University of Cape Town

Master of Science

FireFly: A Bayesian Approach to Source Finding in Astronomical Data.

Author:
Oarabile Hope Moloko

Supervisors:
Dr. Michelle Lochner
Prof. Bruce Bassett

January 28, 2020

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Abstract

Efficient and rigorous source finding techniques are needed for the upcoming large data sets from telescopes like MeerKAT, LSST and the SKA. Most of the current source-finding algorithms lack full statistical rigor. Typically these algorithms use some form of thresholding to find sources, which leads to contamination and missed sources. Ideally we would like to use all the available information when performing source detection, including any prior knowledge we may have. Bayesian statistics is the obvious approach as it allows precise statistical interrogations of the data and the inclusion of all available information.

In this thesis, we implement nested sampling and Monte Carlo Markov Chain (MCMC) techniques to develop a new Bayesian source finding technique called FireFly. FireFly employs a technique of switching 'on' and 'off' sources during sampling to deal with the fact that we don't know how many true sources are present. It therefore tackles one of the critical questions in source finding, which is estimating the number of real sources in the image. We compare FireFly against a Bayesian evidence-based search method and show on simulated astronomical images that FireFly outperforms the evidence-based approach.

We further investigate two implementations of FireFly: the first with nested sampling and the second with MCMC. Our results show that MCMC FireFly has better computational scaling than the nested sampling version FireFly but the nested sampling version of FireFly appears to perform somewhat better than MCMC FireFly. Future work should examine how best to quantify FireFly performance and extend the formalism developed here to deal with multiwavelength data.

Acknowledgements

Dedicated to the memory of my late grandmother, Lelokeng Lydia Seboko, who in dire and happy times supported me throughout my academic journey. You became my psychologist when I started my masters, and I thank you. You are gone but the belief you had in me made this journey possible. Kealeboga Mama.

To my mother, Kealeboga Muriel Moloko.
You have never questioned my abilities. We went through the most difficult times and you taught me that the storm always ends, just soldier through. Your support is everlasting. Thank you for raising Hope. Modimo Oarabile Mmamogolo.

My aunt, Onkabetse Bernice Seboko.
The support you have for me and my brother has never been different from your children, my siblings. Through thick and thin you made sure I had a home. Thank you for forever watching over me. Kealeboga Mmamane.

Keotshepile, Onalenna, Keobakile, Tumelo, Gofaone
my siblings, we fought and now we fight battles together. Thank you for always having hope in big brother.

To my friends, housemates, and collegues.
Thank you for keeping me sane and updated with the world, you taught me so much about humility.

Prof. Bruce Bassett and Dr. Michelle Lochner, my supervisors.
Thank you for letting me exercise your patience to the brim, and it never spilled. I gained an unimaginable experience under your supervision with great knowledge. I learned and got exposed to a vast amount of mind-blowing concepts. Bayesian statistics was a completely new way of thinking for me, but both of you were consistent with teaching me the fundamentals and applications. I thank you.

I would formally like to thank John Skilling, the Bayesian Ninja, for the original idea of FireFly. I also thank John, for personally visiting AIMS and SKA/SARAO, and giving a crash course on the fundamentals of nested sampling, the developer himself, and Bayesian statistics.

Contents

1 Source Finding in Astronomy **1**

1.1 Radio Astronomy 3

1.2 Overview of common source finding methods 5

1.2.1 Thresholding 5

1.2.2 Local peak search 7

1.3 Source Detection in Radio Interferometry 7

1.4 Bayesian Approach 8

1.4.1 Related work 8

2 Bayesian Statistics **11**

2.1 Probability 11

2.1.1 Frequency Theory 11

2.1.2 The Subjective Theory 12

2.2 Brief overview of Probability theory 13

2.3 Bayes' Theorem 14

2.3.1 Bayesian Inference 14

2.3.2 The Likelihood Function 15

2.3.3 The Prior 17

2.3.4 Common Priors . . . 18
2.3.5 The Posterior . . . 20
2.3.6 Marginalization . . . 21
2.3.7 Bayesian Evidence . . . 23
2.4 MCMC . . . 27
2.4.1 Markov chain sampling . . . 27
2.4.2 Metropolis—Hasting algorithm . . . 28
2.4.3 Choosing a step-size . . . 30
2.4.4 Burn-in . . . 31
2.4.5 Convergence . . . 32
2.4.6 Metropolis-Hasting - Chain analysis . . . 34
2.5 Nested Sampling . . . 35
2.5.1 The Nested Sampling algorithm . . . 36
2.5.2 Integration . . . 37
2.5.3 Nested sampling procedure . . . 38
2.5.4 Nested Sampling Termination . . . 39
2.5.5 Evidence error estimation . . . 40
2.5.6 Posterior samples . . . 41
2.5.7 Exploration . . . 42
2.5.8 Applications of Bayesian analysis in astronomy . . . 42

3 FireFly - a new Bayesian approach to source finding 47

3.1 Similar Work . . . 47
3.2 Search Algorithms . . . 48
3.2.1 Evidence-Based Algorithm . . . 48

3.2.2 FireFly Algorithm . . . 49
3.3 Clustering . . . 52

4 Applications **53**
4.1 Gaussian Sources . . . 53
4.2 Simulated data . . . 54
4.3 Method 1 : Evidence-Based approach . . . 55
4.3.1 Results: Evidence-Based Search . . . 58
4.3.2 Issues with Evidence-Based approach . . . 59
4.4 Method 2 : FireFly algorithm . . . 64
4.4.1 Results : FireFly . . . 64
4.5 Computational Requirements . . . 66
4.6 Computational efficiency: FireFly and Evidence-Based Search . . . 66
4.7 FireFly : MCMC vs. Nested Sampling . . . 67
4.7.1 FireFly : Nested sampling . . . 67
4.7.2 FireFly : MCMC . . . 67
4.7.3 MCMC vs. Nested Sampling . . . 68
4.8 Scaling FireFly . . . 68
4.8.1 FireFly Convergence . . . 71

5 Conclusions **73**

List of Figures

1.1 Electromagnetic Spectrum 2

1.2 Optical and Radio image 4

1.3 Thresholding 6

2.1 Evidence Z 25

2.2 MCMC example 31

2.3 MCMC Convergence 33

2.4 Nested Sampling (Multi-dimensional space transform) 37

2.5 Nested Sampling (Posterior Samples) 41

2.6 Related work mcmc (cosmology-planck) 43

2.7 Related work mcmc (cosmology) 44

2.8 Related work mcmc (gravitational waves) 45

2.9 Realted work mcmc (BIRO) 46

4.1 Radio image with and without noise 54

4.2 Simulated data 55

4.3 Evidence-based search (2 reasl sources) 57

4.4 (2 Real sources case) Evidence 58

4.5 Evidence-based search (4 real sources case) 60

4.6 Average Evidence plot (2 real sources) 61
4.7 Average Evidence plot (3 real sources) 61
4.8 Average Evidence plot (4 real sources) 62
4.9 Average Evidence plot (10 real sources) 63
4.10 Simulated data (3 real sources) 64
4.11 FireFly results 65
4.12 Execution time (Evidence-based in parallel vs. FireFly) 66
4.13 Execution time FireFly : MCMC vs. Nested sampling 68

List of Tables

2.1 Jeffrey's scale 26

List of Algorithms

1 Random Walk Metropolis - Hasting 29

2 Nested Sampling Algorithm 39

3 FireFly 51

Chapter 1

Source Finding in Astronomy

Historically, humans have been looking at the sky for aid in navigating through the seas and oceans, for knowledge of different seasons to plant vegetation and keeping track of time. This discipline is known as astronomy, i.e., the study of the space beyond the earth. It is a discipline that opens our eyes to different perceptions of our place in the universe, and it shapes how we see the world.

Astronomy has had a significant impact on our world and what we know in the universe. Early cultures have different interpretations of celestial objects, which are sometimes identified as 'gods' and their movements in the sky explained with prophecies. This coins the term astrology, which is far removed from the hard facts and real instruments of modern astronomy (Rosenberg et al. 2013).

The basic elements we find in the stars, gas and dust, are the same elements we find in our bodies, and that brings further connectedness between us and the cosmos. We use astronomy to study the celestial objects we find ourselves starring at and the unknown depths of the cosmos that's not visible with the naked eye. There is a vast spectrum of known and not well understood celestial objects and phenomena in the cosmos that are studied every day. These include the likes of neutron stars, radio galaxies, optical galaxies, black-holes, the cosmic microwave background, and gravitational waves.

Modern astronomy is divided into numerous specialties. Some of the specialties include cosmology, observational astronomy, and astrochemistry. Most of the time, astronomers study the characteristics of astronomical objects, which is collecting and analyzing light from telescopes to determine details such as elemental composition, temperature, pressure,

magnetic field strength, among other characteristics (Retrê et al. 2019). Astronomical data comes in different forms due to the vast electromagnetic spectrum studied in astronomy. Fig. (1.3) shows the electromagnetic spectrum. However, the data collected from telescopes and detectors is usually contaminated with unknown and unwanted signals. Therefore, to extract meaningful scientific conclusions from noisy and contaminated data, astronomers frequently employ source finding techniques.

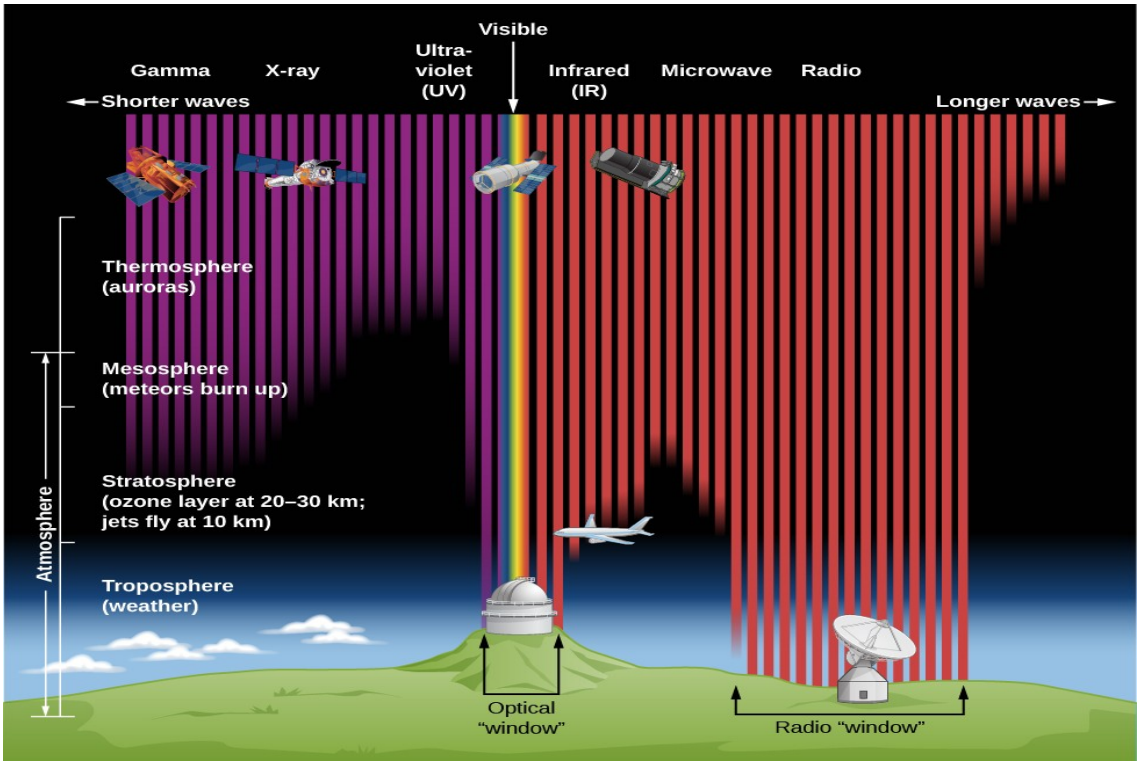


Figure 1.1: The electromagnetic spectrum. Courtesy of *Lumen Astronomy: Radiation and Spectra* 2019

To study these different types of sources after the data has been pre-processed, a key element of finding and classifying these astronomical sources is needed for any inference to be done. This brings about a crucial role of source finding in astronomy, which involves identifying different parts of an astronomical image to an astronomical source, against some background noise (Lukic, Gasperin, and Brüggen 2019).

Source finding or object detection in astronomy is the method of finding and characterizing objects in astronomical images (e.g., from an astronomical survey). The standard analysis for characterizing these objects follows by estimating their properties such as location, intensity, and shape from the images. Once extracted, these properties are summarised to form a

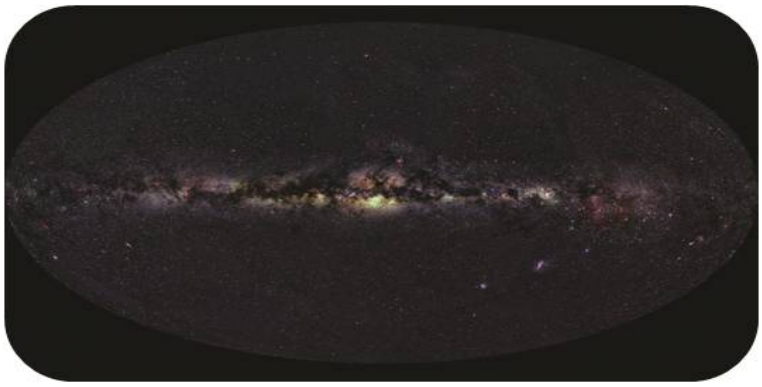
survey catalogue. Catalogues are generally constructed to provide a compressed version of the information contained in the survey and moreover for astrophysical systems represented in those data.

The development of automated source finding algorithms has become a research topic of interest for the astronomical community. Although these algorithms perform the same operations that an experienced astronomer can do with the appropriate tools, their importance lies in the fact that the algorithms can do these things fast, frequently and always with maximum objectivity (Masias et al. 2012).

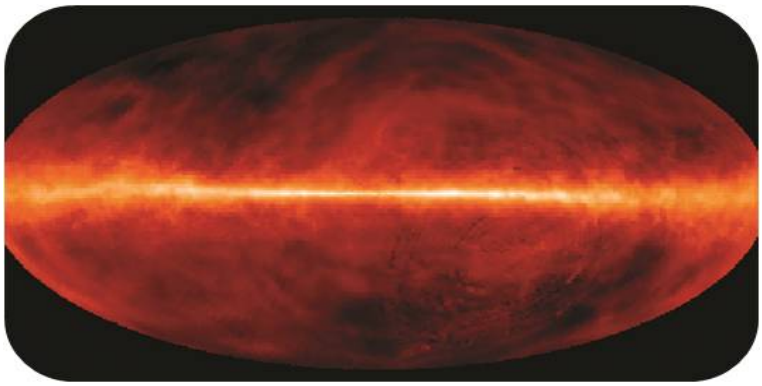
The software packages or methods used for source finding are known as source finders and usually provide object detection and characterization capabilities to produce catalogues of detected sources as their output (Lucas, Staley, and Scaife 2019). The objective of large-scale imaging surveys is to provide an unbiased census of the sky, and therefore an ideal source finder is both complete and reliable, i.e., finds all sources present in the image and all sources found and extracted are real.

1.1 Radio Astronomy

Radio astronomy has the advantage that, radio wavelengths that are from 3 kilohertz up to 300 gigahertz (Mai 2019), unlike visible waves, are immune to dust, that is, they can penetrate through dust without being absorbed or reflected. Radio astronomy allows us to observe things that would otherwise be invisible, such as the centre of the Milky Way. Moreover, one of the most studied topics is the hydrogen element, which is the most abundant in the universe that emits in the radio. One of the prominent disadvantages of radio astronomy is that radio signals do not come only from sources beyond the earth's atmosphere but also from earth's communication systems such as televisions, phones, satellites, and other human-made devices. This induces unwanted noise signals in collected data from radio telescopes. Figure (1.2) shows two sky observations in the optical and radio wavelengths.



(a) Optical



(b) Radio

Figure 1.2: (a) The sky observations at optical wavelengths. Credit: AxelMellinger/NASASkyVie and (b) The sky observed at the radio frequency of neutral hydrogen emission (21cm). Credit: J.Dickey/NASASkyView

The current generation of radio telescopes such as ASKAP(Johnston et al. 2008), MEERKAT (Booth et al. 2009) and eventually, the Square Kilometre Array (Garrett et al. 2010,Maartens et al. 2015) will produce large area, sensitive maps of the sky by probing the deepest parts of the universe. The SKA will be 50 times more sensitive than any existing radio facility, and most of the fundamental science will be addressed through large-area imaging at frequencies from 300 Mhz to a few GHz (Johnston et al. 2008). This will result in more data than previous surveys, just in its first phase, the SKA telescope will produce around 1000 petabytes of raw data per second (SKA 2019). Therefore data processing for these facilities will need to be fully automated, with minimal scope for manual intervention and correction. Hence the small fraction of missed or incorrectly identified sources provided by the current source finders will be a substantial problem. For example, mentioned in P. J. Hancock et al. 2012, given the enormous data rates, a source-finding algorithm that is 99% complete and reliable at a

signal-to-noise ratio (SNR) of ~ 5 will still produce $\sim 10\,000$ false sources and missing $\sim 10\,000$ real sources per day. Source finding algorithms will then not only need to be accurate and reliable but will be required to function efficiently on larger volumes of data.

1.2 Overview of common source finding methods

There are a large number of existing automated source detection algorithms that implement different methods. Some of the commonly used object detection algorithms, such as SExtractor (Bertin, E. and Arnouts, S. 1996), PyBDSF (Mohan and Rafferty 2015) and AGEAN (Paul J. Hancock, Trott, and Hurley-Walker 2018a), are based on background estimation strategies. These types of techniques involve estimating or setting a threshold with some prior knowledge between the sky and background noise in an astronomical image, and separating the two to perform object detection. In this section, we briefly look at two widely used background estimation strategies, which are thresholding and local peak search.

1.2.1 Thresholding

In astronomical images, thresholding is used to determine which regions (connected pixels in an image) are considered as objects (sources) and which are regarded as background noise (Masias et al. 2012). This type of method often requires restricting sources (objects) to those that are at least brighter than some threshold, typically setting the threshold to (e.g., 3σ, 5σ) above the root mean square noise level (Frean et al. 2014). Figure (1.3) shows an example of a thresholding technique called Thresholded Blob Detection (TBD).

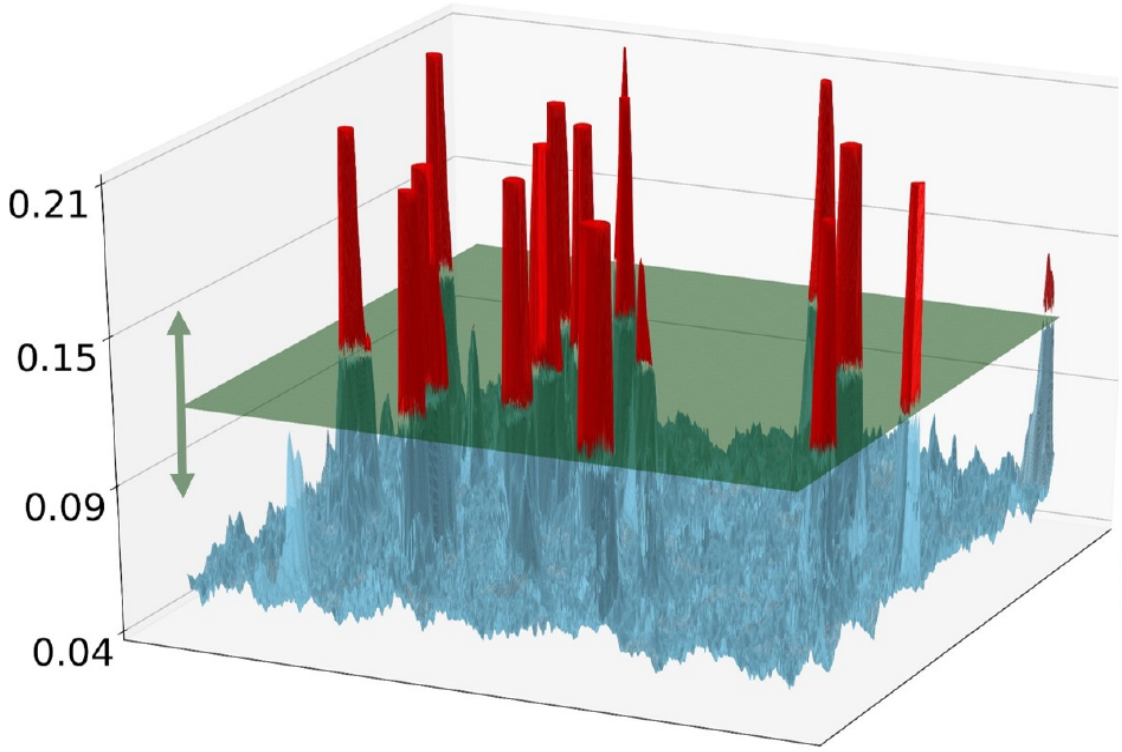


Figure 1.3: Animated example of a thresholding technique, the Thresholded Blob Detection (TBD) procedure developed by Vafaei Sadr et al. 2019.

Several factors can influence the success of a thresholding technique such as variation of noise across an image, and diffuse edges of objects. Any selected threshold may result in some true objects being entirely missed and some spurious objects being considered as real (Masias et al. 2012). Also, varying the threshold to the extremes minimizes one of these types of falsities, but maximizes the other. For example, setting a high threshold cut will allow most of the bright sources to be detected and many faint sources to be undetected, and a low threshold-cut will allow most of the bright and faint sources to be detected but can classify a significant amount of noise as sources. So this becomes a trade-off between the completeness for detecting sources and the precision of finding these sources. Therefore, the difficulty lies in setting the threshold to get these two errors as small as possible by maximizing both completeness and precision, moreover, a science case can influence whether one is more important than the other. For instance, false positive rate is often minimised in cosmology.

1.2.2 Local peak search

The local peak search method operates by searching those pixels that are a local maximum in a neighborhood, in other words, that are considered peaks. Typically, not all pixels are analyzed, and only those peaks that are above a given threshold are considered. An image transformation step such as deconvolution (Cornwell 2008) often accompanies this detection method, which enhances the peaks to be found. Another step follows which corrects the pixels around the peak that belong to the detected object (Masias et al. 2012). Local peak search is suitable for detecting stars and other point sources but is not well suited for detecting complex objects such as galaxies and extended sources. Local peak search methods still require manual tuning of the threshold parameter (Frean et al. 2014).

Common source finding algorithms developed with threshold/ local peak search methods include well-known packages such as SEXTRACTOR (Bertin, E. and Arnouts, S. 1996) and AEGEAN (Paul J. Hancock, Trott, and Hurley-Walker 2018b).

1.3 Source Detection in Radio Interferometry

Radio antennas can be used separately, or they can be combined to create a telescope known as an interferometer. The SKA will be one of the largest interferometers on earth. It will consist of over a square kilometer of collecting area, which is an accomplishment no survey telescope has ever achieved on this scale with a high level of sensitivity. Interferometers produce visibility data rather than pixelated-images produced from optical telescopes.

The difference between maps (visibility maps) and images is that: The cells or pixels of a map are produced by the process of gridding and Fourier transforming visibilities which are a result of correlation (Lucas, Staley, and Scaife 2019). However, in an image, each pixel corresponds to a measurement taken by a single detector element (Hague et al. 2018). Therefore, analysis of the former should be done with understanding the transformation of the image-map in order to make accurate inferences.

Source morphology is also an intricate part of object detection because sources may present different shapes due to the type of astronomical objects they are. Fundamentally, when the

angular size of an object is smaller than the angular resolution[1] of the telescope used in the observation, the object will appear as a point source. The response of a radio interferometer to a point source is called the point spread function (PSF) or sometimes called the synthesized beam (Masias et al. 2012). Sources that exceed the size of the PSF are commonly known as extended or resolved sources.

