



BAYESIAN MODEL SELECTION WITH APPLICATIONS TO RADIO ASTRONOMY

Arrykrishna Mootoovaloo

February 2017

A project submitted in partial fulfilment of the requirements for the degree M.Sc.

in the Department of Astronomy, as part of the National Astrophysics

and Space Science Programme

UNIVERSITY OF CAPE TOWN

Supervisor: Prof. Bruce Bassett

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.

PLAGIARISM DECLARATION

I, Arrykrishna Mootoovaloo, know the meaning of plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is my own.

ACKNOWLEDGEMENT

It was the best decision I took to join the Cosmology Group at AIMS (African Institute for Mathematical Sciences) under the SKA SA (Square Kilometre Array South Africa) Masters Bursary. Above all, I am fortunate to be supervised by Bruce, who has been a phenomenal supervisor. He did not only teach me how to do good research but also prepared me to survive a fast-evolving research community. He gave me the freedom to pursue my ideas and the encouragement to attend various local and international conferences and workshops. During this period, I also had the opportunity to attend the 66th Lindau Nobel Laureate Meeting in Germany under the Horst Köler Fellowship. I am grateful for having been his student.

I would also like to thank Prof Martin Kunz (University of Geneva) and Prof Andrew Jaffe (Imperial College), who have constantly been my best mentors in Bayesian Statistics. I was fortunate to visit both University of Geneva and the ICIC (Imperial Centre for Inference and Cosmology), where Prof Martin and Prof Andrew made the needful to cover my local expenses respectively. Moreover, thanks to Dr Michelle Lochner, I had the opportunity to visit the Astrophysics group at the University College London (UCL).

Moreover, I would also like to thank the SKA SA and NASSP (National Astrophysics and Space Science Programme) for supporting my MSc financially. Kurt (director of NASSP), Nicky and Nuraan (previous NASSP administrators), Olwethu (current NASSP administrator) and Siphelo (NASSP system administrator) have always been very helpful. I want to thank the 2015 and 2016 students, post-docs and staffs for making the Cosmology Group an enthusiastic learning environment: Bruce, David, Diego, Eli, Emmanuel, Ethan, Etienne, Gilad, Kai, Pierre-Yves, Rene, Romeel, Michelle, Shankar and Yabebal. I also want to express my sincere gratitude to Nadeem, Trienko and Griffin have been very helpful in clarifying my doubts in radio astronomy. Most importantly, I am grateful to my parents, brothers and sister who have been instrumental in my life.

ABSTRACT

This thesis consists of two main parts, both of which focus on Bayesian methods and the problem of model selection in particular. The first part investigates a new approach to computing the Bayes factor for model selection without needing to compute the Bayesian evidence, while the second part shows, through an analytical calculation of the Bayesian evidence, that Bayesian methods allow two point sources to be distinguished from a single point source at angular separations that are much smaller than the naive beam size at high signal to noise.

In the first part, the idea is to create a supermodel by combining two models using a hyperparameter, which we call α . Setting $\alpha = 0$ or 1 switches each of the models off. Hence, the ratio of the posterior of α at the two end points (0 or 1) gives the Bayes Factor. This effectively converts the problem of model selection into a Bayesian inference problem. One can then use a standard Markov Chain Monte Carlo method to map the posterior distribution of α and compute the Bayes factor.

In the second part of this thesis, the Bayesian radio interferometry formalism of [Lochner et al. \(2015\)](#) is extended to take into account the gains of the antennae using the StEFCal algorithm, an important part of the calibration pipeline. Finally we study the case of a pair of sources and show that they can be resolved using an analytical computation of the Bayesian evidence. This demonstrates that Bayesian methods allow super-resolution: the pair of sources can be distinguished from a single source at arbitrarily small scales compared to the naive beam size, as long as the measurements have sufficient signal to noise.

THESIS OUTLINE

The first three chapters presented in this thesis summarise the literature and can be regarded as review chapters. As we move on to the next 3 chapters (Chapters 4, 5 and 6), we address two topics which are related to Bayesian Model Selection. In Chapters 4 and 5, we present a new way to do Bayesian Model Comparison while in Chapter 6, we use Bayesian methods to do super-resolution analysis in radio astronomy.

Chapter (1) - An Introduction to Bayesian Statistics: This chapter starts the thesis with the core concepts of Bayesian Parameter Estimation and Bayesian Model Selection. We also explain some of the algorithms used in the astronomy community.

Chapter (2) - Concepts of Radio Astronomy: In this chapter, we provide a brief overview of the concepts used in radio astronomy. We start with the key concepts in positional astronomy followed by an explanation of a simple two-element interferometer.

Chapter (3) - RIME and BIRO: We explain the Radio Interferometric Measurement Equation (RIME), which simplifies the approach to radio interferometry by using Jones matrices. This concept is at the heart of various radio astronomical softwares such as MeqTrees and Montblanc. Through this lens, we also review the key results from the Bayesian Inference for Radio Observation (BIRO) paper, in which [Lochner et al. \(2015\)](#) showed that one could infer both systematic and scientific parameters in the presence of Direction-Dependent Effects (DDEs).

Chapter (4) - A New Approach to Model Selection Using Supermodels: In this chapter, we use existing methods to determine the best-fit parameters given a model and also to select the best-fitting model by calculating the Bayesian Evidence. We then introduce our new methods which deal with the evaluation of the log-Bayes Factor directly. We test and compare our results using the Gaussian Linear Model and we show numerically that they all agree.

Chapter (5) - Testing the Supermodel Approach with MCMC: Here, we provide a simple example to show demonstrate the Metropolis-Hastings algorithm. Additionally, having shown that our method agrees from an analytical point of view from the previous chapter, we now attempt to use MCMC methods to infer the full posterior distribution of the hyperparameter. We show that we obtain reasonable values of the log-Bayes Factor for both the linear and non-linear models. We also investigate reparameterizations of the hyperparameter.

Chapter (6) - Calibration and Super-Resolution in Bayesian Radio Astronomy: In this chapter, we use simulated visibilities not only to recover previous BIRO results but also to show the underlying concept of this new formalism in radio astronomy. We implement a calibration algorithm to correct for the antenna gains followed by Bayesian Inference. Most importantly, we also show that one resolve two point sources at scales much smaller than the beam size at sufficiently high signal-to-noise ratio.

Chapter (7) - Conclusions: We conclude in the last chapter and discuss future work which can be implemented using our methods.

RELATIONSHIP TO ONGOING PUBLICATIONS

Chapter 4 and 5: These two chapters involve joint work with my supervisor, Prof. Bruce Bassett and Prof. Martin Kunz. The paper is available at [arXiv:1609.02186](https://arxiv.org/abs/1609.02186)

Chapter 3 and 6: This work is done collaboratively with my supervisor, Prof. Bruce Bassett and Prof. Andrew Jaffe. It deals mostly with super-resolution in radio astronomy using Bayesian Methods and is still currently under preparation.

Contents

Table of Contents	v
List of Figures	vi
List of Tables	vii
List Of Abbreviations	viii
Symbols	ix
1 An Introduction to Bayesian Statistics	1
1.1 Bayesian Probability	1
1.1.1 What is Probability?	1
1.1.2 A Short Review of Probability Theory	2
1.1.3 Bayes' Theorem	3
1.2 Bayesian Parameter Inference	4
1.2.1 MCMC	5
1.2.2 Samplers	5
1.2.3 Convergence and Burn-In	9
1.2.4 Generating Correlated Samples	9
1.3 Bayesian Model Comparison	11
1.3.1 Bayesian Evidence and Bayes Factor	12
1.3.2 Bayesian Evidence with Nested Sampling	14
1.3.3 Savage-Dickey Density Ratio	15
1.3.4 Approximate Model Selection Methods	15
2 Concepts of Radio Astronomy	17
2.1 Introduction	17
2.1.1 Why Radio Science?	18
2.1.2 Science With Modern Arrays	19

2.2	Coordinate Systems	20
2.2.1	The Horizontal (Alt-Azimuth) System	20
2.2.2	The Equatorial System	21
2.2.3	Direction Cosine System	22
2.2.4	Coordinate Transformation - Equatorial and Direction Cosines	23
2.3	Basic Antenna Theories	24
2.3.1	Beamwidth	25
2.3.2	Beam Solid Angle	25
2.3.3	Directivity	26
2.3.4	Efficiency	26
2.3.5	Effective Aperture	27
2.4	Interferometry	27
2.4.1	The Two-Element Interferometer	28
2.4.2	Relationship Between Visibility and Intensity	30
2.4.3	Earth-Rotation Aperture Synthesis	33
3	RIME and BIRO	36
3.1	Introduction	36
3.2	Deconvolution in Radio Astronomy	37
3.2.1	CLEAN	37
3.3	RIME	39
3.3.1	Stokes Parameters	39
3.3.2	The Visibility Matrix	40
3.3.3	Coherency Matrix	42
3.3.4	RIME for Single Point Source	43
3.3.5	RIME for Multiple Point Sources	44
3.3.6	RIME for Continuous Brightness Distribution	44
3.4	BIRO	46
3.4.1	MeqTrees and Montblanc	46
3.4.2	Parameter Estimation	47
3.4.3	Model Selection	49
4	A New Approach to Model Selection Using Supermodels	51
4.1	Parameter Estimation	51

4.1.1	Chi-Square Fitting	52
4.1.2	Use of Normal Equations	53
4.1.3	MLE and MAP	55
4.1.4	Marginalisation	56
4.2	Model Selection	58
4.2.1	Evidence and Bayes Factor	59
4.2.2	Savage-Dickey Density Ratio	61
4.2.3	Our Method	62
4.2.3.1	General Approach	63
4.2.3.2	Transformation of α and Model Averaged Posteriors	64
4.2.3.3	More than two models	65
4.2.3.4	Using the Same Parameters vs Different Parameters	65
4.2.3.5	Combined Likelihood Approach	66
4.2.3.6	Combined Model Approach	68
5	Testing The Supermodel Approach with MCMC	71
5.1	Introduction	71
5.2	Parameter Estimation - MCMC Method	71
5.3	Bayes Factor Estimation	73
5.3.1	Combined Likelihood	73
5.3.2	Combined Model	76
5.4	Non-Linear Model	78
5.4.1	Data and Models	78
5.4.2	Combined Likelihood Result	79
5.5	Transformation of the Hyperparameter	79
5.5.1	Combined Likelihood	80
5.5.2	Combined Model	83
5.6	Non-Nested Models	84
6	Calibration and Super-Resolution in Bayesian Radio Astronomy	87
6.1	Introduction	87
6.2	Calibration and BIRO	87
6.2.1	Setting up the Measurement Equation	88
6.2.2	Example of Calibration	91

6.2.3	Bayesian Inference	92
6.2.3.1	Measurement Equation	92
6.2.3.2	Likelihood and Prior	93
6.2.3.3	Procedure	93
6.2.3.4	Results and Discussions	94
6.3	Super-Resolution with Bayesian Methods	97
6.3.1	Likelihood and Prior	97
6.3.2	Expected log-Bayes Factor	98
6.3.3	Analysis Without Beam	99
6.3.4	Analysis with WSRT Beam	101
6.3.5	Discussion	105
7	Conclusions	106
A	Topics Related to Bayesian Methods	109
A.1	The Gelman-Rubin Test	109
A.2	The SDDR	110
A.3	Deriving the Bayesian Evidence	112
A.4	Gibbs Sampling	113
A.5	Precision for log-Bayes Factor	115
B	Coordinate Transformations	118
B.1	Spherical Trigonometry	118
B.2	Equatorial and Horizontal Coordinate Systems	119
B.3	Baseline Representation in ENU Coordinates	120
B.4	Baseline Representation in Equatorial System	120
B.5	Baseline Representation in (u, v, w) Space	121
	Bibliography	123

List of Figures

Figure 1.1	Nested sampling illustration	14
Figure 2.1	Atmospheric transmission through different windows	17
Figure 2.2	Different views of Abell 2052	18
Figure 2.3	The horizontal coordinate system	20
Figure 2.4	The Equatorial Coordinate System	22
Figure 2.5	The direction cosines coordinate system	23
Figure 2.6	Coordinate system and pattern for antenna	25
Figure 2.7	The basic two-element interferometer	28
Figure 2.8	Position vectors defining the position of the source	30
Figure 2.9	Transformation from the (θ, ϕ) coordinates to the (l, m) coordinates .	31
Figure 2.10	Example of uv plane sampling	34
Figure 2.11	uv -coverage for an East-West interferometer	35
Figure 3.1	Example of the CLEAN algorithm applied to a dirty image	38
Figure 3.2	The total Jones matrix	40
Figure 3.3	Example of expression tree	47
Figure 3.4	Correlations among a subset of scientific and instrumental parameters	48
Figure 3.5	Resolving power of BIRO	49
Figure 4.1	The data points are shown here with their associated errors	51
Figure 4.2	Best fit to the linear model using MAP method	59
Figure 4.3	The posterior distribution of α for the combined model approach . . .	69
Figure 5.1	The marginalised posterior distributions of the parameters using MCMC	72
Figure 5.2	Likelihood of the hyperparameter α for the linear models	74
Figure 5.3	Likelihood of α for different number of samples	75
Figure 5.4	Normalised Posterior and CDF of α	77
Figure 5.5	The toy non-linear model	78

Figure 5.6	The inferred likelihood for α for the toy non-linear problem	79
Figure 5.7	MCMC run for linear model under the new reparameterisation	80
Figure 5.8	The likelihood of α as we vary ℓ in the linear models	81
Figure 5.9	The likelihood of α given ℓ , for the non-linear models	82
Figure 5.10	The correlation of the parameters ω and ϕ in both methods	82
Figure 5.11	Normalised posterior distribution of reparametrisation of α	83
Figure 5.12	Simulated data using sinusoidal function - The best fits (blue and green) to the data using models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, are shown here. These parameters are determined using the <code>scipy.optimize</code> module where $(a, b, c) = (-0.423, 1.291, -0.049)$ and $(\omega, \phi) = (1.048, -0.025)$	84
Figure 5.13	The autocorrelation time, τ plotted against the thinning factor n . As expected, larger thinning factor leads to less correlation among the samples.	85
Figure 5.14	The likelihood of α in the case of un-nested models - The broken vertical blue line shows the log-Bayes Factor obtained by running PyMultinest twice for the two models. The green curve corresponds to the likelihood of α using our combined-likelihood approach.	86
Figure 6.1	Array of telescopes	89
Figure 6.2	Calibration for a single point source and a single baseline	92
Figure 6.3	The posterior distribution for the flux density of source $S_{(3)}$	95
Figure 6.4	Plot of the inferred flux densities with the true flux densities	95
Figure 6.5	Correlation plot of the inferred flux parameters	96
Figure 6.6	Plot of the log-Bayes Factor against δl in arcsec	101
Figure 6.7	Plot of the log-Bayes Factor against δl in arcsec (with beam)	103
Figure A.1	2D posterior plot using Gibbs Sampling	115
Figure B.1	Spherical triangle representation	118
Figure B.2	Conversion between equatorial and horizontal coordinate system	119
Figure B.3	XYZ baseline representation	121
Figure B.4	Baseline representation in the (u, v, w) space	121

List of Tables

Table 1.3.1	The Jeffreys' scale	13
Table 4.2.1	Bayesian Evidence for each linear model considered	60
Table 5.2.1	Comparing the inferred parameters using different methods	73
Table 6.2.1	Notations for work on Calibration and BIRO	88
Table 6.2.2	Results from calibration and BIRO	94
Table 6.3.1	Notation for work on super-resolution	97

LIST OF ABBREVIATIONS

AIC	: Akaike Information Criteria
BIC	: Bayesian Information Criteria
BIRO	: Bayesian Inference for Radio Observations
DDEs	: Direction-Dependent Effects
DIEs	: Direction-Independent Effects
i.i.d	: independently identically distributed
HMC	: Hamiltonian Monte Carlo
MCMC	: Markov Chain Monte Carlo
MH	: Metropolis-Hastings algorithm
NS	: Nested Sampling
PDF	: Probability Density Function
SDDR	: Savage-Dickey Density Ratio
SKA	: Square Kilometre Array
SNR	: Signal-to-Noise Ratio
StEFCal	: Statistically Efficient and Fast Calibration

SYMBOLS

Symbols	Meanings	Values/Units
λ	Wavelength	m
ν	Frequency	Hz
a	Altitude/Elevation	°
A	Azimuth	°
ϕ	Latitude of Observer	°
δ	Declination	°
H	Hour Angle	° or h
$\mathbf{b}$	Baseline vector	m
c	Speed of light	$3 \times 10^8 \text{ m s}^{-1}$
$\mathcal{D}$	Observed Data	
$\mathcal{P}$	Probability Distribution Function	
$\mathbb{Z}$	Bayesian Evidence	
$\mathcal{M}$	Underlying Model	
$\mathbf{J}$	Jones Matrix	

AN INTRODUCTION TO BAYESIAN STATISTICS

"The theory of probabilities is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it. It is remarkable that this science, which originated in the consideration of games of chance, should have become the most important object of human knowledge." - Pierre-Simon Laplace

1.1 Bayesian Probability

Bayesian tools and methods are increasingly being applied in the field of astrophysics and cosmology. Bayes' theorem, in particular, was discovered by Thomas Bayes ([Bayes 1763](#)) but only published three years after his death. While Bayes used uniform priors in his work, Pierre-Simon Laplace, who was totally unaware of Bayes' work, developed the more general formalism of the theorem ([Tzikas et al. 2008](#)). The applications of Bayesian methods are becoming popular thanks to the fast reliable computing facilities available today. The two most common questions are Bayesian parameter estimation and model comparison. These two are explored later in this work. We first begin by defining probability.

1.1.1 What is Probability?

To undertake statistical analysis one naturally requires the notion of probability. While seemingly intuitive, there are two very different interpretations of what probability actually means. One approach is that of *Bayesian probability* which is actually a measure of the degree of belief of a proposition ([Trotta 2008](#)). It can also be defined as a measure of uncertainty about the occurrence of an event ([Murphy 2012](#)). The other approach is the classical or the frequentist way, where probability is interpreted as the ratio of the number of successful events to the total

number of trials, in the long run of frequencies of the events. For example, after tossing an unbiased coin for a large number of times.

However, the interpretation of probability in the frequentist way has been subject to a great deal of disagreement. It assumes that there is a known probability of occurrence of a specific event while our aim is still to define probability. In short, we are in a vicious circle trying to figure out an exact, accurate definition of probability in the frequentist approach. Moreover, frequentists rely on repeatability of an experiment. If an experiment is not repeatable, then the concept of interpreting probability in the frequentist approach becomes more debatable. Furthermore, two other important criteria in statistical analysis are *prior information* and *nuisance parameters*. Bayesians can deal with nuisance parameters by simply integrating over them, leaving behind just the probability distribution for the relevant parameters. On the other hand, some analyses require prior information to make solid inferences. We will use the Bayesian approach to probability in this thesis.

1.1.2 A Short Review of Probability Theory

Here we briefly review the foundations of probability as a stepping stone to the Bayesian approach to inference and model selection. Let $\mathcal{P}(X)$ indicates the probability that an event X is true. As an example, this may be the event that a die lands on six. On the other hand, $\mathcal{P}(\bar{X})$ indicates that the event is not X and will simply be equal to $1 - \mathcal{P}(X)$. Some of the fundamental rules in probability theory are:

(1) Probability of union of two events X and Y

The probability that either X or Y occurs is defined as

$$\mathcal{P}(X \cup Y) = \mathcal{P}(X) + \mathcal{P}(Y) - \mathcal{P}(X \cap Y) \quad (1.1.1)$$

where $\mathcal{P}(X \cap Y)$ is the probability of both X and Y occurring. If X and Y are mutually exclusive, then

$$\mathcal{P}(X \cup Y) = \mathcal{P}(X) + \mathcal{P}(Y) \quad (1.1.2)$$

(2) Joint event

The probability of the joint event X and Y is $\mathcal{P}(X \cap Y)$ and given by

$$\mathcal{P}(X \cap Y) = \mathcal{P}(X|Y) \mathcal{P}(Y) \quad (1.1.3)$$

where $\mathcal{P}(X|Y)$ implies the probability of X given Y . This is also sometimes referred to as the **product rule**. Moreover, if we have a joint distribution $\mathcal{P}(X, Y)$, then, **the marginal distribution** of X is given by

$$\mathcal{P}(X) = \sum_y \mathcal{P}(X|Y=y) \mathcal{P}(Y=y) \quad (1.1.4)$$

where the sum is over all the possible states of Y . This is also known as the **sum rule**.

(3) Conditional probability

The conditional probability of X , given that Y is true is given by

$$\mathcal{P}(X|Y) = \frac{\mathcal{P}(X, Y)}{\mathcal{P}(Y)} \quad (1.1.5)$$

1.1.3 Bayes' Theorem

Bayes' theorem is simply a result of an easy manipulation of the fundamental rule of joint probability. Suppose, we have some information $\mathcal{I}$ which is assumed to be true and on which both the events X and Y depend upon. The fact that $\mathcal{P}(X, Y|\mathcal{I}) = \mathcal{P}(Y, X|\mathcal{I})$ means that the joint probability of X and Y obeys:

$$\mathcal{P}(X|Y, \mathcal{I}) \mathcal{P}(Y|\mathcal{I}) = \mathcal{P}(Y|X, \mathcal{I}) \mathcal{P}(X|\mathcal{I})$$

$$\mathcal{P}(X|Y, \mathcal{I}) = \frac{\mathcal{P}(Y|X, \mathcal{I}) \mathcal{P}(X|\mathcal{I})}{\mathcal{P}(Y|\mathcal{I})} \quad (1.1.6)$$

where

$$\mathcal{P}(Y|\mathcal{I}) = \mathbb{E}[\mathcal{P}(Y|X, \mathcal{I})] = \begin{cases} \sum_x \mathcal{P}(Y|X, \mathcal{I}) \mathcal{P}(X|\mathcal{I}) & \text{if } X \text{ is discrete} \\ \int_x \mathcal{P}(Y|X, \mathcal{I}) \mathcal{P}(X|\mathcal{I}) dx & \text{if } X \text{ is continuous} \end{cases} \quad (1.1.7)$$

where $\mathbb{E}$ refers to expectation (refer to Equation 1.2.4). If we interpret X as being a set of parameters θ , Y being a set of data $\mathcal{D}$, while $\mathcal{I}$ being a model $\mathcal{M}$ on which our inference is based upon, then,

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M})}{\mathcal{P}(\mathcal{D}|\mathcal{M})} \quad (1.1.8)$$

A detailed explanation of each of quantity in the Bayesian parameter inference context is provided in the next section.

1.2 Bayesian Parameter Inference

Consider the case where we have a new set of data, $\mathcal{D}$ and a model $\mathcal{M}$ which is governed by a set of parameters $\boldsymbol{\theta}$. These parameters can represent some relevant physical quantities which we are interested in as well as nuisance parameters. Our aim is to use the data and the model to find the parameters that best-fit the data. Unlike the frequentist approach which relies on the likelihood function, $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ only, the Bayesian approach strictly requires the likelihood function as well as some prior information on the parameters. For example, the prior for a new data set may be taken as the posterior from a previous experiment. In short, the prior should reflect the state of belief we have on the parameters before we look at the data.

The joint posterior is then given by Bayes' Theorem,

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M})}{\mathcal{P}(\mathcal{D}|\mathcal{M})} \quad (1.2.1)$$

where $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ is the likelihood function which should typically reflect the way the data is obtained. For example, if a measurement is corrupted by Gaussian noise, then a Gaussian distribution is the most suitable function to construct the likelihood. $\mathcal{P}(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution for the parameters while $\mathcal{P}(\mathcal{D}|\mathcal{M})$ is the marginal likelihood or Bayesian evidence which is not relevant in parameter inference as it does not depend on $\boldsymbol{\theta}$ and is simply a proportionality constant. However, the Bayesian evidence is crucial in Bayesian model selection.

If $\boldsymbol{\Phi}$ is the set of relevant physical quantities while $\boldsymbol{\Psi}$ is the set of nuisance parameters, then the marginal posterior of the parameters of interest is given by integrating over $\boldsymbol{\Psi}$:

$$\mathcal{P}(\boldsymbol{\Phi}|\mathcal{D}, \mathcal{M}) \propto \int \mathcal{P}(\mathcal{D}|\boldsymbol{\Phi}, \boldsymbol{\Psi}, \mathcal{M}) \mathcal{P}(\boldsymbol{\Phi}, \boldsymbol{\Psi}|\mathcal{M}) d\boldsymbol{\Psi} \quad (1.2.2)$$

Equation (1.2.2) can be computed analytically in a few cases only, for example in a Gaussian linear model (see Chapter 4). In practice, the posterior distribution is often mapped using Markov Chain Monte Carlo methods.

1.2.1 MCMC

Markov Chain Monte Carlo (MCMC) is an algorithm that allows one to simulate posterior distributions (Metropolis et al. 1953, Licia 2007). It has been applied to a wide range of problems in astrophysics, especially for Bayesian parameter estimation. Bayesian parameter estimation often involves high-dimensional integration which might be difficult to compute (Christensen et al. 2001). If θ is a vector of m parameters, the posterior probability distribution of a single parameter $\theta^{(i)}$ is obtained by integrating out the remaining parameters, that is,

$$\mathcal{P}(\theta^{(i)} | \mathcal{D}) = \int \dots \int \mathcal{P}(\theta | \mathcal{D}) d\theta^{(1)} \dots d\theta^{(i-1)} d\theta^{(i+1)} \dots d\theta^{(m)} \quad (1.2.3)$$

Moreover, the posterior mean is obtained by one further integration,

$$\mathbb{E}_{\mathcal{P}(\theta^{(i)} | \mathcal{D})}(\theta^{(i)}) = \int \theta^{(i)} \mathcal{P}(\theta^{(i)} | \mathcal{D}) d\theta^{(i)} \quad (1.2.4)$$

Equation (1.2.3) and Equation (1.2.4) are referred to as marginalisation and expectation respectively. The main idea behind the Monte Carlo method is to draw an i.i.d (independently identically distributed) set of samples from a target density (Andrieu et al. 2003). One can approximate the integrals in Equation (1.2.4) as:

$$\mathbb{E}_{\mathcal{P}(\theta^{(i)} | \mathcal{D})}(\theta^{(i)}) \approx \frac{1}{N} \sum_j \theta_j^{(i)} \quad (1.2.5)$$

where j is the j^{th} sample of the $\theta^{(i)}$ parameter. If the MCMC is run for a sufficiently long time, then we are guaranteed to reach the stationary value of the parameter, eventually mimicking the actual true posterior distribution of the parameter. It further enables us to compute several other quantities, for example, the mean, variance and confidence interval for the posterior distribution.

1.2.2 Samplers

We now briefly discuss three algorithms for sampling, namely the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970), the Hamiltonian Monte Carlo algorithm (Hajian 2007) and the Gibbs sampler (Andrieu et al. 2003).

Metropolis-Hastings Algorithm: In this particular case, one takes an arbitrary step $\mathbf{u}$ from the current position θ selected from a proposal distribution, $q(\mathbf{u} | \theta)$ also known as a candidate-generating density (Chib & Greenberg 1995) or jump distribution. This step is accepted with

probability

$$\min \left\{ 1, \frac{\mathcal{P}(\mathbf{u}|\mathcal{D}, \mathcal{M}) q(\mathbf{u}|\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) q(\boldsymbol{\theta}|\mathbf{u})} \right\} \quad (1.2.6)$$

If $q(\mathbf{u}|\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\mathbf{u})$, that is we are sampling from a symmetric proposal distribution, for example, a Gaussian distribution, then we recover the original Metropolis algorithm. The Random Walk Metropolis-Hastings algorithm is as follows:

Algorithm 1.1 Random Walk Metropolis-Hastings

```

Initiate  $\boldsymbol{\theta}$ 
for  $i = 1, 2, \dots, N_{\text{steps}}$  do
  Sample  $\Delta\boldsymbol{\theta}$  from a proposal distribution  $q(\Delta\boldsymbol{\theta}|\boldsymbol{\theta})$ .
   $\mathbf{u} = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ 
  Draw  $\alpha \sim \mathcal{U}[0, 1]$ 
  if  $\alpha < \min \left\{ 1, \frac{\mathcal{P}(\mathbf{u}|\mathcal{D}, \mathcal{M}) q(\boldsymbol{\theta}|\mathbf{u})}{\mathcal{P}(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) q(\mathbf{u}|\boldsymbol{\theta})} \right\}$  then
     $\boldsymbol{\theta}(i+1) = \mathbf{u}$ 
  else
     $\boldsymbol{\theta}(i+1) = \boldsymbol{\theta}(i)$ 
  end if
end for

```

The choice of a proper proposal distribution is vital to ensure rapid convergence. If we take very small jumps, then it will take a long time to reach the stationary values of the parameters and will be inefficient. On the other hand, if the jumps are too big, then the peaks of the distributions remain unexplored, thus leading to the case where most of the steps are rejected in the algorithm. Therefore, it is important to monitor the acceptance rate, which is the fraction of the steps being accepted. If the acceptance rate is too high, then the chain is not mixing properly, indicating that it is not moving around the parameter space quickly enough. If the acceptance rate is small, then this implies that many candidate draws are being rejected. In particular, for the Random Walk Metropolis-Hastings algorithm, an acceptance rate of 25 % to 50 % is generally recommended.

Hamiltonian Monte Carlo Algorithm: The Hamiltonian Monte Carlo (HMC) is an algorithm in which a momentum variable is associated with each parameter and Hamiltonian dynamics are invoked. Let $\boldsymbol{\theta}$ and $\mathbf{p}$ represent the positions and the momenta respectively. The Hamiltonian of such a system is then given by the sum of the kinetic energy and the potential energy, that is,

$$\mathcal{H}(\boldsymbol{\theta}, \mathbf{p}) = U(\boldsymbol{\theta}) + K(\mathbf{p}) \quad (1.2.7)$$

The potential energy is chosen to be proportional to the posterior:

$$U(\boldsymbol{\theta}) = -\ln [\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M})] \quad (1.2.8)$$

where $\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})$ and $\mathcal{P}(\boldsymbol{\theta}|\mathcal{M})$ are the likelihood and prior function respectively. On the other hand, the kinetic energy is simply chosen to be:

$$K(\mathbf{p}) = \frac{1}{2} \mathbf{p}^T \mathbf{p} \quad (1.2.9)$$

The Hamiltonian is used to generate samples as follows

$$\begin{aligned} \mathcal{P}(\boldsymbol{\theta}, \mathbf{p}) &\propto \exp[-\mathcal{H}(\boldsymbol{\theta}, \mathbf{p})] \\ &= \exp[-U(\boldsymbol{\theta})] \exp[-K(\mathbf{p})] \\ &= \mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M}) \mathcal{N}(\mathbf{p}; 0, 1) \end{aligned}$$

where $\mathcal{N}(\mathbf{p}; 0, 1)$ is an n -dimensional Gaussian distribution with mean zero and unit variance and n is the number of parameters in our problem. The algorithm is then controlled by the Hamiltonian equations of motion:

$$\frac{d\theta_i}{dt} = p_i \quad (1.2.10)$$

$$\frac{dp_i}{dt} = -\frac{\partial \mathcal{H}}{\partial \theta_i} \quad (1.2.11)$$

Generally, the leapfrog algorithm is implemented to solve the Hamiltonian dynamics. If ϵ is a small step-size, then we can solve these equations using:

$$p_i\left(t + \frac{\epsilon}{2}\right) = p_i(t) - \frac{\epsilon}{2} \left(\frac{\partial U}{\partial \theta_i} \right)_{\boldsymbol{\theta}_i(t)}$$

$$\theta_i(t + \epsilon) = \theta_i(t) + \epsilon p_i\left(t + \frac{\epsilon}{2}\right)$$

$$p_i\left(t + \frac{\epsilon}{2}\right) = p_i(t) - \frac{\epsilon}{2} \left(\frac{\partial U}{\partial \theta_i} \right)_{\boldsymbol{\theta}(t+\epsilon)}$$

For each iteration in the algorithm, the step from the leapfrog move is accepted using the prob-

ability of

$$\min \{1, \exp [-\mathcal{H}(\boldsymbol{\theta}^*, \mathbf{p}^*) - \mathcal{H}(\boldsymbol{\theta}, \mathbf{p})]\} \quad (1.2.12)$$

Algorithm 1.2 Hamiltonian Monte Carlo

```

Initialise  $\boldsymbol{\theta}_0$ 
for  $i = 1$  to  $N_{\text{samples}}$  do
   $\mathbf{p} \sim \mathcal{N}(0, 1)$ 
   $(\boldsymbol{\theta}_{(0)}^*, \mathbf{p}_{(0)}^*) = (\boldsymbol{\theta}_{(i-1)}, \mathbf{p})$ 
  for  $j = 1$  to  $N$  do
    Make a leapfrog move:  $(\boldsymbol{\theta}_{(j-1)}^*, \mathbf{p}_{(j-1)}^*) \rightarrow (\boldsymbol{\theta}_{(j)}^*, \mathbf{p}_{(j)}^*)$ 
  end for
   $(\boldsymbol{\theta}^*, \mathbf{p}^*) = (\boldsymbol{\theta}_{(N)}^*, \mathbf{p}_{(N)}^*)$ 
  Draw  $\alpha \sim \mathcal{U}[0, 1]$ 
  if  $\alpha < \min \left\{ 1, e^{-[\mathcal{H}(\boldsymbol{\theta}^*, \mathbf{p}^*) - \mathcal{H}(\boldsymbol{\theta}_{(j)}^*, \mathbf{p}_{(j)}^*)]} \right\}$  then
     $\boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}^*$ 
  else
     $\boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}_{(i-1)}$ 
  end if
end for

```

The HMC is shown in detail in Algorithm (1.2). Note that in the leapfrog algorithm, an analytical form of the potential energy is required to calculate its gradient. In this particular case, an exploratory MCMC run can be carried out and then used to approximate the potential by a multivariate Gaussian. Moreover, we could also thin the chain to obtain uncorrelated samples.

There are several advantages of the HMC compared to the usual Metropolis-Hastings algorithm. In particular, HMC produces fewer correlated steps and improves acceptance rate and convergence.

The Gibbs Sampler: The Gibbs sampling (Geman & Geman 1984) method updates one of the parameters at a time, but requires the conditional distributions to be tractable. The advantage of Gibbs sampling is that it reduces the need for "tuning" of the proposal distribution as in the case for the Metropolis-Hastings algorithm. In particular, the starting point can be guessed or can be found using an optimisation algorithm. The algorithm for the Gibbs sampler is as follows:

Algorithm 1.3 The Gibbs Sampler

```

Initialise  $x^{(0)}$ 
for iteration  $i = 1$  to  $N_{\text{samples}}$  do
   $x_1^{(i)} \sim \mathcal{P} \left( X_1 = x_1 \mid X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)} \right)$ 
   $x_2^{(i)} \sim \mathcal{P} \left( X_2 = x_2 \mid X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \dots, X_D = x_D^{(i-1)} \right)$ 
   $\vdots$ 
   $x_D^{(i)} \sim \mathcal{P} \left( X_D = x_D \mid X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \dots, X_{D-1} = x_{D-1}^{(i)} \right)$ 
end for

```

An example of the Gibbs sampling method is illustrated in Appendix (A.4).

1.2.3 Convergence and Burn-In

Typically, the early stage of the chains is not in equilibrium. Therefore, this part is thrown away and is referred to as “burn-in” or “warm-up” (Geyer 1992). If the run is long enough to give precise results, then throwing away 1 or 2 % of the chain will typically suffice. Another way to decide how much should be thrown away is to find the first index, i , where the chain reaches half of the maximum posterior probability in the chain, and then remove these first i samples from the chain.

Another issue is that the points in an MCMC chain are correlated and it is important to have a convergence test. Usually, this problem can be reduced by “thinning” the chain, for example, by recording every n^{th} point from the chain instead of the whole chain. In this way, the samples are less correlated than the whole chain. Thinning also helps in reducing the amount of data stored for extremely long chains.

Several convergence diagnostics exist which include Gelman and Rubin, Raftery and Lewis, Geweke, Heidelberg and Welch amongst others (Cowles & Carlin 1996). The Gelman and Rubin (Gelman & Rubin 1992) diagnostic remains the most popular test to verify when it is appropriate to stop sampling and hence use the samples recorded to perform further analysis. The potential scale reduction factor (see Appendix A.1 for further details), $\sqrt{\hat{R}}$ should approach unity when convergence is reached. Otherwise, if it greater than 1.1 or perhaps 1.2, then the chain should be run longer.

1.2.4 Generating Correlated Samples

Parameters can be correlated while running an MCMC. The chains could take a very long time to attain convergence and could be inefficient. One way to nullify this effect is to take correlated movements (meaning generating correlated random numbers) via the covariance

matrix. Suppose we have m chains each with n elements, implying m parameters and n steps, then the matrix $\mathbf{X}$ which has as its *rows* representing each parameter, that is,

$$\mathbf{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & & & \vdots \\ X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{pmatrix}$$

the $(m \times m)$ covariance matrix Σ is given by

$$\Sigma = \mathbb{E} \left[(\mathbf{X} - \mathbb{E}[\mathbf{X}]) (\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \right] \quad (1.2.13)$$

or equivalently as

$$\Sigma_{i,j} = \mathbb{E} [(X_i - \mu_i) (X_j - \mu_j)] \quad (1.2.14)$$

where (i, j) are the positions of the entry elements for Σ and $\mu_i = \mathbb{E}(X_i)$. Note that $\Sigma_{ij} = \Sigma_{ji}$.

We want to find the matrix $\mathbf{U}$ such that $\mathbf{U}^T \mathbf{U} = \Sigma$. Given a set of uncorrelated random numbers $\mathbf{R}$, the correlated random numbers $\mathbf{R}_c$ will be given by

$$\mathbf{R}_c = \mathbf{R} \mathbf{U} \quad (1.2.15)$$

Typically, there are two ways of generating correlated random numbers in Monte Carlo simulations. These are the *Cholesky Decomposition* and the *Eigenvector Decomposition*, which is also known as Spectral Decomposition.

Cholesky Decomposition: The Cholesky factorisation method is simpler to understand. If $\mathbf{U}$ is the Cholesky matrix, in the most simple terms, $\mathbf{U}$ is then the square root of the covariance matrix, Σ . Given $\mathbf{U}$, one can directly find the correlated random numbers.

Eigenvector Decomposition: The eigenvectors $\mathbf{e}_i$ and eigenvalues λ_i of the covariance matrix Σ are given by

$$\Sigma \mathbf{e}_i = \lambda_i \mathbf{e}_i \quad (1.2.16)$$

If we define $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m]$ and $\Lambda = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$, then the eigensystem is

$$\Sigma \mathbf{E} = \Lambda \mathbf{E} \quad (1.2.17)$$

The covariance matrix, Σ can be formed by

$$\Sigma = \mathbf{E} \Lambda \mathbf{E}^T \quad (1.2.18)$$

Therefore,

$$\mathbf{U}^T \mathbf{U} = \mathbf{E} \Lambda \mathbf{E}^T$$

Hence,

$$\mathbf{U} = \sqrt{\Lambda} \mathbf{E}^T \quad (1.2.19)$$

One can use either method to generate correlated samples but the Eigenvector Decomposition is a more robust method for this process (Straka et al. 2013), especially in the case of an ill-conditioned covariance matrix. One would typically allow the chain to be built before carrying out the Eigenvector Decomposition to obtain $\mathbf{U}$ (Equation 1.2.19) followed by generating $\mathbf{R}_c$ (Equation 1.2.15). $\mathbf{R}$ is drawn from a Gaussian with zero mean and unit variance.

1.3 Bayesian Model Comparison

Bayesian model selection is becoming an increasingly important tool to determine whether or not the introduction of a new parameter is justified by the data (Kass & Raftery 1995, Trotta 2008).

One of the key questions underlying science is that of model selection: how do we select between competing theories which purport to explain observed data? The great paradigm shifts in science fall squarely into this domain. In the context of astronomy - as with most areas of science - the next two decades will see a massive increase in data volume through large surveys such as the Square Kilometre Array (SKA) (Hollitt et al. 2016) and LSST (Becla et al. 2006). Robust statistical analysis to perform model selection at scale will be a critical factor in the success of such future surveys.

Bayesian and frequentist answers to hypothesis testing could be quite different (Robert 1993). Through this lens, Lindley (1957) showed that this disagreement between the two approaches can be quite dramatic. In simple term, while a sharp null hypothesis can be strongly

rejected by the standard test of significance, it could be allocated high odds under the influence of a small prior probability in the Bayesian framework. This disagreement was first studied by Jeffreys (1998) but was first referred to as a paradox by Lindley (1957).

In addition, if we have very good data, we may not know when to stop fitting. Two competing models may equally fit the data properly but how do we choose the most appropriate model? The solution in fact lies in the simpler model being preferred. This is known as the *Oc-cam's razor* (Jefferys & Berger 1992, Liddle et al. 2006). The complex model which explains the data somewhat better compared to the simpler model should be penalised for the additional parameters that it introduces. Extra parameters bring about lack of predictability.

1.3.1 Bayesian Evidence and Bayes Factor

The Bayesian Evidence ($\mathbb{Z}$) is also called the *marginal likelihood* or the *model likelihood*. It is interpreted as the probability of the data, given the model, $\mathcal{P}(\mathcal{D}|\mathcal{M})$. It is commonly used as a tool to evaluate the performance of a specific model (Trotta 2008). The evidence is the normalisation integral of the product of the likelihood and the prior. Hence, given a model, the evidence can also be regarded as the average likelihood under the prior. The Bayesian Evidence is given as

$$\mathbb{Z} \equiv \mathcal{P}(\mathcal{D}|\mathcal{M}) = \int \mathcal{P}(\mathcal{D}|\theta, \mathcal{M}) \mathcal{P}(\theta|\mathcal{M}) d\theta \quad (1.3.1)$$

Using Bayes' theorem, we can write

$$\mathcal{P}(\mathcal{M}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}) \mathcal{P}(\mathcal{M})}{\mathcal{P}(\mathcal{D})} \quad (1.3.2)$$

The left-hand side of equation is the posterior probability of the model given the data. Consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$. The ratio of the respective posterior probabilities of the models, given the data is

$$\frac{\mathcal{P}(\mathcal{M}_1|\mathcal{D})}{\mathcal{P}(\mathcal{M}_2|\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_1) \mathcal{P}(\mathcal{M}_1)}{\mathcal{P}(\mathcal{D}|\mathcal{M}_2) \mathcal{P}(\mathcal{M}_2)} \quad (1.3.3)$$

Note that $\mathcal{P}(\mathcal{D})$ is only a constant and can be dropped when calculating the ratio of posterior probability of the models. The ratio $B_{12} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_1)}{\mathcal{P}(\mathcal{D}|\mathcal{M}_2)}$ is the Bayes factor and is, in fact, the ratio of the models' evidences. The ratio $\frac{\mathcal{P}(\mathcal{M}_1|\mathcal{D})}{\mathcal{P}(\mathcal{M}_2|\mathcal{D})}$ is the posterior odds while $\frac{\mathcal{P}(\mathcal{M}_1)}{\mathcal{P}(\mathcal{M}_2)}$ is the prior odds. Therefore, in words, one can describe Equation (1.3.3) as

$$\text{posterior odds} = \text{Bayes Factor} \times \text{prior odds}$$

In the case of non-committal priors on the models, that is, $\mathcal{P}(\mathcal{M}_1) = \mathcal{P}(\mathcal{M}_2)$ the posterior odds are just equal to the Bayes factor. As B_{12} increases, our belief that model $\mathcal{M}_1$ is better than model $\mathcal{M}_2$ increases. Otherwise, model $\mathcal{M}_2$ will be the preferred one. Therefore, from Equation (1.3.3), the Bayes factor indicates how the relative odds vary in the light of new data, irrespective of the prior odds of the models.

