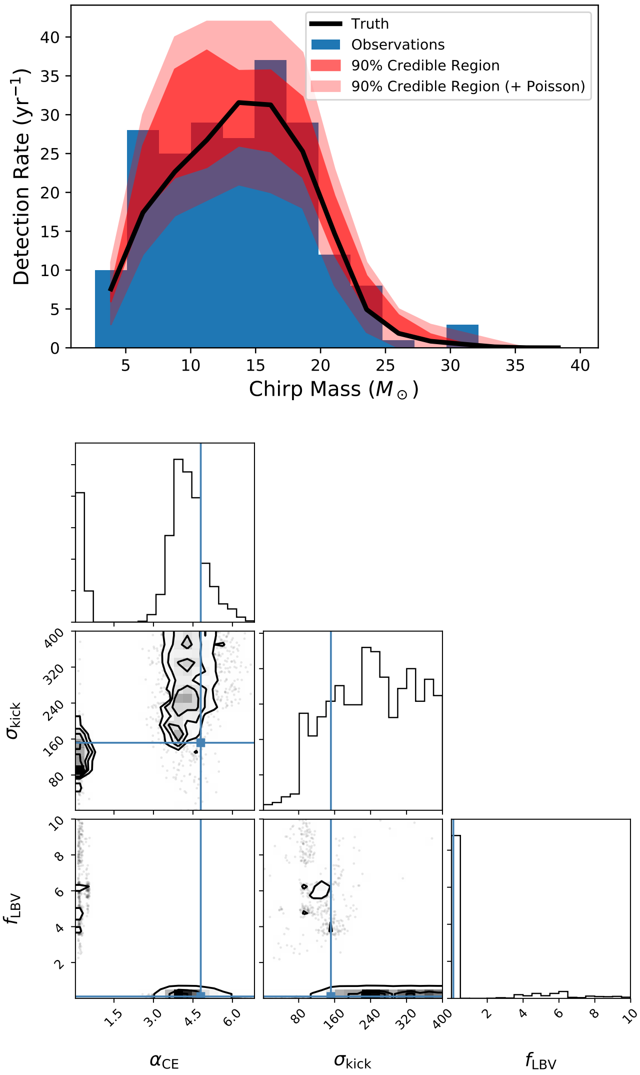


Figure 3.13: The results of agnostically inferring the shape of the chirp mass detection rate distribution for $\alpha_{\rm CE} = 4.8$, $\sigma_{\rm kick} = 152$ kms^{-1} and $f_{\rm LBV} = 0.11$. The mock dataset represented a hypothetical one year of observations, with 209 detections. The top panel shows this dataset, with the true distribution they were drawn from and the inferred 90% credible region in dark red. The lighter red distribution incorporates the region that could be expected to be obsereved, by adding Poisson uncertainty to the 90% region. The bottom panel shows univariate and bivariate marginalised projections of the posterior distribution, with the true population parameter values indicated in blue.

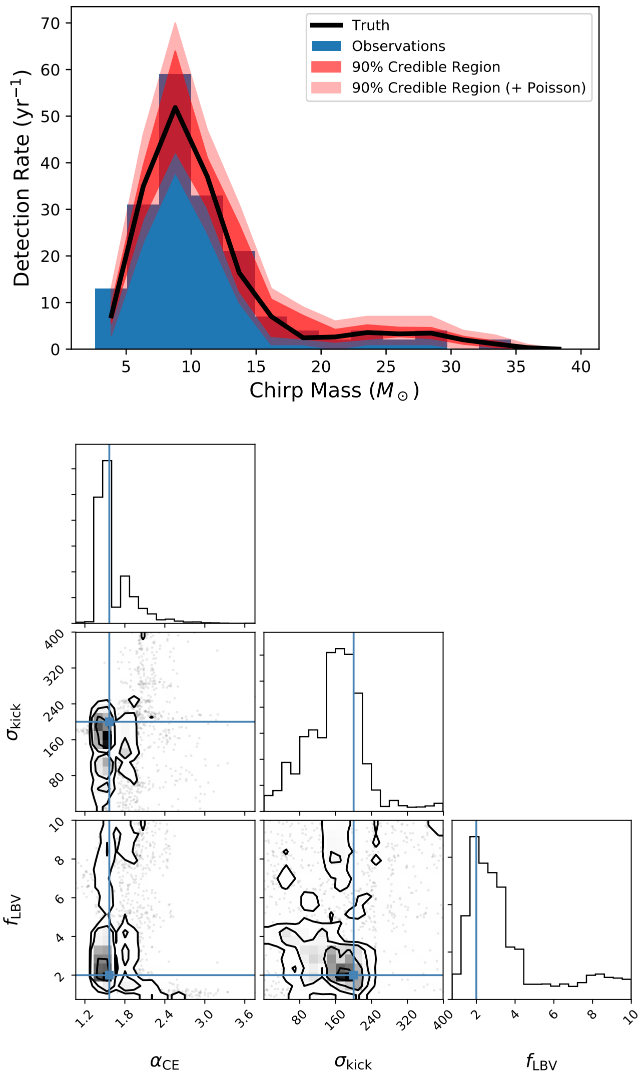


Figure 3.14: The results of agnostically inferring the shape of the chirp mass detection rate distribution for $\alpha_{\rm CE} = 1.57$, $\sigma_{\rm kick} = 200$ kms^{-1} and $f_{\rm LBV} = 2$. The mock dataset represented a hypothetical one year of observations, with 182 detections. The top panel shows this dataset, with the true distribution they were drawn from and the inferred 90% credible region in dark red. The lighter red distribution incorporates the region that could be expected to be obsereved, by adding Poisson uncertainty to the 90% region. The bottom panel shows univariate and bivariate marginalised projections of the posterior distribution, with the true population parameter values indicated in blue.

Event	Chirp Mass	Observing Run	Reference
GW150914	$28.2^{+1.8}_{-1.7}$	O1	[Abbott et al., 2016b,e]
GW151226	$8.9^{+0.3}_{-0.3}$	O1	[Abbott et al., 2016f,e]
GW170104	$21.1^{+2.4}_{-2.7}$	O2	[Abbott et al., 2017d]
GW170608	$7.9^{+0.2}_{-0.2}$	O2	[Abbott et al., 2017c]
GW170814	$24.1^{+1.4}_{-1.1}$	O2	[Abbott et al., 2017e]

Table 3.1: The chirp masses of the gravitational-wave detections of binary black hole coalescences made by advanced LIGO and advanced Virgo so far.

3.9 Gravitational-wave events so far

Since we have demonstrated we are able to successfully recover posteriors which are consistent with the truth, we can finally use the emulator to explore the posterior distribution for the 5 binary black hole coalescences detected so far. In table 3.1 we list these detections with their chirp masses, and which aLIGO observing run they were detected in. Whilst the detector sensitivity in the second observing (O2) run was slightly better than in the first (O1), they are sufficiently similar that we were able to simplify our analysis by only using a single emulator built at O1 sensitivity.

Due to time constraints, we did not take into account measurement uncertainty for our analysis, although this will certainly be the focus of future work. We used the central point estimates for the chirp masses of each event and ran an ensemble sampler for 7000 steps, again using visual inspection of the chains to ensure that they had burned in, and thinning each chain in the ensemble by the autocorrelation length of its log-likelihood evaluations. The results of this inference are displayed in figure 3.15, with the `COMPAS` fiducial population parameter values overplotted. It is clear that the inferences are less precise due to the lack of data, however the posterior distributions appear consistent with the fiducial model in this plot.

However, when drawing samples from the posterior and plotting the emulator's predictions of the detection rate distribution in order to build its 90% credible region shows a clear discrepancy between the fiducial model and the inferred posterior.

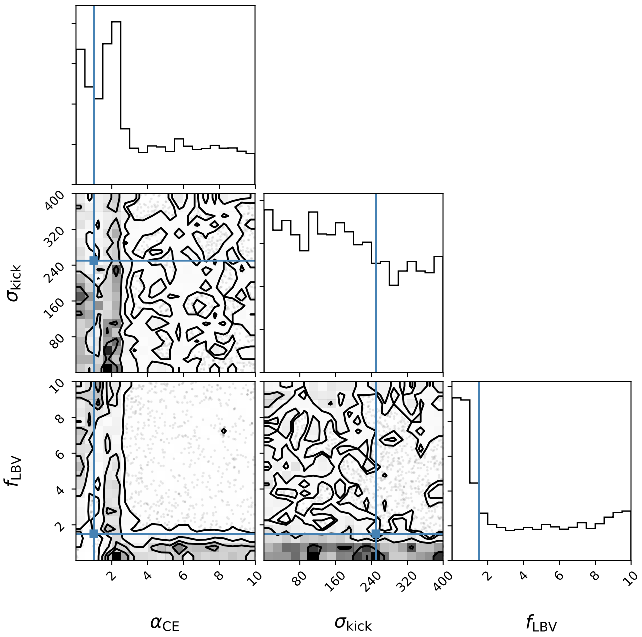


Figure 3.15: Univariate and bivariate projections of the posterior distribution of the COMPAS population parameters inferred from the gravitational wave events so far (table 3.1). The blue lines show the values of the population parameters in the COMPAS fiducial model.

This is displayed in figure 3.16. This plot looks to show detection rates which are too low to be consistent with the data, however when examing the posterior on te overall detection rate (by adding the rates from each bin for each detection rate distribution posterior), the overall detection rate we have observed is just about consistent, although it suggests we may have been lucky with how many events we have detected so far. This posterior is plotted in figure 3.17.

Finally, since we expect the detector sensitivity to improve until it reaches design sensitivity, we can use the emulator built for design sensitivity to plot posteriors for the expected detection rate and distribution once the detectors reach this sensitivity. We compute these posteriors in much the same way as before, by drawing posterior samples of the population parameters from the distribution shown in figure 3.15 and then using the design sensitivity emulator to predict the resultant distribution shapes. The design sensitivity detection rate distribution posterior is shown in figure 3.18.

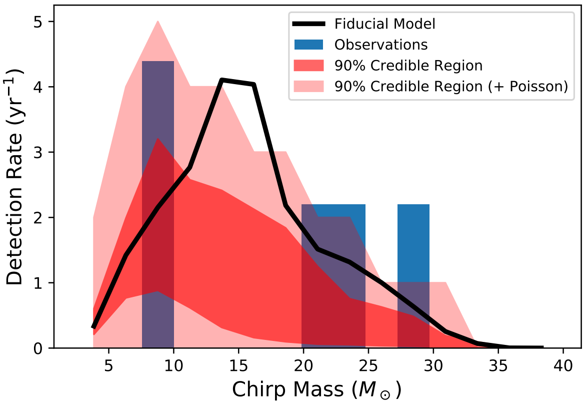


Figure 3.16: The inferred 90% credible region for the chirp mass detection rate distribution given the binary black hole coalescences observed so far, for a detector sensitivity matching the first observing run. The rates of these events are plotted in blue and the predictions of the COMPAS fiducial model are plotted in black.

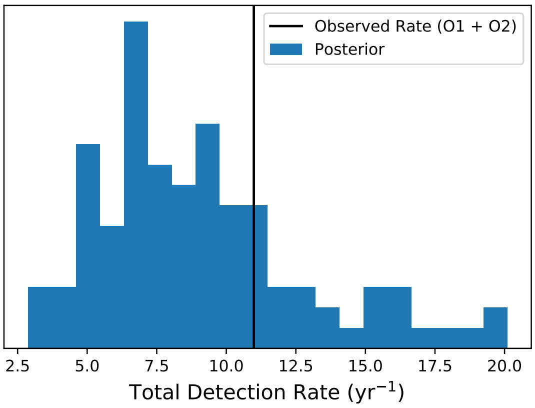


Figure 3.17: The posterior distribution of the overall event detection rate for the O1 and O2 observing runs, inferred from the 5 gravitational events detected so far. The actual detection rate is shown by the vertical black line.

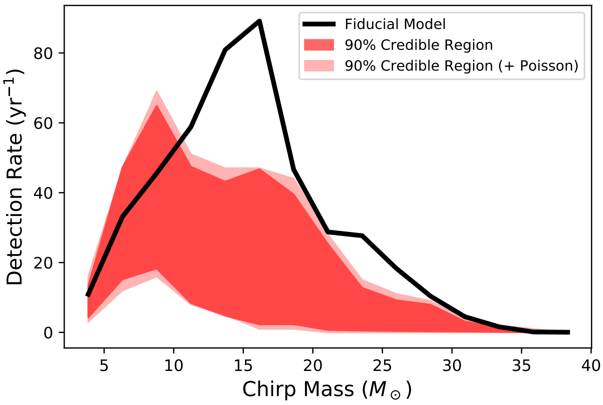


Figure 3.18: The inferred 90% credible region for the chirp mass detection rate distribution given the binary black hole coalescences observed so far, for the projected 'design' sensitivity of the detector. The distribution predicted by the fiducial model is shown in black.

3.10 Summary and Discussion

3.10.1 Chapter Summary

In this chapter we have described how we used two regression techniques to emulate the behaviour of `COMPAS` in predicting populations of merging binary black holes, as a function of the three population parameters; common envelope efficiency α_{CE}, supernova kick velocity dispersion σ_{kick} and the mass loss rate during the luminous blue variable phase f_{LBV}. We described how we designed our training examples to represent the integrands of equation 3.8, which then combine to form the rate at which detections will be made by the current generation of ground based gravitational wave detectors on earth.

We then discussed the question of experimental design, and how we combined simulations from other projects within our group which used the same version of `COMPAS` with new simulations using a latin hypercube experimental design. We described our Gaussian process emulator, which combines approximately 70 independent Gaussian processes, coupled with a random forest classifier, trained using appropriately transformed and scaled representations of the training set to predict the integrands μ^*.

We tested the emulator by using it to predict unseen training examples from a validation set, which showed the Gaussian process emulator to perform inadequately. We then changed algorithms to create a random forest emulator, which performed very well in comparison to the Gaussian process emulator. We then reintroduced the `COMPAS` likelihood function, and used it to demonstrate the efficacy of the emulator to infer the posteriors over population parameters by recovering them from three different mock data sets.