Source finding in radio-synthesis images has evolved concurrently with optical-data processing methods but has faced some challenges. To mention a few : (1) Most astronomical objects do not show clear boundaries since their intensities are similar to the background noise levels, and they get buried in the noise. (2) Since the application of transformation processes is to decrease the impact of background variations and noise, they can also consequently blur sources. (3) PSF with 'side-lobes' can cause bright sources to obscure faint sources even if a large amount of the PSF width separates the source positions. (4) The noise in radio interferometry images is correlated, meaning more spurious sources are detected.

1.4 Bayesian Approach

In this thesis, we implement a Bayesian formalism to source finding. We use the robust nature of Bayesian statistics to infer relevant parameters accurately. We apply two well-known sampling techniques, Markov-Chain Monte Carlo (MCMC) (Hogg and Foreman-Mackey 2017) and Nested Sampling developed by John Skilling 2006.

1.4.1 Related work

Accurate identification and characterization of sources in images and maps are an essential and ongoing problem in astronomy. Besides the many thresholding and local peak search algorithms, there are a small number of Bayesian source finding algorithms (Friedlander 2014).

Most of the Bayesian source finding algorithms use Markov-chain Monte Carlo sampling (MCMC)(Hogg and Foreman-Mackey 2017) techniques to estimate the relative probability that each pixel in an image is a background pixel or source pixel (Friedlander 2014).

M.P. Hobson and McLachlan 2003 introduced a source finding technique for astronomical

[1]How well a telescope can distinguish between two objects separated by small angular distance.

datasets (for any dataset with a localized signal and a diffuse background) using a Bayesian approach. They propose two strategies, which are: the simultaneous detection of all discrete objects in the dataset, and the iterative detection of objects(one-by-one). For both cases, the parameter space of the object(s) is explored using Markov Monte Carlo sampling(MCMC). They use the BAYESYS sampler software developed by John Skilling 2006.

They find that the simultaneous detection approach requires a complicated MCMC (probably due to multimodality (discussed in Feroz and J. Skilling 2013)) and it is also extremely computationally demanding. The iterative approach performs equally well as the simultaneous detection, but it is computationally much faster. The latter approach may have situations where it will perform 'bad,' given that it is usually successful in problems in which the objects of interest are well localized, and the data are such determining the characteristics of one object does not affect the derived properties of any other. This is likely to be problematic in the presence of correlated noise.

Hague et al. 2018 incorporates the Bayesian approach for source finding in a similar manner. They also use the BAYESYS software, which implements the MCMC sampler and yields the posterior of the parameters of interest.

To explore multimodal posterior distributions so that they can study individual sources, they introduce two clustering techniques. The first clustering technique they discuss is the Density-based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996) algorithm. This method assigns points to a cluster or tags them as outliers, based on the local density. The method requires the radius of the clustering search and the minimum number of samples per cluster to be pre-defined. However, it does not require the prior number of clusters to be given. The radius parameter is estimated by the k-distance graph which plots the k-nearest neighbors from all the data points.

The second technique is the Hierarchical Density-based Spatial Clustering of Applications with Noise(HBSCAN) (Campello, Moulavi, and Sander 2013) based on DBSCAN. Unlike DBSCAN, it can find clusters with varying densities and does not need the radius parameter.

Carvalho, Rocha, and M.P. Hobson 2009 introduced an algorithm called PowellSnakes. Instead of performing sampling techniques, they implement multiple local maximizations of the posterior. The evaluation of errors and Bayesian evidence values are performed by making a Gaussian approximation to the posterior at each peak. This method involves replacing the

standard form of the likelihood for the object parameters by an exact alternative form that is much quicker to evaluate, as discussed in Carvalho, Rocha, and M.P. Hobson 2009. It also requires a one-dimensional minimization sub-algorithm.

They found it to be faster than MCMC sampling. They further address that this method depends on several parameters in order to optimize its performance and sensitivity: i.e., no. of launched 'snakes,' the average number of sources per patch, and so. Furthermore, since the detection conditions(background noise, etc.) can display drastic changes, these parameters should be occasionally re-adjusted.

Further work by Lochner et al. 2015 implements a Bayesian technique called Bayesian Inference for Radio Observations (BIRO). The technique uses a Monte-Carlo process to perform a full Bayesian inference on both the source parameters and the systematic noise effects directly on the raw visibility data. They show that this technique allows the full use of all the information available to make accurate inferences.

Chapter 2

Bayesian Statistics

The theory of probability is the fundamental building block on which all of statistics is constructed, giving means for modeling experiments, populations, or about anything that can be regarded as a random phenomenon. Attaining these models gives statisticians the ability to draw inferences about populations, inferences based on examination of only a part of the whole (Casella and L. Berger 2019). *Statistical inference* is described as the process of concluding population parameters based on a sample taken from the population (Rothkopf 2019).

2.1 Probability

There are numerous forms of describing probability in statistics. To perform statistical analysis, a rigorous form of probability is required. We briefly discuss two theories of probability, namely; **Frequency Theory**, and **Subjective Theory**.

2.1.1 Frequency Theory

In the frequency theory, called the *frequentist* approach, probability is described as the ratio of the number of successful occurrences of a particular event to the number of trials, in the limit of an infinite number of repetitions (Tiao 1973a,Aldrich 1997,Trotta 2008). What is

meant by this is that if one conducts an experiment several times, independently and under identical conditions, the percentage of successful outcomes of a particular event will converge to a certain probability. For example, the chance of a "fair" coin landing on tails is 50% means that tossing the coin for a number of times, independently, the ratio of the number of times the coins lands on tails to the total number of tosses converges to 50% as the number of tosses increases.

The *frequentist* approach is usually debatable in the science community. A few of the challenges with this theory are as follows: (1) It makes assumptions that repeatable trials have the same probability of outcomes. (2) It cannot handle events that occur only at once. For instance, in the cosmological context, what is the probability that the universe will end up in a 'big crunch.' is not a repeatable experiment in which the *frequentist* approach can address (Trotta 2008 ,Tiao 1973a,Stark 2015).

2.1.2 The Subjective Theory

The subjective theory, also described as the Bayesian approach, describes probability as a measure of the degree of belief of a proposition. In other words, a measure of uncertainty about the occurrence of an event (Bayes 1763). The Bayesian approach is useful in defining the meaning of probability of events which in principle can occur once. For example, what is the probability that it will rain tomorrow? i.e., tomorrow can only come once.

Furthermore, the Bayesian approach gives a more satisfactory realization of probability. It is subjective, which allows a result of different analyses to be made by different experimenters leading to different conclusions (which is also evident in the frequentist analysis). This form of probability applies to any event, despite whether it is considering repeated experiments or once-off events (Trotta 2008). A crucial advantage of the Bayesian approach is that it deals with uncertainty independently of its origin. (e.g., background noise in astronomical images).

2.2 Brief overview of Probability theory

In this section, we provide a brief overview of the fundamentals of probability, which give a foundation to the Bayesian formalism, drawing from (Tiao 1973a,Trotta 2008,Gregory 2005), we provide a brief overview of the fundamental rules of probability. Let $\mathcal{P}(A)$ denote the probability that event A is true. For example, the event that a 'fair' coin lands on tails. On the other hand, $\mathcal{P}(\bar{A})$ denotes that the event is not A.

1. **The sum rule reads**
$$\mathcal{P}(A) + \mathcal{P}(\bar{A}) = 1 \tag{2.1}$$
which means that all the possible outcomes sum up to unity.

2. **The probability union of two events A and B**
The probability that event A or B occurs is defined by
$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) - \mathcal{P}(A \cap B) \tag{2.2}$$
where $\mathcal{P}(A \cap B)$ is the probability of both A and B occurring simultaneously. Consequently, if A and B are mutually exclusive, then
$$\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) \tag{2.3}$$

3. **Conditional probability**
The conditional probability of event A, given that B is true is defined by
$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} \tag{2.4}$$

4. **Joint probability**
Following the definition of conditional probability, joint probability of A and B is given by
$$\mathcal{P}(A \cap B) = \mathcal{P}(A|B)\mathcal{P}(B) \tag{2.5}$$
where $\mathcal{P}(A|B)$ denotes the probability of A occurring given B has occurred. This is also referred to as the product rule.

5. **The marginal distribution**
   If we have the joint distribution $\mathcal{P}(A \cap B)$ and the interest is the probability of A alone, irrespective of B, then the marginal distribution of A is given by

$$\mathcal{P}(A) = \begin{cases} \sum_B \mathcal{P}(A|B)\mathcal{P}(B), & Discrete \\ \int_B \mathcal{P}(A|B)\mathcal{P}(B), & Continuous \end{cases} \tag{2.6}$$

   where the sum or the integral is taken over all possible states of B.

2.3 Bayes' Theorem

Bayes' Theorem is fundamentally a result of a trivial manipulation of the joint probability. Let $\mathcal{I}$ denote where assumptions have been made for which both events A and B depend on, which we will call the *information*. For example, in conducting a particular scientific experiment, we may assume a set of assumptions on the data, such as in special relativity where the speed of light (in a vacuum) is assumed to be constant. These assumptions will be included in $\mathcal{I}$.

Since $\mathcal{P}(A \cap B|\mathcal{I}) = \mathcal{P}(B \cap A|\mathcal{I})$ it follows that,

$$\mathcal{P}(A|B \cap \mathcal{I})\mathcal{P}(B|\mathcal{I}) = \mathcal{P}(B|A \cap \mathcal{I})\mathcal{P}(A|\mathcal{I}) \tag{2.7}$$

Consequently, we obtain Bayes' Theorem (Bayes 1763),

$$\mathcal{P}(A|B \cap \mathcal{I}) = \frac{\mathcal{P}(B|A \cap \mathcal{I})\mathcal{P}(A|\mathcal{I})}{\mathcal{P}(B|\mathcal{I})} \tag{2.8}$$

2.3.1 Bayesian Inference

Applying Bayes' Theorem in the context of inference, where a vector of unknown parameters θ of a model $\mathcal{I}$ (hypothesis plus any information) are investigated, given data D, Bayes' theorem is given by

$$\mathcal{P}(\theta|D,\mathcal{I}) = \frac{\mathcal{P}(D|\theta,\mathcal{I})\mathcal{P}(\theta)}{\mathcal{P}(D|\mathcal{I})} \tag{2.9}$$

In the expression (2.9), $\mathcal{P}(\theta|D,\mathcal{I})$ is called the *posterior* probability distribution of θ given the model $\mathcal{I}$ and data D, which tells us what is known about θ given knowledge of the data D. $\mathcal{P}(D|\theta,\mathcal{I})$ is called the *likelihood function* of θ and $\mathcal{I}$ given the data D, which tells us how likely the data D is given θ. $\mathcal{P}(\theta|\mathcal{I})$ is called the *prior* of the parameters θ, which tells us what is known about θ without knowledge of the data D . The term $\mathcal{P}(D|\mathcal{I})$ is most commonly called the evidence or the marginal likelihood, which sums over all possible ranges of θ, given by

$$\mathcal{P}(D|\mathcal{I}) = \mathcal{Z} = \int \mathcal{P}(D|\theta,\mathcal{I})\mathcal{P}(\theta|\mathcal{I})d\theta \tag{2.10}$$

Typically in parameter estimation problems, the evidence is ignored since it is independent of θ. The un-normalized posterior then incorporates the full Bayesian inference of the parameter values. However, in model selection problems, the evidence plays a crucial role. It allows model assumptions to be compared to one another. One of the current methods for evaluating the evidence is Nested Sampling developed by John Skilling 2006. In the following sections we now examine each term in Bayes' theorem in turn.

2.3.2 The Likelihood Function

The *likelihood function* of θ for given data D can be written as shortcut notation $\mathcal{L}(\theta) = \mathcal{P}(D|\theta,\mathcal{I})$ (Aldrich 1997). We can therefore write Bayes' theorem as

$$\mathcal{P}(\theta|D,\mathcal{I}) = \mathcal{L}(\theta)\frac{\mathcal{P}(\theta)}{\mathcal{P}(D|\mathcal{I})} \tag{2.11}$$

The likelihood function plays a critical role in Bayes' formula. It is essentially the function in which the data D updates prior knowledge of θ. Moreover, the likelihood function is defined up to a multiplicative constant, that is, the likelihood remains unchanged when multiplied by an arbitrary constant (Tiao 1973a). Consequently, in Bayes' theorem, this does not affect the posterior distribution of θ. Hence, only the relative value of the likelihood is significant.

The likelihood function for independent data is defined as follows. Let $d = \{d_1, d_2, d_3, ...d_n\}$ be a sample of n independent observations from a distribution in which the probability

density function is of the known form $\mathcal{P}(d|\theta)$. Then the likelihood function can be defined as

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \mathcal{P}(d_i|\theta) \tag{2.12}$$

where the likelihood is a product of each probability density of each sample d_i.

Maximum Likelihood Estimation

In practice, the notion of the likelihood function can be used to make statistical inferences about specific populations given some data. This brings about the principle of maximum likelihood estimation(MLE), developed by R. A. Fisher 1921. Myung 2003's review mentions that Fisher states: '*the desired probability distribution is the one that makes the observed data 'most likely,' which means that one must seek the value of the parameter vector that maximizes the likelihood function.*'

The basic idea with MLE is to find an optimal value called an estimate that maximizes the likelihood function $\mathcal{L}(\theta)$ to get the best parameter estimate of a specific model that describes some data. In practice, for this to work, an assumption that $\mathcal{L}(\theta)$ is differentiable and MLE exists has to hold. Therefore, two conditions must be satisfied, which are thoroughly discussed by Myung 2003, which are:

1. $\mathcal{L}(\theta)$ must satisfy the partial differential equation known as the likelihood equation :

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_i} = 0 \tag{2.13}$$

   at $\theta_i = \theta_{i,MLE}$, for all $i = 1, ..., k$

2. to ensure that $\mathcal{L}(\theta)$ at the MLE estimate is a maximum :

$$\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i^2} < 0 \tag{2.14}$$

A practical procedure of obtaining an analytic form solution for the MLE estimate is not always feasible. These techniques implement a basic idea of quickly finding optimal parameters

that maximize the likelihood function. A disadvantage of maximum likelihood estimation is that it only produces the best fit, unlike a full inference analysis, which yields a full posterior output. Moreover, because of random noise, this estimate will be somewhat wrong and you have no way of knowing how wrong it is without the uncertainty.

Maximum Likelihood Estimation vs. Bayesian Estimation

In MLE, only a single value of the parameter θ is sought, in order to make inference about θ, that is by maximizing the likelihood function $\mathcal{L}(\theta)$. This, however, does not allow the insertion of prior information or beliefs, $\mathcal{P}(\theta)$, about the possible values of θ.

However, in Bayesian estimation, θ is treated as a random variable, and instead of getting a single point estimate of θ, a probability distribution of the possible values of θ is presented. A significant value of Bayesian estimation is that it included prior beliefs, $\mathcal{P}(\theta)$, and the evidence, $\mathcal{P}(D)$. The evidence term, $\mathcal{P}(D)$, plays a vital role in model selection (Tiao 1973a,Feroz and J. Skilling 2013,Lochner et al. 2015).

2.3.3 The Prior

The fundamental guidelines of Bayesian inference are that there can be no inference without assumptions. The prior plays an essential role in encoding necessary information about a problem in question before new data is taken into account.

The prior state of knowledge of a particular problem can differ amongst numerous scientists or statisticians, and this has made the choice of a prior to be quite debatable. Consequently, this notion of being able to have a subjective prior, that is, being able to have a different choice of prior distributions, is believed to be an advantage of Bayesian analysis rather than a drawback (Trotta 2008). One of the significant reasons being that Bayesian analysis allows prior beliefs or assumptions to be updated by the evidence of new data. The incorporation of a suitable prior is essential in Bayesian inference as it strengthens inferences and ensures relevant information is not lost.

2.3.4 Common Priors

A general Bayesian analysis assumes that a posterior is a probability distribution. This, however, may be different if an improper prior distribution is used. We will briefly discuss improper and proper prior distributions in this section.

Improper priors

If little or no information is known about a particular problem in question, then we generally need a prior with minimal influence on the inference, such a prior is called an improper prior. The advantage of having an improper prior is to make the data (likelihood function) speak more about the parameters in question when prior knowledge is limited, and usually, the result is a proper posterior distribution (Tak, Ghosh, and Ellis 2018). In mathematical terms, a prior is said to be improper if the integral of the distribution does not integrate to one. An improper prior is often used as an uninformative prior.

A vital issue comes in choosing a suitable improper prior. A discussion by Syversveen 1998 mentions that Bayes/Laplace presented a postulate stated about 200 hundred years ago, which says the following: 'When nothing is known about θ in advance, let the prior $\mathcal{P}(\theta)$ be a uniform distribution, that is, let all possible outcomes of θ have the same probability'. This was further known as the notion of *insufficient reason*.

Syversveen 1998 further discusses that Fisher did not support the Bayes/Laplace postulate. He debated that: 'Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge.' He, therefore, only accepted Bayes' theorem for informative priors.

Proper priors

The incorporation of scientifically motivated priors is one advantage of using Bayesian analysis, because it provides a logical way of incorporating scientific knowledge into inference problems using priors. Proper priors are ideal for this purpose since they ensure proper posterior probability distributions (Tak, Ghosh, and Ellis 2018).

Uniform prior distribution

A standard type of a commonly used proper prior is a uniform prior distribution. A uniform prior distribution is a type of distribution in which all the outcomes within a considered range are equally likely (Tiao 1973a,Syversveen 1998). The uniform prior also known as a 'flat' prior is often a common choice in which the prior is taken to be constant within some minimum and maximum value of the parameters (Tiao 1973a). For a 1-dimensional case, this can be defined as

$$\mathcal{P}(\theta) = \frac{1}{(\theta_{max} - \theta_{min})} \tag{2.15}$$

where,

$$\int_{\theta_{min}}^{\theta_{max}} \mathcal{P}(\theta) d\theta = 1, \quad \theta \in (\theta_{min}, \theta_{max}) \tag{2.16}$$

However, using a 'flat' prior can be quite problematic. The reason being that a uniform prior is not invariant under reparameterization (Syversveen 1998). For instance, a 'flat' prior on a parameter θ does not correspond to a flat prior on a non-linear function of that parameter, $\phi(\theta)$. The two priors can be related (Syversveen 1998,Trotta 2008) by

$$\mathcal{P}(\phi) = \mathcal{P}(\theta) \left| \frac{d(\theta)}{d(\phi)} \right| \tag{2.17}$$

So in other words, if we have no information on θ, then we have no information on ϕ, but a uniform prior on θ does not correspond to a uniform prior for ϕ.

Jeffrey's prior

Another well known proper prior is Jeffrey's prior, named after an English mathematician Sir Harold Jeffreys (Robert, Chopin, and Rousseau 2008). Jeffreys prior is defined by

$$\mathcal{P}(\theta) \propto |I(\theta)|^{1/2} \tag{2.18}$$

where $I(\theta)$ is called the Fisher information (Robert, Chopin, and Rousseau 2008). The Fisher information is described as how well we can measure a parameter θ given a certain amount of data. Robert, Chopin, and Rousseau 2008 states that Jeffrey's prior is a suitable type of non-informative prior in that it is invariant under any transformation.

Gaussian prior

A random variable x is said to have a normal or gaussian distribution with parameters μ and σ if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{2.19}$$

for $-\infty < x < \infty, -\infty < \mu < \infty$ and $\sigma > 0$ (Nadarajah 2005).

This Gaussian distribution, also known as the power exponential distribution, can be used to set up bounds. It sets hard bounds which exclude a specific range of parameters, it forbids values outside the bounds with small but non-zero probability (Tak, Ghosh, and Ellis 2018).

2.3.5 The Posterior

In the Bayesian analysis, analysis involves going from a prior distribution, $\mathcal{P}(\theta)$, to a posterior distribution, $\mathcal{P}(\theta|D)$. Bayesian inference is generally based on the properties of the posterior distribution.

In Bayes' theorem (2.11), inference about the parameters θ is independent of the evidence, $\mathcal{P}(D)$, since inference lies significantly in the shape of the posterior distribution, thus (2.11) is usually written as

$$\mathcal{P}(\theta|D) \propto \mathcal{L}(\theta)\mathcal{P}(\theta) \tag{2.20}$$

alternatively, in words,

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

The posterior distribution incorporates two sources of information about the parameters θ, namely:(1) the prior distribution, $\mathcal{P}(\theta)$, which represents prior information or belief about θ; (2) the likelihood function, $\mathcal{L}(\theta)$, which represents the information about θ provided by the data D.

In principle, the posterior distribution contains all the available information. This fact is useful for estimating parameters, making predictions, and other inferential analyses for which the data were gathered (Tiao 1973a,John Skilling 2006,Trotta 2008,Lochner et al. 2015).

2.3.6 Marginalization

Marginalization is a key concept of dealing with unwanted parameters in Bayesian statistics. Suppose a model consists of multiple parameters θ, but interest is only in the posterior distribution of one of the parameters in θ. For instance, an astronomical sky model which consists of an intensity F and background noise n parameters, whereby interest may only be in the posterior distribution of the intensity F. A Bayesian approach to this problem shows its principal advantage over other methods of inference.

Many problems can include numerous parameters of interest, but conclusions will often be made from one or only a few parameters at a time (Gelman, Carlin, et al. 2014). A direct approach to this case using Bayesian analysis is to obtain the marginal posterior distributions of the particular parameters of interest. This, in principle, is done by acquiring the joint posterior distribution of all the unknown parameters θ and integrating this distribution over all the unknown parameters that are of no interest to obtain the desired marginal distribution (Gelman, Carlin, et al. 2014,Tiao 1973b). Equivalently, numerical simulations are employed in this approach, whereby samples are drawn from the joint posterior distribution, and only the parameters of interest are taken into account while other unwanted parameters are ignored. These kinds of unwanted parameters are called *nuisance* parameters.

To be more formal, given $\theta = \{\theta_1, \theta_2\}$, suppose that we are only interested in inference for θ_1 so that θ_2 may be regarded as a *nuisance* parameter. For example in a normal distribution case,

$$D \sim N(\mu, \sigma^2) \tag{2.21}$$

In which both $\mu \equiv \theta_1$ and $\sigma^2 \equiv \theta_2$ are unknown, interest is typically in the posterior distribution of the mean μ. Therefore if we seek the conditional distribution of the parameter of interest θ_1 given the observed data D, i.e., $P(\theta_1|D)$. This is derived from the joint posterior density (2.5),

$$\mathcal{P}(\theta_1, \theta_2|D) \propto \mathcal{P}(D|\theta_1, \theta_2)\mathcal{P}(\theta_1, \theta_2) \tag{2.22}$$

by integrating over θ_2:

$$\mathcal{P}(\theta_1|D) = \int \mathcal{P}(\theta_1, \theta_2|D)d\theta_2 \tag{2.23}$$

alternatively using eq.(2.6), the joint posterior density can be factored to yield,

$$\mathcal{P}(\theta_1|D) = \int \mathcal{P}(\theta_1|\theta_2, D)\mathcal{P}(\theta_2|D)d\theta_2 \tag{2.24}$$

Bayesian Inference Problem

So far we know that the significant value of Bayesian analysis is to compute the posterior distribution, $\mathcal{P}(\theta|D)$, to perform inference on a particular parameter θ_i. However, this is not always a straightforward procedure. The computation of the posterior can be considerably challenging to evaluate due to the structures of both θ and D, which can vary in complexity and dimensions.

The first difficulty occurs in evidence term, $\mathcal{P}(D)$, which is rarely available in closed form. Although in cases where the constant does not matter, such as in parameter estimation problem, inference can be carried out without the constant. Cases such as model selection and testing the constant becomes the central part of the Bayesian analysis for addressing such inferential problems.