The Bayes factor is typically interpreted by referring to the Jeffreys' scale (Kass & Raftery 1995). It is an empirically determined scale as shown in Table (1.3.1).

$\ln(B_{12})$	B_{12}	Evidence against $\mathcal{M}_2$
0 to 1	1 to 3	Inconclusive
1 to 3	3 to 20	Weak Evidence
3 to 5	20 to 150	Moderate Evidence
> 5	> 150	Strong Evidence

Table 1.3.1 – The Jeffreys' scale - The table shows the relative strengths of one model compared to the other. In particular, there are four categories in which we can classify models. Table adapted from Kass & Raftery (1995), Trotta (2008).

Computing the Bayesian Evidence is a challenging task as it involves multi-dimensional integration over the whole parameter space. If the likelihood is sharply peaked about its maximum ($\hat{\theta}$), one can then use the Laplace approximation to estimate the Bayesian Evidence. In particular, Laplace's method allows to estimate integrals of the form

$$\int e^{-Nh(x)} dx$$

where N is typically the number of data points. If $h = -\frac{1}{N} [\log \mathcal{P}(\mathcal{D}|\hat{\theta}) + \log \mathcal{P}(\hat{\theta})]$, the estimation of the Bayesian Evidence using the Laplace approximation up to the first order is

$$\mathcal{Z} \approx \mathcal{P}(\mathcal{D}|\hat{\theta}) \mathcal{P}(\hat{\theta}) (2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}} N^{-\frac{d}{2}} \quad (1.3.4)$$

where $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} h$ and Σ is the inverse of the hessian of h evaluated at $\hat{\theta}$. Moreover, if one model is nested in another, then one can use the Savage-Dickey Density Ratio, which we discuss in the following sections.

1.3.2 Bayesian Evidence with Nested Sampling

Nested sampling ([Skilling 2004](#)) has been the preferred algorithm for computing the Bayesian Evidence in the cosmology community. The main idea behind nested sampling is to transform the multi-dimensional integration into 1-D integration. Suppose X is the fraction of total prior mass such that $dX = \mathcal{P}(\theta | \mathcal{M}) d\theta$, the Bayesian Evidence, $\mathbb{Z}$ can be written as

$$\mathcal{P}(\mathcal{D} | \mathcal{M}) = \int \mathcal{P}(\mathcal{D} | \theta, \mathcal{M}) \mathcal{P}(\theta | \mathcal{M}) d\theta = \int_0^1 \mathcal{L}(X) dX \quad (1.3.5)$$

where $\mathcal{L}(X) = \mathcal{P}(\mathcal{D} | \theta, \mathcal{M})$. If X is a sequence of decreasing values such that

$$0 < X_m \dots < X_2 < X_1 < 1,$$

the Bayesian Evidence can then be approximated numerically using the trapezoid rule such that

$$\mathbb{Z} = \sum_{j=1}^m \mathbb{Z}_j, \quad \mathbb{Z}_j = \frac{1}{2} \mathcal{L}_j (X_{j-1} - X_{j+1}). \quad (1.3.6)$$

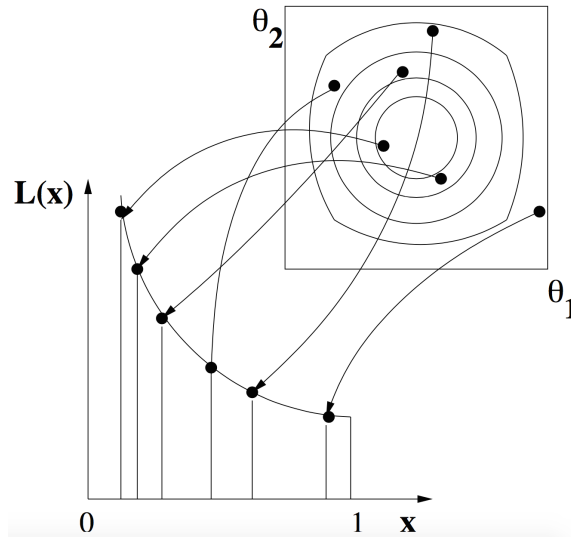


Figure 1.1 – Nested sampling illustration - The idea is to integrate over the likelihood surface from a prior volume by repeatedly updating the likelihood point or in other words, by discarding the previous lower likelihood point (figure adapted from [Mukherjee et al. \(2006\)](#)).

The algorithm proposed to estimate the Bayesian Evidence is as follows:

- (1) N points are sampled from the prior distributions and their likelihoods are then evaluated.

At this point, we will have the full prior range, that is, $X \in [0, 1]$.

- (2) The point with the lowest likelihood $\mathcal{L}(X_j)$ is selected, where $X_j = \exp\left(-\frac{j}{N}\right)$.
- (3) The evidence is incremented by $\mathbb{Z}_j = \frac{1}{2}\mathcal{L}_j(X_{j-1} - X_{j+1})$.
- (4) The lowest likelihood point is discarded and replaced with a new point, which is now uniformly distributed from the remaining prior volume, $X \in [0, X_j]$. However, the condition that $\mathcal{L} > \mathcal{L}_j$ must be satisfied.
- (5) Steps 2 - 4 are repeated until an appropriate accuracy condition is met.

1.3.3 Savage-Dickey Density Ratio

Savage-Dickey Density Ratio (SDDR) (Dickey 1971, Verdinelli & Wasserman 1995, Trotta 2007) is a useful method for comparing nested models and is in fact, a good approximation to the Bayes Factor. For some parameter values, the extended model $\mathcal{M}_1$ reduces to the simpler model $\mathcal{M}_0$. Consider a complex model made up of two sets of parameters $\theta = (\psi, \phi)$ and the simple model is obtained by setting $\phi = \phi_0$. If we further assume that the priors are separable, which is a common scenario:

$$\mathcal{P}(\psi, \phi | \mathcal{M}_1) = \mathcal{P}(\psi | \mathcal{M}_1) \mathcal{P}(\phi | \mathcal{M}_1) \quad (1.3.7)$$

then, the Bayes Factor can, in principle, be written as

$$B_{01} = \frac{\mathcal{P}(\phi | \mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\phi | \mathcal{M}_1)} \Big|_{\phi=\phi_0} \quad (1.3.8)$$

Equation (1.3.8) is referred to as the Savage-Dickey Density Ratio (see Appendix A.2 for its derivation). In short, the SDDR is simply the ratio of the marginal posterior to the prior of the complex model, evaluated at the nested point, that is, at the simpler model's parameters value.

The SDDR is important to judge whether the inclusion of extra parameters in nested models are important when we are trying to fit the observed data. Moreover, in most problems, it is computationally challenging to compute the Bayesian Evidence and hence the Bayes Factor, as it is a multi-dimensional integral. The SDDR provides an alternative avenue to compute the Bayes Factor without computing the Bayesian Evidence provided the models are nested.

1.3.4 Approximate Model Selection Methods

While the Bayesian Evidence remains the preferred choice for model selection, another approach is the use of information criteria. As the Bayesian Evidence is difficult to compute, the

information criteria provides an alternative way to compare models using the likelihood. If one compares the maximum likelihood for two different models, then the model with the higher number of parameters will always be favoured. Information criteria penalises models having extra parameters. While several methods of information criteria exist, for example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), Watanabe-Akaike Information Criterion (WAIC) amongst others (see for instance [Gelman et al. \(2014\)](#)), we discuss here AIC and BIC which have previously been used in the contexts of astrophysics and cosmology, for example, see [Liddle \(2004\)](#) and [Liddle \(2007\)](#).

Akaike Information Criterion (AIC): The AIC [Akaike \(1974\)](#) is a frequentist approach that combines maximum likelihood ($\mathcal{L}_{\max}$) and the number of free parameters (k) into a function that should be minimized:

$$\text{AIC} = -2 \ln \mathcal{L}_{\max} + 2k \quad (1.3.9)$$

where $\mathcal{L}_{\max} = \mathcal{P}(\mathcal{D} | \theta_{\max}, \mathcal{M})$ is the maximum likelihood value. The second term acts as an Occam's factor. As the number of parameters increases, the second term takes over. Hence, the typical shape of the AIC as a function of the number of parameters will be a decrease, a minimum followed by an increase ([Liddle et al. 2006](#)). The best model is the one which minimises the AIC. Moreover, the models need not be nested.

Bayesian Information Criterion (BIC): The BIC ([Schwarz et al. 1978](#)) resembles the AIC and is also referred to as the ‘‘Schwarz Information Criterion’’. The BIC is given by

$$\text{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N \quad (1.3.10)$$

where N is the number of data points. The BIC forces a stricter penalty compared to the AIC. The approximation is valid for large N ([Wit et al. 2012](#)). Again, the best model is the one which minimises the BIC. BIC comes from a Gaussian approximation to the Bayesian Evidence ([Wit et al. 2012](#)). However, this requires various assumptions, for example, the assumption that the data is from an exponential family, thus making the approximation sometimes unreliable.

[Gelman et al. \(2014\)](#) states that ‘‘the current state of the art of measurement of predictive model fit remains unsatisfying’’. These information criteria have their own advantages as well as drawbacks. They often fail in numerous examples and that the Bayesian Evidence is the better approach to model selection where feasible.

CONCEPTS OF RADIO ASTRONOMY

"If there is anything that can bind the heavenly mind of man to this dreary exile of our earthly home and can reconcile us with our fate so that one can enjoy living,—then it is verily the enjoyment of the mathematical sciences and astronomy." - Johannes Kepler

2.1 Introduction

Radio astronomy has a long-established history. Radio waves emanating from the Milky Way was serendipitously discovered by Karl Guthe Jansky in the 1930s ([Jansky 1933](#)) during his time at Bell Telephone Laboratories.

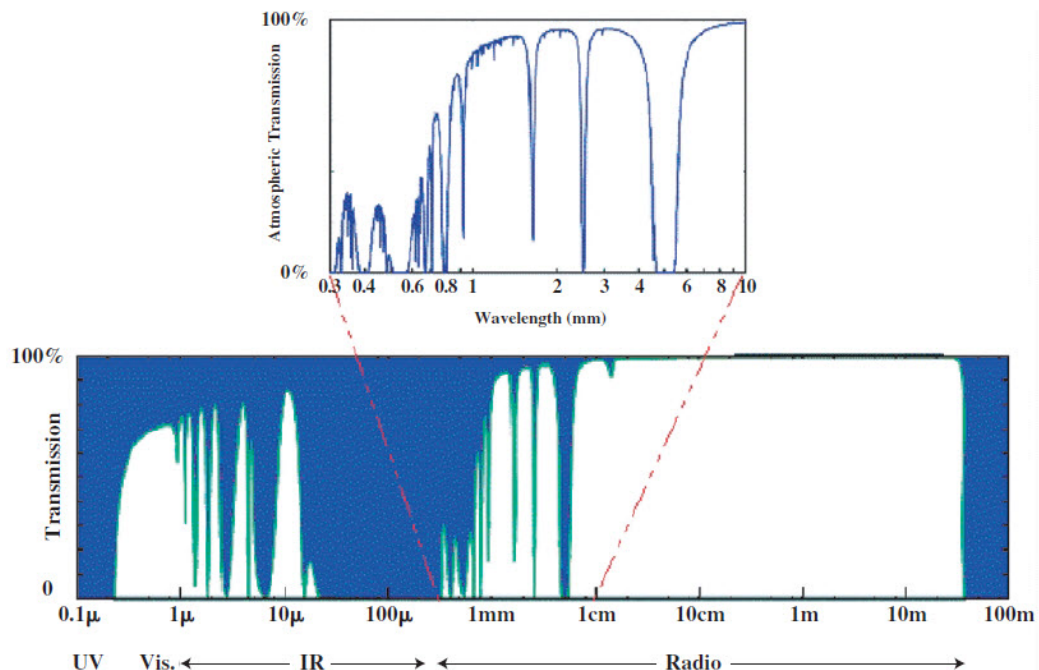


Figure 2.1 – Atmospheric transmission through different windows - The curve shows the fraction of radiation being transmitted. Note that the best transmission is obtained in the radio wavelengths compared to other regions of the electromagnetic spectrum. Figure courtesy of [Kutner \(2003\)](#).

His work was further investigated and extended by Grote Reber who also did the first sky survey in radio astronomy (Kellermann 2005). Martin Ryle, from the University of Cambridge was a pioneer in the development of interferometric techniques, namely aperture synthesis to obtain the location and images of weak radio sources, see Ryle et al. (1950) and Ryle & Hewish (1960).

In order to observe the sky, one can have a single large radio telescope carrying out the observation or we may have multiple telescopes coupled together via interferometric techniques (see §2.4.1). Moreover, the radio window of the electromagnetic spectrum lies in the frequency range of around 10 MHz – 1 THz implying a wavelength range of about 0.3 mm – 30 m (see for example Figure 2.1). It is actually difficult to define an appropriate window for radio observation as it depends highly on current technologies available, atmospheric transparency and noise (Condon & Ransom 2010).

2.1.1 Why Radio Science?

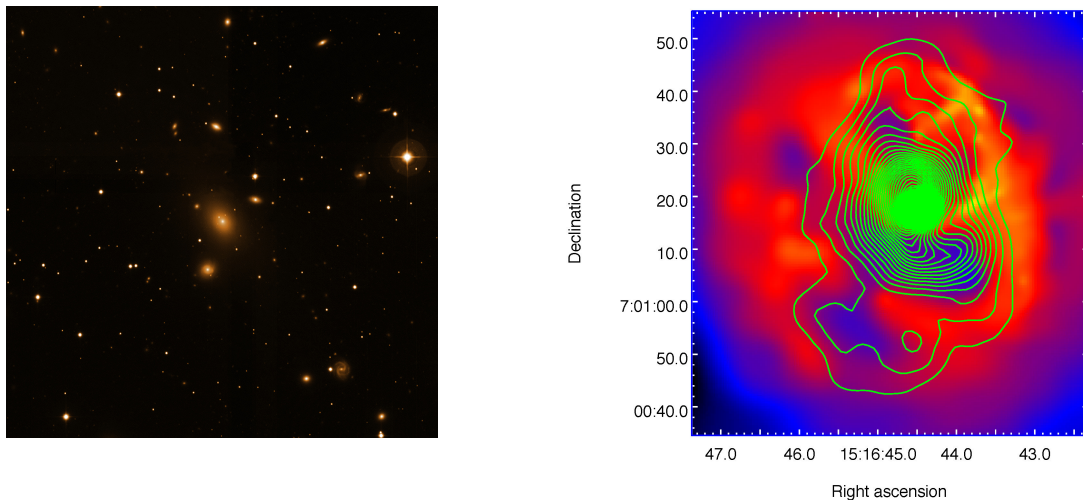


Figure 2.2 – Different views of Abell 2052 - The left panel shows the optical image (figure taken from the Chandra X-ray Observatory [Website*](http://chandra.harvard.edu/photo/2011/a2052/)) of the galaxy cluster, which is located 480 million light years away from Earth. The only things we see are stars and galaxies. Further, the right panel shows an overlay of the radio-X-ray of this cluster. The latter gives further details of the ongoing process towards the central region of this complex system. These cavities (bubbles) are filled with radio emission, which is evident from the superposition of the radio contours (figure adapted from my undergraduate thesis).

Our knowledge of the Universe outside the Earth initially came from purely visual techniques - a field of astronomy called optical astronomy. However, this is only limited to our vision, or in other words, to a restricted wavelength range of about 390 nm to 700 nm of the electromagnetic spectrum. In particular, with the advent of new types of telescopes, it is now possible to

*<http://chandra.harvard.edu/photo/2011/a2052/>

conduct astronomical observations throughout the whole electromagnetic spectrum, ranging from gamma-ray to radio wavelength. The difference between these two ends of the spectrum is roughly 10^{20} in frequency or energy, thus requiring very different techniques to carry out observations.

While different observation windows have their own advantages and disadvantages, here we focus on the radio part of the electromagnetic spectrum. The fact that the radio window is so broad, almost all sources as well as thermal and non-thermal phenomena can be observed at the radio wavelength (Condon & Ransom 2010). Moreover, at such wavelengths, dust scattering is negligible, hence allowing radio astronomers to see through the disk of our Milky Way Galaxy. In addition, low frequency means low photon energy. Hence, it is possible to trace, for example, the 21 cm line of neutral hydrogen in the radio regime (Condon & Ransom 2010). However, observing at radio wavelength is not without issues. Radio Frequency Interference (RFI) is a major concern. Starting from other celestial objects, for example, the Sun itself to the antennas and receivers, almost everything emit radio waves.

In general, the sky looks different in the radio regime compared to the optical window. For example, we may see stars and galaxies in a usual optical image (see Figure 2.2). At the centre of the cluster, several key activities are happening through the agency of a supermassive black hole. Jets emanating from the core create lobes which radiate at the radio wavelength. Moreover, these jets inflate bubbles or cavities which are detected at the radio frequency. Through the same lens, thermal and non-thermal are the two emission mechanisms which are detected at radio frequency, and this allows us to have a better understanding of the core activity sites of radio galaxies.

2.1.2 Science With Modern Arrays

The forthcoming modern interferometric arrays will improve our understanding of the Universe. The SKA (Lazio 2009) will operate at different observing frequencies and will consist of various array types. While the SKA remains a massive project to be accomplished, precursors such as MeerKAT (Booth et al. 2009) and ASKAP (Hopkins et al. 2015) are likely to be merged with the SKA upon completion. The key science projects related to the SKA are those of cosmology, galaxies, dark energy, epoch of reionisation, cosmic magnetism amongst others.

Through the same spectrum, the HERA (DeBoer et al. 2016) is a telescope designed to mark the epoch of reionisation. The frequency range is limited to 100 - 200 MHz which will enable the detection of emissions from neutral hydrogen, present before the birth of stars and galaxies.

In particular, this telescope will be the most sensitive one capable of studying the epoch of reionisation.

2.2 Coordinate Systems

In this section, we discuss two coordinate systems which are typically used in astronomy and we also describe briefly how to transform one coordinate system into the other. For more detail see, for example, [Roy & Clarke \(2003\)](#) and [Thompson et al. \(2008\)](#).

2.2.1 The Horizontal (Alt-Azimuth) System

The Horizontal Coordinate System is commonly used to point a telescope in a particular location of the local sky. In a naive way, the observer considers himself on a flat plane and at the centre of a huge hemisphere across which all celestial objects move. In this case, the observer is strictly at the origin.

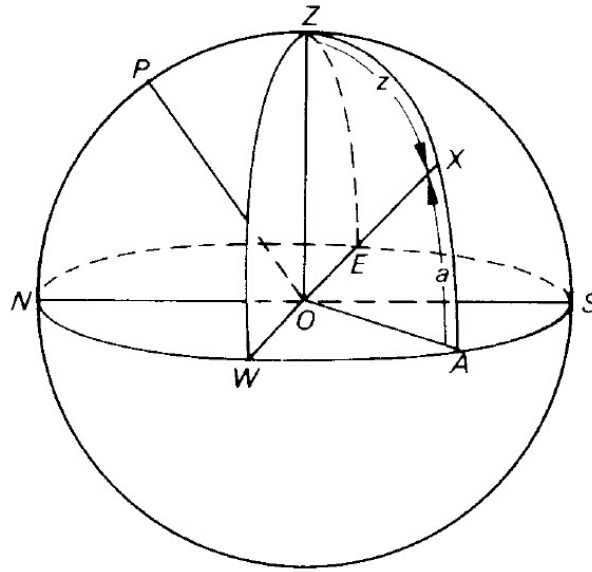


Figure 2.3 – The horizontal coordinate system - Two coordinates which are used in this system are the azimuth, A and the altitude, a as shown. Figure courtesy of [Roy & Clarke \(2003\)](#).

In Figure (2.3), O is the observer, Z is the zenith, P is the North Celestial Pole (NCP) and OX is the direction to the celestial object. The arcs ZN , ZW , ZS and so forth are called the verticals and the points N , W , E and S are referred to as the cardinal points. In particular, the arcs ZE and ZW are called the prime vertical east and prime vertical west respectively.

In the horizontal coordinate system, two quantities which are important in describing the location of a celestial object are the azimuth, A and the altitude, a . It is important to define the azimuth well. The most common way is to measure the azimuth from the north point (N)

eastwards from 0° to 360° for an observer in the northern hemisphere while for an observer in the southern hemisphere, the azimuth is measured eastwards from the south point (S) from 0° to 360° .

The altitude, a , is measure along the arc AXZ in which case the altitude is from A to X and is measured in degrees. An alternative to altitude is the zenith distance, z in which case

$$a = 90 - z \quad (2.2.1)$$

However, the main issue of the Horizontal Coordinate System is that it is a purely local system. Two observers at two different locations will measure the two quantities A and a differently for the same celestial object at the same time. Moreover, the observer will find the coordinates of the celestial body changing with time as the celestial sphere appears to rotate.

2.2.2 The Equatorial System

The Equatorial Coordinate System is one of the most common methods to define the position of a celestial object on the sky. If the plane of the Earth's equator is extended, it will then cut the celestial sphere and this great circle formed is called the celestial equator. This circle coincides with the celestial horizon at two points (W and E) as shown in Figure (2.4). These two points are always 90° from the points P and Z which are the poles for the celestial equator and celestial horizon respectively. Hence, W and E are referred to as the west and east point respectively.

Also, any great circle that goes through P and Q is called the meridian. From Figure (2.4), the heavier line $PZTSQ$ is the observer's meridian. An observer will see that all objects rise in the east and set in the west. In particular, a celestial object will follow the path in a small circle which is parallel to the celestial equator in the direction shown by the arrow. This circle is often called the parallel of declination leading to one of the coordinates in the Equatorial Coordinate System.

The declination, δ , is the angular distance measured from the equator along the meridian through the celestial object. It ranges from 0° to 90° and is positive north of the equator and negative south of the equator. This is analogous to the latitude of a place on Earth. Another quantity which is also often used is the north polar distance (arc PX in Figure 2.4) and is equal to:

$$\text{north polar distance} = 90^\circ - \delta. \quad (2.2.2)$$

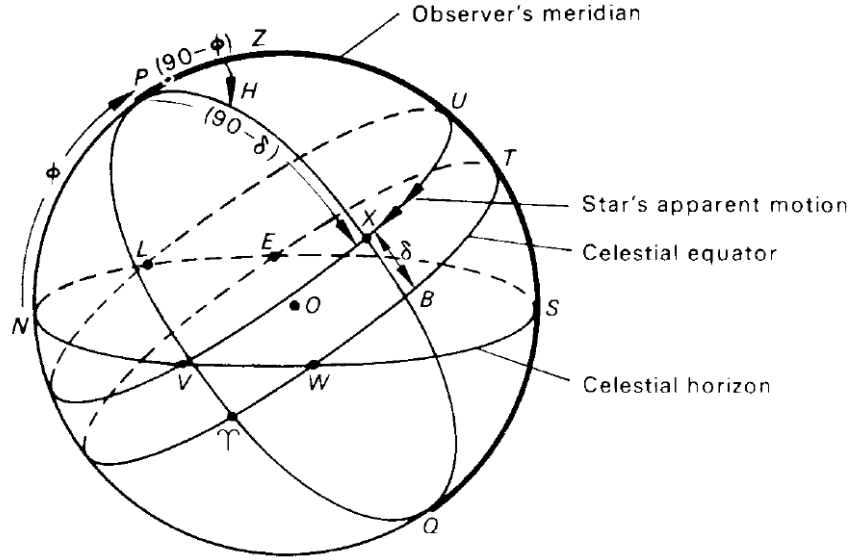


Figure 2.4 – The Equatorial Coordinate System - The two coordinates used to define the position of a celestial body in this frame are the declination, δ and the hour angle, H . ϕ denotes the latitude of the observer. Figure courtesy of [Roy & Clarke \(2003\)](#).

The other coordinate is the hour angle, which is angle ZPX . It is measured from the observer's meridian in the westwards direction to the meridian through the celestial object. It ranges from 0^h to 24^h or equivalently 0° to 360° . However, the hour angle changes with time and hence we need a coordinate such that the Equatorial Coordinate System stays static. One specific point which is chosen is the Vernal Equinox, also called the First Point of Aries, Υ from which the angle measured to the meridian through the celestial object is called the right ascension. Right ascension is measured from Υ eastwards in the direction opposite to which the hour angle is measured.

2.2.3 Direction Cosine System

The Direction Cosine System is not very different from the Equatorial Coordinate System. We do not necessarily need a fixed reference point, for example, Υ as in the Equatorial Coordinate System as we can redefine our own reference point. Moreover, the importance of the quantities used in this coordinate system will become evident when we deal with interferometric data. From Figure (2.5), the direction cosines are

$$\begin{aligned} l &= \cos(\alpha) = \frac{v_x}{|\vec{v}|} \\ m &= \cos(\beta) = \frac{v_y}{|\vec{v}|} \\ n &= \cos(\gamma) = \frac{v_z}{|\vec{v}|} \end{aligned} \quad (2.2.3)$$

If $|\vec{v}| = 1$, then the direction cosines are simply equivalent to the Cartesian coordinates. Also,

$$|\vec{v}|^2 = v_x^2 + v_y^2 + v_z^2 = l^2 + m^2 + n^2$$

If $|\vec{v}| = 1$, then

$$l^2 + m^2 + n^2 = 1 \quad (2.2.4)$$

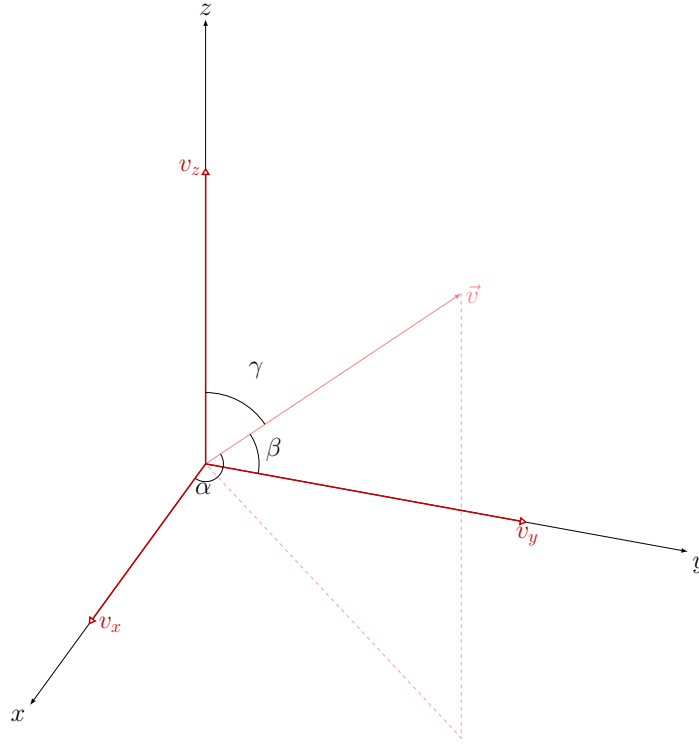


Figure 2.5 – The direction cosines coordinate system - $\cos(\alpha)$, $\cos(\beta)$, $\cos(\gamma)$ are referred to as the direction cosines. v_x , v_y , v_z denote the projection of the $\vec{v}$ along the three axes (x , y , z).

We will later encounter the fact that in radio interferometry, it is very common to express n in terms of the other two direction cosines, that is,

$$n = \sqrt{1 - l^2 - m^2} \quad (2.2.5)$$

2.2.4 Coordinate Transformation - Equatorial and Direction Cosines

Among various other coordinate transformations (refer to other examples in Appendix B), here, we focus on the transformation between the equatorial coordinate system and the direction cosines coordinate system (Taylor et al. 1999), which is important in radio astronomy. If (α_o, δ_o) are the coordinates (right ascension and declination respectively) of the phase centre in the equatorial coordinate frame, then:

$$\begin{aligned}
l &= \cos \delta \sin \Delta\alpha \\
m &= \sin \delta \cos \delta_o - \cos \delta \sin \delta_o \cos \Delta\alpha \\
n &= \sin \delta \sin \delta_o + \cos \delta \cos \delta_o \cos \Delta\alpha
\end{aligned} \tag{2.2.6}$$

On the other hand, the transformation from direction cosines coordinates to the equatorial coordinate system is given by

$$\begin{aligned}
\delta &= \sin^{-1} \left(m \cos \delta_o + \sqrt{1 - l^2 - m^2} \sin \delta_o \right) \\
\alpha &= \alpha_o + \tan^{-1} \left(\frac{1}{\sqrt{1 - l^2 - m^2} \cos \delta_o - m \sin \delta_o} \right)
\end{aligned} \tag{2.2.7}$$

where $\Delta\alpha = \alpha - \alpha_o$.

2.3 Basic Antenna Theories

In its simplest form, an antenna can be defined as a means to transmit or receive radio waves. In particular, for a radio telescope, the antenna acts as a collector of radio waves (Kraus 1966). A single dish in radio astronomy can be viewed as having a circular aperture. The larger the aperture, the larger is the collecting area. This implies a higher sensitivity, which is a telescope's ability to distinguish between the minimum signal and the noise background. For a single dish,

$$A \propto D^2 \tag{2.3.1}$$

where A is the collecting area and D is the diameter of the dish. Moreover, another important quantity is the *angular resolution*. It is simply the telescope's ability to distinguish between two objects separated by a certain distance. For a single dish, the angular resolution, $\Delta\theta$ is given by the Rayleigh's criterion,

$$\Delta\theta = 1.220 \frac{\lambda}{D} \tag{2.3.2}$$

where λ is the observing wavelength.

The response of an antenna as a function of direction is its power pattern. The latter consists of lobes as shown in Figure (2.6). In particular, the lobe having the largest maximum is referred to as the *main lobe*. Any other lobe, excluding the main lobe, is referred to as a minor lobe. In general, a side lobe is one which is adjacent to the main lobe while a back lobe is one which is roughly at an angle of 180° to the main beam (Balanis 2016).

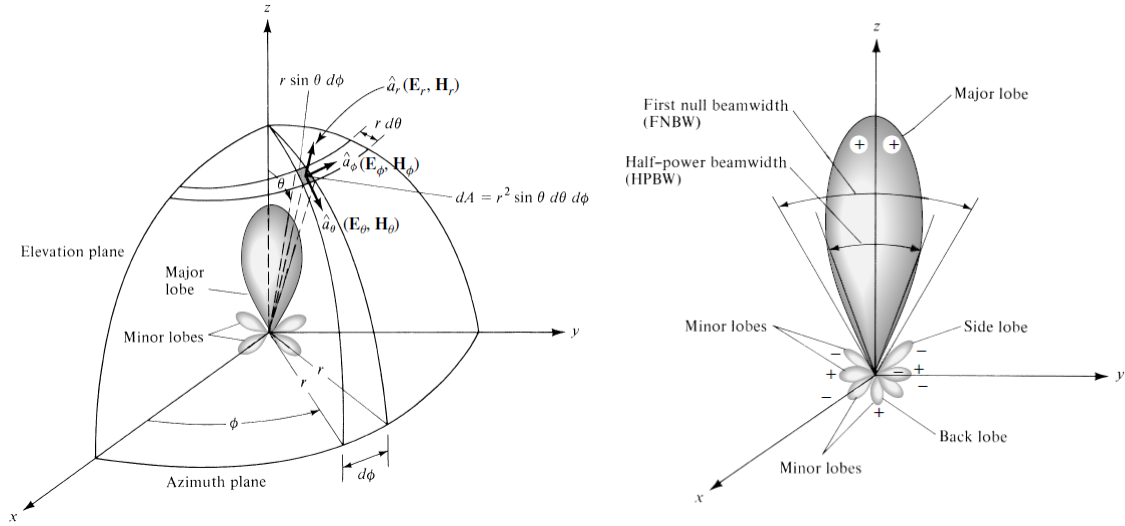


Figure 2.6 – Coordinate system and pattern for antenna - The left panel shows the coordinate system which is typically used for antenna analysis. The x, y, z refers to the Cartesian coordinate system while the θ, ϕ, r are the elevation, azimuth and radial components in the spherical coordinate system. The right panel gives a better description of the antenna pattern, which consists of major, minor, side and back lobes. Moreover, the FNBW and HPBW are also shown. Figure courtesy of [Balanis \(2016\)](#).

2.3.1 Beamwidth

The beamwidth is another important quantity associated with the power pattern of an antenna. It is usually defined as the angular separation between two similar points on the opposite sides of the main lobe. One commonly used beamwidth is the HPBW (half power beamwidth), which is the beamwidth corresponding to an angle between the two directions in which the radiation is half of the maximum. In general, for most single dish telescope, a good approximation of the HPBW is

$$\theta_{\text{HPBW}} \approx 1.2 \frac{\lambda}{D} \quad (2.3.3)$$

Hence, the HPBW is commonly referred to as the resolving power of the dish. Two sources separated by the HPBW can just be resolved using the Rayleigh criterion ([Condon & Ransom 2010](#)). Another common measure of the resolution is first-null beamwidth (FNBW), which is the angular separation between the first nulls of the pattern ([Balanis 2016](#)). Resolution will play an important role in this thesis.

2.3.2 Beam Solid Angle

If the antenna power pattern is a function of the angles θ and ϕ is $P(\theta, \phi)$, then the beam solid angle is the angle through which all the power would stream. This also requires that the power is maximum and constant over this angle. Then, the beam solid angle, Ω_A in rad^2 is

$$\Omega_A = \int_{4\pi} P_n(\theta, \phi) d\Omega \quad (2.3.4)$$

where P_n is the normalised power pattern, that is, $P_n = \frac{P}{P_{\max}}$ and is a dimensionless quantity. $d\Omega$ is an elemental solid angle ($d\Omega = \sin\theta d\theta d\phi$)

2.3.3 Directivity

Another important parameter is directivity, D which is simply the ratio of the maximum antenna intensity to the average antenna intensity, that is,

$$D = \frac{U(\theta, \phi)_{\max}}{U_{\text{avg}}} \quad (2.3.5)$$

$$\text{Since } U_{\text{avg}} = \frac{1}{4\pi} \int_{4\pi} U(\theta, \phi) d\Omega,$$

$$D = \frac{4\pi}{\Omega_A} \quad (2.3.6)$$

Hence, the directivity is defined as the ratio of the solid angle of a sphere, 4π to the beam solid angle.

2.3.4 Efficiency

The ratio of the main beam solid angle to the total beam solid angle is referred to as the beam efficiency, that is,

$$\epsilon_M = \frac{\Omega_M}{\Omega_A} \quad (2.3.7)$$

where $\Omega_M = \int_{\text{Main Beam}} P_n(\theta, \phi) d\Omega$. Further, the ratio of the minor lobe solid angle to the total beam solid angle is the stray factor, ϵ_m given by

$$\epsilon_m = \frac{\Omega_m}{\Omega_A} \quad (2.3.8)$$

Hence,

$$\epsilon_M + \epsilon_m = 1 \quad (2.3.9)$$

2.3.5 Effective Aperture

The effective aperture is related to the beam solid angle and wavelength as

$$A_e = \frac{\lambda^2}{\Omega_A} \quad (2.3.10)$$

for an isotropic and lossless antenna. We also have,

$$D = \frac{4\pi}{\lambda^2} A_e \quad (2.3.11)$$

In general, the gain of an antenna in terms of the effective aperture and efficiency, k is given by

$$G = k \frac{4\pi}{\lambda^2} A_e \quad (2.3.12)$$

2.4 Interferometry

What we typically want from a single dish is high sensitivity and fine resolution. The obvious way to satisfy these two conditions is to have a large aperture area (see Equations 2.3.1 and 2.3.2). The sensitivity is directly proportional to the collecting area while the angular resolution is inversely proportional to the diameter of the dish. For comparison, say we want an amateur telescope which has an angular resolution of 1 arcsec. In the optical regime, for $\lambda = 500$ nm, $D \approx 13$ cm while at a low frequency of 300 MHz ($\lambda = 1$ m), the diameter of the dish is $D \approx 250$ km! This is not only a costly process but also clearly undesirable and is infeasible to be built from an engineering perspective.

If we want to have both mechanical stability and high angular resolution, the best option is to have an interferometric array of telescopes. A modern radio interferometer typically consists of dishes which sample and digitise the incoming wavefronts. The trick in interferometry is that, instead of capturing all the signals at a single point, this can be done using smaller, reflector focus points. We now have an array of smaller dishes with time delays, one for each signal. The digital operation of reconstructing the original signal is referred to as *beamforming*. The easiest way is to use a correlator which computes the visibilities from which an image can be constructed (Stanimirovic 2002).

2.4.1 The Two-Element Interferometer

A simple case of an interferometric system is that of a two-element interferometer which consists of a pair of antennas separated by the baseline vector $\mathbf{b}$. The two voltage outputs from these two elements are then combined together, for example, multiplied and averaged. $\mathbf{s}_0$ is the vector pointing in the direction of the phase centre.

Let us assume that Antenna 2 is taken as the reference antenna (refer to Figure 2.7). The fact that the source is inclined causes an extra optical path length between Antenna 1 and Antenna 2. The projected baseline is given by:

$$|\mathbf{b}| \sin \theta \quad (2.4.1)$$

The optical path difference, also referred to as the *geometric delay*, ΔL is given by

$$\Delta L = |\mathbf{b}| \cos \theta = c \tau_g = \mathbf{b} \cdot \mathbf{s}_0 \quad (2.4.2)$$

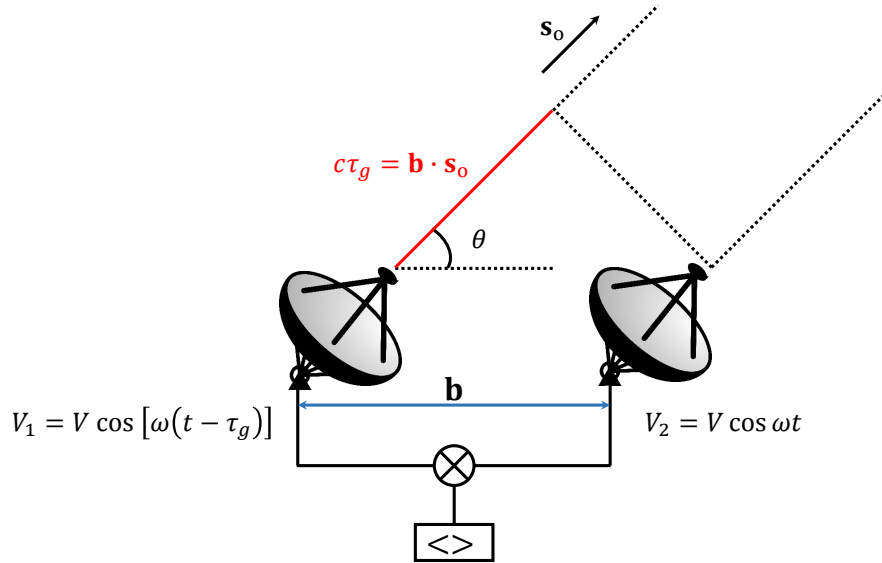


Figure 2.7 – The basic two-element interferometer - Two antennae receive electromagnetic signal from a distant source on the celestial sphere but the output from this system is in voltage format, hence V_1 and V_2 for each antenna. These two signals are multiplied and time-averaged by the correlator. But before doing so, we must take into account the geometric delay shown in red.

In other words, ΔL is simply the projection of the baseline vector $\mathbf{b}$ onto $\mathbf{s}_0$. Before we combine the two signals, we need to compensate for this delay. Signals can be combined in two ways, either

- (1) by adding the two signals, hence we have the sum-interferometer or

(2) by multiplying the two signals in which case we have the product interferometer.

In this work, we will explain the product interferometer. The device which multiplies the two signals is called a correlator and also does averaging in time. Consider the two signals

$$\begin{aligned} V_1 &= V \cos [\omega (t - \tau_g)] \\ V_2 &= V \cos (\omega t) \end{aligned} \quad (2.4.3)$$

The correlation of these two signals can be written as

$$C = \langle V_1 V_2 \rangle_t \quad (2.4.4)$$

The product of these two signals is given by

$$V_1 V_2 = \frac{V^2}{2} [\cos (2\omega t - \omega \tau_g) + \cos \omega \tau_g] \quad (2.4.5)$$

and we assume that the averaging time is long enough such that $\Delta t \gg (2\omega)^{-1}$ (Condon & Ransom 2010) in order to remove the fast-varying component of the signal, that is, the term $\cos (2\omega t - \omega \tau_g)$. The averaging in time is given by:

$$\langle V_1 V_2 \rangle_t = \frac{1}{\Delta t} \int_t^{t+\Delta t} \frac{V^2}{2} [\cos (2\omega t - \omega \tau_g) + \cos \omega \tau_g] dt \quad (2.4.6)$$

such that

$$\langle V_1 V_2 \rangle_t = \frac{V^2}{2} \left[\frac{\sin (2\omega t + 2\omega \Delta t - \omega \tau_g) - \sin (2\omega t - \omega \tau_g)}{2\omega \Delta t} + \cos \omega \tau_g \right] \quad (2.4.7)$$

Under the above assumption,

$$C_{\cos} = \frac{V^2}{2} \cos \omega \tau_g = \frac{V^2}{2} \cos \left(\frac{2\pi \Delta L}{\lambda} \right) \quad (2.4.8)$$

Hence, we see that the correlation depends on the optical path difference, that is, the geometric delay. Note that the above cosine correlator is sensitive to the even part of the sky brightness. In order to have the odd part, we can have a sin correlator such that

$$C_{\sin} = \frac{V^2}{2} \sin \omega \tau_g \quad (2.4.9)$$

2.4.2 Relationship Between Visibility and Intensity

In this section, we are going to derive the relationship between visibility and intensity, which is the cornerstone of radio interferometry. Consider the field centre which is in the direction of the unit vector $\mathbf{s}_0$. In particular, an element of the solid angle $d\Omega$ at position $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\sigma}$ contributes to a power (see Figure 2.8) of

$$dP = \frac{1}{2} A(\sigma) I(\sigma) \Delta\nu d\Omega \quad (2.4.10)$$

at each antenna, where $\Delta\nu$ is the bandwidth of the receiving element and $I(\sigma)$ is the distribution of the source intensity (Thompson et al. 2008). In the above equation, P is measured in Watts (W) whereas the units of $I(\sigma)$ is $\text{W m}^{-2} \text{Hz}^{-1} \text{Sr}^{-1}$. The factor of $\frac{1}{2}$ is due to the fact that the antenna feeds are polarised and hence sensitive to only half of the total power.