We finally used the random forest emulator to infer the posterior distributions of the population parameters for the 5 binary black hole gravitational-wave events detected so far, aquiring a posterior distribution with support spanning the whole prior volume.

3.10.2 Discussion of Gaussian Processes

The poor performance of the Gaussian process emulator was surprising. All of the preliminary investigations we conducted, including those presented in Barrett et al. [2017b], indicated that Gaussian process emulators would be an ideal tool for our problem. Moreover, since the number of training examples was relatively modest and we expected the chirp mass distribution to change smoothly as a function of the three population parameters explored, Gaussian process regression seemed a natural choice. We had also expected the response to be stationary with respect to the metric trained in the Gaussian process kernel. Here we present our hypotheses for why Gaussian processes may have performed badly.

Hyperparameter Optimisation

It is possible that the hyperparameter optimisation problem was not approached in the correct way. For our approach, each GP in the ensemble was optimised independently using a gradient based optimisation routine. Whilst each optimiser was restarted multiple times, with the best result kept in an attempt to avoid local maxima in the likelihood surface, it is possible that this method was insufficient.

In an ideal world, the hyperparameters of every GP in the ensemble would be optimised simultaneously. This simply was not feasible, since our ensemble consisted of ~ 70 gaussian processes, each with its own 4 kernel hyperparameters to be trained, for a potential ~ 280 dimensional optimisation problem. Not only can optimisation problems of this magnitude be intrinsically challenging, the code libraries used for the Gaussian process regression algorithm were not suited to dealing with problems of this type. We believe that with a library designed to operate with ensembles of Gaussian processes, this optimisation problem could potentially by addressed, however it was not possible to robustly develop such a library under the time constraints of a PhD.

We also took the approach of using a singular value decomposition to decorrelate the dimensions of the training set, as well as perform a dimensionality reduction. However, it is possible that by leaving the training set in the 'bin basis', correlations between rates in neighbouring bins could have been utilised to aid with hyperparameter optimisations. This would have involved more careful work in choosing the form of the kernel function, and would have yielded a much higher dimensional optimisation problem without the dimensionality reduction provided by the SVDs.

Choice of Kernel

As alluded to in the previous section, it is possible that with a more careful choice of kernel function the Gaussian process emulator may have performed better. First, a choice of kernel that made a stronger use of the expected corellation between output dimensions may have been fruitful.

The results from the random forest emulator suggest that there may have been some degeneracy in the model predictions. The basic choices of kernel employed in this thesis (squared exponential and matern-3/2) do not always perform well for periodic or degenerate response surfaces. If more work was put into the choice of kernel, the Gaussian processes may have been more successful.

The Training Set

The training set was built using a combination of binaries simulated specifically for this project together with simulations from other projects and our preliminary investigations. This meant that the training set was heterogeneously distributed across the population parameter space. This can be seen in the overdense regions of figure 3.2. The simulations included from other projects were typically only simulated at a single metallicity, meaning that almost 1/3 of the training set was simulated at a single metallicity ($Z = 0.002$). Whilst in principle, more information should always be beneficial to the training, it is possible that this heterogeneity may have contributed to the difficulty of the hyperparameter optimisation problem.

The inclusion of metallicity as an input population parameter may have also been problematic. We discussed in section 2.8 the issue of a 'piling-up' of systems simulated at a single metallicity. When stars enter a luminous blue variable phase, their final masses are a function of the metallicity of the system. This means that when a COMPAS population is simulated at a single metallicity, many massive systems will have the same mass at some point in their evolution. This leads to many of them becoming double compact objects with very similar component and thus chirp masses. This causes the spikes seen in figures 3.7 and 2.3. These spikes are typically much narrower than the width of a bin. The implication of this is that as we move through metallicity space, one of these spikes may cross from one bin into the neighbouring one, causing a discontinuous jump in the response surface.

It was one of our fundamental assumptions when choosing Gaussian process regression that the response surface was smooth with respect to the population parameters. If this assumption does not hold for metallicity, then it could be that Gaussian processes are ill suited for this problen. It is unclear whether these discontinuities transfer in a detrimental way to the reduced dimensionality basis after the singular value decomposition, or whether these metallicity spikes could be dealt with before training, either within COMPAS or in post processing. We believe both of these possibilities warrant further investigation.

Computational Considerations

Gaussian process regression involves storing and inverting an $N_{\text{train}} \times N_{\text{train}}$ covariance matrix for each new set of predicted points. With our training set, $N_{\text{train}} = 2352$, and we have ~ 70 independent Gaussian processes working together, in order to emulate at 8 redshifts. For 32-bit arithmetic, this involves ∼GB of memory to be available for prediction. Moreover, each prediction potentially involves an inversion of this $N_{\text{train}} \times N_{\text{train}}$ matrix, which can be a computationally expensive task.

The speed of inversion can be improved by precomputing factors of the matrix (much like the factorisation used when deriving the update equations in appendix A.1), since much of the matrix involves covariances between training example repsonses, which do not change after training is complete. However, this comes with an increased cost in memory consumption, and a significant expense each time a Gaussian process object must be constructed. this limits the possibility to construct these objects on the fly. Furthermore, if the number of features to predict increases, this also increases the size of the ensemble, creating more computational difficulties.

The existing Gaussian process libraries are not designed with these kind of 'ensemble' applications in mind. It could be possible to implement code which is specialised (in terms of computational resources) to problems such as these, but not within the scope of this project.

3.10.3 Discussion of Random Forests

In contrast to the Gaussian process emulator, the random forest emulator performed surprisingly well on this problem. This is partly because the training set is augmented by the inclusion of populations from over 100 redshifts. However, even when tested at a single redshift the random forest emulator performed well, both on predicting examples from the validation set and the full integral prediction of equation 3.8, as demonstrated in section 3.6.

A potential reason why the random forest emulator worked well simply comes down to the small amount of tuning required to make the algorithm work. We discussed in section 3.10.2 that the full ensemble of Gaussian processes involves $\sim$ 280 model hyperparameters which need to be jointly optimised. For random forests, there are far fewer choices to make.

The issue of system 'pile-ups' due to single metallicity populations discussed in section 3.10.2 also does not affect the performance of the random forest emulator, since decision trees simply partition the input space, without any assumptions of smoothness. The averaging of estimates from the random forest's constituent decision trees may be ideally suited to dealing with the issue of spikes in these distributions, since this averaging will tend to smooth out these spikes, as long as metallicity has been sufficiently sampled.

The downside of random forest emulators is the lack of a formal uncertainty quantification on its predictions. Whilst the distribution of predictions made by each of the constituent decision trees could be used, they do not give a fair nor statistically meaningful impression of the error. This does not compare to the formal joint distribution given by the GPs.

3.10.4 Discussion of Results

The results of using the random forest to infer posteriors on the population parameters are highly encouraging. The results of the posterior predictive checking show that posteriors which are consistent with the parameters of a mock data set can be recovered, and also correspond to credible regions on the chirp mass rate distribution which match the truth well.

The surprising result is that the chirp mass distributions appear to be degenerate between different areas of population parameter space, leading to multimodal posteriors (see e.g., figure 3.12), and posteriors with support spanning the entire prior volume. This has the unfortunate consequence that it is not possible to provide point estimates for the population parameter values. It is possible that when performing inference with more data, this degeneracy will be broken, however it seems more likely that a larger number of features will have to be used in the training set. This will involve either building training set with more bins in the chirp mass histogram, or including predictions of more astrophysical observables than just the chirp mass.

From a physical perspective, it is interesting that the inferred chirp mass rate distributions have such a small credible region, despite the posterior having support across the entire prior distribution, as well as this distribution being highly inconsistent with the `COMPAS` fiducial model. The `COMPAS` fiducial model is designed to be a representation of the rapid population synthesis community's hypothesis for the physics underlying binary evolution. The methods discussed in this chapter provide strong evidence that this model is not the correct one.

3.10.5 Additional Comments

It is the opinion of the author that none of the issues encountered with the Gaussian process emulator are insurmountable. In principle Gaussian processes are still the 'correct' choice of machine learning algorithm for emulating `COMPAS`. However, with the large number of assumptions and concessions involved in rapid population synthesis models, there is the question of pragmatism over idealism. The random forest performs very well, and the lack of robust uncertainty quantification is a small price to pay.

Chapter 4

CARMA Processes

4.1 Introduction

A continuous autoregressive moving average CARMA(p, q) process is defined by the solution to the stochastic differential equation

$$\left[\prod_{j=0}^{p-1}\left(\frac{d}{dt} - r_j\right)\right](y(t) - \mu) = \left[\prod_{i=1}^{q}\left(\frac{d}{dt} - b_i\right)\right]\eta(t). \tag{4.1}$$

Derivatives of the centred observations $(y(t) - \mu)$ form the autoregressive part of the differential equation, characterised by a set of p autoregressive roots, $\{r_j\}$ with units of inverse time. The moving average part is made up of derivatives of a normally distributed white noise process $\eta(t)$ with variance σ^2, characterised by a set of q moving average roots $\{b_i\}$, which also have units of inverse time. In order for the solution of this differential equation to have a finite autocorrelation function, we require that $q < p$, and also that the real part of the roots r_i and b_i is negative.

Because the forcing function has Gaussian statistics and the ODE is linear, the process $y(t)$ is a Gaussian process. Because the ODE is time-invariant, the process is stationary, with a covariance function we derive in Section 4.3.3. The power spectrum of the process can also be inferred from Eq. 4.1 (see Section 4.4).

Suppose now that we have a data set of observations of this process at specified times, t_i, $i = 1, \ldots, N_{\text{obs}}$:

$$y_i \equiv y\left(t_i\right), \quad i = 1, \ldots, N_{\text{obs}}. \tag{4.2}$$

For the moment we assume that the measurements of the y_i are perfect, but in Section 4.3.4 we discuss how to incorporate uncertainty in the measurements of the process values. Given this model it is straightforward to write down a naive likelihood function in terms of the full covariance between measured values $\mathbf{C_{ij}} = \langle y(t_i)y(t_j)\rangle$ and the vector of measurement values $\mathbf{y}$

$$\log(\mathcal{L}) \propto -\frac{1}{2}\left(\log|\mathbf{C}| + \mathbf{y}^T \cdot \mathbf{C}^{-1} \cdot \mathbf{y}\right). \tag{4.3}$$

If N_{obs} is the number of measurements, then since the naive likelihood above involves inverting $\mathbf{C}$, a rank N_{obs} matrix, evaluating it can be achieved using $\mathcal{O}\left(N_{\text{obs}}^3\right)$ computations. However, we will show below that a transformation

$$x(t_k) = \begin{cases} y(t_k) & \text{if } k < p \\ y(t_k) - \sum_{i=k-p}^{k-1} \alpha_{i-(k-p)}^k y(t_i) & \text{if } k \geq p \end{cases}, \tag{4.4}$$

where the α factors satisfy

$$y_h(t_k) - \sum_{i=k-p}^{k-1} \alpha_{i-(k-p)}^k y_h(t_i) = 0, \tag{4.5}$$

where y_h solves the homogeneous part of Eq. (4.1),

$$\left[\prod_{j=0}^{p-1}\left(\frac{d}{dt} - r_j\right)\right]\left(y_h(t) - \mu\right) = 0, \tag{4.6}$$

renders the covariance matrix $\mathbf{C}'$ banded, with only a central band of width $2p-1$ diagonals and zeros elsewhere. Note that the α factors depend only on the sample times, t_i. The likelihood calculation then becomes

$$\log(\mathcal{L}) \propto -\frac{1}{2}\left(\log|\mathbf{C}'| + \mathbf{x}^T \cdot \mathbf{C}'^{-1} \cdot \mathbf{x}\right); \tag{4.7}$$

no Jacobian factor is present because the determinant of the transformation matrix is one. There are well established algorithms, which are able to invert banded matrices in $O(n)$ computations, and these algorithms can be parallelised. We show that this transformation bands the covariance matrix using the Green's function solution to equation (4.1).

4.2 Green's Function Representation

A Green's function is the impulse response solution of an ordinary differential equation, meaning that if, for a linear differential operator satisfying $\hat{D}f(x) = a(x)$ then the Green's function $G(x;\xi)$ of the system defined by

$$f(x) = \int_{-\infty}^{x} G(x;\xi)a(\xi)\mathrm{d}\xi \tag{4.8}$$

satisfies the impulse response differential equation $\hat{D}G(x;\xi) = \delta(x-\xi)$. We wish to determine the Green's function for the CARMA differential equation. For a set of centred observations this means solving

$$\left[\prod_{j=0}^{p-1}\left(\frac{d}{dt} - r_j\right)\right] G(t;\xi) = \delta(t-\xi). \tag{4.9}$$

Since the r_j don't depend on time, the solution to this differential equation will take the form

$$G(t;\xi) = \sum_{k=0}^{p-1} A_k e^{r_k(t-\xi)} \tag{4.10}$$

Where the A_k are constant with respect to time, except in the transition $t < \xi \rightarrow t > \xi$. We have the freedom to choose ξ so that

$$G(t;\xi) = \begin{cases} 0 & t < \xi \\ \sum_{k=0}^{p-1} A_k e^{r_k(t-\xi)} & t > \xi \end{cases}. \tag{4.11}$$

We can expect that all but the highest present derivative of the Green's function are continuous at $t = \xi$. This implies that for some small neighbourhood $\xi \pm \epsilon$ the highest present derivative's contribution to the integral can be computed

$$\int_{\xi-\epsilon}^{\xi+\epsilon} \mathrm{d}t \, \frac{\mathrm{d}^p G(t;\xi)}{\mathrm{d}t^p} = \int_{\xi-\epsilon}^{\xi+\epsilon} \mathrm{d}t \, \delta(t-\xi) = 1, \tag{4.12}$$

so that

$$\left.\frac{\mathrm{d}^{(p-1)} G(t;\xi)}{\mathrm{d}t^{p-1}}\right|_{\xi^+} - \left.\frac{\mathrm{d}^{(p-1)} G(t;\xi)}{\mathrm{d}t^{p-1}}\right|_{\xi^-} = 1. \tag{4.13}$$

Our choice of ξ means that the second term is zero. We therefore have

$$\left.\frac{\mathrm{d}^{(p-1)} G(t;\epsilon^+)}{\mathrm{d}t^{p-1}}\right|_{t=\xi} = \sum_{k=0}^{p-1} A_k r_k^{p-1} e^{r_k \xi} = 1. \tag{4.14}$$

Since we know that the other derivatives do not contribute to the integral in equation (4.12), we can use these to write similar expressions to the ones above, so that we have a further $p-1$ simultaneous equations for $0 \leq x < p$

$$\sum_{k=0}^{p-1} A_k {r_k}^x e^{r_k \xi} = 0. \tag{4.15}$$

So that overall we have a sufficient set of linear simultaneous equations that we can solve for the coefficients A_k. In practise, we make this more numerically stable by changing the definition of the A_k, so that $A_k \rightarrow A_k e^{r_k \xi}$, so that finally the system of equations we solve becomes

$$\sum_{k=0}^{p-1} A_k {r_k}^x = \delta_{xp}, \tag{4.16}$$

for $0 \leq x \leq p$ and δ_{xp} is the Kroenecker delta.