The challenge to solve more problems analytically with the full Bayesian analysis becomes difficult since most integrals do not have analytic solutions. Fortunately, a pool of powerful methods has been developed over the past few decades for approximating and sampling from probability distributions. These numerical methods are applied to Bayesian inference to aid with inferential problems that have intractable posterior distributions.

There are many types of these numerical methods in Bayesian inference, to name a few; Importance Sampling (Zhao, Liu, and Gu 2013), Laplace Approximation (Ruli, Sartori, and Ventura 2015), Markov Chain Monte Carlo (MCMC) (Metropolis et al. 1953, Gregory 2005, Hogg and Foreman-Mackey 2017), Nested Sampling (John Skilling 2006) and Variational Bayes (VB) (Blei, Kucukelbir, and McAuliffe 2016). In this thesis, we will focus on the commonly used MCMC and the Nested sampling technique.

2.3.7 Bayesian Evidence

Bayesian inference is used in a wide range of problems in the rise of a data-driven society. In order to learn patterns or what is assumed to generate a particular type of data that can be statistically analyzed, models or hypotheses are required. Therefore we often want to know which model is best, particularly in science. In the Bayesian formalism, this is achieved by evaluating the Bayesian evidence of each model and comparing them through a ratio commonly referred to as a Bayes factor or the odds ratio (Knuth et al. 2014).

In Bayes' theorem (2.9), the Bayesian evidence is the normalizing constant $\mathcal{P}(D|\mathcal{I})$ or the marginalizing factor, which ensures the posterior probability distribution sums up to one (Tiao 1973b, Mukherjee, Parkinson, and Liddle 2006). The evidence can be described as the total probability of obtaining the data D in favor of model H (MacKay 2003). It is evaluated by summing or integrating (marginalization discussed in 2.3.6) over all possible parameters

values θ of the model H as defined in (2.10) by,

$$\mathcal{Z} = \int \mathcal{P}(D|\theta, H)\mathcal{P}(\theta|H)d\theta \tag{2.25}$$

(2.25) shows that the evidence, $\mathcal{Z}$, is given by the average of the likelihood function with respect to the prior. Therefore a model will have an enormous evidence value if most of its allowed parameter space is more likely, given data. On the contrary, a model will have a small evidence value if large portions of the allowed parameter space have small likelihood values (M.P. Hobson and McLachlan 2003,Murray and Ghahramani 2005).

Tiao 1973b, Murray and Ghahramani 2005 and Lochner et al. 2015 discuss that the value of the evidence embodies what is called Occam's razor[1], that is: A simpler theory, which has a good defined parameter space, will generally have a larger evidence value than a more complex theory, unless the latter explains the data better. A visual representation to support the explanation is provided in fig. (2.1).

[1] 'Simpler solutions are more likely to be correct than complex ones'

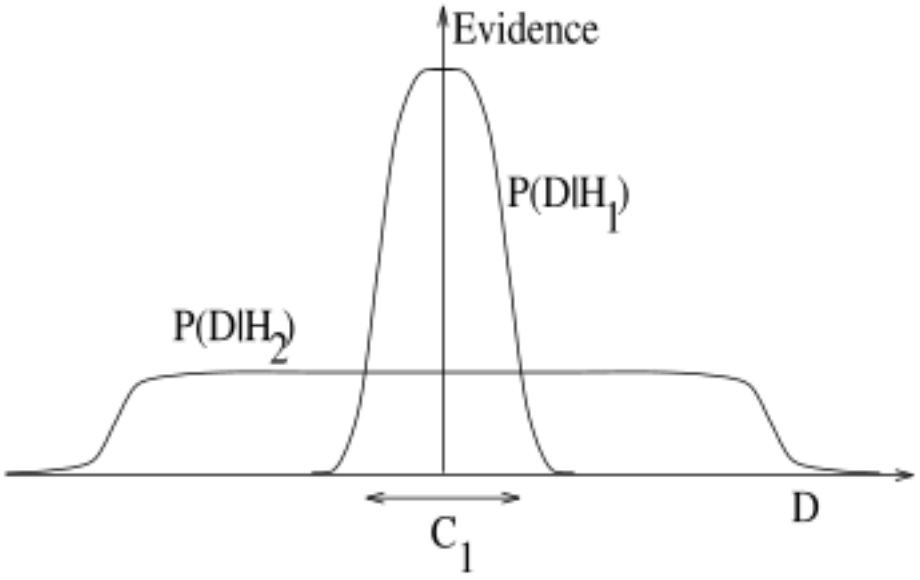


Figure 2.1: The D-axis represents the possible data sets D. Each curve provides a probability distribution over the different portions of the data sets in D. C_1 is assumed to be the region with the most likely data sets that occurred (the given data). H_1 is a simple model that predicts the most likely data sets in C_1. Therefore, H_1 has larger evidence than a more complex model H_2, which would generally be favored given its complexity. For example, if H_2 has more free parameters, it can predict various data sets. However, it does not predict data sets in the region C_1 as confident as H_1 (Murray and Ghahramani 2005).

Given two models H_1 and H_2, which compete to solve a particular problem for given data D, their respective relative evidence can be used for comparison. This provides a significant reason to select which model best suits the problem in question. It leads to the concept of the *Bayes factor* or the odds ratio introduced by Harold Jeffrey (Jeffreys 1946) given by,

$$B_{12} = \frac{\mathcal{P}(D|H_1)}{\mathcal{P}(D|H_2)} = \frac{\int \mathcal{P}(D|\theta_1, H_1)\mathcal{P}(\theta_1|H_1)d\theta_1}{\int \mathcal{P}(D|\theta_2, H_2)\mathcal{P}(\theta_2|H_2)d\theta_2} \tag{2.26}$$

where $\mathcal{P}(D|H_i)$, is the marginal likelihood of the data each model $(i = 1, 2)$, respectively.

Bayesian model selection makes use of the Bayes factor by calculating the posterior odds ratio.

$$\frac{\mathcal{P}(H_1|D)}{\mathcal{P}(H_2|D)} = B_{12}\frac{\mathcal{P}(H_1)}{\mathcal{P}(H_2)} \tag{2.27}$$

The Bayes factor provides a measure of the evidence for model H_1 versus model H_2 (Jeffreys 1946). For example, if $B_{12} = 3$, then the data favors H_1 over H_2 with 3:1 odds ratio. Jeffreys (Jeffreys 1946) recommended a scale of evidence for interpreting Bayes factor, given in table (2.1).

Bayes factor	Interpretation
$B_{12} < 1/10$	Strong evidence for H_2
$1/10 < B_{12} < 1/3$	Moderate evidence for H_2
$1/3 < B_{12} < 1$	Weak evidence for H_2
$1 < B_{12} < 3$	Weak evidence for H_1
$3 < B_{12} < 10$	Moderate evidence for H_1
$B_{12} > 10$	Strong evidence for H_1

Table 2.1: Jeffrey's Scale of Evidence for Bayes Factors

Bayesian model comparison does not replace the parameter inference analysis that is performed with each model independently. Alternatively, model comparison extends the analysis for the assessment of hypotheses in the knowledge of available data (Trotta 2008).

Evidence computation

As mentioned in section (2.3.6), in general, the computation of the Bayesian evidence is a numerically challenging task. This is because it involves a multi-dimensional integration over the whole parameter space. Other problematic issues arise when the likelihood function is complex or multi-modal (Trotta 2008,John Skilling 2006). Until recently, the application of Bayesian model comparison has always suffered due to the difficulty of reliably estimating the evidence. There have been several methods developed to solve this issue, each with its own depths and domain of applicability. Commonly used methods are, thermodynamic integration (Gregory 2005) and nested sampling developed by John Skilling 2006. In this thesis we make use of the nested sampling technique.

2.4 MCMC

Markov Chain Monte Carlo (MCMC) is a numerical technique that is capable of estimating the desired posterior distributions in high-dimensional models and complex systems. MCMC was first introduced in the early 1950s by statistical physicists Metropolis et al. 1953 as a method for the simulation of simple fluids. A Monte Carlo method is a statistical sampling technique that relies on randomly sampling a given distribution to obtain a numerical representation of the distribution (Gregory 2005). Monte Carlo methods are broadly used in all areas of science to simulate complex systems and to evaluate integrals in multi-dimensions.

2.4.1 Markov chain sampling

The most common methods used for sampling from general distributions in high-dimensions are based on Markov Chains (Gregory 2005). The basic principle of a Markov Chain is to form a rule for choosing a sequence of points in parameter space, θ, such that after a sufficient number of samples, the probability of a position θ_i is proportional to the probability at position $\mathcal{P}(\theta_i)$. If the rule for moving from θ_i to θ_{i+1} depends only on the knowledge of θ_i, then the sequence of points is called a Markov chain (Lewis and Bridle 2002). A significant portion of these points can be later considered 'samples' after certain conditions are satisfied.

A Markov chain moves from a position in a parameter space θ_i to the next position θ_{i+1} with a probability distribution called a transition distribution $T(\theta_i, \theta_{i+1})$ (Lewis and Bridle 2002,Gregory 2005). The transition probability determines the probability of the chain moving from θ_i to θ_{i+1}. To ensure that a chain converges to a desired distribution, the probability of moving to a point θ_{i+1} must be equal to the probability of being at that point $\mathcal{P}(\theta_{i+1})$, that is,

$$\mathcal{P}(\theta_{i+1}) = \int \mathcal{P}(\theta_i) T(\theta_i, \theta_{i+1}) d\theta_i \tag{2.28}$$

To ensure that this holds, a condition called *detailed balance* condition should be satisfied. The detailed balance condition mathematically states,

$$\mathcal{P}(\theta_{i+1})T(\theta_{i+1},\theta_i) = \mathcal{P}(\theta_i)T(\theta_i,\theta_{i+1}) \tag{2.29}$$

where T is a normalized transition probability distribution.

In words that is, the probability of being at θ_i and going to θ_{i+1} is the same as the probability of being at θ_{i+1} and coming back to θ_i (Lewis and Bridle 2002,Gregory 2005). There are several ways in which the transition probability T can be chosen so that detailed balance holds (Gregory 2005).

An important note is that even if the position at any given time follows a correct asymptotic distribution, $\mathcal{P}(\theta_i)$ and $\mathcal{P}(\theta_{i+1})$ are not entirely independent of each other. So only running a Markov process for a sufficient number of steps until it reaches equilibrium and then stopping it will generally give a single independent sample from the target distribution. Therefore, ideally, to generate many independent samples, it is necessary to run the chain longer until it has lost all memory of the previous independent sample (Michael Hobson et al. 2009).

2.4.2 Metropolis—Hasting algorithm

The Metropolis-Hasting algorithm developed by Hastings 1970 is the most common way of implementing Markov Chains that satisfies detailed balance. MCMC methods do not require a full analytic description of the pdf for sampling to happen; they only require that a ratio of the pdfs are computable at pairs of locations.

In practice, the M-H MCMC algorithm requires two inputs. The first being the function $f(\theta)$ that is to be sampled, such that the algorithm can evaluate $f(\theta)$ for any value of the parameters θ. In a Bayesian analysis context, this function would be the prior $\mathcal{P}(\theta)$ times the likelihood $\mathcal{P}(D|\theta)$, evaluated at the observed data D (since inference does not depend on the evidence). The second critical input is a *proposal* pdf function $q(\theta_{+1}|\theta_i)$ that can produce samples, such that the algorithm can draw a new positions θ_{i+1} in the parameter space given an 'old' position θ_i (Hogg and Foreman-Mackey 2017). The proposed new point is then accepted with a probability

$$\alpha(\theta_i,\theta_{i+1}) = min\left\{1, \frac{P(\theta_{i+1}q(\theta_{+1},\theta_i)}{P(\theta_i)q(\theta_i,\theta_{i+1})}\right\} \tag{2.30}$$

So that

$$T(\theta_i, \theta_{i+1}) = \alpha(\theta_i, \theta_{i+1}) q(\theta_i, \theta_{i+1})$$

This construction ensures that detailed balance discussed in section (2.4.1) is conserved, that is,

$$\mathcal{P}(\theta_{i+1}) T(\theta_{i+1}, \theta_i) = \mathcal{P}(\theta_i) T(\theta_i, \theta_{i+1}) \tag{2.31}$$

An illustration of the M-H MCMC algorithm is as follows:

Algorithm 1: Random Walk Metropolis - Hasting

Initial starting vector, θ_0
for $i = 1, 2, .., N$ **do**
  Sample $\Delta\theta$ from a proposal distribution $q(\Delta\theta|\theta)$
  $\theta_{trial} = \theta_i + \Delta\theta$
  Draw $\alpha \sim \mathcal{U}[0, 1]$
  if $\alpha < min\left\{1, \frac{\mathcal{P}(\theta_{trial}) q(\theta_{trial}, \theta_i)}{\mathcal{P}(\theta_i) q(\theta_i, \theta_{trial})}\right\}$ **then**
    $\theta_{i+1} = \theta_{trial}$
  else
    $\theta_{i+1} = \theta_i$
  end
end

The series of steps initially starts with a random position in the parameter space. It will take a few steps called *burn-in*, to equilibrate before it starts sampling from the posterior distribution. After that, each chain position is a correlated sample from the posterior. However, by giving the chain enough time to run longer, this allows the chain to move to an uncorrelated position in parameter space, and independent samples can then be obtained (Lewis and Bridle 2002).

The proposal density can either be asymmetrical or symmetrical. If it is symmetrical, it cancels out when evaluating the acceptance probability (2.30), which then becomes a ratio of the posteriors. This is the case when the proposal density is independent of the current posterior of the chain (Hogg and Foreman-Mackey 2017).

Typically the ideal proposal distribution would be of a similar shape to the posterior. However this is generally unknown in advance, so the acceptance ratio can be used to heuristically tune the proposal distribution. Moreover, monitoring the acceptance ratio, determined by the number of accepted samples divided by the number of steps of the chain, also plays a role in the efficient exploration of the target distribution. That is, if the proposal steps (assuming a gaussian proposal distribution) are very small, then it will take longer to explore the target distribution (Hogg and Foreman-Mackey 2017). On the other hand, if the steps are too big, then the peaks of the target distribution remain unexplored. So if the acceptance ratio is too high, for the former case, then the chain is not mixing efficiently, and if it is too small, then the acceptance ratio will be small and many samples get rejected. A typical recommendation of the acceptance ratio range is between 20% - 60% (Hogg and Foreman-Mackey 2017).

2.4.3 Choosing a step-size

The MCMC implementation in this thesis employs a dynamic step size, which is adjusted only before convergence (further discussed in section (2.4.5)) has been checked. The stepsize is changed in accordance with the acceptance ratio. The step-size is refined to let the acceptance ratio converge around 50%. This is given by the following conditions :

- If the number of accepted samples (N_{accept}) is greater than the number of rejected samples (N_{reject}), then the step-size is multiplied by a factor of $\exp(\frac{1.0}{N_{accept}})$.
- Conversely, if the number of rejected samples is greater than the number of accepted samples, then the stepsize is divided by a factor of $\exp(\frac{1.0}{N_{reject}})$.

It is, therefore, with utmost importance to note that adjusting the stepsize in MCMC completely breaks detailed balance, discussed in section (2.4.1). Moreover, only the last fraction of the MCMC chain samples are used for which the stepsize was last adjusted.

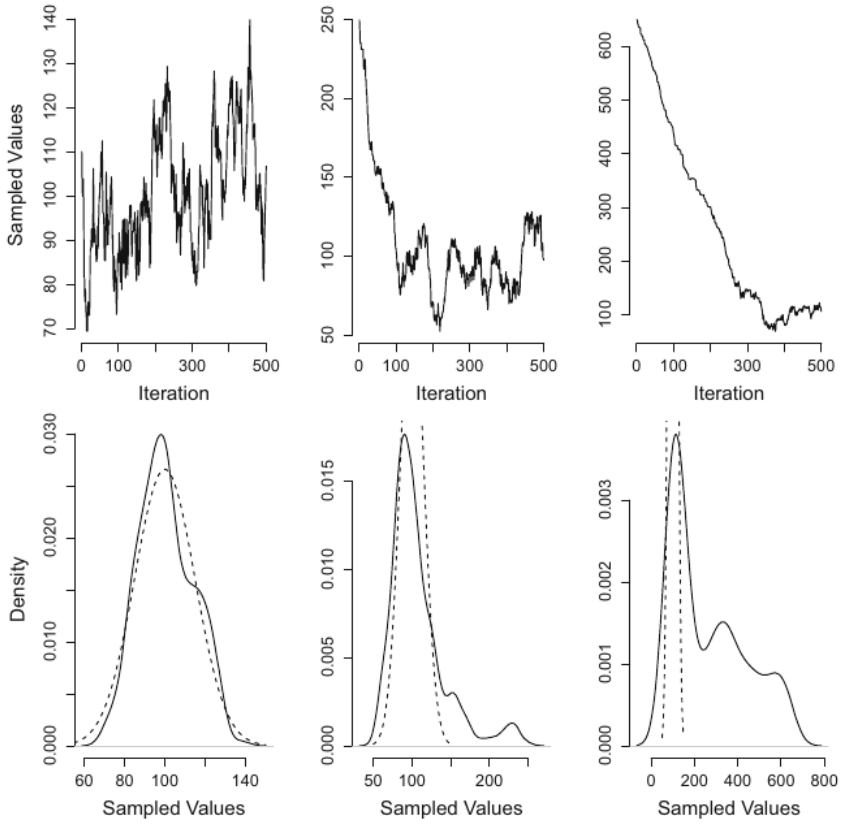


Figure 2.2: An example of MCMC outputs. **Left column**: Shows a sampling chain starting from a good starting value, the mode of the target distribution. **Middle-column**: Shows a sampling chain starting from a starting value in the tails of the target distribution. This chain would require the burn in to be removed before the posterior can be estimated. **Right-column**: Shows a sampling chain starting from a value far from the target distribution. **Top-row**: Markov Chain. **Bottom-row**: Sample density. The dashed line indicates the target distribution, and the solid lines is the sample distribution from the MCMC chain. Courtesy of Ravenzwaaij, Cassey, and Brown 2018

2.4.4 Burn-in

Almost every MCMC method requires a choice about the initialization. The chain/s need to be started somewhere in the parameter space. Ideally, a chain will be initialized in a high posterior region of the parameter space. However, it can happen that at high dimensions or in complex problems, it becomes difficult to set the optimum initialization (Gelman and Rubin 1992). Therefore it is likely the chain will start far from the peak of the distribution

and the first part of the chain will need to be discarded. *Burn-in* refers to discarded initial portion of a Markov Chain sample so that the effect of the initial values is minimal on the posterior inference.

In situations where the target distribution may be multi-modal, then ideally starting at multiple starting values in the parameter space is employed and the resulting chains compared to each other (Gelman and Rubin 1992). In general, MCMC performs poorly in the presence of multimodal posteriors, and a technique like nested sampling is preferable.

2.4.5 Convergence

Whilst the MH-MCMC algorithm is guaranteed to converge to the true posterior after an infinite number of steps, in practice of course only a finite number of steps may be taken. It is nearly impossible to know that the full posterior distribution in an MCMC chain is sampled (Hogg and Foreman-Mackey 2017). For instance, in an inference problem, if a posterior pdf has two modes that are both essential but separated by a region of low probability, a simple sampler could sample one of these modes very well but will take effectively infinite time to find the other mode. Therefore it is generally practical to rely on a heuristic approach to test for convergence.

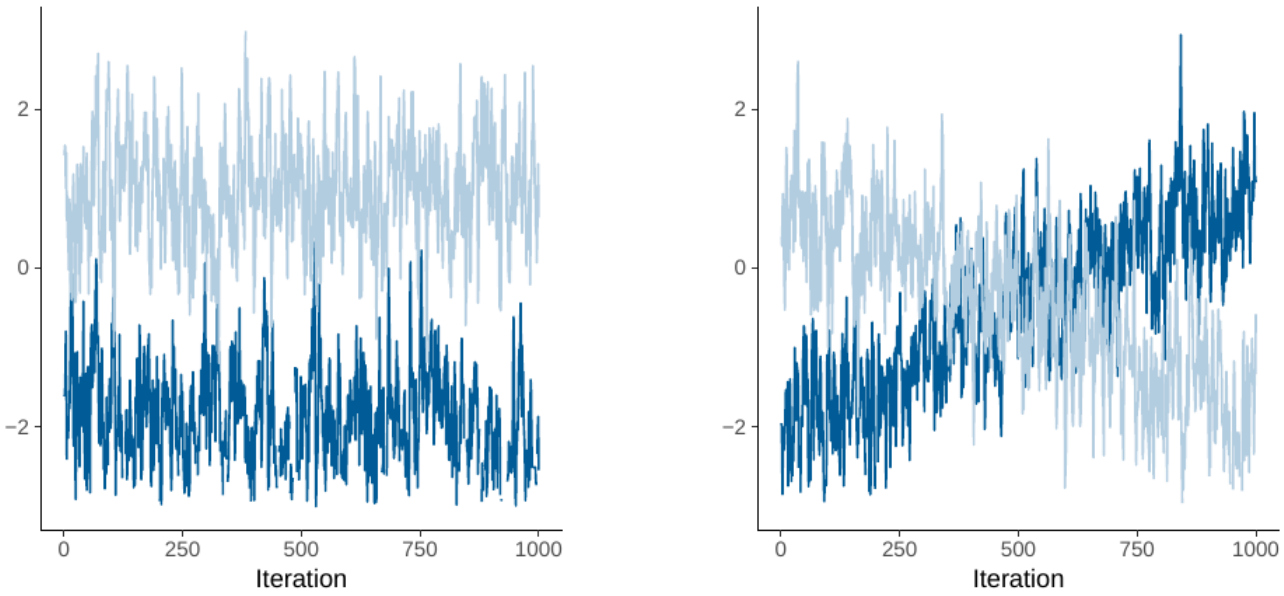


Figure 2.3: Examples of challenging cases in assessing the convergence of an iterative simulation, the light-blue and dark-blue colours shows two MCMC chains. The two chains try to sample the underlying distribution but are started at different places. (a) On the left plot, the two sequences look stable, but the offset of both shows that they have not converged to a common distribution. (b) In the right plot, both the sequences look likely to cover a common distribution, but neither sequence appears stationary. These graphs demonstrate a need for a heuristic approach for assessing convergence. courtesy of Vehtari et al. 2019.

A commonly used method to test for convergence is the Gelman-Rubin diagnostic (Gelman and Rubin 1992). The Gelman-Rubin compares the variance in (one or many parameters) within a chain to the variance across chains (Hogg and Foreman-Mackey 2017,Lochner 2014). The Gelman-Rubin criterion is defined as follows (Gelman and Rubin 1992): Suppose there are M MCMC chains of n length (with burn-in removed). Let $\bar{\theta}_m$ be the mean of the parameter estimates for each chain m and $\bar{\theta}$ the mean of the parameters for all the chains combined. Then let

$$B = \frac{n}{M} \sum_{m=1}^{M} (\bar{\theta}_m - \bar{\theta})^2 \tag{2.32}$$

and

$$W = \frac{1}{M}\sum_{m=1}^{M} S_m^2 \qquad \text{where} \qquad S_m^2 = \frac{1}{1-n}\sum_{t=1}^{n}(\theta_m^t - \bar{\theta})^2 \tag{2.33}$$

B is defined as the between-chain variance, and W is the within-chain variance and the weighted average of W and B, is given by, $\widehat{V}$:

$$\widehat{V} = \frac{n-1}{n}W + \frac{M+1}{nM}B \tag{2.34}$$

If the chains have converged, the ratio $\widehat{V}/W$ called the potential scale reduction factor (PSRF) is close to one, which makes it a useful convergence indicator (Venna, Kaski, and Peltonen 2003). In this thesis, the Gelman-Rubin criteria was used by splitting the chain into three portions and checking that the PSRF is within 1.01.

2.4.6 Metropolis-Hasting - Chain analysis

An advantage of Monte Carlo methods is the ability to obtain samples from a full parameter space (Lewis and Bridle 2002). They allow answering any particular question by examining and computing reliable results from an MCMC output. Eventually, for human understanding and consumption, these results should be presented in a summary of efficient statistics such as expectation values, mean or median, error bars, confidence intervals, and histograms.