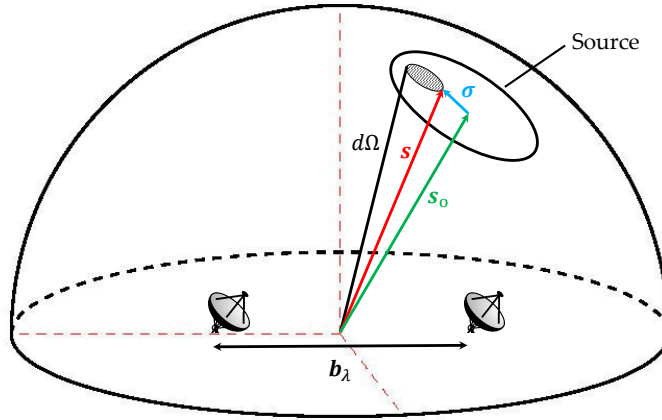


Figure 2.8 – Position vectors defining the position of the source - $\mathbf{s}_0$ is in the direction of the phase reference centre and $\mathbf{s}$ is the vector defining a general area element on the source. $d\Omega$ denotes the solid angle for this area element. The source is the one described by the ellipse shown above.

The output from the correlator is proportional to the received power and to the fringe pattern (refer to Equation 2.4.8), $\cos 2\pi\nu\tau_g$. Since

$$\nu\tau_g = \mathbf{b}_\lambda \cdot \mathbf{s} = \mathbf{b}_\lambda \cdot (\mathbf{s}_0 + \boldsymbol{\sigma})$$

the output from the correlator is given by

$$r(\mathbf{b}_\lambda, \mathbf{s}_0) = \Delta\nu \int_{4\pi} A(\sigma) I(\sigma) \cos [2\pi\mathbf{b}_\lambda \cdot (\mathbf{s}_0 + \boldsymbol{\sigma})] d\Omega \quad (2.4.11)$$

Using the identity

$$\cos (a \pm b) = \cos a \cos b \mp \sin a \sin b$$

We can write:

$$\begin{aligned} r(\mathbf{b}_\lambda, \mathbf{s}_0) &= \Delta\nu \cos (2\pi \mathbf{b}_\lambda \mathbf{s}_0) \int_{4\pi} A(\sigma) I(\sigma) \cos (2\pi \mathbf{b}_\lambda \sigma) d\Omega \\ &\quad - \Delta\nu \sin (2\pi \mathbf{b}_\lambda \mathbf{s}_0) \int_{4\pi} A(\sigma) I(\sigma) \sin (2\pi \mathbf{b}_\lambda \sigma) d\Omega \end{aligned}$$

If A_0 is the collecting area in the direction $\mathbf{s}_0$, the normalised pattern is

$$A_N(\sigma) = \frac{A(\sigma)}{A_0}$$

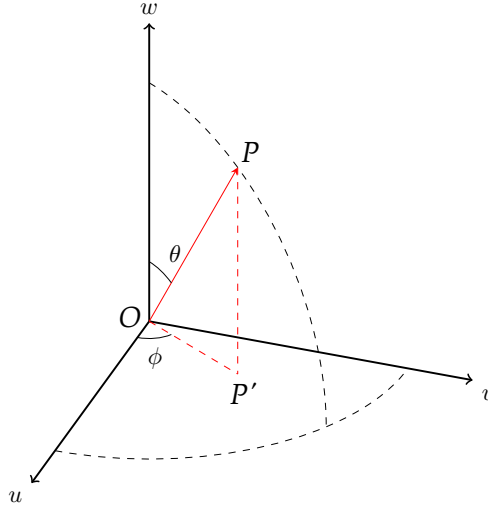


Figure 2.9 – Transformation from the (θ, ϕ) coordinates to the (l, m) coordinates - The vector $\vec{OP}$ is projected on the uv -plane such that it has components l and m along the u and v axes respectively. In particular, we note that $\sin \theta = \frac{m}{l}$ and $\tan \phi = \frac{m}{l}$. These two relationships can be used to transform one coordinate system (θ, ϕ) to another (l, m) by computing the determinant of the Jacobian. Hence, $d\Omega = \sin \theta d\theta d\phi = \frac{dl dm}{\sqrt{1-l^2-m^2}}$.

If we define the visibility as:

$$\mathcal{V} = |\mathcal{V}| e^{j\phi_{\mathcal{V}}} = \int_{4\pi} A_N(\sigma) I(\sigma) e^{-2\pi j \mathbf{b}_\lambda \cdot \sigma} d\Omega \quad (2.4.12)$$

then, the real and imaginary parts of the visibility are given respectively as

$$\begin{aligned} \Re(\mathcal{V}) &= \int_{4\pi} A_N(\sigma) I(\sigma) \cos (2\pi j \mathbf{b}_\lambda \cdot \sigma) d\Omega = |\mathcal{V}| \cos \phi_{\mathcal{V}} \\ \Im(\mathcal{V}) &= \int_{4\pi} A_N(\sigma) I(\sigma) \sin (2\pi j \mathbf{b}_\lambda \cdot \sigma) d\Omega = -|\mathcal{V}| \sin \phi_{\mathcal{V}} \end{aligned} \quad (2.4.13)$$

Hence, the output from the correlator can be re-written as

$$r(\mathbf{b}_\lambda, \mathbf{s}_o) = \Delta\nu |\mathcal{V}| [\cos(2\pi\mathbf{b}_\lambda \cdot \mathbf{s}_o) \cos\phi_\mathcal{V} + \sin(2\pi\mathbf{b}_\lambda \cdot \mathbf{s}_o) \sin\phi_\mathcal{V}]$$

Using the identity again,

$$r(\mathbf{b}_\lambda, \mathbf{s}_o) = \Delta\nu |\mathcal{V}| \cos(2\pi\mathbf{b}_\lambda \cdot \mathbf{s}_o - \phi_\mathcal{V}) \quad (2.4.14)$$

This shows that the output from the correlator can be expressed in terms of the fringe pattern corresponding to the the field centre. Also, the visibility has dimensions of $\text{W m}^{-2} \text{Hz}^{-1}$ or Jy which is the same as I .

Consider a right-handed system of coordinates in which case the baseline vector has components (u, v, w) and (l, m) are the direction cosines with respect to the u and v axes. Also, the unit vector $\mathbf{s}_o$ is aligned with the w axis, such that,

$$\mathbf{s}_o = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

In this frame,

$$\mathbf{b}_\lambda \cdot \mathbf{s}_o = w$$

$$\mathbf{b}_\lambda \cdot \mathbf{s} = ul + vm + w\sqrt{1 - l^2 - m^2}$$

$$\mathbf{b}_\lambda \cdot \boldsymbol{\sigma} = \mathbf{b}_\lambda \cdot (\mathbf{s} - \mathbf{s}_o) = ul + vm + w(\sqrt{1 - l^2 - m^2} - 1)$$

Following the definition of the visibility in Equation (2.4.12),

$$\mathcal{V} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} A(l, m) I(l, m) \exp[-2\pi j \{ul + vm + w(n - 1)\}] \frac{dl dm}{n} \quad (2.4.15)$$

In particular,

$$d\Omega = \frac{dl dm}{n} = \frac{dl dm}{\sqrt{1 - l^2 - m^2}} \quad (2.4.16)$$

This element of solid angle, $d\Omega$ in spherical coordinates is given by

$$d\Omega = \sin \theta \, d\theta \, d\phi$$

By computing the Jacobian determinant, we can move from (θ, ϕ) frame to the (l, m) (see Figure 2.9) as follows:

$$d\theta \, d\phi \rightarrow \begin{vmatrix} \partial\theta/\partial l & \partial\theta/\partial m \\ \partial\phi/\partial l & \partial\phi/\partial m \end{vmatrix} dl \, dm \quad (2.4.17)$$

If the antennas are located in a plane which is perpendicular to the w -axis, then, $w = 0$, hence Equation (2.4.15) reduces to a 2D-Fourier Transform. However, this is not always the case and the assumption we make is that if $l, m \ll 1$ such that the term

$$w \left(\sqrt{1 - l^2 - m^2} - 1 \right) \sim -\frac{1}{2} (l^2 + m^2) w \quad (2.4.18)$$

is negligible, then

$$\mathcal{V}(u, v, w \approx 0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{A(l, m) I(l, m)}{\sqrt{1 - l^2 - m^2}} \exp[-2\pi j(ul + vm)] \, dl \, dm \quad (2.4.19)$$

Hence, the visibility can be considered as a 2-dimensional Fourier Transform under the condition that $w \approx 0$. The inverse Fourier Transform is

$$\frac{A(l, m) I(l, m)}{\sqrt{1 - l^2 - m^2}} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathcal{V}(u, v) \exp[2\pi j(ul + vm)] \, du \, dv \quad (2.4.20)$$

Equation (2.4.20) is also commonly referred to as the Van-Cittert Zernike theorem which was originally derived in the field of optics (Zernike 1938). The Fourier relationship existing between the intensity and the visibility is of great importance in radio interferometry. It is worth specifying that the interferometer is not sensitive to the sky directly but instead to the Fourier transform of the sky. $\mathcal{V}$ is a 2D function in the Fourier space. At a particular time t , baseline $\mathbf{b}_\lambda$ and direction $\mathbf{s}_0$, we will recover one complex sample of the visibility which is a unique (u, v) point in the Fourier domain.

2.4.3 Earth-Rotation Aperture Synthesis

As the Earth rotates, the celestial body rotates gradually on the celestial sphere. The combination of the baseline $\mathbf{b}_\lambda$ and direction $\mathbf{s}_0$ for a specific pair of antennas leads to the projected

baseline $|\mathbf{b}_\lambda \sin \theta|$ (refer to §2.4.1). As a result of the rotation of Earth, the projected baseline and τ_g will vary slowly with time, thus generating a path in the uv -plane.

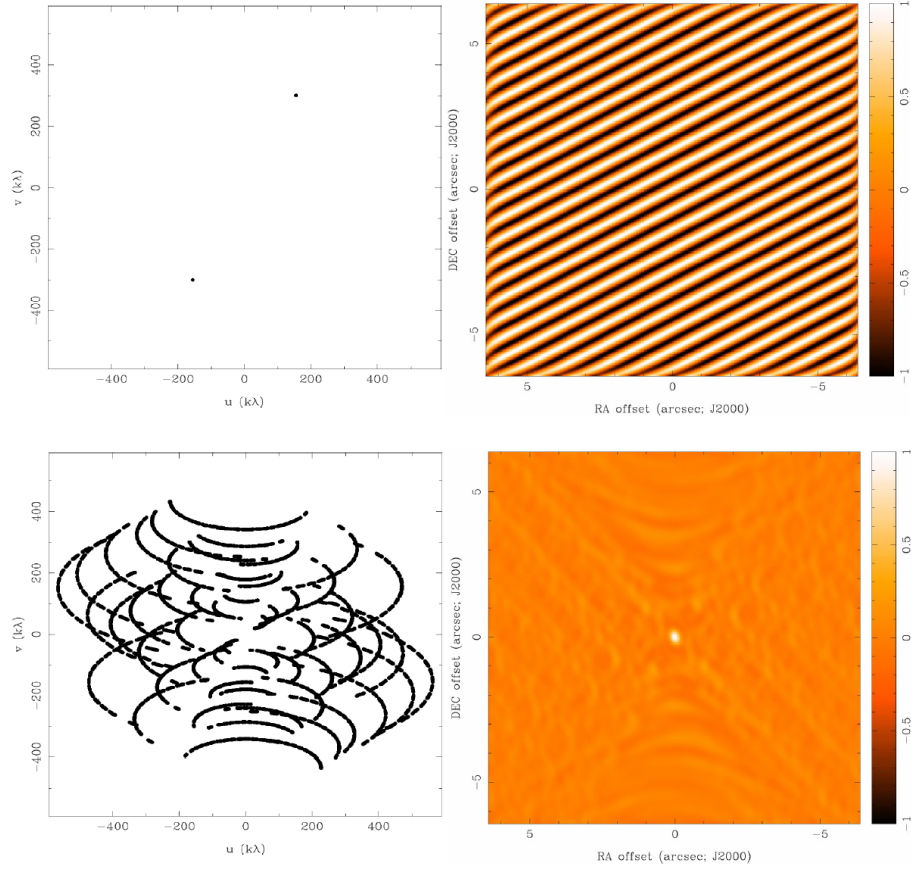


Figure 2.10 – Example of uv plane sampling - The above shows the dirty beam shape on the right panel corresponding to N antennas. In the upper panel, $N = 2$ while in the lower panel, $N = 8$, each recording 480 samples over the observation time. The uv -plane is filled using Earth Rotation but is limited to the number of antennas we have. The number of samples from N antennas is typically equal to $\frac{N(N-1)}{2}$. Figure adapted from (Wilner 2010).

Therefore, we could use the Earth’s rotation itself to improve the baseline coverage for an interferometer. If the baselines are aligned in an East-West configuration, the projected baselines will remain in a plane normal to the North-South rotation axis of the Earth and the projected baselines will follow elliptical paths (see Figure 2.11). In such a case, the baselines are then confined to two dimensions only, hence the brightness distribution is simply equal to the 2D Fourier Transform of the visibility measurements. Moreover, the number of visibility measurements is limited by the number of antennas we have. Figure (2.10) shows that the larger the number of baselines, the better will be the uv -sampling and hence the shape of the dirty beam.

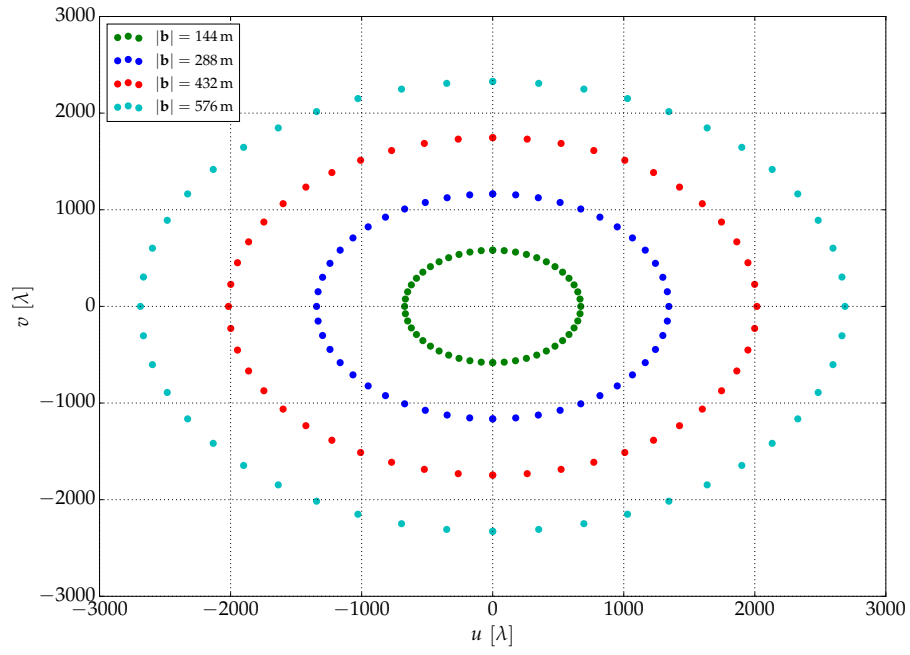


Figure 2.11 – uv -coverage for an East-West interferometer - We simulate the uv -coverage for an East-West interferometer consisting of 4 baselines of length 144 m, 288 m, 432 m and 576 m respectively. The declination of the field centre is $+60^\circ 04' 00''$ and during the 12 h of observation (hour angle from -6 h to +6 h), the projected baselines trace elliptical paths as shown above. The v axis is typically smaller than the u axis by a factor of $\sec \delta$, where δ is the declination (Condon & Ransom 2010).

RIME AND BIRO

"Mathematics is the queen of the sciences and arithmetic [number theory] is the queen of mathematics. She often condescends to render service to astronomy and other natural sciences, but in all relations, she is entitled to first rank" - Carl Friedrich Gauss

3.1 Introduction

The Radio Interferometric Measurement Equations (RIME), is a mathematical formulation which has been developed by [Hamaker et al. \(1996\)](#). Well before the development of this method, other approximate mathematical models were incorporated in various softwares to do calibration. However, with the forthcoming new radio telescopes and even the upgrades of older telescopes, it is of utmost importance to not only deal with Direction-Independent Effects (DIEs) as was the case in the past, but also to deal with Direction-Dependent Effects (DDEs). Fortunately, the RIME provides an elegant way of formulating the problem by dealing with both observation effects and direction-dependent effects ([Smirnov 2011](#)).

While it is common to produce a single final image via deconvolution in radio interferometry, [Lochner et al. \(2015\)](#) showed that it is in fact possible to skip the entire process of image-making and infer both scientific and systematic parameters via a novel method which they refer to as Bayesian Inference for Radio Observations (BIRO). This method is based on Bayesian Statistics (as explained in [Chapter 1](#)).

Novel techniques such as BIRO are able to elegantly address several shortcomings of current radio astronomical approaches to data reduction. In the following sections, we shall elaborate more on the two topics, namely RIME and BIRO and we shall also touch upon briefly on deconvolution methods in radio astronomy.

3.2 Deconvolution in Radio Astronomy

Imaging in the radio astronomy community refers to the process of transforming the measured visibilities into an image of the sky. However, the complication is that the measured visibility is a result of convolution of the true sky and the array response as well as some instrumental noise, that is,

$$I_D = I_{\text{true}} * \text{PSF} + \epsilon \quad (3.2.1)$$

The goal is to create a “clean” image from a “dirty” image. Parameters which are responsible for corrupting the signal are typically known and the Point Spread Function (PSF) is assumed independent of the sky. Therefore, a deconvolution algorithm should be able to separate the effects of the array and the sky. At the end of a deconvolution process, we generally have two outputs, namely the sky model and the residual. The sky model itself typically does not correspond to the true sky model, principally because of incomplete sampling of the visibilities and the finite resolution of our instrument.

The concept of deconvolution is simple to state. Assume we have two functions f and g convolved into a new function h , that is,

$$h = f * g \quad (3.2.2)$$

If we have h and g , we could use the convolution theorem to find f :

$$\mathcal{F}(f) = \frac{\mathcal{F}(h)}{\mathcal{F}(g)} \quad (3.2.3)$$

However, this is applicable only in the noiseless case. Visibilities are generally corrupted by noise, ϵ . If ϵ is small enough such that we can approximate it to zero, the above can be used and the process is referred to as *inverse filtering*.

3.2.1 CLEAN

The first variant of CLEAN was proposed by Högbom (1974). The concept behind CLEAN is that we assume we know the PSF and assume that the sources are point sources. This simplifies the deconvolution process as the PSF can then be interpreted as a point source response. Therefore, the PSF multiplied by some flux values can be iteratively subtracted until we end up with a sky model consisting of δ -functions and a residual image with noise, plus some remaining

sources.

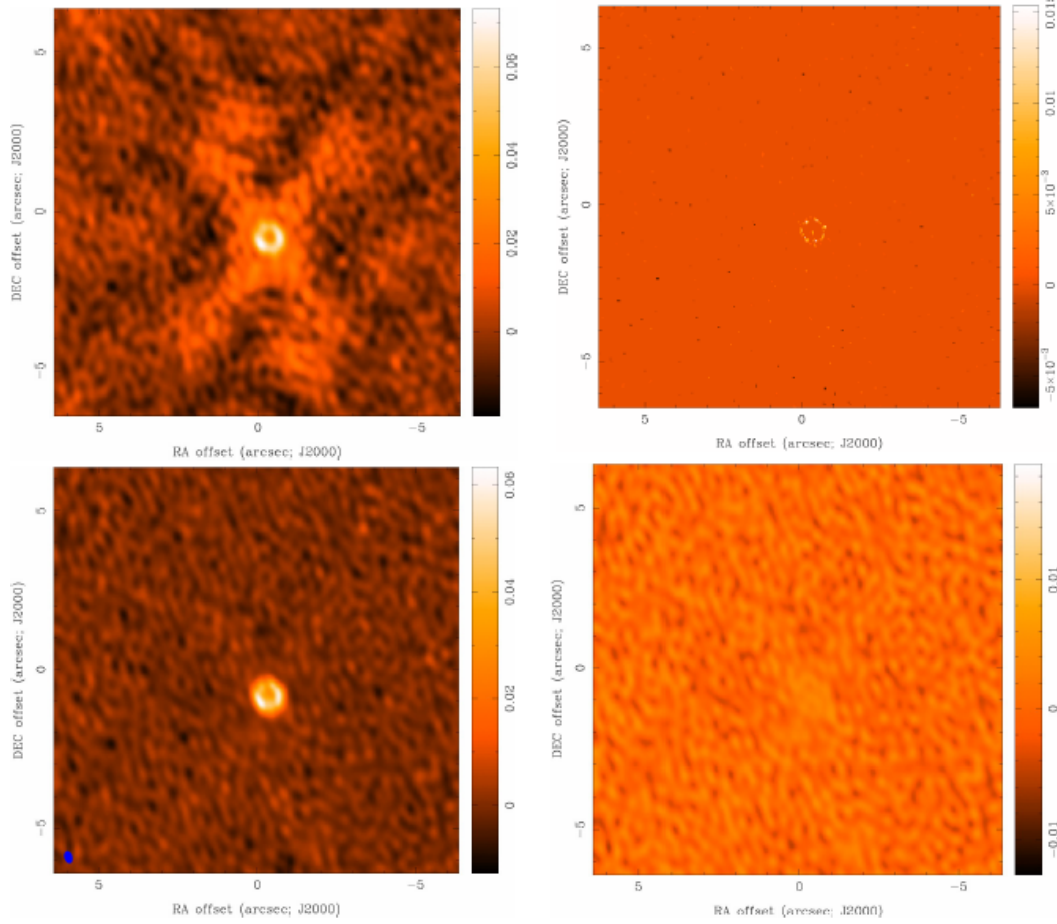


Figure 3.1 – Example of the CLEAN algorithm applied to a dirty image - The upper left and upper right figures show the dirty and clean model respectively. After applying the CLEAN algorithm, the recovered image (sky model) and the residual image are shown in the lower left and right panel respectively. Figure courtesy of [Wilner \(2010\)](#)

The inputs to the CLEAN algorithm are the dirty image, I_D , the PSF, the loop gain γ and the stopping condition which is governed by either f_{thresh} or N . The loop gain γ is between 0 and 1. It controls the amount of flux subtracted as each iteration and good results are obtained for $0.1 \lesssim \gamma \lesssim 0.3$ ([Wilner 2010](#)). If this parameter is too small, the algorithm will require a larger number of iterations. If it is high, noise could even be subtracted which is not of course not desired. In addition, N is the number of iterations and has the advantage of fixing the computation time. However, f_{thresh} is preferred if we specifically want a flux threshold in the residual image.

The CLEAN algorithm is not without issues. However, since the PSF and residual image are both images with a set of pixels, the PSF is offset when subtracting a fraction of flux from the residual image. Moreover, the sky model is fixed to the resolution of pixels in the residual image. This gives rise to artefacts since sources are not exactly found at the centre of a pixel. In addition to this, the PSF used is the one pointing towards the phase centre. In fact, sources

away from the field centre are actually convolved with a different PSF response. Further, deconvolution of extended sources is still a major active research topic (Rau & Cornwell 2011). To improve the results, there are further developments and other variants of CLEAN, for example, see Clark (1980) and Schwab (1984).

3.3 RIME

In this section, we will discuss the RIME. We shall first start by defining the Stokes parameters and then we shall go through three cases using RIME. These are for a single point source, multiple sources and eventually the all-sky RIME.

3.3.1 Stokes Parameters

Consider a signal propagating through space and time from the source to the receiver. We can then define a right-handed and orthonormal set of coordinates, for example, the usual xyz coordinate system. The z -axis, as usual, is defined to be in the direction of propagation, that is, towards the phase centre. The electromagnetic field can be described by a complex vector in a plane perpendicular to the direction of propagation as follows:

$$\epsilon = \begin{pmatrix} \epsilon_x \\ \epsilon_y \end{pmatrix} \quad (3.3.1)$$

The signal goes through a series of effects as it propagates from the source to the receiver, that is, we will in fact measure a corrupted signal which we now call ϵ' . We want a relationship between this corrupted signal ϵ' and the original signal ϵ . Our assumption is that all transformations are linear with respect to the ϵ , such that

$$\epsilon' = \mathbf{J}\epsilon \quad (3.3.2)$$

where $\mathbf{J}$ is a 2×2 Jones matrix (Jones 1941). The very fact that we can establish this relation via the Jones matrix brings about the simplicity and elegance of the further concepts of RIME. Before indulging in the explanation of the RIME, it is first worth discussing the Stokes parameters.

The Stokes parameters are related to the amplitudes of the components of the signal. In particular, there are 4 Stokes parameters (I, Q, U, V) which are defined as follows:

$$\begin{aligned}
I &= \langle \epsilon_x \epsilon_x^* \rangle + \langle \epsilon_y \epsilon_y^* \rangle \\
Q &= \langle \epsilon_x \epsilon_x^* \rangle - \langle \epsilon_y \epsilon_y^* \rangle \\
U &= \langle \epsilon_x \epsilon_y^* \rangle + \langle \epsilon_y \epsilon_x^* \rangle = 2\Re \langle \epsilon_x \epsilon_y^* \rangle \\
V &= -j \left(\langle \epsilon_x \epsilon_y^* \rangle - \langle \epsilon_y \epsilon_x^* \rangle \right) = 2\Im \langle \epsilon_x \epsilon_y^* \rangle
\end{aligned}$$

The angle bracket operator denotes an averaging over time. The Stokes I parameter measures the total intensity of the wave, Q and U represent the components of the linearly polarised signal and V finally is a measure of the circularly polarised component. These quantities all have the same dimension as the flux density or intensity. I is always positive whereas Q , U and V can have negative values (Thompson et al. 2008).

3.3.2 The Visibility Matrix

When the signal is received by the antennas, it is in fact registered as voltages. Assuming we have two feeds (a and b) and also that the voltages, $\mathbf{v}$, are linear with respect to ϵ , then

$$\mathbf{v} = \begin{pmatrix} v_a \\ v_b \end{pmatrix} = \mathbf{J}' \epsilon \quad (3.3.3)$$

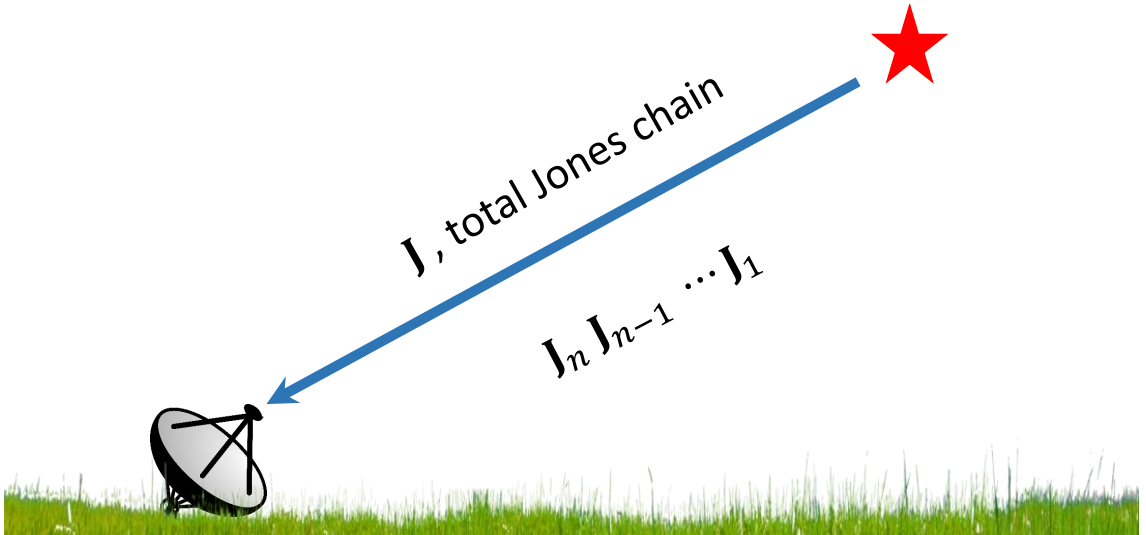


Figure 3.2 – The total Jones matrix - The total Jones matrix is given by the product of the individual Jones matrices in a specific order. J_1 corresponds to the effect when the signal is first emitted from the source while J_n is the final effect at the other extreme end. This could be due to the instrument.

Conceptually, the above relation gives us the the relation between the voltage vector and the original signal with all the propagation effects encoded in a single Jones matrix $\mathbf{J}$. This Jones

matrix will thereafter be referred to as the total Jones matrix (see Figure 3.2). In particular, there are multiple Jones matrices, from the moment the signal is emitted until it is finally converted to a voltage vector:

$$\mathbf{J} = \mathbf{J}_n \mathbf{J}_{n-1} \cdots \mathbf{J}_1 \quad (3.3.4)$$

with $\mathbf{J}_1$ being the first propagation and $\mathbf{J}_n$ is the last effect considered (it could even be electronic effects in the antenna).

Consider now two telescopes p and q which together produce two independent voltage vectors denoted by v_p and v_q . These voltages are fed into the correlator which result in four pairwise correlations $\langle v_{pa} v_{qa}^* \rangle$, $\langle v_{pa} v_{qb}^* \rangle$, $\langle v_{pb} v_{qa}^* \rangle$ and $\langle v_{pb} v_{qb}^* \rangle$. These correlations can be arranged into the visibility matrix, $\mathbf{V}_{pq}$.

$$\mathbf{V}_{pq} = 2 \begin{pmatrix} \langle v_{pa} v_{qa}^* \rangle & \langle v_{pa} v_{qb}^* \rangle \\ \langle v_{pb} v_{qa}^* \rangle & \langle v_{pb} v_{qb}^* \rangle \end{pmatrix} \quad (3.3.5)$$

This matrix can in fact be written as the product of two complex vectors as follows:

$$\mathbf{V}_{pq} = 2 \left\langle \begin{pmatrix} v_{pa} \\ v_{qb} \end{pmatrix} \begin{pmatrix} v_{pa}^* & v_{qb}^* \end{pmatrix} \right\rangle = 2 \langle \mathbf{v}_p \mathbf{v}_q^\dagger \rangle \quad (3.3.6)$$

$\dagger$ denotes conjugate transpose, also known as the Hermitian transpose. The factor of 2 introduced here is a controversial topic in the radio astronomy community. In fact, it is just a convention which has been adopted so as to ensure compatibility with existing software and is discussed in detail by Smirnov (2011). Therefore, from Equation (3.3.3), the visibility matrix can be re-written as

$$\mathbf{V}_{pq} = 2 \langle (\mathbf{J}_p \boldsymbol{\epsilon}) (\mathbf{J}_q \boldsymbol{\epsilon})^\dagger \rangle = \langle \mathbf{J}_p (2\boldsymbol{\epsilon} \boldsymbol{\epsilon}^\dagger) \mathbf{J}_q^\dagger \rangle \quad (3.3.7)$$

Under the assumption that $\mathbf{J}_p$ and $\mathbf{J}_q$ are constant over the averaging time,

$$\mathbf{V}_{pq} = \mathbf{J}_p \langle 2\boldsymbol{\epsilon} \boldsymbol{\epsilon}^\dagger \rangle \mathbf{J}_q^\dagger \quad (3.3.8)$$

However, the term $\langle 2\boldsymbol{\epsilon} \boldsymbol{\epsilon}^\dagger \rangle$ is simply

$$2 \begin{pmatrix} \langle e_x e_x^* \rangle & \langle e_x e_y^* \rangle \\ \langle e_y e_x^* \rangle & \langle e_y e_y^* \rangle \end{pmatrix} = \begin{pmatrix} I + Q & U + jV \\ U - jV & I - Q \end{pmatrix} = \mathbf{B} \quad (3.3.9)$$

where $\mathbf{B}$ is the brightness matrix defined in terms of the Stokes parameters. Hence the visibility matrix, in terms of the total Jones matrix for two antennas p and q and the brightness matrix is given by:

$$\mathbf{V}_{pq} = \mathbf{J}_p \mathbf{B} \mathbf{J}_q^\dagger \quad (3.3.10)$$

The two Jones matrices ($\mathbf{J}_p$ and $\mathbf{J}_q$) can further be broken down into their individual Jones matrices as follows:

$$\mathbf{V}_{pq} = \mathbf{J}_{pn} \left(\dots \left(\mathbf{J}_{p2} \left(\mathbf{J}_{p1} \mathbf{B} \mathbf{J}_{q1}^\dagger \right) \mathbf{J}_{p2}^\dagger \right) \dots \right) \mathbf{J}_{pm}^\dagger \quad (3.3.11)$$

This results in what we call the “onion” form of the RIME. Also note the two different indices n and m . This means that the onion form of the RIME need not be symmetric in the sense that the total propagation effect for the signal to reach antenna p can be completely different to that reaching antenna q . The power of the RIME is that this formalism can easily be extended to VLBI (Very Long Baseline Interferometry) in which case signals have completely different propagation paths.

3.3.3 Coherency Matrix

The RIME discussed previously provides a general approach to model the signal emitted from a source. The fundamental effect, which is the centrepiece of radio interferometry, is the phase delay. We have previously provided an approach to this in §(2.4.2).

As the signal propagates to the two antennas p and q , there exists a geometric path length difference in the two paths. The correlator therefore introduces an additional delay term to take care of this effect, eventually steering the interferometer in a specific direction, referred to as the phase centre. In our case, we will consider the phase difference with respect to a zero point, $\mathbf{u} = (0, 0, 0)$ and the usual convention is to treat the z -axis as pointing towards the phase centre. If the two antennas are located at $\mathbf{u}_p = (u_p, v_p, w_p)$ and $\mathbf{u}_q = (u_q, v_q, w_q)$. Then the phase differences for the signal coming from a direction σ at each antenna with respect to the zero point are given by

$$\begin{aligned} k_p &= \frac{2\pi}{\lambda} (u_p l + v_p m + w_p (n - 1)) \\ k_q &= \frac{2\pi}{\lambda} (u_q l + v_q m + w_q (n - 1)) \end{aligned} \quad (3.3.12)$$

where l , m and n are the direction cosines of σ . The K-Jones matrix which represents the phase delay effect can be expressed by the scalar matrix

$$\mathbf{K} = \begin{pmatrix} e^{j\phi} & 0 \\ 0 & e^{j\phi} \end{pmatrix} = e^{j\phi} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and this type of matrix commutes with everything. Hence,

$$\mathbf{K}_p = e^{-jk_p} \mathbf{I} = \exp [-j2\pi (u_p l + v_p m + w_p (n - 1))] \mathbf{I} \quad (3.3.13)$$

where $\mathbf{I}$ is the identity matrix of size two and note that $\mathbf{u}_p$ and $\mathbf{u}_q$ are expressed in units of the wavelength λ . As a result, the RIME for one point source is given by

$$\mathbf{V}_{pq} = \mathbf{K}_p \mathbf{B} \mathbf{K}_q^\dagger \quad (3.3.14)$$

under the assumption that the phase delay is the only effect considered here. We can re-write the above equation as

$$\mathbf{X}_{pq} = \mathbf{B} e^{-2\pi j (u_{pq} l + v_{pq} m + w_{pq} (n-1))} \quad (3.3.15)$$

where $\mathbf{u}_{pq} = \mathbf{u}_p - \mathbf{u}_q$. $\mathbf{X}_{pq}$ is referred to as the *coherency matrix* and is a function of the baseline $\mathbf{u}_{pq}$. In the next sections, we shall briefly discuss the RIME for different cases.

3.3.4 RIME for Single Point Source

In practice, an interferometer is not a perfect measuring instrument. It will always have some corrupting effects on the signal. Let this effect be represented by the Jones matrix $\mathbf{G}$. Then, the RIME for a single point source (taking phase delay effect and this corrupting effect) is given by the “onion” equation:

$$\mathbf{V}_{pq} = \mathbf{G}_p \mathbf{K}_p \mathbf{B} \mathbf{K}_q^\dagger \mathbf{G}_q^\dagger \quad (3.3.16)$$

The above equation can also be expressed in terms of the source coherency, $\mathbf{X}_{pq}$.

$$\mathbf{V}_{pq} = \mathbf{G}_p \mathbf{X}_{pq} \mathbf{G}_q^\dagger \quad (3.3.17)$$

3.3.5 RIME for Multiple Point Sources

Now, consider the case where we have N point sources instead of only one. The signal emitted from each source will have a different signal path towards each antenna (p or q). Therefore, the propagation effect per source can be described by its own Jones matrix $\mathbf{J}_{sp}$ where the index s represents the s^{th} source. The final visibility matrix is then given by the sum of the individual visibilities due to each source, that is,

$$\mathbf{V}_{pq} = \sum_{s=1}^N \mathbf{J}_{sp} \mathbf{B}_s \mathbf{J}_{sq} \quad (3.3.18)$$

Suppose we have a Jones matrix $\mathbf{G}_p$ describing the direction-independent effects, with a Jones matrix $\mathbf{E}_{sp}$ describing the direction-dependent effects and we have the phase term $\mathbf{K}_{sp}$. Then the visibility matrix can be re-written as

$$\mathbf{V}_{pq} = \sum_{s=1}^N \mathbf{G}_p \mathbf{E}_{sp} \mathbf{K}_{sp} \mathbf{B}_s \mathbf{K}_{sq}^\dagger \mathbf{E}_{sq}^\dagger \mathbf{G}_q^\dagger$$

Note that the effect described by $\mathbf{G}_p$ is the same for all sources, for example, it can be the electronic gain of the antennas. Therefore, we can write:

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\sum_{s=1}^N \mathbf{E}_{sp} \mathbf{K}_{sp} \mathbf{B}_s \mathbf{K}_{sq}^\dagger \mathbf{E}_{sq}^\dagger \right) \mathbf{G}_q^\dagger$$

or in terms of the coherency matrix:

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\sum_{s=1}^N \mathbf{E}_{sp} \mathbf{X}_{pq} \mathbf{E}_{sq}^\dagger \right) \mathbf{G}_q^\dagger \quad (3.3.19)$$

3.3.6 RIME for Continuous Brightness Distribution

In general, the sky is in fact composed of a continuous brightness distribution, which we denote as $\mathbf{B}(\sigma)$. The full-sky RIME can be written as

$$\mathbf{V}_{pq} = \int_{4\pi} \mathbf{J}_p(\sigma) \mathbf{B}(\sigma) \mathbf{J}_q^\dagger(\sigma) d\Omega \quad (3.3.20)$$

The above equation can be re-written in terms of the (l, m, n) coordinates as

$$\mathbf{V}_{pq} = \int_l \int_m \mathbf{J}_p(l, m) \mathbf{B}(l, m) \mathbf{J}_q^\dagger(l, m) \frac{dl dm}{n} \quad (3.3.21)$$

Considering direction-dependent effects, direction-independent effects and the phase term,

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\int_l \int_m \mathbf{E}_p \mathbf{B} \mathbf{E}_q^\dagger e^{-2\pi j(u_{pq}l + v_{pq}m + w_{pq}(n-1))} \frac{dl dm}{n} \right) \mathbf{G}_q^\dagger \quad (3.3.22)$$

If we now define

$$\mathbf{W}_p = \frac{1}{\sqrt{n}} e^{-2\pi j w_p(n-1)} \mathbf{I}$$

then

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\int_l \int_m \mathbf{E}_p \mathbf{W}_p \mathbf{B} \mathbf{W}_q^\dagger \mathbf{E}_q^\dagger e^{-2\pi j(u_{pq}l + v_{pq}m)} dl dm \right) \mathbf{G}_q^\dagger \quad (3.3.23)$$

The term $\mathbf{E}_p \mathbf{W}_p$ can be thought of as a direction-dependent effect and can therefore be incorporated into one single notation

$$\tilde{\mathbf{E}}_p = \mathbf{E}_p \mathbf{W}_p$$

Hence,

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\int_l \int_m \tilde{\mathbf{E}}_p \mathbf{B} \tilde{\mathbf{E}}_q^\dagger e^{-2\pi j(u_{pq}l + v_{pq}m)} dl dm \right) \mathbf{G}_q^\dagger$$

Each baseline will in fact “see” a different brightness. We can write the final visibility matrix as a 2D Fourier Transform of this apparent brightness as

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\int_l \int_m \mathbf{B}_{pq} e^{-2\pi j(u_{pq}l + v_{pq}m)} dl dm \right) \mathbf{G}_q^\dagger \quad (3.3.24)$$

where $\mathbf{B}_{pq} = \tilde{\mathbf{E}}_p \mathbf{B} \tilde{\mathbf{E}}_q^\dagger = \mathbf{E}_p \mathbf{W}_p \mathbf{B} \mathbf{W}_q^\dagger \mathbf{E}_q^\dagger$. The possibility for each baseline to “see” the same brightness is to have a common direction-dependent effect such that $\mathbf{E}_p(l, m) = \mathbf{E}_q(l, m) = \mathbf{E}(l, m)$ across all antennas. Under this assumption, the RIME is

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\int_l \int_m \mathbf{E}(l, m) \mathbf{B}(l, m) \mathbf{E}^\dagger(l, m) e^{-2\pi j(u_{pq}l + v_{pq}m)} dl dm \right) \mathbf{G}_q^\dagger \quad (3.3.25)$$

or in short-hand form as

$$\mathbf{V}_{pq} = \mathbf{G}_p \mathbf{X}_{pq} \mathbf{G}_q^\dagger \quad (3.3.26)$$

$\mathbf{X}_{pq}$ in this case is referred to as the *sky coherency*.

3.4 BIRO

BIRO is a new Bayesian technique which has been recently introduced in the radio astronomy community (Lochner et al. 2015). In common practice, a single image is produced, from which most of the science is done. In particular, this image is usually produced by iteratively applying a deconvolution algorithm such as CLEAN (Högbom 1974).

BIRO takes full advantage of both parameter estimation and model selection from the concepts of Bayesian statistics to address two problems, namely:

- (1) to simultaneously estimate science and nuisance parameters in the Jones matrices, and the correlations between them and
- (2) to apply model selection to determine the correct sky model, such as distinguishing between point or extended sources.

It is also worth mentioning that BIRO operates in the visibility domain directly. Before going into the details of BIRO, it is important to discuss briefly two software, namely MeqTrees and Montblanc.