The Green's function can be used to express the solution to the stochastic CARMA differential equation.

$$y(t) = \int_{-\infty}^{t} \mathrm{d}\xi \; G(t;\xi) \left[\prod_{i=1}^{q} \left(\frac{\mathrm{d}}{\mathrm{d}\xi} - b_i \right) \right] \eta(\xi). \tag{4.17}$$

It is useful to expand the product into its characteristic polynomial

$$y(t) = \int_{-\infty}^{t} \mathrm{d}\xi \; G(t;\xi) \sum_{i=0}^{q} \left[\omega_j \frac{\mathrm{d}^i \eta(\xi)}{\mathrm{d}\xi^i} \right], \tag{4.18}$$

where the ω_i represent the coefficients of the characteristic polynomial. Integrating by parts it can be seen that

$$y(t) = \left[G(t;\xi) \omega_i \frac{\mathrm{d}^{i-1} \eta(\xi)}{\mathrm{d}\xi^{i-1}} \right]_{\xi=-\infty}^{\xi=t} - \int_{-\infty}^{t} \mathrm{d}\xi \; \frac{\mathrm{d}G(t;\xi)}{\mathrm{d}\xi} \sum_{i=0}^{q} \left[\omega_j \frac{\mathrm{d}^i \eta(\xi)}{\mathrm{d}\xi^i} \right]. \tag{4.19}$$

By examining the expression for the Green's function in equation (4.11), and noting that we enforce the roots r_k to be negative, and that equation 4.16 implies that $\sum_{k=0}^{p-1} A_k = 0$, we see that the first term in this expression is zero. Continuing to integrate by parts we eventually see that

$$y(t) = \int_{-\infty}^{t} \mathrm{d}\xi \; \eta(\xi) \sum_{i=0}^{q} \frac{\mathrm{d}^i G(t;\xi)}{\mathrm{d}\xi^i} (-1)^i \omega_i. \tag{4.20}$$

Here it becomes mathematically evident why we require $q < p$, since p^{th} order derivatives of the Green's function are discontinuous by definition. The i^{th} derivative of the Green's function is

$$\frac{\mathrm{d}^i G(t;\xi)}{\mathrm{d}\xi^i} = \sum_{k=0}^{p-1} A_k (-1)^i r_k^i e^{r_k (t-\xi)}, \tag{4.21}$$

so the expression for $y(t)$ becomes

$$y(t) = \int_{-\infty}^{t} d\xi \; \eta(\xi) \sum_{j=0}^{q} \sum_{k=0}^{p-1} \left[r_k^j \omega_j A_k e^{r_k(t-\xi)} \right]. \tag{4.22}$$

We hereby introduce the notation

$$g_{pq}^{i} = \eta(\xi) \sum_{j=0}^{q} \sum_{k=0}^{p-1} \left[r_k^j \omega_j A_k e^{r_k(t_i-\xi)} \right] \tag{4.23}$$

so that

$$y(t_k) = \int_{-\infty}^{t_k} d\xi \; g_{pq}^{k}. \tag{4.24}$$

4.2.1 $\mathcal{G}$ Notation

In order to simplify the algebra that follows this section we introduce a notation for expectations in terms of Green's functions

$$\mathcal{G}_{ab}^{\nu\mu} = \int_{t_a}^{t_b} d\xi \int_{t_a}^{t_b} d\lambda \left\langle g_{pq}^{\nu} g_{pq}^{\mu} \right\rangle. \tag{4.25}$$

By considering the known properties of the underlying noise terms $\eta(\xi)$ we can see that

$$\langle \eta(\xi)\eta(\lambda) \rangle = \sigma^2 \delta(\xi - \lambda), \tag{4.26}$$

and so it is straightforward to evaluate $\mathcal{G}$

$$\mathcal{G}_{ab}^{\nu\mu} = -\sigma^2 \left[\sum_{j,m=0}^{q} \sum_{k,n=0}^{p-1} \left(\frac{r_k^j r_n^m \omega_j \omega_m A_k A_n}{r_k + r_n} \; e^{r_k t_\nu + r_n t_\mu - \xi(r_k + r_n)} \right) \right]_{\xi=t_a}^{\xi=t_b}. \tag{4.27}$$

4.3 The Transformed Covariance Matrix

4.3.1 Isolating the Noise Terms

The autoregressive property of a CARMA model suggests that $y(t)$ can be written as a linear sum of p previous values of $y(t)$ and p noise terms. We make an ansatz that we express the Green's functions as linear sums of past values

$$G(t_k;\xi) = \sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} G(t_i;\xi). \tag{4.28}$$

It is important to see that if this is true for the $\{G(t_n)\}$ then it is also true for $\{g^n_{pq}\}$, since the map $G(t_n;\xi) \to g^n_{pq}$ is time independent. This means that we can write

$$g^k_{pq} = \sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} g^i_{pq}. \tag{4.29}$$

If we separate p regions of integration from the expression for y in equation 4.24,

$$y(t_k) = \int_{-\infty}^{t_{k-p}} d\xi \; g^k_{pq} + \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; g^k_{pq}, \tag{4.30}$$

we can substitute in the weighted sum of previous g^k_{pq}

$$y(t_k) = \left[\sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} \int_{-\infty}^{t_{k-p}} d\xi \; g^i_{pq} \right] + \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; g^k_{pq}. \tag{4.31}$$

Now, we notice that we can pull a term out of the sum

$$y(t_k) = \alpha^k_0 \int_{-\infty}^{t_{k-p}} d\xi \; g^{k-p}_{pq} \quad + \quad \left[\sum_{i=k-p+1}^{k-1} \alpha^k_{i-(k-p)} \int_{-\infty}^{t_{k-p}} d\xi \; g^i_{pq} \right] \tag{4.32}$$

$$+ \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; g^k_{pq}. \tag{4.33}$$

We can immediately identify the first term as $y(t_{k-p})$. Now we can't do this exactly again, since the integral in the brackets only runs up to t_{k-p}. However, we can add a suitably chosen zero to the expression so that we can pull another term out of the sum. If we use

$$0 = \sum_{i=k-p+1}^{k-1} \alpha^k_{i-(k-p)} \left[\int_{t_{k-p}}^{t_{k-p+1}} d\xi \; g^i_{pq} - \int_{t_{k-p}}^{t_{k-p+1}} d\xi \; g^i_{pq} \right], \tag{4.34}$$

then we can see that

$$\begin{aligned} y(t_k) &= \alpha^k_0 y(t_{k-p}) + \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; g^k_{pq} \\ &+ \sum_{i=k-p+1}^{k-1} \alpha^k_{i-(k-p)} \left[\int_{-\infty}^{t_{k-p}} d\xi \; g^i_{pq} + \int_{t_{k-p}}^{t_{k-p+1}} d\xi \; g^i_{pq} - \int_{t_{k-p}}^{t_{k-p+1}} d\xi \; g^i_{pq} \right], \end{aligned} \tag{4.35}$$

which extends the integral

$$\begin{aligned} y(t_k) = \alpha^k_0 y(t_{k-p}) &+ \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; g^k_{pq} \\ &+ \sum_{i=k-p+1}^{k-1} \alpha^k_{i-(k-p)} \left[\int_{-\infty}^{t_{k-p+1}} d\xi \; g^i_{pq} - \int_{t_{k-p}}^{t_{k-p+1}} d\xi \; g^i_{pq} \right]. \end{aligned} \tag{4.36}$$

We can repeat this process until no terms remain in the sum over i. By relabelling indices and rearranging we finally arrive at

$$y(t_k) = \sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} y(t_i) + \sum_{\mu=k-p}^{k-1} \int_{t_\mu}^{t_{\mu+1}} d\xi \; \left[g^k_{pq} - \sum_{i=\mu+1}^{k} \alpha^k_{i-(k-p)} g^i_{pq} \right], \tag{4.37}$$

so that in the end the value of $y(t)$ can be written as a weighted sum of previous p values of the time series (which have no stochastic terms), plus a collection of p 'noise' terms, for which we shall introduce the notation

$$\beta^k_j = \int_{t_j}^{t_{j+1}} d\xi \; \left[g^k_{pq} - \sum_{i=j+1}^{k} \alpha^k_{i-(k-p)} g^i_{pq} \right], \tag{4.38}$$

so that

$$y(t_k) = \sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} y(t_i) + \sum_{j=k-p}^{k-1} \beta^k_j. \tag{4.39}$$

The $\alpha^k_{i-(k-p)}$ can be straightforwardly computed from the time series by solving the set of simultaneous equations implied by equation 4.28. We now argue that by making the transformation from $y(t)$ to $x(t)$ as described in equation 4.4, which is effectively collecting all of the non-stochastic terms from the above, leaving only the sum of β^k_j, so that (for $k > p$)

$$x(t_k) = y(t_k) - \sum_{i=k-p}^{k-1} \alpha^k_{i-(k-p)} y(t_i) = \sum_{j=k-p}^{k-1} \beta^k_j, \tag{4.40}$$

then the covariances between the transformed $x(t_k)$ will be zero except for a small number of $\pm\,(p-1)$ neighbouring observations.

4.3.2 Direct Derivation of the Covariance Matrix

Now we wish to derive an expression for the covariance matrix $\langle x(t_i)x(t_j)\rangle$. We break up this problem into three sections; the case where $i, j \geq p$, the case where $i, j < p$, and the cases where $i < p \leq j$ or $j < p \leq i$.

The elements where $i, j \geq p$ can be calculated by considering cross-correlations in x

$$\langle x(t_k)x(t_m)\rangle = \sum_{i=k-p}^{k-1} \sum_{j=m-p}^{m-1} \left\langle \beta^k_i \beta^m_j \right\rangle . \tag{4.41}$$

By requiring stationarity, only terms where $i = j$ remain, which take the form

$$\begin{aligned}\left\langle \beta^k_j \beta^m_j \right\rangle = {} & \mathcal{G}^{km}_{j(j+1)} + \sum_{\nu=j+1}^{k} \sum_{\mu=j+1}^{m} \alpha^k_{\nu-(k-p)} \alpha^m_{\mu-(m-p)} \mathcal{G}^{\nu\mu}_{j(j+1)} && (4.42)\\ & - \sum_{\nu=j+1}^{k} \alpha^k_{\nu-(k-p)} \mathcal{G}^{\nu m}_{j(j+1)} - \sum_{\mu=j+1}^{m} \alpha^m_{\mu-(m-p)} \mathcal{G}^{k\mu}_{j(j+1)}, && (4.43)\end{aligned}$$

where we have employed the $\mathcal{G}$ notation introduced in equation 4.25. It is important to note that if $|k-m| \geq p$, then there are no terms for which $i = j$, and so the expectation is zero.

By the definition in equation 4.4, $y(t_i) = x(t_i)$ for $i < p$, so the elements of $\mathbf{C}'$

can be computed by directly evaluating covariances between the $y(t_i)$

$$\langle x(t_i)x(t_j)\rangle = \langle y(t_i)y(t_j)\rangle = \begin{cases} \mathcal{G}^{ij}_{(-\infty)(i)} & \text{if } i < j \\ \mathcal{G}^{ij}_{(-\infty)(j)} & \text{if } i > j \end{cases} . \tag{4.44}$$

Finally, for the final $p(p-1)$ elements for which $i < p \leq j$ or $j < p \leq i$, β^k_j takes place on the interval $[j, j+1]$, so it is the noise contributions that happens after the measurement $y(t_j)$, meaning that $\left\langle \beta^k_j y(t_m)\right\rangle = 0$ for all $m \leq j$. For the non zero terms, we find the expression

$$\langle y(t_m)x(t_k)\rangle = \sum_{j=k-p}^{k-1} \left[\mathcal{G}^{mk}_{j(j+1)} - \sum_{i=j+1}^{k} \alpha^k_{i-(k-p)} \mathcal{G}^{mi}_{j(j+1)} \right] . \tag{4.45}$$

4.3.3 Alternative Expression

In the previous section, we gave a detailed derivation of the elements of the covariance matrix resulting from the $y \to x$ transformation defined in equation 4.4. We made no assumptions about the final form of the covariance matrix, and demonstrated that it is banded. However, in practise a computational implementation is not straightforward. A more straightforward approach to calculating the elements of the banded covariance matrix, under the assumption that the result is banded, can be achieved by considering the matrix that represents the transformation of a times series from $y \to x$ space. By making some definitions of the edge cases of the weighting factors, so that $\alpha^n_p \equiv -1$, $\alpha^n_{-ve} \equiv 0$ and $\alpha^n_{>p} \equiv 0$, we can write down a matrix

$$A_{ij} = \begin{cases} \delta_{ij} & i, j < p \\ 0 & i < j \\ -\alpha^i_{j-i+p} & \text{otherwise} \end{cases} , \tag{4.46}$$

so that

$$x(t_i) = A_{ij} y(t_j) \tag{4.47}$$

$$C'_{ij} = A_{il} \left\langle y(t_l) y(t_k) \right\rangle A_{jk}, \tag{4.48}$$

where repeated indices are summed. By combining these expressions we can compute the elements of the banded covariance matrix C'_{ij}. We can also use the fact that the matrix is symmetric to only compute elements for which $i \leq j$ and setting $C'_{ji} = C'_{ij}$. Once again employing $\mathcal{G}$ notation, we see that

$$C'_{ij} = \begin{cases} \mathcal{G}^{ij}_{(-\infty)(i)} & i,j < p \\ -\sum_{k \leq j} \alpha^{j}_{k-j+p} \mathcal{G}^{ik}_{(-\infty)(i)} & i < p, j \geq p \\ \sum_{l \leq i} \sum_{k \leq j} \alpha^{i}_{l-i+p} \alpha^{j}_{k-j+p} \mathcal{G}^{lk}_{(-\infty)(l)} & i,j \geq p \end{cases} \tag{4.49}$$

Since we know from our derivation in the previous section that the resultant matrix is banded, we only need compute this central band.