The choice of posterior summary statistics to report is usually subjective. The following summaries of the posterior samples are generally presented.

- Trace plots: These plots (e.g. figure (2.3)) show the parameter values as a function of step number. They can be used to select burn-in lengths, provide an indication with the model or sampler, and act as a judge of convergence (Hogg and Foreman-Mackey 2017). However, it is important to note that good-looking trace plots do not guarantee converged sampling.

- Posterior predictive plots: The idea behind posterior predictive plots is that if a model is a good fit, it should be simple to generate data that resembles the data that was

observed. This plot is generated by taking N random samples from the resulting MCMC chain, plot the prediction that each sample makes for the data, and overplot the observed data (Hogg and Foreman-Mackey 2017). It can identify problems with convergence or sampling.

- Scatter matrices or histograms : To show low-level covariances and nonlinearities (Hogg and Foreman-Mackey 2017), for instance, in K-dimensional parameter space, plotting all K-one-dimensional sample histograms and two-dimensional histograms (scatter-plots) can show the relationships of the parameters of interest.

2.5 Nested Sampling

Bayesian inference can be divided into two distinct concepts: Parameter estimation and model selection. Parameter estimation is typically performed using MCMC methods based on the Metropolis-Hasting algorithm and other versions (such as the Gibbs sampling algorithm (Chib 2001)). Generally, in Parameter estimation, the Bayesian evidence $\mathcal{Z}$ (2.10) is ignored since it is independent of the parameters θ. Inference is then only made by approximating samples form the unnormalized posterior only (Feroz and M. P. Hobson 2008). Unfortunately, these methods can be highly inefficient in exploring multi-modal or degenerate distributions. Moreover, MCMC techniques require tuning of the proposal distribution and testing for convergence.

In Bayesian model selection, an estimate of the Bayesian evidence is needed, which requires a multi-dimensional integration over the prior density (John Skilling 2006). Moreover, the degree of importance of the evidence for Bayesian model selection is typically higher than that of parameter estimation. However, MCMC methods are computationally expensive and cannot be used directly to compute the evidence. There are evidence evaluation methods based on MCMC such as thermodynamic integration or simulated annealing, which is extremely computational intensive but has been used in astronomical applications (Heavens et al. 2017).

Nested sampling is a Monte Carlo method developed for efficient evaluation of the Bayesian evidence by John Skilling 2006. It can not only calculate the evidence but produces posterior inferences as a by-product (Feroz and M. P. Hobson 2008,John Skilling 2006).

2.5.1 The Nested Sampling algorithm

The calculation of the value $\mathcal{Z}$ allows different model assumptions to be compared through the ratio of Bayes factor discussed in section (2.3.7) in model selection problems.

The primary task is to evaluate the equation (2.25) that can be written as

$$\mathcal{Z} = \int \mathcal{L} dX \tag{2.35}$$

where $\mathcal{L} = \mathcal{L}(\theta)$ is the likelihood function , $dX = P(\theta)d\theta$ is the element of the prior mass and θ the unknown parameters.

John Skilling 2006 proposed the idea of sorting points θ by their likelihood values which are then summed to yield the evidence. He mentions that nested sampling does this operation statistically and allows the evidence to be accompanied by a corresponding numerical uncertainty.

The calculation of $\mathcal{L}dX$ seems like a simple numerical analysis problem, but when considering the underlying coordinate θ to evaluate $\int \mathcal{L}(\theta)P(\theta)d\theta$, this becomes difficult as soon as θ increases in dimension. Instead, the prior mass X is used directly (John Skilling 2006).

The prior mass X can be accumulated from its elements dX defined by

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} P(\theta)d\theta \tag{2.36}$$

where λ is the iso-likelihood contour , $L(\theta)$ as the accumulant prior mass covering all likelihood values greater than λ. As λ increases, the enclosed mass X decreases from 1 to 0. Therefore, with the inverse function of eq.(2.36) as $\mathcal{L}(X)$, the evidence becomes a one-dimensional integral over a unit range (John Skilling 2006,Feroz and M. P. Hobson 2008).

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX \tag{2.37}$$

Figure (2.4) shows that the integrand is positive and decreasing. The transformation of θ to X involves dividing the unit prior volume into small elements and sorting them by likelihood

(Feroz and M. P. Hobson 2008). The idea of transforming a multi-dimensional integral into a one-dimensional integral is a significant element of Nested Sampling.

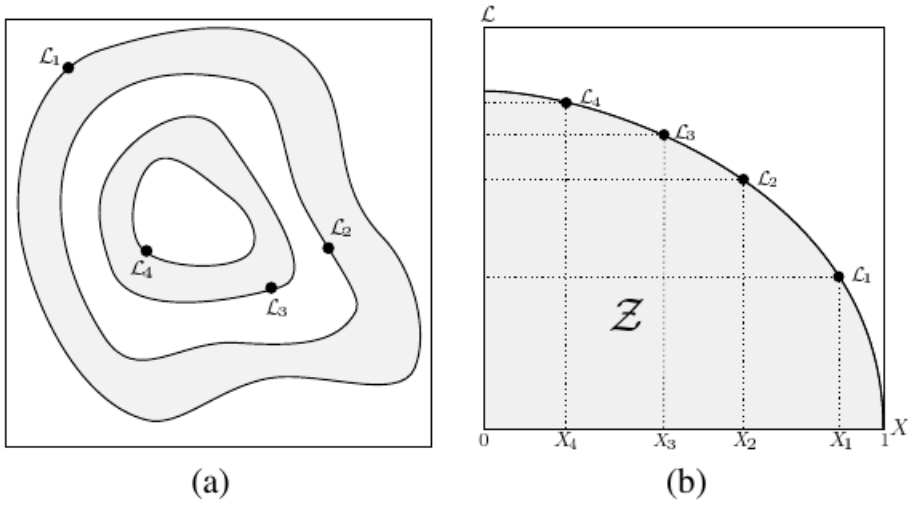


Figure 2.4: The illustration above shows (a) a parameter space of a two-dimensional problem; and (b) the transformed $\mathcal{L}(X)$ function where the prior volumes X_i are associated with each likelihood $\mathcal{L}_i$. Courtesy of Feroz and M. P. Hobson 2008

2.5.2 Integration

Suppose that there is a way of evaluating the likelihood as $\mathcal{L}_i = \mathcal{L}_i(X_i)$ at a right-to-left sequence of m points, i.e

$$0 < X_m < \cdots < X_2 < X_1 < 1$$

Any capable numerical analysis method would be able to estimate $\mathcal{Z}$ as a weighted sum (John Skilling 2006) of these values, given by

$$\mathcal{Z} \longleftarrow \sum_{i=1}^{m} w_i \mathcal{L}_i \tag{2.38}$$

in which the area in fig (2.4,(b)), is approximated as a set of columns of height $\mathcal{L}$ and width $w \sim \Delta X$.

The trapezoid rule is applied, where $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$ (John Skilling 2006,Feroz and M. P. Hobson 2008).

- The condition $w_i = X_i - X_{i+1}$ with $X_{m+1} = 0$, gives the lower bound ,

$$\mathcal{Z} = \int_0^1 \mathcal{L} dX \geq \sum_{i=1}^{m} (X_i - X_{i+1})\mathcal{L}_i$$

- and the condition $w_i = X_{i-1} - X_i$ with $X_0 = 1$, gives the upper bound ,

$$\mathcal{Z} = \int_0^1 \mathcal{L} dX \leq \sum_{i=1}^{m} (X_{i-1} - X_i)\mathcal{L}_i + X_m L_{max}$$

Nested sampling evaluates the summation (2.38) in the following way. In the beginning, an iteration counter i is set to $i = 0$ and N 'live' samples are drawn from the full prior $P(\theta)$ distribution, where the initial prior mass is $X_o = 1$. The samples are then sorted in order of their likelihood $\mathcal{L}$, then the sample with the smallest likelihood value $\mathcal{L}^*$ is removed from the live set and replaced by a new point drawn from the prior subject to the constraint that the point has a likelihood $\mathcal{L} > \mathcal{L}^*$ (John Skilling 2006,Feroz and M. P. Hobson 2008). The prior volume corresponding to the iso-likelihood contour is a random variable distributed as $X_1 = t_1 X_0$, where t_1 follows the distribution for the largest N samples drawn uniformly from the interval [0,1] ($i.e Pr(t) = Nt^{N-1}$). This process is repeated for each iteration i until the prior volume has been examined. The algorithm, therefore, travels through nested shells of likelihood as the prior volume is reduced (John Skilling 2006,Feroz and M. P. Hobson 2008,Speagle 2019).

2.5.3 Nested sampling procedure

Nested Sampling algorithm starts by drawing a population of samples N ('live points') from a prior distribution $P(\theta)$. The likelihood values of these samples are evaluated, and the sample with the lowest likelihood is discarded. The likelihood of the discarded sample is used as a likelihood constraint $\mathcal{L}^*$. An exploration technique such as MCMC is used to generate a new sample to replace the discarded sample, and the new sample is accepted if and only if it has a likelihood higher than that of the discarded sample (i.e., $\mathcal{L} > \mathcal{L}*$). After completion,

the nested sampling algorithm returns the evidence term and the posterior samples as a by-product. A pseudo-code of the nested sampling procedure is shown in algorithm (2).

Algorithm 2: Nested Sampling Algorithm

Input: $\mathcal{N}_{live}, \mathcal{L}_{live}, L$

Output: $\mathcal{Z}$, Posterior Samples

```
/* Initialize 𝒵 and X prior mass                                   */
```
$\mathcal{Z} = 0$

$X_0 = 1$

$i = 0$ // iteration number

while $i <$ iterations **do**

  $\mathcal{L}^* = \min(\mathcal{L}_{live})$ // Minimum loglikelihood

  $\mathrm{X}_i = \exp(\text{-i/N})$ // Prior mass

  $w_i = X_{i-1}$ - X_i // Prior mass width

  $\mathrm{Z} = \mathrm{Z} + \mathcal{L}^* w_i$ // Update Evidence

  Samples $\leftarrow \theta^*, \mathcal{L}^*$ // Save worst sample details

  /* Choose random index j in $[0, i^*) \cup (i^*, N]$ */

  $\theta_j = \mathcal{N}_{live,j}$ // Copy random sample from survivors

  `Update = Explorer(`$\theta_j, \mathcal{L}^*$`)` // Explore , new sample

  /* Update worst sample */

  $\theta^*, \mathcal{L}^*$ = `Update`

  `i = i + 1` */

  /* /* Check whether to stop. */

  `Evaluate stopping criterion`

2.5.4 Nested Sampling Termination

The nested sampling algorithm usually terminates after a pre-set number of steps or when the largest current likelihood taken over the full existing chain, would not increase the current evidence by more than some small fraction (John Skilling 2006). John Skilling 2006 and Feroz and M. P. Hobson 2008 provide a robust explanation and further techniques for implementing a termination condition. The latter approach is employed in this thesis.

John Skilling 2006, Feroz and M. P. Hobson 2008 and Speagle 2019 mention that by selecting the maximum likelihood $\mathcal{L}_{max}$ in the set of live points, it is safe to assume that the largest evidence contribution that can be made by the remaining portion of the posterior can be

given by the product of the remaining prior volume and maximum likelihood value, ie.,

$$\Delta\mathcal{Z}_i \approx \mathcal{L}_{max} X_i \tag{2.39}$$

This can be transformed into a relative stopping criterion by using the ratio between the current estimate evidence $\mathcal{Z}_i$ and the remaining evidence $\Delta\mathcal{Z}_i$ (log-space),

$$ln(\mathcal{Z}_i + \Delta\mathcal{Z}_i) - ln\mathcal{Z}_i < \epsilon \tag{2.40}$$

stopping at a pre-defined user value ϵ.

Fundamentally, this error estimate serves as an upper bound (since X_i is not entirely bound) that can be used for determining when to stop sampling from a target distribution with estimating the evidence (John Skilling 2006, Speagle 2019).

2.5.5 Evidence error estimation

The method proposed and discussed by John Skilling 2006 and Feroz and M. P. Hobson 2008 for error evidence error estimation is applied in this thesis. This type of method provides an error estimation in a single sampling procedure and is far less computationally expensive.

The method proceeds as follows. The general behavior of the evidence increment $\mathcal{L}_i w_i$ is initially to rise with iteration number i, with the likelihood $\mathcal{L}_i$ increasing as the weight w_i decreases. At a certain point, the likelihood $\mathcal{L}$ flattens off sufficiently such that the weight dominates the increase in likelihood, then $\mathcal{L}_i w_i$ reaches a maximum and then starts to drop with the iteration number. The final evidence is then dominated by where the bulk of the posterior mass is found, which usually comes from the exploration around the maximum point (John Skilling 2006,Feroz and M. P. Hobson 2008). This occurs in the region of $X \approx e^{-H}$, where H also known as the information,

$$H = `information' = \int ln\left(\frac{dP}{dX}\right) dX \approx \sum_{i=1}^{m} \frac{\mathcal{L}_i w_i}{\mathcal{Z}} ln\left(\frac{\mathcal{L}_i}{\mathcal{Z}}\right) \tag{2.41}$$

where P is the posterior. Furthermore, since $lnX_i \approx \frac{(-i \pm \sqrt{i})}{N}$, Feroz and M. P. Hobson 2008 mentions that the procedure is expected to take about $NH \pm \sqrt{NH}$ steps to shrink down to the bulk of the posterior. Therefore, the dominant uncertainty in $\mathcal{Z}$ is due to the Poisson variability $NH \pm \sqrt{NH}$ in the number of steps to reach the posterior. Consequently, the acquired lnX_i values will have a standard deviation uncertainty of $\sqrt{H/N}$. This uncertainty is then transmitted to the evidence $\mathcal{Z}$ by equation(2.38), so that $ln\mathcal{Z}$ has a standard deviation uncertainty of $\sqrt{H/N}$ (John Skilling 2006). That is,

$$ln\mathcal{Z} = ln\left(\sum_{i=1}^{m} \mathcal{L}_i w_i\right) \pm \sqrt{\frac{H}{N}}. \tag{2.42}$$

2.5.6 Posterior samples

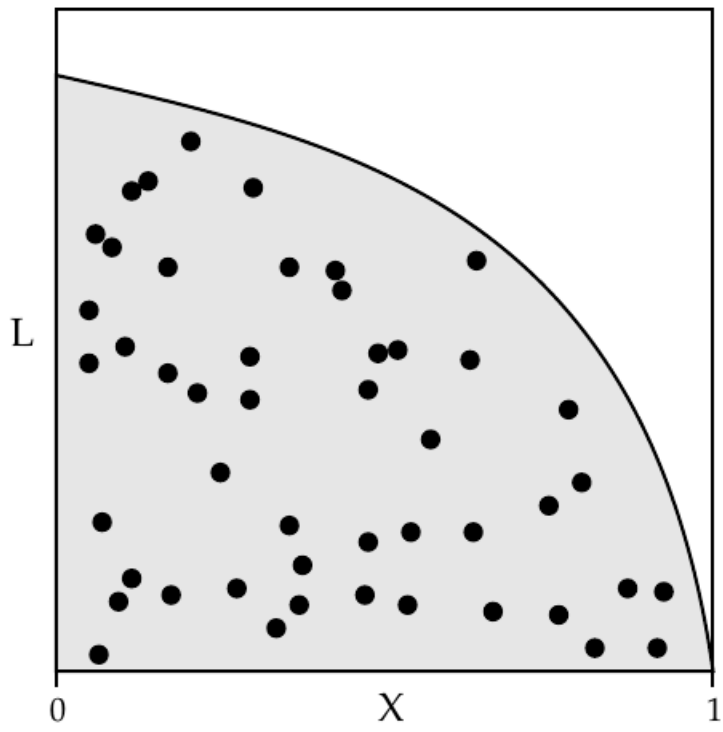


Figure 2.5: Posterior samples scattered randomly over the area $\mathcal{Z}$. Illustration by John Skilling 2006

After obtaining the primary objective, the evidence $\mathcal{Z}$, the posterior can be obtained from the full sequence of the discarded points from the nested sampling procedure. Each sample is assigned an importance weight p_i normalized by $\mathcal{Z}$ (John Skilling 2006), given by

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}. \tag{2.43}$$

Therefore to obtain equally-weighted posterior representatives, only samples with probability p_i/K , where $K \leq max(p_i)$ are accepted. The samples can then be used to present inferences of statistical summaries such as means, standard deviations, confidence intervals, or to obtain marginalized posterior distributions.

Since not all samples are usually used. Therefore the maximum number of representative posterior samples is given by the Shannon channel capacity (John Skilling 2006) as,

$$\mathcal{N} = \exp\left(-\sum_{i=1}^{m} p_i \log(p_i) \right) \tag{2.44}$$

2.5.7 Exploration

One of the main challenges in implementing nested sampling is drawing a sample from the prior within a hard likelihood constraint $\mathcal{L} > \mathcal{L}^*$. Implementing a naive approach that samples blindly from the prior can result in a steady decrease in the acceptance rate of new samples with decreasing volume (Feroz and M. P. Hobson 2008). Different approaches, such as MCMC (John Skilling 2006), ellipsoidal sampling (Feroz and M. P. Hobson 2008), and slice sampling (Handley, M. P. Hobson, and Lasenby 2015), have been implemented to tackle this task for different applications. In this thesis, we focus on this part of nested sampling for object detection.

2.5.8 Applications of Bayesian analysis in astronomy

A vast number of astronomy problems have been solved with MCMC methods.

Lewis and Bridle 2002 implemented a fast MCMC exploration technique in cosmology for parameter estimation called cosmoMC. In their work, they applied cosmoMC to evaluate different varying models with at most 11 parameters. Figure (2.7) shows 1-dimensional marginalised posteriors of the cosmological paramaters. They found that the Monte-Carlo approach works well, though estimating best-fit parameters did not efficiently identify best-fit models to high accuracy. They also found difficulties with investigating posterior distributions with multiple local minima. However, they found that theoretical predictions can be computed essentially correct using the MCMC method as compared to other approaches. Figure (2.6) shows posterior distributions of the Hubble constant and the matter-density computed using cosmoMC using the 2018 planck measurements (Aghanim et al. 2018).

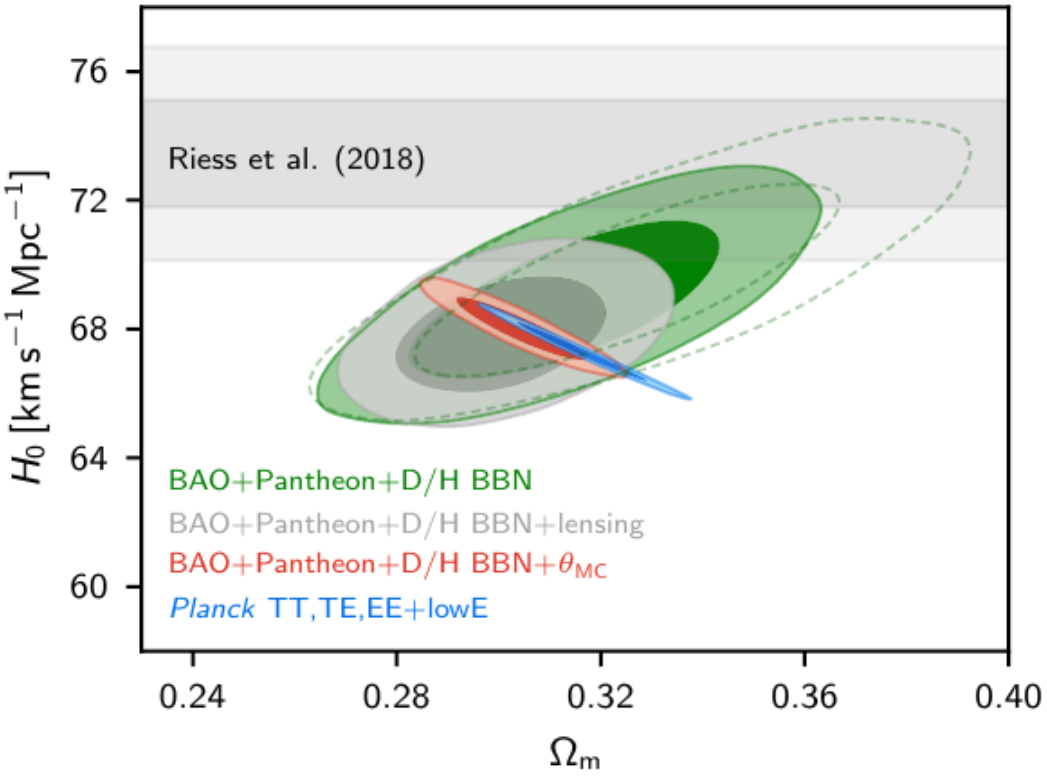


Figure 2.6: Shows posterior distributions for the Hubble, H_0, and matter-density, Ω_m, parameters in the ΛCDM model. The contour intervals are for 68% and 95% probability. The posterior distributions were obtained using cosmoMC (Lewis and Bridle 2002) using the current full Planck measurements of the cosmic microwave background (CMB) (Aghanim et al. 2018). Courtesy of (Aghanim et al. 2018)

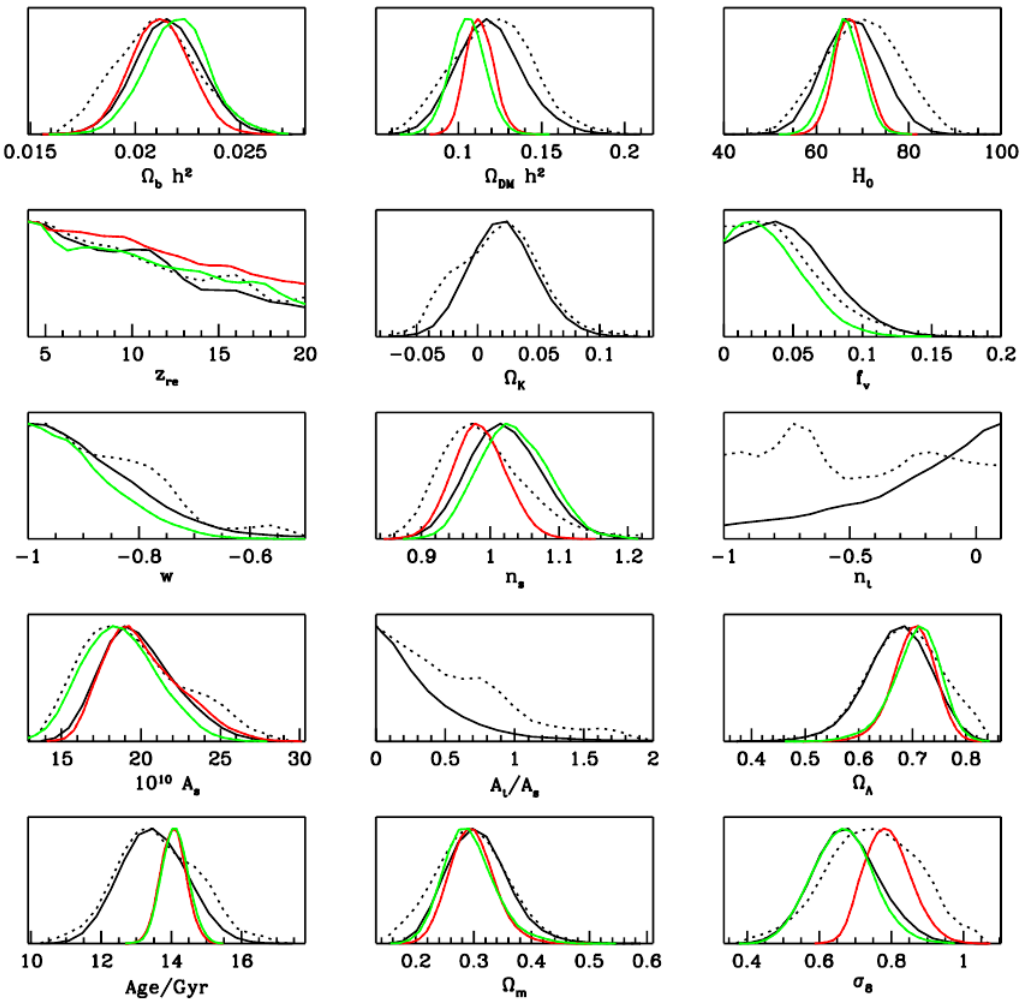


Figure 2.7: An example of how MCMC can be used to constrain fundamental parameters in cosmology. This plot shows the posterior results of the cosmological parameters investigated by Lewis and Bridle 2002. The (black lines) show non-flat models and the (red and green) lines show parameter models (see Lewis and Bridle 2002 for further information). The solid lines show the fully marginalized posterior, and the dotted lines show the relative likelihood values of the samples. The curves were generated using MCMC sampling. Produced by Lewis and Bridle 2002.