3.4.1 MeqTrees and Montblanc

MeqTrees (Noordam & Smirnov 2010) is a software package that implements the RIME. Over and above this, it has various built-in features which can be used for calibration. At the back-end, this software is written in C++ but the modelling can be done in a relatively simple way using Python. MeqTrees is designed using expression trees. A tree consists of nodes, which are connected in a parent-child system. In particular, there are no cycles.

As an example, in Figure (3.3), the aim is to calculate $\sin a + \cos b$. a and b are the leaf nodes, the $\sin$ and $\cos$ are the parent nodes and finally the $+$ is the root node. The expression is calculated by first starting at the leaf nodes through the parent nodes until the final expression is calculated at the root. In this simple example, it can be deduced that a leaf is a node which has no children, and a node which has no parent is the root node. In particular, various trees can be connected to form a forest.

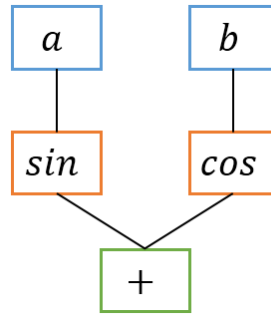


Figure 3.3 – Example of expression tree. Figure adapted from [Noordam & Smirnov \(2010\)](#)

MeqTrees is made up of several major components. The *meqserver*, mostly written in C++ is the bridge between datasets and expression trees which are written using the Python-based Tree Definition Language (TDL). Most importantly, the *meqserver* takes full control in the construction and evaluation of trees. Furthermore, the *meqbrowsers* allows for the control of the *meqserver*. It not only parses TDL scripts but allows for visualisation of the results.

On the other hand, Montblanc ([Perkins et al. 2015](#)) implements the RIME using GPUs. It was originally developed in support of BIRO. As will be discussed in the next section, computing the likelihood is computationally intensive ([Lochner et al. 2015](#)). However, Montblanc is roughly 250 times faster than MeqTrees on a dual hexacore Intel E5 2620v2 CPU. It uses NVIDIA’s CUDA to compute the likelihood which can then be used in the Bayesian Inference scheme. Montblanc is still under development.

3.4.2 Parameter Estimation

BIRO offers an alternative and novel approach to not only find the flux of a source but also to deal with systematics. There are various drawbacks associated with deconvolution methods in radio astronomy (see §3.2.1). Instead of inspecting an image visually to do science, the best approach is to “learn” from the data directly, which is what BIRO framework easily provides us. In this section, we summarise briefly the first part of the BIRO project.

While performing Bayesian inference for the scientific parameters (for example, the fluxes of the radio sources) is relatively straightforward, [Lochner et al. \(2015\)](#) took a step further in modelling the pointing errors associated with the beam of the interferometer as well. This cannot be done within CLEAN for example. In particular, 17 sources were simulated with known positions on the sky. The pointing error can be incorporated in the RIME as follows,

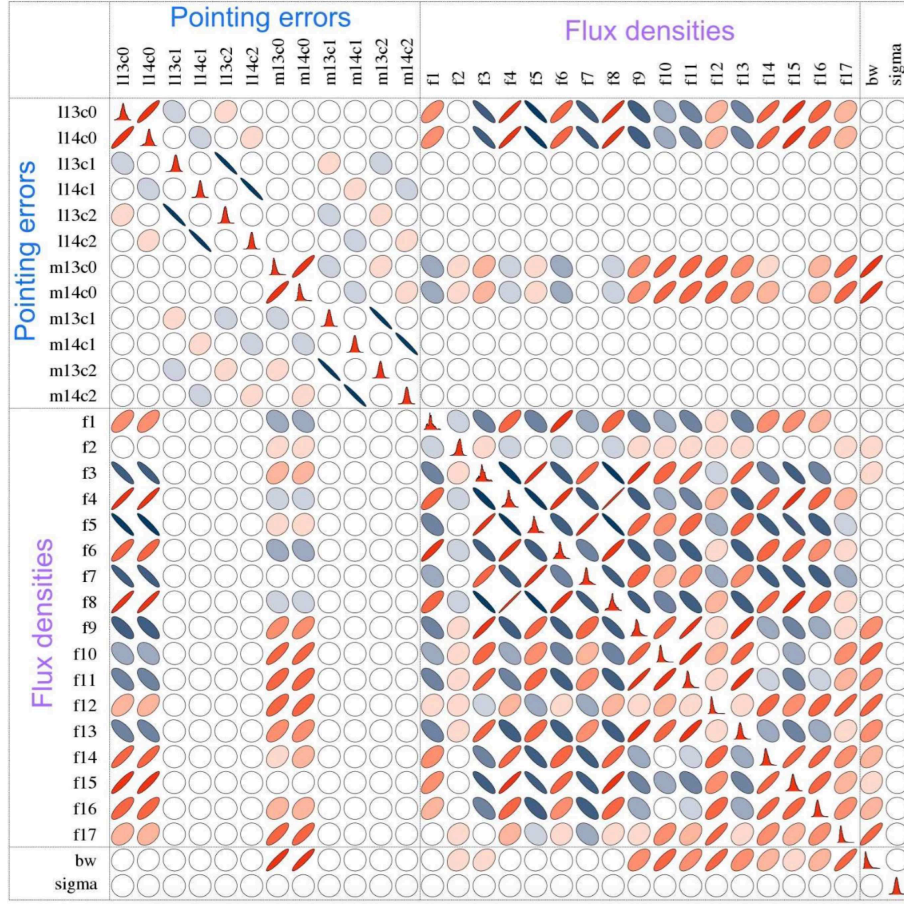


Figure 3.4 – Correlations among a subset of scientific and instrumental parameters - The leading diagonal shows the marginalised posterior distribution for each parameter and the coloured ellipses show the correlation among the respective parameters. Note that in this case, Bayesian inference is performed on 17 sources along with pointing errors which are denoted by $l_j c_k$ and $m_j c_k$ where j and k are the j^{th} source and the k^{th} coefficient of the time-varying pointing error for l and m direction cosines. bw and σ refer to the beam width and noise level respectively. Figure adapted from [Lochner et al. \(2015\)](#).

$$\mathbf{E}_{\text{PE}} = \mathbf{E}_{\text{beam}}(l + \delta l_p, m + \delta m_p) \quad (3.4.1)$$

where δl_p and δm_p are the pointing errors in the l and m direction cosines respectively. Moreover, these pointing errors are modelled using a second-order time-varying polynomial as follows

$$\delta l_p = c_0 + c_1 t + c_2 t^2 \quad (3.4.2)$$

and similarly for δm_p as well. The parameters due to pointing errors are the c_k (84 parameters for the pointing errors due only to 14 WSRT antennas). Furthermore, the beam width, which is a crucial quantity for modelling the flux distribution of the sky is assumed to be unknown and is included as an instrumental parameter in the scheme. The noise level is also considered to be unknown and is included in the inference algorithm. Overall, there are 103 unknown parameters for the inference problem (17 scientific parameters - the fluxes and 84 pointing

errors, the beam width and the noise level). Part of the results, including the correlation among the parameters are shown in Figure (3.4)

To compare the performance of BIRO, Lochner et al. (2015) simulated a data set with pointing errors and then performed an independent analysis using CLEAN followed by source extraction (CLEAN+SE). While BIRO was able to recover the flux densities of all the sources within 3σ confidence interval, CLEAN+SE recovered only 5 out of the 17 sources with estimated fluxes up to $\sim 40\sigma$ away from the true fluxes. This clearly demonstrates the power of Bayesian Parameter Estimation and MCMC in performing both scientific and instrumental parameter inference.

The fact that we should know the number of sources prior to parameter inference, remains one of the criticisms of BIRO. Of course, one could go to the second level of inference, that is, Bayesian Model Comparison, where different models are compared and the model which best describes the observed visibilities is chosen. The forthcoming telescopes/SKA pathfinders themselves are going to be more sensitive than existing radio telescopes. Hence, the sky image will consist of thousands or hundreds of thousands of sources which might as well include new sources that have not been discovered before.

3.4.3 Model Selection

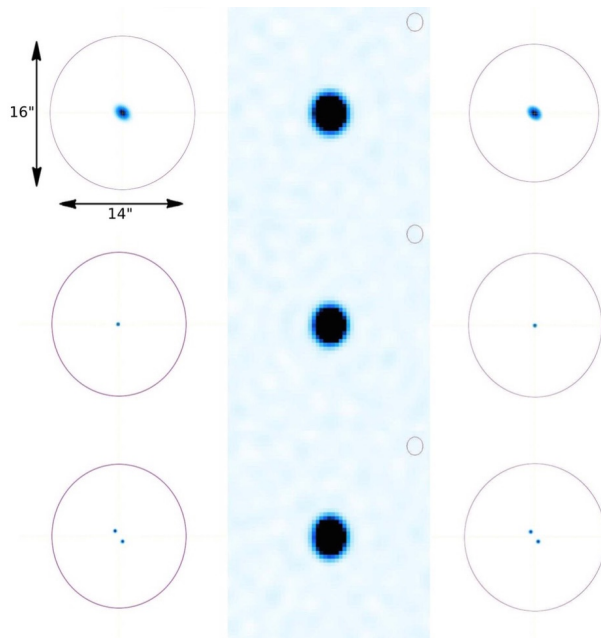


Figure 3.5 – Resolving power of BIRO - The left panel shows the simulated data, in which case, an extended Gaussian, a single point source and a two-point source are considered. These sources are much smaller than the synthesised beam, shown by the purple ellipse. The middle column and the third column compare the performance of CLEAN and BIRO respectively. While BIRO is able to recover all the sources, for a SNR of 1000, CLEAN fails to resolve these sources. Figure courtesy of Lochner et al. (2015).

Moving on the second level of inference, Bayesian Model Comparison, BIRO is able to recover the best model for each scenario when compared with other models.

In particular, Multinest is used for estimating the Bayesian Evidence and it also returns the posterior distributions of the relevant parameters as a by-product of the inference problem. The single point source has three parameters (flux and positions), similarly for the two-point source while the Gaussian extended source has six parameters - flux, positions as well as 3 shape parameters.

These quantities are used for parameterisation in MeqTrees. BIRO is able to choose the correct hypothesis in all cases with odds $10^{593} : 1$, $10^{993} : 1$ and $62 : 1$ for the two-point source, extended source and the single point source respectively. On comparing the performance of BIRO with CLEAN, the latter is unable to resolve the sources within the sub-synthesised beam scales which is easily achieved using BIRO (see Figure 3.5). We will study this super-resolution in detail in a toy model to gain further intuition in Chapter (6).

A NEW APPROACH TO MODEL SELECTION USING SUPERMODELS

"Do not worry about your difficulties in Mathematics. I can assure you mine are still greater." - Albert Einstein

4.1 Parameter Estimation

Very often in data analysis, one wants to fit a model to some data. That particular model depends upon some adjustable parameters and is typically a class of functions. Generally, given the parameters, one uses a merit function that measures the level of agreement between the data and the model ([Press et al. 2007](#)).

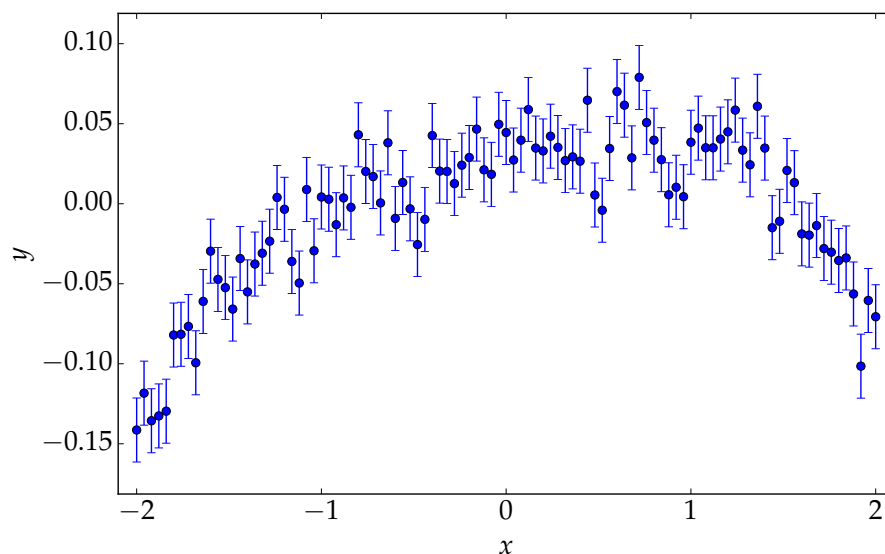


Figure 4.1 – The data points are shown here with their associated errors - A limiting range of $-2 \leq x \leq 2$ in the x -axis is used.

In this section, we shall fit a polynomial model to some data using different methods. We shall first apply a standard chi-square fitting procedure followed by the use of the normal equations method, the maximum likelihood method, the maximum a posteriori method (MAP), marginalisation and finally the MCMC method. We first begin with a brief description of the data. The data spans a range in the x -axis from -2 to 2 with the errors in each data point being equal to 0.02.

4.1.1 Chi-Square Fitting

Suppose we have N data points (x_i, y_i) where $i = 0, 1, 2, \dots, N-1$, then the quantity

$$\chi^2 = \sum_{i=0}^{N-1} \left(\frac{y_i - y(x_i | \theta_0, \theta_1, \dots, \theta_{M-1})}{\sigma_i} \right)^2 \quad (4.1.1)$$

is called the “chi-squared”. Assuming that the noise is Gaussian distributed, then this really is χ^2 distributed. $\theta_0, \theta_1, \dots, \theta_{M-1}$ are the parameters which we want to find and the model being used is $y(x_i | \theta_0, \theta_1, \dots, \theta_{M-1})$. We further assume that the model which will best fit the data is a polynomial of degree 4, that is, a quartic function of type $y(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$. The data is actually generated with the polynomial function $\theta_0 + \theta_1 x + \theta_4 x^4$.

The model parameters are obtained by minimising χ^2 . Hence, the parameters, represented by the vector θ are given by

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^2}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^3}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^4}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^2}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^3}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^4}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^5}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^2}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^3}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^4}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^5}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^6}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^3}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^4}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^5}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^6}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^7}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^4}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^5}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^6}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^7}{\sigma_i^2} & \sum_{i=0}^{N-1} \frac{x_i^8}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} \sum_{i=0}^{N-1} \frac{y_i}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i y_i}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^2 y_i}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^3 y_i}{\sigma_i^2} \\ \sum_{i=0}^{N-1} \frac{x_i^4 y_i}{\sigma_i^2} \end{pmatrix} \quad (4.1.2)$$

Moreover, we estimate the uncertainties in the parameters as the errors in the data points introduce some uncertainty in the determination of those parameters. We assume that the

data points are independent, thus each of them contributes to the overall uncertainty in the parameters. Hence, the propagation error (Press et al. 2007) on one particular parameter, say, θ_f will be

$$\sigma_{\theta_f}^2 = \sum_{i=0}^{N-1} \sigma_i^2 \left(\frac{\partial \theta_f}{\partial y_i} \right)^2 \quad (4.1.3)$$

In our case, since the errors are the same for each data point ($\sigma = 0.02$), we can neglect the index i in the denominator of Equation (4.1.1).

4.1.2 Use of Normal Equations

Let us consider a polynomial fitting function of degree $M - 1$. The model can thus be represented in the form

$$y(x) = \sum_{k=0}^{M-1} \theta_k X_k(x) \quad (4.1.4)$$

where $X(x)$ is the basis function, that is, $x, x^2, x^3, \dots, x^{M-1}$. Instead of writing Equation (4.1.1), we can now re-write the χ^2 as:

$$\chi^2 = \sum_{i=0}^{N-1} \left(\frac{y_i}{\sigma_i} - \frac{1}{\sigma_i} \sum_{k=0}^{M-1} \theta_k X_k(x_i) \right)^2 \quad (4.1.5)$$

We once again minimise the χ^2 to obtain the best-fit parameters. Let $\mathbf{D}$ be a matrix which is formed from the basis functions evaluated at the points x_i such that

$$D_{ij} = \frac{X_j(x_i)}{\sigma_i} \quad (4.1.6)$$

The matrix $\mathbf{D}$ is referred to as the design matrix and has dimensions $N \times M$. $\mathbf{D}$ must in principle have more rows than columns, meaning that we must have more data points than number of parameters ($N \geq M$) and the structure of matrix $\mathbf{D}$ is shown in Equation (4.1.8).

Notice that in the horizontal direction of the matrix, the basis function changes while in the vertical direction, a specific basis function is evaluated at the i^{th} data point. We also define the vector $\mathbf{b}$ of length N such that

$$b_i = \frac{y_i}{\sigma_i} \quad (4.1.7)$$

$$\begin{array}{c}
i = 0 \\
\uparrow \\
\text{datapoints} \\
\downarrow \\
i = N - 1
\end{array}
\left(\begin{array}{ccc}
& \leftarrow \text{Basis Functions} \rightarrow & \\
\frac{X_0(x_0)}{\sigma_0} & \frac{X_1(x_0)}{\sigma_0} & \frac{X_{M-1}(x_0)}{\sigma_0} \\
\vdots & & \vdots \\
\vdots & & \vdots \\
\frac{X_0(x_{N-1})}{\sigma_{N-1}} & \frac{X_1(x_{N-1})}{\sigma_{N-1}} & \dots \frac{X_{M-1}(x_{N-1})}{\sigma_{N-1}}
\end{array} \right) \quad (4.1.8)$$

Our objective is to minimise Equation (4.1.5) with respect to all the M parameters. Therefore, in this case, the derivative with respect to a specific parameter, say θ_k , is given by

$$0 = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[y_i - \sum_{j=0}^{M-1} \theta_j X_j(x_i) \right] X_k(x_i)$$

where $k = 0, 1, \dots, M-1$. By interchanging the position of summation, the above equation can be re-written in the form

$$\sum_{j=0}^{M-1} \left(\sum_{i=0}^{N-1} \frac{X_k(x_i) X_j(x_i)}{\sigma_i^2} \right) \theta_j = \sum_{i=0}^{N-1} \frac{X_k(x_i) y_i}{\sigma_i^2} \quad (4.1.9)$$

This is known as the normal equation and be re-cast in matrix form as:

$$(\mathbf{D}^T \mathbf{D}) \boldsymbol{\theta} = \mathbf{D}^T \mathbf{b} \quad (4.1.10)$$

where $\boldsymbol{\theta}$ is the vector of parameters which we want to find. Hence, the solution to the above is given by the equation

$$\boldsymbol{\theta} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{b} \quad (4.1.11)$$

The matrix $\mathbf{C} = (\mathbf{D}^T \mathbf{D})^{-1}$ is called the covariance matrix and gives the standard uncertainties associated with the parameters determined. In particular, the diagonal elements of this matrix give the variances of the parameters while the off-diagonal elements give the covariances between the parameters θ_j and θ_k . Hence, the errors on the parameters are equal to the

square root of the diagonal elements of the covariance matrix $\mathbf{C}$.

We confirm that using this method, we recover exactly the same result as shown in Table (5.2.1). However, the use of normal equations to solve least square problems is less tedious compared to the usual χ^2 fitting method. The use of normal equations can easily be extended to larger number of parameters but only works for linear equations.

4.1.3 MLE and MAP

Both Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) estimation are methods for estimating parameters given a model. However, MAP relies on the prior information of the parameters compared to MLE, thus immediately invoking the concepts of Bayesian Inference. Moreover, if we assume a uniform prior, then the MAP coincides with the MLE.

To find the parameters using the MLE approach, we assume the data points are independent and identically distributed as a Gaussian distribution around the true model $y(x)$. Hence, the likelihood $\mathcal{P}(\mathcal{D} | \theta, \mathcal{M})$ is given by

$$\mathcal{P}(\mathcal{D} | \theta, \mathcal{M}) = \frac{1}{(2\pi)^{N/2} \prod_i \sigma_i} \exp \left[-\frac{1}{2} (\mathbf{b} - \mathbf{D}\theta)^T (\mathbf{b} - \mathbf{D}\theta) \right] \quad (4.1.12)$$

where $\mathcal{D}$ represents the data, $\mathcal{M}$ is the model used, $\mathbf{b}$ and $\mathbf{D}$ are defined as in §(4.1.2). The most probable parameters are the ones which maximise Equation (4.1.12). This is equivalent to minimising $(\mathbf{b} - \mathbf{D}\theta)^T (\mathbf{b} - \mathbf{D}\theta)$, that is,

$$\frac{\partial \left[(\mathbf{b} - \mathbf{D}\theta)^T (\mathbf{b} - \mathbf{D}\theta) \right]}{\partial \theta} = 0 \quad (4.1.13)$$

We find that the parameters θ are then given by

$$\theta_{\text{MLE}} = \left(\mathbf{D}^T \mathbf{D} \right)^{-1} \mathbf{D}^T \mathbf{b} \quad (4.1.14)$$

which is the same as finding the parameters using normal equations or the simple chi-square method. This method is basically used by frequentists who do not base their investigation on prior knowledge of the parameters.

If instead we use the MAP approach, then instead of maximising the likelihood, the posterior distribution of the parameters is maximised. In particular,

$$\mathcal{P}(\boldsymbol{\theta} | \mathcal{D}, \mathcal{M}) = \frac{\mathcal{P}(\mathcal{D} | \boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta} | \mathcal{M})}{\mathcal{P}(\mathcal{D} | \mathcal{M})} \quad (4.1.15)$$

The denominator (evidence) is not a function of the parameters and hence plays no role in this particular problem. Therefore,

$$\boldsymbol{\theta}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} \mathcal{P}(\mathcal{D} | \boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta} | \mathcal{M}) \quad (4.1.16)$$

We use Gaussian priors, with mean $\bar{\boldsymbol{\theta}}$ and covariance $\mathbf{P}$ respectively:

$$\mathcal{P}(\boldsymbol{\theta} | \mathcal{M}) = \frac{1}{(2\pi)^{5/2} |\mathbf{P}|^{1/2}} \exp \left[-\frac{1}{2} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \mathbf{P}^{-1} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \right] \quad (4.1.17)$$

$|\mathbf{P}|$ and $\mathbf{P}^{-1}$ denote the determinant and inverse of the covariance matrix respectively. In this case, the mean and variance for each parameter are 0 and 1 respectively. Hence, the covariance matrix will just be the identity matrix with dimension 5×5 . Hence, $\mathbf{P} = \mathbf{I}$ and the above equation simplifies to:

$$\mathcal{P}(\mathcal{D} | \boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta} | \mathcal{M}) = k \exp \left[-\frac{1}{2} \left\{ (\mathbf{b} - \mathbf{D}\boldsymbol{\theta})^T (\mathbf{b} - \mathbf{D}\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{I} \boldsymbol{\theta} \right\} \right] \quad (4.1.18)$$

where k is simply the normalisation constant. From Equation (4.1.18), maximising the product of the likelihood and the prior is equivalent to minimising $(\mathbf{b} - \mathbf{D}\boldsymbol{\theta})^T (\mathbf{b} - \mathbf{D}\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{P}^{-1} \boldsymbol{\theta}$ and in this particular case, we find that

$$\boldsymbol{\theta}_{\text{MAP}} = \left(\mathbf{D}^T \mathbf{D} + \mathbf{I} \right)^{-1} \mathbf{D}^T \mathbf{b} \quad (4.1.19)$$

4.1.4 Marginalisation

Another approach to estimate parameters is to first marginalise over the nuisance parameters and then estimate the parameters from the resulting posterior. This can be done analytically in the case of Gaussian linear models. Given the probability density function (PDF) $f(x)$ of a continuous random variable say X , the mean $\bar{x}$ and variance σ^2 are given respectively as

$$\bar{x} = \int x f(x) dx \quad (4.1.20)$$

$$\sigma^2 = \int (x - \bar{x})^2 f(x) dx = \int x^2 f(x) dx - \bar{x}^2 = \overline{x^2} - \bar{x}^2 \quad (4.1.21)$$

The joint posterior distribution of the parameters θ is:

$$\mathcal{P}(\theta | \mathcal{D}, \mathcal{M}) \propto \mathcal{P}(\mathcal{D} | \theta, \mathcal{M}) \mathcal{P}(\theta | \mathcal{M}) \quad (4.1.22)$$

We use a Gaussian prior for each parameter with zero mean and unit variance. Therefore, the marginal posterior density function, for example for θ_0 , can be found using the following integral

$$\mathcal{P}(\theta_0 | \mathcal{D}, \mathcal{M}) = k \int \exp \left[-\frac{1}{2} \left\{ (\mathbf{b} - \mathbf{v}\theta_0 - \mathbf{D}\theta)^T (\mathbf{b} - \mathbf{v}\theta_0 - \mathbf{D}\theta) + \theta^T \mathbf{P}^{-1} \theta + \theta_0^2 \right\} \right] d\theta$$

where the design matrix $\mathbf{D}$ and the covariance matrix $\mathbf{P}$ are now of size $N \times 4$ and 4×4 respectively. In this particular case, k is a constant, $\mathbf{b}$ is the vector $\left[\frac{y_0}{\sigma_0}, \frac{y_1}{\sigma_1}, \dots, \frac{y_{N-1}}{\sigma_{N-1}} \right]$, $\mathbf{v}$ is the vector $\left[\frac{1}{\sigma_0}, \frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_{N-1}} \right]$, θ is the vector of parameters $[\theta_1, \theta_2, \theta_3, \theta_4]$, $\mathbf{P}$ is the covariance matrix for the priors and $\mathbf{D}$ is the design matrix constructed from the remaining basis functions (x, x^2, x^3, x^4) . After evaluating the above integral and normalising the expression found, the posterior for θ_0 is given by

$$\mathcal{P}(\theta_0 | \mathcal{D}, \mathcal{M}) = \sqrt{\frac{s}{2\pi}} \exp \left[-\frac{s}{2} \left(\theta_0 + \frac{t}{2s} \right)^2 \right] \quad (4.1.23)$$

where $s = \mathbf{v}^T \mathbf{v} - \mathbf{v}^T \mathbf{D} (\mathbf{D}^T \mathbf{D} + \mathbf{P}^{-1})^{-1} \mathbf{D}^T \mathbf{v} + 1$ and $t = 2\mathbf{b}^T \mathbf{D} (\mathbf{D}^T \mathbf{D} + \mathbf{P}^{-1})^{-1} \mathbf{D}^T \mathbf{v} - 2\mathbf{b}^T \mathbf{v}$. The above distribution is simply a Gaussian distribution with mean, $\mu_{\theta_0} = -t/2s$ and variance, $\sigma_{\theta_0}^2 = 1/s$.

The other parameters can also be found in a similar way. We only need to change the vector $\mathbf{v}$ and the design matrix $\mathbf{D}$. For example, if we want to find the distribution, mean and variance for the parameter θ_1 , then the vector $\mathbf{v}$ will be given as $\mathbf{v} = \left[\frac{x_0}{\sigma_0}, \frac{x_1}{\sigma_1}, \dots, \frac{x_{N-1}}{\sigma_{N-1}} \right]$. Our results are summarised in Table (5.2.1).

The posterior distributions of the parameters in the linear model are Gaussian distributions. In general, models which are used in most astrophysics and cosmology applications are non-linear. The posterior probability distribution is usually non-gaussian as it depends on the complexity of the model being used. As a result, analytical derivation of the posterior probability distribution function, mean and variance of the parameters may not be possible. We then rely on MCMC methods as explained in §(1.2.1).

4.2 Model Selection

One of the key questions underlying science is that of model selection: how do we select between competing theories which purport to explain observed data? The great paradigm shifts in science fall squarely into this domain. In the context of astronomy - as with most areas of science - the next two decades will see a massive increase in data volume through large surveys such as the Square Kilometre Array (SKA) (Hollitt et al. 2016) and LSST (Becla et al. 2006). Robust statistical analysis to perform model selection at large scale will be a critical factor in the success of such future surveys.

The basic problem of model selection is easy to state. As one considers models with more and more free parameters, one must expect that such models will fit any dataset better and better, irrespective of whether they have anything to do with reality. This problem of overfitting has led to many proposed methods to deal with this kind of situation: that is, finding a way to suitably penalise extra parameters. One method is LASSO (Least Absolute Shrinkage and Selection Operator) (Hastie et al. 2005). Other methods such as Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criteria (BIC) (Schwarz et al. 1978) penalise the best fit likelihood based on the number of free parameters (Gelman et al. 2014).

From a Bayesian point of view, model selection is not viewed as a question to be answered looking only at a single point in the parameter spaces, e.g. the point of maximum likelihood of the models in question, but rather should also depend on the full posterior distribution over the parameters. Hence selection is performed by choosing the model with the maximum model probability $\mathcal{P}(\mathcal{M}|\mathcal{D})$, derived from the Bayesian Evidence (or marginal likelihood) $\mathcal{P}(\mathcal{D}|\mathcal{M})$. This automatically expresses Occam's razor, thus penalising extra parameters which are not warranted by the data. Given two competing models, one would typically compute the Bayesian Evidence for each model and hence the Bayes Factor, which is the ratio of the evidences. There are, however, a number of issues with the Bayesian evidence. It is very sensitive to priors and, of key interest to us, since it involves integrals over the full parameter spaces of each model, is hard to compute efficiently, especially in high dimensions. Techniques such as nested sampling (Skilling 2004) can scale exponentially with the number of parameters and cannot be used for high-dimensionality problems. The prototype example in cosmology is multinest.

However, if one model is nested within the other (i.e. all the parameters of one model are shared by another), we can use the Savage-Dickey Density Ratio (SDDR) (Dickey 1971,

Verdinelli & Wasserman 1995) to directly calculate the Bayes Factor. In this section, we show how we can use the different methods which we have discussed in Chapter 1 (§1.3) to find the best model by calculating the evidence and hence calculating the Bayes Factor. Table (4.2.1) summarises our results.

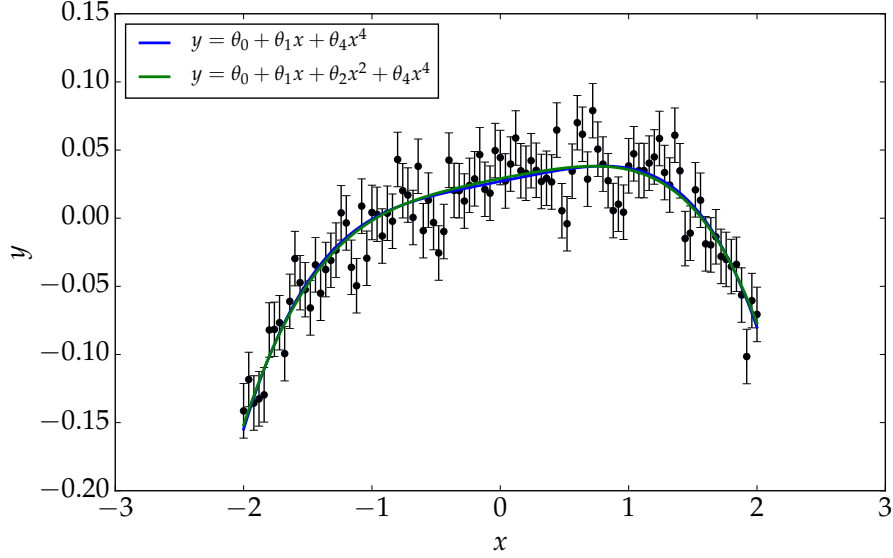


Figure 4.2 – Best fit to the linear model using MAP method - The true model is the quartic without the quadratic term (thick blue line). The errors are normally distributed with $\sigma = 0.02$. The Maximum a Posteriori (MAP) best-fits from the two models $y = \theta_0 + \theta_1x + \theta_4x^4$ and $y = \theta_0 + \theta_1x + \theta_2x^2 + \theta_4x^4$ are shown. Since the two fits are so similar, it is not surprising that the simpler model has the higher Bayesian Evidence.

We are motivated to compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$ where under $\mathcal{M}_1$, $y = \theta_0 + \theta_1x + \theta_2x^2 + \theta_4x^4$ while under $\mathcal{M}_2$, $y = \theta_0 + \theta_1x + \theta_4x^4$. The reason for choosing these two models is that we cannot state which one is the best model just by looking at the fit as shown in Figure (4.2).

4.2.1 Evidence and Bayes Factor

We first calculate the Bayesian Evidence ($\mathbb{Z}$) for all possible models which are contained in the general model $y = \theta_0 + \theta_1x + \theta_2x^2 + \theta_3x^3 + \theta_4x^4$ (see Table 4.2.1). The Bayesian Evidence is given by

$$\mathbb{Z} = \mathcal{P}(\mathcal{D}|\mathcal{M}) = \int \mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \quad (4.2.1)$$

$$\mathbb{Z} = \frac{1}{\prod_i \sqrt{2\pi\sigma_i^2}} \frac{1}{\sqrt{|2\pi\mathbf{P}^{-1}|}} \int \exp \left[-\frac{1}{2} \left\{ (\mathbf{b} - \mathbf{D}\boldsymbol{\theta})^T (\mathbf{b} - \mathbf{D}\boldsymbol{\theta}) + \boldsymbol{\theta}^T \mathbf{P}^{-1} \boldsymbol{\theta} \right\} \right] d\boldsymbol{\theta}$$

Number of Parameters	θ_0	θ_1	θ_2	θ_3	θ_4	$\Delta \log(\mathcal{L}_{\max})$	$\log \mathcal{Z}$	$\log \text{BF}$
1 parameter	✓	-	-	-	-	-262.463	-5.042	248.380
	-	✓	-	-	-	-203.629	53.638	189.700
	-	-	✓	-	-	-163.636	93.184	150.154
	-	-	-	✓	-	-218.616	37.669	205.669
	-	-	-	-	✓	-118.860	136.849	106.489
2 parameters	✓	✓	-	-	-	-202.451	48.597	194.741
	✓	-	✓	-	-	-72.355	178.650	64.688
	✓	-	-	✓	-	-217.438	32.628	210.710
	✓	-	-	-	✓	-60.012	189.700	53.638
	-	✓	✓	-	-	-103.624	146.822	96.516
	-	✓	-	✓	-	-202.676	48.153	195.185
	-	✓	-	-	✓	-58.848	190.487	52.851
	-	-	✓	✓	-	-118.611	130.853	112.485
	-	-	✓	-	✓	-90.087	160.054	83.284
	-	-	-	✓	✓	-73.835	174.518	68.820
3 parameters	✓	✓	✓	-	-	-12.313	232.288	11.050
	✓	✓	-	✓	-	-201.498	43.111	200.227
	✓	✓	-	-	✓	-	243.338	-
	✓	-	✓	✓	-	-27.330	216.319	27.019
	✓	-	-	✓	✓	-14.987	227.369	15.969
	-	✓	✓	✓	-	-102.671	141.336	102.002
	-	✓	✓	-	✓	-30.075	213.692	29.646
	-	✓	-	✓	✓	-57.894	185.001	58.337
	-	-	✓	✓	✓	-45.062	197.723	45.615
	✓	-	✓	-	✓	-59.730	184.820	58.518
4 parameters	✓	✓	✓	✓	-	-11.390	226.803	16.536
	✓	✓	✓	-	✓	0.282	238.458	4.880
	✓	✓	-	✓	✓	0.953	237.852	5.486
	✓	-	✓	✓	✓	-14.705	222.489	20.849
	-	✓	✓	✓	✓	-29.121	208.206	35.132
5 parameters	✓	✓	✓	✓	✓	1.235	232.973	10.366

Table 4.2.1 – Bayesian Evidence for each linear model considered - The table shows all the models we have considered, beginning with the model having the least number of parameters to the most extensive model. The seventh column gives the difference in the maximum log-likelihood, calculated relative to the best model (row in red). Moreover, the log-Bayesian Evidence is shown in the eighth column and the last column shows the Bayes Factor in logarithmic scale, calculated relative to the best model (the row shown in bold red)

Using the same notation as in the previous section, that is, $\mathbf{D}$ for the design matrix and $\mathbf{P}$ for the covariance matrix for the priors, the Bayesian Evidence (see Appendix A.3 for further details) is given by

$$\mathbb{Z} = \frac{\exp \left[-\frac{1}{2} (\mathbf{k} + \mathbf{b}^T \mathbf{b}) \right]}{\prod_i \sqrt{2\pi\sigma_i^2}} \sqrt{\frac{|2\pi (\mathbf{D}^T \mathbf{D} + \mathbf{P}^{-1})^{-1}|}{|2\pi \mathbf{P}^{-1}|}} \quad (4.2.2)$$

where

$$\mathbf{k} = -(\mathbf{D}^T \mathbf{b})^T (\mathbf{D}^T \mathbf{D} + \mathbf{P}^{-1})^{-1} (\mathbf{D}^T \mathbf{b})$$

We also calculate the log-Bayes Factor relative to the model having the highest evidence. In particular, for the two models which we are considering ($\mathcal{M}_1$ and $\mathcal{M}_2$), the log Bayes Factor is

$$\log B_{21} = \log \left[\frac{\mathcal{P}(\mathcal{D} | \mathcal{M}_2)}{\mathcal{P}(\mathcal{D} | \mathcal{M}_1)} \right] = 4.88 \quad (4.2.3)$$

thus showing that $\mathcal{M}_2$ is strongly favoured model over $\mathcal{M}_1$ (see Jeffreys' scale in Chapter 1 - §1.3.1).

4.2.2 Savage-Dickey Density Ratio

On the other hand, given the two models we are considering, $\mathcal{M}_2$ is nested in $\mathcal{M}_1$ at $\theta_2 = 0$. If we define $\boldsymbol{\phi}_1 = (\theta_0, \theta_1, \theta_4)$ and $\boldsymbol{\phi}_2 = (\theta_2)$, we can then write the likelihood and the priors as

$$\mathcal{P}(\mathcal{D} | \mathcal{M}_1, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \frac{1}{\prod_i \sqrt{(2\pi\sigma_i)}} \exp \left[-\frac{1}{2} (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1 - \mathbf{D}_2 \boldsymbol{\phi}_2)^T (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1 - \mathbf{D}_2 \boldsymbol{\phi}_2) \right]$$

$$\mathcal{P}(\boldsymbol{\phi}_1 | \mathcal{M}_1) = \frac{1}{\sqrt{|2\pi \mathbf{P}_1^{-1}|}} \exp \left(-\frac{1}{2} \boldsymbol{\phi}_1^T \mathbf{P}_1^{-1} \boldsymbol{\phi}_1 \right)$$

$$\mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{M}_1) = \frac{1}{\sqrt{|2\pi \mathbf{P}_2^{-1}|}} \exp \left(-\frac{1}{2} \boldsymbol{\phi}_2^T \mathbf{P}_2^{-1} \boldsymbol{\phi}_2 \right)$$

where $\mathbf{P}_1$ and $\mathbf{P}_2$ are covariance matrices of size 3 and 1 respectively and $\mathbf{D}_1$ and $\mathbf{D}_2$ are the appropriate design matrices. In this case, since we are considering unit variance and zero mean priors, $\mathbf{P}_1^{-1} = \mathbf{I}_1$ and $\mathbf{P}_2^{-1} = \mathbf{I}_2$ and are in fact the Fisher Information matrix of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ respectively. The SDDR is given by

$$\text{SDDR} = \left. \frac{\mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{M}_1)} \right|_{\boldsymbol{\phi}_2=0}$$

Therefore,

$$\mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{D}, \mathcal{M}_1) \propto \mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{M}_1) \int \mathcal{P}(\mathcal{D} | \mathcal{M}_1, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \mathcal{P}(\boldsymbol{\phi}_1 | \mathcal{M}_1) d\boldsymbol{\phi}_1$$

The normalised posterior distribution of $\boldsymbol{\phi}_2$ is given by

$$\mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{D}, \mathcal{M}_1) = \frac{1}{\sqrt{|2\pi\mathbf{B}^{-1}|}} \exp \left[-\frac{1}{2} \left(\boldsymbol{\phi}_2^T \mathbf{B} \boldsymbol{\phi}_2 + 2\boldsymbol{\phi}_2^T \mathbf{E} - \mathbf{E}^T \mathbf{B}^{-1} \mathbf{E} \right) \right]$$

where

$$\mathbf{A} = \left(\mathbf{D}_1^T \mathbf{D}_1 + \mathbf{P}_1^{-1} \right)^{-1}$$

$$\mathbf{B} = \mathbf{P}_2^{-1} + \mathbf{D}_2^T \mathbf{D}_2 - \mathbf{D}_2^T \mathbf{D}_1 \mathbf{A} \mathbf{D}_1^T \mathbf{D}_2$$

$$\mathbf{E} = \mathbf{D}_2^T \mathbf{D}_1 \mathbf{A} \mathbf{D}_1^T \mathbf{b} - \mathbf{D}_2 \mathbf{b}$$

Then, finally:

$$\text{SDDR} = \sqrt{\frac{|2\pi\mathbf{P}_2^{-1}|}{|2\pi\mathbf{B}^{-1}|}} \exp \left(-\frac{1}{2} \mathbf{E}^T \mathbf{B}^{-1} \mathbf{E} \right)$$

4.2.3 Our Method

In this section, we discuss the methods that we shall use to calculate the Bayes Factor. The key driver of our interest in these methods is the desire for techniques that do not scale exponentially with the complexity of the models, as occurs for multinet (Feroz et al. 2013). Our method combines multiple models into a single, nested, Supermodel using a hyperparameter which we call α . We can then use the SDDR to evaluate the Bayes Factor.

The embedding of the models can be done in at least two ways. One approach is to embed at the level of the models, another is at the level of the likelihoods. We call these two approaches the Combined Model and Combined Likelihood methods. We test both approaches, finding that the Combined Likelihood approach has significant performance advantages.

MCMC itself is useful as a method exactly because it does not scale exponentially with increasing number of parameters, and hence our goal is to use MCMC-based methods to compute the Bayes factor. Of course, as with any such method, convergence needs to be achieved and there is some evidence that our supermodel methods do make the posterior harder to sample,

with chains that have larger correlation lengths. Nevertheless, since our methods are fundamentally based on MCMC we argue they will still have better scaling properties than nested sampling. Let us now discuss and illustrate the methods in detail.

4.2.3.1 General Approach

In order to use the SDDR for model selection or comparison even in the case of non-nested models, we introduce a hyper-parameter, which we denote α , that takes on particular values for the two models that we want to compare (for example, 0 and 1). So if we want to compare model $\mathcal{M}_1$ with model $\mathcal{M}_2$, we construct a Supermodel that contains the sets of parameters β and γ of the models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, as well as a ‘nesting parameter’ α , and that recovers each of the models at $\alpha = 0, 1$ respectively. Namely it satisfies:

$$\mathcal{P}_S(\mathcal{D} | \beta, \gamma, \alpha = 0) = \mathcal{P}(\mathcal{D} | \beta, \mathcal{M}_1) \quad \mathcal{P}_S(\mathcal{D} | \beta, \gamma, \alpha = 1) = \mathcal{P}(\mathcal{D} | \gamma, \mathcal{M}_2) \quad (4.2.4)$$

where $\mathcal{P}_S(\mathcal{D} | \beta, \gamma, \alpha)$ is the supermodel likelihood. There are a potentially infinite number of supermodels that can achieve this. In this paper we restrict ourselves to study of the simplest linear implementations.