4.3.4 Measurement Uncertainty

As is standard with including measurement error into Gaussian process models, we add a diagonal matrix of the form $\Delta_{ij} = \nu^2 \sigma_i^2 \delta_{ij}$ to the covariance matrix, where σ_i is the measurement error corresponding to the i^{th} measurement, ν is a free parameter to scale the measurement uncertainty and δ_{ij} is the identity matrix.

In order to include this measurement error matrix in the efficient, sparse matrix calculation of the likelihood function, it is necessary to consider the effect of the $y \rightarrow x$ transformation. If we consider the matrix A defined in equation 4.48 then, on the inclusion of measurement errors, $C \rightarrow C + \Delta$, the transformed covariance matrix including measurement errors, which we shall denote by C'' takes the form

$$\begin{aligned} C''_{ij} &= A_{il}(C_{lk} + \Delta_{lk})A_{jk} \quad (4.50) \\ &= C'_{ij} + A_{il}\Delta_{lk}A_{jk}. \quad (4.51) \end{aligned}$$

If we denote a diagonal by Λ_i^k, where k denotes the absolute offset from the main diagonal, so that $k = 0$ denotes the leading diagonal, $k = 1$ denotes the first off-diagonal etc. and $i \in (0, N_{\text{obs}}]$ is the i^{th} element of the diagonal padded with k leading zeros, we can directly calculate the elements of $A_{il}\Delta_{lk}A_{jk}$

$$\begin{aligned} \Lambda_i^0 &= \nu^2 \left(\sigma_i^2 + \sum_{j=0}^{p-1} \sigma_{i-p+j}^2 (\alpha_j^i)^2 \right) \quad (4.52) \\ \Lambda_i^k &= \nu^2 \left(-\sigma_i^2 \alpha_{p-k}^{i+k} + \sum_{j=k}^{p-1} \sigma_{i-p+j}^2 \alpha_j^i \alpha_{j-k}^{i+k} \right) \quad (0 < k \leq p) \quad (4.53) \\ \Lambda_i^k &= 0 \quad (k > p). \quad (4.54) \end{aligned}$$

This extra matrix has the same sparse structure as $\mathbf{C}'$ with an extra nonzero diagonal, so that in general it is a Hermitian matrix with $2p + 1$ non zero diagonals.

4.3.5 Summary

Through the derivations presented in this section, we have demonstrated that by transforming the observations of a time series $y(t)$ to a new basis $x(t)$, as defined in equation 4.4, under the assumption that the time series is a CARMA process with Gaussian statistics, the covariance between observations is only non-zero for a few neighbouring observations. This means that the covariance matrix for the observations in the x basis is banded and can be inverted in $\mathcal{O}(N_{\text{obs}})$ operations.

This is useful for two purposes. First, as was discussed in section 1.4.2, the free parameters in the CARMA model are directly relatable to the shape of the power spectral density of the time series, and thus the physically interpretable features of the power spectrum, such as decay timescales and stochastic periodicities. By allowing the covariance matrix to be rapidly inverted, it becomes tractable to infer posteriors over these parameters and their physical interpretations for long time series.

A second, related benefit is that since we have assumed the noise properties in the CARMA process to be Gaussian, we have in fact written down a physically motivated kernel function for a Gaussian process, so that time series can be efficiently modelled and predicted using the Gaussian process update equations derived in appendix A.1. The hyperparameter optmisation problem in this case is of course equivalent to inferring the free parameters in the CARMA model.

4.4 Power Spectral Density and Autocovariance Function

The power spectral density of a CARMA process, which we derive in appendix A.3.1, is

$$P(f) = \sigma^2 \frac{\left|\prod_{j=1}^{q}(2\pi i f - b_j)\right|^2}{\left|\prod_{j=0}^{p-1}(2\pi i f - r_j)\right|^2}. \tag{4.55}$$

This expression can then be used to derive, in appendix A.3.2, the autocovariance function

$$\varrho(\tau) = \sigma^2 \sum_{k=0}^{p-1} e^{r_k \tau} \frac{\prod_{j=1}^{q}(r_k - b_j)(-r_k - b_j^*)}{(-r_k - r_k^*)\prod_{j=0,j\neq k}^{p-1}(r_k - r_j)(-r_k - r_j^*)}. \tag{4.56}$$

Including measurement uncertainty using the method described in section 4.3.4 introduces a 'white noise floor' to the power spectral density, which is the level below which any features arising from correlated noise are dominated by the measurement uncertainty. The white noise floor is the time average of the measurement uncertainty

$$\text{whitenoise} = \frac{\nu^2}{t_{max} - t_{min}} \int_{t_{min}}^{t_{max}} dt\ \sigma_{obs}^2. \tag{4.57}$$

Which is approximated by quadrature at the sample times. By including a scaling parameter ν on the measurement noise, the overall uncertainty in the inferred parameters of a CARMA model for a given set of data is balanced between both the intrinsic correlated noise of the stochastic process and the measurement uncertainty. This is best understood by considering the extremal cases of maximal and minimal contributions from these two sources.

In the cases that there is infinite amplitude in the measurement noise ($\nu \rightarrow \infty$), then no features of the correlated noise are discernible. Conversely, if there is no amplitude in the measurement noise, so that every observation is directly drawn from the correlated distribution, features will be discernible at arbitrarily low power.

When there is contribution from both correlated and measurement noise, features are only discernible above the white noise level. For CAR models (CARMA(p,0)), this gives a strong constraint on the identifiable features, since the power in the correlated noise is uniformly distributed across all frequencies, and all of the power is given to the autoregressive features of the spectrum that are sufficiently loud to be visible above the measurement noise. However, in the more general case that moving average roots are present, the power is distributed according to the moving average polynomial (i.e. the numerator in equation 4.55), so there is much more flexibility in the overall shape of the PSD. In general, this means that the ν parameter and the overall shape of the PSD are not as well constrained, but that 'quieter' autoregressive features are more easily identified in models with low autoregressive order.

4.5 Implementation

4.5.1 Bayesian Inference

Bayesian inference is a method of inferring the posterior distribution of the parameters controlling a model, given a set of observations. It is centred around Bayes theorem, which relates posterior probabilities to the product of the likelihood of some data D given a parametrised model $M(\theta)$ and an expression of prior knowledge about a problem

$$P(\theta|D, M) = \frac{P(D|\theta, M) \cdot P(\theta|M)}{P(D|M)}. \tag{4.58}$$

The denominator in equation 4.58 is a normalising factor known as the evidence, and is necessary for comparing different models. We set our prior distributions $P(\theta|M)$ to be between a conservative, mathematically valid range for all parameters, but chose uninformative priors between those bounds. We discuss this more in section 4.5.2. The likelihood function $P(D|\theta, M)$ is equation 4.7.

Since model comparison is important for the inferences in this chapter, we explore the posteriors using the multinest implementation of nested sampling, since this gives an estimate of the the evidence term $P(D|M)$. The values of p and q in a CARMA(p,q) model must be chosen using comparison of evidences.

When using the methods described in this chapter to infer characteristics of time series, for a CARMA(p, q) model there are $p + q + 3$ parameters to infer from the data. These are the p roots of the autoregressive polynomial $\{r_i\}$ on the left hand side of equation 4.1, the q moving average roots on the right hand side $\{b_i\}$, the centre of the time series μ. The variance term σ^2, which is central to computing $\mathcal{G}$ terms when computing the transformed covariance matrix, represents the variance of the white noise process ηt in the moving average polynomial. This variance is difficult to interpret physically, so we instead sample over a 'non-stationary variance' V, which is the autocovariance of the CARMA process from equation 4.56 at lag $\tau = 0$. We can then use equation 4.56 to translate this physical variance the the white noise variance σ^2. The final parameter is ν, which scales the measurement uncertainty as described in section 4.3.4.

4.5.2 Choice of Priors

When performing Bayesian inference uninformative priors were used for all of the CARMA model parameters. This implies a uniform prior on the mean μ and log-uniform priors on the non-stationary variance parameter V and the measurement error scaling ν. In practise, these uninformative priors are not normalisable, so conservative upper and lower limits were placed on each of the priors, based on simple estimators.

In the case of the roots, we elected to choose a prior such that the power spectral density of the prior is as flat as possible across all frequencies, indicating that in the absence of time structure, the signal is uncorrelated. This was most effectively achieved by implementing a log-uniform prior on the distance of the root from the origin in the complex plane. Upper and lower bounds for the root parameters were chosen to be proportional to the maximum and minimum plausibly detectable timescales (i.e $r_{min} = \frac{-A}{f_{NY}}$, $r_{max} = \frac{-1}{A(t_{max}-t_{min})}$, where f_{NY} is the Nyquist sampling rate). The upper bound for the imaginary part of complex roots was chosen to simply be $-r_{min}$. The constant scaling parameter A is implemented because for unevenly sampled data the Shannon-Nyquist sampling theorem does not apply and in specfific instances frequencies much higher than f_{NY} could be plausibly detected.

These parameters aren each bijectively mapped from their lower (l) and upper bounds (u) such that $\theta \in [l, u] \rightarrow \theta' \in [-\infty, \infty]$ via a logit transformation, defined by

$$\text{logit}(\theta; l, u) = \log(\theta - l) - \log(u - \theta), \tag{4.59}$$

with prior volume mediated by the appropriate Jacobian density.

$$\rho(\theta) = \left| \frac{\partial \theta}{\partial\left(\mathrm{logit}(\theta)\right)} \right| = \frac{(\theta - l)(u - \theta)}{(u - l)}. \tag{4.60}$$

4.6 Demonstration of Method

In this section we demonstrate the validity of our model and algorithm by recovering injected parameter sets. We also present analyses of some example datasets.

4.6.1 Super-Orbital Periods in the X-Ray Binary - XB158

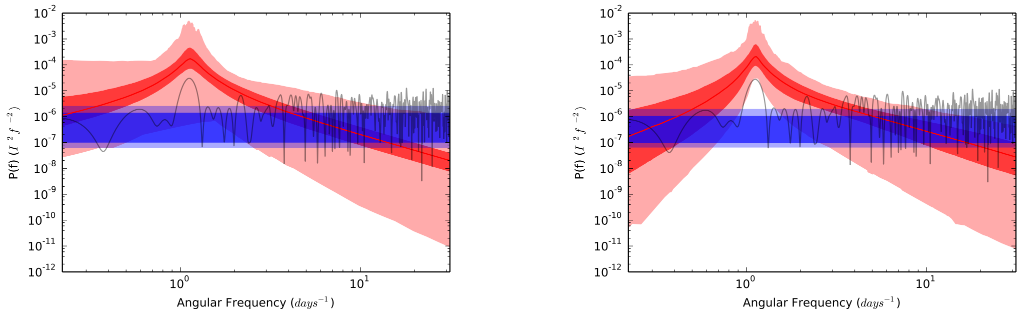
A binary system is a pair of massive objects orbiting their common centre of mass. X-Ray binaries are a particular class of binary system which is luminous in X-Rays, and are thought to typically consist of a compact object, such as a neutron star or black hole, and a normal star. X-Rays are emitted as matter is dragged from the normal star on to the compact companion.

XB158 is an X-Ray binary source in the M31 Globular cluster. The authors of Barnard et al. [2015] give observational data and analysis of this binary, highlighting that a ~ 5.7 day super-orbital period is present in their observations. We analysed the data using CARMA models of a variety of orders. We were then able to compare which orders of the CARMA models provided the best model for the data by comparing evidences. The evidences for all of the models considered can be seen in table 4.1.

Differences in log-evidence of order ~ 1 correspond to odds of order $\sim e$, which is not a significant preference of one model over another. With consideration also to the magnitude of the uncertainties in the computed evidences, there is little to suggest that models of greater order than the CARMA(2,1) need be considered. However, in order to verify this, we here present an analysis of the CARMA(2,1) model and the strictly 'best' model according to raw evidence value CARMA(5,4), so that we can compare the results from each.

p	q	$\log_e P(D\|M)$	$\pm$
1	0	107.24	0.11
2	0	108.67	0.17
2	1	109.99	0.23
3	0	108.15	0.20
3	1	109.74	0.25
3	2	109.92	0.28
4	0	107.55	0.24
4	1	109.49	0.27
4	2	110.28	0.30
4	3	110.31	0.33
5	1	108.80	0.29
5	2	109.65	0.32
5	3	110.52	0.35
5	4	110.94	0.37
6	0	106.94	0.28
6	1	108.50	0.31
6	2	109.54	0.34
6	3	109.86	0.37
7	0	106.71	0.31
7	1	107.95	0.32
7	2	108.93	0.35

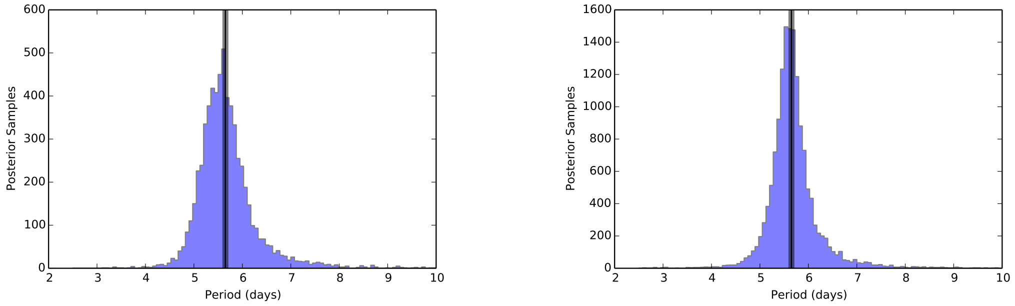
Table 4.1: Evidences for different orders of CARMA models for the XB158 dataset.