Further ground-breaking work was done using MCMC techniques in gravitational waves detection. van der Sluys et al. 2008 implemented an MCMC technique for performing parameter estimation of 12 physical parameters of gravitational waves. They developed these techniques as statistical analysis products for observatories such as LIGO and Virgo. After a few years, the Laser Interferometer Gravitational-Wave Observatory (LIGO) - Virgo collab-

oration made its first direct detection of gravitational waves (see Abbott et al. 2016) where ultimately the MCMC techniques were used to determine the parameters of the system, such as the masses of the two black holes, to high statistical significance.

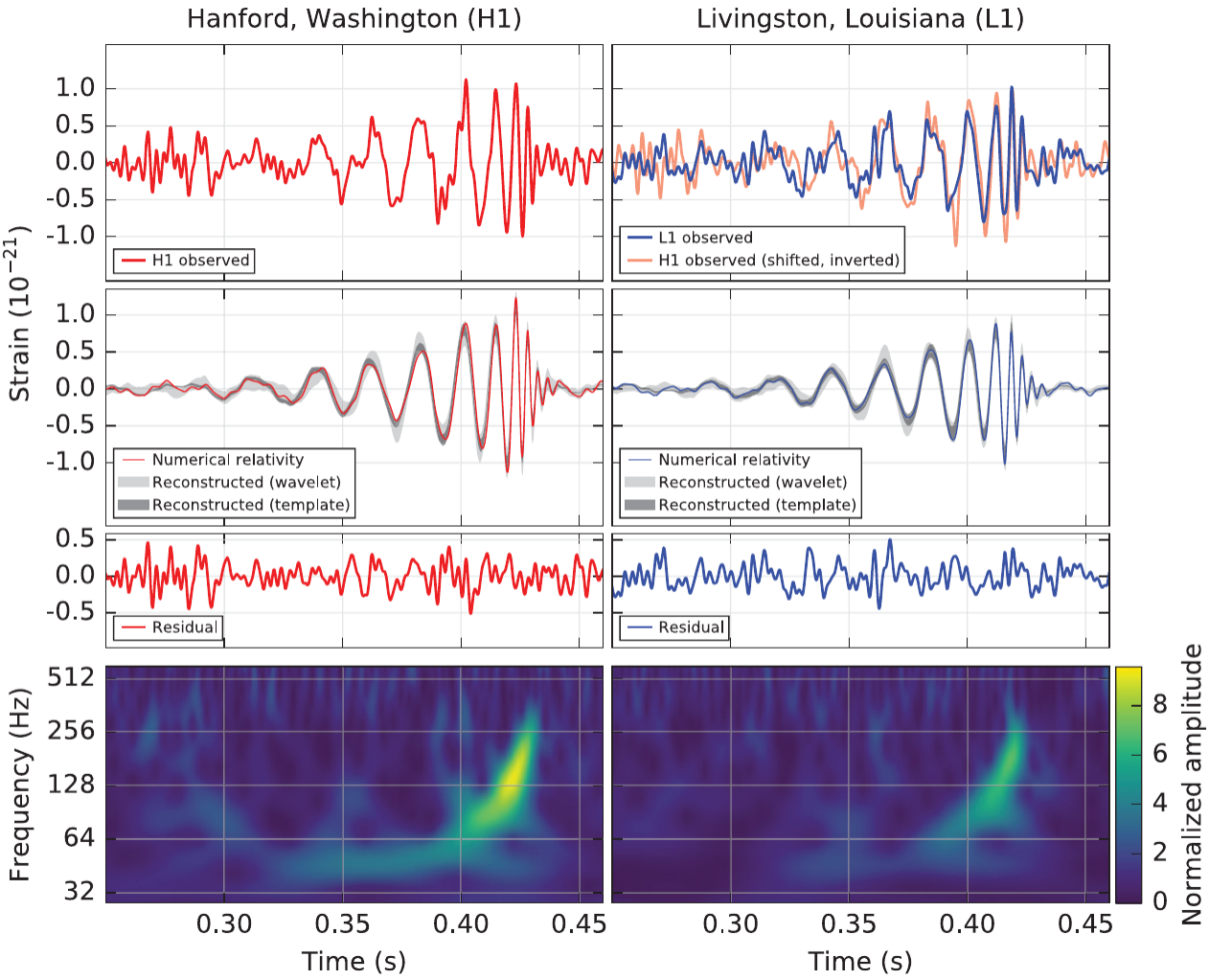


Figure 2.8: Shows the gravitational-wave event observed by LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. The statistical analysis was done with MCMC sampling techniques developed by van der Sluys et al. 2008. Adapted from Abbott et al. 2016

Work in the radio astronomy regime has also been done using MCMC methods. Lochner et al. 2015 developed a technique called Bayesian Inference for Radio Observations (BIRO), which implements a full Bayesian analysis on raw radio data without filtering out any information and dealing efficiently with nuisance parameters (e.g., instrumental noise). The technique

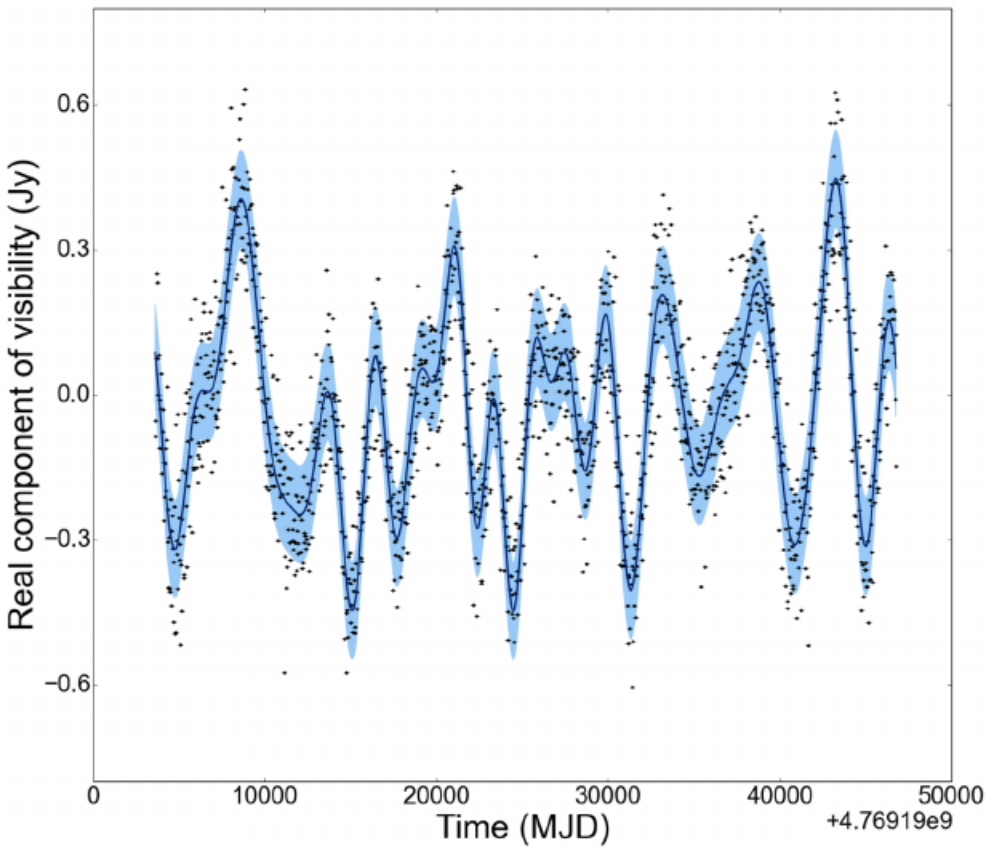


Figure 2.9: An example of the best fitting model and its uncertainties from applying BIRO to raw radio data (Lochner et al. 2015). The black points show the real component of the visibilities for a single baseline and a single channel for the data (raw data). The solid dark-blue line shows the best fitting model with a band of the uncertainty of the original noise added to the simulation in pale blue.

uses a Metropolis-Hasting MCMC and Nested Sampling method which they tested on for at most 103 parameters for parameter inference and model selection. Using a full Bayesian analysis and implementing sampling methods such as MCMC, they found efficient performance parameter inference without any loss of information and accurate model selection. They further used nested sampling to distinguish between extended and point sources.

Chapter 3

FireFly - a new Bayesian approach to source finding

One of the main difficulties with source finding is that in general, the true number of sources present in an astronomical image is unknown. This makes Bayesian source finding in particular quite challenging as a model is always required for Bayesian inference. FireFly tackles this issue by employing a single dynamic model that allows sources to be switched 'on' or 'off' and sufficiently estimating the number of true sources.

3.1 Similar Work

Similar work done by Daylan, Portillo, and Finkbeiner 2017 in probabilistic cataloging implements a similar idea in estimating the source count. They developed a sampler to infer point sources in a given photon count map by sampling from the probability distribution of the underlying catalog space. Daylan, Portillo, and Finkbeiner 2017 assumes that the number of sources is unknown a priori, which allows N, number of sources, to be subject to inference. They use the method known as the Reversible Jump MCMC(RJMCMC) (Fan and Sisson 2010), which allows the freedom to jump across models in a group of models. Then the sampler can to propose updates that change N, by adding or removing sources. This is done within the sampler as a set of possible proposal types (Daylan, Portillo, and Finkbeiner 2017, Portillo et al. 2017). The result then produces an ensemble of catalogs consistent with the count data, which describes the probability distribution of the catalog

given the image. While reverse jump MCMC is very general, it is complex to implement and prone to bugs and coding errors. Here we introduce FireFly which has the benefits of RJMCMC while being very simple.

3.2 Search Algorithms

The main challenging task with doing source finding is detecting the number of 'real' sources in the image without identifying any excessive fake sources or missing them completely, i.e. estimating N_{objs}, number of sources. In this work, two approaches were compared, which are an evidence-based search method and a new technique called FireFly.

3.2.1 Evidence-Based Algorithm

We implemented Nested Sampling (John Skilling 2006) with MCMC (Hogg and Foreman-Mackey 2017) as an evidence-based search algorithm. The nested sampling algorithm performs a death-birth of a sample at each iteration. As described in section (2.5.3), a sample gets discarded 'death' from the live samples, at each iteration, a new sample then gets generated, 'birth', using standard MCMC, and the live samples get updated accordingly. The algorithm then spawns the evidence $\mathcal{Z}$ and a full sequence of discarded points.

The evidence $\mathcal{Z}$ and the discarded points can then be used to construct marginalized posterior distributions. The evidence $\mathcal{Z}$ not only serves as a marginalizing constant, it is also used for model comparison. The aim in this case is to identify and isolate different modes of the posterior which correspond to each source, and the number of modes evident in the posterior distribution is equivalent to the number of sources in the image, N_{objs}. The posterior distribution is an estimate of the underlying model of the real image. So to investigate which model best fits the actual image, the evidence $\mathcal{Z}$ can be assessed for numerous models. This investigation is done by running different models that differ by N_{objs}, and obtaining the evidence for each, then the model with a high $\mathcal{Z}$ is more likely to describe the underlying true sources precisely.

This evidence-based search method allows the best model to be used for estimating the

unknown parameters ($\mathrm{N}_{objs}, \theta_1, \theta_2, ..., \theta_{N_{objs}}$).

3.2.2 FireFly Algorithm

The main challenge with implementing the evidence-based search algorithm is that a significant amount of different models should be evaluated to choose the best one eventually. This approach becomes inefficient with an increasing number of models, which increases the number of unknown parameters as the sources increase with each different model.

An alternative approach is introduced in this work. While it is not possible to know a priori the number of sources, in astronomy it is nonetheless in general possible to estimate the expected number from some prior knowledge.

Our algorithm, FireFly, works by fixing the number of sources to be roughly double the expected number of sources, thus providing a well-defined model, but gives each source the ability to switch on and off (i.e. to set the flux to zero). Eventually the algorithm converges to the real number of sources and provides a full posterior for the position and amplitude of each source.

If an observation of a part of the sky is conducted and a prior can be formulated on the expected number of sources in the observation, denote as $N_{expected}$, then a model can be generated which consists of the double amount of expected sources, $2N_{expected}$. Using eqn. (4.1) and (4.2), we can, therefore, write the model equation as :

$$Model = \sum_{k=1}^{2N_{expected}} \tau(\theta_k) \tag{3.1}$$

However, to get to the 'real' number of sources, $N_{expected} \equiv N_{objs}$, we need to be able to switch off the other half model sources and eventually converge to the 'real' number of sources. Since we need to switch off half of the model sources, then 50% would be our threshold of switching off. Moreover, only switching off is not significant because more sources may get switched off, and other sources could get missed entirely. Therefore, there should be a 50% ratio of either switching on or off a source.

Switching on-off

The 'on' and 'off' switching technique is performed in the generation of a new sample within nested sampling. The basic formulation of nested sampling is that sampling happens within a hard constraint likelihood, that is a new sample is only accepted if the likelihood value of the new sample L is greater than the likelihood of the previous sample L^*, described in section(2.5.1).

FireFly exploits this part of sampling, which is in generating a new sample. The switching on and off of a source is done by either setting the flux to 0 or generating a random sample of the flux by using MCMC or sampling directly from the prior distribution.

The Algorithm

The FireFly algorithm is based on the conditional assessment of the flux values of each source at each iteration. To decide whether a source gets switched on or off, a random value u is chosen from a uniform distribution $U(0, 1)$ and then compared to the threshold on whether to switch the source on or off if u is either higher or lesser than the threshold. After the state of the source has been chosen, then a new sample can be generated depending on several conditions. Algorithm (3) below is an outline of how the FireFly algorithm operates.

Algorithm 3: FireFly

Result: θ_{new}

1 **Initialization**

2 θ_{old} & L^*

3 **for** each source **do**

    // At each step in the MCMC or nested sampling.

4 Choose random variable u from $U(0,1)$ and set state to switch on or off depending on the relation (u $\geq$ threshold);

5 Check the flux of the source;

6 **if** the flux is on and state is on **then**

        // Do a random walk in all the parameter space and update the sample.

7 Do a standard random walk with MCMC;

8 Then update the old sample iff the likelihood constraint ($L^{new} > L^*$) holds.

9 **end**

10 **if** the flux is on and state is off **then**

11 Switch off the source flux, randomly sample the position from the prior and then update the old sample iff the likelihood constraint ($L^{new} > L^*$) holds.

12 **end**

13 **if** the flux is off and state is on **then**

14 Switch on the source flux, randomly sample the flux and position from the prior and then update the old sample iff the likelihood constraint ($L^{new} > L^*$) holds.

15 **end**

16 **if** the flux is off and state is off **then**

17 Repeat the step in line 10.

18 **end**

19 **end**

3.3 Clustering

Due to the random nature of the MCMC and Nested Sampling algorithm, samples in each chain are not necessarily assigned to the same source in every instance. It is, therefore, imperative to use a clustering method in order to characterize individual sources on the posterior results. The use of such a method is because the posterior distribution obtained from both the evidence-based and switcher-based search algorithms is multimodal, and clustering can solve this problem by generating a cluster for each mode and characterizing them. The multimodality of the data will be clear in the next chapter and can be seen e.g. in Fig (4.3).

In this thesis, we implement a hierarchical density-based clustering technique called Hierarchical Density-Based Clustering of Applications with Noise (HDBSCAN) developed by Campello, Moulavi, and Sander 2013.

HDBSCAN is an extension of DBSCAN, which was converted into a hierarchical clustering algorithm that uses a technique to extract flat clustering based on the stability of clusters. HDBSCAN generates a complete density-based clustering hierarchy from which a simplified hierarchy only of the most significant clusters that can easily be extracted (Campello, Moulavi, and Sander 2013).

Chapter 4

Applications

In this section, we demonstrate the evidence-based search and the FireFly-based algorithms on the Gaussian case described in section (4.1). The test example is set up to have features that resemble those that can occur in real inferences problems in astronomical data.

4.1 Gaussian Sources

The objective of this toy problem is to consider detecting and characterizing discrete objects in some background noise using Nested Sampling (John Skilling 2006) with Markov-Chain Monte Carlo 'MCMC' (Hogg and Foreman-Mackey 2017). Consider the data vector K to be pixel values in the image in which to search for these objects. Suppose that a template, $\tau(\theta)$, describes these objects, where θ denotes collectively the (X, Y) position of the object, its amplitude A and a measure R of its spatial extent. In this case, assume circular Gaussian-shaped objects, such that :

$$\tau(\theta) = A \exp\left[- \frac{(x - X)^2 + (y - Y)^2}{2R^2} \right] \tag{4.1}$$

To get K, the contribution from each object is considered to be additive. Therefore if there are N_{objs} objects present then:

$$K = \sum_{k=1}^{N_{objs}} \tau(\theta_k) + n \tag{4.2}$$

where n denotes the generalized noise contribution to the data and $\tau(\theta_k)$ is the contribution to signal from the kth discrete object. Therefore the values of the unknown parameters $(\mathrm{N}_{objs}, \theta_1, \theta_2, ..., \theta_{N_{objs}})$ need to be estimated.

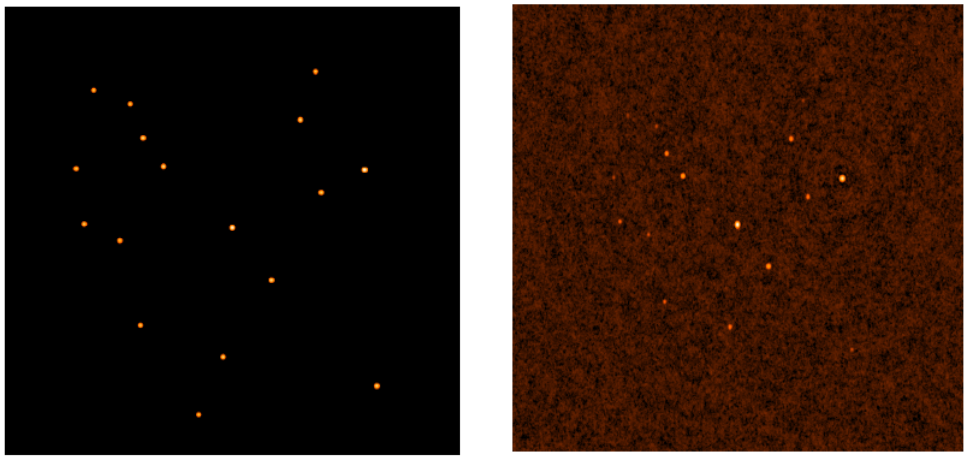


Figure 4.1: Illustration of a simulated radio image. (left) The noise-free sky model of the radio image and (right) shows the radio image with noise. The challenge for object detection techniques is to be able to detect the sources in the radio image with noise (right), which depicts an ideal radio image. Courtesy of (Lochner 2014).

4.2 Simulated data

A number of 8 test cases were simulated, each case consists of (2, 3, 4, 5, 10, 15, 20, 30) true sources. Furthermore, for each case, 40 data maps were produced, whereby in each data map the positions of the true sources were randomly generated. The X and Y coordinates were drawn from the uniform distribution $U(0, 50)$. Similarly, the amplitude A of each object was drawn independently from the uniform distribution $U(5, 10)$, and the spatial extent R was kept constant at 2.0. The simulated data maps (e.g., fig. 4.2) were produced with added independent Gaussian pixel noise of a root mean square (RMS) of 0.7 units. This resembled a signal-to-noise ratio (SNR) of 7-14. Fig (4.2) shows a simulated model and the corresponding data map of 2 real sources case.

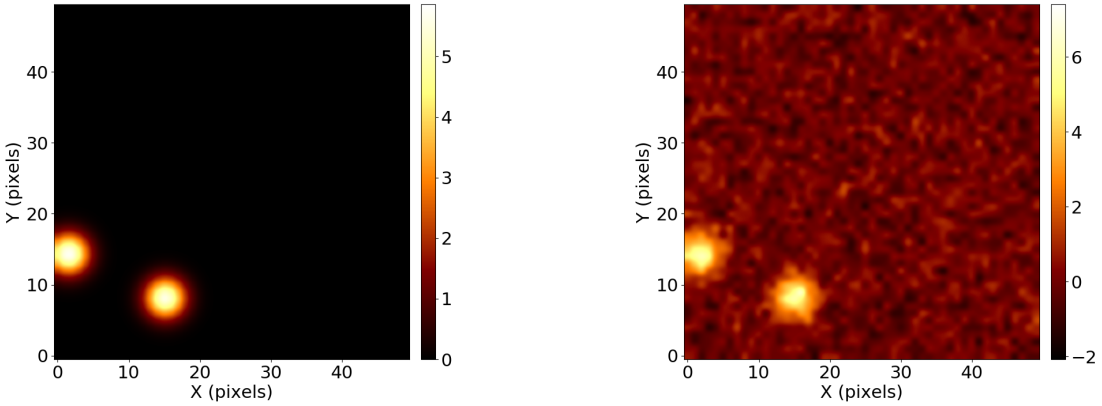
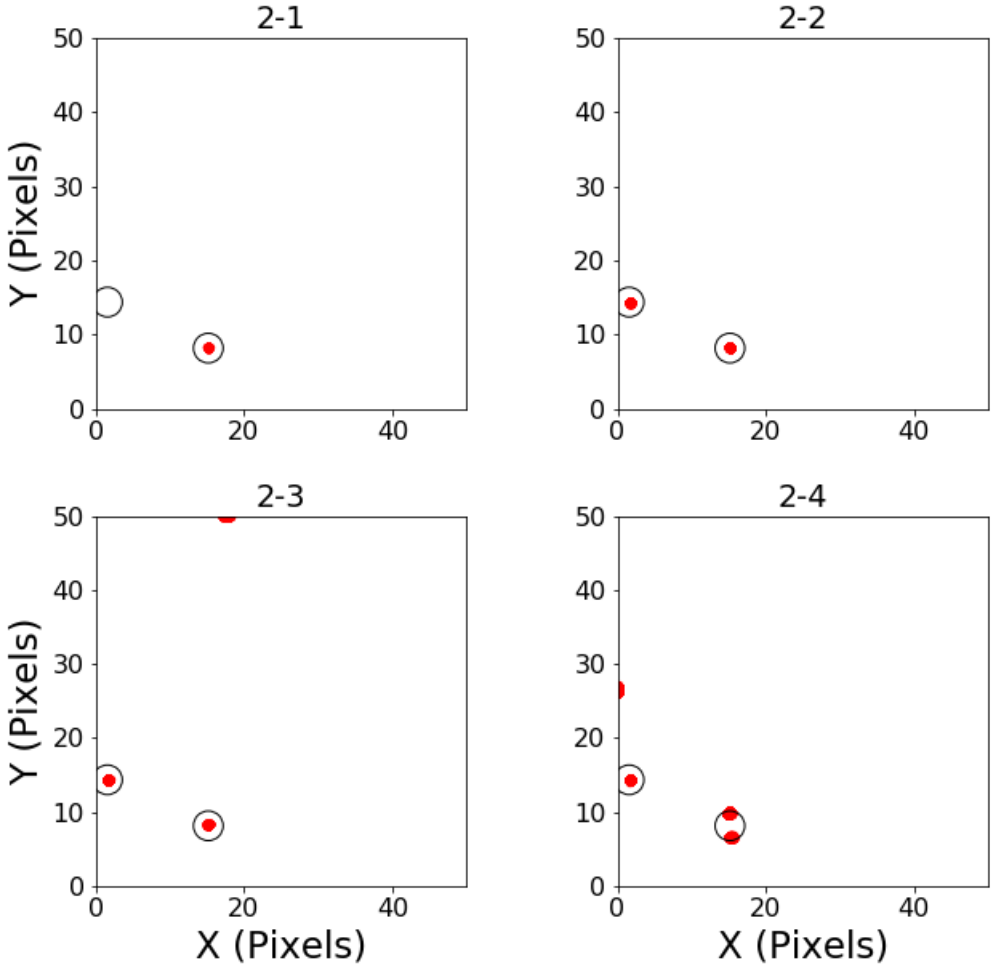


Figure 4.2: Illustration of one of the test cases with 2 real sources. The left image shows the simulated sky model without noise. The right image shows the corresponding data map with independent Gaussian noise added with rms of 0.7 units.