The priors for the Supermodel, $\mathcal{M}_\alpha$ additionally need to be chosen so that they correspond to the desired priors for $\mathcal{M}_1$ and $\mathcal{M}_2$ when $\alpha = 0$ and $\alpha = 1$ respectively. One way to do this is to have separable priors under each model such that the parameters corresponding to a specific model are integrated out relatively easily. Alternatively, one can even combine the models via both the likelihoods and the priors. In this way the models $\mathcal{M}_1$ and $\mathcal{M}_2$ are effectively nested inside the model $\mathcal{M}_\alpha$ for the purpose of the likelihoods, and we can use the SDDR to compute the Bayes factor between these two models,

$$B_{12} = \frac{B_{1\alpha}}{B_{2\alpha}} = \frac{\mathcal{P}_S(\alpha = 0 | \mathcal{D})}{\mathcal{P}_S(\alpha = 0)} \frac{\mathcal{P}_S(\alpha = 1)}{\mathcal{P}_S(\alpha = 1 | \mathcal{D})} \quad (4.2.5)$$

One could also use Bayes’ Theorem to show that the above holds. The Bayes Factor is given by

$$B_{12} = \frac{\mathcal{P}(\mathcal{D} | \mathcal{M}_1)}{\mathcal{P}(\mathcal{D} | \mathcal{M}_2)} \quad (4.2.6)$$

$\mathcal{M}_S$ is $\mathcal{M}_1$ when $\alpha = 0$. Therefore,

$$\mathcal{P}(\mathcal{D}|\mathcal{M}_1) = \mathcal{P}(\mathcal{D}|\mathcal{M}_S, \alpha = 0) = \mathcal{P}_S(\mathcal{D}|\alpha = 0)$$

$\mathcal{M}_S$ is $\mathcal{M}_2$ when $\alpha = 1$. Therefore,

$$\mathcal{P}(\mathcal{D}|\mathcal{M}_2) = \mathcal{P}(\mathcal{D}|\mathcal{M}_S, \alpha = 1) = \mathcal{P}_S(\mathcal{D}|\alpha = 1)$$

Using Bayes' Theorem

$$\mathcal{P}_S(\mathcal{D}|\alpha = 0) = \frac{\mathcal{P}_S(\alpha = 0|\mathcal{D}) \mathcal{P}(\mathcal{D})}{\mathcal{P}(\alpha = 0)} \quad (4.2.7)$$

$$\mathcal{P}_S(\mathcal{D}|\alpha = 1) = \frac{\mathcal{P}_S(\alpha = 1|\mathcal{D}) \mathcal{P}(\mathcal{D})}{\mathcal{P}(\alpha = 1)} \quad (4.2.8)$$

Therefore,

$$B_{12} = \frac{\mathcal{P}_S(\mathcal{D}|\alpha = 0)}{\mathcal{P}_S(\mathcal{D}|\alpha = 1)} = \frac{\mathcal{P}_S(\alpha = 0|\mathcal{D}) \mathcal{P}(\alpha = 1)}{\mathcal{P}_S(\alpha = 1|\mathcal{D}) \mathcal{P}(\alpha = 0)} \quad (4.2.9)$$

4.2.3.2 Transformation of α and Model Averaged Posteriors

Given a Supermodel one can make any transformation, $\alpha \rightarrow f(\alpha)$ as long as $f(\alpha)$ can take the values 0 and 1 within the domain of definition of α , so that Equation (4.2.4) holds. In actual applications these limits do not even need to be strictly verified; for example using $\alpha \rightarrow f(\alpha) = e^\alpha$ for $\alpha \in [-\Lambda, 0]$ is good enough for a large enough Λ , under the (usually true) assumption that the likelihood $\mathcal{P}_S(\mathcal{D}|\beta, \gamma, \alpha)$ tends in a continuous way to the limit $\mathcal{P}(\mathcal{D}|\beta, \mathcal{M}_1)$ as $f(\alpha) \rightarrow 0$. See §(5.5) for a detailed investigation.

In the above we have tacitly assumed that α is a continuous parameter. This is however not necessary, α can also be an index variable that takes discrete values. This case can be seen as the limit of a continuous α that has the form of a step function (or a hyperbolic tangent function with a sharper and sharper transition). In the discrete case, there is not even a need to explicitly construct a supermodel, as we are always only in one of the simpler models $\mathcal{M}_1$ or $\mathcal{M}_2$; see for example (Hee et al. 2015).

This limit is also interesting for another reason. It may be that we are not really interested in precise model probabilities, but rather we want to infer parameter constraints in situations where the model is uncertain. An example could be image reconstruction, for example in astronomy, with an unknown number of point sources. In this situation our object of interest is

the model-averaged posterior for a parameter θ ,

$$\mathcal{P}(\theta | \mathcal{D}) = \frac{\sum_j \mathcal{P}(\theta | \mathcal{D}, \mathcal{M}_j) \mathcal{P}(\mathcal{M}_j | \mathcal{D})}{\sum_j \mathcal{P}(\mathcal{M}_j | \mathcal{D})} \quad (4.2.10)$$

From Equation (4.2.5) we can see that the Bayes Factor B_{12} between two models is given by the probability to find $\alpha = 0$ or $\alpha = 1$ if both have equal prior probabilities. This means that the case where α is an indicator variable will directly give us model-averaged posteriors if we marginalise over all parameters except θ (but including α), without having to compute B_{12} explicitly.

4.2.3.3 More than two models

There are many different possibilities to deal with more than two models. They could be nested at different values of a single parameter α . Alternatively we can introduce a separate parameter α_i for each model together with the global constraint $\sum_i \alpha_i = 1$. In this way the space of the α_i forms a simplex which can be parameterised, for example, with barycentric coordinates on which an MCMC can move. The second approach has the advantage that each model can be reached from any point in the simplex without having to pass through potentially prohibitively bad regions in the global parameter space. On the other hand, we need to introduce nearly as many new parameters as we have models. In general it is unclear which of these two approaches is superior and we leave the study of multiple models to future work.

4.2.3.4 Using the Same Parameters vs Different Parameters

One of the fundamental choices when using the supermodel approach is how to deal with common parameters to the two models. There are again two options: to explicitly share the common parameters or to decouple the models by replicating the shared parameters and treating them as if they are not common. We have verified analytically that it does not matter which approach is taken since the hyper-parameter α is entirely in one of the models at either $\alpha = 0$ or $\alpha = 1$. In practice, when one chooses to replicate the shared parameters so there are no overlapping parameters, then it turns out that the correct model still gets chosen but the posterior distributions of the parameters in the wrong model become very difficult to sample from and hence the autocorrelation time of the α chain is large, making it hard to accurately estimate the log-Bayes Factor. We therefore maintain the common parameters for both models which minimises the total number of parameters.

4.2.3.5 Combined Likelihood Approach

The combined-likelihood method creates the supermodel by combining the two likelihoods via the hyper-parameter α . In this case, the two models are completely distinctive, in the sense that the likelihood $\mathcal{L}_1$ and $\mathcal{L}_2$ only depend on the model parameters β and γ respectively. The combined likelihood is then given by

$$\mathcal{L}_S = f(\alpha) \mathcal{L}_1 + (1 - f(\alpha)) \mathcal{L}_2 \quad (4.2.11)$$

where $\mathcal{L}_1 = \mathcal{P}(\mathcal{D} | \beta, \mathcal{M}_1)$ and $\mathcal{L}_2 = \mathcal{P}(\mathcal{D} | \gamma, \mathcal{M}_2)$. If $f(\alpha) = \alpha$, the posterior probability distribution of α is obtained by marginalising over the parameters β and γ as follows,

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) \propto \int_{\gamma} \int_{\beta} [\alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2] \mathcal{P}(\alpha, \beta, \gamma | \mathcal{M}_1, \mathcal{M}_2) d\beta d\gamma \quad (4.2.12)$$

The condition (4.2.4) applies: setting $\alpha = 1$ yields the Bayesian Evidence of model $\mathcal{M}_1$ while setting $\alpha = 0$ gives the Bayesian Evidence for model $\mathcal{M}_2$. If we assume the priors are separable, which is often the case and a uniform prior on α , then we can write the above posterior as

$$\mathcal{P}(\alpha | \mathcal{M}_1, \mathcal{M}_2) \int_{\gamma} \int_{\beta} [\alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2] \mathcal{P}(\beta, \gamma | \mathcal{M}_1, \mathcal{M}_2) d\beta d\gamma \quad (4.2.13)$$

Since the above integration is independent of α , the posterior will be of the form

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) \propto \mathcal{P}(\alpha | \mathcal{M}_1, \mathcal{M}_2) (m\alpha + c)$$

If a flat or wide Gaussian distribution for the prior (centred on $\alpha = 0.5$ as we assume that $\mathcal{P}(\mathcal{M}_i)$ is equally likely) is imposed on α for $\alpha \in [0, 1]$, then the posterior distribution of α is linear. Even in the case where the prior is not flat in α , the key point is that it is analytically known. The Bayes Factor, B_{21} is then simply given by the ratio of the posterior at the two endpoints. For a flat prior on α this gives:

$$\mathbf{B}_{21} = \frac{c}{m + c} \quad (4.2.14)$$

where m and c are the constants derived from Equation (4.2.13). The posterior distribution of α needs to be normalised, therefore we also have that $m = 2(1 - c)$ and thus $B_{21} = c/2 - c$. We now have a simple Bayesian parameter estimation problem, which is relatively straightforward to solve computationally using a Monte Carlo Markov Chain (MCMC) and a simple Metropolis-

Hastings algorithm (Metropolis et al. 1953, Hastings 1970).

Let us now consider an example to illustrate the combined likelihood method. Since we have been working with the two models $\mathcal{M}_1$ ($\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_4 x^4$) and $\mathcal{M}_2$ ($\theta_0 + \theta_1 x + \theta_4 x^4$), we will consider all common parameters only once. Using $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ (as defined in §4.2.2), the posterior distribution of α

$$c \int_{\boldsymbol{\phi}_1} \int_{\boldsymbol{\phi}_2} [\alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2] \mathcal{P}(\alpha | \mathcal{M}_1, \mathcal{M}_2) \mathcal{P}(\boldsymbol{\phi}_1 | \mathcal{M}_1, \mathcal{M}_2) \mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{M}_1, \mathcal{M}_2) d\boldsymbol{\phi}_1 d\boldsymbol{\phi}_2$$

where

$$\mathcal{L}_1 \sim \exp \left[-\frac{1}{2} (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1 - \mathbf{D}_2 \boldsymbol{\phi}_2)^T (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1 - \mathbf{D}_2 \boldsymbol{\phi}_2) \right]$$

$$\mathcal{L}_2 \sim \exp \left[-\frac{1}{2} (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1)^T (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1) \right]$$

and c is simply is normalisation constant. We can further express $\mathcal{L}_1$ in term of $\mathcal{L}_2$ as

$$\mathcal{L}_1 \sim \mathcal{L}_2 \exp \left[-\frac{1}{2} \left(\boldsymbol{\phi}_2 \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\phi}_2 - 2 \boldsymbol{\phi}_2^T \mathbf{D}_2^T (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1) \right) \right]$$

Then

$$\mathcal{L}_S = \mathcal{L}_2 \left\{ \alpha \exp \left[-\frac{1}{2} \left(\boldsymbol{\phi}_2 \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{\phi}_2 - 2 \boldsymbol{\phi}_2^T \mathbf{D}_2^T (\mathbf{b} - \mathbf{D}_1 \boldsymbol{\phi}_1) \right) \right] + 1 - \alpha \right\}$$

The normalised posterior distribution of α is given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = \frac{2(P - Q)}{P + Q} \alpha + \frac{2Q}{P + Q}$$

where

$$P = \sqrt{|2\pi \Sigma_1^{-1}| |2\pi \Sigma_2^{-1}|} \exp \left[-\frac{1}{2} \left(\mathbf{k}_2 + \mathbf{b}^T \mathbf{b} - \mathbf{b}^T \mathbf{D}_2 \Sigma_1^{-1} \mathbf{D}_2^T \mathbf{b} \right) \right]$$

$$Q = \sqrt{|2\pi \Sigma_3^{-1}| |2\pi \mathbf{P}_2^{-1}|} \exp \left[-\frac{1}{2} \left(\mathbf{k}_3 + \mathbf{b}^T \mathbf{b} \right) \right]$$

and

$$\Sigma_1 = \mathbf{D}_2^T \mathbf{D} + \mathbf{P}_2^{-1} \quad \Sigma_2 = \mathbf{P}_1^{-1} - \mathbf{D}_1^T \mathbf{D}_2 \Sigma_1^{-1} \mathbf{D}_2^T \mathbf{D}_1 + \mathbf{D}_1^T \mathbf{D}_1 \quad \Sigma_3 = \mathbf{D}_1^T \mathbf{D}_1 + \mathbf{P}_1^{-1}$$

$$\mathbf{k}_2 = - \left(\mathbf{D}_1^T \mathbf{D}_2 \Sigma_1^{-1} \mathbf{D}_2^T \mathbf{b} - \mathbf{D}_1^T \mathbf{b} \right)^T \Sigma_2^{-1} \left(\mathbf{D}_1^T \mathbf{D}_2 \Sigma_1^{-1} \mathbf{D}_2^T \mathbf{b} - \mathbf{D}_1^T \mathbf{b} \right)$$

$$\mathbf{k}_3 = - \left(\mathbf{D}_1^T \mathbf{b} \right)^T \Sigma_3^{-1} \left(\mathbf{D}_1^T \mathbf{b} \right)$$

Hence, the Bayes Factor is given by

$$B_{21} = \frac{\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)}{\mathcal{P}(\alpha = 1 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)} = \frac{Q}{P}$$

It also does not matter if we have completely un-nested and partially nested models. The Bayes Factor can still be recovered analytically for the Gaussian Linear Model case.

4.2.3.6 Combined Model Approach

We discussed above the implementation of the Supermodel idea through combining the models at the level of the likelihoods. Here we consider the alternative option: to combine them at the model level via a hyper-parameter α ,

$$\mathcal{M}_S = f(\alpha) \mathcal{M}_1 + (1 - f(\alpha)) \mathcal{M}_2 \quad (4.2.15)$$

In the case where $f(\alpha) = \alpha$, we will in this case usually assume a flat prior in the interval $[0, 1]$ for α , but other choices are possible (for example an enlarged interval which may make it easier to evaluate the posterior for α at $\alpha = 0$ and 1). The posterior distribution of α is then given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = \int_{\gamma} \int_{\beta} \mathcal{P}(\mathcal{D} | \alpha, \beta, \gamma, \mathcal{M}_1, \mathcal{M}_2) \mathcal{P}(\alpha, \beta, \gamma | \mathcal{M}_1, \mathcal{M}_2) d\beta d\gamma \quad (4.2.16)$$

The objective is to find $\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)$ and $\mathcal{P}(\alpha = 1 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)$ because at these two end-points, the posterior of α actually gives the Bayesian Evidence for each model. Hence, the Bayes Factor is given by

$$B_{21} = \frac{\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)}{\mathcal{P}(\alpha = 1 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)} \quad (4.2.17)$$

Although one can show analytically that this is correct, in practice, as we will explicitly show below, the marginal posterior of α can be a complicated and unknown function of α . We can obtain the Bayes Factor only by considering samples with $\alpha \approx 1$ and $\alpha \approx 0$ which means

we need to have a large number of samples in each limit. In contrast, the combined likelihood approach had the advantage that the posterior for α was simply a linear function, which makes it much easier and more accurate in practice to fit for the Bayes factor since all the samples can be used.

As an example, with the two models we are using, the mixture model can finally be written as

$$\mathcal{M}_S = \alpha\theta_2x^2 + \theta_0 + \theta_1x + \theta_4x^4$$

Hence, the likelihood of the mixture model can be written as

$$\mathcal{P}(\mathcal{D}|\mathcal{M}_S, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \alpha) = \frac{1}{\prod_i \sqrt{(2\pi\sigma_i)}} \exp \left[-\frac{1}{2} (\mathbf{b} - \mathbf{D}_1\boldsymbol{\phi}_1 - \alpha\mathbf{D}_2\boldsymbol{\phi}_2)^T (\mathbf{b} - \mathbf{D}_1\boldsymbol{\phi}_1 - \alpha\mathbf{D}_2\boldsymbol{\phi}_2) \right]$$

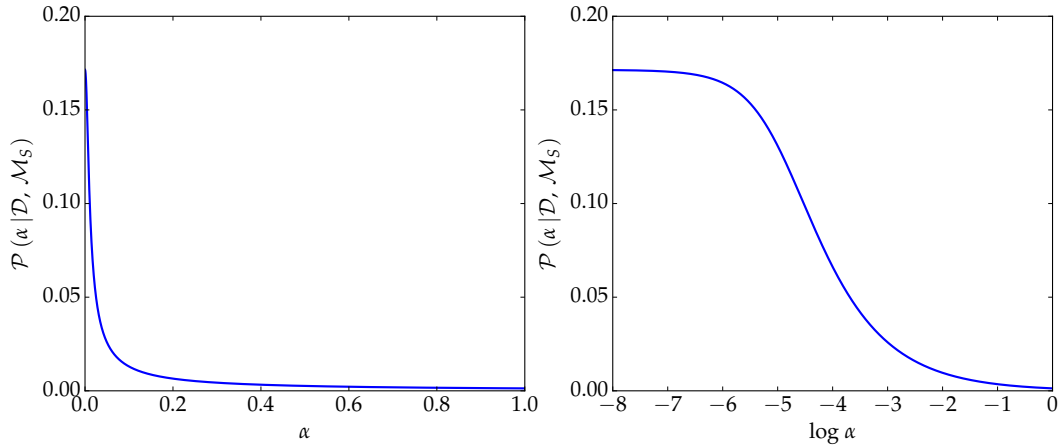


Figure 4.3 – The posterior distribution of α for the combined model approach - The left panel shows the posterior distribution of α for the following case: $\mathcal{M}_S = \alpha\mathcal{M}_1 + (1 - \alpha)\mathcal{M}_2$. Since $\mathcal{M}_2$ is the model with the higher evidence, we do expect the posterior to peak at roughly $\alpha = 0$. We also recover the log-Bayes Factor numerically using this method. One could also try a reparametrisation of the hyper-parameter which is shown on the right. This can be achieved by having a log-scale for α .

The posterior distribution of α is then given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_S) \propto \int_{\boldsymbol{\phi}_2} \int_{\boldsymbol{\phi}_1} \mathcal{P}(\mathcal{D} | \mathcal{M}_S, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \alpha) \mathcal{P}(\boldsymbol{\phi}_1 | \mathcal{M}_S) \mathcal{P}(\boldsymbol{\phi}_2 | \mathcal{M}_S) d\boldsymbol{\phi}_1 d\boldsymbol{\phi}_2$$

The un-normalised posterior distribution of α is given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_S) = k \sqrt{|2\pi\mathbf{M}| |2\pi\mathbf{N}|} \exp \left[\frac{1}{2} \left\{ \alpha^2 \mathbf{A} + \left(\alpha^2 \mathbf{B} - \mathbf{D}_1^T \mathbf{b} \right)^T \mathbf{N} \left(\alpha^2 \mathbf{B} - \mathbf{D}_1^T \mathbf{b} \right) \right\} \right]$$

where

$$\mathbf{A} = \mathbf{b}^T \mathbf{D}_2 \mathbf{M} \mathbf{D}_2^T \mathbf{b} \quad \mathbf{B} = \mathbf{D}_1^T \mathbf{D}_2 \mathbf{M} \mathbf{D}_2^T \mathbf{b}$$

$$\mathbf{M} = \left(\alpha^2 \mathbf{D}_2^T \mathbf{D}_2 + \mathbf{P}_2^{-1} \right)^{-1} \quad \mathbf{N} = \left(\mathbf{D}_1^T \mathbf{D}_1 - \alpha^2 \mathbf{D}_1^T \mathbf{D}_2 \mathbf{M} \mathbf{D}_2^T \mathbf{D}_1 + \mathbf{P}_1^{-1} \right)^{-1}$$

The normalisation constant k is found using Simpson's rule ([Press et al. 2007](#)) as it is difficult to obtain it analytically. In particular, the log-Bayes Factor using all the methods (ratio of Bayesian Evidences, SDDR, Combined Likelihood and Combined Model) all agree numerically, thus showing that our method is feasible. We now proceed to test the Supermodel approach with MCMC.

TESTING THE SUPERMODEL APPROACH WITH MCMC

"That's been one of my mantras - focus and simplicity. Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains." - Steve Jobs

5.1 Introduction

Given the Supermodel approach presented in the previous chapter, the idea is to simply sample the hyperparameter α using an MCMC algorithm to compute the Bayes Factor. In practice, we also have to deal with non-linear models, in which case, the posterior distribution of α cannot be determined analytically. We shall also investigate various other reparametrisation of α , both for the combined likelihood and the combined model. We shall delve into the pros and cons of each method.

5.2 Parameter Estimation - MCMC Method

In this section, we provide a short example of how MCMC works using a linear model. We use the MCMC Metropolis-Hastings method as explained in Chapter 1 to find the best-fit parameters as well as their associated errors. The errors are found at 68% credible interval (1σ level) and our results are summarized in the Table (5.2.1). In particular, the initial guess for all parameters (θ_{initial}) is taken from a Gaussian distribution with mean equal to 0.0 and a standard deviation of 1 and the jump ($\Delta\theta$) is also taken from a Gaussian distribution with mean equal to 0.0 and a standard deviation of 10^{-3} . Moreover, we use 2×10^6 steps and we take every second

point ('thinning') to form the final chain after burn-in. The distributions of all the parameters are shown in Figure (5.1).

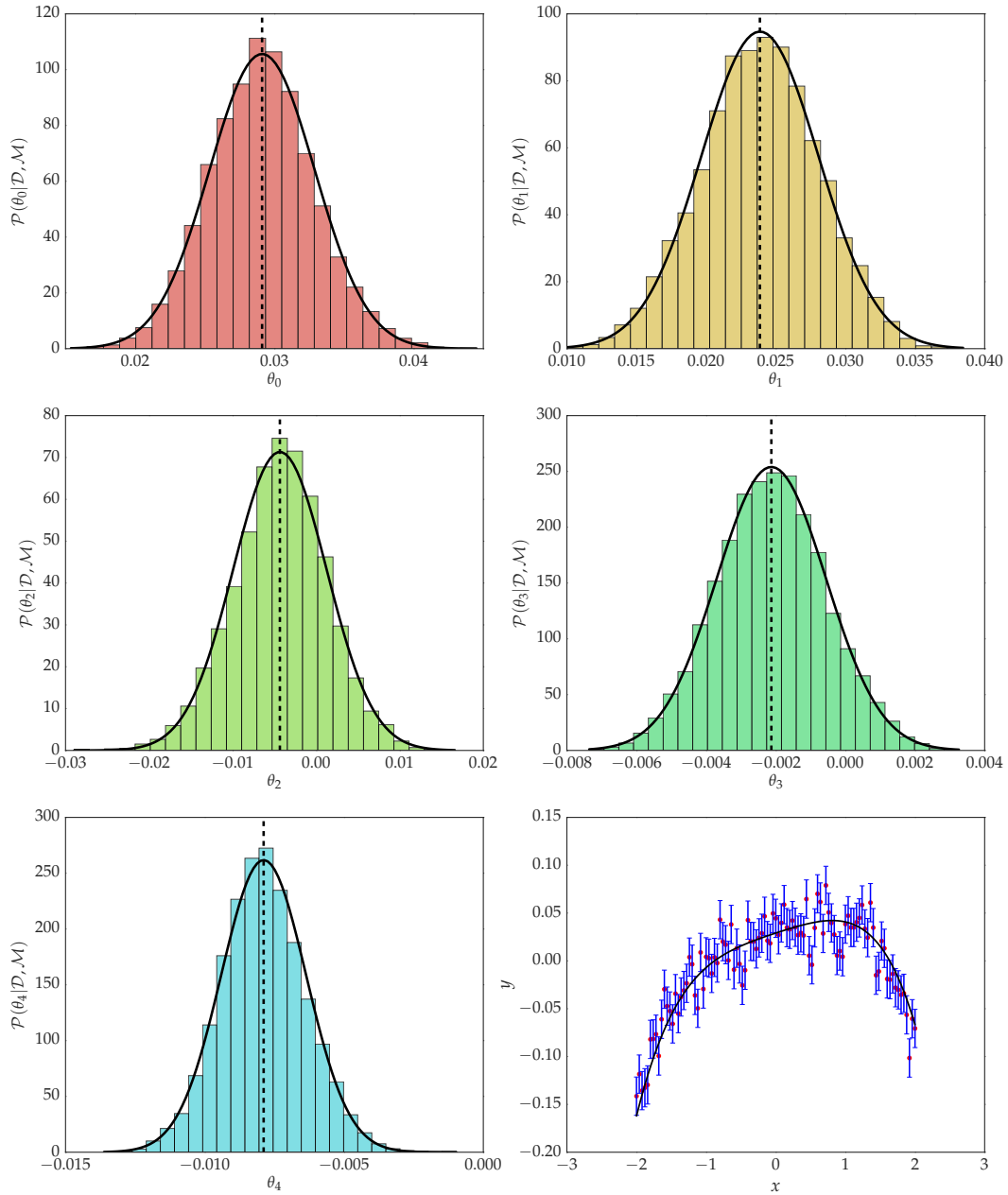


Figure 5.1 – The marginalised posterior distributions of the parameters using MCMC - The vertical broken line shows the mean of the distribution while the black solid curve shows the best-fit Gaussian to the distribution. The bottom right panel shows the fit to the data points where the mean of each parameter is used in the model, that is, $y(x) = \bar{\theta}_0 + \bar{\theta}_1 x + \bar{\theta}_2 x^2 + \bar{\theta}_3 x^3 + \bar{\theta}_4 x^4$.

As shown by this example, given a prior over the parameters, the MCMC is able to map the full posterior distributions of the parameters θ . For a linear model, one can do all the derivations, including the expressions for the posterior distributions analytically as shown in Chapter (4). However, for non-linear models, we have recourse to MCMC methods.

If we do not have a prior distribution, we can use a non-informative prior. However, this might slightly affect the convergence rate. If we start far from the “true” value of the parameter,

it would take long to reach a stationary distribution. Very often, it is a good idea to start with reasonable values of the parameters, which can be found by first invoking an optimisation algorithm. One option is to use an optimization algorithm to find a good starting point of the sampler first, followed by running the MCMC.

	MLE $\times 10^{-3}$	MAP $\times 10^{-3}$	Marginalisation $\times 10^{-3}$	MCMC $\times 10^{-3}$
θ_0	29.081 ± 3.732	29.081 ± 3.732	29.081 ± 3.732	$29.040^{+3.714}_{-3.693}$
θ_1	24.096 ± 4.277	24.096 ± 4.268	24.096 ± 4.268	$24.137^{+4.233}_{-4.203}$
θ_2	-4.306 ± 5.731	-4.305 ± 5.731	-4.305 ± 5.731	$-4.207^{+5.731}_{-5.811}$
θ_3	-2.207 ± 1.598	-2.206 ± 1.598	-2.206 ± 1.598	$-2.218^{+1.568}_{-1.574}$
θ_4	-7.895 ± 1.571	-7.895 ± 1.571	-7.895 ± 1.571	$-7.922^{+1.592}_{-1.569}$

Table 5.2.1 – Comparing the inferred parameters using different methods - The table shows the values of the parameters along with their associated errors using 4 different methods namely, MLE, MAP, Marginalisation and MCMC. In particular, the results from the MLE are similar to the chi-square fitting and normal equation methods (see §(4.1.1) and §(4.1.2)). The remaining methods take into account the priors on each parameter.

5.3 Bayes Factor Estimation

Recall from §(4.2.1) that the two linear models to our data (see Figure 4.1) which we are using are $y = \theta_0 + \theta_1 x + \theta_3 x^3 + \theta_4 x^4$ and $y = \theta_0 + \theta_1 x + \theta_4 x^4$ and are denoted by $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. As proven in the previous chapter, the preferred model is $\mathcal{M}_2$ with the log-Bayes Factor between the two models, $\log B_{21} = 4.88$. In the following sections, we show that we can recover reliable estimate of the log-Bayes Factor using the MCMC method.

5.3.1 Combined Likelihood

Compared to the combined model method which is slightly difficult to deal with (refer to the next section), the combined likelihood approach turns out to be relatively simple but is not without its own issues as we discuss. We will assume a uniform prior for α , that is, $\mathcal{P}(\alpha | \mathcal{M}_1, \mathcal{M}_2) \sim \mathcal{U}[0, 1]$. This leads to the posterior distribution of the hyperparameter α being linear irrespective of the type of models used. As a result, given $\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = m\alpha + c$, the condition $m = 2(1 - c)$ holds for $\alpha \in [0, 1]$ (see §(4.2.3.5)). This follows simply from the normalisation, in which case, m has been expressed in terms of c . Hence,

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = 2\alpha(1 - c) + c \quad (5.3.1)$$

The Bayes Factor is given by the ratio of the normalised posterior at the two end points, that is,

$$B_{21} = \frac{\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)}{\mathcal{P}(\alpha = 1 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)} = \frac{c}{2 - c}$$

If $\ell = \log B_{21}$,

$$c = \frac{2}{1 + e^{-\ell}}$$

The posterior of α in terms of ℓ is therefore given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = 2\alpha + \frac{2(1 - 2\alpha)}{1 + e^{-\ell}} \quad (5.3.2)$$

If we have independent samples for α from an MCMC run, we can treat the above as being the likelihood of α given the parameter ℓ . As a result,

$$\mathcal{P}(\alpha | \ell) = \prod_i \left[2\alpha_i + \frac{2(1 - 2\alpha_i)}{1 + e^{-\ell}} \right] \quad (5.3.3)$$

where α_i are the samples. We tested whether this approach works in practice by sampling directly from the analytically known distribution using Lahiri's method ([Lahiri 1951](#), [Cochran 1977](#)), and found that the result agreed with expectations.

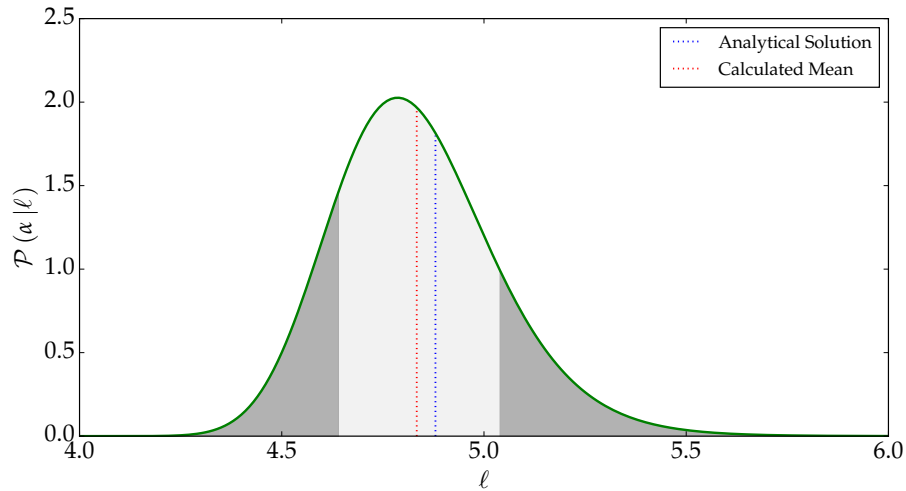


Figure 5.2 – Likelihood of the hyperparameter α for the linear models - For the MCMC, the number of steps is 2×10^7 and every 300 samples are recorded. In the end, we have around 66 000 recorded samples and fit for these samples directly. In the plot, the blue vertical dotted line shows the analytical solution while the red vertical dotted line shows the expected $\log B_{21}$ from the MCMC samples.

However, if the answer is not already known then one will have recourse to a MCMC method. We use the standard Metropolis-Hastings algorithm to obtain the α samples and we then fit for the samples directly using $\mathcal{P}(\alpha | \ell)$. The log-Bayes Factor agrees well with the analytical one at

1σ confidence interval as shown in Figure (5.2), we find

$$\log B_{21} = 4.83^{+0.21}_{-0.19} \quad (5.3.4)$$

Moreover, we can also generically predict the precision with which we can measure $\log B_{21}$ from N samples which are distributed with $\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)$ (the detailed calculation is shown in Appendix A.5). In particular, we can write the normalised posterior distribution as

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = (2\epsilon - 2)\alpha + (2 - \epsilon)$$

and the log-Bayes Factor as

$$\log B_{21} = \log \left(\frac{2 - \epsilon}{\epsilon} \right)$$

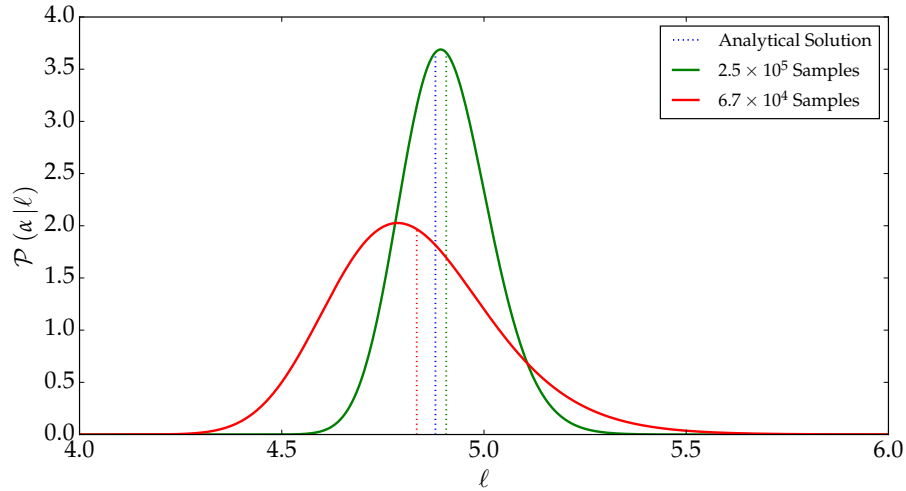


Figure 5.3 – Likelihood of the log-Bayes Factor for different number of samples - As the number of samples increases, the precision with which the log-Bayes Factor is determined increases. In the above plot, with roughly 6.7×10^4 samples, the log-Bayes Factor is $\log B_{21} = 4.83^{+0.21}_{-0.19}$ while with roughly 2.5×10^5 samples, the log-Bayes Factor is $\log B_{21} = 4.91 \pm 0.10$. In other words, in this case for the same log-Bayes Factor, while the number of samples has increased by roughly a factor of 4, the precision improves by a factor of 2, which is in agreement with Equation (5.3.5).

We find that

$$\sigma_{\log B_{21}} \propto \frac{1}{\sqrt{N}} \quad \sigma_{\log B_{21}} \propto \frac{1}{\epsilon(2 - \epsilon)} \quad (5.3.5)$$

We need to the order of 10^5 independent samples to determine the Bayes factor sufficiently accurately; this is comparable to the number of samples needed in other methods like nested

sampling (Skilling et al. 2006). Unfortunately the samples in a MCMC chain are correlated, and in the runs we used for this example, we found that we had to thin the chain by a factor of 300 by computing the autocorrelation length of the chain, in order to obtain effectively uncorrelated samples. In other words, we had to obtain a chain of over 10 million samples, which is computationally very costly. This problem can be alleviated by using other sampling methods to reduce the correlations between samples, for example Hamiltonian MC (HMC) (Neal et al. 2011). The computational cost of single HMC steps is however itself very high except if we can compute the gradients analytically.

The situation changes for problems in very high dimensional spaces relevant, for example, for feature selection. Multinest and polychord do not scale efficiently beyond a few tens or hundreds of dimensions, while MCMC methods have a much better scaling. The number of samples needed for the approaches discussed here depends on the dimensionality of the problem only through the correlation length, as sampling in higher dimensions is more difficult. This is however a relatively weak dependence so that the combined likelihood approach is especially promising for applications to model comparison in thousands of dimensions.

If we use a MCMC method to sample from the posterior, then, as we are dealing with combined likelihoods/models, it is important to have efficient sampling. The MCMC chain for α can be quite strongly autocorrelated. One way to reduce autocorrelation is to thin the chain, by recording only every n^{th} step. In our method, since we fit for the samples directly as we know the mathematical expression for the normalised posterior distribution, it is important to have independent samples as far as possible. We use the Python package `acor*` to monitor the autocorrelation time of the chain. A lower value of the autocorrelation time indicates less correlated samples, hence effectively more independent samples. In the simplest term, an autocorrelation time of τ for N samples shows that the effective sampling size is only N/τ . Therefore an autocorrelation time of $\tau = 1$ indicates that the chain is 100 % efficient. There is no well-defined method of choosing n but we ensure that $\tau < 10$ by thinning the chain by a reasonable value of n such that the chain is at least 10 % efficient.

5.3.2 Combined Model

In this section, we show that the mixture model actually works well but we also consider the limitations of this method. As explained in the previous chapter, the extended model is a mixture of the two models, $\mathcal{M}_1$ and $\mathcal{M}_2$. In this case as well, we have the analytical posterior

* <https://github.com/dfm/acor>

distribution of the hyperparameter α (refer to Figure 5.4).

As the combined model is given by $\mathcal{M}_3 = \alpha\mathcal{M}_1 + (1 - \alpha)\mathcal{M}_2$, we expect the posterior distribution at $\alpha = 0$ to be the highest as $\mathcal{M}_2$ is the preferred model. Nevertheless, we would require the values of the posterior at both ends, that is $\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)$ and $\mathcal{P}(\alpha = 1 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)$ in order to calculate the log-Bayes Factor.

Therefore, the way to proceed would be to obtain samples of α using a MCMC method and then fit the posterior distribution with an appropriate mathematical function. However, this requires a high level of accuracy for the posterior at both ends. However, this is difficult to achieve, given that the samples will typically be more concentrated towards $\alpha = 0$. In short, it is difficult to have the model probability (or weight) α sampling the whole parameter space properly. This eventually also affects the rate of convergence of the chain. An alternative to this is, a reparametrisation of α , which we attempt in §(5.5).

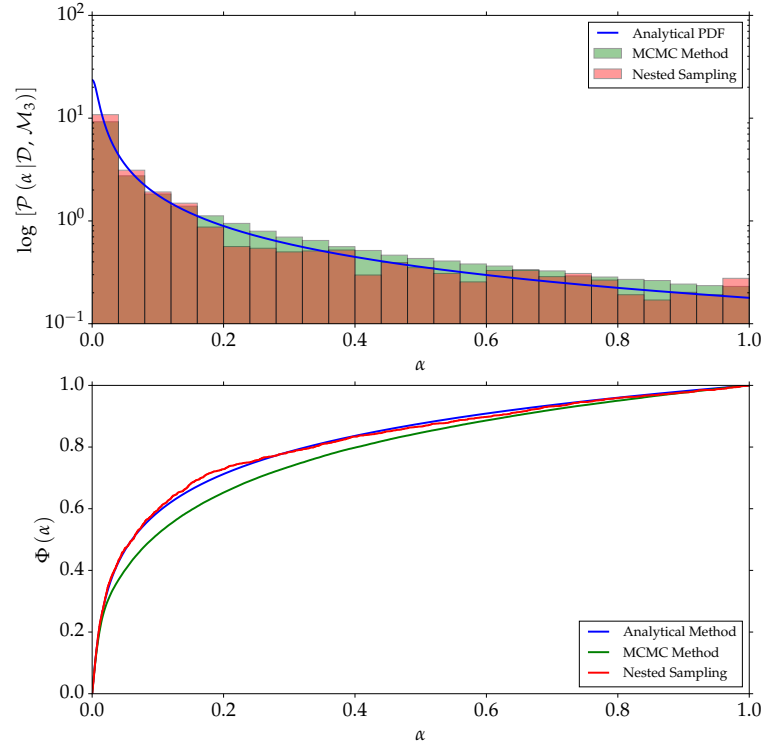


Figure 5.4 – Normalised Posterior and CDF of α - The top panel shows the normalised log-posterior distribution of α using three methods, analytical (shown in blue), MCMC (in green bins) and nested sampling (in red bins). Both nested sampling and MCMC perform badly at the boundaries (at $\alpha = 0$ and $\alpha = 1$). Moreover, it is difficult to find a proper mathematical expression to fit for the posterior distribution. The bottom panel shows the cumulative distribution function (CDF), $\Phi(\alpha)$. Compared to MCMC, nested sampling presumably performs better as it is well suited for dealing with multimodal distributions. However, we still have to deal with the issue of fitting the posterior.

5.4 Non-Linear Model

As we have demonstrated, the combined likelihood method works well in the linear model, we now explore the application to a non-linear model. In this case, the Bayesian Evidence cannot be computed analytically and hence we use Multinest to obtain a sound value for the Bayesian Evidence.

5.4.1 Data and Models

We consider two non-linear models: $\mathcal{M}_1 = \sin(\omega x)$ and $\mathcal{M}_2 = \sin(\omega x + \phi)$. We generate data from $\mathcal{M}_2$

$$y = \sin(\omega x + \phi) \quad (5.4.1)$$

with fiducial values $\omega = 1.0$ and $\phi = 0.06$, for $x \in [0, \pi]$.

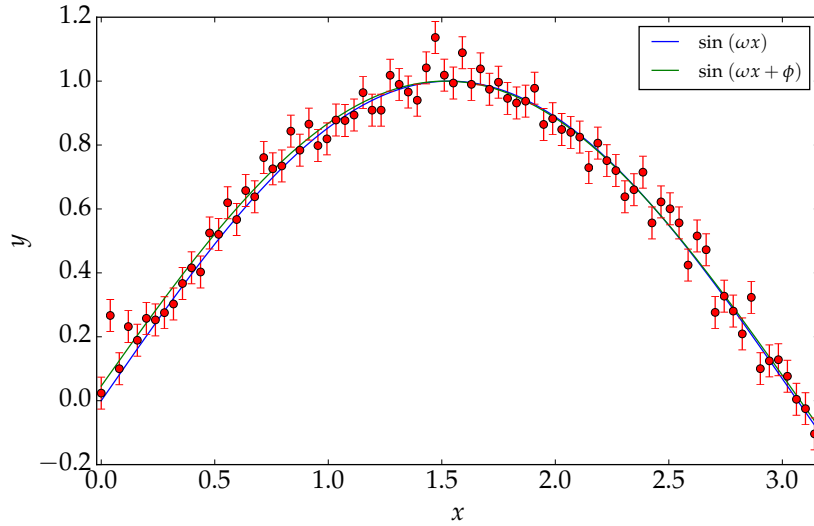


Figure 5.5 – The toy non-linear model - The data has been generated from $\mathcal{M}_2$, that is, $\sin(\omega x + \phi)$ with $\omega = 1$ and $\phi = 0.06$. The best fits corresponding to the two models are also shown. The data is tested in the combined likelihood approach in the case where $f(\alpha) = \alpha$ and $f(\alpha) = e^{-\alpha}$.