(a) CARMA(2,1) model

(b) CARMA(5,4) model

Figure 4.1: The best fit PSDs for two orders of CARMA models. Red areas show the median PSD, 1σ and 3σ confidence intervals. The blue horizontal line shows the 1σ and 3σ confidence intervals for the white noise floor. The Lomb-Scargle periodogram is shown in grey.



(a) CARMA(2,1) model, ∼ 94.7% of posterior.

(b) CARMA(5,4) model, ∼ 98.6% of posterior.

Figure 4.2: Histograms showing the marginalised posterior on the periodicity in the autocorrelation of the XB158 dataset. The black shaded region show the period quoted in Barnard et al. [2015]

Firstly, the PSDs, shown in figure 4.1 show a very clear and well defined peak above the white noise floor for both models, although for the CARMA(5,4) model, there is a clear peak well above the 3σ confidence level, whereas for the CARMA(2,1) model, there are PSDs within the 3σ interval which have a less clearly defined peak or no peak at all.

However, when a histogram of the posterior marginalised to the relevant parameters is plotted, both the CARMA(2,1) and CARMA(5,4) show very clear evidence for a periodicity in the data in the region of 5.7, as claimed by Barnard et al. [2015]. However, our uncertainty on this value is much greater than quoted by the authors. The CARMA(2,0) model shows a period of 5.6 ± 0.6 days, and the CARMA(5,4) model shows a period of 5.6 ± 0.4 days.

4.6.2 X-Ray Variability of ζ Puppis

ζ Puppis is a massive, extremely luminous O-Type star with well tested variability at optical frequencies. It is an open question as to whether ζ Puppis should also have variability at X-ray frequencies. There have been claims of detection of some variability (e.g., Nazé et al. [2013]), however there is still some debate. This is in part due to the challenges involved with gaining high quality X-ray data. We were provided with a lightcurve from observations made by the XMM telescope from

p	q	$\log_e P(D\|M)$	±
1	0	150.10	0.18
2	0	152.41	0.23
2	1	149.57	0.27
3	0	151.19	0.26
3	1	149.64	0.35
4	0	149.41	0.29
4	1	148.71	0.36
4	2	149.25	0.42
4	3	147.98	0.44
5	0	147.90	0.32
5	1	147.45	0.39
5	2	148.30	0.44
6	0	146.33	0.35
6	1	146.24	0.41

Table 4.2: log (ev) for different orders of CARMA models for the ζ Puppis lightcurve.

approximately the last 10 years to examine with the CARMA pipeline. [Howarth and Stevens, 2014]

The ζ Puppis lightcurve serves as an ideal test case for the CARMA pipeline, since it is extremely irregularly sampled. It consists of small collections of relatively high cadence ($\sim 1hr^{-1}$) data collected over more than 10 years. The X-ray lightcurve of ζ Puppis was analysed in the same way as the XB158 data, using the Multinest samplng method to compute evidences for different models. These are presented in table 4.2, and show a clear preference of $\sim$ 2 in $\ln(ev)$ for a CARMA(2,0) model (CAR(2)). The PSD resulting from the CAR(2) model for the ζ Puppis lightcurve and the lightcurve itself are shown in figure 4.4.

The ζ Puppis PSD has large peak in the Lomb-Scargle periodogram which is not present in the CARMA PSD. The X-Ray lightcurve, whilst being very irregularly timescaled on the scale of the whole dataset, consists of a series of regularly sampled epochs with a sampling frequency of $\sim 1hr^{-1}$. This periodicity is responsible for the large peak at $\omega = 150\text{days}^{-1}$. This peak is not present in the CARMA PSD since the sampling cadence is built into the model.

A further question that we cna address is the effect of increasing the number of data points at different cadences. Interpreting the CARMA model as a Gaussian

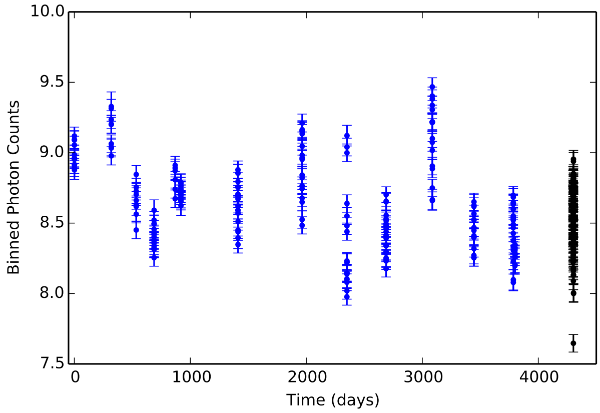


Figure 4.3: The X-ray lightcurve for ζ Puppis. Forecast mock observations are shown in black

processe, some extra data were drawn from the distribution implied by the data, in line with one of the potential observing strategies for XMM. This was a continuous stretch lasting 5 days at a cadence of 1hr^{-1} consistent with the best fit (maximum likelihood) CAR(2) model and appended to the exisiting dataset. The measurement errors were generated to be the mean of all the measurement errors in the original dataset. This new dataset was then analysed in exactly the same way as the original dataset, and then the relative uncertainties on the PSDs calculated (i.e. 1σ and 3σ confidence intervals normalised by the median PSD), and plotted in figure 4.5. We can quantify how another observation epoch at this cadence reduces the uncertainty in the PSD, in particular around the point of the 'knee' feature, which is important for physical interpretation.

4.7 Discussion

4.7.1 Instability of the Transformation

Using CARMA models to interpret stochastic variability is a promising avenue of research, however it is not without its limitations. In the course of our investigation we discovered that in many cases the method is computationally unstable. For example the transformation defined in Eq. (4.4) can become unstable for certain

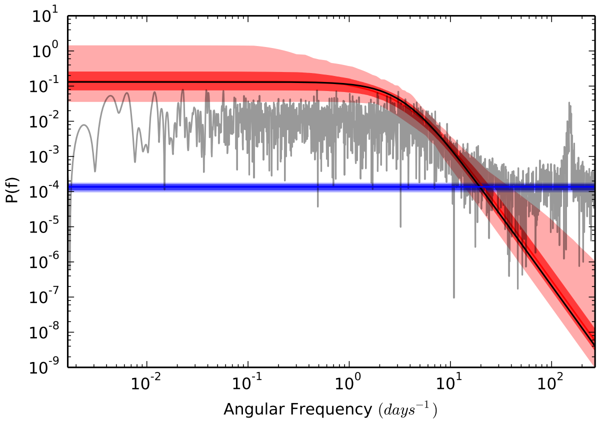


Figure 4.4: The power spectral density for a CAR(2) model for X-ray observations of ζ Puppis, with the same colour scheme in the PSD as figure 4.1, plus the PSD from the maximum likelihood parameters shown in black

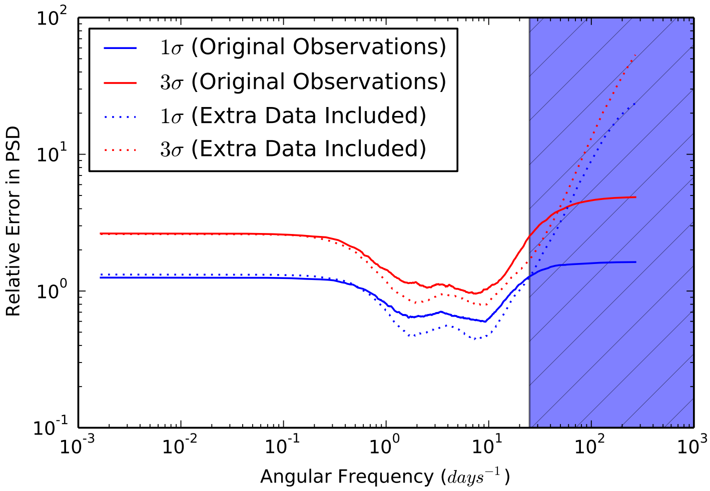


Figure 4.5: The relative error of the CAR(2) model with and without artificial data included. The shaded region represents the frequencies at which the PSDs fall below the white noise floor.

choices of autoregressive roots, r_i, and sample times. The problem is easiest to see in the $p = 2$ case, where

$$y_h(t_k) = \alpha^k_{k-1} y(t_{k-1}) + \alpha^k_{k-2} y(t_{k-2}) \tag{4.61}$$

$$= \frac{e^{r_2 t_k + r_1 t_{k-2}} - e^{r_1 t_k + r_2 t_{k-2}}}{e^{r_2 t_{k-1} + r_1 t_{k-2}} - e^{r_1 t_{k-1} + r_2 t_{k-2}}} y_h(t_{k-1}) \tag{4.62}$$

$$- \frac{e^{r_2 t_k + r_1 t_{k-1}} - e^{r_1 t_k + r_2 t_{k-1}}}{e^{r_2 t_{k-1} + r_1 t_{k-2}} - e^{r_1 t_{k-1} + r_2 t_{k-2}}} y_h(t_{k-2}) . \tag{4.63}$$

The alpha factors become infinite when

$$r_1(t_{k-1} - t_{k-2}) = r_2(t_{k-1} - t_{k-2}) + in\pi, \quad n \in \mathbb{Z}. \tag{4.64}$$

If the roots are real and distinct, then this condition cannot be satisfied. However if the roots r_1 and r_2 are a complex conjugate pair, with $r_{1,2} = \gamma \pm i\omega$ and

$$2\omega(t_{k-1} - t_{k-2}) = n\pi, \tag{4.65}$$

then the corresponding α factors will be infinite and the transformation will be unstable. Eq. (4.65) is effectively a local Nyquist condition for the angular frequency ω. This is particularly problematic in an astrophysical context, since it is a common observing strategy to have unevenly separated epochs of observations at a highly regular cadence, where this regular cadence is often targeted at the Nyquist frequency for the phenomena, in order to maximise efficiency. The ζ-puppis dataset presented in section 4.6.2 is a good example of this issue.

Furthermore, even for datasets where this Nyquist sampling issue wasn't present, we still found the covariance matrix $\mathbf{C}''$ regularly became close to singlular. The condition number of the matrix, defined as the ratio of its largest to its smallest eigenvalues, regularly exceeded 10^16, which can result in intolerable numerical errors when inverting it whilst computing the likelihood function of equation 4.7. Indeed, when we tracked the value of the likelihood of equation 4.7 compared to the slow and expensive (but stable) likelihood of equation 4.3, there were regularly significant deviations between the two. Unfortunately, we never discovered the reasons for this instability, and so were unable to mitigate them.

The Green's function approach for transforming the covariance matrix into something can be inverted efficiently is not the only method. Other closely related methods exist in the literature. In particular the kalman filter method described in Kelly et al. [2014] and the `celerite` method described in Foreman-Mackey et al. [2017] are both equally scalable and do not suffer from the same instability concerns as our method.

4.7.2 Aldebaran Companion

In order to further validate the method presented in this chapter, we took the data from Hatzes et al. [2015], which collates radial velocity measurements of the star Aldebaran, taken over a period of ~ 30 years from six different telescopes. It was the analysis of this dataset which highlighted the instability issues described above. We were unable to reliably reproduce the results detailed in that paper using the Green's function approach. This project was continued using the alternative methods more closely related to those described in Kelly et al. [2014] and Foreman-Mackey et al. [2017], however the author of this thesis had relatively little involvement in this work. The conclusions of this work have been submitted for publication [Farr et al., 2018], and we give a brief summary of the results here.

Aldebaran (also known as α Tauri) is a massive star located approximately 65 light-years from Earth. It is a K-type star, meaning that has exhausted its supply of Hydrogen and has subsequently undergone a Helium flash, where the transition between dominant fusion processes occur in its core, and has thus expanded to approximately 44.2 times larger than the sun (by radius). It is well known that exoplanets can exist around stars of this type, however, since observations of Aldebaran exhibit an intrinsic jitter on the order of $\sim 10^2 \mathrm{ms}^{-1}$, meaning that in order to build up any kind of evidence for exoplanets of a star this size a large number of observations have to be made over a long period of time. Hatzes et al. [2015] present a detailed analysis of their data, principally using the Lomb-Scargle periodogram, to conclude the existence of an exoplanet orbiting Aldebaran.

As well as confirming the presence of this planet, Farr et al. [2018] use CARMA based inference to find that the data also contains evidence for stellar oscillations within Aldebaran, which in turn can be used with asteroseismological techniques to accurately determine Aldebaran's mass.

Chapter 5

Conclusion

In this thesis we have used a variety of statistical techniques to explore several topics in astrophysics.

Since the first detection of gravitational-waves coming from the coalescence of two black holes, there has been renewed interest in understanding how two black holes might come so close together. It is a strong possibility that they are the product of isolated binary evolution, where two stars evolve together and interact with one another. We introduced `COMPAS` , a rapid population synthesis model for isolated binaries, which uses simple parametrisations of the physics of binary evolution so that large populations of binaries can be studied.

In chapter 2, we explored how much we can hope to constrain the parameters in the `COMPAS` model using gravitational-wave observations of coalescing binary black holes. We made small perturbations to the parameters in `COMPAS` in order to compute the Fisher information matrix, which quantifies how much we might learn about these parameters given some observations. We found that within just a few years (once the detectors have reached their design sensitivity), we could place strong constraints on the physics underpinning binary evolution. We also point out that the methodology is not specialised to just binary evolution, or just to gravitational-wave observations. The Fisher matrix techniques could be used to investigate the utility of almost any planned observations of astrophysical phenomena.