4.3 Method 1 : Evidence-Based approach

Ideally, one would like to know how many real sources, N_{objs}, are present in the test image (Fig. 4.2) and also be able to infer all the unknown parameters $(\theta_1, \theta_2, ..., \theta_{N_{objs}})$ simultaneously from the data. A simple, practical approach is to find a model that best describes the data and be able to infer the parameters of the model. This can be done by heuristically taking several models, evaluating the evidence for each model, and choosing the best fitting model to the data.

We use the evidence-based search approach described in section (3.2.1), to evaluate eight models with a varying number of model sources (i.e., 1 to 8 models sources), to compute the evidence for each model of each 2, 3 and 4 real sources test cases. Figure (4.3) shows a posterior predictive plot of the simulated map for a two real sources case.


2-1
2-2
2-3
2-4
Y (Pixels)
X (Pixels)
0
10
20
30
40
50

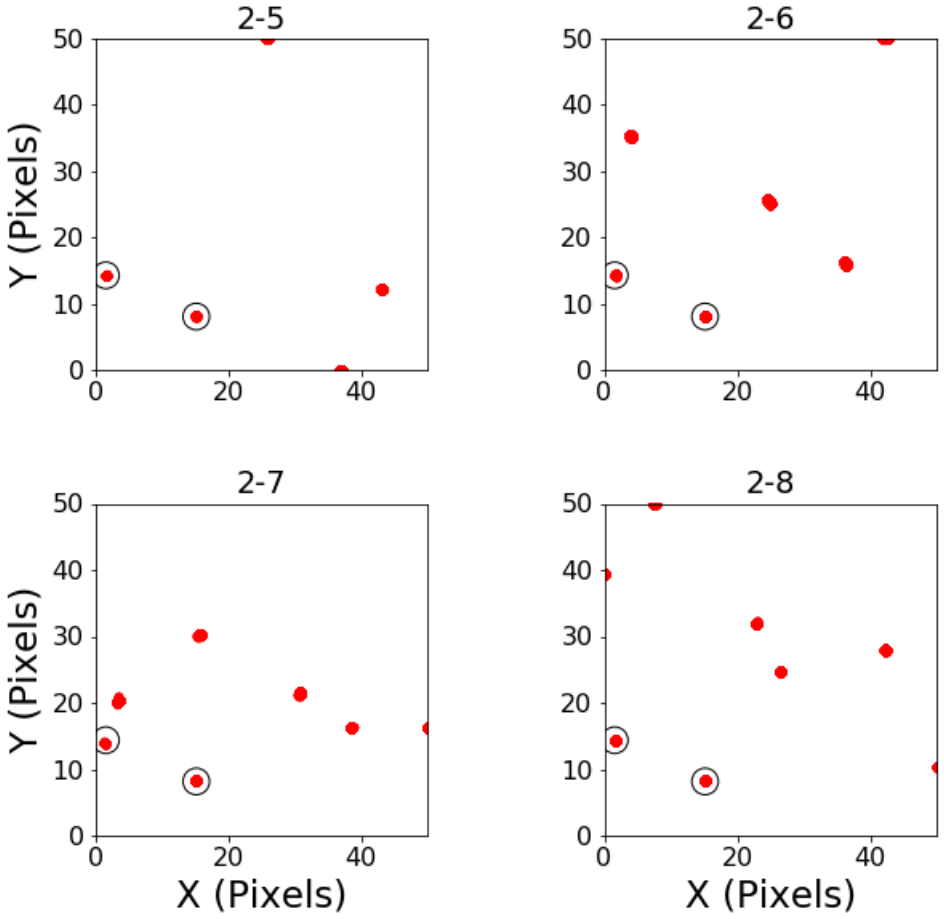


Figure 4.3: Each sub-figure corresponds to a different model, ($N_{objs} = 1, 2, ..., 8$). In each figure, the ○ marker represents the true source positions and the 'red blobs' represent the posterior results from Nested Sampling with MCMC.

4.3.1 Results: Evidence-Based Search

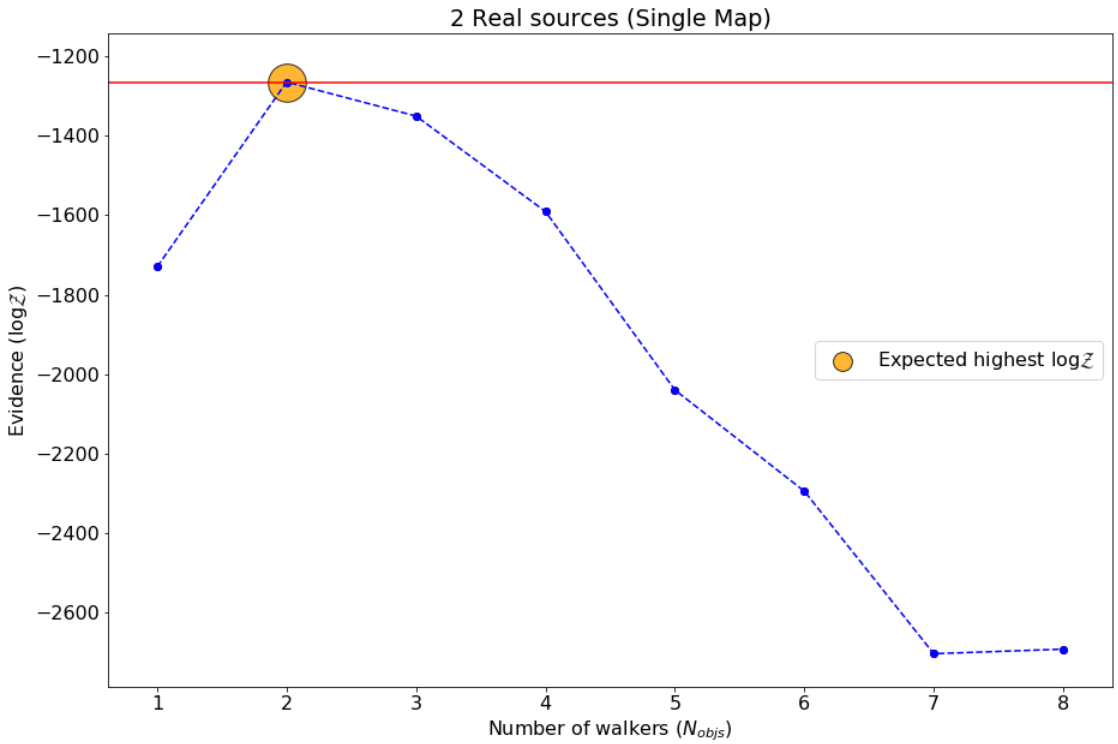
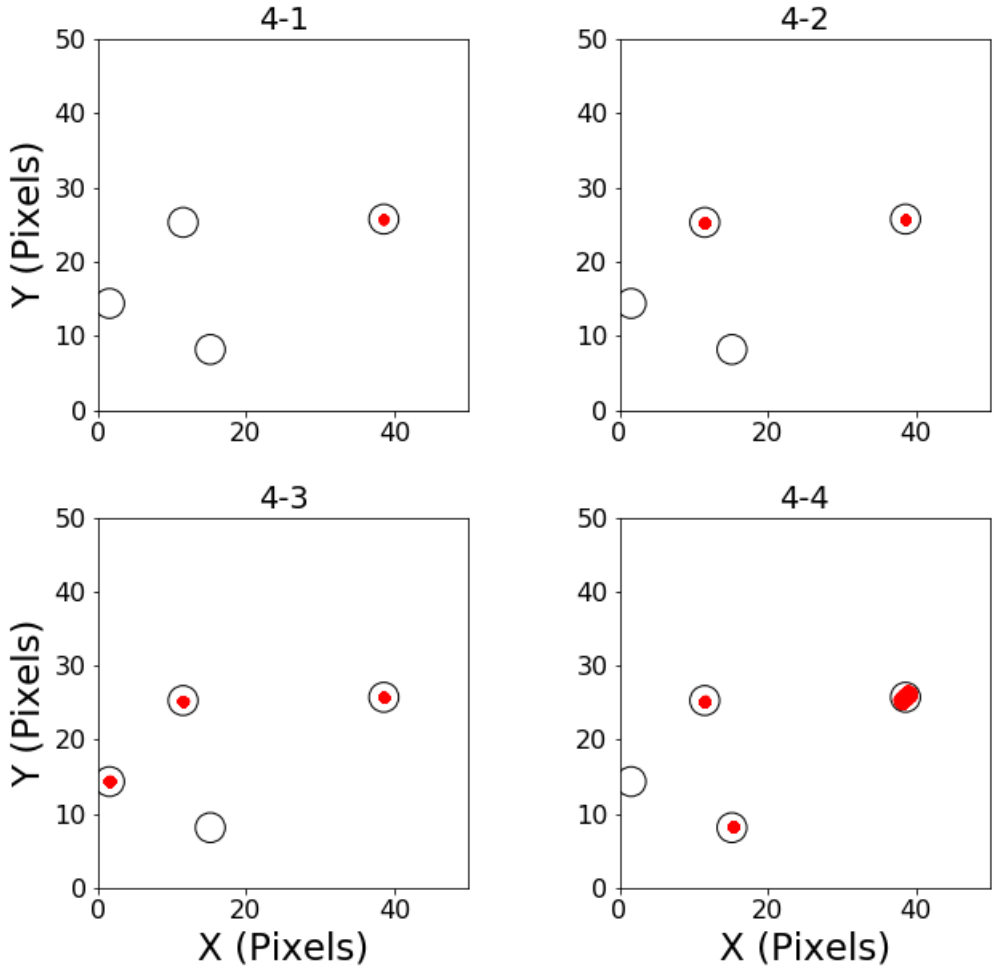


Figure 4.4: The plot shows a relationship between the evidence log$\mathcal{Z}$ and the number of model sources for each model. The graph peaks at two sources ($N_{objs} = 2$), which indicates that the model with two model sources is the best model that describes the simulated data image, Fig. 4.2.

Figure (4.4) illustrates the evidence ($log\mathcal{Z}$) for each model described in figure (4.3). The evidence peaks at a model with two sources and is lower for most; therefore, it shows that the model with two model sources describes the simulated data map in figure (4.2) sufficiently than the other models. Consequently, the unknown parameters $(N_{objs}, \theta_1, \theta_2, ..., \theta_{N_{objs}})$ are then determined using the posterior distribution of the model with two sources.

4.3.2 Issues with Evidence-Based approach

It is clear from figure (4.4) that the best fitting model for the test data map in figure (4.3) is the model with 2 model sources. However, this simple approach is not always sufficient to use for finding the best fitting model. This is due to issues such as degeneracies, whereby more than one model source converging to a single object or an object being entirely missed, figure (4.5, panel (4-4, 4-6 ,4-7 and 4-8)) shows an example of the drawbacks of this simple approach. Moreover, it is not entirely clear on how many models to use to get the best fitting model eventually, and this becomes a challenging problem in large datasets where it is nearly impossible to know how many sources may be present in an astronomical data map.



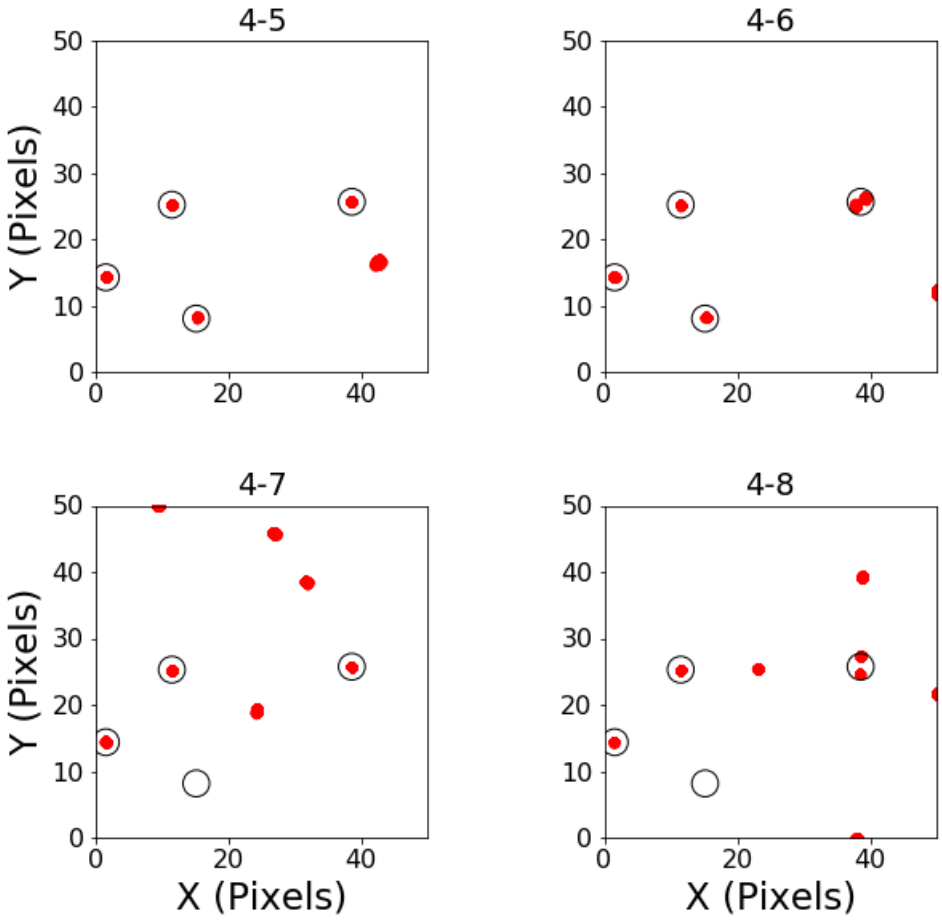


Figure 4.5: Illustration of the 4 real sources test case. Panel (4-4) shows how one object can be entirely undetected. Panel (4-6,7,8) shows that more than one model source can converge to one source, and another source can be undetected.

Figure (4.5, panel (4-4)) shows that more than one model source can converge to a single object, and another object can be entirely undetected. This allows the other model to have a better evidence value than the 'expected' model. As a consequence, a worse model can get chosen to be the best fitting model with the evidence-based search approach.

Therefore we demonstrate that the evidence-based search approach fundamentally works by evaluating the average evidence for the 40 simulated data maps described in section (4.2) for each case. Figures (4.6), (4.7), (4.8) show the average evidence vs. the number of walkers, for each case.

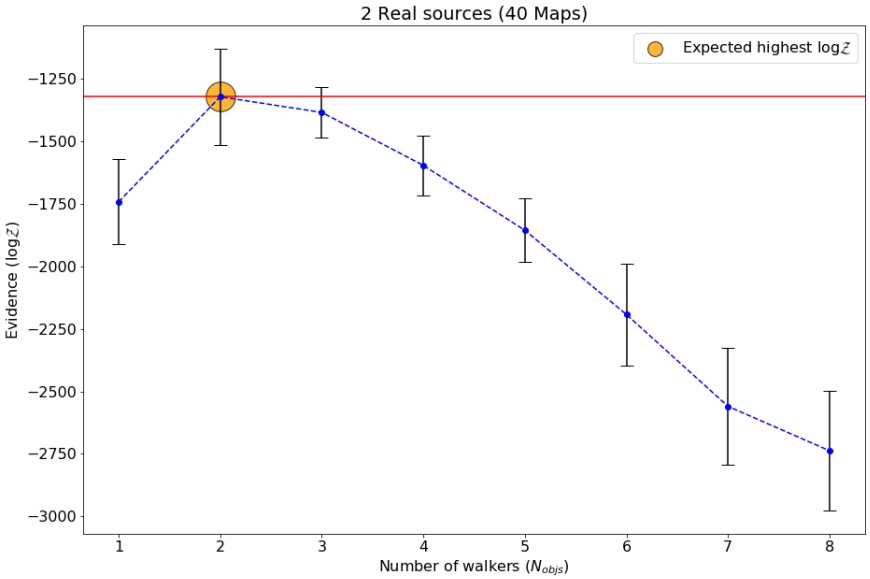


Figure 4.6: Average evidence plot for the 2 real sources test case. The orange circle shows the expected best fit model, and the red line shows where the highest evidence is.

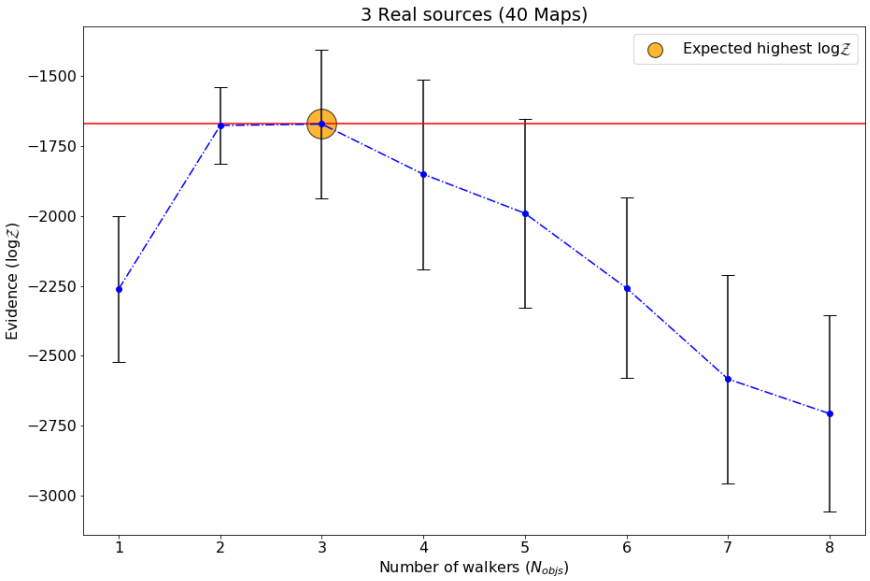


Figure 4.7: Average evidence plot for the 3 real sources test case. The orange circle shows the expected best fit model, and the red line shows where the highest evidence is. We see that two models lie on the red line, in this case, models 2 and 3.

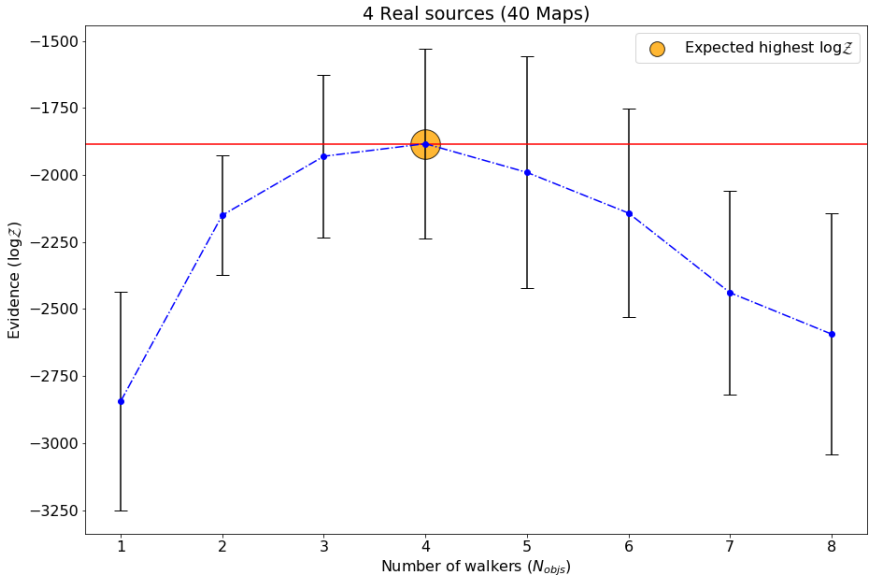


Figure 4.8: Average evidence plot for the 4 real sources test case. The orange circle shows the expected best fit model, and the red line shows where the highest evidence is.

However, even computing the average evidence does not entirely make the approach satisfactory to use. Figure (4.9) shows the average evidence vs. the number of walkers for a 10 real source case, and it is evident that it is not clear that the 10 walkers model is the best model as expected.

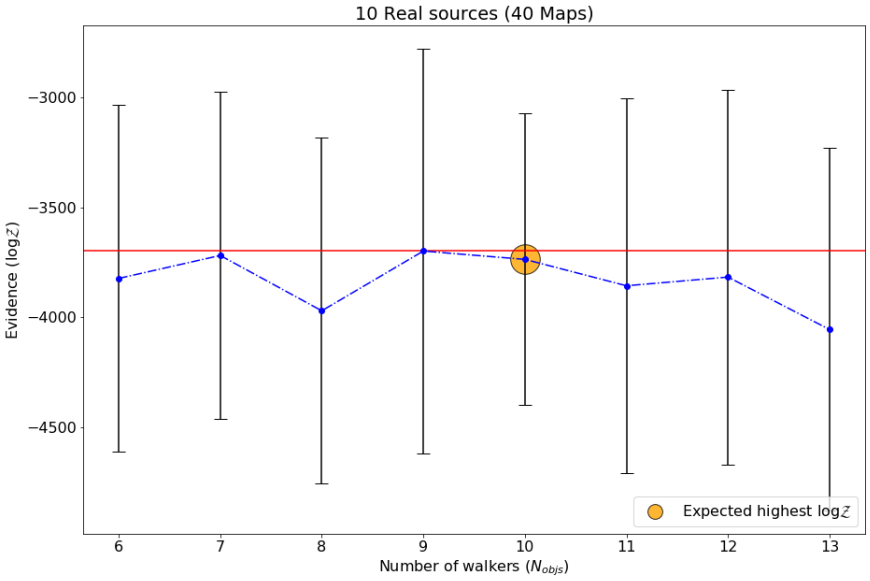


Figure 4.9: Average evidence plot for the 10 real sources test case. The orange circle shows the expected best fit model, and the red line shows where the highest evidence is. A heuristic number of model sources was chosen (i.e., 6-13 model sources) for the 10 real sources case.

One of the main drawbacks of the evidence-based search method is its computational complexity. A heuristic selection of different models to evaluate is quite an inefficient way of looking for the best model. Moreover, as the number of sources increases with each model, the computational time of evaluating each model also increases. Figure (4.9) illustrates how relying only on model selection is not sufficient.

4.4 Method 2 : FireFly algorithm

The simulated data for FireFly is the same, described in section (4.2). Moreover, the addition of four cases was introduced to the test data (i.e., 5, 15, 20, and 30 real source cases). However, instead of using different models for each case, only a single model is used. Each model consists of double the expected number of sources in the simulated data map, as described in section (4.4).