We add Gaussian noise with standard deviation 0.05. Figure (5.5) shows the data, along with the best-fitting functions $\mathcal{M}_1$ and $\mathcal{M}_2$. For the priors on the parameters, we choose independent Gaussian distributions with

$$\mathcal{P}(\omega | \mathcal{M}_1, \mathcal{M}_2) \sim \mathcal{N}(1, 2^2) \quad \mathcal{P}(\phi | \mathcal{M}_1, \mathcal{M}_2) \sim \mathcal{N}(0, 0.05^2)$$

The models being non-linear, we do not have an analytical solution for the Bayesian Evidence, and hence the Bayes Factor, which leads us to use PyMultinest (Buchner et al. 2014) to compute

the Bayes Factor, finding $\log B_{21} = 5.02 \pm 0.08$.

5.4.2 Combined Likelihood Result

Recall that the combined likelihood is given by

$$\mathcal{L}_3 = \alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2$$

We ran a chain of length 4×10^7 and uses the appropriate thinning (in this case a thinning factor of approximately 800) giving approximately 5×10^4 independent samples.

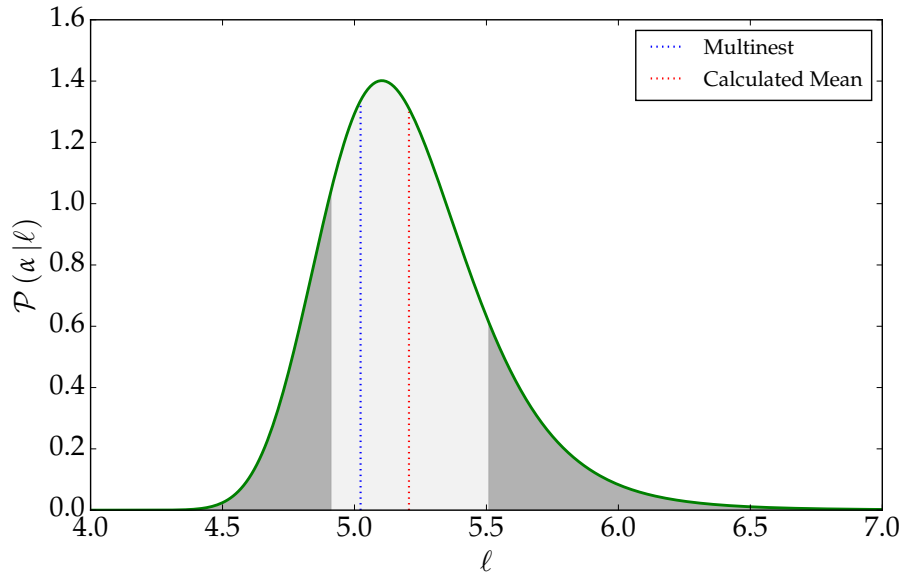


Figure 5.6 – The inferred likelihood for α for the toy non-linear problem - The resulting mean is fully consistent with the PyMultinest result.

The resulting distribution for the log-Bayes Factor is shown in Figure (5.6). The resulting mean of the log-Bayes Factor is given by $\log B_{21} = 5.21 \pm 0.30$, which is consistent with the PyMultinest estimate of $\log B_{21} = 5.02 \pm 0.08$. Of course, for such a small parameter space, Multinest is far superior in performance, however this gives evidence that the combined likelihood method carries over without change for non-linear models.

5.5 Transformation of the Hyperparameter

As demonstrated in the previous chapter, a reparameterisation of α for $\alpha \in [a, b]$ does not have an effect on the log-Bayes Factor as long as the condition $f(\alpha = a) = 0$ and $f(\alpha = b) = 1$. In the simplest case, $f(\alpha) = \alpha$ for $\alpha \in [0, 1]$. In this section, we try one possible reparameterisation of α amongst many. In particular, we choose $f(\alpha) = e^\alpha$ for $\alpha \in [-4, 0]$. We attempt both the

combined likelihood and the combined model using this reparameterisation.

5.5.1 Combined Likelihood

Under the assumption that we have a uniform prior for $\alpha \in [-4, 0]$, the posterior distribution of the hyperparameter is simply given by

$$\mathcal{P}(\alpha | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2) = m e^\alpha + c \quad (5.5.1)$$

As in §(5.3), it is a good idea of having the posterior distribution in terms of the log-Bayes Factor, $\ell = \log B_{21}$. This allows us to have a distribution for ℓ using the α samples from the MCMC. Therefore, using the condition for normalisation,

$$\int_p^0 (m e^\alpha + c) d\alpha = 1$$

We set $p = -4$ as below this value the posterior is nearly constant. This implies that one would need even more samples to determine larger log-Bayes Factor more accurately. We find that setting $p = -4$ allows us to obtain a reliable estimate of the log-Bayes Factor in the following two examples considered. Integrating and expressing m in terms of c ,

$$m = \frac{1 + cp}{1 - e^p} \quad (5.5.2)$$

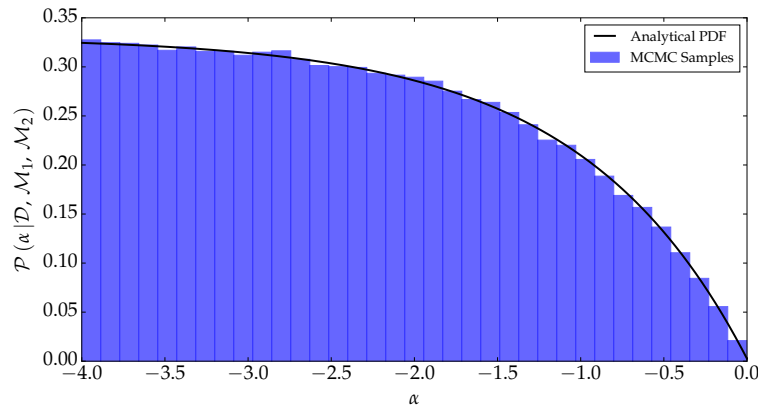


Figure 5.7 – MCMC run for linear model under the new reparameterisation - The analytical posterior distribution of α for the linear model is shown in black and the histogram corresponding to an MCMC run. The total number of steps in the MCMC is 5×10^6 , the thinning factor was set to 15 and eventually we have 330000 recorded samples.

The ratio at the ends of the interval $[p, 0]$ gives the Bayes Factor, hence

$$B_{21} = \frac{\mathcal{P}(\alpha = p | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)}{\mathcal{P}(\alpha = 0 | \mathcal{D}, \mathcal{M}_1, \mathcal{M}_2)} = \frac{m e^p + c}{m + c} \quad (5.5.3)$$

After substituting m using Equation (5.5.2), we can write the log-Bayes Factor as

$$\ell = \log \left[\frac{e^p + c(p e^p + 1 - e^p)}{1 + c(p + 1 - e^p)} \right] \quad (5.5.4)$$

Given the above equation, we can now express c in terms of ℓ as follows,

$$c = \frac{1 - e^{p-\ell}}{e^{-\ell}(p e^p + 1 - e^p) - (p + 1 - e^p)} \quad (5.5.5)$$

Therefore we now have both m and c in terms of ℓ and hence we can fit for the samples directly as we did earlier. We first try the method using the Linear model (which we used earlier) and since we can do everything analytically, we can have an idea of how the posterior of α is and is shown in Figure (5.7). We fit for the α samples directly using Equation (5.5.1) and our result is shown in Figure (5.8).

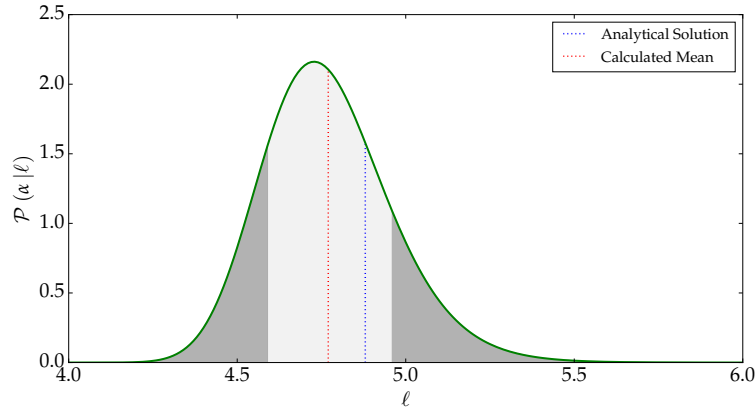


Figure 5.8 – The likelihood of α as we vary ℓ in the linear models - The result is consistent at 1σ confidence interval with the log-Bayes factor determined analytically.

The log-Bayes Factor is given by

$$\log B_{21} = 4.77^{+0.19}_{-0.18} \quad (5.5.6)$$

while the analytically calculated log-Bayes Factor is 4.88, hence our result is consistent at 1σ confidence interval.

We repeat the process with the non-linear models. It follows that the posterior distribution of the hyperparameter α will still follow Equation (5.5.1) as we marginalise over the models' parameters and not the hyperparameter α . In this case, the log-Bayes Factor is

$$4.98 \pm 0.19 \quad (5.5.7)$$

and the likelihood of α given ℓ is shown in Figure (5.9). In this case, the log-Bayes Factor from Multinest is 5.02 ± 0.08 . The advantage of using this transformation is that the sampling gets better. We do not need to have high thinning factors which is a clear advantage as we have a lot more samples in a shorter chain. Thinning factors of $\lesssim 50$ leads to the autocorrelation time being $\lesssim 10$.

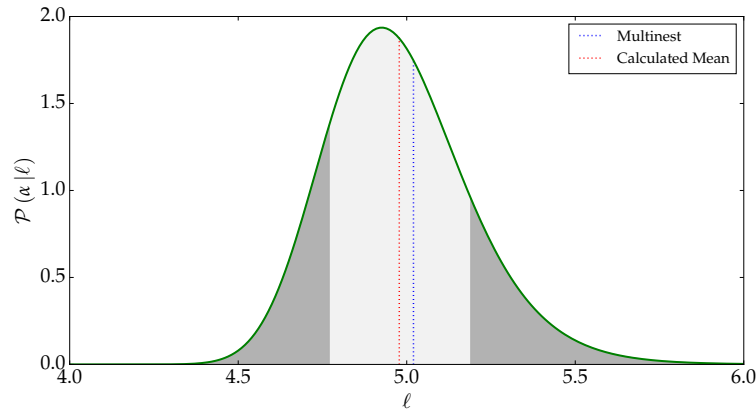


Figure 5.9 – The likelihood of α given ℓ , for the non-linear models - Our result is consistent with the Multinest result at 1σ confidence interval. In the MCMC, the number of steps was fixed to 10^7 with a thinning factor of 25, thus giving around 4×10^5 independent samples of α .

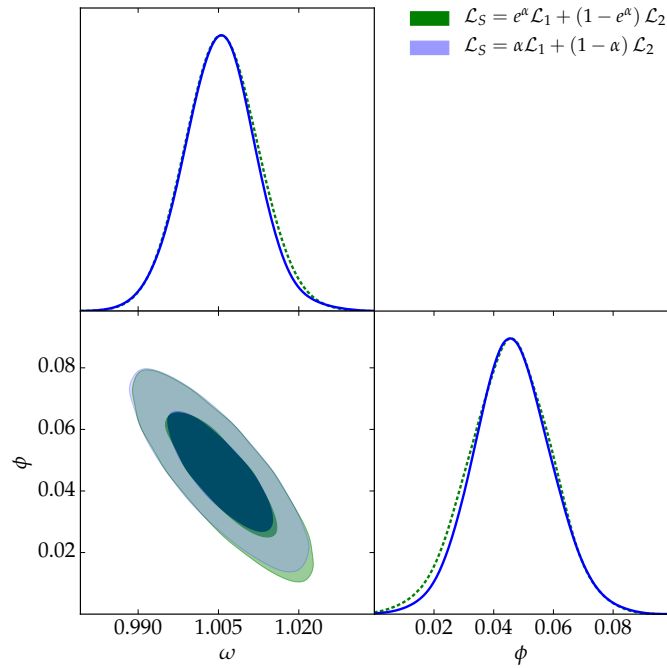


Figure 5.10 – The correlation of the parameters ω and ϕ in both methods - The blue line shows the posterior distribution of α in the combined likelihood method of the form $\mathcal{L}_3 = \alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2$ while the green broken line for $\mathcal{L}_3 = e^\alpha \mathcal{L}_1 + (1 - e^\alpha) \mathcal{L}_2$. The reparametrisation of α does not significantly affect the correlation between the parameters upon marginalisation over α .

Note that $\alpha \in [-4, 0]$. One can decrease the lower limit of α further as the normalised posterior distribution of α follows the generic shape of the function e^α but would require many more samples. Therefore, in short, this method is computationally less expensive but we still have to find the trade-off between the number of samples and the thinning factor.

Furthermore, under this reparametrisation of α , the correlation between the parameter posteriors within the model is invariant as shown in Figure (5.10). Therefore, one can effectively use $f(\alpha)$ for $\alpha \in [a, b]$ to combine the likelihoods of two separate models as long as the conditions $f(\alpha = a) = 0$ and $f(\alpha = b) = 1$ are satisfied.

5.5.2 Combined Model

Further, the combined model under the reparametrisation of α is given by

$$\mathcal{M}_3 = e^\alpha \mathcal{M}_1 + (1 - e^\alpha) \mathcal{M}_2 \quad (5.5.8)$$

The posterior as shown in Figure (5.11) is now better compared to the previous mixture model (refer to Figure (5.4)). It is also easier to sample from this posterior distribution using a MCMC method. As a result, one could possibly try to fit the normalised histogram as generated by the MCMC with a function such as $a \tanh(b\alpha + c) + d$ where a, b, c and d are the new parameters to be determined via, for example, optimisation.

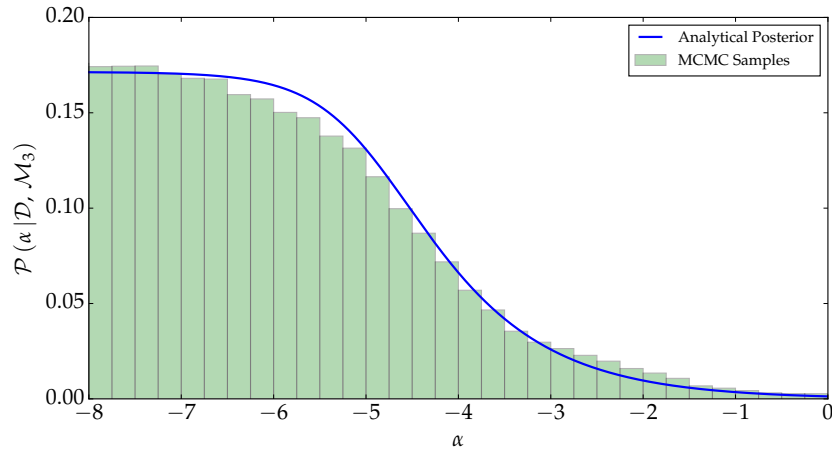


Figure 5.11 – Normalised posterior distribution of reparametrization of α - The number of samples is 2×10^5 , followed by fixing the number of iterations in the MCMC to 2×10^6 , with a thinning factor of 10. The curve in blue shows the analytical posterior distribution. The model which is fitted is $a \tanh(b\alpha + c) + d$ where the parameters a, b, c and d are determined via optimisation using the `scipy.optimize`[†] module. The true log-Bayes Factor is 4.88 while with this method, the log-Bayes Factor is 3.59.

However, two limitations of this are: first, it is complicated to have error bars on the binned histogram though one can presumably attempt to model the samples by a Poisson distribution

[†]<http://docs.scipy.org/doc/scipy/reference/optimize.html>

under the assumption that the samples generated are independent (which is debatable due to correlations in an MCMC chain). Secondly, in cases where we try to compare a different model $\mathcal{M}_1$, the analytical posterior of α can be very complicated. Hence having an analytic mathematical formulation to fit the posterior distribution is undoubtedly difficult.

In general, this is not a recommended approach as we would be biased if we get it wrong. As an example, given the log-Bayes Factor for this combined model is 4.88, the ratio of the posterior at the two ends of the best-fit gives 3.59. This clearly indicates that this method relies highly on the accuracy of the posterior at the two ends. If we get this wrong, then the whole analysis will be flawed. This is surely not a reliable way to proceed.

5.6 Non-Nested Models

In the above test cases, one of the two models is in fact nested in the other. The use of the Supermodel approach is not required as the Bayes Factor can be obtained using the Savage-Dickey Density Ratio. Ideally, what we would prefer is to combine two un-nested models and to illustrate this case, we consider the following two models:

$$\mathcal{M}_1 = ax^2 + bx + c$$

$$\mathcal{M}_2 = \sin(\omega x + \phi)$$

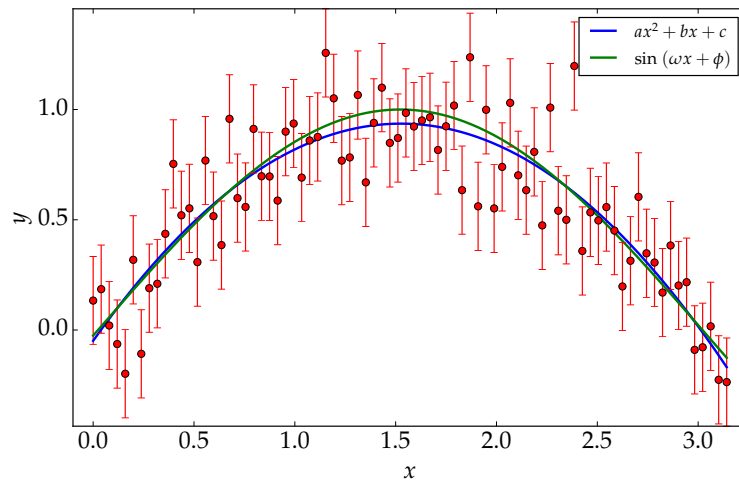


Figure 5.12 – Simulated data using sinusoidal function - The best fits (blue and green) to the data using models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, are shown here. These parameters are determined using the `scipy.optimize` module where $(a, b, c) = (-0.423, 1.291, -0.049)$ and $(\omega, \phi) = (1.048, -0.025)$.

Using the combined likelihood approach, the final likelihood is given by

$$\mathcal{L}_S = \alpha \mathcal{L}_1 + (1 - \alpha) \mathcal{L}_2$$

In particular, the data has been generated from $\mathcal{M}_2$ in which case the true values of the parameters are: $\omega = 1.0$, $\phi = 0.06$ and $\sigma = 0.20$.

We then proceed to estimating the Bayesian Evidence using PyMultinest. The priors used on the parameters are:

$$\mathcal{P}(a | \mathcal{M}_1) \sim \mathcal{N}(-0.5, 2.0^2) \quad \mathcal{P}(b | \mathcal{M}_1) \sim \mathcal{N}(1.5, 2.0^2) \quad \mathcal{P}(c | \mathcal{M}_1) \sim \mathcal{N}(0.0, 0.5^2)$$

$$\mathcal{P}(\omega | \mathcal{M}_2) \sim \mathcal{N}(1.0, 2.0^2) \quad \mathcal{P}(\phi | \mathcal{M}_2) \sim \mathcal{N}(0.0, 0.05^2)$$

The estimated Bayesian Evidences are: $\log \mathbb{Z}_1 = -43.91$ and $\log \mathbb{Z}_2 = -38.77$ and hence the log-Bayes Factor is $\log B_{21} = 5.14$. The aim is now to apply our combined-likelihood approach to this case, by finding the posterior distribution of the hyperparameter α using MCMC methods. We also account for correlated samples by appropriately thinning the chain.

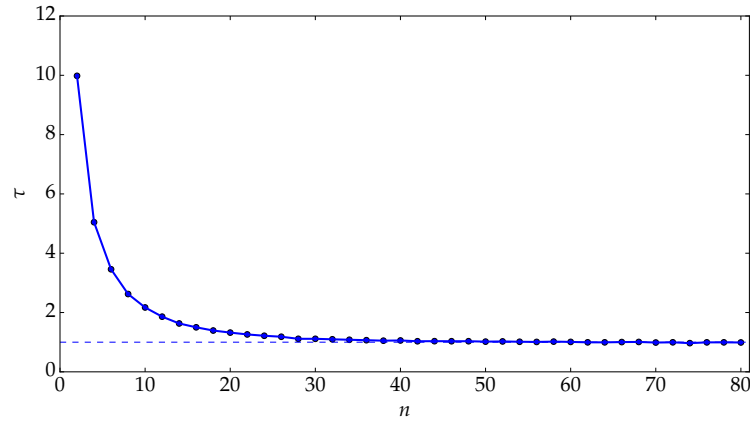


Figure 5.13 – The autocorrelation time, τ plotted against the thinning factor n . As expected, larger thinning factor leads to less correlation among the samples.

We now consider the case where the thinning factor, $n = 72$. We naively expect the likelihood of α , that is, $\mathcal{P}(\alpha | \ell)$ to peak at roughly the value of 5.14. However, we instead have an unbounded likelihood as shown in the figure below.

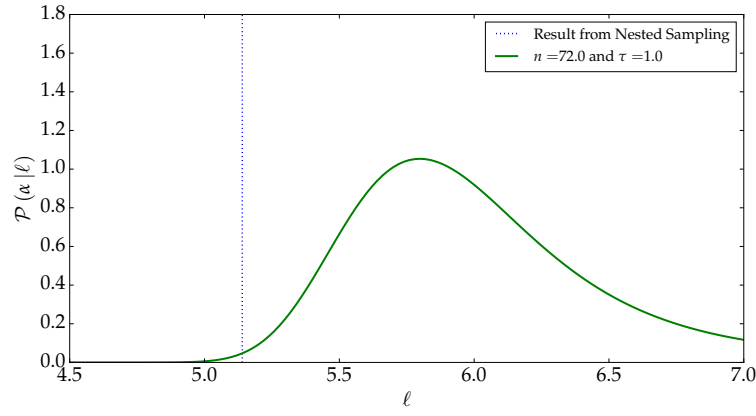


Figure 5.14 – The likelihood of α in the case of un-nested models - The broken vertical blue line shows the log-Bayes Factor obtained by running PyMultinest twice for the two models. The green curve corresponds to the likelihood of α using our combined-likelihood approach.

From the above plot, it appears that in the case of non-nested models, sampling the posterior distribution of α is not very efficient and would presumably require another sampling technique, other than Metropolis-Hastings algorithm to achieve more accurate Bayes Factor.

In this chapter, we have used standard MCMC to estimate the log-Bayes Factor by constructing a Supermodel. The latter can be done in two ways, either by combining models or by combining likelihoods. The combined likelihood approach looks more promising compared to the combined model approach because the posterior distribution of the hyperparameter α is a linear function and hence we have an analytic form of the likelihood of α (see Equation 5.3.3) to estimate the log-Bayes Factor. We also find that the estimation of the log-Bayes Factor using un-nested models is difficult because of the sampling method.

CALIBRATION AND SUPER-RESOLUTION IN BAYESIAN RADIO ASTRONOMY

"Probability does not exist." - Bruno de Finetti

6.1 Introduction

In this chapter, we first demonstrate that we can recover the flux densities of point radio sources, given a known model as previously done by [Lochner et al. \(2015\)](#). We make use of the Bayesian method which gives the whole distribution for the flux densities, thus taking into account the noise present in the received signal. Moreover, before employing these Bayesian techniques, we allow, for the first time, for the possibility that the visibilities are corrupted by antenna gains, that is, they are uncalibrated. Hence, we first perform calibration, followed by Bayesian Inference.

Furthermore, we leverage the power of the Bayesian formalism to distinguish between a point source and a two-point source, which is given by the Bayesian Evidence. Our aim is to compare these two models and obtain an expression for the average log-Bayes-Factor as a function of the angular separation. In general, this technique allows us to go beyond imaging to maximally leverage the sensitivity of radio telescopes. In the following sections, we will address these two topics in further detail.

6.2 Calibration and BIRO

Calibration is a crucial step to correct for the gains of each radio telescope. In simple term, these gains in fact cause a shift between the model and the data visibilities. Existing methods of cal-

ibration are discussed by [Boonstra & Van der Veen \(2003\)](#). Some methods rely on derivatives to solve for the gains of the telescopes. Other algorithms include SAGE ([Kazemi et al. 2011](#)) and another recent method using Wirtinger calculus adopted by [Smirnov & Tasse \(2015\)](#). In particular, the StEFCal (Statistically Efficient and Fast Calibration) algorithm ([Salvini & Wijnholds 2014](#)) uses an Alternate Direction Implicit (ADI) method to correct for the gains of radio telescopes, which does not make use of derivatives.

a	scalar value
$\mathbf{a}$	vector $\mathbf{a}$
$\mathbf{A}$	matrix $\mathbf{A}$
$\mathbf{A}_{:,k}$	k^{th} column of matrix $\mathbf{A}$
$\text{diag}(\cdot)$	converts a vector into a diagonal matrix
$(\cdot)^T$	transpose
$(\cdot)^+$	Hermitian transpose
$\mathbb{E}(\mathbf{A})$	expected value of $\mathbf{A}$
$\mathbf{M}$	model covariance matrix, with diagonal set to zero
$\mathbf{g}$	vector of complex values for the gains of each antenna
$\mathbf{G}$	$\mathbf{G} = \text{diag}(\mathbf{g})$

Table 6.2.1 – Notation for work on Calibration and BIRO

In the previous BIRO work carried out, it was assumed that the gains were all unity, that is, the data is already calibrated ([Lochner et al. 2015](#)). One would ideally solve for the antenna gains simultaneously with the fluxes using BIRO. For simplicity, it is better to first calibrate the data. In the following sections, we provide a brief review of StEFCal with the notations used shown in Table (6.2.1). We further discuss our methods for doing Bayesian Inference.

6.2.1 Setting up the Measurement Equation

Consider an array consisting of N receiver elements. The direction-independent gain can be represented by the complex number g_i ($i = 1, 2, 3, \dots, N$) for each of the receiving elements as shown in Figure (6.1). The output signal from the i^{th} telescope from Q sources, as a function of time, can be described by

$$x_i(t) = g_i \sum_{q=1}^Q a_{i,q} s_q(t) + n_i(t) \quad (6.2.1)$$

where $a_{i,q}$ is the antenna response, which is characterised by the beam pattern, to the q^{th} source,

s_q is the “corrupted” signal from source q and n_i is the noise on the measurement from receiver i . For all the antennae, we can define the following vectors,

- $\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_N(t)]^T$ for the signal received,
- $\mathbf{g} = [g_1, g_2, \dots, g_N]^T$ for the complex gains,
- $\mathbf{a}_q = [a_{1,q}, a_{2,q}, \dots, a_{N,q}]$ for the array response corresponding to the q^{th} source and
- $\mathbf{n}(t) = [n_1(t), n_2(t), \dots, n_N(t)]^T$ for the noise.

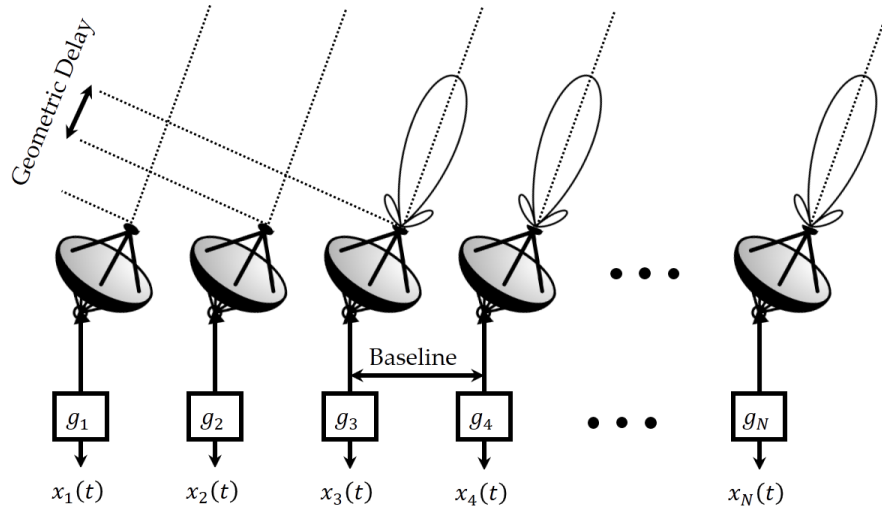


Figure 6.1 – Array of telescopes - The complex gains of the telescopes are given by g_i while the output from each antenna are $x_i(t)$. The geometric delay between the antennae and the baseline is also shown.

Therefore, the array signal received by N antennae from Q sources can be described by

$$\mathbf{x}(t) = \mathbf{g} \odot \sum_{q=1}^Q \mathbf{a}_q s_q(t) + \mathbf{n}(t) \quad (6.2.2)$$

where $\odot$ represents the Hadamard or element-wise product of matrices. If we now define the $N \times N$ matrix $\mathbf{G}$ as $\mathbf{G} = \text{diag}(\mathbf{g})$, the $N \times Q$ matrix $\mathbf{A}$ as $[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_Q]$ and the $Q \times 1$ vector $\mathbf{s}(t)$ as $[s_1(t), s_2(t), \dots, s_Q(t)]$, then the above equation can be re-written in a simpler way as

$$\mathbf{x}(t) = \mathbf{G}\mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \quad (6.2.3)$$

If we have a total of K samples of recorded measurements, then the matrix of “observed” visibilities is given by

$$\mathbf{V} = \mathbb{E} \left(\frac{\mathbf{X}\mathbf{X}^\dagger}{K} \right) = \mathbf{G}\mathbf{A}\Sigma_s\mathbf{A}^\dagger\mathbf{G}^\dagger + \Sigma_n \quad (6.2.4)$$

where $\mathbf{X} = [\mathbf{x}(T), \dots, \mathbf{x}(KT)]$ and KT denotes the duration of observation, $\mathbf{\Sigma}_s = \mathbb{E}[\mathbf{s}(t)\mathbf{s}^\dagger(t)]$ is the covariance matrix of the source signals and $\mathbf{\Sigma}_n = \mathbb{E}[\mathbf{n}(t)\mathbf{n}^\dagger(t)]$ is the noise covariance matrix. We also assume that the noise covariance matrix is uncorrelated, that is, $\mathbf{\Sigma}_n = \text{diag}(\sigma_n^2)$. If we now define $\mathbf{M} = \mathbf{A}\mathbf{\Sigma}_s\mathbf{A}^\dagger$, then our aim is to solve the following optimisation problem,

$$\tilde{\mathbf{g}} = \underset{\mathbf{g}}{\text{argmin}} \left\| \mathbf{V} - \mathbf{G}\mathbf{M}\mathbf{G}^\dagger \right\|_{\text{F}}^2 \quad (6.2.5)$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm, which is defined as the square root of the sum of absolute squares of the elements of a matrix. The way to proceed is to use the alternate direction implicit method (ADI) method, that is, to solve for $\mathbf{G}^\dagger$ first by keeping $\mathbf{G}$ constant and then solve for $\mathbf{G}$ by keeping $\mathbf{G}^\dagger$ constant. However, since $\mathbf{V} - \mathbf{G}\mathbf{M}\mathbf{G}^\dagger$ is Hermitian, the two steps are equivalent. Hence,

$$\mathbf{G}^{[j]} = \underset{\mathbf{G}}{\text{argmin}} \left\| \mathbf{V} - \mathbf{G}^{[j-1]}\mathbf{M}\mathbf{G}^\dagger \right\|_{\text{F}}^2 \quad (6.2.6)$$

Defining $\mathbf{Z}^{[j]} = \mathbf{G}^{[j]}\mathbf{M}$ and since $\|\mathbf{y}\|_{\text{F}} = \|\mathbf{y}\|_2$ for any vector $\mathbf{y}$,

$$\mathbf{G}^{[j]} = \underset{\mathbf{G}}{\text{argmin}} \left\| \mathbf{V} - \mathbf{Z}^{[j-1]}\mathbf{G}^\dagger \right\|_2^2 \quad (6.2.7)$$

$$\left\| \mathbf{V} - \mathbf{Z}^{[j-1]}\mathbf{G}^\dagger \right\| = \sqrt{\sum_{i=1}^N \left\| \mathbf{V}_{:,i} - \mathbf{Z}_{:,i}g_i^* \right\|_2^2}$$

However, given we have to find the parameters $\boldsymbol{\theta}$ in the case where $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$, then, as described in Chapter (4), we can use the normal equation method to find $\boldsymbol{\theta}$ such that the solution is given by $\boldsymbol{\theta} = (\mathbf{A}^\text{T}\mathbf{A})^{-1}\mathbf{A}^\text{T}\mathbf{b}$. In a similar way,

$$(\mathbf{Z}_{:,i})^\dagger \mathbf{Z}_{:,i}g_i^* = (\mathbf{Z}_{:,i})^\dagger \mathbf{V}_{:,i}$$

$$g_i^{[j]} = \left\{ \frac{(\mathbf{Z}_{:,i}^{[j-1]})^\dagger \mathbf{V}_{:,i}}{(\mathbf{Z}_{:,i}^{[j-1]})^\dagger \mathbf{Z}_{:,i}^{[j-1]}} \right\}^* = \frac{(\mathbf{V}_{:,i})^\dagger \mathbf{Z}_{:,i}^{[j-1]}}{(\mathbf{Z}_{:,i}^{[j-1]})^\dagger \mathbf{Z}_{:,i}^{[j-1]}} \quad (6.2.8)$$

Hence, the effects of the antennae gains can be removed by using the following equation,

$$\mathbf{V}^c = \mathbf{G}^{-1}\mathbf{V}\mathbf{G}^{-\dagger} = (\mathbf{g}\mathbf{g}^\dagger)^{-1}\mathbf{V} \quad (6.2.9)$$

where $(\cdot)^{-\dagger}$ denotes the inverse of the Hermitian transpose.

Salvini & Wijnholds (2014) also found that in some cases, the gain solutions do not converge at all and this was dealt with by replacing the gain solution at each iteration by the mean of the current solution and the previous one. It turns out that this simple modification makes the process very robust and fast. The algorithm for the above mathematical derivations is shown below.

Algorithm 6.1 StEFCal

Initiate $\mathbf{G}^{[0]}$; $\mathbf{G}^{[0]} = \mathbf{I}$ is adequate in most cases.

```

for  $j = 1, 2, \dots, j_{\max}$  do
  for  $i = 1, 2, \dots, N$  do
     $\mathbf{z} \leftarrow \mathbf{G}^{[i-1]} \cdot \mathbf{M}_{:,i} \equiv \mathbf{g}^{[i-1]} \odot \mathbf{M}_{:,i}$ 
     $g_i \leftarrow \frac{(\mathbf{V}_{:,i})^\dagger \cdot \mathbf{z}}{\mathbf{z}^\dagger \mathbf{z}}$ 
  end for
  if  $\text{mod}_2(j) = 0$  then
    if  $\frac{\|\mathbf{g}^{[i]} - \mathbf{g}^{[i-1]}\|_F}{\|\mathbf{g}^{[i]}\|_F} \leq \tau$  then
      Convergence Reached
    else
       $\mathbf{G}^{[i]} \leftarrow \frac{\mathbf{G}^{[i]} + \mathbf{G}^{[i-1]}}{2}$ 
    end if
  end if
end for

```

6.2.2 Example of Calibration

To test our code, we first do a preliminary run of the code using a single point source. In particular, we fix the number of iterations ($j_{\max}$) to 300 and the tolerance, τ to 10^{-4} . We do not go into the full details of how to set up these values (see Salvini & Wijnholds (2014) for details of implementation). Given N antennae, the number of baselines is given by

$$n_b = \frac{N(N-1)}{2}$$

As an example, consider the case where $N = 4$ with $i = 0, 1, 2, 3$. Then, $n_b = 6$. In this case, it is easy to construct a 4×4 square matrix as follows:

$$\mathbf{V}(t) = \begin{pmatrix} \mathbf{V}_{00} & \mathbf{V}_{01} & \mathbf{V}_{02} & \mathbf{V}_{03} \\ \mathbf{V}_{10} & \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} \\ \mathbf{V}_{20} & \mathbf{V}_{21} & \mathbf{V}_{22} & \mathbf{V}_{23} \\ \mathbf{V}_{30} & \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} \end{pmatrix}$$

In particular, $\mathbf{V}_{pq} = \mathbf{V}_{qp}$ since $\mathbf{V}$ is Hermitian. The above is obviously true for any number

of baselines. Consider now a single point source of intensity 2 Jy located at $\alpha = 03^{\text{h}}26^{\text{m}}00^{\text{s}}$ and $\delta = 60^{\circ}04'00''$ where (α, δ) are the right ascensions and declination of the source. The field centre is at $\alpha_0 = 03^{\text{h}}25^{\text{m}}00^{\text{s}}$ and $\delta_0 = 60^{\circ}00'00''$. We model an East-West interferometer consisting of only two antennae (hence a single baseline) in which case $|\mathbf{b}| = 144 \text{ m}$ and $\nu = 1400 \text{ MHz}$. The corrupted, model and corrected visibilities are shown below.

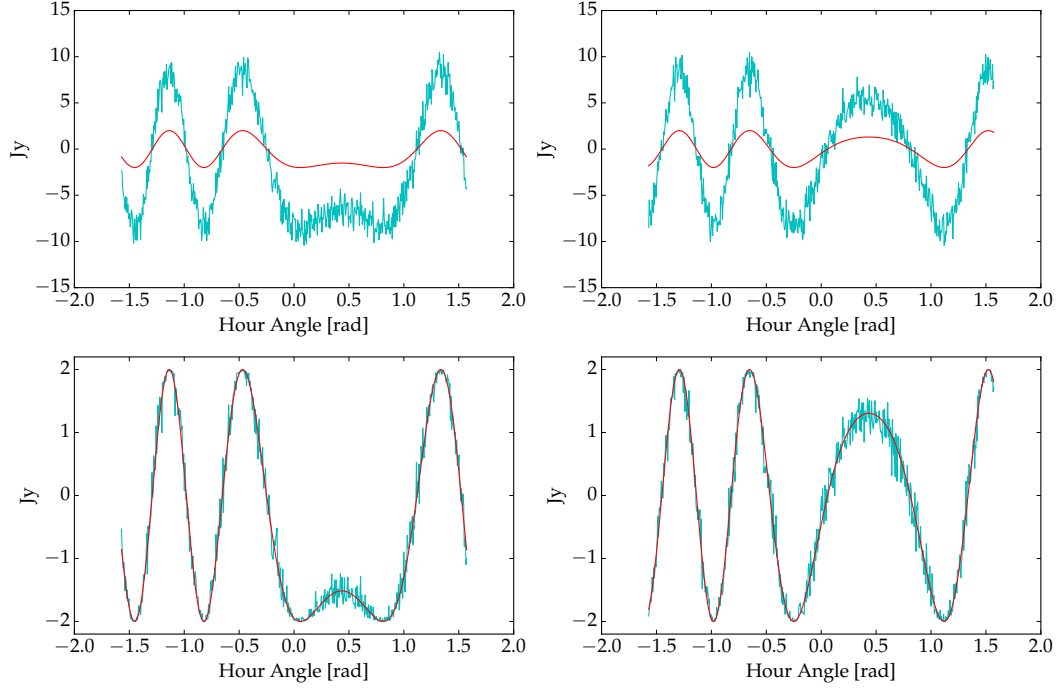


Figure 6.2 – Calibration for a single point source and a single baseline - The real and imaginary parts of the corrupted data are in cyan (generated with $\sigma = 1$) while the red plot shows the model visibilities in the upper left and upper right panel respectively. After calibration (using StEFCal), the real and imaginary parts of the corrected visibilities along with the model visibilities are shown in the lower left and right panels respectively.

6.2.3 Bayesian Inference

Now that we have a method for doing calibration, we can now proceed with Bayesian Inference methods for finding the posterior distributions of the flux densities. In the following sections, we describe how the Measurement Equation, the likelihood and the prior are set up.

6.2.3.1 Measurement Equation

The RIME (Radio Interferometric Measurement Equation) we have used for this work consists of the $\mathbf{G}$ (the matrix for the gains of the antennae - a direction dependent effect) and $\mathbf{K}$ which represents the phase delay effect. For a single point source, the RIME is

$$\mathbf{V}_{pq} = \mathbf{G}_p \mathbf{K}_p \mathbf{B} \mathbf{K}_q^\dagger \mathbf{G}_q^\dagger \quad (6.2.10)$$

where $\mathbf{B}$ is the brightness matrix. For N_s sources, the RIME can be written as

$$\mathbf{V}_{pq} = \sum_{s=1}^{N_s} \mathbf{G}_{sp} \mathbf{K}_{sp} \mathbf{B} \mathbf{K}_{sq}^\dagger \mathbf{G}_{sq}^\dagger \quad (6.2.11)$$

In this simulation, we have assumed for simplicity that there is no direction-dependent effects, in which case, the RIME would have been

$$\mathbf{V}_{pq} = \sum_{s=1}^{N_s} \mathbf{G}_{sp} \mathbf{E}_{sp} \mathbf{K}_{sp} \mathbf{B} \mathbf{K}_{sq}^\dagger \mathbf{E}_{sq}^\dagger \mathbf{G}_{sq}^\dagger$$

where the term $\mathbf{E}$ encapsulates the direction-dependent effects. It could be, for example, the beam pattern.

6.2.3.2 Likelihood and Prior

Suppose we have the observed visibilities, $\mathbf{V}$ which are corrupted by the antennae gains. These visibilities are first calibrated using StEFCal, yielding $\mathbf{V}_c$. The model is given by $\mathbf{V}_t$. The likelihood of the problem is:

$$\mathcal{P}(v_c | \boldsymbol{\theta}, v_t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2} \left(\frac{v_c - v_t}{\sigma} \right)^2 \right] \quad (6.2.12)$$

where $\boldsymbol{\theta}$ are the parameters which we want to infer, and in our case, they are simply the flux densities of each point source. v_c and v_t refer to an element (corresponding to the same index) from the matrices $\mathbf{V}_c$ and $\mathbf{V}_t$ respectively. We assume a uniform prior on the flux densities. The details on the choice for the prior is explained in the procedures below.

6.2.3.3 Procedure

To test the calibration method and our Bayesian Inference technique, we generate visibilities for 10 point sources by modelling an East-West interferometer. These visibilities are corrupted by the gains of the antennae. Since we are using two antennae, that is, only one baseline, the gains are simply encoded in the vector $\mathbf{g}$ where $\mathbf{g} \in \mathbb{C}^2$. Hence, $\mathbf{G} \in \mathbb{C}^{2 \times 2}$ while $\mathbf{V} \in \mathbb{C}^{2 \times 2 \times N}$ where N is the number of “observed” visibilities. We use an observation time of 12 hours and the real and imaginary parts of the visibilities are also corrupted with Gaussian noise, $n \sim \mathcal{N}(0, 1)$.