We then discussed how observations might be used to constrain the `COMPAS` physical parameters in practise in chapter 3. A single population can take tens of hours to simulate with the `COMPAS` model, which makes Bayesian inference using `COMPAS` model predictions in the likelihood computationally infeasible. Instead, we discussed how machine learning methods could be used to produce a cheaper statistical analogue, called an emulator, for the expensive `COMPAS` model. We described how an emulator would be built using Gaussian processes and random forests. The Gaussian process emulator did not perform well on the problem, although we highlight that this does not necessarily rule it out as a good model. The random forest emulator, in contrast, performed very well, and we were able to demonstrate how it could be used to infer the values of the physical parameters in the `COMPAS` model given a set of gravitational wave observations. The emulator approach to inference using gravitational-wave observations is highly promising, with many future avenues of research available, such as including more astrophysical observables.

In chapter 4, we developed a method for characterising stochastic variability in time series. By making the assumption that the process being observed is a continuous autoregressive moving average (CARMA) process, and making a transformation that separates the deterministic from the stochastic parts of the model, we showed that it is possible to write down a covariance function which leads to a covariance matrix which can be rapidly inverted, allowing rapid training of the covariance function hyperparameters and interpolation of the time series. It was unfortunate, however, to find that this transformation is unstable, and so has little utility when applied to real data.

Overall, we have demonstrated how statistical techniques are invaluable for the analysis of astrophysical data, and have paved the way for future studies, especially in the field of binary evolution, as we enter a data-rich era of gravitational-wave astronomy.

Bibliography

J. W. Barrett, S. M. Gaebel, C. J. Neijssel, A. Vigna-Gómez, S. Stevenson, C. P. L. Berry, W. M. Farr, and I. Mandel. Accuracy of inference on the physics of binary evolution from gravitational-wave observations. ArXiv e-prints, November 2017a.

H. Sana, S. E. de Mink, A. de Koter, N. Langer, C. J. Evans, M. Gieles, E. Gosset, R. G. Izzard, J.-B. Le Bouquin, and F. R. N. Schneider. Binary Interaction Dominates the Evolution of Massive Stars. Science, 337:444, July 2012a. doi: 10.1126/science.1223344.

A. Tutukov and L. Yungelson. Evolution of massive close binaries. Nauchnye Informatsii, 27:70, 1973.

P. Podsiadlowski, S. Rappaport, and E. D. Pfahl. Evolutionary Sequences for Low- and Intermediate-Mass X-Ray Binaries. ApJ, 565:1107–1133, February 2002. doi: 10.1086/324686.

Q. Z. Liu, J. van Paradijs, and E. P. J. van den Heuvel. Catalogue of high-mass X-ray binaries in the Galaxy (4th edition). A&A, 455:1165–1168, September 2006. doi: 10.1051/0004-6361:20064987.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, and et al. On the Progenitor of Binary Neutron Star Merger GW170817. ApJ, 850:L40, December 2017a. doi: 10.3847/2041-8213/aa93fc.

W. Hillebrandt and J. C. Niemeyer. Type IA Supernova Explosion Models. ARA&A, 38:191–230, 2000. doi: 10.1146/annurev.astro.38.1.191.

N. Ivanova, S. Justham, J. L. Avendano Nandez, and J. C. Lombardi. Identification of the Long-Sought Common-Envelope Events. Science, 339:433, January 2013a. doi: 10.1126/science.1225540.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Binary Black Hole Mergers in the First Advanced LIGO Observing Run. Phys. Rev. X, 6(4):041015, October 2016a. doi: 10.1103/PhysRevX.6.041015.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, and et al. GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral. Physical Review Letters, 119(16):161101, October 2017b. doi: 10.1103/PhysRevLett.119.161101.

A. Einstein. Die Grundlage der allgemeinen Relativitätstheorie. Annalen der Physik, 354:769–822, 1916a. doi: 10.1002/andp.19163540702.

A. Einstein. Näherungsweise Integration der Feldgleichungen der Gravitation. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), Seite 688-696., 1916b.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Observation of Gravitational Waves from a Binary Black Hole Merger. Phys. Rev. Lett., 116(6): 061102, February 2016b. doi: 10.1103/PhysRevLett.116.061102.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al. GW170608: Observation of a 19-solar-mass Binary Black Hole Coalescence. ArXiv e-prints, November 2017c.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, and et al. GW170104: Observation of a 50-Solar-Mass Binary Black Hole Coalescence at Redshift 0.2. Phys. Rev. Lett., 118(22):221101, June 2017d. doi: 10.1103/PhysRevLett.118.221101.

B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, and et al. GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence. Phys. Rev. Lett., 119(14):141101, October 2017e. doi: 10.1103/PhysRevLett.119.141101.

B. Abbott, R. Abbott, R. Adhikari, A. Ageev, and et al. Detector description and performance for the first coincidence observations between ligo and geo. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 517(1):154 – 179, 2004. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2003.11.124. URL http://www.sciencedirect.com/science/article/pii/S0168900203028675.

V. Cardoso, L. Gualtieri, C. A. R. Herdeiro, and U. Sperhake. Exploring New Physics Frontiers Through Numerical Relativity. Living Reviews in Relativity, 18:1, September 2015. doi: 10.1007/lrr-2015-1.

I. Mandel and S. E. de Mink. Merging binary black holes formed through chemically homogeneous evolution in short-period stellar binaries. MNRAS, 458:2634–2647, May 2016. doi: 10.1093/mnras/stw379.

C. L. Rodriguez, C.-J. Haster, S. Chatterjee, V. Kalogera, and F. A. Rasio. Dynamical Formation of the GW150914 Binary Black Hole. ApJ, 824:L8, June 2016. doi: 10.3847/2041-8205/824/1/L8.

O. De Marco and R. G. Izzard. Dawes Review 6: The Impact of Companions on Stellar Evolution. PASA, 34:e001, January 2017. doi: 10.1017/pasa.2016.52.

S. Stevenson, A. Vigna-Gómez, I. Mandel, J. W. Barrett, C. J. Neijssel, D. Perkins, and S. E. de Mink. Formation of the first three gravitational-wave observations through isolated binary evolution. Nature Communications, 8:14906, April 2017a. doi: 10.1038/ncomms14906.

M. Benacquista. An Introduction to the Evolution of Single and Binary Stars. 2013. doi: 10.1007/978-1-4419-9991-7.

S. Mohamed, R. Booth, and P. Podsiadlowski. The Asymmetric Outflow of RS Ophiuchi. In R. Di Stefano, M. Orio, and M. Moe, editors, Binary Paths to Type Ia Supernovae Explosions, volume 281 of IAU Symposium, pages 195–198, January 2013. doi: 10.1017/S1743921312014998.

R. J. Stancliffe and J. J. Eldridge. Modelling the binary progenitor of Supernova 1993J. MNRAS, 396:1699–1708, July 2009. doi: 10.1111/j.1365-2966.2009.14849.x.

M. Dominik, K. Belczynski, C. Fryer, D. E. Holz, E. Berti, T. Bulik, I. Mandel, and R. O'Shaughnessy. Double Compact Objects. I. The Significance of the Common Envelope on Merger Rates. ApJ, 759:52, November 2012. doi: 10.1088/0004-637X/759/1/52.

P. Kroupa. On the variation of the initial mass function. MNRAS, 322:231–246, April 2001. doi: 10.1046/j.1365-8711.2001.04022.x.

H. Sana, S. E. de Mink, A. de Koter, N. Langer, C. J. Evans, M. Gieles, E. Gosset, R. G. Izzard, J.-B. Le Bouquin, and F. R. N. Schneider. Binary Interaction Dominates the Evolution of Massive Stars. Science, 337:444–, July 2012b. doi: 10.1126/science.1223344.

E. Öpik. Statistical Studies of Double Stars: On the Distribution of Relative Luminosities and Distances of Double Stars in the Harvard Revised Photometry North of Declination -31deg. Publications of the Tartu Astrofizica Observatory, 25, 1924.

H. A. Abt. Normal and abnormal binary frequencies. ARA&A, 21:343–372, 1983. doi: 10.1146/annurev.aa.21.090183.002015.

Kurt W. Weiler and Richard A. Sramek. Supernovae and supernova remnants. Annual Review of Astronomy and Astrophysics, 26(1):295–341, 1988. doi: 10.1146/annurev.aa.26.090188.001455. URL `https://doi.org/10.1146/annurev.aa.26.090188.001455`.

C. L. Fryer, K. Belczynski, G. Wiktorowicz, M. Dominik, V. Kalogera, and D. E. Holz. Compact Remnant Mass Function: Dependence on the Explosion Mechanism and Metallicity. ApJ, 749:91, April 2012. doi: 10.1088/0004-637X/749/1/91.

H.-T. Janka and E. Mueller. Neutron star recoils from anisotropic supernovae. A&A, 290:496–502, October 1994.

S. E. Woosley. The birth of neutron stars. In D. J. Helfand and J.-H. Huang, editors, The Origin and Evolution of Neutron Stars, volume 125 of IAU Symposium, pages 255–270, 1987.

G. Hobbs, D. R. Lorimer, A. G. Lyne, and M. Kramer. A statistical study of 233 pulsar proper motions. MNRAS, 360:974–992, July 2005a. doi: 10.1111/j.1365-2966.2005.09087.x.

B. Paczynski. Common Envelope Binaries. In P. Eggleton, S. Mitton, and J. Whelan, editors, Structure and Evolution of Close Binary Systems, volume 73 of IAU Symposium, page 75, 1976.

N. Ivanova, S. Justham, X. Chen, O. De Marco, C. L. Fryer, E. Gaburov, H. Ge, E. Glebbeek, Z. Han, X.-D. Li, G. Lu, T. Marsh, P. Podsiadlowski, A. Potter, N. Soker, R. Taam, T. M. Tauris, E. P. J. van den Heuvel, and R. F. Webbink. Common envelope evolution: where we stand and how we can move forward. A&A Rev., 21:59, February 2013b. doi: 10.1007/s00159-013-0059-2.

R. F. Webbink. Double white dwarfs as progenitors of R Coronae Borealis stars and Type I supernovae. ApJ, 277:355–360, February 1984. doi: 10.1086/161701.

J. R. Hurley, C. A. Tout, and O. R. Pols. Evolution of binary stars and the effect of tides on binary populations. MNRAS, 329:897–928, February 2002. doi: 10.1046/j.1365-8711.2002.05038.x.

R. M. Humphreys and K. Davidson. The luminous blue variables: Astrophysical geysers. PASP, 106:1025–1051, October 1994. doi: 10.1086/133478.

K. Belczynski, T. Bulik, C. L. Fryer, A. Ruiter, F. Valsecchi, J. S. Vink, and J. R. Hurley. On the Maximum Mass of Stellar Black Holes. ApJ, 714:1217–1226, May 2010. doi: 10.1088/0004-637X/714/2/1217.

W.-R. Hamann and L. Koesterke. Spectrum formation in clumped stellar winds: consequences for the analyses of Wolf-Rayet spectra. A&A, 335:1003–1008, July 1998.

J. S. Vink and A. de Koter. On the metallicity dependence of Wolf-Rayet winds. A&A, 442:587–596, November 2005. doi: 10.1051/0004-6361:20052862.

E.T. Jaynes and G.L. Bretthorst. Probability Theory: The Logic of Science. Cambridge University Press, 2003. ISBN 9780521592710. URL `https://books.google.co.uk/books?id=tTN4HuUNXjgC`.

David J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Copyright Cambridge University Press, 2003.

John Skilling et al. Nested sampling for general bayesian computation. Bayesian Analysis, 1(4):833–859, 2006.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6):1087–1092, 1953.

Jonathan Goodman and Jonathan Weare. Ensemble samplers with affine invariance. Communications in Applied Mathematics and Computational Science, 5(1):65–80, 2010.

D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC Hammer. PASP, 125:306, March 2013. doi: 10.1086/670067.

W. D. Vousden, W. M. Farr, and I. Mandel. Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations. MNRAS, 455:1919–1937, January 2016. doi: 10.1093/mnras/stv2422.

Ben Calderhead and Mark Girolami. Estimating bayes factors via thermodynamic integration and population mcmc. Computational Statistics and Data Analysis, 53(12):4028 – 4045, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2009.07.025. URL http://www.sciencedirect.com/science/article/pii/S0167947309002722.

F Feroz, MP Hobson, and M Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. Monthly Notices of the Royal Astronomical Society, 398(4):1601–1614, 2009.

Farhan Feroz and MP 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, 384(2):449–463, 2008.

F Feroz, MP Hobson, E Cameron, and AN Pettitt. Importance nested sampling and the multinest algorithm. arXiv preprint arXiv:1306.2144, 2013.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL https://books.google.se/books?id=ZXL6AQAAQBAJ.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. Commun. ACM, 18(9):509–517, September 1975. ISSN 0001-0782. doi: 10.1145/361002.361007. URL http://doi.acm.org/10.1145/361002.361007.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133, Dec

1943. ISSN 1522-9602. doi: 10.1007/BF02478259. URL `https://doi.org/10.1007/BF02478259`.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988. ISBN 0-262-01097-6. URL `http://dl.acm.org/citation.cfm?id=65669.104451`.

Carl Edward Rasmussen. Gaussian processes in machine learning. In Advanced lectures on machine learning, pages 63–71. Springer, 2004.

S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil. Fast Direct Methods for Gaussian Processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38, June 2015. doi: 10.1109/TPAMI.2015.2448083.

T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer New York, 2013. ISBN 9780387216065. URL `https://books.google.co.uk/books?id=yPfZBwAAQBAJ`.

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

Leo Breiman. Random forests. Machine Learning, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

Jorge Nocedal and Stephen J. Wright. Numerical Optimization. Springer, New York, NY, USA, 1999.

W.H. Press. Numerical Recipes 3rd Edition: The Art of Scientific Computing. Cambridge University Press, 2007. ISBN 9780521880688. URL `https://books.google.co.uk/books?id=1aAOdzK3FegC`.