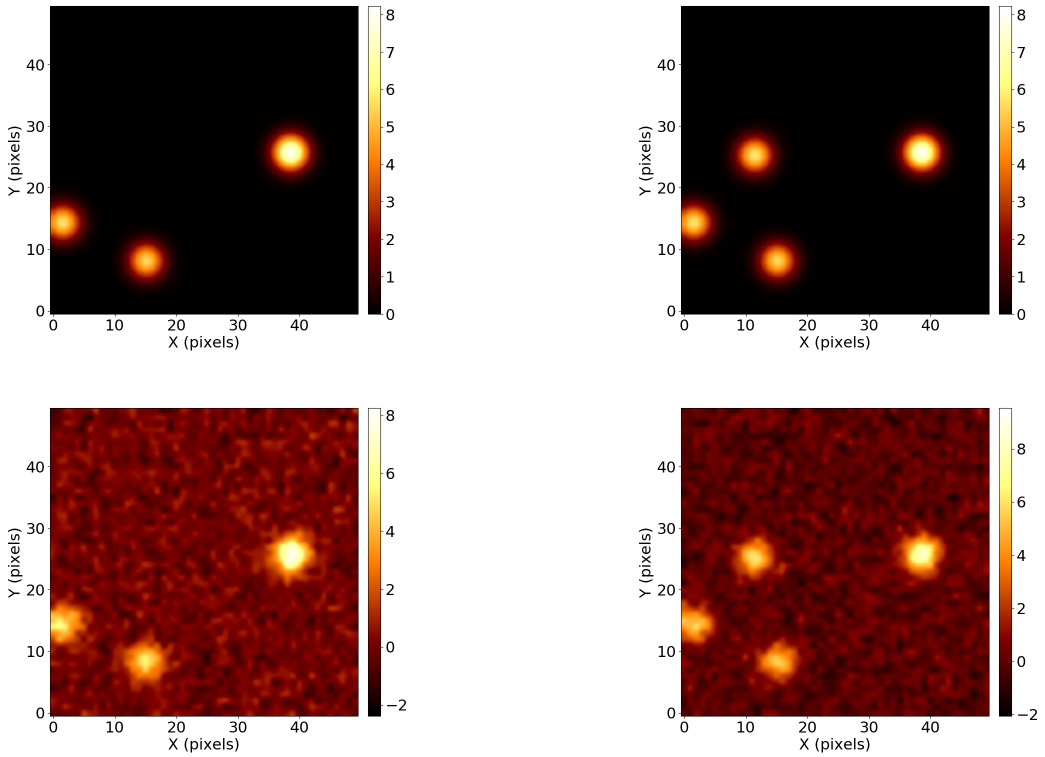


Figure 4.10: The left-column shows the 3 real source case and the right-column shows the 4 real source case. The top-row is the simulated sources without noise using Eqn. 4.1 for each source. The bottom-row is the simulated data map of both cases with added Gaussian pixel noise of rms 0.7 units, using eqn 4.2.

4.4.1 Results : FireFly

The resulting posterior distributions from the FireFly method are shown in figure (4.11), for each 2, 3 and 4 real sources case, respectively. The method can sample the simulated data map distribution in each case quite effectively by employing a dynamical model, which

allows for switching off and on of sources.

It is evident by implementing the FireFly technique that the searching algorithm converges to the expected number of sources in the simulated data map in each test case, fig. 4.11.

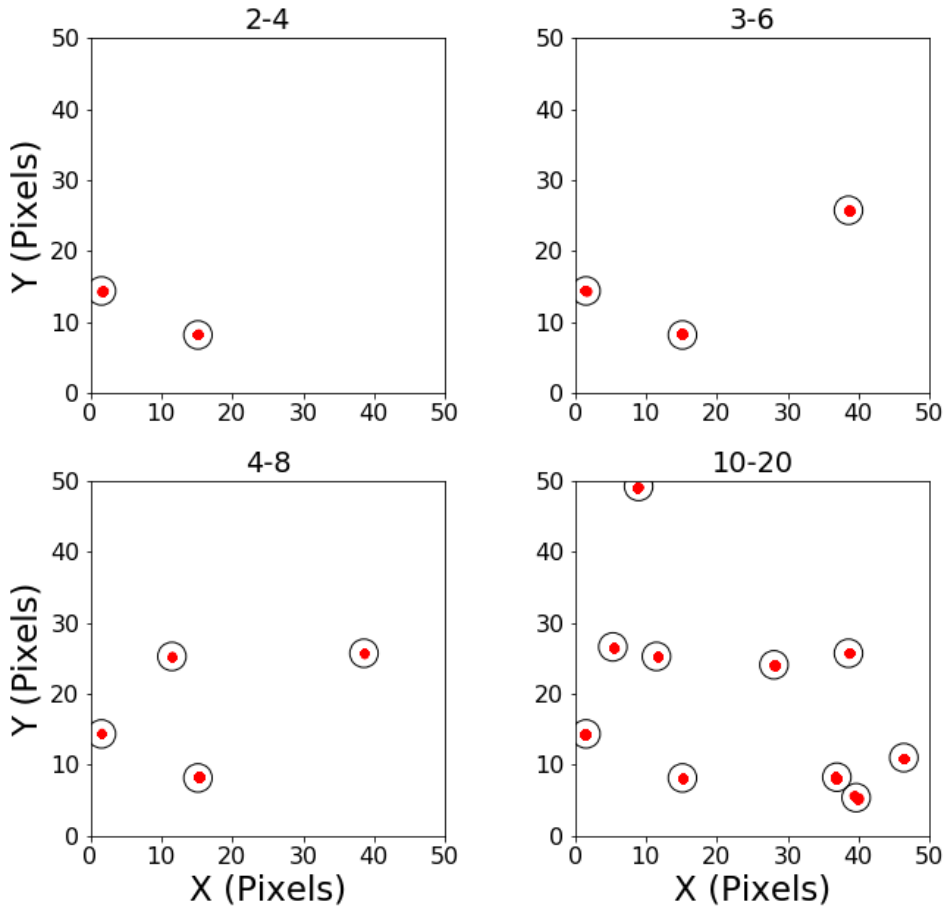


Figure 4.11: Each sub-figure corresponds to 2, 3, 4 and 10 real sources cases. In each case, only a single model (described in section (4.4)) was used. The red blobs represent the posterior distributions of the positions from FireFly and the black circles are the true sources.

We see that both methods can produce a practical source finding technique. However, the efficiency of method 1 drops significantly due to implementing model comparison, which includes evaluating several models to choose the best fitting model eventually. Furthermore, the best-chosen model is not always the fitting model because of issues such as degeneracies, which can lead to worse models having better evidence than the expected best fit model.

4.5 Computational Requirements

All code was run on the CHPC (Centre for high performance computing) cluster, comprising of Intel 5th generation CPUs (CHCP 2019). Running the evidence-based approach and FireFly consisted of using the queue name called 'seriallong' for very long sub-1-node jobs, with a maximum wall time of 144 hours and 12 cores.

4.6 Computational efficiency: FireFly and Evidence-Based Search

Figure (4.12) shows a comparison of the execution time for the evidence search-based method against FireFly with a nested sampling and MCMC implementation. For the evidence-based search method, each case (2, 3, 4, and 10) was evaluated using 8 different models, and only a single dynamic model was used for FireFly.

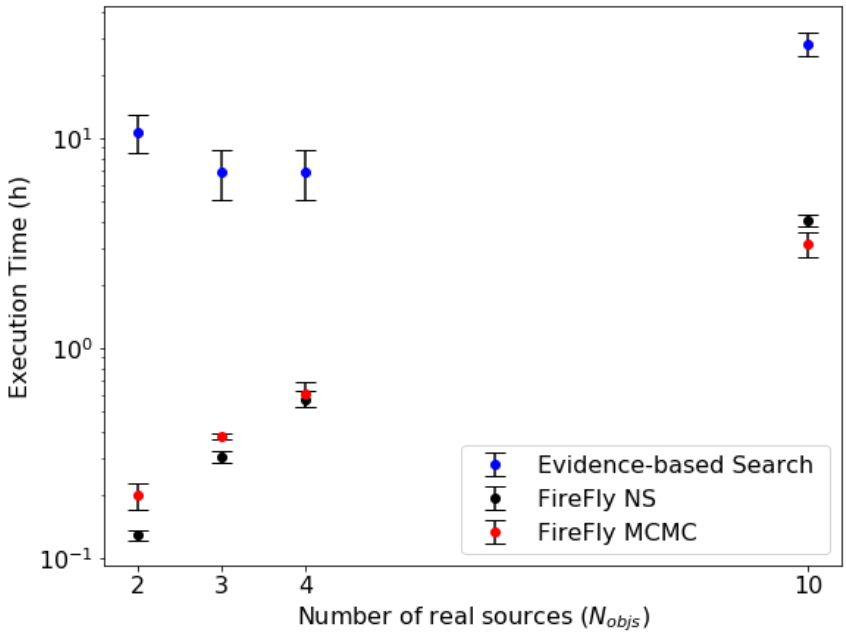


Figure 4.12: Execution time in hours (log-space), for the evidence-based search vs. FireFly nested sampling and FireFly MCMC implementations.

Figure (4.12) shows that the evidence-based approach is slower than FireFly around an order of magnitude. The scaling with the number of true sources for the evidence-based approach was done up to 10 real sources due to the expensive computational time and model complexity.

4.7 FireFly : MCMC vs. Nested Sampling

In this section, we look at the performance of applying FireFly using MCMC and nested sampling. We ran the two implementations on the simulated data for the (2, 3,4,5,10,15,20, and 30) cases as described in (4.2).

4.7.1 FireFly : Nested sampling

In nested sampling, one of the critical challenges is generating a new particle/sample from the prior, subject to a hard likelihood constraint, discussed in section (2.5.7). The fundamental knowledge is that the new sample should be independent of the surviving points. We implement MCMC to generate a new sample. Moreover, the MCMC is then ran for a longer time to ensure that the new sample is effectively independent of the surviving population. The MCMC runs (i.e the number of samples in the MCMC chain.) used for the test cases using FireFly was set to 50, and the stopping threshold for nested sampling, ϵ, described in (2.5.4), was set to 0.1.

4.7.2 FireFly : MCMC

The MCMC implementation uses the basic MCMC algorithm (1), and instead of generating a new sample from a proposal distribution, the new sample is generated with FireFly. However, the implemented MCMC in FireFly is only run for a single step since the full implementation of MCMC with FireFly as a new sample generator already consists of the acceptance criteria (2.4.2). The MCMC implementation uses the Gelman Rubin stopping criteria with a threshold of 1.01.

4.7.3 MCMC vs. Nested Sampling

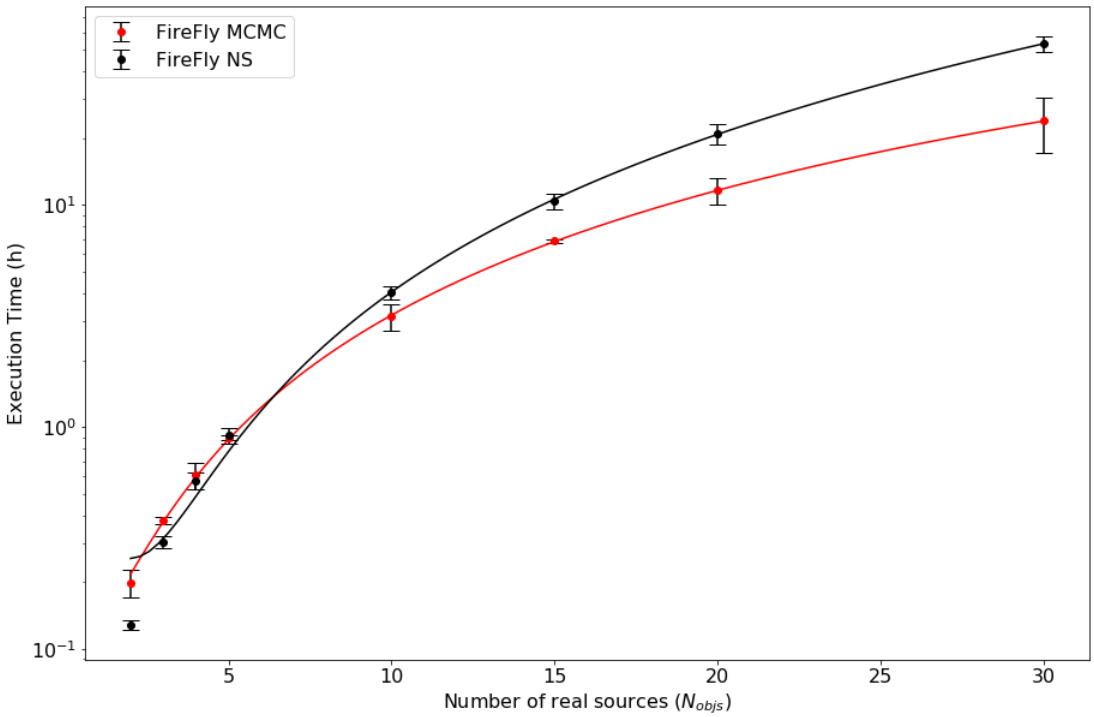


Figure 4.13: Execution time (log-space) for FireFly Nested Sampling and FireFly MCMC implementations.

Figure (4.13) highlights how both versions of FireFly scale sub-exponentially and much better than the evidence-based approach and hence are more suitable for dealing with large numbers of sources. Moreover, FireFly MCMC scales better with computational time than FireFly Nested Sampling as the number of sources increase.

4.8 Scaling FireFly

Source finding algorithms are constantly developed and improved to handle a growing amount of work as more massive amounts of astronomical data are produced. In this work, we scale FireFly by increasing the number of real sources in the simulated data map.

Figure (4.14) and (4.15) shows a posterior predictive plot of 30 real simulated sources in black circles and the posterior results in red blobs from the FireFly nested sampling and MCMC implementations, respectively.

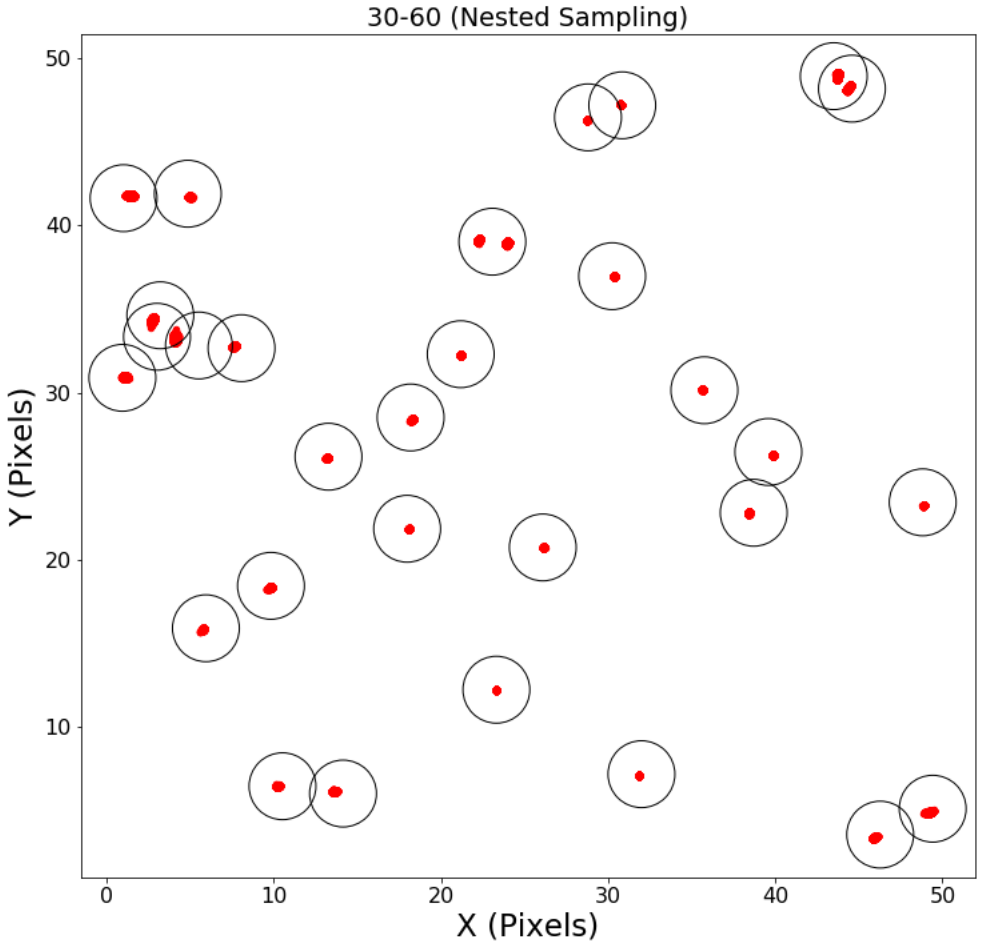


Figure 4.14: FireFly nested sampling. The black circles show the real sources and the red blobs are the posterior results. Test case for 30 real sources and 60 walkers.

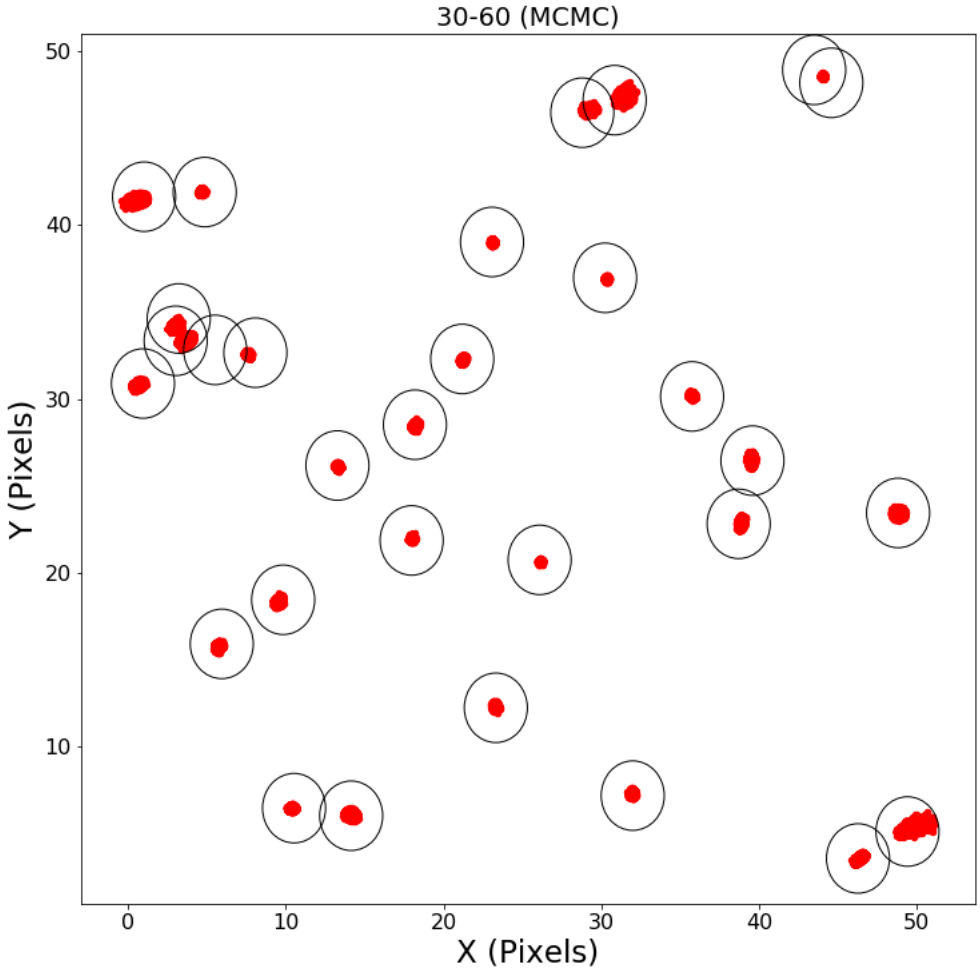


Figure 4.15: FireFly MCMC. The black circles show the real sources and the red blobs are the posterior results. Test case for 30 real sources and 60 walkers

Figure (4.14) and (4.15) shows the clustering of points for both FireFly implementations. The clustered points in MCMC FireFly are less concentrated and more dispersed (Figure 4.15) than in Nested sampling FireFly (Figure 4.14), suggesting that the nested sampling Firefly algorithm is more accurate, despite being slightly more computationally intensive.

4.8.1 FireFly Convergence

The FireFly technique is designed to be capable of switching on and off of sources in its sampling strategy, discussed in section (4.4). This technique allows the algorithm to eventually converge to the real/actual number of sources in the image space. Figure (4.16) and (4.17) shows the distribution of the number of 'on' sources for the whole samples from the FireFly, nested sampling, and mcmc implementations, respectively.

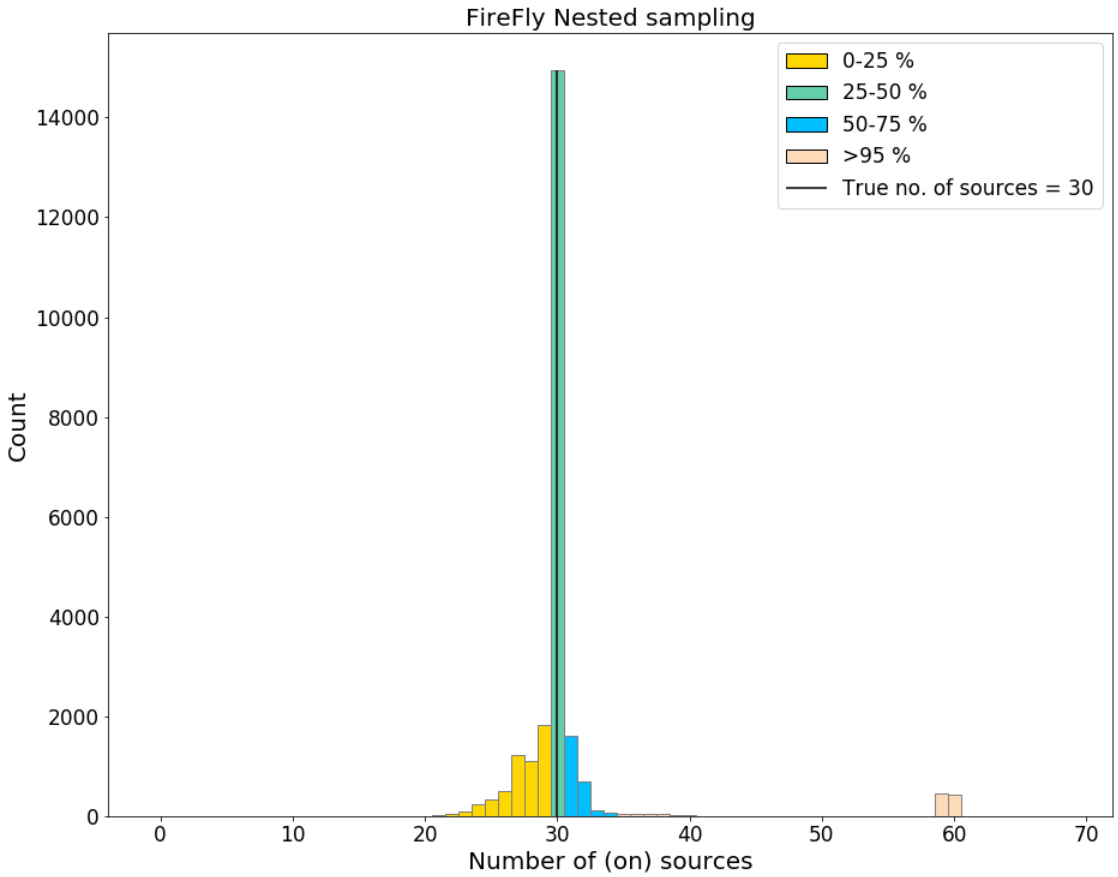


Figure 4.16: Distribution of the number of sources that were switched 'on' from the posterior samples of the FireFly nested sampling implementation. Each bar shows the count and percentage of the total number of switched on sources in the samples. This is for the test case of 30 real sources and 60 walkers.