In particular, we choose PyMultinest (Buchner et al. 2014) to find the distribution of the parameters for it is efficient when dealing with multi-modal distributions (Feroz et al. 2009). We also use the `scipy.optimize` (Jones et al. 2001) module to find the maximum likelihood estimators of the parameters. In particular, we choose Powell’s method (Powell 1964). The great advantage of this method is relevant in finding the local minima of complex continuous

function without requiring derivatives. This is useful even in the case where one does not have an underlying mathematical formulation.

Given that we now have θ_{MLE} , we choose uniform priors on all the parameters with width 0.4, centred on θ_{MLE} . Since we also have small values of the flux densities, we have to be careful when setting up the prior, that is, we cannot have priors with negative flux densities. Hence, in this case, the prior for the flux density is chosen to be $\theta \sim \mathcal{U}[0, \theta_{\text{MLE}} + 0.2]$. When running the sampler, we assume the positions of the sources as well as the noise distribution to be known.

6.2.3.4 Results and Discussions

As shown in Table (6.2.2), our results are very close to the actual true flux densities of the point sources. We define the bias, z as the number of σ away from the true value of the flux density, that is,

$$z = \frac{|F_{\text{true}} - F_{\text{nest}}|}{\sigma}$$

Name	F_{true}	α	δ	F_{nest}	z
$S_{(0)}$	3.00	03 ^h 20 ^m 00 ^s	60°01'00''	$2.943^{+0.027}_{-0.029}$	2.029
$S_{(1)}$	0.08	03 ^h 21 ^m 00 ^s	60°02'00''	$0.092^{+0.027}_{-0.028}$	0.445
$S_{(2)}$	0.15	03 ^h 22 ^m 00 ^s	60°03'00''	$0.139^{+0.026}_{-0.029}$	0.394
$S_{(3)}$	0.40	03 ^h 23 ^m 00 ^s	60°04'00''	$0.395^{+0.028}_{-0.028}$	0.170
$S_{(4)}$	0.55	03 ^h 24 ^m 00 ^s	60°05'00''	$0.574^{+0.027}_{-0.028}$	0.870
$S_{(5)}$	0.25	03 ^h 26 ^m 00 ^s	60°06'00''	$0.262^{+0.027}_{-0.028}$	0.431
$S_{(6)}$	0.75	03 ^h 27 ^m 00 ^s	60°07'00''	$0.785^{+0.027}_{-0.027}$	1.278
$S_{(7)}$	1.00	03 ^h 28 ^m 00 ^s	60°08'00''	$0.995^{+0.028}_{-0.027}$	0.185
$S_{(8)}$	1.60	03 ^h 29 ^m 00 ^s	60°09'00''	$1.581^{+0.027}_{-0.027}$	0.672
$S_{(9)}$	2.00	03 ^h 30 ^m 00 ^s	60°10'00''	$2.026^{+0.027}_{-0.026}$	0.985

Table 6.2.2 – Results from calibration and BIRO - The inferred flux densities (the 50th percentile, that is, the median) using PyMultinest are shown in the 5th column. These values all agree within 1 σ confidence interval except from $S_{(0)}$ and $S_{(6)}$ but are still within 3 σ confidence interval.

All our results agree to within 1 σ confidence interval except from $S_{(0)}$ and $S_{(6)}$ which agree

to 3σ confidence interval. This method of working directly in the visibility plane proves to be very powerful as we recover consistent flux densities of the sources within reasonable error bars.

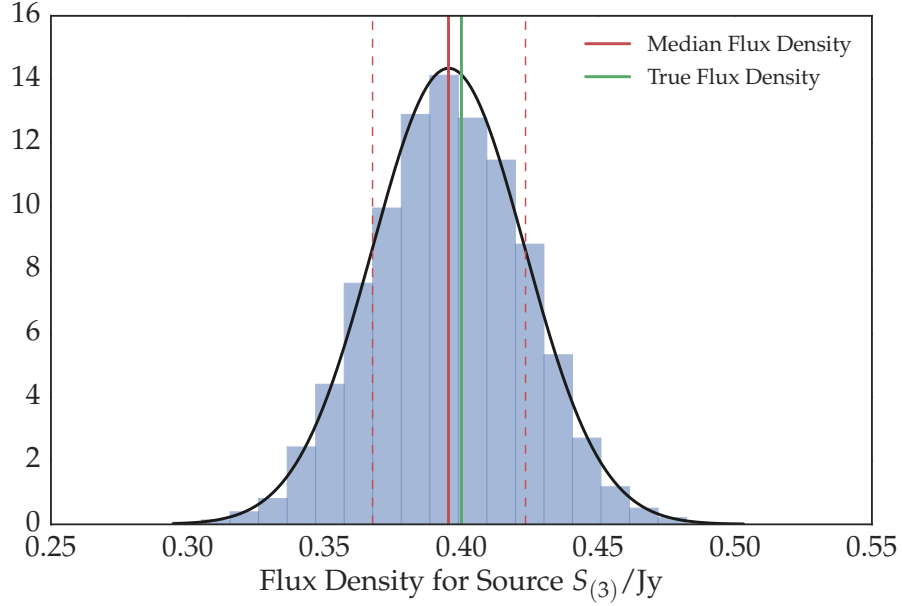


Figure 6.3 – The posterior distribution for the flux density of source $S_{(3)}$ - The red vertical line shows the median flux density while the green solid vertical line is the true flux density. Clearly, the result for this specific source agrees well with the true flux density as the bias is only 0.170.

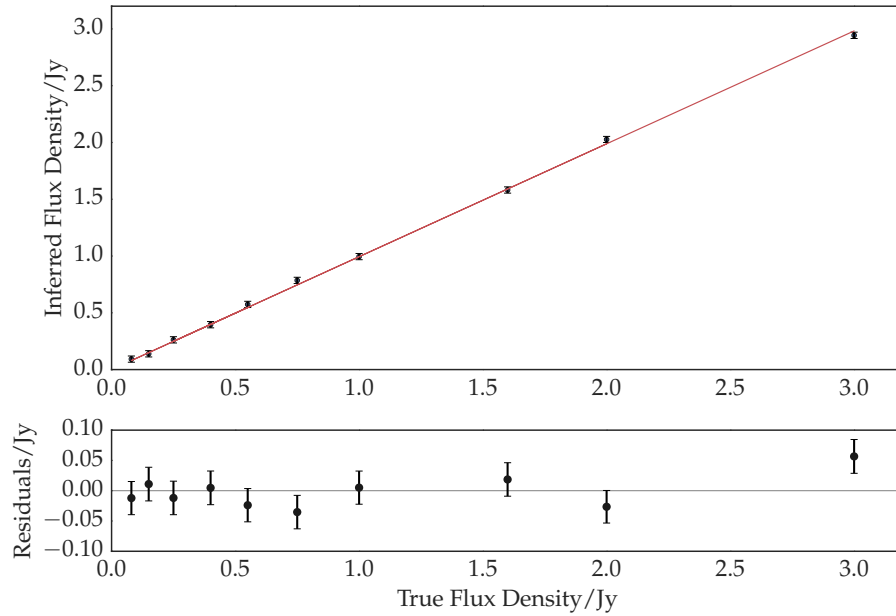


Figure 6.4 – Plot of the inferred flux densities with the true flux densities - The above plot shows the best-fit line to the data-points (the inferred flux densities using PyMultinest and the true flux densities). In particular, we fit the line of the model, $F_{\text{nest}} = mF_{\text{true}}$, the idea behind being that the y -intercept should, in principle, be zero. This also avoids the best-fit line to take negative values. We expect $m = 1$ and our method gives 0.994 ± 0.006 . This value is determined using the `scipy.optimize` module. The bottom plot shows the residuals (the difference between the true flux density and the inferred flux density).

As argued by [Lochner et al. \(2015\)](#), CLEAN performs well in the cases where the flux dens-

ities are high enough but fails in the presence of strong direction-dependent effect.

The Bayesian formalism naturally allows to first, find an estimate of the flux density by summarising the key statistics, for example, median or mean, second, provide a full posterior distribution for the flux density which then enable us to quote the flux density within a given confidence interval and third, the 2D posterior distribution allows us to learn about the correlation between parameters. One posterior distribution of the flux density for source $S_{(3)}$ is shown in Figure (6.3) and the correlation plot for all the inferred parameters is shown in Figure (6.5).

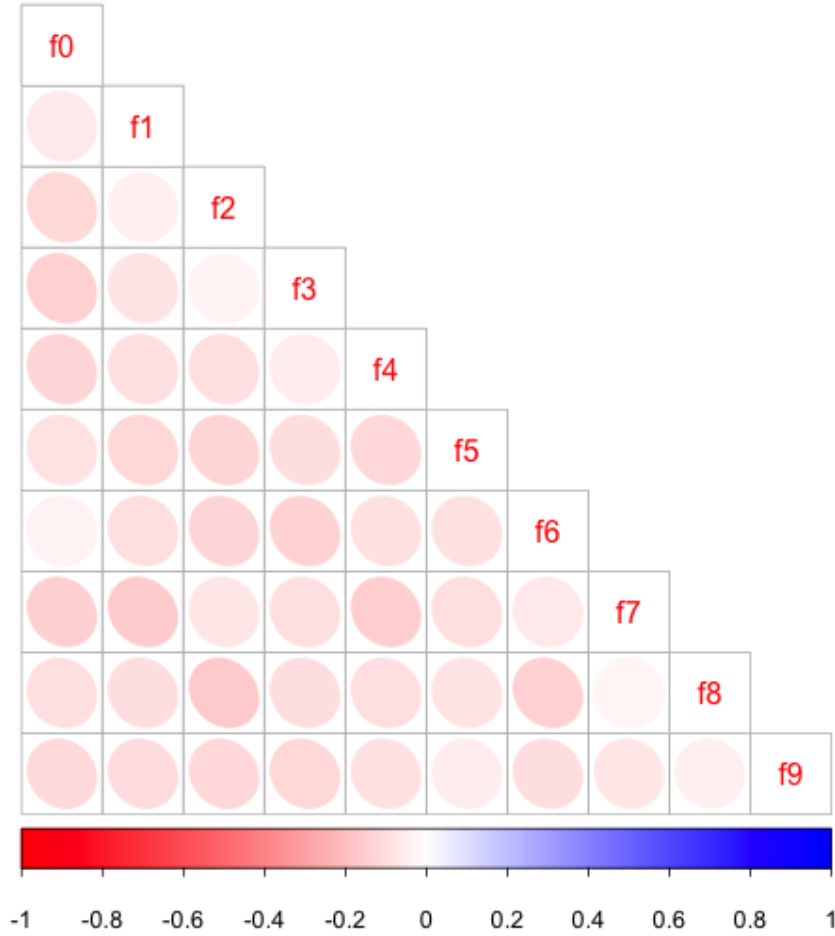


Figure 6.5 – Correlation plot of the inferred flux parameters - The plot shows the strength of correlation among the parameters, which is depicted by the colour and orientation of the ellipse. A positive correlation is indicated by a blue ellipse while a negative correlation by a red ellipse. In our case, the parameters are weakly correlated. However, if we had direction-dependent effect, for example, the beam, it should in principle introduce some degeneracies as shown in Figure (3.4) by [Lochner et al. \(2015\)](#).

Moreover, as seen from Figure (6.4), the Bayesian way of finding the flux densities is very reliable. In a nutshell, this work demonstrates the competence of BIRO. As [Lochner et al. \(2015\)](#) showed, one can even have additional terms in the RIME, for example, beam pattern along with

pointing errors. This illustrates the effectiveness of BIRO.

6.3 Super-Resolution with Bayesian Methods

One further interesting topic is to distinguish between a single point source versus a two-point source, each of intensity I , when they are on top of each other, but are buried in noise. If we were to produce an image, it is very likely that we would interpret it as being a single point source of intensity $2I$ as it is difficult to understand the noise. As pointed out by [Smirnov \(2011\)](#), reaching the level of the noise is the pinnacle of radio astronomy.

$\mathcal{D}_s$	Data for a single point source of intensity $2I$
$\mathcal{M}_t$	Model for two point sources each of intensity I
$\mathcal{M}_s$	Model for a single point source of intensity $2I$
Σ	Noise covariance matrix

Table 6.3.1 – Notation for work on super-resolution

In this section, we again do all analysis in the visibility plane using Bayesian Methods to attempt the above-mentioned problem. In particular, the idea is simple. The visibility pattern for a single point source of intensity $2I$ is the same as the visibility of two single point sources, each of intensity I at the same position on the sky. Given we have data (either $\mathcal{D}_s$ or $\mathcal{D}_t$), one can fit two possible models, either that of a single point source of intensity $2I$ or a Taylor series expansion of the visibility pattern for two single point sources each of intensity I . This will enable us to have an expression of the visibility pattern in terms of the distance of separation (δl and/or δm).

6.3.1 Likelihood and Prior

We assume Dirac delta priors over the parameters (intensity and positions of the source) because we assume they are known perfectly. In other words, the Bayesian Evidence is simply equal to the likelihood as we marginalise over all parameter space. Moreover, given the data $\mathcal{D}_s$ and assuming $\mathcal{M}_s$ is the correct model, we can write

$$\mathcal{D}_s = \mathcal{M}_s + \mathbf{n} \quad (6.3.1)$$

Then, the Bayesian Evidence, $\mathcal{P}(\mathcal{D}_s | \mathcal{M}_s)$ (in this case, the likelihood) is

$$\mathcal{P}(\mathcal{D}_s | \mathcal{M}_s) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp \left[-\frac{1}{2} (\mathcal{D}_s - \mathcal{M}_s)^\dagger \mathbf{\Sigma}^{-1} (\mathcal{D}_s - \mathcal{M}_s) \right]$$

$$\mathcal{P}(\mathcal{D}_s | \mathcal{M}_s) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp \left[-\frac{1}{2} \mathbf{n}^\dagger \mathbf{\Sigma}^{-1} \mathbf{n} \right] \quad (6.3.2)$$

Moreover, the model for the visibility for two point sources, in Taylor series expansion form can be written as

$$\mathcal{M}_t = \mathcal{M}_s + f(\delta l, \delta m) \quad (6.3.3)$$

and the Bayesian Evidence, $\mathcal{P}(\mathcal{D}_s | \mathcal{M}_t)$ is given by

$$\mathcal{P}(\mathcal{D}_s | \mathcal{M}_t) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp \left[-\frac{1}{2} (\mathcal{D}_s - \mathcal{M}_s - f)^\dagger \mathbf{\Sigma}^{-1} (\mathcal{D}_s - \mathcal{M}_s - f) \right]$$

$$\mathcal{P}(\mathcal{D}_s | \mathcal{M}_t) = \frac{1}{\sqrt{|2\pi\mathbf{\Sigma}|}} \exp \left[-\frac{1}{2} (\mathbf{n} - f)^\dagger \mathbf{\Sigma}^{-1} (\mathbf{n} - f) \right] \quad (6.3.4)$$

6.3.2 Expected log-Bayes Factor

Given that we now have two expressions for the Bayesian Evidence (as we considered two possible models), the log-Bayes Factor is given by

$$\log B_{st} = \log \left[\frac{\mathcal{P}(\mathcal{D}_s | \mathcal{M}_s)}{\mathcal{P}(\mathcal{D}_s | \mathcal{M}_t)} \right] = \frac{1}{2} \left[f^\dagger \mathbf{\Sigma}^{-1} f - 2f^\dagger \mathbf{\Sigma}^{-1} \mathbf{n} \right] \quad (6.3.5)$$

Then the expected log-Bayes Factor* is given by

$$\langle \log B_{st} \rangle = \int \mathcal{P}(\mathcal{D}_s | \mathcal{M}_s) \log \left[\frac{\mathcal{P}(\mathcal{D}_s | \mathcal{M}_s)}{\mathcal{P}(\mathcal{D}_s | \mathcal{M}_t)} \right] d\mathbf{n} \quad (6.3.6)$$

Therefore,

$$\langle \log B_{st} \rangle = \frac{1}{2\sqrt{|2\pi\mathbf{\Sigma}|}} \int \exp \left[-\frac{1}{2} \mathbf{n}^\dagger \mathbf{\Sigma}^{-1} \mathbf{n} \right] \left[f^\dagger \mathbf{\Sigma}^{-1} f - 2f^\dagger \mathbf{\Sigma}^{-1} \mathbf{n} \right] d\mathbf{n}$$

The above integration leads to

$$\langle \log B_{st} \rangle = \frac{1}{2} f^\dagger \mathbf{\Sigma}^{-1} f \quad (6.3.7)$$

*In this case, the log-Bayes Factor, as a function of δl , is obtained by marginalising over the noise. Since all other parameters are assumed to be known perfectly, we only have a single integration to do.

The complex visibilities are in fact accompanied by uncorrelated Gaussian noise in the real and imaginary parts (Smirnov 2011). Therefore, the noise covariance matrix, Σ consists of only diagonal entries, such that the expected log-Bayes Factor simplifies to

$$\langle \log B_{st} \rangle = \frac{1}{2} \begin{pmatrix} f_r & f_i \end{pmatrix} \begin{pmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_i^2 \end{pmatrix}^{-1} \begin{pmatrix} f_r \\ f_i \end{pmatrix}$$

and hence

$$\langle \log B_{st} \rangle = \frac{1}{2} \left[\left(\frac{f_r}{\sigma_r} \right)^2 + \left(\frac{f_i}{\sigma_i} \right)^2 \right] \quad (6.3.8)$$

which is at a single time step. In particular, taking all the samples into consideration, the expected log-Bayes Factor is:

$$\langle \log B_{st} \rangle = \frac{1}{2} \sum_T \left[\left(\frac{f_r}{\sigma_r} \right)^2 + \left(\frac{f_i}{\sigma_i} \right)^2 \right] \quad (6.3.9)$$

6.3.3 Analysis Without Beam

In the first case, we consider only the phase and the brightness matrix in the RIME, that is,

$$\mathbf{V}_{pq} = \mathbf{K}_p \mathbf{B} \mathbf{K}_q^\dagger \quad (6.3.10)$$

such that for an unpolarised source of intensity $2I$ at (l_0, m_0) , the visibility is simply given by

$$v_{pq} = 2I e^{-2\pi j(u_{pq}l_0 + v_{pq}m_0)} \quad (6.3.11)$$

whereas for two single point sources at (l_0, m_0) and $(l_0 + \delta l, m_0 + \delta m)$ each of intensity I , the visibility is

$$v_{pq} = 2I e^{-2\pi j(u_{pq}l_0 + v_{pq}m_0)} + f(\delta l, \delta m) \quad (6.3.12)$$

where

$$f(\delta l, \delta m) \approx \delta l \frac{\partial}{\partial l} \left(I e^{-2\pi j(u_{pq}l + v_{pq}m)} \right) \Big|_{(l_0, m_0)} + \delta m \frac{\partial}{\partial m} \left(I e^{-2\pi j(u_{pq}l + v_{pq}m)} \right) \Big|_{(l_0, m_0)}$$

Hence,

$$f(\delta l, \delta m) = -2\pi (u_{pq}\delta l + v_{pq}\delta m) j \mathbf{I} e^{-2\pi j(u_{pq}l_0 + v_{pq}m_0)} \quad (6.3.13)$$

In our analysis, for simplicity, we will assume that $\delta m = 0$ as our aim is to see how the log-Bayes Factor changes as a function of δl . Therefore,

$$\Re[f(\delta l, \delta m = 0)] = -2\pi u_{pq} \mathbf{I} \delta l \sin[2\pi(u_{pq}l_0 + v_{pq}m_0)]$$

$$\Im[f(\delta l, \delta m = 0)] = -2\pi u_{pq} \mathbf{I} \delta l \cos[2\pi(u_{pq}l_0 + v_{pq}m_0)]$$

Under the assumption that $\sigma_r = \sigma_i = \sigma$ and using Equation (6.3.9),

$$\langle \log B_{st} \rangle = 2\pi^2 \delta l^2 \left(\frac{\mathbf{I}}{\sigma} \right)^2 \sum_{\mathbf{T}} u_{pq}^2 \quad (6.3.14)$$

where $\mathbf{I}/\sigma$ is the signal-to-noise ratio. For an East-West interferometer, $u_{pq} = |\mathbf{b}_\lambda| \cos H$, where H is the hour angle. Therefore,

$$\langle \log B_{st} \rangle = 2\pi^2 \delta l^2 \left(\frac{\mathbf{I}}{\sigma} \right)^2 |\mathbf{b}_\lambda|^2 \sum_{\mathbf{T}} \cos^2 H$$

Moreover, the resolution of an interferometer is given by

$$\theta_{\text{res}} \approx \frac{\lambda}{\mathbf{b}_{\text{max}}}$$

Therefore,

$$\langle \log B_{st} \rangle = \frac{2\pi^2 \delta l^2 (\mathbf{I}/\sigma)^2}{\theta_{\text{res}}^2} \sum_{\mathbf{T}} \cos^2 H \quad (6.3.15)$$

Using this equation, for a given beam resolution, θ_{res} and signal-to-noise ($\text{SNR} = \frac{\mathbf{I}}{\sigma}$), we can compute the minimum angular separation, denoted δl^* , for which we have strong evidence for the one point source model versus the two point source model, which we define to occur when $\langle \log B_{st} \rangle = 5$. Therefore,

$$\frac{2\pi^2 \delta l^2 (\mathbf{I}/\sigma)^2}{\theta_{\text{res}}^2} \sum_{\mathbf{T}} \cos^2 H = 5$$

and hence,

$$\delta l^* = \frac{\theta_{\text{res}}}{\pi (1/\sigma)} \sqrt{\frac{5}{2 \sum_T \cos^2 H}} \quad (6.3.16)$$

An interesting insight from this is that as $\sigma \rightarrow 0$ ($\text{SNR} \rightarrow \infty$), we can resolve two point sources at arbitrarily small angular separations in this simple toy model. To illustrate our proposal, we simulate an East-West interferometer with a single baseline. The observing frequency (ν) is, 1400 MHz and the baseline length is 144 m. Moreover, $1/\sigma = 3.0$ (we simply set $\sigma = 1.0$ and $I = 3.0$). The field centre is at $\alpha_0 = 03^{\text{h}}25^{\text{m}}00^{\text{s}}$ and $\delta_0 = 60^{\circ}00'00''$ while the single point source is located at $\alpha = 03^{\text{h}}25^{\text{m}}00^{\text{s}}$ and $\delta = 60^{\circ}04'00''$.

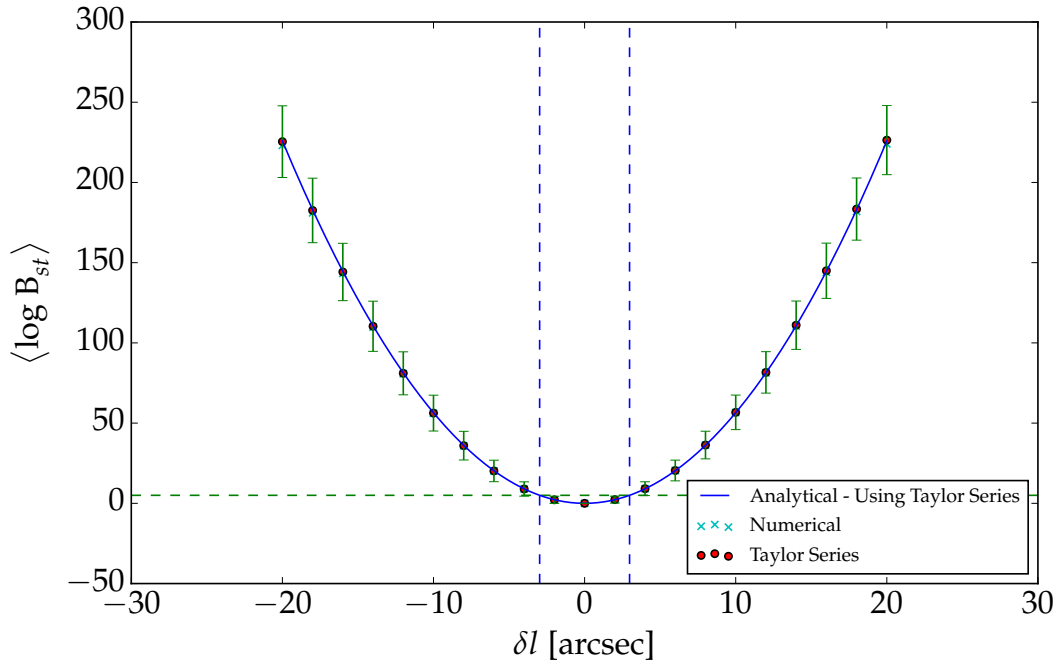


Figure 6.6 – Plot of the log-Bayes Factor against δl in arcsec - The scatter plot in red shows the expected log-Bayes Factor with given error bars. The blue crosses show the log-Bayes Factor if instead we were to be doing the analysis numerically. This also shows that the Taylor expansion works well for $\delta l \lesssim 40$ arcsec. The blue curve shows the “analytical” solution, that is, Equation (6.3.9).

Given this simulated data, we find that

$$\delta l^* = 2.98 \text{ arcsec} \quad (6.3.17)$$

Despite the fact that the resolution of this interferometer is roughly 307 arcsec, we are able to distinguish between a single point source model and a two point source model.

6.3.4 Analysis with WSRT Beam

In this case, we consider an analytic beam which we incorporate in the RIME. We simulate the WSRT beam which is given by

$$\cos^3 \left(cv \sqrt{l^2 + m^2} \right)$$

where c is the beam factor in GHz^{-1} , ν is the observing frequency in GHz, $r = \sqrt{l^2 + m^2}$ is the distance from the pointing centre in degrees (Lochner et al. 2015, Popping & Braun 2008). The RIME can further be written as

$$\mathbf{V}_{pq} = \mathbf{E}_p \mathbf{K}_p \mathbf{B} \mathbf{K}_q^\dagger \mathbf{E}_q^\dagger \quad (6.3.18)$$

where $\mathbf{E}$ now denotes the direction-dependent effect, in this case, the beam. Therefore, for an unpolarised single point source of intensity $2I$ at (l_0, m_0) , the visibility can be written as

$$v_{pq} = 2I e^{-2\pi j(u_{pq}l_0 + v_{pq}m_0)} \cos^6 \left(cv \sqrt{l_0^2 + m_0^2} \right) \quad (6.3.19)$$

On the other hand, for two single point sources, each of intensity I at $(l_0 + \delta l, m_0 + \delta m)$ and (l_0, m_0) , the visibility is

$$v_{pq} = 2I e^{-2\pi j(u_{pq}l_0 + v_{pq}m_0)} \cos^6 \left(cv \sqrt{l_0^2 + m_0^2} \right) + f(\delta l, \delta m) \quad (6.3.20)$$

where

$$f(\delta l, \delta m) \approx \delta l \frac{\partial}{\partial l} \left[I e^{-2\pi j(u_{pq}l + v_{pq}m)} \cos^6 \left(cv \sqrt{l^2 + m^2} \right) \right] \Big|_{(l_0, m_0)} + \\ \delta m \frac{\partial}{\partial m} \left[I e^{-2\pi j(u_{pq}l + v_{pq}m)} \cos^6 \left(cv \sqrt{l^2 + m^2} \right) \right] \Big|_{(l_0, m_0)}$$

If

$$R_1 = \frac{6cvl_0}{\sqrt{l_0^2 + m_0^2}} \sin \left(cv \sqrt{l_0^2 + m_0^2} \right) \cos \left(2\pi (u_{pq}l_0 + v_{pq}m_0) \right)$$

$$R_2 = 2\pi u_{pq} \cos \left(cv \sqrt{l_0^2 + m_0^2} \right) \sin \left(2\pi (u_{pq}l_0 + v_{pq}m_0) \right)$$

$$I_1 = \frac{-6cvl_0}{\sqrt{l_0^2 + m_0^2}} \sin \left(cv \sqrt{l_0^2 + m_0^2} \right) \sin \left(2\pi (u_{pq}l_0 + v_{pq}m_0) \right)$$

$$I_2 = 2\pi u_{pq} \cos \left(cv \sqrt{l_0^2 + m_0^2} \right) \cos \left(2\pi (u_{pq}l_0 + v_{pq}m_0) \right)$$

then,

$$\Re [f(\delta l, \delta m = 0)] = -I \cos^5 \left(cv \sqrt{l_0^2 + m_0^2} \right) (R_1 + R_2) \delta l$$

$$\Im [f(\delta l, \delta m = 0)] = -I \cos^5 \left(cv \sqrt{l_0^2 + m_0^2} \right) (I_1 + I_2) \delta l$$

From Equation (6.3.9),

$$\langle \log B_{st} \rangle = \frac{1}{2} \left(\frac{I}{\sigma} \right)^2 \delta l^2 \cos^{10} \left(cv \sqrt{l_0^2 + m_0^2} \right) \sum_T \left[(R_1 + R_2)^2 + (I_1 + I_2)^2 \right] \quad (6.3.21)$$

Therefore, we again have an equation for the $\langle \log B_{st} \rangle$ as a function of δl . In this case, it is a little bit more difficult to obtain the expression in terms of the resolution, θ_{res} . We again compute the minimum angular separation, as in the previous section.

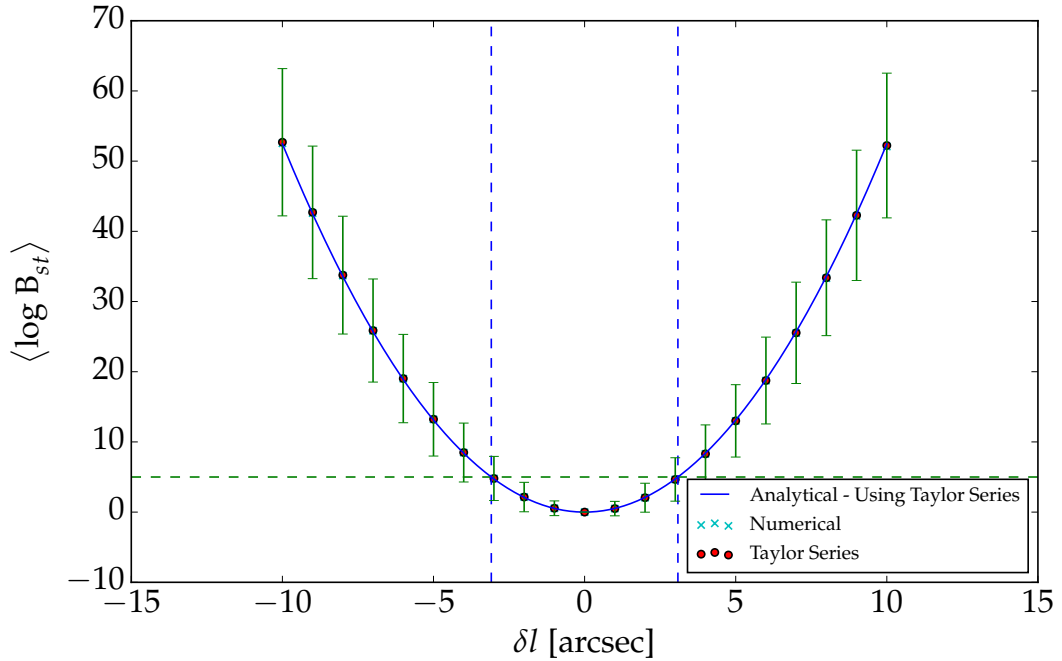


Figure 6.7 – Plot of the log-Bayes Factor against δl in arcsec (with beam) - The curve in blue shows the analytical solution when taking the beam into consideration. The scattered red points, along with the given error bars indicate the log-Bayes factor for a given δl . These are determined by many realisations of data. The cyan cross shows the expected log-Bayes Factor if we were to do this analysis numerically.

Therefore,

$$\frac{1}{2} \left(\frac{I}{\sigma} \right)^2 \delta l^2 \cos^{10} \left(cv \sqrt{l_0^2 + m_0^2} \right) \sum_T \left[(R_1 + R_2)^2 + (I_1 + I_2)^2 \right] = 5$$

and hence,

$$\delta l^* = \frac{\sqrt{10}}{\left(\frac{I}{\sigma}\right) \cos^5 \left(cv \sqrt{l_0^2 + m_0^2} \right) \sqrt{\sum_T \left[(R_1 + R_2)^2 + (I_1 + I_2)^2 \right]}} \quad (6.3.22)$$

We simulate an East-West interferometer with a single baseline as in the previous example. The observing frequency (ν) is 1400 MHz for a beam-factor of 68 (refer to [WSRT Observation Guides](#)[†]). The baseline length is 144 m. Moreover, $I/\sigma = 3.0$ (we simply set $\sigma = 1.0$ and $I = 3.0$). The field centre is at $\alpha_0 = 03^{\text{h}}25^{\text{m}}00^{\text{s}}$ and $\delta_0 = 60^{\circ}00'00''$ while the single point source is located at $\alpha = 03^{\text{h}}25^{\text{m}}00^{\text{s}}$ and $\delta = 60^{\circ}04'00''$. Given this simulated data, we find that

$$\delta l^* = 3.09 \text{ arcsec} \quad (6.3.23)$$

One possible problem is the term

$$\frac{6cvl_0}{\sqrt{l_0^2 + m_0^2}} \sin \left(cv \sqrt{l_0^2 + m_0^2} \right)$$

because as $l_0, m_0 \rightarrow 0$, the denominator will be equal to zero. One way to deal with this is to move to, for example, polar coordinates. Let $l_0 = r_0 \cos\theta$ and $m_0 = r_0 \sin\theta$. Then,

$$\begin{aligned} \frac{6cvl_0}{\sqrt{l_0^2 + m_0^2}} \sin \left(cv \sqrt{l_0^2 + m_0^2} \right) &= \sin\theta \sin (cvr_0) \\ \lim_{l_0, m_0 \rightarrow 0} \left[\frac{6cvl_0}{\sqrt{l_0^2 + m_0^2}} \sin \left(cv \sqrt{l_0^2 + m_0^2} \right) \right] &= \lim_{r_0 \rightarrow 0} [\sin\theta \sin (cvr_0)] = 0 \end{aligned}$$

Therefore, we also have

$$f(\delta l, \delta m = 0) = -2\pi u_{pq} I \delta l_j$$

Notice that, this is similar to treating the problem without the beam (which we already did when we effectively neglected the beam term) at $(l_0 = 0, m_0 = 0)$. As a test, we have a single point source nearly at the phase centre such that $l_0 = 0$ and $m_0 = 4.85 \times 10^{-6}$. In both cases (with and without beam at the same observing frequency), the value of δl^* agrees to the 5th decimal place in either case, hence effectively showing that the beam term has little effect if included correctly.

[†] <http://www.astron.nl/radio-observatory/astronomers/wsrt-guide-observations/wsrt-guide-observations>

6.3.5 Discussion

One of the key questions in radio astronomy is to distinguish between a point source and a two-point source. In this section, using Bayesian methods and a toy model, we have shown that we can find the minimum angular separation, in favour of single point source. We have treated two different scenarios, the first one where we assume no direction-dependent effect while in the second case where we do. Working with the beam is slightly more mathematically tricky but does provide a sound estimate of δl^* .

In the first case, from Equation (6.3.16), for sufficient high SNR, it is possible to resolve two-point sources at small angular separations. Moreover, as we would naively expect, we also see that prediction of the angular distance separation, δl^* improves as the resolution of the interferometer gets better. In the second case (see Equation 6.3.22) where we have considered an analytic expression for the beam, the conclusion drawn is nearly the same as in the first part.

CONCLUSIONS

"The plain fact is that there are no conclusions. If we must state a conclusion, it would be that many of the former conclusions of the nineteenth-century science on philosophical questions are once again in the melting-pot." - James Hopwood Jeans

Starting from the first ever amateur telescope to the most advanced telescope available today, the volumes of data have grown from terabytes to petabytes and are expected to reach the exascale regime by the time when the SKA will be fully operational. This exponential rise in the volume of the data sounds the alarm. At the forefront of statistical analysis, no matter how big the data is, it is of utmost important that we have the right tool for careful data analysis.

In this work we have used the Savage-Dickey Density Ratio (SDDR) to show that we can calculate the Bayes Factor of two non-nested models by introducing a new hyper-parameter that combines the models into a single supermodel. This Savage-Dickey Supermodel (SDSM) method does not need the Bayesian evidence (Marginal Likelihood) to be computed. The core supermodel embedding can be done either at the level of the model (Equation 4.2.15) or at the level of the likelihood (Equation 4.2.11) and effectively makes the the models nested and hence amenable to the SDDR approach to computing the Bayes Factors. In the context of Gaussian linear models we show that the SDDR both analytically and numerically reproduces the Bayes Factors computed analytically. We then consider a nonlinear example and show that our supermodel approach agrees well with that from nested sampling.

Though we have a clever way of avoiding multidimensional integrals to calculate the evidence, this new method requires very efficient sampling and for a small number of dimensions is not faster than individual nested sampling runs. The major reason for this is that we require

independent samples for α and one way to ensure we are doing so is to have a short autocorrelation length. Hence the thinning factor for the MCMC chain needs to be adjusted as well as the number of the steps, especially for large log-Bayes Factor. However, generically the scaling of MCMC methods with the number of dimensions is much more benign than the scaling of nested sampling methods. The approach presented here is thus expected to work also for very high numbers of dimensions where codes such as multineest fail. Additionally, if we only keep in a MCMC chain the elements for which $\alpha = 1$ or $\alpha = 0$ then we obtain a model-averaged posterior. For this application we do not need a very large number of samples, so that the method is competitive with nested sampling for model averaged posteriors also at a smaller number of dimensions.

For future work we note that other, nonlinear, combinations of models/likelihoods are also possible. For example, consider product combined model and likelihood $\mathcal{M}_S = \mathcal{M}_1^\alpha \mathcal{M}_2^{(1-\alpha)}$ and $\mathcal{L}_S = \mathcal{L}_1^\alpha \mathcal{L}_2^{(1-\alpha)}$ in which case, the general condition (4.2.4) still holds for $\alpha \in [0, 1]$.

Such nonlinear supermodels, choices of reparametrisation function $f(\alpha)$ or other innovations (such as using simulated annealing or Gibbs sampling) may greatly simplify some aspects of the sampling and provide a clever way of not only obtaining the log-Bayes Factor, which helps us to understand the relative strength of the models but also to have model averaged posteriors of all the parameters in both models. Study of these generalisations is left to future work.

In Chapter (6), we illustrate the idea behind calibration, which is simply an optimisation problem to solve for the gains of the antennae using StEFCal. In our case, we assumed that these gains are dominated by direction-independent effects, implying they are fixed quantities which need to be optimised depending on the model of the sky. This is followed by the implementation of a Bayesian sampler to reliably recover the full distribution of the flux densities of the radio sources for the calibrated data. However, the idea of knowing the sky model is still a limitation and cries out for more work in this field.

We also use the Bayesian Evidence (and hence the Bayes Factor) to compare two different models (a single point source versus a two-point source). In other words, would it be possible to do a super-resolution analysis in the visibility domain using a Bayesian framework? The answer is yes. We show that we can calculate the minimum angular separation for which there is strong evidence of a single point source model. From our analysis, an interesting observation, which we would expect, is the fact that as the SNR $\rightarrow \infty$, we will be able to resolve two-point-sources at very small angular separation. In addition to this, if the resolving power

θ_{res} of the interferometer gets better, the ability to resolve two point sources at very small angular resolution also gets better. These conclusions are evident from our analysis (see Equation 6.3.16).

The work presented in this thesis could possibly be useful in various aspects of statistical data analysis. In the first case, the motivation is to extend this formalism to models with higher dimensions, and hence apply this concept to implement a quick way to do model selection for the BIRO project, for example, comparing a sky with 1001 sources compared to 1000 in the visibility domain. This simple example might have thousands of parameters which include fluxes, positions and others. Further, while the forthcoming SKA and current pathfinders will map the sky which consists of millions of sources, it is important that we do not misinterpret a source as being two-point source or vice versa. The second part of this thesis provides an in depth statistical analysis of how we could presumably do that. This formalism can easily be extended to any number of baselines in radio astronomy.

TOPICS RELATED TO BAYESIAN METHODS

A.1 The Gelman-Rubin Test

The Gelman-Rubin test is commonly used to assess the convergence of an MCMC chain. Below is a brief description of the Gelman-Rubin convergence test:

- (1) Typically, m chains are run where $m \geq 2$, each of length equal to $2n$.
- (2) The first n samples recorded are discarded (burn-in).
- (3) The within-chain, W and between-chain, B variance are calculated.
- (4) The estimated variance of the parameter is calculated as a weighted sum of the within-chain and between-chain variance.
- (5) Finally, the potential scale reduction factor, which is the ratio between the variance of the means, V and the within-chain variance, W ([Gelman & Rubin 1992](#)), is calculated.

The above is repeated for each parameter involved in the MCMC. Let the indices i and j represent the i^{th} sample and j^{th} chain respectively. The within-chain variance is then calculated as follows. The variance in the j^{th} chain is

$$\sigma_j^2 = \frac{1}{n-1} \sum_{i=1}^n (\theta_{ij} - \bar{\theta}_j)^2 \quad (\text{A.1.1})$$

where $\bar{\theta}_j = \frac{1}{n} \sum_i \theta_{ij}$ and the quantity W is given by

$$W = \frac{1}{m} \sum_{j=1}^m \sigma_j^2 \quad (\text{A.1.2})$$

The quantity W is most likely to be an underestimate because the chains may not have reached the stationary distribution. The between-chain variance is calculated as

$$B = \frac{n}{m-1} \sum_{j=1}^m (\bar{\theta}_j - \bar{\zeta})^2 \quad (\text{A.1.3})$$

where

$$\bar{\zeta} = \frac{1}{m} \sum_{j=1}^m \bar{\theta}_j \quad (\text{A.1.4})$$

The weighted estimate of the variance is then calculated as

$$\hat{\sigma}^2 = \left(\frac{n-1}{n} \right) W + \frac{B}{n} \quad (\text{A.1.5})$$

An estimator for the variance of the means is given by

$$V = \hat{\sigma}^2 + \frac{B}{mn} \quad (\text{A.1.6})$$

While V is an overestimate and W is an underestimate, the ratio between these two yields

$$\hat{R} = \left(\frac{n-1}{n} \right) + \left(\frac{m+1}{mn} \right) \frac{B}{W} \quad (\text{A.1.7})$$

where $\sqrt{\hat{R}}$ is referred to as the potential scale reduction factor and will always be ≥ 1 . If $\sqrt{\hat{R}} \approx 1.01$, we may assume we have a converged chain. Otherwise, more samples are recorded until the convergence criterion is met.