Will M. Farr, Benjamin J. S. Pope, Guy R. Davies, Thomas S. H. North, Timothy R. White, Jim W. Barrett, Andrea Miglio, Mikkel N. Lund, Victoria Antoci, Mads Fredslund Andersen, Frank Grundahl, and Daniel Huber. Aldebaran b's temperate past uncovered in planet search data. In prep, 2018.

R. Barnard, M. R. Garcia, and S. S. Murray. Swift reveals a 5.7 day super-orbital period in the m31 globular cluster x-ray binary xb158. The Astrophysical Journal, 801(1):65, 2015. URL http://stacks.iop.org/0004-637X/801/i=1/a=65.

George B Arfken. Mathematical methods for physicists. Academic press, 2013.

Chris Chatfield. The analysis of time series: an introduction. CRC press, 2013.

Brandon C Kelly, Andrew C Becker, Malgosia Sobolewska, Aneta Siemiginowska, and Phil Uttley. Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets. The Astrophysical Journal, 788(1):33, 2014.

Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and space science, 39(2):447–462, 1976.

Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, 263: 835–853, 1982.

J. T. VanderPlas. Understanding the Lomb-Scargle Periodogram. ArXiv e-prints, March 2017.

J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, K. Ackley, C. Adams, T. Adams, P. Addesso, and et al. Advanced LIGO. Classical and Quantum Gravity, 32(7):074001, April 2015. doi: 10.1088/0264-9381/32/7/074001.

F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou, A. Allocca, J. Amarni, P. Astone, G. Balestri, G. Ballardin, and et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. Classical and

Quantum Gravity, 32(2):024001, January 2015. doi: 10.1088/0264-9381/32/2/024001.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Prospects for Observing and Localizing Gravitational-Wave Transients with Advanced LIGO, Advanced Virgo and KAGRA. ArXiv e-prints, February 2017f.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Astrophysical Implications of the Binary Black-hole Merger GW150914. ApJ, 818:L22, February 2016c. doi: 10.3847/2041-8205/818/2/L22.

M. C. Miller. Implications of the gravitational wave event GW150914. General Relativity and Gravitation, 48:95, July 2016. doi: 10.1007/s10714-016-2088-4.

I. Mandel and A. Farmer. Gravitational waves: Stellar palaeontology. Nature, 547:284–285, July 2017. doi: 10.1038/547284a.

Konstantin A. Postnov and Lev R. Yungelson. The evolution of compact binary star systems. Living Reviews in Relativity, 17(1):3, May 2014. ISSN 1433-8351. doi: 10.12942/lrr-2014-3. URL `https://doi.org/10.12942/lrr-2014-3`.

K. Belczynski, D. E. Holz, T. Bulik, and R. O'Shaughnessy. The first gravitational-wave source from the isolated evolution of two stars in the 40-100 solar mass range. Nature, 534:512–515, June 2016. doi: 10.1038/nature18322.

J. J. Eldridge, E. R. Stanway, L. Xiao, L. A. S. McClelland, G. Taylor, M. Ng, S. M. L. Greis, and J. C. Bray. Binary Population and Spectral Synthesis Version 2.1: Construction, Observational Verification, and New Results. PASA, 34:e058, November 2017. doi: 10.1017/pasa.2017.51.

N. Giacobbo, M. Mapelli, and M. Spera. Merging black hole binaries: the effects of progenitor's metallicity, mass-loss rate and Eddington factor. ArXiv e-prints, November 2017.

S. Stevenson, C. P. L. Berry, and I. Mandel. Hierarchical analysis of gravitational-wave measurements of binary black hole spin-orbit misalignments. MNRAS, 471: 2801–2811, November 2017b. doi: 10.1093/mnras/stx1764.

M. Zevin, C. Pankow, C. L. Rodriguez, L. Sampson, E. Chase, V. Kalogera, and F. A. Rasio. Constraining Formation Models of Binary Black Holes with Gravitational-wave Observations. ApJ, 846:82, September 2017. doi: 10.3847/1538-4357/aa8408.

C. Talbot and E. Thrane. Determining the population properties of spinning black holes. Phys. Rev. D, 96(2):023012, July 2017. doi: 10.1103/PhysRevD.96.023012.

T. Bulik, K. Belczynski, and B. Rudak. Astrophysical significance of detection of coalescing binaries with gravitational waves. Astron. Astrophys., 415:407–414, 2004. doi: 10.1051/0004-6361:20034071.

Tomasz Bulik and Krzysztof Belczynski. Constraints on the binary evolution from chirp mass measurements. The Astrophysical Journal Letters, 589(1):L37, 2003. URL `http://stacks.iop.org/1538-4357/589/i=1/a=L37`.

I. Mandel and R. O'Shaughnessy. Compact binary coalescences in the band of ground-based gravitational-wave detectors. Classical and Quantum Gravity, 27 (11):114007, June 2010. doi: 10.1088/0264-9381/27/11/114007.

D. Gerosa, R. O'Shaughnessy, M. Kesden, E. Berti, and U. Sperhake. Distinguishing black-hole spin-orbit resonances by their gravitational-wave signatures. Phys. Rev. D, 89(12):124025, June 2014. doi: 10.1103/PhysRevD.89.124025.

S. Stevenson, F. Ohme, and S. Fairhurst. Distinguishing Compact Binary Population Synthesis Models Using Gravitational Wave Observations of Coalescing Binary Black Holes. ApJ, 810:58, September 2015. doi: 10.1088/0004-637X/810/1/58.

Rasmus Voss and Thomas M. Tauris. Galactic distribution of merging neutron stars and black holes - Prospects for short gamma-ray burst progenitors and LIGO/VIRGO. Mon. Not. Roy. Astron. Soc., 342:1169, 2003. doi: 10.1046/j.1365-8711.2003.06616.x.

N. Mennekens and D. Vanbeveren. Massive double compact object mergers: gravitational wave sources and r-process element production sites. A&A, 564:A134, April 2014. doi: 10.1051/0004-6361/201322198.

J. R. Hurley, O. R. Pols, and C. A. Tout. Comprehensive analytic formulae for stellar evolution as a function of mass and metallicity. MNRAS, 315:543–569, July 2000. doi: 10.1046/j.1365-8711.2000.03426.x.

A. Burrows and J. Hayes. Pulsar Recoil and Gravitational Radiation Due to Asymmetrical Stellar Collapse and Explosion. Phys. Rev. Lett., 76:352–355, January 1996. doi: 10.1103/PhysRevLett.76.352.

H.-T. Janka. Natal kicks of stellar mass black holes by asymmetric mass ejection in fallback supernovae. MNRAS, 434:1355–1361, September 2013. doi: 10.1093/mnras/stt1106.

G. S. Bisnovatyi-Kogan. Asymmetric neutrino emission and formation of rapidly moving pulsars. Astronomical and Astrophysical Transactions, 3:287–294, 1993. doi: 10.1080/10556799308230566.

A. Socrates, O. Blaes, A. Hungerford, and C. L. Fryer. The Neutrino Bubble Instability: A Mechanism for Generating Pulsar Kicks. ApJ, 632:531–562, October 2005. doi: 10.1086/431786.

G. Hobbs, D. R. Lorimer, A. G. Lyne, and M. Kramer. A statistical study of 233 pulsar proper motions. MNRAS, 360:974–992, July 2005b. doi: 10.1111/j.1365-2966.2005.09087.x.

T.-W. Wong, F. Valsecchi, A. Ansari, T. Fragos, E. Glebbeek, V. Kalogera, and J. McClintock. Understanding Compact Object Formation and Natal Kicks. IV. The Case of IC 10 X-1. ApJ, 790:119, August 2014. doi: 10.1088/0004-637X/790/2/119.

I. Mandel. Estimates of black hole natal kick velocities from observations of low-mass X-ray binaries. MNRAS, 456:578–581, February 2016. doi: 10.1093/mnras/stv2733.

S. Repetto, A. P. Igoshev, and G. Nelemans. The Galactic distribution of X-ray binaries and its implications for compact object formation and natal kicks. MNRAS, 467:298–310, May 2017. doi: 10.1093/mnras/stx027.

J. C. Bray and J. J. Eldridge. Neutron star kicks and their relationship to supernovae ejecta mass. MNRAS, 461:3747–3759, October 2016. doi: 10.1093/mnras/stw1275.

M. de Kool. Common envelope evolution and double cores of planetary nebulae. ApJ, 358:189–195, July 1990. doi: 10.1086/168974.

J. D. M. Dewi and T. M. Tauris. On the energy equation and efficiency parameter of the common envelope evolution. A&A, 360:1043–1051, August 2000.

M. U. Kruckow, T. M. Tauris, N. Langer, D. Szécsi, P. Marchant, and P. Podsiadlowski. Common-envelope ejection in massive binary stars. Implications for the progenitors of GW150914 and GW151226. A&A, 596:A58, November 2016. doi: 10.1051/0004-6361/201629420.

J. J. Eldridge and E. R. Stanway. BPASS predictions for binary black hole mergers. MNRAS, 462:3302–3313, November 2016. doi: 10.1093/mnras/stw1772.

A. Lamberts, S. Garrison-Kimmel, D. R. Clausen, and P. F. Hopkins. When and where did GW150914 form? MNRAS, 463:L31–L35, November 2016. doi: 10.1093/mnrasl/slw152.

J. S. Vink, A. de Koter, and H. J. G. L. M. Lamers. Mass-loss predictions for O and B stars as a function of metallicity. A&A, 369:574–588, April 2001. doi: 10.1051/0004-6361:20010127.

M. Renzo, C. D. Ott, S. N. Shore, and S. E. de Mink. Systematic survey of the effects of wind mass loss algorithms on the evolution of single massive stars. A&A, 603: A118, July 2017. doi: 10.1051/0004-6361/201730698.

N. Smith. Luminous blue variables and the fates of very massive stars. Philosophical Transactions of the Royal Society of London Series A, 375:20160268, September 2017. doi: 10.1098/rsta.2016.0268.

P. A. Crowther. Physical Properties of Wolf-Rayet Stars. ARA&A, 45:177–219, September 2007. doi: 10.1146/annurev.astro.45.051806.110615.

L. A. S. McClelland and J. J. Eldridge. Helium stars: towards an understanding of Wolf-Rayet evolution. MNRAS, 459:1505–1518, June 2016. doi: 10.1093/mnras/stw618.

F. Tramper, H. Sana, and A. de Koter. A New Prescription for the Mass-loss Rates of WC and WO Stars. ApJ, 833:133, December 2016. doi: 10.3847/1538-4357/833/2/133.

S.-C. Yoon. Towards a better understanding of the evolution of Wolf-Rayet stars and Type Ib/Ic supernova progenitors. MNRAS, 470:3970–3980, October 2017. doi: 10.1093/mnras/stx1496.

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\? Phys. Rev. D, 49:2658–2697, March 1994. doi: 10.1103/PhysRevD.49.2658.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Properties of the Binary Black Hole Merger GW150914. Phys. Rev. Lett., 116(24):241102, June 2016d. doi: 10.1103/PhysRevLett.116.241102.

G. Duchêne and A. Kraus. Stellar Multiplicity. ARA&A, 51:269–310, August 2013. doi: 10.1146/annurev-astro-081710-102602.

M. Moe and R. Di Stefano. Mind Your Ps and Qs: The Interrelation between Period (P) and Mass-ratio (Q) Distributions of Binary Stars. ApJS, 230:15, June 2017. doi: 10.3847/1538-4365/aa6fb6.

P. C. Peters. Gravitational Radiation and the Motion of Two Point Masses. Phys. Rev., 136:1224–1232, November 1964. doi: 10.1103/PhysRev.136.B1224.

P. Madau and M. Dickinson. Cosmic Star-Formation History. ARA&A, 52:415–486, August 2014. doi: 10.1146/annurev-astro-081811-125615.

N. Langer and C. A. Norman. On the Collapsar Model of Long Gamma-Ray Bursts:Constraints from Cosmic Metallicity Evolution. ApJ, 638:L63–L66, February 2006. doi: 10.1086/500363.

S. Savaglio, K. Glazebrook, D. Le Borgne, S. Juneau, R. G. Abraham, H.-W. Chen, D. Crampton, P. J. McCarthy, R. G. Carlberg, R. O. Marzke, K. Roth, I. Jørgensen, and R. Murowinski. The Gemini Deep Deep Survey. VII. The Redshift Evolution of the Mass-Metallicity Relation. ApJ, 635:260–279, December 2005. doi: 10.1086/497331.

X. Ma, P. F. Hopkins, C.-A. Faucher-Giguère, N. Zolman, A. L. Muratov, D. Kereš, and E. Quataert. The origin and evolution of the galaxy mass-metallicity relation. MNRAS, 456:2140–2156, February 2016. doi: 10.1093/mnras/stv2659.

E. Vangioni, K. A. Olive, T. Prestegard, J. Silk, P. Petitjean, and V. Mandic. The impact of star formation and gamma-ray burst rates at high redshift on cosmic chemical evolution and reionization. MNRAS, 447:2575–2587, March 2015. doi: 10.1093/mnras/stu2600.

P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al. Planck 2015 results. XIII. Cosmological parameters. A&A, 594:A13, September 2016. doi: 10.1051/0004-6361/201525830.

L. S. Finn. Detection, measurement, and gravitational radiation. Phys. Rev. D, 46: 5236–5249, December 1992. doi: 10.1103/PhysRevD.46.5236.

M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer. Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. Phys. Rev. Lett., 113(15):151101, October 2014. doi: 10.1103/PhysRevLett.113.151101.

S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. Phys. Rev. D, 93(4):044006, February 2016. doi: 10.1103/PhysRevD.93.044006.

S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. Phys. Rev. D, 93(4): 044007, February 2016. doi: 10.1103/PhysRevD.93.044007.

Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi. Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism. Phys. Rev. D, 89(8):084006, April 2014. doi: 10.1103/PhysRevD.89.084006.

S. Babak, A. Taracchini, and A. Buonanno. Validating the effective-one-body model of spinning, precessing binary black holes against numerical relativity. Phys. Rev. D, 95(2):024010, January 2017. doi: 10.1103/PhysRevD.95.024010.