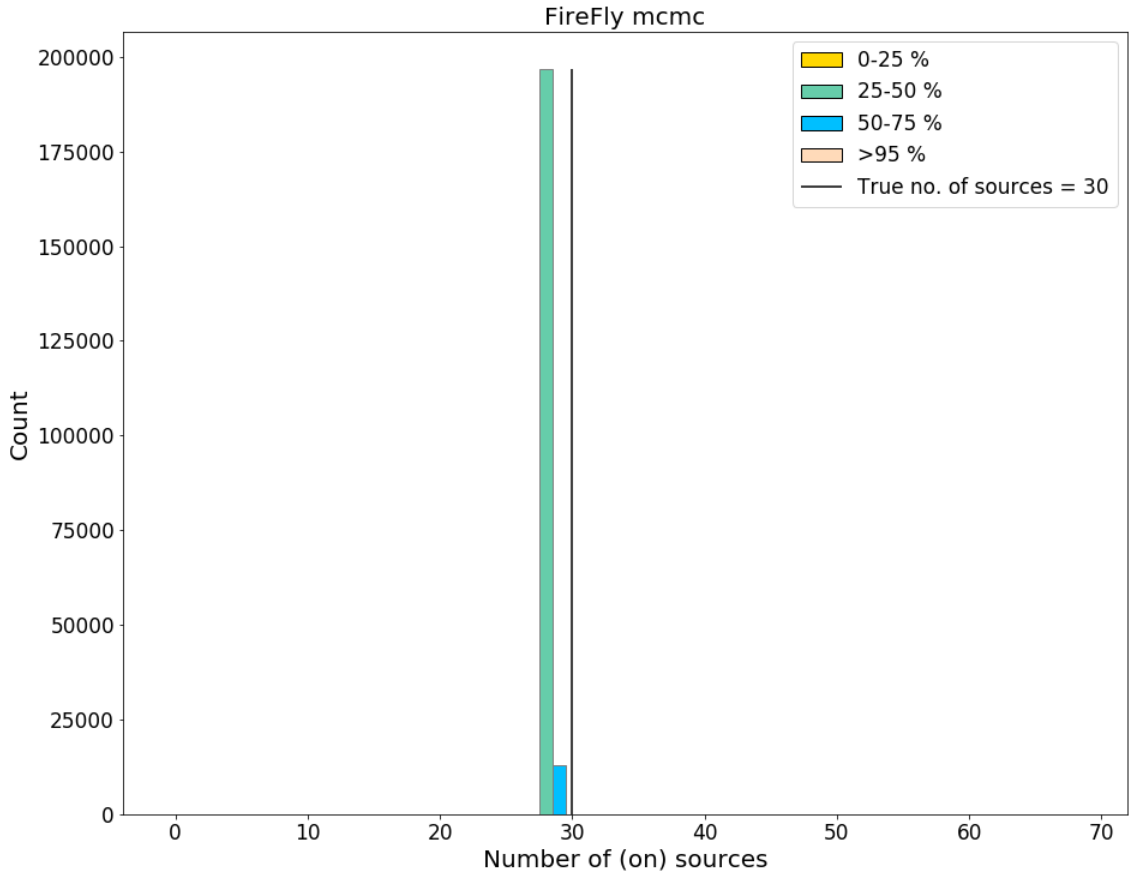


Figure 4.17: Distribution of the number of sources that were switched 'on' from the posterior samples of the FireFly nested MCMC implementation. Each bar shows the count and percentage of the total number of switched on sources in the samples. This is for the test case of 30 real sources and 60 walkers.

The posterior distribution for the number of sources with Nested Sampling Firefly is centered around the true value (see Figure 4.16). However, the majority of samples in the MCMC FireFly run (Figure 4.17) peak at 28 sources rather than the true value of 30. Nested sampling is better at multimodality than MCMC, so when two sources are close together, MCMC cannot easily distinguish them. Therefore, Nested Sampling estimate the number of sources better than MCMC.

Chapter 5

Conclusions

Efficient and rigorous source finding techniques are needed due to the upcoming large data sets from telescopes like MeerKAT, LSST and the SKA. Most of the current source-finding algorithms that have been developed comprise of background estimation strategies such as thresholding and filtering. Typically these algorithms aim to estimate background noise in an image and perform some form of thresholding to find sources. This, however, is a challenging task as the noise can emanate from galactic emissions, instrumental noise, or from emissions of parts of the telescopes itself. Thresholding source-finding algorithms are unable to accurately estimate uncertainties on parameters such as source position, shape and flux. This makes it difficult to fully quantify this source of uncertainty in high precision scientific analyses. Statistically robust techniques are, therefore, essential for sufficient object detection as they are able to use all available data and include any prior knowledge we may have. Bayesian statistics becomes an obvious approach as it allows precise statistical questions of the data. It is also profoundly general, allowing the inclusion of all available information.

Chapter (2) gives a formalism of Bayesian statistics and an introduction to the two sampling techniques which we implement, i.e., nested sampling and MCMC. With these tools, we are able to perform a full Bayesian analysis on the test cases of object detection we investigate. Although Bayesian techniques are not the only techniques that exist in the object detection regime (see chapter 1), we show that a basic implementation of Bayesian statistics using MCMC and nested sampling can perform statistically rigorous source finding.

While Bayesian source-finding algorithms exist, they typically struggle with the problem of

an unknown number of sources. We consider the simplest form of a fully Bayesian approach, which we call the evidence-based method, which repeatedly increases the number of sources in the model until the Bayesian evidence peaks. This method, while rigorous, is very slow.

We thus introduced FireFly, which makes use of a single model consisting of roughly twice the number of expected sources, but gives each source the ability to "switch off" in the sampler itself. Thus FireFly can converge on the correct number of sources, as well as determine their position and flux, without expensive repeated iterations or the need for a changing number of parameters. We show that FireFly can be implemented with any Bayesian sampling algorithm and compare nested sampling and MCMC implementations.

In chapter (4), we apply the two methods, FireFly and the evidence-based search approach to simulated image data. In our test cases, we focused on circularly gaussian shaped objects, where a single object consists of four parameters: flux A, position (X, Y), and spatial extent R. In our simulations, we kept the spatial extent R constant at 2.0, and varied the flux and position. Therefore our inference was based on the three parameters of each source, flux A and position X and Y. We test both methods on simulated data between 2 and 30 real sources, with 40 realisations per test to obtain error bars.

We used our implementation of nested sampling for the evidence-based approach and compared FireFly with both nested sampling and MCMC. In all cases we ran the chains until they were converged and recorded the computation time.

We find that in general the evidence-based approach is able to determine the correct number of sources and their positions and amplitudes (Figures 4.3 and 4.4) but find that it is prohibitively slow, as expected. We find that FireFly is also able to determine the correct number of sources and their full posterior (Figure 4.11) but FireFly is an order of magnitude faster than the evidence-based approach (Figure 4.13) and further optimisation could be done to improve its convergence time.

We further implemented FireFly with nested sampling and MCMC. Evaluating the performance of the two implementations shows that as the number of parameters increases, the MCMC implementation becomes more efficient than nested sampling in absolute computational time, seen in figure (4.13). Although MCMC Firefly has a better computational scaling than nested sampling Firefly, a comparison of the figure (4.14) and (4.15) show that nested sampling Firefly seems to be more accurate than MCMC Firefly.

Our work shows that it is possible for Bayesian source-finding to be reasonably fast and

accurate, obtaining a full posterior without repeated runs. Clustering techniques such as HDBSCAN can be run on the posterior itself to create catalogs for surveys or the full posterior can be used as input for further analyses. Future work would include extending the algorithm to handle extended sources with complex morphology, multiwavelength data and to include a more realistic source distribution, point-spread-function and noise properties.

Bibliography

Abbott, B. P. et al. (2016). "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Phys. Rev. Lett.* 116 (6), p. 061102. DOI: 10.1103/PhysRevLett.116.061102. URL: https://link.aps.org/doi/10.1103/PhysRevLett.116.061102.

Aghanim, N. et al. (2018). *Planck 2018 results. VI. Cosmological parameters*. arXiv: 1807.06209 [astro-ph.CO].

Aldrich, John (1997). "R. A. Fischer and the Making of Maximum Likelihood 1912-1922". In: 12.3, pp. 162–179. URL: http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Likelihood/Fisher%20and%20history%20of%20mle.pdf.

Bayes, T. (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances". In: *Philosophical Transactions* 53, pp. 370–418.

Bertin, E. and Arnouts, S. (1996). "SExtractor: Software for source extraction". In: *Astron. Astrophys. Suppl. Ser.* 117.2, pp. 393–404. DOI: 10.1051/aas:1996164. URL: https://doi.org/10.1051/aas:1996164.

Blei, David M. et al. (2016). "Variational Inference: A Review for Statisticians". In: *arXiv e-prints*, arXiv:1601.00670, arXiv:1601.00670. arXiv: 1601.00670 [stat.CO].

Booth, R. S. et al. (2009). "MeerKAT Key Project Science, Specifications, and Proposals". In: *arXiv e-prints*, arXiv:0910.2935, arXiv:0910.2935. arXiv: 0910.2935 [astro-ph.IM].

Campello, Ricardo J. G. B. et al. (2013). "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172. ISBN: 978-3-642-37456-2.

Carvalho, P. et al. (2009). "A fast Bayesian approach to discrete object detection in astronomical data sets - PowellSnakes I". In: *mnras* 393, pp. 681–702. DOI: 10.1111/j.1365-2966.2008.14016.x. arXiv: 0802.3916.

Casella, George and Roger L. Berger (2019). "Statistical inference / George Casella, Roger L. Berger". In: *SERBIULA (sistema Librum 2.0)*.

CHCP (2019). *Centre for High performance Computing (CHPC)*. http://wiki.chpc.ac.za/quick:start, [Online; accessed 2019].

Chib, S. (2001). "Monte Carlo Methods and Bayesian Computation: Overview". In: *International Encyclopedia of the Social Behavioral Sciences*. Ed. by Neil J. Smelser and Paul B. Baltes. Oxford: Pergamon, pp. 10004–10009. ISBN: 978-0-08-043076-8. DOI: https://doi.org/10.1016/B0-08-043076-7/00467-8. URL: http://www.sciencedirect.com/science/article/pii/B0080430767004678.

Cornwell, Tim J. (2008). "Multiscale CLEAN Deconvolution of Radio Synthesis Images". In: *IEEE Journal of Selected Topics in Signal Processing* 2.5, pp. 793–801. ISSN: 1932-4553. DOI: 10.1109/jstsp.2008.2006388. URL: http://dx.doi.org/10.1109/JSTSP.2008.2006388.

Daylan, Tansu et al. (2017). "Inference of Unresolved Point Sources at High Galactic Latitudes Using Probabilistic Catalogs". In: *The Astrophysical Journal* 839.1, p. 4. ISSN: 1538-4357. DOI: 10.3847/1538-4357/aa679e. URL: http://dx.doi.org/10.3847/1538-4357/aa679e.

Ester, Martin et al. (1996). "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231. URL: http://dl.acm.org/citation.cfm?id=3001460.3001507.

Fan, Y and S A Sisson (2010). *Reversible jump Markov chain Monte Carlo*. arXiv: 1001.2055 [stat.ME].

Feroz, F. and M. P. Hobson (2008). "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses". In: 384.2, pp. 449–463. DOI: 10.1111/j.1365-2966.2007.12353.x. arXiv: 0704.3704 [astro-ph].

Feroz, F. and J. Skilling (2013). "Exploring multi-modal distributions with nested sampling". In: *American Institute of Physics Conference Series*. Ed. by U. von Toussaint. Vol. 1553. American Institute of Physics Conference Series, pp. 106–113. DOI: 10.1063/1.4819989. arXiv: 1312.5638 [astro-ph.IM].

Frean, Marcus et al. (2014). "Source detection in astronomical images by Bayesian model comparison". In: pp. 55–61. DOI: 10.1063/1.4903710.

Friedlander, A. M. (2014). "Dirichlet Methods for Bayesian Source Detection in Radio Astronomy Images". MA thesis. Victoria University of Wellington.

Garrett, M. A. et al. (2010). "Square Kilometre Array: a concept design for Phase 1". In: *arXiv e-prints*, arXiv:1008.2871, arXiv:1008.2871. arXiv: 1008.2871 [astro-ph.IM].

Gelman, Andrew, John Carlin, et al. (2014). *Bayesian Data Analysis, 3rd Ed.*

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". In: *Statist. Sci.* 7.4, pp. 457–472. DOI: 10.1214/ss/1177011136. URL: https://doi.org/10.1214/ss/1177011136.

Gregory, P. C. (2005). *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with 'Mathematica' Support.*

Hague, P. et al. (2018). "Bayesian Source Discrimination in Radio Interferometry". In: *ArXiv e-prints*. arXiv: 1807.01719 [astro-ph.IM].

Hancock, P. J. et al. (2012). "Compact continuum source finding for next generation radio surveys". In: *Monthly Notices of the Royal Astronomical Society* 422.2, pp. 1812–1824. DOI: 10.1111/j.1365-2966.2012.20768.x. arXiv: 1202.4500 [astro-ph.IM].

Hancock, Paul J. et al. (2018a). "Source Finding in the Era of the SKA (Precursors): Aegean 2.0". In: *Publications of the Astronomical Society of Australia* 35, e011. DOI: 10.1017/pasa.2018.3.

– (2018b). "Source Finding in the Era of the SKA (Precursors): Aegean 2.0". In: 35, e011, e011. DOI: 10.1017/pasa.2018.3. arXiv: 1801.05548 [astro-ph.IM].

Handley, W. J. et al. (2015). "polychord: next-generation nested sampling". In: *Monthly Notices of the Royal Astronomical Society* 453.4, pp. 4385–4399. ISSN: 1365-2966. DOI: 10.1093/mnras/stv1911. URL: http://dx.doi.org/10.1093/mnras/stv1911.

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1, pp. 97–109. ISSN: 00063444. URL: http://www.jstor.org/stable/2334940.

Heavens, Alan et al. (2017). *Marginal Likelihoods from Monte Carlo Markov Chains*. arXiv: 1704.03472 [stat.CO].

Hobson, Michael et al. (2009). "Bayesian Methods in Cosmology". In: *Bayesian Methods in Cosmology, by Michael P. Hobson , Andrew H. Jaffe , Andrew R. Liddle , Pia Mukherjee , David Parkinson, Cambridge, UK: Cambridge University Press, 2009*. DOI: 10.1017/CBO9780511802461.

Hobson, M.P. and C. McLachlan (2003). "A Bayesian approach to discrete object detection in astronomical data sets". In: *mnras* 338, pp. 765–784. DOI: 10.1046/j.1365-8711.2003.06094.x. eprint: astro-ph/0204457.

Hogg, David W. and Daniel Foreman-Mackey (2017). "Data analysis recipes: Using Markov Chain Monte Carlo". In: DOI: 10.3847/1538-4365/aab76e. eprint: arXiv:1710.06068.

Jeffreys, Harold (1946). "An Invariant Form for the Prior Probability in Estimation Problems". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Phys-*

ical Sciences 186.1007, pp. 453–461. ISSN: 00804630. DOI: 10.2307/97883. URL: http://dx.doi.org/10.2307/97883.

Johnston, S. et al. (2008). "Science with ASKAP". In: *Experimental Astronomy* 22.3, pp. 151–273. ISSN: 1572-9508. DOI: 10.1007/s10686-008-9124-7. URL: http://dx.doi.org/10.1007/s10686-008-9124-7.

Knuth, Kevin H. et al. (2014). "Bayesian Evidence and Model Selection". In: DOI: 10.1016/j.dsp.2015.06.012. eprint: arXiv:1411.3013.

Lewis, Antony and Sarah Bridle (2002). "Cosmological parameters from CMB and other data: A Monte Carlo approach". In: 66.10, 103511, p. 103511. DOI: 10.1103/PhysRevD.66.103511. arXiv: astro-ph/0205436 [astro-ph].

Lochner, Michelle (2014). "New Applications of Statistics in Astronomy and Cosmology". PhD dissertation. University of Cape Town.

Lochner, Michelle et al. (2015). "Bayesian inference for radio observations". In: 450, pp. 1308–1319. DOI: 10.1093/mnras/stv679. arXiv: 1501.05304 [astro-ph.IM].

Lucas, L. et al. (2019). "Efficient source finding for radio interferometric images". In: *Astronomy and Computing* 27, 96, p. 96. DOI: 10.1016/j.ascom.2019.02.002. arXiv: 1901.04956 [astro-ph.IM].

Lukic, V. et al. (2019). *AutoSource: Radio-astronomical source-finding with convolutional autoencoders*. arXiv: 1910.03631 [astro-ph.IM].

Lumen Astronomy: Radiation and Spectra (2019). https://courses.lumenlearning.com/astronomy/chapter/the-electromagnetic-spectrum/. [Online; accessed 2019].

Maartens, Roy et al. (2015). "Cosmology with the SKA – overview". In: *arXiv e-prints*, arXiv:1501.04076, arXiv:1501.04076. arXiv: 1501.04076 [astro-ph.CO].

MacKay, David J.C (2003). *Information Theory,Inference,and Learning Algorithms*. Cambridge University Press. ISBN: 0-521-64298-1. URL: http://www.journals.cambridge.org/abstract_S026357470426043X.

Mai, Thuy (2019). *What are radio waves?* https://www.nasa.gov/directorates/heo/scan/communications/outreach/funfacts/txt_radio_spectrum.html. [Online; accessed 2019].

Masias, M. et al. (2012). "A review of source detection approaches in astronomical images". In: *Monthly Notices of the Royal Astronomical Society* 422.2, pp. 1674–1689. DOI: 10.1111/j.1365-2966.2012.20742.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2966.2012.20742.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2012.20742.x.

Metropolis, Nicholas et al. (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. DOI: 10.1063/1.1699114. eprint: https://doi.org/10.1063/1.1699114. URL: https://doi.org/10.1063/1.1699114.

Mohan, Niruj and David Rafferty (2015). *PyBDSF: Python Blob Detection and Source Finder*. ascl: 1502.007.

Mukherjee, P. et al. (2006). "A Nested Sampling Algorithm for Cosmological Model Selection". In: 638, pp. L51–L54. DOI: 10.1086/501068. eprint: astro-ph/0508461.

Murray, Iain and Zoubin Ghahramani (2005). *A note on the evidence and Bayesian Occam's razor*. Tech. rep. GCNU-TR 2005-003. Gatsby Computational Neuroscience Unit, University College London.

Myung, In Jae (2003). "Tutorial on maximum likelihood estimation". In: *Journal of Mathematical Psychology* 47.1, pp. 90–100. ISSN: 0022-2496. DOI: https://doi.org/10.1016/S0022-2496(02)00028-7. URL: http://www.sciencedirect.com/science/article/pii/S0022249602000287.

Nadarajah, Saralees (2005). "A generalized normal distribution". In: *Journal of Applied Statistics* 32.7, pp. 685–694. DOI: 10.1080/02664760500079464. eprint: https://doi.org/10.1080/02664760500079464. URL: https://doi.org/10.1080/02664760500079464.

Portillo, Stephen K. N. et al. (2017). "Improved Point-source Detection in Crowded Fields Using Probabilistic Cataloging". In: *The Astronomical Journal* 154.4, p. 132. ISSN: 1538-3881. DOI: 10.3847/1538-3881/aa8565. URL: http://dx.doi.org/10.3847/1538-3881/aa8565.

R. A. Fisher, F. R. S. (1921). "On the Mathematical Foundations of Theoretical Statistics." In:

Ravenzwaaij, Don van et al. (2018). "A simple introduction to Markov Chain Monte–Carlo sampling". In: *Psychonomic Bulletin & Review* 25.1, pp. 143–154. ISSN: 1531-5320. DOI: 10.3758/s13423-016-1015-8. URL: https://doi.org/10.3758/s13423-016-1015-8.

Retrê, João et al. (2019). AU Press Officer Garching bei München, Germany.

Robert, Christian P. et al. (2008). "Harold Jeffreys's Theory of Probability Revisited". In: *arXiv e-prints*, arXiv:0804.3173, arXiv:0804.3173. arXiv: 0804.3173 [math.ST].

Rosenberg, Marissa et al. (2013). *Why is Astronomy Important?* arXiv: 1311.0508 [astro-ph.IM].

Rothkopf, Alexander (2019). *Bayesian techniques and applications to QCD*. eprint: arXiv:1903.02293.

Ruli, Erlis et al. (2015). "Improved Laplace Approximation for Marginal Likelihoods". In: *arXiv e-prints*, arXiv:1502.06440, arXiv:1502.06440. arXiv: 1502.06440 [stat.CO].

SKA (2019). *The SKA is often called a Big Data project. What does it mean?* https://www.skatelescope.org/frequently-asked-questions/, [Online; accessed 2019].

Skilling, John (2006). "Nested sampling for general Bayesian computation". In: *Bayesian Anal.* 1.4, pp. 833–859. DOI: 10.1214/06-BA127. URL: https://doi.org/10.1214/06-BA127.

Speagle, Joshua S (2019). "dynesty: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences". In: *arXiv e-prints*, arXiv:1904.02180, arXiv:1904.02180. arXiv: 1904.02180 [astro-ph.IM].

Stark, Philip B. (2015). *SticiGui "Chapter 13 Probability: Philosophy and Mathematical Background"*. URL: http://www.freetextbooklist.com/sticigui/.

Syversveen, Anne Randi (1998). *Noninformative Bayesian Priors. Interpretation And Problems With Construction And Applications.*

Tak, Hyungsuk et al. (2018). "How proper are Bayesian models in the astronomical literature?" In: 481.1, pp. 277–285. DOI: 10.1093/mnras/sty2326. arXiv: 1712.03549 [astro-ph.IM].

Tiao, George E.P. Box George C. (1973a). *Bayesian Inference in Statistical Analysis*. Philippines: ADDISON-WESLEY PUBLISHING COMPANY. ISBN: 0-201-00622-7. URL: https://www.academia.edu/32086308/Inference_in_Statistical_Analysis_1973_George_E._P._Box_George_C._Tiao-Bayesian_.pdf.

– (1973b). *Bayesian Inference in Statistical Analysis*. Philippines: ADDISON-WESLEY PUBLISHING COMPANY. ISBN: 0-201-00622-7. URL: https://www.academia.edu/32086308/Inference_in_Statistical_Analysis_1973_George_E._P._Box_George_C._Tiao-Bayesian_.pdf.

Trotta, Roberto (2008). "Bayes in the sky: Bayesian inference and model selection in cosmology". In: *Contemporary Physics* 49.2, pp. 71–104. DOI: 10.1080/00107510802066753. arXiv: 0803.4089 [astro-ph].

Vafaei Sadr, A et al. (2019). "DeepSource: point source detection using deep learning". In: *Monthly Notices of the Royal Astronomical Society* 484.2, pp. 2793–2806. ISSN: 1365-2966. DOI: 10.1093/mnras/stz131. URL: http://dx.doi.org/10.1093/mnras/stz131.

van der Sluys, Marc et al. (2008). "Parameter estimation of spinning binary inspirals using Markov chain Monte Carlo". In: *Classical and Quantum Gravity* 25.18, 184011, p. 184011. DOI: 10.1088/0264-9381/25/18/184011. arXiv: 0805.1689 [gr-qc].

Vehtari, Aki et al. (2019). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC". In: *arXiv e-prints*, arXiv:1903.08008, arXiv:1903.08008. arXiv: 1903.08008 [stat.CO].

Venna, Jarkko et al. (2003). "Visualizations for Assessing Convergence and Mixing of MCMC". In: *Machine Learning: ECML 2003*. Ed. by Nada Lavrač et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 432–443. ISBN: 978-3-540-39857-8.

Zhao, Qiang et al. (2013). "Variance Reduction Techniques of Importance Sampling Monte Carlo Methods for Pricing Options". In: *Journal of Mathematical Finance* 03, pp. 431–436. DOI: 10.4236/jmf.2013.34045.