A.2 The SDDR

The Savage-Dickey Density Ratio (SDDR) provides an alternative way of calculating the Bayes Factor without calculating the Bayesian Evidence, provided the models are nested. Consider two competing models $\mathcal{M}_0$ and $\mathcal{M}_1$ where $\mathcal{M}_0$ have parameters $\theta_0 = (\psi, \phi_0)$ while $\mathcal{M}_1$ have parameters $\theta_1 = (\psi, \phi)$. Note that $\mathcal{M}_0$ is nested in $\mathcal{M}_1$ at $\phi = \phi_0$. The probability of model $\mathcal{M}_i$ given the data ($\mathcal{D}$) is given by

$$\mathcal{P}(\mathcal{M}_i | \mathcal{D}) = \frac{\mathcal{P}(\mathcal{D} | \mathcal{M}_i) \mathcal{P}(\mathcal{M}_i)}{\sum_i \mathcal{P}(\mathcal{D} | \mathcal{M}_i) \mathcal{P}(\mathcal{M}_i)} \quad (\text{A.2.1})$$

where $i = 0, 1$ in Equation (A.2.1). Since the denominator is a constant of proportionality, we

can write

$$\frac{\mathcal{P}(\mathcal{M}_0|\mathcal{D})}{\mathcal{P}(\mathcal{M}_1|\mathcal{D})} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_0)}{\mathcal{P}(\mathcal{D}|\mathcal{M}_1)} \frac{\mathcal{P}(\mathcal{M}_0)}{\mathcal{P}(\mathcal{M}_1)} \quad (\text{A.2.2})$$

Assuming non-committal priors of the models, that is $\mathcal{P}(\mathcal{M}_0) = \mathcal{P}(\mathcal{M}_1)$, the Bayes Factor is given by

$$B_{01} = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M}_0)}{\mathcal{P}(\mathcal{D}|\mathcal{M}_1)} = \frac{\mathcal{P}(\mathcal{M}_0|\mathcal{D})}{\mathcal{P}(\mathcal{M}_1|\mathcal{D})} \quad (\text{A.2.3})$$

$$\frac{\mathcal{P}(\mathcal{M}_0|\mathcal{D})}{\mathcal{P}(\mathcal{M}_1|\mathcal{D})} = \frac{\int d\psi \mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_0) \mathcal{P}(\psi|\mathcal{M}_0)}{\int \int d\psi d\phi \mathcal{P}(\mathcal{D}|\psi, \phi, \mathcal{M}_1) \mathcal{P}(\psi, \phi|\mathcal{M}_1)} \quad (\text{A.2.4})$$

where $\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_0)$ and $\mathcal{P}(\psi|\mathcal{M}_0)$ are the likelihood and prior under model $\mathcal{M}_0$ respectively while $\mathcal{P}(\mathcal{D}|\psi, \phi, \mathcal{M}_1)$ and $\mathcal{P}(\psi, \phi|\mathcal{M}_1)$ are the likelihood and prior under model $\mathcal{M}_1$ respectively. Since, $\mathcal{M}_0$ is nested in $\mathcal{M}_1$, the likelihood in $\mathcal{M}_0$ will simply be a slice of the likelihood surface in $\mathcal{M}_1$ at $\phi = \phi_0$.

Let $q = \int \int d\psi d\phi \mathcal{P}(\mathcal{D}|\psi, \phi, \mathcal{M}_1) \mathcal{P}(\psi, \phi|\mathcal{M}_1)$. We first multiply and divide Equation (A.2.4) by the number $\mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) = \mathcal{P}(\phi_0|\mathcal{D})$ which represents the marginalised posterior of the parameter ϕ evaluated at ϕ_0 is model $\mathcal{M}_1$.

$$B_{01} = \mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) \int d\psi \frac{\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_0) \mathcal{P}(\psi|\mathcal{M}_0)}{\mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) \cdot q} \quad (\text{A.2.5})$$

Using

$$\mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) = \frac{\mathcal{P}(\psi, \phi_0|\mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\psi|\phi_0, \mathcal{D}, \mathcal{M}_1)}$$

Equation (A.2.5) can be re-written as

$$B_{01} = \mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) \int d\psi \frac{\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_0) \mathcal{P}(\psi|\mathcal{M}_0)}{q} \frac{\mathcal{P}(\psi|\phi_0, \mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\psi, \phi_0|\mathcal{D}, \mathcal{M}_1)} \quad (\text{A.2.6})$$

However,

$$\mathcal{P}(\psi, \phi_0|\mathcal{D}, \mathcal{M}_1) = \frac{\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_1) \mathcal{P}(\psi, \phi_0|\mathcal{M}_1)}{q}$$

and therefore,

$$B_{01} = \mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) \int d\psi \frac{\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_0) \mathcal{P}(\psi|\mathcal{M}_0) \mathcal{P}(\psi|\phi_0, \mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\mathcal{D}|\psi, \phi_0, \mathcal{M}_1) \mathcal{P}(\psi, \phi_0|\mathcal{M}_1)} \quad (\text{A.2.7})$$

Under the assumption that $\mathcal{P}(\psi|\phi_0, \mathcal{M}_1) = \mathcal{P}(\psi|\mathcal{M}_0)$, which holds in the case where the priors are separable,

$$B_{01} = \mathcal{P}(\phi_0|\mathcal{D}, \mathcal{M}_1) \int d\psi \frac{\mathcal{P}(\psi|\mathcal{M}_0) \mathcal{P}(\psi|\phi_0, \mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\psi|\mathcal{M}_1) \mathcal{P}(\phi_0|\mathcal{M}_1)} \quad (\text{A.2.8})$$

$$B_{01} = \left. \frac{\mathcal{P}(\phi|\mathcal{D}, \mathcal{M}_1)}{\mathcal{P}(\phi|\mathcal{M}_1)} \right|_{\phi=\phi_0} \quad (\text{A.2.9})$$

In Equation (A.2.8), $\mathcal{P}(\psi|\mathcal{D}, \phi_0, \mathcal{M}_1)$ is the normalised marginal posterior for ψ and hence we recover the SDDR, given in Equation (A.2.9).

A.3 Deriving the Bayesian Evidence

In this section, we derive the Bayesian Evidence for the full extended model which we consider, the polynomial of the form $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$. In matrix format, we can write the prior as

$$\mathcal{P}(\boldsymbol{\theta}|\mathcal{M}) = \frac{1}{\sqrt{|2\pi\mathbf{P}^{-1}|}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{P}^{-1}\boldsymbol{\theta}\right) \quad (\text{A.3.1})$$

where $\mathbf{P}^{-1}$ is the inverse of the covariance matrix for the priors. The likelihood is given by

$$\mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) = \frac{1}{\prod_i \sqrt{(2\pi\sigma_i^2)}} \exp\left[-\frac{1}{2}(\mathbf{b} - \mathbf{D}\boldsymbol{\theta})^T(\mathbf{b} - \mathbf{D}\boldsymbol{\theta})\right] \quad (\text{A.3.2})$$

where $\mathbf{b}$ is the vector $\left(\frac{y_0}{\sigma_0}, \frac{y_1}{\sigma_1}, \dots, \frac{y_{N-1}}{\sigma_{N-1}}\right)$ and $\mathbf{D}$ is the design matrix. The Bayesian Evidence, $\mathbb{Z}$ is then given by

$$\mathbb{Z} = \int \mathcal{P}(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}) \mathcal{P}(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$$

$$\mathbb{Z} = \frac{1}{\prod_i \sqrt{(2\pi\sigma_i^2)}} \frac{\exp\left(-\frac{1}{2}\mathbf{b}^T\mathbf{b}\right)}{|2\pi\mathbf{C}^{-1}|} \int \exp\left[-\frac{1}{2}\left\{\boldsymbol{\theta}^T(\mathbf{D}^T\mathbf{D} + \mathbf{P}^{-1})\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{D}^T\mathbf{b}\right\}\right] d\boldsymbol{\theta}$$

If we have a quadratic expression such as $\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{b} + \mathbf{c}$, then this can be expressed as

$$(\mathbf{x} - \mathbf{h})^T \mathbf{A} (\mathbf{x} - \mathbf{h}) + \mathbf{k}$$

where

$$\mathbf{h} = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{h} \quad \mathbf{k} = \mathbf{c} - \frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}$$

Therefore,

$$\mathbb{Z} = \frac{1}{\prod_i \sqrt{(2\pi\sigma_i^2)}} \frac{\exp\left[-\frac{1}{2}(\mathbf{k} + \mathbf{b}^T\mathbf{b})\right]}{|2\pi\mathbf{P}^{-1}|} \int \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{h})^T (\mathbf{D}^T\mathbf{D} + \mathbf{P}^{-1}) (\boldsymbol{\theta} - \mathbf{h})\right] d\boldsymbol{\theta}$$

where

$$\mathbf{k} = -(\mathbf{D}^T\mathbf{b})^T (\mathbf{D}^T\mathbf{D} + \mathbf{P}^{-1})^{-1} (\mathbf{D}^T\mathbf{b}) \quad \mathbf{h} = (\mathbf{D}^T\mathbf{D} + \mathbf{P}^{-1})^{-1} (\mathbf{D}^T\mathbf{b})$$

Hence,

$$\mathbb{Z} = \frac{\exp\left[-\frac{1}{2}(\mathbf{k} + \mathbf{b}^T\mathbf{b})\right]}{\prod_i \sqrt{(2\pi\sigma_i^2)}} \sqrt{\frac{|2\pi(\mathbf{D}^T\mathbf{D} + \mathbf{P}^{-1})^{-1}|}{|2\pi\mathbf{P}^{-1}|}} \quad (\text{A.3.3})$$

A.4 Gibbs Sampling

In this section, we briefly describe how Gibbs sampling works. Gibbs sampling requires that we know the full conditional distribution of the parameters. To illustrate this, we consider here a simple Bayesian Linear Regression, for which we find the full posterior distribution of the two parameters - gradient and y -intercept. The data is available on [Github](https://github.com/Harry45/Self-Taught/tree/master/Gibbs_Sampling)*. Suppose we have the datapoints, $\mathcal{D} = \{x_i, y_i\}$ for $i = 1, 2, \dots, N$, generated from the model $y = \theta_0 + \theta_1 x$. In other words,

$$y = \theta_0 + \theta_1 x + \epsilon$$

where $\epsilon \sim \mathcal{N}(0.0, \sigma_n^2)$. Our aim is to find the full posterior distributions of the parameters θ_0 and θ_1 . Following Bayes' theorem, the joint posterior distribution is simply

$$\mathcal{P}(\theta_0, \theta_1 | \mathcal{D}) \propto \mathcal{P}(\mathcal{D} | \theta_0, \theta_1) \mathcal{P}(\theta_0, \theta_1)$$

* https://github.com/Harry45/Self-Taught/tree/master/Gibbs_Sampling

where we assume factorisable Gaussian priors, that is, $\mathcal{P}(\theta_0, \theta_1) = \mathcal{P}(\theta_0) \mathcal{P}(\theta_1)$:

$$\mathcal{P}(\theta_0) \sim \mathcal{N}(\mu_0, \Sigma_0^2) \quad \mathcal{P}(\theta_1) \sim \mathcal{N}(\mu_1, \Sigma_1^2)$$

We first define the design matrices, $\mathbf{D}_0$, $\mathbf{D}_1$ and the vector $\mathbf{b}$ as follows:

$$\mathbf{D}_0 = \begin{bmatrix} \frac{1}{\sigma_1} \\ \frac{1}{\sigma_2} \\ \vdots \\ \frac{1}{\sigma_N} \end{bmatrix} \quad \mathbf{D}_1 = \begin{bmatrix} \frac{x_1}{\sigma_1} \\ \frac{x_2}{\sigma_2} \\ \vdots \\ \frac{x_N}{\sigma_N} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} \frac{y_1}{\sigma_1} \\ \frac{y_2}{\sigma_2} \\ \vdots \\ \frac{y_N}{\sigma_N} \end{bmatrix}$$

then, the likelihood can be written as:

$$\mathcal{P}(\mathcal{D} | \theta_0, \theta_1) \propto \exp \left[-\frac{1}{2} (\mathbf{b} - \theta_0 \mathbf{D}_0 - \theta_1 \mathbf{D}_1)^T (\mathbf{b} - \theta_0 \mathbf{D}_0 - \theta_1 \mathbf{D}_1) \right]$$

where we assume σ_n to be known and we assume Gaussian priors on the parameters, such that

$$\mathcal{P}(\theta_0, \theta_1) \propto \exp \left[-\frac{1}{2} \left(\frac{\theta_0^2 - 2\mu_0\theta_0}{\Sigma_0^2} \right) \right] \exp \left[-\frac{1}{2} \left(\frac{\theta_1^2 - 2\mu_1\theta_1}{\Sigma_1^2} \right) \right]$$

where we have ignored the term independent of the parameters. Our aim is to find the conditional distribution $\mathcal{P}(\theta_0 | \theta_1, \mathcal{D})$ and $\mathcal{P}(\theta_1 | \theta_0, \mathcal{D})$. The dependence of θ_0 in the log-joint posterior distribution, that is, $\theta_0 | \theta_1, \mathcal{D}$ is simply

$$-\frac{1}{2} \left[\theta_0^2 \left(\mathbf{D}_0^T \mathbf{D}_0 + \frac{1}{\Sigma_0^2} \right) + \theta_0 \left(2\theta_1 \mathbf{D}_0^T \mathbf{D}_1 - 2\mathbf{b}^T \mathbf{D}_0 - \frac{2\mu_0}{\Sigma_0^2} \right) \right]$$

This is simply a quadratic function of θ_0 . If $a = \mathbf{D}_0^T \mathbf{D}_0 + \frac{1}{\Sigma_0^2}$ and $b = 2\theta_1 \mathbf{D}_0^T \mathbf{D}_1 - 2\mathbf{b}^T \mathbf{D}_0 - \frac{2\mu_0}{\Sigma_0^2}$, then

$$\mathcal{P}(\theta_0 | \theta_1, \mathcal{D}) \propto \exp \left[-\frac{1}{2} (a\theta_0^2 + b\theta_0) \right]$$

After completing the square,

$$\mathcal{P}(\theta_0 | \theta_1, \mathcal{D}) \propto \exp \left[-\frac{a}{2} \left(\theta_0 + \frac{b}{2a} \right)^2 \right]$$

The conditional distribution for θ_0 , that is, $\mathcal{P}(\theta_0 | \mathcal{D}, \theta_1)$, is simply a Gaussian distribution with mean, μ and standard deviation, σ given by

$$\mu = -\frac{b}{2a} = \frac{-\theta_1 \Sigma_0^2 \mathbf{D}_0^T \mathbf{D}_1 + \Sigma_0^2 \mathbf{b}^T \mathbf{D}_0 + \mu_0}{\Sigma_0^2 \mathbf{D}_0^T \mathbf{D}_0 + 1}$$

$$\sigma^2 = \frac{1}{a} = \frac{\Sigma_0^2}{\Sigma_0^2 \mathbf{D}_0^T \mathbf{D}_0 + 1}$$

Similarly, it can be shown that the dependence of θ_1 in the joint posterior distribution, that is, $\mathcal{P}(\theta_1 | \theta_0, \mathcal{D})$ is also a Gaussian with

$$\mu = \frac{-\theta_0 \Sigma_1^2 \mathbf{D}_1^T \mathbf{D}_0 + \Sigma_1^2 \mathbf{b}^T \mathbf{D}_1 + \mu_1}{\Sigma_1^2 \mathbf{D}_1^T \mathbf{D}_1 + 1}$$

$$\sigma^2 = \frac{\Sigma_1^2}{\Sigma_1^2 \mathbf{D}_1^T \mathbf{D}_1 + 1}$$

Once we have these two conditional distributions, we start from a random point, say, $(\theta_0^{(i)}, \theta_1^{(i)})$ and we sample from these two distributions over N iterations to obtain the posterior distribution of the parameters as shown in Figure (A.1).

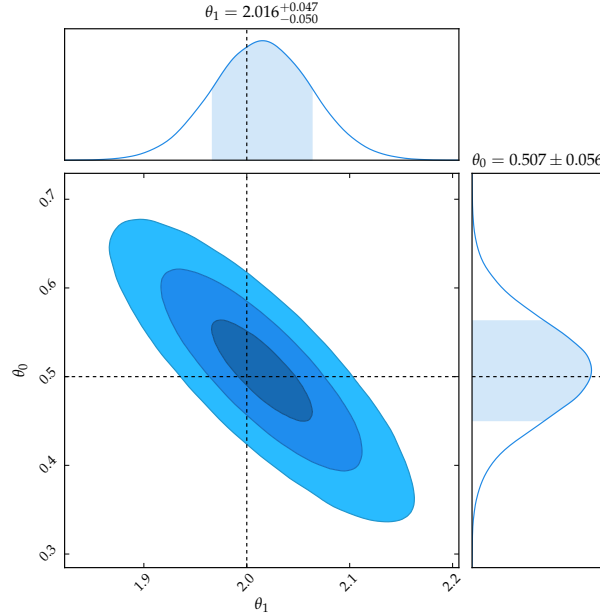


Figure A.1 – 2D posterior plot using Gibbs Sampling - The true values of the parameters are $\theta_0 = 0.5$ and $\theta_1 = 2.0$. These values are recovered to 1σ confidence interval and we also have the correlation between the parameters.

A.5 Precision for log-Bayes Factor

In this section, we show how we can calculate the precision with which we can calculate the log-Bayes Factor (see Chapter 5). Naively, we would expect it to be dependent on the number

of samples recorded for α . The log-Bayes factor can be written as

$$\log B_{21} = \log \left(\frac{2 - \epsilon}{\epsilon} \right) \quad (\text{A.5.1})$$

where $\frac{2 - \epsilon}{\epsilon}$ is imply the ratio of the posterior of α at the two endpoints, 0 and 1. The derivative of the above equation with respect to ϵ is:

$$\frac{d}{d\epsilon} \left[\log \left(\frac{2 - \epsilon}{\epsilon} \right) \right] = - \left[\frac{2}{\epsilon(2 - \epsilon)} \right]$$

Using error propagation method,

$$\sigma_{\log B_{21}}^2 = \left(\frac{d \log B_{21}}{d\epsilon} \right)^2 \sigma_{\epsilon}^2 = \left[\frac{2}{\epsilon(2 - \epsilon)} \right]^2 \sigma_{\epsilon}^2$$

If $f(\alpha)$ is the posterior distribution of α , then

$$f(\alpha_i) = (2\epsilon - 2) \alpha_i + (2 - \epsilon)$$

$$\sigma_{f_i}^2 = \left[\frac{d f(\alpha_i)}{d\epsilon} \right]^2 \sigma_{\epsilon}^2 = (2\alpha_i - 1)^2 \sigma_{\epsilon}^2$$

Assuming we have m bins, then

$$f_i = \frac{m}{N} n_i$$

where N is the total number of samples and n_i is the number of samples in bin i . The error on each bin (assuming Poisson distribution) is given by

$$\sigma_{f_i}^2 = \frac{m}{N} n_i$$

Assuming further that the error on each bin is independent from each other,

$$\sigma_{\epsilon}^2 = \frac{m}{N} \sum_{i=1}^m \frac{n_i}{(2\alpha_i - 1)^2}$$

which then leads to

$$\sigma_{\log B_{21}}^2 = \left[\frac{4m}{N\epsilon^2 (2 - \epsilon)^2} \right] \sum_{i=1}^m \frac{n_i}{(2\alpha_i - 1)^2}$$

In other words,

$$\sigma_{\log B_{21}} \propto \frac{1}{\sqrt{N}} \qquad \sigma_{\log B_{21}} \propto \frac{1}{\epsilon(2-\epsilon)} \qquad (\text{A.5.2})$$

COORDINATE TRANSFORMATIONS

In this section, we will briefly summarise the basic coordinate transformations which are the most relevant to us. In particular, we discuss the transformation between equatorial coordinate system and the horizontal coordinate system. Further, we consider the representation of baseline in different frames.

B.1 Spherical Trigonometry

The fundamental requirement for transforming one spherical coordinate system to another, is spherical trigonometry which is the backbone of spherical astronomy. This is shown in the triangle below:

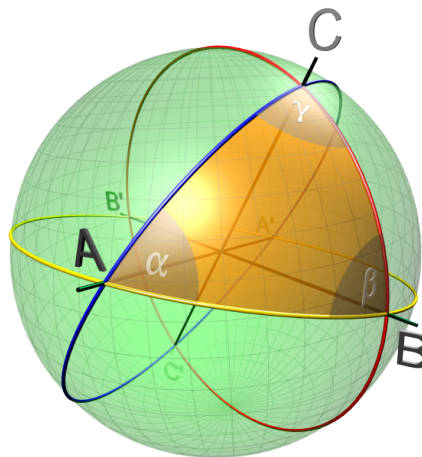


Figure B.1 – Spherical triangle representation - The triangle is defined by three angles (α, β, γ) and the three sides AB, AC and BC. These sides are measured in angular units. The sum of the three angles depends on the triangle. Figure adapted from Wikipedia*.

There are four major relations (Roy & Clarke 2003) which are important in doing transformations. These are

*https://en.wikipedia.org/wiki/Solution_of_triangles

(1) the cosine formula

$$\cos AB = \cos AC \cos BC + \sin AC \sin BC \cos \gamma \quad (\text{B.1.1})$$

(2) the sine formula

$$\frac{\sin AB}{\sin \gamma} = \frac{\sin BC}{\sin \alpha} = \frac{\sin AC}{\sin \beta} \quad (\text{B.1.2})$$

(3) the analogue of the cosine formula

$$\sin AB \cos \beta = \cos AC \sin BC - \sin AC \cos BC \cos \gamma \quad (\text{B.1.3})$$

(4) the four-parts formula

$$\cos AB \cos \alpha = \sin AB \cot AC - \sin \alpha \cot \beta \quad (\text{B.1.4})$$

Equations (B.1.1, B.1.3 and B.1.4) can have different variations depending upon the quantities we want to calculate.

B.2 Equatorial and Horizontal Coordinate Systems

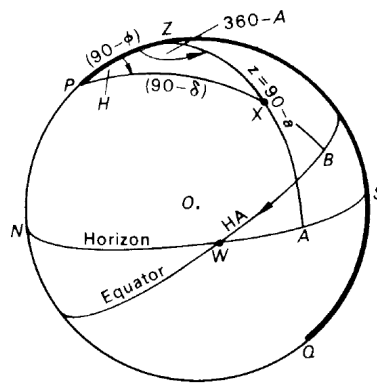


Figure B.2 – Conversion between equatorial and horizontal coordinate system - The spherical triangle PXZ is used to derive the relations between the equatorial and horizontal coordinate systems. X is the location of the celestial body whereas P and Z are the north celestial pole and the zenith respectively.

If ϕ is the latitude of the observer, the following relations can be proven using the above formulae:

$$\begin{aligned}
\cos \delta \cos H &= \cos \phi \sin a - \sin \phi \cos a \cos A \\
-\cos \delta \sin H &= \cos a \sin A \\
\sin \delta &= \sin \phi \sin a + \cos \phi \cos a \cos A
\end{aligned} \tag{B.2.1}$$

B.3 Baseline Representation in ENU Coordinates

The ENU refers to East, North and Up. In this particular scenario, the origin is chosen arbitrarily at some point on Earth but the centre of the array is chosen principally at one of the array elements. The unit vectors $\hat{e}_E$ and $\hat{e}_N$ point respectively towards the geographical East and North whereas $\hat{e}_U$ points towards the local vertical (hence Up). Since, this is a local coordinate frame, it can also be expressed using the horizontal coordinates (A, a) in which case, the azimuth (A) is measured from clockwise from North to East. If $\vec{b}$ is ENU baseline vector, it can be expressed in horizontal coordinates as follows:

$$\vec{b}_{\text{ENU}} = |\vec{b}| \begin{pmatrix} \sin A \cos a \\ \cos A \cos a \\ \sin a \end{pmatrix} \tag{B.3.1}$$

B.4 Baseline Representation in Equatorial System

In this representation, a Cartesian coordinate framework is used to describe the baseline vector. The X and Y axes are in a plane parallel to the equator while the Z axis points towards the north celestial pole. In terms of the hour angle, H and declination, δ , the XYZ axes are defined as

- (1) the X -axis is directed towards $H = 0^h$ and $\delta = 0^\circ$,
- (2) the Y -axis towards $H = -6^h$ and $\delta = 0^\circ$ and
- (3) the Z -axis towards the NCP, that is, at $\delta = 90^\circ$.

The relation between this coordinate system, the equatorial and horizontal coordinate systems ([Thompson et al. 2008](#)) is as follows:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = |\vec{b}| \begin{pmatrix} \cos \delta \cos H \\ -\cos \delta \sin H \\ \sin \delta \end{pmatrix} = |\vec{b}| \begin{pmatrix} \cos \phi \sin a - \sin \phi \cos a \cos A \\ \cos a \sin A \\ \sin \phi \sin a + \cos \phi \cos a \cos A \end{pmatrix} \tag{B.4.1}$$

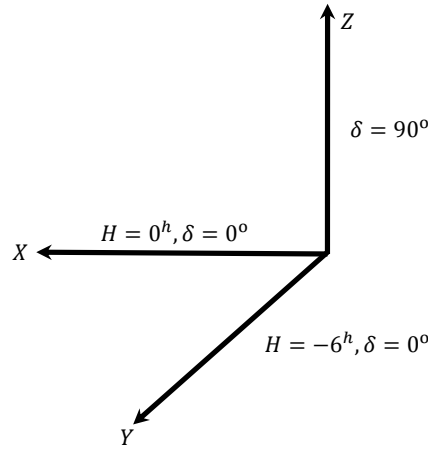


Figure B.3 – XYZ baseline representation - The axes are defined in terms of the equatorial coordinates, that is, hour angle, H and declination, δ .

B.5 Baseline Representation in (u, v, w) Space

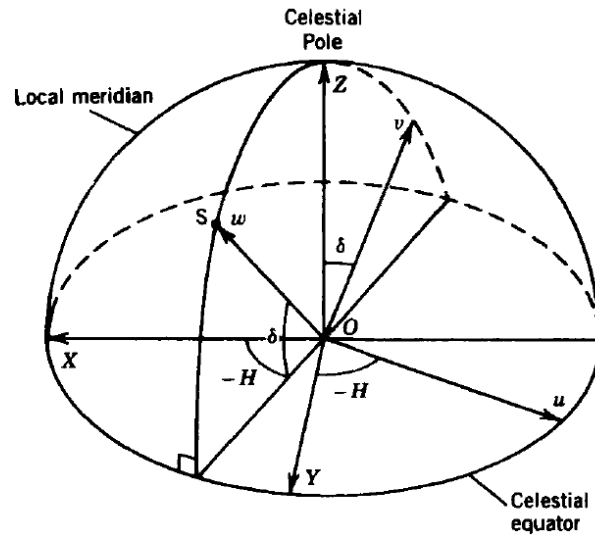


Figure B.4 – Baseline representation in the (u, v, w) space - The point S is the position of the source and the w axis is aligned in the direction $\vec{OS}$. The coordinates of this point S can be expressed in terms of both the hour angle, H and declination, δ and the (u, v, w) frame as well. Figure adapted from [Thompson et al. \(2008\)](#).

In this representation, the source is found at the coordinates (H, α) in the equatorial coordinate system. The X and Y axes are found respectively at $(0^h, 0^\circ)$ and $(-6^h, 0^\circ)$. If the (u, v, w) axes are imagined to be aligned with the (X, Y, Z) axes initially, the final (u, v, w) coordinate frame is formed via two rotations. The first is the rotation of (u, v, w) about the Z axis through $-(H + 6)^h$ and the second rotation is (v, w) through $(90^\circ - \delta)$ about the u axis. If $\vec{b}$, the baseline vector, has components X, Y and Z , then this vector can be represented in the (u, v, w) frame as

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} \sin H & \cos H & 0 \\ -\sin \delta \cos H & \sin \delta \sin H & \cos \delta \\ \cos \delta \cos H & -\cos \delta \sin H & \sin \delta \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (\text{B.5.1})$$

It is also very common to express the baseline vector in units of wavelength, which we will denote by $\vec{b}_\lambda$, hence

$$\vec{b}_\lambda = \frac{\vec{b}}{\lambda} = \begin{pmatrix} u_\lambda \\ v_\lambda \\ w_\lambda \end{pmatrix} = \begin{pmatrix} \sin H & \cos H & 0 \\ -\sin \delta \cos H & \sin \delta \sin H & \cos \delta \\ \cos \delta \cos H & -\cos \delta \sin H & \sin \delta \end{pmatrix} \begin{pmatrix} X_\lambda \\ Y_\lambda \\ Z_\lambda \end{pmatrix} \quad (\text{B.5.2})$$

where the notation r_λ means $\frac{r}{\lambda}$. The projection of $\vec{b}_\lambda$ from the (u, v, w) space to the uv -plane defines ellipses with the hour angle considered as the variable. Considering the two equations below:

$$\begin{aligned} u_\lambda &= X_\lambda \sin H + Y_\lambda \cos H \\ v_\lambda &= -X_\lambda \sin \delta \cos H + Y_\lambda \sin \delta \sin H + Z_\lambda \cos \delta \end{aligned}$$

and eliminating the terms containing H , it can be shown that:

$$u_\lambda^2 + \left(\frac{v_\lambda - Z_\lambda \cos \delta}{\sin \delta} \right)^2 = X_\lambda^2 + Y_\lambda^2 \quad (\text{B.5.3})$$

In particular, the semimajor axis of the ellipse is

$$\sqrt{X_\lambda^2 + Y_\lambda^2}$$

while the semiminor axis is equal to

$$\sin \delta \sqrt{X_\lambda^2 + Y_\lambda^2}$$

The arcs of the ellipse depend on the altitude, azimuth, latitude of the baseline, the declination of the source and the range of hour angle during the observation.

BIBLIOGRAPHY

- Akaike, H. (1974), ‘A new look at the statistical model identification’, *Automatic Control, IEEE Transactions on* **19**(6), 716–723.
- Andrieu, C., De Freitas, N., Doucet, A. & Jordan, M. I. (2003), ‘An Introduction to MCMC for Machine Learning’, *Machine learning* **50**(1-2), 5–43.
- Balanis, C. A. (2016), *Antenna theory: analysis and design*, John Wiley & Sons.
- Bayes, T. (1763), ‘An Essay towards solving a Problem in the Doctrine of Chances’, *Philosophical Transactions of the Royal Society* **53**, 370–418.
- Becla, J., Hanushevsky, A., Nikolaev, S., Abdulla, G., Szalay, A., Nieto-Santisteban, M., Thakar, A. & Gray, J. (2006), Designing a multi-petabyte database for LSST, in ‘SPIE Astronomical Telescopes+ Instrumentation’, International Society for Optics and Photonics, pp. 62700R–62700R. [[arXiv:cs/0604112](https://arxiv.org/abs/cs/0604112)].
- Boonstra, A.-J. & Van der Veen, A.-J. (2003), ‘Gain calibration methods for radio telescope arrays’, *IEEE Transactions on Signal Processing* **51**(1), 25–38.
- Booth, R., De Blok, W., Jonas, J. & Fanaroff, B. (2009), ‘MeerKAT key project science, specifications, and proposals’, *arXiv preprint arXiv:0910.2935* . [[arXiv:0910.2935](https://arxiv.org/abs/0910.2935)].
- Buchner, J., Georgakakis, A., Nandra, K., Hsu, L., Rangel, C., Brightman, M., Merloni, A., Salvato, M., Donley, J. & Kocevski, D. (2014), ‘X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue’, *Astronomy & Astrophysics* **564**, A125. [[arXiv:1402.0004](https://arxiv.org/abs/1402.0004)].
- Chib, S. & Greenberg, E. (1995), ‘Understanding the metropolis-hastings algorithm’, *The american statistician* **49**(4), 327–335.

- Christensen, N., Meyer, R., Knox, L. & Luey, B. (2001), 'Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements', *Classical and Quantum Gravity* **18**(14), 2677.
- Clark, B. (1980), 'An efficient implementation of the algorithm 'CLEAN'', *Astronomy and Astrophysics* **89**, 377.
- Cochran, W. G. (1977), *Sampling Techniques, Third Edition*, 3rd edn.
- Condon, J. J. & Ransom, S. M. (2010), *Essential radio astronomy*, National Radio Astronomy Observatory.
- Cowles, M. K. & Carlin, B. P. (1996), 'Markov chain Monte Carlo convergence diagnostics: a comparative review', *Journal of the American Statistical Association* **91**(434), 883–904.
- DeBoer, D. R. et al. (2016), 'Hydrogen Epoch of Reionization Array (HERA)'. [[arXiv:1606.07473](#)].
- Dickey, J. M. (1971), 'The weighted likelihood ratio, linear hypotheses on normal location parameters', *The Annals of Mathematical Statistics* pp. 204–223.
- Feroz, F., Hobson, M. & Bridges, M. (2009), 'MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics', *Monthly Notices of the Royal Astronomical Society* **398**(4), 1601–1614. [[arXiv:0809.3437](#)].
- Feroz, F., Hobson, M., Cameron, E. & Pettitt, A. (2013), 'Importance nested sampling and the MultiNest algorithm', *arXiv preprint* . [[arXiv:1306.2144](#)].
- Gelman, A., Hwang, J. & Vehtari, A. (2014), 'Understanding predictive information criteria for Bayesian models', *Statistics and Computing* **24**(6), 997–1016. [[arXiv:1307.5928](#)].
- Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* pp. 457–472.
- Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.
- Geyer, C. J. (1992), 'Practical markov chain monte carlo', *Statistical Science* pp. 473–483.
- Hajian, A. (2007), 'Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique', *Physical Review D* **75**(8), 083525. [[astro-ph/0608679](#)].

- Hamaker, J., Bregman, J. & Sault, R. (1996), 'Understanding radio polarimetry. I. Mathematical foundations', *Astronomy and Astrophysics Supplement Series* **117**(1), 137–147.
- Hastie, T., Tibshirani, R., Friedman, J. & Franklin, J. (2005), 'The elements of statistical learning: data mining, inference and prediction', *The Mathematical Intelligencer* **27**(2), 83–85.
- Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.
- Hee, S., Handley, W., Hobson, M. P. & Lasenby, A. N. (2015), 'Bayesian model selection without evidences: application to the dark energy equation-of-state'. [[arXiv:1506.09024](#)].
- Högbom, J. (1974), 'Aperture synthesis with a non-regular distribution of interferometer baselines', *Astronomy and Astrophysics Supplement Series* **15**, 417.
- Hollitt, C., Johnston-Hollitt, M., Dehghan, S., Frea, M. & Bulter-Yeoman, T. (2016), 'An Overview of the SKA Science Analysis Pipeline', *arXiv preprint* . [[arXiv:1601.04113](#)].
- Hopkins, A., Whiting, M., Seymour, N., Chow, K., Norris, R., Bonavera, L., Breton, R., Carbone, D., Ferrari, C., Franzen, T. et al. (2015), 'The ASKAP/EMU Source Finding Data Challenge', *Publications of the Astronomical Society of Australia* **32**, e037. [[arXiv:1509.03931](#)].
- Jansky, K. G. (1933), 'Electrical disturbances apparently of extraterrestrial origin', *Radio Engineers, Proceedings of the Institute of* **21**(10), 1387–1398.
- Jefferys, W. H. & Berger, J. O. (1992), 'Ockham's razor and bayesian analysis', *American Scientist* **80**(1), 64–72.
- Jeffreys, H. (1998), *The theory of probability*, OUP Oxford.
- Jones, E., Oliphant, T., Peterson, P. et al. (2001), 'SciPy: Open source scientific tools for Python'. [Online; accessed 2016-11-20].
URL: <http://www.scipy.org/>
- Jones, R. C. (1941), 'A New Calculus for the Treatment of Optical SystemsI. Description and Discussion of the Calculus', *JOSA* **31**(7), 488–493.
- Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

- Kazemi, S., Yatawatta, S., Zaroubi, S., Lampropoulos, P., de Bruyn, A., Koopmans, L. & Noordam, J. (2011), 'Radio interferometric calibration using the SAGE algorithm', *Monthly Notices of the Royal Astronomical Society* **414**(2), 1656–1666.
- Kellermann, K. (2005), Grote Reber (1911–2002), in 'The New Astronomy: Opening the Electromagnetic Window and Expanding Our View of Planet Earth', Springer, pp. 43–70.
- Kraus, J. D. (1966), 'Radio astronomy', *New York: McGraw-Hill*, 1966 **1**.
- Kutner, M. L. (2003), *Astronomy: A physical perspective*, Cambridge University Press.
- Lahiri, D. (1951), 'A method of sample selection providing unbiased ratio estimates', *Bulletin of the International Statistical Institute* **33**(2), 133–140.
- Lazio, J. (2009), 'The Square Kilometre Array', *arXiv preprint arXiv:0910.0632* . [[arXiv:0910.0632](#)].
- Licia, V. (2007), 'A practical guide to Basic Statistical Techniques for Data Analysis in Cosmology', *arXiv* . [[arXiv:0712.3028](#)].
- Liddle, A. R. (2004), 'How many cosmological parameters', *Monthly Notices of the Royal Astronomical Society* **351**(3), L49–L53.
- Liddle, A. R. (2007), 'Information criteria for astrophysical model selection', *Monthly Notices of the Royal Astronomical Society: Letters* **377**(1), L74–L78.
- Liddle, A. R., Mukherjee, P. & Parkinson, D. (2006), 'Cosmological model selection', *arXiv preprint astro-ph/0608184* . [[astro-ph/0608184](#)].
- Lindley, D. V. (1957), 'A statistical paradox', *Biometrika* pp. 187–192.
- Lochner, M., Natarajan, I., Zwart, J. T., Smirnov, O., Bassett, B. A., Oozeer, N. & Kunz, M. (2015), 'Bayesian inference for radio observations', *Monthly Notices of the Royal Astronomical Society* **450**(2), 1308–1319. [[arXiv:1501.05304](#)].
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.
- Mukherjee, P., Parkinson, D. & Liddle, A. R. (2006), 'A nested sampling algorithm for cosmological model selection', *The Astrophysical Journal Letters* **638**, L51–L54. [[astro-ph/0508461](#)].
- Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, The MIT Press.

- Neal, R. M. et al. (2011), 'MCMC using Hamiltonian dynamics', *Handbook of Markov Chain Monte Carlo* **2**, 113–162. [[arXiv:1206.1901](#)].
- Noordam, J. & Smirnov, O. (2010), 'The MeqTrees software system and its use for third-generation calibration of radio interferometers', *Astronomy & Astrophysics* **524**, A61. [[arXiv:1101.1745](#)].
- Perkins, S., Marais, P., Zwart, J., Natarajan, I., Tasse, C. & Smirnov, O. (2015), 'Montblanc: GPU accelerated radio interferometer measurement equations in support of Bayesian inference for radio observations', *Astronomy and Computing* **12**, 73–85.
- Popping, A. & Braun, R. (2008), 'The standing wave phenomenon in radio telescopes- Frequency modulation of the WSRT primary beam', *Astronomy & Astrophysics* **479**(3), 903–913. [[arXiv:0712.2303](#)].
- Powell, M. J. (1964), 'An efficient method for finding the minimum of a function of several variables without calculating derivatives', *The computer journal* **7**(2), 155–162.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2007), *Numerical recipes 3rd edition: The art of scientific computing*, Cambridge university press.
- Rau, U. & Cornwell, T. J. (2011), 'A multi-scale multi-frequency deconvolution algorithm for synthesis imaging in radio interferometry', *Astronomy & Astrophysics* **532**, A71.
- Robert, C. P. (1993), 'A note on jeffreys-lindley paradox', *Statistica Sinica* pp. 601–608.
- Roy, A. E. & Clarke, D. (2003), *Astronomy Principles and Practice - 4th Edition*, CRC Press.
- Ryle, M. & Hewish, A. (1960), 'The synthesis of large radio telescopes', *Monthly Notices of the Royal Astronomical Society* **120**(3), 220–230.
- Ryle, M., Smith, F. G. & Elsmore, B. (1950), 'A preliminary survey of the radio stars in the Northern Hemisphere', *Monthly Notices of the Royal Astronomical Society* **110**(6), 508–523.
- Salvini, S. & Wijnholds, S. J. (2014), 'Fast gain calibration in radio astronomy using alternating direction implicit methods: Analysis and applications', *Astronomy & Astrophysics* **571**, A97.
- Schwab, F. (1984), 'Relaxing the isoplanatism assumption in self-calibration; applications to low-frequency radio interferometry', *The Astronomical Journal* **89**, 1076–1081.
- Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

- Skilling, J. (2004), ‘Nested sampling’, *Bayesian inference and maximum entropy methods in science and engineering* **735**, 395–405.
- Skilling, J. et al. (2006), ‘Nested sampling for general Bayesian computation’, *Bayesian Analysis* **1**(4), 833–859.
- Smirnov, O. M. (2011), ‘Revisiting the radio interferometer measurement equation. I. A full-sky Jones formalism’, *Astronomy & Astrophysics* **527**, A106. [[arXiv:1101.1764](#)].
- Smirnov, O. & Tasse, C. (2015), ‘Radio interferometric gain calibration as a complex optimization problem’, *Monthly Notices of the Royal Astronomical Society* **449**(3), 2668–2684.
- Stanimirovic, S. (2002), ‘Short-Spacings Correction from the Single-Dish Perspective’, *arXiv preprint*. [[arXiv:astro-ph/0205329](#)].
- Straka, O., Duník, J., Šimandl, M. & Havlík, J. (2013), Aspects and comparison of matrix decompositions in unscented kalman filter, in ‘American Control Conference (ACC), 2013’, IEEE, pp. 3075–3080.
- Taylor, G. B., Carilli, C. L. & Perley, R. A., eds (1999), *Synthesis Imaging in Radio Astronomy II*, Vol. 180 of *Astronomical Society of the Pacific Conference Series*, Astronomical Society of the Pacific, San Francisco.
- Thompson, A. R., Moran, J. M. & Swenson Jr, G. W. (2008), *Interferometry and synthesis in radio astronomy*, John Wiley & Sons.
- Trotta, R. (2007), ‘Applications of Bayesian model selection to cosmological parameters’, *Monthly Notices of the Royal Astronomical Society* **378**, 72–82. [[astro-ph/0504022](#)].
- Trotta, R. (2008), ‘Bayes in the sky: Bayesian inference and model selection in cosmology’, *Contemp. Phys.* **49**, 71–104. [[arXiv:0803.4089](#)].
- Tzikas, D. G., Likas, A. C. & Galatsanos, N. P. (2008), ‘The variational approximation for bayesian inference’, *IEEE Signal Processing Magazine* **25**(6), 131–146.
- Verdinelli, I. & Wasserman, L. (1995), ‘Computing Bayes factors using a generalization of the Savage-Dickey density ratio’, *Journal of the american statistical association* **90**(430), 614–618.
- Wilner, D. (2010), ‘Imaging and Deconvolution’, 12th Synthesis Imaging Workshop.
- Wit, E., Heuvel, E. v. d. & Romeijn, J.-W. (2012), ‘“all models are wrong...”: an introduction to model uncertainty’, *Statistica Neerlandica* **66**(3), 217–236.

Zernike, F. (1938), 'The concept of degree of coherence and its application to optical problems',
Physica 5(8), 785–795.