A. Krolak and B. F. Schutz. Coalescing binaries-Probe of the universe. General Relativity and Gravitation, 19:1163–1171, December 1987. doi: 10.1007/BF00759095.

D. E. Holz and S. A. Hughes. Using Gravitational-Wave Standard Sirens. ApJ, 629: 15–22, August 2005. doi: 10.1086/431341.

L. S. Finn and D. F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. Phys. Rev. D, 47:2198–2219, March 1993. doi: 10.1103/PhysRevD.47.2198.

M. Dominik, K. Belczynski, C. Fryer, D. E. Holz, E. Berti, T. Bulik, I. Mandel, and R. O'Shaughnessy. Double Compact Objects. II. Cosmological Merger Rates. ApJ, 779:72, December 2013. doi: 10.1088/0004-637X/779/1/72.

M. Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. Phys. Rev. D, 77(4):042001, February 2008. doi: 10.1103/PhysRevD.77.042001.

S. Vitale, R. Lynch, R. Sturani, and P. Graff. Use of gravitational waves to probe the formation channels of compact binaries. Classical and Quantum Gravity, 34 (3):03LT01, February 2017a. doi: 10.1088/1361-6382/aa552e.

I. Mandel, C.-J. Haster, M. Dominik, and K. Belczynski. Distinguishing types of compact-object binaries using the gravitational-wave signatures of their mergers. MNRAS, 450:L85–L89, June 2015. doi: 10.1093/mnrasl/slv054.

I. Mandel, W. M. Farr, A. Colonna, S. Stevenson, P. Tiňo, and J. Veitch. Model-independent inference on compact-binary observations. MNRAS, 465:3254–3260, March 2017. doi: 10.1093/mnras/stw2883.

Howard Anton and Chris Rorres. Elementary Linear Analysis. John Wiley & Sons, Ltd, New York, eight edition, 2000. ISBN 0-471-17052-6.

J. R. Gair, C. Tang, and M. Volonteri. LISA extreme-mass-ratio inspiral events as probes of the black hole mass function. Phys. Rev. D, 81(10):104014, May 2010. doi: 10.1103/PhysRevD.81.104014.

S. Vitale, R. Lynch, V. Raymond, R. Sturani, J. Veitch, and P. Graff. Parameter estimation for heavy binary-black holes with networks of second-generation gravitational-wave detectors. Phys. Rev. D, 95(6):064053, March 2017b. doi: 10.1103/PhysRevD.95.064053.

J. J Andrews, A. Zezas, and T. Fragos. dart_board: Binary Population Synthesis with Markov Chain Monte Carlo. ArXiv e-prints, October 2017.

J. W. Barrett, I. Mandel, C. J. Neijssel, S. Stevenson, and A. Vigna-Gómez. Exploring the Parameter Space of Compact Binary Population Synthesis. In M. Brescia, S. G. Djorgovski, E. D. Feigelson, G. Longo, and S. Cavuoti, editors, Astroinformatics, volume 325 of IAU Symposium, pages 46–50, June 2017b. doi: 10.1017/S1743921317000059.

Stefano Conti and Anthony O'Hagan. Bayesian emulation of complex multi-output and dynamic computer models. Journal of Statistical Planning and Inference, 140(3):640 – 651, 2010. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2009.08.006. URL `http://www.sciencedirect.com/science/article/pii/S0378375809002559`.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2):239–245, 1979. ISSN 00401706. URL `http://www.jstor.org/stable/1268522`.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Binary Black Hole Mergers in the First Advanced LIGO Observing Run. Phys. Rev. X, 6(4):041015, October 2016e. doi: 10.1103/PhysRevX.6.041015.

B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Gw151226: Observation of gravitational waves from a 22-solar-mass binary black hole coalescence. Phys. Rev. Lett., 116:241103, Jun 2016f. doi: 10.1103/PhysRevLett.116.241103. URL `https://link.aps.org/doi/10.1103/PhysRevLett.116.241103`.

Yaël Nazé, Lidia M Oskinova, and Eric Gosset. A detailed x-ray investigation of ζ

puppis. ii. the variability on short and long timescales. The Astrophysical Journal, 763(2):143, 2013.

Ian D Howarth and Ian R Stevens. Time-series photometry of the o4 i (n) fp star ζ puppis. Monthly Notices of the Royal Astronomical Society, 445(3):2888–2893, 2014.

D. Foreman-Mackey, E. Agol, S. Ambikasaran, and R. Angus. Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series. AJ, 154:220, December 2017. doi: 10.3847/1538-3881/aa9332.

AP Hatzes, WD Cochran, M Endl, EW Guenther, P MacQueen, M Hartmann, M Zechmeister, I Han, BC Lee, GAH Walker, et al. Long-lived, long-period radial velocity variations in aldebaran: A planetary companion and stellar activity. arXiv preprint arXiv:1505.03454, 2015.

Appendix A

Derivations

A.1 The Gaussian Process Update Equations

In this appendix we present the derivation of the Gaussian process update equations. Note that this derivation was also published on the author's personal blog, at `http://jimbarrett.co.uk/bayesian/gaussian-process-update-equations/`.

The aim of Gaussian process regression is to write down a conditional distribution $P(y^*|y, x^*, x)$ for a set of predicted outputs y^* given a set of training (observed) outputs y, and their respective inputs x^* and x. For notational simplicity, we do not explicitly write the conditioning on x and x^* for the remainder of the derivation. By the product rule

$$P(y^*|y) = \frac{P(y^*, y)}{P(y)} \tag{A.1}$$

Since we have a constant set of data, $P(y)$ is just a constant in this expression.

The underlying assumption in Gaussian process regression is that outputs are jointly Gaussian distributed, so that

$$P(y^*|y) \propto P(y^*, y) \propto \exp\left[-\frac{1}{2}\begin{pmatrix} y \\ y^* \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} y \\ y^* \end{pmatrix}\right] \tag{A.2}$$

Where Σ is the joint covariance matrix. Remember that under the Gaussian process model we have trained a function which computes the elements of the Covariance matrix purely as a function of the inputs, it is only a function of the outputs y^* that we're trying to find. We can define the covariance matrix blockwise

$$\Sigma = \begin{pmatrix} T & C^T \\ C & P \end{pmatrix} \tag{A.3}$$

Where T is the covariance matrix computed using only the training inputs x_t, P is the covariance matrix computed using the prediction inputs x_p, and C is the cross terms (i.e. the covariance between y and y^*, computed using x_t and x_p). It is a well known result that you can blockwise invert a matrix [Press, 2007];

$$\Sigma^{-1} = \begin{pmatrix} T^{-1} + T^{-1}C^T MCT^{-1} & -T^{-1}C^T M \\ -MCT^{-1} & M \end{pmatrix} \tag{A.4}$$

Where $M = (P - CT^{-1}C^T)^{-1}$. So, we can directly compute our Gaussian density

$$P(y^*|y) \propto \exp\left[-\frac{1}{2}y^T(T^{-1} + T^{-1}C^T MCT^{-1})y\right. \tag{A.5}$$

$$\left.+\frac{1}{2}y^T(T^{-1}C^T M)y^* + \frac{1}{2}y^{*T}(MCT^{-1})y - \frac{1}{2}y^{*T}My^*\right] \tag{A.6}$$

However, the only thing that isn't a constant here is y^*, so we can ignore some terms (since we're only interested in the density, not absolute values)

$$P(y^*|y) \propto \exp\left[\frac{1}{2}y^T(T^{-1}C^TM)y^* + \frac{1}{2}y^{*T}(MCT^{-1})y - \frac{1}{2}y^{*T}My^*\right] \tag{A.7}$$

If we take the transpose of the middle term, we can group the terms together a bit more

$$P(y^*|y) \propto \exp\left[\frac{1}{2}y^T(T^{-1}C^TM + (MCT^{-1})^T)y^* - \frac{1}{2}y^{*T}My^*\right] \tag{A.8}$$

Now, in general, a multivariate Gaussian has the form;

$$\mathcal{N}(\tilde{y},\tilde{\Sigma}) \propto \exp\left[-\frac{1}{2}(y-\tilde{y})^T\tilde{\Sigma}^{-1}(y-\tilde{y})\right] \tag{A.9}$$

If we remember that covariance matrices are symmetric, we can expand, drop some constant terms and then rearrange this to

$$\mathcal{N}(\tilde{y},\tilde{\Sigma}) \propto \exp\left[-\frac{1}{2}y^T\tilde{\Sigma}^{-1}y + \tilde{y}^T\tilde{\Sigma}^{-1}y\right] \tag{A.10}$$

we can therefore see that both $P(y^*|y)$ and $\mathcal{N}(\tilde{y}, C)$ have exactly the same form. We can therefore straightforwardly match expressions for $\tilde{\Sigma}$.

$$\tilde{\Sigma} = M^{-1} = P - CT^{-1}C^T \tag{A.11}$$

The expression for $\tilde{y}$ requires a little more work. We start by matching terms

$$\tilde{y}^T\tilde{\Sigma}^{-1} = \frac{1}{2}y^T(T^{-1}C^TM + (MCT^{-1})^T) \tag{A.12}$$

We can rearrange this a little bit

$$\tilde{y}^T\tilde{\Sigma}^{-1} = \frac{1}{2}y^TT^{-1}C^T(M + M^T) \tag{A.13}$$

We know that M is a symmetric matrix (we just showed that its inverse is the covariance matrix). So, if we right multiply by the covariance matrix $\tilde{\Sigma}$ and take the transpose, we finally arrive at

$$\tilde{y} = CT^{-1}y \tag{A.14}$$

And, so, in conclusion we know that

$$P(y^*|y) \sim \mathcal{N}(CT^{-1}y, P - CT^{-1}C^T) \tag{A.15}$$

A.2 Normal Equations

If we have a set of overconstrained simulatneous linear equations, written in matrix form as

$$y = \mathbf{X}\beta \tag{A.16}$$

Where y is the vector of N responses and $\mathbf{X}$ is the $N \times M$ design matrix, where

M is the number of inputs to the model. β is the vector of M unknown parameters. We wish to find the value of $\beta = \hat{\beta}$ which minimises the squared error between the response and the model predictions $y - \mathbf{X}\beta$. The squared error is given by

$$\begin{aligned}
\varepsilon^2 &= (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) && \text{(A.17)}\\
&= y^T y - (\mathbf{X}\beta)^T y - y^T\mathbf{X}\beta + (\mathbf{X}\beta)^T\mathbf{X}\beta && \text{(A.18)}\\
&= y^T y - \beta^T\mathbf{X}^T y - y^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta && \text{(A.19)}
\end{aligned}$$

If we inspect the dimensionality of the middle two terms, we see that they are in fact scalars

$$\varepsilon^2 = y^T y - 2\beta^T\mathbf{X}^T y + \beta^T\mathbf{X}^T\mathbf{X}\beta \tag{A.20}$$

The aim is to find the $\beta = \hat{\beta}$ which minimises this expression. So, we differentiate

$$\frac{\mathrm{d}\varepsilon^2}{\mathrm{d}\beta} = 0 = -2\mathbf{X}^T y + \hat{\beta}^T\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\hat{\beta} \tag{A.21}$$

Which if we rearrange we find

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{A.22}$$

Which are the normal equations.

A.3 PSD and Autocovariance of a CARMA process

A.3.1 PSD

The power spectral density of a CARMA process can be obtained by taking the Fourier transform of equation 4.1, and computing $PSD(f) = \langle|\tilde{y}(t)|^2\rangle$.

$$\tilde{y}(t) = \frac{\prod_{i=1}^{q}(2\pi i f - b_i)}{\prod_{j=0}^{p-1}(2\pi i f - r_j)}\tilde{\eta}(t) \tag{A.23}$$

$$|\tilde{y}(t)|^2 = \frac{\left|\prod_{i=1}^{q}(2\pi i f - b_i)\right|^2}{\left|\prod_{j=0}^{p-1}(2\pi i f - r_j)\right|^2}|\tilde{\eta}(t)|^2 \tag{A.24}$$

$$\left\langle|\tilde{y}(t)|^2\right\rangle = \frac{\left|\prod_{i=1}^{q}(2\pi i f - b_i)\right|^2}{\left|\prod_{j=0}^{p-1}(2\pi i f - r_j)\right|^2}\left\langle|\tilde{\eta}(t)|^2\right\rangle \tag{A.25}$$

$\eta(t)$ is defined to be a stationary white noise process, and so we know its PSD explicitly $\left\langle|\tilde{\eta}(t)|^2\right\rangle = \sigma^2$, so

$$P(f) = \sigma^2\frac{\left|\prod_{i=1}^{q}(2\pi i f - b_i)\right|^2}{\left|\prod_{j=0}^{p-1}(2\pi i f - r_j)\right|^2} \tag{A.26}$$

A.3.2 Autocovariance

The autocovariance of a CARMA process can be calculated from the Weiner-Khinchin theorem, given in equation 1.43. Since all of the poles of the PSD lie in the top left quarter of the complex plane, with non-zero imaginary part, this integral can be computed by considering the integral around a semi-circular contour.

$$= 2\pi i\sum_{k=0}^{p-1}\mathrm{Res}(e^{2\pi i f\tau}P(f), \frac{r_k}{2\pi i}) \tag{A.27}$$

Where *Res* denotes the residue at the pole at $\frac{r_k}{2\pi i}$. Since all of the poles are 2^{nd} order simple poles, their residue can be calculated

$$\text{Res} = \lim_{f \to \frac{r_k}{2\pi i}} \frac{d}{df} \left(f - \frac{r_k}{2\pi i} \right)^2 e^{2\pi i f \tau} P(f) \tag{A.28}$$

These are straightforward to evaluate and so

$$\varrho(\tau) = \sigma^2 \sum_{k=0}^{p-1} e^{r_k \tau} \frac{\prod_{j=1}^{q} (r_k - b_j)(-r_k - b_j^*)}{(-r_k - r_k^*) \prod_{j=0, j \neq k}^{p-1} (r_k - r_j)(-r_k - r_j^*)} \tag{A.29}$$